Vorhersage, Analyse und Bedeutung charakteristischer Transkriptionsfaktor-Bindestellen und deren potentielle Masterkontrollfunktion in der transkriptionellen Genregulation

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm Environmental Informatics (PEI)

der Georg-August University School of Science (GAUSS)

vorgelegt von
Martin Haubrock (geb. Bennemann)
aus Unna (Westfalen)

Göttingen, 2021

Betreuungsauschuss

Prof. Dr. Edgar Wingender,
(Institut für Bioinformatik, Universitätsmedizin Göttingen)

Prof. Dr. Stephan Waack,
(Institut für Informatik, Georg-August-Universität Göttingen)

Mitglieder der Prüfungskommission

Referent:	Prof. Dr. Edgar Wingender, Institut für Bioinformatik, Universitätsmedizin Göttingen
Korreferent:	Prof. Dr. Stephan Waack, Theoretische Informatik und Algorithmische Methoden, Institut für Informatik, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommission

Prof. Dr. Tim Beißbarth
(Institut für Medizinische Bioinformatik, Universitätsmedizin Göttingen)

Prof. Dr. Carsten Damm, Theoretische Informatik und Algorithmische Methoden
(Institut für Informatik, Georg-August-Universität Göttingen)

Prof. Dr. Burkhard Morgenstern, Abteilung für Bioinformatik
(Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen)

Prof. Dr. Ulrich Sax
(Institut für Medizinische Informatik, Universitätsmedizin Göttingen)

Tag der mündlichen Prüfung: 23. September 2021

Zusammenfassung

In Eukaryoten wird die transkriptionelle Genregulation durch eine Menge von Transkriptionsfaktoren (TFs) und Kofaktoren gesteuert. Die meisten TFs binden an bestimmte DNA-Sequenzabschnitte. Diese Bindestellen werden als Transkriptionsfaktor-Bindestellen (TFBSs) bezeichnet. Eine transkriptionsregulatorische Region besteht aus verschiedenen TFBSs, die durch eine definierte Menge an TFs gebunden werden können. Durch Proteininteraktionen zwischen den gebundenen TFs und die Anlagerung weiterer TFs und Kofaktoren wird auf diese Weise ein regulatorisches Modul aufgebaut, welches die Effizienz der Transkription maßgeblich beeinflusst. Verschiedene dieser Module werden am eigentlichen Ort der Transkription zusammengeführt. Durch einen variablen Aufbau dieser Module und deren Kombination wird eine situationsspezifische Aktivität der Transkription eines Gens vermittelt.

In dieser Arbeit wird die prägende Eigenschaft einer TFBS als definitorischer Kern (*Seed*) und deren interagierende TFs als Masterregulatoren eines regulatorischen Moduls untersucht. Auf Grundlage bekannter Bindestellenbeschreibungen (Positionsgewichtungsmatrizen, PWMs) wurden verschiedene Analysestrategien entwickelt, um potenzielle Seeds in unterschiedlichen genomischen Daten zu bestimmen. Das zunächst implementierte Verfahren analysiert Kern-Bindestellen für bekannte Sequenzvariationen in genomischen Sequenzen. Eine zweite Methode wird eingesetzt, um anhand von Sequenzalignments und vergleichender Annotation Seeds auf Grundlage phylogenetischer Ähnlichkeit zu ermitteln. Die dritte Methodik bestimmt qualitative Seed-Bindestellen in genomweit erzeugten regulatorischen Sequenzdaten.

Die Bedeutung der Kern-Bindestellen wurde in dieser Arbeit in vier verschiedenen Teilprojekten untersucht. Die Ergebnisse dieser Untersuchungen liefern neue Erkenntnisse über die Eigenschaften der modulprägenden TFBSs. In der ersten Untersuchung wird auf Grundlage von pathogenen menschlichen Sequenzvariationen gezeigt, dass eine krankheitsbezogene Sequenzvariation in einer TFBS mit einem Wechsel der dort bindenden TFs verbunden ist. Diese bedeutende Eigenschaft konnte durch zwei verschiedene medizinische Studien belegt werden. Die Vorhersage evolutionär konservierter Kern-Bindestellen erlaubt die Erstellung eines globalen regulatorischen Transkriptionsnetzwerks (RTN) und wurde im zweiten Teilprojekt untersucht. Ein RTN besteht aus Seed-Bindestellen und deren interagierende TFs. Der Vergleich verschiedener menschlicher gewebespezifischer RTNs zeigt auf Basis der Knotengradverteilungen ein gemeinsames Konstruktionsprinzip, obwohl die einzelnen Seed-Bindestellen nur sehr gewebespezifisch verwendet werden. Das dritte Teilprojekt dokumentiert die Bedeutung der Seeds und ihrer aktiven Masterregulatoren (TFs) in genomweiten regulatorischen Regionen. Es zeigt sich, dass die verantwortlichen Kern-Bindestellen im Vergleich verschiedener Zelltypen eine vergleichbare Qualität aufweisen. Dies lässt vermuten, dass Modul-prägende TFBSs und ihre interagierenden Masterregulatoren allgemeingültige Bindestellenqualtitäten besitzen. Im letzten Teilprojekt wird die besondere Kontrollfunktion der sequenzdefinierten Kernbindestellen in verschiedenen regulatorischen Regionen analysiert. Die Untersuchung unterscheidet dabei die regulatorische Regionen in der Nähe des Transkriptionsstarts (Promotor) von weiter entfernt liegenden Bereichen (Enhancer). Es kann gezeigt werden, dass definierte Enhancer-Promotor-Interaktionen durch diese Kern-Bindestellen und die interagierenden Masterregulatoren direkt oder indirekt (durch Kofaktoren gebundene Masterregulatoren) vermittelt werden.

Abstract

In eukaryotes, transcriptional gene regulation is controlled by a set of transcription factors (TFs) and co-factors. Most TFs bind to specific DNA sequences. These binding sites are called transcription factor binding sites (TFBSs). A transcriptional regulatory region consists of different TFBSs that can be bound by a defined set of TFs. By protein interactions between the bound TFs and the accumulation of further TFs and co-factors, a regulatory module is formed, which significantly influences the efficiency of transcription. Several of these modules are assembled at the actual transcription start site (TSS). A situation-specific activity of the transcription of a gene is mediated by a variable structure of these modules and their combination at the TSS.

In this work, the characterizing property of a TFBS as a definitional core (seed) and its interacting TFs as master regulators of a potential regulatory module are investigated. Based on known binding site motifs (positional weighting matrices, PWMs) different analysis strategies were developed to determine seeds in various kinds of genomic sequence data. The first implemented method analyzes core binding sites for known sequence variations in genomic sequences. The second method determines seeds based on phylogenetic similarity using sequence alignments and comparative annotation. The third methodology determines defined seed binding sites in genome-wide regulatory sequence data.

The importance of the core binding sites was investigated in this work in four sub-projects. Through these analyses, various properties of the module-forming TFBSs can be determined. In the first study, based on pathogenic human sequence variations, it is shown that a disease-related change in a TFBS is associated with a change in the TFs being bound at that site. This property has been proven in two different clinical studies. Seed binding sites can be used to create regulatory transcription networks (RTNs) and was investigated in the second sub-project. An RTN consists of seed binding sites and their interacting TFs. A comparison of different human tissue-specific RTNs reveals a common design principle based on the node degree distributions, although the individual seed binding sites are used in a very tissue-specific manner. The third sub-project demonstrates the importance of seeds and their active master regulators (TFs). For experimental regions actively bound by a TF, the responsible seed binding sites are shown to be of comparable quality in comparison to different cell types. This suggests that module-forming TFBSs and their interacting master regulators exhibit universal binding site qualities. In the last sub-project, the special control function of the sequence-defined core binding sites in different regulatory regions is investigated. The study distinguishes between regulatory regions close to the TSS (promoters) and more distant regions (enhancers). Overall, a specific control function of sequence-defined seeds in both promoters and enhancers can be demonstrated. Furthermore, it can be shown that defined enhancer-promoter interactions are mediated by these core binding sites and their interacting master regulators directly or indirectly, based on co-factors.

Danksagung

Zuallererst möchte ich Professor Dr. Edgar Wingender danken. Die gemeinsame berufliche Zeit von mehr als 18 Jahren, davon 16 Jahre in Göttingen, haben mich sehr geprägt. Die immer sehr vertrauensvolle Zusammenarbeit habe ich sehr geschätzt und möchte mich dafür an dieser Stelle ausdrücklich bedanken. Die bioinformatische Modellierung und Analyse genregulatorischer Netzwerke und deren Anwendung in der Medizin war der Forschungsschwerpunkt des Instituts für Bioinformatik, welches von Prof. Wingender in Göttingen an der Universitätsmedizin gegründet und geleitet wurde. Eine wichtige Beschreibungsebene in diesen Netzwerken ist die Modellierung der Transkriptionsregulation. Dieses Thema hat mich immer sehr interessiert und ich konnte und kann mich glücklich schätzen mit Prof. Wingender einen Experten auf diesem Gebiet als direkten Ansprechpartner und Lehrer zur Verfügung zu haben. Durch die gemeinsamen Jahre konnte ich so von seinem großen Erfahrungsschatz profitieren. Die wissenschaftlichen Projekte, welche in dieser Zeit am Institut bearbeitet wurden, haben schließlich dazu geführt, dass ich mich mit dem „Projekt" einer eigenen Dissertation befasst habe. Für die Unterstützung dieses Vorhabens und die wissenschaftliche und methodische Begleitung während der gesamten Bearbeitungsphase meiner Doktorarbeit bin ich sehr dankbar.

Professor Dr. Stephan Waack danke ich ganz herzlich für seine Bereitschaft als Betreuer und Korreferent meiner Doktorarbeit zur Verfügung zu stehen. Durch die gemeinsame Betreuung verschiedener Projekt- und Abschlussarbeiten habe ich immer wieder neue Methoden, die in seiner Arbeitsgruppe entwickelt wurden, kennenlernen dürfen. Die Einführung in die Grundbegriffe und mathematischen Konzepte der Informationstheorie, welche ich in seiner Vorlesung kennengelernt habe, sind für mich nach wie vor bei der Bearbeitung verschiedener Projekte eine wertvolle Hilfe. Für die Unterstützung während der gesamten Bearbeitungsphase meiner Dissertation möchte ich mich auf diesem Wege ganz herzlich bedanken.

Mein Dank gilt auch den weiteren Mitgliedern der Prüfungskommission: Prof. Dr. Tim Beißbarth, Prof. Dr. Carsten Damm, Prof. Dr. Burkhard Morgenstern und Prof. Dr. Ulrich Sax. Vielen Dank für Ihre Zeit und Bereitschaft zur Teilnahme an der Kommission.

Außerdem möchte ich allen ehemaligen Mitarbeitern des Instituts für Bioinformatik der Universitätsmedizin danken. Die verschiedenen Aufgaben eines Instituts sind ohne eine kollegiale Zusammenarbeit aller beteiligten Personen nicht möglich. Dafür ist eine gute Arbeitsatmosphäre unerlässlich. Diese habe ich im Institut erfahren und sehr geschätzt. Mein Dank gilt aber auch den neuen Mitarbeitern und dem Leiter des Instituts für Medizinische Bioinformatik Prof. Dr. Tim Beißbarth für die herzliche Aufnahme in seinem neuen Institut.

Besonders möchte ich an dieser Stelle meiner Familie und Freunden danken. Ohne dieses Fundament wäre diese Arbeit nicht entstanden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

bp *Basenpaare*

HGP *Human Genome Project*

ENCODE *Encyclopedia of DNA Elements*

TF *Transkriptionsfaktor*

TFBS *Transkriptionsfaktorbindestelle*

mRNA *messenger RNA*

DNA *Desoxyribonukleinsäure*

TAD *Topologisch assoziierte Domänen*

IUPAC *International Union of Pure and Applied Chemistry*

K *Lysin*

HMM *Hidden-Markov-Modell*

miRNA *mikroRNA*

lncRNA *long non-coding RNA*

rRNA *ribosomale RNA*

tRNA *transfer RNA*

RNA *Ribonukleinsäure*

mRNA *messenger RNA*

PIC *Präinitiationskomplex*

A *Adenin*

C *Cytosin*

G *Guanin*

T *Thymin*

U *Uracil*

CP *core promoter*

TSS *Transcription Start Site*

CPE *Core Promoter Element*

BRE *B Becognition Element*

TATA *TATA box element*

Inr *Initiator element*

TCT *polypyrimidine initiator*

DCE *Downstream Core Element*

DPE *Downstream Promoter Element*

MTE *Motiv Ten Element*

TSS *Transcription Start Site*

CAGE *Cap Analysis of Gene Expression*

GRO-seq *Global Run-On sequencing*

PPI *Protein-Protein-Interaktion*

DBD *DNA-bindende Domäne*

AP-1 *Activator protein 1*

PWM *Positional Weight Matrix*

PFM *Postion Frequency Matrix*

ML *Maximum Likelihood*

EM *Expectation Maximization*

MEME *Multiple EM for Motif Elicitation*

CSS *Core Similarity Score*

MSS *Matrix Similarity Score*

EST *Expressed Sequence Tag*

SNP *Single Nucleotide Polymorphism*

GWAS *Genome-wide Association Study*

EMSA *Electrophretic Mobility Shift Array*

RTN *Reference Transcriptional Network*

PPV *Positive Predictive Value*

SP *Specificity*

TPR *True Positive Rate*

TTN *Tissue-specific Transcription Network*

eTTN *extended Tissue-specific Transcription Network*

eRTN *extended Regulatory Transcription Network*

RTN *Regulatory Transcription Network*

MCC *Matthwes Correlation Coefficient*

AUROC *Area Under Reciever Operator Characterstic*

TP *True Positives*

FP *False Positives*

FN *Fales Negatives*

TN *True Negatives*

ROC *Reciever Operator Characteristic*

DHS *DNase-I Hypersensitive Sites*

TPR *True Positive Rate*

FPR *False Positive Rate*

OR *Odds Ratio*

LTR *Long Terminal Repeat*

EST *Expressed Sequence Tags*

1 Einleitung

Im Jahr 2000 wurde die erste Version der Genomsequenz des Menschen veröffentlicht. Ein Ereignis, das seinerzeit sowohl in politischen als auch wissenschaftlichen Kreisen als gefeierter Meilenstein galt. Auf einer Pressekonferenz, welche dazu im Weißen Haus am 26. Juni 2000 vom damaligen Präsidenten der Vereinigten Staaten, Bill Clinton, durchgeführt wurde, machten die beiden wissenschaftlichen Leiter des Genomprojektes, Craig Venter (*Celera*) und Francis Collins (*Human Genome Project* (HGP)), auf die historische Bedeutung dieses Ereignisses aufmerksam. Francis Collins sprach auf der Veranstaltung von der menschlichen Genomsequenz als einem Buch des Lebens ("*Human Book of Life*"), während Craig Venter in seinen Ausführungen auf die große Sequenzähnlichkeit der Proteinkodierenden menschlichen Genomsequenzen mit anderen bereits bekannten tierischen Genomen aufmerksam machte. Die Ausführungen beider Wissenschaftler wiesen aber insbesondere auf die Bedeutung dieses Meilensteins für die biomedizinische Forschung hin. Beide Forscher betonten gleichermaßen, dass die Entschlüsselung der menschlichen Genomsequenz einen neuen Startpunkt darstelle, welcher ein besseres Verständnis von Krankheiten und daraus abgeleitet auch neue Therapiemöglichkeiten ermöglichen werde. Das Ergebnis der Entschlüsselung des menschlichen Genoms liegt als ein Text vor, aufgebaut aus ungefähr 3,2 Milliarden Buchstaben (*Basenpaare* (bp)), welche auf 24 verschiedene Chromosomen aufgeteilt sind (Lander et al. 2001; Venter et al. 2001).

Eine wesentliche Aufgabe nach der Entschlüsselung des Genoms besteht darin, die funktionstragenden Bereiche in diesem Text zu identifizieren. Dieser Schritt wird auch als Annotation bezeichnet. Als funktionelle Bereiche werden in diesem Zusammenhang zuallererst Gene und deren transkriptionsaktive Bereiche verstanden, welche in verschiedenen Situationen in einer lebenden Zelle gezielt abgelesen werden und so direkt oder indirekt (nach Auf- und eventueller Weiterverarbeitung) definierte Aufgaben in einer Zelle übernehmen

können. Das Auffinden dieser Bereiche ist zu Beginn des Annotationsprozesses vor allem durch bioinformatische Methoden durchgeführt worden. Durch Hinzuziehen experimenteller Daten wird dieser Prozess massiv unterstützt. Eine besondere Rolle nimmt dabei das sogenannte *Encyclopedia of DNA Elements* (ENCODE) Projekt ein (Birney et al. 2007; Dunham et al. 2012). Die Entwicklung und Verbesserung experimenteller Methoden zur Bestimmung dieser funktionstragenden Bereiche und deren einheitliche Beschreibung in einem frei verfügbaren Datenspeicher ist dabei das Ziel dieses Projekts. Am Anfang stand die standardisierte Generierung von Datensätzen im Vordergrund, welche die Transkriptionsaktivität von Genen und deren gewebe- oder zeitpunktspezifische Interpretation analysiert haben. Das Projekt startete 2003 und erste Daten wurden 2007 veröffentlicht.

Im Jahre 2012 wurde durch ENCODE erstmals auch eine neuartige Funktionsinformation veröffentlicht, welche zuvor noch nicht in diesem Umfang für das menschliche Genom zur Verfügung stand: die sogenannten regulatorischen Bereiche. Diese Regionen steuern die Transkription direkt oder indirekt und sind über das gesamte Genom verteilt. Frühere Arbeiten schätzen den Anteil an regulatorischen und transkriptionsaktiven Regionen im menschlichen Genom auf maximal zehn Prozent (Vallania et al. 2009). Die Daten des ENCODE Projekts lassen vermuten, dass die überwiegende Mehrheit des menschlichen Genoms (mehr als 80 Prozent) nachweislich transkribiert oder potentiell regulatorisch aktiv ist (Dunham et al. 2012). Diese Prozentangabe macht deutlich, dass der biologische Funktionsbegriff in diesen beiden Untersuchungen sehr unterschiedlich verwendet worden ist. Graur et al. (2013) bewerten dies in ihrer Veröffentlichung sehr deutlich. Sie kritisieren unter anderem die ihrer Meinung nach überschätzte durchschnittliche Längenangabe für eine regulatorische Region (je nach Labor zwischen 400 und 800 bp) und argumentieren, dass nicht die gesamte Region regulatorische Aktivität besitze, sondern nur darin enthaltene kleine, abgrenzbare und definierte Sequenzbereiche. Diese Bereiche werden durch eine bestimmte Klasse von Proteinen, die sogenannten *Transkriptionsfaktoren* (TFs), erkannt und gebunden. TFs interagieren mit dem basalen Transkriptionsapparat und ermöglichen so die Erkennung der zu transkribierenden Regionen und steuern den Prozess der Transkription maßgeblich. Die meisten TFs erkennen ihre jeweilige Bindestelle, welche auch als *Transkriptionsfaktorbindestelle* (TFBS) bezeichnet wird, in einer sequenzspezifischen Art und Weise. Eine typische TFBS besteht dabei in einer zusammenhängenden 5-15 bp umfassenden DNA-Sequenz. Die Zugänglichkeit einer TFBS, die Abfolge der TFBSs innerhalb einer regulatorischen Region und die vorhandenen interagierenden TFs sind dabei

die hauptverantwortlichen Komponenten der eukaryotischen transkriptionellen Genregulation.

Diese Arbeit untersucht die Bedeutung der TFs auf der Basis ihrer sequenzdefinierten TFBSs in menschlichen genregulatorischen Regionen. Das menschliche Genom kodiert mehr als 1600 unterschiedliche TFs (Lambert et al. 2018). Verschiedene Teilmengen dieser Proteine werden verwendet, um die Transkription der notwendigen Gene und deren Transkriptionseinheiten situations- und zelltypspezifisch zu steuern. Im Vergleich zweier unterschiedlich transkriptionsaktiver Zelltypen kann häufig eine große Schnittmenge der aktiven TFs beobachtet werden. Auf Ebene der aktiven Transkriptionseinheiten kann dabei aber durchaus eine auffalend unterschiedliche Genaktivität festgestellt werden. Wie kann diese Unterschiedlichkeit mit Hilfe einer mehr oder weniger gleiche Menge an TFs erzielt werden? Eine erste genregulatorische Schicht ist dabei von entscheidender Bedeutung. In Eukaryoten ist die Verfügbarkeit der genomischen Information stark reguliert (Cairns 2009). Alle transkriptionsaktiven Zellen besitzen das gleiche genomische Material. Aber die Zugänglichkeit und damit die Lesbarkeit der regulatorischen Bereiche kann zwischen verschiedenen Zellen sehr verschieden sein. Die dort vorhandenen TFBSs definieren durch ihre Abfolge, Distanz und Orientierung die aktive Teilmenge an interagierenden TFs und vermitteln so die zelltypspezifische Transkriptionsaktivität. Einige TFs scheinen diesen Prozess im besonderem Maße zu unterstützen. In der Literatur werden sie als sogenannte Master-TFs oder auch Masterregulatoren beschrieben (Pei 2009). Diese Teilmengen an TFs sind in der Lage, aus einer ausdifferenzierten Zelle eine sogenannte Stammzelle (undifferenzierte Zelle) zu prägen, aus der sich nun wiederum jede andere Körperzelle entwickeln kann. Eine Stammzelle zeichnet sich durch eine weitreichende Aktivität der verschiedenen regulatorischen Regionen aus, welche durch das gesamte verfügbare Repertoire an TFs gebunden werden kann. Durch die vielfältige Öffnung der regulatorischen Regionen können vermutlich alle Entwicklungsmöglichkeiten einer Zelle aktiviert werden. Zusätzlich zum Konzept der Masterregulatoren ist das Modell der Pionier-TFs bekannt geworden. Pionierfaktoren sind TFs, welche an inaktive (kondensierte) DNA binden können und diese allein oder in Kombination mit anderen TFs für die Interaktion mit weiteren vorhandenen TFs vorbereiten können (Zaret und Carroll 2011). Pionierfaktoren öffnen also zuvor nicht zugängliche regulatorische Regionen. Beide Beobachtungen zeigen, dass die Transkriptionsregulation durch ein komplexes und hierarchisches Interaktionsnetzwerk verschiedener TFs gesteuert wird.

In der vorliegenden Arbeit wird der Bedeutung einzelner definierter TFBSs und deren sequenzspezifisch interagierende Menge an TFs für die transkriptionelle Genregulation untersucht. Im ersten Teil der Arbeit wird ein Bewertungsschema vorgestellt und angewendet, welches auf der Grundlage bekannter Bindestellenbeschreibungen diverser TFs für krankheitsbezogende Sequenzvariationen in regulatorischen Regionen potentielle TFBSs bestimmt und bewertet.
Mit Hilfe dieses Verfahrens konnten TFBSs für zwei regulatorisch aktive pathologische Sequenzvariationen identifiziert werden. In zwei unabhängigen medizinischen Studien wurden die jeweiligen TFBS und ihre interagierende TFs experimentell verifiziert. Die Konstruktion eines allgemein gültigen transkriptionsregulatorischen Netzwerks auf der Grundlage evolutionär konservierter TFBSs wird im zweiten Teil der Arbeit vorgestellt. Auf Basis von Genxpressionsdaten und experimentell bestimmten regulatorischen Regionen (ChIP-seq, DNase-seq) werden die Bedeutung und verschiedene Anwendungsfelder dieses Netzwerks gezeigt. Eine Untersuchung der definitorischen Bedeutung von TFBS bei der Kommunikation interagierender regulatorischer Regionen wird in den letzten beiden Teilen der Arbeit vorgestellt. Als Ergebnis dieser Untersuchungen wird gezeigt, dass bestimmte TFs über ihre jeweiligen TFBSs in definierten räumlichen Abhängigkeiten auftreten. Es können sowohl lokale Abstandsbeziehungen einzelner TFBSs, als auch wechselseitige distale Abhängigkeiten definierter Bindestellen und deren interagierende TFs gefunden werden. Auf Grundlage der umfangreichen Datensammlung für genregulatorische Regionen des ENCODE Projekts werden diese Abhängigkeiten für verschiedene menschliche Zelllinien untersucht.

Die Arbeit ist in sechs Kapitel gegliedert. Im folgenden Kapitel werden als Hintergrund die biologischen und bioinformatischen Grundlagen der transkriptionellen Genregulation eingeführt. Der Material- und Methodenteil listet die in dieser Arbeit verwendeten Daten, Datenbanken und Software auf. Der Ergebnisteil beschreibt die verschiedenen Untersuchungen und Beobachtungen, aufgeteilt gemäß der vier Teilprojekte. Im Diskusssionsteil werden die Ergebnisse der Teilprojekte miteinander in Beziehung gesetzt und gemeinsam diskutiert. Im Ausblick werden auf Grundlage der Ergebnisse dieser Arbeit zukünftige Forschungsmöglichkeiten erörtert. Den Abschluss der Arbeit bildet die Auflistung der verwendeten Literaturreferenzen.

2 Hintergrund

Einer der zentralen Sätze der Molekularbiologie besagt, dass Proteine durch definierte RNA-Sequenzen (*messenger RNA* (mRNA)) entstehen, welche zuvor aus der Abschrift von genomischer DNA entstanden sind. Der Entstehungsprozess von RNA aus DNA, welcher auch als Transkription bezeichnet wird, ist ein stark regulierter Prozess. Die Effizienz der Transkription wird dabei wesentlich durch ein Netzwerk von regulatorischen Proteinen, den TFs, gesteuert. Die meisten dieser TFs binden an definierte Sequenzen der DNA und entfalten so ihren regulatorischen Einfluss. Im folgenden Kapitel werden die grundlegenden Begriffe und Konzepte der eukaryotischen Genregulation eingeführt. Anschließend erfolgt eine Vorstellung derjenigen bioinformatischen Konzepte, welche bei der rechnergestützten Vorhersage einer TFBS Verwendung finden.

2.1 Organisation des Genoms

Die beiden Forscher James Watson und Francis Crick haben am 25. April 1953 in der Zeitschrift Nature einen Strukturvorschlag für den Aufbau der *Desoxyribonukleinsäure* (DNA) veröffentlicht (Watson und Crick 1953). Ihr Modell beschreibt die DNA als Doppelhelix, welche aus zwei gegenläufig zueinander verlaufenden DNA-Strängen besteht. Die einzelnen Ketten werden durch ein kovalent gebundenes Zucker-Phosphat-Rückgrat gebildet. Das C_1-Atom des Zuckers (Desoxyribose, siehe auch Abbildung 2.1) ist mit einer der vier möglichen stickstoffhaltigen organischen Basen kovalent verbunden. Die Basen werden als *Adenin* (A), *Cytosin* (C), *Guanin* (G), oder *Thymin* (T) bezeichnet. Die beiden Stränge werden durch Wasserstoffbrückenbindungen zusammengehalten. Dabei stehen sich immer die beiden Nukleotide Adenin und Thymin sowie Cytosin und Guanin gegenüber. Die Abbil-

dung 2.1 verdeutlicht die Struktur der DNA im Detail. Die Paare A-T und G-C werden als komplementäre Nukleotide bezeichnet (auch bp genannt). Die Sequenz des einen DNA-Strangs legt also die Abfolge des anderen DNA-Strangs fest. Diese Tatsache veranlasste schon Watson und Crick (1953) in ihrer Veröffentlichung zu folgendem Satz: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." (Watson und Crick 1953, Seite 737). Diese Aussage verdeutlicht, dass der von den beiden Wissenschaftlern gemachte Strukturvorschlag bereits eine elegante Möglichkeit bereitstellt, wie ein biologisches System die DNA als Grundlage für lebensnotwendige Prozesse nutzen kann. Das menschliche Ge-

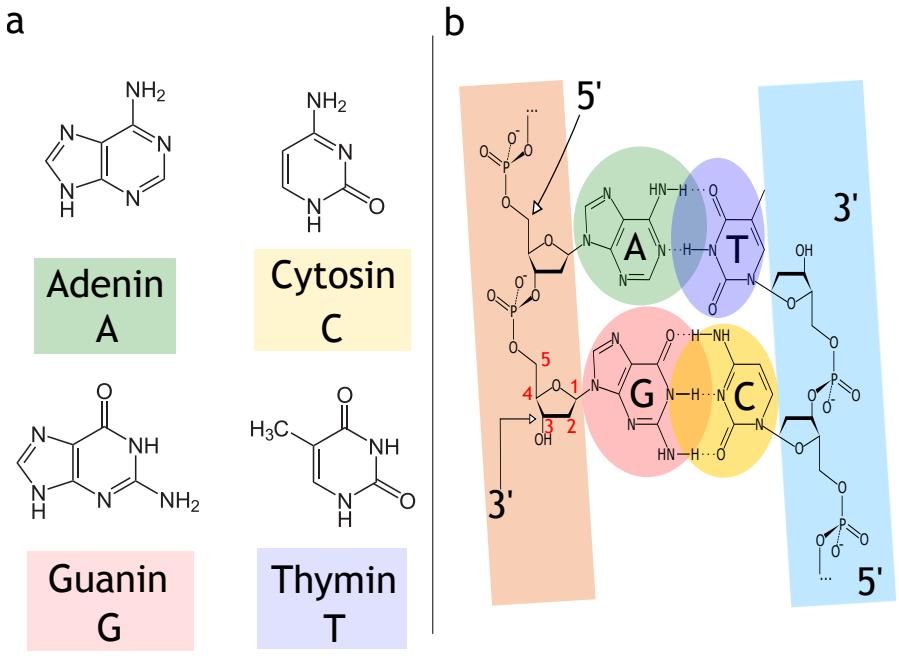


Abbildung 2.1: Aufbau der DNA. Im linken Teil der Abbildung (a) sind die vier Basen der DNA und ihre Strukturformel dargestellt. Die Purinbasen (A,G), sowie die Pyrimidinbasen (C,T) sind hier gezeigt. Der rechte Teil der Abbildung verdeutlicht beispielhaft die Strukturformel eines DNA-Ausschnitts. Das Basenpaar A-T ist durch zwei und das Basenpaar G-C durch drei Wasserstoffbrückenbindungen gekennzeichnet. Die zwei gegenläufigen DNA-Zuckerphosphatstränge (5'-3'/3'-5') bilden das Grundgerüst der DNA.

nom in jeder einzelnen Zelle ergibt, aneinandergereiht als ein zusammenhängender DNA-Strang, eine Länge von ungefähr zwei Metern. Diese Zahl setzt sich zusammen aus der An-

zahl der Nukleotide ($3,2$ Milliarden $= 3,2 \times 10^9$) multipliziert mit dem Abstand, welchen zwei Nukleotide in der DNA zueinander besitzen ($0,34$ Nanometer $= 3,4 \times 10^{-10}$ Meter) (Alberts et al. 2002). Das Genom ist in den meisten menschlichen Zellen vorhanden und wird in einer bestimmten Zellorganelle (abgrenzbarer Bereich einer Zelle), dem Zellkern (engl. *nucleus*), konzentriert. Dieses Organell besitzt modellhaft angenommen eine kugelförmige Gestalt mit einem ungefähren Durchmesser von sechs Mikrometern (6×10^{-6} Meter) (Alberts et al. 2002). Um das gesamte genetische Material im Zellkern aufnehmen zu können, wird es in einer besonderen Art und Weise komprimiert: Ein Proteinkomplex, bestehend aus vier verschiedenen Proteinen, die auch Histonproteine genannt werden, ist dafür hauptverantwortlich. Die Struktur, welche dieser Komplex ausbildet, wird als Nukleosom bezeichnet (engl. *nucleosome*). Er besteht aus den Einzelproteinen H2A, H2B, H3 und H4, die mit jeweils zwei Einheiten die Kernstruktur eines Nukleosoms bilden (siehe auch Abbildung 2.2).

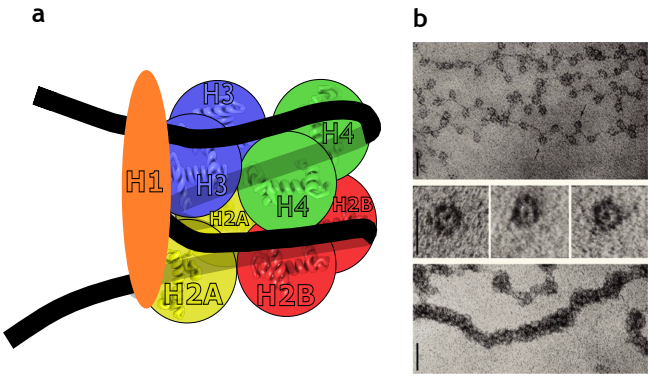


Abbildung 2.2: Aufbau und Organisation des Nukleosoms. Im linken Teil der Abbildung (a) ist die Kernstruktur des Histonkomplexes, bestehend aus je zwei Einheiten der Einzelproteine H2A, H2B, H3 und H4, gezeigt. Das Histonprotein H1 stabilisiert die Kernstruktur. Der rechte Teil der Abbildung (b) zeigt drei verschiedene Elektronenmikroskopieaufnahmen des Chromatins (Olins und Olins 2003). Deren mittlere Abbildung (b, Mitte) entspricht der schematischen Darstellung in (a). Die Strukturen oben stellen die erste Verdichtungsebene der DNA dar (Abbildung b, oben). Die sogenannte 30 nm Chromatinfaser ist im unteren Teil dieser Abbildung dargestellt (Abbildung b, unten).

In der Abbildung 2.3 ist die 3-D Struktur der ersten Organisationseinheit eines Nukleosoms abgebildet. Diese Kernstruktur bindet ungefähr 150 bp. Gezeigt ist die Seiten- (Abb. 2.3

a) und Vorderansicht (Abb. 2.3 b) dieses Komplexes (PDB: 1AOI). Die Darstellung basiert auf einer Veröffentlichung von (Luger et al. 1997). In dieser Struktur ist das Histonprotein H1 noch nicht gebunden. Ähnlich wie die aufgereihten Perlen einer Kette ist die genomische DNA mit Nukleosomen bestückt. Diese Struktur bildet damit die erste Ebene der Verdichtung der genomischen DNA im Nukleus. Das Histonprotein H1 findet Verwendung, um aus dieser Grundstruktur die nächste Verdichtungsstufe zu erreichen (siehe Abbildung 2.3 c).

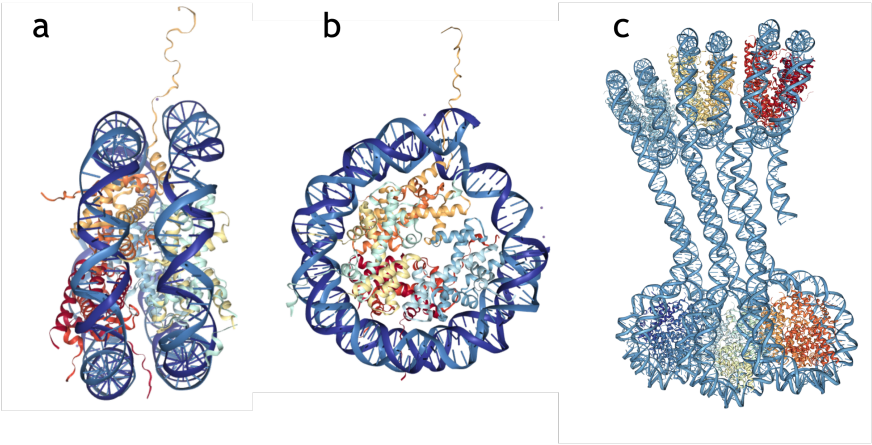


Abbildung 2.3: 3-D-Struktur des Nukleosoms. Es besteht aus jeweils zwei Einheiten der Histonproteine H2A, H2B, H3, H4. Die Vorder- (a) und Seitenansicht (b) wurden aus dem PDB-Eintrag 1AOI erzeugt. Die Darstellung in (c) zeigt die Struktur eines H1-gebundenen 6-Nukleosomen-Arrays (PDB: 6HKT) und entstammt einer Veröffentlichung von Garcia-Saez et al. (2018).

Als Ergebnis dieses Prozesses bilden sich die sogenannten Chromatinfasern aus, welche eine Dicke von ungefähr 30 Nanometern besitzen (siehe auch Abbildung 2.2 b). Weitere Komprimierungen ermöglichen schließlich den Grad der Kompaktheit, der nötig ist, um das gesamte genetische Material im Zellkern aufnehmen zu können. Dies führt schließlich soweit, dass die einzelnen Chromosomen in ihrer maximal kondensierten Form im Zellkern mikroskopisch sichtbar gemacht werden können. Dieser Zustand wird kurz vor der Zellteilung erreicht. Die Abbildung 2.4 (a) zeigt die kondensierten Chromosomen in dieser Situation für das menschliche Genom. Die Darstellung verdeutlicht, dass das genetische Material in höheren Organismen nicht als eine große Einheit vorliegt, sondern in unterschiedlich große Portionen, die sogenannten Chromosomen, aufgeteilt ist. Das Genom des Menschen enthält insgesamt 46 Chromosomen: 22 nicht geschlechtsspezifische

Chromosomen (Autosomen), die als homologe Paare vorliegen und jeweils der mütterlichen und der väterlichen Linie entstammen. Die beiden Geschlechtschromosomen (X bzw. Y) vervollständigen den Chromosomensatz. Auch diese Chromosomen liegen im gewissen Sinne doppelt vor: Das weibliche Geschlecht wird durch zwei X-Chromosomen bestimmt, während das männliche Geschlecht durch die Chromosomen X und Y geprägt wird. Auch diese beiden Chromosomen entstammen jeweils der mütterlichen und väterlichen Linie.

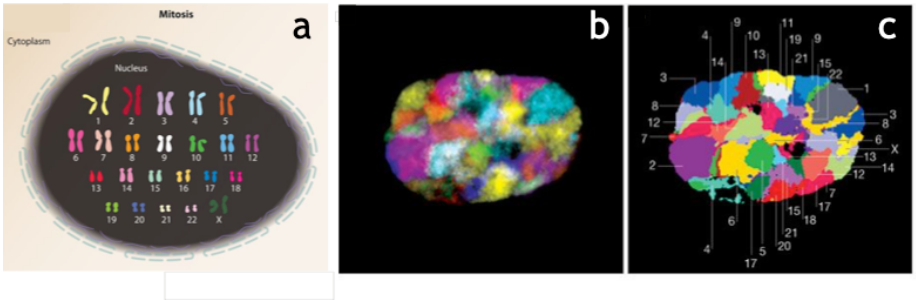


Abbildung 2.4: Organisation des menschlichen Genoms im Zellkern. Der linke Teil der Abbildung (a) zeigt die angefärbten Chromosomen einer mitotischen weiblichen Zelle (Fraser et al. 2015). Die mittlere und rechte Darstellung zeigen die chromosomalen Territorien einer menschlichen Bindegewebszelle im transkriptionsaktiven Zellkern (Bolzer et al. 2005). Die Zahlen in (c) beziehen sich auf die vorliegenden Chromosomen (1-22 und X).

Die zufällige Kombination der beiden Geschlechtschromosomen in einer befruchteten Eizelle entscheidet so über das Geschlecht des entstehenden Nachkommens. Die homologen Chromosomen besitzen eine sehr große Sequenzähnlichkeit. Diese Ähnlichkeit führt dazu, dass im humanen Genomprojekt die väterlichen bzw. mütterlichen Autosomen nicht unterschieden werden. Die Referenzsequenz des Menschen besteht somit nur aus einer Sequenz pro Chromosom für die Autosomen plus die beiden Geschlechtschromosomen X und Y (Lander et al. 2001; Venter et al. 2001). In Interphase-Zellen liegen einzelnen Chromosomen allerdings nicht in einer vollständig komprimierten Form, sondern als Chromatin vor. Sie vermischen sich aber nicht ohne Weiteres, sondern belegen einzelne Bereiche im Zellkern, die als getrennte Territorien sichtbar gemacht werden können (siehe 2.4 b und c). An den Berührungspunkten kann eine Vermischung stattfinden (Dekker et al. 2013). Das Chromatin besteht aus kompakten und weniger kompakten Bereichen, von denen nur die letzteren Transkriptionsaktivität zeigen. Aus diesem Grund sind die Transkriptionseinheiten (Gene), welche bestimmte notwendige Funktionen einer Zelle kodieren, in nicht kompaktem Chromatin zu finden. Die Transkription findet nicht ungeordnet in allen Be-

reichen des Zellkerns statt. Es werden subnukleare Bereiche ausgebildet, in denen sich die transkriptionsaktiven Regionen konzentrieren. Die inaktiven Regionen der Chromosomen interagieren miteinander und sind häufig am Kernrand zu finden (Dekker et al. 2013).

Die transkriptionsaktiven Bereiche innerhalb eines Chromosoms sind in weitere Funktionseinheiten unterteilt, die als *Topologisch Assoziierte Domänen* (TADs) bezeichnet werden. Innerhalb einer TAD lassen sich eine Vielzahl von Interaktionen beobachten, während eine Wechselbeziehung zwischen verschiedenen TADs deutlich seltener zu sehen ist (Dekker et al. 2013; Fudenberg und Mirny 2012; Gibcus und Dekker 2013). Die transkriptionsaktiven Regionen innerhalb einer TAD bilden sogenannte Transkriptionsfabriken aus (Osborne et al. 2004). Die räumliche Konzentration dieser aktiven Bereiche steigert die Transkriptionseffizienz und es wird eine vielfältige Interaktion untereinander beobachtet (Cisse et al. 2013; Rickman und Bickmore 2013; Sutherland und Bickmore 2009). TADs haben eine mittlere Größe von 880 kb und sind innerhalb der Säugetierklasse stark sequenzkonserviert (Dixon, Selvaraj et al. 2012).

Die einzelnen Regionen innerhalb einer TAD lassen sich anhand verschiedener biochemischer Modifikationen der beteiligten Histonproteine weiter unterscheiden. Dabei sind die zugänglichen Aminosäuren aller vier Histonproteine (H2A, H2B, H3, H4) betroffen. Abbildung 2.5 (a) zeigt die bekannten Veränderungen der beteiligten Proteine im Detail. Diese Darstellung ist angelehnt an eine Darstellung von Rodríguez-Paredes und Esteller (2011). Am häufigsten werden dabei Acetylierungen und Methylierungen der Aminosäure *Lysin* (K) beobachtet. Die Einbuchstaben-Kodierung der Aminosäuren wird durch das *International Union of Pure and Applied Chemistry* (IUPAC) Projekt definiert. Die Abbildung 2.5 zeigt die Strukturformel des Lysins und die vorkommenden biochemischen Variationen dieser Aminosäure durch Acetylierung (2.5 (b) bzw. Methylierung (2.5 (c)). Die verschiedenen Histonmodifikationen sind eine wichtige Ebene der Genregulation. So kann z.B. durch eine gezielte Modifikation der verschiedener Histonproteine eine krankhafte Veränderung betroffener Zellen bzw. des Gewebes erzeugt werden (Shilatifard 2012). Die Kombinationen verschiedener Modifikationen aller vier Histonproteine beeinflusst die Kompaktheit des Chromatins und steuert damit die Zugänglichkeit der DNA. Man spricht mittlerweile von einem sogenannten Histon-Code (Jenuwein und Allis 2001). Dieser Begriff deutet an, dass spezifische Kombinationen einzelner Histonmodifikationen kennzeichnend für unterschiedliche Regionen im Genom sind. Der Histon-Code wird in einer dem

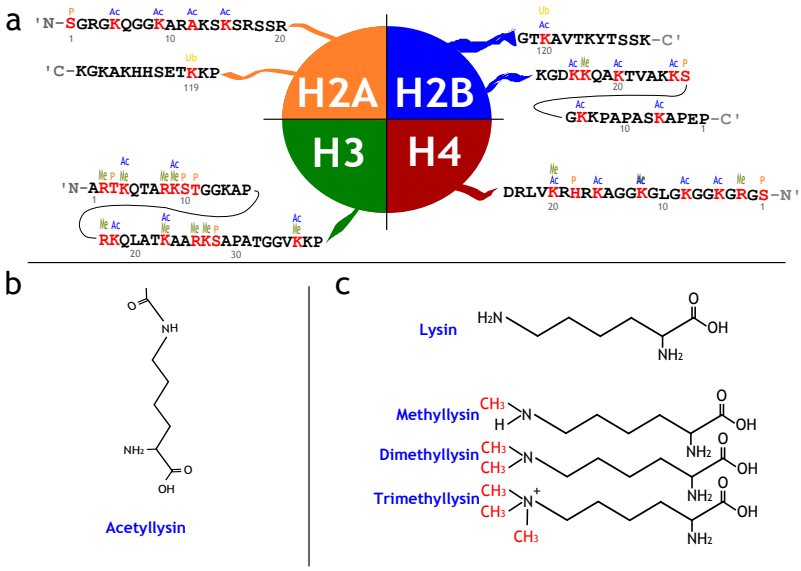


Abbildung 2.5: Bekannte Histonmodifikationen der Histonproteine H2A, H2B, H3 und H4. Die häufigsten biochemischen Veränderungen sind Acetylierungen (Ac), Methylierungen (Me), Phosphorylierungen (P) und Ubiquitinierungen (Ub) von Histonproteinen. In der oberen Teilabbildung (a) sind die verschiedenen Aminosäurenpositionen, welche bekanntermaßen modifiziert werden, gezeigt (Rodríguez-Paredes und Esteller 2011). In der Hauptsache wird die Aminosäure Lysin (Abkürzung K) verändert. Die Actylierung dieser Aminosäure zeigt Teil (b) der Abbildung, während die verschiedenen Methylierungen für Lysin in der Teilabbildung (c) dargestellt sind.

ENCODE Projekt angelehnten Veröffentlichung eindrucksvoll verdeutlicht (Ernst et al. 2011). In dieser Publikation wird die Kombination von neun unterschiedlichen Histonmodifikationen und deren Ausprägung in verschiedenen menschlichen Zelllinien untersucht (siehe Abbildung 2.6). Mit Hilfe eines sogenannten *Hidden-Markov-Modell* (HMM) können spezifische Kombinationen dieser Histonmodifikationen in 15 unterschiedlichen genomischen Regionen als charakterisierendes Merkmal identifiziert werden. Abbildung 2.6 (a) zeigt die Ausprägung dieses Codes am Beispiel des menschlichen WLS-Genlocus. Dabei werden Promotoren (Kategorie 1-3), Enhancer (Kategorie 4-7), Insulatoren (Kategorie 8), transkriptionsaktive Regionen (Kategorie 9-11) und nicht-transkriptionsaktive Regionen (Kategorie 12-15) unterschieden. Zur besseren Unterscheidung werden die einzelnen Kategorien in Abbildung 2.6 farblich unterschieden. Die Verwendung dieses Codes erfolgt in einer zelltypspezifischen Art und Weise: Die transkriptionsaktiven Histonmodifikationen im WLS Genlocus (Kategorie 9-11, grün) werden nur in fünf von neun untersuchten menschlichen Zelllinien gefunden. In zwei für diesen Genlocus nicht transkriptionsakti-

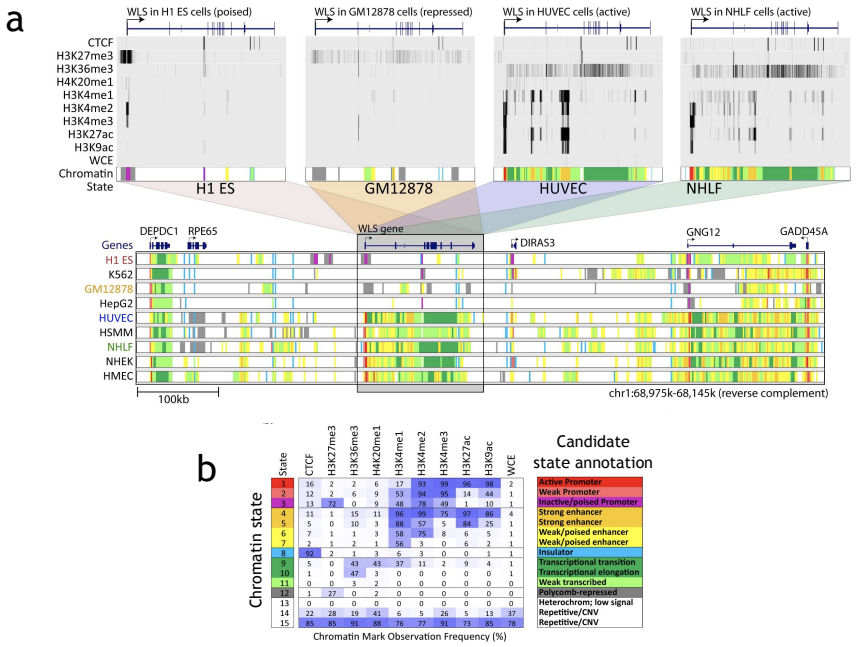


Abbildung 2.6: Chromatin-Zustandsbeschreibung verschiedener funktionaler genomischer Regionen im Menschen. Die Darstellung zeigt den Ausschnitt einer Abbildung von Ernst et al. (2011). In (a) sind die zelltypspezifischen Histonmodifikationen des menschlichen WLS-Gens für vier verschiedene Zelllinien (*H1 ES*, *GM12878*, *HUCEC*, *NHLF*) gezeigt. Der untere Teil der Abbildung (a) zeigt die Situation dieses Gens für die Gesamtheit aller untersuchen Zelllinien. Abbildung (b) listet die unterschiedlichen Histonmodifikationen auf, welche in dieser Untersuchung analysiert wurden. Durch die Kombination dieser verschiedenen *Marker* können 15 verschiedene regulatorische Regionen im menschlichen Genom identifiziert werden. Diese Zustände werden in der Abbildung farblich unterschieden.

ven Zelllinien, (*GM12878* und *H1 ES*), wird eine weitere Besonderheit deutlich: In der *H1 ES* Zelllinie zeigt der Promotor eine vielfältige Modifikation des H3 Histonproteins. Dort werden alle drei Formen der Methylierung des Lysins an der Position 4 (H3K4) gefunden, aber auch die 3fach-Methylierung des Lysins an Position 27 (H4K27me3) ist von Bedeutung (siehe auch Abbildung 2.5 a). Dieser Zustand kennzeichnet sogenannte ausbalancierte Promotoren (engl. *poised promoter*). Gerade in Stammzellen ist dieser Zustande vermehrt anzutreffen und kennzeichnet vorläufig inaktiv geschaltete Promotorregionen (Mikkelsen et al. 2007). Auf dem Weg zu einer weiteren Ausdifferenzierung können diese Promotoren wieder reaktiviert werden, falls die Transkription dieser Regionen in den nachfolgenden Entwicklungsschritten wieder notwendig werden sollte. Die Zelllinie GM12878 hingegen zeigt in der Promotorregion des WLC-Gens fast ausschließlich H3K27me3 Modifizierun-

gen und damit das typische Muster eines inaktiven (gehemmten) Promotors.

2.2 Ebenen der Genregulation

Die eukaryotische Genexpression ist ein vielfältig regulierter Prozess. Ein Gen beschreibt einen definierten Bereich im Genom, welcher durch Transkription in verschiedene *Ribonukleinsäure* (RNA)-Sequenzen übersetzt werden kann. Das bedeutet, dass ein Gen verschiedene Transkriptionseinheiten beschreiben kann, welche zelltyp- oder zeitpunktspezifisch unterschiedlich interpretiert werden können. Das in der Transkription synthetisierte RNA-Molekül besitzt entweder eine direkte biologische Funktion (z.B. *ribosomale RNA* (rRNA), *transfer RNA* (tRNA)), oder wird im Falle der mRNA durch den nachfolgenden Prozess der Translation in eine Proteinsequenz übersetzt. In den letzten Jahren wurde die Bedeutung der regulatorischen RNA immer deutlicher. Dabei sind sowohl kleine, zwischen 20 und 30 bp lange, sogenannte *micro RNA* (miRNAs) (Fire et al. 1998), als auch ca. 200 bp lange *long non-coding RNA* (lncRNAs) (Kapranov et al. 2007) bekannt geworden. Diese regulatorischen RNA-Moleküle inhibieren die Expression von Protein-kodierenden Genen, indem sie mit spezifischen Genen oder mRNA interagieren. Die Klassen von regulatorischen RNA-Molekülen steuern also sowohl die transkriptionelle als auch die posttranskriptionelle Genregulationsebene. Da in dieser Arbeit die transkriptionelle Genregulation untersucht wird, werden die posttranskriptionellen Prozesse hier nicht weiter vertieft.

2.3 Transkription

Das Ablesen bestimmter funktionaler Bereiche in einer genomischen DNA-Sequenz und die gleichzeitige Erstellung einer komplementären RNA-Einzelstrangsequenz wird als Transkription bezeichnet. Dieser Prozess wird durch eine bestimmte Klasse von Enzymen, die sogenannten RNA-Polymerasen, katalysiert. Dabei dient ein definierter DNA-Strang (Matrizenstrang oder auch *template* genannt) als Ausgangspunkt zur Erstellung des komplementären RNA-Strangs. In Bakterien (Prokaryoten, Zellen ohne Zellkern) ist nur

ein einzelnes Enzym für die Transkription von DNA in RNA zuständig. In Eukaryoten (Zellen mit Zellkern) sind drei verschiedene RNA-Polymerasen bekannt (I, II, III), in verschiedenen Pflanzen sind noch zusätzlich zwei weitere RNA-Polymerasen zu finden (IV, V).

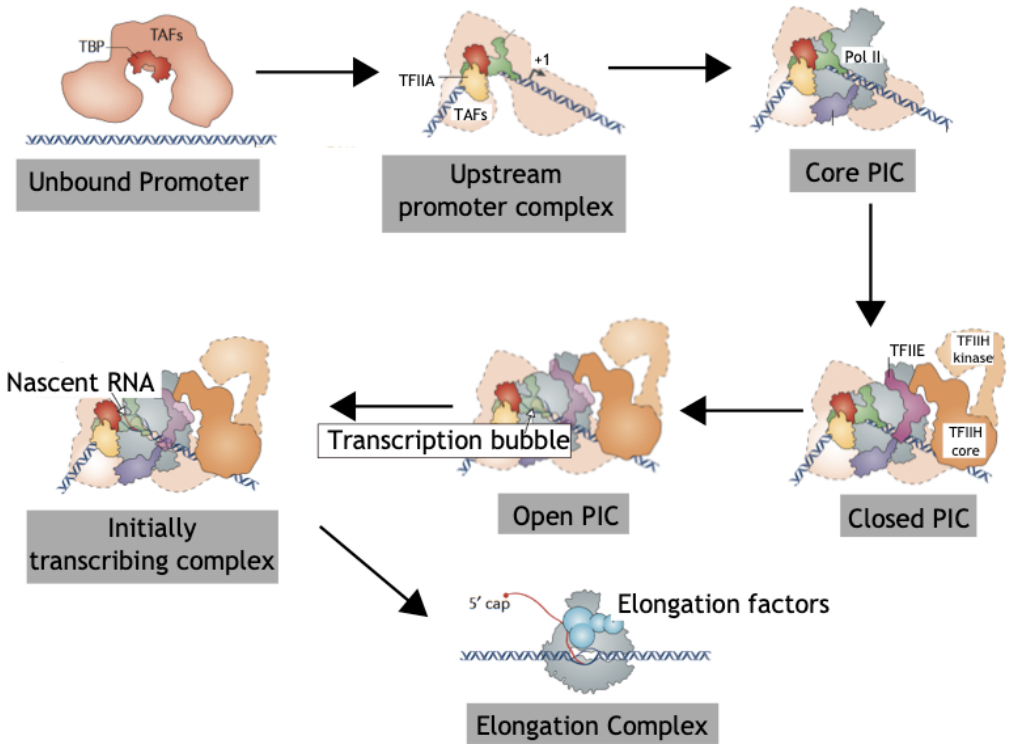


Abbildung 2.7: Der Prozess der Transkription. Der Prozess der Transkription ist ein komplex regulierter Prozess. Das Enzym (Poll II) bindet mit Hilfe allgemeiner TFs an einen definierten Bereich (Promotor) der DNA, kopiert die nachfolgende Transkriptionseinheit und beendet diesen Vorgang an DNA-definierten Sequenzsignalen. Die Abbildung entstammt im Original einer Veröffentlichung von Sainsbury et al. (2015) und wurde zur besseren Lesbarkeit leicht verändert.

Die Abbildung 2.7 (a) zeigt schematisch, welche einzelnen Schritte bei der Transkription durchlaufen werden. Zu Beginn wird der zu kopierende DNA-Strang durch einen sogenannten *Präinitiationskomplex* (PIC) gebunden. Dieser besteht aus generellen TFs und der jeweiligen RNA-Polymerase. Die Transkription unterscheidet drei verschiedene Phasen: als Erstes wird der vorläufige Transkriptionsapparat an die DNA gebunden (engl. *Closed PIC*). Dann wird die DNA lokal in ihre zwei Einzelstränge aufgeteilt. In diesem Schritt entsteht die sogenannte Transkriptionsblase (engl. *Transcription Bubble*) und es entsteht

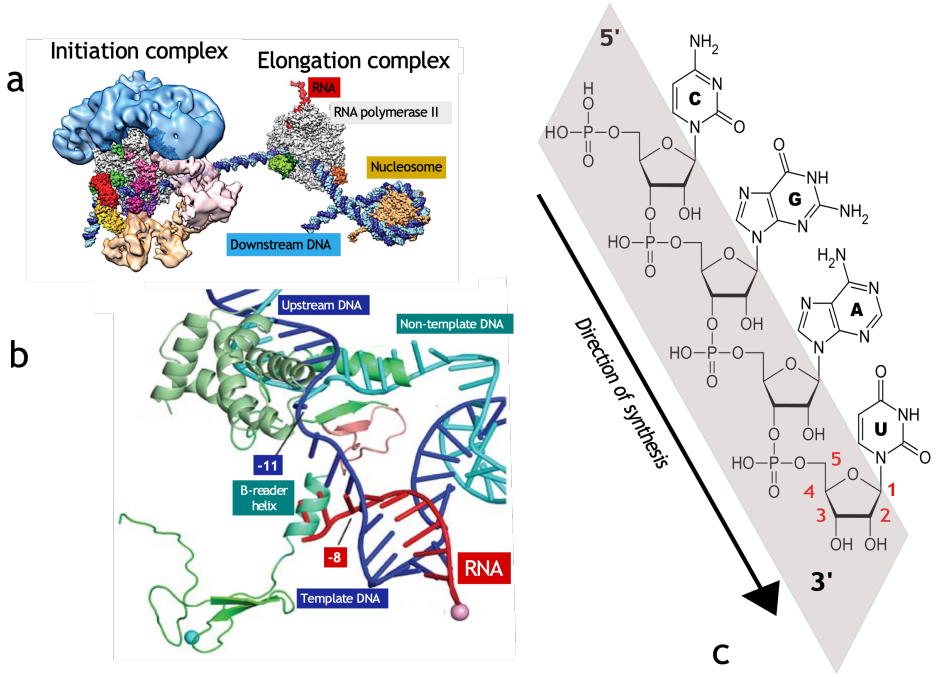


Abbildung 2.8: Details der Transkription. Bildausschnitt (a) zeigt schematisch den Proteinkomplex der Transkription (Cramer 2017). Der Open PIC-Komplex trennt die beiden DNA-Stränge auf (b) (Kostrewa et al. 2009). Der Matrizenstrang (blau) wird von der RNA-Polymerase abgelesen und der komplementäre RNA-Strang wird erzeugt (rot). Im Bildausschnitt (c) der Abbildung ist beispielhaft die Strukturformel eines RNA-Einzelstrangs abgebildet.

der geöffnete PIC (engl. *Open PIC*). Im nächsten Schritt verlassen einige Proteine diesen Komplex wieder und neue Proteine lagern sich an, welche den Fortgang der Transkription ermöglichen. Der Abbruch der Transkription wird bei denjenigen Genen, welche durch die RNA-Polymerase-II transkribiert werden, durch ein sogenanntes Polyadenylisierungssignal eingeleitet. Auf der DNA-Seite wird dieses durch die Basenabfolge `3'-TTATTT-5'` definiert, welche in der entsprechenden RNA-Sequenz dann zur `5'-AAUAAA-3'`-Sequenz transkribiert wird. Stromabwärts dieses Signals wird die RNA-Sequenz geschnitten. Abbildung 2.8 (a) wurde von Kostrewa et al. (2009) veröffentlicht, sie verdeutlicht die Größenverhältnisse des initialen PICs im Vergleich zur aktiv transkribierenden RNA-Polymerase-II (Cramer 2017). Die Funktion des Open PIC Komplexes wird in Abbildung 2.8 (b) detailliert gezeigt. Der zu transkribierende Strang (blau) dient dabei als Matrize und wird in 3'-5'-Orientierung abgelesen. Dieser Strang wird benutzt, um den neu entstehenden RNA-Strang (rot) durch komplementäre Basenpaarung zu bilden. Dieser wird in 5'-3'-Richtung

erzeugt, indem der Matrizenstrang (engl. *template*) der DNA in 3'-5'-Richtung abgelesen wird. Während der Transkription entsteht so ein DNA-RNA-Hybrid. Ähnlich wie in der DNA-Sequenz paaren in diesem Hybrid wieder Guanin und Cytosin miteinander. Die Nukleinsäure Thymin wird in einer RNA-Sequenz allerdings durch *Uracil* (U) (Strukturformel siehe 2.8 c) ausgetauscht, so dass sich nun also A-U-Paare in dem Hybridstrang finden lassen. Im Unterschied zur Desoxyribose, welche bei der Bildung des Zuckerphosphat-Rückgrats der DNA Verwendung findet, wird in der RNA-Sequenz eine Ribose benutzt. Abbildung 2.8 (c) verdeutlicht den Aufbau einer RNA-Sequenz beispielhaft.

2.4 Initiation der Transkription

Die Initiation der Transkription, also die Rekrutierung der jeweiligen RNA-Polymerase an den zu transkribierenden Bereich, ist der entscheidende Schritt der eukaryotischen Transkriptionsregulation. Der Startpunkt der Transkription (*Transcription Start Site* (TSS)) wird durch definierte Sequenzmotive, die auch als Kernpromotorelemente (*Core Promoter Element* (CPEs)) bezeichnet werden, festgelegt. Ein Sequenzmotiv beschreibt dabei eine Menge von DNA-Sequenzen, welche an verschiedenen Stellen im Genom ähnlich vorkommen. Sie bilden somit ein Motiv, dem eine biologische Bedeutung zugewiesen werden kann. Eine gewisse Variabilität wird bei diesen regulatorischen Sequenzsignalen beobachtet. Aus diesem Grund wird ein Sequenzmotiv häufig durch eine sogenannte Konsenussequenz oder alternativ auch durch ein Sequenzlogo beschrieben. Abbildung 2.9 zeigt verschiedene CPEs und deren Positionierung um den TSS in der Form eines Sequenzlogos (Haberle und Lenhard 2016). Eines der bekanntesten CPE ist die sogenannte TATA-Box: Sie befindet sich 25-30 bp stromaufwärts des eigentlichen TSS (Bucher 1990). Der TSS wird in der Literatur häufig mit einem Pfeil gekennzeichnet (siehe Abbildung 2.9), die Pfeilrichtung gibt dabei die Transkriptionsrichtung an. Dieses Motiv ist aber nicht das häufigste CPE, welches in Promotoren gefunden werden kann, da es nur ungefähr in 10 Prozent aller menschlichen Promotoren nachgewiesen wurde (Dreos et al. 2013). Alternativ zur TATA-Box können weitere Sequenzmuster beobachtet werden, z.B. das sogenannte Inr-Motiv (C. Yang et al. 2007). Dieses findet sich in ungefähr 30 Prozent aller menschlichen Promotorregionen, zumeist um den TSS herum (Dreos et al. 2013). Es existieren aber auch stromabwärts gelegene Kernelemente, die für eine Festlegung des TSS wichtig sind.

