THE UNIVERSITY
of EDINBURGH

Conserved temporal ordering of promoter activation implicates common mechanisms governing the immediate early response across cell types and stimuli

Annalaura Vacca

Doctor of Philosophy

The University of Edinburgh 2018

THE UNIVERSITY *of* EDINBURGH

Declaration of Authorship

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. Work done in collaboration with, or with assistance of, others, is indicated as such.

Signed: 

Date: 2/12/2019

To my mum, Simonetta

Abstract

The promoters of immediate early genes (IEGs) are rapidly activated in response to an external stimulus. These genes, also known as primary response genes, have been identified in a range of cell types, under diverse extracellular signals and using varying experimental protocols. Genomic dissection on a case-by-case basis has not resulted in a comprehensive catalogue of IEGs. I completed a rigorous meta-analysis of eight genome-wide FANTOM5 CAGE (cap analysis of gene expression) time-course datasets, and it revealed successive waves of promoter activation in IEGs, recapitulating known relationships between cell types and stimuli. I found a set of 57 (42 protein-coding) candidate IEGs possessing promoters that consistently drive a rapid but transient increase in expression following external stimulation. These genes show significant enrichment for known IEGs reported previously, pathways associated with the immediate early response, and include a number of non-coding RNAs with roles in proliferation and differentiation. There was strong conservation of the ordering of activation for these genes, such that 77 pairwise promoter activation orderings were conserved. Leveraging comprehensive CAGE time series data across cell types, I also observed extensive alternative promoter usage by such genes, which is likely to hinder their discovery from previous, smaller-scale studies. The common activation ordering of the core set of early-responding genes I identified may indicate conserved underlying regulatory mechanisms. By contrast, the considerably larger number of transiently activated genes that are specific to each cell type and stimulus illustrates the breadth of the primary response.

Lay summary

Human cells respond to external events through the transient and rapid activation of a class of genes, the basic units of heredity, called immediate-early genes (IEGs). The same set of IEGs is often activated by the majority of the stimuli in different cell types. However, the IEGs can be started from different points of their sequence in the DNA, called promoters, producing appropriate stimulus-specific responses. In normal and healthy cells, the activation of IEGs peaks and returns to a normal level after a few hours of stimulation. This mechanism is responsible for the execution of important cellular processes, such as becoming specialized to perform specific functions, growing and dividing, and responding to infectious things, such as bacteria and virus. However, the IEGs are often continuously activated in altered conditions, such as cancer, and their characterization constitutes a resource for the design of new therapies. A group of scientists recently started a project, the FANTOM5 project, which has provided a rich amount of data that has captured the activation over time of the IEGs and is well suited to study the preference of different cells for using specific alternative initiation points.

My thesis involved improving a recently developed computational tool, and to identify a set of genes which are rapidly and transiently activated in eight FANTOM5 human time course datasets. Many of them are published IEGs while the others are candidate IEGs. Furthermore, I implemented a method to analyse the common temporal order of gene activation and the different initiation locations of the IEGs across different cell types and stimuli. In this thesis, I discussed a list of 57 known and candidate IEGs, the most interesting being the gene *XBP1,* which is an important gene because it is involved in protecting cells from different sources of damage. I also looked their functional interactions as well as their promoter changes during time.

Abbreviation list

Abbreviation	Description
bp	Base pair
CAGE	Cap Analysis of Gene Expression
CI	Confidence Interval
CTSS	CAGE tag starting site
DPI	Decomposition based Peak Identification
ECDF	Empirical Cumulative Distribution Functions. An estimate of the cumulative distribution function that generated the points in the sample
ER	Endoplasmic Reticulum
EST	Expressed sequence tag
FDR	False Discovery Rate. i.e. the corrected p-values for multiple comparisons
GO	Gene Ontology
HPC	Haematopoietic Progenitor Cells
HSP1	Mitochondrial heavy strand promoter 1
HSP2	Mitochondrial heavy strand promoter 2
ICA	Independent Component Analysis
IEG	Immediate Early Gene
IER	Immediate Early Response
IEP	Immediate Early Protein
JS divergence	Jensen-Shannon measure to quantify how dissimilar are two distributions
KL	Kullback-Leibler measure of divergence between two distributions
KS test	Kolmogorov-Smirnov test
log2FC	Log2 fold change. log-ratio of TSS's expression values at time of peaking and time zero
LSP	Mitochondrial light strand promoter
MAPK	Mitogen-activated protein kinase
MCF7_EGF1	MCF7 breast cancer cells treated with epidermal growth factor serum
MCF7_HRG	MCF7 breast cancer cells treated with Heregulin hormone
mtRNAP	Mitochondrial single-subunit RNA polymerase
NUMT	Nuclear Mitochondrial DNA
OR	Odd ratio
PAC_FGF2	Aortic Smooth Muscle Cells treated with fibroblast growth factor
PAC_IL1B	Aortic Smooth Muscle Cells treated with Interleukin-1beta

PEC_VEGF	Primary lymphatic endothelial cells treated with vascular endothelial growth factor
PMDM_LPS	Macrophage response to LPS dataset
PMSC_MIX	Mesenchymal stem cells differentiation dataset
pv	P-value
SAOS2_OST	Osteosarcoma Stem Cell calcification dataset
SC	Stem Cell
TC	CAGE tag clusters
TF	Transcription Factor
TFbs	Transcription Factor binding site
tp	Time of peaking. Peak model estimated parameter
TPM	Tag Per Million
TSS	Transcription Start Site
tssQTL	TSS-associated Quantitative Trait Loci. Genetic variations associated to changes in promoter shape
UPR	Unfolded Protein Response

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Colin Semple and Stuart Aitken, for their continuous support during my work, for their patience, motivation, and immense knowledge. In addition, I want to thank my dad Umberto, my sister Eleonora and my grandma Bianca, who always believed in me and encouraged me; my beautiful and super cute niece Isabella who turns every moment of sadness and despair in a big big smile; my auntie Monica, my beloved cousins Beatrice, Arianna and Gabriella, who always supported me; and all my friends spread across too many countries who never forgot about me despite the distance, especially Melissa, Flavia and Valentina. A big thank to Joanna for being an amazing flatmate especially during the last months of the writing, and to Victor, Toby, Lana, Cathy and Nefeli, who spent some of their time and knowledge to help me with my thesis.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. The Immediate Early Response (IER)

To function correctly, cells require that the appropriate genes are expressed and translated into proteins at the right time and in the right concentration. Regulation of gene expression is one of the key mechanisms to guarantee that the right level of gene products is produced and depend on the activation of signal-transduction pathways which are mostly stimulus specific (Bahrami and Drabløs 2016). Most cell types, including innate immune system cells, cancer cells and embryonic stem cells react to a variety of stimuli, including mitogens, differentiation agents, cytokines or infection signalling (Fowler, Sen et al. 2011) that induce long term changes in cellular phenotypes in a stimulus and cell specific way. The first cellular processes started by extracellular stimuli happen very quickly and their prompt timing is fundamental for the correct initiation of following cellular process and the survival of the organism. These processes are known as immediate early response (IER) processes. The first genes to be activated and transcribed are now known as immediate early genes (IEGs) (Bahrami and Drabløs 2016).

1.2. Immediate early genes (IEGs)

IEGs, first described in 1983, were believed to be activated during the acquisition of competence by quiescent cells to progression through the cell cycle on treatment with platelet-derived growth factor (PDGF). They were referred to as ‘competence genes’ (Cochran, Reffel et al. 1983, Curran and Morgan 1995). Their name was later changed to immediate early genes, when it became clear that they were activated not only by growth factors but also by a plethora of different external stimuli. The name was inspired by the previously studied ‘viral immediate early genes’, a class of rapidly activated viral genes which do not require *de novo* protein synthesis (Curran and Morgan 1995).

1.1.1. Common features of IEGs

IEGs are characterized by rapid and transient activation of their transcription in the first few hours after stimulation as they do not require *de novo* protein synthesis, indeed they are transcribed even in the presence of protein synthesis inhibitors (Herschman 1991). However, the activation of many IEGs is known to require the binding of specific TFs to particular binding motifs, such as TATA-boxes, serum-response factor (SRF), cyclic AMP response element-binding protein (CREB) and nuclear factor κB (NFκB) motifs, that are frequently found in the upstream promoter regions of IEGs (Tullai, Schaffer et al. 2007, Fowler, Sen et al. 2011, Bahrami and Drabløs 2016). IEGs are short in length in comparison with other genes (19 kb versus 58 kb, on average) and often encode TFs involved in secondary waves of cell activation, secreted proteins and enzymes. Many IEGs are known to have important roles in the regulation of important biological processes such as cell proliferation, differentiation and stress responses, are often found to be deregulated in developmental disease and cancer (Healy, Khan et al. 2013, Lee and Young 2013) and are therefore often well-studied genes. Arner et al. (2015) put together a list of 231 potential human and mouse immediate early genes from the literature (complete list of genes and related literature in supplementary table S5 in (Arner, Daub et al. 2015)). All these genes were characterized by early transient upregulation but only a few experiments included confirmation of IEG status by protein synthesis blockage (Arner, Daub et al. 2015).

1.1.2. Kinetics of IEGs and delayed IEGs induction

Previous literature defined two groups of genes which are induced after cell stimulation: IEGs, also termed primary response genes, and secondary response genes. The primary response genes are thought to be activated immediately after stimulation, peaking at about 30 minutes, as they do not require protein synthesis, and their products, mostly TFs, are thought to activate secondary response genes. However, as reported by Tullai et al. (2007), a large fraction of previously considered secondary response genes, such as *DKK1* and *ESDN,* are expressed in presence of protein synthesis

inhibitors, such as cycloheximide, and are therefore delayed IEGs. The delayed IEGs differ from traditional IEGs in both their gene architecture and functions: they do not prevalently encode for transcriptional regulators and their promoters are not enriched for TATA boxes and other TFbs common among known IEGs. However, they share the same expression profile: a transient and rapid upregulation followed by a return to basal expression. Delayed IEGs usually peak later than IEGs, within 6 hours after cell stimulation (Wick, Burger et al. 1994, Dixon, Sharma et al. 1996, Freter, Alberta et al. 1996, Tullai, Schaffer et al. 2007). Their existence has been recognised in several studies involving stimulation with growth factors in different biological systems, such as human lung fibroblasts (Wick, Burger et al. 1994), rat arterial smooth muscle cells (Dixon, Sharma et al. 1996), and mouse 3T3 cells (Freter, Alberta et al. 1996). Tullai et al. performed microarray gene expression analysis of T98G human glioblastoma cells at 0.5, 1, 2, 3, 4, 5, and 6 h of PDGF treatment. They found 133 genes induced within 4 h of growth factor stimulation, among them 49 known IEGs and 58 delayed IEGs were not inhibited by cycloheximide, and 26 secondary response genes were not expressed when the protein synthesis inhibitor was present. After analysis, their results suggest that the lag in induction of delayed IEGs compared with IEG was caused by delays in both transcription initiation and following elongation and processing (Tullai, Schaffer et al. 2007).

1.1.3. Best studied IEGs

The best characterized group of IEGs includes *FOS*, *JUN* and *MYC*, which were originally found to be homologous to retroviral oncogenes (Boldogh, AbuBakar et al. 1990). *FOS* and *JUN* proto-oncogenes encode the subunits of the TF activator protein-1 (AP-1) heterodimer (Rauscher, Voulalas et al. 1988, Curran and Morgan 1995), which binds a common DNA motif, the AP-1 binding site, and is involved in many important biological processes, such as cell proliferation and the regulation of cell-cycle regulator target gene expression (Karin, Liu et al. 1997). JUN family members encode proteins that can form heterodimers with FOS or homodimers, while FOS members can only bind to

JUN. FOS and JUN also have activation domains that receive cellular signals that increase or attenuate their activity. FOS regulates the secretion of cytokines during inflammatory diseases and in the determination of osteoblast and osteoclasts functions and its gene was first described by Curran et al. (Curran, Peters et al. 1982) to be involved in bone tumours induced by murine sarcoma virus. The oncogenic potential of FOS and JUN is the consequence of dysregulated expression patterns, including aberrant activation and repression, timing and cellular location (Curran and Morgan 1995).
The transcription factor encoded by the *MYC* gene regulates the expression of up to 15% of all cellular genes (Knoepfler 2007), including many which participate in cell adhesion, the cell cycle, growth control, apoptosis and differentiation. MYC also appears to regulate chromatin structure, maintaining widespread active euchromatin (Knoepfler, Zhang et al. 2006), and the *MYC* locus is also known to be altered in many human tumours (Henriksson and Lüscher 1996, Hoffman and Liebermann 2008).

1.1.4. IEGs in cancer and other diseases

IEGs level of expression is transient in physiological cellular processes, but in cancer and other diseases is reported to be aberrantly high and persistent. Abnormally expressed IEGs support cancer progression increasing cell survival, growth, invasion and metastasis (Healy, Khan et al. 2013).

FOS, one of the most studied IEGs, has long been linked to the onset of bone cancer, following the discovery that the viral homologue, v-fos, lacks FOS regulatory regions and is constitutively active in infected laboratory mice, inducing osteosarcoma (Curran, Peters et al. 1982, Curran, Miller et al. 1984). Since its discovery, FOS has been associated with cell transformation in many tumour types and other IEGs have been recognised as potential oncogenes.

IEGs expression in pathological conditions can be altered in many ways, through the misregulation of their transcription and transduction and through the inappropriate destruction of their transcript and protein products (Healy, Khan et al. 2013).

Fittall et al. (2018) report recurrent structural rearrangements of *FOS* and *FOSB* in different human bone tumours. The rearrangements are associated with a disruption of the mechanisms regulating *FOS* and *FOSB* transcript and protein degradation, indicating that the dysregulation of these oncogenes is associated with the pathogenesis of human bone cancer (Fittall, Mifsud et al. 2018). Levin et al. in 1994 compared *FOS*, *JUN*, and *EGR* expression in a cohort of non-small lung cancer (NSCLCs) samples and their surrounding normal tissue and show that their expression was significantly lower in 73% of the tumours samples. This results indicate that the downregulation of these IEGs is involved in lung carcinogenesis (Levin, Casey et al. 1994). Young et al. studied transgenic mice and found that the transactivation of a group of AP-1-dependent genes promote tumour growth and may be targeted for cancer therapy (Young, Li et al. 1999). Furthermore, dysregulation of IEGs is also associated with autoimmune and neurological diseases. Mehic et al. (2005) describe the relationship between the erroneous higher expression of c-Jun, whose protein product controls cytokine expression, in basal keratinocytes to the outset of psoriasis in humans (Mehic, Bakiri et al. 2005, Wagner 2010), while the over expression of c-fos in mouse central nervous system was correlated with irritable bowel syndrome (Zhang, Zou et al. 2011).

1.3. Molecular mechanisms of the IER

Extracellular signalling activated proteins, such as mitogen-activated protein kinases (MAPKs) (Bebien, Salinas et al. 2003), Rho GTPases and extracellular-signal regulated kinases (ERKs), as well as phosphatidylinositol 3-kinase (PI3-kinase) signalling, trigger the transduction of a signal through the phosphorylation of interacting proteins and the relocation and activation of transcription factors (TFs) already present in the cell (Bahrami and Drabløs 2016). Together with the interactions between promoters and enhancers, the transduction of the signal induces the activation of sequential waves of gene expression, known as the IER (Volinsky, McCarthy et al.).

1.3.1. The MAPK Pathway underlies the activation of the IER

The mitogen-activated protein (MAP) kinases are key evolutionary conserved enzymes responsible for orchestrating a variety of fundamental cellular processes, including differentiation, proliferation and apoptosis. The mitogen-activated protein kinase (MAPK) pathway has a central role in activating IEGs and is highly conserved across organisms of very different complexity such as mammals and yeast (Theodosiou and Ashworth 2002). It is probably the most intensely studied signal transduction pathway and is considered a model for the study of other signalling pathways. The MAPK pathway is based on a complicated network of signal transduction and gene activation composed of many kinases, including RAF, MEK and the extracellular-signal-regulated kinases ERK1 and ERK2 (Hindley and Kolch 2002, O'Neill and Kolch 2004), which are activated by consecutively phosphorylating each other (Orton, Sturm et al. 2005, Thalhauser and Komarova 2009). ERKs and other downstream activated kinases, such as the ribosomal S6 kinases (RSKs), can translocate into the nucleus and phosphorylate several important transcriptional regulators, including CREB and histone H3, resulting in the rapid transcription of IEGs (Murphy and Blenis 2006). Of the 811 genes involved in the MAPK cascade (GO:0000165) listed in the Mouse Genome (Roy, Schmeier et al.) Database (Blake, Richardson et al. 2003) 45 are well studied IEGs, such as *ATF3*, *JUN* and *TNF*. After the binding of a specific signalling molecule such as a specific growth factor to the cellular receptors tyrosine kinases (RTKs), the protein kinases in the cascade are activated sequentially, triggering the phosphorylation of transcription factors such as AP1, which is composed by the products of the known IEGs *FOS* and *JUN* (Kitabayashi, Chiu et al. 1991), and the IEG EGR1 (Svaren, Sevetson et al. 1996), downstream regulators of the specific cellular outcome (Katz, Amit et al. 2007).

One of the most interesting characteristics of this and many other signalling pathways involving the activation of IEG is their flexibility: different signals can

elicit very different outcomes. Santos et al. (Santos, Verveer et al. 2007) describe the different response of rat-derived neural progenitor cells to epidermal growth factor, EGF, and neural growth factor, NGF. The two growth factors act through the MAPK pathway engaging different sets of positive and negative feedback control mechanisms, resulting in two opposite fates: inducing cell differentiation and cell proliferation, respectively (Santos, Verveer et al. 2007, Thalhauser and Komarova 2009).

1.3.2. IEGs molecular interpretation of signal duration in cell fate determination

The duration of the signal is critical in dictating the different physiological outcomes in stimulated cells. Marshall et al. (1995) observed that the different duration of ERK activation elicited by nerve growth factor (Lerner et al. 2009) and epidermal growth factor (EGF) in rat PC12 pheochromocytoma cells led to different fates: differentiation to sympathetic-like neurons or proliferation, respectively. They detected persistent ERK activation after NGF stimulation while it was short lived after EGF stimulation (Marshall 1995).

ERK-activation duration was found to affect cell fate in other biological systems (Ebisuya, Kondoh et al. 2005). Murphy et al. studied how sustained, but not transient, activation of ERKs causes murine fibroblasts to proliferate (Murphy, Smith et al. 2002) (Murphy, MacKeigan et al. 2004). This specific cellular outcome is driven by the activation of specific IEGs, which expression level is affected from the duration of the signal but also contribute to its extension in a positive feedback loop. To explain how IEGs contribute to propagate ERK signal amplitude and duration, it has been proposed that IEGs play a role in locally concentrate the active kinase by exposing FXTP (DEF) motifs, functional docking sites for ERK (Murphy and Blenis 2006). In many IEGs, such as FOS JUN, and MYC the carbozy-terminal phosphorylation triggered by sustained ERK signalling not only stabilize the IEG-encoded protein but also exposes the FXTP (DEF) docking site for ERK (Murphy, Smith et al. 2002, Theodosiou and Ashworth 2002). Is widely agreed that the IEGs have a

fundamental role in the molecular interpretation of ERK signal duration and are therefore considered ERK sensors.

1.4. Contribution of non-coding RNA species to the IER

Regulating the expression of IEGs is very important and involves multiple levels of control of both transcriptional and post-transcriptional processes such as the phosphorylation of the transcription factors regulating transcription initiation, alternative promoter usage, alternative splicing and turnover of messenger RNAs and proteins. Furthermore, a number of microRNAs are known to target and suppress the IEGs to maintain a low activation in absence of the appropriate stimulation. Those small non-coding RNA molecules (~22 nucleotides) contain a short sequence which can bind to the target mRNA molecules influencing their translation and contributing to their degradation (Zamore and Haley 2005, Sas-Chen, Avraham et al. 2012). Non-coding RNAs are known to be activated by the same signalling inputs of IEGs in response to external stimuli and are found to share common epigenomic features with mRNAs, such as H3K4me3 and H3K27me3 in lncRNAs (Sati, Ghosh et al. 2012). The action of micro-RNAs is very fast because they do not require translation. Studying epidermal differentiation in mice, Jackson et al. reported that miR-203 abolishes long term cell proliferation and its immediate upregulation promotes the cell cycle exit, implicating an important role in the early stage of stem cell differentiation (Jackson, Zhang et al. 2013). Aitken observed other non-coding species, including MIR99AHG linc-RNA, with similar early overexpression following cell stimulation across different human datasets (MCF7 treated with EGF1 and HRG and vascular smooth muscle primary cells (SMC) treated with FGF2 and IL1b) (Aitken, Magi et al. 2015).
In many reported cases, the expression of non-coding RNA in cancer is aberrant analogously to many IEGs (Calin and Croce 2006, Avraham, Sas-Chen et al. 2010). For example, Epidermal Growth Factor (EGF) is known to activate a set of micro-RNAs, comprising miR-15b, which targets MTSS1, a suppressor of migration in breast cancer and triggers tumour progression (Kedmi, Ben-Chetrit et al. 2015). On the other hand, in MCF10 breast cancer

cell line, EGF promotes tumour progression by downregulating a set of micro-RNAs, including miR-155, which targets and suppress a set of oncogenic TFs, including FOS and EGR1. However, Aitken et al. observed an increase in expression for miR-155 in breast cancer cell line MCF7 treated with heregulin (HRG), possibly indicating a different, stimulus-specific, regulation.

1.5. Alternative TSSs and shifting promoters

Gene promoters are regulatory regions located immediately upstream the coding region and are enriched for the binding sites of transcription factors and other molecules responsible for the recruitment of the transcriptional machinery and the initiation of transcription. Transcription tends to start from local distributions of transcription start sites (TSSs) in the promoter and sometimes these transcription initiation regions are located in multiple promoters associated to the same gene (Hoskins, Landolin et al. 2011), which are called alternative TSSs or alternative promoters in literature. The different mRNAs originating from alternative TSSs of the same gene are characterized by differences in stability and translational efficiency and the resulting protein isoforms can have different life time, functions or localisation (Zhang, Dimont et al. 2017). Alternative TSSs were initially identified by primer extension (Qu, Michot et al. 1983), then later studied with rapid amplification of cDNA ends (Freter, Alberta et al. 1996) (Frohman, Dush et al. 1988) and cap-trapped 5' expressed sequence tag (EST) (Carninci, Kvam et al. 1996) sequencing and finally characterized with cap analysis of gene expression (CAGE) high throughput sequencing (Shiraki, Kondo et al. 2003, Carninci, Sandelin et al. 2006). Using the CAGE technique, the Functional ANnoTation Of the Mammalian genome (FANTOM) consortium (Consortium 2014), created a comprehensive catalogue of TSSs in mouse and human cell lines, primary cells and tissues.

FANTOM also described the different shapes of promoters classifying them into broad promoters characterized by a spread distribution of TSSs, or as sharp promoters with CAGE tags concentrated on nearby start positions (Carninci, Sandelin et al. 2006). Haberle et al. report that not only the different

promoters are characterized by preferential usage across cells, developmental stages, tissues and stimuli, but also the distribution of CAGE tags inside each promoter, which is the individual TSS usage, can vary depending on the context, a phenomena that they called ‘promoter shifting’ (Haberle, Li et al. 2014).

1.6. Alternative TSSs in the IER

The presence of multiple TSSs for the same gene ensure appropriate levels of transcription of important genes, such as the IEGs, and therefore confer robustness to the genome’s transcription program by ‘backing up’ the transcription initiation function of the promoters (Carbajo, Magi et al. 2015). Furthermore, the alternative TSS usage is an important regulatory mechanism in the IER because it increases the specificity of cell response to external stimulation. Different TSSs associated to the same genes can be characterized by tissue and cell-specific expression but also developmental stage-specific expression (Carninci, Sandelin et al. 2006). For example, the FANTOM5 CAGE data show that the gene *SERPINA1*, which encodes for the anti-protease AAT, involved in immunoregulatory processes, has at least four promoters; one is specifically utilized by liver cells, while the other three promoters are used by macrophages and other myeloid cells (Baillie, Arner et al. 2017).

1.7. Conclusions from FANTOM5 time series data

The FANTOM5 project (Lizio, Harshbarger et al. 2015) recently analysed the TSS expression of 33 human and mouse densely sampled CAGE time course datasets. The FANTOM5 CAGE data offer a number of advantages for expression profiling because they are based upon single-molecule sequencing to avoid PCR, digestion and cloning biases. They provide up to single base-pair resolution of TSS and promoter regions, and provide a sensitive, quantitative readout of transcriptional output accounting for the alternative promoters of each gene. The output of individual promoters is not confounded by splicing variation, and many novel lowly expressed transcripts including

non-coding RNAs (ncRNAs) can be readily detected (see http://fantom.gsc.riken.jp/5/). CAGE data are thus ideally suited to studying the strong burst of transcription at promoters seen in IERs. FANTOM5 data include eight CAGE time course datasets employing unusually dense sampling at time points within 300 min of stimulation, for a variety of stimuli treating a variety of human cell types. These heterogeneous datasets, produced using a common experimental platform, should be fertile ground for novel insights into the IER, but a comprehensive meta-analysis has not been performed until now.

Six of these datasets have been the subject of recent publications, while two datasets, the adipogenesis time-series and the calcification in an osteosarcoma cell line time-series, have not. All data is now available on the FANTOM portal and the key points, with regards to their findings on IEGs, are summarized in Table 1.

The general conclusions about these publications are:

1. These studies are not directly related and investigate very different aspects of biology. However, they all study processes where the IER was known or suspected to play an important role.
2. The analysed datasets provide a variety of details about the several genes activated in different biological context which is a great resource to develop algorithms that are relevant to studying the IER.
3. The findings of the thesis are consistent with the observations described in the FANTOM5 publications and emphasise that there are strong commonalities in the IER, even across the very different biological processes profiled.

Table 1 FANTOM5 publications relevant to the study of the IER. This table lists the articles published by FANTOM5 related to the datasets used in this thesis. From left to right: references, a brief description of the cell types and treatments, the dataset names used in this thesis and a description of the key findings of the publication.

Reference	Cell type and treatment	Dataset	Key findings
(Alhendi, Patrikakis et al. 2018)	Vascular smooth muscle cells treated with FGF2 and IL1B	PAC_FGF2 PAC_IL1B	• Three times bigger proportion of promoters differentially expressed in PAC_FGF2 in respect to PAC_IL1B. • A number of promoters, including those of *JUN*, *FOS*, and *EGR1*, differentially expressed in both experiments. • The activation of many TF binding motifs, including IEGs, in target genes is associated to a rapid and transient increment in expression of the TF in the first hour after FGF2-stimulation and a later biphasic response in PAC_IL1B.
(Mina, Magi et al. 2015) (Carbajo, Magi et al. 2015)	MCF7 human breast cancer cells treated with HRG and EGF1	MCF7_HRG MCF7_EGF1	• EGF1 and HRG elicit alternative outcomes relevant to cancer progression: cell proliferation and differentiation, respectively. • Stronger and persistent activation of ErbB receptors in MCF7_HRG with respect to MCF7_EGF1. • Common MAPK regulated response in the early stage associated to the activation of SRF regulated IEGs. • Divergent downstream regulatory events, driven by the activation of stimulus specific IEGs. • Groups of TFs, including many IEGs, are characterized by similar early TSS activation in both datasets. • Lately expressed TFs are characterized by stimulus specific TSS expression profiles.
(Aitken, Magi et al. 2015)	MCF7 human breast cancer cells treated with HRG and EGF1, and Vascular smooth muscle cells treated with FGF2 and IL1B	MCF7_HRG MCF7_EGF1 PAC_FGF2 PAC_IL1B	• Set of 11 IEGs peaking in the first few hours after stimulation, and two peaking lncRNAs, *NEAT1* and *MALAT1*, across the four datasets. • Upregulation of *JUN* in PAC_IL1B, MCF7_EGF1 and MCF7_HRG and downregulation in PAC_FGF2.
(Baillie, Arner et al. 2017)	Response of monocyte derived macrophages to LPS	PMDM_LPS	• Ligation of LPS with its target receptor, TLR4, triggers the transient and rapid expression of ubiquitous IEGs.
(Dieterich, Klein et al. 2015)	Response of human lymphatic endothelial cells (LEC) to VEGF-C	PEC_VEGF	• Of the 241 genes upregulated after stimulation, a large proportion coded for TFs. • These genes include common IEGs such as *FOS*, *JUN* and *EGRs,* and three LECs/VEGF-C specific TFs: MAFB, KLF4 and SOX18.

1.8. Computational analysis of gene expression time series data

Time series experiments are often the best approach to study dynamic processes in complex systems. In recent years, gene expression has been studied genome wide and such data are well suited to understand the function of specific genes and the relationships between them. The earliest time series analysis of gene expression made use of micro-array data (Schena, Shalon et al. 1995, Winkles 1997). Later, high throughput RNA-sequencing technology improved the measurement of gene expression making time series experiments more feasible and relevant (Trapnell, Williams et al. 2010, Pauli, Valen et al. 2012). Nowadays there are huge numbers of deposited time series gene expression data covering many biological systems and organisms, such as the FANTOM5 collection (Kawaji, Kasukawa et al. 2017) and the ENCODE project (Consortium 2012).

1.8.1. Differential expression analysis and clustering techniques.

The computational methods commonly used to analyse time series expression data are differential expression analysis (Nau, Richmond et al. 2002, Bendjilali, MacLeon et al. 2017, Spies, Renz et al. 2017) and clustering (Walker, Volkmuth et al. 1999, McDowell, Manandhar et al. 2018). Differential expression analysis tries to identify the genes which expression significantly changes across time points, while clustering methods try to group together genes with similar expression profiles. Results from both methods have led to important discoveries and increased our understanding of biological dynamic processes. For example Blishack et al. (2015), studying the specific response of macrophages to *Mycobacterium Tuberculosis* infection, identified a group of genes differentially expressed over time and specifically affected by this pathogen (Blischak, Tailleux et al. 2015), while Mina et al. (2015) applied the CIDER pipeline, which integrates hierarchical clustering of time series with motif enrichment analysis, to infer the transcriptional cascades initiated by

ErbB receptors in MCF7 cell line (Mina, Magi et al. 2015). It's noteworthy to mention that differential expression analysis contributed to the discovery of many IEGs (Selvaraj and Prywes 2004, Tullai, Schaffer et al. 2007). However, both differential expression and clustering analysis present challenges and are not appropriate to compare gene expression time-series data from different experiments (Bar-Joseph, Gitter et al. 2012). For example, differential expression methods rely on large amounts of data to detect significant changes in successive time points and lowly expressed genes are discarded as they cannot pass this threshold. Clustering methods do not perform well when encountering genes with unusual expression profiles and tend to group together genes with low correlation and different local behaviours with other members of the cluster, missing global similarities with genes in other clusters. Furthermore, both methods are not appropriate when comparing datasets of different length and different time points and both methods implicates a data mining process where the expression profile of interest is not known.

1.9. Transcriptional dynamics of the IEGs

Using a novel analysis approach for time course datasets, Aitken et al. (2015) identified protein coding and non-coding transcripts with expression profiles that approximate the dynamics of known IEGs. Generally, the method uses a Bayesian model selection approach with nested sampling algorithm to fit the expression patterns to mathematical models of interest, then it select the model that fit better the data (Figure 1).

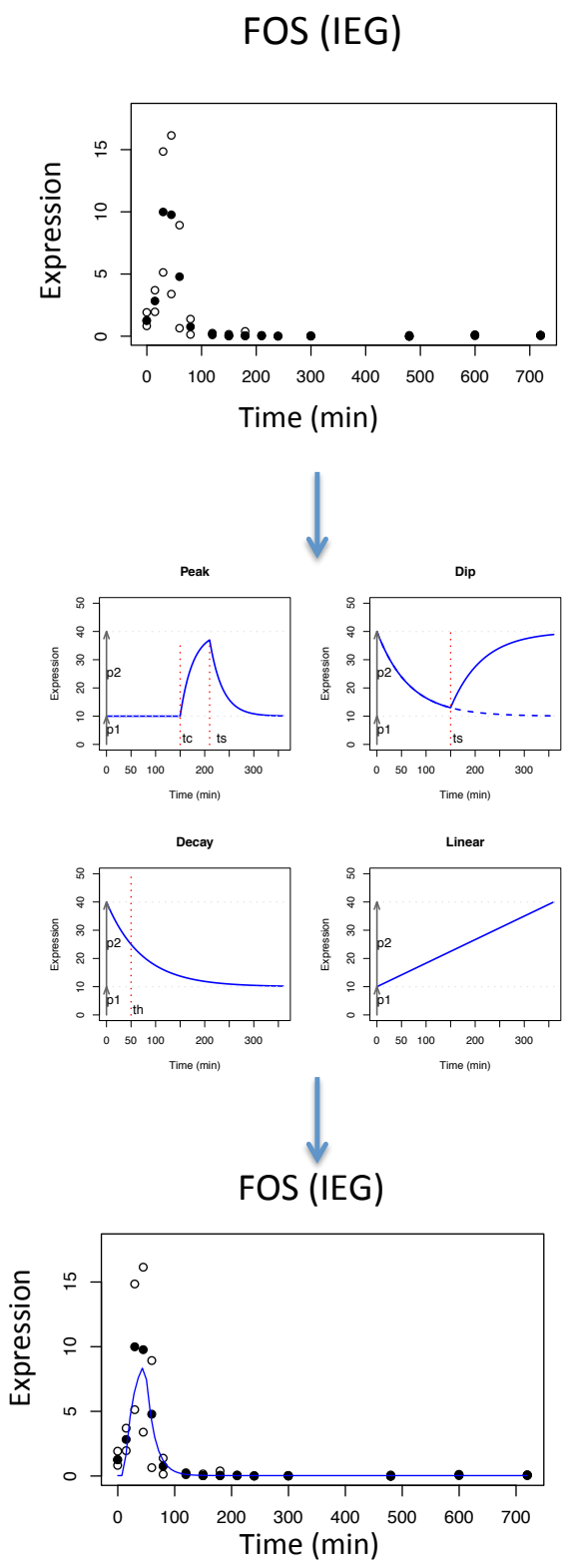


Figure 1 Time-course model-fitting analysis. After defining 4 models of interest (peak, linear, decay, dip), the algorithm calculates the evidence, log Z, for each model and finally selects the model which better explains the expression profile. In this example, the immediate early gene FOS has been assigned to the peak model.

In the paper, Aitken et al. analysed the expression profile of four CAGE time course datasets from FANTOM5 project, the vascular smooth muscle quiescent cells (SMC) treated with FGF2 or IL1b datasets, and the MCF7 human breast cancer cell line samples treated with the epidermal growth factor EGF1 or heregulin (HRG), and classified the genes to four models, a peak model, closing resembling the expression pattern of IEGs, a dip model, a decay model and a linear model.

In the paper, the peak model is characterized by a basal expression level p_1 and an increase up to 90% of the change in expression p_2 at time t_s. The rate δ is defined in terms of t_s.

$$\delta = \frac{\log(0.1)}{t_s}$$

$$y = p_1 + p_2 * \left(1 - e^{\delta t}\right); t \leq t_s$$

$$y = p_1 + 0.9 * p_2 - 0.9 * p_2 * \left(1 - e^{\delta(t - t_s)}\right); t > t_s$$

The dip model is parameterized by p_1, which represents the minimal expression and p_2, which represents the change in expression at t_s. The expressions starts at $p_1 + p_2$ and drops by 90% of p_2:

$$\delta = \frac{\log(0.1)}{t_s}$$

$$y = p_1 + p_2 * e^{\delta t}; t \leq t_s$$

$$y = p_1 + 0.1 * p_2 - 0.9 * p_2 * \left(1 - e^{\delta(t - t_s)}\right); t > t_s$$

In the linear model p_1 is the expression at time 0, while p_2 is the expression at the last time point t_{max}:

$$y = p_1 + \frac{(p_2 - 10) * t}{t_{max}}$$

In the decay model p_1 represents the basal expression, p_2 the maximal change in expression, and t_h the time in which the expression is $\frac{p_2}{2}$(Aitken, Magi et al. 2015):

$$\delta = \frac{\log\ 2}{t_h}$$

$$y = \ p_1 + p_2 * e^{-\delta t}$$

Fitting the expression profiles of the entire set of CAGE TSSs detected in the four datasets, Aitken et al. found that most CAGE TSSs associated to known IEGs were classified to the peak model and identified a set of protein coding-genes, microRNAs and other non-coding RNA species activated similarly to known IEGs.

Although chromatin data are not available for the datasets analysed, Aitken et al. used DNaseI hypersensitivity data for MCF7 cells from ENCODE (Consortium 2011) to investigate the chromatin context were the peak CAGE TSSs were located. They found that protein-coding CAGE TSSs assigned to the peak model are located in accessible chromatin and an association between DNAseI reads between 100 and 1000 and CAGE expression at time 0 higher than 10 TPM suggesting that the rapid activation of IEGs requires minimal levels of chromatin accessibility. In this thesis I am extending the work done by Aitken et al. by adding four FANTOM5 time course datasets, introducing a delay in the peak model to improve the fitting, and focusing on the temporal order of activation of candidate IEGs.

1.10. Meta-analysis: opportunities and limitations

Transcriptomic analysis studies are subject to many challenges, such as the sequencing error rate, reads mappability and data standardization and interpretation. All these factors affect the quality of downstream analysis and introduce variable levels of noise in the results. When the expression of thousands of genes is compared, the error rate increases and the p-values obtained have to be corrected with different techniques, such as the Bonferroni correction (McDonald 2009), decreasing the list of significant genes and therefore the power of the analysis. In a study from 2001, it was estimated that even applying strict multiple–test correction, in 26 over 36 genetic association studies (72%), the significant genomic association were no longer significant after testing them in additional datasets (Ioannidis, Ntzani et al. 2001). Applying meta-analysis, which is the integration of as many pertinent datasets as possible from multiple sources, partially minimizes bias, because it increases the amount of information content. Meta-analysis is a useful approach to estimating and explaining the heterogeneity across datasets. However, comparing different datasets with biological and technical variance is not straightforward. Many different approaches have been proposed to address this issue and guarantee robust and reproducible results, such as combining p-values, z-scores, ranks and or effect sizes (Sweeney, Haynes et al. 2017).

1.11. Aims of the thesis

Although the IER, and a small number of IEGs are well studied, the majority of studies have been performed on single genes or pathways and often considering a specific cell type or stimulus. This means that the results from different studies are difficult to compare computationally because of experimental and technical variations. A number of studies have attempted to compare different time series data using differential expression analysis of time points or clustering of gene expression profiles. However, both techniques present problems, including poor sensitivity for low expressed genes and

difficulties in comparing datasets with different sampling times. A comprehensive analysis of the IER and a catalogue of IEGs is still missing. Furthermore, the relationships between IEGs are still unclear.

In this thesis I perform an extensive meta-analyses of promoter activity in the context of the IER, encompassing unusually diverse cell types and stimuli.

The aims of this work are:

- To improve a recently published method for time series analysis by attempting to find better estimates for model parameters and allowing a delay in the peak model accounting for a delay in the induction of transcription.
- To rigorously classify IEGs and estimate the core IEG repertoire active across cellular responses.
- To identify possible novel protein coding and non-coding genes participating to the IER.
- To analyse the temporal activity patterns of promoters and the relationships between IEGs.
- To discover the extent to which the IER regulatory mechanism is shared among cell types and common to diverse stimuli.
- To document alternative promoter usage patterns for the genes active in the IER and determine whether this contributes to the regulation of the IER across different cell types and stimuli.
- To verify the existence of promoter shifting events in the datasets analysed as a possible additional regulatory mechanism of the IER.

2. Methods

2.1. Data Resources

The data used in this project was produced by the FANTOM collaboration, release 5, phase 2. The eight datasets considered here can be broadly subdivided in 3 groups: Three cell activation datasets, three cell proliferation datasets and finally two cell differentiation datasets. Figure 2 shows the sample collection scheme for the eight datasets. Notably, all the datasets are densely sampled in the first three hours after stimulation. Specifically, they all share the first six time points (0, 15, 30, 45, and 60 minutes). However, the datasets are characterized by overall difference in length and sampling points, with Mesenchymal SC differentiation as the shortest and Macrophage response to LPS as the longest series. In all datasets, time 0 corresponds to inactivated or quiescent cells.

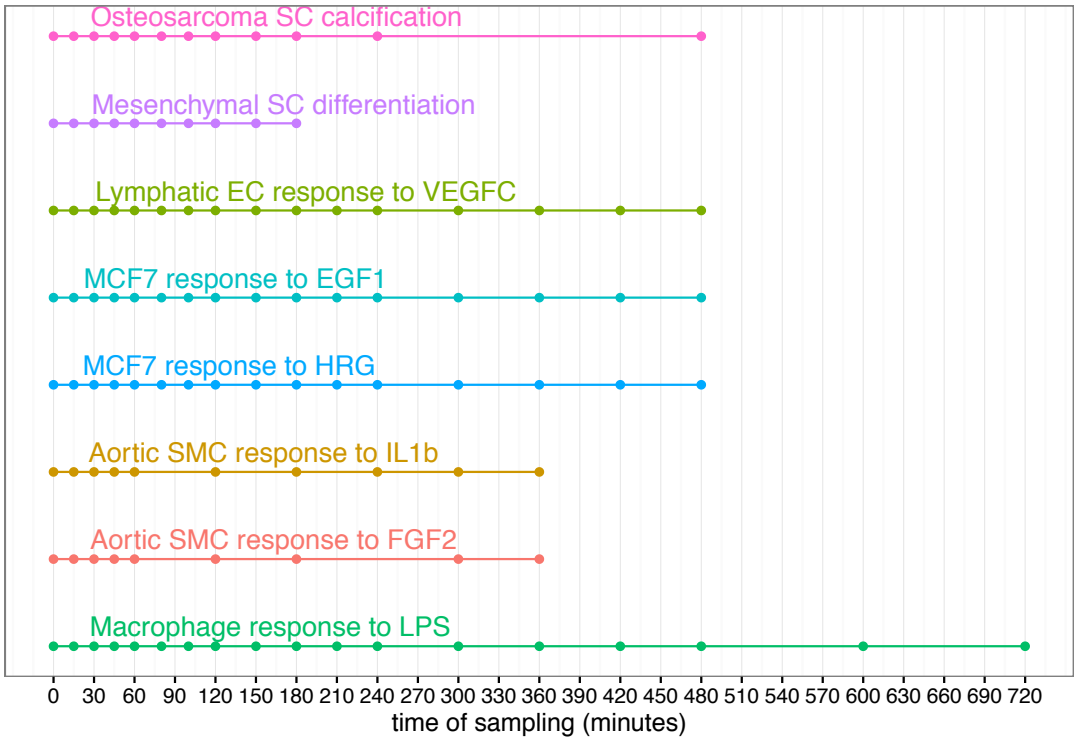


Figure 2 Time course datasets encompassing the immediate early response. A schematic view of the eight time-course datasets, with horizontal lines indicating the span and points representing the times of sampling. For all datasets the time zero corresponds to inactivated or quiescent cells.

2.2. Cell activation series

2.2.1. Macrophage response to LPS - PMDM_LPS

Human macrophages were obtained from CD14-positive monocytes extracted by 3 anonymised donors with approval of the human Ethics Committee of the University of Edinburgh (8/9/09). The CD14-positive monocytes extracted from 320mls of blood from each donor, were cultured in recombinant human CSF1 at 100ng/ml for 7 days. Afterwards, the monocyte-derived macrophages were treated with 100ng/ml of salmonella R595 LPS. The entire dataset is composed of 3 biological replicates sampled at 23 time points along a span of 48 hours but was sparsely sampled after 12 hours. For the meta-analysis I decided to shorten the longer time course, the PMDM_LPS, from the original 48 hours to 12 hours. Samples used in my meta-analysis were taken at 0 time (non-stimulated), 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300, 360, 420, 480, 600 and 720 minutes. The dataset was provided to the FANTOM consortium for sequencing by David Hume (Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, United Kingdom).

2.2.2. Osteosarcoma SC calcification - SAOS_OST

The human osteosarcoma cell line used, Saos-2, was first created extracting osteosarcoma epithelial cells from a 11 years old female patient in 1973. Mineralization was induced by 50 μg/ml ascorbic acid and 2.5 mM Bisphosphoglycerate (BPG) in medium with 10% serum. The dataset consists of 3 biological replicates sampled at 18 time points spanning 28 days, from which we selected samples taken at 0 (untreated), 15, 30, 45, 60, 80, 100, 120, 150, 180, 240, and 480 minutes. The Saos-2 cell line was provided to FANTOM consortium for sequencing by Kim Summers (The Roslin Institute and R(D)SVS, University of Edinburgh, United Kingdom).

2.2.3. AoSMC response to IL1b - PAC_IL1B

Growth arrested human aortic smooth muscle cells were obtained from 3 donors by Cell Applications (CA, USA). The cells were incubated in serum free medium for 24 hours, and therefore rendered cell growth-quiescent, and then treated with IL-1beta (10ng/ml) and sampled at 0 (quiescent and unstimulated cells), 15, 30, 45, 60, 120, 180, 240, 300 or 360 min. The samples were provided to the FANTOM consortium for sequencing by Levon Khachigian (UNSW Centre for Vascular Research, University of New South Wales, Sydney, Australia).

2.3. Cell proliferation series

2.3.1. AoSMC response to FGF2 - PAC_FGF2

Human AoSMC has been collected, treated and finally sampled and submitted to FANTOM for CAGE-sequencing as described above, treating with FGF-2 (50ng/ml) instead of IL1b.

2.3.2. Lymphatic EC response to VEGFC - PEC_VEGF

The primary lymphatic endothelial cells, previously isolated from human foreskin from 3 individual donors, were starved overnight before treating with 1.5 µg/ml recombinant human VEGF-C156S for 0 (unstimulated), 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300, 360, 420, 480 minutes. Michael Detmar (Swiss Federal Institute of Technology, ETH Zurich) provided the samples to FANTOM consortium for sequencing.

2.3.3. MCF7 response to EGF1 - MCF7_EGF1

MCF7 breast cancer cells, originally isolated from the pleural effusion of 69-years old woman with breast cancer, were collected from American Type Culture Collection (ATCC) and maintained DMEM supplemented with 10% fetal bovine serum and analysed in 3 biological replicates. EGF1 was added only after 16-24 hours of serum starving. 16 time-points were taken up to 8 hours from stimulation at 0 time (non-treated), 15, 30, 45, 60, 80, 100, 120,

150, 180, 210, 240, 300, 360, 420, 480 minutes and provided for sequencing to the FANTOM consortium by Mariko Okada-Hatakeyama (Laboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences, IMS, Yokohama, Japan).

2.4. Cell differentiation series

2.4.1. MCF7 response to HRG - MCF7_HRG

MCF7 cell line was collected, treated and sampled as described above, treating with HRG hormone instead of EGF1.

2.4.2. Mesenchymal SC differentiation - PMSC_MIX

Human adipose-derived mesenchymal stem cells were provided to FANTOM consortium for sequencing by Peter Arner (Karolinska Institutet, Stockholm, Sweden) and were derived from the stromal-vascular fraction of human subcutaneous adipose tissue of 4 donors. After propagation in vitro, a stem cell/progenitor population was selected and expanded in multiple passages. The differentiation has been stimulated with a differentiation cocktail containing IBMX, dexamethasone and Rosiglitazone, and sampled in 3 replicates with homogeneous genetic background at 0 (unstimulated), 15, 30, 45, 60, 80, 100, 120, 150 and 180 minutes.

2.5. CAGE technology: opportunities and challenges.

The datasets analysed are all part of FANTOM consortium (Consortium 2014), 5th release, Phase 2. Whole transcriptome abundance analysis and Transcription Start Site identification at single base definition were carried out using Cap Analysis of Gene Expression (CAGE), which is a specialized sequencing technology developed in RIKEN (Shiraki, Kondo et al. 2003, Kodzius, Kojima et al. 2006). CAGE technology (Figure 3) involves the sequencing of DNA tags obtained from the initial nucleotides of the 5’ end of cDNAs and allows to measure transcript abundances and to identify the starting point of transcription at single base resolution. HeliScope single molecule sequencer is used instead of employing polymerase chain reaction

(PCR), with the advantage of avoiding bias introduced by the amplification of the DNA strand. Longer reads (average 33 bases) from HeliScope CAGE minimize the proportion of tags that maps to multiple loci compared to the protocols used in previous versions of FANTOM (Kanamori-Katayama, Itoh et al. 2011). However, highly similar sequences such as transposons and pseudogenes still generate mapping ambiguity and their annotation is often challenging. Furthermore, HeliScope CAGE data have an elevated sequencing error rate (~5%), variable length and lack an assessment of base quality (Consortium 2014).

CAGE ribosomal reads were eliminated using rRNAdust (author: T. Lassmann, software available at fantom.gsc.riken.jp/5/suppl/rRNAdust/), a parallel dynamic non-heuristic programming algorithm that directly aligns the reads to the whole ribosomal DNA, and the remaining reads were mapped to the genome (hg19) using Delve, a probabilistic mapper which makes use of hidden Markov model to iteratively map reads to the genome and assess the probability of a wrong mapping. Individual reads are finally assigned to the genomic location with the highest probability to be true. It is noteworthy that neither rRNAdust nor Delve algorithms have been published so far, although they are both extensively used in all the recent FANTOM CAGE data processing.

