Doctoral theses at NTNU, 2016:243

Shahram Bahrami

Investigation of properties and classification of human transcription factors

ISBN 978-82-326-1824-8 (printed ver.)
ISBN 978-82-326-1825-5 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

NTNU

Norges teknisk-naturvitenskapelige universitet
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine
Department of Cancer Research
and Molecular Medicine

Doctoral theses at NTNU, 2016:243

NTNU
Norwegian University of
Science and Technology

Shahram Bahrami

Investigation of properties and classification of human transcription factors

Thesis for the Degree of Philosophiae Doctor

Trondheim, September 2016

Norwegian University of Science and Technology
Faculty of Medicine
Department of Cancer Research and
Molecular Medicine

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine
Department of Cancer Research
and Molecular Medicine

ISBN 978-82-326-1824-8 (printed ver.)
ISBN 978-82-326-1825-5 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:243

Printed by NTNU Grafisk senter

NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET
DET MEDISINSKE FAKULTET

Shahram Bahrami

Studier av egenskaper og klassifisering av humane transkripsjonsfaktorer

Egenskaper og funksjonen av ulike celler bestemmes gjennom regulering av genuttrykket, og en rekke ulike mekanismer er aktive i den enkelte celle for å øke eller redusere genuttrykket. Blant annet transkripsjonsfaktorer er viktige for genregulering. Hensikten med dette arbeidet har vært å kartlegge spesifikke egenskaper ved transkripsjonsfaktorer for bedre å forstå hvordan de samspiller i genreguleringen, og hvordan deres egenskaper og endringer i disse egenskapene kan påvirke dette samspillet. Dette ble først gjort ved å lage en database med kjente transkripsjonsfaktorer og deres egenskaper, og gjøre statistiske analyser av dette. Den studien viste at et slikt datasett er en nyttig ressurs for å analysere andre datasett på genregulering, men at det også kan være en bias i slike analyser på grunn av ufullstendige data. Denne studien ble så utvidet ved å klassifisere transkripsjonsfaktorer i undergrupper basert på hvordan de åpner opp og binder til kromatin under genregulering, basert på eksperimentelle data. Det klassifiserte datasettet ble brukt til å analysere andre eksperimentelle data, og dette viste blant annet klare forskjeller i hvordan de ulike undergruppene av transkripsjonsfaktorer binder til DNA. Det ble så fokusert på en undergruppe av gener som reagerer veldig raskt på stimulering, for å få en bedre forståelse av hvordan slike gener blir regulert. Dette ble gjort delvis gjennom en litteratur-studie, og delvis ved å gjøre statistisk analyse av et sett av slike gener og deres omgivelser i genomet. Dette studiet identifiserte blant annet flere transkripsjons-faktorer som selv hører til under denne typen av gener, men som også er involvert i reguleringen av slike gener. Dette kan i fremtiden gi grunnlag for en modellering av regulatoriske interaksjoner i gener med rask aktivering.

Kandidat: Shahram Bahrami
Institutt: Institutt for kreftforskning og molekylær medisin
Veiledere: Finn Drabløs (hovedveileder), Pål Sætrom (biveileder)
Finansiering: Samarbeidsorganet mellom Helse Midt-Norge RHF og NTNU

Ovennevnte avhandling er funnet verdig til å forsvares offentlig
for graden PhD i Medisinsk teknologi

Disputas finner sted i Auditoriet, Medisinsk teknisk forskningssenter
torsdag 15. september 2016 kl. 12:15

Abstract

During gene expression information from a gene is transcribed into RNA, which subsequently may be used directly (non-coding RNA), or translated into protein. Regulation of gene expression includes several steps, and these steps are controlled by several elements such as enhancers, activators, and transcription factors. Transcription factors are proteins that bind to specific DNA sequences and control the regulation of gene expression. Transcription factors are generally modular in structure, and often contain one or more domains. Most transcription factors are divided into two major classes; the general TFs and the site-specific TFs. The site-specific TFs bind to specific DNA motifs through their DNA binding domains. The role of transcription factors as a fundamental part of the general regulatory system depends upon specific properties of the factors, including DNA-binding domains, protein-protein interactions (PPIs) and post-translational modifications (PTMs).

Because of the importance of transcription factors in the regulation of gene expression, a transcription factor database including different properties of the transcription factors can be a very useful resource for researchers.

In this thesis, we have first created a comprehensive list of human transcription factors which includes information on Pfam domains, DNA-binding domains, protein–protein interactions and post-translational modifications. Then we have used this data set for enrichment analysis and investigated correlations within this set of features, and between the features.

As part of this work, we have also expanded the annotated set of transcription factors and classified them with respect to their role in chromatin opening as Pioneers, Settlers, positive and negative Migrants. The results showed that the classification is a useful resource for analyzing data on gene expression and for better understanding of how transcription factor expression and the dynamics of chromatin structure are integrated at a functional level.

In the final part of the thesis we have focused on the activation and regulation of immediate-early genes (IEGs), since these genes have several interesting properties with respect to regulation. Immediate early genes are genes which are expressed transiently and quickly within minutes in response to a wide variety of stimuli. These genes play important role in several essential cellular systems, such as the immune system. We have first summarized current knowledge regarding regulation and selected key properties of these genes as a review, including the importance of genetic and epigenetic structure, and the role of poised genes and the importance of in particular strong enhancers. We have then developed a consensus set including 172 immediate-early response genes showing rapid activation with different types of stimulation. We have then done bioinformatics analysis of the gene list, and identified some of the key properties of these genes.

The results showed that the consensus set has a good representation of immediate-early response genes and is largely consistent with previous results. Therefore the consensus set is a

useful resource for analyzing how genes involved in the immediate-early response are regulated.

Preface

The work presented in this thesis was carried out at the Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology in the period from August of 2011 to 2016.

First of all I would like to express my gratitude to Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU) for funding and to give my special thanks to my supervisor, Professor Finn Drabløs, for offering me the opportunity to do my PhD in Bioinformatics. I highly appreciate your enormous knowledge and enthusiasm in science and feel honored for being a member for your group. Thank you for both scientific and non-scientific discussions, magic troubleshooting and all the time you have spent answering my many questions.

In addition, I wish to appreciate my co-supervisor, Professor Pål Sætrom, for the collaboration.

I thank all my colleagues for creating a superb social and scientific atmosphere. I would like to specially mention Rezvan Ehsani, Ane Langkilde-Lauesen Nielsen, Lene Christin Olsen, Pouda Panahandeh, Kjetil Klepper, Morten Beck Rye, Jostein Johansen, Anne Heidi Skogholt, Anna Tarsia, Zehui Qu, and Avinash Achar.

Finally, I would like to thank my friends and family for their love and support.

Trondheim, Norway

Shahram Bahrami

September, 2016

Contents

List of figures

List of papers

This thesis is based on the following three papers, which are referred to in the text by their Roman numerals I-IV.

Paper I) Bahrami S, Ehsani R, Drabløs F: **A property-based analysis of human transcription factors.** BMC Research Notes 2015, 8:82.

Paper II) Ehsani R, Bahrami S, Drabløs F: **Functional classification of human transcription factors based on structural properties.** Submitted.

Paper III) Bahrami S and Drabløs F: **Gene regulation in the immediate-early response process.** Advances in Biological Regulation. Accepted

Paper IV) Bahrami S and Drabløs F: **Identification and Analysis of Genes in Immediate-Early Response Processes.**

Abbreviations

DNA	Deoxyribonucleic acid
TBP	TATA-binding protein
mRNA	messenger RNA
TAF	TBP-associated factor
TF	Transcription Factor
PIC	pre-initiation complex
DBD	DNA-binding domain
HMM	hidden Markov model
ADDA	Automatic Domain Decomposition Algorithm
PTM	Post-translational modification
SSTF	Site-specific, DNA-binding transcription factor
IEG	immediate-early gene
PRG	primary response gene
SRG	secondary response gene
SRF	serum-response factor
NF-kB	nuclear factor-kB
CREB	cyclic AMP response element-binding protein
DSIF	DRB sensitivity-inducing factor
NELF	negative elongation factor
P-TEFb	positive transcription elongation factor
eRNA	enhancer RNA
lncRNA	long non-coding RNA
SVM	support vector machine
PIQ	protein interaction quantitation
PPV	precision
SN	sensitivity
MCC	Matthews's correlation coefficient
TP	True positive
TN	True negative
FP	False positive
FN	False negative
ROC	Receiver operating characteristic
GO	Gene ontology

Background

An important aspect of the project described in this thesis has been to use property-based analysis of transcription factors as a basis for understanding how classes of transcription factors may have different roles in regulation of gene expression. This section will therefore provide an introduction to gene regulation and transcription factors, including relevant properties at the protein level, followed by a brief introduction to immediate-early genes, used as an example of a system where gene regulation is essential, and finally some machine learning methods and statistical approaches used in the analysis are introduced.

Regulation of gene expression

The structure of DNA

Deoxyribonucleic acid (DNA) is a fundamental macromolecule to all living organisms as it carries the genetic instructions used in the development, functioning and reproduction. DNA is a double-stranded helix including two long polynucleotide chains composed of four types of nucleotide subunits. A nucleotide is made of a nucleobase, a five-carbon sugar (deoxyribose), and one or more phosphate groups. There are four different types of nucleobases in normal DNA; adenine (A), cytosine (C), guanine (G), and thymine (T). A nucleobase linked to a sugar is called a nucleoside, and a nucleoside linked to a phosphate group is known as a nucleotide.

The nucleotides are joined together by a phosphodiester bond linking the phosphate groups at the 3´ carbon atom of one sugar to the 5´ carbon of the next sugar. The chain of repeated sugar-phosphate groups makes the backbone of the DNA strand. The direction of the nucleotides in one strand of a double helix is opposite to their direction in the other strand, so that the strands are antiparallel.

In a DNA double helix the two strands are connected by hydrogen bonds, so that one type of nucleobase on one strand bonds with a complementary nucleobase on the other strand; adenine forms two hydrogen bonds to thymine, and cytosine forms three hydrogen bonds to guanine (see Figure 1).

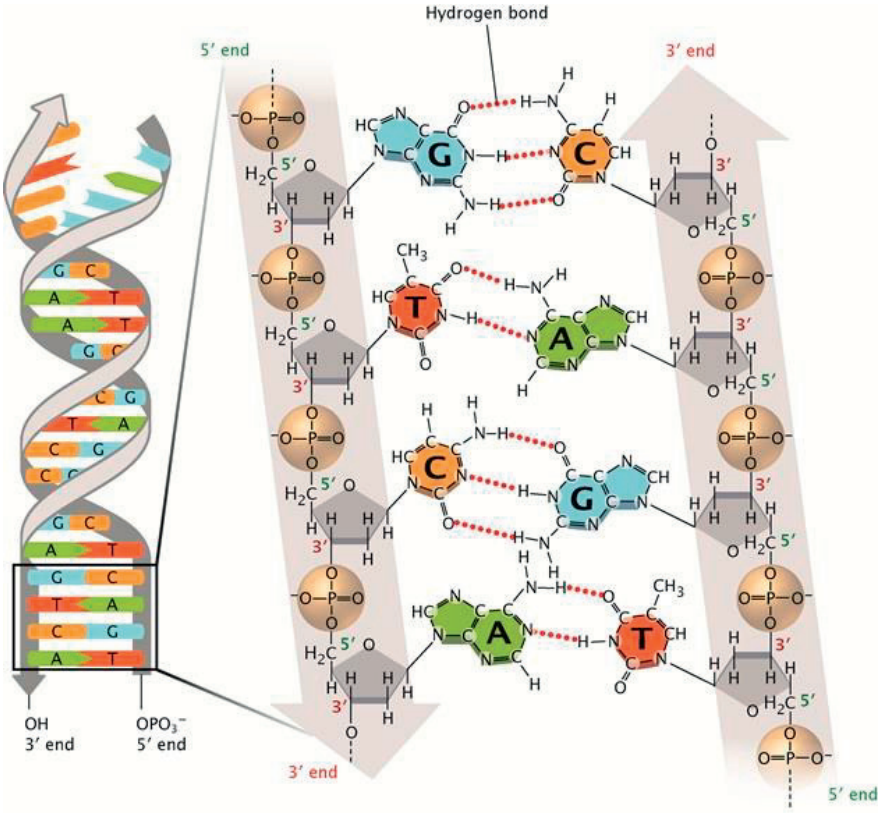


Figure 1: DNA structure, showing the four nucleobases found in DNA.

T is connected with A by two hydrogen bonds, whereas three hydrogen bonds connect G to C. The sugar-phosphate backbones run anti-parallel to each other, so that the 3' and 5' ends of the two strands are aligned and make a double helix. The figure has been adapted from http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397.

From genes to proteins

Genes encode proteins and proteins are the main mediators of cellular function. The process by which information from a gene is used to synthesis a protein is covered by gene expression and translation. The subset of genes expressed in a specific cell determines what that cell can do. Transcription is the first step of gene expression, in this step the information in a strand of DNA is copied by RNA polymerase into messenger RNA (mRNA) for protein production. The mRNAs are the molecules that convey genetic information from DNA out of the cell nucleus to the ribosome, where they determine the amino acid sequence of the protein products. In mRNA the genetic information is arranged into codons consisting of three bases each. But unlike DNA it is a single-stranded molecule, and it is much shorter than genomic DNA. Each nucleotide in mRNA contains a ribose sugar, and the complementary base to adenine in mRNA is not thymine, as in DNA, but rather uracil.

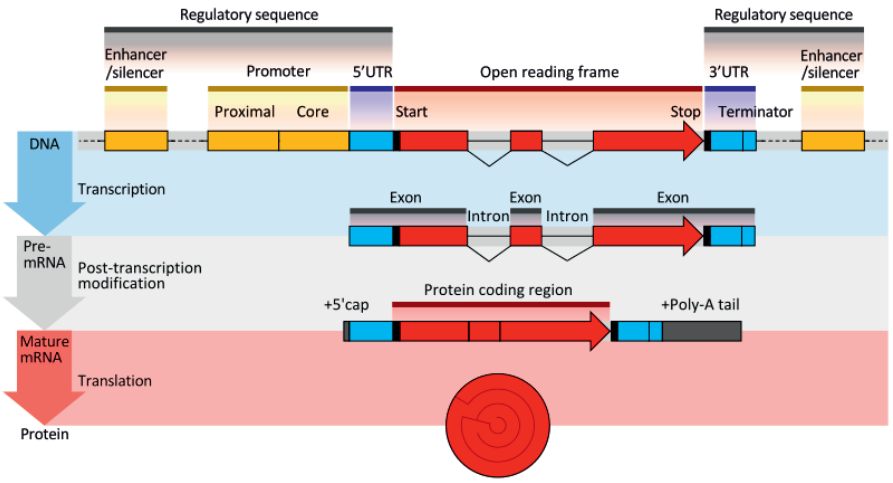


Figure 2: Gene expression, the structure of a eukaryotic protein-coding gene.

Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified by adding a 5' cap and poly-A tail (grey) and removing introns. The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product. The figure has been adapted from Wikipedia.

The primary transcript for protein production is a single-stranded mRNA called pre-mRNA. The pre-mRNA molecule undergoes three main modifications, including RNA splicing, 5' capping, and 3' polyadenylation. Eukaryotic genes contain segments that do not code for proteins, known as introns, and the remaining segments that code for proteins are known as exons. During RNA splicing the introns are removed and the exons are joined. Also a modified guanine nucleotide is added to the 5' end of the pre-mRNA. The 5' cap includes a terminal 7-methylguanosine residue that is connected through a 5'-5'-triphosphate bond to the first transcribed nucleotide. In addition, at the other end of the pre-mRNA strand, a chain of adenosine monophosphates is added, known as a poly-A tail (3' polyadenylation). These modifications protect the mRNA molecule from enzymatic degradation and are important for regulating the export of the mRNA from the nucleus to cytoplasm for translation.

When a pre-mRNA molecule has been correctly processed to an mRNA sequence, it is exported out of the nucleus to be translated into a protein by ribosomes. During translation the resulting mRNA, which is a single-stranded copy of the gene, is translated into a protein molecule.

During translation the information contained in the mRNA is read as three letter words (triplets), called codons. Each codon specifies a particular amino acid (hence, it is a triplet code). During the translation step the amino acids are connected together and form a polypeptide chain which will later be folded into a protein.

The structure of proteins

Proteins have different functions. Some of them have a structural function such as keratin; others are involved in cell signaling, such as hormones and their receptors, and other groups of proteins act as enzymes and catalyze chemical reactions.

Proteins are polymers of amino acids covalently connected via peptide bonds into a chain. There are 20 different amino acids with a common basic structure. A central carbon is bonded to a hydrogen atom, a carboxyl group, an amino group, and a unique side chain or R-group. The chemical properties of amino acids are mainly determined by their unique side chain. A chain containing approximately 50 or fewer amino acids is often called a peptide.

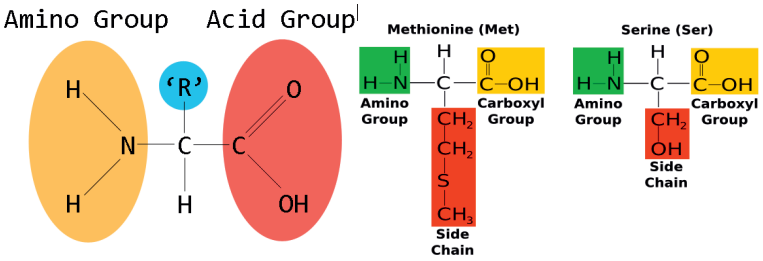


Figure 3: Building blocks for proteins

Left: Proteins are made of amino acids and each amino acid consists of a central carbon, a hydrogen atom, a carboxyl group, an amino group, and a unique side chain or R-group. **Right**: The figure shows two different amino acids. Figure adapted from http://alevelnotes.com/Amino-Acids/59?tree= and http://study.com/academy/lesson/what-are-amino-acids-definition-structure-quiz.html.

Proteins have four main levels of structure; primary, secondary, tertiary and quaternary.

Primary structure: This structure consists of the individual amino acids join together in linear chains by forming peptide bonds between the $-NH_2$ of one amino acid and the -COOH of the next.

Secondary structure: This structure refers to the folding of a polypeptide chain, and involves two main types, the alpha helix and the beta strand. These structures are determined by patterns of hydrogen bonds between the main-chain peptide groups.

Tertiary structure: This represents the overall three-dimensional structure of the polypeptide chain of a protein molecule. The protein chain will twist and bend in such a way as to get maximum stability or lowest energy state.

Quaternary structure: In this structure a protein macromolecule is made up of multiple polypeptide chains (multi-subunit protein) by non-covalent interactions between the multiple polypeptide chains to form a larger aggregated protein complex.

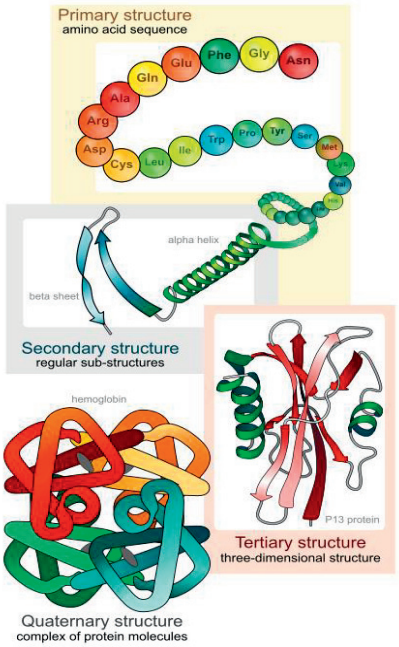


Figure 4: Protein structure.

Different levels of protein structure, showing primary, secondary, tertiary and quaternary structure. The figure has been adapted from Wikipedia.

Proteins can also be represented as consisting of domains. A domain is a part of a protein that can be found in different contexts, independently of the rest of the protein. The function of a protein is determined by its domains and the nature of their interactions. A protein may consist of one or more structural domains. Domains vary in length from about 25 amino acids up to 500 amino acids. Proteins that have the same domains tend to have common functional characteristics and common ancestor [1].

Most proteins also undergo chemical modifications to form the mature protein product in different cells and cellular processes. These modifications are known as post-translational modifications (PTMs). Most PTMs change the properties of a protein by the addition of a specific chemical group to selected amino acid side chains or at the protein's C- or N-termini. Some PTMs, such as phosphorylation may serve to rapidly and transiently activate or deactivate a protein, but other PTMs can be more long-lasting. PTMs can function in several ways; they may for example change site-specific DNA-binding transcription factors (SSTFs) with respect to subcellular localization, stability, secondary structure and DNA binding affinity, or even tertiary structure and association with co-regulatory factors [2].

Pfam: a protein families database

Pfam is a comprehensive database of protein domains that is widely used to annotate and classify proteins. Each domain is defined by an alignment used to make a profile hidden Markov model (HMM). This is an information-rich representation of the set of aligned sequences, initially made from a seed alignment, which can be used to find additional sequences that have the same domain [3]. The database includes two classes of entries: Pfam-A and Pfam-B. Pfam-A families include a seed alignment, a hidden Markov model (HMM), full alignments, associated annotation, literature references, and database links; while Pfam-B families consist of alignments of sequence clusters, derived from the Automatic Domain Decomposition Algorithm (ADDA) database, with no annotation or literature references [3]. Pfam-A families are grouped into clans. A clan contains those families that have a common evolutionary ancestor. Several lines of evidence are used to determine whether or not two families are related [3].

Regulation of gene expression

The properties and function of each cell type is mainly determined by the gene products it contains, in particular the proteins. The kind and amount of the different gene products produced by each cell is regulated. The regulation of gene expression plays a vital role in all organisms and controls the development of the organism. When the gene expression goes awry, cellular properties are changed and the changes can for example lead to development of cancer. Generally gene expression is a bridge to the genotype-phenotype relationship in all organisms. So it is essential to understand the molecular interactions that control gene expression.

Gene regulation involves many molecular events that take place when transcription of a gene occurs, and modification of transcriptional regulation is an important contribution to evolutionary changes in the genotype-phenotype relationship [4]. Transcription of eukaryotic genes are done through several events; including de-condensation of the locus, nucleosome remodeling, histone modifications, binding of transcription factors, activators and coactivators to enhancers and promoters [5]. A promoter merges information about the status of the cell and modifies the rate of transcription initiation of individual genes accordingly [4]. The promoters have two functional features; they have a core promoter, the site upon which the enzymatic machinery of transcription assembles, and they have a collection of different transcription factor binding sites that confer specificity of transcription [6]. Core promotor sequences are different between genes, but for many genes the main binding site is a TATA box, located around 25-30 bp 5' of the transcription start site. Some genes have an initiator element spanning the transcription start site while others contain additional protein binding sites for general transcription factors [7]. The first step in transcriptional initiation of genes with TATA-box is the attachment of TATA-binding protein (TBP) to DNA [8].

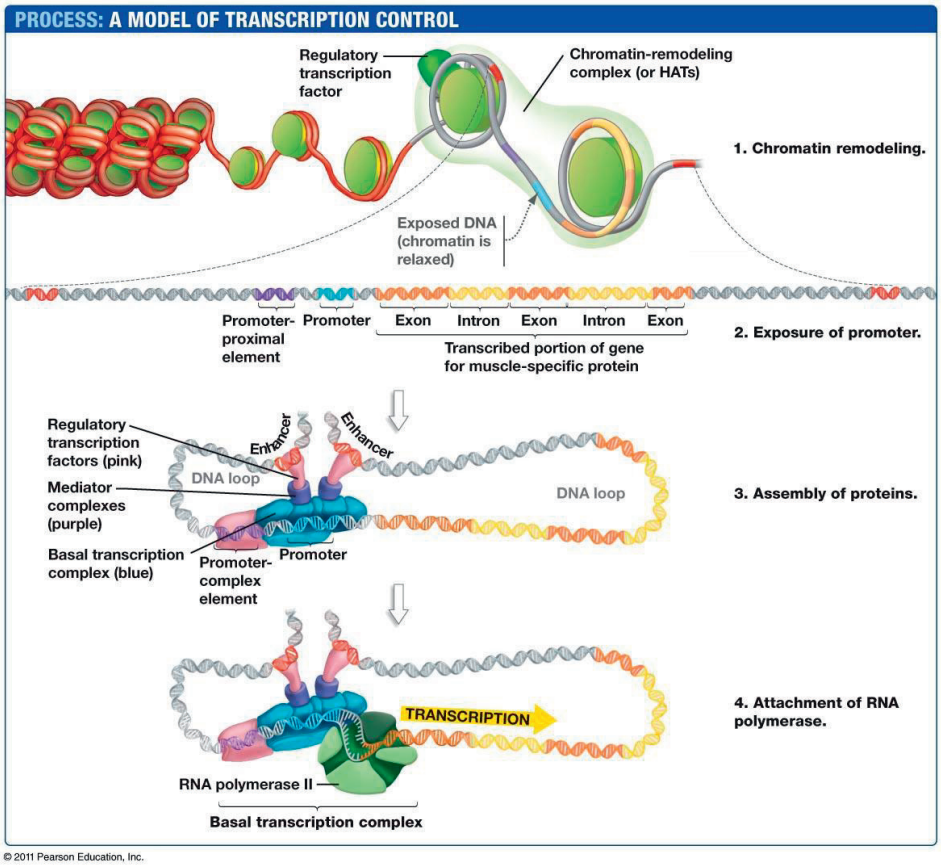


Figure 5: A model of transcription

Figure from http://www.uic.edu/classes/bios/bios100/lectures/genetic_control.htm, used with permission.

In promoters without TATA boxes, there are proteins that associate with other core promoter motifs and simplify association of TBP with DNA. Then several TBP-associated factors (TAFs) guide the RNA polymerase II onto the DNA. This step can be regulated by transcription factors (TFs) bound at other sites and is one of the most important steps of transcriptional regulation [6].

The transcription initiation by eukaryotic RNA polymerase II involves some specific transcription factors. RNA polymerase II itself is regulated and acts within a macromolecular complex, known as the pre-initiation complex (PIC), and includes TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, RNA polymerase II and Mediator [9]. Mediator is a multi-protein complex and functions as a coactivator in regulation of gene expression. Mediator is unable to bind directly to DNA sequences, but can be the main binding interface for DNA-binding transcription factors within the pre-initiation complex, and this TF-Mediator complex is necessary for target gene activation [10, 11].

Structure and properties of transcription factors

Transcription factors (TFs) are proteins that bind to specific sequence motifs of DNA, often close to their target genes, and thereby they modulate transcriptional initiation and regulate gene expression. They have a key role in transcription, and generally the regulation of gene expression involves the binding of multiple transcription factors to the regulatory regions of a given gene. Depending upon where these transcription factors bind relative to the transcription start site of the target gene, they can activate or repress transcription [12, 13]. Transcription factors are generally modular in structure, and they almost always contain one or more DNA-binding domains (DBDs), effector domains and other domain types. The DNA-binding domains are independently folded protein domains that bind to specific sequences of DNA while the effector domains interact with co-activators and other TFs to allow cooperative binding, and also directly or indirectly recruit histone and chromatin modifying enzymes [14].

DNA-binding and protein-protein interactions

Transcription factors are normally divided into two groups; general transcription factors and site-specific transcription factors. General transcription factors, also known as basal transcriptional factors, are involved in transcriptional initiation and elongation but cannot stably bind on their own to promoter and enhancer regions. They can be recruited to cis-regulatory regions via interaction with site-specific transcription factors [14], and normally they do not have any sequence-specific or site-specific DNA-binding domain.

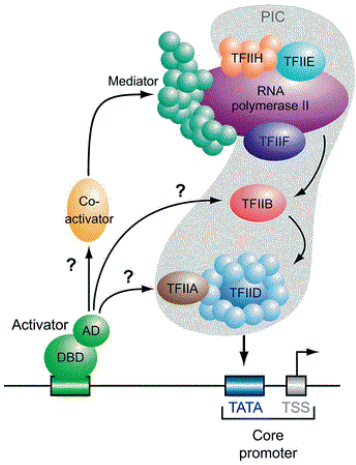


Figure 6: Transcription initiation and general transcription factors.

Figure from Maston GA *et al.* [15], used with permission.

Site-specific transcription factors have the ability to bind to specific sequences of DNA through their DNA-binding domain. These transcription factors have one or more DNA-binding domains, but they can also bind to other transcriptional regulatory proteins via their effector domains [14]. Transcription factors with one or more DNA-binding domains (DBDs) can bind to specific sequences of DNA known as DNA binding motifs. A DNA binding motif is usually 6-12 bp in length. Transcription factors are often classified based on their type of DNA binding domain. In one such classification, known as TFclass, human transcription factors are classified in a six-level hierarchical system [16].

Site-specific transcription factors include main classes such as Zinc finger proteins, Homeodomains, and Helix-Loop-Helix proteins.

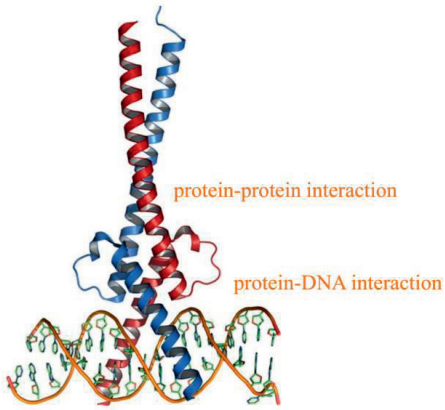


Figure 7: Site-specific transcription factors with both protein-protein interaction and protein-DNA interaction.

Figure from http://www.lookfordiagnosis.com, used with permission.

The general transcription factors can bind mainly by protein-protein interactions, whereas site-specific transcription factors can have both protein-protein and protein-DNA interactions.

Post-translational modifications

Post-translational modifications (PTMs) change the properties of a protein by the addition of a modifying chemical group to one or more of its amino acid residues. PTMs of site-specific DNA-binding transcription factors (SSTFs) are very important as they may alter the transcriptional regulatory activity of these transcription factors [2]. PTMs modulate the access of RNA polymerases to promoter templates and influence the function of transcription factors. In many cases the PTMs are individual events, but in other cases individual PTMs are consecutively linked and may cause and/or inhibit the formation a second-site PTM within the same protein [2].

The most studied PTMs of transcription factors include phosphorylation, sumoylation, ubiquitination, acetylation, glycosylation, and methylation. Filtz *et al.* have shown that most PTMs happen on transcription factors with the same rate as for other proteins. However, ubiquitination, glycosylation, and sumoylation are found on transcription factors with moderately decreased, moderately increased and greatly increased frequencies, respectively [2].

Phosphorylation

Phosphorylation is the addition of a phosphate group by a protein kinase to an amino acid residue [17, 18]. This modification is regulated by the opposing actions of protein kinases and phosphatases, and it controls key cellular processes. Phosphorylation is for example used to transduce extracellular signals to the nucleus and may affect transcription factor stability, location, structure and/or the protein interaction network [19].

Each transcription factor may have multiple phosphorylation sites that may be used in the signaling pathways. This can play an important role in changing the amplitude of gene expression [19].

Acetylation

Acetylation is a key posttranslational modification and can affect several biological features of a transcription factor. Acetylation of transcription factors may increase binding to DNA and may also influence protein-protein interactions [20, 21]. Acetylation regulates the stability of proteins and also intersects with other PTMs [20, 22]. For example, acetylation regulates the function of the Foxo1 transcription factor by altering its affinity towards the target DNA, as well as the sensitivity for phosphorylation [23].

Methylation

Methylation may change the transcriptional regulatory activity of DNA-binding transcription factors by altering the protein interaction network of these factors [24]. The methylation status of target proteins is dynamically regulated by two groups of enzymes, methyltransferases and demethylases. This modification may occur at different levels, such as mono-, di- or trimethylation on the same residue [20].

O-GlcNAcylation

One of the key posttranslational modifications is O-linked GlcNAc modification, which modulates the function of transcription factors by multiple mechanisms in a tissue-specific manner. This modification seems to be catalyzed by a single enzyme called O-linked N-acetylglucosaminyl transferase or OGT [25]. O-linked GlcNAc modifications are dynamic and reversible. In some cases O-linked GlcNAc modifications may compete with phosphorylation of the same residues [25]. O-GlcNAc modification of transcription factors can play a main role in regulation of gene expression in different tissues [26]. O-linked

GlcNAc modification can influence stability, transcriptional activity, protein–protein interaction, localization, and DNA binding ability of transcription factors [25].

Sumoylation

Sumoylation is a dynamic post-translational modification process with a small peptide on a target protein [27]. This post-translational modification affects stability, activity and localization of some specific transcription factors [28]. There are 4 confirmed SUMO isoforms in humans; SUMO-1, SUMO-2, SUMO-3 and SUMO-4. SUMO1–4 can affect the activities, nuclear or sub-nuclear localization, and/or the protein–protein interaction network of DNA-binding transcription factors [29].

Ubiquitination

Ubiquitylation is a modification that leads to the covalent modification of proteins and can serve as a protein mark with distinct signaling functions. There are several similarities between protein ubiquitination and sumoylation [30]. Different types of substrate ubiquitination are known, including the addition of a single ubiquitin molecule (monoubiquitination) or different types of ubiqutin chains (polyubiquitination) [31, 32]. Ubiquitination, like sumoylation, can change the functional properties of DNA-binding transcription factors [31, 32]. The modification can lead to change in the interaction network of the target factors by preventing basal protein-protein interactions, or by promoting interactions between ubiquitinated factors and proteins harboring ubiquitin binding domains [32].

Immediate-early genes

Some eukaryotic genes are induced and expressed by both cell-extrinsic and cell-intrinsic signals without requiring *de novo* protein synthesis, and these are called immediate-early genes (IEGs) or primary response genes (PRGs) [33, 34]. The IEGs have important roles in several biological processes, and several of the genes code for transcription factors or other DNA-binding proteins that govern the growth and differentiation of many cell types by regulating the expression of other genes. Most of these genes are expressed rapidly within minutes after the simulation, and the fact that protein synthesis is not needed indicates that the necessary transcription factors already are available [33].