Grundsätzlich ist festzustellen, dass die Auswahl der CPEs und die Kombination dieser Signale zur Festlegung des TSS noch nicht intensiv genug untersucht worden sind. Die gezeigten CPEs der Abbildung 2.9 sind allein nicht ausreichend, um den TSS positionsgenau festzulegen. Ob eine Kombination dieser Elemente unterschiedliche Auswirkungen z.B. auf die Transkriptionsaktivität hat, muss ebenfalls noch weiter untersucht werden.

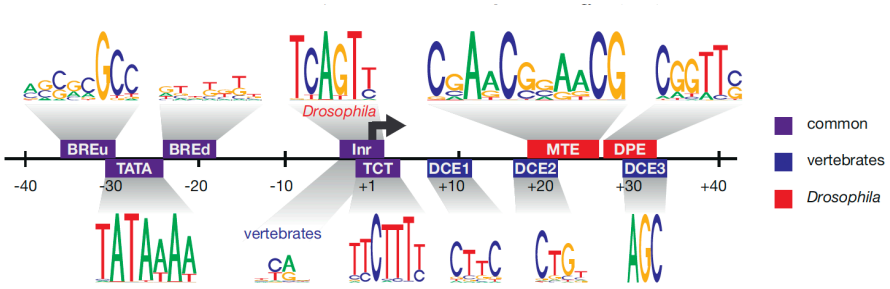


Abbildung 2.9: Bekannte Kernpromotorelemente und ihre Motivbeschreibung. Die Abbildung wurde einer Veröffentlichung von Haberle und Lenhard (2016) entnommen. Sie zeigt die Sequenzlogos verschiedener Kernelemente und deren Positionierung um den Transkriptionsstart (Pfeil). Die Verwendung der verschiedenen CPEs in wirbellosen- bzw. Wirbeltieren kann der Legende auf der rechten Seite entnommen werden. Die Region um den TSS besitzt mehrere überrepräsentierte Sequenzmuster, welche einzeln oder in Kombination vorkommen können. Die folgenden Kernpromoterelemente sind dargestellt: *B Becognition Element* (BRE), *TATA box element* (TATA), *Initiator element* (Inr), *polypyrimidine initiator* (TCT), *Downstream Core Element* (DCE), *Motiv Ten Element* (MTE), *Downstream Promoter Element* (DPE).

Des Weiteren scheint die Positionierung der Nukleosomen bei der Definition eines TSS von Bedeutung zu sein. Durch die Entwicklung neuer experimenteller Methoden konnte die positionsgenaue Bestimmung des TSS wesentlich verbessert werden. Das Hochdurchsatzverfahren *Cap Analysis of Gene Expression* (CAGE) (Shiraki et al. 2003) ist eine der ersten experimentellen Methoden, welche in diesem Zusammenhang eingesetzt wurde. Im FANTOM-Projekt wurden mit Hilfe dieses Verfahrens verschiedene TSSs im Genom der Maus und des Menschen experimentell ermittelt (Forrest et al. 2014). In der Literatur wird die Anwendung der CAGE Technologie durchaus kritisch diskutiert (Haberle und Lenhard 2016). Das experimentelle Protokoll lässt hier den Schluss zu, dass die langlebigen Transkripte die Ergebnismenge eines CAGE Experimentes dominieren. Zusätzlich wird eine ungenaue bzw. unvollständige Abbildung des TSS diskutiert. Auf der Basis des *Global Run-On sequencing* (GRO-seq) Verfahrens (Core, Waterfall et al. 2008) und dessen Weiterentwicklung (Core, Martins et al. 2014) konnte eine deutliche Verbesserung in der Qualität und Quantität bei der positionsgenauen Bestimmung der TSSs erreicht werden. Abbildung

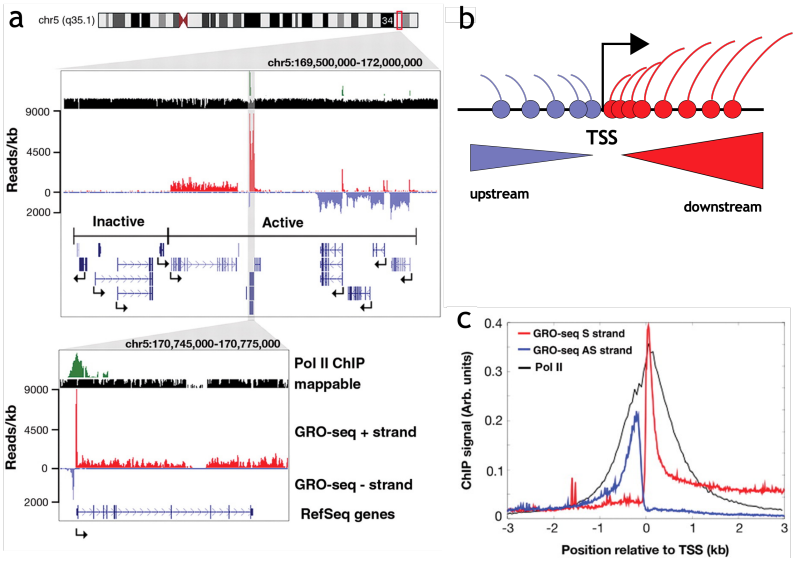


Abbildung 2.10: Transkriptionsinitiation in menschlichen Promotoren. Abbildung (a) zeigt das Ergebnis eines GRO-seq-Experiments und entstammt einer Veröffentlichung von (Core, Martins et al. 2014). Durch die positionsgenaue Kartierung der Sequenzfragmente eines GRO-seq-Experiments kann der TSS eindeutig bestimmt werden. Dabei zeigt sich, dass bei den aktiven Transkriptionseinheiten eine bi-direktionale Transkription am TSS erfolgt. Es werden sowohl RNA-Fragmente des Plusstrangs als auch des Minusstrangs gefunden. Es zeigt sich aber, dass die Anzahl der Fragmente am TSS für den Plusstrang wesentlich größer ist als für den Minusstrang. Dieses Verhalten wird in (b) modellhaft gezeigt. Die Überlagerung der Peak-Position des RNA-Polymerase II-Enzyms und der RNA-Fragmente des Plusstrangs ist in (c) dargestellt.

2.10 (a) zeigt das Ergebnis eines GRO-seq-Experimentes. Ein wesentlicher Vorteil dieser Technik ist, dass die sequenzierten Fragmente strangspezifisch auf das entsprechende Genom kartiert werden können (siehe Abbildung 2.10: rot: Plusstrang; blau: Minusstrang). Diese Abbildung verdeutlicht, dass um den TSS herum sowohl stromaufwärts (-, *antisense*) als auch stromabwärts (+, *sense*) definierte RNA-Fragmente entstehen. Der Plusstrang zeigt allerdings im Vergleich zum Minusstrang eine deutlich höhere Transkriptionsrate (Anzahl an RNA-Fragmenten), was in Abbildung 2.10 (b) modellhaft verdeutlicht wird. Die besondere Verteilung der RNA-Fragmente ist das Ergebnis der gerichteten Transkription (unidirektional): Aktive TSSs zeigen eine hohe Transkriptionsrate über den gesamten zu transkribierenden Bereich (siehe auch Abbildung 2.10 a, unten), die RNA-Fragmente des Minusstrangs sind nur am TSS zu beobachten und sind im Vergleich zu den Plusstrangfragmenten wesentlich kürzer. In der Untersuchung von Core, Martins et al. (2014) wird

weiterhin eine besondere Gruppe von Genen beobachtet. Diese Gruppe zeigt das am TSS vorliegenden typische Verhältnis der Plus- und Minusstrangfragmente. Es lassen sich aber für den gesamten Transkriptionsbereich keine weiteren RNA-Fragmente finden. Die Autoren dieser Studie bezeichnen diese Gruppe als sogenannte pausierende Gene. Die grundsätzliche Aktivierbarkeit ihrer TSSs könnte auf eine zurückliegende und/oder nachfolgende Bedeutung hinweisen. In vielzelligen Tieren (Metazoan) werden verschiedene Promotorklassen beobachtet. Die verschiedenen Transkriptionsaktivitäten deuten auf eine Abhängigkeit von unterschiedlichen CPE in diesen Promotorklassen hin. Weiterhin unterstützen verschiedene Histonmodifikationen der unterschiedlichen Histonproteine der Nukleosomen am TSS die Positionierung der RNA-Polymerase. So kann z.B. für die Gruppe der Protein-kodierenden Gene gezeigt werden, dass sich die GRO-seq definierten Transkriptionsstartpunkte der Fragmente des Plus- und Minusstranges in Übereinstimmung mit den Lokalisationen der RNA-Polymerase-II befinden (siehe Abbildung 10 c).

2.5 Regulation der Transkription

Die Regulation der Transkription wird im Wesentlichen durch zwei unterscheidbare funktionale Regionen gesteuert: Der Promoter (a) initiiert und unterstützt die Transkription, während der (oder die) Enhancer (b) diesen Prozess weiter stimulieren. In der gesamten Promotorregion, die etwa 300-600 bp umfasst, wird häufig noch ein unmittelbar am TSS gelegener Kernpromoter *core promoter* (CP) definiert (Haberle und Stark 2018; Lenhard, Sandelin et al. 2012). Der CP beschreibt eine ungefähr 100 bp umfassende Region (-50 stromaufwärts und 50 bp stromabwärts) um den TSS (Arnold et al. 2017). Die CPEs sind in diesem Bereich lokalisiert und definieren den TSS zusammen mit den dort positionierten Nukleosomen (Haberle und Stark 2018). Der CP zeigt ohne weitere Unterstützung des ihn umgebenden weiter gefassten Promotorbereichs eine nur sehr niedrige Transkriptionsaktivität (Haberle und Stark 2018; Lenhard, Sandelin et al. 2012). Die Enhancer verstärken diesen Effekt: Interagierende Enhancer können mehrere Hunderttausend Basenpaare entfernt vom eigentlichen TSSs lokalisiert sein. Aus diesem Grund werden sie auch als distale regulatorische Bereiche bezeichnet. Die Enhancer- bzw. Promotorbereiche unterscheiden sich im Methylierungs- bzw. Acetylierungsstatus verschiedener Histonproteine der dort lokalisierten Nukleosomen (siehe auch Abbildung 2.5 und 2.6). Es ist nicht ausgeschlossen,

dass einige Promotorregionen auch als distale Enhancer für andere Promotoren wirken können (Dao et al. 2017).

Für die Interaktion der proximalen Promotorregion mit dem Kernpromotor wurde von Carey (1998) der Begriff des Enhanceosoms (engl. *enhanceosome*) ausgearbeitet. Der Begriff findet zum ersten Mal in einem Artikel von Bazett-Jones et al. (1994) Verwendung. Das Modell des Enhanceosoms beschreibt die Kommunikation des proximalen Promotors mit den daran gebundenen TFs und dem PIC sowie den Enhancer-gebundenen Proteinkomponenten. Die Kommunikation wird durch verschiedene definierte *Protein-Protein-Interaktion* (PPI) der beteiligten Proteine ermöglicht (siehe auch Abbildung 2.11 a).

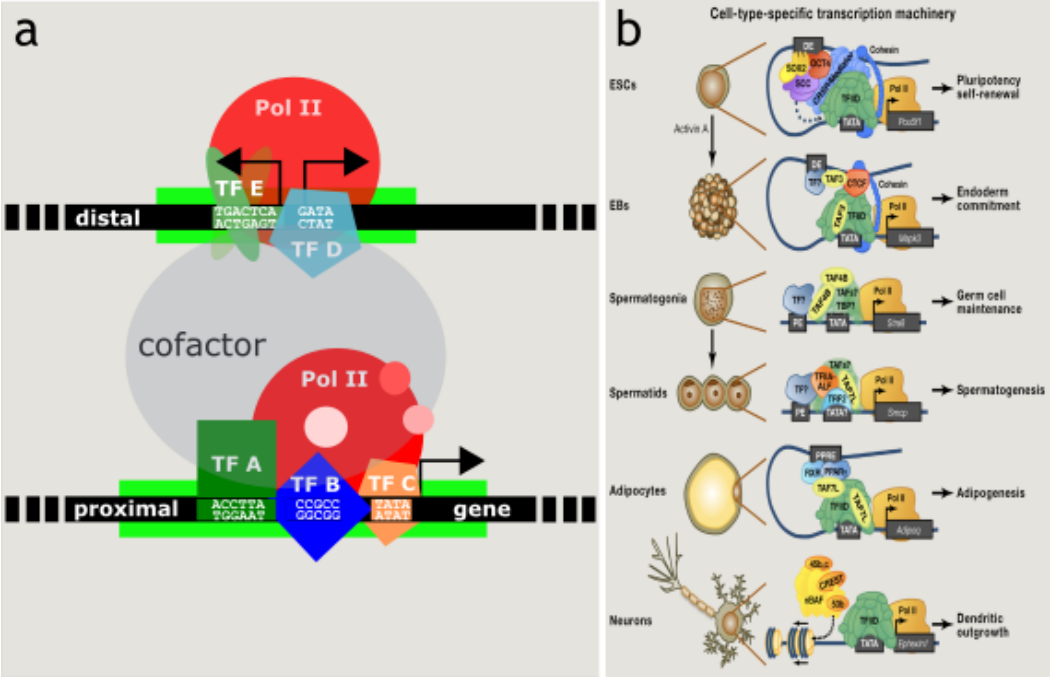


Abbildung 2.11: Das Modell des Enhanceosoms. (a) Das Enhanceosom besteht aus einem komplexen Interaktionsnetzwerk verschiedener sequenzspezifisch- und sequenzunspezifisch-bindender Proteine. Diese können sowohl in distalen regulatorischen Regionen (Enhancern) als auch in proximalen regulatorischen Regionen (Promotoren) binden. Durch die Kombination dieser Regionen entsteht ein Proteinkomplex höherer Ordnung, der einen synergistischen oder antagonistischen Effekt auf die Transkription der jeweiligen Transkriptionseinheit ausübt. (b) zeigt verschiedene regulatorische Beispiele von Interaktionen des PIC mit proximalen und distalen Enhancern (Levine et al. 2014).

Die TFs binden an ihre größtenteils sequenzspezifischen Signale, welche im Kernpromotor oder im proximalen Promoter zu finden sind, und interagieren nun auf der nächsten Ebene durch definierte PPIs direkt oder indirekt miteinander. Das Modell versucht gleichzeitig die verschiedenen Ebenen der Transkriptionsregulation zu berücksichtigen: Als erstes wird die Kooperation verschiedener sequenzspezifisch- und sequenzunspezifisch bindender TFs am proximalen Promoter angenommen. Die Bindung des PICs und die Interaktion dieses Komplexes mit weiteren generellen TFs (GTF) definieren zweitens die nächste Ebene dieses Modells. Weiterhin wird eine abgestimmte Interaktion zwischen den gebundenen TFs und dem PIC, einschließlich der dort gebundenen GTFs, in der Kernpromotorregion berücksichtigt. Die Summe der gesamten kooperativen Interaktionen definiert somit das funktionale Enhanceosom und sorgt für eine messbare Transkriptionsaktivität. Je nach Zelltyp oder Entwicklungsstadium der Zelle kann diese Transkriptionsaktivität unterschiedlich stark ausgeprägt sein. Diese Variation wird durch die unterschiedliche Zusammensetzung des regulatorischen Moduls erreicht. Das bedeutet, dass je nach Zeitpunkt oder Zelltyp unterschiedliche TFs das Enhanceosom einer Transkriptionseinheit definieren können. Die unterschiedlichen kombinatorischen Einflüsse auf die Transkription äußern sich häufig in einer synergistischen (mehr als additiven) Transkriptionsaktivität (Carey 1998). Es sind aber auch antagonistische Einflüsse bekannt.

Die verschiedenen distalen Einflüsse vervollständigen die Ausprägung des funktionalen Enhanceosoms. Auch in diesen Regionen binden verschiedene TFs (Courey und J.-D. Huang 1995; Merika et al. 1998). Diese Enhancer interagieren ebenfalls mit ihren zugehörigen proximalen Regionen und Kernpromotoren durch definierte PPI. Abbildung 2.11 (a) zeigt modellhaft den komplexen Aufbau der Enhancer-Promoter-Interaktion, welcher durch vielfältige PPIs definiert wird. Abbildung 11 (b) entstammt der Veröffentlichung von Levine et al. (2014) und bildet verschiedene bekannte PICs und deren interagierende proximale und distale Enhancer ab. Die dort aufgeführten regulatorischen Module zeigen noch einmal, dass der PIC nicht nur aus ein paar wenigen und immer gleichen GTFs besteht, sondern sehr unterschiedlich aufgebaut sein kann: Ein PIC besteht aus mehr als 85 einzelnen Polypeptiden und unter diesen sind Pol II, TFIID, E, F, H und verschiedene Koaktivatoren mit jeweils zahlreichen Untereinheiten zu finden (Cramer 2002; Goodrich und Tjian 2010; Roeder 1996). Ein Enhancer muss nicht notwendigerweise mit dem nächstgelegenen Promotor interagieren, es können große Distanzen überbrückt werden (Shlyueva et al. 2014; van Arensbergen et al. 2014). Die Wirkungsweise eines Enhancers ist in der Regel aber auf

einen definierten Bereich in einem Chromosom begrenzt. Dieser Bereich wird als TAD bezeichnet (Dixon, Selvaraj et al. 2012), die mehrere Millionen bp umfassen kann. Innerhalb

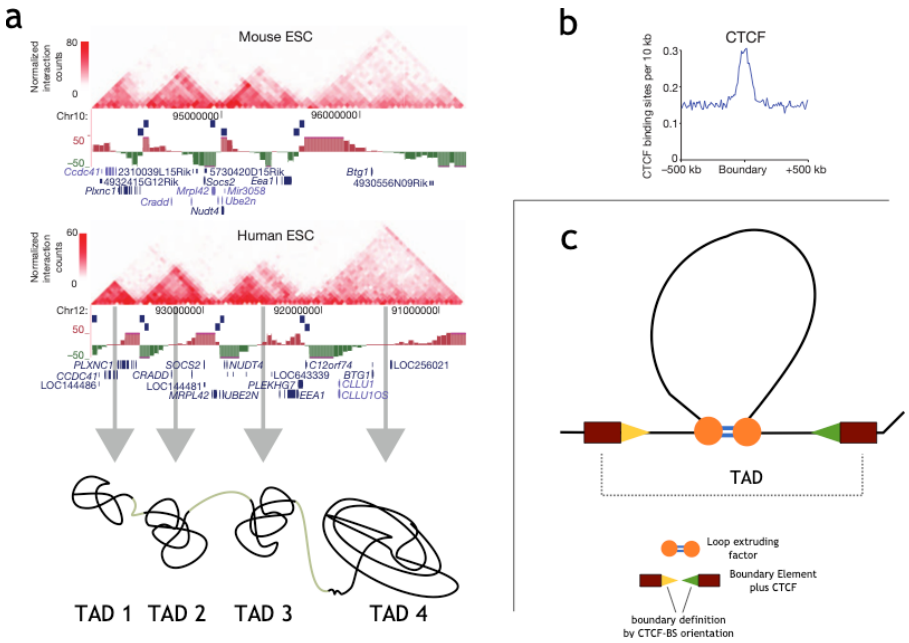


Abbildung 2.12: Die TADs homologer regulatorisch aktiver chromosomaler Regionen des Menschen und der Maus sind sequenzkonserviert. Abbildung (a, oberer Teil) entstammt einer Veröffentlichung von Dixon, Selvaraj et al. (2012). Innerhalb einer TAD können experimentell vielfältige genomische Interaktionen beobachtet werden, die mit einer Heatmap darstellbar sind. Die Intensität des roten Farbtons korreliert mit der Anzahl der beobachteten Interaktionen. Die Menge der Interaktionen ist auf bestimmte Bereiche konzentriert. Diese sind durch die *Heatmap*-Darstellung gut erkennbar. Der bereits vorgestellte Begriff der sogenannten Transkriptionsfabriken (Osborne et al. 2004) fügt sich gut in die beobachtete TAD-Struktur der Chromosomen ein. Die Grenzen zwischen zwei TADs sind durch eine hohe Konzentration des DNA-bindenden Proteins CTCF gekennzeichnet (siehe Abbildung b, (Dixon, Selvaraj et al. 2012)). Im Teil (c) dieser Darstellung wird die Bedeutung dieses TFs modellhaft erläutert: Durch die strangspezifische Bindung dieses Proteins in Kombination mit weiteren Proteinen werden die TAD-Grenzen definiert. Weiter Proteine unterstützen die Ausprägung lokaler Schleifenstrukturen (engl. *loop*) innerhalb einer TAD.

einer TAD lassen sich eine Vielzahl von Interaktionen beobachten, während zwischen verschiedenen TADs eines Chromosoms oder zwischen TADs verschiedener Chromosomen diese Interaktionen eher selten zu beobachten sind. Die Grenzen eines TAD schränken die Wirkungsweise der regulatorischen Regionen ein (Symmons et al. 2014) und das gezielte Entfernen einer TAD definierten Grenzregion ist mit der Schaffung neuer Enhancer-Promotor-Paare verknüpft, welche einen pathologischen Effekt verursachen können (Lu-

piáñez et al. 2015). Abbildung 2.12 zeigt die TAD-Organisation des Chromosom 6 in der Maus (Dixon, Selvaraj et al. 2012).

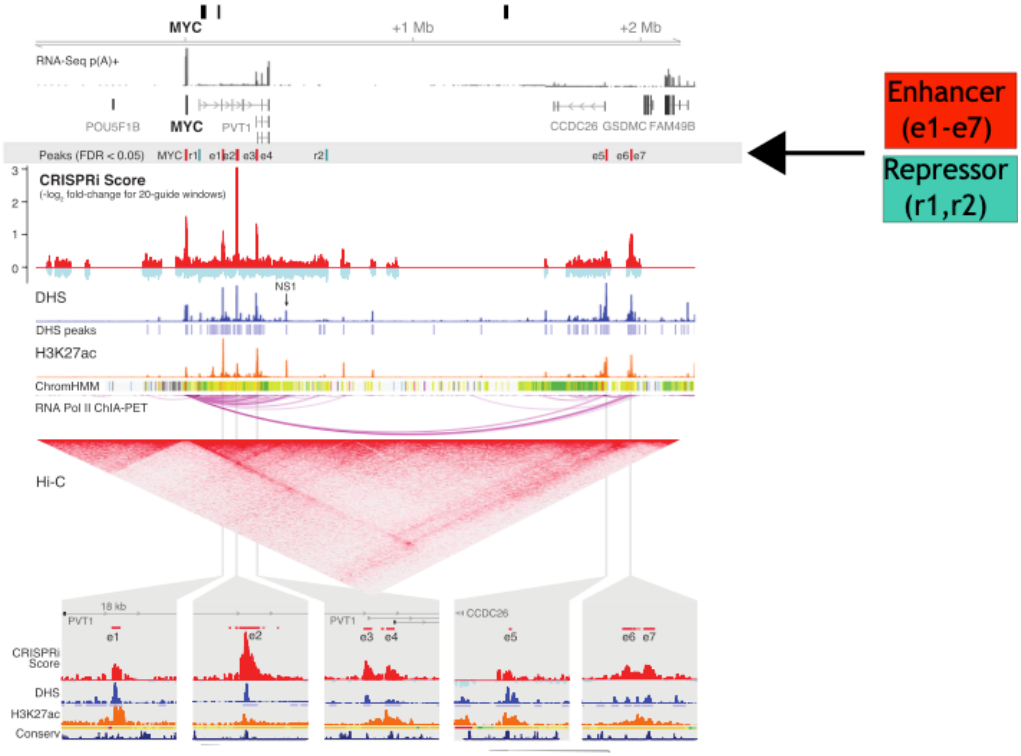


Abbildung 2.13: Diese Abbildung wurde von Fulco und Kollegen erstellt (Fulco et al. 2016). E1-E7 bezeichnen die gefundenen distalen Enhancerregionen (siehe Pfeilmarkierungen). Die Enhancer sind dort rot markiert, während die gefundenen Repressoren (r1 und r2) blau markiert sind. Auch für die Repressoren konnten Fulco und Kollegen eine signifikante biologische Funktion nachweisen.

Die TAD-Grenzen sind zwischen verschiedenen Zelltypen ausgeprägt sequenzkonserviert (Dixon, Jung et al. 2015; Dixon, Selvaraj et al. 2012; Rao et al. 2014), aber auch innerhalb der Säugetierklasse bleiben diese Grenzen erhalten (Dixon, Selvaraj et al. 2012; Vietri Rudan et al. 2015). Weiterhin gilt, dass mehrere Enhancer die Expression eines Gens beeinflussen können (Fulco et al. 2016). Fulco et al. (2016) zeigen diese Eigenschaft z. B. für die Expression des MYC Gens in der menschlichen Zelllinie K562. Die Autoren haben insgesamt sieben verschiedene Enhancer (E1-E7) identifiziert, welche zusammengenommen eine Region von 1,6 Megabasen (Mb) verwenden. Der Einfluss jedes einzelnen Enhancers auf die Transkriptionsaktivität des MYC Gens liegt dabei zwischen 9 und 62 Prozent. Abbildung 2.13 zeigt diese sieben Enhancer in ihrer genomischen Lokalisation (Fulco et al.

2016). Enhancer können auch eine hemmende Wirkung ausüben, sie werden dann häufig auch als Repressor bezeichnet. Auch für das MYC Gen sind zwei Repressoren experimentell bestimmt worden (siehe Abbildung 13, r1 bzw. r2). Abbildung 2.13 (unten) verdeutlicht verschiedene Eigenschaften der distalen regulatorischen Regionen: (1.) Enhancer besitzen eine deutliche regulatorische Funktion (ausgeprägtes DHS-Signal). (2.) Die Nukleosomen in diesen Bereichen zeigen eine spezifische Histonmodifikation der Aminosäure Lysin an Position 27 des Histonprotein 3 (H3K27ac). (3.) Die experimentell bestätigten sieben Enhancer zeigen eine messbare Sequenzkonserviertheit innerhalb der Säugetierklasse (siehe Abbildung 2.13, unten).

2.6 Transkriptionsfaktoren

Transkriptionsfaktoren dekodieren und interpretieren genomische Information. Die meisten dieser Proteine binden an definierte DNA-Sequenzen, interagieren direkt oder indirekt mit dem PIC und steuern so die Transkriptionsaktivität in einer essentiellen Art und Weise. Die Gruppe der DNA-bindenden Transkriptionsfaktoren im menschlichen Genom umfasst ungefähr 1600 Gene (Lambert et al. 2018; Wingender, Schoeps und Dönitz 2013). Die Funktionseinheit in den TFs, welche für die sequenzspezifische Erkennung verantwortlich ist, wird als *DNA-bindende Domänen* (DBDs) bezeichnet. In Eukaryoten werden ungefähr 100 verschiedene DBD Typen unterschieden (Lambert et al. 2018). Diese Domänen sind z. B. in der Pfam-, SMART- oder Interpro-Datenbank beschrieben (Finn, Attwood et al. 2017; Finn, Coggill et al. 2016; Letunic und Bork 2018). Die strukturellen Eigenschaften dieser DBDs können als Unterscheidungskriterium benutzt werden, um die verschiedenen TFs zu klassifizieren. Eine der ersten Taxonomien auf Basis der DBDs wurde bereits im Jahre 1991 vorgenommen (Harrison 1991). Eine vollständige Annotation und Klassifikation aller menschlicher TFs ist durch Wingender, Schoeps und Dönitz (2013) durchgeführt worden. Die dort verwendete Einteilung der DNA-bindenden Domänen geht auf eine Veröffentlichung von E. Wingender (1997) zurück. Die Klassifikation basiert auf der generellen Topologie der DBD und dem Modus dieser Interaktion mit den gebundenen DNA-Sequenzen (Wingender, Schoeps und Dönitz 2013). Die Einteilung wird gegebenenfalls um die Di- bzw. Multimerisierungsbereiche erweitert, um eventuelle, für die DNA-Bindung notwendige, PPIs abbilden zu können (Wingender, Schoeps und Dönitz 2013).

Insgesamt werden auf diese Art und Weise neun definierte Superklassen unterschieden. Tabelle 2.1 zeigt diese Superklassen mit einem Beispiel-TF und dessen 3D-Struktur. Die zusätzliche Superklasse *Null* mit dem Namen, «undefined DNA-binding domains» enthält DBDs, die bisher noch keiner existierenden Einteilung zugordnet werden konnten. Auf der zweiten Ebene (Klasse) werden die einzelnen Superklassen weiter spezifiziert. Dem grundlegenden Gedanken folgend, dass ähnliche DNA-Sequenzen von gleichartigen DBDs erkannt werden, spezifizieren die nächsten beiden Unterebenen (Familie und Unterfamilie) die einzelnen DBDs zusätzlich weiter. Abbildung 2.14 gibt dafür ein Beispiel: Die Superklasse 1 beschreibt die Gruppe der sogenannten Basis domains (TFClass ID: 1). Sie enthält unter anderem eine Klasse mit dem Namen *Basic leucine zipper factors (bZIP)*

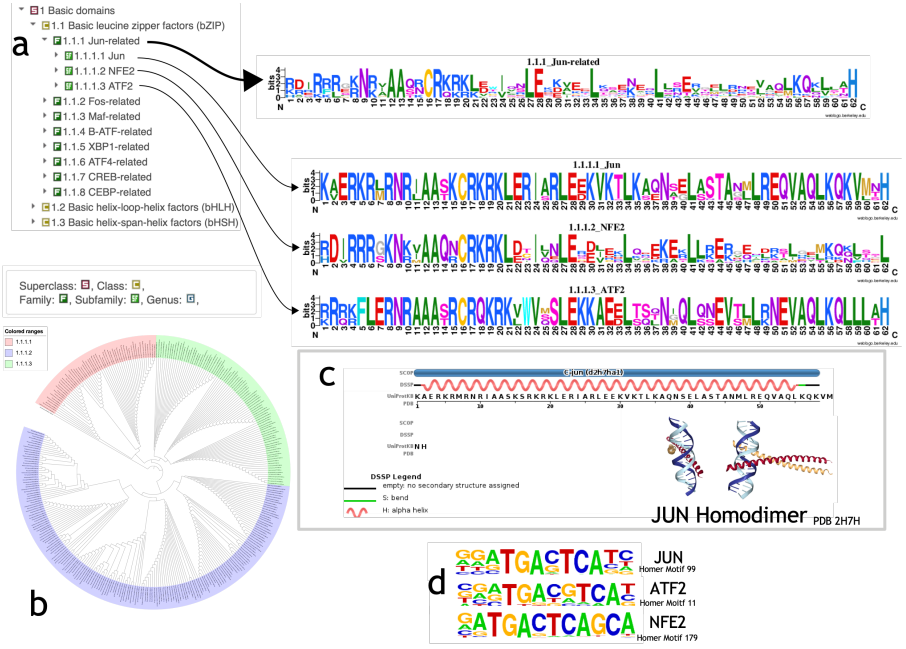


Abbildung 2.14: DNA-bindende Domänen der JUN-Familie. Die Darstellung zeigt den Inhalt des TFClass-Projekts am Beispiel der JUN-Familie (Wingender, Schoeps, Haubrock, Krull et al. 2018). (a) zeigt die Klasse der Basis *leucine zipper factors (bZIP)* mit den acht Superfamilien (1.1.1-1.1.8). Das Sequenzlogo der JUN-related Familie (1.1.1) und die drei Unterfamilien (JUN 1.1.1.1, ATF2 1.1.1.2 und NFE2 1.1.13) werden auf der rechten Seite gezeigt. Der phylogenetische Baum der Säugetiergruppe (b) verdeutlicht die Sequenzkonserviertheit der DBD dieser Familie. Im Bildausschnitt (c) wird die menschliche Domäne des JUN-TF gezeigt und die 3-D Struktur dieses Homodimers (PDB: 2H7H). Die drei Sequenzlogos (d) fassen die Transkriptionsfaktorbindestellen der drei ausgewählten Unterfamilien zusammen (Heinz et al. 2010a).

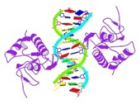
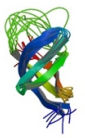
Superclass	Example (PDB)	Visualization
(1) Basic domains	CREB (1DH3)	
(2) Zinc-coordinating DNA-binding domains GR	GR (1R4R)	
(3) Helix-turn-helix domains	Pax-6 (6PAX)	
(4) Other all-alpha-helical DNA-binding domains	SRY (1J46)	
(5) Alpha-Helices exposed by beta-structures	MEF2A (1C7U)	
(6) Immunoglobulin fold	NF-kappaB p50 (1SVC)	
(7) Beta-Hairpin exposed by an alpha/beta-scaffold	SMAD3 (1MHD)	
(8) Beta-Sheet binding to DNA	TBP (1CDW)	
(9) Beta-Barrel DNA-binding domains	YB-1 (1H95)	
(0) Yet undefined DNA-binding domains		

Tabelle 2.1: Superklassen in TFClass.

(TFClass ID:1.1). Die Beschreibung der konservierten DBDs dieser Klasse ist in Abbildung 2.14 (a) dargestellt. Sie umfasst acht Proteinfamilien, die anhand spezieller Ausprägungen dieser DBD unterscheidbar sind. Jede dieser Familien wird, falls möglich, weiter unterteilt und definiert somit die nächste Ebene der Klassifikation. So wird z.B. die Klasse *Jun-related* (TFClass ID 1.1.1, Abbildung 2.14 b) in drei weitere Unterfamilien aufgeteilt: Jun (TFClass ID: 1.1.1.1), NFE2 (TFClass ID: 1.1.1.2) und ATF2 (TFClass ID: 1.1.1.3). Die Grundlage dieser Unterscheidung auf der Subfamilienebene lässt sich gut durch das Ergebnis der phylogenetischen Analyse aller mammalischen DBDs nachvollziehen, welche in Abbildung 2.14 (b) dargestellt ist. Die Subfamilien 1.1.1.1 (rot), 1.1.1.2 (blau) und 1.1.1.3 (grün) zeigen eine größere Ähnlichkeit der Mitglieder innerhalb einer jeden Subfamilie als zu den Mitgliedern anderer Subfamilien. Die farbliche Kennzeichnung betont dieses Ergebnis und definiert damit die Einteilung dieser Gruppen in der TFClass Klassifikation. Eine Subfamilie ist nicht für alle in der TFClass-Hierarchie aufgelisteten DBDs definiert. Die letzten beiden Stufen dieser Einteilung werden durch die Begriffe *genera* und *species* definiert. Dabei beschreibt die *genera*-Ebene die Gene (Ort der Transkription) der in TFClass annotierten TFs. Die letzte Ebene (*species* = *molecular species*) beschreibt dann abschließend mögliche Isoformen (alternative Transkriptionsvarianten) ausgehend von einer Transkriptionseinheit (hier TF-Gen). Alle DNA-bindenden TFs werden folglich auf der fünften Ebene annotiert und die möglichen unterschiedlichen Genprodukte werden auf der sechsten Ebene unterschieden. Angewendet auf die im menschlichen Genom kodierten DNA-bindenden TFs bedeutet das, dass auf der Basis von 1558 TF-Genen 2904 unterschiedliche TFs auf der Protein-Ebene existieren (Wingender, Schoeps und Dönitz 2013). Die Größenordnung der einzelnen Superklassen ist nicht gleich verteilt. Die mit Abstand größte Gruppe der menschlichen DNA-bindenden TFs wird durch die Klasse der *Zinc-coordinating domains* gebildet (52 %). Danach folgenden die *Helix-turn-helix domain* enthaltenden TFs (27 %), gefolgt von den *Basic domain* Faktoren (11 %). Die restlichen zehn Prozent verteilen sich auf die verbleibenden sieben Superklassen inklusive der bisher nicht zugeordneten TFs (Wingender, Schoeps und Dönitz 2013). Wie bereits erwähnt interagieren verschiedene einzelne TFs miteinander. Die Dimerisierung von Transkriptionsfaktoren, also die Zusammenlagerung zweier Transkriptionsfaktoren zu einem funktionalen Paar, ist eine bekannte Fähigkeit von TFs. Diese Eigenschaft ist z.B. für Mitglieder der Klasse der *Helix-loop-helix* oder der *leucine zipper* bekannt. Dabei können sowohl Dimere gleichen Typs gebildet werden (Homodimere) als auch aus zwei verschiedenen TFs (Heterodimere). Die Bildung von Homo- und Heterodimeren kann mit einer unterschiedlichen Sequenzspe-

zifität verbunden sein. Ein bekanntes Beispiel für die Homo- und Heterodimerisierung ist der Transkriptionsfaktor *Activator protein 1* (AP-1) (*basic domain* definierter TF). Dieser Faktor kann z.B. als Homodimer (JUN:JUN) oder als Heterodimer (z.B. FOS:JUN) gebildet werden. Es existieren jedoch keine FOS:FOS Homodimere. Sowohl der JUN:JUN- als auch der FOS:JUN-Komplex binden an die klassische AP-1 Bindestelle: `TGA[C/G]TCA`. Die Affinität des Heterodimers an eine AP-1 Bindestelle ist aber im Vergleich zum Homodimer größer (John et al. 1996; Nakabeppu et al. 1988). Abbildung 2.15 zeigt das bekannte Dimerisierungsnetzwerk der *bZIP* Faktoren (TFCLASS ID: 1.1).

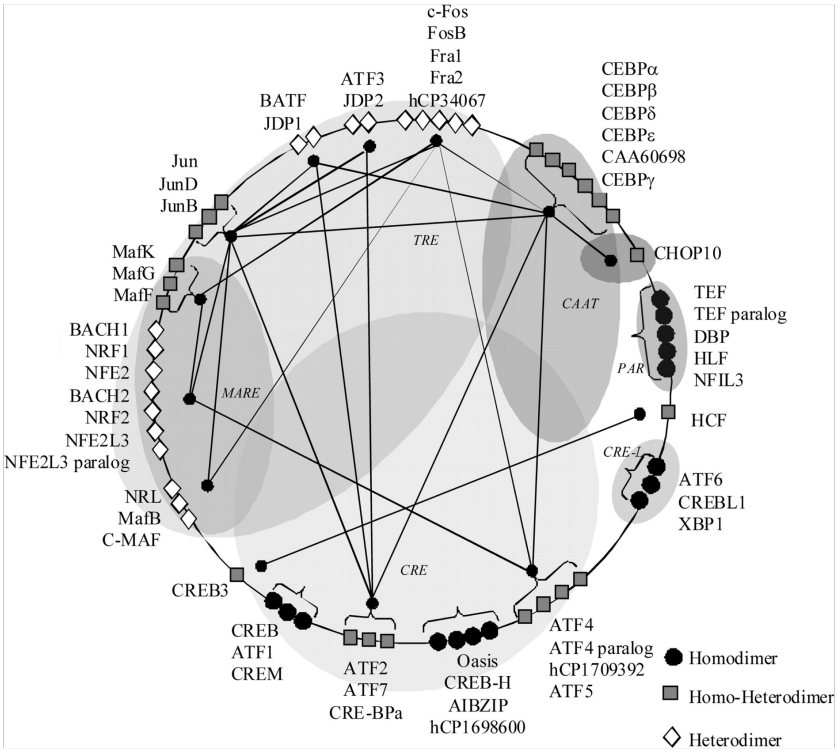


Abbildung 2.15: Dimerisierungsnetzwerk der menschlichen bZIP-Transkriptionsfaktoren (Deppmann et al. 2006). Die Knoten dieses Netzwerkes repräsentieren einzelne TFs dieser Klasse (schwarz: Homodimere, grau: Homo-/Heterodimere, weiß: ausschließlich Heterodimere). Die Klammern stellen bZIP-Faktoren dar, welche mit sich selbst oder anderen Paare ausbilden (dimerisieren). Diese sind in der Abbildung als Verbindungen dargestellt. Die verschiedenen graumarkierten Ellipsen deuten die Verwendung gemeinsamer TFBSs an.

Jeder der in diesem Netzwerk dargestellten Knoten entspricht einem TF dieser Klasse. Ei-

ne Verbindung zweier TFs beschreibt die «Heterodimerisierungsfähigkeit» zu einer unterscheidbaren *Leucine zipper*-Untergruppe dieser Klasse. Die unterschiedlichen Knotenformen der *bZIP*-Faktoren unterscheiden dabei die unterschiedlichen Rollen der TF-Knoten. Dabei beschreiben die schwarzen, runden Knoten solche TFs, die ausschließlich Homodimere bilden. Die grau dargestellten, quadratischen Knoten repräsentieren solche TFs, die sowohl Homo- als auch Heterodimere innerhalb dieser Klasse bilden können, während die weißen TF-Knoten in Form einer Raute die Gruppe der ausschließlich heterodimerisierenden TFs repräsentieren. Die transparenten Ellipsen deuten die Ähnlichkeit der dargestellten Gruppen auf Basis ihrer Sequenzspezifitäten (Erkennungssequenz der TF auf der DNA) an. Der regulatorische Einfluss der entsprechenden TF-Dimere muss nicht immer ein positiver sein. Ein Beispiel dafür stellt der menschliche Transkriptionsfaktor MYOD1 dar (UniProt ID: P15172, TFCLASS ID: 1.2.2.1.1): Das Homodimer dieses Proteins bindet sequenzspezifisch an die DNA. Bildet MYOD1 allerdings ein Heterodimer mit einem Protein namens ID1 (UniProt: P41134), so kann dieser Komplex nicht mehr mit der DNA interagieren (Benezra et al. 1990). Die Bildung dieses Komplexes erzeugt also nicht-funktionale (nicht-DNA-bindende) Proteinkomplexe. Die Komplexbildung ist im Allgemeinen nicht nur auf Paare von TFs beschränkt.

2.7 Modellierung der Sequenzspezifität von Transkriptionsfaktoren

Die Interaktion der TFs mit der DNA erfolgt durch komplexe Wechselwirkungen spezifischer und unspezifischer Nukleotid-Aminosäurekontakte. Die Grundfunktionalität eines TFs kann am besten mit der Erkennung der durch die Nukleotidabfolge geprägten DNA-Struktur beschrieben werden (Mathelier, Xin et al. 2016). Die Strukturerkennung erfolgt auf der Seite des TFs durch eine spezielle Funktionseinheit: der DNA-bindenden Domäne. Diese Region ist ein funktional und strukturell abgrenzbarer Bereich eines TF-Proteins, durch welchen die meist sequenzspezifische Erkennung der DNA-Sequenz erfolgt. Die DNA-Sequenzen, welche durch die TFs erkannt werden, werden als TFBSs, häufig nur Bindestellen genannt, bezeichnet. Es existieren eine Vielzahl verschiedener experimenteller Verfahren, die es erlauben, die spezifisch erkannte DNA-Sequenz eines TFs aufzuklä-

ren. *DNase-Footprinting* (Galas und Schmitz 1978) bzw. die *SELEX*-Technologie (Ellington und Szostak 1990; Oliphant und Struhl 1989; Tuerk und Gold 1990) sind nur zwei bekannte Beispiele, welche in diesem Zusammenhang häufig genutzt werden. Die genomweite Bestimmung der verwendeten TFBSs eines TFs wurde in den letzten Jahren vor allem durch die ChIP-seq Technologie vorangebracht (Johnson et al. 2007). Die z. B. TFs sowie die genutzten regulatorischen Regionen können durch die Verwendung eines Antikörpers genomweit identifiziert werden. Durch die Isolation und Aufreinigung dieser gebundenen Fragmente und eine anschließende Sequenzierung erhält man als Ergebnis eine Liste genomischer Bereiche (Lokalisationen), mit denen der untersuchte TF interagiert hat. Die durchschnittliche Fragmentlänge der durch die ChIP-seq Technologie gefundenen Regionen beträgt mehrere Hundert bp. Eine nachfolgende bioinformatische Untersuchung ist notwendig, um aus dieser Sequenzmenge die eigentliche Menge an statistisch angereicherten Bindestellen zu identifizieren, welche mit dem verwendeten TF direkt oder indirekt in Verbindung stehen. Eine ChIP-seq Region ist durch eine zentrale Region gekennzeichnet, dem sogenannten ChIP-seq *Peak* (deutsch: Höchstwert, Gipfelpunkt). Dieser Punkt wird bei der Auswertung eines ChIP-seq Experimentes für jedes Fragment berechnet und repräsentiert die experimentell identifizierte Region, an welcher das untersuchte Protein am häufigsten gefunden wurde. Im Falle von sequenzspezifisch bindenden TFs sollte in diesem Bereich die potentielle TFBS zu finden sein.

TFs zeigen in der Regel unterschiedliche Präferenzen gegenüber ihren Zielsequenzen. Diese können in verschiedenartiger Form modelliert werden. Eine einfache Beschreibungsform in diesem Zusammenhang ist die Konsensus-Sequenz, welche häufig auch Sequenzmotiv genannt wird: Dabei wird auf der Basis einer alignierten (gruppierten) Sequenzmenge von TFBSs (Alphabet: $\sum_{DNA} = \{A,C,G,T\}$) das häufigste Nukleotid pro Position bestimmt (siehe Abbildung 2.16). Auf diese Weise wird eine einfache Zusammenfassung aller Bindestellen eines TFs bereitgestellt, welche in dieser abstrakten Form in biologischen Sequenzen nur selten vorkommt (D'haeseleer 2006). Eine genauere Beschreibung der Sequenzmotiveigenschaften eines TFs auf der Basis seiner Menge an TFBSs ist die PFM. Diese Matrix modelliert die Sequenzeigenschaften des TFs in einer positionsspezifischen Art und Weise (siehe Abbildung 2.16). Eine PFM kombiniert dabei gleichzeitig die Bedeutung einer Position und die Bestimmtheit/Eindeutigkeit dieser Position auf Basis des verwendeten Alphabets. Bei der Beschreibung von in der DNA vorliegenden Motiven umfasst das Alphabet die bekannten vier Nukleotide: $\sum_{DNA} = \{A,C,G,T\}$. Auf Basis einer

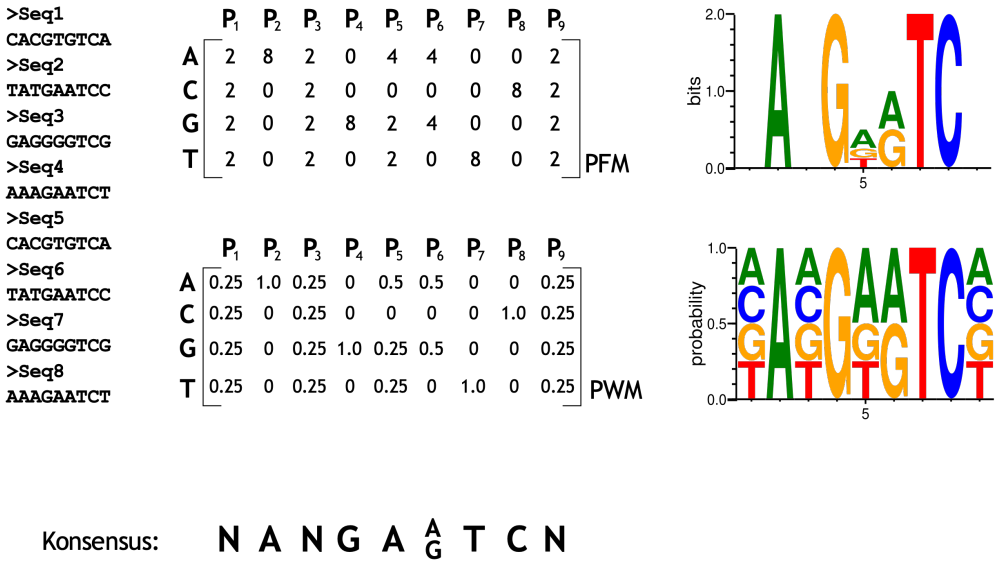


Abbildung 2.16: Beschreibung der Sequenzspezifität von Transkriptionsfaktoren. Die acht Zufallssequenzen (linker Bildausschnitt) können durch eine positionsgenaue Überführung der Sequenzinformation in eine *Postion Frequency Matrix* (PFM) bzw. *Positional Weight Matrix* (PWM) übersetzt werden (siehe Mitte). Diese Information kann durch sogenannte Sequenzlogos visualisiert werden (rechter Bildausschnitt).

PFM kann nun eine PWM erzeugt werden. Das Kennzeichen der PWM (alternativ auch Sequenzprofil genannt) ist, dass die positionsspezifischen Häufigkeiten in relative Häufigkeiten überführt werden. Die Einzelwahrscheinlichkeiten pro Position addieren sich also zu eins auf. Die wohl bekannteste graphische Repräsentation einer PWM ist das sogenannte Sequenzlogo (Schneider und Stephens 1990). Die konservierten Sequenzeigenschaften eines Sequenzmotivs werden darin sehr gut verdeutlicht. Die Abbildung 2.16 zeigt ein Sequenzlogo auf Basis von acht Zufallssequenzen. Die Sequenzlogos werden in der Literatur häufig in zwei verschiedenen Darstellungsarten gezeigt: (a) auf einem der Informationstheorie angelehnten Maß (Maßeinheit bit) und (b) auf Basis der relativen Häufigkeiten (siehe Abbildung 2.16, rechter Bildausschnitt). Die Berechnung des Informationsgehaltes einer jeden Position in einem Sequenzprofil wird mit der folgenden Formel berechnet: $I_j = 2 + \sum_{x \in \{A,C,G,T\}} f_{(x,j)} \; log_2 \; (f_{(x,j)})$. Mit Hilfe dieser Berechnungsvorschrift wird die Bedeutung einer jeden Position in einer PWM abgeschätzt: Existiert eine Gleichverteilung der Nukleotide an einer Position j einer PWM $f_{(A,j)} = f_{(C,j)} = f_{(G,j)} = f_{(T,j)} = 0.25$,

so ergibt sich keine Präferenz dieser Position und der Informationsgehalt dieser Position beträgt 0 bits. Falls jedoch eine Position nur durch ein Nukleotid geprägt wird, diese Position also komplett konserviert ist, ergibt sich ein Informationsgehalt von 2 bits (Beispiel: $f_{(A,j)} = 1,\ f_{(C,j)} = 0,\ f_{(G,j)} = 0,\ f_{(T,j)} = 0$). Für die Berechnung des Informationsgehalts wird bei Auftreten der 0 in der PWM bei der Berechnung des $log_2(0)$ der Wert 0 eingesetzt. Für jede Position einer PWM kann mit Hilfe der oben angegebenen Formel nun der Anteil jedes beliebigen Nukleotids berechnet werden. In einem Sequenzlogo werden diese dann jeweils einzeln graphisch dargestellt. Die Buchstabengröße ergibt sich in der Darstellung relativ zum Anteil am Gesamtinformationsgehalt der jeweiligen Position (siehe Abbildung 2.16). Häufig werden bei der Darstellung bzw. Verwendung dieser PWMs auch sogenannte *Pseudocounts P* verwendet. Das bedeutet, dass auf jede Position der PFM ein bestimmter Betrag aufaddiert wird. In einer Veröffentlichung von G. Stormo (2013) wird ein Wert P=1 für den *Pseudoscore* vorgeschlagen, Wasserman und Sandelin (2004) verwenden einen Wert $P = \sqrt{N}$, wobei N durch die Anzahl der verwendeten TFBSs der Matrix definiert ist und Ambrosini et al. (2018) verwenden einen Wert P=0.001. TFs und ihre zugeordneten TFBSs werden in verschiedenen Datenbanken gesammelt. Häufig stehen diese Bindestellensammlungen als PWMs-Bibliotheken zur Verfügung. Diese PWMs beschreiben damit also bekannte TF-DNA Interaktionen. Die bekanntesten Datenbanken auf diesem Gebiet sind TRANSFAC (Matys, Kel-Margoulis et al. 2006), JASPAR (Mathelier, Fornes et al. 2016), HT-SELEX (Jolma, Yan et al. 2013), UniPROBE (Hume et al. 2015), CisBP (Weirauch et al. 2014) und GRTD/HOCOMOCO (Kulakovskiy et al. 2018; Yevshin et al. 2019). Alle genannten Sammlungen basieren auf experimentell bestimmte Bindestellen.

2.8 Computergestützte Vorhersage potentieller TFBS

Die Vorhersage von TFBSs ist eine häufig verwendete bioinformatische Anwendung bei der Analyse und Interpretation von Genexpressionsdaten. Durch die Weiterentwicklung der Sequenzierungstechnologien und die damit verbundene Kostenreduktion ist die Analyse des gesamten Transkriptoms (Menge aller transkribierten Gene/Transkriptionseinheiten von Genen) mittlerweile zu einer Art Standardlabormethode geworden. Als eine der populärsten Methoden in diesem Zusammenhang hat sich die RNA-seq Methode entwickelt

(Bainbridge et al. 2006; Mortazavi et al. 2008; Wilhelm et al. 2008). Diese Methode erlaubt es, die Gesamtheit aller vorliegenden RNA-Sequenzen quantitativ zu bestimmen. Die vorliegenden RNA-Sequenzen werden molekularbiologisch in stabilere cDNA-Sequenzen übersetzt und anschließend sequenziert. Die Bestimmung von auffällig stark transkribierten Genen ist durch die Unterstützung und Weiterentwicklung verschiedener bioinformatischer und statistischer Verfahren auf Basis dieser RNA-seq Daten möglich geworden. Auf dieser Datengrundlage lassen sich so z. B. krankheitsrelevante Gene detektieren oder auf Basis von Zeitreihen spezifisch transkribierte Gene bestimmen, welche für die Entwicklung verschiedener Zellen, Gewebe oder Organe von Bedeutung sind. Der Vergleich dieser Daten mit krankhaft veränderten Situationen ist die Grundlage vieler medizinischer Fragestellungen und möglicher Therapien. Die Interpretation der vorliegenden spezifisch transkribierten Menge an Genen erfolgt im Anschluss an diesen Analyseschritt.

Grundsätzlich existieren an dieser Stelle zwei unterschiedliche Fragestellungen. Die erste Frage lautet: Welche biologischen Prozesse sind mit der vorliegenden Menge an Transkripten bzw. deren Genprodukten (Proteine, regulatorische RNA, etc.) verbunden? Zu diesem Thema existiert eine Reihe verschiedener statistischer Verfahren, die auf Basis vorliegender biologischer Netzwerke oder biologisch definierter Kategorien statistisch signifikante Prozesse identifizieren. Die biologischen Prozesse, mit denen die in einem Experiment gefundenen exprimierten Gene assoziiert sind, werden als bedeutende Prozesse der untersuchten biologischen oder medizinischen Fragestellung aufgefasst. Die Methoden, welche bei dieser Fragestellung Verwendung finden, können möglicherweise am besten mit dem Begriff der Funktionsanalyse zusammengefasst und beschrieben werden (engl. *functional enrichment analysis*) (D. W. Huang et al. 2009). Die zweite Fragestellung lautet: Welche regulatorischen Einflüsse existieren, die das Vorhandensein des vorliegenden Transkriptoms (Menge aller aktiven Gene) bzw. der daraus abgeleiteten Menge an auffällig exprimierten Transkripten erklären? Diese Fragestellung zielt ab auf das vorhandene regulatorische Netzwerk, welches das gemeinsame Auftreten der exprimierten Gene erklärt. So ist man z. B. an einer Gruppe von Transkripten (gesamte Menge oder auch Teilmenge) interessiert, die gemeinsam durch eine definierte Menge an TFs reguliert wird. Das koordinierte Zusammenspiel verschiedener TFs liefert also die Grundlage für diese Analyse (Davidson 2001). Wie bereits erwähnt, besitzen diese Faktoren den entscheidenden Einfluss auf die Transkriptionsstärke eines Gens (siehe oben). In der Regel binden verschiedene TFs sequenzspezifisch an ihre jeweiligen TFBS. Durch lokale und globale Interaktionen mit anderen

TFs stellt diese Proteinklasse so situationsspezifisch die benötigte Transkriptionsstärke sicher. Gene bzw. deren Genprodukte, die eine gemeinsame biologische Funktion besitzen, zeigen auf Ebene der Transkription häufig eine koordinierte, also abgestimmte, Transkriptionsregulation (Davidson 2006). Auf Basis dieser biologischen Beobachtungen existieren verschiedene Verfahren zur Bindestellen-Vorhersage und deren Anwendung bei der Interpretation von Genexpressionsdaten. Ebenso sind unterschiedliche Verfahren bekannt, welche die statistische Anreicherung potenzieller TFBSs bewerten können. Die Gemeinsamkeit aller existierenden Verfahren besteht immer in einer sequenzbasierten Analyse, welche die regulatorischen Regionen der zu untersuchenden Gene als Grundlage benutzt. Die Menge der zu untersuchenden Gene wird z.B. durch die Gruppe der auffällig exprimierten Gene gebildet. Durch die Bindestellenvorhersage kann dann anschließend direkt oder indirekt auf die dort bindenden TFs geschlossen werden. Existieren keine weiteren Angaben, wird im Allgemeinen eine 500-1000 bp lange Region um den TSS eines Gens als regulatorische Region verwendet (400-900 bp stromaufwärts, 100 bp stromabwärts). Ein Vorteil der RNA-seq Technologie ist in diesem Zusammenhang, dass unterschiedliche Transkripte pro Gen beobachtet und dabei möglicherweise verschiedene TSSs und damit unterschiedliche Promotoren für ein Gen berücksichtigt werden können.

Die Vorhersage von potentiellen TFBSs erfolgt durch sogenannte Muster- oder Motiverkennungsalgorithmen. Grundsätzlich beschränkt sich das Auffinden von Sequenzmotiven nicht nur auf die Analyse von DNA- oder RNA-Sequenzen, sondern kann auch bei der Motivfindung in Aminosäuresequenzen angewendet werden. Funktionale Motive in Aminosäuresequenzen werden auch als Proteindomänen bezeichnet. Die Motivvorhersage unterscheidet zwei verschiedene algorithmische Verfahren: Die erste Gruppe von Algorithmen versucht, ohne Vorkenntnisse angereicherte Sequenzmotive definierter Länge in einer Eingabemenge aufzufinden (Stichwort: Mustererkennung, engl. *pattern recognition*). Eine zweite Gruppe von Algorithmen benutzt bereits bekannte Sequenzprofile, welche z. B. in Form von PWMs zur Verfügung stehen (mit zugeordneten TFs) und untersucht die Sequenzen der Eingabemenge nach ähnlichen Bindestellenmotiven (Stichwort: Mustervergleich, engl. *pattern matching*).
Im Bereich der Mustererkennung ist ein Verfahren bekannt geworden, welches auf Basis des *Expectation Maximization* (EM) Algorithmus Sequenzmotive mit fester Wortlänge auf Grundlage von nicht alignierten Sequenzen lernen kann (Lawrence und Reilly 1990). Der grundlegende Gedanke hinter diesem Verfahren ist der folgende: Jede Sequenz einer Ein-

gabemenge enthalte die Instanz eines gemeinsamen Motivs, welche eine Bindestelle für einen potentiellen TF darstelle (siehe auch Abbildung 2.17). Lawrence und Reilly (1990) erzeugen in ihrem Verfahren durch die Wahl zufälliger Startpunkte von Teilwörtern fester Länge für jede Sequenz aus der Eingabemenge eine PWM. Die Qualität dieser zu Beginn zufälligen PWM als beste Beschreibungsform eines möglichen Sequenzmotivs für alle Sequenzen wird anschließend überprüft. Dazu wird in jeder vorhandenen Sequenz die beste Startposition dieser Matrix identifiziert. Um diese zu finden, wird jedes Teilwort einer vorliegenden Sequenz mit Hilfe der PWM bewertet und die Position bzw. das Teilwort, an dem eine gegebene Bewertungsfunktion S den größten Wert aufweist, legt nun den neuen Startpunkt des Sequenzmotivs fest (siehe Abbildung 2.17 a). Die bisherige PWM wird durch diese Teilwörter neu definiert und die Neubewertung dieser PWM startet (iteratives Verfahren). Die Bewertung der vorliegenden PWM erfolgt mit Hilfe des *pattern-matching*-Verfahrens (siehe unten). Nach einer gewissen Anzahl an Wiederholungen, verbunden mit sehr geringer Veränderung einer Bewertungsfunktion S über die Anzahl der durchgeführten Iterationen (Konvergenz des Wertes S), wird die so berechnete PWM als Endergebnis ausgegeben. Durch die Wahl verschiedener Startpunkte können durch den EM-Algorithmus alternative Sequenzmotive bestimmt werden. Diese finalen PWMs stellen also einzelne Modelle eines gemeinsamen Sequenzmotivs auf Basis der Eingabemenge dar. Für jede PWM wird eine Gesamtbewertung angegeben, welche die maximale logarithmierte Wahrscheinlichkeit (engl. *log likelihood*) dieses Sequenzmotivs für einen Programmdurchlauf repräsentiert. Aus diesem Grund wird dieser Wert auch *Maximum Likelihood* (ML) *Score* genannt und die Literatur benennt das Verfahren als EM Algorithmus. Grundsätzlich garantiert dieses EM Verfahren aber nicht, das globale Maximum zu finden (Dempster et al. 1977; Lawrence, Altschul et al. 1993; Lawrence und Reilly 1990). Die Wahl unterschiedlicher Startpunkte und die anschließende Auswahl der verschiedenen Modelle mit dem größten ML Wert wird von verschiedenen Autoren empfohlen, um diesem globalen Maximum möglichst nahe zu kommen. T. L. Bailey und Elkan (1994) schlagen in ihrem Verfahren (MEME) eine andere Wahl der Startpunkte bzw. die Definition der initialen PWM vor: Für jede Sequenz der Eingabemenge wird durch das Aufzählen aller Teilwörter (fester Länge) eine sequenzspezifische PWM erzeugt. Diese PWM kann nun verwendet werden, um die beste Startposition auf der Grundlage derselben in der jeweiligen Sequenz zu identifizieren (sequenzspezifische Vorauswahl, siehe Abbildung 2.17 b). Diese Startpunkte bzw. die daraus abgeleiteten Teilwörter definieren die initialen Worte für jede Sequenz, welche nun für die Definition der initialen PWM verwendet und durch die wiederholte Anwendung

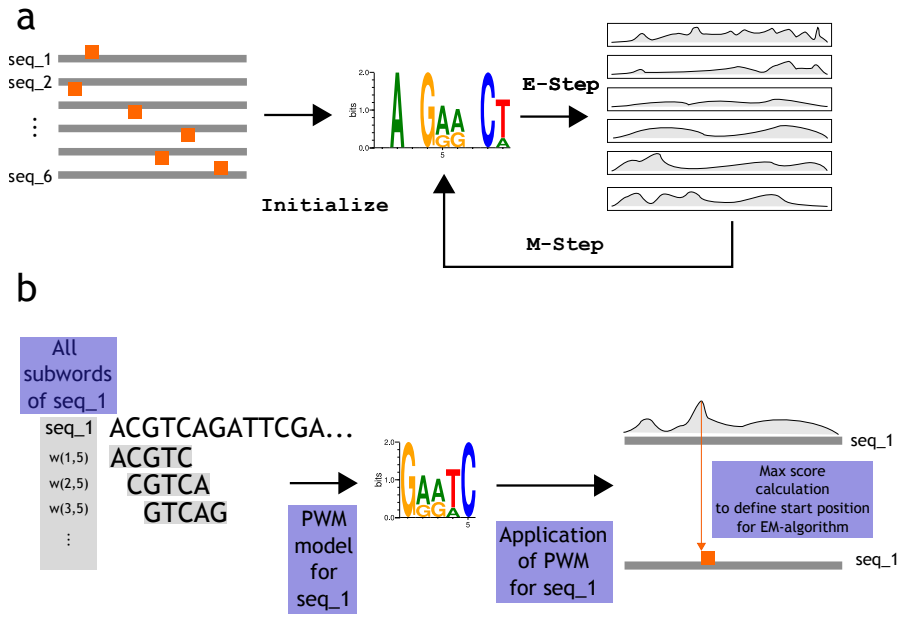


Abbildung 2.17: *Expectation Maximization* (EM) Algorithmus. Abbildung (a) zeigt schematisch den Ablauf des EM-Algorithmus. Zu Beginn wird die Startposition eines Sequenzmotivs fester Länge zufällig für jede Sequenz der Eingabemenge bestimmt (orange markierter Bereich: mögliches Motiv auf Plus- bzw. Minusstrang). Diese Startpositionen definieren eine initiale PWM, welche nun für die Eingabemenge bewertet (*E-step*) und in einem iterativen Verfahren optimiert wird (*M-step*). In jeder Iterationsstufe wird überprüft, inwieweit die Höchstbewertung der aktuellen PWM mit einer anderen Startposition verknüpft ist. Sollte eine solche Abweichung auftreten, wird das bisher verwendete Teilwort der PWM der Teilsequenz durch das höher bewertete Teilwort ersetzt und eine neue Gesamtbewertung eingeleitet. Eine mögliche Erweiterung des EM-Algorithmus ist in (b) dargestellt. T. L. Bailey und Elkan (1994) schlagen in ihrem als *MEME* benannten Verfahren eine Sequenz-spezifische Vorverarbeitung für jede Teilsequenz der Eingabemenge vor, um aussagekräftigere Sequenzmotive zu erhalten. In der Darstellung ist diese Vorverarbeitung für eine Beispielsequenz gezeigt. Auf Grundlage aller Teilworte fester Länge (hier $k = 5$) wird eine sequenzspezifische PWM erstellt, welche nun für die Bestimmung der Startposition in der Ausgangssequenz verwendet wird. Die ähnlichste Sequenz (höchste Bewertung) zur vorliegenden PWM wird verwendet und dient zur Definition der Startposition dieser Sequenz. Diese Einzelbestimmungen werden für alle Teilsequenzen der Eingabemenge durchgeführt und liefern so die Startpositionen für den eigentlichen EM-Algorithmus.

des EM-Algorithmus optimiert wird. Das erste Verfahren (zufällige Wahl der Startpunkte) ist unter dem Namen *Gibbs Sampling* bekannt geworden (Lawrence, Altschul et al. 1993; Lawrence und Reilly 1990). Die Anwendung dieses Algorithmus, unter der Voraussetzung, dass alle Sequenzen der Eingabemenge eine Instanz des gefundenen Sequenzmotivs zeigen sollten, hat einige Limitationen. Diese sind sowohl algorithmischer als auch biologischer Natur. Der *Multiple EM for Motif Elicitation* (MEME) Algorithmus versucht diese Probleme zu berücksichtigen (T. L. Bailey und Elkan 1994). Durch die Wahl unterschiedlicher Modi kann das Fehlen von Motiven oder die Gruppierung von Motiven (potentieller TFBS-Cluster) in einzelnen Sequenzen berücksichtigt werden. Verschiedene Varianten des Gibbs-Sampling-Verfahrens sind in der Zwischenzeit veröffentlicht worden. Eine Übersicht über die Leistungsfähigkeit dieser Werkzeuge und verwandter Verfahren wurde im Jahr 2005 vorgenommen (Tompa et al. 2005).