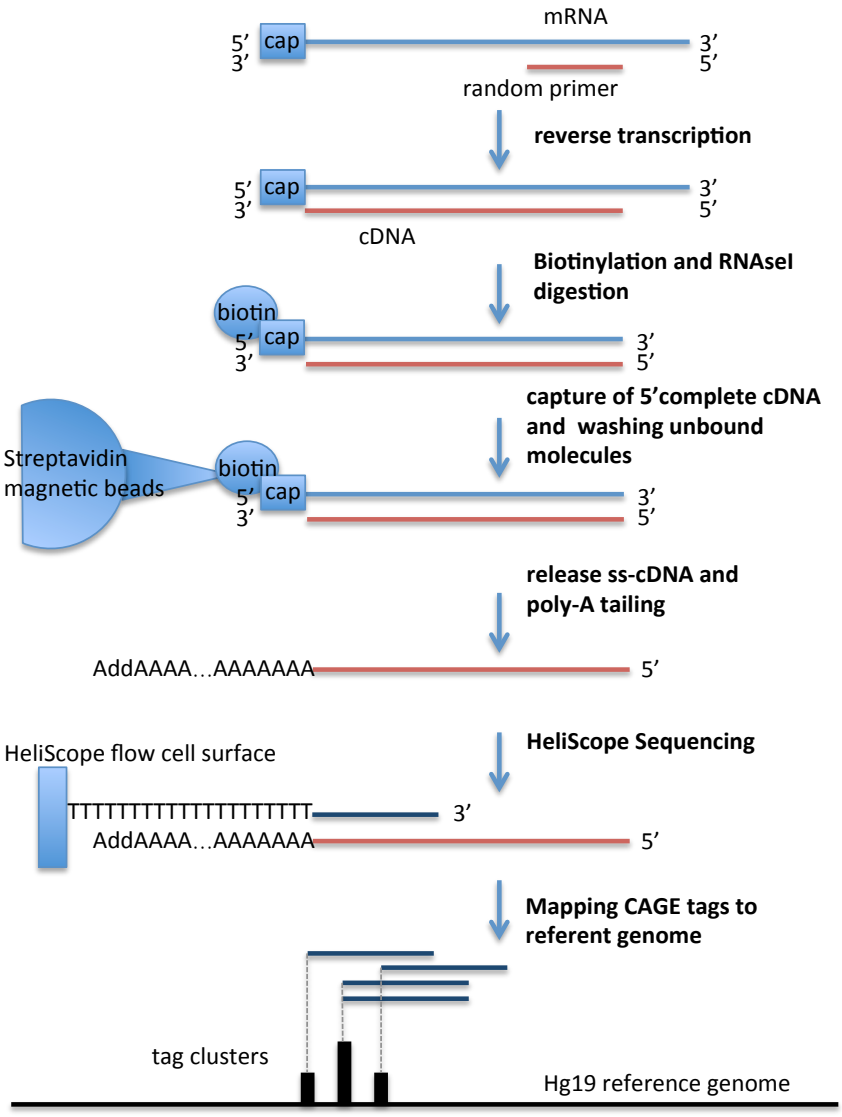


Figure 3 CAGE technology and reads data processing. Schematic view of the protocol for CAGE sequencing. After mRNA reverse transcription, 5' caps are biotinylated to allow selection and the DNA filament is then cut.

2.6. CAGE TSSs clustering and quantification

To assemble nearby TSS into co-regulated units, CAGE peaks were clustered in hierarchical structures of genomic intervals containing more CAGE reads than the surrounding regions (Figure 4).

TSSs were constructed from aligned CAGE tags and the number of tags supporting each TSS counted. Data was filtered and normalized and CAGE tags which map to the same distinct TSS sequences on the same strand of a

chromosome and overlap by at least 1 base pair (bp) were clustered into tag clusters (TC), which were finally aggregated across multiple datasets to build a set of consensus promoters, common to all FANTOM datasets. Clusters longer than 49 bp were decomposed into non-overlapping sub signals using independent component analysis (ICA) (Hyvärinen, Karhunen et al. 2001), and CAGE TSSs were finally validated using supporting evidence from co-localization with expressed sequence tags (ESTs), H3K4Me3 sites and DNase Hypersensitivity regions (Consortium 2014).

Tables with CAGE counts, expressed as tag per million, are available for FANTOM collaborators to download from https://fantom5-collaboration.gsc.riken.jp/wiki/index.php/Timecourses. The method described, defined decomposition based peak identification (DPI), is available to download from github at https://github.com/hkawaji/dpi1.

Furthermore, the FANTOM consortium recently released CAGEr R package (Haberle, Forrest et al. 2015), the first comprehensive CAGE toolbox which contains all the functions necessary to obtain consensus clusters of TSS starting from aligned CAGE tags or files with genomic locations of TSS and number of supporting CAGE tags. It also provides many functions for visualization of statistic over the data and analysis of TSS dynamics and different usage across datasets.

The *CAGEr* package is freely available from Bioconductor at http://www.bioconductor.org/packages/release/bioc/html/CAGEr.html

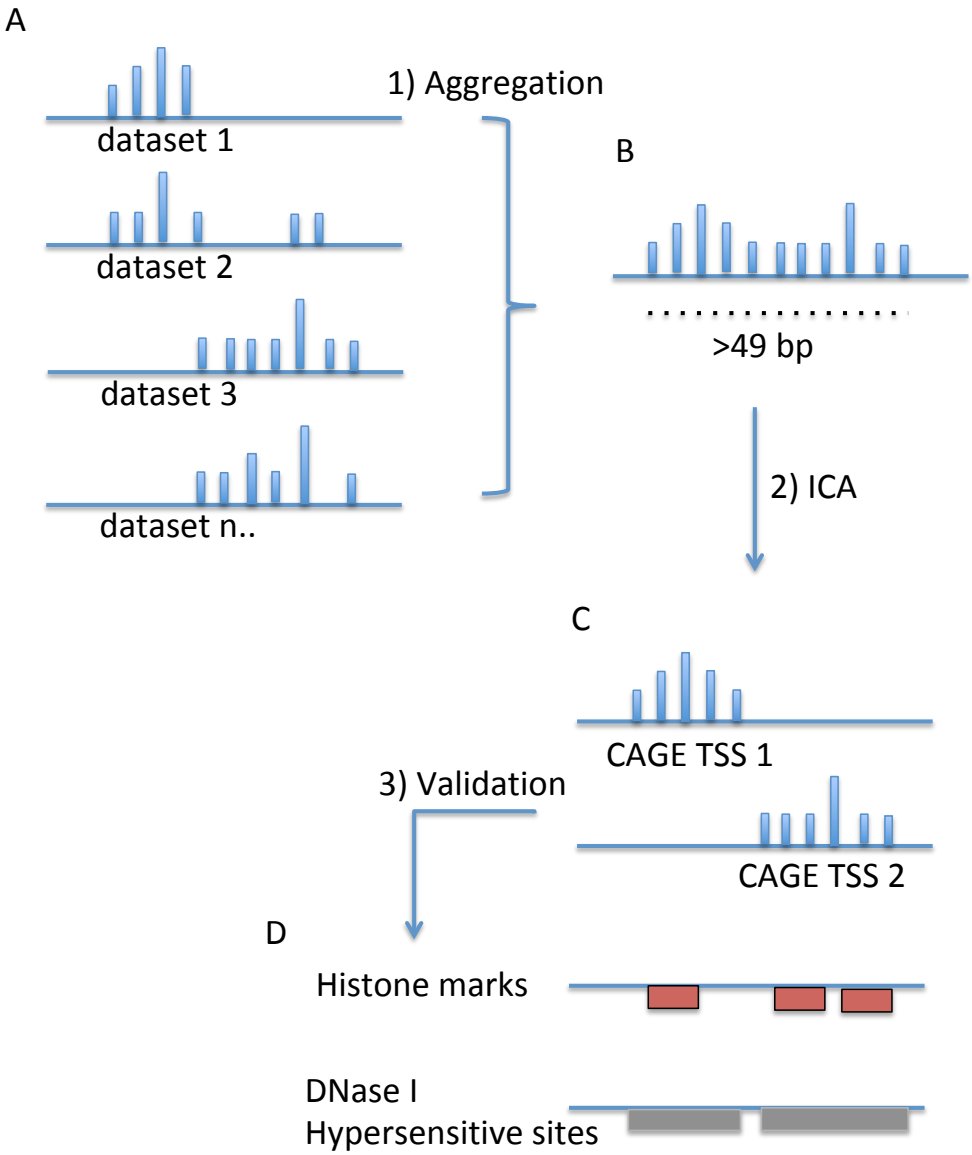


Figure 4 CAGE TSS clustering. (A) Overlapping CAGE tags have been clustered into tag clusters (TC), which have been finally aggregated across multiple datasets to build a set of consensus promoters (1), common to all FANTOM datasets (B). Clusters longer than 49 bp were decomposed (2) into non-overlapping sub signals using independent component analysis (ICA) (Hyvärinen, Karhunen et al. 2001) (C), and CAGE TSSs were finally validated (3) using supporting evidence from co-localization with expressed sequence tags (ESTs), histone marks and DNase Hypersensitivity regions (Consortium 2014) .

2.7. Annotation of CAGE clusters to Gencode V10

As described in (Lizio, Harshbarger et al. 2015) CAGE TSSs have been annotated by the FANTOM consortium, using a hierarchical approach, with respect to Gencode (Harrow, Frankish et al. 2012) V10. The FANTOM5

BioMart (Kasprzyk 2011) web interface (http://biomart.gsc.riken.jp/) can be queried for FANTOM5 CAGE peaks and samples, and contains Ensembl (Hubbard et al. 2002) and other sources id annotations such as NCBI (Maglott, Ostell et al. 2010) and UniProtKB (Boutet, Lieberherr et al. 2007) IDs. However, because of not fully transparent documentation and inconsistency across multiple id references, I decided to re-annotate the FANTOM5 CAGE TSSs using only Gencode V10 Ensembl nomenclature with a simplified annotation procedure (Figure 5). Considering the 5' end of each Ensembl transcript, I computed the distance with each CAGE cluster extremities. For the plus strand I assigned a positive sign where the 5' end of the CAGE cluster lies downstream of the transcript 5' end. Here, the distance value is calculated between the 5' end of CAGE cluster and Ensembl transcript and a negative sign when both the CAGE cluster's 5' and 3' end lies upstream to the transcript 5' end and its values is computed between the CAGE cluster 3' and the transcript 5' ends. The distance value is equal to 0 where the CAGE cluster lies upstream the transcript 5' but the CAGE cluster 3' end is contained in the transcript coordinates or lies downstream to the transcript 3' end.

For the minus strand a positive sign is assigned when the 5' of the CAGE cluster lies upstream the transcript 5' and the distance is computed between the 5' extremities, a negative sign is assigned when the CAGE cluster lies entirely downstream to the transcript and the distance is computed between the 3' of the CAGE cluster and the 5' of the transcript and the value is 0 when CAGE cluster coordinates overlap (completely or partially) the transcript coordinates and the CAGE cluster 5' lies downstream to the transcript 5'.

Finally, for each CAGE cluster I assigned the Ensembl transcript (or group of transcripts when the distance was the same) with absolute distance value <= 500 bp.

Distance was computed using closestBed utility from the bedtools toolset (Quinlan and Hall 2010).

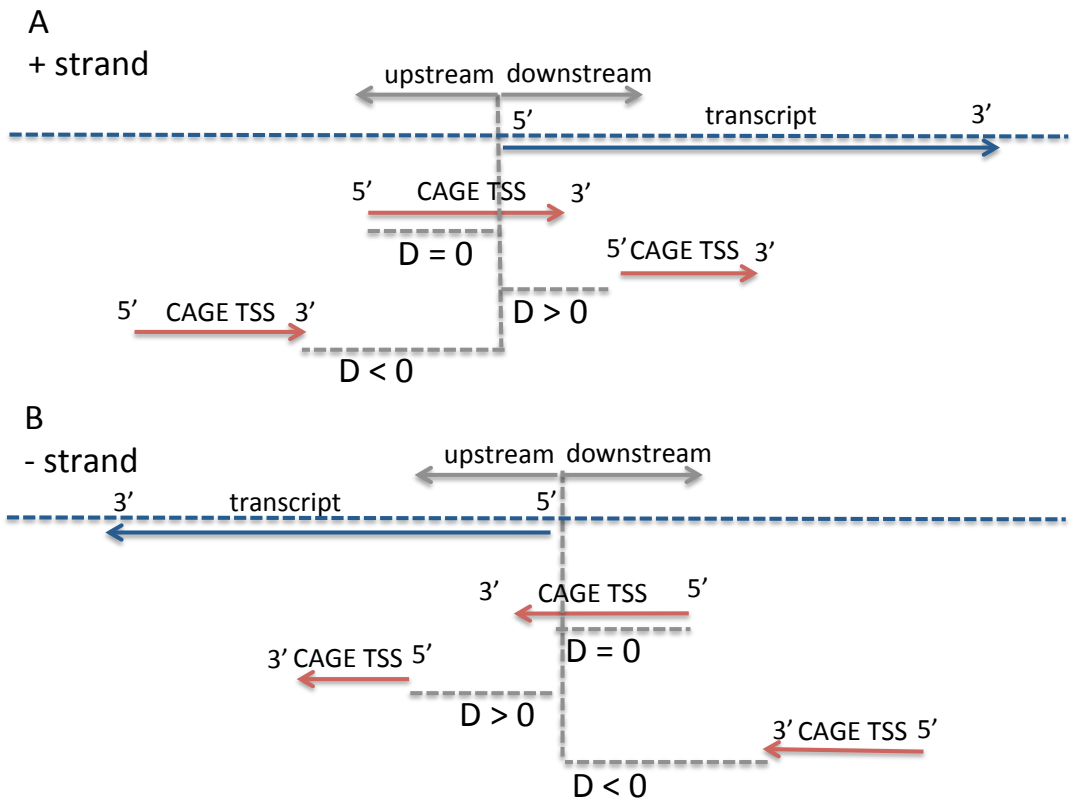


Figure 5 CAGE TSS annotation. Graphic representation of CAGE TSS annotation for plus (A) and minus (B) strands. For the plus strand (A) the distance value, D, is equal to 0 where the CAGE TSS lies upstream of the transcript 5' and the 3' is contained in the transcript coordinates or lies downstream to the transcript 3'. D>0 where the 5' of the CAGE TSS lies downstream of the transcript 5' and D<0 when both the CAGE cluster 5' and 3' lie upstream of the transcript 5'. For the minus strand (B) D=0 when the CAGE TSS overlap the transcript and the 5' lies downstream the transcript 5'. D>0 where the 5' of the CAGE TSS lies upstream the transcript 5' and D<0 when both the CAGE cluster 5' and 3' lies downstream to the transcript 5'.

2.8. Mitochondrial and multigene families' genes are ambiguously mapped by FANTOM CAGE short reads

In humans, mitochondrial DNA exists as a circular molecule about 16kb long, which is transcribed by Mitochondrial single-subunit RNA polymerase (mtRNAP), a specialized RNA polymerase. The transcription of mitochondrial genes only starts from three promoters: one in position 407 on the light strand (light strand promoter, LSP), and two in position 561 and 646 on the heavy strand (heavy strand promoter 1 and 2, HSP1 and HSP2). The resulting polycistronic transcripts are then processed and the mRNAs are extracted.

However, all mitochondrial transcription units are considered uncapped (Grohmann, Amalric et al. 1978, Temperley, Wydro et al. 2010).

In the datasets analysed in this project, I observed 93 CAGE TSSs assigned to 13 protein coding genes and 105 CAGE TSSs assigned to 21 non-coding genes, mapping to the mitochondrial genome (over 37 total genes contained in the mitochondrial DNA). Among these, only two CAGE TSSs were located close to the conventional HSP1 and HSP2 positions, while the others were in the proximity of mitochondrial transcription units. Hoskins et al. (Hoskins, Landolin et al. 2011) observed similar peculiar mapping of CAGE tags in Drosophila melanogaster and in the unpublished data on human cell lines K562 and GM12878 (https://www.genome.gov/26524238/encode-project-common-cell-types/) produced in the ENCODE project (Birney, Stamatoyannopoulos et al. 2007). These results are incompatible with the assumption that CAGE reads involve the TSSs of transcribed genes.

It is likely that the mapping of CAGE reads to mitochondrial genes could be confounded by the presence of NUclear MiTochondrial (NUMT), often represented hundreds of times in the nuclear genome and therefore possibly erroneous targets of short read assignments.

Another class of genes difficult to map reliably with CAGE reads is the set of genes belonging to multigene families, including the small nuclear ribonucleoproteins (snRNPs) involved in gene regulation and alternative splicing, such as U2, U6 and U3 (Will and Lührmann 2011). snRNPs genes are found in the genome in tandem arrays of multiple copies of the same gene flanked by spacer DNA. Furthermore, the majority of the members of this class are pseudogenes, for example U1 snRNA in human cells is represented by 30 functional genes in chromosome 1 and non-functional pseudogenes composed by complete but aberrant U1 gene characterized by extensive flanking homology to the true U1 genes (Lindgren, Bernstein et al. 1985). The mapping of pseudogenes and transposons with short reads technology, such as CAGE, which cannot distinguish identical or highly similar, is error prone.

2.9. Data filtering

From 22,846 tested CAGE TSSs associated with 12,145 Ensemble protein-coding genes and 2,730 tested CAGE TSSs associated with 1,387 Ensembl non-coding genes, we filtered out all the CAGE TSSs with possible ambiguous mapping, that is genes belonging to multigene families, such as the snRNA U1 and 7SK, another component of the snRNP complex, and all the genes annotated to the mitochondrial chromosome. This left 22,753 protein coding CAGE TSSs (12,132 Ensembl genes) and 2,476 non-coding CAGE TSSs (1,226 Ensembl genes).

2.10. Time-course gene expression profile classification

Time-series gene expression data are well suited to study the IER because it involves dynamic processes which cannot be properly characterized with static experiments. The existing methods for time-series gene expression analysis fall in two main groups: differential expression methods and expression profile clustering. Both methods perform well in single experiments but are not well suited to perform comparison across different experiments and neither method is designed to systematically identify expression profiles following a specific trend of interest. To classify time-series data for each CAGE defined TSS, I refine a previously successful Bayesian model selection algorithm (Aitken, Magi et al. 2015) to classify promoter responses to pre-defined mathematical models. I focused on the peak mathematical model which is designed to approximate the rapid and transient expression profile of IEGs, to classify known IEGs and identify possible novel IEGs. This method is designed to handle lowly expressed genes, such as the non-coding RNAs, and to classify gene expression profiles independently from the length and the sampling time of the time-series experiment, and is therefore well suited to compare different experiments, such as the eight FANTOM5 densely sampled time-course datasets previously described which are studied in this project.

2.11. Meta-analysis approach

The IER has been extensively studied in the past years and a number of studies compared the expression of IEGs across multiple and usually related datasets (Mina et al. 2015; Alhendi et al. 2018). Recently Aitken et al. (2015) developed a new method to identify IEGs which is well suited for meta-analysis as it does not require the same number of time points across datasets nor a minimum differential expression level (Aitken et al. 2015). Therefore, in this thesis, Aitken's method was chosen to identify coding and lowly expressed non-coding candidate IEGs in multiple FANTOM time-series of different length. Combining the results from multiple datasets can be achieved using different approaches. In this study I utilized a simplified version of the 'vote-counting' approach described in (Bushman and Wang 1994). I ranked the genes depending on the number of datasets sharing the peak kinetic. I defined a 'robust set' of genes peaking in at least seven out of eight datasets and a 'permissive set' of genes peaking in at least four datasets. Then I computed IEGs and GO terms enrichment analysis for each subset of genes peaking in any different number of datasets. Genes that are higher ranking are expected to be true IEGs. Furthermore, I compared promoter dynamics across datasets and I used the conserved order of time of peaking to build a model of the IER.

2.12. Testing for statistical enrichment

IEG and TF enrichment in the group of genes classified as a peak in each dataset and in the permissive and robust set is performed to validate the peak classification method, as they are expected to contain more known IEGs than the non-peaking genes.

The list of 234 known IEGs (Arner, Daub et al. 2015) was assembled from 20 published human and mouse datasets from the literature. The list of IEGs includes genes identified in cells and/or responses which are not assessed in this project. From this list, 212 known IEGs were detected across the eight datasets during the classification step. I found 204 known IEGs to peak in at least one datasets (Appendix Table 1).

The example in Figure 6 shows how the enrichment of known IEGs CAGE TSSs is calculated for one of the eight datasets (i.e. MCF7_EGF1). I compared the proportion of peaking CAGE TSS assigned to known IEGs with the proportion of peaking CAGE TSSs assigned to candidate IEGs. Fisher's exact test is applied to the contingency matrix of peaking known IEG CAGE TSSs versus peaking candidate IEG CAGE TSSs.

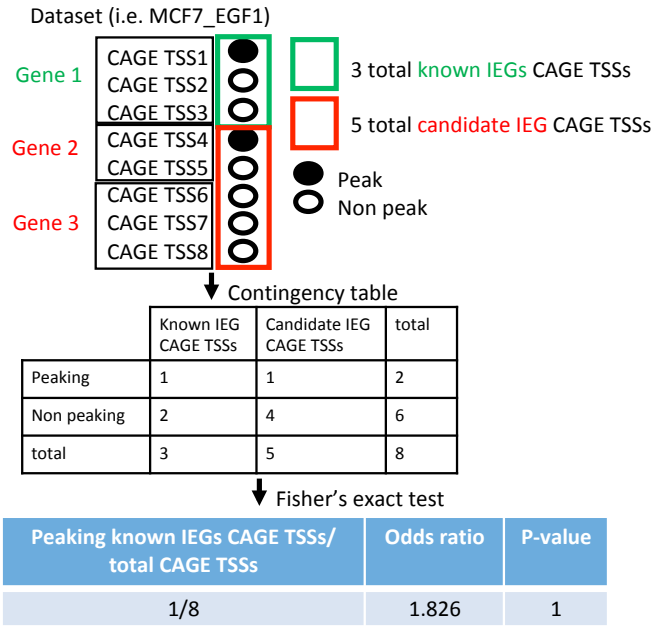
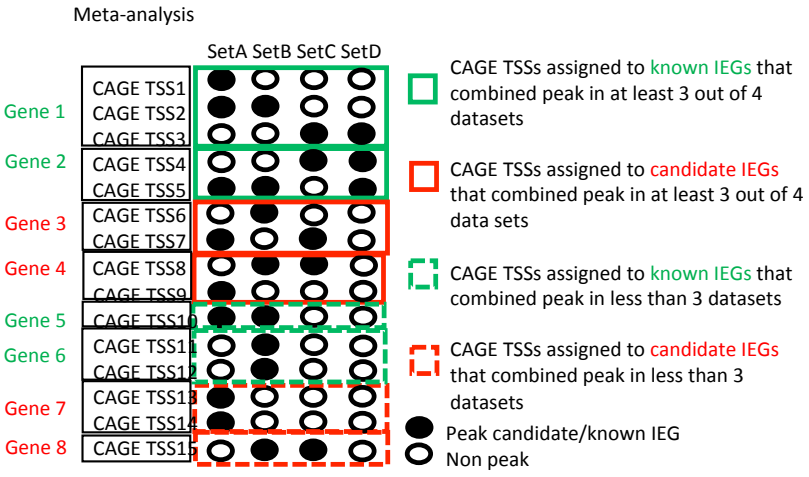


Figure 6 IEGs enrichment in each dataset. Schematic view of the approach used to compute the enrichment for peaking known IEGs CAGE TSSs (in green) in respect to peaking candidate IEGs (in red). The odds ratio and the p-value are calculated with Fisher's exact test using the numbers in the contingency table.

To compute the enrichment of peaking CAGE TSSs assigned to known IEGs in the permissive and robust sets I used a similar approach, which can be visualized in Figure 7. I compared the proportion of peaking CAGE TSSs assigned to the known IEGs in each set of shared peaking genes (in the example 3 out of 4 datasets) with the proportion of peaking CAGE TSSs assigned to IEGs in the remaining set (in the example 1 to 2 datasets). The odds ratio and the p-value was assigned using Fisher 's exact test.



Contingency table

	CAGE TSSs of genes that peak in at least 3 out of 4 data sets	CAGE TSSs of genes that peak in less than 3 datasets	total
Peaking known IEGs' CAGE TSSs	10	4	14
Peaking candidate IEGs' CAGE TSSs + non peaking known IEGs' CAGE TSSs	26	20	46
total	36	24	60

Fisher's exact test

Peaking CAGE TSSs assigned to known IEGs globally peaking in at least 3 out of 4 datasets / total CAGE TSSs	Odds ratio	P-value
10/60	1.903	0.367

Figure 7 IEGs enrichment in sets of shared peaking genes. Schematic view of the approach used to compute the enrichment for peaking known IEGs CAGE TSSs in the set of genes globally peaking in at least 3 out of 4 datasets in respect to the genes globally peaking in less than 3 datasets. The odds ratio and the p-value are calculated with Fisher's exact test using the numbers in the contingency table.

To assess the extent of the bias possibly caused by the higher number of TSSs associated to genes assigned to the peak model across multiple datasets I also performed a more stringent enrichment analysis. To test the IEGs enrichment, only the FANTOM defined canonical TSS, the p1, of each gene was used and the results, as described in Chapter 4.4, support the enrichment analysis described in this paragraph.

I also looked for enrichment in TF coding genes. To this purpose I used a list of 453 manually curated sequence-specific DNA-binding transcription factors extracted from the list created by Vasquerizas et al. (Vaquerizas, Kummerfeld et al. 2009) and applied the method described for IEGs enrichment.

2.13. Transcription factor binding sites enrichment

To assess the enrichment of TF binding profiles for the CAGE TSS regions, I used the entire collection of motifs contained in the JASPAR CORE subcollection of non-redundant profiles (Mathelier, Zhao et al. 2013) (updated on January, 2017). I used bedtools to extract the FASTA files for each region and I used the FIMO function (Grant, Bailey et al. 2011) (MEME version 4.11.2 patch 2) to detect the occurrence of the given motifs. FIMO converts the frequency matrix obtained matching the motif to the sequences in log odds scores and then estimates the p-values from the probability distribution of all possible match scores to the motifs (Grant, Bailey et al. 2011).
The significant TF binding motifs (FDR ≤ 0.05) were selected and Fisher's exact test was applied to calculate the significance (q value <=0.05, Benjamini & Hochberg correction) for the recurrence of each motif in the known IEGs and in the robust set in respect of the total tested genes.

2.14. Functional peaking genes set enrichment analysis

I used GOrilla (Eden, Navon et al. 2009) and InnateDB (Breuer, Foroushani et al. 2012) platforms to infer functional enrichment (FDR <= 0.05) for the set of genes peaking across the eight datasets and in the robust set, respectively. The background gene set for GO terms enrichment analysis consists in the total 12,132 genes analysed across the eight datasets.

3. Comparative analysis of time-course gene expression datasets

3.1. Introduction

RNA-sequencing has been proven to be an excellent procedure to compare cellular whole transcriptome expression and regulation between different conditions, such as disease and health or treatment and control, however, the study of dynamic biological processes such as response to drugs or development, involves the analysis of time-course datasets. In order to classify genes based on the change in the signal across time points, differential expression and clustering analysis, along with appropriate statistical methods to accurately handle count-based data, have been proposed (Bar-Joseph, Gitter et al. 2012).

For example, Nueda et al. (2014) introduced general linear regression in their algorithm, maSigPro, to model RNA-seq time course gene expression profiles. The algorithm measures differences in expression between time points in one or across different time-series and apply clustering analysis to group together similar expression profiles (Nueda, Tarazona et al. 2014). A more appropriate assumption to describe count-based data, the negative binomial distribution, was used by Anders et al. (2010) in the DESeq algorithm for differential expression analysis (Anders and Huber 2010). Despites the advantages of using time courses in gene expression analysis and the described improvements in the methods used, the comparisons across time-course datasets using differential expression analysis and clustering techniques is still challenging due to the impact of lowly expressed genes and differences in procedural variables such as time course length and distribution of sampling time points. Their limitations when used to compare datasets obtained using different protocols makes alternative approaches attractive. Aitken et al. (2015) developed a novel classification method making use of Bayesian model selection to classify longitudinal gene expression profiles to predefined

mathematical functions representing a range of possible trajectories over time (Aitken, Magi et al. 2015). Aitken et al. (2015) successfully classified known IEGs to a peaking model in four FANTOM5 CAGE time-series datasets, characterized by different sampling times. Tullai et al. (2007) found that a large number of genes induced with delayed kinetics was protein synthesis-independent and they are part of the IER (Tullai, Schaffer et al. 2007). However, in Aitken et al. (2015) the peak model doesn't consider the possibility of a delay in the first stage of response. In this chapter, I describe and justify modifications introduced in this method, allowing applications to longer CAGE time course datasets to be characterized by a delay in the activation of transcription. Furthermore, to show that the method of Aitken et al. (2015) can be adapted to the study of different mathematical functions of interest, I present an example of the application of the algorithm using a simulated dataset generated according to a simple piece-wise linear model.

3.2. A refined time-course classification technique

To classify time-series data for each CAGE defined TSS, we refined a previously published method (Aitken, Magi et al. 2015) which fits different mathematical models (or 'kinetic signatures') to individual expression profiles, assessing the fit using nested sampling (Aitken and Akman 2013) to compute the marginal likelihood, log *Z*. The time series were normalized such that the minimum and maximum median expression across the time series was set to 0 and 10, respectively.

The kinetic signatures considered were: linear, decay, dip and peak (Figure 1). The peak kinetic signature considered in the original method (Figure 8.A) was modified to allow a delay (t_d) before expression increases in exponential fashion. For the early Peak (Figure 8.B), parameter t_s is the time duration of the initial increase in expression, p_1 is the expression at time 0, and p_2 is the increase in expression such that expression $y= p_1 + 0.7* p_2$ at the time of peaking, $t_p = t_d + t_s$.

$$\delta = \frac{\log(0.3)}{t_s}$$

$$y = p_1; t \leq t_d$$

$$y = p_1 + p_2 * \left(1 - e^{\delta(t-t_d)}\right); t_d \leq t \leq t_d + t_s$$

$$y = p_1 + p_2 * (1 - e^{\delta(t-t_d)}) - p_2 * \left(1 - e^{\delta(t-t_d-t_s)}\right); t > t_d + t_s$$

Here, t_d is constrained between 1 minutes and 60 minutes and t_s is constrained between 15 and 75 minutes. However, an alternative rate= $\frac{\log(0.1)}{t_d}$ was also used, constraining the upper and lower limits of t_s to be later in time (75 minutes and 135 minutes, respectively), to model the dynamics of transcripts peaking later in time (Figure 8.C), and the best fitting model was selected during the decision step.

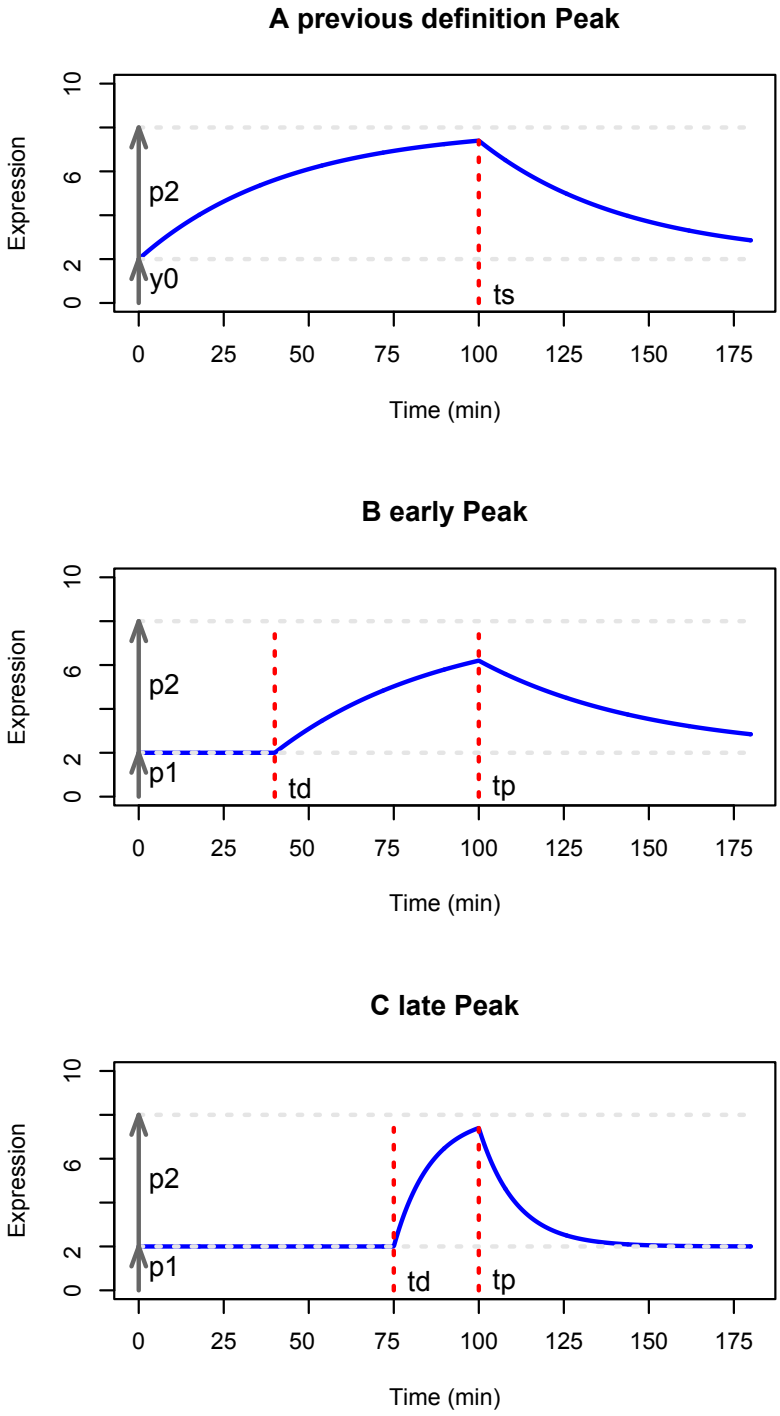


Figure 8 Delayed peak model. Plot A shows the peak model as defined in Aitken et al., while plot B and C show the dynamic of the delayed peaking transcripts with faster and lower dynamics.

The values used to calculate the rates δ for the two alternative delayed peak models (0.1 and 0.3), were chosen empirically observing the results and represent a sharper general trend and a milder general trend. However, as I

show below with an example (Figure 9 and Table 2), changing the rate does not change drastically log *Z*. Here, I compare the results obtained fitting *JUN* expression profile from PMDM_LPS dataset with the models used in the analysis (0.1 and 0.3) and three additional peak models which make use of three different rates: 0.4, 0.2 and 0.01. In this example the evidence is generally much higher for the peak models than the other models, with a better score for the peaking rate = 0.3.

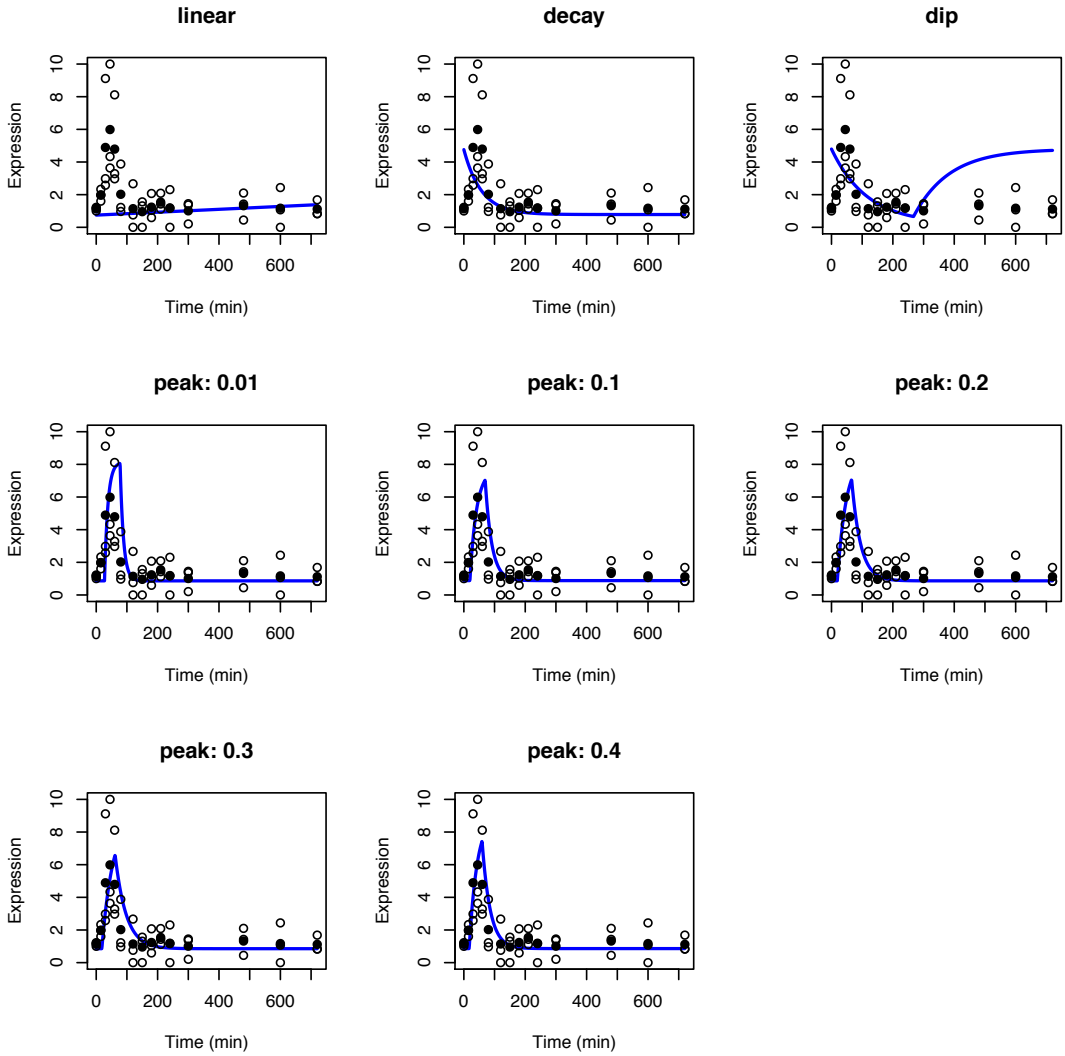


Figure 9 Model comparison. JUN expression profile for the three replicates (empty circles) and the average expression (filled black circles) for JUN three replicates. All the peaking models in this example are constrained to have 15<=ts<=75 minutes.

Table 2 Model's evidence comparison for JUN expression profile. Log Z* assumes parameters in the range of 0 and 1, therefore to obtain the evidence log Z for the data I summed it with the log volume.

Model	Log *Z**	Log SD	Log Volume	Log *Z*
Linear	-31.58	0.30	4.61	-26.98
Decay	-30.72	0.18	6.85	-23.87
Dip	-39.78	0.30	8.72	-31.06
Peak: 0.01	-25.56	0.23	11.98	-13.58
peak: 0.1	-25.34	0.23	11.98	-13.36
peak: 0.2	-25.31	0.24	11.98	-13.33
peak: 0.3	-25.25	0.23	11.98	-13.27
peak: 0.4	-25.77	0.23	11.98	-13.79

Normalising the data such that expression lies in the range 0-10 allowed the prior probability distribution of parameters to be restricted to plausible values that applied to all time series. The fit of models to data was improved as a result. To account for any impact on the log *Z* calculation, I generated synthetic time series datasets using parameter values selected uniformly at random between the prior ranges (upper and lower limits) to generate one replicate (r_1), and generated two other replicates by adding and subtracting a given amount of noise to the first replicate.

The noise was generated by randomly drawing a value from a negative binomial distribution BNM(*size,μ*), with *size* =1/(bcv^2) and $\mu = r_1$, using the biological coefficient of variation (BCV) = 0.2 (McCarthy, Chen et al. 2012) to calculate the dispersion parameter. Negative binomial distribution is the preferred method to statistically analyse sequencing data because read counts distribution is not normal and is generally characterized by greater variance than the mean (Di 2015).

Model fitting was applied to 1000 such synthetic datasets per model (using the same noise values for each model on each of the 1000 iterations) and log *Z* was calculated for each dataset and each model.

Comparing the obtained log *Z* distribution for each model, I observed an advantage for each of the complex models (peak, dip and decay) that was consistent over the range of log *Z* values obtained for the linear model across the 1000 iterations:

- $\text{Mean}\,(\log Z_{peak\ model}) - \text{Mean}\,(\log Z_{linear\ model}) = 6.8$
- $\text{Mean}\,(\log Z_{decay\ model}) - \text{Mean}\,(\log Z_{linear\ model}) = 3.1$
- $\text{Mean}\,(\log Z_{dip\ model}) - \text{Mean}\,(\log Z_{linear\ model}) = 3.2$

To offset the advantage of the complex models, an advantage value was subtracted from the log *Z* values calculated for CAGE TSS data when making the categorisation decision. The advantage value for each complex model in respect to the linear model was calculated as the mean difference of the log *Z* distribution, plus two standard deviations (0.9 for the peak model, 0.6 for the dip model and 0.7 for the decay model) to compensate in at least the 95% of the cases. After the adjustment, the time series was assigned to the model with higher log *Z*.

We also defined a margin between the linear model and the more complex models such that if the difference in log *Z* between the linear model and the more complex model was greater than an empirically chosen value (margin=4 in our analysis) the classification of the time series was considered linear, else it was assigned to the 'No decision' category. High variability between replicates, which is particularly evident for lowly expressed genes make it difficult to confidently assign them to any particular signature, therefore they are also assigned to the 'No decision' category.

3.2.1. Example of a simple model applied to synthetic data

The marginal likelihood Z corresponds to the probability of observing the data D given that the data corresponds to the considered model H_i, $P(D|H_i)$, and can be obtained multiplying the likelihood function and the prior, integrating over the space of parameter values. LogZ can be computed by a brute force approach, which is very slow and computationally expensive, or using the more efficient nested sampling algorithm which exploits statistical properties of the prior volume reduction. As explained in (Skilling 2006), on each iteration of the nested algorithm, the prior mass shrinks by removing the sample with the lowest likelihood Li and replacing it with another sample with likelihood higher than Li, until the termination criterion is reached (when the change in log Z become negligible).

Here, I designed a simple piecewise linear model to illustrate the log *Z* calculation across different models with synthetic exemplificative data, and to compare the brute force approach and the nested sampling approach.

The piecewise linear model below is parameterized by t_s, the time of switching, p_1, the minimal expression and p_2, which represents the change in expression at t_s.

$$y = p_1 + t * \frac{p_2}{t_s}; t \leq t_s$$

$$y = p_1 + p_2 - (t - t_s) * \left(\frac{p_2}{max_t - t_s}\right); t > t_s$$

I created the synthetic data at 15 time points by using this function with parameters: p1=3, p2=10 and ts=150 and adding noise to simulate 2 additional replicates of the data.

After running the nested sampling algorithm for all the models as defined in (Aitken and Akman 2013, Aitken, Magi et al. 2015) on the synthetic data, I obtained a set of values which are listed in Table 3.

Table 3 Model parameters for synthetic data. The values in the table are inferred running nested sampling algorithm on synthetic data from a piece wise linear function. Log Z* assumes parameters in the range of 0 and 1, therefore to obtain the log Z for the data I summed it with the log volume.

Model	Log *Z**	Log volume	Log *Z*	P_1	P_2	T_s		P_1 SD	P_2 SD	T_s SD	
Piecewise linear	-14.77	9.00	-5.77	2.77	10.51	144.87		0.18	0.39	9.10	
Dip	-35.97	8.72	-27.25	3.16	7.73	50.12		0.56	0.93	0.92	
Decay	-43.97	6.85	-37.12	2.50	4.13	T_h		0.36	0.85	T_h SD	
						88.60				23.44	
Delayed peak	-38.74	11.98	-26.76	2.84	8.78	T_d	T_s	0.20	1.87	T_d SD	T_s SD
						65.76	17.53			9.13	10.01
Linear	-39.44	4.61	-34.83	P_1		P_2		P_1 SD		P_2 SD	
				7.35		5.89		1.52		3.10	

Higher log *Z* assigned to the piecewise linear model indicates an expected better fitting in comparison with the other models (Figure 10).

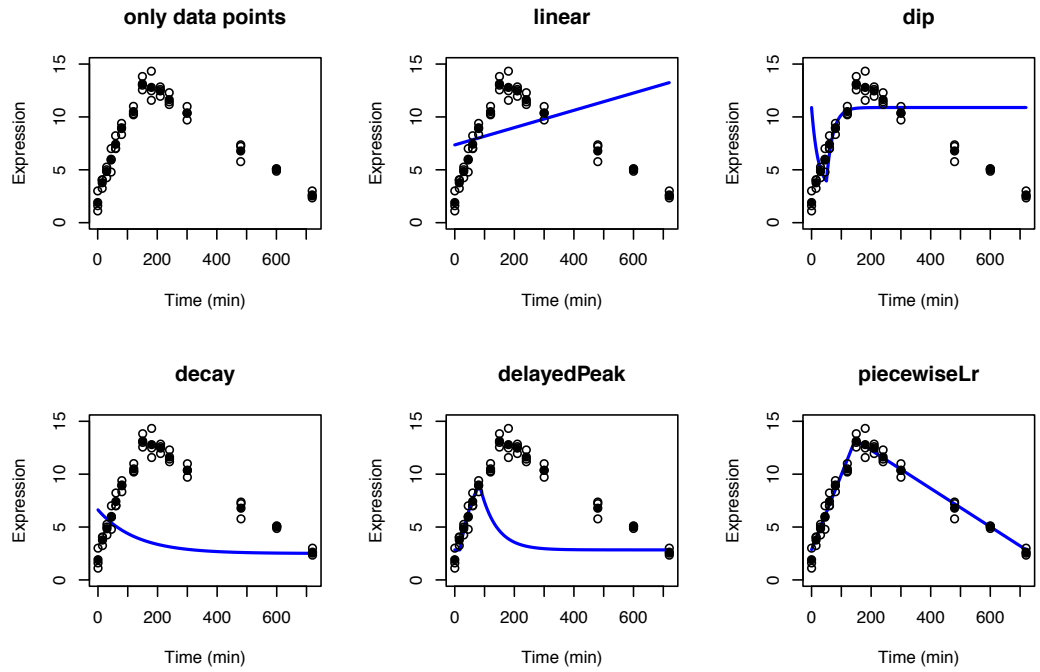


Figure 10 Fitted models for synthetic data. The plots show the expression of synthetic data for the 15 time points selected (circles, median value is filled). The kinetic signature function (in blue) is computed using the parameter means.

For the piecewise linear model the results obtained with a brute force approach, computing log *Z* for each cell of a 3-dimensional grid (one dimension for each parameter), is comparable to the nested sampling approach. I fixed the minimum and maximum values for the three dimensions as: min p_1=0, min p_2=0, min t_s=15, max p_1=3, max p_2=15, max t_s=195; I separated the grid in 150*150*150 cells, and then I computed the log *Z* as the sum of $\log likelihood * \log volume$ over the grid, where the summation properly accounts for the addition of log values:

$$logZ = \sum_{p_1}\sum_{p_2}\sum_{p_3} \log\left(P(x|p_1, p_2, t_s)\right)$$

The results show that both approaches converge on very similar values of parameter estimates, but with the brute force approach the computing time is two orders of magnitude higher (Table 4).

Table 4 Comparison between nested sampling and brute force approaches. For each approach the log Z and the computed parameters are reported. p1 , p2 and ts are the estimates of the parameters corresponding to the highest local log Z value (i.e. lighter colour in the grid in Figure 11).

Method	Log *Z*	p_1	p_2	t_s	Time
Nested sampling	-5.77	2.77	10.51	144.87	18.82
Brute force	-5.26	2.96	10.07	149.09	2043.20

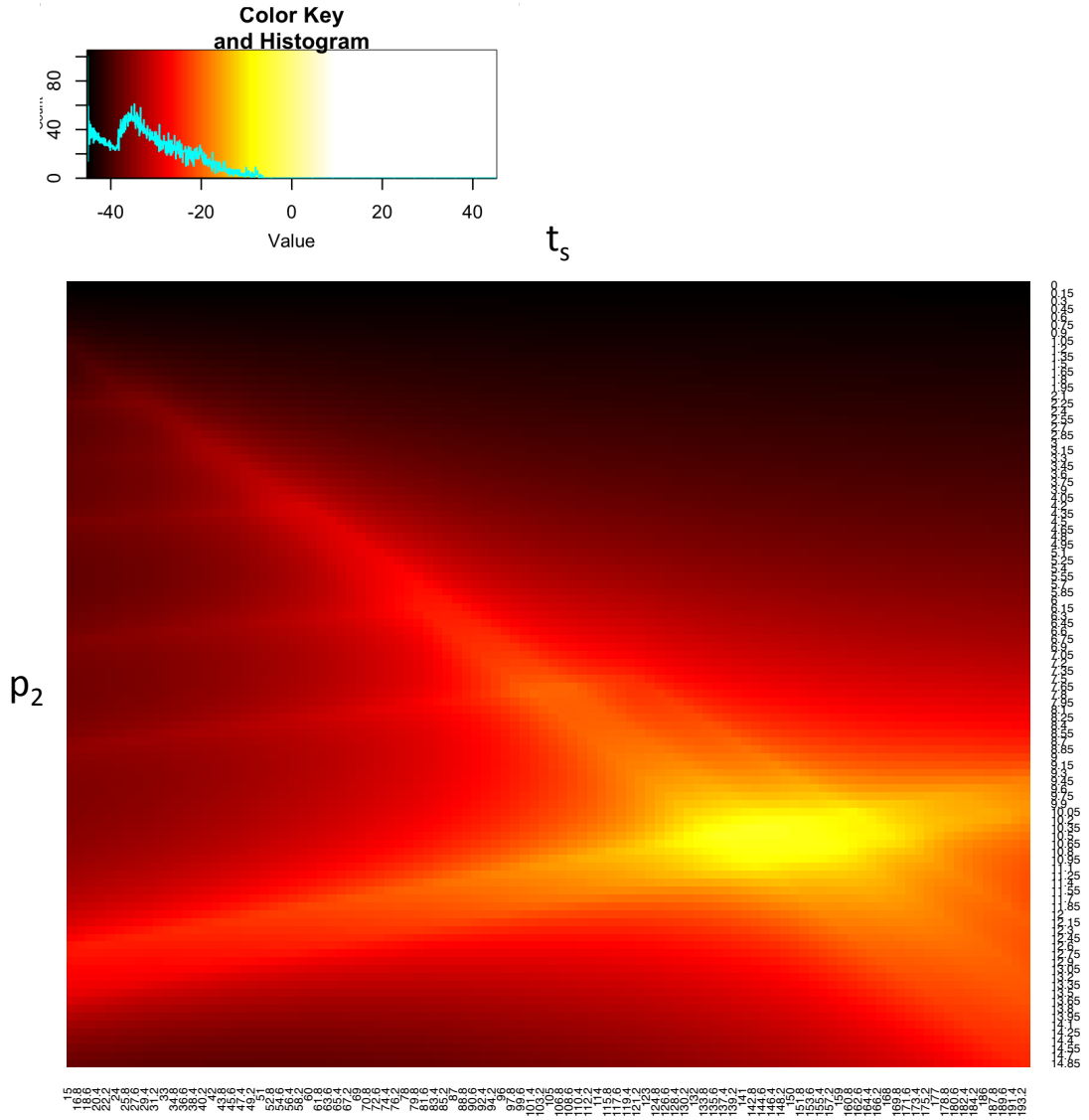


Figure 11 A slice through the likelihood function. The figure shows a 2-dimensional likelihood distribution. The colour of each cell in the grid represent the likelihood for 100 different values of p_2 and t_s and p_1 fixed to 2.77. Lighter colours correspond to higher likelihood.