Another group of genes requires protein synthesis for activation; they are more abundant and are called secondary response genes (SRGs).

A third group of genes is called delayed immediate early genes. They differ from immediate early genes in both their genomic architecture and in functions. These genes can be expressed by several of the same stimulants as the immediate early genes, but show delayed expression.

Although transcription factors are encoded by many IEGs, they are not prevalent for these delayed genes [35].

Immediate early genes are important in a wide range of biological activities, including differentiation, metabolism and proliferation, and have a recognized role as transcriptional effectors in the expression of secondary response genes [34-36]. Expression of these genes is fast and transient and their protein products are also typically unstable and rapidly targeted for proteolytic degradation by proteasome [37]. Differences between immediate early genes and delayed immediate early genes is seen in both primary transcript length and in exon frequencies; immediate early genes have shorter transcripts with fewer exons [34, 35]. There are also some specific transcription factor binding sites such as serum-response factor (SRF), nuclear factor-kB (NF-kB) and cyclic AMP response element-binding protein (CREB) binding sites that are over-represented in the upstream promoter region of this kind of genes. In contrast, binding sites for these transcription factors are not over-represented upstream of delayed primary response genes. Delayed primary response genes also have fewer clusters of transcription factor binding sites near their promoters, and the transcription factor binding sites upstream of delayed primary response genes are generally lower affinity sites compared to those upstream of immediate early genes [35].

Another difference between immediate early genes and delayed primary response genes is seen in the core promoter. Promoters of the immediate-early genes include higher affinity TATA boxes than those of the delayed primary response genes. These characteristics show that genomic features of immediate early genes are selected for rapid simulation based on their regulatory functions [35].

In addition to the differences in both upstream transcription factor binding sites and core promoters of immediate-early and delayed primary response genes, there is a difference in the binding of RNA polymerase II to the promoter regions of these genes. The amount of RNA pol II bound to the promoters of immediate-early genes is significantly greater than that bound to the delayed primary response gene promoter [35].

Regulation of immediate-early genes

Transcription elongation factors play an important role in regulation of immediate early genes (IEGs), they act at the elongation step and are necessary for development in higher eukaryotes [38]. They involve three factors; DRB sensitivity-inducing factor (DSIF), negative elongation factor (NELF) and positive transcription elongation factor (P-TEFb) [39]. The DSIF/NELF complex acts as a negative regulator complex and induces transcriptional pausing by binding to RNA polymerase II at the promoter-proximal region of IEGs. During stimulation, P-TEFb phosphorylates CTD Ser-2 of RNA polymerase II and this leads to dissociation of NELF and releases the transcriptional pausing [40]. The main function of NELF on IEG transcription seems to be to stall RNA polymerase II and block its elongation [38].

The type of stimulation has an important role on the function of NELF. It has been shown that NELF knock-down reduced TRH-induced transcription of IEGs, while it maintained or increased EGF-induced transcription of IEGs. So some stimuli, such as EGF, could increase transcription of IEGs, while others, such as TRH, seem to require NELF. Possibly NELF can affect transcription of IEGs directly via RNA polymerase II elongation on IEGs as well as indirectly via activation of the ERK1/2 MAP kinase pathway after stimulations such as by TRH [38].

Also enhancer RNAs (eRNAs) play an important role in regulation of immediate early genes (IEGs) at the elongation step. They are a class of long non-coding RNAs (lncRNA) expressed from active enhancers, and they seem to influence RNA Polymerase II pausing and release in the IEGs [41].

Machine learning and statistics

A subfield of computer science is machine learning, which explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms function by building a model from inputs to make data-driven predictions or decision [42].

Machine learning tasks are typically classified into categories, such as *supervised learning* and *unsupervised learning*. Supervised learning consists of inferring a function from labeled training data. The training data involve a set of training examples and each example consists of an input object and a desired output. The aim of supervised learning is to produce an inferred function by analyzing the training data such that the function eventually can be used for mapping new examples [42, 43], while in unsupervised learning the data are unlabeled. Machine learning can group data points into groups according to the basis of a similarity measure, or it can be used to facilitate data mining [42, 43]. Machine learning includes many different methods. In this thesis we used in particular two classifiers, Support Vector Machine and Random Forest.

The support vector machine

The support-vector machine is a supervised machine learning algorithm that is used for classification and regression [44]. In classification the support vector machine algorithm classifies input using basically a geometric idea where it expresses the data as elements of some vector space, and then constructs a hyperplane that appropriately separates the data into its two classes. The SVM algorithm functions by finding the hyperplane that gives the largest minimum distance to the training examples [44].

In addition to performing linear classification, SVMs can also do a non-linear classification by applying “the kernel trick”, implicitly mapping their inputs into high-dimensional feature

spaces where data becomes linearly separable. Then the SVM can find the optimal hyperplane that separates the classes [44].

The random forest classifier

Random forest is an ensemble learning method for classification and regression that execute by constructing many decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) over the individual trees. In the classification cases, the ensemble of simple trees votes for the most popular class. For the regression cases, it is based on an average on their responses to obtain an estimate of the dependent variable. The prediction accuracy can be improved significantly by using tree ensembles [45]. Random forest can be used effectively on large data bases with thousands of input variables and gives good predictions of which variables are important in the classification. This classifier can also provide effective methods for estimating missing data and for balancing the error in unbalanced data sets. Random forest can also be used for unlabeled data, leading to unsupervised clustering.

Evaluation of classification methods

Once a model has been built based on a training data set, then the validity of the model should be evaluated in an independent testing sample, for the same reasons and using the same methods as is usually done in most cases of predictive modeling. We should look at how well different methods do on the test set and evaluate the performance of methods on that. There are several approaches to evaluate the quality of predictions of a model.

Cross-validation

Cross validation is a standard method of assessing the accuracy and validity of a statistical model. The available data set is divided into two parts, called the training set and the testing set. It is common to hold out some parts of the data for testing and use the remaining parts for training. We normally make a decision on a fixed number of folds of the data [46]. For example we divide the data into 10 equal folds, each fold is in turn used for testing and the remainder is used for training. This means that we use nine-tenths of the data for training and one-tenth for testing and repeat the procedure ten times, so that in the end every instance has been used just once for testing. This is called tenfold cross validation. Then the error rate is calculated on the test set. Since the learning procedure is done 10 times on the different splits, we get 10 error estimates that are averaged to yield an overall error estimate. It has been shown that 10 is a reasonable number of folds to get a good error estimate. However, a single tenfold cross validation may not be enough to get a reliable error estimate, and it is standard procedure to repeat also the cross validation process 10 times. This is called 10 times tenfold cross validation [46].

Tenfold cross-validation can be defined as the standard way of measuring the error rate of a learning scheme on a dataset, but another method is also used, known as bootstrap. The

bootstrap is an estimation method based on the statistical procedure of random sampling with replacement. The idea of the bootstrap is to randomly sample the dataset with replacement to form a training set. The bootstrap procedure is repeated several times, and finally the results are averaged [47].

Matthew's correlation coefficient

The Matthews correlation coefficient is a method that is used in machine learning as an evaluation of the quality of binary classifications. The MCC function is used to evaluate the performance of the predictors as a correlation coefficient between the observed and predicted two class classifications and is measured based on true and false positives and negatives. The equation for calculating MCC is written as:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

In this formula, TP and TN are the number of true positives and true negatives respectively, and FP and FN are the number of false positives and false negatives, respectively. The MCC measure gives a value between −1 and +1, where −1 corresponds to all predictions being incorrect, 0 to random predictions, and +1 to a perfect prediction.

Precision and recall

The measures precision (also called positive predictive value (PPV)) and recall (also known as sensitivity (SN)) are the basic measures used in evaluating search strategies. The precision is a measure of result relevancy, whereas recall is a measure of how many truly relevant results that are returned [48]. They are measured based on true and false positives and negatives, and precision and recall are then defined as:

$$PPV = p = \frac{TP}{(TP + FP)}$$

$$SN = r = \frac{TP}{(TP + FN)}$$

In a classification, the precision is defined as the number of true positives (TP) divided by the number of true positives plus the number of false positives (FP) which are items incorrectly labeled as belonging to the class, while recall is defined as the number of true positives divided by the sum of true positives and false negatives (FN), which are items which were not labeled as belonging to the positive class but should have been.

Receiver operating characteristic

Receiver operating characteristic (ROC) is a graphical representation that shows the performance of a binary classifier as its distinction threshold is varied. The ROC plots the true

positive rate (Sensitivity) as the function of the false positive rate for different cut-off points of a parameter [48].

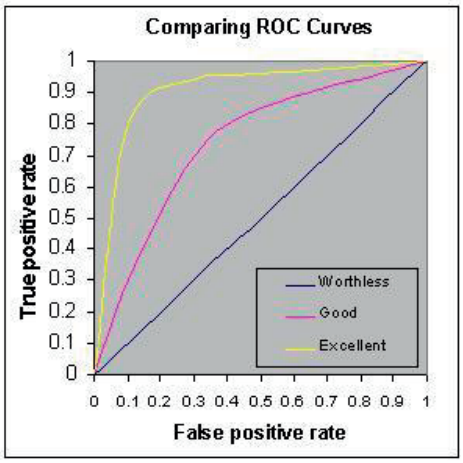


Figure 8: ROC curves.

Figure from http://gim.unmc.edu/dxtests/roc3.htm.

The closer the ROC curve is to the upper left corner the higher is the accuracy of the test. The performance according to a ROC curve is represented by the Area Under Curve (AUC), which is an evaluation of how well a parameter can distinguish between two diagnostic groups.

Enrichment analysis

Enrichment analysis is an approach to identify properties that are over-represented in a set of genes or proteins by using statistical methods. Gene ontology (GO) is a key resource for representing information on genes and gene products in a form that is suitable for enrichment analysis [49]. GOrilla and DAVID are two important tools for identifying enriched GO terms, as well as other properties (at least for DAVID), in gene lists [50, 51].

DAVID consists of integrated biological information and analytic tools aimed to extract biological concepts from large gene/protein lists. DAVID uses a comprehensive set of functional annotation tools for the researcher to gain a better understanding of the biological themes that are enriched in lists of genes [50].

GOrilla is an efficient GO analysis tool that identifies enriched GO terms in ranked gene or protein lists, without requiring the user to provide both target and background sets [51].

Aims of the study

This project consists of two main parts. The first part was initiated to improve our understanding of complementary regulatory roles exercised by different classes of transcription factors. In this part we tried to contribute to more precise knowledge by using bioinformatic and statistical methods to study correlations between structural features of transcription factors and their functional roles in gene regulation. This part includes two papers (I and II). The second part of the project was to focus on a specific process where transcription factors and gene regulation seems to be essential, and we decided to focus on key aspects of the activation and regulation of immediate-early genes, by collecting and analyzing a novel consensus set of immediate-early genes, also using some of the tools developed in Paper I and II. This part includes two papers (III and IV).

Paper I

This paper investigates correlations between TF properties and TF function, in order to identify properties and correlations that are important for understanding the role of different TFs in gene regulation. It also shows how such a resource can be used to identify properties that are enriched in a set of TFs

Paper II

This paper investigates how to predict the functional classification of TFs by using our set of properties (from Paper I) and methods in machine learning to design, select, and evaluate classifiers and feature sets.

Paper III

Paper III is a review paper describing key properties of the IER pathways and genes. It establishes a context for the analyses performed in Paper IV, and the discussion of these analyses.

Paper IV

Paper IV describes the process of making an improved consensus gene set of IER genes, and the results of extensive bioinformatics analyses of this set, including analyses with tools developed in Paper I and II.

Approaches and main findings

This part of the thesis discusses important aspects of the papers, the overall work and key contributions with respect to the aim and the original research questions that were posed. The thesis spans many types of problems, from both the bioinformatical and the biological perspective.

Computational analysis of transcription factors (Paper I)

Paper I shows that creating a good resource on properties of human transcription factors is challenging. Since transcription factors are proteins that have a key role in the general regulatory system of any cell, having a comprehensive resource on human transcription factors may facilitate our understanding of cell regulation. In this paper we have used a list of human transcription factors, originally published by Ravasi *et al.* [52], to make an annotated data set including information on Pfam domains, DNA binding domains, protein-protein interaction domains and post-translational modifications, and have worked on understanding any complementary regulatory roles exercised by these transcription factors.

Evaluating DNA binding domains

We have first assigned Pfam domains to all entries of the list of TFs, and these domains were manually reviewed and curated for evidence strongly suggesting DNA binding, in order to add annotation on Pfam domains acting as DNA-binding domains (DBDs). However, it is likely that there are additional Pfam domains with DNA-binding properties that are not annotated as such. We therefore started to identify additional Pfam domains as DNA-binding by using the threading-based method implemented in DBD-Threader [53]. We did the DBD predictions over all Pfam domains in the set of TF proteins. We estimated the overall prediction quality over all occurrences for each Pfam domain at three different levels; protein level, domain level, and residue level. Then we trained a support vector machine (SVM) with a linear kernel function [54] to distinguish between true positive and false positive prediction of DNA-binding Pfam domains, based on how consistent the predictions were across all occurrences of a given Pfam domain. The SVM had best performance on data at the residue level, so residue level %Sn and %PPV were used as features for classification. Finally we determined the final set of DBDs based on the SVM output.

Analysis of protein-protein interaction domains

In the study by Ravasi *et al.* they were able to capture cDNA clones for most human transcription factors, and used this to map actual protein-protein interactions between transcription factors [52]. We used this set and tested for correlation against other features by using a general enrichment analysis. This was done as a Fisher's exact test on a 2×2 contingency table.

Protein-protein interactions often take place via interactions between specific domains. A general enrichment analysis was done for specific Pfam domains, as well as pairs of Pfam domains, on the PPI data. The enrichment analysis showed 73 Pfam domains as enriched in protein-protein interaction, and the analysis of all possible pairs for the 73 domains showed 227 enriched pairs of Pfam domains.

Post-translational modifications of transcription factors

The modification of transcription factors by post-translational modifications may affect their activity. In this project we used information from Phosphosite for mapping of post-translational modifications, and imported data for six post-translational modifications types including phosphorylation, acetylation, methylation, O-GlcNAc, sumoylation and ubiquitination [55]. We investigated correlations between these modifications and correlations of these modifications with other properties. The results showed significant associations between most of the PTMs. However, this is most likely due to an experimental bias in the data set, where TFs tested for a given PTM also are more likely to have been tested for other PTMs, thereby creating artificially strong associations.

In this paper we identified 27 new DBDs and 318 additional TFs that have at least one Pfam DBD. We also identified 347 pairs of Pfam domains that are enriched in PPI between TFs. We used the database to identify sub-groups of TFs which are correlated with specific functions or properties, and analysis showed for example clear differences between TFs with and without a DBD.

The results show that such a comprehensive list of transcription factors properties is a useful resource for extensive data analysis; both of transcription factor properties in general and of properties associated with specific processes.

Classification of transcription factors (Paper II)

In Paper I, we created a well annotated database of 1975 human transcription factors using a set of machine learning methods. In the next study we extended an experimental classification of transcription factors based on chromatin-associated properties, and used the classification for investigation of properties and functions in each TF class.

Classification of transcription factors on chromatin opening

Wingender *et al.* have made a comprehensive classification of human transcription factors known as TFClass, which classifies transcription factors according to a hierarchy of six levels [16]. In a study from 2014, Sherwood *et al.* classified human transcription factors based on

functional properties. They used protein interaction quantitation (PIQ) for description of properties of transcription factor binding sites, and used this to classify transcription factors into three different groups; Pioneers, Settlers, and Migrants [56]. In the present study we extended their classification using a feature vector-based approach as input to machine learning methods. We used the TFClass classification and our set of TF properties, including frequent Pfam domains, DNA binding, number of DNA binding domains (DBDs), PPI, number of PPIs, PTMs (generally and individually), and number of positions for phosphorylation as the feature vector. We used multiclass classification to classify TFs, and in particular a one-vs-rest strategy for reducing the problem of multiclass classification into a binary classification problem [57].

For the final classification we applied the random forest classifier as an optimal classifier. This classifier had the best performance based on several evaluation measures, including precision (PPV), recall (sensitivity or SN), F-score, MCC (Matthews's correlation coefficient), and AUC. Finally we classified additional transcription factors into four groups based on chromatin opening, including: Pioneers, Settlers, positive and negative Migrants, based on the assumption that there are functional and structural differences between Migrants with negative and positive chromatin opening index. This classification was used together with previously published data on interactions between transcription factors, based on DNA co-binding and protein-protein interactions. This showed that there are complementary differences between the subclasses, where Pioneers often interact with other transcription factors through DNA co-binding, whereas Migrants to a larger extent are involved in protein-protein interactions. This analysis illustrates how the expanded classification is a useful resource that can be used to analyze other datasets on transcription factors and their role in gene regulation.

Identification and investigation of genes in immediate-early response processes (Paper III and IV)

Immediate-early genes (IEGs) are very rapidly expressed in response to both cell-extrinsic and cell-intrinsic signals. During stimulation extracellular signals are transduced via activity of a chain of proteins in the cell, such as extracellular-signal-regulated kinases (ERKs), mitogen-activated protein kinases (MAPKs) and members of the RhoA-actin pathway. These genes play a key role in several essential cellular systems such as the immune system. They also play key roles in different diseases, like cancer. Therefore we have tried to summarize in a review some new advances on key aspects of the regulation and activation of this kind of genes. Most previous work on immediate-early genes is based on data for single cell types. However, a larger consensus set may make it possible to distinguish between general properties and cell type specific properties. Therefore we made a robust consensus set of genes showing an early response pattern after different types of stimulation. In the present study we used a number of published time course experiments. We used gene lists for up to 60 minutes after stimulation. In most cases we selected and ranked genes based on fold

change but in some cases the selection and ranking was based on significance of change, estimated with e.g. edgeR [58]. Finally we determined a final list of 172 potential immediate-early response (IER) genes. The final list of IER genes was analyzed for enrichment of relevant properties by estimating whether the set of IER genes was significantly more enriched for that property than a reference set of protein coding genes from UniProt [59]. The analysis of IER gene properties showed consistent results with our current understanding of IER genes which confirmed that the consensus set gave a good representation of immediate-early response genes and can be a good resource for analysis of genes involved in rapid responses.

Conclusions and future perspectives

Our overall aim of this study has been to make an integrated resource on transcription factor properties using heterogeneous experimental data from different sources. This has been complemented with predicted data and used to expand an experimental classification of proteins associated with chromatin modeling, using bioinformatics and supervised machine learning methods to study correlations between structural features of transcription factors and their functional roles in gene regulation. The comprehensive list is a useful resource for researchers working on gene regulation, and it can improve our understanding of complementary regulatory roles exercised by different classes of transcription factors. This may be used to analyze e.g. expression data related to normal and disease-associated regulatory networks. There is an essential need for such studies being undertaken. Overall, a future challenge for bioinformatics will be to integrate many different properties of transcription factors that will be studied and relate these to gene expression.

Since immediate-early genes play a key role in several essential cellular systems, it is important to have a good understanding of the properties of these genes. In the last two papers we summarized some new advances in our understanding of key aspects on the activation and regulation of these genes. We compared previous observations to data from a new consensus data set of immediate-early genes. The results confirm that these genes are often in a poised state which is maintained by repressive TFs, histone modifiers and the DISF/NELF complex, and that they are in contact with enhancers through DNA looping, stabilized by cohesin and insulators. The data set is a very useful resource for evaluating important properties of immediate-early genes and may offer an interesting direction for further research, for example with links to other cellular processes such as the circadian rhythms, or on the roles of these genes in diseases like cancer. It is also an important fact that most of the experiments so far have not looked into possible roles of ncRNAs in IER, at least partly due to lack of data. However, the rapid increase in relevant experiments based on CAGE, RNA-seq and similar technologies makes this increasingly relevant, and this will be an important area for future research into the process of gene regulation in IER.

References

1. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA: **Structure, function and evolution of multidomain proteins**. *Curr Opin Struct Biol* 2004, **14**(2):208-216.
2. Filtz TM, Vogel WK, Leid M: **Regulation of transcription factor activity by interconnected post-translational modifications**. *Trends Pharmacol Sci* 2014, **35**(2):76-85.
3. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**(Database issue):D222-230.
4. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes**. *Mol Biol Evol* 2003, **20**(9):1377-1419.
5. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter**. *Annu Rev Biochem* 2003, **72**:449-479.
6. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes**. *Annu Rev Genet* 2000, **34**:77-137.
7. Ohler U, Niemann H: **Identification and analysis of eukaryotic promoters: recent computational approaches**. *Trends Genet* 2001, **17**(2):56-60.
8. Kuras L, Struhl K: **Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme**. *Nature* 1999, **399**(6736):609-613.
9. Thomas MC, Chiang CM: **The general transcription machinery and general cofactors**. *Crit Rev Biochem Mol Biol* 2006, **41**(3):105-178.
10. Borggrefe T, Yue X: **Interactions between subunits of the Mediator complex with gene-specific transcription factors**. *Semin Cell Dev Biol* 2011, **22**(7):759-768.
11. Casamassimi A, Napoli C: **Mediator complexes and eukaryotic transcription regulation: an overview**. *Biochimie* 2007, **89**(12):1439-1446.
12. Browning DF, Busby SJ: **The regulation of bacterial transcription initiation**. *Nat Rev Microbiol* 2004, **2**(1):57-65.
13. Kerschner JL, Gosalia N, Leir SH, Harris A: **Chromatin remodeling mediated by the FOXA1/A2 transcription factors activates CFTR expression in intestinal epithelial cells**. *Epigenetics* 2014, **9**(4):557-565.
14. Frietze S, Farnham PJ: **Transcription factor effector domains**. *Subcell Biochem* 2011, **52**:261-277.
15. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome**. *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
16. Wingender E, Schoeps T, Donitz J: **TFClass: an expandable hierarchical classification of human transcription factors**. *Nucleic Acids Res* 2013, **41**(Database issue):D165-170.
17. Fischer EH, Krebs EG: **Conversion of phosphorylase b to phosphorylase a in muscle extracts**. *J Biol Chem* 1955, **216**(1):121-132.
18. Walsh DA, Perkins JP, Krebs EG: **An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle**. *J Biol Chem* 1968, **243**(13):3763-3765.
19. Manning G, Plowman GD, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man**. *Trends Biochem Sci* 2002, **27**(10):514-520.
20. Deribe YL, Pawson T, Dikic I: **Post-translational modifications in signal integration**. *Nat Struct Mol Biol* 2010, **17**(6):666-672.
21. Kouzarides T: **Acetylation: a regulatory modification to rival phosphorylation?** *EMBO J* 2000, **19**(6):1176-1179.
22. Ianari A, Gallo R, Palma M, Alesse E, Gulino A: **Specific role for p300/CREB-binding protein-associated factor activity in E2F1 stabilization in response to DNA damage**. *J Biol Chem* 2004, **279**(29):30830-30835.

23. Matsuzaki H, Daitoku H, Hatta M, Aoyama H, Yoshimochi K, Fukamizu A: **Acetylation of Foxo1 alters its DNA-binding ability and sensitivity to phosphorylation**. *Proc Natl Acad Sci U S A* 2005, **102**(32):11278-11283.
24. Zhao X, Jankovic V, Gural A, Huang G, Pardanani A, Menendez S, Zhang J, Dunne R, Xiao A, Erdjument-Bromage H *et al*: **Methylation of RUNX1 by PRMT1 abrogates SIN3A binding and potentiates its transcriptional activity**. *Genes Dev* 2008, **22**(5):640-653.
25. Ozcan S, Andrali SS, Cantrell JE: **Modulation of transcription factor function by O-GlcNAc modification**. *Biochim Biophys Acta* 2010, **1799**(5-6):353-364.
26. Comer FI, Hart GW: **O-GlcNAc and the control of gene expression**. *Biochim Biophys Acta* 1999, **1473**(1):161-171.
27. Girard M, Goossens M: **Sumoylation of the SOX10 transcription factor regulates its transcriptional activity**. *FEBS Lett* 2006, **580**(6):1635-1641.
28. Gill G: **SUMO and ubiquitin in the nucleus: different functions, similar mechanisms?** *Genes Dev* 2004, **18**(17):2046-2059.
29. Cubenas-Potts C, Matunis MJ: **SUMO: a multifaceted modifier of chromatin structure and function**. *Dev Cell* 2013, **24**(1):1-12.
30. Jadhav T, Wooten MW: **Defining an Embedded Code for Protein Ubiquitination**. *J Proteomics Bioinform* 2009, **2**:316.
31. Geng F, Wenzel S, Tansey WP: **Ubiquitin and proteasomes in transcription**. *Annu Rev Biochem* 2012, **81**:177-201.
32. Husnjak K, Dikic I: **Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions**. *Annu Rev Biochem* 2012, **81**:291-322.
33. Fowler T, Sen R, Roy AL: **Regulation of primary response genes**. *Mol Cell* 2011, **44**(3):348-360.
34. Healy S, Khan P, Davie JR: **Immediate early response genes and cell transformation**. *Pharmacol Ther* 2013, **137**(1):64-77.
35. Tullai JW, Schaffer ME, Mullenbrock S, Sholder G, Kasif S, Cooper GM: **Immediate-early and delayed primary response genes are distinct in function and genomic architecture**. *J Biol Chem* 2007, **282**(33):23981-23995.
36. O'Donnell A, Odrowaz Z, Sharrocks AD: **Immediate-early gene activation by the MAPK pathways: what do and don't we know?** *Biochem Soc Trans* 2012, **40**(1):58-66.
37. Gomard T, Jariel-Encontre I, Basbous J, Bossis G, Moquet-Torcy G, Piechaczyk M: **Fos family protein degradation by the proteasome**. *Biochem Soc Trans* 2008, **36**(Pt 5):858-863.
38. Fujita T, Piuz I, Schlegel W: **Negative elongation factor NELF controls transcription of immediate early genes in a stimulus-specific manner**. *Exp Cell Res* 2009, **315**(2):274-284.
39. Yamaguchi Y, Shibata H, Handa H: **Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond**. *Biochim Biophys Acta* 2013, **1829**(1):98-104.
40. Yamada T, Yamaguchi Y, Inukai N, Okamoto S, Mura T, Handa H: **P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation**. *Mol Cell* 2006, **21**(2):227-237.
41. Schaukowitch K, Joo JY, Liu X, Watts JK, Martinez C, Kim TK: **Enhancer RNA facilitates NELF release from immediate early genes**. *Mol Cell* 2014, **56**(1):29-42.
42. de Ridder D, de Ridder J, Reinders MJ: **Pattern recognition in bioinformatics**. *Brief Bioinform* 2013, **14**(5):633-647.
43. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S: **Machine learning and its applications to biology**. *PLoS Comput Biol* 2007, **3**(6):e116.
44. MA Hearst SD, Osman E, J. P, Scholkopf B: **Support vector machines**. *Intelligent Systems and their Applications* 1998, **13**(4):10.
45. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.

46. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995: 1137-1143.
47. Efron B: **Second thoughts on the bootstrap**. *Statistical Science* 2003, **18**(2):135-140.
48. Powers DMW: **EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION**. *Machine Learning Technologies* 2011, **2**(1):37-63.
49. Gene Ontology C: **The Gene Ontology project in 2008**. *Nucleic Acids Res* 2008, **36**(Database issue):D440-444.
50. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
51. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009, **10**:48.
52. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N *et al*: **An atlas of combinatorial transcriptional regulation in mouse and man**. *Cell* 2010, **140**(5):744-752.
53. Gao M, Skolnick J: **A threading-based method for the prediction of DNA-binding proteins with application to the human genome**. *PLoS Comput Biol* 2009, **5**(11):e1000567.
54. Cortes C, Vapnik V: **Support-Vector Networks**. *Machine Learning* 1995, **20**(3):273-297.
55. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse**. *Nucleic Acids Res* 2012, **40**(Database issue):D261-270.
56. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK: **Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape**. *Nat Biotechnol* 2014, **32**(2):171-178.
57. Duan K-B, Rajapakse J, Nguyen M: **One-Versus-One and One-Versus-All Multiclass SVM-RFE for Gene Selection in Cancer Classification**. In: *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics.* Edited by Marchiori E, Moore J, Rajapakse J, vol. 4447: Springer Berlin Heidelberg; 2007: 47-56.
58. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
59. UniProt C: **Activities at the Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2014, **42**(Database issue):D191-198.