Die zweite große Klasse von Algorithmen zur Vorhersage potentieller TFBS wird durch die Gruppe der Sequenzvergleichsalgorithmen gebildet. Ausgangspunkt der Verfahren dieser Klasse ist eine Sammlung von experimentellen TFBS, die durch ein Sequenzalignment in eine positionsspezifische Motivbeschreibung überführt wurden. Diese Positionsabhängigkeiten können, wie bereits gezeigt wurde, als PWMs modelliert werden (siehe Abbildung 2.16). Eine derartige PWM kann nun benutzt werden, um in einer vorliegenden DNA-Sequenz nach Bindestellen zu suchen, die diesem Sequenzprofil ähneln. Diese Sequenzähnlichkeit kann über folgende Formel berechnet werden:

$$I_{seq}(i) = \sum_{b=\{A,C,G,T\}} f_{b,i} \cdot log_2 \frac{f_{b,i}}{p_b}$$

Dabei gibt $f_{b,i}$ die relative Häufigkeit des Nukleotids b ($\sum_{DNA=\{A,C,G,T\}}$) an der Position i an und p_b ist die Hintergrundwahrscheinlichkeit des betrachteten Nukleotids (G. Stormo 2013). Häufig wird bei der Berechnung dieses Wertes eine Gleichverteilung der vier Nukleotide angenommen. Alternativ dazu kann der Sequenzhintergrund aber auch aus dem genomischen GC/AT-Gehalt oder auf Grundlage der zu untersuchenden Sequenzen bestimmt werden. Der Term $log_2 \frac{f_{b,i}}{p_b}$ ist auch als *log-odds* Matrix bekannt. Der Einfluss des Sequenzhintergrunds auf den *log-odds Score* mit und ohne Sequenzhintergrund im Vergleich mit und ohne *Pseudocount* ist in Abbildung 2.18 gezeigt. Die vier Matrizen auf der rechten Seite dieser Abbildung zeigen die *log-odds* Matrizen (LO) für die Beispiel PWM,

die als Matrize und als Sequenzlogo gezeigt ist. Aus dieser PFM wurden zwei verschiedene PWMs berechnet (Bildmitte): Die obere PWM beschreibt die relativen Häufigkeiten, während die untere durch *Pseudocounts* ($P = 1$ pro Position) erweitert wurde. Die

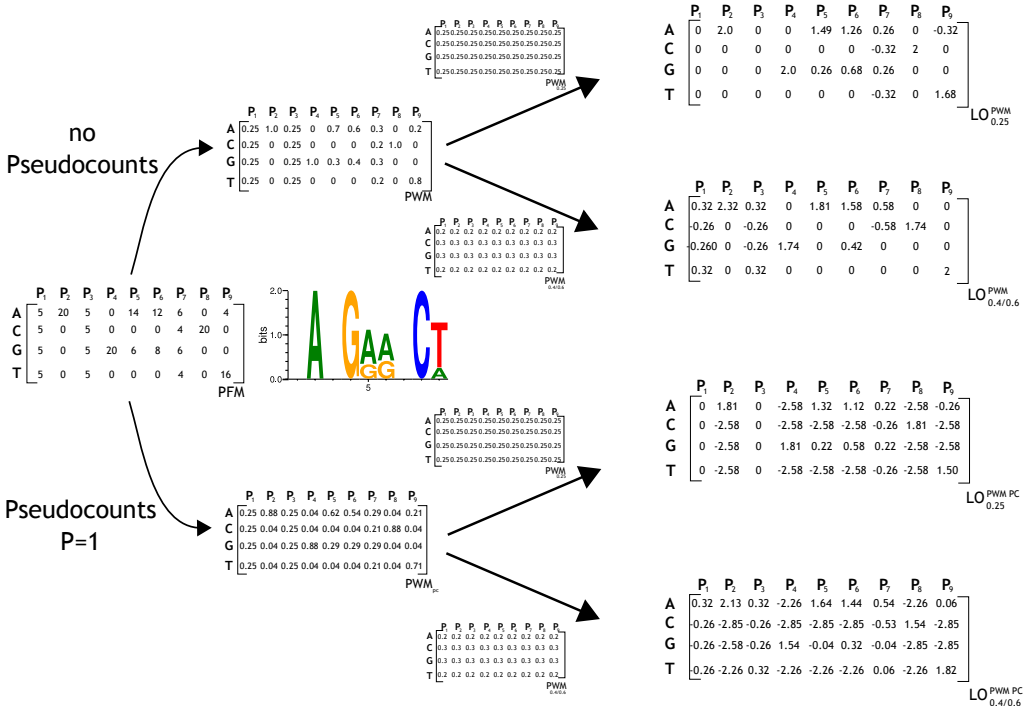


Abbildung 2.18: Einfluss des Sequenzhintergrunds auf die entsprechenden log-odds-Matrizen. Die vier Matrizen $LO_{0.25}$, $LO_{0.2/0.3}$, $LO_{PC\_0.25}$ und $LO_{PC\_0.2/0.3}$ ergeben sich auf Grundlage der Verwendung eines unterschiedlichen Sequenzhintergrunds (0.25 = Gleichverteilung, 0.2/0.3 AT- bzw. CG-Gehalt von 40:60). Ausgangspunkt ist die Matrize auf der rechten Seite, welche einmal mit (oben) und einmal ohne (unten) Verwendung von Pseudoscounts berechnet wird.

vier Matrizen $LO_{0.25}$, $LO_{0.2/0.3}$, $LO_{PC_0.25}$ und $LO_{PC_0.2/03}$ zeigen die jeweiligen log-odds Matrizen auf Basis dieser zwei verschiedenen PWMs (siehe Abbildung 2.18 rechts). Die PWM ohne und mit Pseudocounts wurde dazu jeweils durch zwei verschiedene Matrizen geteilt, welche einen unterschiedlichen Sequenzhintergrund modellieren (siehe Bildmitte, Pfeilmarkierungen): Die obere Matrize modelliert eine Gleichverteilung aller vier möglichen Nukleotide der DNA, während die untere Matrize ein AT-/GC-Verhältnis von 40:60 verwendet. Die entsprechenden log-odds-Matrizen zeigen den Einfluss dieser beiden unterschiedlichen Hintergrundmatrizen: Positionen der PWM, welche in AT-armen Regionen

ein A oder T in der PWM aufweisen, werden in ihrer log-odds-Bewertung verstärkt. Dies kann z.B. an Position zwei der PWM nachvollzogen werden. Beide Matrizen ($LO_{40/60}$ bzw. $LO_{PC40/60}$) zeigen im Vergleich zu ihrer jeweils gleich verteilten Situation eine Vergrößerung des *log-odds Scores*.

Die Berechnung der Sequenzähnlichkeit wird in verschiedenen Veröffentlichungen auch als relative Entropie bezeichnet. In der Informationstheorie ist diese Formel als Kullback-Leibler-Divergenz bekannt (G. D. Stormo 2000). Bei der Vorhersage von TFBSs in regulatorischen DNA-Sequenzen ist die Tatsache zu berücksichtigen, dass potentielle Bindestellen auf beiden Strängen der DNA vorliegen können. Das bedeutet in der Praxis, dass sowohl der Plus- als auch der Minusstrang einer DNA-Sequenz nach potentiellen TFBSs untersucht werden muss. Abbildung 2.18 zeigt dies für eine Beispielmatrix. Der rechte

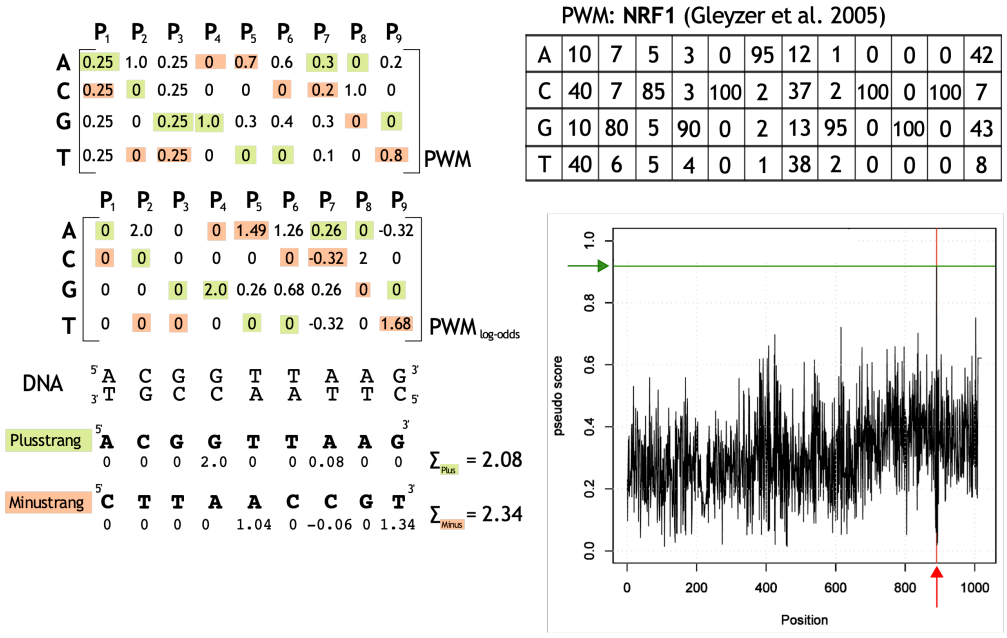


Abbildung 2.19: Beispielanwendung der Sequenzähnlichkeitsberechnung. Auf Basis der Beispiel-PWM und der daraus abgeleiteten log-odds-Matrize (Gleichverteilung der vier Nukleotide) kann mit Hilfe der Kullback-Leibler-Divergenz ein Ähnlichkeitswert berechnet werden. Da eine DNA-Sequenz einen Doppelstrang besitzt und potenzielle Bindestellen auf beiden Strängen zu finden sind, existieren pro Position zwei mögliche Bindestellen-Vorhersagen. Auf der rechten Seite der Abbildung ist die Anwendung dieses Bewertungsschemas für eine experimentell verifizierte Bindestelle gezeigt (Gleyzer et al. 2005). Die mit rotem Pfeil markierte Position entspricht der im Experiment nachgewiesenen TFBS.

Teil der Abbildung demonstriert die Aussagekraft der Bindestellenvorhersage auf Basis einer bekannten PWM. Als Beispiel wird hier der menschliche Transkriptionsfaktor NRF1 mit seinen in der Literatur beschriebenen TFBS verwendet (Gleyzer et al. 2005). Die in der Abbildung dargestellte Analyse zeigt die Bewertung der menschlichen Promotorregion des *TFBM1*-Gens (-1000 bp stromaufwärts des TSS). Es wird nur jeweils die beste Bewertung pro Position im Vergleich des Plus- bzw. Minusstrangs gezeigt. Gleyzer et al. (2005) zeigen in ihrer Veröffentlichung, dass der Transkriptionsfaktor NRF1 in der Promotorregion dieses Gens bindet. Die durch den rot markierten Pfeil gekennzeichnete Position verdeutlicht die am höchsten bewerte Position dieses Promotors durch die verwendete NRF1-Matrix, welche ebenfalls der Veröffentlichung entnommen wurde (siehe Abbildung 2.19). Sie entspricht exakt der in der Publikation angegebenen Bindestelle. Der grüne Pfeil markiert die Höhe der berechneten Bewertungen (engl. *Scores*), welche Min-Max-normalisiert wurde. Bei der Erstellung der zugrundeliegenden PWM wurde ein Pseudoscore $P = 1$ verwendet und mit Hilfe der Kullback-Leibler-Divergenz der beste Score pro Position berechnet. Als Hintergrundwahrscheinlichkeit wurde die Gleichverteilung aller vier Nukleotide angenommen (siehe oben). Wie bereits erwähnt, modellieren die PWMs die Präferenz der Nukleotide, welche für die Bindung des TFs von Bedeutung sind, in einer positionsunabhängigen Art und Weise. Für einige Faktoren sind aber Abhängigkeiten zwischen den Positionen beschrieben (Bulyk et al. 2002; Jolma, Yan et al. 2013). Die unterschiedliche Konfiguration eines TF-Proteins kann auch mit unterschiedlichen Motiven verbunden sein (Badis et al. 2009; Rohs et al. 2009). Des Weiteren existieren Beispiele, die abbilden, dass die Kooperation von TFs mit veränderten Motivbeschreibungen korreliert (Jolma, Yin et al. 2015). Außerdem zeigen Methylierungen spezifische Veränderungen von TFBS einiger TF (Yin et al. 2017). Aus diesem Grund sind in den letzten Jahren komplexere Modelle und Verfahren entwickelt worden, um eine verbesserte Genauigkeit bei der Bindestellenvorhersage zu erreichen. Eine Auswahl an veröffentlichten Verfahren ist in Tabelle 2.2 aufgelistet.

Für einzelne TFs bzw. Transkriptionsfaktorfamilien konnte so eine Verbesserung gezeigt werden. Die Modelle verwenden z. B. Dinukleotidmatrizen, um die Abhängigkeit von benachbarten Positionen in Bindestellen zu modellieren, oder sie verwenden Wörter gegebener Länge k (sogenannte k-mere), um eine weitreichende, abstraktere Positionsabhängigkeit zu modellieren. In einem Vergleich von Weirauch et al. (2014) konnte allerdings kein großer Vorteil dieser Methoden gegenüber den klassischen Sequenzanalyseverfah-

Modell	Literaturreferenz
PWM	(T. L. Bailey und Elkan 1994; Kel et al. 2003; Lawrence und Reilly 1990; Lenhard und Wasserman 2002; Linhart et al. 2008; Luehr et al. 2012; Pavesi et al. 2001; Quandt et al. 1995; Roth et al. 1998; G. D. Stormo 2000; Ward und Bussemaker 2008)
Dinukleotidmatrix (DWM)	(Siddharthan 2010; Zhao et al. 2012)
Baysche Netzwerk	(Ben-Gal et al. 2005)
Markov Modell	(Eggeling et al. 2017; Grau et al. 2013; Siebert und Söding 2016)
Wort-basierte (k-mere)	(Annala et al. 2011; Kähärä und Lähdesmäki 2013)
Markov Netzwerk	(Keilwagen et al. 2011; Sharon et al. 2008)
Neuronales Netzwerk	(Alipanahi et al. 2015; Avsec et al. 2021; J. Zhou und Troyanskaya 2015; Q. Zhou und Liu 2008)
Random Forrest	(Chen et al. 2007; Hooghe et al. 2012)
Support Vector Machine (SVM)	(Agius et al. 2010; Arvey et al. 2012; Gordân et al. 2013; Ma et al. 2017)
Thermodynamisches Verfahren	(Djordjevic et al. 2003)
Strukturbasiertes Verfahren	(Maienschein-Cline et al. 2012; Mathelier, Fornes et al. 2016; L. Yang et al. 2014)

Tabelle 2.2: Modelle zur Vorhersage potenzieller TFBS.

ren auf Basis der PWM festgestellt werden, insbesondere, wenn verschiedene Datensätze in die Untersuchung einbezogen wurden. Insgesamt betrachtet scheinen die existierenden Modelle die vielfältigen Einflüsse auf die transkriptionelle Genregulation noch nicht gut genug abzubilden (Slattery et al. 2014). Auf Grundlage dieser Beobachtungen verwendet diese Arbeit das Modell der PWM für die Vorhersage von potentiellen TFBS.

3 Material und Methoden

3.1 TRANSFAC

TRANSFAC ist eine Datenbank, welche eukaryotische regulatorische DNA Elemente (TFBSs) und deren interagierende TFs beschreibt. Die erste Version dieser Datenbank ist im Jahr 1988 veröffentlicht worden (E. Wingender 1988). Jeder Eintrag in der *TRANSFAC SITE Table* basiert auf einem experimentellen Nachweis einer sequenzspezifischen DNA-Interaktion eines bekannten TFs. Falls möglich, werden weitere Informationen verknüpft. So werden z.B. die regulierte Transkriptionseinheit (Gen), die Spezies sowie weitere bekannte experimentelle Informationen gespeichert.

Falls für einen TF bzw. eine TF-Familie eine ausreichende Menge an TFBSs in TRANSFAC vorhanden ist, werden diese durch ein definiertes Verfahren zusammengefasst und als positionsspezifische Mononukleotid Matrize PWM gespeichert. Diese PWMs können für die Vorhersage potenzieller TFBSs verwendet werden (siehe Kapitel 2). Es sind aber auch publizierte PWMs in der TRANSFAC-Datenbank gespeichert. Mit fast 10000 verschiedenen PWMs (Stand 2019.2) stellt die TRANSFAC-Datenbank eine der umfangreichsten Sammlungen auf diesem Gebiet dar. Es werden PWMs für Insekten, Pflanzen, Pilze und Wirbeltiere unterschieden.

Da in den verschiedenen Projekten dieser Arbeit menschliche genregulatorische Fragestellungen im Vordergrund standen, wurde hier aus der TRANSFAC-Datenbank die Teilmenge der Wirbeltier-PWMs (Vertebraten) genutzt; unterschiedliche Versionen der TRANSFAC-Datenbank wurden in den verschiedenen Projekten zugrunde gelegt. Die nachfolgende Tabelle 3.1 listet die einzelnen Versionen und deren Projektbezüge auf.

TRANSFAC	Projekt-Beschreibung	Kapitel
2014.4	SNP-Analyse SP-1	4.1
2014.4	SNP-Analyse LHX4	4.1
2009.4	Regulatorisches TF-Netzwerk (1. Version)	4.2
2013.1	Regulatorisches TF-Netzwerk (Aktuelle Version)	4.2
2013.3	Enhancer-Promotor-Projekt	4.3, 4.4

Tabelle 3.1: Auflistung der verwendeten TRANSFAC-Versionen.

3.2 MATCH

MATCH ist ein Programm zur Vorhersage potenzieller TFBSs auf der Grundlage von definierten PWMs (Modell: positionsspezifische Mononukleotid-Matrizen). Es ist Teil der TRANSFAC Datenbank und kann direkt mit den dort gespeicherten PWMs verwendet werden. Ausgangspunkt einer Bindestellenanalyse sind eine oder mehrere DNA-Sequenzen. Als gültige Sequenzformate werden dabei die beiden textbasierten Formate FASTA bzw. EMBL unterstützt. MATCH verwendet für die Vorhersage eine sogenannte Profil-Datei *profile.prf*. Diese listet alle für die jeweilige Analyse zu verwendenden PWMs und deren Schwellenwerte auf (siehe Beispiel: Tabelle 3.2).

```
Profile
test.prf
MIN_LENGTH 300
0.0
1.000000 0.678 0.778 M00001 V$MAT1
1.000000 1.000 0.885 M00002 V$MAT2
...
//
```

Tabelle 3.2: Beispiel-Datei eines MATCH-Profils.

Die ersten beiden Zeilen einer Profil-Datei bezeichnen einen frei wählbaren Text gefolgt vom Dateinamen der Profil-Datei (siehe Tabelle 3.2, Zeilen 1 und 2). Die folgenden beiden Zeilen sind vorgegeben, sie werden für eine fehlerfreie Ausführung des MATCH Programms benötigt. Ab der fünften Zeile werden die für die Analyse zu verwendenden PWMs aufgeführt. Das Ende dieser Angaben wird durch das Zeichen «//»definiert.

Eine PWM wird immer durch die Angabe von fünf verschiedenen Werten näher bestimmt (Angaben durch Leerzeichen getrennt). Die erste Angabe gibt den Wertebereich einer MATCH-Vorhersage an (Anzahl der Nachkommastellen). Der zweite und dritte Parameter definieren zwei verschiedene Schwellenwertsituationen, welche für jede verwendete PWM individuell angegeben werden müssen: Der erste Schwellenwert (*core-similarity cut-off*) bezieht sich auf den *Core Similarity Score* (CSS) für die fünf aufeinanderfolgenden und stark sequenz-konservierten Positionen einer PWM. Für eine finale TFBS Vorhersage durch MATCH muss dieser Schwellenwert überschritten werden. Ist diese Bedingung erfüllt, wird anschließend der MATCH *Score* (*Matrix Similarity Score* (MSS)) für die gesamte Bindestelle berechnet. Die Gesamtlänge ist durch die untersuchte PWM vorgegeben. Übersteigt diese Berechnung nun die zweite Schwellenwert-Angabe (*matrix-similarity cut-off*), wird diese potenzielle Bindestelle in einer textbasierten Ergebnisdatei aufgeführt. Diese Datei beschreibt alle potenziellen TFBSs einer jeden PWM aus der Profil-Datei für jede DNA-Sequenz der Eingabe-Datei. Es werden für jede Vorhersage die Startposition der Vorhersage in der jeweiligen Sequenz, der CSS- sowie der MSS-Wert und der Strang-Bezug aufgeführt. Alle Bindestellenvorhersagen werden für eine DNA-Sequenz der Eingabemenge sortiert ausgegeben.

In dieser Arbeit findet der CSS Wert keine Verwendung. Das bedeutet, dass in den verschiedenen Projekten nur die Gesamtbewertung einer Bindestellenvorhersage berücksichtig wird. Für einige TFs bzw. deren Motivbeschreibung ist die Berechnung des CSS unpassend: Gerade für einige längere Motive/PWMs, welche ein in sich wiederholendes Muster aufweisen, kann die eindeutige Bestimmung einer Kernregion nur ungenau erfolgen. Als Konsequenz daraus könnten einige potenzielle TFBSs übersehen werden. Die Tabelle 3.3 zeigt das Berechnungsverfahren, welches im MATCH-Algorithmus angewendet wird (Kel et al. 2003). Durch die Anwendung der Min-Max-Normalisierung wird der finale MSS Wert in den Wertebereich [0,1] überführt (siehe Auflistung 3.3). Die Anwendung des Informationsvektors (engl. *Information vector*) ist eine Besonderheit des MATCH-Algorithmus. Der Anwendung dieser Berechnungsvorschrift sorgt für eine größere Toleranz von sogenannten *Mismatches* in nicht stark ausgeprägten sequenzkonservierten Positionen einer PWM im Vergleich zu stark konservierten Positionenf (Kel et al. 2003). Damit unterscheidet sich dieses Berechnungsverfahren von der klassischen Log-Odds-Score Berechnung bei Gleichverteilung der Nukleotide (siehe auch Kapitel Grundlagen).

Matrix-Score	$MSS = \frac{(Cur-Min)}{(Max-Min)}$
Current Score	$Cur = \sum_{i=1}^{L} I(i) \cdot f_{i,b_i}$
Minimal Score	$Min = \sum_{i=1}^{L} I(i) \cdot f_{i,b_i}^{min}$
Maximal Score	$Max = \sum_{i=1}^{L} I(i) \cdot f_{i,b_i}^{max}$
Information Vector	$I(i) = \sum_{B\varepsilon\{A,C,G,T\}} f_{i,B} \cdot ln(4 f_{i,B})$

Tabelle 3.3: Bewertung einer potenziellen TFBS durch den MATCH-Algorithmus.

Wie bereits erwähnt wurde, ist das MATCH-Programm eng verknüpft mit der TRANSFAC-Datenbank: Dort werden die eigentlichen PWMs aufgelistet, die in dieser Anwendung benutzt werden. Die anzuwendenden Schwellenwerte werden in unterschiedlichen Qualitätsstufen als verschiedene Profil-Dateien mitgeliefert. In der Standardanwendung von MATCH werden drei verschiedene Qualitätsstufen verwendet: das Profil mit der höchsten Schwellenwertdefinition für jede PWM wird in MATCH/TRANSFAC als sogenanntes minFP-Profile aufgeführt. In dieser Einstellung wird versucht, den Anteil der falsch-positiven Vorhersagen zu minimieren. Als unterste Stufe findet das sogenannte minFN-Profil Verwendung. Dieser Einstellung hat zum Ziel, durch Minimierung der falsch-negativen Ergebnisse möglichst viele potenzielle Bindestellen zu erkennen. Als Kompromiss zwischen diesen beiden Einstellungen wird das sogenannte minSUM-Profil aufgeführt. Die Schwellenwerteinstellungen dieses Profils versuchen, die Extreme beider Profile auszugleichen (Kel et al. 2003). Die Wahl der Schwellenwerte in den einzelnen Projekten sind in den einzelnen Ergebniskapiteln angegeben.

3.3 MEME

MEME ist ein Programm zur Vorhersage von Motiven, welche wiederholend in einer Menge von Nukleotid- oder Aminosäuresequenzen, ohne Verwendung des sogenannten

Lückensymbols (Gap: −), vorkommen. Als textbasiertes Eingabeformat der Sequenzen wird das FASTA-Format unterstützt. Als Motiv ist dabei ein Wort mit einer definierten Länge zu verstehen (Alphabet abhängig von der Eingabemenge), das annähernd gleich in der Eingabemenge aufzufinden ist. MEME repräsentiert dabei ein Motiv in einer positionsspezifischen Art und Weise. Dabei wird jede Position eines Motivs aus den Einzelwahrscheinlichkeiten der Symbole dieser Position gebildet und als PWM repräsentiert (siehe auch Kapitel 2). Durch verschiedene statistische Teilschritte wird die optimale Länge sowie deren Häufigkeit pro Sequenz bzw. innerhalb aller Sequenzen für jedes Motiv bestimmt (T. L. Bailey und Elkan 1994). Die Wortlänge, die Gesamtanzahl und die Häufigkeit pro Sequenz eines Motivs können durch die Wahl verschiedener Parameter voreingestellt werden. Die Analyse von potenziellen TFBSs erfolgt auf der Grundlage von DNA Sequenzen. Als Ergebnis listet diese Anwendung die verschiedenen gefundenen Sequenzmotive in Form von PWMs auf und bewertet sie mit Hilfe eines sogenannten P- bzw. des E-Wertes. In dieser Arbeit wurde MEME in der Version 4.10.0 verwendet. Für die relevanten ChIP-seq bzw. DNase-seq Daten, die durch ihre chromosomalen Positionen definiert sind, wurden diese genomischen Lokalisationen in die entsprechenden DNA-Sequenzen überführt (FASTA-Format). Für die MEME Analysen fanden die Parametereinstellungen 'DNA' und 'revcomp' Anwendung. Der Parameter 'DNA' zeigt an, dass die Analyse auf der Grundlage von DNA-Sequenzen erfolgen soll. Die zweite Angabe legt fest, dass beide DNA-Stränge untersucht werden sollen. Um die Laufzeit des Verfahrens zu beschränken, werden bei der Motivsuche durch MEME maximal zehn verschiedene Motive betrachtet (Angabe: -motive 10), welche eine maximale Länge von 15 bp nicht überschreiten sollen (Angabe: -maxw 15). Weiterhin wurde bei den einzelnen MEME Untersuchungen die Einstellung 'zoops' gewählt. Durch diese Angabe wird die Möglichkeit des Fehlens eines Motivs in einzelnen Sequenzen toleriert.

3.4 UCSC

Die Universität von Kalifornien Santa Cruz (UCSC) hat in den vergangenen Jahren eine vielfältige Internetressource aufgebaut, welche die Visualisierung und Bearbeitung von experimentell erzeugten genomweiten Daten auf Basis von zuvor sequenzierten Genomen erlaubt. Diese Ressource wurde in den verschiedenen Projekten dieser Arbeit zur Darstel-

lung, zur Kontrolle der eigenen Implementierungen und zur Datenbeschaffung verwendet. Unter der URL https://genome.ucsc.edu/ kann diese Ressource gefunden werden. Das Visualisierungswerkzeug dieses Projekts (Genom-Browser/-Navigator) ist neben dem ENSEMBL Projekt (siehe Abschnitt 3.5) eines der bekanntesten Softwareprojekte in diesem Themenfeld. Die übersichtliche Darstellung genomischer Information ist eine wesentliche Aufgabe dieses Softwareprojekts. Außerdem erlaubt dieses System in einfacher Art und Weise verschiedene genomweite Datensätze komfortabel dem existierenden Internetbasierten System hinzuzufügen und einheitlich darzustellen. So können auch Daten des ENCODE Projektes gezielt über dieses System angezeigt werden, aber auch eigene Daten sind durch einfache Import-Möglichkeiten in dieser Ressource darstellbar. Als Basis des Systems dient eine relationale Datenbank, welche auch unabhängig von der Darstellung im Genom-Browser z.B. per SQL-Schnittstelle angefragt werden kann. Alternativ zu einer SQL-gestützten Datenabfrage können diese Abfragen auch durch eine eigene Webseite erfolgen. Auch diese Abfragemöglichkeit wurde in dieser Arbeit verwendet. Die URL der verwendeten Webseite lautet: https://genome.ucsc.edu/cgi-bin/hgTables.

3.5 ENSEMBL

Das ENSEMBL Projekt ist ein Softwaresystem, welches, ähnlich zum UCSC Werkzeug, versucht, eine übersichtliche und effiziente Darstellung von genomischer Information zu erreichen (Zerbino et al. 2018). Eigene sowie auch verschiedene frei verfügbarer Projekte (z.B. die des ENCODE Projektes) können in dieser Ressource dargestellt werden. Die Grundlage des Projektes bildet ebenfalls ein relationales Datenbanksystem, welches auch unabhängig von der Visualisierung z.B. per SQL angefragt werden kann. Zusätzlich steht eine auf der Programmiersprache Perl basierte Schnittstelle zur Verfügung. Unter folgender URL kann diese Ressource gefunden werden: https://www.ensembl.org. Es existiert aber auch eine in der Programmiersprache R implementierte Zugriffsmöglichkeit. Das biomaRt Paket (Durinck et al. 2009) wurde in den verschiedenen Teilprojekten dieser Arbeit verwendet, welches auf verschiedene Teilbereiche der ENSEMBL Datenbank aufbaut.

3.6 Projekt-bezogene ENCODE Daten dieser Arbeit

Das ENCODE-Projekt ist ein Forschungsprojekt, welches eine Vielzahl von funktionalen Elementen im menschlichen Genom charakterisiert. Ziel des ENCODE-Projekts ist sowohl die Entwicklung/Weiterentwicklung diverser Methoden, welche die genomweite Bestimmung verschiedener funktionaler Regionen erlauben, als auch die Anwendung dieser Methoden auf z.B. definierte menschliche Zellen. Die Veröffentlichung der Daten des ENCODE-Projekts erfolgte zu Beginn über den Web-Server des UCSCs (siehe Abschnitt 3.4). Mittlerweile werden die Daten durch einen eigenen Web-Service veröffentlicht. Die URL dieser Ressource lautet: https://www.encodeproject.org.

Alle in den Kapiteln 4.2, 4.3 und 4.4 verwendeten Datensätze wurden durch das ENCODE Projekt erzeugt und sind frei verfügbar, sie wurden in dieser Arbeit durch die UCSC Web-Ressource bezogen. Die verschiedenen experimentellen Datensätze beziehen sich alle auf die menschliche Referenzsequenz in der Version GRCh37 (hg19). In den beiden Ergebnisteilen wurden ChIP-seq und DNase-seq Daten der Zelllinien HUVEC, HeLa S3, K562 und GM12878 verwendet. Für alle Datensätze wurden die im ENCODE-Projekt angefertigten, zusammengefassten und normalisierten Daten verwendet, welche als textbasierte BED Dateien bezogen wurden (Format: narrowPeak). Als allgemeine Ressource für die Speicherung und Verwaltung verschiedener experimenteller Hochdurchsatzdaten wird die GEO-Datenbank angesehen (Barrett et al. 2013). Aus diesem Grund werden im Folgenden die in dieser Arbeit verwendeten Datensätze durch ihre von der GEO-Datenbank definierten Zugangskennungen angegeben.

Für die HUVEC Zelllinie wurde der ChIP-seq Datensatz für den c-Fos Antikörper verwendet (GEO: GSM935585) und DNase-seq Daten benutzt (GEO: GSM816646). Die Analysen der HeLa S3 Zellen wurden auf der Basis der ENCODE-Datensätze (ChIP-seq) für c-Fos (GEO: GSM935317) und NF-YB (GEO: GSM935408) sowie einen DNase-seq Datensatz (GEO: GSM816643) durchgeführt. Für K562 Zellen wurden die ChIP-Seq Daten für c-Fos, NF-YB und DNase-seq Daten mit den folgenden GEO Zugangsnummern untersucht: GSM93535355, GSM935429 und GSM816655. Die GM12878-Daten entstammen ChIP-seq Daten für c-Fos (GEO: GSM935409), NF-YB (GEO: GSM935507) und DNase-seq (GEO: GSM816665).

3.7 Ref-Seq

Die RefSeq-Datenbank ist eine frei verfügbare Datenbank, welche am NCBI (engl. *National Center of Biotechnology Information*) gepflegt wird. Sie speichert für verschiedene Organismen auf der Grundlage von öffentlichen Nukleotidsequenzen (DNA/RNA) eine nicht redundante Sammlung aller genomischen DNA, genomkodierter RNA und daraus abgeleitete Proteinsequenzen (Pruitt et al. 2007). Diese Sammlung wird aus allen übermittelten Sequenzen einer Spezies der GenBank-Datenbank erzeugt. Für eine speziesspezifische Sammlung in RefSeq gilt: es werden alternative Transkripte annotiert. Im Falle von proteinkodierenden Genen bedeutet das zum Beispiel, dass möglichereweise unterschiedliche Transkripte pro Gen auftreten, welche für alternative Proteine kodieren können (sogenannte Isoformen). Die RefSeq-Datensammlung des Menschen wurde als Ausgangsbasis für die Rekonstruktion bzw. Vorhersage des allgemeinen transkriptionsregulatorischen Netzwerks in dieser Arbeit verwendet (siehe Ergebnisteil 4.2). Durch RefSeq steht ein beschreibender Datensatz aller menschlichen Gene und deren alternativer Transkripte zur Verfügung. Für jedes einzelne Transkript ist in dieser Datenbank der sogenannte Transkriptionsstart verzeichnet. Ausgehend von dieser Angabe wurde nun für die Erstellung des Transkriptionsnetzwerks die menschliche Promotorregion definiert. Abhängig von der Lokalisation der Transkriptionseinheit (Plus- bzw. Minusstrang) wurde dazu der -1000 bp umfassende Bereich stromaufwärts des TSS berechnet. Für jede einzelne Transkriptionseinheit wird die Promotorregion so festgelegt und für die weiteren Analysen verwendet. Durch die UCSC Ressource ist dazu der folgende RefSeq-definierte Datensatz für den Menschen bezogen worden: refGene; (Apr. 14, 2010, hg19).

3.8 UniGene

Die UniGene-Datenbank ist eine Datenbank des NCBI. Diese Ressource beschreibt für verschiedene Spezies auf Basis von gewebespezifischen *Expressed Sequence Tags* (EST)-Clustern und exprimierten mRNA Sequenzen aktiv transkribierte Gene für eine Gruppe der jeweiligen Spezies. Um das allgemeine regulatorische Transkriptionsnetzwerk dieser Arbeit für eine Gruppe menschlicher Gewebe genauer untersuchen zu können, wurde die-

se Ressource verwendet. Das Web-Interface des NCBI erlaubt die Abfrage transkribierter Gene für z.B. verschiedene Gewebe/Organe des Menschen. Das Ergebnis dieser Anfrage kann als textbasierte Datei gespeichert werden. Die einzelnen Dateien wurden mit Hilfe des biomaRt Pakets auf die Verwendung des offiziellen HGNC-definierten Gennamens (Braschi et al. 2019) hin überprüft (siehe Abschnitt 3.5). Das allgemeine Transkriptionsnetzwerk verwendet ebenfalls diese HGNC-definierten Gennamen, so dass durch eine Filterung über den Gennamen die entsprechenden gewebe- bzw. organ-spezifischen Transkriptionsnetzwerke direkt vorliegen. Diese Filterung wurde mit Hilfe der R-Programmierumgebung durchgeführt (siehe Abschnitt 3.12). Die Gewebe- bzw. Organ-spezifischen Genlisten wurden am 18.06.2012 von der UniGene-Datenbank über das NCBI Webinterface bezogen.

3.9 ChIP-seq

ChIP-seq ist eine Methode zur Bestimmung definierter Protein-DNA Interaktionen im genomischen Maßstab. Sie kombiniert die Chromatin-Immunopräzipitation mit der anschließenden DNA-Sequenzierung und wurde im Jahre 2007 von verschiedenen Laboren veröffentlicht (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). Durch den Einsatz von Formaldehyd können DNA-gebundene Proteine reversibel kovalent mit dieser verknüpft werden (*cross linking*). Abschließend wird die genomische DNA z.B. mit Ultraschall oder Nukleasen fragmentiert. Mit Hilfe der Immunopräzipitation werden nun aus der Menge aller vorliegenden Protein-DNA Komplexe durch die Verwendung eines für ein bestimmtes DNA-bindendes-Protein spezifischen Antikörpers nur die Protein-DNA-Fragmente isoliert, welche von diesem Antikörper spezifisch erkannt werden. Diese Fragmente können nun durch einen nachfolgenden Reinigungsschritt von den übrigen Protein-DNA-Komplexen getrennt werden. Abschließend erfolgt eine thermische Freisetzung der durch den Antikörper isolierten Protein-DNA-Fragmente, welche dann in einem weiteren Schritt sequenziert werden. Eine anschließende bioinformatische Auswertung dieser Sequenzierungsdaten liefert final die genomischen Positionen, an denen die Protein-spezifischen Interaktionen gefunden worden sind. Die Fragmente eines ChIP-seq Experiments haben eine durchschnittliche Länge von mehreren hundert Basenpaaren. Dabei wird jeder Region ein experimenteller Qualitätswert (engl. *Peak Score*) und eine sogenannte *Peak* Position zugeordnet. Diese Position beschreibt die Örtlichkeit, an der das zu untersu-

chende Protein am häufigsten gefunden worden ist. In den letzten Jahren wurde eine große Menge verschiedener sogenannter *Peak Calling* Algorithmen veröffentlicht (Nakato und Shirahige 2017). Ein unabhängiger Vergleich der Methoden erscheint schwierig, da ein allgemeingültiger Testdatensatz nicht existiert. Da im ENCODE-Projekt standardisierte Protokolle verwendet werden, welche sowohl die experimentelle als auch die nachgeschaltete bioinformatische Analyse einbeziehen, wurden in dieser Arbeit die vom ENCODE-Projekt erzeugten Daten verwendet (BED-Format, narrowPeak-Daten).

3.10 DNase-seq

DNase-seq ist ein experimentelles Verfahren zur Bestimmung unspezifischer, potenziell regulatorisch aktiver Regionen im genomischen Maßstab. Die Methode verwendet das Enzym DNase-I. Dieses Enzym schneidet bevorzugt in frei-zugänglichen DNA-Bereichen, also solchen Regionen, die nicht durch kompaktes Chromatin geschützt sind. Diese Bereiche werden bevorzugt durch TFs gebunden. Außerhalb dieser durch Protein-DNA-Interaktionen geschützten Bereiche kann das DNase-I-Enzym den DNA-Doppelstrang unspezifisch schneiden. Damit charakterisiert die DNase-seq-Methode allgemein gültige regulatorische Regionen, welche im Gegensatz zur ChIP-seq-Technologie TF-unspezifisch sind (Boyle et al. 2008). Die geschnittenen genomischen DNA-Regionen werden extrahiert, die anschließende Sequenzierung dieser Regionen ergibt die genomischen Lokalisationen der potenziell regulatorisch aktiven Bereiche. Durch das experimentelle Protokoll bestehen die DNase-seq-Bibliotheken in der Regel aus mehreren Tausend gleichlangenr DNA-Fragmente. Die im ENCODE-Projekt vorliegenden Fragmente sind einheitlich 150 bp lang. Ein Genom kann in jeder Zelle unterschiedlich interpretiert werden. Das bedeutet, dass die regulatorischen Regionen sich von Zelltyp zu Zelltyp unterscheiden können (Tim R. Mercer et al. 2013).

3.11 Referenzgenom des Menschen

Ein Referenzgenom bezeichnet die bekannte, durch Sequenzierung eines einzelnen Individuums oder einer kleinen Anzahl von Individuen erzeugte, genomische Nukleotidabfolge. Über 180000 prokaryotische (ohne Zellkern) und mehr als 7000 eukaryotische Genome stehen mittlerweile als Referenzsequenz zur Verfügung und können z.B. über das NCBI bezogen werden. Einige Referenzgenome werden nach ihrer ersten Veröffentlichung kontinuierlich überarbeitet. Durch experimentelle und bioinformatische Nachbearbeitungen werden also fortschreitend die anfänglich noch ungenau bestimmten Bereiche besser untersucht und die Ergebnisse dann in Form der jeweils aktuellen Form an die Öffentlichkeit weitergegeben. Aus diesem Grund stehen verschiedene Versionen der menschlichen Genomsequenz beim NCBI als Download zur Verfügung. Bei der Analyse und Interpretation der verschiedenen genomweiten Daten ist darauf zu achten, mit welcher dieser unterschiedlichen Versionen gearbeitet wurde. In den verschiedenen Projekten dieser Arbeit wurde die Version hg19 (NCBI: GRCh37) verwendet.

3.12 Programmiersprache R

R ist eine freie Programmiersprache und Programmierumgebung, welche die statistische Datenverarbeitung und Visualisierung großer Datensätze unterstützt. Sie wurde von Ross Ihaka und Robert Gentleman entwickelt (Ihaka und Robert Gentleman 1996). R-Code wurde in allen Teilprojekten dieser Arbeit für die verschiedenen statistischen Berechnungen, zur Datengenerierung sowie zur Darstellung der berechneten Eigenschaften verwendet. Die finalen Berechnungen der AUROC-basierten Analyse (siehe Kapitel 4.3 und 4.4) und Darstellungen sind vollständig in R durchgeführt worden.

3.13 Programmiersprache Java

Die objektorientierte Programmiersprache Java, ursprünglich vom Unternehmen Sun Microsystems entwickelt, befindet sich mittlerweile im Eigentum der Firma Oracle. In dieser Arbeit wurde Java-Code in allen vier Teilprojekten entwickelt und angewendet. Durch die umfangreiche Java-API konnten auf diese Weise zeit- und speicherintensive Berechnungen durchgeführt werden. Zwei Projekte dieser Arbeit haben besonders durch verschiedene implementierte JAVA-Programme profitiert: die Vorhersage und Erstellung des regulatorischen Transkriptionsnetzwerks (Kapitel 4.2) und die AUROC-Analysen (Kapitel 4.3 bzw. 4.4). In beiden Projekten wurden große Datenmengen erzeugt und nachbearbeitet. Die anfallenden Aufgaben sind gleichzeitig, verteilt auf mehreren Rechnern, berechnet worden. Dadurch konnte die Berechnungs- und Bearbeitungszeit der Projekte massiv reduziert werden.

3.14 Programmiersprache Perl

Perl ist eine freie, plattformunabhängige Programmiersprache. Sie wurde von Larry Wall 1987 entwickelt und erlaubt eine sehr einfache und effektive Verarbeitung großer Textdateien. Diese Eigenschaft hat sie vor allem auf dem Gebiet der Bioinformatik sehr populär gemacht. Teile der *Single Nucleotide Polymorphism* (SNP) Analysen wurden in der Programmiersprache Perl durchgeführt (Kapitel 4.1). Um die Ergebnisse der MATCH-Vorhersagen in der Programmiersprache R weiterverarbeiten zu können, wurde ein einfaches Computerprogramm entwickelt. Das Text-basierte Ausgabeformat der verschiedenen MATCH-Analysen wurde durch die Anwendung dieses Programmes in ein tabellarisches Format überführt. Die finalen Berechnungen der beiden SNP-Analysen wurden dann anschließend in R weitergeführt (siehe Kapitel 4.1).

4 Ergebnisse

4.1 Bewertung pathogener Einzelnukleotidvariationen in regulatorischen Sequenzen

Die Entwicklung eines Verfahrens zur Analyse potenzieller regulatorisch wirksamer Einzelnukleotidvariationen (SNPs) steht in diesem Kapitel im Vordergrund. Die Anwendung dieses Verfahrens wird für zwei verschiedene SNPs vorgestellt, deren pathologische Wirkung durch klinische Studien belegt werden konnte. Die Ergebnisse des implementierten Verfahrens fanden in zwei verschiedenen Veröffentlichungen Berücksichtigung (Dalila et al. 2015; Schirmer et al. 2016).

4.1.1 Bedeutung regulatorischer Einzelnukleotidvariationen

Die mit Abstand häufigsten Sequenzvariationen im menschlichen Genom sind SNPs (Auton et al. 2015)). Ein SNP bezeichnet dabei eine Position im Genom, welche im Vergleich zu einer vorliegenden Referenzsequenz einer untersuchten Spezies in mindestens zwei Varianten existiert (Polymorphismus) und mit einer Häufigkeit von einem Prozent oder mehr in einer untersuchten Population vorkommt (Auton et al. 2015). Im Unterschied zu Punktmutationen werden SNPs stabil über mehrere Generationen vererbt. Statistisch beobachtet man alle 500 bis 1000 bp einen SNP. Durch genomweite Assoziationsstudien (*Genome-wide Association Studies* (GWASs)) können häufig Tausende verschiedene Regionen im Genom identifiziert werden, welche in der Summe kennzeichnend für eine bestimmte Ausprägung (Phänotyp) sind (Visscher et al. 2017).

In der Medizin können auf Grundlage dieser Assoziationsstudien und der Korrelation dieser Daten mit verschiedenen Patienten-bezogenen Daten z.B. krankheitsrelevante Variationen detektiert werden. Als sogenannte Marker können dabei sowohl Gene aber auch SNPs Verwendung finden (Claussnitzer et al. 2015; Kichaev et al. 2014; Pickrell 2014). Im Falle von SNP-basierten Analysen zeigt sich, dass häufig Sequenzvariationen in nichtkodierenden regulatorischen Regionen gefunden werden (Farh et al. 2015; Finucane et al. 2015; Maurano et al. 2012a).

Die Bewertung der regulatorischen Sequenzvariationen erfordert im Vergleich zu proteinkodierenden Regionen eine aufwendigere Analyse, da die Sequenzvariation nicht direkt analysiert und interpretiert werden kann. Um für diese Regionen mögliche sequenzspezifisch interagierende TFs zu bestimmen, muss eine SNP-Position mit einer potenziellen TFBS in Verbindung gebracht werden. Eine Sequenzvariation kann dabei eine existierende Bindestelle erzeugen und/oder eine existierende Bindestelle zerstören. Auf der Basis von PWMs und einer damit verbundenen Bewertungsfunktion können für diese regulatorischen SNPs potenzielle TFBSs vorhergesagt werden. Durch einen Vergleich der unterschiedlichen Bewertungen einer Bindestellenvorhersage auf Grundlage der vorliegenden Polymorphismen einer SNP-Position können so mögliche funktionale TFs für einen oder mehrere SNPs identifiziert werden. Anhand der beiden veröffentlichten Studienergebnisse soll das in dieser Arbeit entwickelte und angewendete Verfahren vorgestellt werden.

4.1.2 Analyse der regulatorischen Bedeutung der Sequenzvariation rs11644322

Im Jahr 2012 ist eine Array-basierte GWAS veröffentlicht worden, welche für eine Gruppe von 351 an Bauchspeicheldrüsenkrebs erkrankten Patienten auf der Basis von ungefähr 550.000 Einzelnukleotidvariationen pro Patient krankheitsrelevante SNPs bestimmt hat (Innocenti et al. 2012). Untersuchungsschwerpunkt dieser Studie war die Analyse des Krankheitsverlaufs der Patienten unter Anwendung einer Chemotherapie und die Bestimmung krankheitsrelevanter SNPs, welche den Therapieerfolg der Behandlung beeinflussen. Auf Basis der verfügbaren Daten sind zur Krankheit und der angewendeten Therapie korrelierte Einzelnukleotidvariationen berechnet worden. Alle Patienten dieser Studie sind

während der Chemotherapie mit dem Zytostatikum *Gemcitabin* behandelt worden (Goldstein et al. 2015). Das genomische Material (DNA) zur SNP-Analyse wurde aus dem Blut der Patienten extrahiert (periphere Blutzellen). Die Verteilung aller signifikanten SNPs auf die verschiedenen Chromosomen ist in Abbildung 4.1 (a) gezeigt und entstammt der Veröffentlichung von Innocenti und Kollegen (Innocenti et al. 2012).

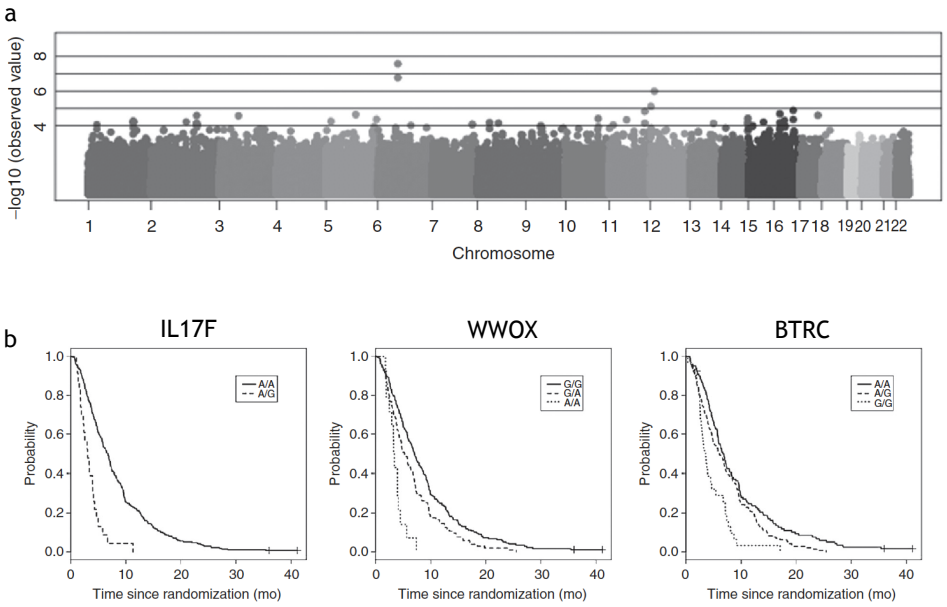


Abbildung 4.1: Krankheitsrelevante Gene des Bauchspeicheldrüsenkrebs. (a) Manhattan-Plot aller SNPs, welche mit der Überlebensrate der Studienteilnehmer assoziiert sind. Abbildung (b) zeigt die drei signifikantesten krankheitsrelevanten Gene der Studie mit den jeweils assoziierten Polymorphismen (Innocenti et al. 2012). Die Darstellung stellt die Überlebensrate der Patienten nach Studienaufnahme (*Time since randomization*) in Abhängigkeit der vorliegen Polymorphismen als Wahrscheinlichkeit für den Untersuchungszeitraum dar.

Die Überlebensrate (engl. overall survival, kurz OS) gibt den Zeitraum an, welcher ab der Diagnosestellung bzw. dem Therapiebeginn die Wahrscheinlichkeit des Überlebens repräsentiert. In Abbildung 4.1 wird der Zeitraum der Überlebensrate und deren Wahrscheinlichkeit auf die Aufnahme in die Studie (*Time since randomization*) bezogen. Mit Hilfe eines nach Kaplan und Meier vorgestellten Verfahrens kann diese Wahrscheinlichkeit geschätzt und dargestellt werden (Kaplan und Meier 1958). Abbildung 4.1 zeigt die auf der

Grundlage der Studie von Innocenti et al. (2012) beobachtete OS für die drei Gene IL17F (Interleukin 17-F, UniProt: Q96PD4), WWOX (WW domain-containing oxidore-ductase, UniProt: Q9NZC7), BTRC (F-Box/WD repeat containing protein 1A, UniProt: Q9Y297) im Zusammenhang der beobachteten SNPs dieser Gene. Die beobachteten Polymorphismen sind jeweils in der Legende angegeben. Diese Gene sind als sogenannte kennzeichnende Gene (engl. marker) klassifiziert worden, da die signifikanten SNPs (bedeutende Sequenzvariationen oder Allele), welche sich in dieser Untersuchung finden ließen, in den Transkriptionseinheiten dieser Gene lokalisiert sind. Abbildung 4.1 (b) präsentiert den Einfluss, welche die Sequenzvarianten und deren homozygotes (väterliches und mütterliches Allel sind gleich) bzw. heterozygotes Auftreten (unterschiedliche Ausprägung der beiden elterlichen Merkmale) auf die Überlebensrate besitzen. So zeigt z.B. die homozygote Ausprägung des SNPs rs11644322 (AA) eine geringere Überlebensrate als die heterozygote (AG)- bzw. die homozygote (GG)-Variante für das WWOX-Gen (siehe Abbildung 4.1 (b), mittlere Darstellung). Das Markergen IL17F mit dem dort auftretenden SNP rs763780 ist laut der Studie von Innocenti et al. (2012) klinisch bedeutend (signifikantes Testergebnis), während die dargestellten SNPs für die Gene WWOX und BTRC die gegebene Signifikanzschwelle nicht unterschreiten (Innocenti et al. 2012).

Die klinische Bedeutung dieser Beobachtungen ist in einer unabhängigen Studie in Göttingen von Schirmer et al. (2016) neu bewertet worden. Dazu wurden in einer Kooperation der Unikliniken Hamburg, Heidelberg und Göttingen drei unabhängig erfasste Patientengruppen (Kohorten) gemeinsam untersucht. Alle Patienten dieser zusammengefassten Studie sind ebenfalls an Bauchspeicheldrüsenkrebs erkrankt und mit dem Chemotherapeutikum *Gemcitabin* behandelt worden. *Gemcitabin* ist ein chemisches Analogon zum Nukleotid Cytosin. Durch den Einbau dieses Analogons in die DNA wird die DNA-Synthese unterbrochen und der Zelltod initiiert. Durch diese Behandlung sollen also die unkontrolliert wachsenden Krebszellen abgetötet werden. Aus den insgesamt 381 Patienten wurden durch verschiedene pathologische Untersuchungen vier unterschiedliche Teilgruppen definiert. Verschiedene molekularbiologische und medizinische Parameter wurden für diese Einteilung verwendet (Schirmer et al. 2016). Die einzelnen Gruppen repräsentieren damit unterschiedliche Ausprägungen/Schweregrade der Bauchspeicheldrüsenkrebserkrankung. Die OS-Kurven der vier Teilgruppen sind auf der Grundlage des SNP rs11644322 in Abhängigkeit der Überlebenszeit gezeigt (Abbildung 2, B-E). Abbildung 2 (A) zeigt die Situation aller Studienteilnehmer, während Abbildung 2 (F) die palliativ behandelten Patienten

zusammenfasst. Für alle Studienteilnehmer, aber auch die vier Subtypen, kann eine signifikante Korrelation dieses SNPs mit der OS gezeigt werden. Nur für die Gruppe der palliativ

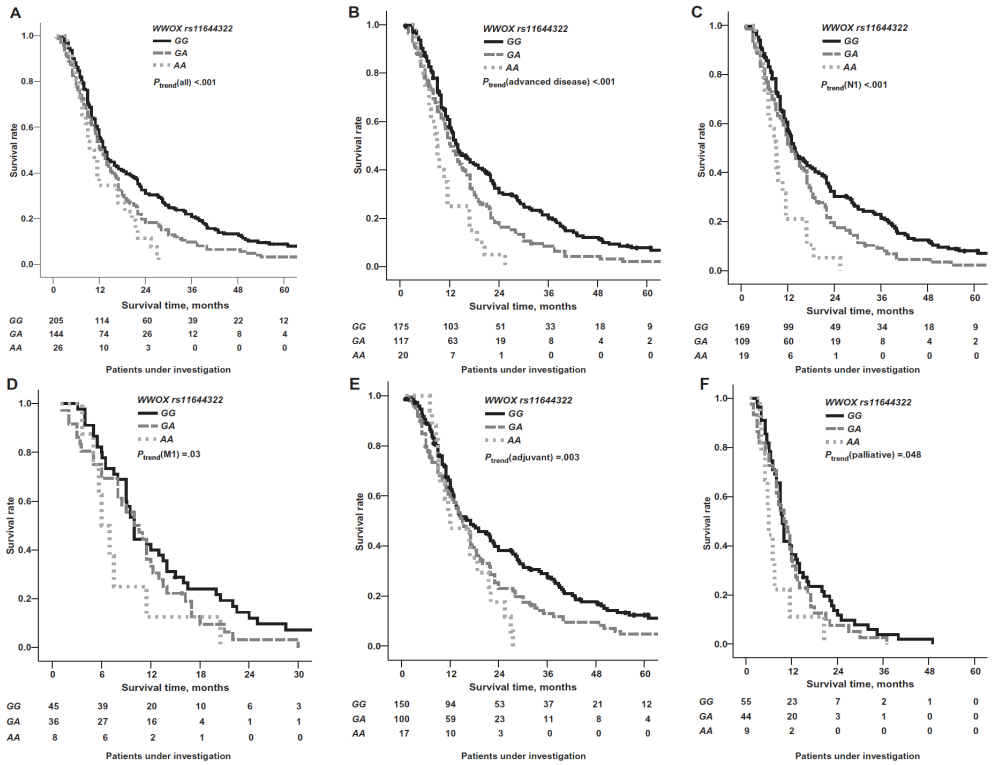


Abbildung 4.2: Pathologische Bedeutung des SNPs rs11644322 für das WWOX-Gen. Die Darstellung zeigt die Überlebensrate der 381 an Bauchspeicheldrüsenkrebs erkrankten Patienten für den Genlocus WWOX. Die Patientendaten stammen aus drei unabhängig erfassten Patientenkohorten der Universitätskliniken Hamburg, Heidelberg und Göttingen. Die Abbildung entstammt einer Veröffentlichung von Schirmer et al. (2016). Die folgenden Teilgruppen werden in der Abbildung unterschieden: (A) zeigt die gesamte Patientenkohorte, (B) stellte eine Untergruppe der fortgeschrittenen Kranheitsstadien dar, (C) ist die Teilgruppe an Patienten mit nachweisbaren Tumorzellen, (D) ist die Kohorte der Patienten, welche Metastasen in anderen Organen aufweisen, (E) zeigt die Patienten mit unterstützenden Therapiemaßnahmen und (F) die palliativ behandelten Patienten.

behandelten Patienten ist dieser signifikante Zusammenhang nicht gegeben (siehe Abbildung 4.2). Im Vergleich der drei beobachteten Merkmalsausprägungen wird aber auch für diesen Datensatz deutlich, dass über alle Gruppen hinweg die schlechteste Überlebensrate mit dem homozygot vorliegenden Merkmal (AA) assoziiert ist. Das AA-Allel zeigt sich in

26 Prozent der in Europa beheimateten sogenannten kaukasischen Population (Schirmer et al. 2016).

Die Genprodukte der WWOX-Transkriptionseinheit wirken als potenzielle Tumorsuppressoren. Die Genprodukte dieser Gene kontrollieren den Zellzyklus oder sind Regulatoren der Apoptose. Alle drei beobachteten Merkmalsausprägungen (GG, GA, AA) zeigen eine insgesamt geringe Lebenserwartung. Der Median der OS für das GG-Allel beträgt 14 Monate. Für eine heterozygote Situation (GA) beträgt sie 13 Monate. Der niedrigste Wert mit im Mittel 9 Monaten ist für die AA-Merkmalsausprägung zu beobachten. Die niedrige Prognose für das AA-Allel geht einher mit der Beobachtung, dass diese Patientengruppe am schlechtesten auf die Behandlung mit *Gemcitabin* reagiert (Schirmer et al. 2016); die AA-Patientengruppe zeigt eine erhöhte Resistenz gegen das Chemotherapeutikum *Gemcitabin*. In verschiedenen Krebszellenlinien konnte ein Einfluss von *Gemcitabin* auf die Genexpressionstärke des WWOX-Gens für den AA-Genotyp gezeigt werden, während dieser Effekt für die Varianten AG- bzw. GG nicht deutlich erkennbar ist (Schirmer et al. 2016).