3.2.2. Long macrophage time series

In contrast to the peak model defined in (Aitken, Magi et al. 2015), real time series datasets, can show an initial period of time in which expression is constant, followed by a rapid and transient increase in expression after this delay. Adding a delay to the peak model can significantly increase the number of classified genes in such datasets. This was particularly evident when I analysed the long PMDM_LPS time course (0 to 1,440 minutes) (Table 5) using both the standard peak and delayed peak models. The minimum and maximum values for the three parameters of the standard peak model in the long time course were set to: min p_1=0, min p_2=0, min t_s=15, max p_1=3, max p_2=15, max t_s=960 while the minimum and maximum values for the four parameters of the delayed peak model were p_1=0, min p_2=0, min t_d=75, min t_s=1, max p_1=3, max p_2=15, max t_d=480, min t_s=480.

Table 5 shows that adding the delay to the parameters allowed an additional 3,881 CAGE TSSs to be assigned to the delayed peak model, of which 271 were originally assigned to another model and 3,610 could not be assigned to any model and therefore were classified as 'No decision'. Figure 12 shows representative examples where the delayed peak model describes the behaviour of the data better than the peak model. At the same time the enrichment of known IEGs classified to the peak model remains similar when applying the delayed peak model instead of the peak model (OR = 4.3, p-value = $2.2e^{-16}$ and OR = 4.7, p-value = $2.2e^{-16}$, respectively). This suggests that the additional 3,881 CAGE TSSs assigned to the peak model do not markedly increase the proportion of false positives classified with this model.

Table 5 Comparison between delayed peak and peak model. Contingency matrix showing the frequency of classification with the different models when delayed peak or peak functions are considered.

PMDM_LPS 0 to 1,440 min	Peak	Dip	Decay	Linear	NO DECISION	Tot
Delayed peak	3,239	205	40	26	3,610	7,120
Dip	15	1,271	0	0	45	1,331
Decay	203	0	1,774	0	337	2,314
Linear	25	0	0	21	30	76
NO DECISION	339	129	37	62	3,051	3,618
Total	3,821	1,605	1,851	109	7,073	14,459

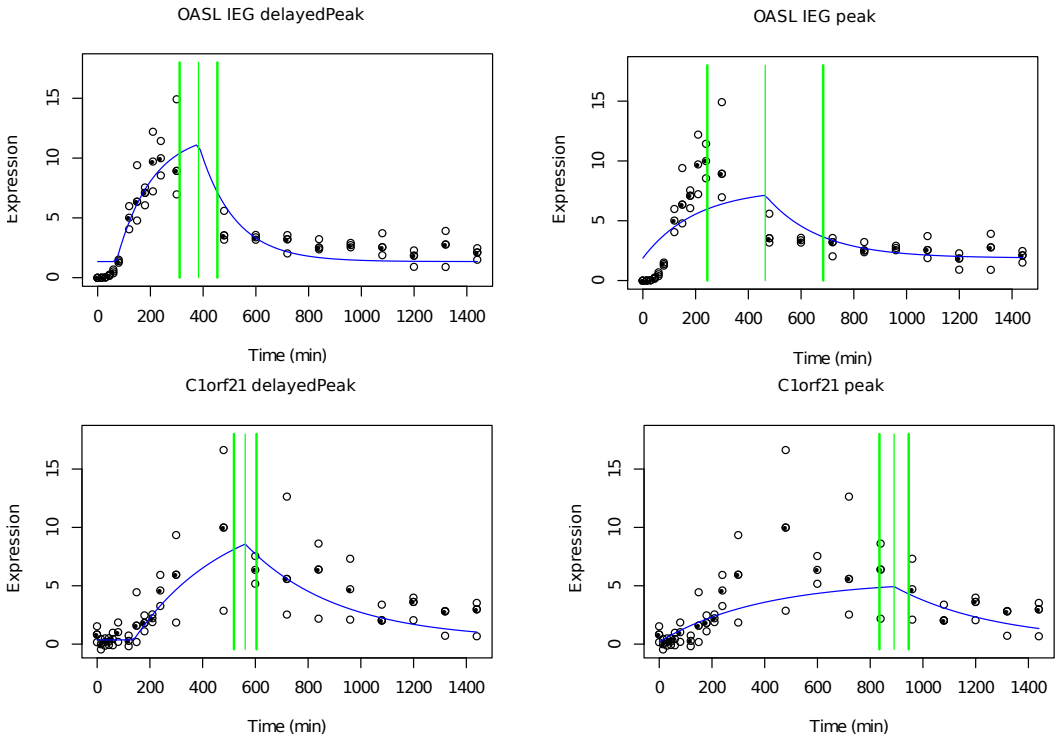


Figure 12 Delayed peak and peak model fitted data. Expression data and kinetic signatures for the OASL and C1orf21 genes for delayed peak (left) and standard peak (right) models. The expression data are plotted as empty circles, median values as filled circles, while the model function is indicated with a blue line. Green lines indicate the mean t_p (for the delayed peak model) and t_s and 1 standard deviation below and above.

3.3. Discussion

The IER mediates cellular changes affecting a plethora of physiological and pathological cellular processes induced by external stimuli such as cytokines, stress, growth factors and proliferation agents. However, the temporal

dynamics and the specificity of different cell and stimulus specific gene activation patterns are not well understood. Aitken et al. (Aitken, Magi et al. 2015) developed a method to identify candidate and known IEGs in four FANTOM5 CAGE datasets, PAC_FGF2, PAC_IL1B, MCF7_EGF1 and MCF7_HRG, exploiting the classical IEG expression dynamic which involves a brief and transient peak in expression during the first few hours after cell stimulation (Greenberg and Ziff 1984, Bahrami and Drabløs 2016).

Here we describe modifications to the method, so it can be used to classify temporal expression profiles in eight densely sampled CAGE time course datasets, including four additional datasets: PMDM_LPS, PEC_VEGF, PMSC_MIX and SAOS2_OST. The peak model was adapted to allow a delay in expression before an exponential increase is observed. Data were also normalized such that expression ranges between 0 to 10 units, allowing a better comparison of the parameters across all time-series. As a result, I see evidence of an improvement in the fit of the peak model to real datasets, especially for long time courses such as PMDM_LPS, in the terms of a higher number of CAGE TSSs assigned to the delayed peak model including many associated to known IEGs such as IL6 and CSF2. We also show that the approach can be applied to other cases, as customised mathematical functions of interest can be included.

4. Meta-analysis of the time course expression of protein-coding and non-coding genes

4.1. Introduction

The recent improvements on high-throughput gene sequencing techniques and longitudinal gene expression analysis made available a big number of time course gene expression datasets from experiments, such as the FANTOM project. A comparison across datasets is useful because it can shed light on core events common across biological systems as well as specific to a cell type and/or stimulus. However, until now, comparison across datasets has been mainly limited to datasets from similar systems obtained by a single laboratory. The reason of this is that the comparison of gene expression time course datasets with traditional tools requires the datasets to have similar sampling times and length. Furthermore, given the complexity and cost of time course sequencing experiments, the majority of the studies include few time points very distant between each other, which is a limitation when studying the immediate early response.

In this chapter I will describe an expansion of the IER meta-analysis undertook by Aitken et al. (2015) on four FANTOM5 time course datasets. The objective here is to provide a unique comprehensive comparison across different cell types and stimuli for both protein coding and non-coding genes and to gain insights on the core events of the immediate early response common to different biological systems.

I considered eight densely sampled, and well replicated, FANTOM5 CAGE time course datasets (Figure 2). These datasets consist of a variety of primary human cell samples and cell lines responding to a range of stimuli: growth factors, hormones, drugs, pro-inflammatory cytokines and bacterial endotoxin. Many of these time course datasets relate to widely used biomedical model

systems where the IER is known to occur, but had not previously been studied by deep sequencing of CAGE libraries. The datasets include the four datasets compared by Aitken et al. (2015): MCF7_EGF1, MCF7_HRG, PAC_IL1B and PAC_FGF2, with the addition of other four datasets: SAOS2_OST, PMSC_MIX, PEC_VEGF and PMDM_LPS (2.1 Data Resources, for more details about these time courses). These diverse data provided a potent resource to discover core features of the IER conserved across human cell types and stimuli. As exemplified by FOS and JUN (Figure 13), the responses of known immediate early genes often show characteristic expression peaks early in the time series datasets, though even for these well-studied genes I observed substantial variation in the magnitude, timing and duration of peaks across cell types and stimuli. This emphasises the challenges presented in IEG detection, even when studying known IEGs using a uniform experimental platform. These challenges increase when examining data for lowly expressed genes, such as the non-coding RNA species that appear to be part of the immediate early response (Aitken, Magi et al. 2015).

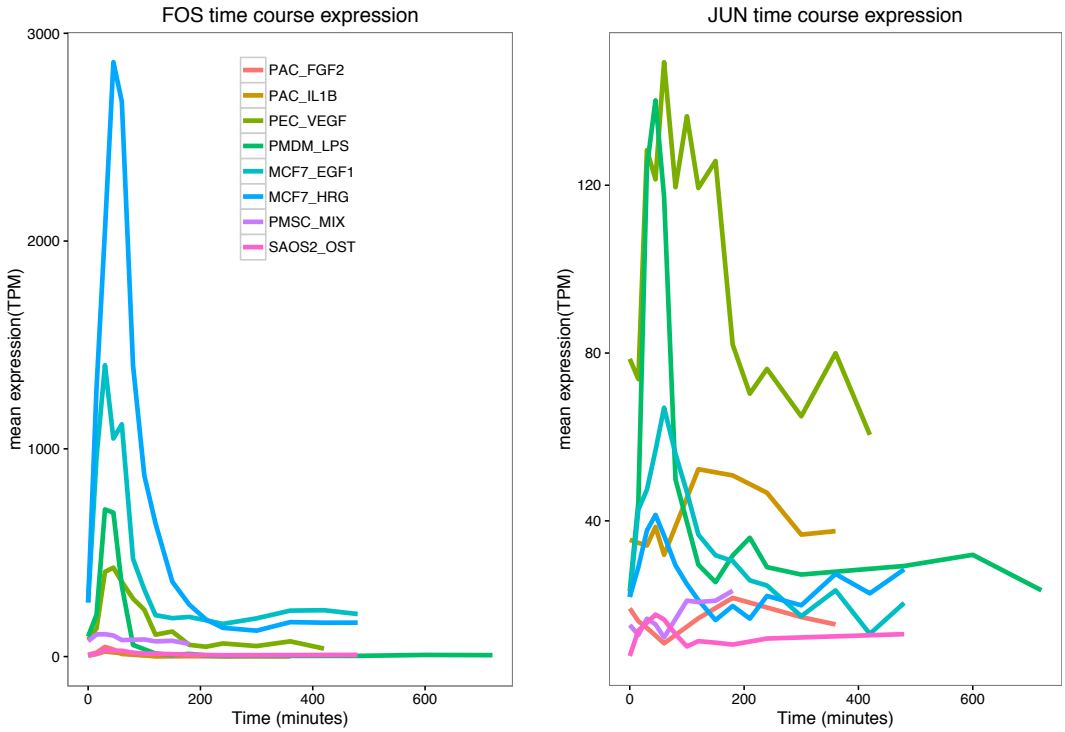


Figure 13 The expression profiles of FOS and JUN gene promoters in the eight datasets. Cage cluster expression (mean TPM in three replicates) is plotted over time for FOS (left) and JUN (right). Line colour defines the datasets

4.2. Classification of protein-coding and non-coding genes by a model fitting procedure

Optimising and refining the approach developed by Aitken et al (2015), I defined four mathematical models representing archetypal expression profiles of interest over time: peak, linear, dip and decay, and I assessed the fit of each model to the expression profile of each gene using nested sampling to compute the marginal likelihood, log Z, from which the best model was selected.

Across the eight time series datasets, I considered all CAGE clusters corresponding to the TSSs of known Ensembl transcripts (Hubbard, Barker et al. 2002), encompassing between 10,513 (corresponding to 7,706 Ensembl genes) and 14,376 (8,951 genes) protein coding CAGE TSSs, and between

1,202 (692 genes) and 1,640 (858 genes) non-coding RNA (ncRNA) CAGE TSSs (Table 6 and Table 7).

Between 15% and 42% of protein-coding CAGE TSSs, and between 15% and 33% of non-coding TSSs were confidently classified to one of the four models, depending on the dataset (Figure 14 A and B, Table 6 and Table 7). The remainder ('No decision') could not be rigorously classified to a single model and were omitted from further analysis.

The differences between the number of genes classified in each class and the total number of tested genes is caused by the fact that a gene is represented by multiple CAGE TSS that are classified independently. Therefore, alternative CAGE TSSs assigned to the same gene can be classified to different models.

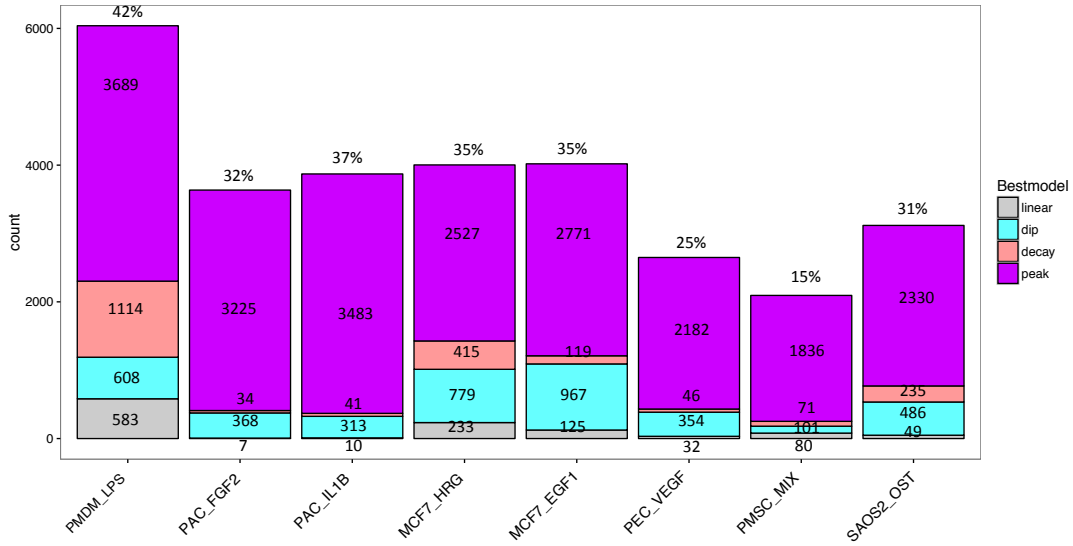
Table 6 Classification of promoter dynamics for protein coding genes. The table summarizes the number of CAGE clusters (representing TSSs) and Ensembl genes in each dataset and the number of protein coding CAGE clusters and Ensembl genes classified to each model (or to no models: 'No decision') in each dataset.

Protein coding genes	PMDM_LPS		PAC_FGF2		PAC_IL1B		MCF7_HRG	
	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes
Total tested	14,376	8,951	11,235	8,112	10,513	7,706	11,513	8,511
No decision	8382	5890	7601	5894	6666	5356	7559	6045
Peak	3689	2731	3225	2852	3483	3053	2527	2200
Decay	1114	998	34	34	41	41	415	400
Dip	608	585	368	360	313	306	779	740
Linear	583	541	7	7	10	10	233	231
	MCF7_EGF1		PEC_VEGF		PMSC_MIX		SAOS2_OST	
	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes
Total tested	11,352	8,458	10,686	8,040	13,881	9,197	11,521	8,639
No decision	7370	5906	8072	6469	11793	8228	8421	6731
Peak	2771	2424	2182	1941	1836	1672	2330	2105
Decay	119	115	46	45	71	67	235	226
Dip	967	904	354	353	101	100	486	470
Linear	125	123	32	32	80	79	49	49

Table 7 Classification of promoter dynamics for noncoding RNA genes. The table summarizes the number of CAGE clusters (representing TSSs) and Ensembl genes in each dataset and the number of non-protein coding CAGE clusters and Ensembl genes classified to each model (or to no models: 'No decision') in each dataset.

Non coding genes	PMDM_LPS		PAC_FGF2		PAC_IL1B		MCF7_HRG	
	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes
Total tested	1,209	673	1,202	692	1,210	681	1,342	763
No decision	949	589	856	551	861	550	893	594
Peak	183	140	324	254	322	251	381	288
Decay	63	57	3	2	2	2	29	26
Dip	6	6	19	18	25	25	33	33
Linear	8	8	0	0	0	0	6	6
	MCF7_EGF1		PEC_VEGF		PMSC_MIX		SAOS2_OST	
	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes	#CAGE TSSs	#genes
Total tested	1,345	769	1,377	772	1,252	692	1,640	858
No decision	965	624	1,086	665	1,068	612	1,255	725
Peak	306	235	254	205	178	146	338	256
Decay	6	6	5	5	2	2	8	8
Dip	59	56	28	24	2	2	38	37
Linear	9	8	4	3	2	2	1	1

A



B

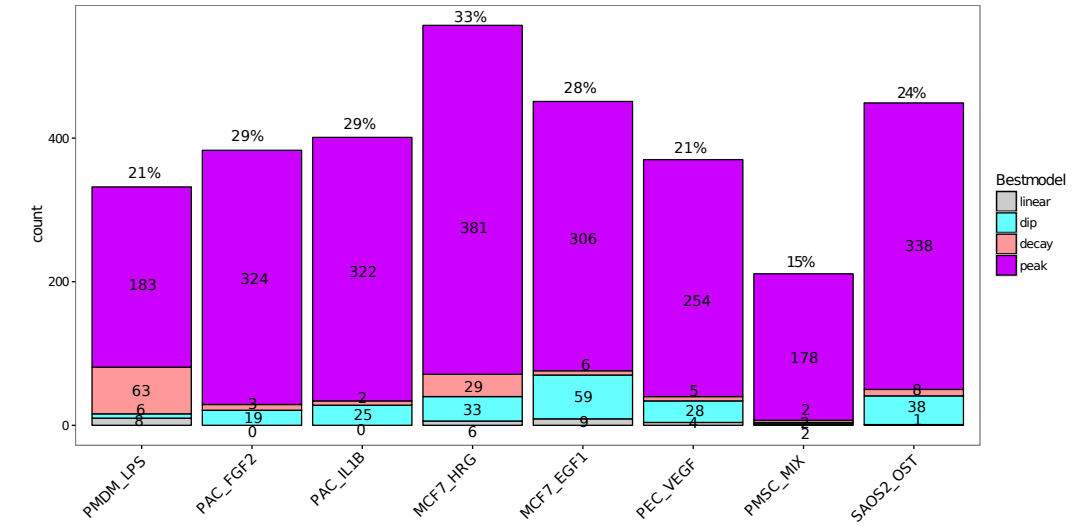


Figure 14 Model based classifications of CAGE TSSs. Histogram presentation of the CAGE derived TSS classification, with columns representing datasets and colours representing the four models of interest (peak, decay, dip, linear), for protein coding (A) and non-coding (B) CAGE TSSs.

The peak model had the highest number of assignments in all the datasets for both protein-coding and non-coding RNA genes. Of 12,132 total Ensembl protein-coding genes tested and 1,226 non-coding RNAs, I found 8,785 (72%) of protein coding and 779 (64%) of non-coding Ensembl genes to peak in at

least one of the datasets (with at least one CAGE TSS). In contrast, relatively few genes were classified to the peak model in multiple datasets. Only 42 protein coding and 15 non-coding RNA genes shared peaking model across at least seven out of eight datasets (Figure 15), underlining the high variability of transcriptional responses seen for the same promoters across time series. These 42 protein coding and 15 non-coding genes RNA constituted the 'robust set' of candidate IEGs. I also defined a less stringent 'permissive' set of 1,304 protein coding IEG candidates shared across four (797 genes), five (325 genes), six (140 genes), seven (37 genes) or eight (5 genes) out of eight datasets.

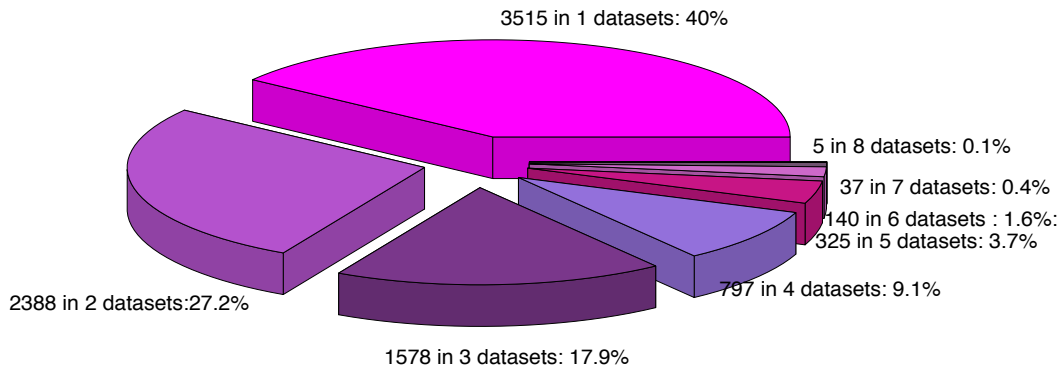


Figure 15 Distribution of peak classified TSSs shared across multiple datasets. Each slice represents the proportion of genes associated to peaking CAGE TSSs shared by 1 to 8 datasets. The robust set includes 42 genes peaking in at least 7 out of eight datasets (37 genes in 7 datasets and 5 genes in 8 datasets).

4.3. Known IEGs are particularly enriched in the group of peaking genes of the robust set

I assessed promoter classifications by testing the enrichment of known IEGs and TFs CAGE TSSs (Chapter 2.12) within each class, for each dataset. The peak class was enriched for known IEGs in all datasets (Figure 17, Table 8), but failed to reach statistical significance in PMSC_MIX (OR = 1.3, p = 0.2). The decay class was strongly enriched for both PMSC_MIX (OR = 19.7, p = 1.1e-14) and PAC_FGF2 (OR = 6.4, p = 5.1e-3). This is likely caused by a

failing of the sampling for the time series analysed. If there are not enough time-points sufficiently early, the initial induction of IEG expression could be missed and instead the first thing detected might be the IEG's peak expression followed by its decline in expression. In that case an exponential decline in expression (i.e. the decay signature) will provide the best fit. Figure 16 show the data and the fitted model for the 4 known IEGs classified as decay in PAC_FGF2. In *IL6* and *SOCS3* time-courses, the first time point (time 0) is characterized by a lower expression than the model starting point. In *CLIC4* time-course the expression peak at about one hour; however, the sampling around this time is not dense enough to catch the peaking trajectory and the data are instead classified as 'decays'. Additional sampling could improve the performance of the model fitting in detecting all the peaking expression profiles.

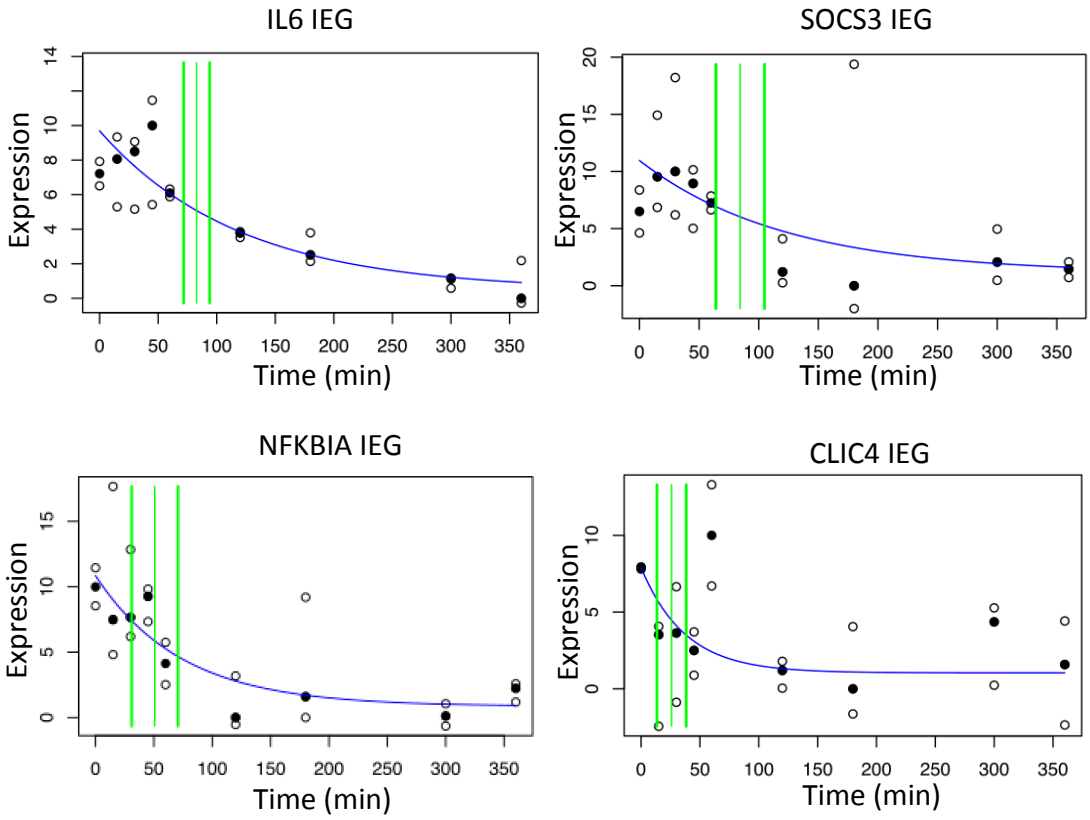


Figure 16 Decay IEGs. The plots show the expression profile for the 4 IEGs classified as decay in PC_FGF2 dataset (circles, median value is filled), the fitted decay model (blue line) t_h and 1 standard deviation above and below (green lines).

In contrast, the 'no decision' class, which contains all the CAGE TSSs that could not be assigned to any model, is significantly depleted for IEG CAGE

TSSs across all the datasets. We might expect that genes with a non-defined expression trajectory to be disproportionately non-IEG as the algorithm assign the majority of the known IEGs to the IEG archetype, the peak model.

Table 8 Known IEG enrichment across models and datasets. The significance of enrichments (Odds Ratios > 1) and depletions (Odds Ratios < 1) were computed using Fisher's exact test.

Dataset	Peak			Dip			Decay		
	# IEG CAGE TSSs / total CAGE TSSs	OR	p	# IEG CAGE TSSs / total CAGE TSSs	OR	p	# IEG CAGE TSSs / total CAGE TSSs	OR	p
PMDM_LPS	264/ 3,689	5.8	2.2e-16	2/ 608	0.1	7.3e-6	8/ 1,114	0.2	6.2e-7
PAC_FGF2	87/ 3,225	1.5	4.2e-3	8/ 368	1.1	0.9	4/ 34	6.4	5.1e-3
PAC_IL1B	116/ 3,483	2.1	9.1e-8	3/ 313	0.4	0.2	2/ 41	2.3	0.2
MCF7_HRG	140/ 2,527	6.4	2.2e-16	10/ 779	0.6	0.2	7/ 415	0.9	0.9
MCF7_EGF1	71/ 2,771	1.9	3.3e-5	21/ 967	1.4	0.2	2/ 119	1	0.7
PEC_VEGF	127/ 2,182	4.3	2.2e-16	6/ 354	0.7	0.6	1/ 46	0.9	1
PMSC_MIX	36/ 1,836	1.3	0.2	2/ 101	1.3	0.7	16/ 71	19.7	1.1e-14
SAOS2_OST	72/ 2,330	3.6	4.0e-14	5/ 486	0.8	0.8	2/ 235	0.6	0.8
Dataset	**Linear**			**No decision**			**Total tested**		
	# IEG CAGE TSSs / total CAGE TSSs	**OR**	**p**	**# IEG CAGE TSSs / total CAGE TSSs**	**OR**	**p**	**# IEG CAGE TSSs / total CAGE TSSs**		
PMDM_LPS	1/ 582	0.1	1.3e-6	130/ 8,382	0.3	2.2e-16	405/14,376		
PAC_FGF2	1/ 7	7.9	0.1	133/ 7,601	0.6	6.6e-4	233/11,235		
PAC_IL1B	1/ 10	4.9	0.2	110/ 6,666	0.5	5.7e-7	232/10,513		
MCF7_HRG	1/ 233	0.2	0.1	63/ 7,559	0.2	2.2e-16	221/11,513		
MCF7_EGF1	2/ 125	1	1	91/ 7,370	0.5	4.4e-6	187/11,352		
PEC_VEGF	0/ 32	0	1	114/ 8,072	0.3	2.2e-16	248/10,686		
PMSC_MIX	3/ 80	2.5	0.1	160/ 11,793	0.5	1.4e-5	217/13,881		
SAOS2_OST	0/ 49	0	1	73/ 8,421	0.3	6.0e-11	152/11,521		

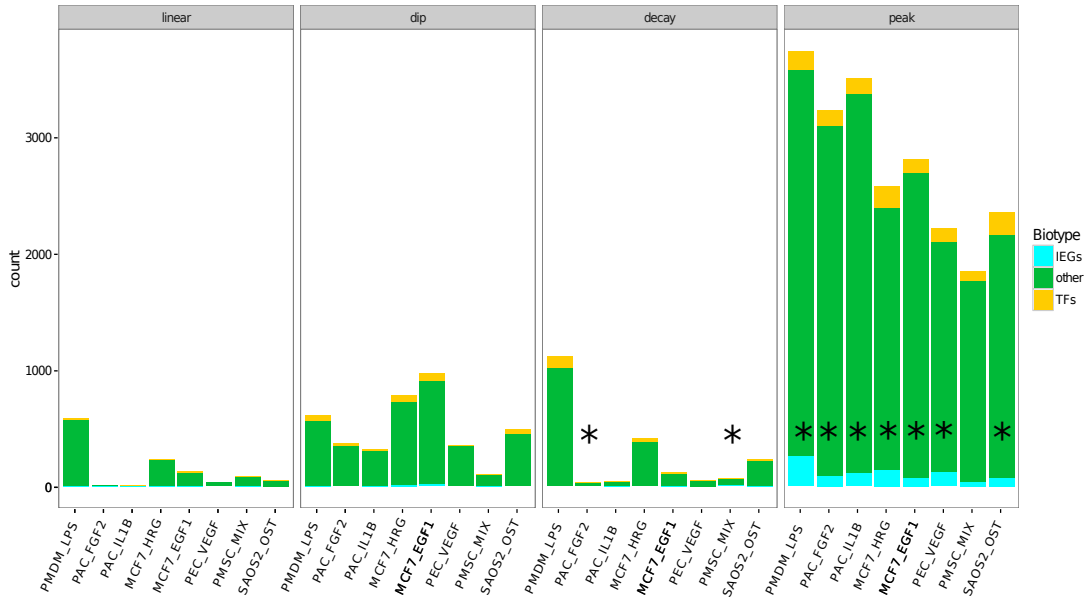


Figure 17 Known IEGs are enriched in genes classified to the peak model. Barcharts show the number of known IEGs (light blue) and TFs (yellow) recovered for the four models of interest (linear, dip, decay and peak) in each time series dataset (x-axis). Asterisks indicate the datasets which are characterized by known IEG enrichment (Fisher's exact test p-value<0.01).

Likewise, I analysed the enrichment for TF genes (Table 9), which includes many known IEGs, and I found significant enrichment (Fisher's exact test p-value<=0.01; Odds Ratio >1) in the peak subset of PMDM_LPS, MCF7_HRG, PEC_VEGF and SAOS2_OST. These results indicate that, similarly to the known IEGs, many TFs are rapidly and transiently activated across different datasets. Furthermore, a number of TFs show dip or decay expression in PMDM_LPS which indicates a PMDM_LPS-specific downregulation of these genes which could have a role in the IER.

Table 9 Known TF enrichment across models and datasets. The significance of enrichments (Odds Ratios > 1) and depletions (Odds Ratios < 1) were computed using Fisher's exact test. TF genes in this table include known IEGs.

Dataset	Peak			Dip			Decay		
	# TF CAGE TSSs/ total CAGE TSSs	OR	p	# TF CAGE TSSs/ total CAGE TSSs	OR	p	# TF CAGE TSSs/ total CAGE TSSs	OR	p
PMDM_LPS	223/ 3,689	1.2	0.2e-2	49/ 608	1.6	4.0e-3	97/ 1,114	1.8	9.5e-7
PAC_FGF2	152/ 3,225	1.1	0.3	26/368	1.7	0.2e-1	3/ 34	2.1	0.2
PAC_IL1B	157/ 3,483	1	0.8	17/ 313	1.2	0.4	3/ 41	1.7	0.4
MCF7_HRG	238/ 2,527	1.7	6.3e-11	55/ 779	1.1	0.5	29/ 415	1.1	0.6
MCF7_EGF1	142/ 2,771	0.8	0.5e-1	76/ 967	1.4	0.1e-1	10/ 119	1.5	0.2
PEC_VEGF	160/ 2,182	1.9	4.9-10	10/ 354	0.6	0.1	2/46	0.9	1
PMSC_MIX	91/ 1,836	0.9	0.6	7/ 101	1.3	0.4	6/ 71	1.7	0.3
SAOS2_OST	214/ 2,330	1.8	1.5e-11	38/ 486	1.3	0.1	17/ 235	1.2	0.4
Dataset	**Linear**			**No decision**			**Total tested**		
	# TF CAGE TSSs/ total CAGE TSSs	**OR**	**p**	**# TF CAGE TSSs/ total CAGE TSSs**	**OR**	**p**	**# TF CAGE TSSs/ tot CAGE TSSs**		
PMDM_LPS	9/ 582	0.3	2.9e-6	368/ 8,382	0.6	7.2e-9	764/14,376		
PAC_FGF2	0/ 7	2.4	0.4	311/ 7,601	0.8	0.03	492/11,235		
PAC_IL1B	1/ 10	4.9	0.2	290/ 6,666	0.9	0.5	468/10,513		
MCF7_HRG	8/ 233	0.5	0.06	415/ 7,559	0.6	6.5e-9	745/11,513		
MCF7_EGF1	13/ 125	1.9	0.05	438/ 7,370	1	0.8	679/11,352		
PEC_VEGF	0/ 32	0	1	337/8,072	0.7	3.9e-5	509/10,686		
PMSC_MIX	5/ 80	1.2	0.6	622/ 11,793	1	1	731/13,881		
SAOS2_OST	3/49	1	1	426/ 8,421	0.6	9.6e-13	698/11,521		

Next I analysed the enrichment of known IEG's CAGE TSSs in sets of genes that peak in one or more datasets, in two or more datasets, etc. Genes classified to the peak model in multiple datasets generally show significant enrichment for known IEGs (Table 10), with the robust set characterized by the strongest enrichment (Odds Ratio= 12.6; p-value < 2.2e-16) and therefore expected to contain fewer false positives.

Table 10 Enrichment of known IEGs for genes classified to the peak model in multiple datasets. Enrichment (expressed as odds ratios) and p values for genes classified across different numbers of time series datasets, computed using Fisher's exact test. The enrichment in the different subsets (genes shared by 2 to 8 datasets up to genes shared by all 8 datasets) is computed using the whole set of 8,785 genes as background. Furthermore, the table lists the median number of CAGE clusters belonging to known IEGs peaking in each subset.

Shared datasets	IEGs enrichment						# Known IEG CAGE clusters (median)
	# genes	# IEGs	# clusters (across 8 datasets)	# IEG clusters (across 8 datasets)	Odds Ratio	p-value	
1 to 8 (all peaking genes)	8,785	204	102,496	913	-	-	1
2 to 8	5,270	171	71, 384	853	6.3	2.2e-16	1
3 to 8	2,882	128	45,360	751	5.9	2.2e-16	2
4 to 8	1,304	86	24,616	590	5.9	2.2e-16	2
5 to 8	507	56	11,528	433	7.4	2.2e-16	3
6 to 8	182	35	4,896	299	10.3	2.2e-16	3
7 to 8	42	13	1,376	124	12.6	2.2e-16	4
8	5	2	264	18	8.3	4.6e-11	5

The robust set includes 13 known IEGs (*FOS, FOSB, FOSL1, JUN, KLF6, SGK1, PPP1R15A, BHLHE40, DUSP1, PHLDA1, NAB2, SDC4* and *EHD1*) and 29 candidate IEGs (*CUL3, ITM2B, PTGES3, SCL3A2, TMBIM6, SEC11A, SEC22B, SMARCA5, UBE2D3*, *OCIAD1, DKC1, MATR3, SRSF11, B4GALT1, FLNA, PLK1S1, NME2, RPSA, GNB2L1, PFKFB3, THBS1, PTP4A1, GNAS, PLEC, ARF4, ATG12, TMEM185B, HNRNPL, XBP1*).

4.4. Assessing bias in the robust set selection approach

I defined the robust set to be a shortlist of 13 known IEGs and 29 protein coding genes that are most promising candidate IEGs on the basis of the IEGs enrichment analysis described in Chapter 2.12. A limitation of the approach used to define the robust set is that I assumed independence for the CAGE TSSs associated with each gene. In reality, different genes are characterized by a different number of CAGE TSSs. Genes with higher number of TSSs are more likely to be classified as possessing a TSS with some characteristic, such as peaking expression pattern. Known IEGs and the robust set do indeed have a higher median number of TSSs per gene than all genes covered by the time series data, as documented in Chapter 6.2.

To assess the potential of TSS number to inflating the detection of IEGs by our protocol, I repeated the analysis strictly using only the canonical p1 TSS, the CAGE TSS with the highest number of observations annotated by FANTOM5 (Lizio, Harshbarger et al. 2017, Noguchi, Arakawa et al. 2017), for each gene (Table 11).

Table 11 Enrichment of known IEGs for single canonical p1 CAGE TSS classified to the peak model in multiple datasets. Enrichment (expressed as odds ratios) and p values for canonical p1 CAGE TSSs classified to the peak model across different numbers of time series datasets, computed using Fisher's exact test. The enrichment in the different subsets (genes shared by 1 to 8 datasets) is computed using the whole set of 11,744 p1 CAGE TSSs as background.

Peaking in at least	Total	Known IEGs	Candidate IEGs	p-value	Odds Ratio
8 datasets	2	2	0	3.20E-04	Inf
7 datasets	10	10	0	2.80E-18	Inf
6 datasets	46	27	19	4.20E-36	88.60
5 datasets	177	46	131	3.30E-41	24.20
4 datasets	654	76	578	3.20E-42	10.70
3 datasets	1904	115	1789	6.50E-38	6.50
2 datasets	4277	158	4119	1.90E-30	5.40
1 dataset	7936	198	7738	9.90E-21	7.50
0 datasets	3808	13	3795	1.00E+00	0.10

Comparing the 654 p1 TSSs peaking in at least 4 datasets with the 42 genes in the robust set (all TSSs for each gene peaking in at least 7 out of 8 datasets) I observe that 29 genes (13 known and 16 candidate IEGs) have the same p1 TSS peaking in at least 4 datasets.

Table 12 shows that all the 13 known IEGs in the robust set are characterized by the p1 TSS peaking in at least 4 datasets. Thus, I believe that the success of my protocol is not simply a result of the number of TSSs possessed by the input genes. Instead, the enrichment of known IEGs in the robust set appears to be a result of the accurate classification of expression profiles over time. However, the FANTOM5 p1 collection of canonical TSSs highlights another problem: since one single p1 TSS was defined across all FANTOM5 libraries, inevitably some genes in our smaller collection of time series datasets (which constitute a minority of FANTOM5 libraries) do not express their p1 promoters

at a detectable level in these time series. By choosing a restricted set of TSSs (such as the p1 set) I therefore exclude many (non-p1) TSSs from my analysis, and restrict the number of new candidate IEGs I can discover. Similarly, if I randomly select a TSS for each gene I do not interrogate the complete dataset, and I limit my ability to discover new candidate IEGs.

Table 12 Robust set p1 CAGE TSS. The table lists the 29 p1 CAGE TSSs peaking in at least four datasets and associated to genes in the robust set.

P1 CAGE TSS	Gene Ensid	Gene name	IEG	Shared datasets
chr1:59249707..59249727,-	ENSG00000177606	JUN	Known	6
chr10:3827389..3827408,-	ENSG00000067082	KLF6	Known	7
chr10:6244887..6244904,+	ENSG00000170525	PFKFB3	Candidate	4
chr11:64646086..64646101,-	ENSG00000110047	EHD1	Known	7
chr11:65667846..65667868,-	ENSG00000175592	FOSL1	Known	7
chr12:57082060..57082083,-	ENSG00000110958	PTGES3	Candidate	4
chr12:57482882..57482907,+	ENSG00000166886	NAB2	Known	6
chr12:76425368..76425384,-	ENSG00000139289	PHLDA1	Known	7
chr13:48807334..48807354,+	ENSG00000136156	ITM2B	Candidate	4
chr14:75745523..75745537,+	ENSG00000170345	FOS	Known	8
chr19:39340563..39340634,-	ENSG00000104824	HNRNPL	Candidate	4
chr19:45971246..45971265,+	ENSG00000125740	FOSB	Known	8
chr19:49375669..49375684,+	ENSG00000087074	PPP1R15A	Known	7
chr2:120980939..120980983,-	ENSG00000226479	TMEM185B	Candidate	6
chr2:225450013..225450068,-	ENSG00000036257	CUL3	Candidate	4
chr20:43977055..43977073,-	ENSG00000124145	SDC4	Known	7
chr22:29196511..29196541,-	ENSG00000100219	XBP1	Candidate	4
chr3:39448201..39448222,+	ENSG00000168028	RPSA	Candidate	6
chr3:5021113..5021180,+	ENSG00000134107	BHLHE40	Known	6
chr3:57583064..57583079,-	ENSG00000168374	ARF4	Candidate	4
chr4:103748696..103748723,-	ENSG00000109332	UBE2D3	Candidate	4
chr4:48833070..48833118,+	ENSG00000109180	OCIAD1	Candidate	4
chr5:115177247..115177279,-	ENSG00000145782	ATG12	Candidate	5
chr5:138629417..138629446,+	ENSG00000015479	MATR3	Candidate	4
chr5:172198190..172198206,-	ENSG00000120129	DUSP1	Known	7
chr6:134495992..134496010,-	ENSG00000118515	SGK1	Known	7
chr8:145013711..145013786,-	ENSG00000178209	PLEC	Candidate	4
chr9:33167149..33167170,-	ENSG00000086062	B4GALT1	Candidate	4
chrX:153991088..153991168,+	ENSG00000130826	DKC1	Candidate	4

4.5. GO term enrichment for peaking genes in each dataset is consistent with the function of known IEGs

Genes possessing TSSs assigned to the peak class showed enrichments for gene ontology (GO) processes associated with transcription, cell activation, cell proliferation, cell differentiation and cancer-related terms such as cell death and apoptosis (FDR = 0.05). These terms were also consistent with previous studies of IEGs (Bahrami and Drabløs 2016). Genes peaking across the eight dataset showed enrichment for 285 GO terms, over 30% (88, Appendix Table 2) of which were shared with the list of 773 GO terms of all known IEGs (Figure 18 and Figure 19, Appendix Table 3). In contrast, the genes classified to other models were not enriched for any GO terms.

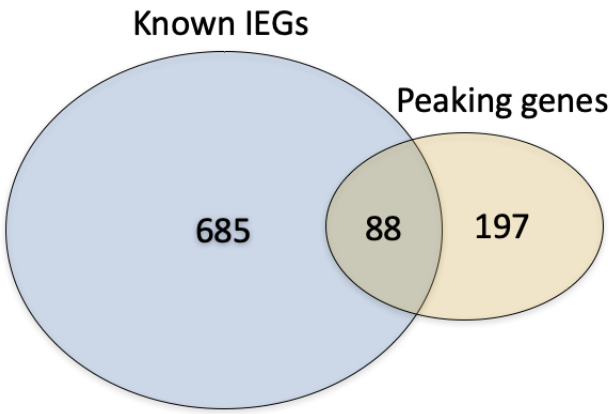


Figure 18 GO terms shared between peaking genes and known IEGs. GO terms significantly enriched (FDR corrected hypergeometric test q-value <0.05) in all the 8785 genes classified as peak across the eight datasets and in the 212 known IEGs.

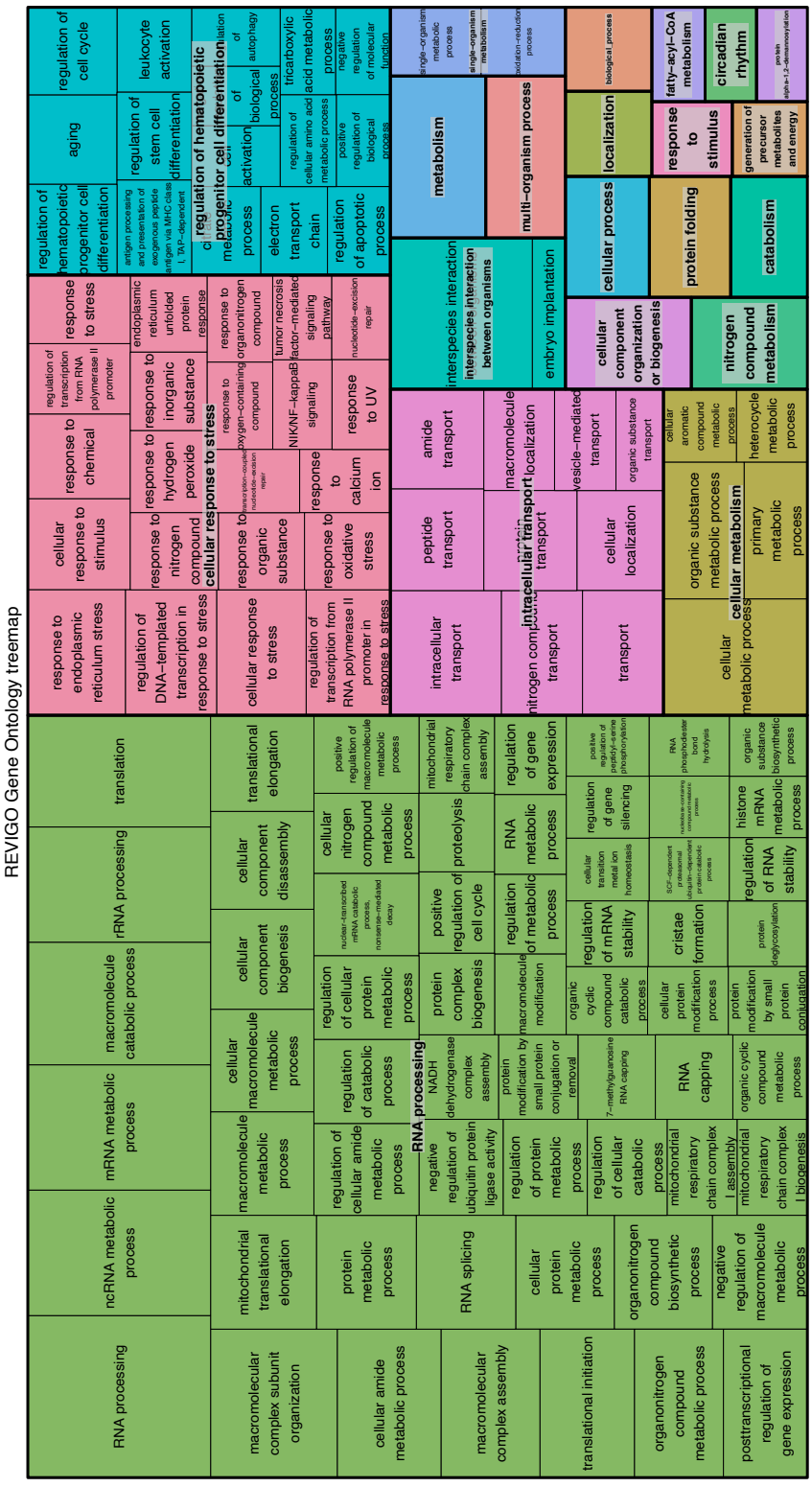


Figure 19 GO term enrichment analysis. GO terms significantly enriched (FDR corrected hypergeometric test q-value <0.05) in all the genes classified as peak and in known IEGs. Each rectangle represents a cluster of terms with similar and related description, and clusters of related terms are grouped together in 'superclusters' to reduce redundancy. The size of the rectangles reflects the frequency of the terms in the resulting output (the enrichment.

4.6. Peak expression times are often similar between datasets

This model fitting approach provided parameter estimates for all promoters assigned to any given model, providing a straightforward and intuitive basis for meta-analysis. For example, comparison of the peak times (t_p) for all protein coding promoters classified as peaks in at least four datasets (the permissive set) readily demonstrated common patterns across datasets (Figure 20). Waves of promoter activation were evident, with certain promoters, particularly in known IEGs, activated in the same early time window in multiple datasets. For this analysis I decided to use only CAGE TSSs peaking in at least 4 datasets in the first 3 hours, which is the length of the shortest time series, PMSC_MIX, thus removing tissue specific CAGE TSSs and CAGE TSSs with late dynamics which could be detected only in the longest time courses. Hierarchical clustering of the datasets based on the t_p of permissive set's promoters also recapitulated known relationships between cell types and stimuli (Figure 20). The two datasets derived from the MCF7 breast cancer cell line, stimulated with different ligands of the same ErbB receptor family (EGF1 and HRG) clustered together. Similarly, the two primary aortic cell samples exposed to a growth factor or activated by a pro-inflammatory cytokine (PAC_FGF2 and PAC_IL1B, respectively) also clustered together. Thus, similarities in promoter activation dynamics (reflected in t_p parameter estimates) between datasets may reflect commonalities in their underlying biology.