Appendix

A Python script to do enrichment analysis on the human TFs.

```
import xlrd
from math import *
from fisher import pvalue
from numpy import array, empty

class TFs_DataBase(object):

    """ Holds information from the main database.

    Members:
    TF_Names            Name of TF entry, e.g. AATF.
    Approved symbol     Approved name of TF entry, e.g. AATF.
    UniProt_IDs         Uniprot ID of TF, e.g AATF_HUMAN.

    All Domains         List of all Pfam A domains in database.

    DB_TFs              List DNA Binding TFs in database.
    PPI TFs             List Protein-Protein Interaction TFs in database.
    Experimented_PPI_TFs List Experimented PPI by Ravasi et al.

    Phosphorylation TFs List PHOSPHORYLATION TFs in database.
    Ubiquitination_TFs  List UBIQUITINATION TFs in database.
    Methylation TFs     List METHYLATION TFs in database.
    Acetylation_TFs     List ACETYLATION TFs in database.
    Sumoylation_TFs     List SUMOYLATION TFs in database.
    O GlcNAc TFs        List O-GlcNAc TFs in database.

    """
    def __init__(self):

        self.TF_Names = []
        self.Approved symbol = []
        self.UniProt_IDs = []

        self.All Domains = []

        self.DB TFs = []
        self.PPI_TFs = []
```

```
        self.Experimented_PPI_TFs = []

        self.Phosphorylation TFs = []
        self.Ubiquitination_TFs = []
        self.Methylation_TFs = []
        self.Acetylation TFs = []
        self.Sumoylation_TFs = []
        self.O GlcNAc TFs = []


DataBase = TFs DataBase()
def TF_Names(handle_Table):

    for i in range(1, sh.nrows):
        DataBase.TF_Names.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.TF_Names

def Approved_symbol(handle_Table):

    for i in range(1, sh.nrows):
        DataBase.Approved_symbol.append(sh.cell_value(rowx = i, colx = 1))
    return DataBase.TF_Names


def UniProt_IDs(handle_Table):

    for i in range(1, sh.nrows):
        DataBase.UniProt IDs.append(sh.cell value(rowx = i, colx = 2))
    return DataBase.UniProt_IDs


def All_Domains(handle_Table):

    for i in range(1, sh.nrows):
        if sh.cell_value(rowx = i, colx = 7) != '':
            DBD = sh.cell_value(rowx = i, colx = 7).split('; ')
            for name in DBD:
                DataBase.All_Domains.append(name.split('   ')[0])
        if sh.cell_value(rowx = i, colx = 8) != '':
            N DBD = sh.cell value(rowx = i, colx = 8).split('; ')
            for name in N_DBD:
                if 'Pfam-B' not in name:
                    DataBase.All_Domains.append(name.split('   ')[0])
    return list(set(DataBase.All_Domains))
```

```
def DB_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if sh.cell_value(rowx = i, colx = 6) == '+':
            DataBase.DB_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.DB_TFs


def PPI_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if sh.cell_value(rowx = i, colx = 9) == '+':
            DataBase.PPI_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.PPI_TFs


def Experimented_PPI_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if sh.cell_value(rowx = i, colx = 9) == '+'\
           or sh.cell_value(rowx = i, colx = 9) == '-':
              DataBase.Experimented_PPI_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.Experimented_PPI_TFs


def Phosphorylation_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if 'PHOSPHORYLATION' in sh.cell_value(rowx = i, colx = 11):
            DataBase.Phosphorylation_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.Phosphorylation_TFs


def Ubiquitination_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if 'UBIQUITINATION' in sh.cell_value(rowx = i, colx = 11):
            DataBase.Ubiquitination TFs.append(sh.cell value(rowx = i, colx = 0))
    return DataBase.Ubiquitination_TFs


def Methylation_TFs(handle_Table):
```

```
    for i in range(1, sh.nrows):
        if 'METHYLATION' in sh.cell_value(rowx = i, colx = 11):
            DataBase.Methylation_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.Methylation_TFs


def Acetylation_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if 'ACETYLATION' in sh.cell_value(rowx = i, colx = 11):
            DataBase.Acetylation_TFs.append(sh.cell_value(rowx = i, colx = 0))
    return DataBase.Acetylation_TFs


def Sumoylation_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if 'SUMOYLATION' in sh.cell_value(rowx = i, colx = 11):
            DataBase.Sumoylation TFs.append(sh.cell value(rowx = i, colx = 0))
    return DataBase.Sumoylation_TFs


def O_GlcNAc_TFs(handle_Table):

    for i in range(1, sh.nrows):
        if 'O-GlcNAc' in sh.cell_value(rowx = i, colx = 11):
            DataBase.O GlcNAc TFs.append(sh.cell value(rowx = i, colx = 0))
    return DataBase.O_GlcNAc_TFs


def Properties():

    """ Return a list of all TFs with specific properties from the database
    (8 properties including: DNA-Binding, PPI, Phosphorylation, Ubiquitination,
    Methylation, Acetylation, Sumoylation, O_GlcNAc respectively).
    """
    List_Names = []
    List Names.append(DB TFs(handle Table))
    List Names.append(PPI TFs(handle Table))
    List_Names.append(Phosphorylation_TFs(handle_Table))
    List_Names.append(Ubiquitination_TFs(handle_Table))
    List_Names.append(Methylation_TFs(handle_Table))
    List_Names.append(Acetylation_TFs(handle_Table))
```

```
    List_Names.append(Sumoylation_TFs(handle_Table))
    List Names.append(O GlcNAc TFs(handle Table))
    return List_Names


def Check_By_Genenames(Gene_List):

    """ Checking for a list of input genes or proteins with different existing
    names in the database (e.g. Uniprot ID, Gene names, ...). Return a modified
    list of input based on existing names.
    """
    Modify_Gene_List = []
    for i in range(len(Gene_List)):
        i  = 0
        for j in range(1, sh.nrows):
            if Gene_List[i] == sh.cell_value(rowx = j, colx = 0)\
               or Gene_List[i] == sh.cell_value(rowx = j, colx = 1)\
               or Gene_List[i] == sh.cell_value(rowx = j, colx = 2)\
               or Gene_List[i] == sh.cell_value(rowx = j, colx = 2)[:-6]\
               or Gene_List[i] in sh.cell_value(rowx = j, colx = 3).split('; '):
                    Modify_Gene_List.append(sh.cell_value(rowx = j, colx = 0))
                    i_ += 1
                    break
         if i_ == 0:
             Modify Gene List.append(Gene List[i])
    return Modify_Gene_List


def Info_from_confusion_matrix(cm):

    """ Return Observed, Expected, Pvalue, and Mcc from confusion matrix in
    overrepresentation analysis.
    """
     Info = []
    _Info.append(cm[0])
    _Info.append((cm[0]+cm[1]) * (cm[0]+cm[2])/( cm[0]+cm[1]+cm[2]+cm[3] ))
    p = pvalue(cm[0], cm[1], cm[2], cm[3])
    _Info.append(p.two_tail)
    M = (cm[0]+cm[1]) * (cm[2]+cm[3]) * (cm[0]+cm[2]) * (cm[1]+cm[3])
    if M != 0:
        _Info.append(float( (cm[0]*cm[3]) - (cm[1]*cm[2]) )/sqrt(M))
    else:
        _Info.append('-')
    return _Info
```

```
def Benjamini_Correction(Pvalues):

    """ Return corrected Pvalues (Benjamini) for a list of Pvalues. """
    Pvalues = array(Pvalues)
    n = float(Pvalues.shape[0])
    New Pvalues = empty(n)
    values = [ (pvalue, i) for i, pvalue in enumerate(Pvalues) ]
    values.sort()
    values.reverse()
    new_values = []
    for i, vals in enumerate(values):
        rank = n - i
        pvalue, index = vals
        new_values.append((n/rank) * pvalue)
    for i, vals in enumerate(values):
        pvalue, index = vals
        New Pvalues[index] = new values[i]
    return New_Pvalues

def Final_List(List_Info_Overrepresentations):

    """ Return a list of all significant overrepresentation properties
    and Pfam domains with Benjamini correction.

    """
    Pvalues = [ List Info Overrepresentations[i][3]\
                for i in range(len(List_Info_Overrepresentations)) ]

    Benjamini = Benjamini Correction(Pvalues)
    Index_Sorted_Pvalues = sorted( range(len(Pvalues)), key = lambda k: Pvalues[k] )
    Sorted_Benjamini = [ Benjamini[i] for i in  Index_Sorted_Pvalues ]
    Info with Benjamini = [ List Info Overrepresentations[Index Sorted Pvalues[i]]\
                            for i in range(len(Index_Sorted_Pvalues)) ]

    # Adding  corrected Benjamini
    for i in range(len(Info_with_Benjamini)):
        Info with Benjamini[i].insert(4, Sorted Benjamini[i])

    # Sorting and filtering based on Pvalues
    Benjamini_Final_List = [  Info_with_Benjamini[i] for i in range(len(Info_with_Benjamini))\
                              if Info_with_Benjamini[i][4] < 0.05  ]
    return Benjamini_Final_List
```

```
def _Print(Sorted_List):

    """ To pretty print the significant overrepresentation items. """

    for i in range(len(Sorted_List)):
        print "{0:<20} {1:<24} {2:<8} {3:<8} {4:<11} {5:<11} {6:<10}"\
              .format(
                      ''*20, Sorted_List[i][0],\
                      Sorted List[i][1],\
                      Sorted_List[i][2],\
                      '%.2E' %Sorted_List[i][3],\
                      '%.2E' %Sorted List[i][4],\
                      '%.3f' %Sorted_List[i][5]\
                     )


# Initial lists of database for analysis
print
Path_Reference = raw_input('Please enter the path of the reference database (xls, with title
row): ')
handle_Table = xlrd.open_workbook(Path_Reference)
sh = handle Table.sheet by index(0)
List_Name_Properties = [
                          'DNA_Binding',\
                          'PPI',\
                          'Phosphorylation',\
                          'Ubiquitination',\
                          'Methylation',\
                          'Acetylation',\
                          'Sumoylation',\
                          'O GlcNAc'
                        ]  # List of the property names


List Properties = Properties()                          # List of the properties
Background_TFs = TF_Names(handle_Table)                 # List of all TFs in database
all domain = All Domains(handle Table)                  # List all Pfam A domains
Background PPI = Experimented PPI TFs(handle Table)     # List of TFs for experimental PPIs


#List Genes or Proteins from the Test set for analysis
Path Test Set = raw input('Please enter the path of the test set (xls, with title row): ')
Number_Col = raw_input('Please enter the column number for TFs in your table (starting at 0):
')
External Data = xlrd.open workbook(Path Test Set)
Sheet_Names = External_Data.sheet_names()               # List of the gene/protein list names
```

```
List_External_Data = []                                  # List of the gene/protein lists
for i_sheet in range(External_Data.nsheets):
    Sheet book i = []
    sh_ = External_Data.sheet_by_index(i_sheet)
    for j in range(1, sh_.nrows):
        Sheet book i.append(sh .cell value(rowx = j, colx = int(Number Col)))
    Sheet_book_i = Check_By_Genenames(Sheet_book_i)
    List External Data.append(Sheet book i)

print '\n\n'

# Print headline
print "{0:<20} {1:<24} {2:<8} {3:<8} {4:<11} {5:<11} {6:<10}"\
      .format( 'Category', 'Term', 'Observed', 'Expected', 'Pvalue', 'Benjamini', 'MCC' )
print "-"*93

# Overrepresentation analysis
for i in range(len(List_External_Data)):

    Per info Properties = []
    Per_info_Domains = []

    # Overrepresentation of properties (DB/PPI/Individual PTMs)
    for j in range(len(List_Properties)):

        cm = confusion_matrix = [0, 0, 0, 0]
        # Needs to setup for PPI property separately
        if List_Name_Properties[j] == 'PPI':
                Background = Background_PPI
        else:
                Background = Background TFs
        cm[0] += len( set(Background) & set(List_External_Data[i]) & set(List_Properties[j]) )
        cm[1] += len( (set(Background)&set(List_External_Data[i])) - set(List_Properties[j]) )
        cm[2] += len( (set(Background)&set(List Properties[j])) - set(List External Data[i]) )
        cm[3] += len( set(Background)-(set(List_Properties[j]) | set(List_External_Data[i])) )
        Per_Property = Info_from_confusion_matrix(cm)
        Per Property.insert(0, List Name Properties[j])
        Per_info_Properties.append(Per_Property)

    # Overrepresentation of Pfam domains
    for domain in all_domain:
        Per_Domain = []
        cm = confusion_matrix = [0, 0, 0, 0]
        for i_sh in range(1, sh.nrows):
```

```
            if sh.cell_value(rowx = i_sh, colx = 0) in List_External_Data[i]:
                if domain in sh.cell_value(rowx = i_sh, colx = 7)\
                   or domain in sh.cell_value(rowx = i_sh, colx = 8):
                            cm[0] += 1
                else:
                            cm[2] += 1
            else:
                if domain in sh.cell_value(rowx = i_sh, colx = 7)\
                   or domain in sh.cell_value(rowx = i_sh, colx = 8):
                            cm[1] += 1
                else:
                            cm[3] += 1
          Per_Domain = Info_from_confusion_matrix(cm)
        Per Domain.insert(0, domain)
        Per_info_Domains.append(Per_Domain)

    F Properties = Final List(Per info Properties)
    F_Domains = Final_List(Per_info_Domains)

    # Print final result
    print Sheet_Names[i]
    if F_Properties == [] and F_Domains == []:
            print ' '*21 + 'There is no overrepresented item for this Gene/Protein list'
    if F_Properties != []:
        if F_Domains != []:
            _Print(F_Properties)
            print ' '*21 +  '-'*10
             Print(F Domains)
        else:
            _Print(F_Properties)
    else:
        _Print(F_Domains)
    print
```

Paper I

Bahrami et al. BMC Research Notes (2015) 8:82
DOI 10.1186/s13104-015-1039-6

BMC Research Notes

RESEARCH ARTICLE Open Access

A property-based analysis of human transcription factors

Shahram Bahrami[1,2†], Rezvan Ehsani[1†] and Finn Drabløs[1*]

Abstract

Background: Transcription factors are essential proteins for regulating gene expression. This regulation depends upon specific features of the transcription factors, including how they interact with DNA, how they interact with each other, and how they are post-translationally modified. Reliable information about key properties associated with transcription factors will therefore be useful for data analysis, in particular of data from high-throughput experiments.

Results: We have used an existing list of 1978 human proteins described as transcription factors to make a well-annotated data set, which includes information on Pfam domains, DNA-binding domains, post-translational modifications and protein–protein interactions. We have then used this data set for enrichment analysis. We have investigated correlations within this set of features, and between the features and more general protein properties. We have also used the data set to analyze previously published gene lists associated with cell differentiation, cancer, and tissue distribution.

Conclusions: The study shows that well-annotated feature list for transcription factors is a useful resource for extensive data analysis; both of transcription factor properties in general and of properties associated with specific processes. However, the study also shows that such analyses are easily biased by incomplete coverage in experimental data, and by how gene sets are defined.

Keywords: Transcription factor, DNA-binding domain, Protein–protein interaction, Post-translational modification, Enrichment analysis

Background

Transcription Factors (TFs) are proteins that in most cases bind to specific DNA sequences known as Transcription Factor Binding Sites (TFBSs), in particular in enhancer regions or in promoter regions near their target genes [1]. The transcription factors modulate transcription initiation and regulate gene expression, and are thereby an essential part of the general regulatory system of any cell. Normally regulation of gene expression involves the binding of multiple transcription factors to the regulatory regions of a given gene. However, the definition of TFs is not always very clear-cut, and may include DNA-binding proteins that do not recognize any specific DNA motif, proteins that do not bind DNA, but influence transcription through protein–protein interactions (PPIs), and proteins that influence transcription in more indirect ways, for example by mediating chromatin remodeling [2].

Transcription factors are typically modular in structure, and will often contain effector domains and other domain types, in addition to (in most cases) one or more DNA-binding domains (DBDs). A DBD is typically a protein domain with a characteristic fold that can recognize a specific DNA sequence (motif), and thereby regulate transcription of specific target genes, although there are also examples of TFs with a more general (less motif-specific) affinity to DNA [3,4]. The interaction between a TF and its TFBSs defines the specificity of the TF, which is mediated by non-covalent interactions between the structural motif of the TF DBD and the surface of the DNA bases and backbone atoms [5,6].

Most TFs belong to one of two major classes; the general TFs and the site-specific TFs. The general TFs are important components of the basal transcriptional machinery around transcription start sites. The general TFs cannot stably bind to promoter or enhancer regions on

* Correspondence: finn.drablos@ntnu.no
†Equal contributors
[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, P.O. Box 8905, NO-7491 Trondheim, Norway
Full list of author information is available at the end of the article

their own. In most cases they are bound to regulatory regions through interaction with site-specific DNA-binding TFs. These site-specific TFs bind to DNA through their DBDs, and at the same time they bind to other transcriptional regulatory proteins via effector domains [7], thereby stabilizing the whole complex.

Protein–protein interactions are important for the function of proteins and the processes they are involved in, and such interactions are often facilitated by specific protein domains interacting with each other. Therefore, understanding protein interactions at the domain level can provide a generalized understanding of protein interaction, and thereby protein function. As an example, Gao *et al.* constructed a protein–protein network of transcription factors involved in regulation of liver cell proliferation and regeneration [8]. They identified 64 interactions in a regulatory network, providing additional information on the regulatory aspects of liver regeneration.

An important group of regulatory mechanisms available to the cell is post-translational modifications (PTMs). The PTMs are highly dynamic and often reversible, and they may occur on almost all proteins. Most PTMs change the properties of a protein by the addition of a specific chemical group to one or more of its amino acid residues [9,10]. The PTMs make possible diverse signaling that is suitable for relaying rapid messages throughout the cell. Some PTMs, such as phosphorylation, can be quite transient, and may serve to rapidly activate or deactivate a protein, whereas other PTMs may be more long-lasting. PTMs may create further signaling through modular protein domains that recognize particular types of PTMs located on specific residues. A relevant example of how PTMs may modify TF function is the MEF-2A factor which regulates gene expression in neuronal cells, where it can act as either a transcriptional activator or a repressor. This switch is controlled by post-translational modification of MEF-2A, with acetylated MEF-2A acting as a transcriptional activator, whereas the factor acts as a transcriptional repressor when it is modified by sumoylation and phosphorylation [11].

This shows that the regulatory roles of TFs can be modified by the properties of the TFs, including DNA-binding and effector domains, PPIs and PTMs. Therefore there is a need to increase our knowledge about TF domains and other properties, in addition to their binding sites in target genes, and this makes a collection of well-curated annotation data of TFs highly relevant.

There are some existing TF databases, but in general they contain very limited information about TF properties, except for DNA motif specificity, most often through a Position Weight Matrix (PWM), and links to more general protein databases with additional information. For example, JASPAR is an open-access database of DNA binding site profiles, based on collections of position frequency matrices (PFMs) that are mainly derived from published data, including chromatin immunoprecipitation and sequencing (ChIP-seq) experiments. The newest JASPAR version includes interfaces to several packages (BioPython, Rtool, R/Bioconductor) to facilitate access for both manual and automated methods [12,13].

Zhang *et al.* published in 2012 a comprehensive animal transcription factor database based on DNA-binding domains, where they collected and curated 71 animal TF families [14]. Although this includes detailed annotations for each TF (basic information, gene structure, functional domain, 3D structure hit, Gene Ontology, pathway, protein–protein interaction, paralogs, orthologs, potential TF-binding sites and targets), it is not very suitable for detailed analysis of TF properties. Fulton *et al.* made in 2009 a catalog of mouse and human TFs (called TFCat), where TFs were classified according to evidence supporting DNA-binding and transcriptional activation [15]. TFCat was based on information from four transcription factor data sets, and categorized DNA-binding TFs into 9 protein groups with 39 protein families. It is a very useful resource for TF classification, but with limited information on TF properties. Vaquerizas *et al.* used a set of 1391 manually curated sequence-specific DNA-binding transcription factors to investigate function, genomic organization and evolutionary conservation [16]. Ravasi *et al.* identified almost 2000 proteins from the human genome that are potential TFs [17]. They built a global atlas of combinatorial transcriptional regulation in mouse and human and screened for physical interactions between the majority of human and mouse DNA-binding transcription factors. This is again a useful resource, but with limited additional information.

In this paper we describe the collection and curation of a list of properties for human TFs, using the list of TFs published by Ravasi *et al.* The main reason for using this particular data set was that it also includes a consistent set of protein–protein interaction data, with a clear distinction between missing data and lack of interaction. The properties that were added include DNA-binding domains, protein–protein interactions, and post-translational modifications. We then show how this can be used for example to identify sub-groups of TFs and to correlate these with specific functions, and to identify TF properties that are associated with specific processes. However, we also show that such analyses are easily biased by data set composition and incomplete annotations, and therefore have to be interpreted with great care. The TF property data set and software for data analysis is available with the paper as additional data.

Methods

Initial definition of a data set of human TFs

We used a list of 1988 human transcription factors, originally used by Ravasi *et al*. to build an atlas of combinatorial transcriptional regulation [17]. The gene names were checked against HGNC [18] and UniProt [19], and duplicates were removed. This gave a final list of 1978 TFs. Initial annotation of the TFs was based on database entries downloaded from UniProt (last update done using release 2012_07).

Comparison to other TF collections

The gene list from Ravasi *et al.* was compared to previously published gene lists from Zhang *et al.* [14] and Vaquerizas *et al.* [16]. These additional gene lists were downloaded from supplementary material. DAVID does not accept HGNC gene names for explicit definition of background, therefore the gene names were remapped to UniProt IDs for DAVID analysis, using the ID converter of BioMart (http://www.biomart.org/) [20].

General domain annotation

Specific domains, as defined for example in Pfam [21], are often associated with specific functions, and are therefore an important annotation resource. Unfortunately the Pfam annotation in UniProt does not include information about sequence position of Pfam domains. Therefore we downloaded the most recent swisspfam list from Pfam (last update done using release 12.03.2013), and searched the list for UniProt IDs [19,21].

Our annotation data include both levels of Pfam families; Pfam-A and Pfam-B. Both entry types are made from the most recent release of UniProtKB at a given time and produced automatically from the non-redundant clusters after sequence clustering. Pfam-A entries can be successfully annotated by profile HMM searches of primary sequence databases, whereas Pfam-B entries are un-annotated [21].

Adding annotation on DNA-binding domains

In the following description we try to distinguish between the domains as defined by Pfam (*Pfam domains*), and the individual occurrences of these domains in a set of proteins (*domain occurrences*). In order to add annotation on Pfam domains acting as DNA-binding domains (DBDs), all entries for Pfam domains assigned to the list of TFs were first manually reviewed and curated for evidence strongly suggesting DNA binding, using Pfam descriptions and associated literature references. In order to get a more complete annotation of DBDs in these proteins, we then used a DBD prediction method to identify additional Pfam domains as DNA-binding. In order to distinguish between sporadic and consistent predictions we did the DBD predictions over all Pfam domains in the set of TF proteins, including domains assumed not to be DNA-binding. We then estimated the overall prediction quality over all occurrences for each Pfam domain, on the hypothesis that it was a DBD, and used a support vector machine (SVM) [22] to distinguish between true positive and false positive cases. Ideally, Pfam domains where individual occurrences frequently overlap with DBD predictions should be accepted as true positive cases, whereas Pfam domains with few overlaps should be rejected as false positives. The challenge is to find a suitable cutoff between these two alternatives.

We used the threading-based method DBD-Threader [23] for the prediction of DNA-binding domains. In this method DNA-binding propensity is calculated using a statistical DNA–protein pair potential. The sequence of a target protein is compared against an experimentally determined template library of DNA-binding protein domains, using threading. Any significant template hits are further evaluated using the DNA–protein interaction energy, calculated using the alignment of the target template and the corresponding DNA structure in complex with the template protein. If there is at least one significant template for a target protein according to the specified Z-score and energy threshold conditions, the protein is predicted to be DNA-binding, otherwise it is classified as non-DNA-binding [23]. It has been shown that DBD-Threader has significantly improved performance when both threading Z-score and protein–DNA interaction propensity are taken into account, leading to a sensitivity of 56% and a precision of 86% on a benchmark set with 179 DNA-binding and 3797 non-DNA-binding proteins [23]. The method has also shown good performance in an independent benchmark study, in particular with respect to specificity [24].

We used a reference set of TFs with Pfam domains where we knew from manual curation that these specific Pfam domains were DNA-binding. On this set we predicted DBDs using DBD-Threader. We then compared annotated and predicted DNA-binding regions, and estimated the quality of the predictions at three different levels; protein level, domain level, and residue level, in order to find optimal criteria for identifying false positive predictions.

The protein level

At this level we predicted whether a protein was DNA-binding or not, irrespective of domain overlap. We used the set of proteins where curated annotation data showed that they were DNA-binding because they contained a Pfam domain annotated as DNA binding [Additional file 1]. We then counted the number of TFs with a known DBD that also were predicted to have a DBD, and estimated the rate of true positive predictions, or sensitivity (Sn, Equation 1).

$$Sn = TP/(TP + FN) \quad (1)$$

The domain level

At the domain level we tested how often the predicted DBD (for proteins correctly predicted to have a DBD) showed overlap with the known DBD (from curated annotation data), see Figure 1 for details. For each known DBD we compared it to the predicted DBD and estimated the amount of overlap relative to the Pfam domain. An overlap of at least 1 residue was counted as significant, and the values for TP, FN and FP were used to estimate sensitivity (Sn, Equation 1) and positive predictive value (PPV, Equation 2).

$$PPV = TP/(TP + FP) \quad (2)$$

The residue level

At the residue level we measured the amount of overlap between known and predicted DBDs for the actual overlaps that were identified above. This was done according to Figure 2, and used to estimate Sn and PPV as for the domain level.

Predicting new DBDs

DBD-Threader was run on all TFs, and occurrences of Pfam domains showing any overlap with DBD predictions were used as an indication of potential DNA-binding. In order to distinguish between random overlaps and true DBDs we used the Support Vector Machine method (SVM) [22] as implemented in scikit-learn version 0.15.0 [25], with a linear kernel function, and used it to separate false positive from true positive cases, based on prediction quality according to the hypothesis that each Pfam domain is a DBD. The Pfam domains annotated as DBDs after manual curation were considered as positive data, and for negative data we identified any additional Pfam domains in the DNA-binding proteins with at least one known DBD, arguing that most likely the majority of the remaining domains of these proteins are non-DBDs. These Pfam domains were evaluated by manual curation (scientific literature and Pfam entry annotation), and were separated into 2 groups; Pfam domains with *unknown* DBD status, and *non-DBD* Pfam domains [Additional file 1]. Obviously, only non-DBD Pfam domains that showed some overlap with DBD-Threader predictions could actually be used as negative data for the SVM classifier. Initial tests showed that the SVM had best performance on data at the residue level, leading to better separation of positive and negative cases (data not shown), so we used residue level %Sn and %PPV as features for classification. We then determined the final set of DBDs based on the SVM output.

PTM annotation

For data on post-translational modifications (PTMs) we used information from PhosphoSite (last update done using release 01.01.2014) [26]. We imported data for 6 PTM types; acetylation, methylation, O-GlcNAc, phosphorylation, sumoylation and ubiquitination.

GOrilla and DAVID

We used GOrilla [27,28] and DAVID [29] for enrichment analysis of TF subsets on a broad range of annotation data. The reason for using both tools is that although DAVID can analyze a broader range of properties, the information in GOrilla is more up to date. In general we used a specific subset as the positive set, and the full set of TFs as background. In cases where we could identify the subset of TFs for which we had reliable data (e.g. the PPI data) we used this subset as background. In most cases (e.g. for PTMs) it was difficult to identify TFs for which we actually had a lack of data (rather than negative data), and in these cases the full TF set was used.

Protein–Protein Interactions

Ravasi *et al.* were able to capture cDNA clones for 1222 TFs in human, in order to map PPIs [17]. The number of possible interactions (including homodimers) is $\frac{n(n+1)}{2} = \frac{1222 \times 1223}{2} = 747253$, but based on the data from Ravasi *et al.* only 762 out of these (0.1%) were observed as actual interactions. This set was tested for correlation against other features, using a general enrichment analysis.

Enrichment analysis

The enrichment analysis was implemented as a Fisher's exact test on a 2×2 contingency table. Observations

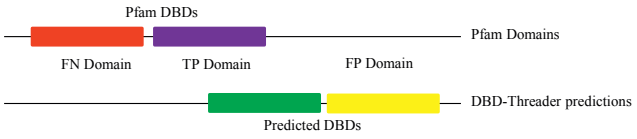


Figure 1 Prediction quality at the domain level. Domains are classified as TP, FN and FP as shown, relative to the curated Pfam domains. TNs are not included in this comparison, as negative domains are not well defined.

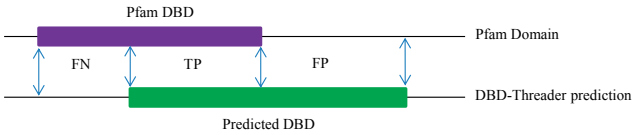


Figure 2 Prediction quality at the nucleotide level. Regions are classified as TP, FN and FP as shown, relative to overlap with the curated Pfam domains. TNs are not included in this comparison, as they represent a very large fraction of the comparison, which may bias the analysis.

were grouped according to pairs of properties, like being involved in PPIs (yes/no) and having a DBD (yes/no). This was then tested using the Fisher's exact test, in most cases with a threshold for p-value at 0.05 after Benjamini correction for multiple testing. In addition to the p-value, the expected number of occurrences and the Matthew's correlation coefficient (MCC, Equation 3) was estimated for cases with significant p-values. The testing was implemented using the full set of TFs (1978) as background for all properties except PPI. For the PPI case we used the 1222 TFs actually mapped for PPI in the Ravasi *et al*. paper as background. For calculation of MCC, a TF was considered as TP if it had both properties, as TN if it had none of properties and as FN or FP if just had one of the properties (based on the 2×2 contingency table).

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / \mathrm{sqrt}\left((\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})\right) \quad (3)$$

Python scripts were used to extract subgroups of TFs with specific properties for enrichment analysis [30]. Biopython was used to extract all gene names for each TF from the UniProt files [31]. The p-values were estimated using the Fisher 0.1.4 package [32]. The software for enrichment analysis is available with the paper.

Ethical approval and consent

This study is based on human data. However, all data have been downloaded from open data repositories (UniProt, Pfam, PhosphoSite) or from supplementary material from existing publications (see text), and cannot be linked to individuals. Ethical approval and consent is therefore not required.

Results and discussion

Making an initial set of TFs

The starting point for the annotated TF list was the set of 1988 TFs by Ravasi *et al*. [17]. These TFs were then supplemented with annotation data as described below and in Methods, in particular with respect to UniProt IDs, Pfam domains including DBDs, PPI data and PTMs.

Comparison to other TF collections

We wanted to use the data set by Ravasi *et al.* in order to utilize the consistent set of PPI data generated for that particular data set. However, alternative data sets have been used in other studies, and in order to put the set from Ravasi *et al.* into context, we compared it to the sets from Zhang *et al.* [14] and Vaquerizas *et al.* [16]. The set by Zhang *et al.* is based on manual curation of animal TF families, and includes a separation into DNA-binding TFs, TF cofactors and chromatin remodeling factors. The set by Vaquerizas *et al.* is based on curation of a list of potential TFs identified from InterPro database entries.

We first tested for overlap between the different lists based on unique HGNC gene names (see below). This showed a quite similar overlap of 1253 genes between Ravasi and Vaquerizas, 1374 between Ravasi and Zhang, and 1404 between Vaquerizas and Zhang. These numbers are on average 10% lower if we focus on DNA-binding TFs (1132, 1100, and 1359, respectively (see below for definition of DBDs in the Ravasi set)). Of the genes included in the Ravasi set, 186 and 66 are classified in the Zhang set as TF cofactors and chromatin remodeling factors, respectively. This overlap is reduced to just 14 and 10 if we focus on DNA-binding TFs in the Ravasi set.

The similarity between the data sets from Ravasi and Vaquerizas is further confirmed by comparing the distribution of domain types. The Vaquerizas set is strongly dominated by the InterPro domains ZNF-C2H2, Homeodomain, HLH and bZip, in that order. This is very similar to the distribution of Pfam domains in the Ravasi set (see below for how they were mapped), which is dominated by the Pfam domains for zinc fingers, homeobox, HLH and bZIP (Figure S1 [see Additional file 2]). The Ravasi set may be somewhat enriched in rare Pfam domains (i.e. domains found less than 5 times), but this may also be caused by differences between InterPro and Pfam.

In order to highlight the differences between these collections we used unique genes from each collection as input to DAVID and GOrilla, in each case using the full gene list for that collection as background. The genes that are unique to Ravasi compared to Vaquerizas are enriched for histone-related properties and transcription

co-factor activity (results not shown), indicating that it contains some cases that are not classical TFs. The Vaquerizas set is, on the other hand, enriched for RNA binding activity, but also catalytic activity, indicating that also this data set may contain cases that are not TFs according to a strict definition. Comparison of the Ravasi data to the Zhang data shows a similar pattern, with some enrichment for RNA binding and histone-related properties in the Ravasi set. This shows that the gene set defined by Ravasi *et al.* may have some inherent biases, but that this may be a problem also in other gene sets.

Mapping of UniProt IDs and Pfam domains

The gene names by Ravasi *et al.* were mapped to unique HGNC and UniProt IDs. In total 1978 TFs (99.5%) could be mapped to unique IDs. Mapping of Pfam domains was done using the annotations from Pfam (in swisspfam) [21]. The list of 1978 human TFs had 1664 unique Pfam domains, which included 936 Pfam-B domains and 728 Pfam-A domains. However, most of the Pfam domains have few occurrences in the set of human TFs (see later).

Mapping of DBDs

Verification on known Pfam DBDs

The ability for motif-specific DNA binding is an important property of most TFs. However, it is not necessarily an essential property, as TFs also can interact through PPIs. The observation of TFs that may bind to regions without any apparent binding site motifs highlights this. Motif-specific vs motif-less binding may have functional relevance, and it is therefore important to identify TFs with and without DNA-binding domains.

Less than 1% of all proteins have an experimentally determined structure, which makes it difficult to assign function based on structure. However, significantly similar sequences may share function, although functional roles of related proteins can change during evolution [33]. Therefore prediction methods based on sequence/structure similarity can be used to try to identify DNA-binding domain types when annotation is lacking. However, such predictions will contain some false positive and false negative predictions. It is difficult to correct for false negative predictions, i.e. to recognize something that was missed by the prediction method. However, it may be possible to correct for false positive predictions by estimating prediction quality over a set of predictions. Here we used Pfam domains as a basis, and tried to predict individual occurrences of DNA-binding for these Pfam domains. We could then estimate the consistency of prediction over all occurrences of a given Pfam domain as a quality measure, and use this to identify predictions that are likely to be false positive.

As a first step the 728 Pfam-A entries were checked for DNA-binding properties from scientific literature and Pfam entry annotation. This showed that after manual curation 70 of the Pfam-A domains were confirmed to be DNA-binding [see Additional file 1], and the proteins that had at least one of these DNA-binding domains were classified as DNA-binding proteins. These 70 DNA-binding Pfam domains were found in 907 proteins, whereas 1071 proteins did not have a reliably annotated DNA-binding domain at this stage.

We then used DBD-Threader to predict additional Pfam domains as DBDs [23] (please see Methods for details). As an initial estimate of the expected reliability of predictions, we started by doing prediction on the 907 TFs with known DBDs. These predictions were evaluated at three different levels. At the *protein level* we just checked whether the protein was predicted to be DNA binding or not. This may be useful for classification of TFs, but it does not identify new DNA-binding domains. Therefore, for the true positive predictions at the protein level we also evaluated the predictions at the *domain level*, by checking whether the prediction was able to identify the correct Pfam domain as DBD. This was evaluated both for each domain type, and over all domain occurrences. For the true positive predictions at the domain level, we finally evaluated the predictions at the *residue level*, by checking how well the predictions overlap with the Pfam domain annotated as DBD. The results (Table 1) showed that 776 out of the 907 TFs had been correctly predicted by DBD-Threader as DNA-binding. At the domain level, 40 out of the 70 known DNA-binding domains were correctly predicted by DBD-Threader at least 50% of the time, giving a sensitivity of 57%. We then considered the domains with correct prediction frequency of less than 50% as FN domains. Statistics based on domain occurrences rather than domain types gave a higher sensitivity (74%), showing that

Table 1 Prediction results for DNA-binding domains on positive data

Level	Unit	NPfam	Npredicted	TP	FP	TN	FN	Sn	PPV
Protein	proteins	907	776	718	-	-	189	79.16	-
Domain	domains	70	46	40	-	-	30	57.14	-
Domain	occurrences	1159	872	863	519	-	296	74.46	62.45
Nucleotide total	nucleotides	69320	43326	42783	16899	-	26537	61.72	71.68
Nucleotide average	nucleotides	59	49	49	32	-	89	35.51	60.49

performance is better on frequently occurring domains. Doing the statistics at the level of residues gave a somewhat lower sensitivity (62%). The most likely reason for this is shown in the average values, with a relatively high FN rate. This shows that the Pfam domains on average are longer than the predicted DBDs.

The results in Table 1 show that DBD-Threader in general works quite well, with sensitivity of almost 75% for the identification of DNA-binding domains. In particular it seems to work well for frequent DBDs, which means that a large fraction of DBD-containing proteins will be correctly identified, whereas rare cases are more likely to be missed.

Some predictions were checked in more detail, based on high FP/FN rates or large differences in Sn and PPV. This involved three domain types (LAG1-DNAbind (PF09271), BTD (PF09272), and HNF-1_N (PF04814)), and two of these (PF09271 and PF09272) did illustrate a potential problem, as there was one predicted continuous DBD overlapping two Pfam domains (Figure 3). This gives a low overlap when each domain is treated individually. The manual evaluation also showed that the HNF-1_N domain is likely to be an outlier. However, this constitutes a small fraction of the actual domains, and has minor impact on the analysis.

Identification of additional DBDs

For identifying additional Pfam domains as DBDs we used DBD-Threader predictions as a starting point. We then used the average overlap over all occurrences of each Pfam domain as input for a Support Vector Machine (SVM) [22], in order to identify Pfam domains that had too low overlap with DBD predictions to be classified as DNA-binding. As positive data we used the 40 Pfam domains that were correctly predicted by DBD-Threader as DNA-binding. As negative data we used any additional Pfam domains co-occurring with the 40 Pfam domains in the positive set [Additional file 1], based on the assumption that most TFs only have one type of DBD. This may be an oversimplification in some cases, but the SVM approach is supposed to be robust with respect to outliers. The negative data also had to show some overlap with DBD-Threader predictions in order to be useful for defining a classification cutoff between true positive and false positive cases (all Pfam domains without any overlap with DBD predictions will be zero in both Sn and PPV). This left only 6 Pfam domains as negative data. However, this should be a reliable data set of non-DBD Pfam domains in DNA-binding proteins, despite the small size.

The SVM classifier was used with the %Sn and %PPV values for DBD-Threader predictions on each Pfam domain, over all occurrences (i.e. for the hypothesis that the Pfam domain is a DBD). The performance of the classifier was assessed on the 46 Pfam domains with known classification by using a two-way cross-validation with five re-samplings, in addition to a leave-one-out cross-validation. This gave an average performance of 98% for both Sn and PPV. We then used this SVM to classify the remaining Pfam domains, based on overlap (or lack of overlap) of individual occurrences of each domain with the DBD-Threader predictions (Figure S2 [see Additional file 2]). For prediction of new DBDs we focused on Pfam-A domains, and 38 Pfam domains not included in the training set showed a non-zero overlap with DBD-Threader predictions. According to the SVM step 27 of these Pfam domains could be reliably identified as DNA-binding whereas 11 Pfam domains were more likely to be non-DNA-binding (Table 2).