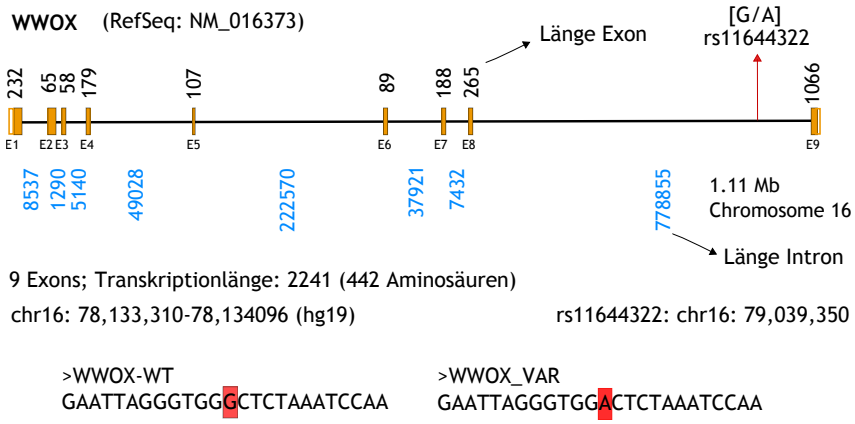


Abbildung 4.3: WWOX Genstruktur. Die orange-markierten Flächen zeigen die kodierenden Anteile des Gens (Exon E1-E9). Die in Schwarz angegebenen Zahlen geben die Länge des jeweiligen Exons an, während die blauen Zahlenwerte die Länge der Introns darstellen. Der SNP rs11644322 befindet sich zwischen Exon 8 und 9 (siehe auch rote Pfeilmarkierung).

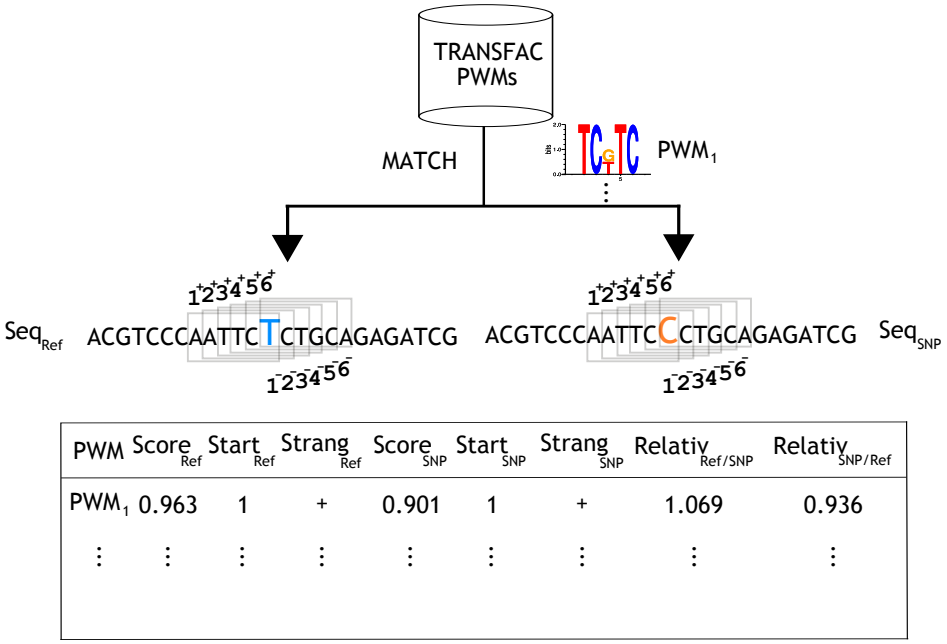
Diese Beobachtungen sind der Ausgangspunkt für die nachfolgenden bioinformatischen Analysen, welche in diesem Projekt durchgeführt wurden. Die Einzelnukleotidvariation rs11644322 ist in einem Intron des WWOX-Gens lokalisiert, welches sich zwischen

dem 8. und 9. kodierenden Exon befindet (siehe Abbildung 4.3). Abbildung 4.3 zeigt die Genstruktur dieser Transkriptionseinheit im Detail. Die RefSeq-Datenbank (siehe Kapitel 3.7) listet fünf bekannte Transkriptionsvarianten dieses Gens auf, wobei nicht alle beschriebenen Transkriptionseinheiten die letzten beiden Exons (E8, E9) enthalten. Weitere vorhergesagte Transkriptionsvarianten sind in der Datenbank aufgelistet. Schirmer et al. (2016) zeigen in ihrer Veröffentlichung, dass diese Intronregion, in welcher dieser SNP lokalisiert ist, an 67 % aller Transkriptionsereignisse beteiligt ist. Da eine unterschiedliche Transkriptionsaktivität in Abhängigkeit der verschiedenen Allele dieser SNP-Position für das WWOX-Gen beobachtet werden konnte, wird eine regulatorische Bedeutung dieses Sequenzbereichs angenommen. Diese Vermutung wird durch weitere experimentelle Analysen unterstützt. Mit Hilfe der *Electrophretic Mobility Shift Array* (EMSA)-Methode kann die regulatorische Bedeutung dieser SNP-Region experimentell bestätigt werden. Das EMSA-Verfahren wird im Allgemeinen angewendet, um die sequenzspezifische Interaktion von DNA-bindenden Proteinen mit einer bekannten DNA-Sequenzen zu untersuchen. In diesem Projekt konnten Schirmer et al. (2016) mit Hilfe dieses Verfahrens nachweisen, dass eine für die SNP-Region rs11644322 definierte Protein-DNA-Wechselwirkung vorliegt. Mit Hilfe eines Zellkernextrakts, welcher aus *Jurkat*-Zellen isoliert wurde, konnte für ein unbekanntes Protein die spezifische Protein-DNA-Interaktion nachgewiesen werden. Weiterhin wurde gezeigt, dass das G-Allel im Vergleich zur A-Variante eine signifikant stärkere Interaktion aufweist (Schirmer et al. 2016).

Das Auffinden einer möglichen TFBS und des daraus abzuleitenden TFs, welcher mit dieser vorhergesagten Bindestelle interagiert, ist die konkrete bioinformatische Aufgabe, die sich nun im Anschluss an diese experimentellen Beobachtungen stellt. Die TRANSFAC-Datenbank bietet für die Beantwortung dieser Fragestellung eine gute Grundlage. Sie beschreibt in Form der PWMs verschiedene Bindestellenmotive für bekannte TFs (Matys, Fricke et al. 2003; Matys, Kel-Margoulis et al. 2006; Edgar Wingender 2008). Diese Motiv-Sammlung kann insgesamt verwendet werden, um mittels eines neu zu entwickelnden Bewertungsverfahrens für die vorliegende Sequenzregion inklusive der SNP-Region rs11644322 potenzielle TFBS vorherzusagen. Die Vorhersage der potentiellen TFBSs ist mit Hilfe des MATCH-Algorithmus durchgeführt worden, der diePWM Bibliothek der TRANSFAC-Datenbank nutzt und potenzielle TFBSs anhand des MSS (siehe Kapitel 3.3) bewertet (Kel et al. 2003). Bei der vorliegenden Fragestellung muss die Vorhersage aller möglichen TFBSs allerdings im Vergleich der beiden Einzelnukleotidvariationen erfolgen.

Gleichzeitig ist durch die experimentelle Voruntersuchung bekannt, dass beide Sequenzvariationen die Interaktion eines TFs erlauben. Die stärkere Präferenz des G-Allels im Vergleich zur A-Variante sollte dabei ebenfalls Berücksichtigung finden. Weiterhin gilt, dass auf Grundlage des EMSA-Experiments kein bestimmter TF favorisiert werden kann.

Ausgehend von diesen Beobachtungen ist für das Projekt eine vergleichende Analysemöglichkeit zur Bewertung regulatorischer SNPs entwickelt worden. Diese Methode analysiert



PWM	Score Ref	Start Ref	Strang Ref	Score SNP	Start SNP	Strang SNP	Relativ Ref/SNP	Relativ SNP/Ref
PWM1	0.963	1	+	0.901	1	+	1.069	0.936
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Abbildung 4.4: Schematische Darstellung der regulatorischen SNP-Analyse. Die Darstellung zeigt die vergleichende Analysestrategie des implementierten Verfahrens. Eine regulatorische Sequenzvariation wird durch die Einzelbewertung der jeweiligen SNP-Situation mit Hilfe von in der TRANSFAC-Datenbank beschriebenen PWMs untersucht. Jede Position einer Vorhersage, welche die SNP-Position enthält, wird für die Analyse betrachtet. Potenzielle TFBSs können sowohl auf dem Plus-, als auch auf dem Minusstrang gefunden werden. Aus diesem Grund wird bei der vergleichenden Analyse auch diese Annotation berücksichtigt.

alle möglichen TFBSs einer SNP Position, indem sie die zwei unterschiedlichen Sequenzvariationen einer Position in einer vorhergesagten TFBS auf Grundlage der MATCH-Vorhersage berücksichtigt. Der Vergleich einer TFBS-Vorhersage für zwei unterschiedli-

che Sequenzvariationen an der gleichen Position und unter Berücksichtigung des Strangbezugs erlaubt eine direkte Bewertung des Einflusses der SNP-Position. Für die relativen Vergleiche der einzelnen Matrizenbewertungen kann der MSS-Sequenzähnlichkeitswert des MATCH-Algorithmus verwendet werden, der auf den Wertebereich [0,1] skaliert ist (siehe Kapitel 3). Das bedeutet, dass diese PWM-Bewertung direkt für eine vergleichende Analyse herangezogen werden kann. Ausgangspunkt der Untersuchung ist eine nicht-redundante Datensammlung der TRANSFAC-Datenbank (vertebrate_non-_redundant_minFN.prf, Version 2014.4). Um ein möglichst umfangreiches Bild aller möglichen vorhergesagten TFBSs zu erhalten, wird in dieser Analyse für die Ähnlichkeitsbewertung mittels des MATCH-Algorithmus der für jede Matrix vorgegebene Schwellenwert minFN verwendet (zur Minimierung der false negatives siehe Material- und Methodenteil 3.3). Auf Grundlage der zuvor beschriebenen Beobachtungen sind nur die potenziellen TFs

Matrize	Start	Strang	Site	MSS.G	MSS.A	G/A Ratio
V$RHOX11_01	8	+	ggtggG/Actctaaatcca	0.798	0.655	1.218
V$HOXD12_01	9	+	gtggG/Actctaaatccaa	0.624	0.563	1.108
V$SP1_Q6_01	6	+	agggtggG/Act	0.900	0.854	1.054
V$HOXC13_01	9	+	gtggG/Actctaaatcca	0.607	0.597	1.017
V$NKX25_Q6	3	-	attagggtggG/A	0.698	0.690	1.012
V$TBX5_01	5	+	tagggtggG/Actc	0.772	0.764	1.010
V$HOXB13_01	10	+	tggG/Actctaaatccaa	0.651	0.647	1.006
V$SREBP_Q6	2	-	aattagggtggG/Actc	0.91	0.906	1.004
V$CPHX_01	5	-	tagggtggG/Actcta	0.511	0.510	1.002
V$PBX_Q3	11	-	ggG/Actctaaatc	0.769	0.773	0.995
V$RHOX11_01	6	+	agggtggG/Actctaa	0.500	0.518	0.965
V$RHOX11_01	9	-	attagggtggG/Actctaa	0.559	0.619	0.903

Tabelle 4.1: Vorhergesagte TFBSs des SNPs rs11644322. Die aufgelisteten Bindestellen beschreiben die positions- und stranggefilterten Vorhersagen, welche sowohl in der A, als auch in der G-Variante gefunden worden sind.

untersucht worden, welche für das G-Allel im Vergleich zur A-Situation eine größere vorhergesagte Bindungsaffinität zeigen.

Die beiden Sequenzvarianten (G/A) werden einzeln mit der entsprechenden SNP-Variation untersucht (siehe auch Abbildung 4.3). Die mittlere Position einer 25 bp langen DNA-Sequenz beschreibt die entsprechende SNP-Position (Position 13). Die Länge der untersuchten Sequenz ist nicht festgelegt und kann im implementierten Verfahren dynamisch angepasst werden. Die beiden Eingabesequenzen unterscheiden sich also nur in der mittle-

ren Position, welche entweder durch ein G oder durch ein A bestimmt ist. Es werden nun alle Bindestellenvorhersagen betrachtet, die den gegebenen Schwellenwert des minFN-Profils überschreiten (siehe Material- und Methodenteil 3.3). Das verwendete Profil enthält insgesamt 164 Positionsgewichtungsmatrizen. Für die 25 bp lange DNA-Sequenz werden für die G-Variante 120 bzw. für die A-Situation 121 potentielle TFBSs vorhergesagt. Da die SNP-Position einen im Experiment nachgewiesenen Einfluss auf die Transkription des WWOX-Gens besitzt (Schirmer et al. 2016), werden die Bindestellenvorhersagen nun auf die Bereiche gefiltert, die diese SNP-Position in mindestens einer Position überlappen (siehe auch Abbildung 4.4). Dieser Schritt reduziert die möglichen TFBS-Vorhersagen für die beiden Sequenzvarianten auf 34 bzw. 37 für die G- bzw. A-Sequenzvariation. Diese unabhängig durchgeführten Analysen werden nun weiterverarbeitet. Es werden nur diejenigen Bindestellenvorhersagen betrachtet, die sich in beiden Situationen wiederfinden. Als Bedingung für einen Vergleich gilt: gleiche PWM, gleiche Startposition und gleicher DNA-Strang. Tabelle 4.2 zeigt die Ergebnisse der Analyse und Filterung für die analysierte Sequenzvariation. Zur besseren Orientierung ist die jeweilige SNP-Position als Großbuchstabe in der jeweils vorhergesagten TFBS angegeben, während der Rest der Bindestelle in Kleinbuchstaben notiert ist. Auf Grundlage der zuvor beschriebenen Beobachtungen sind

Matrize	**Spezies in Matrix**	**TF(s)**
V$RHOX11_01	1	1
V$HOXD12_01	3	3
V$SP1_Q6_01	16	22
V$HOXC13_01	2	4
V$NKX25_Q6	5	9
V$TBX5_01	3	8
V$HOXB13_01	2	2
V$SREBP_Q6	9	40
V$CPHX_01	1	2
V$PBX_Q3	3	36
V$RHOX11_01	1	1
V$RHOX11_01	1	1

Tabelle 4.2: Auflistung der relevanten TRANSFAC-Matrizen für die Sequenzvariation rs11644322 und die damit verbunden unterschiedlichen Spezies bzw. TFs.

nur die potenziellen TFs untersucht worden, welche für das G-Allel im Vergleich zur A-Situation eine größere vorhergesagte Bindungsaffinität zeigen.

Die Ergebnisse dieser Vorhersage wurden anschließend zur experimentellen Validierung an das beteiligte Labor übermittelt. Zur Evaluierung kam wiederum das EMSA-Verfahren zum Einsatz (Lüske 2015; Roppel 2013). Der repräsentative EMSA-Plot dieser Untersuchung entstammt der Veröffentlichung von Schirmer et al. (2016) und ist in Abbildung 4.5 (a) gezeigt. Das EMSA-Diagramm belegt die allgemeine Protein-DNA-Interaktion dieser regulatorischen SNP-Region: Ein allgemeiner Kernprotein-Extrakt, welcher aus *Jurkat*-Zellen isoliert wurde, bindet in unterschiedlicher Konzentration sowohl mit der G-Variante als auch mit der A-Variante (siehe Abbildung 4.5 a). Diese Experimente belegen die allgemeine regulatorische Aktivität dieser genomischen Region. Die vorletzten beiden Bahnen des EMSA-Plots zeigen die spezifische Interaktion des Transkriptionsfaktors SP1 (UniProt: P08047) mit der häufigeren G-Variante (Wildtyp): Durch Verwendung eines Antikörpers für diesen TF erscheint eine weitere Bande (siehe Abbildung 4.5 a, schwarzer Pfeil). Diese zeigt sich, da die spezifische Interaktion des Antikörpers mit dem SP1-Protein und der untersuchten regulatorischen Region einen neuen Proteinkomplex definiert, welcher in der Kontrollvariante nicht erscheint (siehe Abbildung 4.5 a, letzte Bande). Abbildung 4.5 (c) bestätigt, dass die regulatorische Region auch mit Pankreas-spezifischen Kernproteinen interagiert (*MIA-PaCa-2* Zelllinie, siehe (Schirmer et al. 2016)) und dort wahrscheinlich auch vom Transkriptionsfaktor SP1 gebunden ist. Durch die Positionierung des SNPs rs11644322 im letzten bekannten Intron des WWOX Gens (siehe auch Abbildung 4.3) ist eine weitere mögliche biologische Funktion dieser regulatorischen SNP-Situation untersucht worden. Durch Lokalisation und die in der Ref-Seq-Datenbank vorgeschlagenen verschiedenen Transkriptionsvarianten ist nicht auszuschließen, dass diese SNP-Region einen alternativen Promotor für das WWOX-Gen definiert. Diese Möglichkeit konnte jedoch experimentell ausgeschlossen werden (Roppel 2013; Schirmer et al. 2016). Die EMSA-Experimente zeigen jedoch eindeutig eine unterschiedliche Affinität der A-Variante im Vergleich zum Wildtyp (G-Variante). Weiterhin ist durch nachfolgende Experimente belegt, dass die Region dieses SNPs die Expressionsstärke der aus der WWOX-Transkriptionseinheit exprimierten Transkripte beeinflusst (Lüske 2015; Schirmer et al. 2016).

Wie im Grundlagenkapitel bereits beschrieben, wird die Transkription eines Gens sowohl durch proximale als auch distale Regionen wesentlich beeinflusst. Die experimentellen Ergebnisse lassen vermuten, dass eine distale regulatorische Funktion von dieser SNP-enthaltenden Intronregion ausgeht. Unter Beteiligung des Transkriptionsfaktors SP1

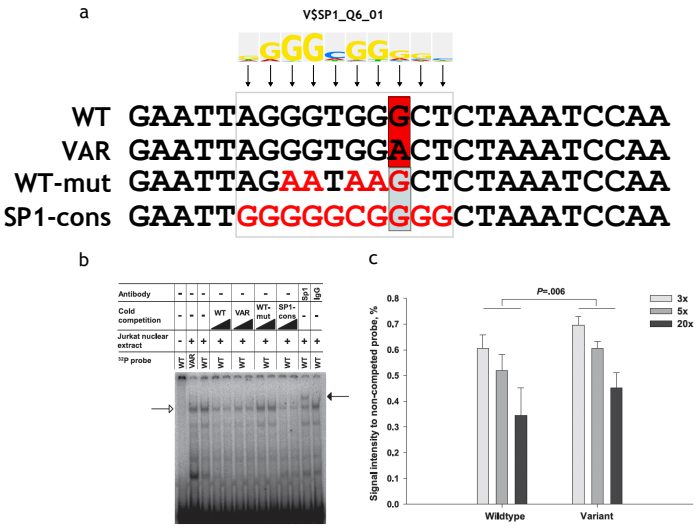


Abbildung 4.5: Bedeutung des Transkriptionsfaktors SP1 in der SNP-Region. Die Abbildung (a) zeigt das Sequenzlogo der Bindestellen für den Transkriptionsfaktors SP1 aus der TRANSFAC-Datenbank (V$SP1_Q6_01). In der vergleichenden Analyse kann mit Hilfe dieser PWM die SNP-Region als potenziell SP1-bindend vorausgesagt werden. Im Teil (b) dieser Abbildung ist die experimentelle Bestätigung durch die EMSA-Analyse gezeigt. Das Diagramm belegt die Bindung von TFs aus Kernproteinextrakten (aus *Jurkat*-Zellen) für den SNP rs11644322 im WWOX-Gen. Die erste Bahn zeigt die Negativkontrolle ohne Kernproteine. Die Bahnen 2 und 3 veranschaulichen die Komplexbildung für die A-Variante (VAR) bzw. G-Situation (Wildtyp, WT). Die Banden ergeben sich auf Grundlage der verwendeten radioaktiven Markierung (32P-markiert). In den Bahnen 4-11 wird die Wechselwirkung zwischen Kernproteinen und dem G-Allel (4,5), dem A-Allel (6,7), G-Allel-mutiert (8,9; siehe Sequenz 3 aus a) und der Sequenzvariante des SP1-Konsensus (10,11; siehe Sequenz 1 aus a) gezeigt. Jede dieser Bedingungen wird in zwei verschiedenen experimentellen Bedingungen untersucht (3- bzw. 5-fachen Überschuss mit nicht-radioaktiv markierter Sonde). In der vorletzten Bande wird die Bedeutung des Transkriptionsfaktor SP1 deutlich: durch die Verwendung eines Antikörpers für SP1 belegt diese Untersuchung, dass der Transkriptionsfaktor SP1 an die G-Variante bindet: die Zugabe des Antikörpers erklärt die neue Bande.

könnte diese Region als Enhancer mit den am Transkriptionsstart gebundenen Transkriptionsfaktoren und Kofaktoren interagieren. Es gibt Hinweise in der Literatur, welche für verschiedene Transkriptionseinheiten einen solchen Einfluss unter Verwendung des Transkriptionsfaktors SP1 gezeigt haben (Bianchi et al. 2009; Deshane et al. 2010).

4.1.3 Analyse der regulatorischen Bedeutung der Sequenzvariation rs3857080 für die Transkription des NR3C2-Gens

In einem vergleichbaren Projekt wurde die regulatorische Bedeutung des SNPs rs3857080 (A/G) untersucht. Dieser Polymorphismus befindet sich im dritten Intron des NR3C2-Gens (Mineralokortikoidrezeptor, UniProt: P08235). Diese Transkriptionseinheit kodiert für einen sogenannten nukleären Rezeptor (NR), der durch Interaktion mit einem Steroidhormon als Transkriptionsfaktor wirken kann (Arriza et al. 1987).

Harntreibende Medikamente kontrollieren den Salzhaushalt, die Wasserausscheidung und den Blutdruck. Sie werden daher häufig bei der Behandlung von Bluthochdruck, Herzkreislauferkrankungen und Herzkranzgefäßerkrankungen verwendet (Chobanian et al. 2003; Mancia et al. 2013; McMurray et al. 2012; Yancy et al. 2013). Der Mineralokortikoidrezeptor (MR, auch Aldosteron-Rezeptor genannt) spielt eine wichtige Rolle bei der Regulation des Natriumhaushalts (Vormfelde et al. 2007). Dalila et al. (2015) haben in ihrer Studie die in der NR3C2-Transkriptionseinheit vorliegenden SNPs in zwei freiwilligen Kohorten untersucht. Die eine Gruppe, bestehend aus 96 Personen, ist mit einer Einzeldosis von *Bumetanid*, *Furosemid* und *Torsemid* behandelt worden. Die zweite Gruppe setzt sich aus 107 Teilnehmern zusammen. Diese Patienten wurden mit Einzeldosen der Medikamente *Hydrochlorothiazid* und *Triamteren* behandelt. Beide Probandengruppen haben sich strikt an eine Natriumchlorid-arme Diät gehalten und jede Testsperson war eindeutig einer der beiden Gruppen zugeordnet. Durch Literaturrecherche wurden 12 verschiedene SNPs identifiziert, welche im NR3C2-Gen zu finden sind und mit einem klinischen Phänotypen der untersuchten Krankheiten assoziiert werden können (Dalila et al. 2015). Durch einen Vergleich der Salzkonzentration (Natrium- und Kaliumionenkonzentration) des Urins der beiden untersuchten Gruppen sowie der Assoziation dieser Ergebnisse mit den zwölf untersuchten SNPs für jeden Studienteilnehmer aus beiden Gruppen konnte der SNP rs3857080

Matrize	Start	Strang	Site	MSS.A	MSS.G	A/G Ratio
V$DLX_01	7	+	gtgtcA/Gttttaatgt	0.824	0.622	1.325
V$DLX_01	9	-	gtcaA/Gttttaatgt	0.716	0.540	1.326
V$MSX1_02	5	+	gagtgtcaA/Gttttaat	0.780	0.588	1.327
V$HOXB_01	6	+	agtgtaA/Gttttaatgt	0.707	0.532	1.329
V$BARX1_01	5	-	gagtgtcaA/Gttttaat	0.846	0.635	1.332
V$LHX4_01	5	-	gagtgtcaA/Gttttaatg	0.717	0.538	1.333
V$HOXA7_02	5	-	gagtgtcaA/Gttttaatg	0.759	0.569	1.334
V$BARX1_01	7	+	gtgtcaA/Gttttaatgt	0.841	0.630	1.335
V$HOXD3_01	7	-	gtgtcaA/Gttttaatgt	0.705	0.528	1.335
V$HOXC6_01	5	-	gagtgtcaA/Gttttaatg	0.720	0.539	1.336
V$HOXC6_01	6	+	agtgtcaA/Gttttaatgt	0.714	0.533	1.340

Tabelle 4.3: Vorhergesagte TFBS des SNP rs3857080. Die aufgelisteten Bindestellen beschreiben die positions- und stranggefilterten Vorhersagen, welche sowohl in der A- als auch in der G-Variante gefunden worden sind.

(p-Wert < 0.001) als klinisch relevant gefunden werden. Diese Einzelnukleotidvariation taucht in den Varianten GG, GA, AA auf. Dieser SNP zeigt in beiden unabhängigen Kohorten eine Korrelation mit hohen Kalium- bzw. Natriumchlorid-Ausscheidungen und tritt mit einer Häufigkeit von fünf bis zehn Prozent in der kaukasischen Bevölkerungsgruppe auf (Dalila et al. 2015). Durch experimentelle Voruntersuchungen konnte gezeigt werden, dass der SNP rs3857080 einen Einfluss auf die Expression des NR3C2-Gens besitzt. Analog zu dem WWOX-bezogenen Projekt (siehe oben) ist die implementierte Analysestrategie auch in diesem Projekt angewandt worden. Tabelle 4.3 listet die zehn stärksten relativen Veränderungen potentieller Bindestellenvorhersagen für die beiden untersuchten Einzelnukleotidvariation auf (TRANSFAC-Version 2014.4). Der auf dem sechsten Rang aufgelistete Transkriptionsfaktor LHX4 konnte durch nachfolgende Experimente als bindender TF in diesem Bereich nachgewiesen werden (siehe Tabelle 4.3). Für das A-Allel ließ sich eine stärkere Bindung des Transkriptionsfaktors LHX4 ermitteln. Anhand einer Mutation der vier wichtigsten Positionen der vorliegenden LHX4-Bindestelle in beiden vorliegenden SNP-Varianten wurde in einem nachfolgendem Experiment gezeigt, dass in der mutierten Form, unabhängig von der vorliegenden Ausprägung des SNP, der Transkriptionsfaktor LHX4 nicht mehr an die so veränderte DNA-Sequenz bindet (Dalila et al. 2015).

Abbildung 4.6 (b) zeigt die Situation des SNPs in den beiden Sequenzausprägungen. Die Genstruktur der untersuchten NR3C2-Transkriptionseinheit ist in 4.6 (a) dargestellt. Abbildung 4.6 (c) gibt die regulatorische Bedeutung des SNPs rs3857080 wieder (Dalila

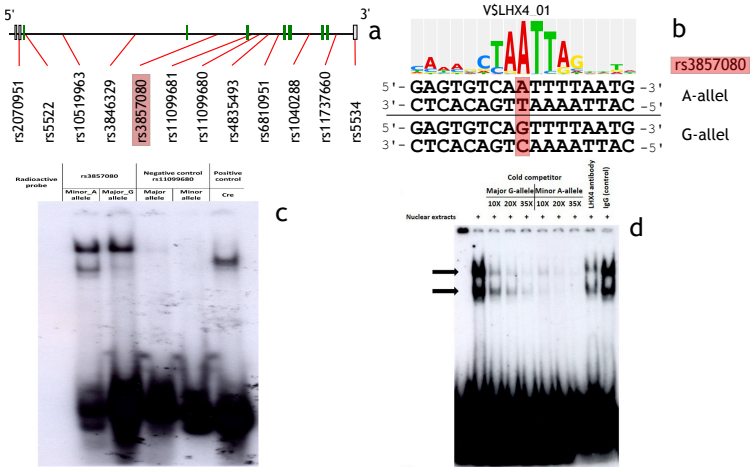


Abbildung 4.6: Bedeutung des Transkriptionsfaktors LHX4 in der SNP-Region rs3857080. (a) zeigt die Struktur der NR3C2 Transkriptionseinheit und die Positionierung der 12 untersuchten SNPs aus der Untersuchung von Dalila et al. (2015). Die rechteckigen Strukturen beschreiben die bekannten Exons dieses Gens (grün: kodierende; grau: nicht-kodierend). (b) zeigt das Sequenzlogo für die PWM V$LHX4_01 aus TRANSFAC. (c) stellt den experimentellen Nachweis für die Interaktion von Kernproteinen (aus *HEK293*-Zellen) für diesen SNP dar (EMSA-Plot). Das A-Allel zeigt ein stärkeres Signal als die G-Variante. Die Negativkontrolle zeigt keinerlei Kernproteininteraktion, während die Positivkontrolle ein vergleichbares Signal liefert. In einem weiteren EMSA-Diagramm (d) wird die Bedeutung des A-Allels deutlich. In drei verschiedenen Konzentrationen (10x, 20x und 35x) wurde nicht-32P-markierte DNA auf Basis des A-Allel hinzugegeben. Dadurch schwächt sich das Signal ab (siehe Pfeilmarkierung, Bahnen 3-8). Die letzten beiden Bahnen belegen die Bedeutung des Transkriptionsfaktors LHX4. Durch die Verwendung eines Antikörpers für diesen TF wird das Signal (Bahn 9) im Vergleich zu einer Kontrolle (Bahn 10) deutlich reduziert.

2014; Dalila et al. 2015). Die experimentelle Bestätigung der Interaktion des Transkriptionsfaktors LHX4 ist im Abbildungsteil 4.6 (d) dargestellt. Vergleichbar mit der Untersuchung der Sequenzvariation im WWOX-Gen wird auch für die Einzelnukleotidvariation rs3857080 eine im Intron vorliegende Enhancer-Funktion vermutet. Diese lokale Interaktion mit dem proximalen Promoter dieses Gens und den dort gebundenen Transkriptionsfaktoren und Kofaktoren könnte eine verstärke Transkriptionsaktivität bewirken. Für die regulatorische Bedeutung von LHX4 in Intronregionen lassen sich in der Literatur Beispiele finden (Gergics et al. 2015). Eine stärkere Bindung des Transkriptionsfaktors LHX4, wie sie in der A-Variante vermutet wird, sorgt möglicherweise für eine stabilere Ausprägung dieser Enhancer-Promoter-Interaktion.

4.1.4 Eigenschaften transkriptionsregulatorischer Sequenzvariationen

Die beiden Untersuchungen zeigen verschiedene Eigenschaften klinisch bedeutender regulatorisch aktiver Einzelnukleotidvariationen und deren Bedeutung in der transkriptionellen Genregulation. Der Vergleich von mindestens zwei verschiedenen genomischen Sequenzvariationen kann mit Hilfe von PWMs analysiert und bewertet werden und somit direkt auf eine potenzielle TFBS und deren interagierende Menge an TFs bezogen werden. Dabei erscheint die Vorhersage der Bindestellenaffinität ein geeignetes Maß zu sein, um potenziell funktionale TFBS vorherzusagen. Die beiden Untersuchung zeigen weiterhin, dass TFBSs nicht nur in Promotoren, sondern auch in sogenannten intragenischen regulatorischen Regionen zu finden sind. Diese werden in der Literatur auch alternativ als Intron-definierte Enhancer bezeichnet. Die Untersuchungen für das WWOX-Gen zeigen, dass eine weitere TFBS, die bereits auch im Promoter dieses Gens zu finden ist, einen messbaren Einfluss auf die Transkriptionsstärke des Gens besitzt und damit eine krankheitsbezogene Bedeutung ausüben kann. Für das NR3C2-Gen konnte gezeigt werden, dass eine Einzelnukleotidvariation eine Bindestelle für einen neuen TF erzeugen kann.

4.2 Regulatorische Transkriptionsnetzwerke

In diesem Teil der Arbeit wird die Erstellung und Anwendung eines *Regulatory Transcription Network* (RTN) vorgestellt. Die Konstruktion dieses Netzwerks ist im Jahr 2012 erstmals vorgestellt und untersucht worden (Haubrock, Li et al. 2012; Li et al. 2012). Im Jahr 2013 wurde das RTN neu erstellt und ist in verschiedenen Projekten verwendet worden (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013; Bhar, Haubrock, Mukhopadhyay und Edgar Wingender 2015; Daou et al. 2020; Wingender, Schoeps, Haubrock und Dönitz 2015; Wingender, Schoeps, Haubrock, Krull et al. 2018)

4.2.1 Bedeutung regulatorischer Transkriptionsnetzwerke

Bisherige Schätzungen vermuten mehr als 300 verschiedene Zelltypen im Menschen (Alberts et al. 2002). Die Entwicklung dieser Zelltypen und deren koordiniertes Zusammenspiel in einem Organ oder biologischen System unterliegt spezifischen regulatorischen Einflüssen. Eine grundlegende Fragestellung in diesem Zusammenhang ist: Welche Prozesse kontrollieren bzw. koordinieren die unterschiedlichen situationsspezifischen Genaktivitäten? Den Transkriptionsfaktoren kommt hier eine besondere Rolle zu. Sie sind unter den ersten regulatorisch wirkenden Proteinen zu finden, welche durch ihr koordiniertes Zusammenwirken dafür Sorge tragen, dass situationsspezifisch die notwendigen Gene und deren spezifische Transkriptionseinheiten transkribiert werden. Durch die Verwendung verschiedener Hochdurchsatzdaten zur Messung der Genaktivität, wie z.B. Microarrays oder RNA-seq, können für diese Ebene der Genregulation effizient experimentelle Daten erzeugt werden. Die Menge der exprimierten Gene wird im Allgemeinen auch als Transkriptom bezeichnet.

Es existieren verschiedene statistische Verfahren, welche auf Grundlage einer Transkriptommessung die für einen solchen experimentell erzeugten Datensatz bedeutende TFs bestimmen können. Die Gemeinsamkeit dieser Verfahren besteht darin, dass auf Basis der regulatorischen Bereiche der auffällig exprimierten Gene die statistisch signifikant angereicherten TFBSs vorhergesagt werden. Anhand dieser Bindestellen kann dann in einem

zweiten Schritt auf die möglichen interagierenden TFs geschlossen werden. Diese Analyse erfolgt z.B. auf Grundlage bekannter PWMs, oder aber sie verwendet die regulatorischen Sequenzen ohne Vorwissen und analysiert dort angereicherte Motive, welche als Bindestellen von TFs interpretiert werden (siehe auch Kapitel 2). Die Ergebnisse dieser Untersuchungen zeigen häufig ein sehr unspezifisches Bild. Das bedeutet, dass eine große Menge an TFs als regulatorisch bedeutend (statistisch signifikant) gefunden wird. Die notwendigen biologischen Prozesse einer lebenden Zelle sind im Genom kodiert. Durch die gezielte Transkription dieser funktionalen Bereiche stehen die verschiedenen Entitäten (Transkriptionseinheiten) eines biologischen Prozesses direkt (in Form von RNA) oder indirekt (nach Translation als Proteine) zur Verfügung. TFs steuern also durch die Regulation der Transkription das Vorhandensein dieser biologischen Entitäten und regulieren damit die notwendigen allgemeinen und spezifischen Zellfunktionen. Um die Bedeutung eines einzelnen TFs im Kontext der Regulation besser bewerten zu können, wird in dieser Arbeit ein allgemeingültiges sogenanntes Referenznetzwerk der Transkriptionsregulation erstellt. Dieses basiert auf vorhergesagten TFBSs, die gleichzeitig auch in einer Gruppe von vier ausgewählten Säugetiergenomen sequenzkonserviert sind. Die Erstellung, Bewertung und Anwendung dieses allgemeinen Transkriptionsnetzwerks wird in den weiteren Unterkapiteln vorgestellt.

Durch die Vorhersage von qualitativen TF-Gen-Beziehungen auf Basis von konservierten potentiellen TFBSs steht ein allgemeines Werkzeug zur Beschreibung bedeutender regulatorischer Signale zur Verfügung. Die Abbildung und Kombination dieser Signale als regulatorisches Netzwerk weist außerdem die vielfältigen Einflüsse verschiedener TFs aus, die bei der statistischen Analyse so nicht betrachtet bzw. beschrieben werden können. Ähnlich zu den in der Literatur viel diskutierten Konzepten der Master- oder Pionier-Transkriptionsfaktoren (Heinz et al. 2010b; Mullen et al. 2011; Whyte et al. 2013; Zaret und Carroll 2011) können die konservierten potenziellen TFBSs des Referenznetzwerks als prägende regulatorische Schaltstellen im Genom der untersuchten Spezies aufgefasst werden. In den proximalen (Promotor-nahen) regulatorischen Sequenzen lassen sich so einige wenige definierte und sequenzkonservierte TFBSs finden, welche im weiteren Verlauf dieser Arbeit als sogenannte Seed-Sites (Kernbindestellen, kurz Seeds) bezeichnen werden. Diese Seeds charakterisieren regulatorisch bedeutende Positionen in einem Promotor, welche mit Hilfe der dort interagierenden TFs den Ausgangspunkt für charakteristische regulatorische Module einer Transkriptionseinheit definieren.

4.2.2 Vorhersage evolutionär konservierter TFBSs

Das vorliegende RTN ist in einem mehrstufigen Prozess erstellt worden. Als erstes sind auf der Grundlage der in der RefSeq-Datenbank (O'Leary et al. 2016) definierte menschliche Promotorregionen bestimmt worden. Dazu wurde für jede Transkriptionseinheit aus Ref-Seq der 1000 bp umfassende Promotor stromaufwärts eines TSS identifiziert. Für die so definierten Promotoren können nun die zu diesen Regionen sequenzähnlichen Bereiche bestimmt werden. Dazu wird ein vorberechnetes sogenanntes Whole-Genome-Alignment

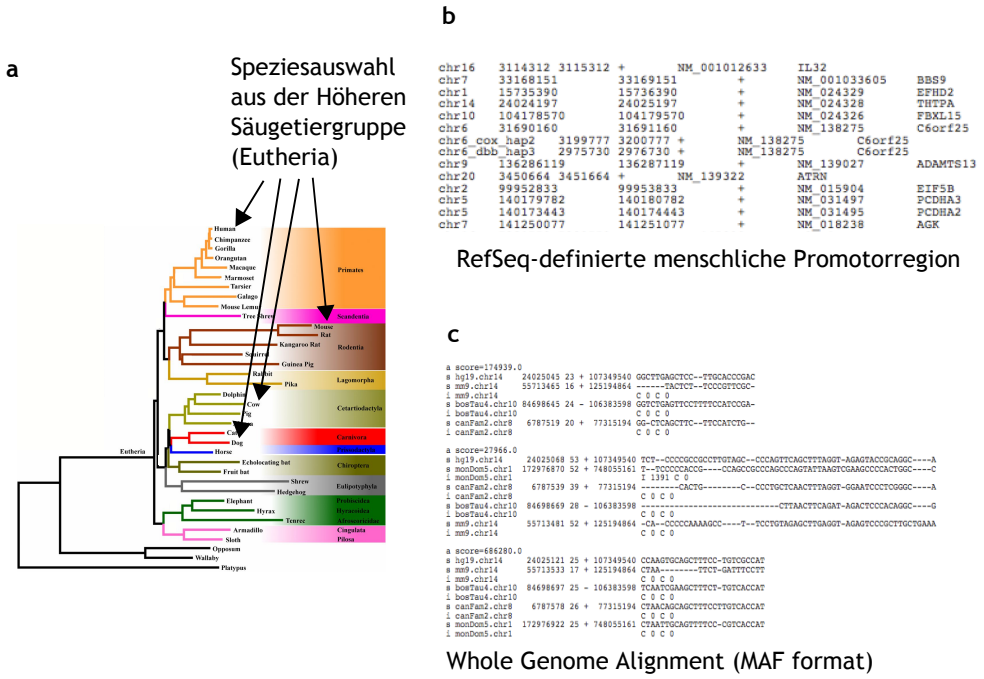


Abbildung 4.7: Phylogenetische Konservierung regulatorischer Regionen. (a) Phylogenetischer Baum innerhalb der Gruppe ausgewählter Säugetiergenome. Die Abbildung entstammt der Publikation von Song et al. (2012). In Abschnitt (b) ist ein Ausschnitt der verwendeten RefSeq-definierten menschlichen Promotoren dargestellt. (c) zeigt ein Beispiel der identifizierten Sequenz-konservierten Regionen eines Promotors des Menschen (hg19), der Maus (mm9), des Rinds (bosTau4) und des Hundes (canFam2). Pro Block sind die vier ausgewählten Spezies und deren homolog gefundene Promotorbereiche gezeigt.

(WGA) verwendet, welches auf Grundlage von 46 ausgewählten Wirbeltiergenomen erstellt wurde (Blankenberg et al. 2011). Dieses Alignment steht durch einen Service der

Universität Santa Cruz zur Verfügung (Kent et al. 2002). Auf dieser Datengrundlage können die sequenzkonservierten Bereiche zwischen dem Menschen, der Maus, dem Hund und dem Rind identifiziert werden, welche zu den zuvor bestimmten menschlichen Promotorregionen als sequenzähnlich identifiziert worden sind. Dazu wurde die Funktionalität des Galaxy-Servers verwendet (Afgan et al. 2018). Die vier Spezies wurden gewählt, da sie einen ungefähr äquidistanten phylogenetischen Abstand aufweisen (siehe auch Abbildung 4.7 a). Mit Hilfe des MATCH-Algorithmus (Kel et al. 2003) und der in der TRANSFAC-Datenbank (Matys, Fricke et al. 2003; Edgar Wingender 2008) gespeicherten, für Wirbeltiere definierten, PWMs, können nun die vorliegenden konservierten Promotorbereiche dieser vier Spezies als Basis für die TFBSs-Vorhersagen verwendet werden. Für die Vorhersage ist die TRANSFAC-Datenbank in der Version 2009.4 benutzt worden. Eine konservierte TFBS liegt vor, wenn die verwendete PWM in allen vier Spezies an den im WGA als zueinander ähnlich identifizierten Positionen eine mindestens dem minFN-Schwellenwert genügende Vorhersagequalität aufweist (siehe Kapitel 3). Die Abbildung 4.7 zeigt beispielhaft drei identifizierte konservierte Sequenzregionen, welche Ausgangspunkt für die Bindestellenvorhersage sind.

Abbildung 4.8 zeigt zwei Bespiele menschlicher TFBSs aus der TRANSFAC-Datenbank. Die Sequenzkonservierung dieser Bindestellen ist im Vergleich zu den orthologen Promotorregionen der Maus (*Mus musculus*, kurz mmu) dargestellt. Die Beispiele beschreiben unterschiedliche Ausprägungen der Sequenzkonservierung einer TFBS: (a) vollständige Sequenzkonservierung bzw. (b) Musterkonservierung einer TFBS. Im oberen Teil der Abbildung ist die TFBS für den Transkriptionsfaktor MYOD1 (UNIPROT: P15172) im menschlichen FOS-Promoter gezeigt. Der Sequenzvergleich mit dem orthologen Mauspromotor zeigt die vollständige Sequenzkonservierung dieser Bindestelle. Die Konservierung erstreckt sich über eine größere Region: Die rot markierten Positionen (siehe Abbildung 4.8 a) definieren die Nukleotide der eigentlichen Bindestelle, welche in der TRANSFAC-Datenbank für diese Sequenz als TFBS annotiert sind. Eine Sequenzkonservierung muss allerdings nicht alle Positionen einer Bindestelle betreffen. Hierfür gibt Abbildung 4.8 (b) ein Beispiel: Zehn von dreizehn Positionen der Bindestelle für den Transkriptionsfaktor C/EBP (UniProt: P49715) im PTGS2-Promotor des Menschen sind im Vergleich zwischen den orthologen Promotoren des Menschen und der Maus sequenzkonserviert. Dieses Beispiel macht deutlich, dass nicht alle Positionen einer TFBS gleichbedeutend sind und als Konsequenz daraus nicht zwangsläufig in einem definierten Sequenzalignment zu hun-

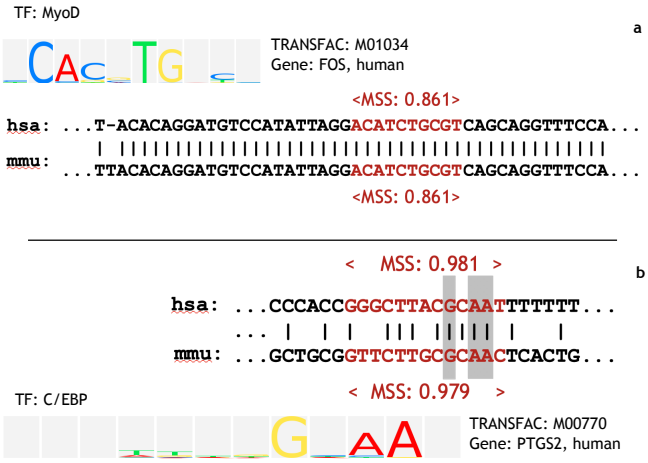


Abbildung 4.8: Beispiel der Sequenz- und Musterkonserviertheit von TFBSs. Im oberen Teil der Abbildung (a) ist die Sequenzkonserviertheit einer TFBS in der c-Fos-Promotorregion des FOS-Gens für eine Gruppe von MyoD-TFs aus der TRANSFAC-Datenbank dargestellt. Gezeigt sind die Promotorsequenzen des Menschen (hsa) und der Maus (mmu). Der Bereich weist eine ausgeprägte Sequenzkonserviertheit zwischen den orthologen Positionen der beiden Spezies auf. Diese Situation wird durch das 'I'-Symbol betont. Die eigentliche konservierte Bindestelle ist rot eingefärbt. (b) zeigt eine im Vergleich zu (a) nicht so ausgeprägte Sequenz-konservierte TFBS für den Transkriptionsfaktor C/EBP für das PTGS2 Gen. Die Bindestellensituation in (a) wird auch als sequenz- und musterkonserviert bezeichnet, während die Bindestelle in (b) als musterkonserviert beschrieben wird, da der MSS erhalten ist.

dert Prozent identisch sein müssen. Dieser spezielle Fall der Konservierung einer TFBS wird als Musterkonservierung bezeichnet (Sauer et al. 2006). Vor allem die für die Ausprägung der DNA-spezifischen Kontakte notwendigen Positionen eines TFs sind in einer TFBS sequenzkonserviert. Diese werden in der PWM bzw. der Visualisierung der PWM-Information in Form eines Sequenzlogos betont (siehe Abbildung 4.8). Der MATCH-Algorithmus, welcher in dieser Arbeit zur Vorhersage potenzieller TFBS verwendet wurde, zeigt eine vergleichbare Bewertung für musterkonservierte TFBS (siehe Abbildung 4.8 b). Grundsätzlich werden die in Abbildung 4.7 (c) gezeigten WGA-definierten Sequenzkonservierungsblöcke verwendet, um in diesen Regionen potenzielle TFBSs vorherzusagen. Die beispielhafte Sequenzkonservierung für zwei Spezies, welche in Abbildung 4.8 gezeigt ist, wird im implementierten Verfahren auf vier Spezies ausgeweitet. Dazu wird die Sequenzinformation des Alignments des Menschen, der Maus, des Hundes und des Rinds verwendet. Um die Extreme der Muster- bzw. Sequenzkonservierung berücksichtigen zu können, wird das minFN-Profil für die PWM-Bibliothek aus der TRANSFAC-Datenbank verwendet (siehe Material- und Methodenteil 3). Die gleichzeitige Bedingung der Sequenzkonservierung und einer minimalen Bindestellenqualität (Verwendung des minFN-Schwellenwerts, unterschiedlich für jede PWM) sichert mit einer gewissen Verlässlichkeit die grundsätzliche Funktionalität dieser Region als potenzielle TFBS (Sauer et al. 2006). Im Allgemeinen wird die Vorhersage von TFBSs unter Zuhilfenahme orthologer Sequenzinformation auch als phylogenetisches Footprinting bezeichnet (Tagle et al. 1988).

Ausganspunkt der phylogenetischen Analyse sind 35.750 verschiedene Transkriptionseinheiten, welche in der RefSeq-Datenbank (UCSC track: refGene, Apr. 14, 2010, hg19) aufgelistet sind (Haubrock, Li et al. 2012). Diese Einheiten beschreiben insgesamt 21.532 verschiedene Gene. Damit lassen sich im Mittel 1,7 Transkripte pro Gen beobachten. Für jede Transkriptionseinheit wird der entsprechende -1000 bp Bereich stromaufwärts des TSS als potenzieller Promoter verwendet. Auf Basis dieser Angaben werden nun die orthologen Sequenzen im WGA (46_Way_MULTIZ, hg19) für Mensch (hg19), Maus (mm9), Hund (canFam2) und Rind (bosTau4) bestimmt. Da die Bindestellenvorhersagen auf Basis von TRANSFAC bzw. MATCH lückenlose (Gap-freie) Sequenzen benötigen, werden die vorhandenen Gaps entfernt und durch die nachfolgenden Nukleotide in den Grenzen eines orthologen Sequenzblocks aufgefüllt (siehe Abbildung 4.7 c). Die Vorhersage der potenziellen Bindestellen wurde auf Basis von 854 Wirbeltier-definierten Matrizen aus der TRANSFAC-Datenbank (Version 2009.4) durchgeführt. Mit diesem Ansatz können so ins-

gesamt $4,3 \cdot 10^9$ verschiedene potenzielle TFBSs in menschlichen Promotorregionen vorhergesagt werden, die sequenzkonserviert und gleichzeitig in den übrigen drei Spezies mit mindestens einem dem minFN-Profil folgenden Schwellenwert vorliegen. Die Verteilung der vom Menschen konservierten MATCH-definierten Bindestellenbewertungen ist für ein Prozent der besten bzw. für alle potenziellen TFBS in Abbildung 4.9 gezeigt. Durch An-

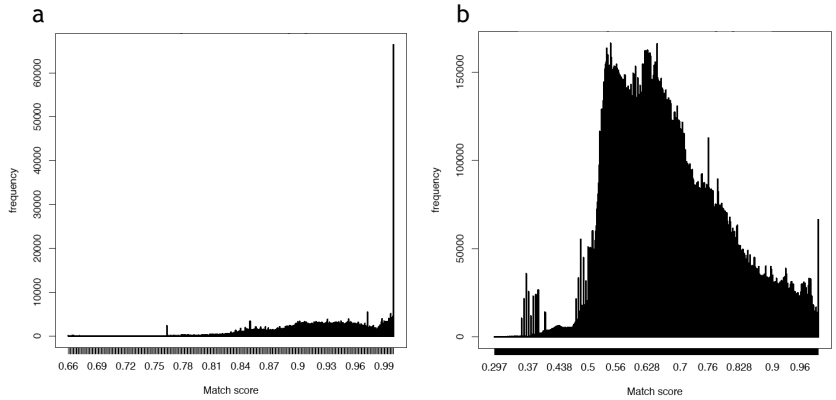


Abbildung 4.9: Verteilung vorhergesagter konservierter potentieller TFBS im Menschen. Gezeigt sind die MATCH-Bewertungen eines Prozents der besten (a) bzw. aller (b) konservierten TFBS Vorhersagen des Menschen. Die konservierten Bindestellenvorhersagen sind durch die gemeinsame Analyse und den anschließenden Vergleich der sequenz- bzw. musterkonservierten TFBS Vorhersagen erstellt worden.

wendung des implementierten Verfahrens kann für 16.900 Gene (78,5 %) des Menschen mindestens eine konservierte TFBS identifiziert werden. Eine Filterung von einem Prozent der besten Vorhersagen pro PWM zeigt immer noch für 15.619 (43,7 %) menschliche Gene mindestens eine hochkonservierte Bindestellenvorhersage. In der Summe beschreibt dieses Filterung 490.277 potenzielle TFBS im menschlichen Genom.

4.2.3 Konstruktion des regulatorischen Transkriptionsnetzwerks

In einem weiteren Schritt kann auf Basis aller vorhergesagten TFBS ein RTN erstellt werden. Die Bindestellenvorhersage basiert auf zwei unterschiedlichen Informationsressourcen: Die Promotorregionen, welche für die Vorhersage verwendet werden, basieren auf RefSeq-definierten Transkriptionseinheiten. In dieser Datenquelle ist verzeichnet, welches

Gen mit welchen unterschiedlichen Transkriptionseinheiten verbunden ist. Die Gennamen werden in der RefSeq-Datenbank mit ihrem offiziellen Gennamen aufgeführt (Braschi et al. 2019). xDie zweite Datenquelle ist die TRANSFAC-Datenbank mit den dort aufgeführten Bindestelleninformationen, die in Form einer PWM zur Verfügung stehen. Für jede dieser PWM ist dort verzeichnet, welche experimentell identifizierten TF-TFBS-Paare Teil der PWM-Definition sind. Für die Erstellung des Referenznetzwerks werden alle in TRANSFAC annotierten TFs einer jeden PWM auf die offiziellen menschlichen Gensymbole zurückgeführt. Diese Übersetzung findet entweder direkt oder indirekt statt: Die direkte Überführung erfolgt, indem alle menschlichen TFs 1:1 auf die menschlichen Gennamen abgebildet werden (Braschi et al. 2019). Dazu wird der offizielle Genname aus der HGNC-Datenbank verwendet (siehe Kapitel 3). Die Gruppe der TFs einer in TRANSFAC-definierten PWM kann aber auch nicht-menschliche TFs enthalten. Die Mehrheit der Wirbeltier-bezogenen regulatorischen Informationen in TRANSFAC stammt aus Experimenten, welche in der Spezies Maus oder Ratte durchgeführt worden sind. Für diese TF wird eine indirekte Übersetzungsstrategie angewendet. Die Angabe orthologer Gene für Maus, Ratte und Mensch ist in der MGI-Datenbank gegeben (Bult et al. 2019). Diese Ressource wird hier verwendet, um das orthologe menschliche Gen zu identifizieren und so die Menge der möglichen TFs einer gegebenen PWM zu vervollständigen. Als Ergebnis dieser Übersetzungsstrategie erhalten wir eine Liste aller vorhergesagten TF-Zielgene, repräsentiert durch das jeweilige HGNC-definierte menschliche Gensymbol.

Die finale TF-Zielgen-Auflistung entspricht damit fast direkt einer sogenannten Adjazenzliste. Diese ist in der Graphentheorie als eine mögliche Datenstruktur bekannt, um eine Menge von Objekten und deren bestehende Verbindungen mathematisch korrekt zu modellieren. Da die transkriptionelle Genregulation ein gerichteter Prozess ist, in dem ein TF die Transkription eines Gens bzw. verschiedene definierte Transkriptionseinheiten eines Gens aktiv reguliert, wird die Adjazenzliste als ein sogenannter gerichteter Graph modelliert. Ein Graph wird durch zwei unterschiedliche Mengen (V, E) beschrieben. Die Menge V beschreibt dabei die Knoten des Graphen (engl. *Vertex*). Das bedeutet hier, dass alle Gene des RTN (Knoten des Netzwerks) durch ihren eindeutigen Gennamen repräsentiert werden. Die Menge der Beziehungen/Kanten (engl. *edge*) zwischen zwei Knoten wird durch die Menge E repräsentiert. Durch das Löschen von möglichen Mehrfacheinträgen entspricht diese Menge dann automatisch der Anforderung der mathematischen Mengendefinition. Die Mehrfacheinträge sind z.B. vorhanden, da für jede Transkriptionsein-

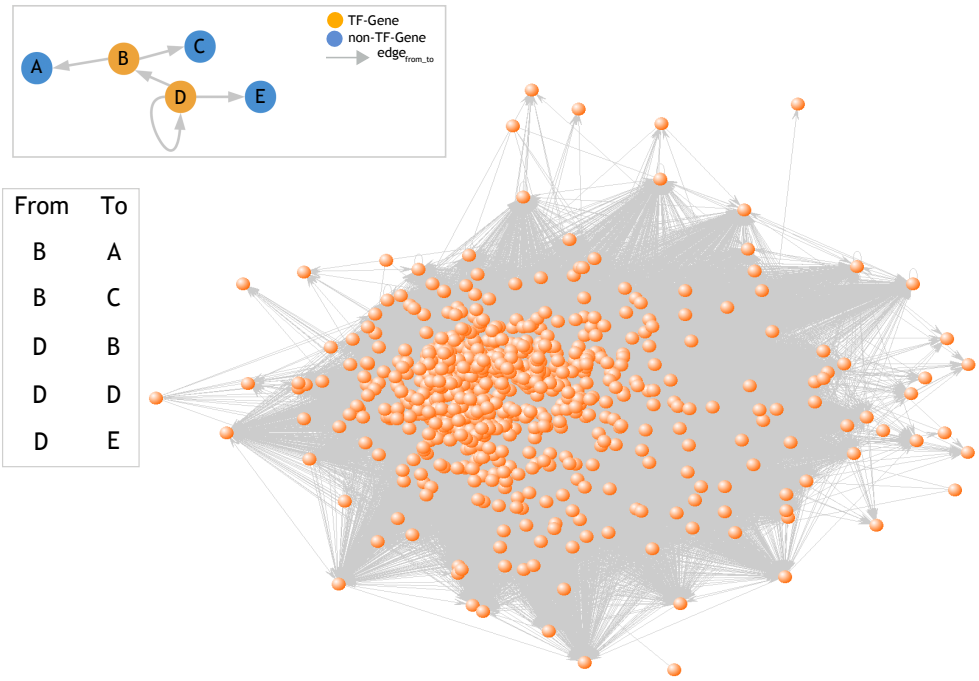


Abbildung 4.10: Allgemeines regulatorisches Transkriptionsnetzwerk der menschlichen TF. Auf Basis eines Prozents der besten Vorhersagen können 442 verschiedene menschliche TF-Gene identifiziert werden, welche sich auf 854 verschiedene PWMs beziehen. Im Durchschnitt lassen sich also zwei PWMs pro TF in der TRANSFAC-Datenbank auffinden. Verschiedene Schätzungen gehen von ungefähr 1500 bis 2000 menschlichen TF-Genen aus (Lambert et al. 2018; Wingender, Schoeps und Dönitz 2013). Ein Grund für die niedrige Anzahl an TFs, die in der TRANSFAC-Datenbank verfügbar sind, ist auf die Qualitätsvorschriften dieser Datenbank zurückzuführen: TRANSFAC verwendet nur TFBSs und fasst diese zu einer PWM zusammen, wenn für sie auch ein experimenteller und publizierter Nachweis existiert (Edgar Wingender 2008). Aus diesem Grund sind für einige TFs noch keine PWMs vorhanden.

heit ein eigener Promotor definiert wurde, für den die Vorhersage der potenziellen Bindestellen erfolgt. Sollte allerdings die Promotordefinition für mehrere Transkriptionseinheiten überlappen, werden mögliche konservierte TFBSs mehrfach vorhergesagt. Zusätzlich existiert die Möglichkeit, dass für einige TFs verschiedene unterschiedliche PWMs in der TRANSFAC-Datenbank vorhanden sind. Für beide Fälle wird bei der Übersetzung der PWM-Gen-Beziehung in eine TF-Zielgen-Zuordnung eine redundante Relation in der Übersetzung erzeugt. Durch die Löschung dieser Mehrfacheinträge kann diese eindeutige Kantenmenge erhalten werden. Die Kantenmenge E beschreibt damit alle eindeutigen Knotenpaare, zwischen denen eine beeinflussende (gerichtete) Beziehung besteht und stellt somit die graphentheoretisch korrekte Definition einer Kantenmenge dar. Die Klasse der TFs (Regulatoren der Transkription) ist genomkodiert. Das bedeutet, dass jeder TF selbst auch Zielgen eines Transkriptionsfaktors ist. Aus diesem Grund wird im zugrundeliegenden Graphen nur ein Knotentyp verwendet. Dieser ist durch die Verwendung des menschlichen Gennamens eindeutig zu identifizieren. Abbildung 4.10 zeigt modellhaft die Situation für das vorliegende Referenznetzwerk: ein Beispielgraph (oben links) und die dazugehörige Adjazenzliste (unten links). Das gezeigte Netzwerk auf der rechten Seite stellt alle regulatorischen Beziehungen innerhalb der Gruppe der TFs dar. Gezeigt ist ein Prozent der besten Bindestellenvorhersagen innerhalb dieser Gruppe. Die Abbildung wurde mit Hilfe des *visANT*-Tools erzeugt (Granger et al. 2016).