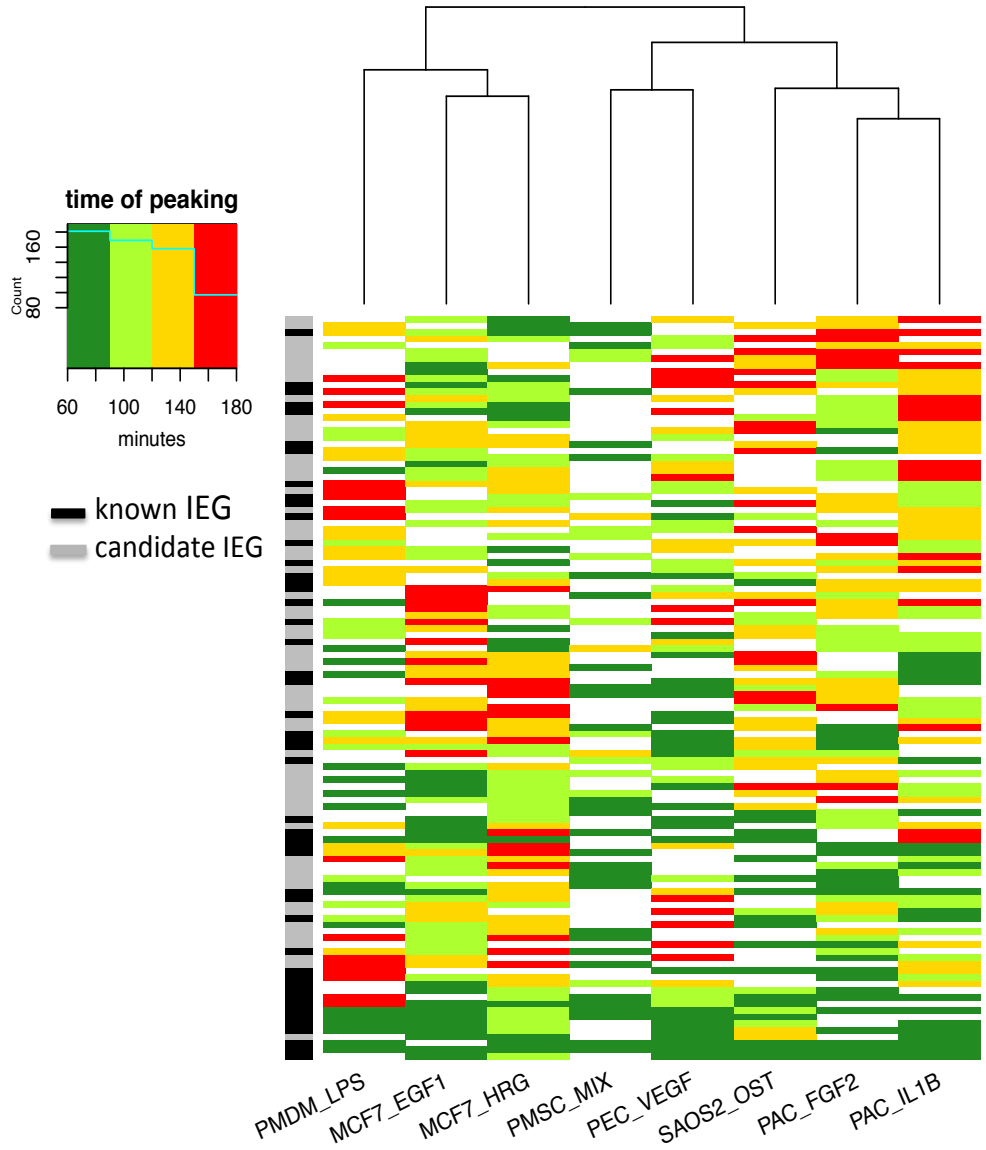


Figure 20 Broad trends in peak expression times across datasets. The time of peak TSS expression (tp) for CAGE TSS from the permissive set of peak classified genes for all datasets. Heatmap colours reflect the tp for each CAGE TSS, from dark green within 90 min, to red for peak expression up to 180 min after stimuli. In the left column black cells indicate TSSs of known IEGs. The dendrogram reflects the similarity among datasets based upon peak expression times for all TSSs in the permissive set.

4.7. Known IEGs and candidate IEGs participate in common signalling pathways

Having shown that the peak model described the behaviour of known IEGs, I speculated that the other genes assigned to this model might include novel candidate IEGs. Of the 42 genes in the robust set, more than two thirds (29 genes) are not known to be IEGs and can, therefore, be considered to be candidate novel IEGs (henceforth candidate IEGs). Pathway analysis (Breuer, Foroushani et al. 2012) recovers many known relationships among known IEGs, as expected, centred on heavily studied IEGs such as FOS and JUN. However, the same analysis suggests that more than half (ATG12; UBE2D3; THBS1; FLNA; GNB2L1; ITM2B; B4GALT1; GNAS; ARF4; PLEC; PTGES3; SLC3A2; XBP1; RPSA; PFKFB3; CUL3 and DKC1) of candidate IEGs also participate in common pathways with known IEGs, involving a densely inter-connected network of 83 significantly over-represented pathways (Table 13), including signalling cascades known to mediate the IER, such as the Ca2+-dependent pathways and the mitogen-activated protein (MAP) kinase network (Treisman 1996, Schratt, Weinhold et al. 2001).

For example, the candidate IEG UBE2D3 participates with the known IEGs PPP1R15A, FOS and JUN in the Transforming growth factor-β (TGFβ) signalling pathway, which is known to drive many biological processes, such as cell proliferation, differentiation and morphogenesis in animal cells (Massagué 2012). A chromatin immunoprecipitation assay also reported physical association between JUN and PPP1R15A, suggesting that induction of PPP1R15A is mediated by JNK/JUN pathway as documented by Xu et al. (Xu, Xiao et al. 2015). UBE2D3 also participates with JUN and FOS in the TRIF-mediated TLR signalling pathway, which has a fundamental role in the induction of the innate immune response (Ahmed, Maratha et al. 2013), and with ATG12, a candidate IEG involved in autophagy, to negatively regulate the RIG-I and MDA5 innate immune receptors. Similarly, the known IEG *SDC4*, which encodes a member of the Syndecans transmembrane receptors family, is linked to the candidate *THBS1* gene, which encodes an adhesive

glycoprotein, both participating in signalling events in cell proliferation and cell–matrix and cell–cell adhesion (Carey 1997, Cheng, Montmasson et al. 2016).

Table 13 Biological pathways overrepresentation. Pathway analysis of the 42 protein coding genes in the robust set, obtained using InnateDB database. Known IEGs are indicated in bold.

Genes	Overrepresented pathways (P-value corrected <0.05)
FOS; JUN	Tsp-1 induced apoptosis in microvascular endothelial cell Pertussis toxin-insensitive ccr5 signaling in macrophage Activation of the AP-1 family of transcription factors S1P2 pathway Calcium signaling by hbx of hepatitis b virus Repression of pain sensation by the transcriptional regulator dream Cadmium induces dna synthesis and proliferation in macrophages Nerve growth factor pathway (Lerner, Harada et al.) Mets affect on macrophage differentiation Oxidative stress induced gene expression via nrf2 Igf-1 signaling pathway PDGFR-alpha signaling pathway Inhibition of cellular proliferation by gleevec Tpo signaling pathway IL12 signaling mediated by STAT4 Pdgf signaling pathway Endothelins ErbB2/ErbB3 signaling events Osteopontin-mediated events Fc epsilon receptor i signaling in mast cells Bcr signaling pathway RhoA signaling pathway Role of egf receptor transactivation by gpcrs in cardiac hypertrophy MAPK targets/ Nuclear events mediated by MAP kinases Signal transduction through il1r Toll-like receptor pathway Angiotensin ii mediated activation of jnk pathway via pyk2 dependent signaling IL2-mediated signaling events IL6-mediated signaling events LPA receptor mediated events Presenilin action in Notch and Wnt signaling FCERI mediated MAPK activation Keratinocyte differentiation T cell receptor signaling pathway Mapkinase signaling pathway MAP kinase activation in TLR cascade BCR signaling pathway Colorectal cancer Regulation of nuclear SMAD2/3 signaling Leishmaniasis B cell receptor signaling pathway
FOS; JUN; DUSP1	Fc-epsilon receptor I signaling in mast cells
FOS; JUN; FOSL1	Calcium signaling in the CD4+ TCR pathway Downstream signaling in naive CD8+ T cells
FOS; JUN; FOSB; FOSL1	CD4 T cell receptor signaling Osteoclast differentiation

FOS; JUN; FOSB	BCR
FOS; JUN; SGK1	Glucocorticoid receptor regulatory network IL6
DUSP1; JUN	Mechanism of gene regulation by peroxisome proliferators via ppara
FOSL1; JUN	Validated transcriptional targets of AP1 family members Fra1 and Fra2
FOS; FOSL1	Bone remodeling
FOS; JUN; *PTGES3; DKC1*	Regulation of telomerase
FOS; JUN; DUSP1; *CUL3*	ATF-2 transcription factor network
FOS; JUN; DUSP1; *ARF4*	ErbB1 downstream signalling
FOS; JUN; DUSP1; *FLNA*	MAPK signaling pathway
FOS; JUN; BHLHE40; *PFKFB3*	HIF-1-alpha transcription factor network
FOS; JUN; *PLEC; RPSA*	Alpha6Beta4Integrin
FOS; JUN; *XBP1*	FOXA1 transcription factor network
FOS; JUN; FOSL1; *SLC3A2*	Calcineurin-regulated NFAT-dependent transcription in lymphocytes
FOS; JUN; SDC4*;*	FGF signaling pathway
JUN; FOS; *FLNA*	Prolactin
FOS; JUN; *UBE2D3*	MyD88-independent cascade TRIF-mediated TLR3/TLR4 signaling Toll Like Receptor 3 (TLR3) Cascade Activated TLR4 signaling Toll Like Receptor 4 (TLR4) Cascade Toll-Like Receptors Cascades
FOS; JUN; FOSB; PPP1R15A*; UBE2D3*	TGF_beta_Receptor
FOS; JUN; *PTGES3; UBE2D3*	Cellular responses to stress
FOS; JUN; SDC4; DUSP1; *ARF4; PLEC*	EGFR1
FOS; JUN; *GNAS*	Chagas disease (American trypanosomiasis)
JUN; *B4GALT1*	Pre-NOTCH Expression and Processing
JUN; *GNB2L1*	Regulation of Androgen receptor activity IL5
JUN; SGK1; *ITM2B*	IL2
JUN; *FLNA; GNB2L1; UBE2D3*	TNFalpha
SDC4*; THBS1*	Syndecan interactions Syndecan-4-mediated signaling events Beta3 integrin cell surface interactions Non-integrin membrane-ECM interactions
ATG12; UBE2D3	Negative regulators of RIG-I/MDA5 signaling

4.8. Novel non-coding RNA candidates in the immediate early response

I next classified the promoters of non-coding transcripts and found peak classified promoters driving the expression of 20 non-coding RNA genes across at least seven datasets, constituting the robust set of non-coding RNA

candidate IEGs. These included genes associated with the cellular splicing machinery, such as small nuclear RNA multi-gene families, which are part of the spliceosome, SCARNA17, a small nuclear RNA which contributes to the post transcriptional modification of many snRNPs, and SNORD65 and SNORD82, snRNAs involved in rRNA modification and alternative splicing. Kalam et al. (2017) have shown that macrophage infection with *Mycobacterium tuberculosis* results in the systematic perturbation in splicing patterns (Kalam, Fontana et al. 2017), and these results appear to suggest more general roles for alternative splicing in the IER. However, repetitive multigene families, such as these small nuclear RNAs (U1, U2, U3, U4 and 7SK) present particular challenges for reliable sequence read mapping (Chapter 2.8). Although probabilistic approaches to mapping ambiguously mapped reads were developed and applied in FANTOM5 (Arner, Daub et al. 2015), I chose to conservatively remove these genes from the robust set (Chapter 2.9), leaving a group of 15 noncoding genes (Table 14).

Table 14 Noncoding RNA genes peaking in at least 7 out of 8 datasets. The number of datasets in which the genes are classified to the peak model is provided with a short description of the molecular function attributed by the literature (via GeneCards database).

Gene ID	Nº Datasets	Description (PubMed ref.)
LINC00478 (MIR99AHG)	7	Has a role in cell proliferation and differentiation and considered a regulator of oncogenes in leukaemia (PMID: 25027842)
LINC00263	7	Regulation of oligodendrocyte maturation (PMID: 25575711)
LINC-PINT	8	Putative tumour suppressor (PMID: 24070194)
LINC00963	7	Involved in the prostate cancer transition from androgen-dependent to androgen-independent and metastasis via the EGFR signalling pathway (PMID: 24691949)
LINC00476	8	Non characterized lincRNA
LINC00674	7	Non characterized lincRNA
STX18-AS1	7	Non characterized lincRNA
DLEU2	7	Critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1 (PMID:19591824)
MiR-29A	7	The expression of the MiR-29 family has antifibrotic effects in heart, kidney, and other organs. The family have also been shown to induce apoptosis and regulate cell differentiation (PMID: 22214600)
MiR-3654	7	***Involved in Prostate Cancer progression (PMID: 27297584)***
MiR-21	7	Oncogenic potential (PMID: 18548003)
AL928646	7	Non characterized ncRNA
SCARNA17	7	scaRNA Involved in the maturation of other RNA molecules (PMID: 12032087)
SNORD65	7	Belongs to the Small nucleolar RNAs, C/D bof family. Involved in rRNA modification and alternative splicing (PMID: 26957605)
SNORD82	7	Belongs to the Small nucleolar RNAs, C/D bof family. Involved in rRNA modification and alternative splicing (PMID: 26957605)

Three miRNAs are present in the robust set (Table 14). MiR-21 has been demonstrated to have oncogenic potential by inhibiting the expression of phosphatases, limiting the activity of signalling pathways such as AKT and MAPK, which are involved in regulating cellular proliferation, differentiation and survival. This miRNA was previously reported to show IEG-like behaviour in the PAC_FGF2, PAC_FGF2 and MCF7_HRG time series (Aitken, Magi et al. 2015). Here, I found similar behaviour for the same gene in the MCF7_EGF1, PEC_VEGF, PMSC_MIX and SAOS2_OST datasets. This extends previous studies reporting that the MiR-21 mature transcript is upregulated on EGF treatment in MCF10A (Avraham, Sas-Chen et al. 2010) and HeLa (Llorens, Hummel et al. 2013) cells. MiR-29A has been associated with the viability and proliferation of mesenchymal stem cell and gastric cancer cells (Liu, Cai et al.

2015, Zhang and Zhou 2015). MiR-3654, is reported to be involved in prostate cancer progression (Saravanan, IH Islam et al. 2016) and in the immune response in Myasthenia gravis patients (Barzago, Lum et al. 2016). DLEU2 is a putative tumour suppressor gene that hosts two miRNAs, MiR-15A and MiR-16-1 which are known to inhibit cell proliferation and the colony-forming ability of tumour cell lines, and to induce apoptosis (Cimmino, Calin et al. 2005, Lerner, Harada et al. 2009, Gao, Xing et al. 2011). Seven long non-coding RNAs (lncRNAs) also appear in the robust set (Table 14) and among them LINC00478 is particularly interesting, as it has already been reported to show IEG-like behaviour (Aitken, Magi et al. 2015). It is implicated in breast cancer and hosts an intronic cluster of miRNAs comprising let-7c, MiR-99A, and MiR-125B (Gökmen-Polar, Zavodszky et al. 2016). Although poorly characterised, LINC00263, LINC-PINT and LINC00963 are thought to be involved in biological processes often associated with IEGs induction, such as cell maturation, cell proliferation and the expression of growth factor receptors (Marín-Béjar, Marchese et al. 2013, Wang, Han et al. 2014, Mills, Kavanagh et al. 2015, Müller, Raulefs et al. 2015). In addition, I found four functionally uncharacterized non-coding RNAs (LINC00476, LINC00674, STX18-AS1 and AL928646) showing peaking behaviour in at least seven out of eight datasets.

4.9. Exploration of the set of genes modelled by dip, decay and linear functions

The classification of transcriptional responses across datasets shows that the IER is often associated with promoter dynamics that are clearly different from the behaviours of IEGs, such as those classified to the dip, decay and linear models (Table 15). However, these classifications are less consistent across datasets than the peak model: 15% of the peaking genes show the same kinetics in more than four out of eight datasets, while only 0.5% 0.01% and no genes at all share dip, decay and linear model, respectively, in more than four datasets. Nevertheless, when I analysed the GO terms enrichment, I found that a big proportion of the terms associated to dip and decay sets (which includes genes which are also classified as peaks using different CAGE TSSs)

is shared with the group of 212 known IEGs. A possible explanation could be that the IER results not only in the quick and transient activation of the IEGs but also in the down regulation of other classes of genes which could be involved in the same processes.

Table 15 Functional comparisons of different models. Classification and GO term enrichment (FDR corrected hypergeometric test q-value <0.05) for the genes assigned to different mathematical models.

Model	# genes	# genes shared by at least 4 datasets	# IEGs	# GO terms	Proportion GO terms shared with known IEGs
Peak	8,785	1,304 (15%)	204	285	30%
Dip	3,093	16 (0.05%)	36	81	53%
Decay	1,769	1 (0.01%)	36	47	66%
Linear	1,025	0 (0%)	9	52	3%

4.10. Assessing limitations of gene set enrichment analysis

Annotation enrichment procedures are subjected to a number of potentially limiting factors and artefacts which should be considered. For example, the choice of the wrong background could result in false positives, wrongly considering significant genes which are not significant (Khatri and Drăghici 2005). This bias is caused by the fact that all the genes which are in the background but are not present in the pool of genes available for comparison won't have the opportunity of being in the list of genes tested for enrichment. For this reason, I used only the genes detected across the eight time-course datasets instead of the whole human genes list.

Another limitation is caused by the heterogeneity in the quality of annotation for the genes in the list respect to the genes in the background. Some genes are very poorly annotated and can never be detected as significantly enriched for anything, while others are more thoroughly annotated and better described (Khatri and Drăghici 2005). Unfortunately, this is a limitation common to all the studies involving gene set enrichment analysis and associated bias can't be easily avoided.

4.11. Discussion

Here I collate and describe an unusual collection of FANTOM5 CAGE times series datasets capturing the IER genome-wide to various stimuli at the level of individual promoters. A statistically rigorous classification of individual promoters across these datasets has not been performed until now. Refining an existing method (Aitken, Magi et al. 2015), I classify time course gene expression profiles to one of several predefined, mathematical models of time course dynamics: peak, dip, decay and linear. The peak model definition reflects current knowledge of the immediate early response, which is characterized by the transient and rapid (i.e. protein synthesis-independent) activation of the IEGs, which are known to be involved in many fundamental cellular processes such as cell proliferation, differentiation, apoptosis and survival. Many IEGs encode secreted proteins and TFs and exert their cellular activities triggering the activation of secondary response genes, involved themselves in complex and tightly regulated signalling pathways (Tullai, Schaffer et al. 2007, Bahrami and Drabløs 2016).

I found that the peak model had the highest number of assignments in all the datasets for both protein-coding and non-coding RNA genes (72% and 64%, respectively), of the total genes detected by CAGE sequencing in the eight datasets. A plausible explanation for this higher proportion could be that the algorithm has been optimized to most effectively discover peaking genes, since I focus on IEGs, and therefore many genes with other kinetics of interest could be undiscovered in the unclassified set.

I describe a permissive set of 1,304 genes peaking in at least four out of eight datasets. The majority of the promoters assigned to the 84 known IEGs in this set, peak early in most datasets. However, comparing the time of peaking for the permissive set across the eight datasets I observe a higher similarity between the promoter dynamics of analogous biological systems, such as PAC_IL1B and PAC_FGF2, which may suggest system specific regulatory mechanisms of the IER, as previously suggested by Aitken et al (Aitken, Magi et al. 2015). In addition, I define a robust set of 42 protein-coding genes

peaking in at least seven out of eight datasets. This set contains 13 previously known IEGs, and 29 candidate IEGs, which participate in common signalling pathways and are likely to be core components of the IER. I also performed a more stringent analysis using only the single p1 CAGE TSSs and I still detected significant enrichments of known IEGs present in the robust set, obtained using the original protocol. This result addresses the problem of multiple CAGE TSSs associated to the set of genes compared in the enrichment analysis and supports the robust set.

Applying my approach to the CAGE TSSs of non-coding RNAs, I also discovered a set of 15 non-coding RNAs peaking across at least seven datasets, comprising miRNAs and lncRNAs, suggesting regulatory roles for particular non-coding RNA species in the IER (Aitken, Magi et al. 2015).
Many known IEGs are not present in robust set of shared peaking genes, peaking exclusively in one or few datasets. This variability may reflect the ability of different biological system to react in a stimulus-specific manner, integrating unique mechanisms. It follows that a number of undiscovered IEGs could be present in the set of genes peaking in less than four datasets. Future experimental analysis could exploit the protein synthesis-independent activation nature of the IEGs to support the candidate IEGs in the robust and permissive set.

5. Temporal patterns of gene activation are conserved across datasets

5.1. Introduction

The interaction of biological molecules in complex networks is the basis of all biological processes, including cellular responses to stimuli, and modeling these interactions can reveal the mechanisms underlying fundamental biological processes (Price and Shmulevich 2007).
Signalling molecules promote specific temporal patterns of signalling-transduction and gene expression, triggering the execution of generalized and specific biological responses (Perrimon, Pitsouli et al. 2012). Many biological responses, such as immune-system response or cellular stress, need a prompt and specific activation of the regulatory systems to successfully accomplish the appropriate function (Bahrami and Drabløs 2016).

Fully understanding the normal behaviour of a living cell or the responses of cells to changes in external conditions requires not only the characterization of the genes involved in the transduction between the input and the specific output or the identification of the transcription factors and the effectors of the response, but also the knowledge of the magnitude and timing of the activation of each gene involved (Murphy and Blenis 2006).
The majority of studies of regulatory networks have been focused on static data, mostly due to the shortage of appropriate high resolution time-series expression data and the challenges in the analysis of such data. Time-series datasets can be massive and high-dimensional, so that incorporating them into a coherent network can be computationally intensive (Price and Shmulevich 2007, Sima, Hua et al. 2009, Streit, Tambalo et al. 2013, Thorne 2018).
IEGs are thought to represent the core of the complex and finely orchestrated systems responding to external stimuli, and often use common activation systems across signalling pathways and transcription factors (Saito, Uda et al.

2013). However, the temporal regulatory network of the IER is still poorly characterized and will be analysed in this chapter. The expression of the genes encoding FOS and JUN proteins are surely the most studied among the IEGs and also the first to be temporally characterized. Their time course expression has been described in many systems in different studies, aiming to disentangle the basis of different biological processes, such as the activation of FOS by the growth factors in quiescent 3T3 cell (Greenberg and Ziff 1984), the inhibition of androgen-induced PSA promoter activity (Sato, Sadar et al. 1997) by FOS and JUN induction. All these studies reported the fast and transient kinetics of FOS and JUN activation.

The activation dynamics of the IEGs are difficult to generalize because many regulatory patterns are cell, system or biological process specific (O'Donnell, Odrowaz et al. 2012). Aitken et al. (2015) compared the time of peaking of several IEGs in four cell systems: aortic smooth muscle cells treated with IL1B and FGF2 and the MCF7 cell line treated with EFG1 and HRG (Aitken, Magi et al. 2015), but did not examine the conservation of the order of activation of IEGs. Here, I provide a systematic statistical analysis of the temporal order of gene expression, and in particular IEG expression, to identify conserved patterns across cell types and stimuli.

5.2. Distribution of peaking time and change in expression parameters

The dynamics of the expression of peak-classified genes can be visualized by a scatterplot of expression fold change against peak expression time (t_p).

Figure 21 plots the fold change expression (in tag per millions (TPM)) versus t_p for the PMDM_LPS (Figure 21A) and MCF7_EGF1 datasets (Figure 21B) in an interval up to 300 minutes, for a subset of known IEGs as well as candidate protein coding and non-coding IEGs from the robust set. The directional arrows illustrate the temporal ordering of gene activation. *FOS* stands out with its early and high expression, preceding the majority of the other genes in both datasets. This is consistent with its role as principal

regulator in the IER. Together with JUN, which follows FOS in timing in our analysis (Figure 21 and Figure 24), they are subunits of the AP1 transcription factor, which regulates the expression of many other genes involved in physiological and pathological cellular processes, in response to a variety of different stimuli, such as cytokines and growth factors (Hess, Angel et al. 2004). Among the candidate IEGs in the robust set, *XBP1* is especially noteworthy. This gene encodes a transcription factor and is relatively short in length (6Kb compared with the mean of 58Kb for all Ensembl protein coding genes) consistent with the IEG archetype (Fowler, Sen et al. 2011). It is activated by cellular stress processes affecting the normal functions of the Endoplasmic Reticulum (Xu, Bailly-Maitre et al. 2005). The TF XBP1 gene, together with miR-3654 and LINC00476 non-coding RNAs, appear to be consistently upregulated after FOS and before the activation of another known IEG: EHD1, which is known to have a role in the endocytosis and recycling of various receptors (Naslavsky, Rahajeng et al. 2006). XBP1 is known to form a heterodimer with FOS which regulate the human class II major histocompatibility complex genes (Ono, Liou et al. 1991), and Martinez et al. found a correlation between *XBP1* and *FOS* upregulation in the human hippocampus related to memory formation (Martínez, Vidal et al. 2016); however, their interaction in other biological systems is less studied. The functions of miR-3654 and LINC00476 non-coding RNAs in the IER are unstudied and their association with *FOS* or *EHD1* expression has not been mentioned before. However, other miRNAs and lincRNAs are known to be involved in the regulation of protein coding genes. Thus the conserved timing of these non-coding RNAs could reflect possible roles in the regulation of the IER. The temporal ordering of known and candidate IEGs will be analysed in more detail below using a formal definition of conserved temporal ordering.

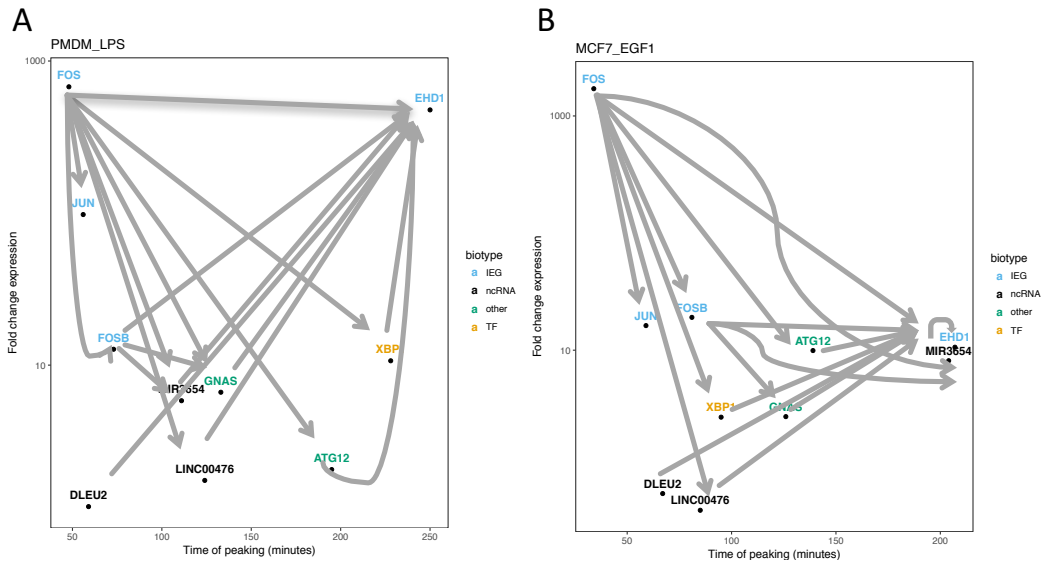


Figure 21 Transcriptional dynamics of genes classified to the peak model. Scatterplots of log fold change against the time of peaking for selected genes of interest, with conserved temporal ordering indicated by arrows for (a) PMDM_LPS and (b) MCF7_EGF1. *FOS* peaks earliest and has many conserved temporal relations to later peaking genes, while *EHD1* peaks late and has many conserved temporal orderings with earlier peaking genes.

When examining the magnitude of expression changes across time course datasets (Table 16), known IEGs CAGE TSSs tend to show the greatest fold changes (Figure 22A, all datasets merged, Wilcoxon p-value < 2.2e-16, difference in sample medians = 1.85, confidence interval calculated with Wilcoxon test = (1.72 , 1.99)). However, candidate protein coding IEGs promoters do not show a notable difference in timing (Figure 22C, all datasets merged, Wilcoxon p-value = 0.89, difference in sample medians= -0.26, confidence interval calculated with Wilcoxon test = (-4.41 , 3.9)). There is no evidence of a statistical difference between population medians, which is reflected by the confidence interval spanning zero. Comparing candidate and known IEGs in single datasets (Table 16), the time of peaking is significantly earlier for known IEGs relative to the other candidate IEGs in four time-series: MCF7_EGF1, PEC_VEGF SAOS2_OST and PAC_FGF2, while is slightly earlier for PMDM_LPS and PAC_IL1B candidate IEGs respect to known IEGs. These results suggest that known IEGs show a higher magnitude of change in expression and may be easier to detect, explaining their widespread presence in literature.

Fold changes in peaking non-coding RNA promoters tend to be lower than for known IEGs (Figure 22B, Table 16) (all datasets merged Wilcoxon p-value <

2.2e-16, difference in sample medians= 4.55, confidence interval calculated with Wilcoxon test = (4.39 , 4.72)) but they occur earlier than known IEGs (Figure 22D, Table 16) in 3 datasets, PMDM_LPS, MCF7_HRG and PAC_IL1B and later in MCF7_EGF1 and PEC_VEGF (all datasets merged Wilcoxon p-value < 2.2e-16, difference in sample medians= 21.5, confidence interval calculated with Wilcoxon test = (16.8 , 26.3)). I speculate that many non-coding RNAs could be expressed in the earliest stages of the IER to regulate the specific IEGs in a stimulus-specific manner.

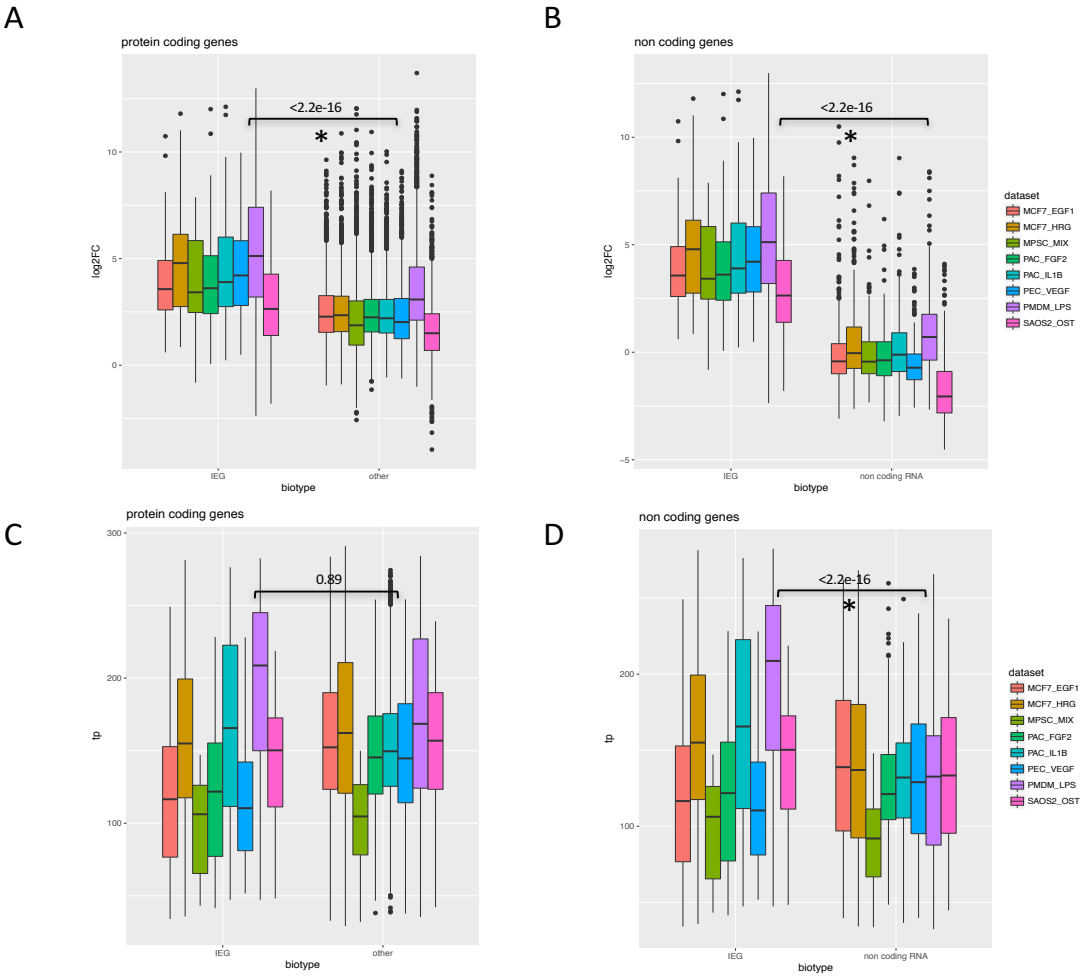


Figure 22 Distributions of known IEG expression change and t_p across datasets. (A) and (B) box plots show a higher log fold change between maximal and basal expression for known IEGs than for other protein coding and non-coding candidate IEGs, respectively. (C) and (D) boxplots indicate that the tie of peaking is comparable between known IEGs and protein-coding candidate IEGs but is significantly different between known IEGs and non-coding RNAs (significant difference after performing the independent variable t-test evaluation is indicated with an asterisk).

Table 16 Unpaired two-samples Wilcoxon test values of the comparisons between known and candidate IEGs for time of peaking and log fold change. For each dataset the log fold change in expression (in the left) and time of peaking (in the right) medians are compared, between known and candidate IEGs, with Wilcoxon Signed Rank Test. For both protein coding (top table) and non-coding RNAs (bottom table) genes, a confidence interval for the difference between two measures of location (sample medians) is provided. The null hypothesis can't be rejected when the confidence interval for the difference in sample medians spans zero.

	Protein coding genes					
	expression change (log fold change)			time of peaking		
	difference in sample medians	95% Confidence Interval	p-value	difference in sample medians	95% Confidence Interval	p-value
MCF7_EGF1	1.29	(0.92 , 1.68)	4.2E-11	-37.90	(-50 , -25)	1.4E-08
MCF7_HRG	2.16	(1.80 , 2.54)	4.2E-32	-5.66	(-16.48 , 4.98)	0.29
MPSC_MIX	1.89	(1.20 , 2.61)	8.8E-07	-1.63	(-13.19 , 8.81)	0.75
PAC_FGF2	1.26	(0.91 , 1.63)	2.1E-11	-28.00	(-37.8 , -18)	1.1E-07
PAC_IL1B	1.79	(1.44 , 2.16)	5.6E-24	13.40	(1.5 , 25.3)	2.8E-02
PEC_VEGF	2.03	(1.69 , 2.38)	1.3E-27	-33.60	(-41.6 , -25.4)	3.2E-14
PMDM_LPS	1.80	(1.47 , 2.09)	1.1E-29	23.00	(14.6 , 31.6)	3.5E-08
SAOS2_OST	1.07	(0.7 , 1.47)	4.3E-07	-12.90	(-23.97 , -1.93)	2.2E-02

	non-coding RNA					
	expression change (log fold change)			time of peaking		
	difference in sample medians	95% Confidence Interval	p-value	difference in sample medians	95% Confidence Interval	p-value
MCF7_EGF1	3.93	(3.54 , 4.34)	3.5E-32	-21.90	(-36.68 , -7.42)	3.3E-03
MCF7_HRG	4.29	(3.86 , 4.73)	2.6E-53	19.80	(7.76 , 31.75)	1.5E-03
MPSC_MIX	4.01	(3.38 , 4.72)	1.5E-15	9.06	(-2.59 , 21.84)	0.13
PAC_FGF2	3.94	(3.54 , 4.34)	2.8E-41	-5.68	(-16.95 , 5.63)	0.32
PAC_IL1B	4.13	(3.72 , 4.55)	2.0E-47	35.70	(22.1 , 49.1)	2.3E-07
PEC_VEGF	4.80	(4.42 , 5.21)	1.9E-55	-16.80	(-27.28 , -6.31)	2.1E-03
PMDM_LPS	4.27	(3.84 , 4.71)	2.7E-55	73.40	(61.6 , 84.4)	1.4E-26
SAOS2_OST	4.55	(4.1 , 4.97)	2.6E-35	9.94	(-2.69 , 22.82)	0.01

5.3. Discovery of a conserved IER activation network

Having established comparable patterns of peak gene induction at similar times across datasets for certain known IEGs (chapter 5.2), I hypothesised that many IEGs may be induced in a conserved order over time. To our knowledge, the extent of conserved ordering in gene induction is unstudied in general, and in the IER it is of particular interest for two main reasons. Firstly, the presence of conserved gene orderings, in addition to common gene classifications, may provide an additional indication of functional similarity between different cell types and ERI stimuli. Secondly, strongly conserved ordering may suggest the existence of conserved regulatory mechanisms governing the induction of these genes, motivating further studies of these underlying mechanisms. To analyse the relative order of activation across the eight datasets I compared the peak time of each gene to that of all others in the peak class, adopting a permutation strategy to assess significance.

A total of 57 protein coding and non-coding candidate IEGs (all corresponding to known Ensembl genes) from the robust candidate set were considered for construction of the conserved activation network. For genes with multiple peaking CAGE TSSs I chose the earliest peaking CAGE TSS (lowest tp) in each dataset, thus removing any possible statistical artefact due to multiple TSSs associated to genes of the robust set, then the relative pairwise order of activation for each gene was computed with respect to all the other genes in the robust set. I define the relative activation ordering between Gene1 and Gene2 to be conserved if $t_{p,\,Gene1} < t_{p,\,Gene2}$ in at least 7 of the 8 datasets (Figure 23).

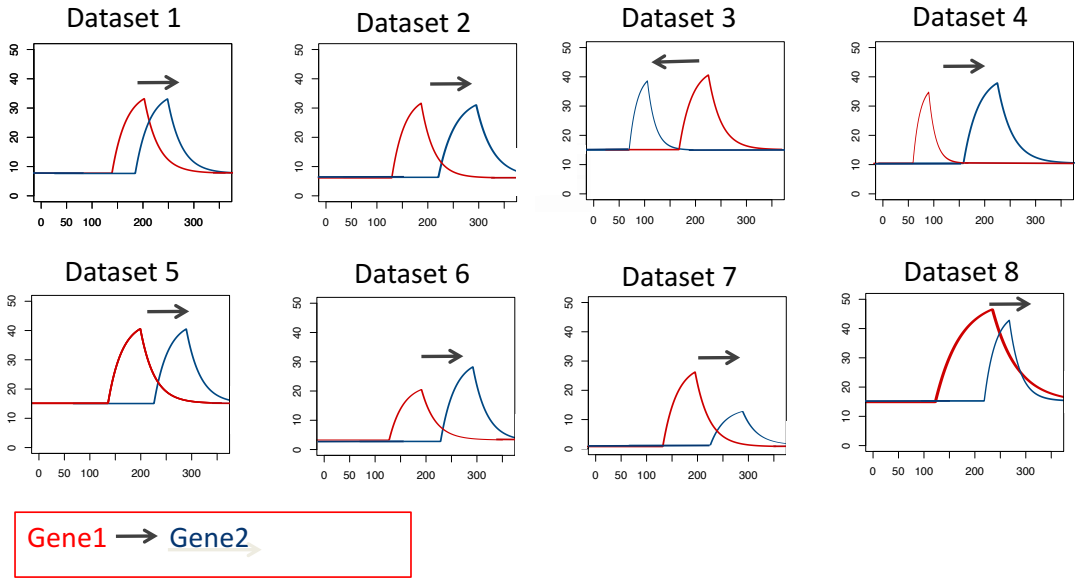


Figure 23 Defining conserved activation order. The eight plots show the time course expression of Gene1 (in red) and Gene2 (in blue) in the eight datasets. The arrows over the plots connect the smaller t_p with the greater t_p. Gene1 t_p is smaller than Gene2 t_p in seven out of eight datasets (all datasets except dataset 3), therefore I conclude that Gene1 had a conserved activation order respect to Gene2 and I connect them with an arrow directed from the earliest gene, Gene1, to the latest gene, Gene2.

Applying this procedure to all 57 coding and non-coding genes of the robust set I discovered 40 genes temporally connected by 77 conserved relative orderings (Figure 24). *FOS* was the first gene to be activated, in the sense that it lacks a predecessor in these data, and *SDC4*, *EHD1* and *TMEM185B* were the last. Many genes in this network are known to participate in well-studied pathways active in the IER such as FOS, JUN, DUSP1 and FLNA which are part of the MAPK signalling pathway, one of the most studied signaling pathways with vital roles in many cellular processes in normal and cancerous cells (Orton, Sturm et al. 2005). In addition, FOS, JUN, SDC4, DUSP1, ARF4 and PLEC are part of the EGFR1 regulated signaling pathway, which plays roles in cellular proliferation and survival and is often deregulated in cancer (Fromm, Johnson et al. 2008). The significance of the number of temporal connections observed was measured relative to a null distribution constructed by permuting t_p for all the CAGE TSSs 1,000,000 times and repeating the pairwise ordering at each iteration (Figure 25); with the proportion of permuted

datasets with at least as many conserved orderings as the observed taken as an empirically derived p-value. The observed value (77 conserved orderings) was detected or exceeded in 4,516 out of 1,000,000 permutations indicating that the number of temporal connections was significantly higher than expected by chance (p-vale < 5e-3). The average number of connections for the permutations is 36.56 with a confidence interval of '36.53 ; 36.58', and the experimental value, 77, is 3.2 standard deviations from the permutation average value. Thus, there is evidence in these diverse time course datasets for a conserved coordination of promoter activation during the IER, that includes known IEGs and further supports the candidacy of the novel IEGs detected, such as the strong candidate XBP1 TF coding gene.

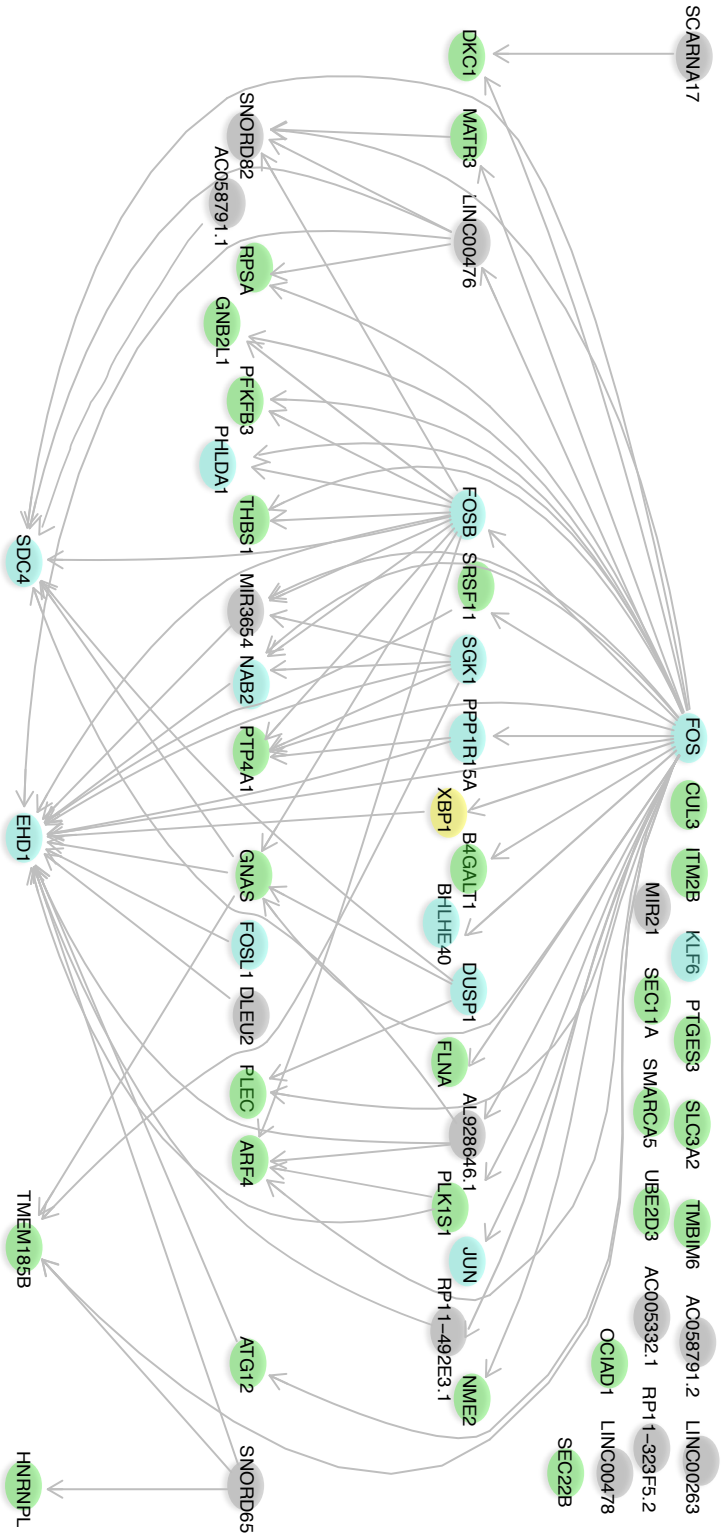


Figure 24 Conserved activation network. The network describes the conserved order of activation for known IEGs (in light blue), TF XBP1 (in yellow), ncRNA genes (in black) and other protein coding genes (in green). The directional arrows connect the earlier peaking genes with the genes peaking later. The 17 genes in the right top of the plot are part of the protein coding and non-coding robust set but do not show conserved temporal ordering with any other gene in the network and are therefore not connected with any arrow.

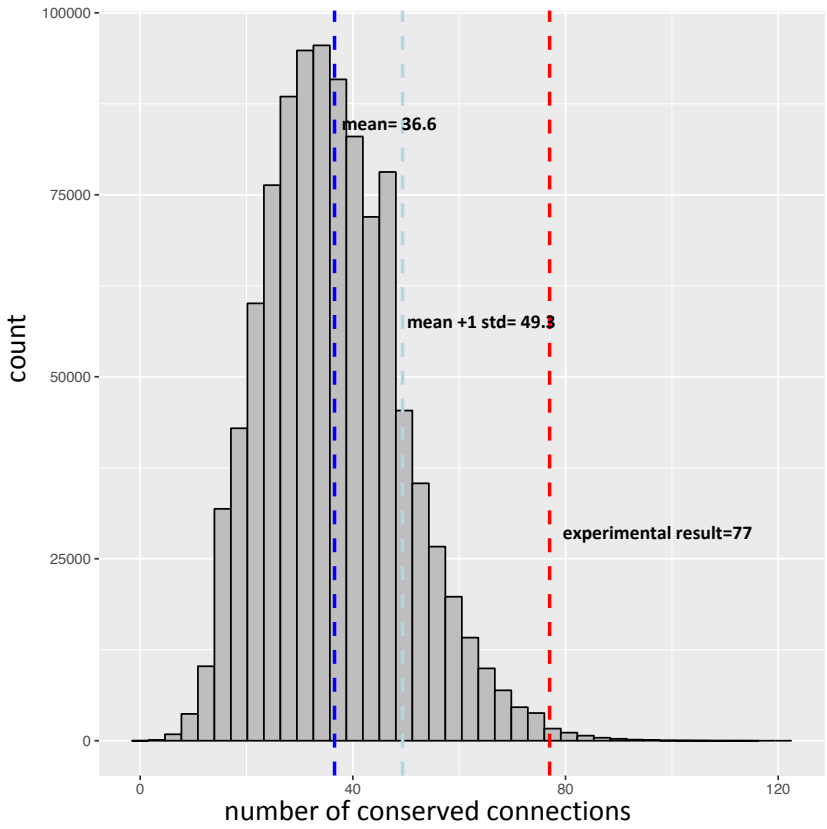


Figure 25 Pair wise connections permutation test. Distribution for 1,000,000 permutations of the conserved pairwise connections between couple of genes across at least 7 out of 8 datasets. Permuting the data, I found an average of 36.56 conserved connections (blue dashed line). The experimental result, 77 conserved pair-wise connections, corresponds to 3.2 standard deviations from the mean of the permutation distribution. Only 4,516 permutations exceeded the experimental result, which corresponds to p-value < 5e-3.

5.4. Canonical IEG TF binding sites

IEG promoters are known to be enriched for specific TF binding sites such as those bound by serum-response factor (SRF), nuclear factor-kB (NFKB) and cyclic AMP response element-binding protein (CREB), suggesting that IER transcriptional regulation mechanisms are shared and possibly redundant (Tullai, Schaffer et al. 2007, Healy, Khan et al. 2013). To test for similarities between the known IEGs and the candidate IEGs I firstly verified the enrichment of the IEG specific TF motifs in the FANTOM CAGE TSS regions of the 212 known IEGs and then I compared with their enrichment in the robust set. The sequences corresponding to 200 bases upstream and downstream of

the centre (400 base window) of the CAGE TSSs associated to the tested genes were matched to the collection of JASPAR CORE TFBS matrices (Sandelin, Alkema et al. 2004), comprising a curated collection of experimentally defined TF binding sites for multi-cellular eukaryotes.