Following the above analysis we had in total 97 Pfam-A domains annotated as DNA-binding, including the 30 domains that were annotated as DBD in literature, but not reliably predicted by DBD-Threader in the initial analysis. A total of 1225 proteins had at least one occurrence of a Pfam domain annotated and/or classified as DBD, and were therefore considered to be DNA-binding, whereas the remaining 753 proteins could not be identified as DNA-binding. This means that at least 61% of the TFs are DNA-binding, and this number seems to be comparable to the result from Fulton *et al.* [15].

Pfam-B domains were not included in the final prediction process for new DBDs. Such domains are generated by an automatic process, which means that they do not have a stable definition, and they will often be of low quality. Also, they had only minor impact on the actual TF classification. 45 Pfam-B domains showed at least some overlap with DBD-Threader predictions. Following the SVM-based analysis 25 out of them were confirmed as DNA-binding, whereas 20 Pfam-B domains were

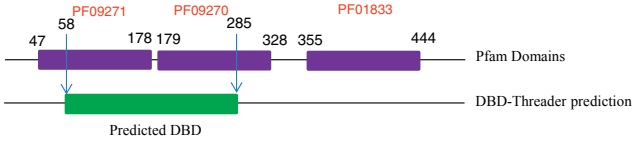


Figure 3 Example of a challenging DBD prediction. The predicted region overlaps with two independent Pfam domains.

Table 2 New DNA-binding and non-DNA-binding domain types

DBD	DBD	DBD	non-DBD*
Homeobox_KN	zf-C2H2_6	Maf1	PBC
MCM2_N	zf-C2H2_4	zf-H2C2_5	zf-C2H2_2
CBFD_NFYB_HMF	TFIID-18 kDa	Exo_endo_phos	TFIIA
SKIP_SNW	TFIIB	DUF3432	SCAN
Ku	DNA_methylase	Toprim	Prox1
Pax2_C	TFIID_20kDa		SSXRD
TAFII28	ResIII		HJURP_C
DUF2028	FAD_binding_7		Ku_N
Histone	RNA_pol_Rpb1_1		DNA_photolyase
zf-H2C2_2	SOXp		SNF2_N
zf-met	DNA_topoisoIV		TIG

*After filtering predicted DBDs for false positives.

identified as non-DNA-binding. The 25 possibly DNA-binding Pfam-B domains were found in 27 TFs, but 24 of these TFs had at least one DNA-binding Pfam-A domain, and had therefore already been identified as DNA-binding TFs.

The number of TFs with a clear DBD is certainly a conservative estimate, as DBD-Threader could not reliably identify all Pfam domains that are known DBDs according to literature annotation. However, as we also have shown that this affects mainly the less frequently occurring DNA-binding domains, we believe that the estimate is at least close to the real value.

Mapping of PPIs and PTMs

Ravasi *et al.* tested 1222 TFs experimentally for protein–protein interactions and found 762 actual interactions for 482 TFs [17]. These interactions were included in the data set. For the mapping of PTMs, we retrieved information for each TF from the PTM-specific files from Phosphosite [26]. The distribution of PTMs is shown in Figure 4.

Based on these data sources, including the analysis of DBDs described above, we then made a final annotated set of transcription factors. The main properties are listed in Table 3, and the full table is available [see Additional file 3].

Using the annotated TFs for data analysis

We now want to illustrate how such data can be used to analyze sets of TFs. We used two main approaches. In the first approach we used properties in the TF table to split the set of TFs into subsets, and analyzed these subsets using either enrichment analysis against other properties in the TF table, or against Gene Ontology data or annotation-based property data, using GOrilla [27,28] and DAVID [29]. As a more general approach we also used external data to define subsets of TFs, and then analyzed these subsets using enrichment analysis against properties in the TF table.

Subsets analyzed with GOrilla and DAVID

Here subsets were defined based on properties in the TF table, like DNA-binding or acetylation, and these subsets were analyzed with GOrilla and DAVID, using the full set of relevant TFs as background. Selected results for GOrilla are shown in Table 4, and comprehensive results for GOrilla and DAVID are given in Table S1 and S2 [see Additional file 2].

The results show a particularly clear difference between TFs with and without a DBD. The DNA-binding TFs are enriched in sequence-specific DNA-binding, receptor properties, dimerization and core promoter interactions. The non-DNA-binding TFs are enriched in RNA-binding and cofactor activity, but also in catalytic activity, histone binding and related processes.

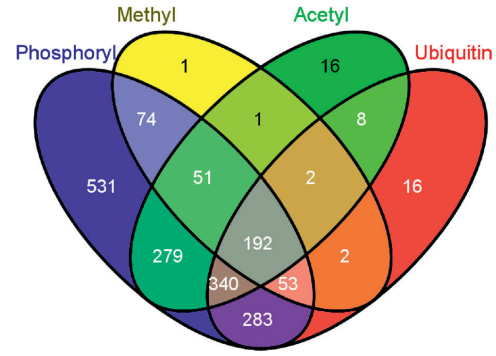


Figure 4 A Venn diagram for distribution of PTMs across TFs. The diagram shows that PTMs tend to co-occur, possibly due to experimental bias.

Table 3 Overview of TF annotation data

Information	Type	TFs with data	Positives*	Average**
Uniprot ID	protein ID	1978	1978	1
Pfam non-DBD	domain IDs	1978	753	2.16
Pfam DBD	domain IDs	1978	1225	1.33
PPI	protein IDs	1222	482	1.58
PTM - acetylation	positions	1978	884	3.55
PTM - methylation	positions	1978	376	3.22
PTM - O-GlcNAc	positions	1978	41	2.90
PTM - phosphorylation	positions	1978	1797	13.12
PTM - sumoylation	positions	1978	190	1.77
PTM - ubiquitination	positions	1978	896	4.38

*Number of TFs that actually have the property. **Average number of occurrences in the positive TFs.

This shows that the list of TFs includes some epigenetic factors. In order to verify this we compared the TF list used here to a list of epigenetic factors (F. Drabløs, unpublished data). This indicates that the list included 322 genes (16%) that also could be classified as epigenetic factors. This is probably an overestimate, as the list of epigenetic factors includes some TFs that recruit epigenetic factors. However, it confirms that subsets of genes on the list from Ravasi *et al.* are not classical TFs.

Table 4 Selected enriched terms according to GOrilla

Description		P-value	FDR q-value	Enrichment (N, B, n, b)
DNA_Binding	DNA binding	2.11E-185	1.72E-182	1.28 (1939,1475,1206,1174)
	core promoter sequence-specific DNA binding	7.87E-5	1.79E-3	1.37 (1939,60,1206,51)
	protein dimerization activity	4.00E-8	1.13E-6	1.24 (1939,254,1206,196)
Non_DNA_Binding	catalytic activity	1.07E-49	8.75E-47	2.01 (1939,305,735,232)
	RNA binding	3.95E-34	1.62E-31	2.00 (1939,222,735,168)
	transcription cofactor activity	9.56E-12	4.61E-10	1.42 (1939,359,735,193)
	histone binding	1.03E-10	3.39E-9	2.07 (1939,60,735,47)
	ubiquitin-protein transferase activity	2.29E-10	7.21E-9	2.40 (1939,33,735,30)
	methylated histone binding	3.80E-10	1.11E-8	2.54 (1939,26,735,25)
Acetylation	transcription factor binding	2.12E-6	2.17E-4	1.28 (1939,292,879,169)
	structure-specific DNA binding	2.27E-5	7.76E-4	1.38 (1939,136,879,85)
Non_Acetylation	sequence-specific DNA binding	1.36E-6	1.11E-3	1.11 (1939,887,1061,537)
Methylation	protein binding	2.67E-8	3.12E-6	1.21 (1939,1135,372,264)
	chromatin binding	3.93E-7	2.48E-5	1.62 (1939,264,372,82)
O-GlcNAc	protein binding	6.83E-6	2.80E-3	1.54 (1939,1133,41,37)
	histone deacetylase binding	2.71E-4	7.41E-2	6.31 (1939,45,41,6)
Phosphorylation	protein binding	4.93E-5	2.02E-2	1.02 (1939,1133,1782,1065)
PTM	protein binding	3.12E-6	2.55E-3	1.02 (1939,1135,1827,1093)
Sumoylation	sequence-specific DNA binding	3.00E-12	4.1E-10	1.73 (1939,617,189,104)
	core promoter binding	1.86E-7	8.03E-6	2.90 (1939,92,189,26)
	chromatin binding	1.92E-7	7.86E-6	1.98 (1939,264,189,51)
Ubiquitination	protein binding	3.71E-30	3.04E-27	1.24 (1939,1133,888,641)
	transcription cofactor activity	3.27E-8	8.12E-7	1.28 (1939,359,888,211)
Non_Ubiquitination	DNA binding	6.99E-14	5.73E-11	1.09 (1939,1473,1052,869)
PPI	transcription factor binding	1.38E-4	4.83E-2	1.31 (1203,185,475,96)

Associations between individual PTM properties

The modification of transcription factors by PTMs like phosphorylation, acetylation, methylation, ubiquitination, sumoylation and O-GlcNAc may affect their activity. It is therefore relevant to see how these modifications are correlated, and whether they are correlated with other properties. This is shown in Table 5, and in Table S5 [see Additional file 2].

The results show significant associations between most of the PTMs. It is likely that this shows an experimental bias in the data set, where TFs tested for a given PTM also are more likely to have been tested for other PTMs, thereby creating artificially strong associations. Figure 4 seems to indicate this, as for example almost all proteins that are methylated are also phosphorylated. We also see that there is in general a negative correlation between PTMs and DNA-binding properties, possibly indicating that PTMs are less important for classical TFs than for TFs involved for example in chromatin organization. This may indicate that processes at the chromatin level are more actively regulated at the PTM level than TF binding itself, which seems reasonable based on current knowledge.

Association between DNA-binding and PPI

It is relevant to look further into possible associations between DNA-binding and PPI propensity, as stabilization through PPI is a possible mechanism for stable binding despite lack of strong DBDs in TFs. As seen from Table 5, there is not any significant non-random association between having a DNA-binding domain and participating in PPI (p-value 0.343).

However, this is a rather general analysis, and it may be relevant to look closer into more specific cases, where one, both or none of the TFs have a DBD. These results are shown in Table 6. The results show that all cases are significant after Benjamini correction, in particular for cases with no DBD in any of the partners, where we see more pairs than expected. For the other two cases, where at least one TF is DNA-binding, we see fewer pairs than expected. A reasonable initial hypothesis would have been that TFs without a DBD will tend to associate with TFs with a DBD, in order to recognize regulatory regions, but this analysis indicates the opposite. The data make sense for cases where both TFs have DBD, and therefore do not need PPI to bind, but we do not have a good explanation for the other two cases, although participation in large complexes may be a possible hypothesis.

Table 5 Associations between property-based subgroups

Property pair		P-value	Benjamini	Corr.
Phosphorylation	Acetylation	1.84E-10	5.15E-09	0.190
Phosphorylation	Ubiquitination	1.94E-10	2.72E-09	0.190
DNA_Binding	Methylation	2.08E-10	1.94E-09	−0.156
Phosphorylation	Methylation	2.42E-10	1.70E-09	0.127
Methylation	Acetylation	2.78E-10	1.56E-09	0.202
Ubiquitination	Methylation	2.85E-10	1.33E-09	0.204
DNA_Binding	Ubiquitination	3.16E-10	1.26E-09	−0.280
Ubiquitination	Acetylation	3.39E-10	1.19E-09	0.289
Acetylation	Sumoylation	5.99E-09	1.86E-08	0.131
Ubiquitination	Sumoylation	4.03E-08	1.13E-07	0.124
DNA_Binding	Acetylation	6.30E-08	1.60E-07	−0.122
Methylation	O-GlcNAc	1.24E-05	2.90E-05	0.110
Phosphorylation	Sumoylation	1.51E-05	3.24E-05	0.086
Acetylation	O-GlcNAc	3.45E-04	6.91E-04	0.083
Ubiquitination	O-GlcNAc	3.82E-03	7.13E-03	0.067
PPI	Sumoylation	1.23E-02	2.16E-02	0.072
Methylation	Sumoylation	1.49E-02	2.46E-02	0.056
Phosphorylation	O-GlcNAc	2.85E-02	4.43E-02	0.046
DNA_Binding	PPI	3.43E-01	4.37E-01	−0.027

Enrichment of domains and domain pairs in PPI

PPIs are often achieved through interactions between specific domains. It is therefore interesting to see whether specific Pfam domains, or pairs of Pfam domains, are enriched in the PPI data.

As previously described there were 762 PPIs involving 482 transcription factors, and these TFs contained 518 different Pfam domains. Each Pfam domain was tested for association with PPI. This identified 73 enriched Pfam domains [see Additional file 4].

Subsequently we tested pairs of Pfam domains, rather than individual occurrences. First we tested all possible pairs for the 73 Pfam domains (see above), which identified 227 enriched pairs of Pfam domains. However, there is a risk that some interactions are significant as pairs even though they are not significant individually. We therefore relaxed the criteria so that at least one of the two Pfam domains had to be significantly associated with PPI [see Additional file 4]. In total we identified 347 pairs of Pfam domains as enriched in PPI data after Benjamini correction. However, 177 out of the 347 pairs were observed just once [see Additional file 4]. The main pairwise interactions, except for RNA polymerases and Pfam-B domains, are plotted in Figure 5. All interactions are shown in Figure S3 [see Additional file 2]. The plot

Table 6 Occurrence of DBDs in 762 PPI pairs

DBD found in	Expected	Observed	P-value	Benjamini
both TFs	255	229	0.046	4.58E-02
only one TF	371	343	0.042	4.58E-02
none TFs	135	190	7.50E-07	2.25E-06

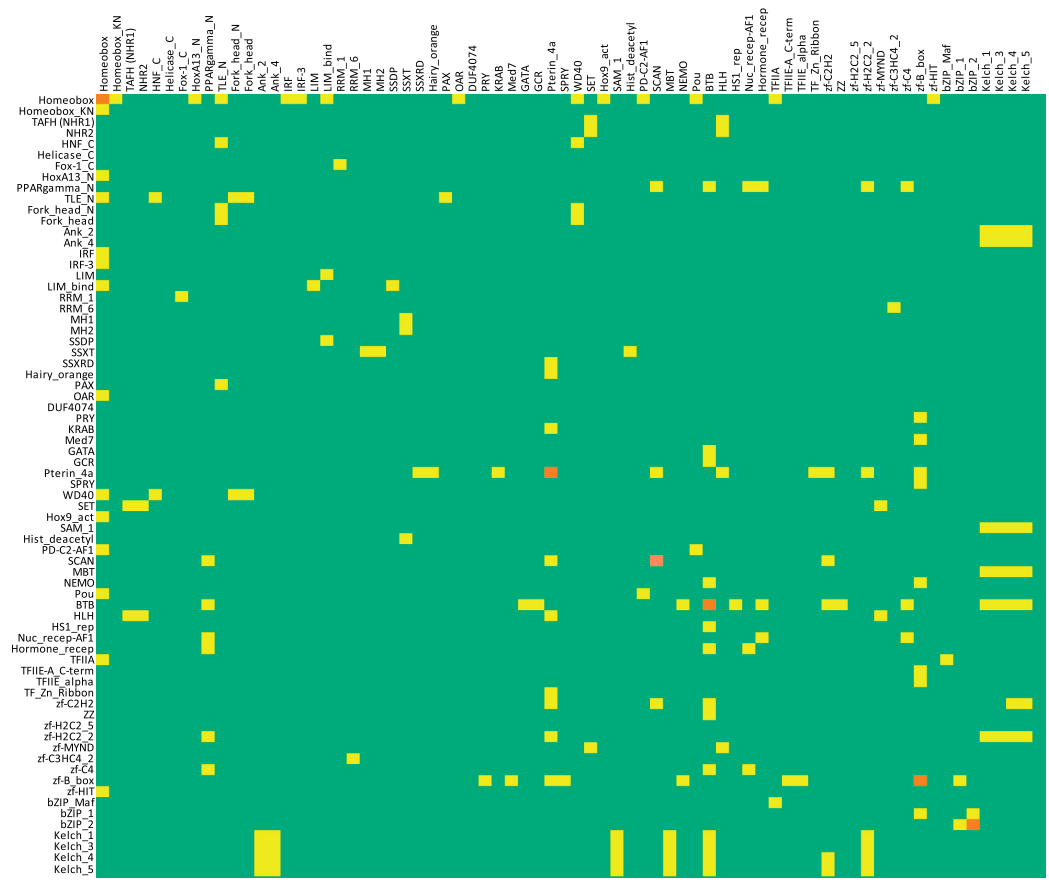


Figure 5 A matrix representation of enriched domain pairs in PPI data. Homodimers are indicated in orange. RNA polymerases and Pfam-B domains are not included; please see Figure S2 [in Additional file 2] for the full data set.

shows that the network of domains that are enriched (and possibly involved) in PPI is quite sparse. Although more than half (66%) of the domain pairs are found in pair with more than one other domain type, this is in most cases limited to two different domains, and often involve related types (like Kelch domains).

Analysis of externally defined sets of TFs

To illustrate how such annotated lists can be used to analyze data from different types of experiments, we analyzed gene lists from three recent papers. The software used for this analysis is available with the paper [Additional file 5].

A paper by Tuomela *et al.* discusses early changes in gene expression during differentiation of human Th17 cells from $CD4^+$ T-cells [34]. Expression levels were measured with microarrays, and differentially expressed genes were identified. One of the largest groups of differentially expressed genes was transcription factors. Groups of genes with similar temporal changes in expression patterns were identified by clustering into 10 groups (see the paper for details). Some of these groups showed similar general trends, like groups 1, 2 and 3 (up-regulation), 4, 5 and 6 (down-regulation), and 7, 8, 9 and 10 (no change). All the individual groups, as well as the indicated combinations, were tested for enrichment [see Additional file 6]. The results (Table 7; full results in Table S4 [see Additional file 2]) show that in particular ubiquitination is clearly enriched, in particular in the combined cluster with down-regulated expression pattern (4, 5, and 6). It may make sense that proteins of down-regulated genes are ubiquitinated, in order to speed up the process of down-regulation. It is also interesting that there is a clear depletion of DNA-binding in

Table 7 Results for TF expression changes during cell differentiation

Category*	Term	Observed	Expected	Pvalue	Benjamini	MCC
1	Ubiquitination	4	1	4.20E-02	3.36E-01	0.049
	Sumoylation	2	0	4.84E-02	1.93E-01	0.062
6	O-GlcNAc	3	0	9.69E-03	7.75E-02	0.086
	Ubiquitination	16	9	1.58E-02	6.31E-02	0.058
	Methylation	9	4	2.35E-02	6.27E-02	0.059
8	Ubiquitination	17	9	1.36E-03	1.09E-02	0.074
	PPI	10	5	2.39E-02	9.58E-02	0.070
1,2,3	PPI	9	4	1.57E-02	1.26E-01	0.072
4,5,6	Ubiquitination	43	29	1.48E-03	1.18E-02	0.074
	Methylation	21	12	1.03E-02	4.11E-02	0.061
	Sumoylation	12	6	2.98E-02	7.93E-02	0.054
	O-GlcNAc	4	1	4.53E-02	9.07E-02	0.052
7,8,9,10	DNA_Binding	28	38	7.49E-03	5.99E-02	−0.062
	Ubiquitination	38	28	1.32E-02	5.29E-02	0.058

*Indicates TFs with similar expression profiles: 1, 2, 3 - Up-regulated; 4, 5, 6 - Down-regulated; 7, 8, 9, 10 - No clear change.

genes with a stable (housekeeping-like) expression pattern. It is possible that these transcription factors rely on interaction with open chromatin initiated by other transcription factors, and are therefore less actively regulated than such key factors.

A paper by Lawrence *et al.* identified somatic point mutations in exome sequences from 4742 human cancers with matched normal-tissue samples across 21 cancer types [35]. Frequently mutated genes were identified and analyzed according to whether the gene was mainly mutated in a single cancer, or across many cancers. This made it possible to identify subsets of genes, here identified as gene set I (mainly mutated across many cancers), II (highly mutated in a few cancers), and III (highly mutated across many cancers). The last set could further be divided into IIIA and IIIB, where B consists of the genes that are most broadly mutated [see Additional file 7]. The analysis shows that many features are enriched, but often represented by a small number of genes (Table 8, full results in Table S5 [see Additional file 2]). The most significant enrichments are for PTMs. However, it is possible that this is influenced by experimental bias, as known cancer genes may have been more frequently tested for PTMs. We also see that DNA-binding again is depleted, possibly indicating that TFs with a strong and easily identified DBD are more essential to cellular function, and therefore less frequently mutated. Also some Pfam domains show a small enrichment, in particular for the SET and PHD domains. These domains are found frequently for example in members of the MLL family, which catalyze H3K4 methylation as part of a large multiprotein complex containing several chromatin remodeling factors. More than 70% of infant leukemia and approximately 10% of adult human leukemia display chromosomal translocations of the MLL (KMT2A) gene, and 450 functionally diverse MLL fusions having been identified. However, it is interesting that in all fusion proteins the C-terminal SET domain is lost and consequently they lack H3K4 methyltransferase activity [36]. The PLU-1/JARID1B is a nuclear protein which is expressed in a high proportion of breast cancers. Two PHD domains in PLU-1/JARID1B are involved in transcriptional repression. Indeed the interaction between the class II HDACs (histone deacetylase) and PLU-1/JARID1B depends on functional PHD domains, and is responsible for transcriptional repression [37].

Vaquerizas *et al.* [16] have published an analysis of 1391 manually curated sequence-specific DNA-binding transcription factors. They looked into the tissue distribution of TF expression, and identified a bi-modal distribution; 37% of the TFs showed significant expression in at least one tissue, 32% of these were expressed in most tissues, whereas the majority was expressed only in a subset (typically 1–3 tissues). We used these three subsets (general tissue distribution, specific distribution, and unknown; [see Additional file 8]) as input for analysis. The results are shown in Table 9 (full results in Table S6 [see Additional file 2]). They show an expected enrichment for DNA-binding, since this particular dataset has been selected for DNA-binding TFs. They also show a depletion of PTMs and PPIs in the set with unknown tissue distribution. This most likely indicates the same problem as before with respect to data bias; many of these TFs have been less studied, and the lack of PTMs most likely reflects a lack of experimental data, and not that they are less frequently modified. It is probably more relevant that the tissue-specific TFs are more likely to be sumoylated or be hormone receptors than the

Table 8 Selected results for TFs that are frequently mutated in cancer

Category*	Term	Observed	Expected	Pvalue	Benjamini	MCC
II + IIIAB	Acetylation	48	26	1.823E-08	1.46E-07	0.125
	Ubiquitination	47	27	2.08E-07	8.31E-07	0.118
	Methylation	26	11	1.21E-05	3.23E-05	0.110
	PF00856(SET)	6	0	1.21E-05	8.80E-03	0.164
	PF13771(zf-HC5HC2H)	4	0	4.90E-05	1.78E-02	0.175
	PF00628(PHD)	8	1	1.78E-04	2.58E-02	0.114
	Sumoylation	14	5	1.14E-03	2.28E-0	0.082
	O-GlcNAc	5	1	7.03E-03	1.12E-02	0.078
II	Acetylation	21	12	1.67E-03	1.33E-02	0.073
	Ubiquitination	20	12	6.61E-03	2.64E-02	0.063
	Methylation	11	5	1.23E-02	3.28E-02	0.062
	O-GlcNAc	3	0	1.89E-02	3.78E-02	0.073
IIIB	Sumoylation	5	1	3.51E-03	2.80E-02	0.085
I + IIIAB	Ubiquitination	32	16	2.69E-07	2.15E-06	0.114
	Acetylation	30	16	6.44E-06	2.12E-05	0.101
	Methylation	19	7	7.94E-06	2.12E-05	0.114
	PF00856(SET)	5	0	1.67E-05	5.78E-03	0.178
	PF00628(PHD)	7	1	4.70E-05	8.56E-03	0.136
	PF13771(zf-HC5HC2H)	3	0	3.17E-04	2.25E-02	0.168
	Sumoylation	10	3	1.82E-03	3.64E-03	0.082
	DNA_Binding	15	22	9.56E-03	1.53E-02	−0.061
IIIAB	Acetylation	27	14	5.47E-06	2.56E-05	0.102
	Ubiquitination	27	14	6.39E-06	2.56E-05	0.101
	PF00628(PHD)	6	0	1.82E-04	2.64E-02	0.125
	PF00439(Bromodomain)	4	0	4.78E-04	4.35E-02	0.132
	Methylation	15	6	2.79E-04	7.44E-04	0.091
	Sumoylation	9	3	2.28E-03	4.56E-03	0.081
	DNA_Binding	12	19	5.45E-03	8.71E-03	−0.065
I	Methylation	4	0	5.47E-03	4.38E-02	0.078

*Indicates TFs with similar mutation profiles: I - Mainly mutated across many cancers; II - Highly mutated in a few cancers; IIIA - Highly mutated across many cancers; IIIB - Even more highly mutated across many cancers.

Table 9 Selected results for TFs with differences in tissue specificity

Category*	Term	Observed	Expected	Pvalue	Benjamini	MCC
General	DNA_Binding	126	85	1.36E-10	1.09E-09	0.166
Specific	DNA_Binding	306	205	1.92E-10	1.53E-09	0.280
	Sumoylation	57	31	1.85E-06	7.39E-06	0.115
	PF00104(Hormone_recep)	28	7	2.32E-11	1.69E-08	0.179
	PF01352(KRAB)	16	40	9.20E-07	1.67E-04	−0.103
Unknown	DNA_Binding	702	486	2.82E-10	2.26E-09	0.459
	Ubiquitination	229	355	3.19E-10	1.28E-09	−0.263
	Methylation	105	149	1.68E-07	4.47E-07	−0.116
	PPI	146	172	1.48E-03	2.37E-03	−0.091
	PF01352(KRAB)	200	96	2.11E-10	7.66E-08	0.324

*Indicates TFs found in many tissues (general), a few tissues (specific), or unknown (due to very low or no expression).

general ones, as this may reflect mechanisms for tissue-specific regulation (see e.g. [38]). It is also interesting that the KRAB domain is depleted in the tissue-specific set, but enriched in the unknown (not expressed) set, as KRAB is a known transcriptional repressor domain [39].

Conclusions

A combination of literature-based curation and prediction methods has been used to build a comprehensive list of transcription factor properties, and this list has been applied towards investigating relationships between TF properties, TF–TF (protein–protein) interactions, and external data, and used to find significant correlations and enriched or depleted features. The results show that the comprehensive list is a useful data analysis resource for researchers working on gene regulation. However, it also shows that such analyses are easily biased by incomplete data or by how the gene sets have been selected. This mirrors to some extent the recent results by Rolland *et al.* [40], where they identified a strong bias in existing PPI data towards well-studied proteins.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files, or were downloaded from open sources as shown in Methods.

Additional files

Additional file 1: Manually checked Pfam domains for DNA-binding.

Additional file 2: Figure S1. Distribution of Pfam domain types. Domains with less than 5 occurrences are grouped under "Others". **Figure S2.** Plot of data used to predict new DNA-binding domain types. Only domains that overlap with DBD-Threader predictions are shown. The classification line for SVM-based classification with a linear kernel is indicated. **Figure S3.** Enriched PPI domain pairs. **Table S1.** Selected enriched terms according to GOrilla. **Table S2.** Selected enriched terms according to DAVID. **Table S3.** Associations between property-based subgroups. **Table S4.** Output from enrichment analysis of data from Tuomela et al. **Table S5.** Output from enrichment analysis of data from Lawrence et al. **Table S6.** Output from enrichment analysis of data from Vaquerizas et al.

Additional file 3: Main table of transcription factor properties.

Additional file 4: Enriched domains and domain pairs in PPI.

Additional file 5: Software for data analysis.

Additional file 6: Data from Tuomela et al. for Table S4.

Additional file 7: Data from Lawrence *et al.* for Table S5.

Additional file 8: Data from Vaquerizas *et al.* for Table S6.

Abbreviations

TF: Transcription factor; DBD: DNA-binding domain; PPI: Protein–protein interaction; TFBS: Transcription factor binding site; PTM: Post-translational modifications; PWM: Position weight matrix; PFM: Position frequency matrix; HMM: Hidden Markov model; FN: False negative; FP: False positive; TP: True positive; TN: True negative; Sn: Sensitivity; Sp: Specificity; PPV: Positive predictive value; MCC: Matthews correlation coefficient; Bp: Base pair; SVM: Support vector machine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SB and RE collected all data, performed the analysis and drafted the initial manuscript. FD initiated and supervised the project. All authors contributed to and approved the final manuscript.

Acknowledgements

This work was supported by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU).

Author details

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, P.O. Box 8905, NO-7491 Trondheim, Norway. [2]St. Olavs Hospital, NO-7006 Trondheim, Norway.

Received: 26 September 2014 Accepted: 2 March 2015
Published online: 14 March 2015

References

1. Latchman DS. Transcription factors: an overview. Int J Biochem Cell Biol. 1997;29(12):1305–12.
2. Kerschner JL, Gosalia N, Leir SH, Harris A. Chromatin remodeling mediated by the FOXA1/A2 transcription factors activates CFTR expression in intestinal epithelial cells. Epigenetics. 2014;9(4):557–65.
3. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: A structural analysis. J Mol Biol. 1999;287(5):877–96.
4. Hughes TR. A handbook of transcription factors. Dordrecht Heidelberg London New York: Springer; 2011.
5. Reddy DA, Prasad BVLS, Mitra CK. Functional classification of transcription factor binding sites: Information content as a metric. J Integr Bioinform. 2006;3(1):20.
6. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011;25(21):2227–41.
7. Frietze S, Farnham PJ. Transcription factor effector domains. Subcell Biochem. 2011;52:261–77.
8. Gao J, Li WX, Feng SQ, Yuan YS, Wan DF, Han W, et al. A protein-protein interaction network of transcription factors acting during liver cell proliferation. Genomics. 2008;91(4):347–55.
9. Kho Y, Kim SC, Jiang C, Barma D, Kwon SW, Cheng J, et al. A tagging-via-substrate technology for detection and proteomics of farnesylated proteins. Proc Natl Acad Sci U S A. 2004;101(34):12479–84.
10. Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol. 2003;21(3):255–61.
11. Beg AA, Scheiffele P. Neuroscience. SUMO wrestles the synapse. Science (New York, NY). 2006;311(5763):962–3.
12. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014;42(1):D142–7.
13. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):D91–4.
14. Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, et al. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. 2012;40(Database issue):D144–9.
15. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, et al. TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. 2009;10(3):R29.
16. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10(4):252–63.
17. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010;140(5):744–52.
18. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(D1):D1079–1085.
19. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42:D191–8.

20. Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database. 2011;2011:bar049.
21. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222–30.
22. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
23. Gao M, Skolnick J. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput Biol. 2009;5(11):e1000567.
24. Lou W, Wang X, Chen F, Chen Y, Jiang B, Zhang H. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. PLoS One. 2014;9(1):e86703.
25. scikit-learn. [http://scikit-learn.org/].
26. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012;40(Database issue):D261–70.
27. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinf. 2009;10:48.
28. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 2007;3(3):e39.
29. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
30. Python. [https://www.python.org/].
31. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England). 2009;25(11):1422–3.
32. Fisher's Exact Test. [https://pypi.python.org/pypi/fisher/].
33. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol. 2009;10(2):207.
34. Tuomela S, Salo V, Tripathi SK, Chen Z, Laurila K, Gupta B, et al. Identification of early gene expression changes during human Th17 cell differentiation. Blood. 2012;119(23):e151–60.
35. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495–501.
36. Sarris M, Nikolaou K, Talianidis I. Context-specific regulation of cancer epigenomes by histone and transcription factor methylation. Oncogene. 2014;33(10):1207–17.
37. Barrett A, Santangelo S, Tan K, Catchpole S, Roberts K, Spencer-Dene B, et al. Breast cancer associated transcriptional repressor PLU-1/JARID1B interacts directly with histone deacetylases. Int J Cancer. 2007;121(2):265–75.
38. Ward JD, Yamamoto KR, Asahina M. SUMO as a nuclear hormone receptor effector: New insights into combinatorial transcriptional regulation. Worm. 2014;3:e29317.
39. Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, Rauscher 3rd FJ. Kruppel-associated boxes are potent transcriptional repression domains. Proc Natl Acad Sci U S A. 1994;91(10):4509–13.
40. Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–26.

Paper II

Functional classification of human transcription factors based on structural properties

Rezvan Ehsani[1,2,+], Shahram Bahrami[1,3,+], Finn Drabløs[1,*]

[1] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

[2] Department of Mathematics, University of Zabol, Zabol, Iran

[3] St. Olavs Hospital, Trondheim University Hospital, NO-7006 Trondheim, Norway

+ These authors contributed equally

* Corresponding author

Email addresses:

Rezvan Ehsani - <rezvan.ehsani@ntnu.no>

Shahram Bahrami - <shahram.bahrami@ntnu.no>

Finn Drabløs - <finn.drablos@ntnu.no>

Address for correspondence:

Finn Drabløs

Department of Cancer Research and Molecular Medicine

Norwegian University of Science and Technology

P.O. Box 8905

NO-7491 Trondheim

Norway

ABSTRACT

Background: Transcription factors are key proteins in the regulation of gene transcription. An important step in this process is the opening of chromatin in order to make genomic regions available for transcription. A subset of transcription factors has previously been classified into Pioneers, Settlers and Migrants with respect to their role in this process. This can be a useful classification for understanding the different steps in gene regulation, and we therefore wanted to use machine learning to expand the set of classified transcription factors in order to include as many known factors as possible.