4.2.4 Erweiterung des regulatorischen Transkriptionsnetzwerks

Die konservierten Bindestellenvorhersagen bilden die Grundlage für das zu erstellende RTN. Jede vorhergesagte TFBS bezieht sich auf eine in TRANSFAC definierte PWM, welche das Bindestellenmotiv für eine Gruppe verschiedener TFs beschreibt. Für eine PWM aus TRANSFAC kann die Zahl der TFs signifikant erhöht werden. Die Funktionseinheit der DNA-bindenden Domäne (DBD) eines TFs ist der relevante, für die spezifische DNA-Erkennung verantwortliche, Teil des Proteins bzw. Proteinkomplexes. Einige TFs zeigen auf Basis ihrer DBD eine sehr ausgeprägte Sequenzähnlichkeit. Diese Information ist systematisch in der TFClass-Datenbank erfasst und bildet damit die Grundlage für die hierarchische Struktur dieser Datenbank (Wingender, Schoeps und Dönitz 2013). Um die Zahl der im Netzwerk vorhandenen TF-Gene zu erhöhen, wird das TFClass-Projekt als zusätz-

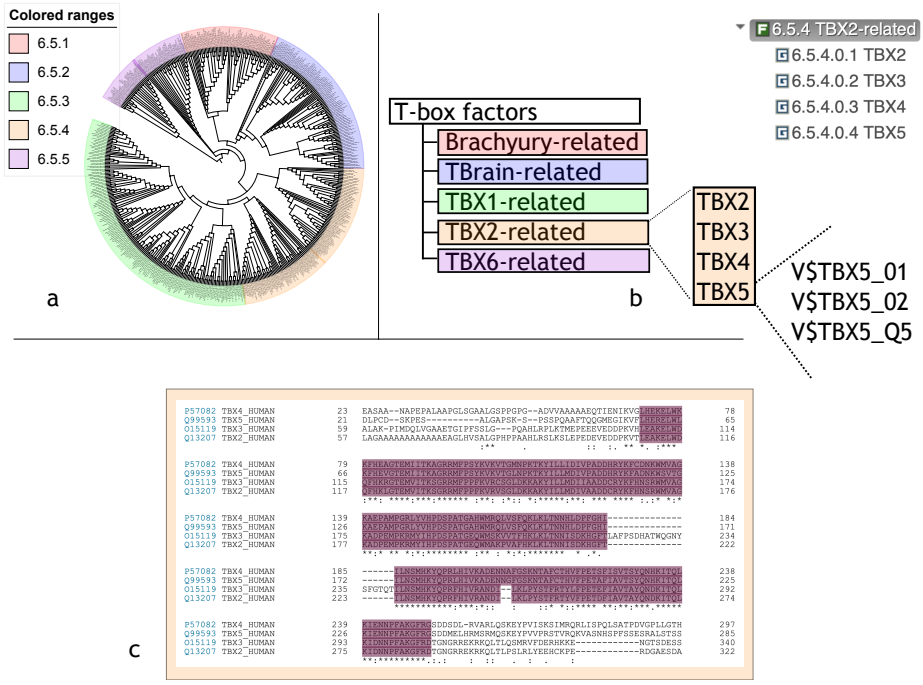


Abbildung 4.11: Hierarchie der T-box Transkriptionsfaktoren des TFClass-Projekts. (a) Darstellung der Sequenzähnlichkeit der TFClass-definierten Klasse 6.5 'T-box factors'). Der phylogenetische Baum zeigt die Sequenzähnlichkeit aller Säugetierproteine dieser Klasse. Die Farbeinteilung entspricht den Familiendefinitionen dieser Klassen (6.5.1-6.5.5). (b) zeigt die Mitglieder (Genera) der Familie 'TBX-2-related' und die Matrizen aus TRANSFAC, welche für die Vorhersage dieser Familie verwendet wurden. (c) zeigt die Domänenkonservierung der menschlichen TFs dieser Familie (TBX2-TBX5). Das Alignment wurde mit Hilfe der UniProt-Webressource erstellt (UniProt Consortium 2019)

liche Informationsressource verwendet: TFClass unterscheidet vier generelle Hierarchieebenen: superclass (Superklasse), class (Klasse), family (Familie) oder subfamily (Unterfamilie). Sie geben die verschiedenen Stufen des Projekts vor, in welche alle verfügbaren Säugetier-TFs des Projekts einsortiert wurden (Wingender, Schoeps, Haubrock, Krull et al. 2018). Auf der Hierarchiestufe der Familien bzw. Unterfamilien sind die verschiedenen DBDs auf Basis der gruppierten DBD-Sequenzalignments kaum mehr unterscheidbar. Diese Gruppen zeigen also häufig eine sehr ausgeprägte Sequenzkonserviertheit für die entsprechenden funktionalen DBDs. Das bedeutet, dass die Sammlung der Proteine der nachfolgenden Ebene in TFClass (Ebene Genus) als funktionsähnlich bei der Erkennung ihrer Bindestellen angesehen werden kann. Existiert nun eine PWM für mindestens einen TF dieser Gruppe in TFClass, so werden alle vorliegenden konservierten Bindestellenvorhersagen dieses TFs auf die gesamten Mitglieder der Familie bzw. Unterfamilie übertragen.

Als Beispiel für diese Netzwerkerweiterung kann die 'TBX-2 related'-Familie aus TFClass benutzt werden (siehe Abbildung 4.11 b). Diese Familie umfasst vier verschiedene Genera (TBX2, TBX3, TBX4 und TBX5). Die TRANSFAC-Datenbank enthält in der verwendeten Version nur PWMs für den Transkriptionsfaktor TBX5. Im Einzelnen sind das die Matrizen: V$TBX5_01, V$TBX5_02 und V$TBX5_Q5. Es werden nun alle vorhergesagten potenziellen TFBSs für diese PWMs um die Faktoren dieser Familie erweitert und gelten somit auch für die Transkriptionsfaktoren TBX2, TBX3 und TBX4. Die Grundlage für die Einteilung in dem TFClass Projekt wird noch einmal in Abbildung 4.11 (a) verdeutlicht. Dort ist die Klasse der 'T box factors' gezeigt (TFCLASS-ID: 6.5), welche aus fünf Familien besteht (6.5.1-6.5.5). Abbildung 4.11 (b) gibt die PWMs auch noch einmal namentlich wieder. Der phylogenetische Baum dieser Klasse zeigt die Einteilung der vorhandenen Säugetierproteine dieser Klasse auf Basis der gesamten Proteinsequenz (Wingender, Schoeps, Haubrock, Krull et al. 2018). Die Farbgebung der Darstellung markiert die Gruppe dieser fünf verschiedenen Familien, welche sich aus dieser Klasse ableiten lassen. Die orange eingefärbte Gruppe repräsentiert dabei die 'TBX-2-bezogene' Familie. Diese Darstellung zeigt noch einmal die Sequenzähnlichkeit dieser Gruppe, welche der verwendeten Einteilung in TFClass als gegeben zugrunde liegt. Da sowohl die Gesamtproteinsequenz als auch die DBDs dieser fünf Unterfamilien durch Sequenzvergleich ausreichend unterschiedlich, aber gleichzeitig die DBDs einer jeden Familie ununterscheidbar sequenzähnlich untereinander sind, wird diese Information als Erweiterung des Referenznetzwerks benutzt.

Durch Anwendung der Domäneninformation entsteht das sogenannte erweiterte regulatorische Transkriptionsnetzwerk (*extended Regulatory Transcription Network* (eRTN)). Die Erweiterung verwendet paraloge Funktionsinformation aus TFClass. Der Begriff 'paralog' beschreibt hier die Eigenschaft von weitgehend funktionsgleichen Genen innerhalb einer Spezies, während 'ortholog' funktionsgleiche Gene zwischen verschiedenen Spezies charakterisiert. Als Ergebnis dieser Erweiterung folgt, dass die Menge des besten einen Prozents der Vorhersagen, bezogen auf jede verwendete PWM, nun von 442 TF-Genen des RTN auf 742 (Erhöhung um 76,9 Prozent) TF-Gene im eRTN ansteigt (Haubrock, Li et al. 2012). Durch diese Expansion wird also eine deutliche Vergrößerung der Kantenanzahl im sogenannten '1-Prozent-Profil'-Netzwerk erreicht (162 Prozent). Die Kantenanzahl steigt dadurch von 277.661 auf 728.667. Durch diese Zahlen wird bereits deutlich, dass die paraloge Erweiterung die Größe des Referenznetzwerks deutlich verändert.

4.2.5 Architektur regulatorischer Transkriptionsnetzwerke

Um die Eigenschaften des RTN bzw. des eRTN zu bestimmen und so Aussagen über die Vergleichbarkeit der beiden Netzwerke zu erhalten, können verschiedene topologische Maße verwendet werden. Einen ersten Eindruck über die Auswirkungen der paralogen Expansion erhält man z.B., wenn die Knotengradverteilung, unterteilt nach ein- (engl. *in-degree*) bzw. ausgehenden (engl. *out-degree*) Kanten, benutzt wird (Lima-Mendez und van Helden 2009). Dazu kann z.B. die *inverse cumulative distribution function* (iCDF) für die Berechnung der Wahrscheinlichkeiten des Auftretens eines bestimmten Knotengrads herangezogen werden (Lima-Mendez und van Helden 2009; Tanaka et al. 2005). Die Knotengradbezogene Verteilungen für die beiden Netzwerke sind in Abbildung 4.12 aufgezeigt. Die vier Abbildungen im Teil (a) geben die Knotengradverteilung des RTNs wieder, während die vier Abbildungen im Teil (b) die Knotengradverteilung des eRTN darstellen. In Abbildung 6 (a) zeigen die oberen beiden Darstellungen die Verteilungen der eingehenden Kanten des regulatorischen Transkriptionsnetzwerks (RTN) sowohl ohne als auch mit logarithmierten Werten der Wahrscheinlichkeiten des Knotengrads (y-Achse). Die beiden unten abgebildeten Kurven zeigen die gleichen Situationen für die ausgehenden Kanten des RTN. Abbildung 4.12 (b) stellt in vergleichbarer Art und Weise die vier Situationen für das erweiterte Netzwerk (eRTN) dar. Insgesamt zeigt sich, dass die beiden regulatori-

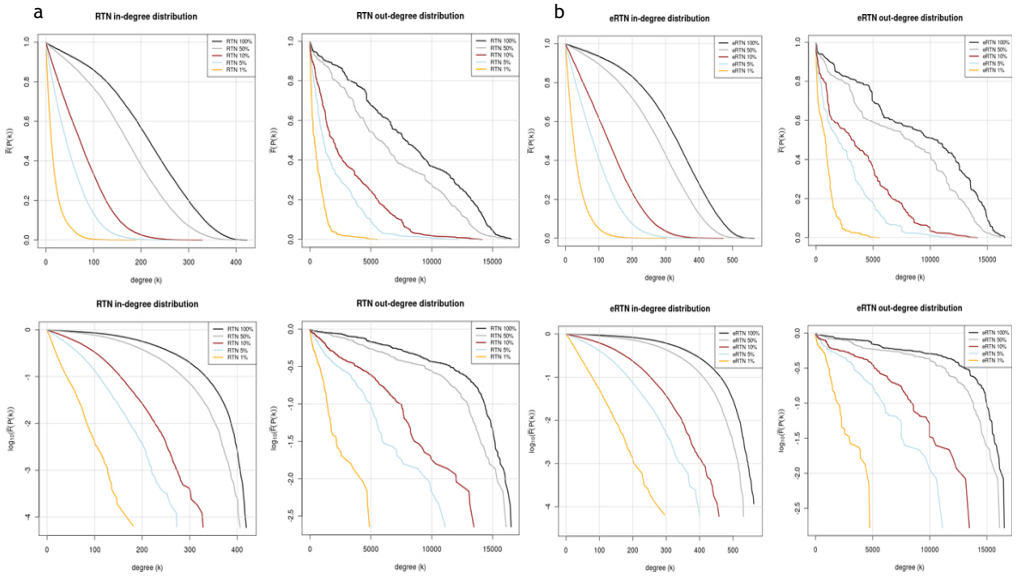


Abbildung 4.12: Knotengrad-Verteilung des regulatorischen Transkriptionsnetzwerks. Dargestellt sind die Knotengradverteilungen, unterteilt nach ein- bzw. ausgehenden Kanten, für das regulatorische Netzwerk RTN (a) und das erweiterte regulatorische Netzwerk eRTN (b). Die verschiedenen Kurven unterscheiden das beste eine Prozent (orange) sowie die besten fünf Prozent (blau), zehn Prozent (rot), 50 Prozent (grau) PWM-Bewertungen bzw. alle Vorhersagen (100 Prozent, schwarz). Die oberen Darstellungen zeigen die kumulative Knotengradverteilung der ein- bzw. ausgehenden Kanten (y-Achse) für einen gegeben Knotengrad (x-Achse). Die unteren Darstellungen zeigen den gleichen Datensatz in semi-logarithmierter Form ($\log_{10}$).

schen Netzwerke eine vergleichbare Knotengradverteilung aufweisen. Auch die Anwendung der paralogen Netzwerkerweiterung verändert nicht die grundsätzliche Netzwerktopologie. Die Skalenfreiheit, welche für verschiedene Arten von biologischen Netzwerken eine häufig beschriebene Eigenschaft ist, zeigt sich in beiden regulatorischen Netzwerken. Am deutlichsten wird diese Eigenschaft in dem einen Prozent der besten PWM-bezogenen Vorhersagen repräsentiert (siehe Abbildung 4.12, orange Kurven). Durch Hinzunahme weiterer konservierter Bindestellenvorhersagen kann eine gewisse Art der Sättigung in den verschiedenen regulatorischen Netzwerken beobachtet werden. Dies betrifft sowohl die Klasse der Regulatoren (TF, ausgehende Kanten) als auch die regulierten Transkriptionseinheiten (eingehende Kanten).

4.2.6 Evaluierung der regulatorischen Transkriptionsnetzwerke

Die Bewertung der vorliegenden Transkriptionsnetzwerke kann nur exemplarisch erfolgen. Es existieren keine experimentellen Datensätze, mit deren Hilfe die Gültigkeit aller vorhergesagten regulatorischen Interaktionen des RTN bzw. eRTN als Ganzes bewertet werden könnten. Für eine Gruppe von TF liegen jedoch verschiedene ChIP-seq-Daten für z.B. unterschiedliche menschliche Zelllinien vor, welche im ENCODE-Projekt erzeugt wurden (siehe Kapitel 3.6. Diese Datensätze können benutzt werden, um beispielhaft das Leistungsverhalten der TFBSs-Vorhersagen auf Basis der z.B. vorliegenden Bindestellenqualitäten zu untersuchen.

Der verwendete MATCH-Algorithmus bewertet eine Bindestellenvorhersage auf Grundlage einer PWM. Je Sequenz-ähnlicher die vorliegende regulatorische Region einer potenziellen Bindestelle zu einer in TRANSFAC definierten PWM ist, desto größer ist der berechnete Ähnlichkeitswert in MATCH. Die MATCH-Bewertung einer Bindestellenvorhersage ist auf den Wertebereich zwischen 0 und 1 skaliert (siehe auch Kapitel 3.3). Das bedeutet, dass eine zu einer PWM perfekt passende Sequenz eine MATCH-Bewertung von 1 erhält, während eine komplett unähnliche DNA-Sequenz einen Wert nahe 0 annimmt. Für eine Bewertung der vorliegenden konservierten Bindestellen auf Grundlage der MATCH-Bewertung ist für jede einzelne PWM aller verwendeten PWMs des Netzwerks eine prozentuale Einteilung (*Perzentile*) vorgenommen worden. Insgesamt wurden

zwanzig verschiedene Perzentile für jede Matrize gebildet. Das erste Perzentil beschreibt dabei die besten Hundert aller konservierten Bindestellen pro PWM. Das zwanzigste Perzentil enthält alle konservierten Vorhersagen (100 Prozent) und die dazwischen liegenden Perzentile beschreiben entsprechende Teilmengen.

Als Evaluierungskriterien werden dabei drei verschiedene Gütekriterien verwendet: der *Positive Predictive Value* (PPV), die *True Positive Rate* (TPR) und die Spezifität *Specificity* (SP). Diese drei Bewertungen können auf Grundlage der wahr-positiven- (*True Positives* (TP)), falsch-positiven- (*False Positives* (FP)), falsch-negativen- (*Fales Negatives* (FN)) und den wahr-negativen- (*True Negatives* (TN)) Klassifikationen auf Basis der vorhergesagten Bindestellendaten und den jeweiligen im ENCODE-Projekt durchgeführten ChIP-seq-Experimenten bestimmt werden. In dieser Analyse wird eine vorhergesagte TFBS, welche in einer experimentellen ChIP-seq-Region gefunden wurde, als TP bewertet. Falls eine vorhergesagte Bindestelle nicht im untersuchten ChIP-seq-Experiment lokalisiert ist, wird diese als FP-Ergebnis gezählt. Ein FN-Ereignis wird beobachtet, wenn eine ChIP-seq-Region existiert, die eine menschliche Promotorregion überlappt, diese aber keine Bindestellenvorhersage für den entsprechenden TF zeigt. Zusätzlich muss die ChIP-seq-Region aber mit mindestens 50 Prozent des experimentellen Bereichs den betrachteten Promotorbereich überlappen. Ein TN-Ereignis liegt vor, falls keine TFBS-Vorhersage und keine ChIP-seq-Region mit mindestens 50-prozentiger Überlappung zwischen dem experimentell identifizierten Bereich und dem Promotorbereich gefunden werden kann.

Das ENCODE-Projekt stellt eine gruppierte Sichtweise aller durchgeführten ChIP-seq-Experimente in einer speziesspezifischen Zusammenfassung zur Verfügung (Cluster von TF gebundenen Regionen). Diese Datei (Dateiname: wgEncodeRegTfbsClusteredV2.bed, Bezugsquelle UCSC) gruppiert die in verschiedenen Zelllinien durchgeführten genomischen Lokalisationen und fasst sie zu einer gemeinsamen übergeordneten regulatorischen Region zusammen. Daraus ergeben sich im Durchschnitt ChIP-seq-Regionen mit einer gemittelten Länge von 250 bp. Auf Basis dieser vier Klassifikationen können nun beispielhaft drei Gütekriterien bestimmt werden. Die benutzten drei Berechnungsvorschriften sind in Tabelle 4.4 angegeben. Zusätzlich ist in dieser Auflistung auch die Berechnungsvorschrift der Korrelation nach Matthews gezeigt. Abbildung 4.13 zeigt den Verlauf dieser drei Gütekriterien für die Transkriptionsfaktoren NFκ und CTCF. Die Berechnung der Spezifität auf Grundlage des NFκ-Datensatzes zeigt eine kontinuierliche Erniedrigung über das gesamte

Positive Predictive Value	$PPV = \frac{TP}{TP+FP}$
Specificity	$SP = \frac{TN}{TN+FP}$
True Positive Rate	$TPR = \frac{TP}{TP+FN}$
Matthews Correlation Coefficient	$MCC = \frac{(TP+TN)-(FP+FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Tabelle 4.4: Angewendete Gütekriterien zur Bewertung des regulatorischen Transkriptionsnetzwerks.

Bewertungsprofil. Für das 100-Prozent-Profil wird eine knapp unter 70 Prozent liegende Spezifität erreicht. Die umgekehrte Reaktion wird für die TPR beobachtet. Dieser Wert steigt von einem sehr kleinen Wert (0,01) im ein-Prozent-Perzentil auf 0,3 für alle Vorhersagen (100 Prozent Perzentil). Der Anteil der PPV-Werte für alle Perzentile bewegt sich im Rahmen von 0,2 bis 0,37. Er ist am größten für die höchsten MATCH-Bewertungen (ein-Prozent-Perzentil). Insgesamt kann für alle untersuchten Datensätze ein ähnlicher Trend beobachtet werden. Für einige TFs, welche eine besondere regulatorische Bedeutung in Stammzellen zukommt, wird ein insgesamt doch sehr niedriger PPV beobachtet (Haubrock, Li et al. 2012). Im Vergleich dazu beobachten wir für den Transkriptionsfaktor CTCF einen PPV von 0,82 und dieser stellt damit den größten Wert aller untersuchten TFs.

Um eine allgemeingültige Bewertungsmöglichkeit über die Bedeutung der verschiedenen Profile zu erhalten, wird die Berechnung des nach Matthews definierten Korrelationskoeffizienten verwendet (siehe Tabelle 4.4). Der *Matthwes Correlation Coefficient* (MCC) wird als ein ausgewogenes Maß betrachtet, selbst wenn die einzelnen Klassifikationen (TP, FP, FN und TN) sehr unterschiedliche Größenordnungen besitzen. Abbildung 4.14 zeigt die verschiedenen Werte des MCC über die zwanzig untersuchten Bewertungsprofile. Es wird deutlich, dass für die ersten zehn Profile der besten Bindestellenbewertungen ein positiver mittlerer MCC zu beobachten ist. Die vier hier durchgeführten Leistungsbewertungen zeigen insgesamt, dass vor allem die hochbewerteten sequenzkonservierten TFBS-Vorhersagen eine besondere Bedeutung besitzen. Die Vermutung besteht, dass diese Bindestellen wichtige regulatorische Module prägen. Diese Bindestellen stellen damit den

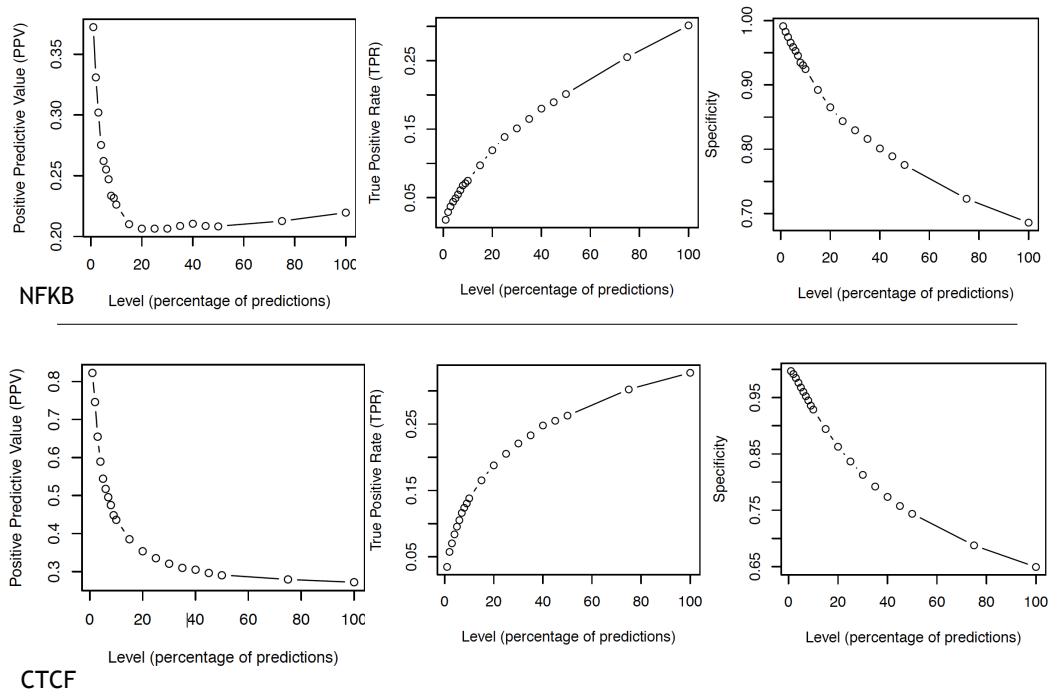


Abbildung 4.13: Leistungsverhalten der konservierten menschlichen Vorhersagen. Die Abbildung zeigt den Positive Predictive Value (PPV), die True Positive Rate (TPR) und die Spezifität der Transkriptionsfaktoren NFκ und CTCF.

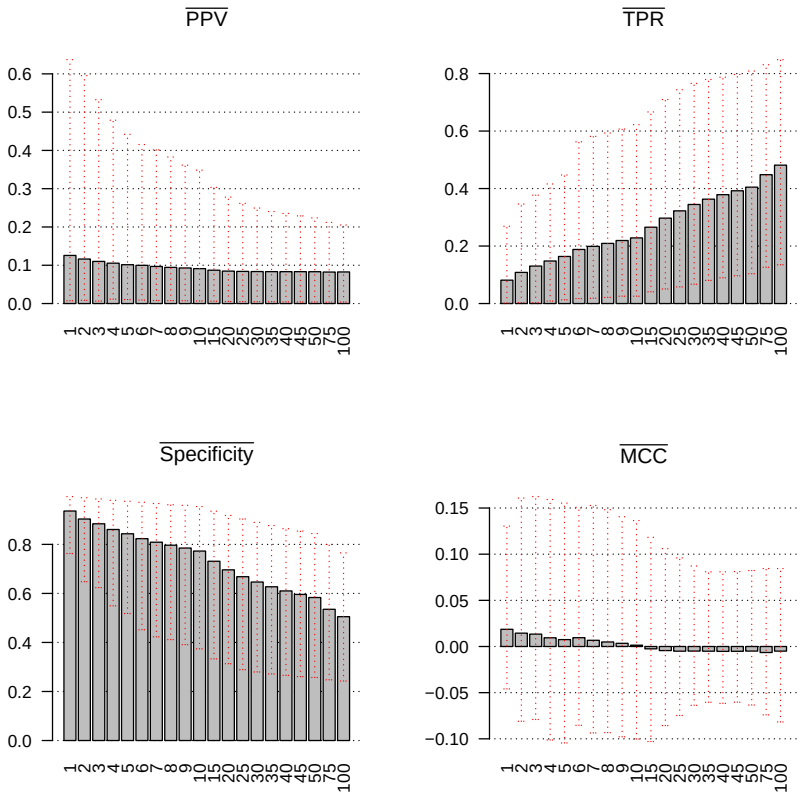


Abbildung 4.14: Evaluierung der RTN durch ENCODE-definierte ChIP-seq-Daten. Bewertung der Bindestellenvorhersagen des RTN auf Basis von ChIP-seq-Experimenten des ENCODE-Projekts für 22 verschiedene menschliche TFs

.

definitorischen Kern dar, welche durch die interagierenden TFs definiert gebunden werden. Insgesamt zeigt aber gerade der MCC-Wert eine sehr niedrige Korrelation. Diese Beobachtung überrascht indes nicht sehr, da die Netzwerke nur aus Promotor-bezogenen Vorhersagen gebildet wurden. Die ChIP-seq-Daten beschreiben aber genomweite regulatorische Regionen und die Mehrheit dieser Regionen werden nicht in Promotor-definierten Regionen gefunden. Außerdem ist zu beachten, dass die verschiedenen Leistungsbewertungen nur auf einer Menge von 22 verschiedenen TFs durchgeführt worden sind. Die einzelnen ChIP-seq-Experimente wiederum wurden nur in einer kleinen Anzahl unterschiedlicher menschlicher Zellen für diese TFs getestet. Diese Tatsache erklärt die hier beobachtete hohe Anzahl an FP-Vorhersagen (Haubrock, Li et al. 2012).

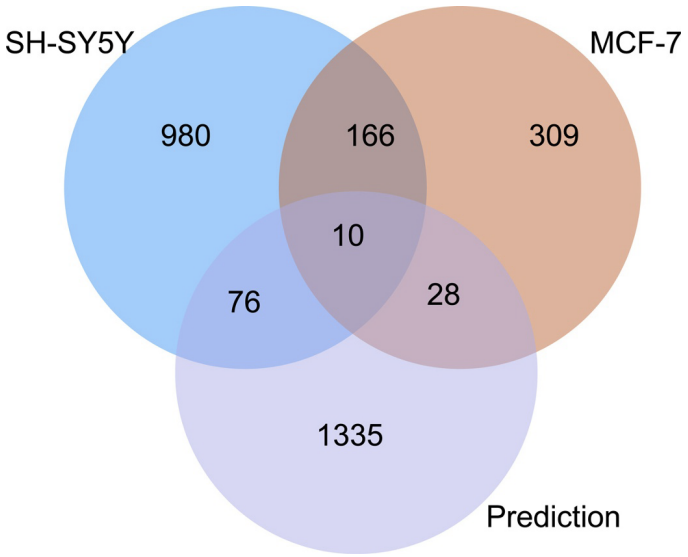


Abbildung 4.15: Verteilung der GATA3 Vorhersagen in menschlichen ChIP-seq Regionen. Das dargestellte Venn-Diagramm vergleicht die verschiedenen ChIP-seq-definierten regulatorischen DNA-Fragmente der menschlichen Zelllinien SH-SY5Y und MCF-7 mit den konservierten Bindestellenvorhersagen des Transkriptionsfaktors GATA3.

Um die vermeintlich hohe Zahl der FP-Vorhersagen besser bewerten zu können, sind für fünf verschiedene menschliche TFs (GATA3, MYC, JUN, MAX und FOS) die Verteilungen der konservierten Bindestellenvorhersage auf Grundlage des ein-Prozent-Perzentils genauer untersucht worden. Abbildung 4.15 zeigt die Situation für den Transkriptions-

faktor GATA3, während Abbildung 4.16 die Situation der restlichen vier TFs zusammenfasst. Der Vergleich verwendet die genomischen Lokalisationen der einzelnen ChIP-seq-Experimente aller vorhergesagten konservierten Bindestellenvorhersagen in einem Mengenvergleich und stellt die Situation in einem sogenannten Venn-Diagramm dar. Die Daten entstammen dem ENCODE-Projekt. Für den Transkriptionsfaktor GATA3 existierten zum Analysezeitpunkt zwei verschiedene Zelllinien: *SH-SY5Y* und *MCF-7*. Insgesamt zeigen 176 genomische Lokalisationen eine experimentell identifizierte Kolokalisation (Überlappung; Bedingung: 50-prozentige Überlappung der Fragmente). Für zehn dieser Fragmente konnte eine konservierte Bindestelle im '1-Prozent-Profil' gefunden werden (Seed). Weitere 76 Seeds überlappen eindeutig mit ChIP-seq-Fragmenten aus der *SH-5YSY* Zelllinie, während 28 weitere Seeds in *MCF-7*-spezifischen ChIP-seq-Fragmenten gefunden werden können. Die verbleibende große Mehrheit von 1335 Seeds für den Transkriptionsfaktor GATA3 kann keiner der beiden untersuchten Zelllinien zugeordnet werden. Ähnliche Situationen sind auch für die anderen vier TFs zu beobachten (siehe Abbildung 4.16). Die

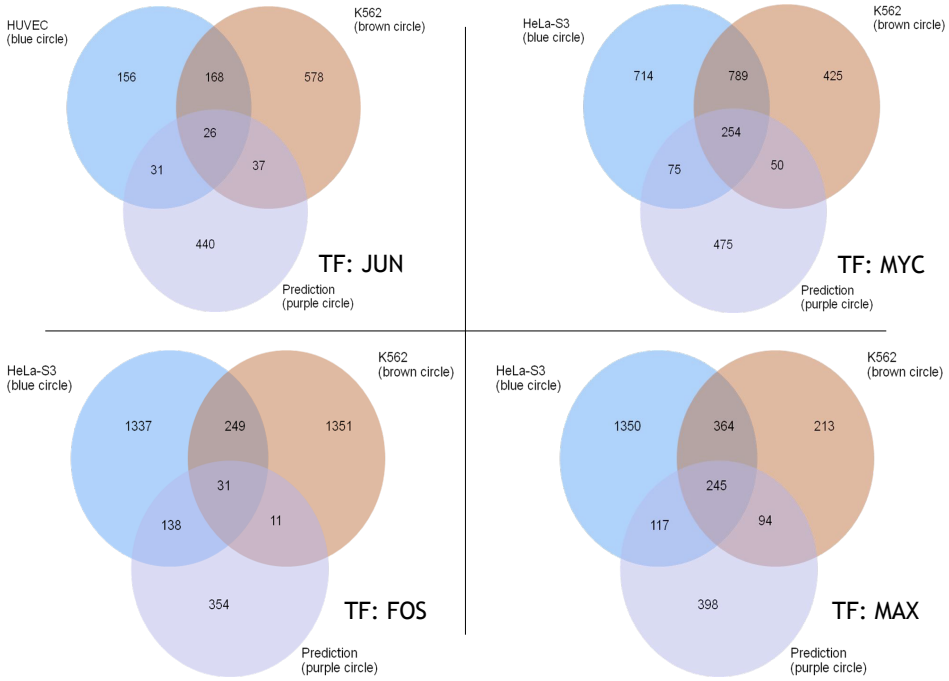


Abbildung 4.16: Zelltypspezifischen Überlappungen der konservierten Bindestellenvorhersagen mit ChIP-seq-Daten aus ENCODE.

Verwendung der Seeds des '1-Prozent-Profils' erfolgt in den fünf Beispielen auf eine zelltypspezifische Art und Weise. Es besteht die Vermutung, dass diese hochbewerteten Seeds wichtige regulatorische Masterbindestellen definieren, welche durch die entsprechenden TFs gebunden werden und so zelltypspezifisch die Transkription der entsprechenden Gene steuern. Unterstützt wird diese Beobachtung durch die Tatsache, dass die Überlappung der ChIP-seq-Regionen mit der Größe der Schnittmenge korreliert: Je größer die Überlappung der ChIP-seq-Regionen aus zwei verschiedenen Zelllinien ist, desto größer ist auch die Schnittmenge zu den vorliegenden Seed-Bindestellen. Da das ENCODE-Projekt nur eine kleine Menge an menschlichen Zellen untersucht, ist zu erwarten, dass die Verfügbarkeit weiterer experimenteller ChIP-seq-Daten den Anteil der Seeds, welche in keinem ENCODE-Datensatz zu finden sind, zunehmend verringern wird. Die vermeintlich beobachtete hohe FP-Rate könnte also durch die geringe Anzahl an zelltypspezifischen ChIP-seq-Daten zu erklären sein.

4.2.7 Rekonstruktion gewebespezifischer Transkriptionsnetzwerke

Bei der Erstellung des RTN stand die allgemeine robuste Abbildung aller möglichen unterschiedlichen regulatorischen Einflüsse auf Basis der verwendeten PWM-Bibliothek im Vordergrund. Die gewebespezifische Interpretation bzw. Verwendung dieses Netzwerks und ein Vergleich dieser Netzwerke untereinander bzw. mit dem allgemeinen Referenznetzwerk wird in diesem Ergebniskapitel untersucht.

Mit der UniGene-Datenbank (Wheeler et al. 2003) steht eine Datenressource zur Verfügung, welche eine einfache Möglichkeit bereitstellt, exprimierte Gene des Menschen für verschiedene Gewebe oder Organe zu erhalten. Auf Basis von EST und mRNA-Sequenzen werden für verschiedene Organismen gewebespezifische Genexpressionsdaten in dieser Datenbank erfasst. Durch Verwendung des *biomaRt* Pakets (siehe Material- und Methodenteil 3) in Kombination mit den eindeutigen Bezeichnern aus der UniGene-Datenbank kann ein UniGene-Eintrag in den offiziellen Gennamen für eine Spezies überführt werden. Auf diesem Wege kann eine direkte Filterung der regulatorischen Kanten des RTN bzw. eRTN erfolgen. Die beiden berechneten regulatorischen Netzwerke liegen als Adjazenzlis-

Typ	**RTN**			**eRTN**		
	TF	**nonTF**	**Kanten**	**TF**	**nonTF**	**Kanten**
Referenz	442	15177	277661	742	15107	728667
Gehirn	343	11750	167560	555	11831	442658
Herz	230	7496	69185	334	7602	169985
Niere	295	9865	118145	447	10044	292301
Leber	253	8662	88575	380	8787	213675
Eierstock	240	7783	78274	366	7986	185617
Prostata	265	9099	99767	411	9136	243260
Milz	161	5266	33962	230	5728	77877
Hoden	282	10468	119221	468	10629	320805

Tabelle 4.5: Angewendete Gütekriterien zur Bewertung des regulatorischen Transkriptionsnetzwerks.

ten vor (siehe oben). Eine Kante im RTN bzw. eRTN wird in ein gewebespezifisches Netzwerk übernommen, falls sowohl der TF als auch das von ihm regulierte Zielgen (TF- oder nicht-TF-Gen) Teile der UniGene-Auflistung für ein untersuchtes Organ/Gewebe sind. So

können die gewebespezifischen Kanten des Netzwerks direkt aus dieser Datenquelle extrahiert werden. Tabelle 4.5 zeigt die Anzahl der TFs, nicht-TFs und die Gesamtkantenanzahl für acht verschiedene Gewebe/Organe des Menschen für das RTN bzw. das eRTN.

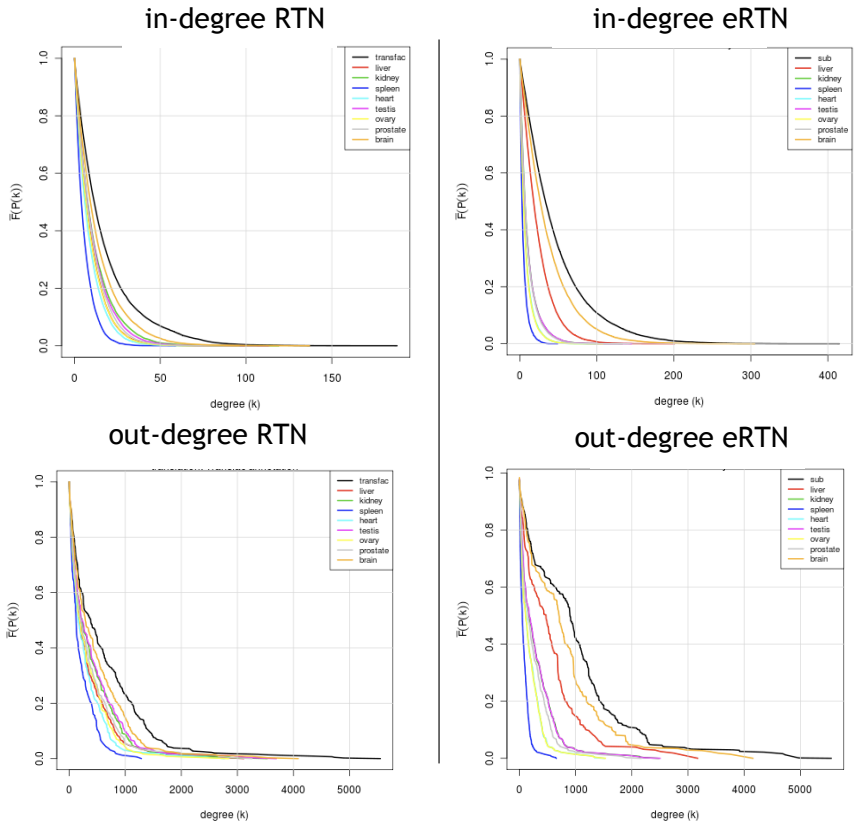


Abbildung 4.17: Knotengradverteilung menschlicher gewebespezifischer regulatorischer Netzwerke. Gezeigt sind acht verschiedene gewebespezifische Rekonstruktionen auf Basis der UniGene-Datenbank und das jeweilige allgemeine regulatorischen Transkriptionsnetzwerks in der Ausprägung des $RTN_{transfac}$ (links) bzw. $eRTN_{sub}$ (rechts). Die oben dargestellten Diagramme zeigen die eingehende Knotengradverteilungen (*in-degree*), die beiden unteren Darstellung beschreiben die ausgehenden Knotengradverteilungen (*out-degree*). Die einzelnen gewebespezifischen Knotengradverteilungen zeigen einen sehr ähnlichen Kurvenverlauf und folgenden damit dem Trend des entsprechenden Referenznetzwerks (schwarze Kurve).

In Abbildung 4.17 ist die Knotengradverteilung, unterteilt nach ein- bzw. ausgehenden Kanten, auf Basis des besten einen Prozents der Vorhersagen gezeigt. Es wurden die folgenden acht Gewebe ausgewählt: Eierstock, Herz, Gehirn, Hoden, Leber, Milz, Niere und

Prostata. Als Orientierung ist auch die jeweilige Verteilung des allgemeinen Referenznetzwerks in den beiden Versionen (RTN, eRTN) in schwarz gezeigt. Die Knotengradverteilungen sind wie in Abbildung 4.12 auf Basis der inversen kumulativen Wahrscheinlichkeiten (y-Achse) in Abhängigkeit zum jeweiligen Knotengrad (x-Achse) berechnet worden (Tanaka et al. 2005). Die Darstellung unterscheidet die Verteilung der ein- bzw. ausgehenden Kanten für das RTN und eRTN. Alle gewebespezifischen Netzwerke folgen in der entsprechenden Knotengradverteilung dem allgemeinen Trend des Referenznetzwerks. Bezogen auf die Anzahl der verwendeten TFs, nicht-TFs und der regulatorischen Kanten ist das rekonstruierte Netzwerk der Milz (engl. *spleen*) das kleinste aller acht untersuchten Netzwerke. Das RTN dieses Organs enthält 161 TF-Knoten und 5266 nicht-TF-Knoten, welche durch 33.962 Kanten miteinander verbunden sind. Im Vergleich dazu besteht das eRTN der Milz aus 230 TF-Knoten und 5728 nicht-TF-Knoten. Diese Knoten sind durch 77.877 verschiedene Kanten miteinander verbunden. Die Daten zum Gehirn (engl. *brain*) zeigen im Vergleich der acht untersuchten Organe die größte Knoten- und Kantenanzahl in beiden Netzwerkausprägungen: 343 bzw. 555 TF-Knoten für das RTN bzw. eRTN; 11750 bzw. 11831 nicht-TF-Knoten können für das RTN bzw. eRTN gefunden werden. 167.560 Kanten werden für das RTN gefunden, während das eRTN im Gehirn durch 442.658 Kanten gebildet wird. Das eRTN im Gehirn verwendet im Vergleich zum RTN 2,6-mal so viele Kanten. Für die Milz beträgt dieser Faktor 2,3.

Die Bedeutung der gewebespezifischen TFs auf Basis der beiden Referenznetzwerke kann bewertet werden, indem das Verhältnis der TF-Gene in einem Gewebe im Vergleich zu allen gewebespezifischen verwendeten TF-Genen dieser acht Gewebe ermittelt wird. Die Ergebnisse dieser Untersuchung sind in Abbildung 4.18 (links) gezeigt. Es wird deutlich, dass durch die Verwendung des eRTNs eine wesentliche Erhöhung der TFs erreicht wird. Für die eTTNs zeigt sich ein Mittelwert von 73,2 Prozent, für die eTTNs (581 TFs) im Vergleich zu 44,5 Prozent im TTN-Gewebe (381 TFs). Wenn man die Verhältniswerte der Knotengrade der eTTNs im Vergleich zu den TTNs näher betrachtet wird deutlich, dass sich die Knotengrade der ein- und ausgehenden Kanten verändern: Der Knotengrad der ausgehenden Kanten, bezogen auf die TFs, erhöht sich in den eTTNs im Vergleich zu den TTNs um das 1,6-Fache. Diese Erhöhung ist über alle acht Gewebe zu beobachten. Die eingehende Kantengradverteilung, unabhängig ob TFs oder nicht-TFs betrachtet werden, erhöht sich sogar um das 2,2- bis 2,6-Fache (siehe Abbildung 12, rechte Seite). Auch dieser Trend lässt sich in den Werten zu allen acht untersuchten Organen beobachten.

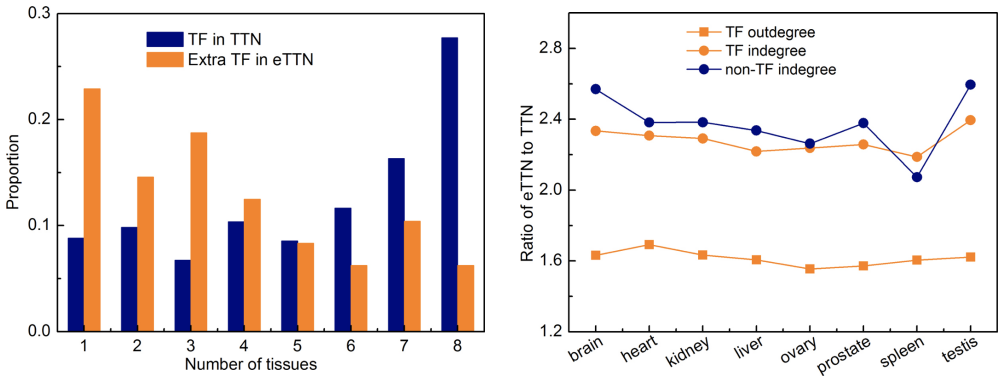


Abbildung 4.18: Vergleich der gewebespezifischen regulatorischen Transkriptionsnetzwerke. Auf der linken Seite ist das Verhältnis der gewebespezifischen TFs bezogen auf die TFs aller acht aus der UniGene-Datenbank rekonstruierten Gewebe gezeigt. Die orange eingefärbten Säulen stellen die gewebespezifischen Netzwerke auf Basis des regulatorischen Transkriptionsnetzwerks dar und werden als *Tissue-specific Transcription Network* (TTN) bezeichnet. Die Verhältniswerte des erweiterten regulatorischen Netzwerks (*extended Tissue-specific Transcription Network* (eTTN)) sind in blauer Farbe dargestellt. Auf der rechten Seite der Abbildung ist das Verhältnis der beiden Situationen bezogen auf die ein- bzw. ausgehenden Kanten gezeigt. Runde Knoten beschreiben die eingehenden Kantenwerte für TFs- (orange) bzw. nicht-TFs (blau). Die viereckigen Knoten zeigen die gewebespezifischen Verhältniswerte für die ausgehenden Kanten. Die Abbildung auf der rechten Seite entstammt einer Veröffentlichung von Li et al. (2012).

4.2.8 Interpretation von Genexpressionsdaten

Die Bestimmung von bedeutenden TFs, welche die Transkription einer Gruppe von Genen reguliert, ist eine häufig verwendete Analysestrategie bei der Interpretation von Genexpressionsdaten. Ausgangspunkt dieser Analyse ist ein Satz an differenziell exprimierten Genen, welche auf Grundlage von Microarray- oder RNA-seq-Experimenten gefunden worden sind und die im Vergleich zweier Bedingungen auffallend unterschiedlich stark exprimiert werden (Anders und Huber 2010; Lönnstedt und Speed 2002; Love et al. 2014). Es existieren verschiedene Methoden, welche anhand einer bekannten Menge an differentiell exprimierten Genen bedeutende TFs auf Basis der regulatorischen Sequenzen dieser Gene bestimmen können.

Das vorliegende regulatorische Transkriptionsnetzwerk kann ebenfalls zur Interpretation von Genexpressionsdaten verwendet werden: Es enthält, wie bereits erwähnt wurde, konservierte TF-Zielgen-Zuordnungen auf Grundlage verschiedener PWMs. Diese vorhergesagten regulatorischen PWM-definierten Kategorien können verwendet werden, um z.B. die Liste der experimentell auffällig exprimierten Gene/Transkripte nach statistisch auffälligen (signifikanten) Überlappungen zu untersuchen. Mit Hilfe der hypergeometrischen Verteilung kann die Wahrscheinlichkeit einer bedeutenden Überlappung auf Grundlage einer Eingabemenge (z.B. Gene aus Expressionsdatenanalyse) mit den vorliegenden PWM-definierten Gengruppen bestimmt werden (Roider et al. 2009).

$$P(k \geq X) = 1 - \sum_{k=0}^{X-1} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Die Wahrscheinlichkeit P einer Überlappung von mindestens X Genen wird mit Hilfe des Binomialkoeffizienten berechnet. N beschreibt dabei die Menge aller Gene, für die eine konservierte Bindestellenvorhersage existiert (Anzahl Gene aus der Datenbank). M ist gleich der Anzahl der Gene der aktuell betrachteten PWM-definierten Kategorie. n gibt die Stichprobengröße an. Mit Hilfe dieser Summenformel unter Anwendung des Binomialkoeffizienten kann nun die Wahrscheinlichkeit für die Überlappung einer PWM-definierten Genliste mit der Eingabemenge berechnet werden. Es werden also die Einzelwahrscheinlichkeiten aller Überlappungen von $X-1$ berechnet und von der Gesamtwahrscheinlichkeit

1 abgezogen. Dies ergibt dann die Wahrscheinlichkeit, mindestens eine Überlappung von X ($k \geq X$) zu beobachten.

Die einzelnen PWMs können in unterschiedliche Bindestellenqualitäten auf Grundlage der Sequenzähnlichkeitsbewertung durch den MATCH-Algorithmus eingeteilt werden. In den nachfolgenden zwei Projekten ist jeweils das sogenannte '1-Prozent-Profil' jeder PWM der verwendeten PWM-Bibliothek aus TRANSFAC für die Analyse verwendet worden. Da die Eingabemenge bei der Anreicherungsanalyse mehrfach genutzt wird (für jede der vorliegenden PWM-definierten Kategorien), muss dies bei der statistischen Bewertung berücksichtigt werden (Noble 2009). Diese Situation ist auch unter dem Begriff des 'multiplen Testens' in der Literatur beschrieben. Durch die Anpassung des p-Wertes kann dieser Fehler (Fehler 1. Art, auch α Fehler genannt) korrigiert werden. Wir verwenden hierzu die durch Benjamini und Yekutieli (2001) vorgeschlagene FDR-Methode: Alle Ergebnisse, die nach der p-Wert-Korrektur kleiner oder gleich dem Signifikanzniveau $\alpha = 0.05$ sind, werden als signifikante PWMs dieses Datensatzes aufgefasst.

Als Anwendungsbeispiel dieser Methode wurden Genexpressionsdaten eines Zeitreihenexperiments ausgewählt, welche für die Brustkrebszelllinie *MCF7* erzeugt wurden (Carroll et al. 2006). Unter Zugabe von Östrogen wurde auf Grundlage von Genaktivitätsmessung mit Hilfe von Microarrays zu 4 verschiedenen Zeitpunkten ($0, 3, 6, 12$ Stunden) die Menge der exprimierten Gene bestimmt, welche durch dieses Hormon direkt oder indirekt aktiviert wurden. Mit Hilfe eines im Institut für Bioinformatik in Göttingen entwickelten Clustering-Verfahrens können nun die koexprimierten Gene bestimmt werden, welche sich in unterschiedlicher Kombination der vier verfügbaren Zeitpunkte bestimmen lassen (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013). Jeder einzelne Koexpressionscluster kann nun mit Hilfe des implementierten Testverfahrens auf die Bedeutung der sequenzkonservierten TFBSs, welche für einen TF oder eine Gruppe von TFs vorhergesagt wurden, untersucht werden. Dieser Zusammenhang ist in den PWM-definierten Gengruppen abgebildet. Tabelle 4.6 zeigt die Ergebnisse dieser Analyse.

Die verwendete Version der TRANSFAC-Datenbank (2009.4) beschreibt 854 verschiedene Positionsgewichtungsmatrizen. Insgesamt liegen mehr als 42 Millionen konservierte Bindestellenvorhersagen auf Grundlage dieser PWM-Bibliothek vor. Für jede dieser PWMs ist das beste eine Prozent aller vorhergesagten TFBS pro PWM bestimmt worden. Da die-

Cluster (Anzahl Gene)	**PWM** (Top 20)	**P-Wert** (FDR korrigiert)
3 (875)	VNCX_02, VMSX1_02, V$PAX4_02, V$POU3F2_01, VTBP_01,VBRN3C_01, V$BARX2_01, V$HB24_02, V$HOXD10_01, V$BARX1_01,V$DBX1_01, V$HMBOX1_01, VHDX_01, VBSX_01, V$NKX52_01,V$HMX3_-02, V$LBX2_01, V$HOXD13_01, V$NFAT1_Q6, V$HOXD8_01	4.29e-08
1 (4477)	VNCX_02, VHDX_01, V$BCL6_01, V$ZNF333_-01, V$DLX2_01,V$DLX7_01, V$DLX5_01, V$SRY_02, V$BARX1_01, V$SOX4_01,V$NKX24_-01, V$HOXD3_01, V$LBX2_01, V$LHX61_02, VSRY_01,VTST1_01, V$DLX3_01, V$XVENT1_-01, V$EVX1_01, V$BARX2_01	1.27e-05
26 (3177)	V$E2F_Q2, V$ZF5_01, V$USF2_Q6, V$SP1_Q6_01, V$KID3_01, V$CHCH_01	2.99e-05
4 (3482)	V$BCL6_01, V$HOXA10_01, VSRY_01, VNKX23_01, V$WT1_Q6,V$HOXB9_01, V$ISL2_01, V$HOXD10_01, V$HOXD8_01, V$NCX_02,V$X1_02, V$PAX4_04, V$BARHL2_01, V$DLX1_01, VSRY_02, VOCT1_03,V$DLX5_01, V$LHX9_01, V$DBX2_01, V$HMGIY_Q6	9.51e-05
2 (2186)	V$CHCH_01, V$MOVOB_01, VMAZ_Q6, VPAX4_03, V$CACD_01,V$GEN_INI3B_B, VGEN_INI_B, VCKROX_Q2	0.0001
12 (476)	V$SRY 02, V$NCX 02, V$BCL6_01, V$HB24_-01, V$HOXA10_01,V$NKX25_02, VSRY_01, VPBX1_02, V$HOXD10_01	0.002
17 (999)	V$CREB_01, V$CREBATF_Q6, V$SP1_Q6_01, V$ATF3_Q6,V$CREBP1CJUN_01	0.004
50 (182)	V$ETF_Q6	0.006
18 (260)	V$STAT1STAT1_Q3	0.042
31 (2465)	V$SP1_Q6_01	0.046

Tabelle 4.6: Bedeutende TRANSFAC-definierte PWMs in Koexpressionscluster in der Brustkrebszelllinie *MCF-7* (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013)

.

se Bindestellen in den Promotoren gegebener RefSeq-definierter Transkriptionseinheiten vorliegen, wird für die Anwendung der Anreicherungsanalyse nun eine eindeutige Genliste erstellt. Diese Liste beschreibt das eine Prozent bester Bindestellenvorhersagen für jede PWM. Durch Anwendung des Testverfahrens auf Basis der hypergeometrischen Verteilung und der Verwendung dieser PWM-definierten Genlisten können nun die signifikanten Überlappungen für jede vorliegende PWM-Menge mit der Eingabemenge bestimmt werden. Die Liste der Eingabemenge wird durch 115 verschiedene koexprimierte Gencluster gebildet (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013). Diese konnten auf Basis der Östrogen-aktivierten Genexpressionsdaten identifiziert werden. Zehn dieser Cluster (neun Prozent) enthalten mindestens eine signifikante Überlappung zu einer der 854 gegebenen PWM-definierten Kategorien, sind also angereichert an Bindestellen für mindestens einen TF, welche der jeweiligen PWM in der TRANSFAC-Datenbank zugeordnet ist. Die Liste der gefundenen TFs kann durch gezieltes Literaturstudium überprüft werden. So zeigt sich z.B., dass die im 'Cluster 26' (siehe Tabelle 4.6) gefundene TRANSFAC-Matrize V$E2F_Q2, welche unter anderem mit dem menschlichen Transkriptionsfaktor E2F1 (UniProt: Q01094) verbunden ist, bekanntermaßen die Zellteilung von Östrogen-induzierten Brustkrebszelllinien auslöst (Stender et al. 2007). Weitere Beispiel sind in der Veröffentlichung von Bhar und Kollegen aufgelistet (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013).
Es bleibt festzuhalten, dass für den Großteil der gefundenen PWMs bzw. den damit verknüpften TFs ein Bezug zu Hormon-induzierten Veränderungen im Kontext dieser Zelllinien oder allgemein mit der Brustkrebserkrankung hergestellt werden kann. Weiterhin kann gezeigt werden, dass einige in der KEGG-Datenbank (Kanehisa et al. 2019) definierte Stoffwechsel- und Signaltransduktionswege mit der Gruppe der differenziell exprimierten Gene pro Zeitpunkt und den signifikanten TFs in Übereinstimmung gebracht werden können (Bhar, Haubrock, Mukhopadhyay, Maulik et al. 2013).

Eine zweite Anwendung dieses Testverfahrens untersucht verschiedene Koexpressionscluster, welche auf Grundlage von Zeitreihendaten berechnet worden sind. Die Genexpressionsdaten entstammen einer veröffentlichten Studie von Babiarz et al. (2012). In dieser werden zwölf verschiedene menschliche Genexpressionsprofile bestimmt, welche unterschiedliche Zeitpunkte in der Entwicklung menschlicher Herzmuskelzellen beschreiben. Die Daten dieser Studie bilden einen Zeitraum von insgesamt 120 Tagen ab. Durch die Weiterentwicklung eines dreidimensionalen Clustering-Verfahrens wurden die unterschiedli-

chen Genexpressionscluster bestimmt (Bhar 2015). Abbildung 4.19 gibt einen Überblick über die experimentelle Studie und fasst die Ergebnisse der Clusteranalyse zusammen. Vier unterschiedliche zusammenhängende Zeitintervalle konnten auf Basis der Clusteranalyse gefunden werden. Auf Grundlage einer Gene-Ontology-definierten Anreichungsanalyse mit Hilfe des R-Pakets *GOstats* (Falcon und R. Gentleman 2007) konnten für diese Intervalle verschiedene biologische Prozesse berechnet werden. Die signifikanten GO-definierten biologischen Prozesse dieser Intervalle sind in Abbildung 4.19 aufgeführt. Für die so qualifizierten Koexpressionscluster können nun mit Hilfe der vorhergesagten konservierten Bindestellenvorhersagen auf Grundlage des hypergeometrischen Testverfahrens angereicherte TFBSs und deren interagierende TF bestimmt werden. Die Datenbasis in

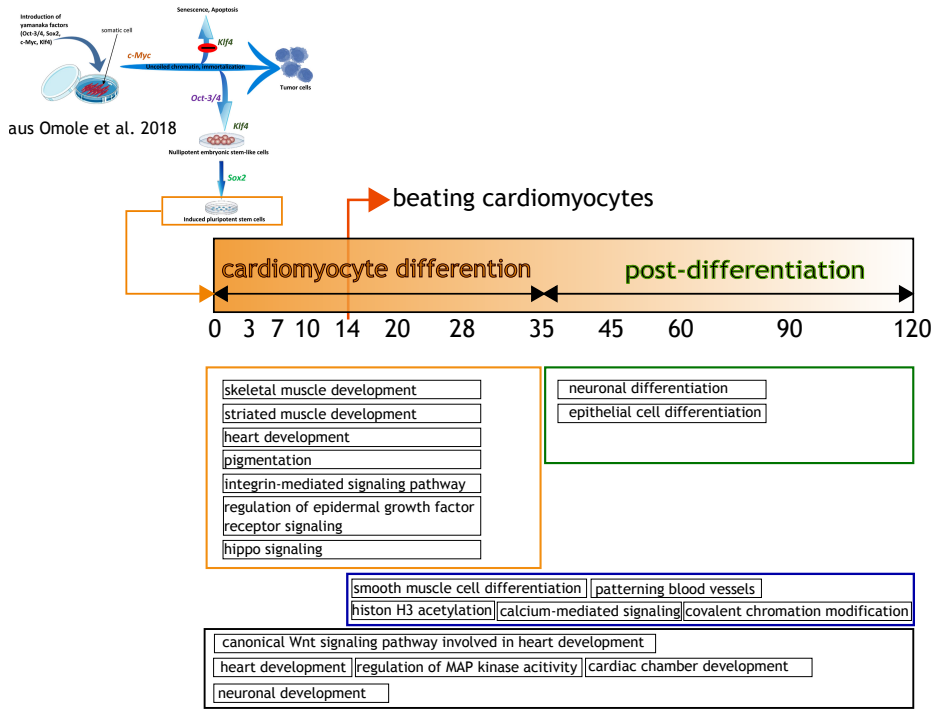


Abbildung 4.19: Zeitreihenanalyse der Herzmuskelzellenentwicklung im Menschen, basierend auf Genexpressionsdaten. Die Abbildung zeigt den Versuchsaufbau der Zeitreihendaten von Babiarz et al. (2012). Ausdifferenzierte (menschliche) Zellen können mit Hilfe definierter Transkriptionsfaktoren in sogenannte pluripotente Stammzellen überführt werden. Aus diesen können dann verschiedene Zellen des Organismus erzeugt werden. Diese Technik ist in der Studie von Babiarz et al. (2012) angewendet worden, um die Herzmuskelzellen in ihrer Entwicklung zu unterschiedlichen Zeitpunkten zu studieren. Die Abbildung oben links wurde aus der Veröffentlichung von Omole und Fakoya (2018) entnommen.

dieser Analyse ist mit Hilfe einer neuen TRANSFAC-Bibliothek (2013.1) durchgeführt worden. Der finale Datensatz beschreibt insgesamt 52 Millionen konservierte Bindestellendaten, von denen in diesem Projekt wiederum das '1-Prozent-Profil' aller Vorhersagen pro PWM verwendet wurde. Die Sequenzkonservierung der Bindestellen zwischen den vier ausgewählten Säugetiergenomen Mensch, Maus, Hund und Rind ist auf Basis der gleichen menschlichen Promotordefinition durchgeführt worden. Für jede dieser Intervall-definierten Gencluster konnten mit Hilfe des implementierten Testverfahrens signifikante angereicherte TFBS-Vorhersagen für verschiedene TRANSFAC-definierte PWMs gefunden werden. Die so identifizierten TF werden in verschiedenen Publikationen als Masterfaktoren der Herzentwicklung beschrieben (Bhar 2015; Bhar, Haubrock, Mukhopadhyay und Edgar Wingender 2015).

4.3 Analyse potenzieller Masterregulatoren

Gewebespezifische Expressionsmuster werden in Vielzellern (Metazoan) durch eine ausgeprägte Kombination verschiedener genomkodierter DNA-definierter Sequenzmotive reguliert (Davidson 2006). Die Bindung dieser Motive mit den entsprechenden sequenzspezifisch interagierenden TFs und deren Kombination zu sogenannten regulatorischen Modulen definieren so eine der ersten Ebenen der eukaryotischen Genregulation. Die ChIP-seq-Technologie ist eine experimentelle Methode, die es erlaubt, definierte Protein-DNA-Interaktionen genomweit zu bestimmen (siehe auch Kapitel 3.9). Die Anwendung dieser Technologie ist gerade für die Gruppe der TFs besonders interessant, da aus den für einen TF experimentell bestimmten regulatorischen Regionen auf die potenziell verwendeten TFBSs geschlossen werden kann. Ein ChIP-seq-Fragment kann eine Länge von mehreren Hundert bp umfassen. Im Unterschied zu der Interpretation von Genexpressionsdaten, bei denen die Definition einer Promotorregion in der Nähe des Transkriptionsstarts nur sehr ungenau erfolgt, stehen durch ein ChIP-seq-Experiment die vorliegenden regulatorischen Sequenzen direkt zur Verfügung. In diesem Teil der Arbeit wird ein Verfahren vorgestellt, welches die Bedeutung aller in der TRANSFAC-Datenbank definierten PWMs in einem ChIP-seq-Experiment bewertet. Grundsätzlich ist dieses Verfahren aber nicht auf die Analyse von ChIP-seq-Daten beschränkt.

4.3.1 AUROC-Analyse von ChIP-seq-Experimenten

Das Auffinden und Bewerten bedeutender PWMs in ChIP-seq-Experimenten wird in diesem Projekt auf Grundlage der Flächenberechnung unter einer ROC-Kurve (*Area Under Reciever Operator Characterstic* (AUROC)) durchgeführt. Eine ROC-basierte Analyse ist eine Technik zur Visualisierung, Organisation und Auswahl eines Klassifikators (Fawcett 2006). Die Anwendung des ROC-Verfahrens wird in dieser Arbeit in folgender Art und Weise durchgeführt: Das ChIP-seq-Experiment liefert die experimentelle Grundlage, also den positiven Satz an zu untersuchenden DNA-Sequenzen. Eine Liste von nicht mit dem ChIP-seq-Experiment überlappenden regulatorischen Regionen stellt den Kontrolldatensatz zur Verfügung und definiert damit den negativen Satz an regulatorischen DNA-

Sequenzen. Für jede dieser MATCH-definierten Grenzen (im Intervall zwischen 0,5 und

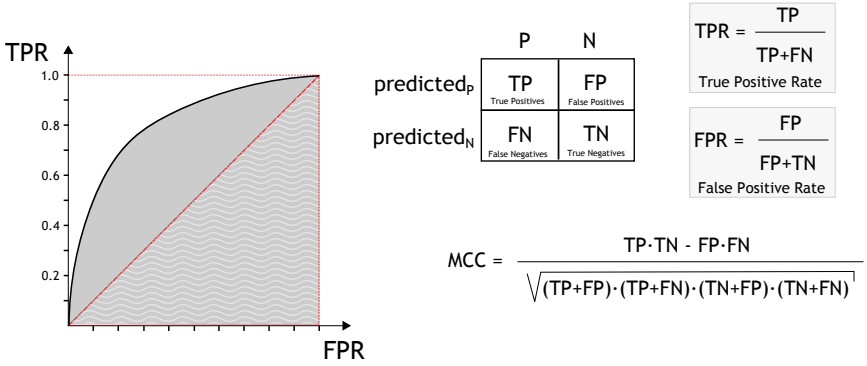


Abbildung 4.20: Beispielhafter Verlauf einer AUROC Kurve.