Using a stringent threshold (adjusted motif occurrence p-values lower than 0.05), I found significant matches for a total of 352 motifs in the set of all CAGE TSSs tested (corresponding to 12,132 tested genes) and to 157 motifs within the CAGE TSSs associated with the 212 known IEGs. As expected, in this IEG set I found motifs belonging to the known IEGs regulators: SRF, CREB5, NFKB1, NFKB2 (complete list in Appendix Table 4). Furthermore, using Fisher's exact test to compute the enrichment of each motif in the IEG set relative to the total gene set, I found 145 enriched motifs (FDR adjusted p-value < 0.05) including the known IEGs regulators already mentioned. The protein-coding robust set was enriched for 266 motifs, and I found all the known regulators and another 51 binding sites enriched in both the known IEGs and the robust set (Table 17, Figure 26). Remarkably, the promoter region of the *XBP1* candidate IEG is also enriched for NFKB1 and NFKB2, two known IEG regulators.

Table 17 Enriched transcription factor binding sites in IEG promoters. TF binding sites significantly enriched in the known IEGs as well as in the robust set. The p-values, odd ratios and q-values are referred to the robust set.

Motif gene name	pv	Odd ratios	q-value
RELA	9.2e-05	3.5	0.0061
TBP	2.7e-04	4.3	0.0098
Nfe2l2	9.7e-04	3.1	0.0132
MAF::NFE2	2.4e-03	3.1	0.0227
NFKB2	9.5e-04	2.9	0.0131
blmp-1	2.2e-03	2.6	0.0216
RREB1	3.9e-04	3.7	0.0120
Trl	1.0e-02	2.2	0.0477
AGL15	7.4e-03	2.5	0.0402
AP3	3.1e-03	2.9	0.0255
FLC	9.4e-03	2.5	0.0461
AGL27	2.9e-03	2.9	0.0247
HSF1	6.4e-03	2.6	0.0366
AGL1	5.3e-05	4.9	0.0053
CEBPA	2.5e-04	4.3	0.0093
JUN	1.6e-03	3.0	0.0181
JUND(var.2)	1.3e-04	3.9	0.0066
su(Hw)	5.7e-03	3.1	0.0343
dl(var.2)	2.4e-04	3.2	0.0093
HMG-I/Y	1.6e-03	2.8	0.0182
SCRT2	1.3e-04	3.8	0.0066
REL	4.6e-03	2.4	0.0291
IRF2	3.6e-03	2.8	0.0267
AZF1	8.6e-03	2.3	0.0431
TBR1	1.7e-03	3.3	0.0185
GIS1	1.5e-04	4.1	0.0074
CDF3	4.2e-03	3.0	0.0285
DOF2.4	2.8e-03	3.1	0.0241
TGA7	2.6e-03	2.8	0.0233
Stat4	2.5e-03	2.7	0.0233
FOSL1	5.2e-04	3.6	0.0122
FOSL2	3.3e-03	2.8	0.0260
JUNB	4.5e-04	3.6	0.0120
JUND	1.0e-02	2.7	0.0474
SCRT1	7.2e-04	3.4	0.0127
Stat92E	2.8e-03	2.9	0.0242
bZIP910	2.1e-03	3.7	0.0214
TGA1	8.5e-03	2.5	0.0431
AP1	3.4e-03	3.0	0.0260
JDP2(var.2)	3.4e-04	3.7	0.0114
BATF3	7.3e-04	3.4	0.0127
Creb5	4.7e-03	3.0	0.0297
BZIP60	3.8e-03	3.1	0.0275
OAF1	1.5e-03	3.0	0.0177
YNR063W	3.4e-03	2.8	0.0261
SEP1	1.4e-03	3.9	0.0172
MEF2A	3.8e-03	3.3	0.0275
MEF2C	5.3e-04	3.6	0.0122
MGA	1.0e-03	3.3	0.0134
squamosa	1.2e-04	4.4	0.0066
T	9.1e-03	3.7	0.0452
HLF	2.6e-03	4.9	0.0233
DBP	5.3e-03	4.9	0.0328
SRF	5.5e-04	5.0	0.0122
SPT15	7.9e-05	6.7	0.0057

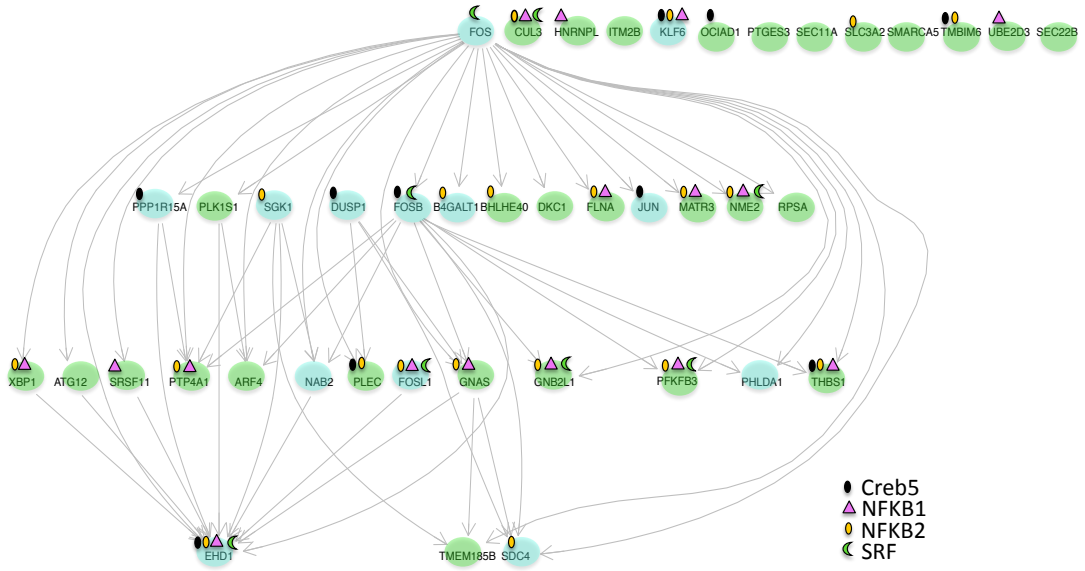


Figure 26 IEG known regulators in the robust set. Each symbol and colour represents one of the 4 known regulators of IEGs (Creb5, NFKB1, NFKB2, SRF). Known IEGs are indicated in blue while the candidate IEGs are green.

I found occurrences of the 1,007 tested motifs in 382 promoter regions associated to non-coding RNA genes detected in the eight datasets. However, the non-coding robust set was not enriched for any motif in respect to the other non-coding genes tested across all the datasets. This could indicate that the activation of non-coding genes in the IER relies on regulatory mechanisms that are not shared by the protein coding known and candidate IEGs.

5.5. XBP1 binding site enrichments

Among the candidate IEGs in the robust set, *XBP1* is particularly interesting. It encodes for a TF which is a key component of the Unfolded Protein Response (UPR) signalling pathways involved in the response of cells to ER stress (Yoshida, Matsui et al. 2001). The ER plays many important roles in functions such as calcium storage and gated release, folding and processing of membrane and secretory protein, biosynthesis of lipids and metabolism. Furthermore, the ER is highly sensitive to intracellular and extracellular stimuli.

When misfolded proteins accumulate in the ER, the ER homeostasis is altered (ER stress), and many important cellular signalling processes are affected, such as apoptosis, differentiation and energy production (Cao and Kaufman 2014). The main regulatory mechanism for XBP1 is alternative splicing mediated by the ER transmembrane kinase IRE1, however, although much less studied, its transcription is regulated by many tissue-specific and developmentally regulated TFs, such as ATF6 (Tsuru, Imai et al. 2016), in a highly dynamic manner, suggesting it as a compelling therapeutic target for ER-related disorders, including the majority of metabolic diseases (He, Sun et al. 2010). As discussed previously, *XBP1* is transiently activated after *FOS* and before *EDH1* across the eight datasets, and is therefore a strong candidate IEG. Interestingly, I found a significant enrichment (p-value < 0.05) for the XBP1 binding site in the promoter regions (see Methods) of 51 known IEGs (Table 18) and in the robust set of genes (Figure 27, Table 18).

Table 18 XBP1 binding site enrichment. List of known IEGs and genes in the robust set associated to promoter regions (400 bases window) enriched for XBP1 binding sites.

Known IEGs' promoter region
NFKB2; DUSP5; CREM; SPTY2D1; EHD1; FOSL1; UBC; TNFAIP2; NFKBIA; ZFP36L1; FOS; ARL4D; SOCS3; TGIF1; ICAM1; JUNB; NFKBIB; BCL3; PPP1R15A; SLC16A1; RGS4; IER5; RGS1; RGS2; CLIC4; KLHL21; CYR61; ADAMTS1; ETS2; ATF4; SLC20A1; IL1B; NFE2L2; KLF7; REL; PELI1; NFKBIZ; SIAH2; CCNL1; CSRNP1; NFKB1; DUSP1; SQSTM1; ELL2; SGK1; VEGFA; IL6; MYC; NR4A3; KLF4; ZFAND5
Robust set genes' promoter region
PFKFB3; EHD1; FOSL1; TMBIM6; PTGES3 ; FOS; PPP1R15A; SRSF11; GNAS; TMEM185B

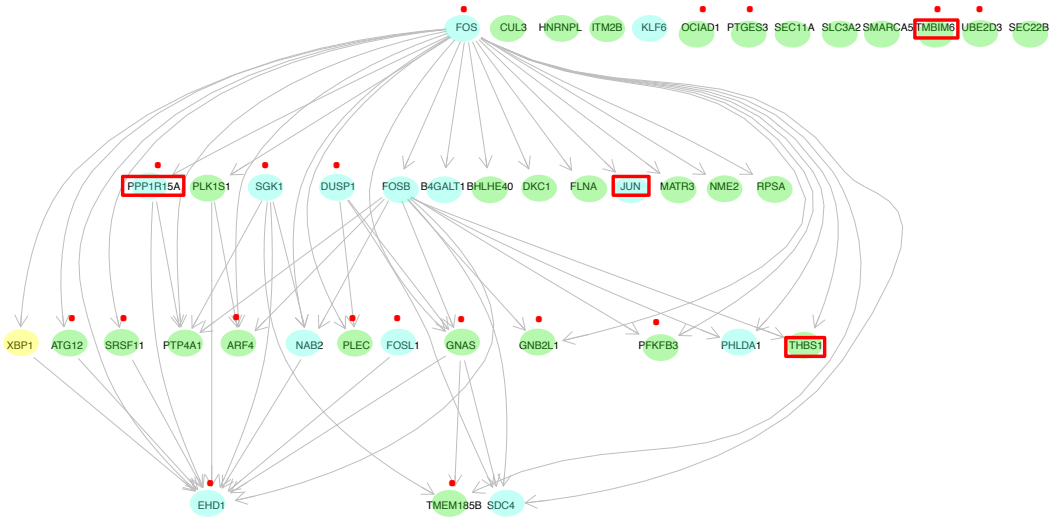


Figure 27 The conserved activation network of the candidate IEG XBP1. Conserved time of peaking across at least 7 datasets, showing genes sharing the enriched GO term GO:003497 response to endoplasmic reticulum stress (red rectangles) in the network. Red points indicate genes with promoters enriched for XBP1 transcription factor binding sites. Light blue, yellow and green ovals highlight known IEGs, the TF XBP1 and other protein coding genes, respectively.

Furthermore, genes in the robust set are significantly enriched for the GO term GO:003497 'response to endoplasmic reticulum stress' (q-value <0.05, all tested genes as the background), and four of the five genes in the robust set sharing this term peak in conserved order across the datasets (Figure 27). These results support the candidacy of XBP1 novel IEG and suggest a role for the IER in the UPR pathway.

5.6. Discussion

In this chapter I discussed the extent of conserved ordering in gene induction and in the IER. *FOS* expression has long been considered to lead the IER after cell stimulation (Hu, Mueller et al. 1994, Fei, Viedt et al. 2000), and the IER conserved activation network support this, but the network also includes similarly conserved relationships extending to an additional 39 coding and noncoding RNA genes. Furthermore, I observed many known and candidate IEGs in this network known to be involved in a range of signalling pathways active in the IER, such as the MAPK and the EGF/EGFR signalling pathways. This suggests that the variable constellations of genes involved in the IER to

any particular stimulus may be underpinned by a deeper level of conservation in the regulation of the IER across stimuli.

One of the most interesting candidate IEGs, XBP1, can be rapidly activated by alternative splicing minutes after cell stimulation with mitogenic hormones, activating peptides such as LPS and cytokines (Skalet, Isler et al. 2005, Andruska, Zheng et al. 2015, Shapiro, Livezey et al. 2016). This key event of the induced unfolded protein response (UPR) pathway is a conserved eukaryotic response to cellular stress, and is thought to cooperate in the regulation of immediate early gene expression (Shapiro, Livezey et al. 2016). However, the dynamics of *XBP1* promoter induction in the context of the IER have not been studied previously. Interestingly I found a significant enrichment for XBP1 TF binding sites in the promoter regions of 14 genes in the IER conserved activated network. The presence of XBP1 and XBP1-responding genes in the temporally conserved network supports this candidate IEG suggesting that it may act as an important novel link between the IER and the UPR pathway.

The observed strongly conserved ordering of activation and the presence of characteristic IEG-regulators in the robust set, may suggest the existence of core ubiquitous IEGs and conserved regulatory mechanisms controlling the induction of these genes, motivating further studies of the dynamics e underlying mechanisms of the IER.

6. Changes in the choice of promoter and in read patterns across promoters over time

6.1. Introduction

Promoters are defined as modulatory structures that contain the necessary regulatory elements required for cells to initiate transcription and to control gene expression (Molina and Grotewold 2005). It is becoming increasingly evident that promoters contribute heavily to the diversity and flexibility of gene expression through the regulated choice of alternative promoters and transcription initiation sites (Ayoubi and Van De Ven 1996, Trinklein, Aldred et al. 2003). Promoter choice can result in tissue-specific or stimulus-specific levels of transcription, translational efficiency and the generation of alternative protein isoforms (Ayoubi and Van De Ven 1996, Consortium 2014). Despite being the most extensively studied category of regulatory sequence in eukaryotes and prokaryotes (Myers, Tilly et al. 1986, Landolin, Johnson et al. 2010), a systematic and comprehensive annotation of the cell-type and condition specific promoter expression profile has not been performed until recently due to a lack of adequate sequencing data and the low sensitivity of transcript quantification and identification tools. In 2014, the FANTOM consortium, making use of single molecule CAGE-seq technology, mapped the TSSs of human and mouse primary cells, cell lines and tissues, providing the most complete annotation of mammalian promoters to date (Consortium 2014). Schoer et al. (2017) recently demonstrated the existence of thousands of genetic variations associated with changes in the shape of TSSs in 81 drosophila lines (Schor, Degner et al. 2017). Those variations, which they called tssQTLs, are thought to increase expression noise; however, in many cases their effect is buffered by other heteroallelic variations, indicating that the reduction of expression noise is a key factor in promoter evolution.

Haberle et al. (Haberle, Li et al. 2014) identified dynamic changes in promoter shape and choice during maternal to zygotic development in zebrafish,

indicating a switch between two distinctive modes of transcription initiation in maternal and zygotic transcriptomes.

However, a comparative analysis of alternative promoter choice and shape in a range of stimulated cell types has not yet been performed and many questions are still open, including the extent to which promoter choice and shape is conserved across human cell type/stimuli. In this chapter, I will analyse the extent of alternative promoter choice for genes belonging to the robust set and other subsets previously identified. Furthermore, I will attempt to reproduce the Haberle et al. (2014) approach to investigate the extent of promoter shape changes across time in some of the best characterized datasets of the FANTOM5 human time course collection, including the PMDM_LPS, MCF7_HRG, MCF7_EGF1 datasets described in previous chapters, and an additional time course dataset for human H1 embryonic stem cells differentiated into CD34-positive hematopoietic cells (H1_CD34). Overall this chapter aims to use these deeply sequenced samples to explore human promoter dynamics, and study the extent of alternative promoter choice, as well as change in promoter shape (Figure 28).

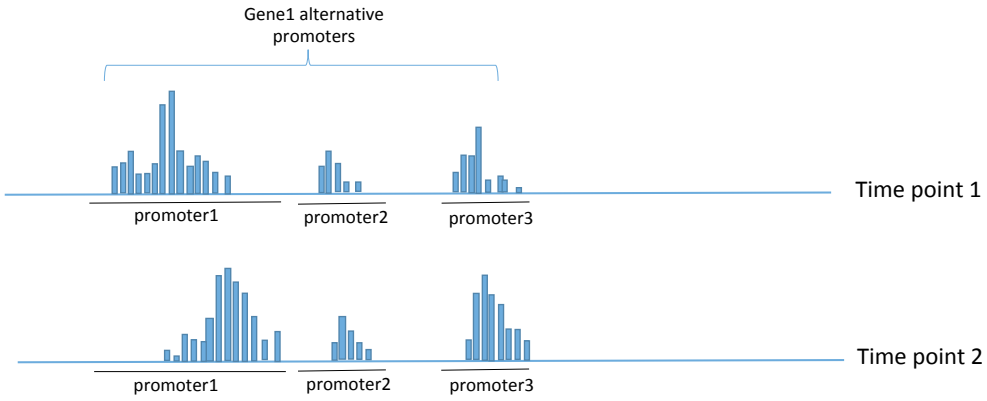


Figure 28 Schematic view of alternative promoters and choice and change in promoter shape. Different datasets can choose different promoters to start transcription of the same gene. Gene1 can be initiated by three alternative promoters: promoter1, promoter2 or promoter3. In this graph we can appreciate the shape of promoter1 'shifting' to the right in time point2.

6.2. Promoter choice in the IER

The existence of alternative promoters and alternative TSSs, is believed to be important for the control of gene expression under different conditions and

ensures that important genes are expressed at optimal levels by having multiple promoters available which can substitute for one another (Carbajo, Magi et al. 2015). The IER consists of only a few hundred participant genes and only a few are active across multiple cell types and stimuli. However, this relatively small group of genes participates in a wide range of fundamental biological processes. Carbajo et al. (2015) observed that FOS and EGR1 IEGs are characterized by the presence of alternative TSSs and some of these alternative TSS are expressed in MCF7 cells treated with both HRG or EGF1 while other alternative TSSs were activated only in one of the two experiments (Carbajo, Magi et al. 2015).

Comparing the number of alternative TSSs associated with the 212 known IEGs extracted from the literature with the number of alternative TSSs associated with the total set of genes detected across the eight datasets, I observed a higher (t-test p-value = 1.1e-3) number of TSSs for the known IEGs (mean = 2.5) with respect to the whole set (mean = 1.9). The higher number of TSSs of the known IEGs could constitute an additional regulatory mechanism which generates diversity in the IER and ensure the occurrence of the proper initiation of these critical genes.

To support the idea that the genes in the robust set participate in the core regulatory mechanisms of the IER, I would expect that the genes in the robust set would have more TSSs than the other peaking genes which are transiently activated in only few datasets. I found that the genes in the robust set are characterized by a significantly higher number of TSSs than the whole peaking dataset (t-test p-value = 1.6e-4). However, candidate IEGs in the robust set show only slightly greater number of alternative TSSs they activate across datasets compared with known IEGs (Table 19) and a test of significance could not be performed due to the small size of the dataset (42 genes in the robust set). Comparing the distribution of the number of alternative TSSs for known and candidate IEGs in the permissive set, I didn’t observe a significant difference (t-test p-value>0.1), although the permissive set genes have

significantly higher number of TSSs (t-test p-value < 2.2e-16) than the whole peaking set. These results could indicate a relationship between a high number of alternative TSSs and a central role in the IER, or could be simply related to the higher chance of genes associated with many TSSs to be classified to the peak model in more datasets (as discussed in Chapter 4.4).

Table 19 Alternative TSSs. Summary of the distribution of alternative TSSs in different subsets of the data: the whole set of peaking genes and the candidate and known IEGs in the robust and permissive set. All the detected alternative TSSs, regardless their signature, are considered here.

Subset	Min	Median	Mean	max	Number of genes
All peaking genes	1.0	2.0	2.1	48.0	8,785
Robust set	1.0	4.0	5.6	29.0	42
Robust set known IEGs	1.0	3.0	2.5	6	13
Robust set candidate IEGs	1.0	5.0	7	29	29
Permissive set	1.0	2.0	3.1	48.0	1,304
Permissive set known IEGs	1.0	2.0	2.7	20.0	86
Permissive set candidate IEGs	1.0	2.0	3.1	48.0	1,218

To analyse the extent to which the different TSSs show a peaking behaviour across multiple experiment datasets (PMDM_LPS, MCF7_HRG etc.), I tested the TSS associated with the genes in the whole peaking set, the robust set and the permissive set which exhibit peaking behaviour (which is a portion of the total number of TSS associated to each of this gene). The extent of promoter choice across the robust set of IEGs, candidate IEGs and non-coding RNA is shown in Figure 29, Appendix Figure 1, Appendix Figure 2 and Appendix Figure 3.

For both the permissive and the robust sets, known IEGs tend to possess CAGE TSSs that are successfully classified to the peak model across a larger number of datasets (median number of datasets classified as peak per TSS for known IEGs in the robust set = 4, in the permissive set = 3; candidate IEG median proportion = 2 for both robust and permissive set, Table 20). Again, given the smaller number of TSSs tested in the robust set I could calculate significance only for the permissive set and I found significant difference between the means of the two groups (t-test p-value < 0.001).

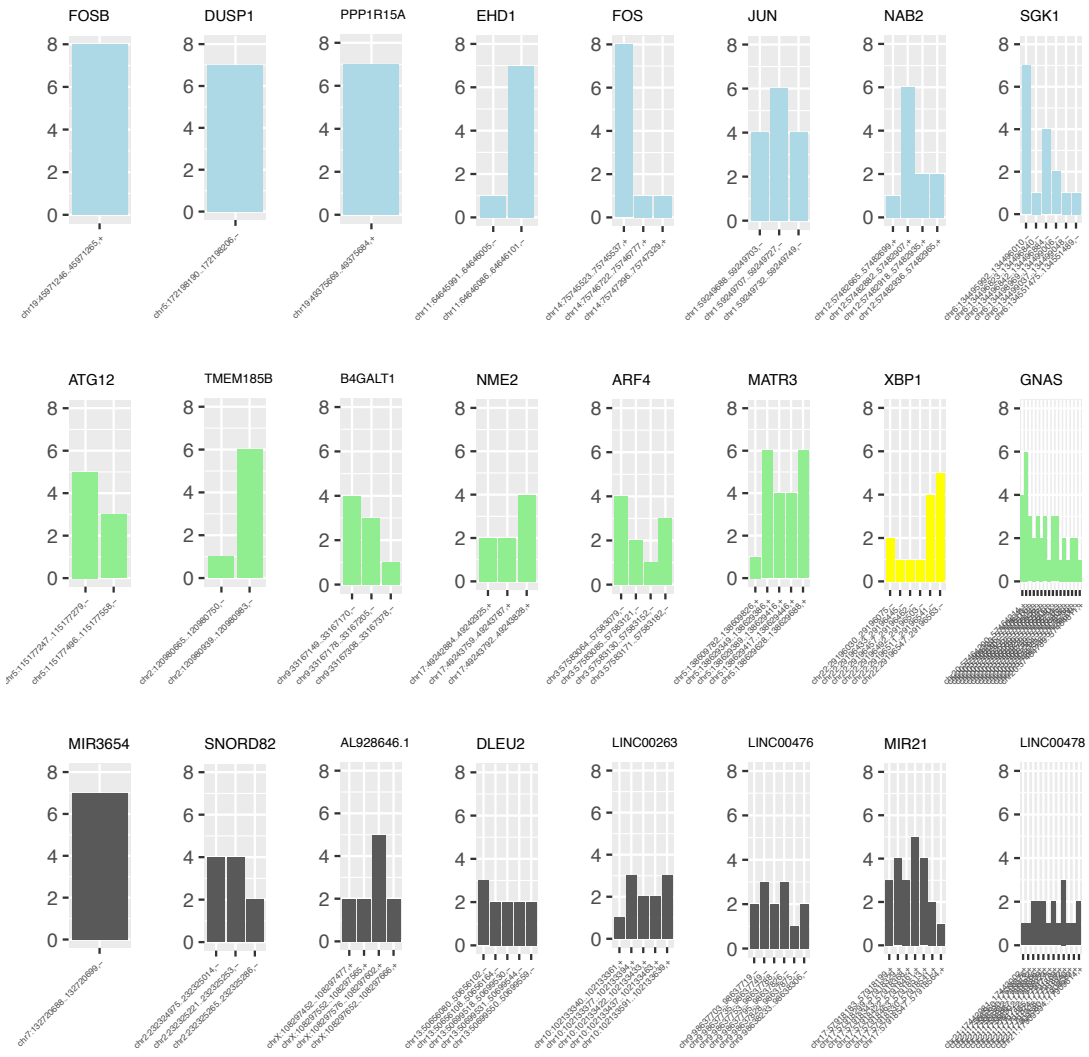


Figure 29 Promoter choice across time series datasets. For representative genes, bar charts show the number of datasets where each TSS peaks to illustrate the diversity of TSS choice and commonality of the peaking response. Known IEGS are shown in blue, TFs in yellow, non-coding RNA in grey and other genes in green. FOSB has a single TSS that peaks in eight datasets, MIR3654 has a single TSS that peaks in 7 datasets, JUN has three TSS each peaking in four or more datasets and XBP1 has six TSS that peak in between one and six datasets.

Table 20 Known IEGs promoter initiation sites are conserved across datasets. The table shows a summary of the number of datasets sharing the same peaking alternative TSS in the whole peaking dataset, the robust set and the permissive set for candidate and known IEGs.

Subset	Min	Median	Mean	max	Number of peaking TSSs
All peaking genes	1.0	1.0	1.7	8.0	12,812
Robust set	1.0	2.0	2.8	8.0	172
Robust set known IEGs	1.0	4.0	4.0	8.0	31
Robust set candidate IEGs	1.0	2.0	2.5	6.0	141
Permissive set	1.0	2.0	2.5	8.0	3,077
Permissive set known IEGs	1.0	3.0	3.1	8.0	191
Permissive set candidate IEGs	1.0	2.0	2.4	6.0	2,886

Thus, in the robust and permissive sets, known IEGs tend to possess comparable numbers of alternative TSSs to candidate IEGs, but tend to show discernible peaks in the majority of the time series datasets, while candidate IEGs appear to peak in different biological systems using specific alternative TSSs. For example, in the robust set, both *SGK1*, a known IEG involved in the regulation of many fundamental processes, such as apoptosis, inflammation and cell growth (Lang and Shumilina 2013), and XBP1, a transcription factor and candidate IEG, which contributes to regulate cell stress responses (Yoshida, Nadanaka et al. 2006), are associated with 6 alternative TSSs peaking across the datasets; I observed a primary TSS peaking in 7 out of 8 datasets for *SGK1* and a more dispersed distribution of peaking alternative TSSs used by *XBP1* across different datasets (Figure 29).

It is possible that these relatively stereotypical transcriptional characteristics of known IEGs may, in some cases, have led to their status as well-established IEGs. Similarly, the stimulus-specific signature seen for the TSSs of candidate IEGs could have led them to a failure to be detected previously. Another possible implication of these results is that some genes in the robust set may be statistical artefacts. However, this issue has been discussed previously in this thesis (Chapter 4.4) and the results of additional analysis support the

candidate IEGs in the robust set; all the p1 TSSs associated to the 13 known IEGs in the robust set are found to peak in at least four datasets as well as the p1 TSSs of 16 of the 29 candidate IEGs of the robust set.

6.3. Shape changes at TSSs

The distribution of CAGE tags inside TSS regions defines the promoter shape. Using CAGE single base TSS mapping, promoters were classified into two major classes: 'sharp' (or 'narrow') promoters, where the majority of the tags are concentrated in a narrow region with a single dominant TSS, and 'broad', with a wide-spread distribution of TSSs (Carninci, Sandelin et al. 2006). Genes activated in different samples (cell types, time points after external stimulation, developmental stages) can be characterized by alternative promoter choice and also by variances in the promoter shape. However, traditional analysis of expression profiling cannot characterize changes in the distribution of TSSs whereas the overall transcript abundance does not change and therefore other approaches have been proposed. Haberle et al. (Haberle, Li et al. 2014) used CAGE TSS mapping to analyse the change in TSSs shape across 12 early embryonic developmental stages in zebrafish, from the unfertilized egg to organogenesis. between the maternal transcriptome stage and the mid-blastula transition, they observed a change in promoter shape for about 900 promoters (Haberle, Li et al. 2014).

In this work, I analysed the changes in the shape of promoter across different time points in three short and densely sampled time series and one long and sparsely sampled embryonic developmental time series from FANTOM5. The goal of the study was to test whether I could recapitulate in human cells the differences in promoter shape previously observed during embryonic transition in zebrafish (Haberle, Li et al. 2014). To measure the change in CAGE tag distribution inside TSS regions, I applied the CAGEr pipeline developed by Haberle et al (2014). It consists of two major steps: the definition of 'shifting score' representing the proportion of reads in shifted locations, and a statistical

significance assessed by corrected p-values (FDRs) generated applying a two-sample Kolmogorov-Smirnov (KS) test.

6.4. Shifting score and KS analysis

After tag count normalization over the bam file of each dataset, the cumulative distribution of CAGE signal along all TSS clusters at all the time points was calculated. The shifting score and the P-value of Kolmogorov-Smirnov tests for all CAGE tag starting sites (CTSS) clusters was computed between time 0 and all the other time points. The shifting score is defined as:

$$Score = \frac{\max(F_1 - F_2)}{\max(F_1)}$$

Where F_1 represents the cumulative sum of the CAGE signal of the TSS cluster in the sample with lower total signal for the considered TSS cluster (Figure 30, blue line), and F_2 represents the cumulative sum for the opposite group (Figure 30, red line).

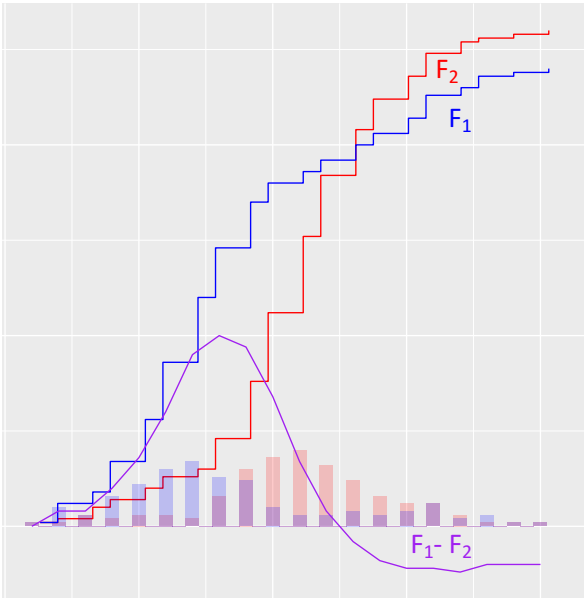


Figure 30 Shifting CAGE TSSs. The shifting (in purple) of the CAGE TSSs is computed as the difference between the cumulative distributions (lines) of the CAGE tags (columns) for the dataset with smaller total counts (F_1, in blue) and the bigger total counts (F_2, in red).

Shifting score is computed for both forward (5' → 3') and reverse (3' → 5') direction and the bigger value is selected. The shifting score spans the values between -∞ and 1. The value of 1 is assigned when there is complete physical separation between the TSS cluster in the two samples.

The positive values represent the proportion of counts in the lower expressed sample which not covered by the reads in the other sample.

The Kolmogorov-Smirnov test is a non-parametric statistic for comparing two empirical distributions which measures the divergence between two cumulative distribution curves computing their maximum absolute difference $D_{m,n}$ (Lopes, Reid et al. 2007, Filion 2015) :

$$D_{m,n} = max_x|F(x) - G(x)|$$

Where *m* is the size of the cumulative distribution of the function *F(x)* and *n* is the size of the cumulative distribution of the function *G(x)*.

The shifting score and the p-value for the KS test along with the values adjusted for multiple testing (FDR) are computed using an empirical

cumulative distribution functions (ECDF) which represent sampling of the distribution of reads for all the TSS consensus cluster in the two samples (Haberle, Haberle et al. 2013, Haberle, Forrest et al. 2015).

6.5. Shifting promoters: variation in spatial promoter activity

Even when the transcription of a gene originates from the same promoter over a time course, there may be substantial changes in the patterns of transcription initiation within that promoter region (Figure 31). I applied the approach described by Haberle et al. (2014) to investigate the extent to which promoters change shape in three time-course datasets previously analysed in this work, PMDM_LPS, MCF7_HRG and MCF7_EGF1, and an additional FANTOM time-course dataset consisting of the differentiation of embryonic stem cells to hematopoietic cells, H1_CD34. The H1_CD34 dataset was also generated by the FANTOM5 project and consists of 3 time points: Day 0, Day 3 and Day 9 during haematopoietic differentiation of H1 embryonic stem cells and has not been considered before in this thesis because it is not densely sampled in the first few hours after stimulation.

The goal is to examine the evidence for shifting promoters and to compare the extent of promoter shifting in datasets with different magnitude of change in the transcriptome repertoire. H1_CD34 is expected to show the most drastic changes as it involves the switch between embryonic pluripotent cells to differentiated haematopoietic progenitor cells (HPC). Whereas Haberle et al. (2014) found about 900 significant “shifting promoters” (KS FDR <= 0.01 and shifting score > 0.6) in Zebrafish embryogenesis, I observed few shifting promoters in MCF7_HRG and H1_CD34 (Table 21). Even relaxing the thresholds (FDR <= 0.05 and shifting scores >= 0.4 and 0.2) the number of shifting promoters in all four datasets is still very small (Table 21).

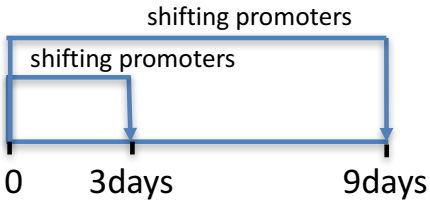


Figure 31 Shifting promoters. Promoters characterized by differences in read distribution between the beginning of the time series (time 0) and each sampling time points are defined as "shifting promoters".

Table 21 Shifting promoters for different thresholds of KS FDR and shifting scores. Each cell contains the number of shifting promoters across time points for increasingly stringent shifting scores and KS FDRs.

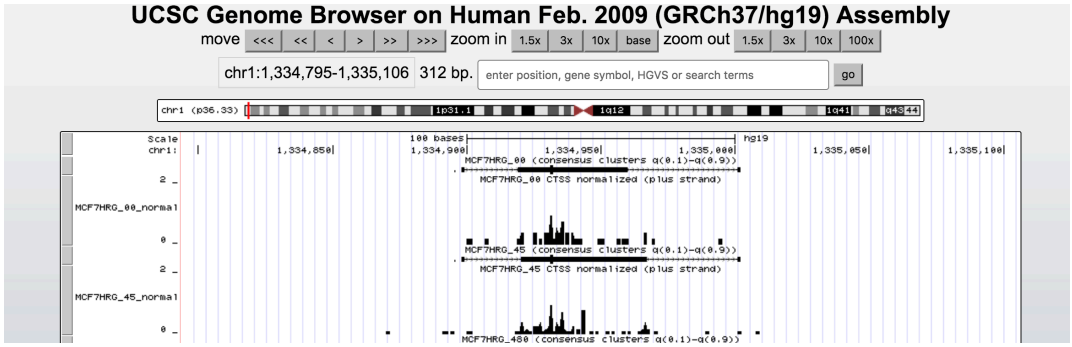
No. shifting promoters	KS FDR 0.05			KS FDR 0.01		
	Score> 0.2	Score>0.4	Score>0.6	Score>0.2	Score>0.4	Score>0.6
PMDM_LPS	21	3	0	15	3	0
MCF7_EGF1	15	4	0	9	2	0
MCF7_HRG	23	8	5	10	5	4
H1_CD34	211	59	23	105	23	8

Table 22 shows the difference in the resulting output for one exemplificative shifting promoter and a non-shifting promoter arbitrarily chosen in MCF7_HRG (complete list of shifting promoters in Appendix Table 6). The dominant TSSs location changes between compared data points in the shifting promoter (shifting score = 0.8 and KS FDR <= $2.6e^{-3}$), while it is stable in the non-shifting promoter. A table with the 8 shifting promoters of H1_CD34 and 4 of MCF7_HRG can be found in the appendix (Appendix Table 5). A genome browser can be then utilized to visualize the change in read distributions (Figure 32).

Table 22 Comparison of shifting and non-shifting promoters in MCF7_HRG. Shifting scores, KS FDRs, as well as coordinates and CAGE signal (in TPM) of dominant TSSs in the shifting promoter chrX:141155504..141155563,- and non-shifting promoter chr1:1334898..1335002,+ detected in MCF7_HRG.

CAGE cluster	Dominant TSS (00 min)	Dominant TSS (45 min)	Dominant TSS TPM (00 min)	Dominant TSS TPM (45 min)	Shifting score	KS FDR
chr1:1334898..1335002,+	1334932	1334932	44.65	46.24	0.05	1
chrX:141155504..141155563,-	141155551	141155524	8.65	17.55	0.80	$2.6e^{-3}$

A



B

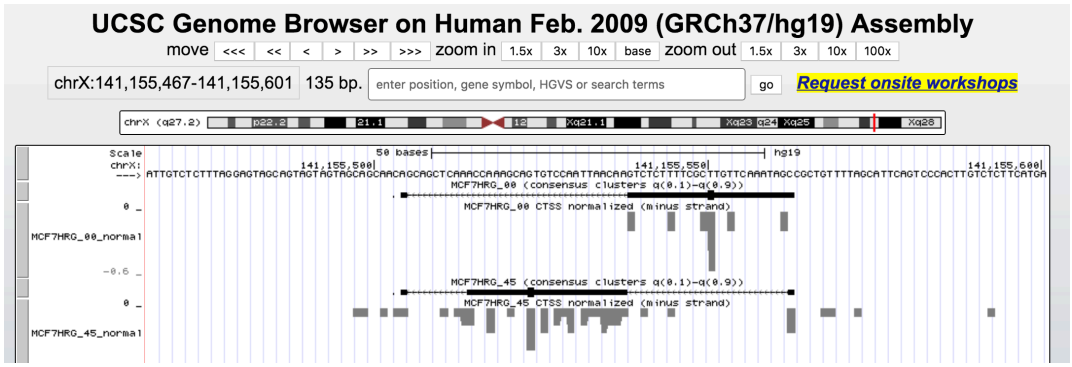


Figure 32 CAGE reads visualization in the genome browser. Read distribution at two time points (0 and 45 minutes) for the non-shifting promoter chr1:1,334,795-1,335,106,+ (A) and the shifting promoter chrX:141,155,467-141,155,601,- (B). Black horizontal lines indicate the interquartile width of the promoter region and the location of the dominant TSSs.

For these two promoters, I also performed wavelet analysis. The method, discussed in the appendix together with the preliminary results, supported the shifting of chrX:141,155,467-141,155,601,- and the non-shifting of chr1:1,334,795-1,335,106,+ promoters in MCF7_HRG between 0 and 45 minutes from stimulation.

6.6. Discussion

In this chapter I investigated extent of alternative TSS choice in IEGs over time, and the dynamics of transcription initiation within human promoters.

I observed that the IEGs, and specifically the participants in the core repertoire of the IER, are associated with a higher number of TSSs than other genes. This could reflect their varied functional requirements, as different regulatory molecules involved in different signalling pathways can activate, repress or have no effects on alternative TSSs of the same gene (Kimura, Wakamatsu et al. 2006). The stringent analysis of the p1 single TSSs discussed in Chapter 4.4 support the hypothesis that the genes in the robust set are genuine candidate IEGs and are not statistical artefacts, suggesting that the existence of multiple TSSs could be a functional feature of the IER. Comparing the alternative TSS choice for known and candidate IEGs in the robust set I found that known IEGs are characterized by a conserved group of CAGE TSSs peaking, on stimulation, across the majority of the datasets while the candidate IEGs make use of diverse cell or treatment-specific alternative TSSs to be rapidly and transiently activated. This could have led to the status of the well-established IEGs, while the increased variability seen for the TSSs of candidate IEGs could have led to a failure to detect their IEG-like behaviour in former studies.

To gain deeper understanding on promoter dynamics during the IER, I also compared the change in promoter shape across time points in a group of datasets densely sampled in the first few hours after stimulation, PMDM_LPS, MCF7_EGF1, MCF7_HRG, and a sparse differentiation time-series, H1_CD34, which describes long term changes in the promoter activation of differentiating embryonic stem cells. In general, I detected a much smaller number of shifting promoters than previously reported in a Zebrafish unfertilized egg to embryo time-series Haberle et al (2014), with a few shifting promoters in the differentiation series MCF7_HRG and H1_CD34 in respect to the other datasets. Therefore, I speculate that promoter change in shape could be a phenomenon distinctive of more drastic cellular events such as egg to embryo transition and stem cell differentiation while it doesn't seem to strongly

affect the IER. A preliminary analysis was conducted to investigate the changes in the shape of selected TSS at different resolution level using wavelet decomposition (Chapter 8.1.1). Preliminary results show that wavelet analysis of TSSs helps to understand local changes in promoter shape. It is important for future work to involve the systematic wavelet analysis of all the TSSs detected in the different datasets because it could identify shifting promoters which are not detected by other methods as well as confirming previous findings.

7. Conclusions

The immediate early response mediates essential biological processes involved in the wide range of cellular responses to external stimulation and to internal signalling. However, commonalities in the responses of known IEGs across different stimuli, the role of non-coding RNA and the general mechanisms and temporal patterns of promoter activation in the IER were not fully appreciated prior to the investigations reported here. I used eight time-series datasets from the FANTOM5 project to explore the core IEG transcriptomic repertoire and to examine the breadth of the primary response in different cell types and stimuli.

7.1. Computational modelling

Existing methods for gene expression analysis are not well suited for comparison across different time-course experiments, or to handle lowly expressed genes. They are better suited to unsupervised data mining than to model specific dynamics such as the early peak in expression of IEGs. I improved an existing method for time-series analysis that classifies expression profiles to one of a set of mathematical models of interest including the peak model describing the typical transient and rapid activation of known IEGs (Aitken, Magi et al. 2015). The updated version of the algorithm performs better when comparing time-course datasets with very different lengths and is able to handle delay in IEG activation and is therefore suitable for analysing the IER in CAGE time-series expression data. I propose that future studies of the IER and more generally of any biological process involving genes characterized by a specific expression profile could benefit by adopting this optimized method, especially when involving lowly expressed genes and the comparison of different time-course experiments.

7.2. Meta-analysis of the transcriptome identifies candidate immediate early genes and non-coding RNAs

I performed a comprehensive meta-analysis of eight densely sampled FANTOM5 time-series expression datasets, covering differentiation, growth and activation in a range of primary cells and cell lines. I discovered a permissive set of 1,301 protein coding genes whose promoters are classified as peaking in half of the datasets and a robust set of 42 protein coding genes, containing 13 known IEGs and 29 candidate IEGs, and 15 non-coding RNAs peaking in at least seven out of eight datasets. The common classification across multiple datasets is useful to identify the core participants to the IER. These results not only confirm the core role of known IEGs in the biological systems investigated by the FANTOM project but also suggest the existence of additional coding and non-coding genes involved in the IER. Known IEGs are often employed to identify cell activation, for example the up-regulation of *FOS* is used in functional anatomical mapping in the neuroendocrine system (Hoffman, Smith et al. 1993), or regarded as prognostic markers for different diseases and as targets for cancer therapy, such as the IEGs involved in the ERK signalling pathway (Kohno and Pouyssegur 2006). The list of candidate IEGs, once verified, could provide additional molecular indicators to identify cell activation, cancer progression, and the activation of specific cellular processes and could help the development of new therapeutic strategies.

7.3. Conserved temporal patterns of transcriptomic activation

Most studies of the IER dynamics are focused on the activation of a few known IEGs in different stimulated cells, and their temporal patterns of activation are well characterized. For example, the expression of FOS is known to occur in the first hour after stimulation, preceding the transcription of other IEGs and delayed response genes (Hu, Mueller et al. 1994, Fei, Viedt et al. 2000, Tullai, Schaffer et al. 2007). My work confirmed the leading role of FOS gene activation, and I observed a similar early timing for known and candidate IEGs

in the robust set, indicating shared dynamics of activation. Comparing the order of activation of known and candidate IEGs across the eight datasets I built a conserved activation network which recapitulates the regulatory landscape of the IER and supports the candidate IEGs and their relationship with other IEGs. The IER network includes 39 coding and non-coding candidate IEGs with conserved temporal relationship. Interestingly, I found that many genes in this network are activated by a range of signalling pathways associated with the IER, such as the MAPK and the EGF/EGFR signalling pathways. These results reveal commonalities in the regulatory mechanisms governing the IER across cells and stimuli and support the hypothesis that the protein coding genes in the robust set are likely to be real IEGs. Furthermore, the IER network supports previous studies on the role of the non-coding RNA species in the regulation of the IER and suggest novel insights into the interaction with other members of the IER. For example, I observed that the non-coding RNAs are not activated by the common TFs activating the known IEG, implying a different regulatory mechanism.

7.4. XBP1 links the IER to the UPR pathway

Among the candidate IEGs in the robust set, *XBP1* is the only gene that encodes for a transcription factor and shares a number of properties common to known IEGs such as the short length and the enrichment of NFKB1 and NFKB2 TF binding sites in its promoter region (Bahrami and Drabløs 2016). It is also a highly conserved component of the Unfolded Protein Response (UPR) signalling pathways which is activated by unconventional splicing upon Endoplasmic Reticulum (ER) or nonclassical anticipatory activation (Skalet, Isler et al. 2005, Andruska, Zheng et al. 2015, Shapiro, Livezey et al. 2016), and regulates a diverse array of genes involved in ER homeostasis, adipogenesis, lipogenesis and cell survival (He, Sun et al. 2010, Piperi, Adamopoulos et al. 2016). UPR activation is known to be a protective response to the accumulation of unfolded or misfolded proteins in the ER by reducing protein synthesis and increasing the protein folding, transport and degradation (Wang and Kaufman 2014). Together with the described mode of

UPR activation, the 'anticipatory' UPR, an alternative mechanism which does not involve ER stress, has been proposed. It has been described in many different human cell types subjected to different stimuli, such as B-lymphocytes treated with interleukin 4 and lipopolysaccharide, (Skalet, Isler et al. 2005), endothelial cells (EC) treated with VEGF (Karali, Bellou et al. 2014) and breast cancer cells stimulated with EGF (Yu, Andruska et al. 2016), where it is proposed to trigger the activation of IRE1, ATF6 and PERK and the consequent transcription and splicing of *XBP1*. However, the role of XBP1 in the early stage of the general response of cells to external stimulation and the dynamics of promoter activation are not fully investigated yet.

Remarkably, I observed evidence of conservation of *XBP1* activation after *FOS* and before *EHD1* known IEGs across the eight datasets. Furthermore, I detected a significant enrichment for XBP1 TF binding sites in the promoter regions of 14 genes in the temporally conserved network. These results suggest novel roles for XBP1 connecting the IER and the ER stress signalling pathways making this gene worthy of further experimental investigation at the RNA and protein level.

7.5. Promoter dynamics in the IER

FANTOM5 CAGE data have expanded our knowledge about the promoter dynamics of diverse cells in different conditions such as developmental stage or treatment (Kawaji, Kasukawa et al. 2017). However, a broad comparison of the alternative promoter usage across multiple time series data sets in the context of the IER has not been performed previously to this work. I investigated the promoter usage of the 212 known IEGs and I observed a significantly higher number of alternative TSSs in comparison to the whole set of genes detected across the eight datasets. IEGs are known to regulate many different biological processes in a specific manner and the existence of alternative promoters increases diversity of the response of cells to different stimuli (Carbajo, Magi et al. 2015). Similarly, I observed a higher number of alternative TSSs for known and candidate IEGs in the robust set, which are thought to regulate the IER in multiple datasets, than for the genes peaking in

fewer datasets and therefore active in stimulus or cell-specific IER processes. This result supports the hypothesis that genes in the robust set are characterized by many TSSs which can be activated in a specific way in different datasets. However, known IEGs in the robust set tend to show distinct peaks in the majority of the time series datasets, while candidate IEGs seems to peak in different biological system using specific alternative TSSs. These results suggest a possible explanation for the status of well-established IEGs in contrast with candidate IEGs which could be harder to detect due to the increased variability in TSS's usage.

Additionally, I investigated the variability in the exact CAGE tag composition inside each TSS region. The distribution of tags in each TSS region is called promoter shape, and Haberle et al. in (Haberle, Li et al. 2014) demonstrated that it can change over time depending on the embryos developmental stage in zebrafish. I compared the promoter shape between time points in four different systems, three densely sampled datasets (MCF7_HRG, MCF7_EGF1 and PMDM_LPS) and an additional dataset describing the differentiation of embryonic stem cells to haematopoietic cells (H1_CD34). At the same FDR and shifting score thresholds used by Haberle et al., I found only a small number of shifting promoters in H1_CD34 and MCF7_HRG differentiation series, which suggests a much smaller effect on the IER in comparison to zebrafish embryogenesis. These results provide new insights into the diversity of the regulatory processes underlying promoter activation between species, suggesting that changes in promoter shape previously observed in zebrafish embryogenesis are not observed in the human datasets analysed. However, further research is needed in order to verify these findings across multiple species and multiple biological systems.

7.6. Summary and final remarks

In this work, I performed a comprehensive meta-analysis of eight densely sampled time course datasets to identify a subset of known IEGs and sets of putative novel IEGs and non-coding RNAs which are activated across multiple cells right after a range of stimuli. Furthermore, I discovered a conserved

temporal ordering among these genes which could indicate an underlying conserved regulatory mechanism in the IER. Additionally, I explored the modalities in which alternative TSSs are used across different datasets, and I observed that, whereas both known and candidate IEGs are associated with a higher number of TSSs which are likely to increase diversity in the regulation in different biological systems, known IEGs TSS usage is conserved across datasets, with few dominant TSSs overexpressed across datasets, candidate IEGs use a broader set of cell or stimulus specific TSSs. Finally, I repeated a previously published analysis on the change in promoter shape in three of the IER datasets, plus an additional differentiation time-series, and I found a much smaller effect than the original study by Haerle et al. (Haberle, Li et al. 2014), suggesting that the promoter shifting events characterizing the transition between maternal and zygotic expression in zebrafish, is not present in the transcriptional changes of the FANTOM5 IER datasets.