Results: We have used a well-annotated dataset of 1978 transcription factors as input to supervised machine learning methods in order to classify transcription factors with respect to their role in the process of chromatin opening. We then used this classification to investigate associated properties and functions of the transcription factors in each class, including an analysis of interaction data for transcription factors based on DNA co-binding and protein-protein interactions. We also used the classification to analyze a previously published set of gene lists associated with a time course experiment on cell differentiation.

Conclusions: The results showed that the classification of transcription factors with respect to their role in chromatin opening largely was determined by how they bind to DNA. Each subclass of transcription factors was enriched for properties that seemed to characterize the subclass relative to its role in gene regulation, with very general functions for Pioneers, whereas Migrants to a larger extent were associated with specific processes. Further analysis showed that the expanded classification is a useful resource for analyzing other datasets on transcription factors and their role in gene regulation. The analysis of transcription factor interaction data showed complementary differences between the subclasses, where Pioneers often interact with other transcription factors through DNA co-binding, whereas Migrants to a larger extent use protein-protein interactions. The analysis of time course data on cell differentiation indicated a shift in the regulatory program associated with Pioneer transcription factors during differentiation.

Keywords: Transcription factors; Chromatin opening; Machine learning; Classification

BACKGROUND

Cells recognize and respond to internal and external signals, often leading to changes in the transcription level of specific genes. Transcriptional regulatory systems play a key role in many biological processes, such as cell cycle progression, maintenance of intracellular metabolism, physiological balance, and cellular differentiation in developmental time courses [1, 2]. The regulatory system for transcription involves several proteins, in particular transcription factors (TFs), which can coordinate a diversity of regulatory processes. Many diseases arise from errors in the regulatory system for transcription; TFs are overrepresented among oncogenes [3], and a third of human developmental disorders have been related to dysfunctional TFs [4]. However, alterations in the activity and regulatory pathway of TFs are also likely to be a source for phenotypic diversity and evolutionary adaptation [5-7].

Most TFs bind to DNA by recognizing specific DNA sub-sequences known as transcription factor binding sites (TFBSs), and thereby they control the transcription of nearby genes through their promoters, or more distant genes through enhancers. However, it has been realized that binding of TFs to TFBSs is not enough to fully explain the regulatory program of gene expression [8]. The set of cis-regulatory regions (promoters, enhancers) is identical at the DNA sequence level in all cell types of a given species. Therefore the transcriptional program specific to each cell type must be the result of the set of TFs expressed in that cell type, and how genes are selected for transcriptional activation or repression. The same TFs can be expressed at the same rate in different cell types, but may have separate binding sites, as TF function and regulatory pathways also depend upon chromatin structure and epigenetic modifications [9].

TFs can be classified according to functional properties. Sherwood *et al.* (2014) introduced PIQ (protein interaction quantitation), a computational method for modeling the magnitude and shape of genome-wide DNase I hypersensitivity profiles used for identification of transcription factor binding sites [10]. They identified binding sites for more than 700 transcription factors from an experiment using DNase I hypersensitivity analysis followed by sequencing, and used the data to classify transcription factors into three groups; Pioneers, Settlers and Migrants. Pioneer TFs are distinguished by their ability to bind to DNA target sites, even in inaccessible regions, and were found bound to chromatin before activation of enhancers and gene expression modulation. The binding of Settler TFs is dependent on the openness of chromatin at their binding sites. They almost always bind to sites matching their DNA-binding motif, but they do not enable binding to inaccessible DNA sites [10, 11]. And finally Migrant TFs only bind to a subset of their target sites, even in accessible DNA [10].

TFs can also be classified based on structural properties, and the most common classifications are based on the structure of their DNA-binding domains (DBDs) [12]. In some instances the structural classification may also indicate the function of TFs. For example, TFs with homeo-domain are often associated with developmental processes, and those with a “winged” helix-turn-helix (HTH) motif are frequently associated with the interferon regulatory factor family and triggering of immune responses against viral infections [12].

There are several TF databases. One of the most frequently used databases is TRANSFAC (TRANScription FACtor database), a manually curated database of eukaryotic transcription factors with their TFBSs and DNA binding profiles. The content of the database is suitable for a comprehensive analysis of genomic sequences for potential TFBSs. The database lists the target sequences and the regulated genes for each TF, and can also be used for benchmarking of TFBS recognition tools or as training sets for new TFSB recognition algorithms. TF classifications based on properties of the DNA-binding domains may for example be used to analyze data on the regulatory function of these TFs [13]. A related TF database is JASPAR, which is an open-access database of binding site profiles, based on position frequency matrix (PFM) profiles derived from literature, including chromatin immunoprecipitation and sequencing (ChIP-seq) experiments. However, except for the DNA motif specificity such databases contain very limited information about other TF properties, although they have links to more general protein databases with additional information [14, 15].

Wingender *et al*. (2013) have made a comprehensive classification of 1558 human TFs based on a hierarchy of general topology (Superclass), similar structures of the DBD (Class), sequence and functional similarities (Family), sequence-based subgroupings (Subfamily), TF gene (Genus), and TF polypeptide (Factor 'species'), and this classification is known as TFClass. TFs are classified according to this six-level classification scheme, where four levels are abstractions according to different criteria, while the fifth level shows the TF genes, and the sixth level individual gene products. They collected and curated 71 animal TF families. Altogether, ten superclasses have been identified, comprising of 40 classes and 111 families [16].

In a previous paper we presented a comprehensive list of properties for 1978 human TFs. We identified 1225 DNA binding TFs, based on existing annotation of Pfam domains and identification of additional Pfam DBDs. Annotated properties included DBDs, protein–protein interactions, and post-translational modifications. The paper demonstrated how such a resource can be used to identify properties that are enriched in a set of TFs [17].

1175 TFs in our collection are also classified in the TFClass classification. But the fraction of our collection that could be mapped to TF function as defined by Sherwood *et al*. was much smaller (459 TFs). We therefore decided to predict the classification of TFs according to Sherwood *et al.* by using our set of properties, including the TFClass classification, as a feature vector. We used well-known methods in machine learning to design, select, and evaluate classifiers and feature vectors [18-21]. We then used this to predict TF function for TFs not classified by Sherwood *et al.*, and analyzed the result. We used the full set of functionally classified TFs to analyze data on TF-TF interactions, and also to analyze a time course data set on cell differentiation.

RESULTS AND DISCUSSION

The transcription factor dataset

We used the comprehensive collection of TF properties from our previous work on 1978 TFs [17]. This included information on DNA binding domains (DBDs), protein–protein interactions (PPIs), and post-translational modifications (PTMs). The information on DNA binding domains was based on Pfam annotation and literature, plus a DBD prediction method for identification of additional DNA-binding Pfam-domains [22, 23]. The original list of 1978 human transcription factors was taken from Ravasi *et al.*, where they generated experimental data on PPIs to build an atlas of combinatorial regulation [24], and this information on PPIs was included in our data set. Finally we added information about PTMs from Phosphosite [25].

We extended the initial annotation by mapping data on TF classification (TFClass), and on TF function for the subset of the TFs analyzed by Sherwood *et al*., in order to train prediction methods for TF function. We used the set of 1558 TFClass TFs by Wingender *et al.* [16] ("TF class"), and 1175 of these were also found in our set of 1978 TFs. We then used the set of TFs classified according to chromatin activity by Sherwood *et al.* [10] ("TF function"). We could identify 459 of these TFs in our database, and 457 of these had intersection with the 1175 TFs with TFClass annotation. These 457 TFs included 45 TFs with function as Pioneers, 47 as Settlers, and 365 as Migrants.

Transcription factor properties

We initially tested a large set of available properties in feature vectors for TF classification: TFClass, frequent Pfam domains, DNA binding (0/1), number of DBDs, PPI (0/1), number of PPIs, PTMs (generally and individually), and number of positions for phosphorylation. Since the properties initially were a mixture of quantitative and qualitative features (e.g. having a specific Pfam domain (yes/no) versus the number of phosphorylation sites (0, 1, 2, …)), they were converted to a more consistent binary representation before analysis, as described below and in Table 1.

(Table 1)

Encoding of properties for TFs

Encoding of TFClass (TF_Class)

TFClass uses a hierarchical classification, with 10 superclasses at the top level, and a varying number of classes, families, subfamilies etc. below that. We encoded the superclasses (10) and classes (37 in total, ignoring 3 classes with no overlap with our set of TFs) as a 47 bit binary vector, with 1 for the corresponding superclass and class, and 0 elsewhere. This is a reasonable encoding because it will give a Hamming distance of 2 between TFs belonging to different classes within the same superclass, whereas it will be 4 for TFs belonging to different superclasses. Thereby TFs from the same superclass (but different class) will be more similar than TFs from different superclasses. The largest class was class 2.3, the C2H2

zinc finger factors with 475 TFs (the second largest was the homeodomain (3.1) with 199 TFs). To get more balanced subset sizes for the feature vector we extended this into four subclasses (families 2.3.1-4, as family 2.3.5 did not have any overlap with our data).

Encoding of frequent Pfam domains (PD)

Transcription factors are typically modular in structure, and will often contain effector domains, one or more DNA-binding domains and other domain types. Type and frequency of domains may reflect the function of a transcription factor.

There were twenty domains with occurrence frequency of more than 20 in the set of TFs mapped to TFClass (see Table S1 in Additional File 1). We encoded this as a 20 bit binary vector with each bit corresponding to one of the frequent Pfam domains.

Encoding of DNA binding, PPI, and PTM (DBD, PPI, and PTM)

For encoding of DNA binding, PPI, and PTM as TF properties we encoded each of these properties individually as bits in a binary feature vector, indicating whether it had this property or not (1/0), without taking the number of occurrences into account, e.g. whether it is known to have a PTM or not, and not type or frequency. However, see below for a more detailed encoding.

Encoding of individual PTMs (Ind_PTM)

There were six different types of PTMs annotated in our TF collection; phosphorylation, acetylation, methylation, ubiqutination, sumoylation, and O-GlcNAc. We encoded this for each TF as bits in a binary vector of length six.

Encoding of number DBDs, PPI interactions, and phosphorylations (N_DBD, N_PPI, and N_PhS)

We extracted the number of DBDs for each TF from our collection. We also used the BioGRID database [26] to extract the number of PPIs for each TF, and added this to our set of properties. From Phosphosite we used the number of sites for each individual modification. We encoded each TF in binary as [1 1] if the number of sites (e.g. for phosphorylations) was higher than the average, [1 0] if it was between one and the average, and [0 0] otherwise. This was done for the number of DBDs (average = 4), the number of PPIs (average = 9), and the number of phosphorylation sites (average = 14). The average for other PTMs was less than 1 and was therefore not considered for extended encoding. This can be seen as a reduced resolution encoding of counts (zero, below average, above average), which is more robust against non-relevant variation than a direct binary encoding of the individual counts.

Encoding of the number of frequent zinc finger domains (N_ZFD)

The zinc finger domains are very frequent in TFs, and may therefore require special treatment to get good classification. The Pfam domains zf_C2H2 and zf_H2C2_2 had the highest frequency among zinc fingers. These domains were therefore encoded for the TFs as [1 1] if

the TF had more than three of these zinc finger domains, [1 0] if it had between one to three of these domains, and as [0 0] if had none.

The general classification strategy

The main functional TF classes

In the functional classification on chromatin activity there are three main functions; Pioneers, Settlers, and Migrants. An initial enrichment analysis based on DAVID [27] indicated a functional and structural difference between the Migrants with negative chromatin opening index and the Migrants with positive chromatin opening index (see Table S2 in Additional File 1). For classification we therefore considered them separately, and divided our database into four functions: Pioneers, Settlers, positive Migrants and negative Migrants (Figure 1). Positive and negative Migrants are in this paper sometimes annotated as Migrants+ and Migrants- (or M+ and M-), respectively.

(Figure 1)

Multiclass classification

The functional classification of additional human TFs is a multiclass classification problem, i.e. classification of patterns into more than two classes. Some classification algorithms are binary algorithms that can be adapted to multiclass classification, whereas other classification algorithms can handle more than two classes by design. There are general strategies for handling the problem of multiclass classification as a binary classification problem [28], and we used a well-known one-vs-rest strategy, which involves training a single classifier per class, with the patterns of that class as positive patterns and all other patterns as negatives. However, this strategy requires that the base classifier produces a real-valued confidence score for its decision, rather than just a class label, as discrete class labels alone can result in ambiguities, where multiple classes are predicted for a single sample [28]. We used four different cases: Pioneers vs Rest, Settlers vs Rest, positive Migrants vs Rest, and negative Migrants vs Rest.

Handling imbalanced data

Any data set that shows an unequal distribution between its classes can be considered as an imbalanced database. Studies have shown that for several base classifiers, a balanced data set improves the overall classification performance, compared to an imbalanced data set [29, 30]. Using sampling methods on an imbalanced data set, in order to make a balanced one, will therefore normally improve the performance [31].

In this paper we used random under-sampling on training data, without replacement [31]. Specifically, in the Pioneers vs Rest case we randomly split the Migrants and Settlers into 9 subclasses, making 9 different cases, each of them balanced (see Table 2). Random splitting in the Settlers vs Rest case was handled in the same way. For the positive Migrants vs Rest case we randomly split the Rest case into 5 subclasses. For the negative Migrants vs Rest case

we randomly split the negative Migrants case into 2 subclasses, and used all Pioneers, Settlers, and positive Migrants as the Rest class.

(Table 2)

Evaluation measures

We used several common evaluation measures to evaluate the performance of the classifiers, including: precision (positive prediction value or PPV), recall (sensitivity or SN), F-score, MCC (Matthews's correlation coefficient), and AUC (area under curve for receiver operating characteristics (ROC)), with the following formulas:

$$PPV = p = \frac{TP}{(TP + FP)}$$

$$SN = r = \frac{TP}{(TP + FN)}$$

$$F\text{-}score = \frac{2TP}{(2TP + FP + FN)}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

In these formulas, TP (true positive) and TN (true negative) correspond to TFs correctly predicted as belonging to the specific class or the "Rest" class, respectively, FP (false positive) denote cases where TFs from the Rest class where predicted as belonging to the specific class, and FN (false negative) denote cases where TFs from the specific class were predicted to the Rest class. The MCC measure ranges between -1 and 1, where -1 corresponds to all predictions being incorrect, 0 to random predictions, and 1 to all predictions being correct.

The F-score can be explained as a harmonic mean of the precision and recall. The AUC is the area under a receiver operating characteristic (ROC). The ROC curve is a graphical plot which shows the performance of a binary classifier system as its discrimination threshold is changed. It is made by using the fraction of true positives out of the total actual positives (true positive rate (TPR), or SN) vs the fraction of false positives out of the total actual negatives (false positive rate (FPR)), at various threshold settings.

The performance of classifiers was tested using n-fold bootstrap cross-validation with multiple runs on all possible rebalanced TF subsets on TF function, and for each individual property [32]. In each fold, TFs were randomly sampled as two separate sets: one subset to test the prediction model (test set) and the rest of the TFs to establish the model (training set). The precision, recall, F-score, MCC, and AUC were computed for each run and then averaged over runs for each classifier. It was applied for the four classification cases separately (Pioneers vs Rest, Settlers vs Rest, positive Migrants vs Rest, and negative Migrants vs Rest).

The most frequently used statistical tests to determine significant differences between two machine learning algorithms are the t-test and the Wilcoxon test [19]. The t-test is a

parametric one and requires that the necessary conditions for using it are true, i.e. independence, normality, and heteroscedasticity. This is not the case in the majority of experiments in machine learning [33]. Thus, we investigated the statistical significance of the differences on performance using the nonparametric Wilcoxon test; we kept the result of the AUC measure for each fold and each classifier, and then compared them using Wilcoxon [19].

After identification of the locally best classifier we evaluated the performance on the properties using the same process as above (bootstrap cross-validation) on the individual properties. Again the performance measures were computed for each fold and then averaged on runs for the four classification cases separately.

The performance was evaluated by a forward best-first search on the list of properties for the four classification cases separately. We executed the runs of each cross-validation and used the average AUC to rank the properties. We started with the property giving the largest AUC, and at each step added the property (among the remaining properties) which results in the best average AUC [34]. We also investigated the statistical significance of the differences of the average AUC in each step after adding a new property, using the paired Wilcoxon tests.

Selection of classifier

Several classifiers including random forest (RF), Support Vector Classification (SVC) with different kernels (linear, radial basis function (RBF), polynomial), k-nearest neighbor (kNN) and Gaussian naïve Bayes (GNB) were applied to each of the individual properties to identify classifiers with good performance.

RF is one of the most successful ensemble learning techniques in machine learning and bioinformatics for high-dimensional classification. The RF algorithm makes a large number of individual decision tree classifiers (i.e. a forest) where each tree gives a classification, and the final classification is based on the votes over all the trees in the forest. The rule to generate a tree is through splits at each node based on the yes and no answer of the predictors. The split selection is performed by using decrease of Gini impurity in each step, where Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

In addition to the RF classifier, SVC with different kernels (linear, RBF, polynomial), kNN, and GNB were initially tested in order to find a good classifier. Also additional approaches were tested, like AdaBoost, but they did not improve the performance, and only results for the methods listed above are shown here. Some of the methods, in particular kNN and SVC, were first applied to individual properties. The kNN methods need a specification of the number of neighbors, and the SVC requires parameterization of the complexity constant C and the kernel function. The number of neighbors for kNN was limited to the set $\{3, 5, 7, 9\}$ [35]. The kNN was performed over all the allowable number of neighbors and the one that had the highest AUC score (see below) was kept. For SVC we considered two kinds of kernel; RBF $K(x_i, x_j) = \exp(-\gamma \|(x_i - x_j)\|^2)$ where γ is the width of the RBF function, and polynomial

$K(x_i, x_j) = (x_i . x_j)^d$ where d is the degree. A grid search was performed to optimize the parameters of support vector machine (SVM) classifiers. For the RBF kernel, C = $\{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$ and $\gamma = \{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$ and for polynomial kernel C = $\{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$ and $d = \{2, 3\}$.

We applied five-fold bootstrap cross validation with ten runs (10×5) on the rebalanced training set. In each fold, 80% of data was considered as training set and 20% as test set by bootstrap sampling. The average AUC score was used as criterion. The RF classifier was selected for further analysis as it had the best performance compared to the other classifiers; together with the SVC-RBF classifier it was always ranked as one of the best classifiers in the different cases, but with a higher average AUC score than the SVC-RBF classifier (see Table S3 in Additional File 1 for detailed results).

Table 3 shows the AUC score for the individual properties in the four classification cases for the RF classifier (complete measures of performance on the individual properties are shown in Table S4 in Additional File 1). The results showed that the TF_Class and PD (*i.e.*, Pfam domains) properties gave the highest AUC performance for each of four binary classification cases. The N_ZFD, PPI and Ind_PTM properties gave the next highest performance for the Pioneers vs Rest, the Settlers vs Rest and Migrants vs Rest cases respectively. The remaining properties gave roughly the same (and lower) performance, often close to random classification (AUC 0.5), even though all the properties were initially selected as potentially relevant for TF function.

(Table 3)

The RF classifier can estimate a feature importance score, which also has been included in Table 3. This score is highly correlated with the AUC score, and highlights the same properties. The most important properties are in particular TF_class and PD, followed by Ind_PTM, whereas N_PPI and N_ZFD are important in specific cases.

To find the best list of properties for the final model we used the forward best-first algorithm with the RF classifier, which should be robust with respect to redundancy between properties. We started with the property that gave the highest AUC score in each case. The process was continued by adding all remaining properties separately to the pervious step, and selecting the property that produced the highest AUC score on a five-fold cross validation over ten runs. Figure 2 shows the general improvement of AUC while stepwise increasing the number of properties.

(Figure 2)

Table 4 shows the significance of change in AUC when adding properties. The results show that a list including the TF_Class, PD and N_PPI properties was the best list for the Pioneers vs Rest case; a list including the PD, N_PPI and TF_Class properties was best for the Settlers vs Rest case; a list including the PD, TF_Class, and N_ZFD properties was best for the positive Migrants vs Rest case; and finally for the negative Migrants vs Rest the best list included the TF_Class and PD properties. Interestingly the Ind_PTM property was not

selected despite a good ranking based on the feature importance (see Table 3), indicating that the information may be redundant when other properties are included. The result shows that in particular TF classification and frequent Pfam domains are important features for correct prediction of functional roles. The individual subsets of properties that are listed above were used for the final classification.

(Table 4)

Final classification

The random forest classifier was used for the final classification, with the balanced set of 457 TFs with known chromatin function as training data. 718 TFs were predicted over all splits of balanced data as shown in Table 2, and the optimal list of properties was used for each case separately. This means that each TF was predicted 9 times for Pioneer vs Rest and Settler vs Rest, 5 times for Positive Migrant vs Rest and 2 times for Negative Migrant vs Rest. Average probability was then computed for each TF for each case separately. The highest probability was used for final classification.

By this strategy we identified 289 TFs as Pioneers, 169 TFs as Settlers, 211 TFs as positive Migrants, and 49 TFs as negative Migrants (see Table S5 in Additional File 2). The average and margin probability from the adaptive boosting on random forest were used to evaluate the quality of the final prediction. An overview of the classification result is given in Table 5.

(Table 5)

Discussion of the classification result

Several interesting observations can be made from the classification results**.** It is clear that the original dataset by Sherwood *et al.* is highly biased, in particular with respect to TFClasses 2.3.3 (*Zinc-coordinating DNA-binding domains* / *C2H2 zinc finger factors* / *More than 3 adjacent zinc finger factors*) and 2.3.4 (… / … / *Factors with multiple dispersed zinc fingers*), where the training set contains only 3.7% and 4.9% of the TFs from each class, respectively. Still the classification is quite robust, in particular for 2.3.3 where the average p-value is 0.86 and margin is 0.21. This means that the classification presented here provides an important extension of the initial functional classification.

The results show clearly that the functional role of a TF to a large extent is determined by how it binds DNA, as the main properties for successful classification are TFClass and Pfam domains. It is also clear that although TFClass is very important for the classification, it is not sufficient by itself, as other properties, in particular Pfam, provide complementary information. This is seen for example in TFClass 2.3.2 (… / … / *Other factors with up to three adjacent zinc fingers*), with 0 cases in the training set, where 15 cases in the full set still could be classified as Pioneers (12) or positive Migrants (3) with reasonable performance (average p-value 0.77, margin 0.20). A similar example is TFClass 3.7 (*Helix-turn-helix domains* / *ARID domain factors*). However, the number of TFs in classes with no cases in the training set is very low (76 TFs in total), and in general there is a good distribution of cases in the training set. This is reflected in the classification result as several cases of high average p-

values. An example is TFClass 3.5 (… / *Tryptophan cluster factors*), with an average p-value of 0.94 and a margin of 0.26. This indicates a quite reliable classification.

However, the classification is clearly not perfect, with an AUC score in the range of 0.82-0.92. It is difficult to say whether the main reason is imperfect training set, key features lacking in the property set, or non-optimal classifiers. However, it is clear that the best performance is seen for the negative Migrants, which also represent the most well-defined subset (see Figure 1). This may indicate that the initial classification by Sheerwood *et al.* can be improved. However, this aspect has not been investigated further.

Analyses based on TF functions

We then used the full set of TFs that could be classified as Pioneers, Settlers or Migrants for three types of data analysis; enrichment analysis on TF functional classes, analysis of properties associated with TF-TF and TF-DNA interactions for the individual TF functional classes, and finally a time course experiments where several TFs show changes in expression level.

General properties of the TF classes

Each functional class was analyzed for enriched properties using DAVID [27] and GOrilla [36], using the 1175 TFs with functional classification as background. The set of unclassified TFs was analyzed separately with DAVID to test for any potential bias, using the full set of 1978 TFs as background.

For DAVID we looked in particular into the Functional Annotation Clustering output, which groups together associated terms from different annotation sources into functional clusters, based on enrichment (see Table S6 in Additional File 3). For the unclassified TFs the most enriched terms where “chromatin modification” and similar terms, followed by for example “protein complex assembly”, “transcription initiation”, “RNA processing” and “DNA repair”. This seems to indicate that the unclassified TFs were not enriched for regulatory TFs, but rather general TFs and proteins associated with TFs and other regulatory functions. There was no strong indication of any problematic bias caused by the unclassified TFs.

For Pioneers all the most enriched clusters were related to DNA-binding domains and their properties, such as “zinc fingers”, “metal binding”, KRAB, BTB, ETS etc. More functional terms like “cell cycle” or “cancer” were found only at quite low enrichment. This was confirmed by GOrilla (see Table S7 in Additional File 1), where the most enriched term was “metal ion binding”, and no terms related to function were enriched. For Settlers the picture was similar, with highest enrichment for domains like HLH, PAS, TBOX and bZIP, and this was confirmed by GOrilla, where the most enriched term was “protein dimerization activity”. However, in DAVID also more functional terms like “tissue morphogenesis” and “response to stimulus” were clearly enriched. This trend was even clearer for the positive Migrants. Although terms like “zinc finger” and “metal ion binding” were strongly enriched, so were also terms like “ligand binding”, “hormone receptor”, “lipid binding” and “signaling pathway”, and for GOrilla the most enriched term was “receptor activity”. A similar picture

was seen in negative Migrants. Here “fork head” and POV was strongly enriched, but for example “cell motility”, “cell migration”, “cell morphogenesis” and “axonogenesis” showed a similar enrichment. For GOrilla “chromatin binding” was the most enriched term.

This seems to indicate functional roles of these classes of TFs that are consistent with what previously has been assumed, where Pioneers may have a very general role in initiating gene regulation, independent of specific biological processes. Then Settlers may be somewhat closer to biological process, whereas positive and negative Migrants are even more closely linked to specific processes, in particular signaling and differentiation, respectively.

The fact that similar DBDs tend to be associated with the same functional class may be consistent with a hypothesis suggesting that TFs from different functional classes bind to cis-regulatory regions in a hierarchical process, rather than by competing for the same binding site(s) [10].

Properties related to TF binding and interactions

Binding of several TFs to a regulatory region is an important process for gene activation. It is therefore relevant to look into the interactions between TFs. Such interactions can take place either because the TFs tend to bind to neighboring binding sites in DNA, or because they tend to interact through protein-protein interactions, or both, and it seems likely that the relative importance of this may differ between TF classes. We therefore compared data from four different sources of information on TF-TF interactions. Jolma *et al.* [37] used SELEX with a two-step affinity purification to map TF-TF-DNA interactions, showing that interactions between TFs were predominately mediated by DNA. They identified 315 TF-TF pairs showing cooperative binding (out of 9,400 potential pairs), and we used these pairs to identify significant ($p<0.05$) enrichment (or depletion) for TF-TF interactions within and between the different TF classes, using a 2x2 matrix for statistical analysis (*i.e.*, testing for each possible pair of TFs whether the pair represented a specific combination of TF functions, and whether there was a known interaction between the pair of TFs). The results (and results for the subsequent analyses) are illustrated in Figure 3. Further, we used results from the ENCODE Consortium [38], where they used ChIP-seq data to identify pairs of TFs that tend to be co-located to the same genomic regions. For PPIs we used data from Human Protein Reference Database (HPRD) [39], which includes data on pairwise PPIs for a large number of proteins, including TFs. We also used data from Ravasi *et al.* [24], where they used a M2H system to systematically screen for PPIs between TFs. This should represent a more coherent dataset than the collection in HPRD, but limited to the experimental conditions used in the M2H system.

(Figure 3)

As expected, the results are somewhat noisy, as in particular PPI data are known to be affected by large fractions of false positives, e.g. due to non-specific binding. However, the results are fairly consistent within the two main interaction types (TF co-localization and PPI), and the overall trend is also quite clear; TF co-localization and PPI are to some extent mutually exclusive features. Whereas Pioneers and Settlers tend to be enriched for TF-TF-

DNA interactions and depleted for PPI, the Migrants (and in particular positive Migrants) are enriched for PPI and depleted for TF-TF-DNA.

In order to explain this observation we looked at three properties that may influence TF-binding to DNA; the number of DNA-binding domains, the GC-content of the binding site, and the Information Content (IC) of the binding site motif. In all three cases we split the set of TFs into two, based on a suitable cutoff value for the relevant property, and tested each TF-class for enrichment or depletion. The results are shown in Table 6.

(Table 6)

For the number of TFs with more than one DBD, we saw a very significant enrichment in Pioneers, and a clear depletion in Settlers and negative Migrants, which means that Pioneers often will have a very strong and specific binding, compared to the other TFs. With respect to GC content, the negative Migrants were strongly depleted for binding sites with high GC content, whereas the other TFs were enriched. The general enrichment is consistent with previous results [40], which strengthens the significance of the depletion seen in negative migrants. It has been shown that high GC-content favors binding and positioning of nucleosomes. The high GC-content of Pioneer binding sites is consistent with this, as the Pioneers are more likely to be involved in chromatin opening and repositioning of nucleosomes, whereas the low GC-content in binding sites of negative Migrants may be consistent with their preference for open chromatin without stably bound nucleosomes. Finally, the analysis of information content showed that Pioneers are enriched for high IC, whereas positive Migrants are depleted. The result for Pioneers is consistent with their role in initiating chromatin opening at specific genomic positions.

This may suggest that the chromatin opening index of Sherwood *et al.* is associated with IC and the number of DBDs, whereas the chromatin dependence is associated with GC content of the binding sites. This is a reasonable result. In this context the positive Migrants seem to represent a special case; they have binding sites with high GC-content, but low IC and no enrichment for multiple DBDs. Therefore their preference for PPIs may be due to a need for additional support and specificity during binding to regulatory regions, whereas in particular Pioneers and Settlers to a larger extent may have to rely upon DNA binding without any additional support through PPI from TFs already present within the regulatory region.

Analysis of TFs in a time course experiment

Finally we analyzed data from an *in vitro* differentiation time course experiment, generated by Soichi Ogishima and analyzed for expression levels by the FANTOM5 consortium using CAGE (cap analysis of gene expression) [41]. The experiment follows the transition from epithelial cells to mesenchymal cells after induction with TGF-β and TNF-α. The expression levels of individual genes have been compared to time zero using edgeR [42], and TFs with significantly changed expression level ($p<0.05$) where assigned to functional classes. The number of TFs showing significant changes in expression level at each time point are shown in Table S8 (see Additional File 1). We then asked whether there were any significant differences between categories of TFs throughout the time course.

As can be seen in Figure 4, all groups of TFs follow a similar time course where they are rapidly upregulated, followed by a relaxation leading to what seems to be a net downregulation in number of expressed genes. This may reflect a rapid activation of new genes in the regulatory network, with no clear distinction between functional classes of TFs. Since Pioneers, Settlers and Migrants are all needed for this, it is not surprising that they seem to be regulated in parallel.

(Figure 4)

We then used our list of TF properties [17], including e.g. Pfam domains and post-translational modifications, to check whether there were significant differences in properties for the TFs in this experiment. This was done as an enrichment analysis with a Fisher exact test, asking whether a given property was significantly enriched (or depleted) in a given set of TFs, compared to the full set of annotated TFs (Table 7). The p-values were corrected for multiple testing using the Benjamini correction. It should be noted that although we are partly using the same properties as for the classification, we are here testing for enrichment in the specific subset of TFs that show significant changes in expression levels in this particular experiment, rather than across all TFs.

(Table 7)

This analysis revealed several interesting features, in particular for Pioneers where there seems to be a shift in the regulatory program. The Pfam Ets domain is enriched in up-regulated Pioneers, whereas KRAB is enriched in down-regulated. The KRAB domain is associated with transcriptional repression, as it interacts with a corepressor protein (KAP-1) which recruits histone deacetylases and chromatin remodeling complexes to chromatin [43], maintaining a repressed status. This indicates that transcriptional repression is actively released in the differentiation process, enabling activation of new genes. The Ets domain, which is enriched in up-regulated Pioneers, can act both as an activator and a repressor [44]. However, another point here could be that many Ets-containing TFs are down-stream targets of signal transduction cascades [44], indicating an up-regulation of responses to signaling.

As already indicated, most of the observed changes are linked to down-regulation. For example, down-regulated Settlers are enriched in ubiquitination, a post-translational modification that may target proteins for degradation, possibly leading to a more rapid and efficient down-regulation of TFs than by changing the transcript level alone. For up-regulated positive Migrants there is enrichment for PPIs, supporting the observation above regarding PPIs and positive Migrants. This is possibly linked to stabilization of protein complexes involved in regulation of transcription.

A couple of Pfam domains (zf-H2C2_2 and HLH) are enriched in both up-regulated and down-regulated sets. The zinc-finger domain zf-H2C2, which is often involved in sequence-specific targeting of other domains, including KRAB [43], is strongly enriched in Pioneers, indicating site-specific changes in gene regulation. This is not seen for the Settlers, although the HLH domain may play a similar role here. For the positive Migrants it is seen only for the down-regulated TFs, whereas the negative Migrants actually are strongly depleted for zf-

H2C2 domains. This illustrates a clear difference between the sequence-specific targeting of Pioneers, compared to other TFs where additional interactions may be important.

We also did the same analysis over all TFs, independent of functional classification (Table 7). This identified most of the same terms as enriched, but at lower significance, and not the two depletions. Also, the analysis using functional classes clearly linked several changes in enrichment of properties to specific functional subclasses, such as Ets to Pioneers and PPI to positive Migrants. This is additional information that may help in interpretation of results, and underlines the added benefit of including data on functional classes.

The results described above seem to support a general picture of these TFs that is consistent with their assumed roles. The Pioneers are rapidly regulated to modify the transcriptional program, mainly by removing repressing TFs and up-regulating activating TFs that bind in a sequence-specific manner. This process is supported by the regulation of Settlers, many of which are rapidly degraded and removed, possibly to close up the regulatory regions that are being de-activated. The up-regulated Migrants are enriched for protein-protein interactions, which may support the formation of clusters of TFs in open regulatory regions.