1,0) wird nun der Anteil der Fragmente bestimmt, die in dem positiven Datensatz gefunden worden sind. Diese Trefferanzahl wird als sogenannter wahr-positiver Wert TP gewertet. Die Fragmente, die keine Bindestellenvorhersage mit dem gegebenem Schwellenwert im Datensatz aufweisen, werden als falsch-negative (FN) Situationen verzeichnet. In der Kontrolle wird gleichzeitig der Anteil der FP und der wahr-negativen TN bestimmt. Der Anteil der FP-Fragmente kennzeichnet also die Anzahl an Kontrollsequenzen, die mindestens eine Vorhersage einer Bindestelle in der gegebenen Qualität besitzen, während sich die TN-Fragmente aus der Anzahl der Kontrollsequenzen zusammensetzen, die keine Bindestellenvorhersage in der vorgegebenen Qualität zeigen.

Durch die Bestimmung dieser vier Werte sind alle Voraussetzungen erfüllt, welche für die Berechnung einer *Reciever Operator Characteristic* (ROC) notwendig sind. Die Auswahl eines geeigneten regulatorischen Kontrolldatensatzes ist für die Analyse der bedeutenden PWMs wichtig. In diesem Projekt werden regulatorische Regionen des gleichen Zelltyps verwendet, welche nicht mit dem ChIP-seq-Datensatz überlappen. Diese Regionen sind für einige menschliche Zelllinien im ENCODE-Projekt bestimmt und basieren z.B. auf *DNase-I Hypersensitive Sites* (DHS) (Hesselberth et al. 2009). Die Bedeutung dieser DHS Bibliotheken als Zusammenfassung aller möglichen regulatorischen Regionen, wurde in verschiedenen Veröffentlichungen gezeigt (Neph et al. 2012; Thurman et al. 2012). Eine ROC zeigt die Abhängigkeit der TPR im Vergleich zur *False Positive Rate* (FPR). Die Formel zur Berechnung dieser beiden Werte ist in Abbildung 4.20 zu sehen. Ein beispielhafter Verlauf einer AUROC ist ebenfalls in dieser Abbildung abgebildet. Die

verschiedenen MATCH-bezogenen Schwellenwerte definieren die einzelnen Punkte einer ROC-Kurve. Mit anderen Worten, jeder der untersuchten MATCH Schwellenwerte ist ein diskreter Klassifikator und bewertet durch das TPR- und FPR-Wertepaar die Qualität dieses Klassifikators für einen gegebenen Schwellenwert. Die hohen MATCH-Bewertungen in der ROC-Darstellung prägen den Beginn der ROC-Kurve (Punkt: 0,0). Mit abnehmendem Schwellenwert werden zunehmend alle als positiv und negativ charakterisierten Fragmente erkannt, so dass die Darstellung im Punkt (1,1) endet (siehe auch Abbildung 4.20). Um eine Kurvenverlauf zu erhalten, werden die einzelnen Klassifikationen verbunden und ergeben so die finale ROC.

Wie schon betont wurde, können durch ein ChIP-seq-Experiment alle genomweiten regulatorischen Regionen bestimmt werden, welche direkt oder indirekt durch das präzipitierte Protein gebunden worden sind. Für die Analyse und Interpretation der Daten ist diese Tatsache von besonderer Bedeutung. Eine direkte Interaktion eines TFs bedeutet, dass der TF sequenzspezifisch an die DNA bindet. Diese Sequenzspezifität äußert sich dann in der Anreicherung dieser Bindestellen für diesen TF in den untersuchten ChIP-seq-Fragmenten. Die TRANSFAC-Datenbank beschreibt die Sequenzspezifität eines TFs in Form von PWMs. Falls die Bindestellen in den Fragmenten des ChIP-seq-Experiments zu einer in TRANSFAC beschriebenen PWM ausreichend ähnlich sind, äußert sich diese Übereinstimmung in einer hohen MATCH-Bewertung der untersuchten PWM für diese Fragmente (Maß der Sequenzähnlichkeit zwischen PWM und DNA-Sequenz). Zeigen also eine Vielzahl von Fragmenten eines ChIP-seq-Experiments im Vergleich zu einem Kontrolldatensatz eine hohe MATCH-Bewertung für eine untersuchte PWM, äußert sich dies in einer hohen AUROC-Bewertung für die untersuchte Matrix. Die mit der PWM verbundenen TFs können somit als potenzielle Masterregulatoren dieses ChIP-seq-Experiments aufgefasst werden. Für den Fall, dass eine PWM sowohl im ChIP-seq-Experiment als auch im Kontrolldatensatz eine ähnliche MATCH-Bewertung zeigt, wird der AUROC-Wert für diese Matrix im Bereich von $0,5$ liegen. Die TFs dieser PWM besitzen also keine ausreichende Bedeutung für diesen Datensatz und zeigen somit keine ausreichende Masterkontrollfunktion. Eine potenzielle Masterkontrollfunktion für TFs des Kontrolldatensatz kann aus AUROC-Bewertungen kleiner $0,5$ geschlossen werden.

Alternativ dazu können auch sogenannte indirekte Masterkontrollfunktionen verschiedener TFs mit Hilfe der AUROC-Bewertung identifiziert werden. Sollte der untersuchte TF

indirekt an eine genomische Region binden, geschieht auch dies nicht unkoordiniert. Die Regulation der Transkription wird durch eine Menge an verschiedenen TFs und Kofaktoren gesteuert. Dieser Proteinkomplex baut sich aber nicht zufällig auf, sondern die zugrunde liegende DNA-Sequenz definiert die möglichen regulatorischen Module. Sollte der untersuchte TF eines ChIP-seq-Experiments in einer definierten, aber unabhängig von seiner eigenen DNA-bindenden Domäne, an die DNA binden, könnte alternativ ein anderer sequenzspezifisch bindender Transkriptionsfaktor die Grundlage dieses Moduls bilden (Protein-Protein-Interaktion). Diese Situation kann ebenfalls durch das hier implementierte AUROC-Verfahren untersucht werden. Für diesen Fall lassen sich möglicherweise auch hohe AUROC-Werte für eine PWM oder eine Menge von PWMs finden, die nicht direkt mit der DNA-bindenden Domäne des präzipitierten TF in Übereinstimmung zu bringen sind.

4.3.2 AUROC Analyse für den Transkriptionsfaktor FOS

Das ENCODE-Projekt stellt alle benötigten Daten zur Verfügung, welche für eine AUROC-basierte Analyse eines ChIP-seq-Experiments notwendig sind. Es sind verschiedene dieser Experimente für eine Vielzahl menschlicher TFs in unterschiedlichen Zelllinien vorhanden. Gleichzeitig stehen für diese gewebespezifischen Daten häufig auch die passenden DHS-Bibliotheken zur Verfügung, aus denen die Kontrolldatensätze generiert werden können (siehe Kapitel 3.10).

Rang	Matrix_ID	AUROC	Sequenzlogo
1	V$NFYC_Q5	0.791	
2	V$NFYA_Q5	0.791	
3	V$NFY_Q6_01	0.789	
4	V$NFY_Q6	0.788	
5	V$NFY_01	0.787	
6	V$YB1_Q4	0.780	
7	V$CAAT_01	0.780	
8	V$ALPHACP1_01	0.777	
9	V$NFY_C	0.771	
10	V$YB1_Q3	0.769	
11	V$LHX8_01	0.740	
12	V$VSX1_03	0.740	
13	V$ISX_01	0.737	
14	V$DUXBL_01	0.735	
15	V$DUXL_01	0.735	
16	V$ACAAT_B	0.733	
17	V$AP1_Q2_01	0.733	
18	V$AP1_Q4	0.732	
19	V$CFOSCUN_Q5	0.732	
20	V$AP1_Q6	0.730	

Tabelle 4.7: AUROC-Analyse für den Transkriptionsfaktor FOS in der Zelllinie *K562*. Gezeigt sind die 20 besten, auf Grundlage der AUROC-Werte sortierten, PWM aus TRANSFAC (2012.2), welche in der *K562*-Zelllinie im Vergleich zu einem DHS-definierten Kontrolldatensatz angereichert sind.

Die nachfolgende Analyse ist für den Transkriptionsfaktor FOS in der Zelllinie *K562* durchgeführt worden. Dieser Datensatz wurde über den UCSC Table-Browser bezogen (GEO: GSM935355). Insgesamt sind dort 7646 genomische Regionen für den Transkriptionsfaktor FOS im menschlichen Genom notiert. Diese Regionen lassen sich zu 7640 eindeutigen Regionen zusammenfassen. Als Kontrolldatensatz kann der DHS-Datensatz aus der gleichen Zelllinie verwendet werden (GEO: GSM816655). Dieser beschreibt insgesamt 202.266 DHS-sensitive Regionen. 194.829 der DHS-Fragmente überlappen nicht mit ChIP-seq-Fragmenten und bilden so den eindeutigen Kontrolldatensatz. Die 7640 Fragmente des ChIP-seq-Experiments und die 194.829 nicht-überlappende genomischen Regionen aus der DHS-Bibliothek bzw. die daraus abgeleiteten DNA-Sequenzen werden nun für die Vorhersage der potenziellen TFBSs verwendet (TRANSFAC-Version 2012.2). In dieser TRANSFAC-Version sind 2181 verschiedene PWM für die Gruppe der Vertebraten vorhanden. Es werden alle mögliche Bindestellen auf Basis des MATCH-Algorithmus für die Matrizen im Wertebereich zwischen $0,5$ und 1 vorhergesagt. Diese beiden Datensätze werden mit Hilfe des implementierten AUROC-Verfahrens analysiert. Tabelle 4.7 zeigt das Ergebnis dieser Analyse. Dort sind die 20 größtenAUROC-Werte dieses Datensatzes aufgeführt. Die Tabelle ist nach dem AUROC-Wert sortiert und listet die Matrix-ID (eindeutiger Name der PWM), den AUROC-Wert und das Sequenzlogo der PWM auf. Auffällig ist, dass die ersten 16 PWMs keine Bindestellenmotive der bZIP-Familie sind, zu denen der untersuchte Transkriptionsfactor FOS zuzuordnen ist.

Die ersten 10 PWMs definieren das sequenzspezifische Bindestellenmuster für den Transkriptionsfaktor NF-Y (`TFClass-ID: 4.2.1 Heteromeric CCAAT-binding`). Die Positionen 11-15 der Tabelle beschreiben Sequenzmotive der HOX-Proteinfamilie (`TFClass-ID: 3.1.1`). Erst die letzten vier Einträge der Tabelle dieser FOS-basierten Datenanalyse beziehen sich auf eine Transkriptionsfaktorklasse, zu der auch die Bindestellen des Transkriptionsfaktors FOS zugeordnet werden können. In Tabelle 4.7 sind zusätzlich zu den AUROC-Werten auch die Sequenzlogos dargestellt. Der Vergleich der ersten 16 Sequenzlogos untereinander zeigt, dass eine große Sequenzähnlichkeit der beschriebenen Sequenzmotive innerhalb dieser Gruppe von Positionsgewichtungsmatrizen vorliegt. Die Mehrheit dieser Matrizen beschreibt die Konsensussequenz `CCAAT`. Zwei dieser Matrizen zeigen das reverse Komplement dieses Konsensus: `ATTGG`. Dazu gehören die Matrizen V$NFY_C (Rang 9) und V$ACCAT_B (Rang 16). Somit repräsentieren die ersten 16 Ränge dieser Auflistung angereicherte potenzielle TFBSs, welche nicht direkt

mit dem präzipitierten Transkriptionsfaktor FOS des analysierten ChIP-seq-Experiments in Verbindung zu bringen sind. Auf Grundlage dieser Ergebnisse kann eine mögliche Protein-Protein-Interaktion zwischen den Transkriptionsfaktoren NF-Y und FOS angenommen werden: Der Transkriptionsfaktor NF-Y bindet sequenzspezifisch an die DNA und der Transkriptionsfaktor FOS bindet an diesen Proteinkomplex. Die Ränge 17 bis 20 der Tabelle 4.7 zeigen aber, dass auch der Transkriptionsfaktorkomplex AP-1, welcher durch diese Matrizen beschrieben wird, eine bedeutende Rolle in diesem Datensatz spielt. Diese beiden Beobachtungen machen deutlich, dass der Transkriptionsfaktor FOS in mindestens zwei verschiedenen Modi mit seinen experimentell gefundenen regulatorischen Regionen binden kann.

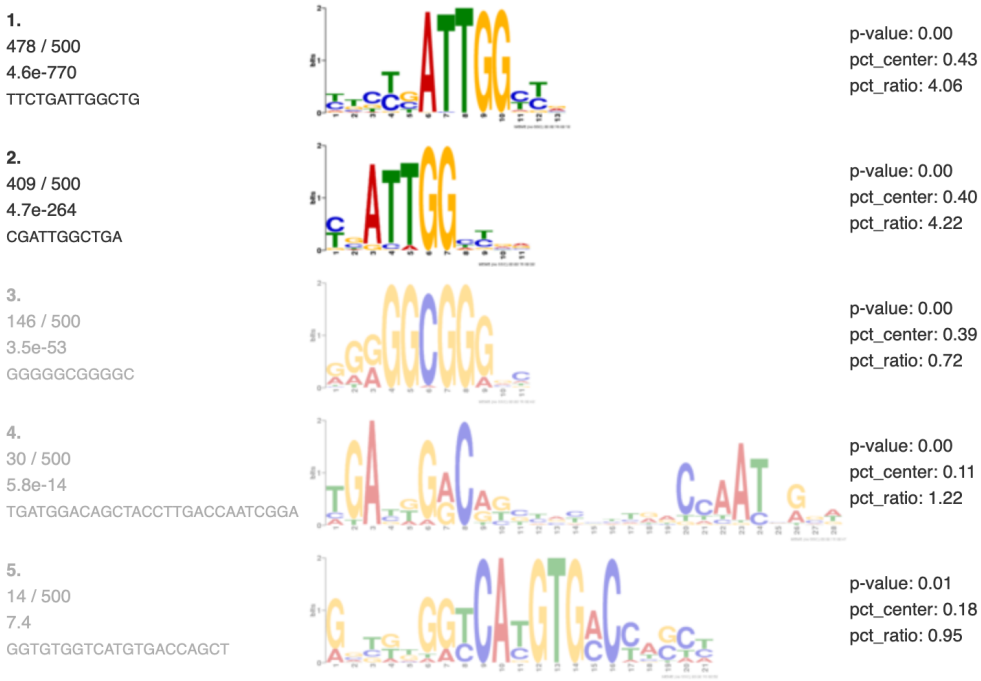


Abbildung 4.21: Proximale und distale AUROC-Analyse von FOS-präzipitierten ChIP-seq-Experimenten.

Um die Ergebnisse dieser AUROC-Analyse bewerten zu können, werden sie mit zwei unterschiedlichen Ressourcen verglichen. Die Webressource Factorbook zeigt verschiedene Eigenschaften von im ENCODE-Projekt erstellten ChIP-seq-Datensätzen (J. Wang

et al. 2012). Dieses Projekt analysiert unter anderem die angereicherten Sequenzmotive auf Basis der besten 500 ChIP-seq-Regionen einiger ENCODE-Experimente durch Anwendung des MEME-Algorithmus (Timothy L. Bailey et al. 2009) und repräsentiert die Ergebnisse der Analyse in Form verschiedener PWM und einer Bewertung dieser auf Grundlage von berechneten Wahrscheinlichkeiten (P-Werte). Abbildung 4.21 zeigt

Name	**ID**	**Lokal_PV**	**Global_PV**	**COR**	**POS**	**POS_PV**
NFYA	MA0060.1	0	0	0.0126	[3,13]	1.12832e-42
NFYB	MA0502.1	0	0	0.0058	[2,12]	4.39034e-48
NFYA	MA0060.2	0	0	0.0068	[4,14]	1.8722e-33
Klf4	MA0039.2	9.92153e-30	0	0.0623	[40,50]	1
KLF5	MA0599.1	5.64931e-20	0	0.0532	[50,60]	1
SP2	MA0516.1	3.20504e-19	0	0.0542	[62,72]	0.0055959
SP1	MA0079.3	1.03668e-13	0	0.0532	[60,70]	0.00184424
EGR1	MA0162.2	2.74547e-07	0	0.06	[-63,-53]	0.0997676
SP1	MA0079.2	0.000322602	0	0.058	[50,60]	0.00621518
Ets1	MA0098.2	0.016821	0.000395519	0.0024	[57,67]	0.0201454

Tabelle 4.8: Pscan-ChIP-Anreicherungsanalyse des Transkriptionsfaktors FOS in *K562*. Die Tabelle zeigt die besten zehn PWMs aus TRANSFAC (2014.1), die mit Hilfe der PscanChIP-Ressoure für den FOS-Datensatz aus ENCODE in K562 berechnet wurden. Die Tabelle wurde nach dem globalen P-Wert sortiert.

das Ergebnis dieser Analyse aus Factorbook für den von uns untersuchten Datensatz (URL: http://www.factorbook.org/human/chipseq/tf/FOS). Das beste Sequenzmotiv beschreibt ebenfalls die Konsensussequenz `ATTGG` (Top 1 und Top 2). Das reverse Komplement dieser Sequenz lautet `CCAAT`. Die weiteren drei Motivbeschreibungen (Rang 3 bis 5) zeigen eine nicht ausreichende Signifikanz, welche durch den MEME-Algorithmus berechnet wird. Diese wird durch die transparente Darstellung verdeutlicht (siehe Abbildung 4.21). Wie bereits in der AUROC-Analyse zeigt sich auch hier das klassische Bindestellenmotiv des NF-Y-Transkriptionsfaktors als dominantes Sequenzmotiv dieses Datensatzes und bestätigt damit die Ergebnisse der AUROC-Analyse.

Durch das Programm PscanChip (Zambelli et al. 2013) erfolgte eine zweite Evaluierung des FOS Datensatzes. Das Programm ermöglicht das Auffinden von überrepräsentierten TF-Bindungsmotiven und deren Korrelationen in Sequenzen aus ChIP-seq-Experimenten. Tabelle 4.8 zeigt die besten zehn Ergebnisse dieser Ressource für den von uns untersuchten FOS-ChIP-seq-Datensatz. Auch hier wird die Bedeutung des Transkriptionsfaktors NF-

Y deutlich. Die ersten drei Positionen der Ergebnistabelle werden durch PWMs dieses TF belegt. Außerdem zeigen sich sechs weitere PWMs aus der Superklasse der Zink-koordinierenden Transkriptionsfaktoren (`TFClass-ID: 2`) und auf Position zehn wird eine PWM geführt, welche der Helix-Turn-Helix-Superklasse (`TFClass-ID: 3`) zugeordnet werden kann. Unter den zehn besten PWMs lassen sich mit Hilfe dieser Ressource keine Matrizen finden, welche mit dem Transkriptionsfaktor FOS in Verbindung zu bringen sind. In den nachfolgenden 10 Positionen werden diese Matrizen dann aber gefunden (Daten nicht gezeigt).

Das hier vorgestellte AUROC-Verfahren in Kombination mit in der TRANSFAC-Datenbank gespeicherten PWMs kann verwendet werden, um ChIP-seq-Experimente zu analysieren. Es liefert im Vergleich zu anderen Methoden vergleichbare Ergebnisse. Die Beobachtung, dass nicht ausschließlich bZIP-PWMs in einem FOS-ChIP-seq-Datensatz auf den ersten Plätzen zu finden sind, ist nicht zufällig und kann auch durch andere Programme bestätigt werden. Diese Beobachtungen werden in den nachfolgenden Ergebnisteilen weiter vertieft.

4.3.3 Zelltypspezifische Analyse für den Transkriptionsfaktor FOS

Das ENCODE-Projekt stellt für den Transkriptionsfaktor FOS vier verschiedene Zelltyps-pezifische Datensätze zur Verfügung: *HUVEC*, *K562*, *HeLa S3* und *GM12878*. Für diese Zelllinien existieren sowohl ChIP-seq- als auch DHS-Datensätze. Um die generelle Bedeutung des Transkriptionsfaktors NF-Y in FOS-regulierten regulatorischen Regionen im Detail zu untersuchen, werden für diese Zelllinien die entsprechenden AUROC-Analysen einzeln durchgeführt und miteinander verglichen. Um die möglichen unterschiedlichen Einflüsse der distalen (Enhancer-bezogenen) und proximalen (promotor-bezogenen) regulatorischen Regionen unterscheiden zu können, wird jedes einzelne untersuchte ChIP-seq-Experiment in unterschiedliche regulatorische Kategorien unterteilt. Die Kategorien werden durch die bekannten, in der RefSeq-Datenbank definierten, Gene bzw. deren Transkriptionseinheiten definiert (O'Leary et al. 2016). Insgesamt werden vier verschiedene Kategorien unterschieden: Die erste Klasse wird durch die Gruppe der potenziellen Enhancer gebildet. Diese Gruppe charakterisiert genomische Bereiche, die nicht mit dem

Promoter (-1000 bp Bereich) oder einem in der RefSeq-Datenbank definierten transkriptionsaktiven Bereich überlappen. Die zweite Klasse wird durch alle bekannten potenziellen Promotoren gebildet. Dabei wird die 1000 bp stromaufwärts eines TSS gelegene Region als potentielle Promotorregion aufgefasst. Der TSS wird durch die RefSeq-Datenbank definiert. Die letzten beiden Kategorien heißen Exon bzw. Intron. Ein Exon beschreibt den

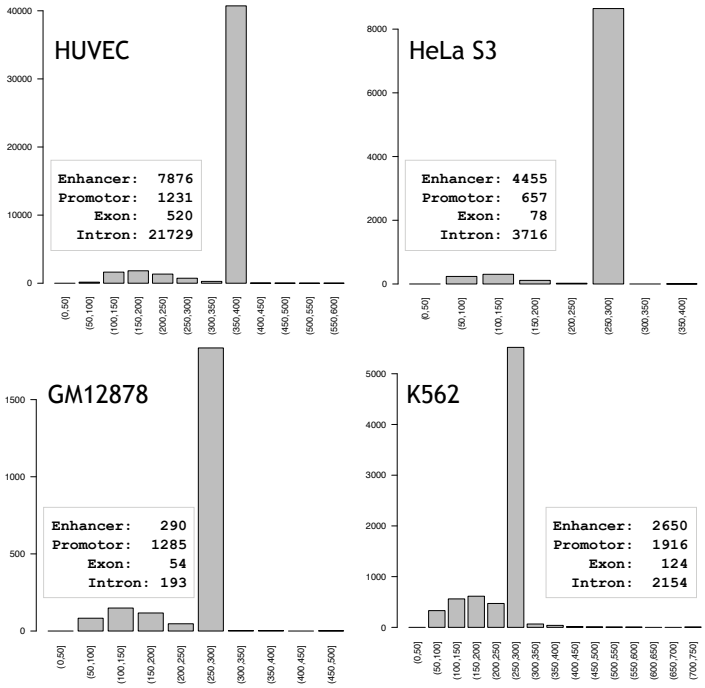


Abbildung 4.22: Zelltypspezifische Situation für den Transkriptionsfaktor FOS in vier verschiedenen Zelllinien. Die Abbildung zeigt die Längenverteilung der ChIP-seq-Datensätze vierer Zelllinien *HUVEC*, *HeLa S3*, *GM12878* und *K562*. In den jeweiligen Kästen ist die Verteilung der Fragmente in die vier von uns betrachteten Kategorien aufgelistet (Enhancer, Promotor, Exon und Intron).

Teil der kodierenden Region eines Transkripts, welcher während der Translation in eine Aminosäuresequenz übersetzt wird. Ein Transkript enthält direkt nach der Transkription sowohl Exon-, als auch Intron-Bereiche. Die Introns werden durch einen nachfolgenden Reifungsschritt herausgeschnitten, so dass die finale mRNA-Sequenz nur noch aus Exons besteht. Die Introns definieren die vierte Kategorie. Die Überlappungen eines ChIP-seq-Fragments mit einer Intron- bzw. Exon-Region muss jeweils innerhalb der durch diese beiden Gruppen definierten Grenzen erfolgen. Eine Promotorüberlappung kann die Grenzen überschreiten. Dabei muss die Überlappung allerdings zu mindestens 50 Prozent mit

einem Promotorbereich erfolgen. Die Charakterisierung eines ChIP-seq-Fragments als Enhancer erfolgt durch die überlappungsfreie Abbildung eines Fragments mit allen weiteren genomischen Regionen, welche nicht als Promotor, Exon oder Intron klassifiziert worden sind. Abbildung 4.22 zeigt die Ausgangssituation für die vier von uns betrachteten Zelllinien. Die Darstellung zeigt die Anzahl der Fragmente und deren Längenverteilung aller vier untersuchten Zelltypen. In den kleinen Fenstern sind die Überlappungssituationen der vier unterschiedlichen Kategorien gezeigt. Die Mehrheit der ChIP-seq-Fragmente

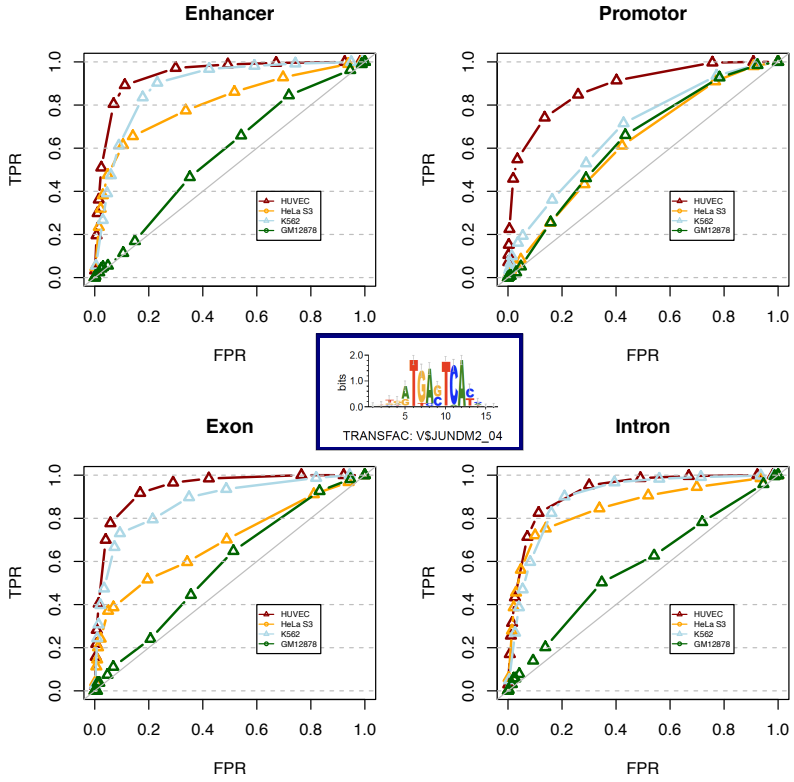


Abbildung 4.23: Zelltypspezifische Bedeutung von AP-1-Bindestellen in FOS-definierten ChIP-seq-Experimenten.

lässt sich in der Gruppe der potenziellen Enhancer finden. Bis auf die *GM12878*-Zelllinie ist die Gruppe der distalen regulatorischen Bereiche größer als die der proximale Gruppe (Promotor-Gruppe). Auffällig ist, dass die *GM12878*-Zelllinie eine sehr kleine Anzahl an FOS-gebundenen Enhancern aufweist (290 Fragmente). Bemerkenswert ist außerdem die hohe Anzahl an Intron-definierten Fragmenten in der *HUVEC*-Zelllinie, die derart ausgeprägt nicht in den anderen drei Datensätzen zu beobachten ist. Mit mehr als 21.000

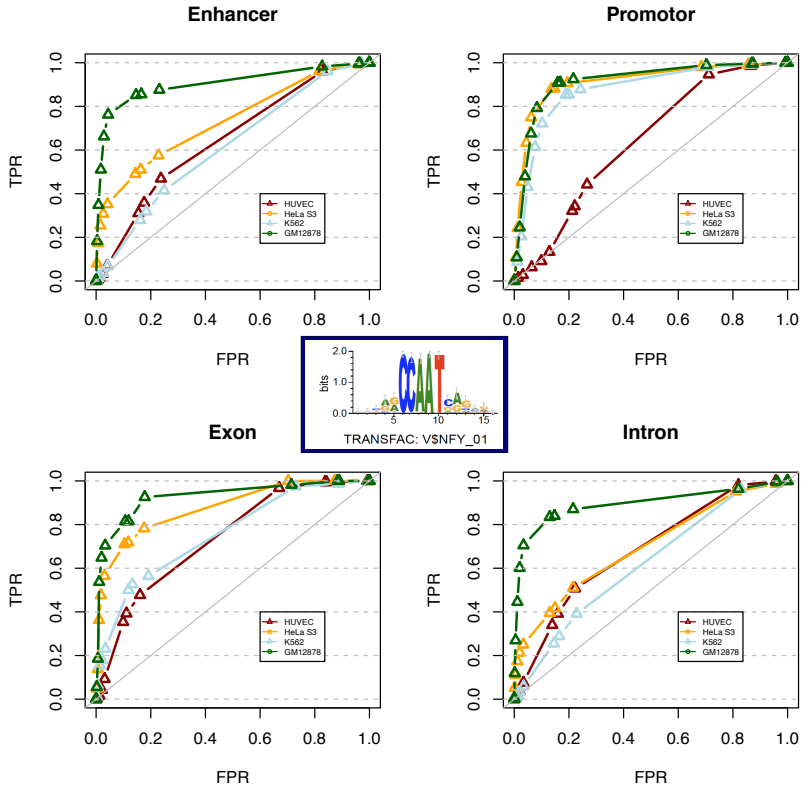
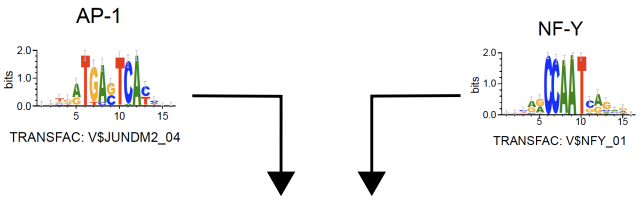


Abbildung 4.24: Zelltypspezifische Bedeutung von NFY-Bindestellen in FOS-gebundenen ChIP-seq-Experimenten.

Fragmenten ist diese Gruppe 2,8-mal so groß wie die Enhancer-definierte Gruppe dieser Zelllinie. Aber auch in den *HeLa S3*- und *K562*-Zellen ist diese Kategorie größer als die jeweilige proximale Gruppe in diesen Zellen. Die proximalen Fragmente sind mit 657 Fragmenten in der *HeLa S3*-Zelllinie am geringsten und mit 1916 Fragmenten in der *K562*-Zelllinie am größten. Auf Basis der AUROC-Analyse konnten für die vier Zelllinien die besten zwei Matrizen aus der TRANSFAC-Datenbank bestimmt werden, die in allen vier ChIP-seq-Experimenten das beste Leistungsverhalten zeigen (siehe auch Abbildung 4.25). Wie aus den Voruntersuchungen bereits vermutet werden konnte, handelt es sich um Matrizen, welche Bindestellen des Transkriptionsfaktors AP-1 bzw. NF-Y beschreiben. Als AP-1-bezogene Matrize wurde die TRANSFAC-PWM mit dem Namen V$JUNDM2_-04 gefunden und als NF-Y bezogene beste PWM aus TRANSFAC konnte die Matrize V$NFY_01 identifiziert werden. Diese beiden Matrizen zeigen in allen vier Datensätzen ein auffälliges Verhalten: In der Gruppe der Enhancer wird die AP-1-bezogene Matrize als

die bedeutendste aller 2176 Matrizen dieses TRANSFAC-Profils (2013.3) in den Zelllinien *HUVEC*, *K562* und *HeLa S3* gefunden, während für Zellen aus *GM12878* kein besonderer regulatorischer Einfluss in den Enhancer dieser Zelllinie messbar ist (siehe Abbildung 4.23, Kategorie Enhancer, grüne Kurve). Dafür steht die ausgewählte NF-Y-definierte PWM aus TRANSFAC in *GM12878*-Zellen an Platz eins der AUROC-sortierten Liste in der Gruppe der Enhancer (siehe Abbildung 4.24, Kategorie Enhancer, grüner Kurve). Das Verhalten ist so auch in der Gruppe der Introns- und Exons-definierten ChIP-seq-Fragmente zu beobachten. Bei den Promotoren dreht sich das Bild um. Dort zeigen ChIP-seq-Fragmente aus den Zelllinien *K562*, *HeLa S3* und *GM12878* einen deutlichen Einfluss von potenziellen NF-Y Bindestellen, welcher durch die hohen AUROC-Werte der NF-Y-Matrize deutlich wird. Einzige Ausnahme hier ist die *HUVEC*-Zelllinie (Abbildung 4.23, rote Kurve). Die pro-



class	cell-line	MEME motif	AP-1$_{AUC}$	NF-Y$_{AUC}$
proximal	HUVEC	3.9e-192 411/657	0.884[1]	0.639
	HeLa S3	1.3e-654 531/657	0.685	0.878[2]
	K562	7.8e-741 578/657	0.623	0.915[1]
	GM12878	7.1e-786 595/657	0.634	0.916[1]
distal	HUVEC	4.4e-366 657/657	0.942[1]	0.646
	HeLa S3	1.8e-683 657/657	0.891[1]	0.614
	K562	4.0e-431 410/657	0.810[1]	0.731
	GM12878	1.1e-319 247/290	0.575	0.904[1]

Abbildung 4.25: Proximale und distale AUROC-Analyse von FOS-präzipitierten ChIP-seq-Experimenten.

ximalen ChIP-seq-Fragmente dieser Zelllinie sind durch AP-1-Bindestellenvorhersagen gekennzeichnet. Abbildung 4.25 stellt die Ergebnisse der AUROC-Analyse noch einmal im Vergleich dar. Die dort gezeigten Matrizen für die Faktoren AP-1 und NF-Y zeigen

die LogoPlots der beiden ausgewählten TRANSFAC-Matrizen für diese Untersuchung. Auf der rechten Seite sind die AUROC-Werte in den proximalen und distalen ChIP-seq-Fragmenten gezeigt; die hochgestellten Zahlen repräsentieren den Rang der beiden Matrizen in den jeweiligen Untersuchungen. Um einen unabhängigen Vergleich zu bekommen, wurde in einer vergleichenden Analyse auf Grundlage von MEME das beste Sequenzmotiv für jeden Datensatz bestimmt. Ausgewählt wurden jeweils 657 ChIP-seq-Fragmente einer jeden Gruppe. Diese Zahl entspricht der Fragmentanzahl der proximalen Gruppe in der *HeLa S3*-Zelllinie. Diese Anzahl wurde gewählt, um ein akzeptables Laufzeitverhalten des MEME-Algorithmus zu erhalten. Die Ergebnisse der MEME-basierten Analyse sind ebenfalls in dieser Auflistung erfasst (Spalte: MEME motif). Der Vergleich der beiden Untersuchungen bestätigt die Beteiligung der beiden Faktoren AP-1 und NF-Y. Die proximalen, also Promotor-definierten, Fragmente zeigen einen deutlichen Einfluss des Faktors NF-Y in den Zelllinien *HeLa S3*, *K562* und *GM12878*. Die distalen Fragmente dieser Zelllinien werden mehrheitlich durch AP-1-Bindestellen geprägt. Im Unterschied zu diesen drei Zelllinien ist die *HUVEC*-Zelllinie in den proximalen und distalen FOS-ChIP-seq-Fragmenten durch AP-1-Bindestellen gekennzeichnet. Die MEME-bezogenen Kontrollanalysen bestätigten die AUROC-Analysen.

4.3.4 Schwellenwertbestimmung für Positionsgewichtungsmatrizen

Mit Hilfe des nach Matthews benannten sogenannten MCC (Matthews 1975) ist es möglich, für eine vorliegende PWM auf der Grundlage einer Menge an regulatorischen Sequenzen (positiver Datensatz) und einem Satz an geeigneten Kontrollsequenzen (negativer Datensatz) einen optimalen Schwellenwert zu bestimmen. Die Formel des MCCs ist

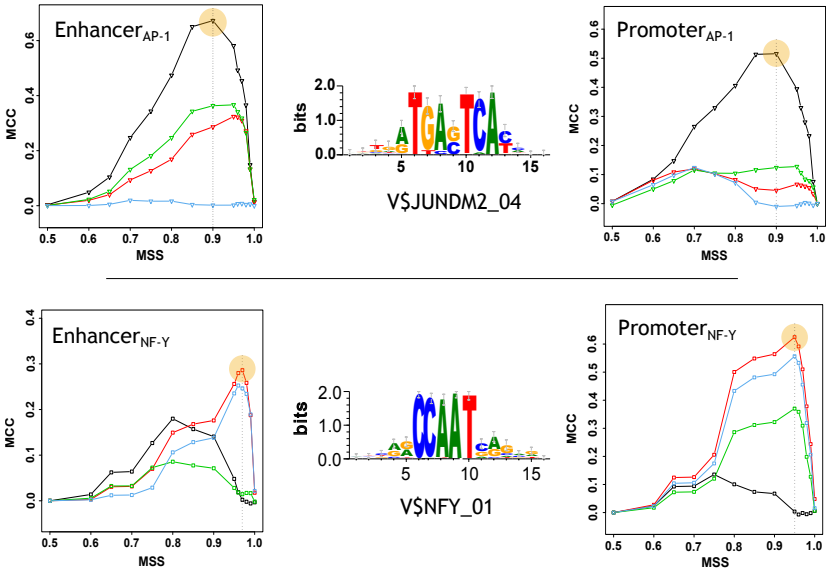


Abbildung 4.26: Schwellenwertdefinition einer PWM durch die Anwendung des MCC. Für die prägenden Bindestellensituationen von AP-1 bzw. NF-Y können innerhalb der jeweiligen Matrizensituationen vergleichbare optimale MATCH-definierte Schwellenwerte beobachtet werden. Diese Schwellenwerte zeigen z.B. in der NF-Y-Situation (siehe unten) einen über verschiedene Zelltypen hinweg vergleichbaren optimalen Schwellenwert. Die vier verschiedenen Farben zeigen den Verlauf des MCC in den Zelllinien *HUVEC* (schwarz), *HeLa S3* (grün), *K562* (rot) und *GM12878* (blau). Die optimalen Schwellenwerte sind auch für die unterschiedlichen regulatorischen Regionen (z.B. Enhancer- und Promotorregionen) vergleichbar (siehe oben, schwarzer Kurvenverlauf). Auf der x-Achse ist die MATCH-Score-Bewertung gezeigt, die Höhe des MCCs ist auf der y-Achse verzeichnet.

in Abbildung 4.20 dargestellt. Die Abbildung 4.20 zeigt diese Situation für die Matrize V$JUNDM2_04 und V$NFY_01 aus TRANSFAC für den Datensatz der distal- (a) bzw. der proximal-definierten (b) ChIP-seq-Fragmente für FOS-ChIP-seq-Datensätze der Zelllinien *HUVEC* (schwarz), *HeLa S3* (grün), *K562* (rot) und *GM12878* (blau).

Die Abbildung 4.26 zeigt die Bedeutung des MCC bei der Bestimmung der Modul-prägenden Seeds. Das wesentliche Kennzeichen einer Seed-Bindestelle ist eine ausgeprägte Sequenzähnlichkeit der Bindestelle zu einer gegebenen PWM. Die Ähnlichkeit kann z.B. durch MATCH bewertet werden. Die Abbildung 4.26 zeigt das Profil des MCC (y-Achse) im Abhängigkeit des angewendeten MATCH-Schwellenwertes (x-Achse, Wertebereich $0,5-1$). Es wird deutlich, dass der maximale MCC für eine bedeutende PWM in den untersuchten regulatorischen Regionen der vier Zelllinien eine vergleichbar hohe Match-Bewertung aufweist (siehe Abbildung 4.26, gelbe Markierungen). Die FOS-präzipitierten Promotoren zeigen z.B. für die Zelllinien *HeLa S3*, *K562* und *GM12878* ein Anreicherung von Kernbindestellen des Transkriptionsfaktor NF-Y. Für diese drei Zelllinien kann der gleiche optimale MATCH-Schwellenwert für die PWM V$NFY_01 von 0.945 beobachtet werden. Die Promotoren der FOS-präzipitierten HUVEC-Zelllinie zeigen keine Anreicherung von NF-Y-Seeds. Im Vergleich dazu weisen nur die Promotoren der *HUVEC* Zelllinie Seed-Bindestellen für den Transkriptionsfaktor AP-1 auf (siehe Abbildung 4.26, oben rechts). Als Konsequenz aus diesen Beobachtungen liegt die Vermutung nahe, dass eine prägende Bindestelle in einer regulatorischen Region eine bestimmte Ähnlichkeitsbewertung erfüllen sollte, anhand derer sich dann im Nachgang die beteiligten TF zuerst direkt und in einem zweiten Schritt dann auch indirekt anlagern können.

4.4 Beziehung der Masterregulatoren NF-Y und AP-1

ChIP-seq-Experimente erfassen detailliert die Belegung des Chromatins für DNA-bindende Proteine. Im vorherigen Ergebnisteil wurde eine Analysemöglichkeit basierend auf der AUROC-Bewertung vorgestellt, welche für eine gegebene Sammlung von PWM in einem ChIP-seq-Experiment die bedeutenden TFs in einem Datensatz bestimmen kann. Die Vergleiche mehrerer speziesspezifischer ChIP-seq-Bibliotheken für verschiedene TFs zeigen häufig ein komplexes kombinatorisches und kontextabhängiges Überlappungsverhalten (Kolokalisationen). Als Beispiel dafür kann die Situation der im ENCODE-Projekt erzeugten Fos-ChIP-seq-Experimente herangezogen werden, welche im vorherigen Kapitel untersucht wurden. Als Ergebnis der AUROC-Analyse konnte eine potenzielle Masterkontrollfunktion des Transkriptionsfaktors NF-Y für FOS-gebundene regulatorische Regionen gezeigt werden. In diesem Ergebniskapitel wird der beobachtete Zusammenhang zwischen dem Transkriptionsfaktor FOS/AP-1 und dem Transkriptionsfaktor NF-Y weiter untersucht. Die Ergebnisse sind in der Zeitschrift PLOS ONE veröffentlicht worden (Haubrock, Hartmann et al. 2016).

4.4.1 NF-Y-definierte Bindestellen-Architektur in FOS-gebundenen regulatorischen Regionen

Die Bedeutung des NF-Y-Bindestellenmotivs in FOS-präzipitierten regulatorischen Regionen ist bereits im vorherigen Ergebnisteil vorgestellt worden. Mit Ausnahme der *HUVEC*-Zelllinie zeigen sich in den proximalen (Promotor) regulatorischen Regionen der Zelllinien *HeLa S3*, *K562* und *GM12878* deutliche Motiv-Anreicherungen für den Transkriptionsfaktor NF-Y. Die distalen FOS-präzipitierten Regionen sind bis auf die *GM12878*-Zelllinie durch keine hochbewertete Bindestellenvorhersagen dieses TFs gekennzeichnet. Um die Bedeutung von NF-Y in FOS-ChIP-seq-Regionen bewerten zu können, wird die Überlappungssituation der beiden TFs näher untersucht. Es zeigen sich verschiedene Datensätze in ENCODE, welche für den Transkriptionsfaktor NF-Y durchgeführt wurden. Da eine zelltypspezifische Unterscheidung zu beobachten ist, sollten die verfügbaren Datensätze möglichst in den gleichen untersuchten Zelllinien durchgeführt worden sein. Der Transkripti-

onsfaktor NF-Y ist ein Heterotrimer und besteht aus den Teilproteinen NFYA, NFYB und NFYC. Jede dieser drei Komponenten ist für die DNA-spezifische Erkennung der Bindestellen (Konsensussequenz `CCAAT`) mitverantwortlich (Sinha, Maity et al. 1995). NF-Y ist in allen Eukaryoten konserviert (Romier et al. 2003). NFYA enthält die DNA-bindende Domäne, welche für die sequenzspezifische Erkennung des CCAAT-Motivs verantwortlich ist, aber auch NFYB und NFYC bilden unspezifische DNA-Kontakte aus (Kim et al. 1996; Sinha, Kim et al. 1996; Zemzoumi et al. 1999). In ENCODE konnten für drei der vier untersuchten Zelllinien ChIP-seq-Experimente für die Faktoren NYYA und NFYB gefunden werden. Für den Transkriptionsfaktor NFYB existiert die größte Anzahl an Fragmenten, so dass die ChIP-seq-Experimente dieses TF für die Kolokalisationsanalyse verwendet wurden. Tabelle 4.9 zeigt die gefundenen Überlappungen für die Zelllinien *HeLa S3*, *K562* und *GM12878*. Für die *HUVEC*-Zelllinie ist kein ChIP-seq-Experiment für den Transkriptionsfaktor NFYB in ENCODE vorhanden.

Sequenztyp	**Zelllinie**	**NFYB**	**FOS**
proximal	*HUVEC*	-	1231
	HeLa S3	2037	657
	K562	2296	1912
	GM12878	2924	1285
distal	*HUVEC*	-	20181
	HeLa S3	2385	4455
	K562	4409	2649
	GM12878	4828	290

Tabelle 4.9: Distale und proximale Fos-präzipitierte Regionen aus dem ENCODE-Projekt. Gezeigt sind genomischen Regionen, welche durch FOS bzw. NFYB gebunden wurden. Im ENCODE-Projekt konnten insgesamt vier verschiedene Zelllinien identifiziert werden, die gleichzeitig ChIP-seq- und DHS-Daten beschreiben. Für die *HUVEC*-Zelllinie liegen keine NFYB-ChIP-seq-Daten vor, so dass die Felder für diese Spalte mit einem Bindestrich (-) markiert wurden.

Die Auflistung unterscheidet nach proximalen und distalen Überlappungen und zeigt die regulatorischen Bereiche der Transkriptionsfaktoren NFYB und FOS. Der Transkriptionsfaktor NFYB interagiert im Vergleich zu FOS in den drei Zelllinien mit einer größeren Anzahl an proximalen Fragmenten. In der distalen Situation werden in den Zelllinien *GM12878* und *K562* viel häufiger NFYB interagierende Bereiche gebunden, als das für den Transkriptionsfaktor FOS zu beobachten ist. Nur in der *HeLa S3*-Zelllinie lassen sich im Vergleich zu NFYB wesentlich mehr FOS interagierende Regionen finden. Auf Basis der in Tabelle 4.9 aufgelisteten regulatorischen Regionen der beiden Faktoren können nun

die überlappenden ChIP-seq-Fragmente bestimmt werden. Es zeigt sich, dass die meisten der proximalen FOS-Regionen auch durch NFYB gebunden sind (*HeLa S3*: 74,7 Prozent, *K562* 79,0 Prozent, *GM12878*: 85,7 Prozent).

Zelllinie	Typ	FOS	FOS+NFYB	NFYB
HeLa S3	p	166 (7,6%)	491 (22,3%)	1540 (70,1%)
	d	4104 (63,3%)	351 (5,4%)	2030 (31,3%)
K562	p	401 (14,9%)	1511 (56,1%)	779 (28,9%)
	d	1745 (28,4%)	904 (14,7%)	3488 (56,8%)
GM12878	p	184 (6,0%)	1101 (35,8%)	1790 (58,2%)
	d	39 (1,0%)	251 (5,2%)	4560 (94,0%)

Tabelle 4.10: Überlappung der FOS- und NFYB-präzipitierten ENCODE-ChIP-seq-Datensätze. Proximale (p) bzw. distale (d) genomische Intervalle werden durch diese Analyse in nur FOS-, FOS+NFYB- und nur NFYB-bindend unterteilt. Die Bedeutung der einzelnen Situationen wurde durch den prozentualen Wert bezogen auf die Summe aller Fragmente berechnet.

Die genomischen Intervalle, die ausschließlich mit FOS interagieren, werden im Weiteren als FOS(only) bezeichnet. Für eine Teilmenge dieser Fragmente existiert eine hochaffine Bindestellenvorhersage für ein AP-1-bezogenes Sequenzmotiv. Die Daten-getriebene Schwellenwertdefinition wurde mit Hilfe des MCC realisiert (siehe 4.3.4). Durch die Anwendung dieses Schwellenwertes bei der Bindestellenvorhersage der ausgewählten TFs kann die Anzahl der Fragmente bestimmt werden, welche eine ausreichend hohe Bindestellenqualität für diesen TF zeigt. Um die genomischen Regionen zu benennen, für die eine hochaffine Bindestellenvorhersage existiert, wird der Namenszusatz (+) verwendet. Die Regionen, für die dies nicht zutrifft, werden mit dem Namenszusatz (-) markiert. Da das Bindestellenmotiv für NFY für die hoch-bewerteten V$NFY_01 Situationen direkt mit der Konsensussequenz `CCAAT` korreliert, werden im Folgenden diese Situationen mit (CCAAT+) bzw. (CCAAT-) für Vorhersagen unterhalb dieser Schwelle bezeichnet. Tabelle 4.11 und Tabelle 4.12 fassen die Motivausstattungen der FOS(only)-Gruppe für diese beiden TRANSFAC-Matrizen für die proximale und distale Gruppe zusammen. Die beiden Tabellen zeigen, dass das Hauptmerkmal von FOS(only, proximal)-Regionen das Vorhandensein von CCAAT-Box-Motiven ist, während die FOS(only, distal)-Regionen hauptsächlich durch ein AP-1-Motiv gekennzeichnet sind. Dieses Missverhältnis wird durch die entsprechenden Odds- bzw. Odds-Verhältniswerte (*Odds Ratio* (OR)) für alle drei Zelllinien unterstützt.

Zelllinie	Typ	AP-1+	AP-1−	Odds	OR
HELA S3	p	85	81	1,049	0,276
	d	3478	626	5,556	
K562	p	77	324	0,238	0,038
	d	1505	240	6,271	
GM12878	p	5	179	0,028	0,008
	d	31	8	3,875	

Tabelle 4.11: AP-1-Motiv in FOS-gebundenen proximalen Regionen. Gezeigt sind Fragment-bezogenen Häufigkeiten des AP-1-Motivs für FOS-gebundene proximale Regionen.

Zelllinie	Typ	CCAAT+	CCAAT−	Odds	OR
HELA S3	p	52	114	0,325	49,890
	d	37	4067	0,009	
K562	p	203	198	1,025	32,622
	d	53	1692	1,139	
GM12878	p	98	86	1,139	42,838
	d	1	38	0,026	

Tabelle 4.12: Häufigkeit des `CCAAT`-Motivs in FOS-gebundenen proximalen Regionen. FOS-gebundene proximale Regionen besitzen eine Anreicherung des `CCAAT`-Box-Bindestellenmotivs.

Die Gruppe der FOS(only, AP-1−, proximal)-Fragmente aller drei Zelllinien ist durch das `CCAAT`-Motiv geprägt. Durch die Anwendung von MEME kann die dominante Rolle dieses Motivs in dieser Gruppe bestätigt werden (siehe Abbildung 4.27).

4.4.2 Wechselseitiger Ausschluss von NF-Y- und AP-1-Motiven in FOS gebundenen Regionen

Um ein robusteres Kriterium für die gemeinsame Auftrittswahrscheinlichkeit dieser beiden regulatorischen Elemente zu erhalten, wird die Bedeutung von potenziellen AP-1- und NF-Y-Bindungsstellen in den proximalen und distalen Regionen durch die Anwendung der MCC-definierten Schwellenwerte für die besten Matrizen der beiden Transkriptionsfaktoren aus TRANSFAC bewertet. Die Matrize V$JUNDM2_04 zeigt in der Analyse der AP-1-Vorhersagen das beste Leistungsverhalten in allen untersuchten ChIP-seq-Experimenten, während sich die Matrize V$NFY_01 bei der Untersuchung der NF-Y-

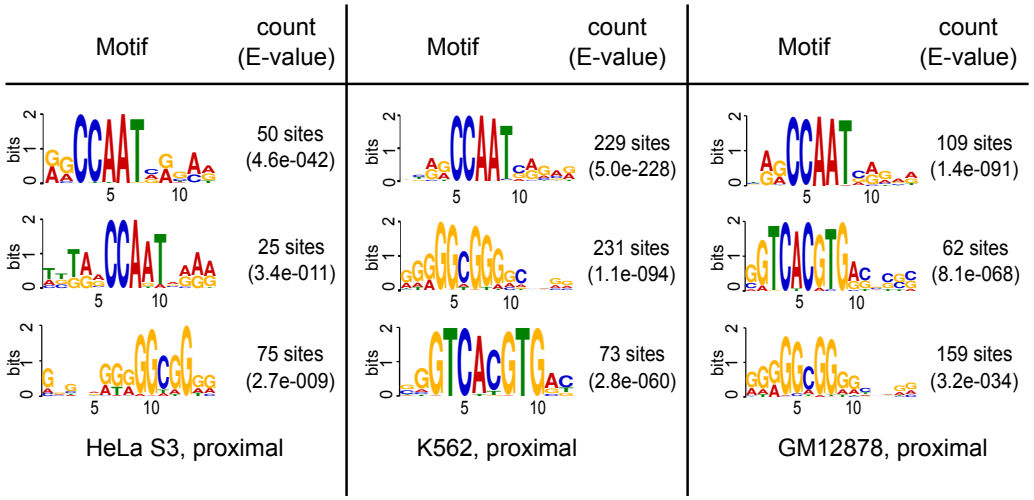


Abbildung 4.27: Anreicherung des CCAAT-Motivs in FOS(AP-1−)-ChIP-seq-Fragmenten. Die proximalen ChIP-seq-Regionen aus *HeLa S3*, *K562* und *GM12878*, welche keine hochaffine Bindestellenvorhersage des AP-1-Motivs aufweisen, zeigen eine Anreicherung des CCAAT-Motivs.

Bindestellensituationen über alle drei Zelllinien als die Beste aller NF-Y-definierten PWMs aus TRANSFAC erweist.

Wie in der Tabelle 4.13 zu sehen ist, besitzen die meisten FOS-gebundenen Regionen in den *HUVEC*-Zellen ein AP-1-Motiv und zeigen gleichzeitig nur wenige NF-Y-Bindungsmotive. Für die Zelllinie *GM12878* gilt das Gegenteil: Erstens existieren nur sehr wenige FOS-gebundene distale Bereiche (290 im Vergleich zu mehr als 20.000 in *HUVEC*), in denen praktisch keine AP-1-Motive zu finden sind. Zweitens zeigen auch die proximalen Regionen dieser Zelllinie keine hochbewerteten AP-1 Motive. Die Situation der FOS-gebundenen ChIP-seq-Regionen, welche gleichzeitig auch durch NFYB gebunden sind, ist für die drei Zelllinien sehr unterschiedlich (siehe auch Tabelle 4.10). Während die proximalen Regionen (Promotoren) eine größere Anzahl an Überlappungen aufweisen

Zelllinie	Typ	AP-1	NF-Y	Odds	OR (p-Wert)	Beide	AP-1/ NF-Y	Sum
HUVEC	p	500	98	5,1	0,034	64	569	1231
	d	14270	95	150,2	(7,8e-86)	281	5535	20181
Hela S3	p	86	451	0.191	0,002	20	100	657
	d	3661	38	88,4	(0)	60	696	4455
K562	p	79	1392	0.057	0,021	32	409	1912
	d	1536	574	2.676	(0)	91	448	2649
GM12878	p	4	989	0.004	0.029	27	265	1285
	d	18	130	0.138	(1,9e-13)	15	127	290

Tabelle 4.13: Häufigkeiten von AP-1- bzw. NF-Y-Motiven in den Zelllinien *HUVEC*, *HeLa S3*,*K562*,*GM12878*. Hochbewertete potentielle Bindestellen für die Transkriptionsfaktoren AP-1 und NF-Y schließen sich gegenseitig aus.

(*HeLa S3*: 22 %, *K562*: 56 % und *GM12878*: 36 %), ist die Situation für die distalen Bereiche (Enhancer) geringer (*HeLa S3*: 5 %, *K562*: 15 % und *GM12878*: 5 %). Die Analyse der hochaffinen Bindestellen zeigt einen noch deutlicheren Trend: Je höher der Anteil definierter AP-1-Bindestellenvorhersagen ist, desto niedriger ist der Anteil hochbewerteter NF-Y-Bindestellen und umgekehrt. Abbildung 4.29 zeigt den linearen und antikorrelierten Zusammenhang dieser beiden Motivsituationen eindeutig ($R = -0.957$, p-Wert: $3,317e-281$). Dieser deutliche Wert zeigt, dass das gemeinsame Auftreten hochaffiner Bindestellemotive für den Transkriptionsfaktor AP-1 und NF-Y nahezu ausgeschlossen scheint. Abbildung 4.3 zeigt die Bedeutung der hochaffinen Bindestellenvorhersagen für die Transkriptionsfaktoren NF-Y (blau) und AP-1 (rot) im gesamten Datensatz aller proximalen FOS-präzipitierten und NFYB-überlappenden ChIP-seq-Fragmente aus *HeLa S3*, *K562*, und *GM12878*. Im Vergleich zu den nicht-überlappenden DHS-Fragmenten (Kontrollgruppe) aus allen drei Zellen zeigt sich die deutliche Anreicherung hochbewerteter NF-Y Bindestellenmotive in dieser Gruppe (Abbildung 4.3, (a)). Die hochaffinen AP-1-Bindestellenmotive spielen in diesem Datensatz keine Rolle. Der Austausch des experimentellen Hintergrunds durch NF-YB(only)-Fragmente bestätigt dieses Bild noch einmal (Abbildung 4.29, (b)). Hochaffine NF-Y-Motive zeigen sich nun sowohl im Vorder-, als auch im Hintergrund. Die Abbildung verdeutlicht dieses Verhalten: die dargestellte AUROC-Kurve zeigt für hohe Bindestellenbewertungen (Punkt 0,0) keine Unterschiede zwischen dem Vorder- bzw. Hintergrund (Kontrolle). Die AP-1-Vorhersagen spielen in diesem Datensatz abermals keine Rolle (rote Kurve).

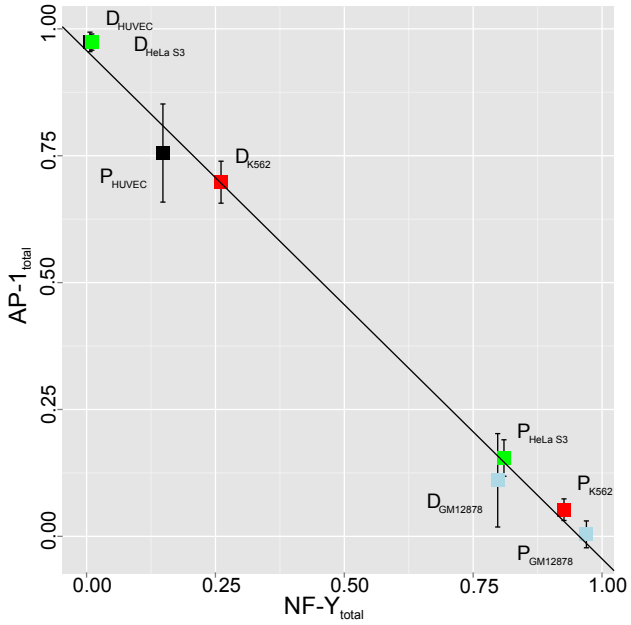


Abbildung 4.28: Antikorrelation des Auftretens von hochaffinen AP-1- und NF-Y-Bindungsmotiven. Die Abbildung zeigt die distalen (D) und proximalen (P) prozentualen Anteile der genomischen Intervalle, welche ein definiertes AP-1- bzw. ein definiertes NF-Y-Motiv für die Zelllinien *HeLa S3* (schwarz), *K562* (rot) und *GM12878* (blau) enthalten. Zu beachten ist, dass die Punkte D_{HUVEC} und $P_{HeLa\ S3}$ fast deckungsgleich sind (siehe oben links).