7.6.1. Opportunities

Overall, the computational analyses described in this thesis provide a better understanding of the IER by reporting a set of new candidate IEGs and non-coding genes likely to be involved in the IER and investigating the temporal relationships among candidate and known IEGs and the regulation of alternative promoters associated to IEGs. Furthermore, because of the oncogenic potential of IEGs and their involvement in fundamental cellular processes in normal and unhealthy cells, the additional insights on the IER discussed in this thesis may help advance relevant biological and medical studies of cancer. As the IER plays such important roles in biology and medicine it is perhaps surprising that there have not been more detailed comparative studies of IER induction, dynamics and downstream impact of other systems. Understanding how the IER differs between tissues/cell types could lead to new tissue or disease specific interventions, for example via the design of drugs that perturb IER components active only in the cells of interest. There is currently great interest in the roles of non-coding RNAs in cellular systems, including the potential for lncRNAs to mediate IER induction and

carcinogenesis (Gao, Xing et al. 2011, Pei, Hu et al. 2018). My results show that there is still much to be learned about the IER, and even previously published data can be mined to extract new insights, including novel candidate IEGs and their interactions.

7.6.2. Limitations

One of the main limitations of studying mRNA transcription as a proxy for the abundance and activities of the corresponding proteins, is the role of regulatory events occurring after mRNA transcription, such as post-translational modifications, mRNA stability and protein degradation (Vogel and Marcotte 2012). Recent studies suggest that as little as ~40% of the change in protein abundance can be explained by changes in the levels of gene expression depending on the system (de Sousa Abreu, Penalva et al. 2009, Maier, Güell et al. 2009). Nevertheless, gene expression analysis has already provided important insights and is undoubtedly a significant source of information.

Protein synthesis blockage is known as the most direct method to identify IEGs (Tullai, Schaffer et al. 2007). However, as discussed in the supplementary material of Arner et al. (2015), only a few of the available published experiments have actually inhibited protein synthesis and the majority of the potential IEGs have been identified indirectly, by detecting an early up-regulation signature, and are strictly speaking not verified IEGs. Arner et al. propose that, in the absence of more extensive access to protein synthesis blockage data, most of the potential IEGs that are up-regulated in multiple experiments are likely to be true IEGs (Arner, Daub et al. 2015). In Aitken et al. (2015) and in this thesis we used a peaking signature to identify genes strongly up-regulated across multiple datasets. An additional limitation of our method is that we focus on core ubiquitously expressed IEGs, while potentially losing IEGs specific to single biological systems.

The most important caveat of the work presented in this thesis is the possible bias in the statistical analysis used to identify the set of candidate IEGs, induced by the existence of multiple TSSs associated to a single gene. Using

CAGE TSS expression data to identify the upregulation of IEGs across multiple datasets, it is likely that genes possessing a larger number of CAGE TSSs are more likely to be detected in multiple experiments, and therefore have a greater chance to be annotated as an IEG. On the other hand, those possessing a small number of TSSs should be less likely to be identified here as an IEG, generating false negatives and potentially other statistical artefacts among the candidate IEGs identified in this thesis. However, using more stringent analysis, relying on the selection of only one CAGE TSS for each gene, the FANTOM primary promoter p1 (Kawaji, Kasukawa et al. 2017), I demonstrated that the robust set of candidate IEGs peaking across at least 7 out of 8 datasets is likely to include many true IEGs.

7.6.3. Future work

In common with all computational analysis of high throughput sequencing data, the results presented in this thesis will require supporting experimental studies in order to validate the candidate IEGs and the IER conserved temporal network.

These experiments could primarily include the blockage of protein synthesis, using inhibitors such as cycloheximide, to confirm that the candidate IEGs are induced by the activation of pre-existing transcription factors and do not require *de novo* protein synthesis (Tullai, Schaffer et al. 2007).

Among the additional experiments required to verify the mechanisms underlying the activation of candidate IEGs, I propose the execution of longitudinal (time-course) ChIP-seq or ATAC-seq data analysis to investigate the context and binding events required for the transcriptional regulation that characterises cell and stimulus specific transcription of the candidate IEGs.

Importantly, the expected increase in the concentrations of Immediate Early Proteins (IEPs) encoded by candidate IEGs should also be verified by time-course protein assay experiments. The temporal interactions between different IEPs could then also be studied by looking at the conserved temporal ordering

of protein regulation using the same approach described in this thesis for IEG activation.

Future work could also involve the addition of supplementary human and non-human developmental time-course datasets and the analysis of the shifting score together with the wavelet decomposition analysis to achieve a more complete and systematic description of the changes in promoter expression shape at different time points across different biological systems. Future studies of IER dynamics could discover new mechanistic layers underpinning cellular responses to stimuli by directly profiling the transcriptome and proteome in parallel.

8. Appendix

8.1. Appendix analysis

8.1.1. Wavelet representation

Wavelet analysis decomposes signals into component waves called wavelets, subsignals of different resolution levels. Wavelet decomposition has been used in many different fields, predominantly to compress complex signals. Wavelet decomposition can be used to analyse the trend of a complex biological signal at different resolution levels (Shim and Stephens 2015), for example Schor et al. (2017) applied wavelet-based analysis to raw CAGE data signals to assess the effect of tssQTLs on promoter shape changes at individual bases across three embryogenesis stages in 81 Drosophila lines (Schor, Degner et al. 2017).

In this project, I proposed that decomposing the signal resulting from the difference between the cumulative distributions of the CAGE reads at the two time points ($d_{1,2}$) might provide insights on the size of the change, and the spatial scale where any change is most apparent. I was able to analyse the entire window, half of the window and so on to single base resolution and this can help to identify which scale is most affected by the change.
As a simple example (Appendix Figure 4), consider a window of 16 bases centred on a CAGE TSS where a signal of 16 values is obtained by the difference in cumulative reads count at two different time points. In this case the wavelet transform makes use of a 'mean and difference' function which is applied to the values vector to decompose the signal (Appendix Table 7). No information is lost when the bottom row is transformed into the top row (Jensen and la Cour-Harbo 2001). It can generally be assumed that there is some correlation between successive values, and so the difference between them will be smaller than their magnitudes and so the signal will be compressed. The next row of the table is computed as the mean from a pair of values in the

current row (left hand side), and by subtracting the mean to the first value of the pair (differences, in bold in the right side of the table). This procedure goes on until there is only one cell containing the mean value left (in the example = 23.13 in level 0).

The calculation can be reversed from the first row to the last row simply summing the first difference to the first value (23.13 + 9.62 = 32.75) and subtracting the first difference to the first value (23.13-9.62= 13.51) and so on until the original vector of signal values is reconstructed. Other transforms use different approaches to decompose the signal, for example the Haar transform (Appendix Table 8) divides the sum (to compute the trend a) and the difference (to compute the fluctuation f) of each couple of value by $\sqrt{2}$ (i.e. for two values 4 and 6, $a = \frac{4+6}{\sqrt{2}} = 5\sqrt{2}$ and $f = \frac{4-6}{\sqrt{2}} = -\sqrt{2}$), as in the simpler example with means and differences, the Haar calculation is loss-free and can be reversed by summing and subtracting the trend and the fluctuation values for each couple of signal values and dividing by $\sqrt{2}$ (Walker 2008) (i.e. $\frac{5\sqrt{2}-\sqrt{2}}{\sqrt{2}} = 4$ and $\frac{5\sqrt{2}+\sqrt{2}}{\sqrt{2}} = 6$). The method is described in Walker et al. (2008) and in Appendix Table 8 I applied it to an exemplificative series of 8 values (4, 6, 28, 24, 48, 48, 52).

In addition, the Haar transform and many other transforms have the important property that the energy ε_s of the signal s, the sum of the squares of the data points, is conserved (Walker 2008):

$$\varepsilon_s = s_1^2 + s_2^2 + \cdots + s_N^2 = \varepsilon_{a^1} + \varepsilon_{f^1} = a_1^2 + a_2^2 + \cdots + a_N^2 + f_1^2 + f_2^2 + \cdots + f_N^2$$

In the example in Appendix Table 8:

$\varepsilon_s = 4^2 + 6^2 + 28^2 + 24^2 + 48^2 + 48^2 + 56^2 + 52^2 = 11860$

and $\varepsilon_{a^1} + \varepsilon_{f^1} = 5^2*2 + 26^2*2 + 48^2*2 + 54^2*2 + 2 + 2^2*2 + 0*2 + 2^2*2 = 11860$

This means that I can compute the contribution of the trend $\frac{\varepsilon_{a^1}}{\varepsilon_s}$ and the fluctuation $\frac{\varepsilon_{f^1}}{\varepsilon_s}$ values to the total energy for each level of decomposition. $\frac{\varepsilon_{a^1}}{\varepsilon_s}$ will be bigger than $\frac{\varepsilon_{f^1}}{\varepsilon_s}$ every time that the magnitudes of the fluctuation values

is smaller than the magnitude of the trend values (Walker 2008). In our example $\frac{\varepsilon_{a1}}{\varepsilon_s} = \frac{11{,}842}{11{,}860} = 0.998$, which means that the energy of the trend at the first level of decomposition accounts for almost total energy of the transformed energy.

In this work, shifting and non-shifting promoters were compared testing for discrete wavelet transform (DWT) of $d_{1,2}$, the difference between the cumulative distributions of CAGE tags in pairs of time points, d_1 and d_2:

$$d_{1,2} = \frac{d_1}{\max(d_1)} - \frac{d_2}{\max(d_2)}$$

$$WTe = DWT(d_{1,2})$$

I used the *wd* R function (from the package 'wavethresh', version 3.5.3), which makes use of the Mallat Pyramidal Algorithm to compute the DWT coefficients (Nason 1993, Li, Li et al. 1997), with standard parameters.

Next, I considered the wavelet energy to constitute a distribution and computed the Jensen-Shannon (JS) divergence in distributions between time points (i.e. 0 to 45 minutes).

The Jensen-Shannon divergence quantifies how dissimilar are two distributions. It is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence (Lin 1991).

After computing the KL divergence for each couple of comparison as:

$$KL_1 = \sum p * log\frac{p}{q}$$

$$KL_2 = \sum q * log\frac{q}{p}$$

The JS divergence is computed as:

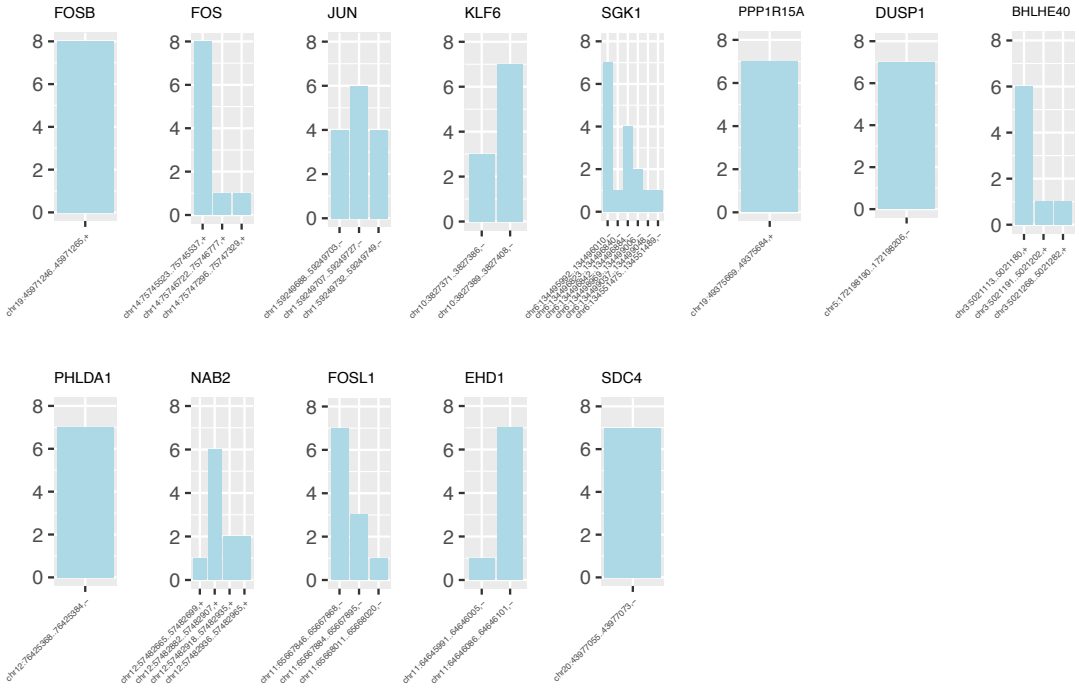
$$JS = \sqrt{\frac{KL_1}{2} + \frac{KL_2}{2}}$$

Where p is the normalized WTe for the first compared time points (i.e. 0 and 15) and q is the normalized WTe for the second compared time points (i.e. 00 and 30).

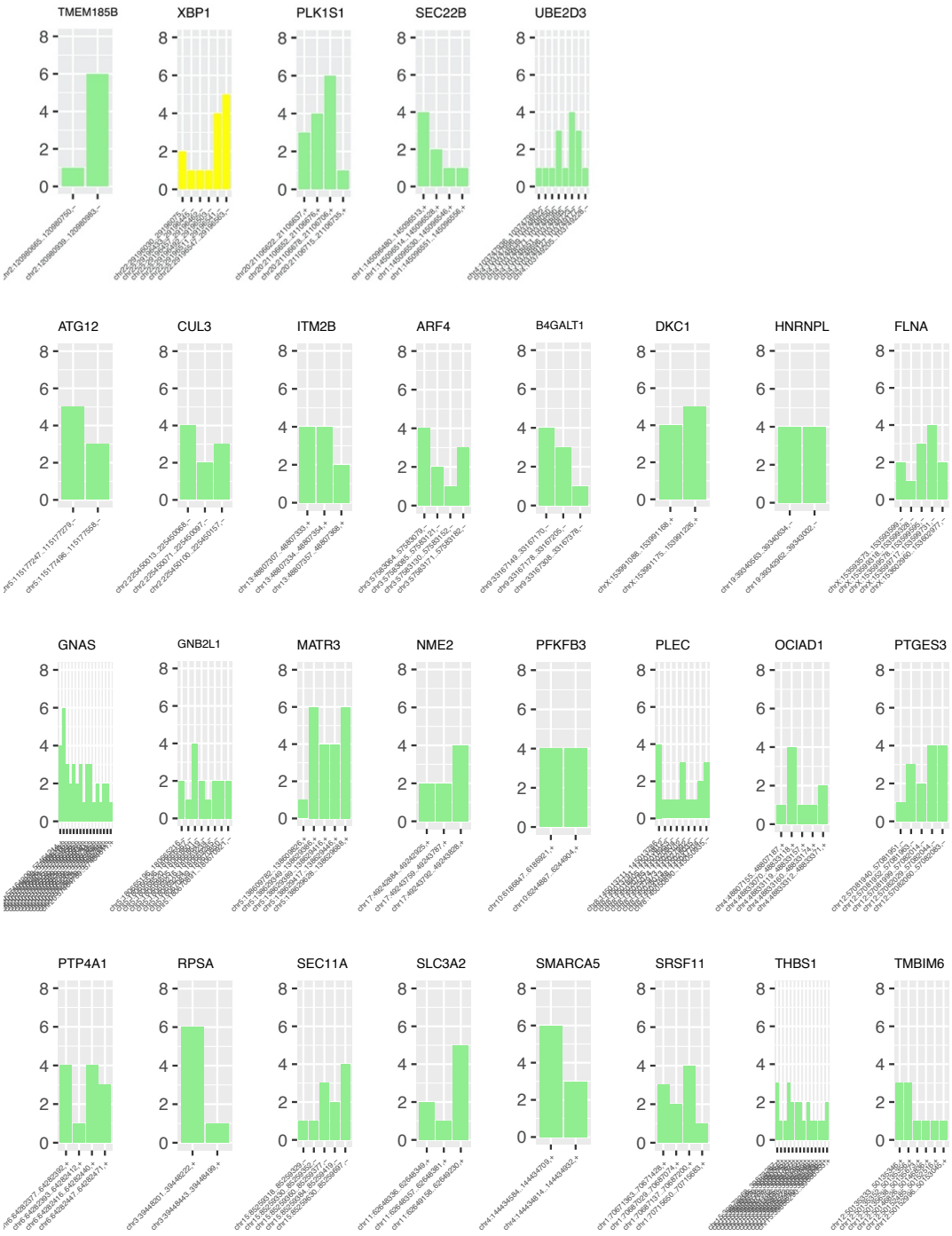
To evaluate the effect size of the change in CAGE tag distribution across time points by an alternative method, wavelet signal decomposition and Jensen Shannon distance (JS) between energy distributions were calculated. In the first analysis (Appendix Figure 5 and Appendix Figure 6), the wavelet transform is applied to the cumulative counts at two time points and the transforms are examined. Whereas wavelet energy and signal, obtained by decomposing the CAGE counts cumulative distribution of the non-shifting promoter are substantially unchanged between 0 minutes and 45 minutes (Appendix Figure 5, JS = 0.04), they are significantly different for the shifting promoter (Appendix Figure 6, JS = 0.69).

In a second analysis, the counts that result from subtracting the cumulative counts at 0 minutes from the cumulative counts at 45 minutes are used as the input to wavelet analysis (Appendix Figure 7). This is comparable to the calculation of the shifting score.
The derived wavelet energies are higher for the shifting promoter (average cumulative counts= 0.32, average wavelet energy= 0.95, Appendix Figure 7C and D) than for the non-shifting promoter (average cumulative counts = $2.24*10^{-3}$, average wavelet energy= $8.31*10^{-3}$, Appendix Figure 7A and B). These results are consistent with those obtained from the shifting score Haberle et al. (2014) and show the extent of the change in tag distribution at different resolution levels with small ‘high frequency’ changes observed in non-shifting promoters at resolution levels 2 and 3 and large ‘low frequency’ changes in shifting promoters at resolution levels 0 and 1.
I conclude that these preliminary analyses demonstrate the potential of wavelet analysis as a way to understand the local changes in promoter shape.

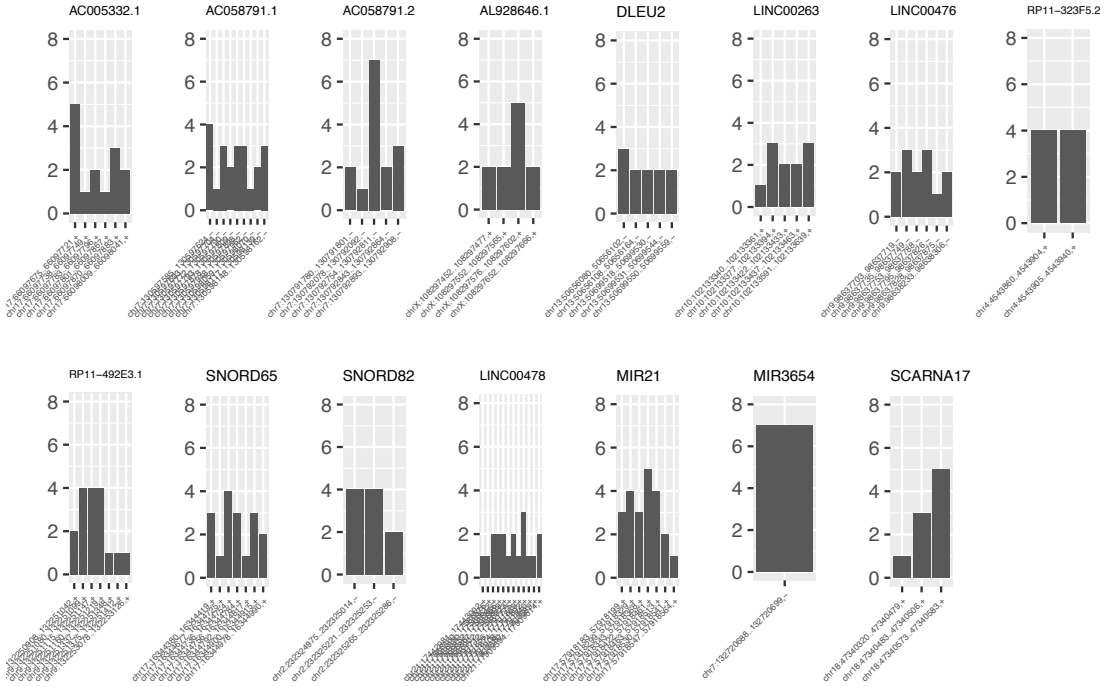
8.2. Appendix Figures



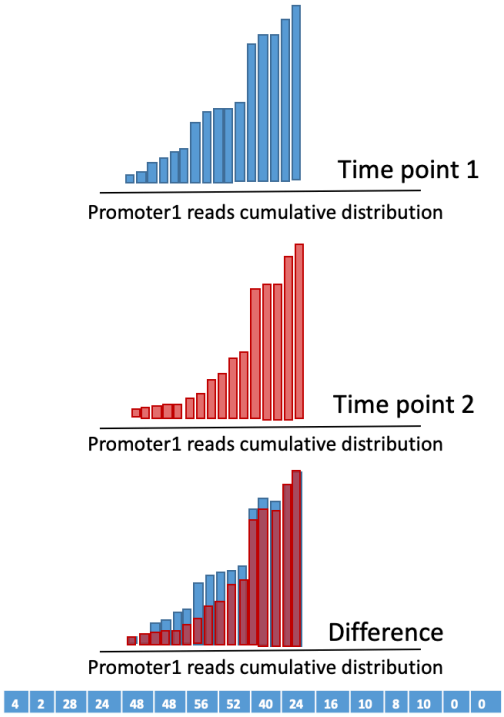
Appendix Figure 1 Promoter choice across time series datasets. For known IEGs, bar charts show the number of datasets where each TSS peaks to illustrate the diversity of TSS choice and commonality of the peaking response.



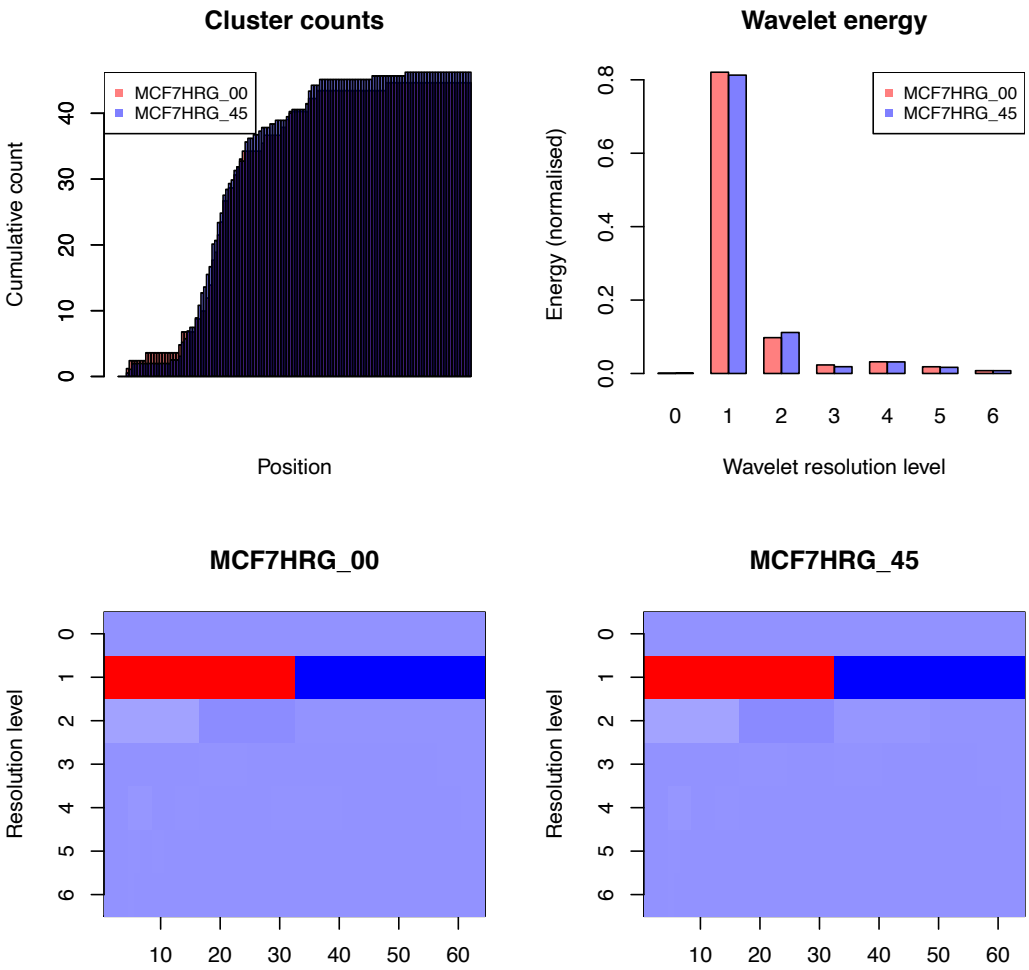
Appendix Figure 2 Promoter choice across time series datasets. For candidate IEGs, bar charts show the number of datasets where each TSS peaks to illustrate the diversity of TSS choice and commonality of the peaking response.



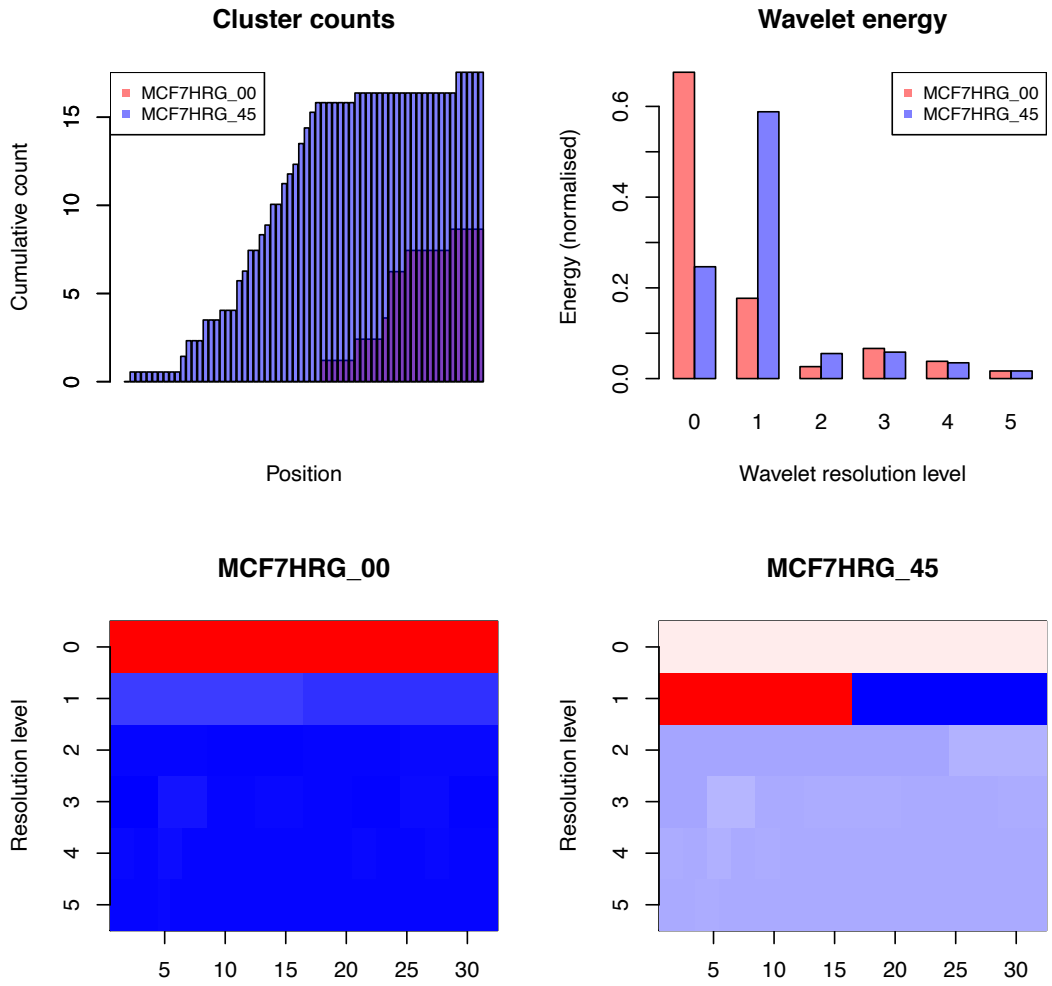
Appendix Figure 3 Promoter choice across time series datasets. For non-coding IEGs, bar charts show the number of datasets where each TSS peaks to illustrate the diversity of TSS choice and commonality of the peaking response.



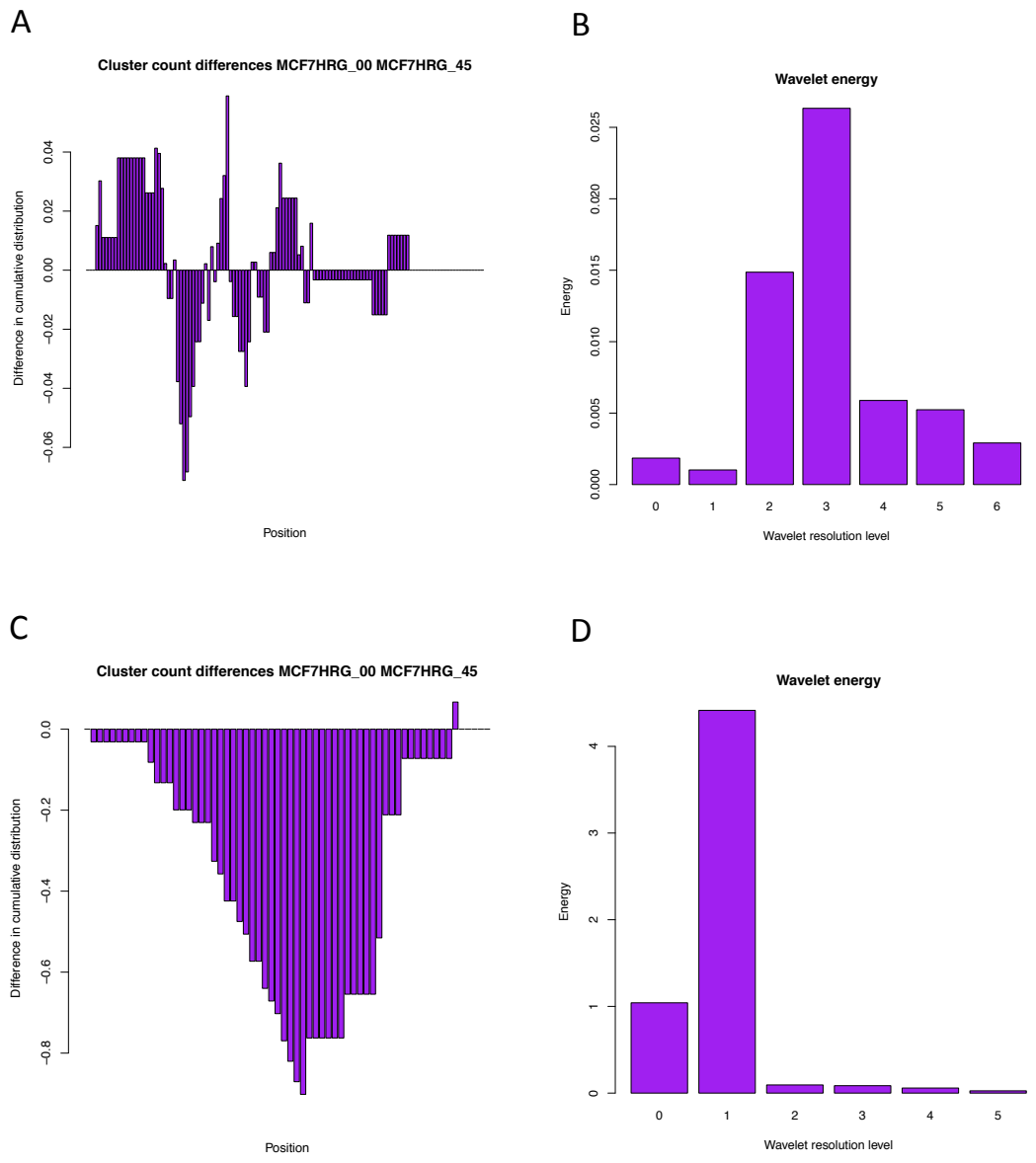
Appendix Figure 4 Change in promoter shape. Distribution of reads for the same promoter at two time points (time point 1 in blue and time point 2 in red), and the raw signal (bottom numerical vector) obtained by subtracting the number of reads at time 2 from the number of reads at time 1.



Appendix Figure 5 Wavelet decomposition of non-shifting promoter. Cumulative counts distribution and wavelet energy at time= 0 (red) and 45 minutes (blue), for MCF7_HRG non-shifting promoter chr1:1334898..1335002,+. Wavelet signals at different resolution levels are represented by coloured cells (dark blue to red for increasing values).



Appendix Figure 6 Wavelet decomposition of shifting promoter. Cumulative counts distribution and wavelet energy at time= 0 (red) and 45 minutes (blue), for MCF7_HRG non shifting promoter chrX:141155504..141155563,-. Wavelet signals at different resolution levels are represented by coloured cells (dark blue to red for increasing values).



Appendix Figure 7 Wavelet analysis of the difference in cumulative counts between time points. Cumulative count differences and energy distributions at increasing wavelet resolution levels for non-shifting and shifting promoters, chr1:1334898..1335002,+ (A and B) and chrX:141155504..141155563,- (C and D).

8.3. Appendix Tables

Appendix Table 1 Known IEGs. List of known IEGs detected in the eight dataset, grouped depending on the classification to the peak model (in at least one out of eight datasets) or other model.

Known IEGs peaking in at least one dataset
VCAM1; CSF1; SLC16A1; MCL1; S100A9; IER5; PTGS2; RGS1; RGS2; BTG2; NUAK2; IL10; ATF3; ID3; CLIC4; ZC3H12A; JUN; KLHL21; PDE4B; GADD45A ; ERRFI1; BCL10; GBP1; ISG15; F3; NFKB2; DUSP5; MAP3K8; CREM; KLF6; ARID5B; EGR2; PLAU; CH25H; IFIT1; BIRC3; ADM; AMPD3; CASP4; ZC3H12C; SPTY2D1; EHD1; CDC42EP2; FOSL1; OLR1; OASL; UBC; NR4A1; GPR84; ITGA5; IL23A; NAB2; MDM2; PHLDA1; SLC2A3; CLEC4E; DUSP6; CD69; IRS2; TSC22D1; TRIM13; SPRY2; TNFAIP2; NFKBIA; FBXO33; ZFP36L1; FOS; BCL2A1; CCL2; CCL7; CCL5; CCL3; CCL4; DUSP14; CSF3; ARL4D ; EIF5A; SOCS3; CBX4; PER1; TGIF1; PMAIP1; ICAM1; JUNB; DNAJB1; KLF2; GDF15; GADD45B; NFKBID; NFKBIB; ZFP36; SERTAD1; PVR; BCL3; RELB; FOSB; PPP1R15A TNFSF9; SLC20A1; IL1A; IL1B; RND3; NR4A2; IFIH1; NFE2L2; CFLAR; RHOB; KLF7; CCL20; FOSL2; ZFP36L2; REL; PELI1; SERTAD2; ID2; DUSP2; SDC4; CEBPB; RCAN1; ETS2; ICOSLG; LIF; MAFF; ATF4; NFKBIZ; SIAH2; TIPARP; CCNL1; SKIL; B3GNT5; HES1; CSRNP; CCRL2; TREX1; BHLHE40; NFKB1; CCRN4L; IL8; CXCL1; CXCL3; CXCL2; CXCL10; RASGEF1B; SRFBP1; CSF2; IRF1; EGR1; HBEGF; CD14; TNIP1; DUSP1; SQSTM1; PTGER4; ELL2; PRDM1; NEDD9; EDN1; SGK1; CITED2; CD83; IER3; TNF; HMGA1; PIM1; SRF; VEGFA; NFKBIE; PNRC1; SERPINE1; IFRD1; MAFK; IL6; KLF10; TRIB1; MYC; EGR3; CEBPD; ZBTB10; OSGIN2; GEM; NR4A3; KLF4; IFNB1; ZFAND5; GADD45G; EPHA2; CYR61; TBX3; FRMD6; IER2; CXCR7; ID1; ADAMTS1; PLK2; CTGF; SELE; FOXC2 ANKRD57; DUSP4; EGR4; FLRT3; SIK1; LMCD1; ARC
Known IEGs assigned to other models
IKBKE; SLC2A1; TOB1; IL12B; IRF4; GS4; C1orf51; BMP2

Appendix Table 2 GO term enrichment analysis. GO terms significantly enriched (FDR corrected hypergeometric test q-value <0.05) in all the genes classified as peak and in known IEGs.

GO term id	**Description**	**FDR q-value** (peaking genes)	**Enrichent (peaking genes)**
GO:0044237	cellular metabolic process	0.0000000226	1.03
GO:0008152	metabolic process	0.000000121	1.03
GO:0051704	multi-organism process	0.00000344	1.11
GO:0009892	negative regulation of metabolic process	0.0000135	1.07
GO:0043620	regulation of DNA-templated transcription in response to stress	0.000021	1.29
GO:0033554	cellular response to stress	0.0000255	1.09
GO:0071704	organic substance metabolic process	0.0000302	1.03
GO:0010605	negative regulation of macromolecule metabolic process	0.0000381	1.07
GO:0043170	macromolecule metabolic process	0.0000451	1.03
GO:0006807	nitrogen compound metabolic process	0.0000459	1.03
GO:0044260	cellular macromolecule metabolic process	0.0000497	1.03

GO:0043618	regulation of transcription from RNA polymerase II promoter in response to stress	0.0000517	1.29
GO:0009987	cellular process	0.0000883	1.01
GO:0051716	cellular response to stimulus	0.000227	1.06
GO:0010629	negative regulation of gene expression	0.000278	1.08
GO:0032268	regulation of cellular protein metabolic process	0.00045	1.06
GO:0042221	response to chemical	0.000533	1.06
GO:0044238	primary metabolic process	0.000538	1.02
GO:0010604	positive regulation of macromolecule metabolic process	0.000961	1.05
GO:0051246	regulation of protein metabolic process	0.0011	1.06
GO:0007568	aging	0.0012	1.19
GO:0051726	regulation of cell cycle	0.00151	1.08
GO:0010628	positive regulation of gene expression	0.00164	1.07
GO:0009893	positive regulation of metabolic process	0.0029	1.05
GO:0045862	positive regulation of proteolysis	0.00296	1.14
GO:0008150	Biological process	0.003	1.01
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.0031	1.06
GO:0031324	negative regulation of cellular metabolic process	0.00374	1.05
GO:0006950	response to stress	0.00374	1.05
GO:0048585	negative regulation of response to stimulus	0.00386	1.07
GO:1901698	response to nitrogen compound	0.00482	1.10
GO:0030162	regulation of proteolysis	0.00481	1.10
GO:0010033	response to organic substance	0.00577	1.06
GO:0006979	response to oxidative stress	0.00584	1.14
GO:0045321	leukocyte activation	0.00806	1.09
GO:0048522	positive regulation of cellular process	0.00814	1.03
GO:0042542	response to hydrogen peroxide	0.00822	1.23
GO:1901360	organic cyclic compound metabolic process	0.00834	1.03
GO:0023057	negative regulation of signaling	0.00871	1.07
GO:0009968	negative regulation of signal transduction	0.00943	1.08
GO:0045787	positive regulation of cell cycle	0.00983	1.13
GO:0010648	negative regulation of cell communication	0.00978	1.07
GO:0043900	regulation of multi-organism process	0.0104	1.13
GO:0051172	negative regulation of nitrogen compound metabolic process	0.0109	1.05
GO:0010035	response to inorganic substance	0.0115	1.12
GO:0050896	response to stimulus	0.0117	1.03
GO:0031325	positive regulation of cellular metabolic process	0.012	1.04
GO:1904951	positive regulation of establishment of protein localization	0.0121	1.11
GO:0080134	regulation of response to stress	0.0124	1.07
GO:2001234	negative regulation of apoptotic signaling pathway	0.0124	1.17
GO:0010557	positive regulation of macromolecule biosynthetic process	0.0126	1.06
GO:0043069	negative regulation of programmed cell death	0.0145	1.09
GO:2001233	regulation of apoptotic signaling pathway	0.0156	1.12
GO:0051592	response to calcium ion	0.0159	1.24
GO:1901700	response to oxygen-containing compound	0.0163	1.07
GO:0043066	negative regulation of apoptotic process	0.0165	1.09
GO:0016070	RNA metabolic process	0.0165	1.04
GO:0019222	regulation of metabolic process	0.0164	1.03
GO:0010243	response to organonitrogen compound	0.0164	1.09

GO:0010468	regulation of gene expression	0.0168	1.03
GO:0051173	positive regulation of nitrogen compound metabolic process	0.0173	1.04
GO:0000302	response to reactive oxygen species	0.0176	1.18
GO:0006915	apoptotic process	0.0176	1.10
GO:0043067	regulation of programmed cell death	0.0181	1.06
GO:0042981	regulation of apoptotic process	0.0191	1.06
GO:0034599	cellular response to oxidative stress	0.0197	1.16
GO:0051222	positive regulation of protein transport	0.0234	1.11
GO:0001775	cell activation	0.0235	1.08
GO:0006725	cellular aromatic compound metabolic process	0.025	1.03
GO:0032269	negative regulation of cellular protein metabolic process	0.0249	1.07
GO:0038061	NIK/NF-kappaB signaling	0.0252	1.26
GO:0060548	negative regulation of cell death	0.0263	1.08
GO:0048519	negative regulation of biological process	0.0272	1.03
GO:0046483	heterocycle metabolic process	0.0297	1.03
GO:0033138	positive regulation of peptidyl-serine phosphorylation	0.0339	1.24
GO:0010941	regulation of cell death	0.0339	1.06
GO:0060255	regulation of macromolecule metabolic process	0.0353	1.02
GO:0051248	negative regulation of protein metabolic process	0.0363	1.07
GO:0048518	positive regulation of biological process	0.0371	1.03
GO:0006139	nucleobase-containing compound metabolic process	0.0392	1.03
GO:0051247	positive regulation of protein metabolic process	0.0394	1.05
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	0.04	1.15
GO:2001236	regulation of extrinsic apoptotic signaling pathway	0.0401	1.17
GO:0007623	circadian rhythm	0.0425	1.20
GO:0009891	positive regulation of biosynthetic process	0.045	1.05
GO:0044092	negative regulation of molecular function	0.0451	1.07
GO:1901576	organic substance biosynthetic process	0.047	1.03
GO:0032270	positive regulation of cellular protein metabolic process	0.0471	1.06

Appendix Table 3 GO terms enrichment for known IEGs. GO terms significantly enriched (FDR corrected hypergeometric test q-value <0.05) in the known IEGs.

GO Term									
GO:0048518	GO:0050793	GO:0051094	GO:0001934	GO:0044763	GO:0050727	GO:0045639	GO:0032946	GO:0001704	GO:0042990
GO:0050896	GO:0009889	GO:0030334	GO:0010648	GO:0006366	GO:0045936	GO:0048525	GO:0043407	GO:0050870	GO:1901655
GO:0048522	GO:0010557	GO:0009890	GO:0023057	GO:0006928	GO:0010563	GO:0043085	GO:0031668	GO:0090022	GO:0051591
GO:0010033	GO:0045893	GO:0019220	GO:0071219	GO:0032103	GO:1903708	GO:0043434	GO:0000302	GO:0003006	GO:0071548
GO:0050794	GO:1903508	GO:0010558	GO:0042327	GO:1901342	GO:0012501	GO:0001666	GO:0036499	GO:0007267	GO:0043331
GO:0007165	GO:1902680	GO:0031399	GO:0040011	GO:0006351	GO:0008219	GO:0007623	GO:0008283	GO:0006807	GO:0002526
GO:0048519	GO:0031326	GO:0019221	GO:0010562	GO:0097659	GO:0043901	GO:0070482	GO:1904407	GO:0071214	GO:0042345
GO:0050789	GO:0002376	GO:0060326	GO:0045937	GO:0050790	GO:0001775	GO:0042129	GO:0045429	GO:1901623	GO:0030217
GO:0010604	GO:0045935	GO:0009719	GO:0016477	GO:0022407	GO:0071417	GO:0045321	GO:0051049	GO:0051385	GO:0033138
GO:0042221	GO:0014070	GO:0051174	GO:0010942	GO:0001933	GO:0043410	GO:0044260	GO:0009058	GO:0071674	GO:0050792
GO:0006950	GO:0031328	GO:0071407	GO:0009725	GO:0044699	GO:1990266	GO:0051726	GO:0001525	GO:0044702	GO:0044238
GO:0009605	GO:0048584	GO:0031327	GO:0048870	GO:0043409	GO:0032944	GO:0045859	GO:0051348	GO:0042542	GO:0045580
GO:0051173	GO:1903506	GO:0030335	GO:0002685	GO:0042326	GO:0071621	GO:1900745	GO:0048729	GO:0010038	GO:0034405
GO:0031325	GO:0051252	GO:2000145	GO:0009968	GO:0007154	GO:1901653	GO:0008285	GO:0071456	GO:0032680	GO:0043534
GO:0009893	GO:0009966	GO:0051270	GO:0002687	GO:1901699	GO:0071774	GO:0045596	GO:0070665	GO:2001234	GO:2000403
GO:0010941	GO:0051240	GO:0051090	GO:1903706	GO:0032774	GO:1902107	GO:0009653	GO:0043281	GO:0032102	GO:0006979
GO:0065007	GO:0060548	GO:0051241	GO:0045597	GO:0048511	GO:0070663	GO:1901654	GO:0000165	GO:0023014	GO:0007611
GO:0043207	GO:0006355	GO:2000147	GO:0030155	GO:0044092	GO:0001763	GO:1903039	GO:0070374	GO:0050671	GO:0071704
GO:0042981	GO:0042127	GO:0009628	GO:0045934	GO:0002694	GO:0097530	GO:2000501	GO:0050729	GO:0050776	GO:0045087
GO:0043067	GO:0044767	GO:0051247	GO:0070098	GO:1901652	GO:0043170	GO:1900744	GO:2001236	GO:0071375	GO:0090304
GO:0051171	GO:0065009	GO:0040012	GO:0071216	GO:0006915	GO:0097529	GO:0043618	GO:0042110	GO:0032649	GO:0048754
GO:0009607	GO:0001817	GO:0051272	GO:0071356	GO:0030595	GO:0045637	GO:0007249	GO:0071622	GO:0007186	GO:0042531
GO:0051716	GO:0032502	GO:0044707	GO:0033554	GO:0050865	GO:0009991	GO:0048513	GO:0022414	GO:0050731	GO:0001706
GO:0006952	GO:0001932	GO:0048585	GO:0001819	GO:0051707	GO:0070373	GO:1902624	GO:0031667	GO:0050707	GO:0045598
GO:0051239	GO:0010646	GO:0035556	GO:0048856	GO:0050921	GO:0045766	GO:0033673	GO:0090023	GO:0090257	GO:1904035
GO:0080090	GO:0043066	GO:0071495	GO:0071347	GO:0071496	GO:0044057	GO:0071383	GO:0000188	GO:2000503	GO:0060426
GO:0002682	GO:2000112	GO:0008284	GO:0031960	GO:0007568	GO:0050867	GO:2000377	GO:0045619	GO:0043435	GO:0052548
GO:0071310	GO:0009892	GO:0040017	GO:0042035	GO:0002521	GO:0022409	GO:0009615	GO:0046683	GO:0071376	GO:0030856
GO:0060255	GO:0010605	GO:2000113	GO:0043065	GO:0051249	GO:0071385	GO:0008150	GO:2001233	GO:0042536	GO:0030278
GO:0002237	GO:0023051	GO:0070372	GO:0031401	GO:0034654	GO:0036293	GO:0051093	GO:0043903	GO:1903555	GO:0002573
GO:0048523	GO:0031324	GO:0009967	GO:0070848	GO:0050900	GO:0051091	GO:0043900	GO:0044249	GO:0071453	GO:0032675
GO:0031323	GO:0043069	GO:0032501	GO:0043068	GO:0009059	GO:2000379	GO:0043620	GO:0042752	GO:1904892	GO:0071364
GO:0048583	GO:0019219	GO:1901698	GO:0032870	GO:0050863	GO:0044271	GO:0043086	GO:0032479	GO:0046425	GO:1903524
GO:0032496	GO:0051246	GO:0071345	GO:0022603	GO:0031347	GO:0051704	GO:0009314	GO:1903707	GO:0002286	GO:0032642
GO:0045944	GO:0042325	GO:0071222	GO:0002688	GO:0050920	GO:1903426	GO:0048660	GO:0009888	GO:0071346	GO:0042102
GO:0070887	GO:0051172	GO:0010629	GO:1903037	GO:0045765	GO:0031400	GO:1902622	GO:0043433	GO:0050679	GO:0032897
GO:0051254	GO:0030154	GO:0034612	GO:0010243	GO:0002683	GO:0009617	GO:0051251	GO:0051607	GO:0032677	GO:0071900
GO:1901700	GO:0048869	GO:0080134	GO:0051248	GO:0019438	GO:0050670	GO:0051338	GO:0071624	GO:0050678	GO:0002042
GO:0006357	GO:1901701	GO:0000122	GO:0051384	GO:0009612	GO:0042742	GO:0006469	GO:0002548	GO:0080135	GO:0061158
GO:0006954	GO:0010468	GO:0032270	GO:0044093	GO:1904018	GO:0042036	GO:0007159	GO:0036003	GO:0043123	GO:0035994
GO:1902531	GO:0045595	GO:0009987	GO:1902532	GO:0043122	GO:0061138	GO:0048661	GO:0002252	GO:0045787	GO:0031669
GO:0002684	GO:0043408	GO:0071396	GO:0048646	GO:0034645	GO:0002761	GO:0065008	GO:0043405	GO:0051336	GO:0034614
GO:0034097	GO:0006955	GO:1902679	GO:1902105	GO:0018130	GO:0071384	GO:1901576	GO:0042108	GO:0044700	GO:0032663
GO:0010628	GO:0006935	GO:1903507	GO:0032879	GO:0071363	GO:0045428	GO:0002009	GO:1903428	GO:0051147	GO:0050730
GO:0033993	GO:0042330	GO:0010647	GO:0002690	GO:0098542	GO:0002763	GO:0014074	GO:0016070	GO:0051092	GO:0048534
GO:0010556	GO:1902533	GO:0051253	GO:0032269	GO:0030593	GO:0031349	GO:0042493	GO:0036294	GO:0023052	GO:0061014
GO:0019222	GO:0007166	GO:0032101	GO:0070555	GO:0010035	GO:0043549	GO:0051128	GO:0046649	GO:0048771	GO:0035914
GO:0009891	GO:0032268	GO:0023056	GO:0001818	GO:1901362	GO:0044344	GO:0051050	GO:0035239	GO:1901214	GO:0001570
GO:2001141	GO:2000026	GO:0045892	GO:0048545	GO:0045785	GO:0002696	GO:0030098	GO:2000116	GO:2000401	GO:0007610
GO:2000117	GO:0045862	GO:0046632	GO:0048844	GO:0048514	GO:0051924	GO:0042832	GO:0031663	GO:0048245	GO:0032729
GO:0030855	GO:0006937	GO:0038061	GO:0030949	GO:0002819	GO:0003013	GO:0048745	GO:0071409	GO:0060708	GO:0022408
GO:0071470	GO:1905952	GO:0043523	GO:1990440	GO:0001959	GO:0009267	GO:0097305	GO:0034242	GO:0031622	GO:0070304
GO:0032743	GO:1903532	GO:0002218	GO:0033598	GO:0051347	GO:0010656	GO:0030336	GO:0030223	GO:0051046	GO:0009266
GO:0045682	GO:1903313	GO:0002221	GO:0045621	GO:0002697	GO:0051897	GO:0048468	GO:0045415	GO:0009408	GO:0050714
GO:0001101	GO:0016337	GO:1900180	GO:1903531	GO:0031175	GO:0006959	GO:0044089	GO:0032765	GO:0010332	GO:0001938
GO:0009416	GO:0045667	GO:0002793	GO:0045786	GO:0046822	GO:1901216	GO:0051153	GO:0031077	GO:1901722	GO:1904589
GO:0010660	GO:0002711	GO:0002718	GO:0001569	GO:0050708	GO:0050728	GO:0031334	GO:0033141	GO:0072216	GO:0000288
GO:1903900	GO:0008015	GO:0043124	GO:2001240	GO:0040013	GO:0010744	GO:1904036	GO:2000097	GO:0060707	GO:0045076
GO:0009743	GO:0050691	GO:0045604	GO:1901099	GO:0043542	GO:0043922	GO:0048247	GO:0072677	GO:0031098	GO:0042534
GO:0001660	GO:0040007	GO:0002268	GO:0007507	GO:0007569	GO:0050710	GO:0060670	GO:0051412	GO:0009648	GO:0007179
GO:1904582	GO:0030099	GO:0010611	GO:0040008	GO:0090049	GO:0090183	GO:0035767	GO:0090050	GO:0050878	GO:1901360
GO:1904580	GO:0097193	GO:0003151	GO:0001709	GO:0060142	GO:1904951	GO:0010959	GO:0050868	GO:0043535	GO:0033674
GO:0061394	GO:2000027	GO:0072676	GO:0003281	GO:0002285	GO:0007613	GO:0061469	GO:0051051	GO:0051896	GO:0006984
GO:0045362	GO:1901740	GO:0051222	GO:0061157	GO:0008016	GO:0007612	GO:0045657	GO:0006953	GO:0008152	GO:0042149
GO:0050848	GO:0045360	GO:0002266	GO:0070301	GO:0002709	GO:0045651	GO:0045414	GO:0043153	GO:0090287	GO:0044237
GO:0051345	GO:0032682	GO:0045860	GO:0045600	GO:0002831	GO:0051250	GO:0042594	GO:0007178	GO:0090303	GO:0097084

GO:0001944	GO:0032868	GO:0009611	GO:0060337	GO:0045655	GO:0034341	GO:0003008	GO:0035690	GO:0048731	GO:0097012
GO:0045933	GO:1902041	GO:0061013	GO:0034599	GO:1903034	GO:0032652	GO:0002292	GO:0046483	GO:0006725	GO:0097011
GO:0007167	GO:0045071	GO:1903670	GO:0006468	GO:0007157	GO:0010595	GO:0071359	GO:0050866	GO:0050778	GO:0045073
GO:0045672	GO:0044708	GO:0050715	GO:0033135	GO:0001568	GO:0043279	GO:0045823	GO:0045069	GO:0071499	GO:0010830
GO:0070302	GO:0010575	GO:0050779	GO:0032874	GO:0007275	GO:0050864	GO:0071498	GO:0031086	GO:0014743	GO:0032922
GO:0097190	GO:0002764	GO:0002758	GO:0042306	GO:0043154	GO:0010212	GO:0070371	GO:0032872	GO:1903036	
GO:1903038	GO:0045670	GO:1900151	GO:0070849	GO:0060759	GO:1903522	GO:0030162	GO:1904627	GO:0022610	
GO:0071277	GO:0051592	GO:1900153	GO:0071549	GO:0007155	GO:0010634	GO:0052547	GO:1901724	GO:2001237	
GO:0046427	GO:0043269	GO:0097201	GO:0010574	GO:0051403	GO:0002695	GO:1904628	GO:0006139	GO:0090184	
GO:1901739	GO:0072678	GO:1903672	GO:0051130	GO:1903901	GO:0045089	GO:0043491	GO:0042509	GO:0045088	
GO:0070542	GO:1902106	GO:0035924	GO:0009649	GO:0019722	GO:0045444	GO:0060674	GO:0071901	GO:0010632	
GO:0042592	GO:0045165	GO:0050890	GO:0071260	GO:1904894	GO:0010594	GO:1901215	GO:0032480	GO:0043502	

Appendix Table 4 TF binding sites significantly occurring in IEG promoters.