CONCLUSIONS

Data on properties of transcription factors has been used as input for supervised machine learning methods in order to expand an experimental classification of transcription factors associated with chromatin opening, as Pioneers, Settlers, positive and negative Migrants. Our results support the hierarchical relationship between the transcription factors indicated by the original classification, and shows that this is associated with specific properties of these transcription factors, in particular their DNA binding domains. The expanded classification is a useful resource for analyzing other data, as for example transcription factor interaction data or time course experiments. This has been demonstrated on several sets of experimental data.

MATERIAL AND METHODS

The initial list of TFs and properties was taken from Bahrami *et al.* [17]. This list includes data on Pfam domains, PPIs and PTMs, please see the original paper for details. Data on TF classification was taken from the TFClass database [16]. Data on chromatin function of TFs was taken from Sherwood *et al.* [10]. Data on TF co-binding to DNA was taken from Jolma *et al.* [37] and from the ENCODE Consortium [38]). Data on PPI pairs for TFs was taken from Ravasi *et al.* [24] and from the Human Protein Reference Database (HPRD) [39] release 9. Data on the number of DNA-binding domains were generated from the list by Bahrami *et al.* [17]. Data on GC content and IC was generated from matrices downloaded from the Jaspar database [14]. Time course data were made available by the FANTOM Consortium [41, 45], and the assignment of genes to CAGE TSS clusters and edgeR analysis performed by the consortium was used for the project. An adjusted p-value of at least 0.05 was used as cutoff for the edgeR output.

All machine learning methods were implemented using scikit-learn [46], and all scripts used in the analysis were based on python 2.7 [47]. Estimation of p-values on 2x2 matrices was done using the fisher.test in R.

DECLARATIONS

List of abbreviations

AUC - area under curve (normally receiver operating characteristic curve); CAGE - cap analysis of gene expression; ChIP-seq - chromatin immunoprecipitation and sequencing; DBD - DNA-binding domain; FPR - false positive rate; GNB - Gaussian naïve Bayes; HTH - helix-turn-helix; IC - information content; kNN - k-nearest neighbor; LinearSVC - linear Support Vector Classification; MCC - Matthews correlation coefficient; PFM - position frequency matrix; PIQ - protein interaction quantitation; PPI - Protein-Protein Interaction; PPV - positive prediction value; PTM - post-translational modifications; RBF - radial basis function; RF - Random forest; ROC - receiver operating characteristic; SN - sensitivity; SVM - support vector machine; TF - Transcription factor; TFBS - transcription factor binding site; TN - true negative;.TP - true positive; TPR - true positives rate; TRANSFAC - TRANScription FACtor database;

Ethics approval and consent to participate

All data used in this project are from open sources, and do not require ethics approval or consent.

Availability of of data and material

The data sets supporting the results of this article are included within the article and its additional files.

Additional files

Additional file 1. Tables S1 to S4 and S7 to S8 (pdf)
Additional file 2. Table S5 (xls)
Additional file 3. Table S6 (xls)

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by funding from the Faculty of Medicine, Norwegian University of Science and Technology (NTNU) to RE, and by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU) to SB.

Authors' contributions

SB and RE collected all data, performed the analysis and drafted the initial manuscript. FD initiated and supervised the project, and participated in analysis of the interaction and differentiation data. All authors contributed equally to this work and approved the final manuscript.

Acknowledgements

This work was supported by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU). We thank Erik Arner (RIKEN Center for Life Science Technologies) for access to edgeR data on FANTOM5 time courses.

REFERENCES

1. Dynlacht BD: **Regulation of transcription by proteins that control the cell cycle**. *Nature* 1997, **389**(6647):149-152.
2. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS *et al*: **Serial regulation of transcriptional regulators in the yeast cell cycle**. *Cell* 2001, **106**(6):697-708.
3. Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N: **Structural and functional properties of genes involved in human cancer**. *BMC Genomics* 2006, **7**:3.

4. Boyadjiev SA, Jabs EW: **Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders**. *Clin Genet* 2000, **57**(4):253-266.
5. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD *et al*: **Natural selection on protein-coding genes in the human genome**. *Nature* 2005, **437**(7062):1153-1157.
6. De S, Lopez-Bigas N, Teichmann SA: **Patterns of evolutionary constraints on genes in humans**. *BMC Evol Biol* 2008, **8**:275.
7. Lopez-Bigas N, De S, Teichmann SA: **Functional protein divergence in the evolution of Homo sapiens**. *Genome Biol* 2008, **9**(2):R33.
8. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R: **Absence of a simple code: how transcription factors read the genome**. *Trends Biochem Sci* 2014, **39**(9):381-399.
9. Choukrallah MA, Matthias P: **The Interplay between Chromatin and Transcription Factor Networks during B Cell Development: Who Pulls the Trigger First?** *Front Immunol* 2014, **5**:156.
10. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK: **Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape**. *Nat Biotechnol* 2014, **32**(2):171-178.
11. Magnani L, Eeckhoute J, Lupien M: **Pioneer factors: directing transcriptional regulators within the chromatin environment**. *Trends Genet* 2011, **27**(11):465-474.
12. Luscombe NM, Austin SE, Berman HM, Thornton JM: **An overview of the structures of protein-DNA complexes**. *Genome Biol* 2000, **1**(1):REVIEWS001.
13. Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation**. *Brief Bioinform* 2008, **9**(4):326-332.
14. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H *et al*: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles**. *Nucleic Acids Res* 2014, **42**(Database issue):D142-147.
15. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, **32**(Database issue):D91-94.
16. Wingender E, Schoeps T, Donitz J: **TFClass: an expandable hierarchical classification of human transcription factors**. *Nucleic Acids Res* 2013, **41**(Database issue):D165-170.
17. Bahrami S, Ehsani R, Drablos F: **A property-based analysis of human transcription factors**. *BMC Res Notes* 2015, **8**:82.
18. Ben-Hur A, Weston J: **A user's guide to support vector machines**. *Methods Mol Biol* 2010, **609**:223-239.
19. Demsar J: **Statistical Comparisons of Classifiers over Multiple Data Sets**. *J Mach Learn Res* 2006, **7**:1-30.
20. Dietterich TG: **Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms**. *Neural Comput* 1998, **10**(7):1895-1923.
21. Salzberg SL: **On Comparing Classifiers: Pitfalls toAvoid and a Recommended Approach**. *Data Min Knowl Discov* 1997, **1**(3):317-328.
22. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**(Database issue):D222-230.
23. Cortes C, Vapnik V: **Support-Vector Networks**. *Mach Learn* 1995, **20**(3):273-297.
24. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N *et al*: **An atlas of combinatorial transcriptional regulation in mouse and man**. *Cell* 2010, **140**(5):744-752.
25. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function**

of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012, **40**(Database issue):D261-270.
26. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2013 update**. *Nucleic Acids Res* 2013, **41**(Database issue):D816-823.
27. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
28. Duan K-B, Rajapakse JC, Nguyen MN: **One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification**. In: *Proceedings of the 5th European conference on Evolutionary computation, machine learning and data mining in bioinformatics.* Valencia, Spain: Springer-Verlag; 2007: 47-56.
29. Estabrooks A, Jo T, Japkowicz N: **A Multiple Resampling Method for Learning from Imbalanced Data Sets**. *Computational Intelligence* 2004, **20**(1):18-36.
30. Laurikkala J: **Improving Identification of Difficult Small Classes by Balancing Class Distribution**. In: *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine.* Springer-Verlag; 2001: 63-66.
31. He H, Garcia EA: **Learning from Imbalanced Data**. *IEEE Trans on Knowl and Data Eng* 2009, **21**(9):1263-1284.
32. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995: 1137-1143.
33. Graczyk M, Lasota T, Telec Z, Trawiński B: **Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems** *14th International Conference, KES* 2010, **6276**:9.
34. Kohavi R, John GH: **Wrappers for feature subset selection**. *Artif Intell* 1997, **97**(1-2):273-324.
35. Miralles F, Posern G, Zaromytidou AI, Treisman R: **Actin dynamics control SRF activity by regulation of its coactivator MAL**. *Cell* 2003, **113**(3):329-342.
36. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009, **10**:48.
37. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J: **DNA-dependent formation of transcription factor pairs alters their binding specificity**. *Nature* 2015, **527**(7578):384-388.
38. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.
39. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.
40. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y *et al*: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors**. *Genome Res* 2012, **22**(9):1798-1812.
41. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M *et al*: **Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells**. *Science* 2015, **347**(6225):1010-1014.
42. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
43. Lupo A, Cesaro E, Montano G, Zurlo D, Izzo P, Costanzo P: **KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions**. *Curr Genomics* 2013, **14**(4):268-278.
44. Oikawa T, Yamada T: **Molecular biology of the Ets family of transcription factors**. *Gene* 2003, **303**:11-34.
45. **FANTOM5** [**http://fantom.gsc.riken.jp/5/**]

46. **scikit-learn** [**http://scikit-learn.org/**]
47. **Python** [https://**www.python.org/**]

TABLES

Table 1 - Summary of property encodings

Property	Description	Encoding
TF_Class	Encoded 2-3 top levels of five digit code based on TFClass classification; i.e. superclass followed by class (see text)	A 47-dimensional vector where i^{th} position (superclass) and j^{th} position (class) are 1, other positions are 0
PD	TF has a frequent Pfam domain (yes/no)	1 / 0
DBD	TF has a DBD (yes/no)	1 / 0
N_DBD	Number of DBDs (see text)	11 / 10 / 00
PPI	TF has a PPI (yes/no)	1 / 0
N_PPI	Number of PPIs (see text)	11 / 10 / 00
N_PhS	Number of Phosphorylation sites (see text)	11 / 10 / 00
PTM	TF has a PTM (yes/no)	1 / 0
Ind_PTM	TF has a specific PTM (yes/no)	An ordered 6-dimensional vector where position i corresponding to PTM i is 1 / 0
N_ZFD	Number of the zinc finger domains (see text)	11 / 10 / 00

Table 2 - Summary of criteria for balanced datasets

Case	Specific class	Rest class	Splits	Average size
Pioneers vs Rest	45	47+77+288	1+9	45+46
Settlers vs Rest	47	45+77+288	1+9	47+46
Positive Migrants vs Rest	77	45+47+288	1+5	77+76
Negative Migrants vs Rest	288	45+47+77	2+1	144+169

Table 3 - AUC score and feature importance by RF classifier on individual properties.

Properties	Pioneers vs Rest		Settlers vs Rest		Migrants+ vs Rest		Migrants- vs Rest	
	AUC score	Feature importance	AUC score	Feature importance	AUC score	Feature importance	AUC score	Feature importance
TF_Class	*0.824*	*0.414*	*0.798*	*0.517*	*0.791*	*0.454*	*0.909*	*0.533*
PD	*0.825*	*0.351*	*0.803*	*0.256*	*0.804*	*0.286*	*0.868*	*0.308*
Ind_PTM	0.520	*0.122*	0.494	*0.118*	*0.603*	*0.161*	*0.612*	*0.110*
N_PPI	0.504	0.040	0.575	0.053	0.599	0.049	0.536	0.019
N_ZFD	*0.591*	0.028	0.502	0.009	0.491	0.005	0.512	0.009
PPI	0.508	0.019	*0.582*	0.028	0.570	0.021	0.506	0.008
PTM	0.498	0.006	0.490	0.007	0.532	0.008	0.525	0.004
DBD	0.495	0.008	0.502	0.004	0.499	0.005	0.498	0.002
N_DBD	0.504	0.007	0.497	0.002	0.499	0.005	0.500	0.002
N_PhS	0.525	0.000	0.496	0.000	0.536	0.000	0.547	0.000

The three best scores for each case are shown in *italics*. The individual properties are explained in Table 1.

Table 4 - Significance of change in AUC score for stepwise addition of properties.

Pioneers vs Rest										
Properties	TF_Class	PD	N_PPI	N_DBD	N_ZFD	DBD	PPI	PTM	Ind_PTM	N_PhS
AUC-Score	0.8223	0.8315	0.8404	0.8459	0.8444	0.8456	0.8430	0.8451	0.8373	0.8330
P-value		0.0008	0.0004	0.0229	0.4139	0.8206	0.1055	0.1378	0.0122	0.0395
Settlers vs Rest										
Properties	PD	N_PPI	TF_Class	PTM	N_DBD	N_ZFD	N_PhS	PPI	DBD	Ind_PTM
AUC-Score	0.8002	0.8180	0.8224	0.8211	0.8220	0.8251	0.8230	0.8168	0.8129	0.8024
P-value		6.3e-06	0.0222	0.8165	0.4149	0.3442	0.3659	0.0197	0.0258	0.0040
Positive Migrants vs Rest										
Properties	PD	TF_Class	N_ZFD	PPI	DBD	N_DBD	PTM	N_PPI	Ind_PTM	N_PhS
AUC-Score	0.8053	0.8238	0.8290	0.8269	0.8280	0.8241	0.8133	0.8059	0.7993	0.8039
P-value		1.0e-05	0.0365	0.5172	0.6968	0.0438	0.0003	0.0232	0.0586	0.7215
Negative Migrants vs Rest										
Properties	TF_Class	PD	PTM	N_ZFD	N_DBD	DBD	PPI	N_PhS	N_PPI	Ind_PTM
AUC-Score	0.9063	0.9131	0.9157	0.9172	0.9176	0.9139	0.9101	0.9117	0.9071	0.8946
P-value		0.0182	0.3988	0.1448	0.8565	0.0351	0.2052	0.0940	0.0475	0.0004

Table 5 - Experimental data and classification results according to TFClass

		Total	Experimental					Classified					Average p-value	
TFClass	Code	#TFs	#TFs	P	S	M+	M-	#TFs	P	S	M+	M-	All	Margin
Uncharacterized	0.0	6	0	0	0	0	0	6	0	5	1	0	0.682	0.027
NonO domain factors	0.2	2	0	0	0	0	0	2	0	2	0	0	0.679	0.022
Leucine-rich repeat proteins	0.3	1	0	0	0	0	0	1	0	1	0	0	0.679	0.022
NFX1-type putative zf factors	0.4	3	0	0	0	0	0	3	1	2	0	0	0.673	0.016
bZIP	1.1	54	25	2	9	3	11	29	10	9	0	10	0.720	0.150
bHLH	1.2	84	29	1	23	3	2	55	0	55	0	0	0.943	0.539
bHSH	1.3	4	1	1	0	0	0	3	3	0	0	0	0.759	0.128
Nuclear receptors with C4 zfs	2.1	47	47	1	1	45	0	0	0	0	0	0	-	-
Other C4 zfs	2.2	17	4	0	0	1	3	13	0	0	0	13	0.738	0.056
Three-zf Kruppel-related factors	2.3.1	26	5	4	1	0	0	21	21	0	0	0	0.958	0.273
Other factors with 3 adjacent zf	2.3.2	15	0	0	0	0	0	15	12	0	3	0	0.773	0.203
More than 3 adjacent zf	2.3.3	327	12	6	1	4	1	315	201	0	114	0	0.862	0.213
Factors multiple dispersed zf	2.3.4	103	5	1	2	1	1	98	35	58	5	0	0.765	0.086
DM-type intertwined zf factors	2.5	6	0	0	0	0	0	6	0	0	6	0	0.704	0.166
CXXC zf factors	2.6	7	0	0	0	0	0	7	0	0	7	0	0.704	0.112
C2HC zf factors	2.7	8	0	0	0	0	0	8	0	0	8	0	0.704	0.116
C3H zf factors	2.8	2	0	0	0	0	0	2	0	0	2	0	0.704	0.116
C2CH THAP-type zf factors	2.9	1	0	0	0	0	0	1	0	0	1	0	0.704	0.084
Homeo domain factors	3.1	198	198	0	0	3	195	0	0	0	0	0	-	-
Paired box factors	3.2	9	5	0	1	1	3	4	0	1	2	1	0.735	0.181
Fork head	3.3	56	56	2	3	1	50	0	0	0	0	0	-	-
Heat shock factors	3.4	5	0	0	0	0	0	5	0	5	0	0	0.683	0.120
Tryptophan cluster factors	3.5	50	34	23	3	7	1	16	0	0	16	0	0.937	0.264
TEA domain factors	3.6	4	1	0	0	1	0	3	0	0	3	0	0.808	0.152
ARID domain factors	3.7	12	1	0	0	0	1	11	1	10	0	0	0.631	0.012
HMG domain factors	4.1	41	19	0	0	2	17	22	0	0	0	22	0.741	0.338
Het. CCAAT-binding factors	4.2	4	2	2	0	0	0	2	2	0	0	0	0.817	0.357
MADS box factors	5.1	5	2	0	0	0	2	3	0	0	0	3	0.650	0.014
SAND domain factors	5.3	7	3	1	0	2	0	4	0	0	4	0	0.906	0.304
RHR factors	6.1	21	3	0	1	2	0	18	0	0	18	0	0.907	0.125
STAT domain factors	6.2	7	0	0	0	0	0	7	0	0	7	0	0.777	0.076
p53 domain factors	6.3	3	0	0	0	0	0	3	0	0	3	0	0.777	0.062
Runt domain factors	6.4	3	0	0	0	0	0	3	0	0	3	0	0.777	0.062
T-Box factors	6.5	16	1	0	1	0	0	15	0	12	3	0	0.814	0.071
Grainyhead domain factors	6.7	4	0	0	0	0	0	4	0	0	4	0	0.777	0.122
SMAD/NF-1 DBD factors	7.1	8	1	1	0	0	0	7	3	4	0	0	0.761	0.211
GCM domain factors	7.2	2	1	0	1	0	0	1	0	1	0	0	0.844	0.268
TATA-binding proteins	8.1	2	1	0	0	1	0	1	0	0	1	0	0.805	0.171
AT hook factors	8.2	2	1	0	0	0	1	1	0	1	0	0	0.634	0.001
Cold-shock domain factors	9.1	3	0	0	0	0	0	3	0	3	0	0	0.679	0.022
Sum	-	1175	457	45	47	77	288	718	289	169	211	49		

The results are shown for Pioneers (P), Settlers (S), positive Migrants (M+) and negative Migrants (M-).

Table 6 – Enriched or depleted features related to TF-TF-DNA interactions

	Obs/Exp	P	Average
TFs with #DBDs > 1			#DBDs
Pioneers	278/152	<2.2e-16	7.05
Settlers	*61/98*	*1.1e-08*	*2.87*
Migrants+	145/131	0.07	4.80
Migrants-	*51/153*	*<2.2e-16*	*1.23*
TFs with GC > 40%			%GC
Pioneers	45/16	1.4e-11	60.7
Settlers	47/17	6.9e-11	58.8
Migrants+	71/28	5.3e-11	52.6
Migrants-	*9/108*	*1.8e-10*	*23.7*
TFs with IC > 9.0			IC
Pioneers	38/23	3.2e-06	10.4
Settlers	29/24	0.21	10.3
Migrants+	*20/40*	*3.4e-07*	*7.7*
Migrants-	153/151	0.77	9.3

Significantly depleted features are highlighted in *italics*.

Table 7 - Enriched or depleted features in significantly regulated TFs

	Up			Down		
Class	Term	Obs/Exp	P(Benj)	Term	Obs/Exp	P(Benj)
Pioneers	Ets	9/0	2.2e-07			
	zf-H2C2_2	25/9	1.2e-05	zf-H2C2_2	69/18	1.0e-08
				KRAB	36/10	1.1e-08
				zf-H2C2_4	14/3	9.2e-05
Settlers	HLH	20/1	2.5e-09	HLH	18/2	3.9e-09
				Ubiquitination	33/18	4.7e-04
				Sumoylation	11/5	4.6e-02
				PPI	*5/11*	*4.9e-02*
Migrants+	PPI	20/11	6.1e-03			
				Hormone_recep	19/1	6.6e-09
				zf-C2H2	33/8	7.3e-09
				zf-C4	19/1	9.7e-09
Migrants-				HMG_box	11/1	1.1e-05
				zf-H2C2_2	*1/14*	*3.6e-04*
				Homeobox	17/5	9.3e-04
				Methylation	16/8	4.2e-02
All	Ubiquitination	74/49	5.4e-05	Ubiquitination	112/90	8.4e-03
	bZIP_1	12/2	1.8e-04			
	HLH	20/6	2.7e-04			
	Ets	10/2	1.1e-03			
	Sumoylation	27/15	5.1e-03			
	PPI	50/37	1.8e-02			
	Phosphorylation	144/136	4.9e-02	Phosphorylation	269/252	7.7e-04
				zf-H2C2_2	113/63	5.7e-08
				zf-C2H2	56/29	6.0e-05
				zf-C4	20/6	4.8e-05
				Hormone_recep	20/6	8.4e-05
				KRAB	58/34	8.9e-04
				zf-C2H2_4	24/10	2.5e-03

Only features with at least 9 occurrences (as observed for enrichments or expected for depletions) are listed. Depleted features are highlighted in *italics*. The All category shows enrichment analysis of all significantly up- or down-regulated genes, independent of functional classification.

FIGURE LEGENDS

Figure 1 - Distribution of TFs based on the classification by Sherwood *et al*.

The classification by Sherwood *et al.* [10] has been extended to four different functional classes; Pioneers, Settlers, positive Migrants and negative Migrants. The points in light olive show the TFs that had intersection with TF classification (TFClass), TF function (Sherwood *et al.*), and our database of TF properties. The figure has been adapted from Sherwood *et al*.

Figure 2 - AUC scores for forward best-first search.

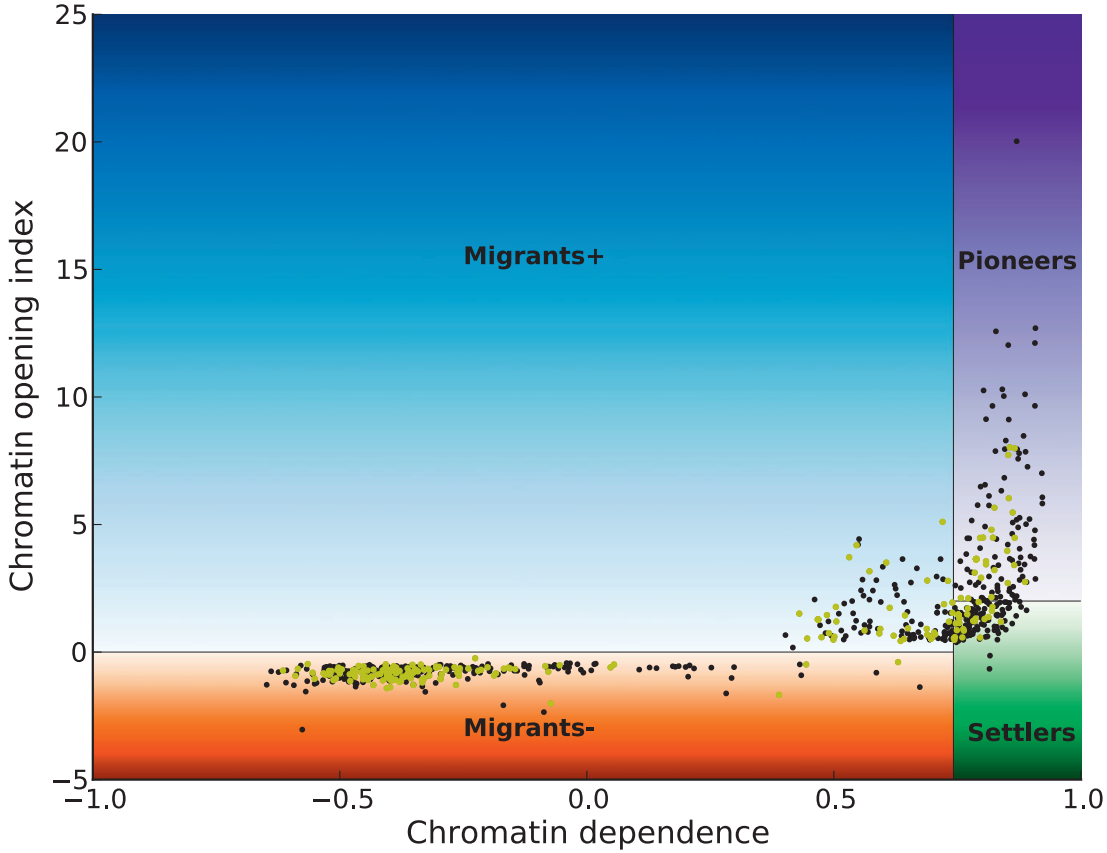
The forward best-first algorithm with the RF classifier was used to find a list of properties with good classification performance. The AUC scores while adding properties stepwise are shown, at each step adding the property that gave the highest AUC score.

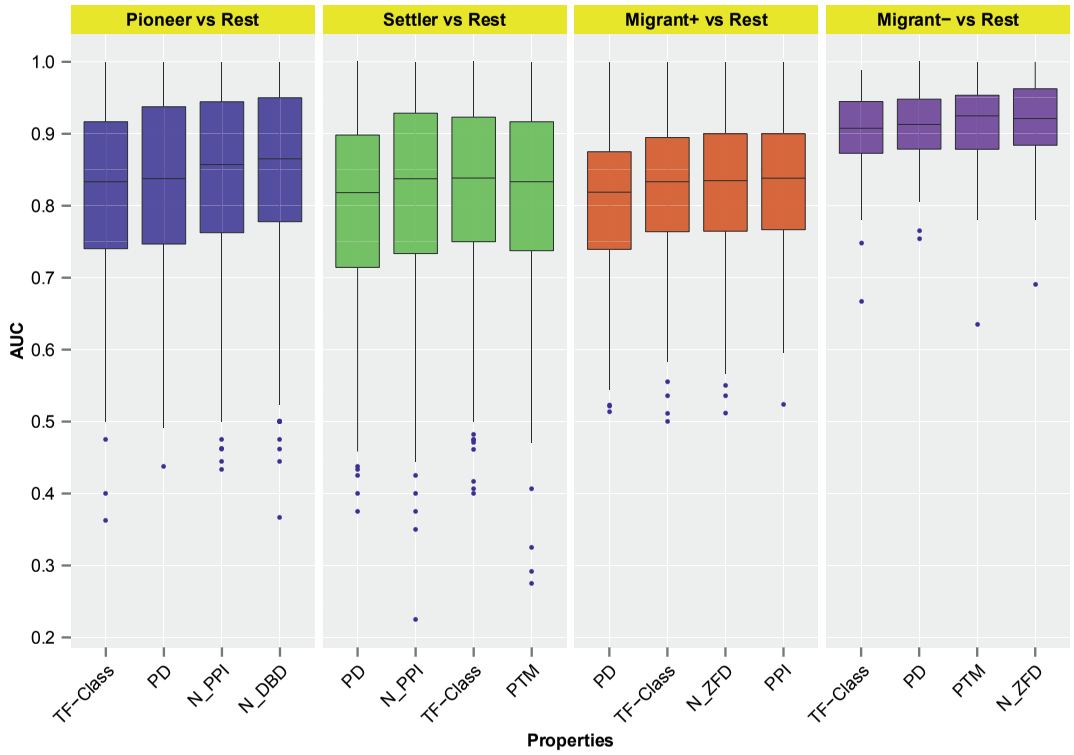
Figure 3 – Enrichment and depletion in TF-TF interactions

The figure illustrates cases of enrichment (red plus) and depletion (blue minus) relative to random expectation for TF-TF interactions between functional classes, based on data related to a) DNA binding and b) PPI. For each pair of functional classes (Pioneers (P), Settlers (S), positive Migrants (M+) and negative Migrants (M-)) the enrichment or depletion is indicated for each of the data sources, 2x DNA binding in a) and 2x PPI in b). The strongest tendency is that interactions involving in particular Pioneers tend to be enriched in DNA-based interactions, whereas interactions involving in particular positive Migrants tend to be depleted in DNA-based interactions and enriched in PPIs.

Figure 4 - Log ratio of number of up-regulated versus down-regulated TFs.

The graph shows the log ratio of number of significantly up-regulated genes versus number of down-regulated genes at each time point for each of the TF classes. The different functional classes show similar trends, indicating that they are needed in combination for activation of new genes, at least within the time resolution provided by the experimental data.


Migrants+
Pioneers
Migrants-
Settlers
Chromatin opening index
Chromatin dependence
25
20
15
10
5
0
−5
−1.0
−0.5
0.0
0.5
1.0


Pioneer vs Rest
Settler vs Rest
Migrant+ vs Rest
Migrant− vs Rest
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
TF−Class
PD
N_PPI
N_DBD
PD
N_PPI
TF−Class
PTM
PD
TF−Class
N_ZFD
PPI
TF−Class
PD
PTM
N_ZFD
Properties

a)

	P		S		M+		M-	
P		+	+	+		−	+	
S				+				+
M+					−	−	−	−
M-	DNA (SELEX)	DNA (ChIP-seq)					−	

b)

	P		S		M+		M-	
P			−					−
S				+	−	+	−	−
M+					+	+		−
M-	PPI (TFs)	PPI (general)					+	+


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Log ratio up vs. down
Pioneer
Settler
Migrant+
Migrant-
no net change
00h15m
00h30m
00h45m
01h00m
01h20m
01h40m
02h00m
02h30m
03h00m
03h30m
04h00m
05h00m
06h00m
07h00m
08h00m
12h00m
16h00m
24h00m
42h00m
60h00m
Time

Paper III

Accepted Manuscript

Gene regulation in the immediate-early response process

Shahram Bahrami, Finn Drabløs

PII: S2212-4926(16)30001-X

DOI: 10.1016/j.jbior.2016.05.001

Reference: JBIOR 148

To appear in: *Advances in Biological Regulation*

Received Date: 11 January 2016

Accepted Date: 3 May 2016

Please cite this article as: Bahrami S, Drabløs F, Gene regulation in the immediate-early response process, *Advances in Biological Regulation* (2016), doi: 10.1016/j.jbior.2016.05.001.

Gene regulation in the immediate-early response process

Shahram Bahrami[a, b], Finn Drabløs[a,*]

[a] Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

[b] St. Olavs Hospital, Trondheim University Hospital, NO-7006 Trondheim, Norway

[*] Corresponding author

Email addresses:

Shahram Bahrami - <shahram.bahrami@ntnu.no>

Finn Drabløs - <finn.drablos@ntnu.no>

Address for correspondence:

Finn Drabløs

Department of Cancer Research and Molecular Medicine

Norwegian University of Science and Technology

P.O. Box 8905

NO-7491 Trondheim

Norway

Gene regulation in the immediate-early response process

Abstract

Immediate-early genes (IEGs) can be activated and transcribed within minutes after stimulation, without the need for *de novo* protein synthesis, and they are stimulated in response to both cell-extrinsic and cell-intrinsic signals. Extracellular signals are transduced from the cell surface, through receptors activating a chain of proteins in the cell, in particular extracellular-signal-regulated kinases (ERKs), mitogen-activated protein kinases (MAPKs) and members of the RhoA-actin pathway. These communicate through a signaling cascade by adding phosphate groups to neighboring proteins, and this will eventually activate and translocate TFs to the nucleus and thereby induce gene expression. The gene activation also involves proximal and distal enhancers that interact with promoters to simulate gene expression. The immediate-early genes have essential biological roles, in particular in stress response, like the immune system, and in differentiation. Therefore they also have important roles in various diseases, including cancer development. In this paper we summarize some recent advances on key aspects of the activation and regulation of immediate-early genes.

Keywords

Immediate-early response; Signaling cascades; Poised genes; Transcription factors; Enhancers

Abbreviations

IER - immediate-early response; IEG - immediate-early gene; PRG - primary response gene; SRG - secondary response genes; PDGF - platelet-derived growth factor; EGF - epidermal growth factor; SRF - serum-response factor; NF-kB - nuclear factor-kB; CREB - cyclic AMP response element-binding protein; AP-1 - activator protein-1; TCF - ternary complex factor; ERK - extracellular signal-regulated kinase; MAPK - Mitogen-activated protein kinases; ELK1- E26-like kinase; MRTF - myocardin related transcription factor; NF1 - nuclear factor 1; PARP1 - Poly (ADP-ribose) polymerase 1; RSK - p90 ribosomal S6 kinase; JNK - c-Jun N- terminal kinase; ERK5 - extracellular signal regulated kinase-5; BMK1- Big MAP kinase-1; MSK - Mitogen/stress activated protein kinase; RNA Pol II - RNA polymerase II; GO - Gene Ontology; TBP - TATA binding protein; TSS - Transcription Start Site; HAT - histone acetyl transferase; IRF3 - interferon regulatory factor 3; TLR - Toll-like receptor; NGF - nerve growth factor; G protein - guanine nucleotide binding protein; TF - Transcription Factor; MKL - megakaryoblastic leukemia; ESC - embryonic stem cells; DSIF - DRB sensitivity-inducing factor; NELF - negative elongation factor; P-TEFb - positive transcription elongation factor; CTD - C-terminal domain; eRNA- enhancer RNA

Introduction

Regulation of gene transcription is one of the main mechanisms that are used by cells to increase or decrease the concentration of specific gene products (RNA and protein) (Lewin, 2004). Gene transcription is controlled through many layers of regulation, where the choice of specific pathways affects the timing of induced gene expression as a response to an external signal. A specific group of genes seems to be able to respond very quickly to regulatory signals, for example in immune responses or cellular stress. Such processes are often known as immediate-early response (IER) processes, and the genes involved are therefore known as immediate-early genes (IEGs).

There are many relevant questions regarding IEGs. For example, how are IEGs activated, since they are able to respond very rapidly to external signals? What are the key aspects of their promoters? Do they interact with enhancers? How important is the epigenetic profile of the IEGs? This paper tries to summarize and provide updated information on some of these questions.

Early gene responses

Primary responses

Several genes respond rapidly to cellular signals, and such signal-responsive primary response genes (PRGs) are expressed following a wide range of different stimuli, linked to diverse signaling pathways. They can be divided into two main classes; the immediate-early response genes, and the delayed primary response genes.