4.4.3 NF-Y-definierte Bindestellen-Konfiguration in FOS-gebundenen Regionen

Durch die MEME-Analysen der FOS-gebundenen proximalen Fragmente aller drei Zelllinien deutet sich an, dass zwei unterscheidbare Varianten von `CCAAT`-definierten Sequenzmotiven in den proximalen Regionen aller drei Zelllinien existieren (siehe auch Abbildung 4.27). In den nachfolgenden Experimenten wird untersucht, ob diese `CCAAT`-Boxen verschiedene Teilmengen dieser proximalen Gruppen prägen. Um die statistische Basis für die weiteren Experimente zu verbessern, werden die Sequenz-Datensätze für die drei Zelllinien *HeLa S3*, *K562* und *GM12878* zusammengeführt. Als Ergebnis stehen somit zwei eindeutige Datensätze zur Verfügung, welche die eindeutigen proximalen bzw. die

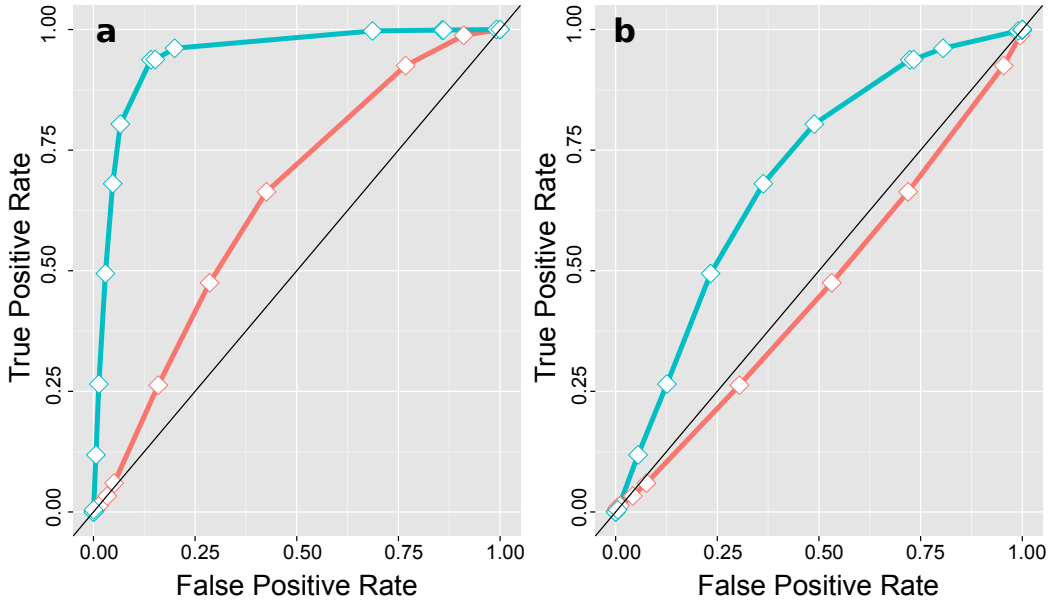


Abbildung 4.29: Abhängigkeit der gemeinsamen FOS-gebundenen proximalen Fragmente aus den Zelllinien *HeLa S3*, *K562* und *GM12878*. Gezeigt werden die AUROC Kurven der TRANSFAC Matrize V$NFY_01 (blau) und der Matrize V$JUNDM2_04 (rot). Die erste AUROC Kurve (a) zeigt die Qualität dieser Matrizen bezogen auf den allgemeinen DHS-definierten Hintergrund aller drei Zelllinien. Die AUROC Kurve (b) zeigt die Qualtität dieser Matrizen bezogen auf die NFYB(only)-definierten genomischen Intervalle aller drei Zelllinien.

distalen Situationen für die Transkriptionsfaktoren FOS und NFYB beschreiben. Abbildung 4.30 zeigt die Situation für beide Gruppen (links proximal, rechts distal). Die jeweiligen Mengen werden mit FOS(only), NFYB(only) bzw. Inter(NFYB,FOS) bezeichnet. Die FOS(only)-Gruppe besteht aus ChIP-seq-Intervallen, die mit FOS interagieren und sich nicht mit einer NFYB-gebundenen Region der drei untersuchten Zelllinien überschneiden. Die NFYB(only)-Gruppe ist analog nur für NFYB-ChIP-seq-Regionen definiert. Die Inter(NFYB,FOS)-Kategorie enthält eindeutige genomische Regionen aus den drei verwendeten Zelllinien, welche mit den beiden untersuchten TFs überlappen. Die Venn-Diagramme der Abbildung 4.30 (jeweils links dargestellt) zeigen die Größenordnungen für diese drei Sequenzsätze, getrennt nach proximalen (Abbildung 4.30, links) und dis-

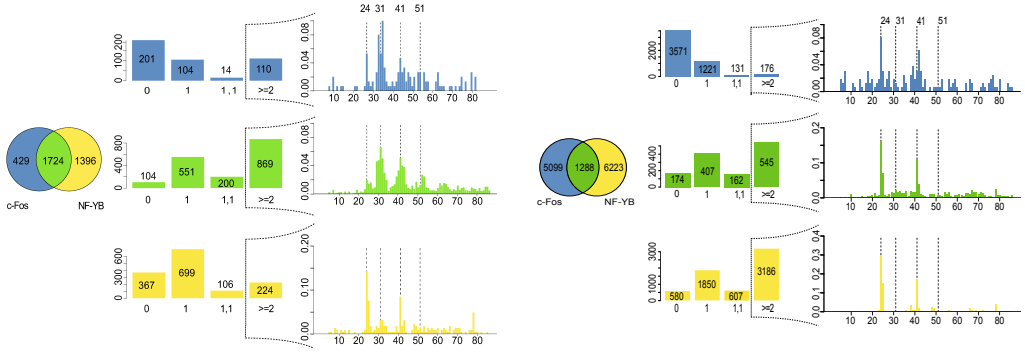


Abbildung 4.30: Verteilung der CCAAT Boxen in FOS-gebundenen Regionen. Die Darstellung auf der linken Seite zeigt die proximalen Regionen als Zusammenfassung der drei Zelllinien *HeLa S3*, *K562* und *GM12878*. Auf der rechten Seite sind die distalen Regionen dieser drei Zelllinien gezeigt.

talen (Abbildung 4.30, rechts) Bereichen. Die zentralen Histogramme zeigen an, wie viele Intervalle keine CCAAT-Box enthalten, die als regulärer Ausdruck (CCAAT bzw. ATTGG) identifizierbar ist (Spalte 0), eine einzelne CCAAT-Box (oder reverse Komplement) besitzen (Spalte 1), oder zwei CCAAT-Boxen in gleichgesinnter (CCAAT:CCAAT,ATTGG:ATTGG) oder in nicht-gleichgesinnter Ausrichtung (CCAAT:ATTGG,ATTGG:CCAAT) enthalten. Die zweite Situation wird im jeweiligen Histogramm als (1,1) bezeichnet, während die erste Situation mit (>=2) markiert ist. Die beobachteten Häufigkeiten dieser vier Gruppen werden in den entsprechenden Histogrammen gezeigt und wiederum für die proximalen und distalen Sequenzmengen unterschieden.

Die Ergebnisse zeigen, dass das Auftreten des CCAAT-Motivs eine prägende Eigenschaft der Inter(NFYB,FOS)- und der NFYB(only)-Gruppe ist. Nur sechs bzw. vierzehn Prozent dieser beiden Gruppen besitzen kein CCAAT-Motiv. Lediglich die FOS(only)-Gruppe enthält mehrheitlich kein CCAAT-Motiv. Eine Besonderheit zeigt sich in den proximalen und distalen Inter(NFYB,FOS)-Gruppen: Die direkten Wiederholungen des CCAAT-Motivs (Kategorie: ≥ 2) ist die häufigste Kategorie in den beiden Schnittmengen. Diese wurde für alle drei Sequenzmengen weiter betrachtet. In der Untersuchung wurde für diese Kategorie in allen proximalen und distalen Datensätzen die Abstandsverteilung der gleichgesinnten CCAAT-Boxen bestimmt. Dabei wurde jeweils der Abstand zwischen der ersten

Position des Motivs zur Anfangsposition der nächsten gleichgesinnten CCAAT-Box berechnet. Die in Abbildung 4.30 gezeigten Verteilungen zeigen die Abstandssituation der drei verschiedenen Fragmentsituationen. Auffällig ist, dass in allen drei Gruppen (FOS(only), Inter(NFYB,FOS) und NFYB(only)) eine Anreicherung bestimmter Abstände der CCAAT-Tandems zu beobachten ist. Als deutliches Merkmal der FOS-gebundenen proximalen Regionen ist zu beobachten, dass unabhängig davon, ob sie in der Gruppe FOS(only) oder Inter(NFYB, FOS) vorkommen, ein Maximum (Peak) um den Distanzwert von 31 bp zu beobachtbar ist. Begleitet wird dieses Maximum durch eine Reihe von abnehmenden Maxima, welche periodisch mit einer Amplitude von zehn bis elf bp auftreten. Zwei Spitzen, die einen Abstand von 24 und 41 markieren, werden in der Gruppe der distalen Bereiche für die NFYB(only)-Gruppe gefunden. Dieses charakteristische Profil findet sich auch in den distalen Inter(NFYB, FOS)-Bereichen. Bei näherer Betrachtung stellte sich heraus, dass sie einer bestimmten repetitiven Elementklasse, den sogenannten *Long Terminal Repeat* (LTR), insbesondere der LTR12-Familie, entsprechen (Haubrock, Hartmann et al. 2016). Ihr Auftreten in den FOS-gebundenen proximalen Regionen kann von diesen LTR-Sequenzen, die bekanntermaßen mit dem Transkriptionsfaktor NF-Y interagieren, als *Kontamination* dieser Promotorklasse angesehen werden (Fleming et al. 2013). Aus diesen Beobachtungen kann gefolgert werden, dass FOS-gebundene Promotoren, die kein AP-1-Bindungsmotiv zeigen, aber ein direktes CCAAT-Box-Tandem enthalten, ein neuartiges regulatorisches Modul beschreiben. Dieses wird direkt durch den Transkriptionsfaktor NF-Y gebunden.

4.4.4 Das AP-1/NF-Y-Enhancer-Promotor-Modell

Die beiden Wissenschaftler Mercer und Mattick diskutieren in ihrer Veröffentlichung aus dem Jahre 2013 bereits die Möglichkeit, dass Enhancer-gebundene Proteine durch das ChIP-seq-Verfahren auch zu Promotoren vernetzt werden können (Tim R Mercer und Mattick 2013). Diese Überlegung motiviert folgende in diesem Abschnitt untersuchte Fragestellung: Können FOS-gebundene und durch das AP-1-Bindungsmotiv charakterisierte Enhancer mit FOS- und NFYB-gebundenen Promotoren interagieren? Die Vermutung besteht, dass eine Teilmenge dieser Promotoren, welche ein direktes CCAAT-Tandem-Sequenzmotiv enthalten, gezielt durch diese Enhancer gebunden werden. Um diese Hypo-

these zu untersuchen, werden die Abstandsbeziehung zwischen den distalen Fragmenten (potentielle Enhancer) und der Gruppe der FOS- und NFYB-interagierenden proximalen genomischen Fragmente analysiert. Dabei wird, unabhängig von der Positionierung des Enhancers (stromaufwärts oder stromabwärts), die kürzeste Distanz zwischen einer Enhancerregion und einer Promotorregion des gleichen Chromosoms analysiert. Für den nächsten stromaufwärts liegenden Enhancer wird dazu der Abstand zwischen dem Ende der Enhancerregion und dem Startpunkt der proximalen Region berechnet. Für den nächsten stromabwärts liegenden Enhancer wird der Abstand zwischen dem Endpunkt des proximalen Fragments und dem Startpunkt der Enhancer-Region bestimmt. Die kürzeste Distanz dieser beiden Werte wird als Abstandsmaß für die weiteren Untersuchungen verwendet. Als potenzielle Promotoren stehen 1724 Fragmente zur Verfügung (siehe Venn-Diagramm Abbildung 4.30, linke Hälfte). Die Gruppe der potenziellen Enhancer wird durch zwei Datensätze gebildet: Der erste Datensatz besteht aus 3571 Fragmenten und wird durch die distale FOS(only,CCAAT-)-Gruppe gebildet (siehe Abbildung 4.30, rechts oben (blau)). Als Kontrollgruppe dienen 3186 distale NFYB(only,CCAAT$\geq$2)-Enhancer-Fragmente (siehe Abbildung 4.30, rechts unten (gelb)). Auf Basis dieser beiden Enhancergruppen wurde nun 1000 Mal eine Teilmenge von 1500 zufälligen Enhancern gezogen. Für jede Promotorregion kann nun für die beiden zufälligen Enhancermengen die kürzeste Distanz des Promotors zum nächstliegenden potenziellen Enhancer bestimmt werden. Die sich ergebende Distanzverteilung ist in Abbildung 4.31 dargestellt. Es zeigt sich, dass die FOS-gebundenen und das AP-1-Bindungsmotiv enthaltenden potentiellen Enhancer mit deutlich näher an FOS-gebundenen und CCAAT-Box-Tandems-definierten Promotoren liegen, als das für die Kontrollgruppe der distalen NF-YB(only, CCAAT$\geq$2)-Gruppe der Fall ist. Die Unterschiedlichkeit beider Verteilungen wird durch den signifikanten P-Wert durch Anwendung des Mann-Whitney-Tests unterstützt (P-Wert: 3.317e-281). Die beiden Transkriptionsfaktoren FOS und NFYB interagieren mit dem Koaktivatorprotein p300 (Faniello et al. 1999; W.-M. Wang et al. 2011). Das p300-Protein (UniProt: Q09472) ist eine sogenannte Histonacetyltransferase, welche mit einer Reihe verschiedener TF interagieren kann und durch Veränderung des Chromatins einen allgemeinen positiven Einfluss auf die Transkription ausübt (Delvecchio et al. 2013; Ogryzko et al. 1996; Tropberger et al. 2013). NFYB bzw. NF-Y als Komplex und auch FOS interagieren mit p300, indem sie entweder mit dem C-terminalen Ende oder dem zentralen Bereich von p300 interagieren (Faniello et al. 1999; Preston et al. 2000). Da eine bemerkenswerte Anzahl an FOS-präzipitierten Promotoren auch mit NF-Y interagiert und davon eine definierte Teilmenge in einer besonde-

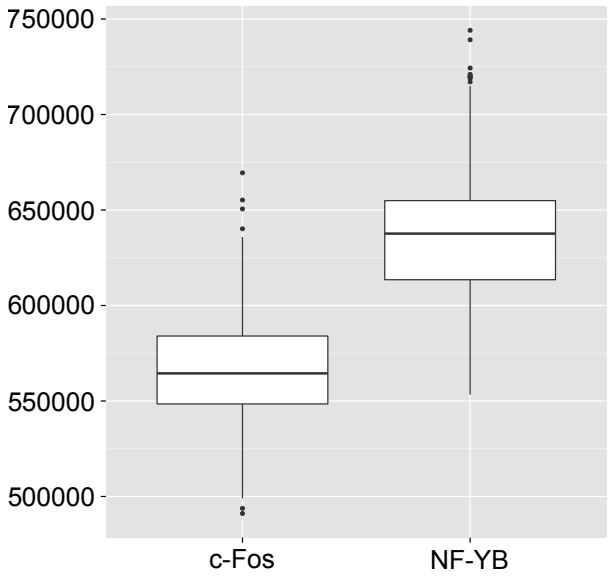


Abbildung 4.31: Distanzverteilung von ausschließlich c-Fos- bzw. NF-YB-interagierenden distalen Bereichen zu c-Fos- und NF-YB-regulierten proximalen Fragmenten.

ren CCAAT-Tandem-Konfiguration vorliegt, welche möglicherweise durch AP-1-definierte Enhancer zusätzlich reguliert wird, liegt die Vermutung nahe, dass der Koaktivator p300 an der Ausprägung dieses regulatorischen Moduls aktiv beteiligt ist. Um die Bedeutung dieser unterschiedlichen Einflüsse bewerten zu können, werden die einzelnen Datensätze auf eine besondere biologische Funktion getestet. Klassischerweise werden diese Untersuchungen durch die Anwendung der Gene Ontology (Ashburner et al. 2000) in Kombination mit einem statistischen Testverfahren durchgeführt. Die GREAT-Plattform kann direkt eine GO-basierte Anreicherungsanalyse auf Basis von ChIP-seq-definierten Regionen verwenden (McLean et al. 2010). Als erstes wird die gesamte Inter(NFYB,FOS)-Überlappungsgruppe getestet. Sie enthält insgesamt 1724 unterschiedliche genomische Regionen. Die zweite Gruppe ist eine Teilmenge dieser Gruppe. Die Gruppe wird als Inter(CCAAT2+) bezeichnet, da alle Fragmente dieser Gruppe mindestens ein direktes CCAAT-Repeat enthalten. Sie beschreibt insgesamt 869 genomische Fragmente. Die dritte und kleinste Gruppe enthält 470 genomische Regionen. Sie wird aus Fragmenten der Inter(CCAAT2+)-Gruppe gefiltert, die zusätzlich noch mit p300-interagierenden Regionen aus den drei Zelllinien *Hela S3*, *K562* und *GM12878* überlappen. Die Inter(NFYB,FOS)-Fragmentmenge ist die einzige der untersuchten drei proximalen Gruppen, für die eine signifikante Anreicherung

GO Kategorie (Anzahl Gene)	**ChIP-seq Datensatz** (Anzahl Gene)		
	Inter	**Inter**(CCAAT+)	(**CCAAT+p300**)
Nucleosome assembly (144)	8.3661e-23 (66)	1.0935e-30 (59)	9.6057e-32 (48)
Protein-DNA complex assembly (166)	2.0001e-22 (71)	6.4928e-30 (62)	4.5491e-31 (50)
Chromatin assembly (156)	3.1488e-21 (67)	6.9986e-30 (60)	2.2627e-30 (48)
Nucleosome organization (167)	4.3924e-20 (68)	3.3856e-27 (59)	5.8612e-29 (48)
DNA conformation change (232)	1.4360e-18 (80)	1.2123e-27 (70)	7.3449e-29 (55)
Protein-DNA complex subunit organization (189)	4.6583e- (73)	9.6572e-27 (62)	1.5335e-28 (50)
DNA packing (193)	6.6237e-17 (39)	3.2159e-26 (62)	3.4329e-28 (50)
Chromatin assembly or disassembly (175)	7.2675e-19 (68)	6.0315e-27 (60)	3.6575e-28 (48)
Response to endoplasmic reticulum stress (127)	3.4353e-07 (39)	7.1109e-05 (24)	0.0007 (16)
Positive regulation of nuclease activity (67)	1.3394e-07 (27)	0.0010 (15)	0.0302 (9)
Regulation of nuclease activity (73)	2.4459e-07 (28)	0.0025 (15)	0.0497 (9)

Tabelle 4.14: Auflistung der signifikanten GO-definierten *Biological Process*-bezogenen Prozesse in FOS/NFYB-kolokalisierten regulatorischen ChIP-seq-Fragmente aus den Zelllinien *HeLa S3*, *K562* und *GM12878*.

verschiedener biologischer Prozesse von in Gene Ontology (GO)-annotierten Funktionskategorien gefunden werden kann. Die FOS(only)-Gruppe zeigt keine signifikanten GO-Kategorien nach der P-Wert-Korrektur, während die Menge der NFYB(only)-Fragmente, bedingt durch ihre Größe, nur sehr allgemeine und unspezifische GO-Kategorien aufweist. Tabelle 4.14 zeigt die Ergebnisse der Anreicherungsanalyse der Inter(NFYB,FOS)-Situation. Es werden die besten elf biologischen Prozesse aufgelistet, welche sich unter der strengsten Filterung noch als signifikante GO-Kategorien finden lassen. Neben der Kolokalisation von FOS und NFYB bedingt diese Filterung das Vorhandensein eines gleichgesinnten CCAAT-Tandems und die Überlappung mit dem Koaktivator p300. Die beste GO-Kategorie für diese Fragment-Menge lautet: *Nukleosom Assemblierung* (*engl. nucleosome assembly*, GO:0006334). Die Gene/Proteine dieses biologischen Prozesses sind an der Aggregation, Anordnung und Bindung des Nukleosoms an der DNA beteiligt. Die Inter(NFYB,Fos, CCAAT2+)-Gruppe zeigt eine vergleichbare statistische Signifikanz der

dargestellten GO-Kategorien. Auch die größte Gruppe (Inter(NFYB,FOS)) verhält sich dazu vergleichbar. Die explizite Anreicherung der Chromatin- und Nukleosom-bezogenen Prozesse in den proximalen Regionen der Inter(NFYB,FOS)-Gruppe und das deutliche Auftreten von direkten CCAAT-Tandem-Repeats in dieser Gruppe, welche gleichzeitig mit dem p300 Koaktivatorprotein interagieren, lässt vermuten, dass die Promotoren dieser biologischen Prozesse durch dieses neuartige regulatorische Modul geprägt sind. Durch dieses CCAAT-Dimer könnte ein regulatorisches Interface in den Promotoren definiert werden, welches durch Vermittlung von p300 gezielt mit FOS als Komponente des AP-1-Transkriptionsfaktors durch hochaffine AP-1-Transkriptionsfaktorbindestellen in den Enhancern interagiert.

5 Diskussion

In dieser Arbeit ist die prägende Eigenschaft einer TFBS als definitorischer Kern (Seed) und deren interagierende TFs als Masterregulatoren eines regulatorischen Moduls in vier verschiedenen Projekten untersucht worden. Das implementierte Verfahren zur Analyse und Bewertung regulatorischer Einzelnukleotidvariationen (SNPs) konnte für zwei krankheitsassoziierte SNPs erfolgreich angewendet werden. Genetische Veränderungen in regulatorischen Regionen können Krankheitsrisiken in vielerlei Hinsicht verstärken. Durch strukturelle Veränderungen des genetischen Materials, wie z.B. Deletionen oder Translokationen von chromosomalen Bereichen, können regulatorische Elemente von ihren eigentlichen Zielgenen getrennt oder neue Zielgene geschaffen werden (Kleinjan und van Heyningen 1998). So wurde z.B. von Erikson et al. (1983) gezeigt, dass die Translokation des MYC-Gens von Chromosom 8 in den Genbereich der Immunglobuline auf Chromosom 14 zu einer außergewöhnlichen Expression (ektopische Genexpression) von MYC führt und so die Ausbildung des sogenannten Burkitt-Lymphoms verursacht wird. Nicht nur die Veränderung des regulatorischen Kontexts kann nachteilige Effekte haben, auch die Verdoppelung einer regulatorischen Region kann eine starke Veränderung bewirken: So führt z.B. eine Verdoppelung des 600 KB regulatorischen Bereiches stromaufwärts des menschlichen SOX9-Gens dazu, dass sich bei Frauen männliche anatomische Merkmale ausprägen (Cox et al. 2011). In den beiden genannten Beispielen sind jeweils die Transkriptionsfaktoren MYC und SOX9 Ursache der pathologischen Veränderung.

Auch für Einzelnukleotidvariationen sind Beispiele in der Literatur beschrieben, welche einen krankheitsassoziierten Einfluss dieser SNPs in den regulatorischen Regionen zeigen (Maurano et al. 2012b; Roadmap Epigenomics Consortium et al. 2015). In einer Studie von Musunuru et al. (2010) konnte gezeigt werden, dass die Sequenzvariation rs1274037 mit der Schaffung einer Transkriptionsfaktorbindestelle für den Transkripti-

onsfaktor CEBP einhergeht. Die Studie zeigt, dass durch diese neue Bindestelle die Genaktivität des SORT1-Gens in der Leber stark hochreguliert wird. Das Protein diese Gens ist Mitglied eines Stoffwechselweges, welcher den Cholerstinspiegel im Blutplasma reguliert und erklärt damit das krankheitsassoziierte Risiko dieses SNPs (Myokardinfarkt). Der Polymorphismus ist im 3'-UTR-Bereich eines anderen Gens zu finden (CELSR2). Das 3'-Ende des SORT1-Gens liegt mehr als 46 kb stromabwärts dieses SNP.

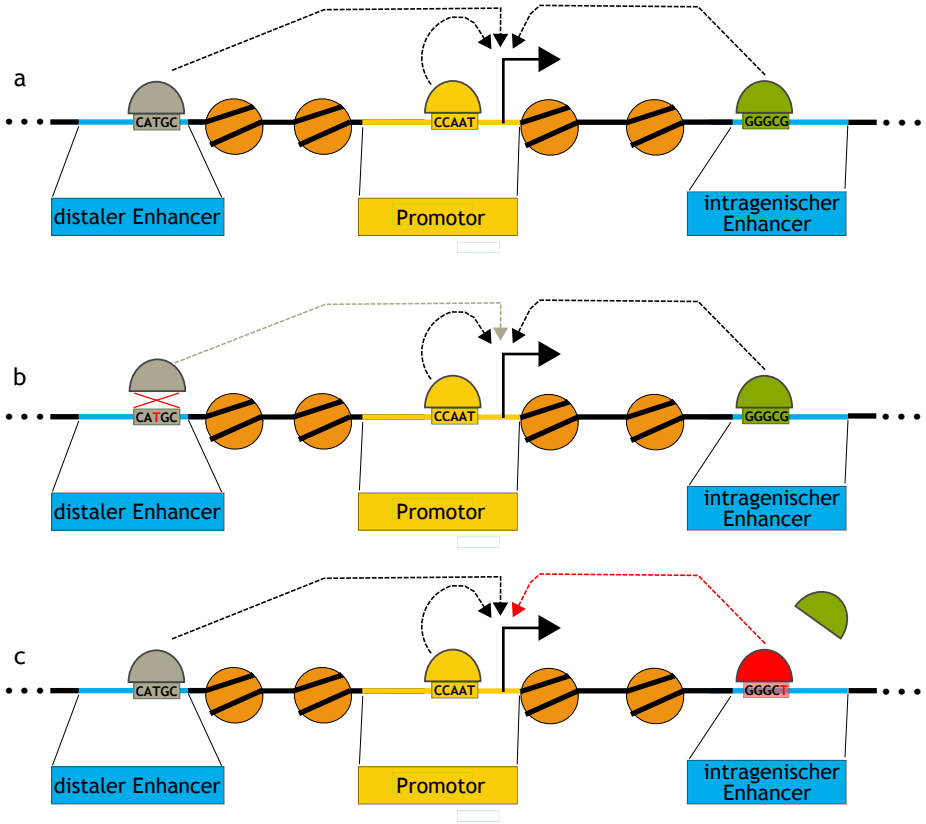


Abbildung 5.1: Die unterschiedlichen regulatorischen Einflüsse auf eine Transkriptionseinheit. Die beispielhafte Darstellung (a) zeigt die proximalen (promotornahen) und distalen (Enhancer-definierten) regulatorischen Einflüsse, welche die verschiedenen Transkriptionsfaktoren durch ihre sequenzdefinierten Bindestellen auf eine Transkriptionseinheit ausüben können. Eine Einzelnukleotidvariation kann dazu führen, dass ein Transkriptionsfaktor nicht mehr an seine entsprechende Bindestelle binden kann (b). Eine alternative Möglichkeit ist im Abbildungsteil (c) dargestellt: Durch die Variation einer Bindestelle kann eine neue Bindestelle für einen anderen Transkriptionsfaktor geschaffen werden. Dieser verdrängt den bisherigen Transkriptionsfaktor und kann so einen völlig neuen Einfluss auf die Transkriptionseinheit ausüben.

Die Bedeutung dieses SNPs für die Transkription des SORT1-Gens lässt auf eine

Enhancer-definierte Funktion dieses 3'-UTR-Genbereichs schließen. Die Erzeugung einer TFBS in einer Enhancerregion durch eine Sequenzvariation muss nicht zwangsläufig nur mit einer Erhöhung der Transkriptionsstärke der regulierten Transkriptionseinheit verbunden sein. Das Brustkrebs-assoziierte SNP rs4784227 (C/T) zeigt zum Beispiel, dass die seltener vorkommende T-Variante mit der Erzeugung einer TFBS für den Transkriptionsfaktor FOXA1 verbunden ist (Cowper-Sal·lari et al. 2012). Dieser TF sorgt in Kombination mit einem anderen TF dafür, dass das Zielgen TOX3 bis zu fünffach niedriger exprimiert wird, als in der C-Variante. Alle aufgeführten Beispiele zeigen, wie bedeutend eine Einzelnukleotidvariation im Kontext der transkriptionellen Genregulation sein kann.
Abbildung 5.1 (a) zeigt die unterschiedlichen proximalen (Promotor-definierten) bzw. distalen (Enhancer-definierten) regulatorischen Einflüsse, welche für eine Transkriptionseinheit existieren können. Eine Variation in den beteiligten TFBSs kann eine Veränderung der Transkriptionsaktivität bewirken, indem z.B. der bisherige TF nicht mehr an seine Bindestelle binden kann (Abbildung 5.1, b). Da eine Transkriptionseinheit in der Regel immer durch eine Vielzahl von proximalen und distalen TFs reguliert wird, kann der Ausfall einer einzelnen Regulationseinheit möglicherweise kompensiert werden. Eine größere Bedeutung können allerdings die Einzelnukleotidvariationen ausüben, welche eine neue TFBS erzeugen oder eine existierende Bindestelle so verändern, dass ein neuer TF gebunden werden kann (Abbildung 5.1, c). Auf diese Art und Weise kann eine Transkriptionseinheit auf neue regulatorische Einflüsse reagieren und so eine veränderte Transkription des Gens in verschiedenen Zellen ermöglichen oder verhindern. Analog zum Beispiel der chromosomalen Umlagerung des MYC-Gens bedeutet dies, dass in so einer Situation die Regulation der Transkription des Gens neuen Einflüssen unterliegen und so ein möglicher pathogener Einfluss ausgeübt werden kann. Die in dieser Arbeit analysierten Beispiele zeigen jedenfalls, dass die pathogenen nicht-kodierenden SNPs mit hochaffinen neu geschaffenen Transkriptionsfaktorbindestellen assoziiert werden können. Die Lokalisation der beiden Sequenzvariationen ist jeweils in nicht-kodierenden Regionen (Intron) einer Transkriptionseinheit zu finden.
Die beiden in dieser Arbeit untersuchten SNPs sind jeweils in einer Intronregionen der untersuchten Gene gefunden worden. Es gibt für beide Situationen Anhaltspunkte, dass die pathogenen SNPs auf die Transkriptionsaktivität der eigenen Transkriptionseinheit Einfluss besitzen, also eine Intron-definierte Enhancerregion prägen könnten. Für den Transkriptionsfaktor SP1, dessen Bindestelle mit der krankheitsassoziierten Einzelnukleotidvariation rs11644322 des WWOX-Gens als prägende neue Bindestelle vorhergesagt wurde,

ist in nachfolgenden Untersuchungen dieser Einfluss experimentell bestätigt worden. In einer Untersuchung von Deshane et al. (2010) ist für das menschliche Hämoxygenase-Gen (engl. Heme oxygenase I, kurz: HMOX1) gezeigt worden, dass der Transkriptionsfaktor SP1 die Interaktion eines Intron-definierten und von diesem TF-gebundenen Enhancer mit dem stromaufwärts gelegenen Promotor unterstützt wird. Für die Sequenzvariation rs3857080, für die eine neue LHX4-Transkriptionsfaktorbindestelle vorhergesagt und die Bindung dieses Transkriptionsfaktors ebenfalls experimentell nachgewiesen wurde, existieren auch verschiedene veröffentlichte Belege für eine Enhancer-definierte Interaktion dieses TF (Fuxman Bass et al. 2015; Lee et al. 2004). Die Gemeinsamkeit der beiden analysierten SNP ist, dass durch die jeweils krankheitsassoziierte Einzelnukleotidvariation eine hochaffine neue TFBS in einem Enhancer geschaffen wird. Durch Rückkopplung mit dem eigentlichen Transkriptionsapparat (Stichwort Enhanceosom) kann der neue Transkriptionsfaktor seine pathogene Wirkung ausüben. Als Konsequenz dieser neuen regulatorischen Interaktion erscheint sowohl eine Verstärkung als auch eine Erniedrigung der Transkriptionsaktivität möglich. Ob allerdings gerade Intron-definierte Enhancer besonders geeignet sind, um regulatorische bedeutende Sequenzvariationen auszubilden, müssen weitere Untersuchungen zeigen.

Wie bereits bei der Analyse der Einzelnukleotidvariationen deutlich geworden ist, üben die hochbewerteten TFBS verschiedener TFs in proximalen und distalen regulatorischen Regionen einen besonderen Einfluss auf die Transkriptionsregulation aus. Die Kombination der prägenden TFBSs einer regulatorischen Region tritt nicht zufällig auf. Als Beispiel hierfür kann die Kombination der Transkriptionsfaktoren FOS als Mitglied des Heterodimers AP-1 und NFYB im Proteinkomplex (Heterotrimer) des Transkriptionsfaktors NF-Y herangezogen werden. In ChIP-seq-präzipitierten regulatorischen Regionen erscheinen die hochaffinen Bindestellen dieser beiden TFs innerhalb einer regulatorischen Region wechselseitig ausgeschlossen zu sein. Das belegt die Beobachtung, dass ähnlich regulierte Fragmente einer Gruppe von Promotoren (proximale Fragmente) durch `CCAAT`-Motive geprägt sein können, für die gleichzeitig das Fehlen hochbewerteter AP-1-Bindestellenmotive zu beobachten ist. Fleming et al. (2013) berichten in ihren Untersuchungen von einer gemeinsamen Lokalisierung von FOS und NFYB im Zusammenhang mit dem `CCAAT`-Motiv und einem Fehlen des AP-1-Motivs. In der Literatur konnte für das Gen GNRHR (*Gonadotropin-releasing hormon receptor*) gezeigt werden, dass AP-1 und NF-Y im Promotor dieses Gens durch eine definierte Protein-Protein-Interaktion zwischen JUN und

NFYA interagieren (Coss et al. 2004). Xie et al. (2013) zeigen in ihrer Veröffentlichung, dass FOS und NFYB in ChIP-seq-Experimenten kolokalisieren.

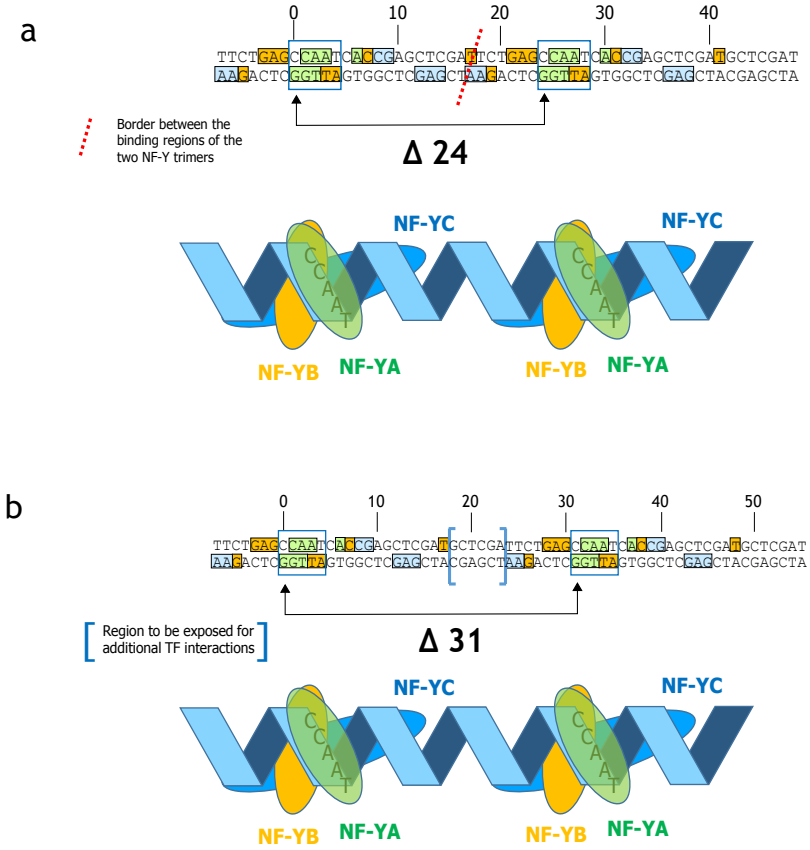


Abbildung 5.2: Abstände verschiedener CCAAT-Box-Tandems.

All diese Beispiele deuten an, dass die grundsätzliche Kombinierbarkeit dieser beiden TFs möglich ist. Sie erklären aber nicht, warum manche FOS-präzipitierte Promotoren ausschließlich `CCAAT`-Motive (Konsensusmotiv einer NF-Y-Bindestelle) enthalten und gleichzeitig keine hochaffinen AP-1-definierten Transkriptionsfaktorbindestellen aufweisen. Eine Besonderheit zeichnet eine Teilmenge dieser Promotoren zusätzlich aus: Das `CCAAT`-Motiv tritt in Form einer direkten Wiederholung auf (`CCAAT-CCAAT`- oder `ATTGG-ATTGG`-Tandemwiederholungen). Gleichzeitig zeigt sich, dass diese `CCAAT`-Tandems in einem Abstand von zehn bis elf Nukleotiden auftreten. Der häufigste Abstand dieser Gruppe von Promotoren zeigt sich bei 31 bp. Bei Abständen von 41 und 51 Basenpaaren werden mit abnehmender Amplitude weitere Maxima beobachtet. Diese Situation deutet auf

ein strukturelles Merkmal der CCAAT-Box-definierten Promotorgruppe hin. Auf Grundlage von Strukturanalysen ist bekannt, dass NF-Y-gebundene DNA zwischen dem zweiten A und dem nachfolgenden T stark geknickt wird Nardini et al. (2013). Mit Hilfe der Veröffentlichung von Nardini et al. (2013) kann der Mindestabstand zweier gebundener NF-Y-Transkriptionsfaktoren abgeleitet werden. Abbildung 5.2 zeigt diese Situation schematisch (Haubrock, Hartmann et al. 2016). Es wird deutlich, dass ein Mindestabstand von 24 bp zwischen zwei DNA-gebundenen NF-Y-Proteinkomplexen existiert (siehe Abbildung 5.2, a). Die häufigste beobachtete Distanz von 31 bp zeigt einen nicht-gebundenen Zwischenraum von sechs Basenpaaren (siehe Abbildung 5.2, b).

Aus den strukturellen Eigenschaften dieser speziellen Promotorklasse und der Beobachtung, dass sich keine hochaffine AP-1-Bindestelle in diesen Promotoren zeigt, kann geschlossen werden, dass FOS bzw. AP-1 durch indirekte Protein-Protein-Interaktion mit den durch NF-Y-strukturgeprägten Promotoren interagiert. Diese Vermutung wird durch

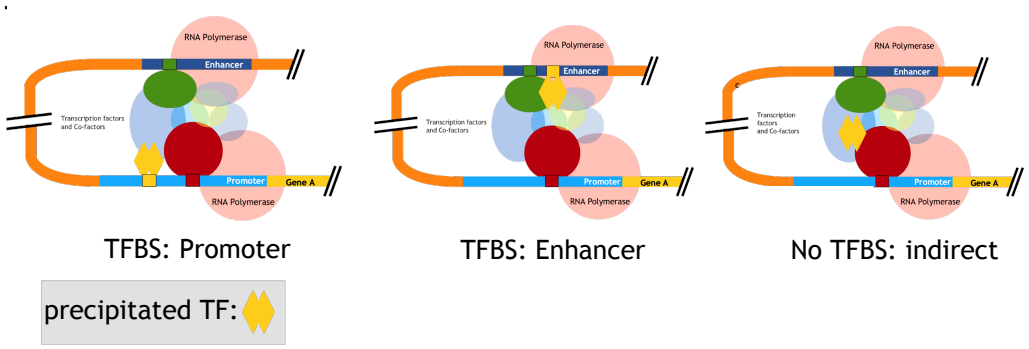


Abbildung 5.3: Mögliche ChIP-seq-Fragmente des Enhanceosoms.

zwei Beobachtungen unterstützt. Erstens zeigt die Teilmenge der CCAAT-Tandems enthaltenden Promotoren eine deutliche Anreicherung für eine im Gene Ontology-Projekt definierte Funktionskategorie (GO-Term: *Nucleosome assembly*). Zweitens zeigen die Promotoren dieser Klasse eine ausgeprägte signifikante Nähe zu AP-1-Motiv-enthaltenen und FOS/AP-1-regulierten Enhancern. Tim R Mercer und Mattick (2013) diskutierten bereits in ihrer Veröffentlichung über die Möglichkeit, dass die Fragmente eines ChIP-seq-Experiments auch Fragmente enthalten können, welche nicht nur durch eine direkte Interaktion des präzipitierten Transkriptionsfaktor mit der DNA geprägt wurden, sondern auch indirekt gebundene Fragmente, welche z.B. durch die Interaktion von funktionalen

Enhancer-Promotor-Paare entstanden sind. Abbildung 5.3 zeigt diese drei Möglichkeiten im Enhanceosom beispielhaft. Das experimentelle Protokoll eines ChIP-seq-Experiments kann durch eine mögliche Enhanceosomstruktur nicht zwischen einer direkten Interaktion eines Transkriptionsfaktors (sequenzspezifisch), oder einer indirekten Interaktion (sequenzunspezifisch) unterscheiden. Da eine TFBS sowohl im Promotor als auch im Enhancer lokalisiert sein kann, sind auch diese beiden Fälle bei der Interpretation von ChIP-seq-Experimenten zu beachten (siehe Abbildung 5.3).
Auf Basis dieser gesamten Beobachtungen kann ein hypothetisches Modell entwickelt werden, welches die Interaktion der direkten CCAAT-Dimer-geprägten und NF-Y-gebunden Promotoren mit den AP-1-gekennzeichneten Enhancern beschreibt. Abbildung 5.4 zeigt

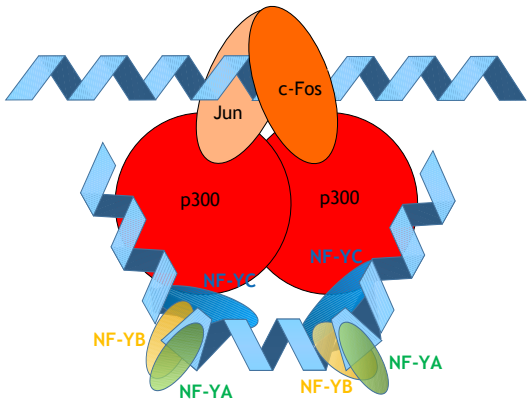


Abbildung 5.4: Hypothetisches Enhancer-Promotormodell der FOS/NFYB-Interaktion. Ausgehend von der Beobachtung, dass die zusätzliche Filterung der CCAAT-Box-Dimeren mit p300 ChIP-seq-Überlappungen eine stärkere Anreicherung eines GO-definierten biologischen Prozesses zeigt, fasst dieses Modell alle Ergebnisse dieser neuen NFYB-geprägten und FOS-gebundenen Promotorklasse zusammen. Unterstützt wird die Rolle des p300-Proteins in diesem Zusammenhang von einer aktuellen Veröffentlichung von Ortega et al. (2018).

dieses Modell. Die Rolle des Proteins p300 wird in diesem Modell berücksichtigt. Unterstützt wird die Rolle des p300-Proteins in diesem Zusammenhang von einer aktuellen

Veröffentlichung von Ortega et al. (2018). In dieser Untersuchung wird gezeigt, dass eine TF-Dimerisierung die Acetylierungseigenschaft des p300-Proteins aktiviert. In dieser Untersuchung wird außerdem auch die Bindung zweier p300-Proteine an dimerisierende TFs diskutiert. Da am Enhancer der AP-1-Komplex auf der Grundlage von FOS und vermutlich JUN gebildet wird (Heterodimer), könnte diese Eigenschaft auch für diese Gruppe der Enhancer gelten.

Alle bisher diskutierten Beispiele zeigen die herausragende Bedeutung einzelner TFBSs und den dort interagierenden TFs als Masterregulatoren. Durch definierte Proteininteraktionen zwischen den gebundenen Masterregulatoren und weiteren TFs und Kofaktoren wird ein regulatorisch aktives Modul vervollständigt. Die Erstellung des regulatorischen Transkriptionsnetzwerks (RTN) folgt diesem Gedanken. Das RTN wurde in dieser Arbeit durch die Anwendung bekannter Bindestellenbeschreibungen aus der TRANSFAC-Datenbank, den sogenannten PWMs, auf Grundlage der potenziellen Promotorregionen des Menschen erzeugt. Die einzelnen vorhergesagten TFBSs wurden zusätzlich durch die Anwendung des *phylogenetischen footprinting* gefiltert. Mit Hilfe von vorberechneten Sequenzalignments können auf diese Weise konservierte Seed-Bindestellen berechnet werden. Durch die Anwendung des menschlichen Transkriptionsnetzwerkes in den verschiedenen Projekten zeigt sich die Bedeutung dieser Kernbindestellen und der dort interagierenden TFs als definitorische Basis des Enhanceosoms. Verschiedene zelltypspezifische im ENCODE-Projekt erstellte ChIP-seq-Experimente deuten an, dass die Seeds sehr spezifisch verwendet werden: So zeigt z.B. der Vergleich zweier ChIP-seq-Experimente, welche für den GATA3-Transkriptionsfaktor durchgeführt worden sind, dass nur 9 % (10 von 114) dieser GATA3-definitorischen TFBS-Vorhersagen zwischen zwei untersuchten Zelllinien überlappen. Diese Zahlenwerte zeigen sich auch für andere Beispiele in vergleichbarer Art und Weise.
Die Rekonstruktion gewebespezifischer Netzwerke lässt vermuten, dass die regulatorischen Transkriptionsnetzwerke der acht untersuchten Gewebe trotz unterschiedlicher Netzwerkgrößen im Vergleich der topologischen Netzwerkeigenschaften einem gemeinsamen Konstruktionsprinzip folgen. Damit weisen sie Eigenschaften auf, die so bereits vielfach in der Literatur diskutiert worden sind (Lima-Mendez und van Helden 2009). Die Erweiterung der regulatorischen Netzwerke durch die Anwendung der in der TFClass-Datenbank gespeicherten DBD-Sequenzähnlichkeiten weist auf eine Vergleichbarkeit der zur Verfügung stehenden TFs zwischen den verschiedenen Geweben hin. Diese Beobachtung macht

deutlich, dass die Verwendung der TFs sehr generisch erfolgt. Auch diese Beobachtung wird durch verschiedene veröffentlichte Arbeiten belegt (Jolma, Yan et al. 2013; Jolma, Yin et al. 2015; Lambert et al. 2018). Dies scheint auf den ersten Blick der Beobachtung der zelltypspezifischen Verwendung prägender TFBSs in z.B. ChIP-seq-Experimenten zu widersprechen. Allerdings bedeutet die geringe Überlappung von ChIP-seq-Regionen, dass auf Basis der größtenteils gleichen Menge von exprimierten TFs unterschiedliche regulatorische Regionen spezifisch gebunden werden. Aber wie wird diese Spezifität erreicht? Die Verpackung der DNA spielt hier eine besondere Rolle. Die im Zellkern vorliegende DNA ist an sogenannte Nukleosome gebunden. Dieser Proteinkomplex, bestehend aus den Histonproteinen H2A, H2B, H3 und H4, komprimiert die DNA. Die Positionierung der Nukleosome spielt bei der Genregulation eine wichtige Rolle. In den aktiven Promotoren sind z.B. weniger Nukleosome zu finden als beispielsweise in nicht-aktiven Promotoren (Bai und Morozov 2010; Haberle und Lenhard 2016; Haberle und Stark 2018; Lenhard, Sandelin et al. 2012). Die Histonproteine des Nukleosoms können durch verschiedene biochemische Veränderungen stark verändert werden (Rivera und Ren 2013; Tan et al. 2011). Durch diese Veränderung wird die Zugänglichkeit des Chromatins reguliert. Das Zusammenspiel der Prozesse, welche die Verfügbarkeit der DNA regulieren und das Auslesen der regulatorischen Information steuern, muss durch die Klasse der TFs in abgestimmter Art und Weise erfolgen. Das von Zaret und Carroll (2011) eingeführte Pionierfaktorenkonzept lieferte wichtige Erkenntnisse in diesem Zusammenhang. In dieser Veröffentlichung wird für die Familie der der FOXA-Transkriptionsfaktoren (FOXA1, FOXA2, FOXA3) gezeigt, dass die Mitglieder dieser Familie mit kompaktem Chromatin interagieren können. Sie binden dabei sequenzspezifisch an ihre Nukleosomen-gebundenen TFBSs. Die Transkriptionsfaktoren OCT4, SOX2, KLF4 und MYC zeigen ebenfalls diese Pioniereigenschaft (Soufi et al. 2015). Im Unterschied zur FOXA-Familie erkennen diese Transkriptionsfaktoren schon Teilbereiche ihrer Sequenzmotive, welche in der Nukleosomengebundenen DNA exponiert werden. Durch die Kombination von DNase-seq-Daten mit ChIP-seq-Experimenten für verschiedene Transkriptionsfaktoren und deren zeitpunktspezifische Betrachtung, ausgehend von embryonischen Stammzellen der Maus und deren Differenzierung zu Bauchspeicheldrüsengewebe oder Darmgewebe, wurde das Konzept der Pionier-TF weiter verfeinert (Sherwood et al. 2014). Die Autoren dieser Veröffentlichung legen eine hierarchische Abhängigkeit der verwendeten TFs dar. Die untersuchten Pionierfaktoren in dieser Studie zeigen eine abgestimmte Kombination mit allgemeinen zelltyp- bzw. entwicklungsspezifischen TFs. Sowohl die Kombination der Pionierfaktoren

mit anderen nicht-Pionierfaktoren als auch deren Verdrängung bzw. Ablösung durch andere nachfolgende TFs wird in dieser Studie beschrieben.
Die in dieser Arbeit untersuchte Rolle der hochaffinen Transkriptionsfaktoren fügt sich gut in diese Beobachtungen ein. Die Besiedlung einer regulatorischen Region mit den möglichen TFs wird im Wesentlichen durch die hochaffinen/hochbewertete TFBS bestimmt. Die zelltypspezifischen regulatorischen Regionen werden durch die Pionierfaktoren geöffnet. Der Satz an TFs, welcher in einer Zelle vorhanden ist, kann nun sequenzspezifisch mit den exponierten TFBSs in diesen Regionen interagieren. Die hochaffinen Bindestellen werden am stabilsten mit den entsprechenden Masterregulatoren gebunden und bilden damit den Startpunkt eines regulatorischen Moduls in diesen regulatorischen Regionen. Weitere sequenzspezifische aber auch nicht-sequenzspezifische Faktoren und Kofaktoren lagern sich an und bilden so das vollständige regulatorische Modul aus. Eine jede mit regulatorischen Proteinen ausgestattete Sequenz definiert so eine mögliche Schnittstelle für weitere Interaktionen, welche dann das finale Enhanceosom bilden. Dabei können mehrere distale Bereiche mit einem proximalen Bereich interagieren und so die notwendige situationsspezifische Transkriptionsstärke regulieren.

Die besondere Kontrollfunktion des Promotors als entscheidende Schnittstelle für die Transkriptionsregulation wird bei der Analyse von Genexpressionsdaten in Zeitreihenexperimenten besonders deutlich. Abbildung 5.5 zeigt das aktive transkriptionsregulatorische Netzwerk der Masterregulatoren, welches bei der Differenzierung von menschlichem Herzmuskelgewebe zu beobachten ist. Ausgehend von einem Kontrollpunkt (D_{ctrl}) wird am Tag D_0 mit einem definierten experimentellen Protokoll gezielt die Entwicklung künstlich hergestellten menschlichen Herzmuskelgewebes und dessen fortschreitende Reifung gesteuert (Tiburcy et al. 2017). Die Genexpression aller aktiven Gene wird mit Hilfe von RNA-seq für acht verschiedene Zeitpunkte bestimmt (siehe Abbildung 5.5). Durch eine Korrelation der Genexpressionsprofile und deren Kombination mit dem in dieser Arbeit erstellten Transkriptionsnetzwerk können auf diese Weise TFs als zeitpunktspezifische Masterregulatoren identifiziert werden (Daou et al. 2020). Falls zwischen den identifizierten Regulatoren eine konservierte Bindestellenvorhersage im verwendeten Transkriptionsnetzwerk existiert, wird diese regulatorische Beziehung als Kante ins Netzwerk aufgenommen. Das Netzwerk zeigt, dass die zeitpunktspezifische Expression von Masterregulatoren durch vorherige Masterregulatoren reguliert werden. Die so aktivierten Masterregulatoren können nun nachfolgende Prozesse gezielt beeinflussen. Das dargestellte Netzwerk reprä-

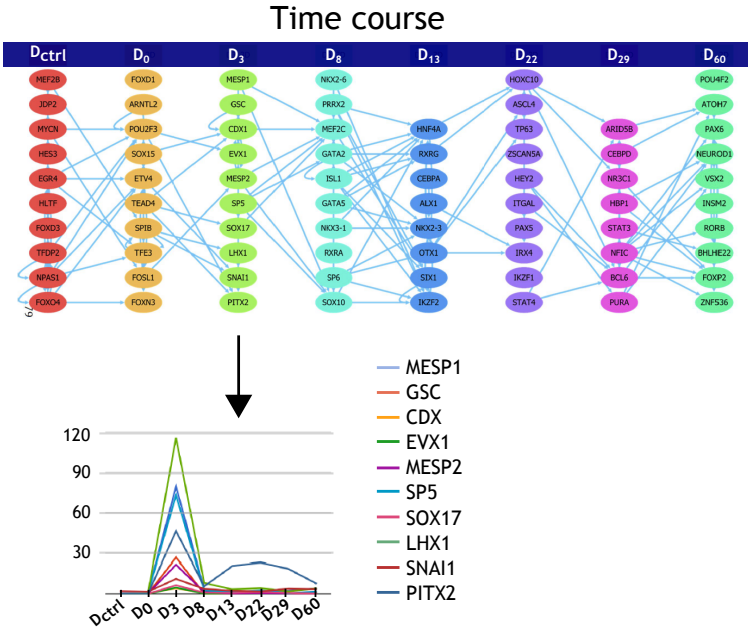


Abbildung 5.5: Zeitpunktspezifische Masterregulatoren im menschlichen Herzmuskelgewebe. Das regulatorische Netzwerk zeigt die Masterregulatoren für künstlich hergestelltes menschliches Herzmuskelgewebe. Ausgehend von definierten menschlichen Stammzellen des Herzmuskels kann durch Anwendung eines neuen experimentellen Protokolls differentiertes Herzmuskelgewebe erzeugt werden (Tiburcy et al. 2017). Insgesamt sind acht verschiedene Zeitpunkte gezeigt, an denen mit Hilfe von RNA-seq die Aktivität aller exprimierten Gene bestimmt worden ist. Der erste Zeitpunkt beschreibt dabei die Genaktivität auf Grundlage einzelner Herzmuskelzellen (Zeitpunt (D_{ctrl}). Die nachfolgenden Zeitpunkte (D_0, D_3, D_8, D_{13}, D_{22}, D_{60}) beschreiben feste Zeitpunkte, an denen unterschiedliche experimentelle Beeinflussungen vorgenommen werden, um das finale Herzmuskelgewebe in ausreichender Qualität erhalten zu können. Durch Anwendung der Korrelationsanalyse auf Grundlage der RNA-seq-Genexpressionsdaten können zeitpunktspezifische Masterregulatoren (TFs) identifiziert werden, welche ein Expressionsmaximum für einen vorliegenden Zeitpunkt aufweisen (Daou et al. 2020). In der Abbildung ist dies beispielhaft für den Zeitpunkt D_3 dargestellt. Die gelisteten TFs weisen ein Maximum für diesen Zeitpunkt auf und zeigen darüber hinaus ein ausreichend korreliertes Expressionsprofil über das gesamte Profil. Für die so identifizierten Masterregulatoren kann auf Grundlage des vorliegenden transkriptionsregulatorischen Netzwerks, welches in dieser Arbeit erstellt wurde, eine Verbindung zwischen zwei Masterregulatoren in das Netzwerk aufgenommen werden.

sentiert also die bedeutenden Masterregulatoren der menschlichen Herzentwicklung. Das besondere (zeitpunktspezifische) Expressionsprofil der Masterregulatoren ist ebenfalls in der Abbildung 5.5 dargestellt. Beispielhaft ist das Expressionsprofil der zehn Masterregulatoren für den Zeitpunkt D_3 gezeigt. Die in dieser Anwendung identifizierten TFs scheinen mehrheitlich ausschließlich zu einem Zeitpunkt maximal exprimiert zu werden. Diese Eigenschaft lässt sich auch für die anderen Zeitpunkte feststellen.

6 Ausblick

In dieser Arbeit wurde die Bedeutung einzelner charakteristischer TFBSs für die menschliche transkriptionelle Generegulation untersucht. In den verschiedenen Teilprojekten konnte die prägende Eigenschaft einer Bindestelle als definitorischer Kern (*Seed*) und deren interagierende TFs als Masterregulatoren eines potenziellen regulatorischen Moduls gezeigt werden. Diese besondere Kontrollfunktion der sequenzdefinierten Signale wurde dabei sowohl in Promotoren als auch in Enhancern nachgewiesen. Weiterhin konnte gezeigt werden, dass definierte Enhancer-Promoter-Interaktionen durch diese Masterregulatoren direkt (zwischen Masterregulatoren) und/oder indirekt (durch Kofaktoren gebundene Masterregulatoren bzw. der beteiligten Module) vermittelt werden.
In Eukaryoten wird die transkriptionelle Genregulation durch eine Menge von sequenzspezifisch und -unspezifisch bindenden TFs und Kofaktoren gesteuert. Ein solches Modul beeinflusst die Effizienz der Transkription in essentieller Art und Weise. Inwieweit die in dieser Arbeit untersuchten Masterregulatoren das gesamte regulatorische Modul (Menge aller lokal gebunden und/oder interagierende Proteine) beeinflussen, ist eine noch offene Fragestellung, welche in nachfolgenden Arbeiten weiter untersucht werden könnte.
Die vorliegende Datenbank mit ungefähr 42 Millionen potenziellen proximalen TFBSs kann dabei eine wertvolle Grundlage sein. Um ein möglichst vollständiges Bild der transkriptionellen Genregulation zu erhalten, ist es aber notwendig, Enhancer-definierte Masterregulatoren und deren Seeds zu identifizieren. Dazu kann der in dieser Arbeit entwickelte *Workflow* benutzt werden. Da die zu erwartende Datenmenge im Vergleich zu der im Promotor identifizierten Seeds um ein Vielfaches größer sein wird, sollte für weitere Projekte im Vorfeld über eine effiziente Speicherung und Anfragemöglichkeit der anfallenden Datenmengen nachgedacht werden.
Neu entwickelte Labormethoden erlauben es in naher Zukunft, zelltyp- und zeitpunktspezifische Enhancer-Promoter-Paarungen effizient und in guter Auflösung zu bestimmen.

Eine integrierte Analyse dieser Daten in Kombination mit der hier vorgestellten Analyse beteiligter Masterregulatoren könnte ein weiteres lohnenswertes Projekt darstellen. Als Analysewerkzeug in diesem Zusammenhang könnten sich verschiedene maschinelle Lernverfahren anbieten. Einige erfolgreiche Beispiele des *Deep Learnings* auf genomischen Daten deuten an, dass diese Methoden auch für die hier erwähnten Fragestellungen geeignet sein könnten.
Die Fortschritte auf dem Gebiet des *single cell sequencing* stellen eine vielversprechende weitere Anwendungsmöglichkeit des in dieser Arbeit erzeugten regulatorischen Netzwerks dar. Auf der Grundlage dieser Daten könnte es vielleicht möglich sein eine vielfältige Verwendung der Masterregulatoren zu untersuchen.

Literatur

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., … Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. doi:10.1093/nar/gky379

Agius, P., Arvey, A., Chang, W., Noble, W. S. & Leslie, C. (2010). High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS computational biology*, *6*(9). doi:10.1371/journal.pcbi.1000916

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *Molecular Biology of the Cell* (4th). Garland Science.

Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. Number: 8 Publisher: Nature Publishing Group. doi:10.1038/nbt.3300

Ambrosini, G., Groux, R. & Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, *34*(14), 2483–2484. doi:10.1093/bioinformatics/bty127

Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Annala, M., Laurila, K., Lähdesmäki, H. & Nykter, M. (2011). A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PloS One*, *6*(5), e20059. doi:10.1371/journal.pone.0020059

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T. & Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature biotechnology*, *35*(2), 136–144. doi:10.1038/nbt.3739

Arriza, J. L., Weinberger, C., Cerelli, G., Glaser, T. M., Handelin, B. L., Housman, D. E. & Evans, R. M. (1987). Cloning of human mineralocorticoid receptor complementary DNA: structural and functional kinship with the glucocorticoid receptor. *Science (New York, N.Y.) 237*(4812), 268–275. doi:10.1126/science.3037703

Arvey, A., Agius, P., Noble, W. S. & Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, *22*(9), 1723–1734. doi:10.1101/gr.127712.111

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Auton, A., Abecasis, G., Altshuler, D. & Durbin, R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., … Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, *53*(3), 354–366. doi:10.1038/s41588-021-00782-6

Babiarz, J. E., Ravon, M., Sridhar, S., Ravindran, P., Swanson, B., Bitter, H., … Kolaja, K. L. (2012). Determination of the human cardiomyocyte mRNA and miRNA differentiation network by fine-scale profiling. *Stem Cells and Development*, *21*(11), 1956–1965. doi:10.1089/scd.2011.0357

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., … Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science (New York, N.Y.) 324*(5935), 1720–1723. doi:10.1126/science.1162327

Bai, L. & Morozov, A. V. (2010). Gene regulation by nucleosome positioning. *Trends in genetics: TIG*, *26*(11), 476–483. doi:10.1016/j.tig.2010.08.003

Bailey, T. L. [T. L.] & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36.

Bailey, T. L. [Timothy L.], Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(suppl_2), W202–W208. doi:10.1093/nar/gkp335

Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., … Jones, S. J. M. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using

a sequencing-by-synthesis approach. *BMC genomics*, *7*, 246. doi:10.1186/1471-2164-7-246

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. doi:10.1093/nar/gks1193

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., ... Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. doi:10.1016/j.cell.2007.05.009

Bazett-Jones, D. P., Leblanc, B., Herfort, M. & Moss, T. (1994). Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF.

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., ... Grosse, I. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics (Oxford, England)*, *21*(11), 2657–2666. doi:10.1093/bioinformatics/bti410

Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. (1990). The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. *Cell*, *61*(1), 49–59.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. doi:10.1214/aos/1013699998

Bhar, A. (2015). Application of A Novel Triclustering Method in Analyzing Three Dimensional Transcriptomics Data. Zugriff 24. Juni 2019 unter https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0022-602C-1

Bhar, A., Haubrock, M., Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. & Wingender, E. (2013). Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. *Algorithms for molecular biology: AMB*, *8*(1), 9. doi:10.1186/1748-7188-8-9

Bhar, A., Haubrock, M., Mukhopadhyay, A. & Wingender, E. [Edgar]. (2015). Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes. *BMC bioinformatics*, *16*, 200. doi:10.1186/s12859-015-0635-8

Bianchi, M., Crinelli, R., Giacomini, E., Carloni, E. & Magnani, M. (2009). A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene. *Gene*, *448*(1), 88–101. doi:10.1016/j.gene.2009.08.013

Birney, E. P., Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. doi:10.1038/nature05874

Blankenberg, D., Taylor, J. & Nekrutenko, A. (2011). Making whole genome multiple alignments usable for biologists. *Bioinformatics*, *27*(17), 2426–2428. doi:10.1093/bioinformatics/btr398

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., … Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, *3*(5), e157. doi:10.1371/journal.pbio.0030157

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., … Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, *132*(2), 311–322. doi:10.1016/j.cell.2007.12.014

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., … Bruford, E. (2019). Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Research*, *47*(D1), D786–D792. doi:10.1093/nar/gky930

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, *212*(4), 563–578. doi:10.1016/0022-2836(90)90223-9

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E. & Mouse Genome Database Group. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, *47*(D1), D801–D806. doi:10.1093/nar/gky1056

Bulyk, M. L., Johnson, P. L. F. & Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, *30*(5), 1255–1261. doi:10.1093/nar/30.5.1255

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature*, *461*(7261), 193–198. doi:10.1038/nature08450

Carey, M. (1998). The enhanceosome and transcriptional synergy. *Cell*, *92*(1), 5–8.

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., … Brown, M. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics*, *38*(11), 1289–1297. doi:10.1038/ng1901

Chen, X., Hughes, T. R. & Morris, Q. (2007). RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics (Oxford, England)*, *23*(13), i72–79. doi:10.1093/bioinformatics/btm224

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., ... National High Blood Pressure Education Program Coordinating Committee. (2003). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA*, *289*(19), 2560–2572. doi:10.1001/jama.289.19.2560

Cisse, I. I., Izeddin, I., Causse, S. Z., Boudarene, L., Senecal, A., Muresan, L., ... Darzacq, X. (2013). Real-time dynamics of RNA polymerase II clustering in live human cells. *Science (New York, N.Y.) 341*(6146), 664–667. doi:10.1126/science.1239053

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., ... Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine*, *373*(10), 895–907. doi:10.1056/NEJMoa1502214

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. doi:10.1038/ng.3142

Core, L. J., Waterfall, J. J. & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.) 322*(5909), 1845–1848. doi:10.1126/science.1162228

Coss, D., Jacobs, S. B. R., Bender, C. E. & Mellon, P. L. (2004). A novel AP-1 site is critical for maximal induction of the follicle-stimulating hormone beta gene by gonadotropin-releasing hormone. *The Journal of Biological Chemistry*, *279*(1), 152–162. doi:10.1074/jbc.M304697200

Courey, A. J. & Huang, J.-D. (1995). The establishment and interpretation of transcription factor gradients in the Drosophila embryo. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, *1261*(1), 1–18. doi:10.1016/0167-4781(94)00234-T

Cowper-Sal·lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., ... Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature genetics*, *44*(11), 1191–1198. doi:10.1038/ng.2416

Cox, J. J., Willatt, L., Homfray, T. & Woods, C. G. (2011). A SOX9 duplication and familial 46,XX developmental testicular disorder. *The New England Journal of Medicine*, *364*(1), 91–93. doi:10.1056/NEJMc1010311

Cramer, P. (2002). Multisubunit RNA polymerases. *Current Opinion in Structural Biology*, *12*(1), 89–97.