TFAP2A; Klf4; MZF1(var.2); SP1; EWSR1-FLI1; EGR1; MIG2; MIG3; NHP10; RSC30; STP1; SUT1; E2F4; Myod1; SP2; ZNF263; Mad; ERF1B; KLF5; TFAP2B(var.3) ; TFAP2C(var.3) TFAP2A(var.3); CRF2; CRF4; ERF008; RAP2-10; ERF4; ERF7; ERF8; ERF069; ERF096; ERF098; ERF11; ERF112; ERF13; ERF3; Os05g0497200; ERF094; RAP2-3; RAP2-6; ERF109; daf-12; Zfx; PLAG1; MIG1; RDS1; UGA3; YGR067C; btd; E2F6; SP4; EGR3; EGR4; KLF16; SP3; TFAP2A(var.2); KLF14; SP8; NRF1; ERF105; ERF6; REST; RREB1; abi4; brk; RSC3; STP2; SOC1; CDC5; SRF; CTCF; Klf12; Tcfl5; PAX5; eor-1; IRF1; PIF1; Trl; IRF2; STAT1::STAT2; TFAP2C; TFAP2B; Gabpa; ELK4; SPI1; ZNF740; h; FUS3; opa; INSM1;

Appendix Table 5 Shifting promoters for H1_CD34. List of promoters with shifting score >=0.6 and FDR K-S <0.01. The dominant TSS is the TSS with higher number of tags (expressed in TPM) and its location with TPM counts are reported for the compared time points.

CAGE cluster	Shifting score	Dominant TSS (Day 0)	Dominant TSS (Day 3)	Dominant TSS TPM (day 0)	Dominant TSS TPM (day 3)	FDR KS
chr11:64536521..64536541,-	0.66	64536531	64536527	14.4	5.4	0.0014
chr14:43088147..43088201,-	0.64	43088153	43088166	12.8	15.5	0.0097
chr19:19050980..19051133,+	0.60	19051124	19051037	9.2	20.9	0.0007
chr3:14530430..14530508,+	0.83	14530433	14530508	5.4	24.7	0.002
chr6:27806485..27806592,+	0.95	27806486	27806587	8.7	8.6	0.0004
chr6:133136125..133136157,+	0.62	133136156	133136156	22.6	23.9	9.5E-05
chr6:15663151..15663311,-	0.63	15663310	15663211	16.1	17.2	0.002
CAGE cluster	**Shifting score**	**Dominant TSS (Day 0)**	**Dominant TSS (Day 9)**	**Dominant TSS TPM (Day 0)**	**Dominant TSS TPM (day 3)**	**FDR KS**
chr4:111120306..111120363,-	0.68	111120307	111120353	17.3	14.6	0.003

Appendix Table 6 Shifting promoters for MCF7_HRG. List of promoters with shifting score >=0.6 and FDR K-S <0.01. The dominant TSS is the TSS with higher number of tags (expressed in TPM) and its location with TPM counts are reported for the compared time points.

CAGE cluster	Shifting score	Dominant TSS (0 min)	Dominant TSS (15 min)	Dominant TSS TPM (0 min)	Dominant TSS TPM (15 min)	FDR KS
chr4:176770514..176770525,+	0.86	176770523	176770524	13.4	6.1	1.2E-4
CAGE cluster	**Shifting score**	**Dominant TSS (0 min)**	**Dominant TSS (30 min)**	**Dominant TSS TPM (0 min)**	**Dominant TSS TPM (30 min)**	**FDR KS**
chrX:141155504..141155563,-	0.91	141155551	141155534	8.6	14.8	0.007
CAGE cluster	**Shifting score**	**Dominant TSS (0 min)**	**Dominant TSS (45 min)**	**Dominant TSS TPM (0 min)**	**Dominant TSS TPM (45 min)**	**FDR KS**
chrX:141155504..141155563,-	0.80	141155551	141155524	8.6	17.6	0.003
chr17:61911871..61911940,-	0.92	61911887	61911835	10.4	6.7	0.0096
CAGE cluster	**Shifting score**	**Dominant TSS (0 min)**	**Dominant TSS (60 min)**	**Dominant TSS TPM (0 min)**	**Dominant TSS TPM (60 min)**	**FDR KS**
chrX:141155504..141155563,-	0.80	141155551	141155529	8.6	16.9	0.002
CAGE cluster	**Shifting score**	**Dominant TSS (0 min)**	**Dominant TSS (240 min)**	**Dominant TSS TPM (0 min)**	**Dominant TSS TPM (360 min)**	**FDR KS**
chr2:74776000..74776263,+	0.61	74776243	74776243	25.4	26.9	0.001
CAGE cluster	**Shifting score**	**Dominant TSS (0 min)**	**Dominant TSS (360 min)**	**Dominant TSS TPM (0 min)**	**Dominant TSS TPM (360 min)**	**FDR KS**
chr2:74776000..74776263,+	0.73	74776243	74776232	25.4	21.2	0.005

Appendix Table 7 Wavelet decomposition. Mean and difference (in boldface type) computation, beginning with the data in the top row, and transformations from levels 1 to 4.

Level 0	23.1	**9.6**	**-18.3**	**9**	**-11.5**	**-3**	**9.5**	**4.5**	**1**	**2**	**0**	**2**	**8**	**3**	**-1**	**0**
Level 1	32.8	13.5	**-18.3**	**9**	**-11.5**	**-3**	**9.5**	**4.5**	**1**	**2**	**0**	**2**	**8**	**3**	**-1**	**0**
Level 2	14.5	51	22.5	4.5	**-11.5**	**-3**	**9.5**	**4.5**	**1**	**2**	**0**	**2**	**8**	**3**	**-1**	**0**
Level 3	3	26	48	54	32	13	9	0	**1**	**2**	**0**	**2**	**8**	**3**	**-1**	**0**
Data	4	2	28	24	48	48	56	52	40	24	16	10	8	10	0	0

Appendix Table 8 Haar transform decomposition. Trend and fluctuation (in boldface type) computation.

Level 0	$20\sqrt{2}$	$\mathbf{-5\sqrt{2}}$	$\mathbf{-10.5\sqrt{2}}$	$\mathbf{-1.5\sqrt{2}}$	$\mathbf{-\sqrt{2}}$	$\mathbf{2\sqrt{2}}$	**0**	$\mathbf{2\sqrt{2}}$
Level 1	$15.5\sqrt{2}$	$25.5\sqrt{2}$	$\mathbf{-10.5\sqrt{2}}$	$\mathbf{-1.5\sqrt{2}}$	$\mathbf{-\sqrt{2}}$	$\mathbf{2\sqrt{2}}$	**0**	$\mathbf{2\sqrt{2}}$
Level 2	$5\sqrt{2}$	$26\sqrt{2}$	$48\sqrt{2}$	$54\sqrt{2}$	$\mathbf{-\sqrt{2}}$	$\mathbf{2\sqrt{2}}$	**0**	$\mathbf{2\sqrt{2}}$
Data	4	6	28	24	48	48	56	52

8.4. Publications

Some materials contained in this thesis have previously been published in:

Vacca, A., et al. (2018). “Conserved temporal ordering of promoter activation implicates common mechanisms governing the immediate early response across cell types and stimuli.” Open biology, **8**(8), 180011.

OPEN BIOLOGY

rsob.royalsocietypublishing.org

Research

Cite this article: Vacca A *et al*. 2018 Conserved temporal ordering of promoter activation implicates common mechanisms governing the immediate early response across cell types and stimuli. *Open Biol.* **8**: 180011. http://dx.doi.org/10.1098/rsob.180011

Received: 22 January 2018
Accepted: 4 July 2018

Subject Area:
bioinformatics/genetics/systems biology

Keywords:
immediate early response, promoter activity, CAGE data, time series analysis

Author for correspondence:
Colin A. Semple
e-mail: colin.semple@igmm.ed.ac.uk

Electronic supplementary material is available online at rs.figshare.com.

THE ROYAL SOCIETY PUBLISHING

Conserved temporal ordering of promoter activation implicates common mechanisms governing the immediate early response across cell types and stimuli

Annalaura Vacca[1], Masayoshi Itoh[2], Hideya Kawaji[3], Erik Arner[4], Timo Lassmann[5], Carsten O. Daub[6], Piero Carninci[4], Alistair R. R. Forrest[7], Yoshihide Hayashizaki[2], the FANTOM Consortium[4], Stuart Aitken[1] and Colin A. Semple[1]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK
[2]RIKEN Preventive Medicine and Diagnosis Innovation Program, 2F Main Research Building, 2-1 Hirosawa, Wako, Japan
[3]RIKEN Advanced Center for Computing and Communication, and [4]RIKEN Center for Life Sciences Technologies, RIKEN Yokohama Campus, Yokohama 230-0045, Japan
[5]Telethon Kids Institute, The University of Western Australia, Roberts Road, Subiaco, Western Australia, Australia
[6]Department of Biosciences and Nutrition, Karolinska Institutet, 141 86 Stockholm, Sweden
[7]Harry Perkins Institute of Medical Research, 6 Verdun Street, Nedlands, Western Australia 6009, Australia

SA, 0000-0003-4867-4568; CAS, 0000-0003-1765-4118

The promoters of immediate early genes (IEGs) are rapidly activated in response to an external stimulus. These genes, also known as primary response genes, have been identified in a range of cell types, under diverse extracellular signals and using varying experimental protocols. Whereas genomic dissection on a case-by-case basis has not resulted in a comprehensive catalogue of IEGs, a rigorous meta-analysis of eight genome-wide FANTOM5 CAGE (cap analysis of gene expression) time course datasets reveals successive waves of promoter activation in IEGs, recapitulating known relationships between cell types and stimuli: we obtain a set of 57 (42 protein-coding) candidate IEGs possessing promoters that consistently drive a rapid but transient increase in expression over time. These genes show significant enrichment for known IEGs reported previously, pathways associated with the immediate early response, and include a number of non-coding RNAs with roles in proliferation and differentiation. Surprisingly, we also find strong conservation of the ordering of activation for these genes, such that 77 pairwise promoter activation orderings are conserved. Using the leverage of comprehensive CAGE time series data across cell types, we also document the extensive alternative promoter usage by such genes, which is likely to have been a barrier to their discovery until now. The common activation ordering of the core set of early-responding genes we identify may indicate conserved underlying regulatory mechanisms. By contrast, the considerably larger number of transiently activated genes that are specific to each cell type and stimulus illustrates the breadth of the primary response.

1. Introduction

Human cells respond to a broad range of extracellular stimuli with a characteristic burst of transcription within minutes at many sites across the genome, known as

the immediate early response (IER). The IER has been observed as an initiating event in many cellular processes, notably during differentiation, in responses to cellular stress and in inflammation. The earliest events in the IER involve the activation of the promoters of a particular set of genes, known as immediate early genes (IEGs). The promoters of IEGs are activated rapidly, and their activation is transient in normal cells [1]. However, IEGs are often dysregulated in cancers where they can become continuously activated; accordingly, some of the best-studied IEGs are known oncogenes [2]. For example, the expression of the FOS proto-oncogene normally peaks within 60 min of a stimulus and subsides after 90 min [3], in contrast with its continuous overexpression in many cancers.

IEGs possess unusually accessible promoters that allow rapid transcriptional activation in response to a stimulus without the requirement of de novo protein synthesis. Various features are thought to discriminate IEGs such as the shorter transcripts they generate and enrichments of certain transcription factor (TF) binding sites at their promoters [4]. However, current knowledge about IEGs is derived mainly from studies of individual genes or pathways, and often considers a specific cell type and stimulus. This means that comparison across studies can be confounded by experimental and technical variation, and a comprehensive catalogue of IEGs remains elusive. There is also controversy about the regulatory mechanisms governing the response of even relatively well-studied IEGs [5]. Beyond the induction of protein-coding IEG promoters, the features and underlying mechanisms of the IER are even less well understood. Some studies have implicated altered patterns of IEG splicing as playing important roles in the IER [6], while others have suggested a prominent role for lncRNA activation [7] and transcribed enhancers [8]. Approximately 20% of known IEGs are TFs, including some of the best characterized: EGR1–EGR4, FOS, FOSB, FOSL1, JUN, JUNB and MYC.

The FANTOM5 cap analysis of gene expression (CAGE) data offer a number of advantages for expression profiling because they are based upon single-molecule sequencing to avoid PCR, digestion and cloning biases. They provide up to single base-pair resolution of transcription start sites (TSSs) and promoter regions, and provide a sensitive, quantitative readout of transcriptional output accounting for the alternative promoters of each gene. The output of individual promoters is not confounded by splicing variation, and many novel lowly expressed transcripts including non-coding RNAs (ncRNAs) can be readily detected (see http://fantom.gsc.riken.jp/5/). CAGE data are thus ideally suited to studying the strong burst of transcription at promoters seen in IERs. FANTOM5 data include eight CAGE time course datasets employing unusually dense sampling at time points within 300 min of stimulation, for a variety of stimuli treating a variety of cell types. These heterogeneous datasets, produced using a common experimental platform, should be fertile ground for novel insights into the IER, but a comprehensive meta-analysis has not been performed until now.

Many previous approaches to time series analysis of expression data have been based upon differential expression between successive time points, or have clustered genes according to the similarity of their expression profiles over time [9]. Both of these approaches present problems for the analysis of CAGE data. Differential expression between time points provides poor sensitivity for lowly expressed transcripts (possessing too few reads to generate significant differences in expression), and presents serious difficulties when comparing expression profiles from datasets with somewhat different sampling points over time. Clustering approaches often rely upon arbitrary thresholds (e.g. based upon cluster size or significant enrichment of functional annotation terms) and, by definition, will miss transcripts that cannot be assigned to a cluster but may nevertheless show dynamics of interest. Hence, we refine a previously successful Bayesian model selection algorithm to classify promoter responses to pre-defined mathematical models [7].

Here, we perform extensive meta-analyses of promoter activity in the human IER, encompassing unusually diverse cell types and stimuli, to rigorously classify IEGs and estimate the core IEG repertoire active across cellular responses. We show that computational classification of the temporal activity patterns of promoters provides a potent basis for meta-analyses across time courses, exposing the combined activity of known IEGs and compelling new IEG candidates in the IEG core repertoire. We also show that the timing of the peak expression of a core set of transiently activated genes has a conserved order. This surprising outcome indicates a previously unidentified regulatory mechanism that is shared among cell types and common to diverse stimuli.

2. Results

We considered eight densely sampled, and well-replicated, FANTOM5 CAGE time course datasets obtained following diverse stimuli: calcification in an osteosarcoma cell line in response to osteocalcin (SAOS2_OST), differentiation of adipose-derived primary mesenchymal stem cells in response to a drug mixture (3-isobutyl-1-methylxanthine, dexamethasone and rosiglitazone) (PMSC_MIX), differentiation of primary lymphatic endothelial cells in response to VEGF (PEC_VEGF), MCF7 breast cancer cell line responses to EGF1 (MCF7_EGF1) and to HRG (MCF7_HRG), primary aortic smooth muscle cells response to IL1b (PAC_IL1B) and FGF2 (PAC_FGF2), and primary monocyte-derived macrophage cells activation in response to LPS (PMDM_LPS). Thus, we included a variety of primary and cell line samples, tracking responses to a range of stimuli: growth factors, hormones, drugs, pro-inflammatory cytokines and bacterial endotoxin (figure 1*a*). These diverse data provided a potent resource to discover core features of the IER conserved across cell types and stimuli. All TSSs for protein-coding transcripts were represented by conservatively selected CAGE read clusters (at least 10 TPM) following Arner *et al*. [10]. As expected, the responses of known IEGs often showed characteristic expression peaks early in the time series datasets—as exemplified by FOS and JUN—though even for these well-established IEGs, we observed substantial variation in the magnitude, timing and duration of peaks across cell types and stimuli (figure 1*b*). These observations illustrate the challenges presented in IEG detection, even when studying known IEGs using a uniform experimental platform.

Optimizing and refining the approach developed by Aitken *et al*. [7] (see Material and methods), we defined four mathematical models representing archetypical expression profiles of interest over time—peak, linear, dip and decay (electronic supplementary material, figure S1)—and assessed the fit of each model to the expression profile of each gene using nested sampling to compute the marginal likelihood, log Z [7]. Where sufficient evidence exists (given the variation between

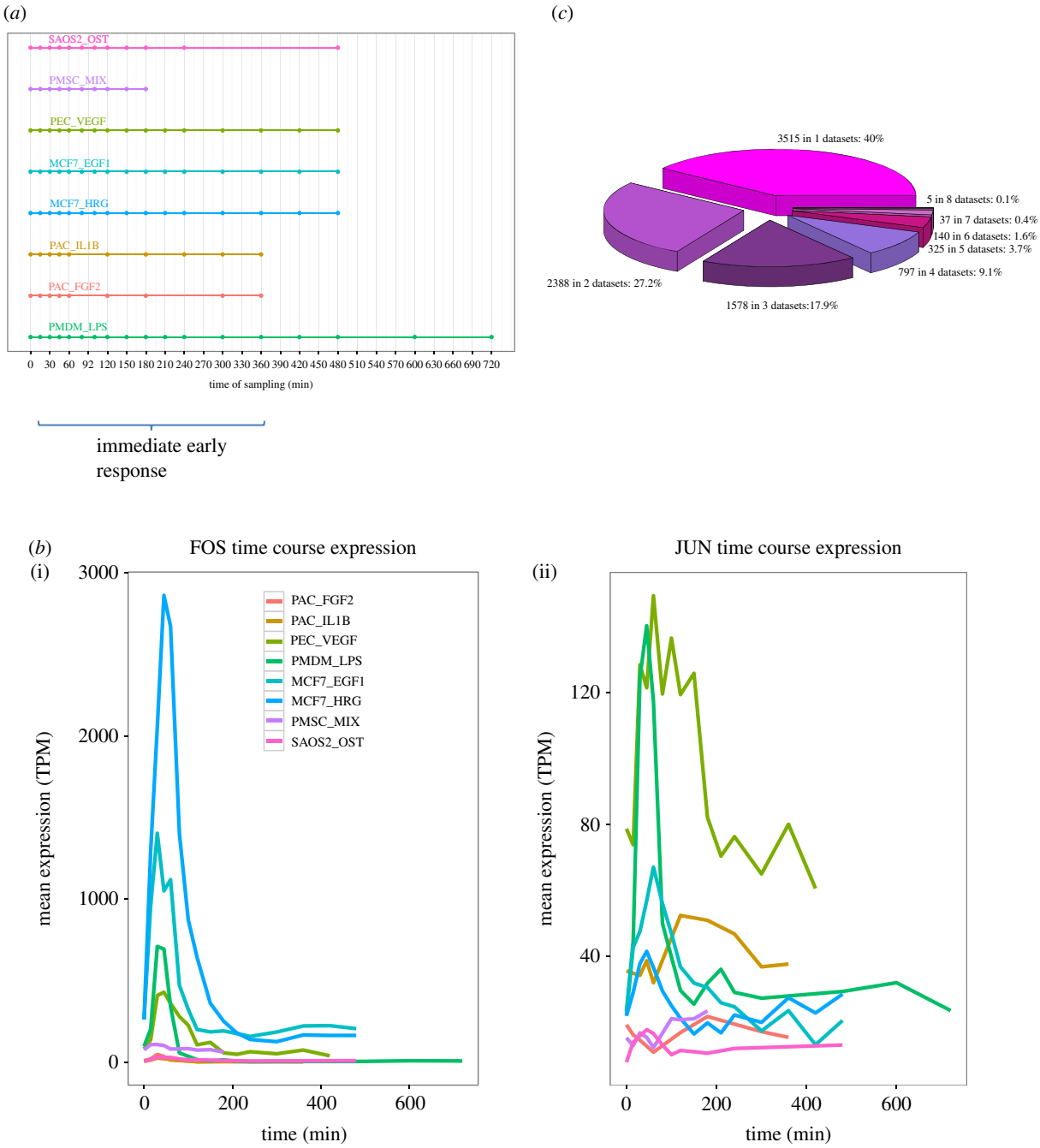


Figure 1. Time course datasets demonstrating the immediate early response. (*a*) Schematic of the eight time course datasets considered. Horizontal lines indicate the time span and symbols show the sampling times. Time zero corresponds to inactivated or quiescent cells in all cases. (*b*) The time course expression profile of FOS (i) and JUN (ii) in all eight datasets. Cage cluster expression (mean TPM of three replicates) is plotted against time. (*c*) The extent to which the classification of a TSS as a peak is unique to one dataset (3515 TSS) or shared between two or more datasets.

replicates), the algorithm returns a classification of an input transcript to a model, and also computes relevant parameters of the fitted models (e.g. time and magnitude of peak expression). These parameter estimates provide a reliable basis for comparisons across time series datasets, even with different sampling densities [7], as they are not restricted to sampling times or expression values at those times.

2.1. Cap analysis of gene expression time series meta-analysis reveals a core complement of transiently activated promoters

Across the eight time series datasets, we considered all CAGE clusters corresponding to the TSSs of known Ensembl [11] transcripts, encompassing between 10 513 (corresponding to 7706 Ensembl genes) and 14 376 (8951 genes) protein-coding CAGE TSSs, depending on the dataset, and between 1202 (692 genes) and 1640 (858 genes) ncRNA CAGE TSSs (electronic supplementary material, table S1). Between 15 and 42% of protein-coding CAGE TSSs, and between 15 and 33% of non-coding TSSs were confidently classified to one of the four models, depending on the dataset (electronic supplementary material, figure S2 and table S2). The remainder could not be rigorously classified to a single model and were omitted from further analysis. The peak model had the highest number of assignments in all the datasets for both protein-coding and ncRNA genes; for example, of 12 132 total Ensembl protein-coding genes tested, we found 8785 Ensembl genes (72%) to peak in at least one of the datasets. By contrast, few genes

were classified to the peak model in multiple datasets, with only 42 such genes shared across at least seven datasets (figure 1*c*), underlining the high variability of transcriptional responses seen for the same promoters across time series. These 42 genes constituted our 'robust' set of candidate IEGs (genes, TSSs and peak times listed in electronic supplementary material, File S1). We also defined a less stringent 'permissive' set of 1304 candidates shared across at least four out of eight datasets.

We then explored the overlap in peaking genes outside of the robust set (electronic supplementary material, table S3) and found that, for each dataset, at least 8% of peaking genes are shared with another dataset (range 8–16%) and up to 52% of peaking genes are shared (range 19–52%). The intersections between sets of three datasets became smaller consequently. Notably, approximately 50% of peaking genes are shared between datasets where the cell type is the same (MCF7 and PAC).

Our model fitting approach provided parameter estimates for all promoters assigned to the same model, providing a straightforward and intuitive basis for meta-analysis. For example, comparison of the peak times (t_p) (figure 2*a*) for all promoters classified as peaks in at least four datasets (the permissive set) readily demonstrated common patterns across datasets (figure 2*b*). Waves of promoter activation were evident, with certain promoters, particularly known IEGs, activated in the same early time window in multiple datasets. Hierarchical clustering of the datasets based on these peak class promoters (9% of all promoters assayed) also recapitulated known relationships between cell types and stimuli (figure 2*b*). The two datasets derived from the same breast cancer cell line (MCF7_EGF1 and MCF7_HRG) and stimulated with different ligands of the same ErbB receptor family clustered together as might be expected. We observed similar behaviour for the two primary aortic cell samples exposed to a growth factor or activated by a pro-inflammatory cytokine (PAC_FGF2 and PAC_IL1B, respectively). Thus, similarities in promoter activation dynamics (reflected in t_p parameter estimates) between datasets may reflect underlying commonalities in their underlying biology.

The extent of alternative promoter usage across the robust set of IEGs and candidate IEGs is shown in figure 3 (see also electronic supplementary material, figure S3). Candidate IEGs show slightly greater variability in the TSSs they activate across datasets compared with known IEGs, with a greater median number TSS found to peak (3.5 compared with 2 for known IEGs). In addition, known IEGs tend to possess TSSs that are successfully classified to the peak model across a larger number of datasets (mean proportion of datasets classified as peak per TSS for known IEGs in the robust set = 4; candidate IEG mean proportion = 2.5). Thus, known IEGs tend to possess smaller numbers of alternative TSSs that also tend to show discernible peaks in the majority of the time series datasets. It is possible that these relatively stereotypical transcriptional characteristics of known IEGs may, in some cases, have led to their status as well-established IEGs. Similarly, the increased variability seen for the TSSs of candidate IEGs could have led to a failure to detect their IEG-like behaviour in former studies.

We investigated the nature of our promoter classifications by testing the enrichment of known IEGs (see Material and methods) within each class, for each dataset. The peak class was enriched for known IEGs in all datasets (electronic

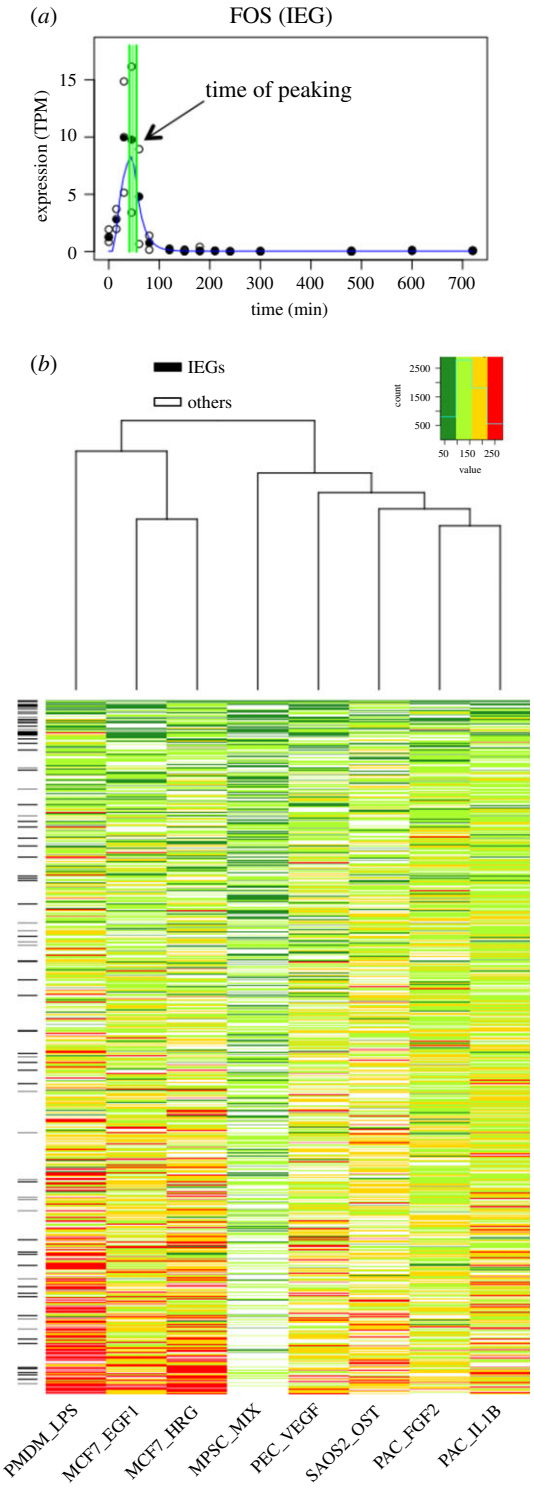


Figure 2. Broad trends in peak expression times across datasets. (*a*) Identification of the peak time parameter (t_p) of FOS estimated from the PMDM_LPS time series (filled symbols indicate the median TPM; unfilled symbols are individual replicates; green lines represent t_p and one standard deviation above and below). (*b*) Heatmap of the times of peak TSS expression (t_p) for TSSs in the permissive set for all datasets. Heatmap colours reflect the t_p for each CAGE TSS (within 100 min: dark green; 100–150 min: light green; 150–200 min: yellow; beyond 200 min: red). Known IEGs are indicated on the left by black cells.

supplementary material, figure S4 and table S4), but failed to reach statistical significance in PMSC_MIX (OR = 1.3, p = 0.2). Peaking genes shared across datasets were generally associated with significant enrichments of known IEGs (table 1), with the permissive set (shared across four or more

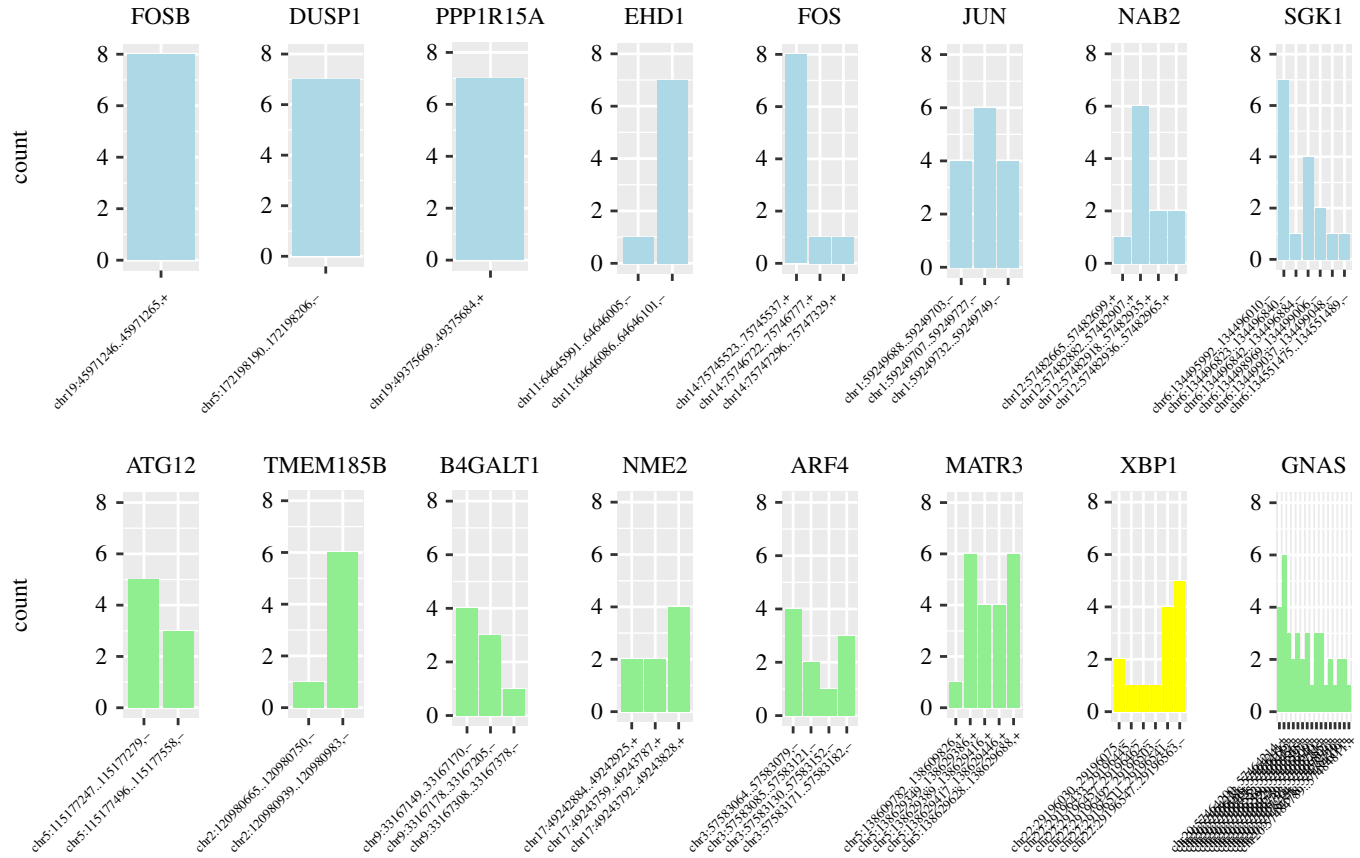


Figure 3. Promoter usage across time series datasets. For representative genes, bar charts show the number of datasets where each TSS peaks to illustrate the diversity of TSS usage and commonality of the peaking response. Known IEGS are shown in blue, TFs in yellow and other genes in green. FOSB has a single TSS that peaks in eight datasets, JUN has three TSS each peaking in four or more datasets and XBP1 has six TSS that peak in between one and six datasets.

Table 1. Enrichment of known IEGs for genes classified to the peak model in multiple datasets. Enrichment (expressed as odds ratios) and *p*-values for genes classified across different numbers of time series datasets.

	IEGs enrichment						
shared datasets	no. genes	no. IEGs	no. CAGE TSSs (across eight datasets)	no. IEG CAGE TSSs (across eight datasets)	OR	*p*-value	no. CAGE TSSs (median)
1–8 (all peaking genes)				8785	204	102 496	913
—	—	1					
2–8	5270	171	71 384	853	6.3	2.2×10^{-16}	1
3–8	2882	128	45 360	751	5.9	2.2×10^{-16}	2
4–8	1304	86	24 616	590	5.9	2.2×10^{-16}	2
5–8	507	56	11 528	433	7.4	2.2×10^{-16}	3
6–8	182	35	4896	299	10.3	2.2×10^{-16}	3
7–8	42	13	1376	124	12.6	2.2×10^{-16}	4
8	5	2	264	18	8.3	4.6×10^{-11}	5

datasets) expected to contain higher numbers of false positives than the robust set (seven or more datasets). Genes possessing TSSs assigned to the peak class showed enrichments for gene ontology (GO) processes associated with transcription, cell activation, cell proliferation, cell differentiation and cancer-related terms such as cell death and apoptosis (FDR < 0.05; Material and methods) [12,13]. These terms were also consistent with previous studies of IEGs [4] as genes in the robust set showed enrichment for 285 GO terms, over 30% (88) of which were shared with the list of 773 GO terms of all known IEGs (electronic supplementary material, table S5).

2.2. Novel non-coding RNA candidates in the immediate early response

We next applied our classification to promoters driving the expression of non-coding transcripts and found peak

Table 2. Non-coding genes peaking in at least seven out of eight datasets. The short descriptions of the molecular function are from the genecard database [15].

gene ID	no. of shared datasets	description (PubMed ref.)
LINC00478 (MIR99AHG)	7	it has a role in cell proliferation and differentiation and it is considered a regulator of oncogenes in leukaemia (PMID: 25027842)
LINC00263	7	regulation of oligodendrocyte maturation (PMID: 25575711)
LINC-PINT	8	putative tumour suppressor (PMID: 24070194)
LINC00963	7	involved in the prostate cancer transition from androgen-dependent to androgen-independent and metastasis via the EGFR signalling pathway (PMID: 24691949)
LINC00476	8	uncharacterized lincRNA
LINC00674	7	uncharacterized lincRNA
STX18-AS1	7	uncharacterized lincRNA
DLEU2	7	critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1 (PMID:19591824)
MiR-29A	7	the expression of the miR-29 family has antifibrotic effects in heart, kidney and other organs; miR-29s have also been shown to induce apoptosis and regulate cell differentiation (PMID: 22214600)
MiR-3654	7	involved in prostate cancer progression (PMID: 27297584)
MiR-21	7	oncogenic potential (PMID: 18548003)
AL928646	7	uncharacterized ncRNA
SCARNA17	7	scaRNA involved in the maturation of other RNA molecules (PMID: 12032087)
SNORD65	7	belongs to the small nucleolar RNAs, C/D family; involved in rRNA modification and alternative splicing (PMID: 26957605)
SNORD82	7	belongs to the small nucleolar RNAs, C/D family; involved in rRNA modification and alternative splicing (PMID: 26957605)

promoters driving the expression of 20 ncRNA genes (across at least seven datasets), constituting the robust set of ncRNA candidate IEGs. These included promoters associated with the cellular splicing machinery, such as small nuclear RNA multi-gene families (U1, U2 and U4), which are part of the spliceosome, and SCARNA17, a small nuclear RNA which contributes to he post transcriptional modification of many snRNPs. Kalam *et al*. [14] have shown that macrophage infection with *Mycobacterium tuberculosis* results in the systematic perturbation in splicing patterns, and our results suggest more general roles for alternative splicing in the IER. However, multigene families, such as these small nuclear RNAs, present particular challenges for reliable sequence read mapping. Although probabilistic approaches to mapping ambiguously mapped reads were developed in FANTOM5 [10], we have chosen to conservatively remove these genes from the robust set, leaving a group of 15 non-coding genes with a median of five peaking TSS (table 2; electronic supplementary material, figures S5 and S6).

Three miRNAs are present in the robust set (table 2) including the oncogene miR-21 which was previously reported to show IEG-like behaviour in the PAC_FGF2, PAC_IL1B and MCF7_HRG time series [7]. Here, we find similar behaviour in the MCF7_EGF1, PEC_VEGF, PMSC_MIX and SAOS2_OST datasets. This extends previous studies reporting that the miR-21 mature transcript is upregulated on EGF treatment in MCF10A [16] and HeLa [17] cells. miR-29A has been associated with the viability and proliferation of mesenchymal stem cell and gastric cancer cells [18,19] and DLEU2 is a putative tumour suppressor gene that hosts two miRNAs, miR-15A and miR-16-1 which are known to inhibit cell proliferation and the colony-forming ability of tumour cell lines, and to induce apoptosis [20–22]. Seven lncRNAs also appear in the robust set (table 2), and among them, LINC00478 is particularly interesting, as it has already been reported to show IEG-like behaviour [7], is implicated in breast cancer and hosts an intronic cluster of miRNAs comprising let-7c, miR-99a and miR-125b [23]. Although poorly characterized, LINC00263, LINC-PINT and LINC00963 are thought to be involved in biological processes often triggered by IEGs, such as cell maturation, cell proliferation and the expression of growth factor receptors [24–27].

2.3. Known immediate early gene promoters show conserved temporal order of activation across datasets

Having established common patterns of peak gene induction at similar times across datasets (figure 2*b*), we hypothesized that IEGs may also be induced in a conserved order over time. To our knowledge, the extent of conserved ordering in gene induction is unstudied in general, and in the IER, it is of particular interest for two main reasons. First, the presence of conserved gene orderings, in addition to common gene classifications, provides an additional test for functional similarity between datasets. Second, strongly conserved ordering may suggest the existence of conserved regulatory mechanisms governing the induction of these genes. To analyse the relative order of activation across the eight datasets, we compared the peak time of each gene to that of all others in the peak class. If the

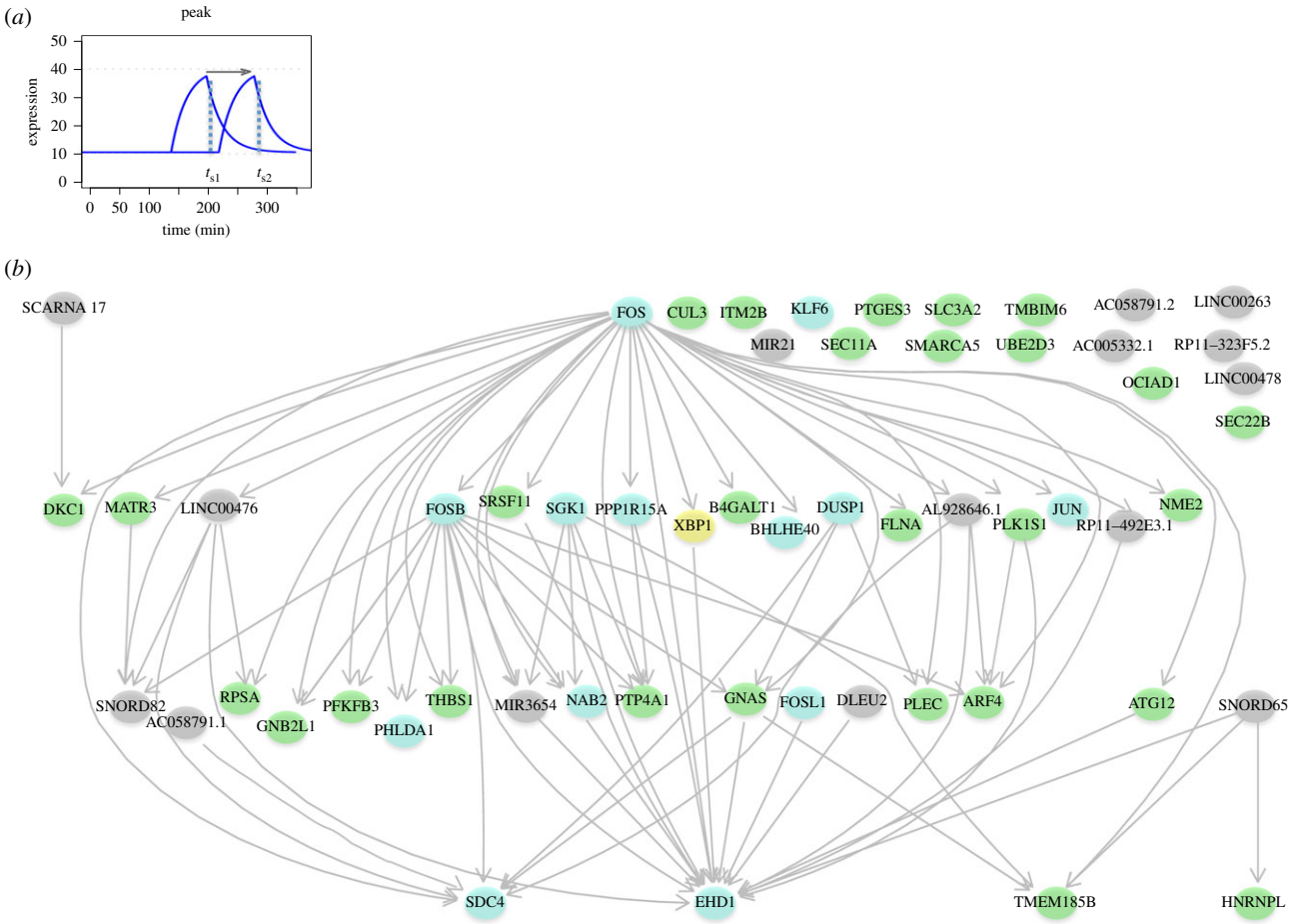


Figure 4. Conserved activation network. (*a*) Schematic profiles of two peaking genes, with temporal precedence indicated by the arrow. (*b*) Conserved temporal precedence between IEGs (light blue nodes), TFs (yellow nodes) ncRNA (grey nodes) and other protein-coding genes (green nodes) is shown by directed edges. A subset of IEGs in this network are also TFs (FOS, KLF6, FOSB, BHLHE40, JUN and FOSL1).

relative temporal order of two genes was conserved in at least seven of the eight datasets, the ordering for this pair was considered conserved and represented by an edge in the conserved activation network (figure 4).

We found 77 pairs of genes showing conserved ordering in their activation, involving 40 of the 57 genes in the robust set. FOS was the first gene to be activated (lacking a predecessor in the ordering) and SDC4, EHD1 and TMEM185B were the last. The number of conserved temporal connections observed overall is statistically significant ($p < 5 \times 10^{-3}$) by comparison with the distribution of expected connections, given 1 000 000 permuted datasets (Material and methods). This appears to reflect a conserved coordination in promoter activation during the IER and further supports the candidacy of the novel IEGs detected. Many genes in this network are known to participate in well-studied pathways active in the IER such as the MAPK signalling pathway as we now discuss.

2.4. Known immediate early genes and candidate immediate early genes participate in common signalling pathways

Having shown that the peak model described the behaviour of known IEGs, we speculated that the other genes assigned to this model might include novel candidate IEGs. Of the 42 genes in the robust set, more than two-thirds (29 genes) are not known to be IEGs and can therefore be considered to be candidate novel IEGs (henceforth candidate IEGs). Pathway analysis [28] recovers many known relationships among known IEGs as expected, centred on heavily studied IEGs such as FOS and JUN. However, the same analysis suggests that more than half (17) of candidate IEGs also participate in common pathways with known IEGs, involving a densely inter-connected network of 83 significantly over-represented pathways (electronic supplementary material, table S6), including signalling cascades known to mediate the IER, such as the Ca^{2+}-dependent pathways and the mitogen-activated protein (MAP) kinase network [29,30].