Immediate-early response genes

The mRNA for IEGs may appear in cells within minutes after stimulation. Even more important, cells can transcribe mRNA for IEGs in the presence of protein synthesis inhibitors, indicating that the proteins required for their synthesis (including e.g. the transcription factors) are already available in the cell, and not synthesized as part of the activation process (Herschman, 1991, Morgan and Curran, 1991). These genes respond to a wide variety of extrinsic stimuli and in multiple cell types (Fowler et al., 2011), indicating a very general response mechanism. There are probably a few hundred genes in this group. These genes were first identified in cells exposed to mitogens, and have an important role in the regulation of the cell cycle (Greenberg and Ziff, 1984). Many IEGs are proto-oncogenes and their sustained expression can have profound effects on cellular growth.

Delayed primary response genes

Many of the primary response genes encode transcription factors, which again regulate secondary response genes (Winkles, 1998) (see subsection Secondary responses). However, it has been shown that some of the delayed inductions do not require protein synthesis, and therefore represent delayed induction of primary response genes rather than induction of secondary response genes. This group of genes is called delayed primary response genes, and they are different from IEGs both in function and in genomic architecture (Tullai et al., 2007).

Secondary responses

This group of genes is also expressed in response to signaling, but requires *de novo* protein synthesis. These genes are much more abundant than the genes in the first group, and are called secondary response genes (SRGs) (Herschman, 1991, Serrat et al., 2014).

General properties of IEGs

Expression of IEGs is quick and mainly transient, it does not require protein synthesis, and therefore translational inhibitors have no effect on their expression. Their expression in interphasic cells is initiated by an extracellular signal, such as growth factors (e.g. platelet-derived growth factor (PDGF) and epidermal growth factor (EGF)), mitogens and phorbol esters, immunological and neurological signals, developmental, and stress (e.g. UV, toxins) (Herschman, 1991, Morgan and Curran, 1991, O'Donnell et al., 2012). For example, expression of the FOS gene peaks 30 to 60 minutes after stimulation, and returns to basal expression after 90 minutes (Greenberg and Ziff, 1984). IEG protein products are usually unstable and they are sometimes targeted for proteolytic degradation by the proteasome without prior ubiquitination (Gomard et al., 2008). For IEG transcripts, downregulation is suggested to follow an additional mechanism through the actions of targeted microRNAs (Aitken et al., 2015, Avraham et al., 2010), where a family of microRNAs target the 3′ UTR region of several transcripts. Multiple microRNAs may target multiple IEGs, which provides some redundancy. After stimulation of IEG expression the production of these microRNAs is blocked, but then comes quickly back to normal levels (Aitken et al., 2015, Avraham et al., 2010). The combination of several mechanisms for rapid degradation and inactivation enables very transient signaling after IEG activation.

IEGs have on average shorter length than other genes (19 kb versus 58 kb), and they have significantly fewer exons. They have a high prevalence of TATA boxes and CpG islands. There is an enrichment for some specific transcription factor binding sites within regulatory regions of IEGs, including serum-response factor (SRF), nuclear factor kappa B (NF-kB) and cyclic AMP response element-binding protein (CREB) binding sites. This suggests a consistent and maybe redundant mechanism of transcription regulation (Healy et al., 2013).

Important IEGs and pathways

Our current knowledge about IEGs and how they are activated is to a large extent based on studies of individual genes and pathways. Here we describe some representative examples.

Important Immediate Early Genes

Two of the most famous and well-characterized immediate-early genes are FOS and JUN (Healy et al., 2013, O'Donnell et al., 2012). They can be rapidly and transiently induced by a variety of stimuli, including serum, growth factors, cytokines, tumor promoters, and UV radiation. FOS plays a key role in cellular events, including proliferation, differentiation and survival, and is also regulated by posttranslational modification such as phosphorylation by different kinases like MAP kinases, which influences protein stability, DNA-binding activity and the trans-activating potential of the transcription factors (O'Donnell et al., 2012).

The FOS and JUN proteins have a leucine zipper-containing domain (Pfam bZIP_1) used for dimerization and DNA-binding. The JUN protein also includes a JUN domain, which can be modified by posttranslational modifications such as phosphorylation and acetylation (Bahrami et al., 2015, Finn et al., 2014). The FOS transcription factor is not independently active, and must form a heterodimer with a member of the JUN family to form the active transcription factor activator protein (AP-1). This interaction happens via the leucine zipper motif, forming a bipartite DNA-binding domain (Healy et al., 2013). AP-1 regulates the expression of target genes by binding DNA at the consensus sequence known as the TPA responsive element (TRE), which is found within the upstream promoter region of AP-1 target genes (Healy et al., 2013). This transcription factor plays an important role during both normal development and disease states such as cancer (Ozanne et al., 2006).

Early Growth Response gene 1 (EGR-1) is another member of the immediate early genes family. EGR-1 is also known as Zif268 and encodes a nuclear phosphoprotein, also known as Krox24 (Kukushkin et al., 2005). EGR-1 has both DNA-binding and non-DNA-binding domains (Bahrami et al., 2015), it has interaction with CEBPB, PSMA3 and P53 and is involved in the regulation of cell growth and differentiation in response to signals such as mitogens, growth factors, and stress stimuli (Bae et al., 2002, Liu et al., 2001, Zhang et al., 2003).

Signaling pathways for Immediate Early Genes

Extracellular signals will promote activation of an assortment of pathways within the cell, leading to activation of transcription factors and induction of gene expression, in particular IEGs. There are several pathways that lead to the activation of regulatory proteins involved in IEG expression, such as the RhoA-actin, ERK and p38 MAPK and PI3K pathways. Here we will mainly focus on the RhoA-actin pathway and the ERK and p38 MAPK pathways (Figure 1). These pathways lead to phosphorylation and activation of regulatory proteins involved in IEG expression, such as members of the ETS-domain family, for example transcription factors ELK1 and ETS1/2, which bind to the promoter of relevant genes and form complex with lysine acetyltransferases. Also, these pathways lead to activation of other regulatory factors that are essential for induction, such as the SRF and the Mediator complex (Fowler et al., 2011). They will also initiate changes in post-translational modifications of histones, leading to changes in the chromatin structure (Ciccarelli and Giustetto, 2014, Flouriot et al., 2014, Sawicka et al., 2014). Multiple pathways may be activated in parallel for a given signal (Bebien et al., 2003).

(Figure 1)

Rho GTPases regulate the activity of SRF, one of the transcription factors that regulate many immediate-early genes, through their ability to induce actin polymerization. The Rho GTPases is a family of small signaling G proteins, and one of the major Rho GTPases involved in for example spine morphogenesis is RhoA, which modulates the regulation and timing of cell division. The major receptors of RhoA are GPCR (G-Protein Coupled Receptor), EphA (Ephrin A), IGF (Insulin-like Growth Factor) and Ktn1 (Kinectin-1).

There is a cycle between an active GTP-bound state and an inactive GDP-bound state for Rho proteins. Their activation state is controlled by regulatory proteins such as GEFs (guanine exchange

factors), which catalyze the exchange of GDP for GTP and thereby activates Rho, as well as GDIs (guanine dissociation inhibitors) and GAPs (GTPase activating proteins).

The Rho-associated kinases (ROCKs) are principal mediators of RhoA activity. ROCK leads to the stimulation of LIMK (LIM-kinase). Both LIMK1 and 2 phosphorylate and inactivate Cofilin, an actin-depolymerizing factor, and Cofilin reorganizes the actin cytoskeleton of the cell, leading to polymerization of G-actin into F-actin. G-actin binds MKL1 through N-terminal RPEL motifs (Miralles et al., 2003, Vartiainen et al., 2007), and a reduction in G-actin therefore leads to more free MKL1 in the nucleus (Vartiainen et al., 2007). MKL1/2 forms a complex with SRF and activates SRF target gene expression in the nucleus, including the SRF gene itself (Cen et al., 2003, Miralles et al., 2003).

IEG expression can also be induced by one of the MAPK (mitogen-activated protein kinase) effector cascades. There are different MAPK cascades, with five major groups of MAPKs in mammalian cells, including ERK (extracellular signal regulated kinase), RSK (p90 ribosomal S6 kinase), JNK (c-Jun N-terminal kinase), p38, and ERK5 (extracellular signal regulated kinase-5, also called Big MAP kinase-1 (BMK1)) (Raman et al., 2007, Yang et al., 2003, Yasuda and Kurosaki, 2008). JNK and p38 are activated by UV or stress stimuli, ERK and RSK are mainly activated by mitogenic stimuli such as growth factors and hormones, whereas ERK5 is activated by both stress stimuli and growth factors (Yang et al., 2003).

Here we will focus on two important MAPK cascades; the ERK-MAPK and the p38-MAPK pathways (Figure 1). In the ERK-MAPK pathway, signals lead to phosphorylation of ELK-1 by ERK1/2, and ELK-1, which is a ternary complex factor (TCF), acts as a co-factor for SRF (Yang et al., 2003). Phosphorylation of ELK-1 leads to alternation of the complex with p300 and facilitates transcriptional activation (Li et al., 2003). Phosphorylated ELK-1 binds to SRE target sites and is associated with transcriptional co-activators like CREB-binding protein and/or p300 (Hazzalin and Mahadevan, 2005, Li et al., 2003). The p38-MAPK pathway can be stimulated by both growth factors and general stress, and leads to activation of the p38 MAPK kinase, which subsequently activates several transcription factors, including ELK1. MSK1/2 (mitogen- and stress-activated protein kinase 1and 2) are downstream targets that can be phosphorylated by both ERK1/2 and p38 MAPK, and therefore this represents a link between these two pathways. MSK1/2 phosphorylates several proteins such as transcription factors of CREB and NF-kB, which regulate IEG expression, and also histone H3 at serine 10 and serine 28 at the upstream promoter region of IEGs. It has been shown that these kinases are active as negative regulators of acute inflammation, and for example MSK1/2 is involved in the activation of feedback mechanisms that dampen oxazolone-induced skin inflammation (Bertelsen et al., 2011, Soloaga et al., 2003).

A binding site for the phosphoserine binding protein 14-3-3 is created by MSK1/2 (Macdonald et al., 2005), and this protein connects components of the transcription activation machinery, such as the lysine acetyltransferase PCAF and the SWI/SNF ATPase BRG1 (Drobic et al., 2010). These components produce an open promoter complex which allows transcription to proceed. Extracellular signaling via activation of MSK1/2 leads to direct chromatin modification, and this regulation is called the nucleosomal response. If MSK1/2 is knocked out or blocked the expression of IEGs is reduced (Soloaga et al., 2003). It has on the other hand been observed that

phosphorylated histone H3 at serine10 (H3S10ph) has a significant role in transcription initiation. For example, following induction and MSK1/2-induced phosphorylation of histone H3, this modification site acts towards the lysine acetyltransferase MOF. This transferase acetylates lysine 16 on histone H4 (H4K16), which is bound by the bromodomain of BRD4. BRD4 recruits the kinase PTEF-b. Then the kinase PTEF-b phosphorylates and releases stalled RNA polymerase from the proximal promoter region, which finally results in transcription elongation (Zippo et al., 2009).

The molecular events during FOS expression can be used as an example of IEG regulation. A key transcription factor complex consisting of SRF and a member of the TCF family of ETS transcription factors is responsible for transduction of a signal from the ERK-associated MAPK pathway. The TCF component is a receptor of this signal, being a direct MAPK phosphorylation target (Selvaraj et al., 2015, Shaw and Saxton, 2003, Yang and Sharrocks, 2006). TCFs can be multiple phosphorylated (Bahrami et al., 2015), although the exact role of this is unclear. ELK1 is example of an ETS/TCF-type transcription factor containing a carboxy-terminal MAPK-controlled transcriptional activation domain activated by MAPKs (Mylona et al., 2011). TCFs have a high affinity to DNA (Bahrami et al., 2015), and the affinity of TCFs for the binary SRF-DNA complex increases upon phosphorylation by MAPKs and decreases markedly upon treatment with phosphatases (Price et al., 1995).

SRF acts as a platform for TCF. SRF has been fused to the C-terminal region of ELK1, which has been used to show that the TCF component signaling through SRF is enough to couple ERK pathway signaling *in vivo* to T-cell development (Mylona et al., 2011). In other signaling situations, SRF can co-operate with other co-regulatory factors such as members of the MRTF (myocardin-related transcription factor) family, and thereby affect the regulation of FOS expression (Cen et al., 2003, Knoll and Nordheim, 2009, Posern and Treisman, 2006).

Several other transcription factors that bind up- and downstream from the TCF-SRF binding site may play a potential role in FOS expression. ELK1 is one the TCF proteins that is located upstream of the positioned -1 nucleosome where there is a binding site for the TCF-SRF complex. The transcription factor is modified through sumoylation in the absence of growth factor signaling. This can recruit histone deacetylase (HDAC)-containing co-repressor complexes to the FOS promoter to maintain a low basal expression level (Khan and Davie, 2013, Yang and Sharrocks, 2006). Upon growth factor-mediated activation of the ERK MAPK pathway, a p300-dependent pathway leads to increased histone acetylation levels. This occurs through allosteric activation of p300 by ELK1 phosphorylation (Li et al., 2003). ELK1 leads to recruitment of MSKs to the promoter and thereby H3S10 phosphorylation (Zhang et al., 2008). The changes in histone acetylation lead to access of NF1 (nuclear factor 1) to a binding site occluded by the -1 nucleosome, and thereby transcriptional activation can take place. Also PARP1 (poly (ADP-ribose) polymerase 1) is recruited and can trigger the binding of additional regulators to the FOS promoter. PARP1 in FOS regulation functions through directly enhancing ERK-mediated ELK1 phosphorylation (Cohen-Armon et al., 2007).

Once the chromatin remodeling (modification) steps are completed, Mediator can be added by undergoing a phosphorylation-dependent interaction with ELK1, and finally RNA polymerase activity can increase at the FOS promoter (Wang et al., 2005). This has been shown for the EGR1

promoter, and it has been indicated that the process for the FOS promoter is similar (O'Donnell et al., 2012). The main components in transcriptional activation of FOS as an IEG are shown in Figure 2.

(Figure 2)

The immune system as a model

The immune system is a well-studied system where rapid response is essential, and many IEGs have an important role there. The activation of B and T lymphocytes is generally initiated by signaling through the antigen receptor, and it is often regulated by other cell surface proteins such as adhesion molecules, co-stimulatory molecules, and cytokine receptors. The transcription factor products of IEGs play an important role in dictating patterns of expression of downstream, function-related genes. Several studies indicate that a well-known IEG such as EGR1 may be of particular importance in response of the immune system (McMahon and Monroe, 1996), but many other genes are also involved. It has for example been shown that stimulation of airway epithelial cells with house dust mite extract leads to rapid up-regulation of ATF3, EGR1, DUSP1 and FOS, and a later strong up-regulation of JUN (Golebski et al., 2014). Stimulation with a viral double stranded RNA analogue leads to a similar response. Stimulation of mouse bone marrow derived macrophages with LPS, which will activate genes via Toll-like receptors, leads to strong induction of e.g. NR4A1, EGR1, EGR2, JUN, JUNB, FOS and FOSB (Ramirez-Carrozzi et al., 2009). However, the actual picture is sensitive both to the cellular system, how it is stimulated, and how rapid the measurement is done. For example, activation of lymphocytes with concanavalin A and measurement after 4 hours identifies e.g. EGR1, EGR2, EGR3 and ATF3 as up-regulated, but FOS and JUNB as down-regulated (Ellisen et al., 2001). Infection of human epithelial lung cells with influenza virus leads to a strong down-regulation of e.g. FOS, EGR1, EGR2, FOSB, JUN, NR4A1 and NR4A2 after 8 and 24 hours (Tatebe et al., 2010). This shows that the identification of IEGs is sensitive to the experimental conditions.

Characterization of IER gene sets

Although there are both general IEGs that are expressed in almost all cell types, and more cell-type specific IEGs, they are likely to share some key properties. It may be useful to have a good understanding of these properties as general principles of IEG activation and regulation. Shared properties of IEGs can be identified from collections (lists) of genes displaying IEG behavior in various contexts. Most analyses of IEG properties have focused on identified IEGs from experiments for specific processes or pathways, and we will first present such a study in some detail (subsection Identification and analysis of IEGs). This is followed by a more general overview of IEG properties, also largely based on studies of individual systems (subsection General properties of IEG-like genes).

Identification and analysis of IEGs

Several studies have characterized IEGs based on experimental data for specific cell types and conditions. For example, Tullai et al. (2007) have done an extensive analysis of genes induced within four hours after growth factor stimulation, using T98G human glioblastoma cells and PDGF (platelet-derived growth factor). They identified 49 IEGs, 58 delayed primary response genes, and

26 secondary response genes. An analysis of gene ontology showed that the IEGs were enriched in terms for molecular function related to transcriptional regulation, in particular "transcription factor activity" and "DNA binding". However, these terms were not significantly enriched in either delayed primary response or secondary response genes. The immediate-early genes were also highly enriched in the cellular component term "nucleus", but again the term was not enriched in the delayed primary response or secondary response genes. On the other hand, both the delayed primary response and secondary response genes were highly enriched in the cellular component term "extracellular region", but this was not seen for the IEGs. This is consistent with the assumption that many IEGs encode transcription factors that in turn regulate the secondary response genes (Tullai et al., 2007).

Analysis of promoters and upstream regions of the genes showed that the difference in induction between the IEGs and the delayed primary response genes could be caused by a variety of factors, including differences in transcription initiation, elongation, pre-mRNA processing, or mRNA stability. The analysis of human sequences showed that in upstream sequences of the IEGs, four transcription factors were significantly overrepresented; SRF, NF-κB, PAX-3 and KROX. However, for delayed primary response genes no transcription factor was found to be overrepresented (Tullai et al., 2007). Also the analysis was extended with phylogenetic footprinting to identify over-represented binding sites that were conserved in orthologous genomic regions, and this showed that conserved occurrences of binding sites of SRF, NF-κB, CREB (cyclic AMP response element-binding protein) and AP-1 were significantly overrepresented in the upstream regions of IEGs (Tullai et al., 2007).

Comparison of the core promoter sequences of IEGs and the delayed primary response genes with respect to binding sites for general transcription factors indicated that there on average is a significantly higher score for a TATA box (subsection The promoter structure - CpG and TATA) for the IEGs in comparison to delayed primary response genes. Also, it was shown that the IEGs may have a greater tendency to initiate transcription from an initiation site than the delayed primary response genes, indicating that the lag in delayed primary response gene expression could be caused by RNA Pol II (RNA polymerase II) abundance and/or recruitment at target gene promoters, and that the delay in mRNA induction for these genes occurs after the recruitment of RNA Pol II (Tullai et al., 2007).

Comparison of mRNA processing of IEGs and delayed primary response genes showed that there was no significant difference between the splice site characteristics of these groups of genes. But there was a significant difference in both the primary transcript length and exon frequency; the primary transcripts of the IEGs were significantly shorter than the primary transcripts of the delayed primary response genes and contained significantly fewer exons (Tullai et al., 2007).

General properties of IEG-like genes

The promoter structure - CpG and TATA

Many genes in mammalian genomes start transcription from regions of the genome with an elevated content of CpG dinucleotides and G+C base pairs referred to as 'CpG islands'. CpG islands have a high frequency of CpG sites and are typically 300-3000 base pairs long. They have been found

within or close to almost 40% of all promoters of mammalian genes (Deaton and Bird, 2011, Fatemi et al., 2005). Also, the core promoter of eukaryotic genes often includes a short motif around 30 nucleotides before transcription start, known as the TATA-box. During transcription the TATA binding protein (TBP) normally binds to the TATA-box sequence, and this unwinds the DNA. The AT-rich sequence of the TATA-box facilitates easy unwinding (Kutyavin et al., 2000, Yang et al., 2007).

A major class of IEGs has been associated with CpG-island promoters. The promoters of these genes assemble into unstable nucleosomes, and therefore they do not need nucleosome remodeling complexes to facilitate induction from active chromatin. There is also another major class of IEGs with non-CpG-island promoters and stable nucleosomes, which results in dependence on nucleosome remodeling and transcription factors that promote this. However, both classes are induced by the same signaling cascade initiated from Toll-like receptors (Ramirez-Carrozzi et al., 2009).

As already mentioned, promotors of IEGs have more high-affinity TATA boxes than other gene classes. This can play an important role in transcriptional activity at the promoter of IEGs, and high affinity of the TBP binding site may also lead to rapid re-initiation.

Chromatin structure

IEGs have a special chromatin structure which seems to contribute to the rapid activation of transcription. A genome-wide mapping of repressed intergenic and intragenic transcription start sites (TSSs) enriched with active chromatin marks and RNA polymerase II showed strong association with IEGs (Rye et al., 2014). Such promoters are often bivalent, which means that they have both repressive and activating histone modifications. They are therefore silenced, but still poised for rapid activation. An important repressive mark is methylation at histone H3 lysine 27 (H3K27me3), whereas methylation at histone H3 lysine 4 (H3K4me3) is an important activating mark (Bernstein et al., 2006, Spaapen et al., 2013).

It has been shown that histone acetylation remains consistently present both prior to and after stimulation of gene expression, and this generates a constitutively permissive and open promoter structure (Healy et al., 2012, Soloaga et al., 2003). There is a high level of H3K4me3 marks across the promoter region of IEGs, a mark normally found around the transcription start site of actively transcribed genes, as well as H3K36me3 in the coding region, indicating actively transcribed gene bodies. The promoter regions are also enriched in the repressive H3K27me3 mark, creating a bivalent promoter. However, this is different from a silenced promoter with inactive chromatin marks. These are enriched in H3K9me3 and H3K27me2/me3 and are correlated with transcriptional repression (Bernstein et al., 2006, Rosenfeld et al., 2009). It has also been shown that there is a dynamic turnover of histone acetylation by the action of histone acetyl transferases (HATs) and histone deacetylases (HDACs), which affects all K4me3-modified H3s. This is detectable also in the absence of signaling (Edmunds et al., 2008), and it has been shown that a specific HAT (p300/CBP) mediates the dynamic acetylation of IEG regions (Crump et al., 2011). Lysine acetyltransferase p300 transfers an acetyl group to specific histone lysines, and bookmarks the proximal promoter region of IEGs when the transcription is finished, and reactivates it again

following gene induction. Also RNA polymerase II is accumulated and "poised" at the proximal promoter region of IEGs (Byun et al., 2009, Tullai et al., 2007).

Maintenance of histone acetylation seems to be important for IEGs. Crump et al. (2011) have shown that fibroblasts taken from a p300/CBP double knockout mouse display inhibition of signal-induced acetylation of H4K5, K8, K12 and K16 at IEGs. However, for efficient expression of IEGs a high level of acetylation is not enough, and reduction in transcription of such genes as a result of p300 ablation cannot be overcome by pre-acetylating nucleosomes before inhibition.

Also other histone modifications are important. PIM1 kinase phosphorylates H3 at serine 10 (H3S10ph) at the FOSL1 enhancer, and recruits the HAT protein MOF (Zippo et al., 2009). Then MOF promotes H4K16Ac by generating a histone crosstalk and increased recruitment of bromodomain-containing protein BRD4 via interaction with P-TEFb. Enhanced recruitment of P-TEFb is accompanied by release of paused RNA Pol II and continuation of elongation. So H3S10ph stimulates a relay switch, which connects changes in chromatin landscape with transcriptional elongation via P-TEFb (Zippo et al., 2009). Also the modification H3S28ph has been linked to this process (Lau and Cheung, 2011).

It has been shown that poly(ADP-ribosyl)ation is required to modulate chromatin changes, for example at the MYC promoter during emergence from quiescence. Poly(ADP-ribosyl)ation is a post-translational modification found in several types of proteins, and it has an important role in the regulation of chromatin structure and transcription. PARP-1 is a major family member of poly(ADP-ribose)polymerases, participate in the cell cycle reactivation of resting cells by regulating the expression of several IEGs, such as MYC, FOS, JUNB and EGR-1 (Mostocotto et al., 2014). Inhibition of PARP activity along with serum stimulation, by preventing the accumulation of histone H3 phosphoacetylation, damages MYC induction, and this can be a specific chromatin mark for the activation of IEGs (Mostocotto et al., 2014).

Chromatin remodeling

Chromatin can exist in different structural states, and dynamic modification of chromatin structure through 'chromatin remodeling' can be accomplished by covalent histone modifications, utilization of histone variants, DNA methylation and/or by the action of ATP-dependent remodeling complexes. Chromatin remodeling allows proteins of the regulatory transcription machinery access to condensed genomic DNA, and thereby control of gene expression (Teif and Rippe, 2009).

An important factor in chromatin remodeling is remodeling complexes. These use ATP hydrolysis to alter the state of chromatin by moving, ejecting, or restructuring the nucleosome. There are four important families of chromatin remodeling complexes, including the SWI/SNF family, ISWI family, CHD family, and INO80 family remodelers (Clapier and Cairns, 2009).

The assembly of CpG-island promoters into unstable nucleosomes contributes to their independence of chromatin remodeling complexes (SWI/SNF). The unstable nucleosomes, in the absence of transcription factor targeting, are sensitive to acetylation and methylation, although it is possible that expressed transcription factors play an important role in targeting histone modifications (Ramirez-Carrozzi et al., 2009). SWI/SNF-independent genes are in general induced more quickly

than SWI/SNF-dependent genes (Ramirez-Carrozzi et al., 2006). It has also been shown that nucleosomes associated with inducible CpG-island promoters are structurally different from nucleosomes associated with non-CpG-island promoters in unstimulated cells. It is possible that the CpG-island sequence is responsible for the low nucleosome occupancy (Ramirez-Carrozzi et al., 2009).

Most LPS-induced primary response genes are SWI/SNF independent, but some of them show a substantial SWI/SNF dependence. Some of these genes, for activation in LPS-stimulated macrophages, require IRF3 (interferon regulatory factor 3), which is induced by a subset of TLRs (Toll-like receptors) such as TLR4. It has been shown that most primary response genes that require IRF3 for expression in LPS inducted macrophages are SWI/SNF dependent, and these IRF3-dependent primary response genes do in general not have CpG-island promoters (Ramirez-Carrozzi et al., 2009).

There are also SWI/SNF-dependent primary response genes that do not require IRF3 for expression. It has been hypothesized that one or more specialized LPS-induced transcription factors other than IRF3 promote nucleosome remodeling at promoters within this class, contributing to their selective activation (Ramirez-Carrozzi et al., 2009). Ramirez *et al*. have shown that TNFα signaling does not induce IRF3, and may also not directly induce any other transcription factors for nucleosome remodeling in macrophages (Ramirez-Carrozzi et al., 2009), which limits activation to SWI/SNF-independent primary response genes. Therefore, IFN-induced factors might be suitable for the selective activation of SWI/SNF-dependent genes assembled into stable nucleosomes. On the other hand, IFNβ induces transcription via IRFs and STAT proteins, and both of these protein families promote nucleosome remodeling by SWI/SNF complexes. This shows that perhaps some stimuli preferentially induce SWI/SNF independent CpG-island genes during a primary response, but that these stimuli cannot activate transcription factors capable of promoting nucleosome remodeling (Ramirez-Carrozzi et al., 2009).

Initiation of transcription - Transient and sustained signals

Transcription of IEGs is initiated by signaling cascades, and such signals can be either short-term (transient) or long-term (sustained). Depending on the kind of cell type and the duration of signaling, the biological outcome may be different (Murphy and Blenis, 2006). For example, studies with PC12 cells showed that sustained signaling with nerve growth factor (NGF) led to neurite outgrowth in tissue culture, while transient signaling in these cells resulted in proliferation (Marshall, 1995). Both transient and sustained signaling leads to ERK activation in PC12 cells, but corresponding nuclear translocation is associated only with sustained signaling. Nuclear accumulation of active ERK will result in phosphorylation of transcription factors, leading to different outcomes of transient and sustained signaling (Marshall, 1995). ERK-dependent phosphorylation of the FOS protein protects it from degradation and results in cell cycle entry (Fowler et al., 2011, Murphy and Blenis, 2006, Yamamoto et al., 2006).

ERKs in transient versus sustained signaling can regulate PRGs and affect cell fate choices in several ways. For example, angiotensin II-mediated signaling involves heteromeric guanine nucleotide binding protein (G-protein) and β-arrestin. The G-protein dependent pathway produces a transient ERK activation, nuclear accumulation, and activation of IEGs. However, the β-arrestin-

dependent pathway results in a sustained ERK activation and restricts localization to cytosol and endosomes (Shenoy and Lefkowitz, 2005).

Glauser and Schlegel have shown that almost 90% of the genes regulated by sustained signaling were not regulated by transient signaling (Glauser and Schlegel, 2006). Indeed, only a few genes were regulated by transient signaling, while many genes were regulated by sustained signaling, and some genes were regulated by both mechanisms. There were several IEGs (e.g., FOS and EGR1), which were rapidly induced by transient signaling. Both the duration of signaling and cell type context are important for biological responses, and the levels of expression of IEGs might have distinct effects in determining these responses (Damdinsuren et al., 2010, Fowler et al., 2011, Spaapen et al., 2013).

Transcription factors

Regulation of gene expression includes the binding of multiple transcription factors to the regulatory regions of a given gene (Gill, 2001). However, in IEGs the role of TFs is somewhat more unclear. There is no need for *de novo* synthesis of TFs to activate IEGs. On the other hand there are some specific transcription factors such as serum-response factor (SRF), nuclear factor κB (NFκB), cyclic AMP response element-binding protein (CREB) and Zeste-like that are frequently found in the upstream promoter region of IEGs (Fowler et al., 2011, Pintchovski et al., 2009, Tullai et al., 2007). Serum response factor (SRF) belongs to the MADS family of transcription factors, and it is essential for the induction of many IEGs through signaling cascades such as the RAS-MAPK signaling pathway (Yang et al., 2003) and the RhoA actin pathway (Hill et al., 1995).

Selvaraj and Prywes (2004) suggested that TCF and MKL/MRTF family factors might function in an antagonistic fashion, so that SRF target gene regulation and cell fate choices are likely to be determined by the specificity of these cofactors (Lee et al., 2010, Selvaraj and Prywes, 2004). Also some of the IEGs that are SRF targets (e.g., FOS, EGR1 and EGR2) are MKL1 independent, while others like JUNB and FOSL1 (FOS-like 1) are MKL1-dependent targets (Lee et al., 2010, Selvaraj and Prywes, 2004). Lee et al. (2010) showed that some IEGs need just MKLs for serum induction, while other IEGs could be activated by either the TCFs or MKLs (Lee et al., 2010).

The importance of the control of MKL1 activation by TCFs or other factors is clear in megakaryoblastic leukemia, where MKL1 is fused to the RBM15 protein and activated due to constitutive nuclear localization (Cen et al., 2003, Guettler et al., 2008). Phosphorylation of MKL1 inhibits its activity, while SUMO-modification of MKL1 and myocardin has the opposite effect (Nakagawa and Kuzumaki, 2005, Wang et al., 2007).

The role of enhancers and the Mediator complex in regulation of IEGs

An enhancer is a short region of DNA that can be bound by transcription factors to activate gene transcription. Pintchovski et al. (2009) showed that there are both distal and proximal enhancer regions for IEGs. The proximal enhancer contains one or more DNA elements. For example the Zeste-like factor binds to such sites and plays a key role for some IEGs, such as the Arc gene (Pintchovski et al., 2009). Here the distal enhancer has a functional and conserved serum response

element (SRE), this binds SRF and ELK-1, which are important transcription factors for the induction of many IEGs through the ERK signaling pathway (Pintchovski et al., 2009).

It has also been shown that most IEGs are in an epigenetically poised state (Bahrami and Drabløs, 2015). They may be activated through interaction with enhancers, and it has been hypothesized that such enhancers may produce eRNA, which may play a key role in active elongation of transcription as described below.

Mediator is a multi-protein complex that is evolutionarily conserved, and it is an important transcriptional regulator of protein-coding genes by forming an interface between gene-specific activator proteins and the preinitiation complex with RNA Pol II (Malik and Roeder, 2010). In particular, it may mediate long-range interactions between promoters and enhancers, together with cohesin. The Mediator subunit MED23 is very important for regulation of EGR1 in the context of ERK/MAPK signaling through the serum response pathway (Balamotis et al., 2009). MED23 knockout leads to elimination of EGR1 expression in embryonic stem cells (ESCs) with paused RNA Pol II at the promoter, while the same effect was not observed in differentiated fibroblasts (Balamotis et al., 2009). This shows that the mechanism of regulation of IEGs in embryonic stem cells might differ from differentiated cells in a cell type specific manner (Balamotis et al., 2009). A missense mutation in MED23 leads to change in interaction of the Mediator complex with ELK1 and TCF4 and altered regulation of IEGs FOS and JUN. Deregulation of these IEGs was also observed in neurocognitive deficits. This shows that MED23 is important for regulation of IEGs (Hashimoto et al., 2011). Also the CDK8 subunit of Mediator regulates IEGs in response to serum stimulation by enhancing transcription elongation (Galbraith and Espinosa, 2011). After stimulation a CDK8-containing Mediator subcomplex is recruited to the IEG promoters where it functions as a co-activator (Donner et al., 2010). Positive transcription elongation factor, P-TEFb, plays an essential role in the regulation of transcription by pausing of RNA Pol II soon after transcription initiation in eukaryotes (Cheng et al., 2012, Zhou et al., 2012). Signal-dependent CDK8 recruitment to IEGs increases ultimately the recruitment of P-TEFb, so damage to CDK8 results in a decrease of induction of these genes by impacting both RNA Pol II and P-TEFb recruitment (Donner et al., 2010).

The elongation step of transcription

Eukaryote transcription consists of a series of steps. First a preinitiation complex assembles at the promoter, leading to DNA separation and initiation of transcription. After a short initial transcript has formed the process may move into elongation. This elongation continues until the final step, termination, where the transcript and the polymerase are released. However, there may also be pausing of the transcription at the start of the elongation step.