Cramer, P. (2017). Structural Molecular Biology-A Personal Reflection on the Occasion of John Kendrew's 100th Birthday. *Journal of Molecular Biology*, *429*(17), 2603–2610. doi:10.1016/j.jmb.2017.05.007

D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, *24*(4), 423–425. doi:10.1038/nbt0406-423

Dalila, N. (2014). Genetic polymorphisms in genes regulating renal ion excretion and diuretic drug effects. Zugriff 8. Oktober 2019 unter https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0022-5F1E-7

Dalila, N., Brockmöller, J., Tzvetkov, M. V., Schirmer, M., Haubrock, M. & Vormfelde, S. V. (2015). Impact of mineralocorticoid receptor polymorphisms on urinary electrolyte excretion with and without diuretic drugs. *Pharmacogenomics*, *16*(2), 115–127. doi:10.2217/pgs.14.163

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., ... Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, *49*(7), 1073–1081. doi:10.1038/ng.3884

Daou, R., Beißbarth, T., Wingender, E., Gültas, M. & Haubrock, M. (2020). Constructing temporal regulatory cascades in the context of development and cell differentiation. *PLOS ONE*, *15*(4), e0231326. Publisher: Public Library of Science. doi:10.1371/journal.pone.0231326

Davidson, E. (2001). *Genomic Regulatory Systems*. Academic Press.

Davidson, E. (2006). *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press.

Dekker, J., Marti-Renom, M. A. & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews. Genetics*, *14*(6), 390–403. doi:10.1038/nrg3454

Delvecchio, M., Gaucher, J., Aguilar-Gurrieri, C., Ortega, E. & Panne, D. (2013). Structure of the p300 catalytic core and implications for chromatin targeting and HAT regula-

tion. *Nature Structural & Molecular Biology*, *20*(9), 1040–1046. doi:10.1038/nsmb.2642

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Zugriff 5. Juni 2019 unter https://www.jstor.org/stable/2984875

Deppmann, C. D., Alvania, R. S. & Taparowsky, E. J. (2006). Cross-Species Annotation of Basic Leucine Zipper Factor Interactions: Insight into the Evolution of Closed Interaction Networks. *Molecular Biology and Evolution*, *23*(8), 1480–1492. doi:10.1093/molbev/msl022

Deshane, J., Kim, J., Bolisetty, S., Hock, T. D., Hill-Kapturczak, N. & Agarwal, A. (2010). Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells. *The Journal of Biological Chemistry*, *285*(22), 16476–16486. doi:10.1074/jbc.M109.058586

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., … Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. doi:10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. doi:10.1038/nature11082

Djordjevic, M., Sengupta, A. M. & Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, *13*(11), 2381–2390. doi:10.1101/gr.1271603

Dreos, R., Ambrosini, G., Cavin Périer, R. & Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Research*, *41*(Database issue), D157–164. doi:10.1093/nar/gks1233

Dunham, I., Kundaje, A., Aldred, S., Collins, P. & Davis, C. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Durinck, S., Spellman, P. T., Birney, E. & Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. *Nature protocols*, *4*(8), 1184–1191. doi:10.1038/nprot.2009.97

Eggeling, R., Grosse, I. & Grau, J. (2017). InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics (Oxford, England)*, *33*(4), 580–582. doi:10.1093/bioinformatics/btw689

Ellington, A. D. & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, *346*(6287), 818–822. doi:10.1038/346818a0

Erikson, J., Nishikura, K., ar-Rushdi, A., Finan, J., Emanuel, B., Lenoir, G., … Croce, C. M. (1983). Translocation of an immunoglobulin kappa locus to a region 3' of an unrearranged c-myc oncogene enhances c-myc transcription. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(24), 7581–7585. doi:10.1073/pnas.80.24.7581

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., … Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. doi:10.1038/nature09906

Falcon, S. & Gentleman, R. [R.]. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)*, *23*(2), 257–258. doi:10 . 1093 / bioinformatics/btl567

Faniello, M. C., Bevilacqua, M. A., Condorelli, G., de Crombrugghe, B., Maity, S. N., Avvedimento, V. E., … Costanzo, F. (1999). The B subunit of the CAAT-binding factor NFY binds the central segment of the Co-activator p300. *The Journal of Biological Chemistry*, *274*(12), 7623–7626. doi:10.1074/jbc.274.12.7623

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. doi:10.1038/nature13835

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861–874.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., … Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, *45*(D1), D190–D199. doi:10.1093/nar/gkw1107

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–285. doi:10.1093/nar/gkv1344

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., … Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide

association summary statistics. *Nature Genetics*, *47*(11), 1228–1235. doi:10.1038/ng.3404

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806–811. doi:10.1038/35888

Fleming, J. D., Pavesi, G., Benatti, P., Imbriano, C., Mantovani, R. & Struhl, K. (2013). NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Research*, *23*(8), 1195–1209. doi:10.1101/gr.148080.112

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., Hoon, M. J. L. d., Haberle, V., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. doi:10.1038/nature13182

Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and molecular biology reviews: MMBR*, *79*(3), 347–372. doi:10.1128/MMBR.00006-15

Fudenberg, G. & Mirny, L. A. (2012). Higher-order chromatin structure: bridging physics and biology. *Current Opinion in Genetics & Development*, *22*(2), 115–124. doi:10.1016/j.gde.2012.01.006

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., … Engreitz, J. M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, *354*(6313), 769–773. doi:10.1126/science.aag2445

Fuxman Bass, J. I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., … Walhout, A. J. M. (2015). Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, *161*(3), 661–673. doi:10.1016/j.cell.2015.03.003

Galas, D. J. & Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, *5*(9), 3157–3170. doi:10.1093/nar/5.9.3157

Garcia-Saez, I., Menoni, H., Boopathi, R., Shukla, M. S., Soueidan, L., Noirclerc-Savoye, M., … Dimitrov, S. (2018). Structure of an H1-Bound 6-Nucleosome Array Reveals an Untwisted Two-Start Chromatin Fiber Conformation. *Molecular Cell*, *72*(5), 902–915.e7. doi:10.1016/j.molcel.2018.09.027

Gergics, P., Brinkmeier, M. L. & Camper, S. A. (2015). Lhx4 Deficiency: Increased Cyclin-Dependent Kinase Inhibitor Expression and Pituitary Hypoplasia. *Molecular Endocrinology*, *29*(4), 597–612. doi:10.1210/me.2014-1380

Gibcus, J. H. & Dekker, J. (2013). The hierarchy of the 3D genome. *Molecular Cell*, *49*(5), 773–782. doi:10.1016/j.molcel.2013.02.011

Gleyzer, N., Vercauteren, K. & Scarpulla, R. C. (2005). Control of mitochondrial transcription specificity factors (TFB1M and TFB2M) by nuclear respiratory factors (NRF-1 and NRF-2) and PGC-1 family coactivators. *Molecular and Cellular Biology*, *25*(4), 1354–1366. doi:10.1128/MCB.25.4.1354-1366.2005

Goldstein, D., El-Maraghi, R. H., Hammel, P., Heinemann, V., Kunzmann, V., Sastre, J., ... Von Hoff, D. D. (2015). nab-Paclitaxel plus gemcitabine for metastatic pancreatic cancer: long-term survival from a phase III trial. *Journal of the National Cancer Institute*, *107*(2). doi:10.1093/jnci/dju413

Goodrich, J. A. & Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nature Reviews. Genetics*, *11*(8), 549–558. doi:10.1038/nrg2847

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. & Bulyk, M. L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports*, *3*(4), 1093–1104. doi:10.1016/j.celrep.2013.03.014

Granger, B. R., Chang, Y.-C., Wang, Y., DeLisi, C., Segrè, D. & Hu, Z. (2016). Visualization of Metabolic Interaction Networks in Microbial Communities Using VisANT 5.0. *PLoS computational biology*, *12*(4), e1004875. doi:10.1371/journal.pcbi.1004875

Grau, J., Posch, S., Grosse, I. & Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, *41*(21), e197. doi:10.1093/nar/gkt831

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A. & Elhaik, E. (2013). On the immortality of television sets: "functionïn the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, *5*(3), 578–590. doi:10.1093/gbe/evt028

Haberle, V. & Lenhard, B. (2016). Promoter architectures and developmental gene regulation. *Seminars in Cell & Developmental Biology*, *57*, 11–23. doi:10.1016/j.semcdb.2016.01.014

Haberle, V. & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 1. doi:10.1038/s41580-018-0028-8

Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature*, *353*(6346), 715–719. doi:10.1038/353715a0

Haubrock, M., Hartmann, F. & Wingender, E. (2016). NF-Y Binding Site Architecture Defines a C-Fos Targeted Promoter Class. *PloS one*, *11*(8), e0160803. doi:10.1371/journal.pone.0160803

Haubrock, M., Li, J. & Wingender, E. (2012). Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks. *BMC systems biology*, *6 Suppl 2*, S15. doi:10.1186/1752-0509-6-S2-S15

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010a). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. doi:10.1016/j.molcel.2010.05.004

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010b). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. doi:10.1016/j.molcel.2010.05.004

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., … Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, *6*(4), 283–289. doi:10.1038/nmeth.1313

Hooghe, B., Broos, S., van Roy, F. & De Bleser, P. (2012). A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Research*, *40*(14), e106. doi:10.1093/nar/gks283

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. doi:10.1093/nar/gkn923

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. (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*(Database issue), D117–122. doi:10.1093/nar/gku1045

Ihaka, R. & Gentleman, R. [Robert]. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. doi:10.1080/10618600.1996.10474713

Innocenti, F., Owzar, K., Cox, N. L., Evans, P., Kubo, M., Zembutsu, H., … Ratain, M. J. (2012). A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *18*(2), 577–584. doi:10.1158/1078-0432.CCR-11-1387

Jenuwein, T. & Allis, C. D. (2001). Translating the histone code. *Science (New York, N.Y.) 293*(5532), 1074–1080. doi:10.1126/science.1063127

John, M., Leppik, R., Busch, S. J., Granger-Schnarr, M. & Schnarr, M. (1996). DNA binding of Jun and Fos bZip domains: homodimers and heterodimers induce a DNA conformational change in solution. *Nucleic Acids Research*, *24*(22), 4487–4494.

Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.) 316*(5830), 1497–1502. doi:10.1126/science.1141319

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., … Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, *152*(1-2), 327–339. doi:10.1016/j.cell.2012.12.009

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., … Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, *527*(7578), 384–388. doi:10.1038/nature15518

Kähärä, J. & Lähdesmäki, H. (2013). Evaluating a linear k-mer model for protein-DNA interactions using high-throughput SELEX data. *BMC bioinformatics*, *14 Suppl 10*, S2. doi:10.1186/1471-2105-14-S10-S2

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, *47*(D1), D590–D595. doi:10.1093/nar/gky962

Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481. doi:10.1080/01621459.1958.10501452

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., … Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function

for pervasive transcription. *Science (New York, N.Y.) 316*(5830), 1484–1488. doi:10.1126/science.1138341

Keilwagen, J., Grau, J., Paponov, I. A., Posch, S., Strickert, M. & Grosse, I. (2011). De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS computational biology*, *7*(2), e1001070. doi:10.1371/journal.pcbi.1001070

Kel, A. E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. & Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic acids research*, *31*(13), 3576–3579.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., ... Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, *10*(10), e1004722. doi:10.1371/journal.pgen.1004722

Kim, I. S., Sinha, S., de Crombrugghe, B. & Maity, S. N. (1996). Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule. *Molecular and Cellular Biology*, *16*(8), 4003–4013. doi:10.1128/mcb.16.8.4003

Kleinjan, D. J. & van Heyningen, V. (1998). Position effect in human genetic disease. *Human Molecular Genetics*, *7*(10), 1611–1618. doi:10.1093/hmg/7.10.1611

Kostrewa, D., Zeller, M. E., Armache, K.-J., Seizl, M., Leike, K., Thomm, M. & Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, *462*(7271), 323–330. doi:10.1038/nature08548

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., ... Makeev, V. J. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, *46*(D1), D252–D259. doi:10.1093/nar/gkx1106

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., ... Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. doi:10.1016/j.cell.2018.01.029

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.) 262*(5131), 208–214.

Lawrence, C. E. & Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, *7*(1), 41–51. doi:10.1002/prot.340070105

Lee, S.-K., Jurata, L. W., Funahashi, J., Ruiz, E. C. & Pfaff, S. L. (2004). Analysis of embryonic motoneuron gene regulation: derepression of general activators function in concert with enhancer factors. *Development (Cambridge, England)*, *131*(14), 3295–3306. doi:10.1242/dev.01179

Lenhard, B., Sandelin, A. & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews. Genetics*, *13*(4), 233–245. doi:10.1038/nrg3163

Lenhard, B. & Wasserman, W. W. (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics (Oxford, England)*, *18*(8), 1135–1136. doi:10.1093/bioinformatics/18.8.1135

Letunic, I. & Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, *46*(D1), D493–D496. doi:10.1093/nar/gkx922

Levine, M., Cattoglio, C. & Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. *Cell*, *157*(1), 13–25. doi:10.1016/j.cell.2014.02.009

Li, J., Hua, X., Haubrock, M., Wang, J. & Wingender, E. (2012). The architecture of the gene regulatory networks of different tissues. *Bioinformatics (Oxford, England)*, *28*(18), i509–i514. doi:10.1093/bioinformatics/bts387

Lima-Mendez, G. & van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, *5*(12), 1482–1493. doi:10.1039/b908681a

Linhart, C., Halperin, Y. & Shamir, R. (2008). Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome research*, *18*(7), 1180–1189. doi:10.1101/gr.076117.108

Lönnstedt, I. & Speed, T. (2002). Replicated microarray data. Zugriff 23. Juni 2019 unter http://www3.stat.sinica.edu.tw/statistica/j12n1/j12n12/j12n12.htm

Love, M. I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. doi:10.1186/s13059-014-0550-8

Luehr, S., Hartmann, H. & Söding, J. (2012). The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. *Nucleic Acids Research*, *40*(Web Server issue), W104–109. doi:10.1093/nar/gks602

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, *389*(6648), 251–260. doi:10.1038/38444

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., ... Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5), 1012–1025. doi:10.1016/j.cell.2015.04.004

Lüske, C. (2015). Molecular determinants for the outcome in gemcitabine-treated pancreatic cancer. Zugriff 13. Juni 2019 unter https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0028-865F-A

Ma, W., Yang, L., Rohs, R. & Noble, W. S. (2017). DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics (Oxford, England)*, *33*(19), 3003–3010. doi:10.1093/bioinformatics/btx336

Maienschein-Cline, M., Dinner, A. R., Hlavacek, W. S. & Mu, F. (2012). Improved predictions of transcription factor binding sites using physicochemical features of DNA. *Nucleic Acids Research*, *40*(22), e175. doi:10.1093/nar/gks771

Mancia, G., Fagard, R., Narkiewicz, K., Redon, J., Zanchetti, A., Böhm, M., ... Wood, D. A. (2013). 2013 ESH/ESC guidelines for the management of arterial hypertension: the Task Force for the Management of Arterial Hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). *European Heart Journal*, *34*(28), 2159–2219. doi:10.1093/eurheartj/eht151

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-Y., Denay, G., Lee, J., ... Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *44*(D1), D110–115. doi:10.1093/nar/gkv1176

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R. & Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, *3*(3), 278–286.e4. doi:10.1016/j.cels.2016.07.001

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta*, *405*(2), 442–451. doi:10.1016/0005-2795(75)90109-9

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., … Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, *31*(1), 374–378.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., … Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, *34*(Database issue), D108–D110. doi:10.1093/nar/gkj143

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012a). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.) 337*(6099), 1190–1195. doi:10.1126/science.1222794

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012b). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.) 337*(6099), 1190–1195. doi:10.1126/science.1222794

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. doi:10.1038/nbt.1630

McMurray, J. J. V., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., Dickstein, K., … ESC Committee for Practice Guidelines. (2012). ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European Journal of Heart Failure*, *14*(8), 803–869. doi:10.1093/eurjhf/hfs105

Mercer, T. R. [Tim R.], Edwards, S. L., Clark, M. B., Neph, S. J., Wang, H., Stergachis, A. B., … Stamatoyannopoulos, J. A. (2013). DNase I–hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics*, *45*(8), 852–859. doi:10.1038/ng.2677

Mercer, T. R. [Tim R] & Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome research*, *23*(7), 1081–1088. doi:10.1101/gr.156612.113

Merika, M., Williams, A. J., Chen, G., Collins, T. & Thanos, D. (1998). Recruitment of CBP/p300 by the IFN Enhanceosome Is Required for Synergistic Activation of Transcription. *Molecular Cell*, *1*(2), 277–287. Publisher: Elsevier. doi:10.1016/S1097-2765(00)80028-3

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., … Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–560. doi:10.1038/nature06008

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. doi:10.1038/nmeth.1226

Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., … Young, R. A. (2011). Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell*, *147*(3), 565–576. doi:10.1016/j.cell.2011.08.050

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., … Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714–719. doi:10.1038/nature09266

Nakabeppu, Y., Ryder, K. & Nathans, D. (1988). DNA binding activities of three murine Jun proteins: stimulation by Fos. *Cell*, *55*(5), 907–915.

Nakato, R. & Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*, *18*(2), 279–290. doi:10.1093/bib/bbw023

Nardini, M., Gnesutta, N., Donati, G., Gatta, R., Forni, C., Fossati, A., … Mantovani, R. (2013). Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell*, *152*(1-2), 132–143. doi:10.1016/j.cell.2012.11.047

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414), 83–90. doi:10.1038/nature11212

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*(12), 1135–1137. doi:10.1038/nbt1209-1135

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–745. doi:10.1093/nar/gkv1189

Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell*, *87*(5), 953–959.

Oliphant, A. R. & Struhl, K. (1989). An efficient method for generating proteins with altered enzymatic properties: application to beta-lactamase. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(23), 9094–9098. doi:10.1073/pnas.86.23.9094

Omole, A. E. & Fakoya, A. O. J. (2018). Ten years of progress and promise of induced pluripotent stem cells: historical origins, characteristics, mechanisms, limitations, and potential applications. *PeerJ*, *6*, e4370. doi:10.7717/peerj.4370

Ortega, E., Rengachari, S., Ibrahim, Z., Hoghoughi, N., Gaucher, J., Holehouse, A. S., ... Panne, D. (2018). Transcription factor dimerization activates the p300 acetyltransferase. *Nature*, *562*(7728), 538–544. doi:10.1038/s41586-018-0621-1

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., ... Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, *36*(10), 1065–1071. doi:10.1038/ng1423

Pavesi, G., Mauri, G. & Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics (Oxford, England)*, *17 Suppl 1*, S207–214. doi:10.1093/bioinformatics/17.suppl_1.s207

Pei, D. (2009). Regulation of pluripotency and reprogramming by transcription factors. *The Journal of Biological Chemistry*, *284*(6), 3365–3369. doi:10.1074/jbc.R800063200

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, *94*(4), 559–573. doi:10.1016/j.ajhg.2014.03.004

Preston, G. A., Srinivasan, D. & Barrett, J. C. (2000). Apoptotic response to growth factor deprivation involves cooperative interactions between c-Fos and p300. *Cell Death and Differentiation*, *7*(2), 215–226. doi:10.1038/sj.cdd.4400637

Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl_1), D61–D65. doi:10.1093/nar/gkl842

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research*, *23*(23), 4878–4884. doi:10.1093/nar/23.23.4878

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., ... Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Rickman, C. & Bickmore, W. A. (2013). Transcription. Flashing a light on the spatial organization of transcription. *Science (New York, N.Y.) 341*(6146), 621–622. doi:10.1126/science.1242889

Rivera, C. M. & Ren, B. (2013). Mapping human epigenomes. *Cell*, *155*(1), 39–55. doi:10.1016/j.cell.2013.09.011

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., ... Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. doi:10.1038/nature14248

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., ... Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. doi:10.1038/nmeth1068

Rodríguez-Paredes, M. & Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, *17*(3), 330–339. doi:10.1038/nm.2305

Roeder, R. G. (1996). [14] Nuclear RNA polymerases: Role of general initiation factors and cofactors in eukaryotic transcription. In *Methods in Enzymology* (Bd. 273, S. 165–171). RNA Polymerase and Associated Factors Part A. doi:10.1016/S0076-6879(96)73016-1

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. & Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature*, *461*(7268), 1248–1253. doi:10.1038/nature08473

Roider, H. G., Manke, T., O'Keeffe, S., Vingron, M. & Haas, S. A. (2009). PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, *25*(4), 435–442. doi:10.1093/bioinformatics/btn627

Romier, C., Cocchiarella, F., Mantovani, R. & Moras, D. (2003). The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor

NF-Y. *The Journal of Biological Chemistry*, *278*(2), 1336–1345. doi:10.1074/jbc.M209635200

Roppel, S. (2013). Functional Assessment of Biomarkers in Gemcitabine-Treated Pancreatic Cancer with Specific Focus on Nucleoside Transporter ENT1. Zugriff 13. Juni 2019 unter https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0001-BC18-5

Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, *16*(10), 939–945. doi:10.1038/nbt1098-939

Sainsbury, S., Bernecky, C. & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, *16*(3), 129–143. doi:10.1038/nrm3952

Sauer, T., Shelest, E. & Wingender, E. (2006). Evaluating phylogenetic footprinting for human-rodent comparisons. *Bioinformatics (Oxford, England)*, *22*(4), 430–437. doi:10.1093/bioinformatics/bti819

Schirmer, M. A., Lüske, C. M., Roppel, S., Schaudinn, A., Zimmer, C., Pflüger, R., … Ghadimi, B. M. (2016). Relevance of Sp Binding Site Polymorphism in WWOX for Treatment Outcome in Pancreatic Cancer. *Journal of the National Cancer Institute*, *108*(5). doi:10.1093/jnci/djv387

Sharon, E., Lubliner, S. & Segal, E. (2008). A Feature-Based Approach to Modeling Protein–DNA Interactions. *PLOS Computational Biology*, *4*(8), e1000154. doi:10.1371/journal.pcbi.1000154

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., … Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, *32*(2), 171–178. doi:10.1038/nbt.2798

Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annual Review of Biochemistry*, *81*, 65–95. doi:10.1146/annurev-biochem-051710-134100

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., … Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15776–15781. doi:10.1073/pnas.2136655100

Shlyueva, D., Stampfel, G. & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews. Genetics*, *15*(4), 272–286. doi:10.1038/nrg3682

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS One*, *5*(3), e9722. doi:10.1371/journal.pone.0009722

Siebert, M. & Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, *44*(13), 6055–6069. doi:10.1093/nar/gkw521

Sinha, S., Kim, I. S., Sohn, K. Y., de Crombrugghe, B. & Maity, S. N. (1996). Three classes of mutations in the A subunit of the CCAAT-binding factor CBF delineate functional domains involved in the three-step assembly of the CBF-DNA complex. *Molecular and Cellular Biology*, *16*(1), 328–337. doi:10.1128/mcb.16.1.328

Sinha, S., Maity, S. N., Lu, J. & Crombrugghe, B. d. (1995). Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3. *Proceedings of the National Academy of Sciences*, *92*(5), 1624–1628. doi:10.1073/pnas.92.5.1624

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R. & Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, *39*(9), 381–399. doi:10.1016/j.tibs.2014.07.002

Song, S., Liu, L., Edwards, S. V. & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 14942–14947. doi:10.1073/pnas.1211733109

Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M. & Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, *161*(3), 555–568. doi:10.1016/j.cell.2015.03.017

Stender, J. D., Frasor, J., Komm, B., Chang, K. C. N., Kraus, W. L. & Katzenellenbogen, B. S. (2007). Estrogen-regulated gene networks in human breast cancer cells: involvement of E2F1 in the regulation of cell proliferation. *Molecular Endocrinology (Baltimore, Md.) 21*(9), 2112–2123. doi:10.1210/me.2006-0474

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, *16*(1), 16–23. doi:10.1093/bioinformatics/16.1.16

Stormo, G. (2013). *Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics*. Google-Books-ID: BweZMQEACAAJ. Cold Spring Harbor Laboratory Press.

Sutherland, H. & Bickmore, W. A. (2009). Transcription factories: gene expression in unions? *Nature Reviews. Genetics*, *10*(7), 457–466. doi:10.1038/nrg2592

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., … Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, *24*(3), 390–400. doi:10.1101/gr.163519.113

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., … Zhao, Y. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, *146*(6), 1016–1028. doi:10.1016/j.cell.2011.08.008

Tanaka, R., Yi, T.-M. & Doyle, J. (2005). Some protein interaction data do not exhibit power law statistics. *FEBS letters*, *579*(23), 5140–5144. doi:10.1016/j.febslet.2005.08.024

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. doi:10.1038/nature11232

Tiburcy, M., Hudson, J. E., Balfanz, P., Schlick, S., Meyer, T., Chang Liao, M.-L., … Zimmermann, W.-H. (2017). Defined Engineered Human Myocardium With Advanced Maturation for Applications in Heart Failure Modeling and Repair. *Circulation*, *135*(19), 1832–1847. Publisher: American Heart Association. doi:10.1161/CIRCULATIONAHA.116.024145

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., … Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, *23*(1), 137–144. doi:10.1038/nbt1053

Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., … Schneider, R. (2013). Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell*, *152*(4), 859–872. doi:10.1016/j.cell.2013.01.032

Tuerk, C. & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.)* *249*(4968), 505–510.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. doi:10.1093/nar/gky1049

Vallania, F., Schiavone, D., Dewilde, S., Pupo, E., Garbay, S., Calogero, R., … Poli, V. (2009). Genome-wide discovery of functional transcription factor binding sites by comparative genomics: the case of Stat3. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(13), 5117–5122. doi:10.1073/pnas.0900473106

van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. (2014). In search of the determinants of enhancer-promoter interaction specificity. *Trends in Cell Biology*, *24*(11), 695–702. doi:10.1016/j.tcb.2014.07.004

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.) 291*(5507), 1304–1351. doi:10.1126/science.1058040

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A. & Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, *10*(8), 1297–1309. doi:10.1016/j.celrep.2015.02.004

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Vormfelde, S. V., Sehrt, D., Toliat, M. R., Schirmer, M., Meineke, I., Tzvetkov, M., … Brockmöller, J. (2007). Genetic variation in the renal sodium transporters NKCC2, NCC, and ENaC in relation to the effects of loop diuretic drugs. *Clinical Pharmacology and Therapeutics*, *82*(3), 300–309. doi:10.1038/sj.clpt.6100131

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., … Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, *22*(9), 1798–1812. doi:10.1101/gr.139105.112

Wang, W.-M., Wu, S.-Y., Lee, A.-Y. & Chiang, C.-M. (2011). Binding site specificity and factor redundancy in activator protein-1-driven human papillomavirus chromatin-dependent transcription. *The Journal of Biological Chemistry*, *286*(47), 40974–40986. doi:10.1074/jbc.M111.290874

Ward, L. D. & Bussemaker, H. J. (2008). Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, *24*(13), i165–i171. doi:10.1093/bioinformatics/btn154

Wasserman, W. W. & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews. Genetics*, *5*(4), 276–287. doi:10.1038/nrg1315

Watson, J. D. & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*, 737. Zugriff unter http://dx.doi.org/10.1038/171737a0

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., ... Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, *158*(6), 1431–1443. doi:10.1016/j.cell.2014.08.009

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., ... Wagner, L. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, *31*(1), 28–33. doi:10.1093/nar/gkg033

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., ... Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. doi:10.1016/j.cell.2013.03.035

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., ... Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*(7199), 1239–1243. doi:10.1038/nature07002

Wingender, E. [E.]. (1988). Compilation of transcription regulating proteins. *Nucleic Acids Research*, *16*(5), 1879–1902. doi:10.1093/nar/16.5.1879

Wingender, E. [E.]. (1997). Classification of eukaryotic transcription factors. *Molekuliarnaia Biologiia*, *31*(4), 584–600.

Wingender, E., Schoeps, T. & Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, *41*(D1), D165–D170. doi:10.1093/nar/gks1123

Wingender, E., Schoeps, T., Haubrock, M. & Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Research*, *43*(Database issue). doi:10.1093/nar/gku1064

Wingender, E., Schoeps, T., Haubrock, M., Krull, M. & Dönitz, J. (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, *46*(D1), D343–D347. doi:10.1093/nar/gkx987

Wingender, E. [Edgar]. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics*, *9*(4), 326–332. doi:10.1093/bib/bbn016

Xie, D., Boyle, A. P., Wu, L., Zhai, J., Kawli, T. & Snyder, M. (2013). Dynamic trans-acting factor colocalization in human cells. *Cell*, *155*(3), 713–724. doi:10.1016/j.cell.2013.09.043

Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Drazner, M. H., ... Wilkoff, B. L. (2013). 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation*, *128*(16), 1810–1852. doi:10.1161/CIR.0b013e31829e8807

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, *389*(1), 52–65. doi:10.1016/j.gene.2006.09.029

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R. & Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research*, *42*(Database issue), D148–155. doi:10.1093/nar/gkt1087

Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. (2019). GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Research*, *47*(D1), D100–D105. doi:10.1093/nar/gky1128

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., ... Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (New York, N.Y.) 356*(6337). doi:10.1126/science.aaj2239

Zambelli, F., Pesole, G. & Pavesi, G. (2013). PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic acids research*, *41*(Web Server issue), W535–W543. doi:10.1093/nar/gkt448

Zaret, K. S. & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, *25*(21), 2227–2241. doi:10.1101/gad.176826.111

Zemzoumi, K., Frontini, M., Bellorini, M. & Mantovani, R. (1999). NF-Y histone fold alpha1 helices help impart CCAAT specificity. *Journal of Molecular Biology*, *286*(2), 327–337. doi:10.1006/jmbi.1998.2496

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., ... Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. doi:10.1093/nar/gkx1098

Zhao, Y., Ruan, S., Pandey, M. & Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, *191*(3), 781–790. doi:10.1534/genetics.112.138685

Zhou, J. & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931–934. doi:10.1038/nmeth.3547

Zhou, Q. & Liu, J. S. (2008). Extracting sequence features to predict protein-DNA interactions: a comparative study. *Nucleic Acids Research*, *36*(12), 4137–4148. doi:10.1093/nar/gkn361

A Appendix

A.1 The architecture of the gene regulatory networks of different tissues

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i509–i514
doi:10.1093/bioinformatics/bts387

The architecture of the gene regulatory networks of different tissues

Jie Li[1,2], Xu Hua[1], Martin Haubrock[2], Jin Wang[1,*] and Edgar Wingender[2,*]

[1]The State Key Laboratory of Pharmaceutical Biotechnology and Jiangsu Engineering Research Center for MicroRNA Biology and Biotechnology and [2]Department of Bioinformatics, University Medical Center Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany

ABSTRACT

Summary: The great variety of human cell types in morphology and function is due to the diverse gene expression profiles that are governed by the distinctive regulatory networks in different cell types. It is still a challenging task to explain how the regulatory networks achieve the diversity of different cell types. Here, we report on our studies of the design principles of the tissue regulatory system by constructing the regulatory networks of eight human tissues, which subsume the regulatory interactions between transcription factors (TFs), microRNAs (miRNAs) and non-TF target genes. The results show that there are in-/out-hubs of high in-/out-degrees in tissue networks. Some hubs (strong hubs) maintain the hub status in all the tissues where they are expressed, whereas others (weak hubs), in spite of their ubiquitous expression, are hubs only in some tissues. The network motifs are mostly feed-forward loops. Some of them having no miRNAs are the common motifs shared by all tissues, whereas the others containing miRNAs are the tissue-specific ones owned by one or several tissues, indicating that the transcriptional regulation is more conserved across tissues than the post-transcriptional regulation. In particular, a common bow-tie framework was found that underlies the motif instances and shows diverse patterns in different tissues. Such bow-tie framework reflects the utilization efficiency of the regulatory system as well as its high variability in different tissues, and could serve as the model to further understand the structural adaptation of the regulatory system to the specific requirements of different cell functions.

Contact: edgar.wingender@bioinf.med.uni-goettingen.de; jwang@nju.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

There are some hundred of cell types in the human body harboring the same genome, but showing quite diverse behaviors in morphology and biological functions. This suggests that the cells of different types are developed by eliciting different gene regulatory networks, all encoded in the human genome (Ben-Tabou de-Leon and Davidson, 2007). This raises a number of questions: what is the mechanism that governs the activation of different gene regulatory networks? Are the gene regulatory networks accounting for the different types of cells designed under a same principle or architectural framework? What kinds of differences in regulatory networks contribute to the diversity of cell types? These are key questions in understanding the mechanisms underlying the specific functions of different cell types and are essential for cell regeneration that closely relates to the treatment of tissue damage and injury involved in various diseases.

The gene regulatory network of a cell is supposed to contain the comprehensive regulatory information governing the gene expression in this cell. In previous works, the gene regulatory network mainly referred to the transcriptional network which depicts the regulations at transcriptional level (Babu *et al.*, 2004; Cosentino Lagomarsino *et al.*, 2007; Yu *et al.*, 2003). Nevertheless, the microRNA (miRNA), a small RNA that negatively regulates the gene expression at post-transcriptional level by binding to the mRNA sequences of its target gene, was found at the end of the 20th century (Lee *et al.*, 1993; Wightman *et al.*, 1993). So far, >20 000 microRNAs of different species have been identified (Griffiths-Jones *et al.*, 2008), and many useful databases and effective tools that are used to predict the targets of miRNAs have been developed (Krek *et al.*, 2005; Lewis *et al.*, 2003; Sethupathy *et al.*, 2006). This gives the possibility to construct a more comprehensive map of gene regulatory networks for certain cell types by integrating transcriptional and post-transcriptional regulations of gene expression.

Nowadays, there are several studies on the gene regulatory networks covering both the transcriptional and post-transcriptional regulation of genes. For example, Pilpel and co-workers constructed a global miRNA–TF regulatory network for mammalian and studied the combinatory regulations between TFs and miRNAs according to the local and global architecture of the network (Shalgi *et al.*, 2007). Qian *et al.* investigated the biological functions of miRNAs by pursuing the miRNA motif profiles in the miRNA-TF regulatory network (Yu *et al.*, 2008). Gersten *et al.* developed an integrated strategy to construct and analyze the gene regulatory network from high-throughput sequencing data (Cheng *et al.*, 2011). However, all of these works are focused on the genome-wide level. Studies about the regulatory networks of specific tissues and their design principles are largely missing.

Here, we constructed the large-scale gene regulatory networks of eight human tissues and investigated their design principles from the perspective of three levels, i.e. the local structure of vertices (i.e. degrees), the small circuits (i.e. network motifs) and the assembly of small circuits. Our results show that the tissue-specific regulatory networks (TRNs) constructed and analyzed here contain hub nodes, the specific features of which may vary considerably among the tissues investigated. The different TRNs also vary significantly with regard to the composition of topological motifs, the instances of which are organized in a bow-tie structure. The bow-tie structures

*To whom correspondence should be addressed.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

A.2 Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks

Haubrock et al. BMC Systems Biology 2012, 6(Suppl 2):S15
http://www.biomedcentral.com/1752-0509/6/S2/S15

BMC Systems Biology

PROCEEDINGS Open Access

Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks

Martin Haubrock, Jie Li, Edgar Wingender*

From 23rd International Conference on Genome Informatics (GIW 2012)
Tainan, Taiwan. 12-14 December 2012

Abstract

Background: Transcriptional networks of higher eukaryotes are difficult to obtain. Available experimental data from conventional approaches are sporadic, while those generated with modern high-throughput technologies are biased. Computational predictions are generally perceived as being flooded with high rates of false positives. New concepts about the structure of regulatory regions and the function of master regulator sites may provide a way out of this dilemma.

Methods: We combined promoter scanning with positional weight matrices with a 4-genome conservativity analysis to predict high-affinity, highly conserved transcription factor (TF) binding sites and to infer TF-target gene relations. They were expanded to paralogous TFs and filtered for tissue-specific expression patterns to obtain a reference transcriptional network (RTN) as well as tissue-specific transcriptional networks (TTNs).

Results: When validated with experimental data sets, the predictions done showed the expected trends of true positive and true negative predictions, resulting in satisfying sensitivity and specificity characteristics. This also proved that confining the network reconstruction to the 1% top-ranking TF-target predictions gives rise to networks with expected degree distributions. Their expansion to paralogous TFs enriches them by tissue-specific regulators, providing a reasonable basis to reconstruct tissue-specific transcriptional networks.

Conclusions: The concept of master regulator or seed sites provides a reasonable starting point to select predicted TF-target relations, which, together with a paralogous expansion, allow for reconstruction of tissue-specific transcriptional networks.

Background

Regulation of transcription is mediated through complex arrays of transcription factor binding sites (TFBSs), which constitute promoter and enhancer regions. In spite of the advent of high-throughput approaches to identify TFBSs in a given cellular context, the available information, most comprehensively collected in the TRANSFAC® database [1], is still fragmented and biased with regard to the systems selected. Consequently, any transcriptional network reconstructed from the available experimental data is highly incomplete. This situation deteriorates further when filtering such a transcriptional "reference" network for gene expression data in order to generate tissue-specific networks. Therefore, constructing comprehensive gene regulatory networks still depends on reliable algorithms for predicting individual TFBSs as a basis for inferring TF-target gene relations. These predictions, however, depend on the availability of information about the DNA-binding specificity of ideally all TFs encoded by a genome. Unfortunately, we are far from this ideal situation, so that we can do such predictions only for a subset of, e.g., human TFs. Although promising methods have been reported for inferring DNA-binding specificities by homology modeling [2,3], the required 3D

* Correspondence: Edgar.Wingender@bioinf.med.uni-goettingen.de
Department of Bioinformatics, University Medical Center Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany

A.3 Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell

Bhar *et al. Algorithms for Molecular Biology* 2013, **8**:9
http://www.almob.org/content/8/1/9

ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell

Anirban Bhar[1*], Martin Haubrock[1], Anirban Mukhopadhyay[2], Ujjwal Maulik[3*], Sanghamitra Bandyopadhyay[4*] and Edgar Wingender[1*]

Abstract

Background: Estrogen is a chemical messenger that has an influence on many breast cancers as it helps cells to grow and divide. These cancers are often known as estrogen responsive cancers in which estrogen receptor occupies the surface of the cells. The successful treatment of breast cancers requires understanding gene expression, identifying of tumor markers, acquiring knowledge of cellular pathways, etc. In this paper we introduce our proposed triclustering algorithm δ-TRIMAX that aims to find genes that are coexpressed over subset of samples across a subset of time points. Here we introduce a novel mean-squared residue for such 3D dataset. Our proposed algorithm yields triclusters that have a mean-squared residue score below a threshold δ.

Results: We have applied our algorithm on one simulated dataset and one real-life dataset. The real-life dataset is a time-series dataset in estrogen induced breast cancer cell line. To establish the biological significance of genes belonging to resultant triclusters we have performed gene ontology, KEGG pathway and transcription factor binding site enrichment analysis. Additionally, we represent each resultant tricluster by computing its eigengene and verify whether its eigengene is also differentially expressed at early, middle and late estrogen responsive stages. We also identified hub-genes for each resultant triclusters and verified whether the hub-genes are found to be associated with breast cancer. Through our analysis *CCL2, CD47, NFIB, BRD4, HPGD, CSNK1E, NPC1L1, PTEN, PTPN2 and ADAM9* are identified as hub-genes which are already known to be associated with breast cancer. The other genes that have also been identified as hub-genes might be associated with breast cancer or estrogen responsive elements. The TFBS enrichment analysis also reveals that transcription factor *POU2F1* binds to the promoter region of *ESR1* that encodes estrogen receptor α. Transcription factor *E2F1* binds to the promoter regions of coexpressed genes *MCM7, ANAPC1 and WEE1*.

Conclusions: Thus our integrative approach provides insights into breast cancer prognosis.

Keywords: Time series gene expression data, Tricluster, Mean-squared residue, Eigengene, Affirmation score, Gene ontology, KEGG pathway, TRANSFAC

*Correspondence: anirban.bhar@bioinf.med.uni-goettingen.de; umaulik@cse.jdvu.ac.in; sanghami@isical.ac.in; edgar.wingender@bioinf.med.uni-goettingen.de
[1]Institute of Bioinformatics, University Medical Center Goettingen, University of Goettingen, Goldschmidtstrasse 1, D-37077 Goettingen, Germany
[3]Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
[4]Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700108, India
Full list of author information is available at the end of the article

A.4 Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes

Bhar *et al. BMC Bioinformatics* (2015) 16:200
DOI 10.1186/s12859-015-0635-8

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes

Anirban Bhar[1], Martin Haubrock[1], Anirban Mukhopadhyay[2] and Edgar Wingender[1*]

Abstract

Background: Exploratory analysis of multi-dimensional high-throughput datasets, such as microarray gene expression time series, may be instrumental in understanding the genetic programs underlying numerous biological processes. In such datasets, variations in the gene expression profiles are usually observed across replicates and time points. Thus mining the temporal expression patterns in such multi-dimensional datasets may not only provide insights into the key biological processes governing organs to grow and develop but also facilitate the understanding of the underlying complex gene regulatory circuits.

Results: In this work we have developed an evolutionary multi-objective optimization for our previously introduced triclustering algorithm δ-TRIMAX. Its aim is to make optimal use of δ-TRIMAX in extracting groups of co-expressed genes from time series gene expression data, or from any 3D gene expression dataset, by adding the powerful capabilities of an evolutionary algorithm to retrieve overlapping triclusters. We have compared the performance of our newly developed algorithm, EMOA-δ-TRIMAX, with that of other existing triclustering approaches using four artificial dataset and three real-life datasets. Moreover, we have analyzed the results of our algorithm on one of these real-life datasets monitoring the differentiation of human induced pluripotent stem cells (hiPSC) into mature cardiomyocytes. For each group of co-expressed genes belonging to one tricluster, we identified key genes by computing their membership values within the tricluster. It turned out that to a very high percentage, these key genes were significantly enriched in Gene Ontology categories or KEGG pathways that fitted very well to the biological context of cardiomyocytes differentiation.

Conclusions: EMOA-δ-TRIMAX has proven instrumental in identifying groups of genes in transcriptomic data sets that represent the functional categories constituting the biological process under study. The executable file can be found at http://www.bioinf.med.uni-goettingen.de/fileadmin/download/EMOA-delta-TRIMAX.tar.gz.

Keywords: Microarray gene expression data, Developmental biology, Tricluster, Multi-objective optimization, Eigen gene, Affirmation score, TRANSFAC

Background

One of the main aims of functional genomics is to understand the dynamic features encoded in the genome such as the regulation of gene activities. It often refers to high-throughput approaches devised to gain a complete picture about all genes of an organism in one experiment. Several steps, such as transcription, RNA splicing and translation are involved in the process of gene expression, which is subject to a great many of regulatory mechanisms. Analysis of such gene expression data provides enormous leverages to understand the principles of cellular systems, diseases mechanisms, molecular networks etc. Genes having similar expression profiles are frequently found to be regulated by similar mechanisms. Previous studies elucidated the impact of highly connected intra-modular hub genes on such regulations [1–3]. Detecting hub genes and analyzing their roles may facilitate understanding the

*Correspondence: edgar.wingender@bioinf.med.uni-goettingen.de
[1] Institute of Bioinformatics, University Medical Center, Georg August University, Goettingen, Goldschmidtstrasse 1, D-37077 Goettingen, Germany
Full list of author information is available at the end of the article

A.5 Impact of mineralocorticoid receptor polymorphisms on urinary electrolyte excretion with and without diuretic drugs

Research Article
For reprint orders, please contact: reprints@futuremedicine.com

Pharmacogenomics

Impact of mineralocorticoid receptor polymorphisms on urinary electrolyte excretion with and without diuretic drugs

Nawar Dalila*[,1], Jürgen Brockmöller[1], Mladen Vassilev Tzvetkov[1], Markus Schirmer[1], Martin Haubrock[2] & Stefan Viktor Vormfelde[1]

[1]Institute of Clinical Pharmacology, University Medical Center Göttingen, Göttingen, Germany
[2]Institute of Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
*Author for correspondence:
Tel.: +49 551 39 5796
Fax: +49 551 39 12767
nawar82@gmail.com

Aim: Polymorphisms in the mineralocorticoid receptor may affect urinary sodium and potassium excretion. We investigated polymorphisms in the MR gene in relation to urinary electrolyte excretion in two separate studies. **Patients & methods:** The genotype–phenotype association was studied in healthy volunteers after single doses of bumetanide, furosemide, torsemide, hydrochlorothiazide, triamterene and after NaCl restriction. **Results:** High potassium excretion under all conditions except torsemide, and high NaCl excretion after bumetanide and furosemide were associated with the A allele of the intron-3 polymorphism (rs3857080). This polymorphism explained 5–10% of the functional variation and *in vitro*, rs3857080 affected DNA binding of the transcription factor LHX4. **Conclusion:** rs3857080 may be a promising new candidate for research in cardiac and renal disorders and on antialdosteronergic drugs like spironolactone.

Original submitted 23 June 2014; Revision submitted 5 November 2014

Keywords: antihypertensive agents • aldosterone • hydrochlorothiazide • LHX4 • loop diuretics • mineralocorticoid receptor • *NR3C2* • pharmacogenetics • polymorphisms • triamterene

Diuretic drugs control salt, water excretion and blood pressure and are one of most important drugs in hypertension and heart failure [1–4]. Rare genetic variants and frequent polymorphisms in the sodium–potassium–dichloride cotransporter (NKCC2), the sodium–chloride cotransporter (NCC) and the epithelial sodium channel (ENaC) have been reported to affect electrolyte excretion when diuretics were applied [5]. The mineralocorticoid receptor (MR, aldosterone receptor) plays an important role in sodium reabsorption and potassium excretion and is partially mediated by transcriptional regulation of ENaC. Although less investigated, modulation of the renal outer medullary potassium channel (ROMK) by MR has also been suggested [6]. Twelve polymorphisms in the MR gene, *NR3C2*, have been implicated with *in vivo* phenotypes (Table 1). The Ile180Val polymorphism (rs5522) has been especially implicated with neuropsychiatric phenotypes [7–9], but diuretic drug effects have been poorly investigated in relation to these 12 polymorphisms given in Table 1.

Here we investigated, whether *NR3C2* polymorphisms are implicated in urinary electrolyte excretion in two single-dose crossover studies: One with 96 volunteers on bumetanide, furosemide and torsemide, and one in 107 volunteers, on hydrochlorothiazide, triamterene and moderate sodium chloride restriction.

Patients & methods

We investigated the urinary electrolyte excretion in two single-dose crossover studies in healthy, male Caucasian volunteers [5,22]. The first study was on the loop diuretics bumetanide, furosemide and torsemide. The second study was on two doses of hydrochlorothiazide, on triamterene and on moderate sodium chloride restriction. The two

Future Medicine part of fsg

A.6 Relevance of Sp Binding Site Polymorphism in WWOX for Treatment Outcome in Pancreatic Cancer

OXFORD

JNCI J Natl Cancer Inst (2016) 108(5): djv387

doi:10.1093/jnci/djv387
First published online February 8, 2016
Article

ARTICLE

Relevance of Sp Binding Site Polymorphism in *WWOX* for Treatment Outcome in Pancreatic Cancer

Markus A. Schirmer*, Claudia M. Lüske*, Sebastian Roppel*, Alexander Schaudinn, Christian Zimmer, Ruben Pflüger, Martin Haubrock, Jacobe Rapp, Cenap Güngör, Maximilian Bockhorn, Thilo Hackert, Thomas Hank, Oliver Strobel, Jens Werner, Jakob R. Izbicki, Steven A. Johnsen, Jochen Gaedcke, Jürgen Brockmöller†, B. Michael Ghadimi†

Affiliations of authors: Institute of Clinical Pharmacology (MAS, CML, SR, AS, CZ, RP, JB), Institute of Bioinformatics (MH), Clinic of General and Visceral Surgery (CML, SR, JR, SAJ, JG, BMG), and Clinic of Radiotherapy and Radiation Oncology (MAS), University Medical Center Göttingen, Göttingen, Germany; Department of General, Visceral, and Thoracic Surgery, University Hospital Hamburg-Eppendorf, Hamburg, Germany (CG, MB, JRI); Department of General, Visceral, and Transplantation Surgery, University of Heidelberg, Heidelberg, Germany (THac, THan, OS, JW).

*Authors contributed equally to this work.
†Authors contributed equally to this work.

Correspondence to: Markus A. Schirmer, MD, MSc, Institute of Clinical Pharmacology and Clinic of Radiotherapy and Radiation Oncology, University Medical Center Göttingen, Robert-Koch-Strasse 40, 37075 Göttingen, Germany (e-mail: mschirmer@med.uni-goettingen.de).

Abstract

Background: A genome-wide association study (GWAS) suggested inherited genetic single-nucleotide polymorphisms (SNPs) affecting overall survival (OS) in advanced pancreatic cancer. To identify robust clinical biomarkers, we tested the strongest reported candidate loci in an independent patient cohort, assessed cellular drug sensitivity, and evaluated molecular effects.

Methods: This study comprised 381 patients with histologically verified pancreatic ductal adenocarcinoma treated with gemcitabine-based chemotherapy. The primary outcome was the relationship between germline polymorphisms and OS. Functional assays addressed pharmacological dose-response effects in lymphoblastoid cell lines (LCLs) and pancreatic cancer cell lines (including upon RNAi), gene expression analyses, and allele-specific transcription factor binding. All statistical tests were two-sided.

Results: The *A* allele (26% in Caucasians) at SNP rs11644322 in the putative tumor suppressor gene *WWOX* conferred worse prognosis. Median OS was 14 months (95% confidence interval [CI] = 12 to 15 months), 13 months (95% CI = 11 to 15 months), and nine months (95% CI = 7 to 12 months) for the *GG*, *GA*, and *AA* genotypes, respectively (P_{trend} < .001 for trend in univariate log-rank assuming a codominant mode of inheritance; advanced disease subgroup P_{trend} < .001). Mean OS was 25 months (95% CI = 21 to 29 months), 19 months (95% CI = 15 to 22 months), and 13 months (95% CI = 10 to 16 months), respectively. This effect held true after adjustment for age, performance status according to Eastern Cooperative Oncology Group classification, TNM, grading, and resection status and was comparable with the strongest established prognostic factors in multivariable analysis. Consistently, reduced responsiveness to gemcitabine, but not 5-fluorouracil, along with lower *WWOX* expression was demonstrated in LCLs harboring the *AA* genotype. Likewise, RNAi-mediated *WWOX* knockdown in pancreatic cancer cells confirmed differential cytostatic drug sensitivity. In electrophoretic mobility shift assays, the *A* allele exhibited weaker binding of Sp family members Sp1/Sp3.

Received: October 3, 2014; **Revised:** July 29, 2015; **Accepted:** November 16, 2015

A.7 NF-Y Binding Site Architecture Defines a C-Fos Targeted Promoter Class

PLOS ONE

RESEARCH ARTICLE

NF-Y Binding Site Architecture Defines a C-Fos Targeted Promoter Class

Martin Haubrock*, Fabian Hartmann, Edgar Wingender

Institute of Bioinformatics, University Medical Center Göttingen (UMG), Georg-August-University Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, Germany

* mhaubro@uni-goettingen.de

Abstract

ChIP-seq experiments detect the chromatin occupancy of known transcription factors in a genome-wide fashion. The comparisons of several species-specific ChIP-seq libraries done for different transcription factors have revealed a complex combinatorial and context-specific co-localization behavior for the identified binding regions. In this study we have investigated human derived ChIP-seq data to identify common cis-regulatory principles for the human transcription factor c-Fos. We found that in four different cell lines, c-Fos targeted proximal and distal genomic intervals show prevalences for either AP-1 motifs or CCAAT boxes as known binding motifs for the transcription factor NF-Y, and thereby act in a mutually exclusive manner. For proximal regions of co-localized c-Fos and NF-YB binding, we gathered evidence that a characteristic configuration of repeating CCAAT motifs may be responsible for attracting c-Fos, probably provided by a nearby AP-1 bound enhancer. Our results suggest a novel regulatory function of NF-Y in gene-proximal regions. Specific CCAAT dimer repeats bound by the transcription factor NF-Y define this novel cis-regulatory module. Based on this behavior we propose a new enhancer promoter interaction model based on AP-1 motif defined enhancers which interact with CCAAT-box characterized promoter regions.

Citation: Haubrock M, Hartmann F, Wingender E (2016) NF-Y Binding Site Architecture Defines a C-Fos Targeted Promoter Class. PLoS ONE 11(8): e0160803. doi:10.1371/journal.pone.0160803

Editor: Roberto Mantovani, Università degli Studi di Milano, ITALY

Received: April 26, 2016

Accepted: July 25, 2016

Published: August 12, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Transcription factors (TFs) are proteins that control gene expression through a variety of mechanisms such as enhancing the efficiency of the basal transcription complex to assemble or re-model chromatin. Most of them act by recognizing cis-regulatory elements or TF binding sites (TFBS) in gene proximal (promoter) or distal (enhancer) regions in a sequence-specific way. Each regulatory region is defined by an array of such TFBSs. The cooperative binding of multiple TFs to these closely located TFBSs determines the transcription activity of their target genes [1, 2]. Potential TFBSs can be found everywhere in the genome, but only a minority of them appears to be functional in a given cellular context. Moreover, the activity of a proximal or a distal region in a certain cellular context is associated with its epigenetic status such as DNA- or histone modification, and can be monitored by its sensitivity against DNase I attack [3, 4]. However, it is unclear whether the binding of a TF is the prerequisite or the consequence

A.8 TFClass: a classification of human transcription factors and their rodent orthologs

Published online 31 October 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D97–D102***
doi: 10.1093/nar/gku1064

TFClass: a classification of human transcription factors and their rodent orthologs

Edgar Wingender[1,2,*], Torsten Schoeps[1], Martin Haubrock[1] and Jürgen Dönitz[3]

[1]Institute of Bioinformatics, University Medical Center Göttingen, Georg August University, D-37077 Göttingen, Germany, [2]geneXplain GmbH, D-38302 Wolfenbüttel, Germany and [3]Johann-Friedrich-Blumenbach Institute of Zoology and Anthropology, Georg August University, D-37077 Göttingen, Germany

Received September 28, 2014; Accepted October 15, 2014

ABSTRACT

TFClass aims at classifying eukaryotic transcription factors (TFs) according to their DNA-binding domains (DBDs). For this, a classification schema comprising four generic levels (superclass, class, family and subfamily) was defined that could accommodate all known DNA-binding human TFs. They were assigned to their (sub-)families as instances at two different levels, the corresponding TF genes and individual gene products (protein isoforms). In the present version, all mouse and rat orthologs have been linked to the human TFs, and the mouse orthologs have been arranged in an independent ontology. Many TFs were assigned with typical DNA-binding patterns and positional weight matrices derived from high-throughput in-vitro binding studies. Predicted TF binding sites from human gene upstream sequences are now also attached to each human TF whenever a PWM was available for this factor or one of his paralogs. TFClass is freely available at http://tfclass.bioinf.med.uni-goettingen.de/ through a web interface and for download in OBO format.

INTRODUCTION

DNA-binding transcription factors (TFs) regulate transcription by binding to genomic sites in regions of regulatory impact. In a complex interaction with enzymes that modify chromatin structure, mostly by methylation and acetylation events, they support the formation of the transcription preinitiation complex and direct the RNA polymerase to the transcription start site (TSS). The key function of TFs is to read out regulatory sequence signals in the genome and help transmitting them into the process of gene activation.

TFs recognize short specific sequence elements in a relaxed manner, which is frequently represented by positional weight matrices (PWMs). The way how the DNA-binding domains (DBDs) of TFs interact with their target sequences depends highly on the specific structural features of these domains. Different DBDs seem to have developed their own DNA–protein recognition code, which renders a systematic classification of DBDs a necessary prerequisite for any systematic characterization and prediction of protein–DNA interactions.

Exceeding the scope of previous catalogs of TFs and DBDs (1–6), we have introduced TFClass as a classification of human TFs based on their DBDs (7), which was a new version of a much older scheme that became part of the TRANSFAC® database (8,9). With the recent updates to be reported here, we have introduced a number of smaller revisions in the structure, added mouse and rat orthologs of the human TFs in the classification, and present an independent ontology of mouse TFs, so far confined to the orthologs of the human TFs. Moreover, the information about TFs targets was enhanced by linking PWMs from a systematic in vitro screen (10), and lists of target sites and genes predicted with the TRANSFAC® matrix library (11,12).

DATA SOURCES

Domain assignments, protein sequences and information about isoforms were taken from UniProt, last update done using release 2014_07 (13), and TRANSFAC® (BIOBASE, Germany), with the last update using release 2014.2 (11). 3D structures were obtained from the PDB database (14), generally used as entry point to retrieve the original publications. The linked PWMs were taken from Jolma *et al.* (10). Domain annotations from UniProt were manually validated and edited where necessary. Sequence comparisons were basically done as described previously (7). In most cases, we used the BLAST option provided by the Expasy server (15) and the Clustal Omega tool implemented on the same server to check the similarities of presumed orthologs (16). As reported previously, protein expression signatures were composed using data from Protein Atlas, with the original data sets linked (17), and associated with the genus entries. Newly added are links to the corresponding BioGPS entries (18).

*To whom correspondence should be addressed. Tel: +49 551 3914911; Fax: +49 551 3914914; Email: edgar.wingender@bioinf.med.uni-goettingen.de

A.9 TFClass: expanding the classification of human transcription factors to their mammalian orthologs

Published online 26 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D343–D347***
doi: 10.1093/nar/gkx987

TFClass: expanding the classification of human transcription factors to their mammalian orthologs

Edgar Wingender[1,2,*], Torsten Schoeps[1], Martin Haubrock[1], Mathias Krull[2] and Jürgen Dönitz[1,3]

[1]Institute of Bioinformatics, University Medical Center Göttingen, Georg August University, D-37077 Göttingen, Germany, [2]geneXplain GmbH, D-38302 Wolfenbüttel, Germany and [3]Dpt. of Evolutionary Developmental Genetics, Johann-Friedrich-Blumenbach Institute of Zoology and Anthropology, Georg August University, D-37077 Göttingen, Germany

Received September 15, 2017; Revised October 09, 2017; Editorial Decision October 10, 2017; Accepted October 12, 2017

ABSTRACT

TFClass is a resource that classifies eukaryotic transcription factors (TFs) according to their DNA-binding domains (DBDs), available online at http://tfclass.bioinf.med.uni-goettingen.de. The classification scheme of TFClass was originally derived for human TFs and is expanded here to the whole taxonomic class of mammalia. Combining information from different resources, checking manually the retrieved mammalian TFs sequences and applying extensive phylogenetic analyses, >39 000 TFs from up to 41 mammalian species were assigned to the Superclasses, Classes, Families and Subfamilies of TFClass. As a result, TFClass now provides the corresponding sequence collection in FASTA format, sequence logos and phylogenetic trees at different classification levels, predicted TF binding sites for human, mouse, dog and cow genomes as well as links to several external databases. In particular, all those TFs that are also documented in the TRANSFAC® database (FACTOR table) have been linked and can be freely accessed. TRANSFAC® FACTOR can also be queried through an own search interface.

INTRODUCTION

Transcription factors (TFs) are proteins that regulate transcription, e.g. by directing RNA polymerase to the transcription start site of a gene. Most TFs do so by recognizing regulatory elements in promoters and enhancers in a sequence-specific way through their DNA-binding domains (DBDs). These DBDs are organized by few structural principles, which can be used to classify DNA-binding TFs as we have done with TFClass. Previously, we have described the underlying classification scheme and its application to human TFs (1) and to their rodent orthologs (2). Criteria for classifying TFs have been discussed elsewhere (3). In this report, we present the extension of TFClass to all mammals as far as annotated genomic information is available. The TFs assigned in TFClass are linked to entries in the FACTOR table of the TRANSFAC® database, the oldest actively maintained resource for TFs and their DNA-binding sites and properties (4).

METHODS

Data sources

Using the catalog of human TFs from previous versions of TFClass as starting point, we retrieved the corresponding ortholog clusters from OrthoDB. v8 (5). The collected entries from up to 41 mammalian species were semi-manually pruned for any entries that were highly truncated or contained stretches of >5 undefined positions or any undefined positions within their DNA-binding domains (DBDs). If necessary, further paralog assignments were done manually after several iterative phylogenetic analyses.

Domain annotation

As reported for the previous releases, domain assignments, protein sequences, and information about isoforms are taken from UniProt, last update done using release July 2014 (6), and from TRANSFAC®, with the last update using release 2017.2 (4). By searching for the orthologous domain boundaries in the alignments and subsequent extensive manual editing, the DBD sequences were compiled as FASTA files. For the visualization of the isoforms and for marking the DBD the entries are retrieved dynamically from UniProt as RDF file (Resource Description Framework). The downloaded file is cached and invalidated after 3 months to reflect updates in the database of UniProt in TFClass. For the marking of DBD the sequence of the compiled DBD FASTA file is mapped to the isoform sequences

*To whom correspondence should be addressed. Tel: +49 551 3914911; Fax: +49 551 3914914; Email: edgar.wingender@bioinf.med.uni-goettingen.de

A.10 Constructing temporal regulatory cascades in the context of development and cell differentiation

PLOS ONE

RESEARCH ARTICLE

Constructing temporal regulatory cascades in the context of development and cell differentiation

Rayan Daou[1], **Tim Beißbarth**[1], **Edgar Wingender**[1], **Mehmet Gültas**[2,3], **Martin Haubrock**[1]*

1 Department of Medical Bioinformatics, University Medical Center Göttingen, Goettingen, Niedersachsen, Germany, **2** Breeding Informatics Group, Department of Animal Science, Georg-August University, Goettingen, Niedersachsen, Germany, **3** Center for Integrated Breeding Research (CiBreed), Georg-August University, Goettingen, Niedersachsen, Germany

* martin.haubrock@bioinf.med.uni-goettingen.de

Citation: Daou R, Beißbarth T, Wingender E, Gültas M, Haubrock M (2020) Constructing temporal regulatory cascades in the context of development and cell differentiation. PLoS ONE 15(4): e0231326. https://doi.org/10.1371/journal.pone.0231326

Editor: Roberto Mantovani, Università degli Studi di Milano, ITALY

Received: October 9, 2019

Accepted: March 20, 2020

Published: April 10, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0231326

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Abstract

Cell differentiation is a complex process orchestrated by sets of regulators precisely appearing at certain time points, resulting in regulatory cascades that affect the expression of broader sets of genes, ending up in the formation of different tissues and organ parts. The identification of stage-specific master regulators and the mechanism by which they activate each other is a key to understanding and controlling differentiation, particularly in the fields of tissue regeneration and organoid engineering. Here we present a workflow that combines a comprehensive general regulatory network based on binding site predictions with user-provided temporal gene expression data, to generate a a temporally connected series of stage-specific regulatory networks, which we call a temporal regulatory cascade (TRC). A TRC identifies those regulators that are unique for each time point, resulting in a cascade that shows the emergence of these regulators and regulatory interactions across time. The model was implemented in the form of a user-friendly, visual web-tool, that requires no expert knowledge in programming or statistics, making it directly usable for life scientists. In addition to generating TRCs the tool links multiple interactive visual workflows, in which a user can track and investigate further different regulators, target genes, and interactions, directing the tool along the way into biologically sensible results based on the given dataset. We applied the TRC model on two different expression datasets, one based on experiments conducted on human induced pluripotent stem cells (hiPSCs) undergoing differentiation into mature cardiomyocytes and the other based on the differentiation of H1-derived human neuronal precursor cells. The model was successful in identifying previously known and new potential key regulators, in addition to the particular time points with which these regulators are associated, in cardiac and neural development.