The dynamics of the expression of peak-classified genes can be visualized by a scatterplot of fold change against peak time (electronic supplementary material, figure S7). These quantitative features along with the conserved temporal orderings described above show FOS as the earliest peaking IEG, EHD1 as the last, with an array of conserved orderings subsequent to, and prior to the peaking of these genes, respectively (selected genes plotted in figure 5). The TSSs of known IEGs CAGE tend to show the greatest fold changes (electronic supplementary material, figure S8*a*; Wilcoxon $p < 2.2 \times 10^{-16}$); however, some candidate IEGs promoters show notably similar timing (electronic supplementary material, figure S8*c*; Wilcoxon $p = 0.89$). The time of peaking is significantly earlier for known IEGs relative to the other protein-coding promoters in only three time series: PMDM_LPS, MCF7_EGF1 and PEC_VEGF. Fold changes in peak ncRNA promoters tend to be lower than for known IEGs (electronic supplementary

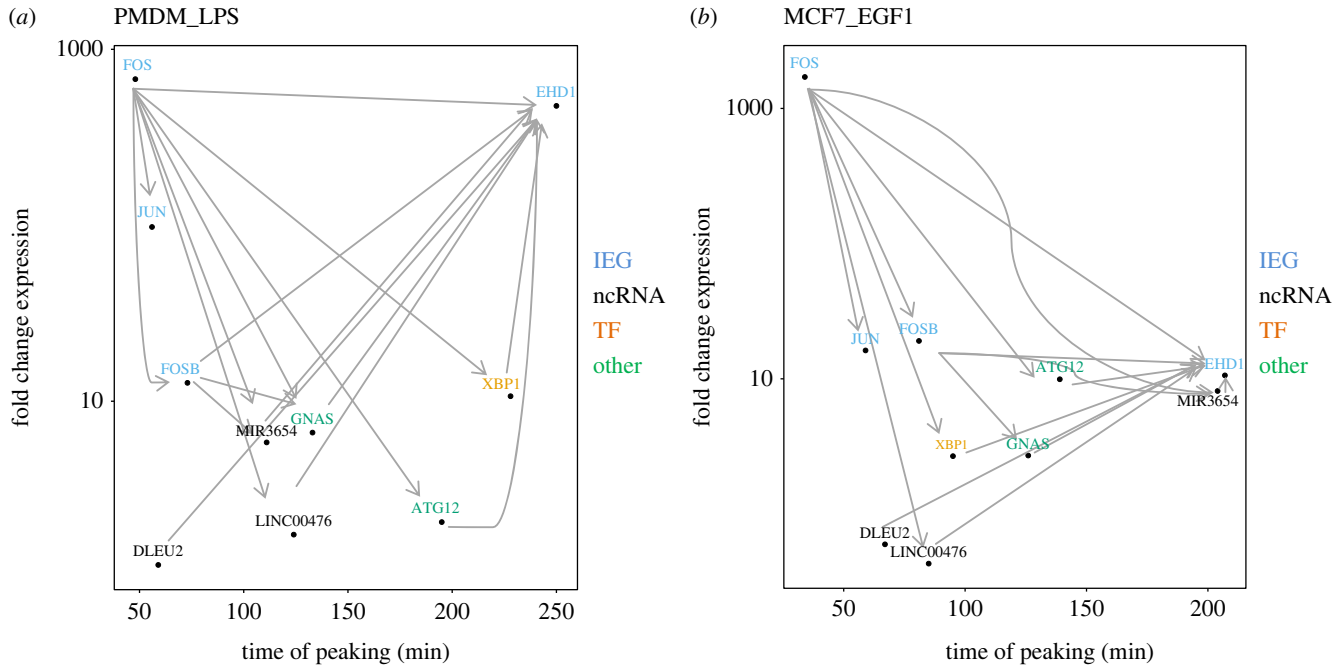


Figure 5. Transcriptional dynamics of genes classified to the peak model. Scatterplots of log fold change against the time of peaking for selected genes, with conserved temporal precedence indicated by arrows for (*a*) PMDM_LPS and (*b*) MCF7_EGF1. FOS peaks earliest and has many conserved temporal relations to later peaking genes, while EHD1 peaks late and has many conserved temporal orderings with earlier peaking genes.

material, figure S8*b*; Wilcoxon $p < 0.05$), but they occur earlier than known IEGs (electronic supplementary material, figure S8*d*, Wilcoxon $p < 0.05$ for all datasets).

Among the candidate IEGs in the robust set, XBP1 is especially noteworthy. This gene encodes a TF and is relatively short in length (6 kb compared with the mean of 58 kb for all Ensembl protein-coding genes), consistent with the IEG archetype [1]. XBP1 is a highly conserved component of the unfolded protein response (UPR) signalling pathways, activated by unconventional splicing upon endoplasmic reticulum (ER) stress or non-classical anticipatory activation [31–33], and regulates a diverse array of genes involved in ER homeostasis, adipogenesis, lipogenesis and cell survival [34,35]. Interestingly, genes in the robust set are significantly enriched for the GO term GO:003497 *response to ER stress* (FDR < 0.05, all tested genes as the background), and four of the five genes in the robust set sharing this term peak in conserved order across the datasets. Furthermore, we found a significant enrichment (FDR < 0.05) of the XBP1 binding motif in the promoter regions (see Material and methods) of the robust set of genes (electronic supplementary material, figure S9).

3. Discussion

Exploiting the precision of FANTOM5 CAGE times series data, we discover a robust set of 42 protein-coding genes driven by promoters showing rapid and transient activation in response to multiple stimuli. This set contains 13 previously known IEGs and 29 candidate IEGs, which are likely to be core components of the IER. Applying our approach to the CAGE TSSs of ncRNAs, we also discovered a set of 15 ncRNAs peaking across at least seven datasets, comprising miRNAs and lncRNAs, suggesting regulatory roles for particular miRNAs and lncRNAs species in the IER [7].

FOS expression has long been considered to lead the IER after cell stimulation [36,37]. Our results on the IER conserved activation network support this, but also similarly conserved relationships extending to an additional 39 coding and non-coding genes. Furthermore, we observed many known and novel IEGs in this network known to be involved in a range of signalling pathways active in the IER, such as the MAPK and the EGF/EGFR signalling pathways. This suggests the variable constellations of genes involved in the IER to any particular stimulus may be underpinned by a deeper level of conservation in the regulation of the IER across stimuli.

One of the most interesting candidate IEGs, XBP1, can be rapidly activated by alternative splicing minutes after cell stimulation with mitogenic hormones, activating peptides such as LPS and cytokines [31–33]. This key event of the induced UPR pathway is a conserved eukaryotic response to cellular stress, and is thought to cooperate in the regulation of IEG expression [32]. However, the dynamics of XBP1 promoter induction in the context of the IER have not been studied previously. Interestingly, we found a significant enrichment for XBP1 TF binding sites in the promoter regions of 11 genes in the IER conserved activated network. The presence of XBP1 and XBP1-responding genes in this network suggests this gene may act as an important link between the IER and the UPR pathway.

4. Material and methods

4.1. Datasets

The eight datasets used (figure 1) are the most densely sampled human time series produced by the FANTOM5 Project, with all time points represented by three replicates [38]. Detailed information on the generation of these datasets is available from Arner *et al*. [10], including CAGE library preparation, quality control, sequencing and qRT–PCR validation, as well as protocols for CAGE read clustering and TSS detection. All CAGE clusters representing TSSs of protein-coding genes were

conservatively thresholded to more than 10 TPM (tags per million), while CAGE clusters corresponding to ncRNA were thresholded to greater than 2 TPM, allowing for their generally lower expression levels. FANTOM5 data downloads, browsers and genomic tools are available from the project website (http://fantom.gsc.riken.jp/5/).

4.2. Model-based classification of transcription start site expression profiles

To classify time series data for each CAGE-defined TSS, we refined a previously published method [7] which fits different mathematical models (kinetic signatures) to individual expression profiles, assessing the best fit using nested sampling [39] to compute the marginal likelihood, log Z. All time series were normalized such that the medium minimum and maximum across the time series was set to 0 and 10, respectively.

The kinetic signatures considered are: linear, decay, dip and delayed peak (electronic supplementary material, figure S1). The peak kinetic signature considered in the previous method was modified to allow a delay before expression starts to increase in exponential fashion (t_d). Parameter t_s is the time duration of the initial increase in expression, p_1 is the expression at time 0, and p_2 is the increase in expression at the time of peaking, $t_p = t_d + t_s$.

$$\delta = \frac{\log(0.3)}{t_s}, \tag{4.1}$$

$$y = p_1; t \leq t_d, \tag{4.2}$$

$$y = p_1 + p_2(1 - e^{\delta(t-t_d)}); t_d < t \leq t_d + t_s \tag{4.3}$$

$$\text{and} \quad y = p_1 + p_2(1 - e^{\delta(t-t_d)}) - p_2(1 - e^{\delta(t-t_d-t_s)}); t > t_d + t_s. \tag{4.4}$$

However, an alternative rate $\delta = \log(0.1)/t_d$ was also used to model the slower dynamics of transcripts peaking later in time, and the best fitting model selected during the decision step. Normalizing the data such that expression lies in the range 0–10 allowed the prior ranges of parameters to be restricted to plausible values that applied to all time series. The fit of models to data was improved as a result. To account for any impact on the log Z calculation, we generated synthetic time series datasets using parameter values drawn at random from the prior ranges to generate one replicate, and generated two other replicates by adding and subtracting (respectively) a given amount of noise to the first. Model fitting was applied to 1000 such datasets per model (using the same noise values for each model on each of the 1000 iterations) and we observed an advantage for each of the more complex models in comparison with the linear model that was consistent over the range of log Z values obtained for the linear model. To offset this effect for each complex model, the advantage (mean difference plus two standard deviations observed in synthetic data) was subtracted from the log Z values calculated for CAGE TSS data when making the categorization decision.

4.3. Transcription factor binding site identification

We assessed the enrichment of transcription factor binding site (TFBS) motifs in the JASPAR database [40] (January 2017 release) for all CAGE TSS assigned to genes in the robust set relative to those assigned to the 12 132 genes tested across all the datasets. Motif matches (FDR $\leq$ 0.05) were sought in flanking 400 bp windows centred on the middle of each CAGE TSS analysed), using FIMO [41] from the MEME package (v. 4.11.2 patch 2). Enrichment of each motif in the robust set relative to the total set was assessed with Fisher's exact tests, correcting for multiple testing (FDR $\leq$ 0.05).

4.4. Pathway and gene ontology enrichment

Functional and pathway enrichments were assessed using GOrilla [13] and InnateDB [28], respectively (FDR $\leq$ 0.05), using the total 12 132 genes analysed across the eight datasets as the background set.

The list of 234 known IEGs [10] was assembled from 20 published human and mouse datasets from the literature; it is expected to contain few false positives but does include a number of IEGs only reported in cells and/or responses not examined in this study. To compute the enrichment of known IEGs in each dataset, we compared the proportion of peaking CAGE TSSs assigned to IEGs with the proportion of peaking CAGE TSSs assigned to candidate IEGs. For the enrichment of known IEGs in each set of shared peaking genes, we compared the proportion of peaking CAGE TSSs assigned to the IEGs shared in each group of shared genes with the proportion of peaking CAGE TSSs assigned to IEGs in the remaining tested genes. The odds ratio and the *p*-value were assigned using Fisher's exact test.

4.5. Network conservation

A total of 57 protein-coding and non-coding candidate IEGs (corresponding to known Ensembl genes) were considered for construction of the conserved activation network. For genes with multiple peaking CAGE TSS, we chose the earliest peaking CAGE TSS (smallest t_p) in each dataset, then the relative pairwise order of each gene was computed with respect to all the other genes in the robust set. For example, if in dataset-1, gene-A peaks before gene-B ($t_{p\ \text{gene-A}} < t_{p\ \text{gene-B}}$), and this order is observed in six or more of the other seven dataset, the temporal precedence is defined to be conserved. Applying this procedure to all 57 coding and non-coding genes of the robust set, we discovered 40 genes temporally connected by 77 conserved relative orderings (figure 4). The significance of the number of temporal connections observed was measured relative to null distribution, constructed by permuting t_p for all the CAGE TSSs 1 000 000 times; with the proportion of permuted datasets with at least as many conserved orderings as the observed taken as an empirically derived *p*-value. The observed value (77) was observed or exceeded in 4516 out of 1 000 000 permutations, indicating that the number of temporal connections was statistically significant ($p < 5 \times 10^{-3}$).

Data accessibility. This article has no additional data.

Authors' contributions. A.V. carried out the computational analysis and drafted the manuscript; C.A.S. and S.A. designed the study and drafted the manuscript; M.I., H.K., E.A., T.L., C.O.D., P.C. and A.R.R.F. generated and processed the CAGE data coordinated by Y.H. All authors gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. This research was funded by MRC Core funding to the MRC Human Genetics Unit.

Acknowledgments. We thank all members of the FANTOM Consortium for contributing to generation of samples and analysis of the

FANTOM5 dataset and thank GeNAS for data production. The core members of the FANTOM5 phase 2 project were: Alistair R. R. Forrest, Albin Sandelin, Carsten O. Daub, Christine Wells, David A. Hume, Erik Arner, Hideya Kawaji, Kim M Summers, Kristoffer Vitting-Seerup, Piero Carninci, Robin Andersson, Yoshihide Hayashizaki. Finn Drabløs (Department of Clinical and Molecular Medicine, Laboratory Center, Erling Skjalgsons gt. 1, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway) curated the list of IEGs within the FANTOM5 Consortium.

References

1. Fowler T, Sen R, Roy AL. 2011 Regulation of primary response genes. *Mol. Cell* **44**, 348–360. (doi:10.1016/j.molcel.2011.09.014)
2. Healy S, Khan P, Davie JR. 2013 Immediate early response genes and cell transformation. *Pharmacol. Ther.* **137**, 64–77. (doi:10.1016/j.pharmthera.2012.09.001)
3. Greenberg ME, Ziff EB. 1984 Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. *Nature* **311**, 433–438. (doi:10.1038/311433a0)
4. Bahrami S, Drabløs F. 2016 Gene regulation in the immediate-early response process. *Adv. Biol. Regul.* **62**, 37–49. (doi:10.1016/j.jbior.2016.05.001)
5. O'Donnell A, Odrowaz Z, Sharrocks AD. 2012 Immediate-early gene activation by the MAPK pathways: what do and don't we know? *Biochem. Soc. Trans.* **40**, 58–66. (doi:10.1042/BST20110636)
6. Fowler T, Suh H, Buratowski S, Roy AL. 2013 Regulation of primary response genes in B cells. *J. Biol. Chem.* **288**, 14 906–14 916. (doi:10.1074/jbc.M113.454355)
7. Aitken S *et al.* 2015 Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response. *PLoS Comput. Biol.* **11**, e1004217. (doi:10.1371/journal.pcbi.1004217)
8. Andersson R *et al.* 2014 An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461. (doi:10.1038/nature12787)
9. Bar-Joseph Z, Gitter A, Simon I. 2012 Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564. (doi:10.1038/nrg3244)
10. Arner E *et al.* 2015 Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014. (doi:10.1126/science.1259418)
11. Hubbard T *et al.* 2002 The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41. (doi:10.1093/nar/30.1.38)
12. Supek F, Bošnjak M, Škunca N, Šmuc T. 2011 REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800. (doi:10.1371/journal.pone.0021800)
13. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009 GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 1. (doi:10.1186/1471-2105-10-48)
14. Kalam H, Fontana MF, Kumar D. 2017 Alternate splicing of transcripts shape macrophage response to Mycobacterium tuberculosis infection. *PLoS Pathog.* **13**, e1006236. (doi:10.1371/journal.ppat.1006236)
15. Stelzer G *et al.* 2016 The GeneCards suite: from Gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* **54**, 1–30.
16. Avraham R *et al.* 2010 EGF decreases the abundance of microRNAs that restrain oncogenic transcription factors. *Sci. Signal.* **3**, ra43. (doi:10.1126/scisignal.2000876)
17. Llorens F *et al.* 2013 Microarray and deep sequencing cross-platform analysis of the mirRNome and isomiR variation in response to epidermal growth factor. *BMC Genomics* **14**, 371. (doi:10.1186/1471-2164-14-371)
18. Liu X *et al.* 2015 MicroRNA-29a inhibits cell migration and invasion via targeting Roundabout homolog 1 in gastric cancer cells. *Mol. Med. Rep.* **12**, 3944–3950. (doi:10.3892/mmr.2015.3817)
19. Zhang Y, Zhou S. 2015 MicroRNA 29a inhibits mesenchymal stem cell viability and proliferation by targeting Roundabout 1. *Mol. Med. Rep.* **12**, 6178–6184. (doi:10.3892/mmr.2015.4183)
20. Cimmino A *et al.* 2005 miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl Acad. Sci. USA* **102**, 13 944–13 949. (doi:10.1073/pnas.0506654102)
21. Lerner M *et al.* 2009 DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1. *Exp. Cell Res.* **315**, 2941–2952. (doi:10.1016/j.yexcr.2009.07.001)
22. Gao S-M, Xing C-Y, Chen C-Q, Lin S-S, Dong P-H, Yu F-J. 2011 miR-15a and miR-16-1 inhibit the proliferation of leukemic cells by down-regulating WT1 protein level. *J. Exp. Clin. Cancer Res.* **30**, 1. (doi:10.1186/1756-9966-30-1)
23. Gökmen-Polar Y, Zavodszky M, Chen X, Gu X, Kodira C, Badve S. 2016 Abstract P2-06-05: LINC00478: a novel tumor suppressor in breast cancer. *Cancer Res.* **76**(4 Suppl.), P2-06-5-P2-5. (doi:10.1158/1538-7445.SABCS15-P2-06-05)
24. Mills JD, Kavanagh T, Kim WS, Chen BJ, Waters PD, Halliday GM, Janitz M. 2015 High expression of long intervening non-coding RNA OLMALINC in the human cortical white matter is associated with regulation of oligodendrocyte maturation. *Mol. Brain* **8**, 1. (doi:10.1186/s13041-014-0091-9)
25. Müller S *et al.* 2015 Next-generation sequencing reveals novel differentially regulated mRNAs, lncRNAs, miRNAs, sdRNAs and a piRNA in pancreatic cancer. *Mol. Cancer* **14**, 1. (doi:10.1186/1476-4598-14-1)
26. Marín-Béjar O *et al.* 2013 Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2. *Genome Biol.* **14**, 1. (doi:10.1186/gb-2013-14-9-r104)
27. Wang L, Han S, Jin G, Zhou X, Li M, Ying X, Wang L, Wu H, Zhu Q. 2014 Linc00963: a novel, long non-coding RNA involved in the transition of prostate cancer from androgen-dependence to androgen-independence. *Int. J. Oncol.* **44**, 2041–2049. (doi:10.3892/ijo.2014.2363)
28. Breuer K *et al.* 2012 InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228–D1233. (doi:10.1093/nar/gks1147)
29. Schratt G, Weinhold B, Lundberg AS, Schuck S, Berger J, Schwarz H, Weinberg RA, Ruther U, Nordheim A. 2001 Serum response factor is required for immediate-early gene activation yet is dispensable for proliferation of embryonic stem cells. *Mol. Cell. Biol.* **21**, 2933–2943. (doi:10.1128/MCB.21.8.2933-2943.2001)
30. Treisman R. 1996 Regulation of transcription by MAP kinase cascades. *Curr. Opin. Cell Biol.* **8**, 205–215. (doi:10.1016/S0955-0674(96)80067-6)
31. Andruska N, Zheng X, Yang X, Helferich WG, Shapiro DJ. 2015 Anticipatory estrogen activation of the unfolded protein response is linked to cell proliferation and poor survival in estrogen receptor α positive breast cancer. *Oncogene* **34**, 3760. (doi:10.1038/onc.2014.292)
32. Shapiro DJ, Livezey M, Yu L, Zheng X, Andruska N. 2016 Anticipatory UPR activation: a protective pathway and target in cancer. *Trends Endocrinol. Metab.* **27**, 731–741. (doi:10.1016/j.tem.2016.06.002)
33. Skalet AH, Isler JA, King LB, Harding HP, Ron D, Monroe JG. 2005 Rapid B cell receptor-induced unfolded protein response in nonsecretory B cells correlates with pro-versus antiapoptotic cell fate. *J. Biol. Chem.* **280**, 39 762–39 771. (doi:10.1074/jbc.M502640200)
34. He Y, Sun S, Sha H, Liu Z, Yang L, Xue Z, Chen H, Qi L. 2010 Emerging roles for XBP1, a sUPeR transcription factor. *Gene Expr.* **15**, 13–25. (doi:10.3727/105221610X12819686555051)
35. Piperi C, Adamopoulos C, Papavassiliou AG. 2016 XBP1: a pivotal transcriptional regulator of glucose and lipid metabolism. *Trends Endocrinol. Metab.* **27**, 119–122. (doi:10.1016/j.tem.2016.01.001)
36. Hu E, Mueller E, Oliviero S, Papaioannou V, Johnson R, Spiegelman B. 1994 Targeted disruption of the c-fos gene demonstrates c-fos-dependent and-independent pathways for gene expression stimulated by growth factors or oncogenes. *EMBO J.* **13**, 3094.
37. Fei J, Viedt C, Soto U, Elsing C, Jahn L, Kreuzer J. 2000 Endothelin-1 and smooth muscle cells.

Arterioscler. Thromb. Vasc. Biol. **20**, 1244–1249. (doi:10.1161/01.ATV.20.5.1244)

38. Lizio M *et al.* 2015 Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22. (doi:10.1186/s13059-014-0560-6)
39. Aitken S, Akman OE. 2013 Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC Syst. Biol.* **7**, 72. (doi:10.1186/1752-0509-7-72)
40. Mathelier A 2013 JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147. (doi:10.1093/nar/gkt997)
41. Grant CE, Bailey TL, Noble WS. 2011 FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018. (doi:10.1093/bioinformatics/btr064)

9. References

Ahmed, S., et al. (2013). "TRIF-mediated TLR3 and TLR4 signaling is negatively regulated by ADAM15." The Journal of Immunology **190**(5): 2217-2228.

Aitken, S. and O. E. Akman (2013). "Nested sampling for parameter inference in systems biology: application to an exemplar circadian model." BMC systems biology **7**(1): 72.

Aitken, S., et al. (2015). "Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response." PLoS Comput Biol **11**(4): e1004217.

Alhendi, A. M., et al. (2018). "Promoter Usage and Dynamics in Vascular Smooth Muscle Cells Exposed to Fibroblast Growth Factor-2 or Interleukin-1β." Scientific reports **8**(1): 13164.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome biology **11**(10): R106.

Andruska, N., et al. (2015). "Anticipatory Estrogen Activation of the Unfolded Protein Response is Linked to Cell Proliferation and Poor Survival in Estrogen Receptor α Positive Breast Cancer." Oncogene **34**(29): 3760.

Arner, E., et al. (2015). "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells." Science **347**(6225): 1010-1014.

Avraham, R., et al. (2010). "EGF decreases the abundance of microRNAs that restrain oncogenic transcription factors." Sci Signal **3**(124): ra43.

Ayoubi, T. and W. Van De Ven (1996). "Regulation of gene expression by alternative promoters." The FASEB journal **10**(4): 453-460.

Bahrami, S. and F. Drabløs (2016). "Gene regulation in the immediate-early response process." Advances in Biological Regulation.

Baillie, J. K., et al. (2017). "Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease." PLoS genetics **13**(3): e1006641.

Balwierz, P. J., et al. (2014). "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." Genome Research **24**(5): 869-884.

Bar-Joseph, Z., et al. (2012). "Studying and modelling dynamic biological processes using time-series gene expression data." Nature Reviews Genetics **13**(8): 552-564.

Barzago, C., et al. (2016). "A novel infection-and inflammation-associated molecular signature in peripheral blood of myasthenia gravis patients." Immunobiology **221**(11): 1227-1236.

Bebien, M., et al. (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**(12): 1836.

Bendjilali, N., et al. (2017). "Time-course analysis of gene expression during the Saccharomyces cerevisiae hypoxic response." G3: Genes, Genomes, Genetics **7**(1): 221-231.

Birney, E., et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blake, J. A., et al. (2003). "MGD: the mouse genome database." Nucleic Acids Research **31**(1): 193-195.

Blischak, J. D., et al. (2015). "Mycobacterial infection induces a specific human innate immune response." Scientific reports **5**: 16882.

Boldogh, I., et al. (1990). "Activation of proto-oncogenes: an immediate early event in human cytomegalovirus infection." Science **247**(4942): 561-564.

Boutet, E., et al. (2007) "Uniprotkb/swiss-prot." Plant bioinformatics. Humana Press: 89-112.

Breuer, K., et al. (2012). "InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation." Nucleic Acids Research: gks1147.

Bushman, B. J. and M. C. Wang (1994). "Vote-counting procedures in meta-analysis." The handbook of research synthesis **236**: 193-213.

Calin, G. A. and C. M. Croce (2006). "MicroRNA signatures in human cancers." Nature Reviews Cancer **6**(11): 857.

Cao, S. S. and R. J. Kaufman (2014). "Endoplasmic reticulum stress and oxidative stress in cell fate decision and human disease." Antioxidants & redox signaling **21**(3): 396-413.

Carbajo, D., et al. (2015). "Application of Gene Expression Trajectories Initiated from ErbB Receptor Activation Highlights the Dynamics of Divergent Promoter Usage." PloS one **10**(12): e0144176.

Carey, D. J. (1997). "Syndecans: multifunctional cell-surface co-receptors." Biochemical Journal **327**(1): 1-16.

Carninci, P., et al. (1996). "High-efficiency full-length cDNA cloning by biotinylated CAP trapper." Genomics **37**(3): 327-336.

Carninci, P., et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nature genetics **38**(6): 626-635.

Cattoretti, G., et al. (2005). "PRDM1/Blimp-1 is expressed in human B-lymphocytes committed to the plasma cell lineage." The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland **206**(1): 76-86.

Cheng, B., et al. (2016). "Syndecans as Cell Surface Receptors in Cancer Biology. A Focus on their Interaction with PDZ Domain Proteins." Frontiers in pharmacology **7**.

Cimmino, A., et al. (2005). "miR-15 and miR-16 induce apoptosis by targeting BCL2." Proceedings of the National Academy of Sciences of the United States of America **102**(39): 13944-13949.

Cochran, B. H., et al. (1983). "Molecular cloning of gene sequences regulated by platelet-derived growth factor." Cell **33**(3): 939-947.

Consortium, E. P. (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS biology **9**(4): e1001046.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57.

Consortium, T. F. (2014). "A promoter-level mammalian expression atlas." Nature **507**(7493): 462-470.

Curran, A. T., et al. (1984). "Viral and cellular fos proteins: a comparative analysis." Cell **36**(2): 259-268.

Curran, T. and J. I. Morgan (1995). "Fos: An immediate-early transcription factor in neurons." Developmental Neurobiology **26**(3): 403-412.

Curran, T., et al. (1982). "FBJ murine osteosarcoma virus: identification and molecular cloning of biologically active proviral DNA." Journal of virology **44**(2): 674-682.

Dieterich, L. C., et al. (2015). "DeepCAGE transcriptomics reveal an important role of the transcription factor MAFB in the lymphatic endothelium." Cell reports **13**(7): 1493-1504.

Di, Y. (2015). "Single-gene negative binomial regression models for RNA-Seq data with higher-order asymptotic inference." Statistics and its interface **8**(4): 405.

Dixon, B. S., et al. (1996). "The bradykinin B2 receptor is a delayed early response gene for platelet-derived growth factor in arterial smooth muscle cells." Journal of Biological Chemistry **271**(23): 13324-13332.

Ebisuya, M., et al. (2005). "The duration, magnitude and compartmentalization of ERK MAP kinase activity: mechanisms for providing signaling specificity." Journal of cell science **118**(14): 2997-3002.

Eden, E., et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC bioinformatics **10**(1): 1.

Fei, J., et al. (2000). "Endothelin-1 and Smooth Muscle Cells." Arteriosclerosis, thrombosis, and vascular biology **20**(5): 1244-1249.

Filion, G. J. (2015). "The signed Kolmogorov-Smirnov test: why it should not be used." GigaScience **4**(1): 9.

Fittall, M. W., et al. (2018). "Recurrent rearrangements of FOS and FOSB define osteoblastoma." Nature communications **9**(1): 2150.

Fowler, T., et al. (2011). "Regulation of primary response genes." Molecular cell **44**(3): 348-360.

François, M., et al. (2008). "Sox18 induces development of the lymphatic vasculature in mice." Nature **456**(7222): 643.

Freter, R. R., et al. (1996). "Platelet-derived growth factor induction of the immediate-early gene MCP-1 is mediated by NF-κB and a 90-kDa phosphoprotein coactivator." Journal of Biological Chemistry **271**(29): 17417-17424.

Frohman, M. A., et al. (1988). "Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer." Proceedings of the National Academy of Sciences **85**(23): 8998-9002.

Fromm, J. A., et al. (2008). "Epidermal growth factor receptor 1 (EGFR1) and its variant EGFRvIII regulate TATA-binding protein expression through distinct pathways." Molecular and cellular biology **28**(20): 6483-6495.

Gao, S.-m., et al. (2011). "miR-15a and miR-16-1 inhibit the proliferation of leukemic cells by down-regulating WT1 protein level." Journal of Experimental & Clinical Cancer Research **30**(1): 1.

Geiger, T. L., et al. (2014). "Nfil3 is crucial for development of innate lymphoid cells and host protection against intestinal pathogens." Journal of Experimental Medicine **211**(9): 1723-1731.

Gökmen-Polar, Y., et al. (2016). "Abstract P2-06-05: LINC00478: A novel tumor suppressor in breast cancer." Cancer Research **76**(4 Supplement): P2-06-05-P02-06-05.

Grant, C. E., et al. (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Greenberg, M. E. and E. B. Ziff (1984). "Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene." Nature **311**: 433-438.

Grohmann, K., et al. (1978). "Failure to detect “cap” structures in mitochondrial DNA-coded poly (A)-containing RNA from HeLa cells." Nucleic Acids Research **5**(3): 637-651.

Haberle, V., et al. (2013). "Package ‘CAGEr’."

Haberle, V., et al. (2014). "Two independent transcription initiation codes overlap on vertebrate core promoters." Nature **507**(7492): 381-385.

Haberle, V., et al. (2015). "CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses." Nucleic Acids Research: gkv054.

Harrow, J., et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Research **22**(9): 1760-1774.

Hart, G. T., et al. (2012). "Krüppel-like factors in lymphocyte biology." The Journal of Immunology **188**(2): 521-526.

He, Y., et al. (2010). "Emerging roles for XBP1, a sUPeR transcription factor." Gene expression **15**(1): 13-25.

Healy, S., et al. (2013). "Immediate early response genes and cell transformation." Pharmacology & therapeutics **137**(1): 64-77.

Henriksson, M. and B. Lüscher (1996). "Proteins of the Myc network: essential regulators of cell growth and differentiation." Advances in cancer research **68**: 109-182.

Herschman, H. R. (1991). "Primary response genes induced by growth factors and tumor promoters." Annual review of biochemistry **60**(1): 281-319.

Hess, J., et al. (2004). "AP-1 subunits: quarrel and harmony among siblings." Journal of cell science **117**(25): 5965-5973.

Hindley, A. and W. Kolch (2002). "Extracellular signal regulated kinase (ERK)/mitogen activated protein kinase (MAPK)-independent functions of Raf kinases." Journal of cell science **115**(8): 1575-1581.

Hoffman, B. and D. Liebermann (2008). "Apoptotic signaling by c-MYC." Oncogene **27**(50): 6462.

Hoskins, R. A., et al. (2011). "Genome-wide analysis of promoter architecture in Drosophila melanogaster." Genome Research **21**(2): 182-192.

Hu, E., et al. (1994). "Targeted disruption of the c-fos gene demonstrates c-fos-dependent and-independent pathways for gene expression stimulated by growth factors or oncogenes." The EMBO journal **13**(13): 3094.

Hubbard, T., et al. (2002). "The Ensembl genome database project." Nucleic Acids Research **30**(1): 38-41.

Ioannidis, J. P., et al. (2001). "Replication validity of genetic association studies." Nature genetics **29**(3): 306-309.

Jackson, S. J., et al. (2013). "Rapid and widespread suppression of self-renewal by microRNA-203 during epidermal differentiation." Development **140**(9): 1882-1891.

Jensen, A. and A. la Cour-Harbo (2001). Ripples in mathematics: the discrete wavelet transform, Springer Science & Business Media.

Joukov, V., et al. (1998). "A recombinant mutant vascular endothelial growth factor-C that has lost vascular endothelial growth factor receptor-2 binding, activation, and vascular permeability activities." Journal of Biological Chemistry **273**(12): 6599-6602.

Kalam, H., et al. (2017). "Alternate splicing of transcripts shape macrophage response to Mycobacterium tuberculosis infection." PLoS pathogens **13**(3): e1006236.

Kanamori-Katayama, M., et al. (2011). "Unamplified cap analysis of gene expression on a single-molecule sequencer." Genome Research **21**(7): 1150-1159.

Karin, M., et al. (1997). "AP-1 function and regulation." Current opinion in cell biology **9**(2): 240-246.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management, Oxford University Press.

Katz, M., et al. (2007). "Regulation of MAPKs by growth factors and receptor tyrosine kinases." Biochimica et Biophysica Acta (BBA)-Molecular Cell Research **1773**(8): 1161-1176.

Kawaji, H., et al. (2017). "The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types." Scientific data **4**.

Khatri, P. and S. Drăghici (2005). "Ontological analysis of gene expression data: current tools, limitations, and open problems." Bioinformatics **21**(18): 3587-3595.

Kedmi, M., et al. (2015). "EGF induces microRNAs that target suppressors of cell migration: miR-15b targets MTSS1 in breast cancer." Sci. Signal. **8**(368): ra29-ra29.

Kimura, K., et al. (2006). "Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes." Genome Research **16**(1): 55-65.

Kitabayashi, I., et al. (1991). "E1A dependent up-regulation of c-jun/AP-1 activity." Nucleic Acids Research **19**(3): 649-655.

Knoepfler, P. S. (2007). "Myc goes global: new tricks for an old oncogene." Cancer Research **67**(11): 5061-5063.

Knoepfler, P. S., et al. (2006). "Myc influences global chromatin structure." The EMBO journal **25**(12): 2723-2734.

Kodzius, R., et al. (2006). "CAGE: cap analysis of gene expression." Nature methods **3**(3): 211-222.

Kühl, A. A., et al. (2015). "Diversity of intestinal macrophages in inflammatory bowel diseases." Frontiers in immunology **6**: 613.

Landolin, J. M., et al. (2010). "Sequence features that drive human promoter function and tissue specificity." Genome Research **20**(7): 890-898.

Lang, F. and E. Shumilina (2013). "Regulation of ion channels by the serum-and glucocorticoid-inducible kinase SGK1." The FASEB journal **27**(1): 3-12.

Lee, T. I. and R. A. Young (2013). "Transcriptional regulation and its misregulation in disease." Cell **152**(6): 1237-1251.

Lenhard, B., et al. (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." Nature Reviews Genetics **13**(4): 233.

Lerner, M., et al. (2009). "DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1." Experimental cell research **315**(17): 2941-2952.

Levin, W. J., et al. (1994). "Tumor suppressor and immediate early transcription factor genes in non-small cell lung cancer." Chest **106**(6): 372S-376S.

Li, X., et al. (1997). "A remark on the Mallat pyramidal algorithm of wavelet analysis wavelet analysis." Communications in Nonlinear Science and Numerical Simulation **2**(4): 240-243.

Lin, J. (1991). "Divergence measures based on the Shannon entropy." IEEE Transactions on Information theory **37**(1): 145-151.

Lindgren, V., et al. (1985). "Human U1 small nuclear RNA pseudogenes do not map to the site of the U1 genes in 1p36 but are clustered in 1q12-q22." Molecular and cellular biology **5**(9): 2172-2180.

Liu, X., et al. (2015). "MicroRNA-29a inhibits cell migration and invasion via targeting Roundabout homolog 1 in gastric cancer cells." Molecular medicine reports **12**(3): 3944-3950.

Lizio, M., et al. (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas." Genome biology **16**(1): 22.

Lizio, M., et al. (2017). "Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals." Nucleic Acids Research **45**(D1): D737-D743.

Llorens, F., et al. (2013). "Microarray and deep sequencing cross-platform analysis of the mirRNome and isomiR variation in response to epidermal growth factor." BMC genomics **14**(1): 371.

Lopes, R. H., et al. (2007). "The two-dimensional Kolmogorov-Smirnov test."

Marín-Béjar, O., et al. (2013). "Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2." Genome biology **14**(9): 1.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2010). "Entrez Gene: gene-centered information at NCBI." Nucleic acids research, **39**(suppl_1), D52-D57.

Marshall, C. (1995). "Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation." Cell **80**(2): 179-185.

Martínez, G., et al. (2016). "Regulation of memory formation by the transcription factor XBP1." Cell reports **14**(6): 1382-1394.

Massagué, J. (2012). "TGFβ signalling in context." Nature reviews Molecular cell biology **13**(10): 616-630.

Mathelier, A., et al. (2013). "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles." Nucleic Acids Research **42**(D1): D142-D147.

McCarthy, D. J., et al. (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." Nucleic Acids Research **40**(10): 4288-4297.

McDonald, J. H. (2009). Handbook of biological statistics, Sparky House Publishing Baltimore, MD.

McDowell, I. C., et al. (2018). "Clustering gene expression time series data using an infinite Gaussian process mixture model." PLoS computational biology **14**(1): e1005896.

Mehic, D., et al. (2005). "Fos and jun proteins are specifically expressed during differentiation of human keratinocytes." Journal of investigative dermatology **124**(1): 212-220.

Mills, J. D., et al. (2015). "High expression of long intervening non-coding RNA OLMALINC in the human cortical white matter is associated with regulation of oligodendrocyte maturation." Molecular brain **8**(1): 1.

Mina, M., et al. (2015). "CIDER: a pipeline for detecting waves of coordinated transcriptional regulation in gene expression time-course data." bioRxiv: 012518.

Mina, M., et al. (2015). "Promoter-level expression clustering identifies time development of transcriptional regulatory cascades initiated by ErbB receptors in breast cancer cells." Scientific reports **5**: 11999.

Molina, C. and E. Grotewold (2005). "Genome wide analysis of Arabidopsis core promoters." BMC genomics **6**(1): 25.

Müller, S., et al. (2015). "Next-generation sequencing reveals novel differentially regulated mRNAs, lncRNAs, miRNAs, sdRNAs and a piRNA in pancreatic cancer." Molecular cancer **14**(1): 1.

Murphy, L. O. and J. Blenis (2006). "MAPK signal specificity: the right place at the right time." Trends in biochemical sciences **31**(5): 268-275.

Murphy, L. O., et al. (2004). "A network of immediate early gene products propagates subtle differences in mitogen-activated protein kinase signal amplitude and duration." Molecular and cellular biology **24**(1): 144-153.

Murphy, L. O., et al. (2002). "Molecular interpretation of ERK signal duration by immediate early gene products." Nature cell biology **4**(8): 556.

Myers, R. M., et al. (1986). "Fine structure genetic analysis of a beta-globin promoter." Science **232**(4750): 613-618.

Nason, G. (1993). "The wavethresh package; wavelet transform and thresholding software for S." Available from the StatLib archive.

Nau, G. J., et al. (2002). "Human macrophage activation programs induced by bacterial pathogens." Proceedings of the National Academy of Sciences **99**(3): 1503-1508.

Naslavsky, N., et al. (2006). "Interactions between EHD proteins and Rab11-FIP2: a role for EHD3 in early endosomal transport." Molecular biology of the cell **17**(1): 163-177.

Noguchi, S., et al. (2017). "FANTOM5 CAGE profiles of human and mouse samples." Scientific data **4**: 170112.

Nueda, M. J., et al. (2014). "Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series." Bioinformatics **30**(18): 2598-2602.

O'Donnell, A., et al. (2012). Immediate-early gene activation by the MAPK pathways: what do and don't we know?, Portland Press Limited.

O'Neill, E. and W. Kolch (2004). "Conferring specificity on the ubiquitous Raf/MEK signalling pathway." British journal of cancer **90**(2): 283.

Ono, S. J., et al (1991). "Human X-box-binding protein 1 is required for the transcription of a subset of human class II major histocompatibility genes and forms a heterodimer with c-fos." Proceedings of the National Academy of Sciences **88**(10): 4309-4312.

Orton, R. J., et al. (2005). "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway." Biochemical Journal **392**(2): 249-261.

Park, D.-Y., et al. (2014). "Lymphatic regulator PROX1 determines Schlemm's canal integrity and identity." The Journal of clinical investigation **124**(9): 3960-3974.

Pauli, A., et al. (2012). "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis." Genome Research **22**(3): 577-591.

Perrimon, N., et al. (2012). "Signaling mechanisms controlling cell fate and embryonic patterning." Cold Spring Harbor perspectives in biology **4**(8): a005975.

Price, N. D. and I. Shmulevich (2007). "Biochemical and statistical network models for systems biology." Current opinion in biotechnology **18**(4): 365-370.

Qu, L. H., et al. (1983). "Improved methods for structure probing in large RNAs: a rapid 'heterologous' sequencing approach is coupled to the direct mapping of nudease accessible sites. Application to the 5′ nal domain of eukaryotic 28S rRNA." Nucleic Acids Research **11**(17): 5903-5920.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Rabbani, M. and R. Joshi (2002). "An overview of the JPEG 2000 still image compression standard." Signal processing: Image communication **17**(1): 3-48.

Raines, E. W., et al. (1989). "Interleukin-1 mitogenic activity for fibroblasts and smooth muscle cells is due to PDGF-AA." Science **243**(4889): 393-396.

Raj, T., et al. (2006). "Inhibition of fibroblast growth factor receptor signaling attenuates atherosclerosis in apolipoprotein E-deficient mice." Arteriosclerosis, thrombosis, and vascular biology **26**(8): 1845-1851.

Rauscher, F. d., et al. (1988). "Fos and Jun bind cooperatively to the AP-1 site: reconstitution in vitro." Genes & development **2**(12b): 1687-1699.

Roy, S., et al. (2015). "Redefining the transcriptional regulatory dynamics of classically and alternatively activated macrophages by deepCAGE transcriptomics." Nucleic Acids Research: gkv646.

Rushworth, L. K., et al. (2014). "Dual-specificity phosphatase 5 regulates nuclear ERK activity and suppresses skin cancer by inhibiting mutant Harvey-Ras (HRasQ61L)-driven SerpinB2 expression." Proceedings of the National Academy of Sciences **111**(51): 18267-18272.

Saito, T. H., et al. (2013). "Temporal decoding of MAP kinase and CREB phosphorylation by selective immediate early gene expression." PLoS One **8**(3): e57037.

Sandelin, A., et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic acids research **32**(suppl_1): D91-D94.

Santos, S. D., et al. (2007). "Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate." Nature cell biology **9**(3): 324.

Saravanan, S., et al. (2016). "In Silico Identification of Human miR 3654 and its Targets Revealed its Involvement in Prostate Cancer Progression." MicroRNA **5**(2): 140-145.

Sas-Chen, A., et al. (2012). "A crossroad of microRNAs and immediate early genes (IEGs) encoding oncogenic transcription factors in breast cancer." Journal of mammary gland biology and neoplasia **17**(1): 3-14.

Sati, S., et al. (2012). "Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci." Nucleic Acids Research **40**(20): 10018-10031.

Sato, N., et al. (1997). "Androgenic induction of prostate-specific antigen gene is repressed by protein-protein interaction between the androgen receptor and AP-1/c-Jun in the human prostate cancer cell line LNCaP." Journal of Biological Chemistry **272**(28): 17485-17494.

Shapiro, D. J., et al. (2016). "Anticipatory UPR activation: A protective pathway and target in cancer." Trends in Endocrinology & Metabolism **27**(10): 731-741.

Schena, M., et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schor, I. E., et al. (2017). "Promoter shape varies across populations and affects promoter evolution and expression noise." Nature genetics.

Schratt, G., et al. (2001). "Serum response factor is required for immediate-early gene activation yet is dispensable for proliferation of embryonic stem cells." Molecular and cellular biology **21**(8): 2933-2943.

Selvaraj, A. and R. Prywes (2004). "Expression profiling of serum inducible genes identifies a subset of SRF target genes that are MKL dependent." BMC molecular biology **5**(1): 13.

Severa, M., et al. (2014). "The transcriptional repressor BLIMP1 curbs host defenses by suppressing expression of the chemokine CCL8." The Journal of Immunology: 1301799.

Sheikh, F., et al. (2014). "An essential role for IFN-β in the induction of IFN-stimulated gene expression by LPS in macrophages." Journal of leukocyte biology **96**(4): 591-600.

Shim, H., & Stephens, M. (2015). "Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays." The annals of applied statistics **9**(2), 655.

Shiraki, T., et al. (2003). "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." Proceedings of the National Academy of Sciences **100**(26): 15776-15781.

Sima, C., et al. (2009). "Inference of gene regulatory networks using time-series data: a survey." Current genomics **10**(6): 416-429.

Skalet, A. H., et al. (2005). "Rapid B cell receptor-induced unfolded protein response in nonsecretory B cells correlates with pro-versus antiapoptotic cell fate." Journal of Biological Chemistry **280**(48): 39762-39771.

Skilling, J. (2006). "Nested sampling for general Bayesian computation." Bayesian analysis **1**(4): 833-859.

Spies, D., et al. (2017). "Comparative analysis of differential gene expression tools for RNA sequencing time course data." Briefings in bioinformatics.

Streit, A., et al. (2013). "Experimental approaches for gene regulatory network construction: the chick as a model system." genesis **51**(5): 296-310.

Suzuki, H., et al. (2009). "The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line." Nature genetics **41**(5): 553.

Svaren, J., et al. (1996). "NAB2, a corepressor of NGFI-A (Egr-1) and Krox20, is induced by proliferative and differentiative stimuli." Molecular and cellular biology **16**(7): 3545-3553.

Sweeney, T. E., et al. (2017). "Methods to increase reproducibility in differential gene expression via meta-analysis." Nucleic Acids Research **45**(1): e1-e1.

Temperley, R. J., et al. (2010). "Human mitochondrial mRNAs—like members of all families, similar but different." Biochimica et Biophysica Acta (BBA)-Bioenergetics **1797**(6): 1081-1085.

Thalhauser, C. J. and N. L. Komarova (2009). "Specificity and robustness of the mammalian MAPK-IEG network." Biophysical journal **96**(9): 3471-3482.

Theodosiou, A. and A. Ashworth (2002). "MAP kinase phosphatases." Genome biology **3**(7): reviews3009. 3001.

Thorne, T. (2018). "Approximate inference of gene regulatory network models from RNA-Seq time series data." BMC bioinformatics **19**(1): 127.

Trapnell, C., et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature biotechnology **28**(5): 511.

Treisman, R. (1996). "Regulation of transcription by MAP kinase cascades." Current opinion in cell biology **8**(2): 205-215.

Trinklein, N. D., et al. (2003). "Identification and functional analysis of human transcriptional promoters." Genome Research **13**(2): 308-312.

Tsuru, A., et al. (2016). "Novel mechanism of enhancing IRE1α-XBP1 signalling via the PERK-ATF4 pathway." Scientific reports **6**: 24217.

Tullai, J. W., et al. (2007). "Immediate-early and delayed primary response genes are distinct in function and genomic architecture." Journal of Biological Chemistry **282**(33): 23981-23995.

Vaquerizas, J. M., et al. (2009). "A census of human transcription factors: function, expression and evolution." Nature reviews. Genetics **10**(4): 252.

Volinsky, N., et al. (2015). "Signalling mechanisms regulating phenotypic changes in breast cancer cells." Bioscience reports **35**(2): e00178.

Wagner, E. (2010). "Bone development and inflammatory disease is regulated by AP-1 (Fos/Jun)." Annals of the rheumatic diseases **69**(Suppl 1): i86-i88.

Walker, J. S. (2008). A primer on wavelets and their scientific applications, CRC press.

Walker, M. G., et al. (1999). "Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes." Genome Research **9**(12): 1198-1203.

Wang, L., et al. (2014). "Linc00963: A novel, long non-coding RNA involved in the transition of prostate cancer from androgen-dependence to androgen-independence." International journal of oncology **44**(6): 2041-2049.

Wick, M., et al. (1994). "Identification of serum-inducible genes: different patterns of gene regulation during G0--> S and G1--> S progression." Journal of cell science **107**(1): 227-239.

Wigle, J. T. and G. Oliver (1999). "Prox1 function is required for the development of the murine lymphatic system." Cell **98**(6): 769-778.

Will, C. L. and R. Lührmann (2011). "Spliceosome structure and function." Cold Spring Harbor perspectives in biology **3**(7): a003707.

Winkles, J. A. (1997). Serum-and polypeptide growth factor-inducible gene expression in mouse fibroblasts. Progress in nucleic acid research and molecular biology, Elsevier. **58:** 41-78.

Xu, C., et al. (2005). "Endoplasmic reticulum stress: cell life and death decisions." The Journal of clinical investigation **115**(10): 2656-2664.

Xu, N., et al. (2015). "Induction of GADD34 Regulates the Neurotoxicity of Amyloid β." American journal of Alzheimer's disease and other dementias **30**(3): 313-319.

Yamamoto, M., et al. (2004). "Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein IκBζ." Nature **430**(6996): 218.

Yoshida, H., et al. (2001). "XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor." Cell **107**(7): 881-891.

Yoshida, H., et al. (2006). "XBP1 is critical to protect cells from endoplasmic reticulum stress: evidence from Site-2 protease-deficient Chinese hamster ovary cells." Cell structure and function **31**(2): 117-125.

Yoshida, T. and K. Georgopoulos (2014). "Ikaros fingers on lymphocyte differentiation." International journal of hematology **100**(3): 220-229.

Young, M. R., et al. (1999). "Transgenic mice demonstrate AP-1 (activator protein-1) transactivation is required for tumor promotion." Proceedings of the National Academy of Sciences **96**(17): 9827-9832.

Zamore, P. D. and B. Haley (2005). "Ribo-gnome: the big world of small RNAs." Science **309**(5740): 1519-1524.

Zhang, P., et al. (2017). "Relatively frequent switching of transcription start sites during cerebellar development." BMC genomics **18**(1): 461.

Zhang, R., et al. (2011). "Elevated expression of c-fos in central nervous system correlates with visceral hypersensitivity in irritable bowel syndrome (IBS): a new target for IBS treatment." International journal of colorectal disease **26**(8): 1035-1044.

Zhang, Y. and S. Zhou (2015). "MicroRNA-29a inhibits mesenchymal stem cell viability and proliferation by targeting Roundabout 1." Molecular medicine reports **12**(4): 6178-6184.