The elongation step of IEGs, and thereby also transcription, seems to be controlled by transcription elongation factors (Fujita et al., 2009). This includes factors such as DSIF (DRB sensitivity-inducing factor), NELF (negative elongation factor) and P-TEFb (positive transcription elongation factor). DRB is a nucleoside analog that inhibits transcription elongation by RNA Pol II. DSIF is a heterodimeric protein complex consisting of the Spt4 and Spt5 subunits, and is essential for cell growth and survival at the single-cell level. DSIF may act as a negative or positive elongation factor according to the phosphorylation state of Spt5 (Komori et al., 2009, Wada et al., 1998, Yamada et

al., 2006). NELF is a DSIF cofactor that consists of four subunits (A, B, C/D and E). P-TEFb is a protein kinase composed of Cdk9 and Cyclin T, and it phosphorylates the C-terminal domain (CTD) of the largest RNA Pol II subunit in a DRB-sensitive manner (Peng et al., 1998).

Transcription elongation factors are necessary for development in higher eukaryotes, and many of the IEGs, such as FOS and JUNB, are controlled by these elongation factors (Aida et al., 2006).

As noted above, NELF and DSIF may pause RNA Pol II at the promoter-proximal regions by binding directly to it, and Spt5 of DSIF binds to the clamp domain of RNA Pol II (Hirtreiter et al., 2010, Martinez-Rucobo et al., 2011). Since the clamp is a flexible domain that tightly holds DNA and RNA (Cramer et al., 2001), any structural changes in this region are likely to have an important influence on elongation kinetics, possibly by affecting the translocation step of the elongation cycle. NELF is also likely to bind to the RNA Pol II clamp (Yamaguchi et al., 2001).

But how do transcription elongation factors regulate overall transcription elongation of IEGs during a specific stimulus? The complex of DSIF/NELF directly acts as a negative regulator complex to pause RNA Pol II at the promoter-proximal regions of IEGs. But during stimulation, RNA Pol II elongation proceeds together with the continuous association of P-TEFb and DSIF as a positive regulator, where P-TEFb allows DSIF to function as an accelerative elongation factor, and NELF to separate from the IEGs (Fujita et al., 2009, Rogatsky and Adelman, 2014).

DSIF requires NELF to induce promoter-proximal pausing. On the other hand, NELF probably requires DSIF to repress transcription fully because NELF only binds to RNA Pol II with low affinity (Yamaguchi et al., 1999). Within the paused RNA Pol II complex, CTD Ser-2 of RNA Pol II is hypophosphorylated, and then P-TEFb phosphorylates CTD Ser-2 of RNA Pol II to repress transcriptional pausing. So, CTD Ser-2 phosphorylation results in dissociation of NELF and the transcription to leave from pausing (Rogatsky and Adelman, 2014, Yamada et al., 2006). The mechanism is illustrated in Figure 3. However, the role of NELF seems to depend upon the type of stimulation. Stable knock-down of NELF by RNAi showed very little effect on activation by EGF, whereas THR-induced activation of the MAP kinase pathway was clearly down-regulated (Fujita et al., 2009).

(Figure 3)

Thus stable NELF knock-down affects transcription of IEGs both directly via RNA Pol II elongation on IEGs as well as indirectly via activation of the ERK1/2 MAP kinase pathway after stimulations such as by TRH. This shows that the regulation of transcription of IEGs by the NELF is both direct and indirect and that it is stimulation-specific (Fujita et al., 2009).

Enhancer RNAs (eRNAs) seem to play an important role in the early transcription elongation step that involves RNA Pol II pausing and release in the IEGs. The eRNAs probably destabilize the association of the DSIF-NELF complex with RNA Pol II and facilitate the transition of paused RNA Pol II into productive elongation by interaction with the NELF complex upon induction of IEGs (Schaukowitch et al., 2014).

Conclusions

IEGs have an important role in several essential cellular systems, for example the immune system, and they are also important in serious diseases like cancer. It is therefore highly relevant to have a good understanding of the properties of IEGs, including gene structure, how they are activated and regulated, and how they affect downstream processes. In this paper we have summarized some key elements of our current understanding of IEGs, including the importance of genetic and epigenetic structure, and the role of poised genes and how IEGs may interact with strong enhancers.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work was funded by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU). A special thanks to Markus Haug, researcher in adaptive immunity at Norwegian University of Science and Technology (NTNU), for his valuable comments on an early version of the manuscript.

References

Aida, M., Chen, Y., Nakajima, K., Yamaguchi, Y., Wada, T., Handa, H., 2006. Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. Mol Cell Biol 26, 6094-6104.

Aitken, S., Magi, S., Alhendi, A.M., Itoh, M., Kawaji, H., Lassmann, T., et al., 2015. Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response. PLoS Comput Biol 11, e1004217.

Avraham, R., Sas-Chen, A., Manor, O., Steinfeld, I., Shalgi, R., Tarcic, G., et al., 2010. EGF decreases the abundance of microRNAs that restrain oncogenic transcription factors. Sci Signal 3, ra43.

Bae, M.H., Jeong, C.H., Kim, S.H., Bae, M.K., Jeong, J.W., Ahn, M.Y., et al., 2002. Regulation of Egr-1 by association with the proteasome component C8. Biochim Biophys Acta 1592, 163-167.

Bahrami, S., Drabløs, F., 2015. Identification and analysis of genes in immediate-early response processes. Submitted.

Bahrami, S., Ehsani, R., Drabløs, F., 2015. A property-based analysis of human transcription factors. BMC Res Notes 8, 82.

Balamotis, M.A., Pennella, M.A., Stevens, J.L., Wasylyk, B., Belmont, A.S., Berk, A.J., 2009. Complexity in transcription control at the activation domain-mediator interface. Sci Signal 2, ra20.

Bebien, M., Salinas, S., Becamel, C., Richard, V., Linares, L., Hipskind, R.A., 2003. Immediate-early gene induction by the stresses anisomycin and arsenite in human osteosarcoma cells involves MAPK cascade signaling to Elk-1, CREB and SRF. Oncogene 22, 1836-1847.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., et al., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315-326.

Bertelsen, T., Iversen, L., Riis, J.L., Arthur, J.S., Bibby, B.M., Kragballe, K., et al., 2011. The role of mitogen- and stress-activated protein kinase 1 and 2 in chronic skin inflammation in mice. Exp Dermatol 20, 140-145.

Byun, J.S., Wong, M.M., Cui, W., Idelman, G., Li, Q., De Siervi, A., et al., 2009. Dynamic bookmarking of primary response genes by p300 and RNA polymerase II complexes. Proc Natl Acad Sci U S A 106, 19286-19291.
Cen, B., Selvaraj, A., Burgess, R.C., Hitzler, J.K., Ma, Z., Morris, S.W., et al., 2003. Megakaryoblastic leukemia 1, a potent transcriptional coactivator for serum response factor (SRF), is required for serum induction of SRF target genes. Mol Cell Biol 23, 6597-6608.
Cheng, B., Li, T., Rahl, P.B., Adamson, T.E., Loudas, N.B., Guo, J., et al., 2012. Functional association of Gdown1 with RNA polymerase II poised on human genes. Mol Cell 45, 38-50.
Ciccarelli, A., Giustetto, M., 2014. Role of ERK signaling in activity-dependent modifications of histone proteins. Neuropharmacology 80, 34-44.
Clapier, C.R., Cairns, B.R., 2009. The biology of chromatin remodeling complexes. Annu Rev Biochem 78, 273-304.
Cohen-Armon, M., Visochek, L., Rozensal, D., Kalal, A., Geistrikh, I., Klein, R., et al., 2007. DNA-independent PARP-1 activation by phosphorylated ERK2 increases Elk1 activity: a link to histone acetylation. Mol Cell 25, 297-308.
Cramer, P., Bushnell, D.A., Kornberg, R.D., 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science 292, 1863-1876.
Crump, N.T., Hazzalin, C.A., Bowers, E.M., Alani, R.M., Cole, P.A., Mahadevan, L.C., 2011. Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP. Proc Natl Acad Sci U S A 108, 7814-7819.
Damdinsuren, B., Zhang, Y., Khalil, A., Wood, W.H., 3rd, Becker, K.G., Shlomchik, M.J., et al., 2010. Single round of antigen receptor signaling programs naive B cells to receive T cell help. Immunity 32, 355-366.
Deaton, A.M., Bird, A., 2011. CpG islands and the regulation of transcription. Genes Dev 25, 1010-1022.
Donner, A.J., Ebmeier, C.C., Taatjes, D.J., Espinosa, J.M., 2010. CDK8 is a positive regulator of transcriptional elongation within the serum response network. Nat Struct Mol Biol 17, 194-201.
Drobic, B., Perez-Cadahia, B., Yu, J., Kung, S.K., Davie, J.R., 2010. Promoter chromatin remodeling of immediate-early genes is mediated through H3 phosphorylation at either serine 28 or 10 by the MSK1 multi-protein complex. Nucleic Acids Res 38, 3196-3208.
Edmunds, J.W., Mahadevan, L.C., Clayton, A.L., 2008. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J 27, 406-420.
Ellisen, L.W., Palmer, R.E., Maki, R.G., Truong, V.B., Tamayo, P., Oliner, J.D., et al., 2001. Cascades of transcriptional induction during human lymphocyte activation. Eur J Cell Biol 80, 321-328.
Fatemi, M., Pao, M.M., Jeong, S., Gal-Yam, E.N., Egger, G., Weisenberger, D.J., et al., 2005. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res 33, e176.
Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., et al., 2014. Pfam: the protein families database. Nucleic Acids Res 42, D222-230.
Flouriot, G., Huet, G., Demay, F., Pakdel, F., Boujrad, N., Michel, D., 2014. The actin/MKL1 signalling pathway influences cell growth and gene expression through large-scale chromatin reorganization and histone post-translational modifications. Biochem J 461, 257-268.
Fowler, T., Sen, R., Roy, A.L., 2011. Regulation of primary response genes. Mol Cell 44, 348-360.
Fujita, T., Piuz, I., Schlegel, W., 2009. Negative elongation factor NELF controls transcription of immediate early genes in a stimulus-specific manner. Exp Cell Res 315, 274-284.

Galbraith, M.D., Espinosa, J.M., 2011. Lessons on transcriptional control from the serum response network. Curr Opin Genet Dev 21, 160-166.
Gill, G., 2001. Regulation of the initiation of eukaryotic transcription. Essays Biochem 37, 33-43.
Glauser, D.A., Schlegel, W., 2006. Mechanisms of transcriptional regulation underlying temporal integration of signals. Nucleic Acids Res 34, 5175-5183.
Golebski, K., Luiten, S., van Egmond, D., de Groot, E., Roschmann, K.I., Fokkens, W.J., et al., 2014. High degree of overlap between responses to a virus and to the house dust mite allergen in airway epithelial cells. PLoS One 9, e87768.
Gomard, T., Jariel-Encontre, I., Basbous, J., Bossis, G., Moquet-Torcy, G., Piechaczyk, M., 2008. Fos family protein degradation by the proteasome. Biochem Soc Trans 36, 858-863.
Greenberg, M.E., Ziff, E.B., 1984. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. Nature 311, 433-438.
Guettler, S., Vartiainen, M.K., Miralles, F., Larijani, B., Treisman, R., 2008. RPEL motifs link the serum response factor cofactor MAL but not myocardin to Rho signaling via actin binding. Mol Cell Biol 28, 732-742.
Hashimoto, S., Boissel, S., Zarhrate, M., Rio, M., Munnich, A., Egly, J.M., et al., 2011. MED23 mutation links intellectual disability to dysregulation of immediate early gene expression. Science 333, 1161-1163.
Hazzalin, C.A., Mahadevan, L.C., 2005. Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: analysis at c-fos and c-jun. PLoS Biol 3, e393.
Healy, S., Khan, P., Davie, J.R., 2013. Immediate early response genes and cell transformation. Pharmacol Ther 137, 64-77.
Healy, S., Khan, P., He, S., Davie, J.R., 2012. Histone H3 phosphorylation, immediate-early gene expression, and the nucleosomal response: a historical perspective. Biochem Cell Biol 90, 39-54.
Herschman, H.R., 1991. Primary response genes induced by growth factors and tumor promoters. Annu Rev Biochem 60, 281-319.
Hill, C.S., Wynne, J., Treisman, R., 1995. The Rho family GTPases RhoA, Rac1, and CDC42Hs regulate transcriptional activation by SRF. Cell 81, 1159-1170.
Hirtreiter, A., Damsma, G.E., Cheung, A.C., Klose, D., Grohmann, D., Vojnic, E., et al., 2010. Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. Nucleic Acids Res 38, 4040-4051.
Khan, D.H., Davie, J.R., 2013. HDAC inhibitors prevent the induction of the immediate-early gene FOSL1, but do not alter the nucleosome response. FEBS Lett 587, 1510-1517.
Knoll, B., Nordheim, A., 2009. Functional versatility of transcription factors in the nervous system: the SRF paradigm. Trends Neurosci 32, 432-442.
Komori, T., Inukai, N., Yamada, T., Yamaguchi, Y., Handa, H., 2009. Role of human transcription elongation factor DSIF in the suppression of senescence and apoptosis. Genes Cells 14, 343-354.
Kukushkin, A.N., Svetlikova, S.B., Pospelov, V.A., 2005. [Effect of anisomycin on activation of early response genes c-fos, c-jun, Egr-1 in cells transformed by E1A and cHa-ras oncogenes]. Mol Biol (Mosk) 39, 80-88.
Kutyavin, I.V., Afonina, I.A., Mills, A., Gorn, V.V., Lukhtanov, E.A., Belousov, E.S., et al., 2000. 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. Nucleic Acids Res 28, 655-661.
Lau, P.N., Cheung, P., 2011. Histone code pathway involving H3 S28 phosphorylation and K27 acetylation activates transcription and antagonizes polycomb silencing. Proc Natl Acad Sci U S A 108, 2801-2806.
Lee, S.M., Vasishtha, M., Prywes, R., 2010. Activation and repression of cellular immediate early genes by serum response factor cofactors. J Biol Chem 285, 22036-22049.

Lewin, B., 2004. Genes VIII. Pearson Prentice Hall, New Jersey.
Li, Q.J., Yang, S.H., Maeda, Y., Sladek, F.M., Sharrocks, A.D., Martins-Green, M., 2003. MAP kinase phosphorylation-dependent activation of Elk-1 leads to activation of the co-activator p300. EMBO J 22, 281-291.
Liu, J., Grogan, L., Nau, M.M., Allegra, C.J., Chu, E., Wright, J.J., 2001. Physical interaction between p53 and primary response gene Egr-1. Int J Oncol 18, 863-870.
Macdonald, N., Welburn, J.P., Noble, M.E., Nguyen, A., Yaffe, M.B., Clynes, D., et al., 2005. Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. Mol Cell 20, 199-211.
Malik, S., Roeder, R.G., 2010. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat Rev Genet 11, 761-772.
Marshall, C.J., 1995. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. Cell 80, 179-185.
Martinez-Rucobo, F.W., Sainsbury, S., Cheung, A.C., Cramer, P., 2011. Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. EMBO J 30, 1302-1310.
McMahon, S.B., Monroe, J.G., 1996. The role of early growth response gene 1 (egr-1) in regulation of the immune response. J Leukoc Biol 60, 159-166.
Miralles, F., Posern, G., Zaromytidou, A.I., Treisman, R., 2003. Actin dynamics control SRF activity by regulation of its coactivator MAL. Cell 113, 329-342.
Morgan, J.I., Curran, T., 1991. Stimulus-transcription coupling in the nervous system: involvement of the inducible proto-oncogenes fos and jun. Annu Rev Neurosci 14, 421-451.
Mostocotto, C., Carbone, M., Battistelli, C., Ciotti, A., Amati, P., Maione, R., 2014. Poly(ADP-ribosyl)ation is required to modulate chromatin changes at c-MYC promoter during emergence from quiescence. PLoS One 9, e102575.
Murphy, L.O., Blenis, J., 2006. MAPK signal specificity: the right place at the right time. Trends Biochem Sci 31, 268-275.
Mylona, A., Nicolas, R., Maurice, D., Sargent, M., Tuil, D., Daegelen, D., et al., 2011. The essential function for serum response factor in T-cell development reflects its specific coupling to extracellular signal-regulated kinase signaling. Mol Cell Biol 31, 267-276.
Nakagawa, K., Kuzumaki, N., 2005. Transcriptional activity of megakaryoblastic leukemia 1 (MKL1) is repressed by SUMO modification. Genes Cells 10, 835-850.
O'Donnell, A., Odrowaz, Z., Sharrocks, A.D., 2012. Immediate-early gene activation by the MAPK pathways: what do and don't we know? Biochem Soc Trans 40, 58-66.
Ozanne, B.W., Spence, H.J., McGarry, L.C., Hennigan, R.F., 2006. Invasion is a genetic program regulated by transcription factors. Curr Opin Genet Dev 16, 65-70.
Peng, J., Zhu, Y., Milton, J.T., Price, D.H., 1998. Identification of multiple cyclin subunits of human P-TEFb. Genes Dev 12, 755-762.
Pintchovski, S.A., Peebles, C.L., Kim, H.J., Verdin, E., Finkbeiner, S., 2009. The serum response factor and a putative novel transcription factor regulate expression of the immediate-early gene Arc/Arg3.1 in neurons. J Neurosci 29, 1525-1537.
Posern, G., Treisman, R., 2006. Actin' together: serum response factor, its cofactors and the link to signal transduction. Trends Cell Biol 16, 588-596.
Price, M.A., Rogers, A.E., Treisman, R., 1995. Comparative analysis of the ternary complex factors Elk-1, SAP-1a and SAP-2 (ERP/NET). EMBO J 14, 2589-2601.
Raman, M., Chen, W., Cobb, M.H., 2007. Differential regulation and properties of MAPKs. Oncogene 26, 3100-3112.
Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., et al., 2009. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell 138, 114-128.

Ramirez-Carrozzi, V.R., Nazarian, A.A., Li, C.C., Gore, S.L., Sridharan, R., Imbalzano, A.N., et al., 2006. Selective and antagonistic functions of SWI/SNF and Mi-2beta nucleosome remodeling complexes during an inflammatory response. Genes Dev 20, 282-296.

Rogatsky, I., Adelman, K., 2014. Preparing the first responders: building the inflammatory transcriptome from the ground up. Mol Cell 54, 245-254.

Rosenfeld, J.A., Wang, Z., Schones, D.E., Zhao, K., DeSalle, R., Zhang, M.Q., 2009. Determination of enriched histone modifications in non-genic portions of the human genome. BMC Genomics 10, 143.

Rye, M., Sandve, G.K., Daub, C.O., Kawaji, H., Carninci, P., Forrest, A.R., et al., 2014. Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. BMC Genomics 15, 120.

Sawicka, A., Hartl, D., Goiser, M., Pusch, O., Stocsits, R.R., Tamir, I.M., et al., 2014. H3S28 phosphorylation is a hallmark of the transcriptional response to cellular stress. Genome Res 24, 1808-1820.

Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., Kim, T.K., 2014. Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell 56, 29-42.

Selvaraj, A., Prywes, R., 2004. Expression profiling of serum inducible genes identifies a subset of SRF target genes that are MKL dependent. BMC Mol Biol 5, 13.

Selvaraj, N., Kedage, V., Hollenhorst, P.C., 2015. Comparison of MAPK specificity across the ETS transcription factor family identifies a high-affinity ERK interaction required for ERG function in prostate cells. Cell Commun Signal 13, 12.

Serrat, N., Sebastian, C., Pereira-Lopes, S., Valverde-Estrella, L., Lloberas, J., Celada, A., 2014. The response of secondary genes to lipopolysaccharides in macrophages depends on histone deacetylase and phosphorylation of C/EBPbeta. J Immunol 192, 418-426.

Shaw, P.E., Saxton, J., 2003. Ternary complex factors: prime nuclear targets for mitogen-activated protein kinases. Int J Biochem Cell Biol 35, 1210-1226.

Shenoy, S.K., Lefkowitz, R.J., 2005. Angiotensin II-stimulated signaling through G proteins and beta-arrestin. Sci STKE 2005, cm14.

Soloaga, A., Thomson, S., Wiggin, G.R., Rampersaud, N., Dyson, M.H., Hazzalin, C.A., et al., 2003. MSK2 and MSK1 mediate the mitogen- and stress-induced phosphorylation of histone H3 and HMG-14. EMBO J 22, 2788-2797.

Spaapen, F., van den Akker, G.G., Caron, M.M., Prickaerts, P., Rofel, C., Dahlmans, V.E., et al., 2013. The immediate early gene product EGR1 and polycomb group proteins interact in epigenetic programming during chondrogenesis. PLoS One 8, e58083.

Tatebe, K., Zeytun, A., Ribeiro, R.M., Hoffmann, R., Harrod, K.S., Forst, C.V., 2010. Response network analysis of differential gene expression in human epithelial lung cells during avian influenza infections. BMC Bioinformatics 11, 170.

Teif, V.B., Rippe, K., 2009. Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. Nucleic Acids Res 37, 5641-5655.

Tullai, J.W., Schaffer, M.E., Mullenbrock, S., Sholder, G., Kasif, S., Cooper, G.M., 2007. Immediate-early and delayed primary response genes are distinct in function and genomic architecture. J Biol Chem 282, 23981-23995.

Vartiainen, M.K., Guettler, S., Larijani, B., Treisman, R., 2007. Nuclear actin regulates dynamic subcellular localization and activity of the SRF cofactor MAL. Science 316, 1749-1752.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., et al., 1998. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev 12, 343-356.

Wang, G., Balamotis, M.A., Stevens, J.L., Yamaguchi, Y., Handa, H., Berk, A.J., 2005. Mediator requirement for both recruitment and postrecruitment steps in transcription initiation. Mol Cell 17, 683-694.

Wang, J., Li, A., Wang, Z., Feng, X., Olson, E.N., Schwartz, R.J., 2007. Myocardin sumoylation transactivates cardiogenic genes in pluripotent 10T1/2 fibroblasts. Mol Cell Biol 27, 622-632.

Winkles, J.A., 1998. Serum- and polypeptide growth factor-inducible gene expression in mouse fibroblasts. Prog Nucleic Acid Res Mol Biol 58, 41-78.

Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T., Handa, H., 2006. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. Mol Cell 21, 227-237.

Yamaguchi, Y., Filipovska, J., Yano, K., Furuya, A., Inukai, N., Narita, T., et al., 2001. Stimulation of RNA polymerase II elongation by hepatitis delta antigen. Science 293, 124-127.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., et al., 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell 97, 41-51.

Yamamoto, T., Ebisuya, M., Ashida, F., Okamoto, K., Yonehara, S., Nishida, E., 2006. Continuous ERK activation downregulates antiproliferative genes throughout G1 phase to allow cell-cycle progression. Curr Biol 16, 1171-1182.

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, 52-65.

Yang, S.H., Sharrocks, A.D., 2006. Convergence of the SUMO and MAPK pathways on the ETS-domain transcription factor Elk-1. Biochem Soc Symp, 121-129.

Yang, S.H., Sharrocks, A.D., Whitmarsh, A.J., 2003. Transcriptional regulation by the MAP kinase signaling cascades. Gene 320, 3-21.

Yasuda, T., Kurosaki, T., 2008. Regulation of lymphocyte fate by Ras/ERK signals. Cell Cycle 7, 3634-3640.

Zhang, F., Lin, M., Abidi, P., Thiel, G., Liu, J., 2003. Specific interaction of Egr1 and c/EBPbeta leads to the transcriptional activation of the human low density lipoprotein receptor gene. J Biol Chem 278, 44246-44254.

Zhang, H.M., Li, L., Papadopoulou, N., Hodgson, G., Evans, E., Galbraith, M., et al., 2008. Mitogen-induced recruitment of ERK and MSK to SRE promoter complexes by ternary complex factor Elk-1. Nucleic Acids Res 36, 2594-2607.

Zhou, Q., Li, T., Price, D.H., 2012. RNA polymerase II elongation control. Annu Rev Biochem 81, 119-143.

Zippo, A., Serafini, R., Rocchigiani, M., Pennacchini, S., Krepelova, A., Oliviero, S., 2009. Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. Cell 138, 1122-1136.

Figure legends

Figure 1 - Important signaling pathways.

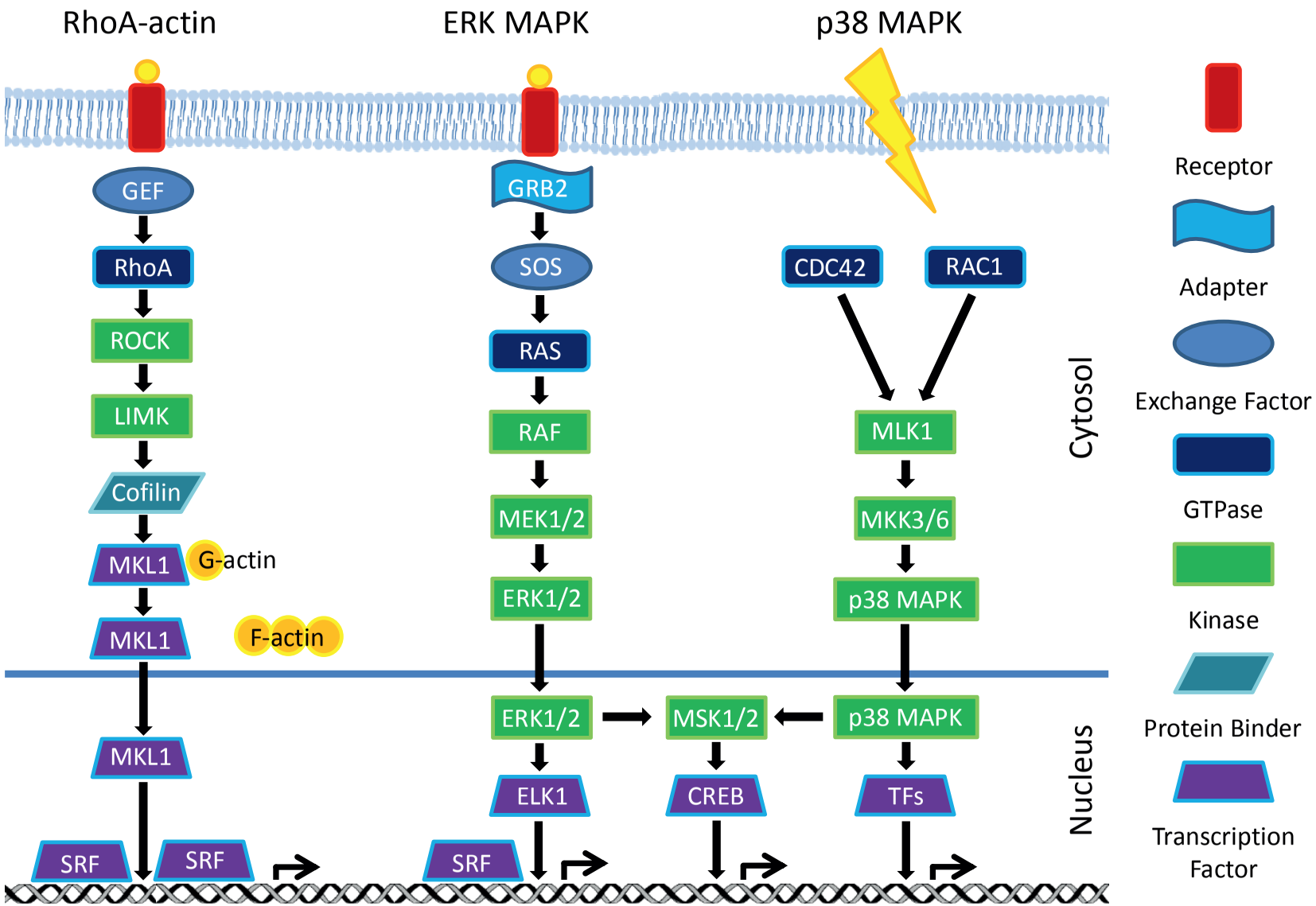
The RhoA-actin, ERK-MAPK and p38-MAPK pathways initiated by different external stimuli are shown. RhoA-actin and ERK are in particular activated by mitogenic stimuli such as growth factors and hormones while p38 is activated by stress stimuli. These pathways will also initiate chromatin modifications. The pathways are simplified, and only selected components are shown. The figure is based on data from several sources, in particular Healy et al. (2013).

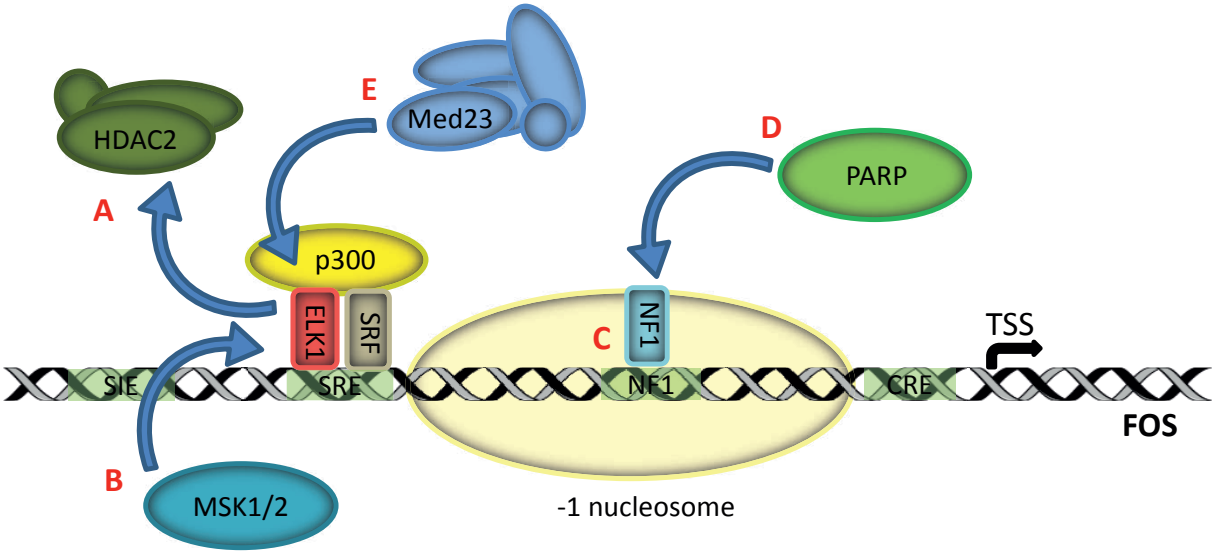
Figure 2 - Molecular events during FOS promoter activation.

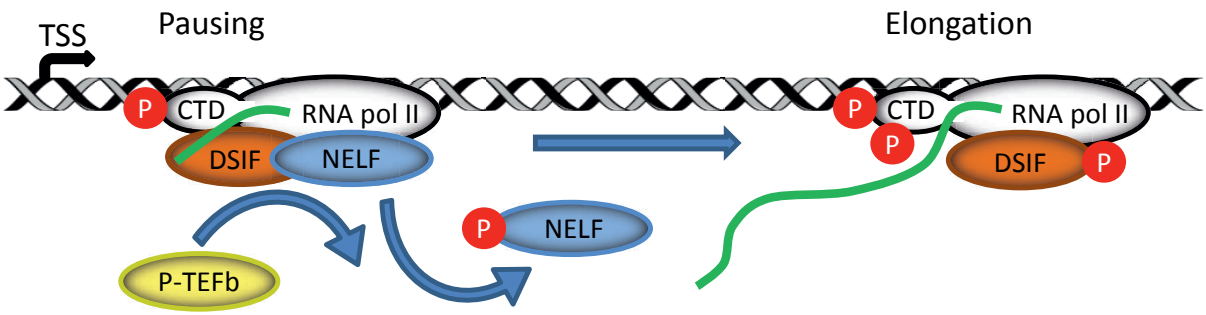
The promoter with pre-bound SRF, EKL1 and p300 is in a poised condition. ELK1 is maintained in an inactive form via SUMO-modification, and this permits recruitment of the repressive modifier HDAC2. During ERK pathway activation loss of SUMO-modification and HDAC2 from ELK1 (A) leads to recruitment of MSKs to the promoter (B). This promotes histone modification and the -1 nucleosome becomes acetylated, which facilitates NF1 recruitment (C). The NF1 then recruits PARP, which will open up for recruitment of other chromatin remodeling complexes (D). Then ELK1 recruits the Mediator complex. This enables basal transcription factors and RNA polymerase, and initiation of transcription (E). See the text for more details. The figure is adapted from O'Donnell et al. (2012)

Figure 3 - A model of stimulation-specific activation of IEG transcription.

Transcription starts with initiation at the transcription start site (TSS). The DSIF/NELF complex then directly stalls RNA Pol II at the promoter-proximal regions of IEGs. After stimulation, P-TEFb activates DSIF as an accelerative elongation factor and NELF to detach from the promoter, and this reactivates the transcription. NELF also stimulates directly or indirectly the expression of genes coding for factors which maintain TRH-dependent activation of the ERK1/2 MAP kinase pathway. The figure is adapted from Fujita et al. (2009).


RhoA-actin
ERK MAPK
p38 MAPK
GEF
RhoA
ROCK
LIMK
Cofilin
MKL1
G-actin
MKL1
F-actin
MKL1
SRF
SRF
GRB2
SOS
RAS
RAF
MEK1/2
ERK1/2
ERK1/2
ELK1
SRF
CDC42
RAC1
MLK1
MKK3/6
p38 MAPK
MSK1/2
p38 MAPK
CREB
TFs
Cytosol
Nucleus
Receptor
Adapter
Exchange Factor
GTPase
Kinase
Protein Binder
Transcription Factor


HDAC2
E
Med23
D
PARP
A
p300
ELK1
SRF
C
NF1
TSS
SIE
SRE
NF1
CRE
FOS
B
MSK1/2
-1 nucleosome


TSS
Pausing
Elongation
P
CTD
RNA pol II
DSIF
NELF
P
CTD
P
RNA pol II
DSIF
P
P
NELF
P-TEFb

Paper IV

Identification and Analysis of Genes in Immediate-Early Response Processes

Shahram Bahrami[1,2], Finn Drabløs[1*]

[1] Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

[2] St. Olavs Hospital, Trondheim University Hospital, NO-7006 Trondheim, Norway

[*]Corresponding author

Email: finn.drablos@ntnu.no (FD)

Short title: Genes in Immediate-Early Response Processes