THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

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

OPEN ACCESS

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

Short title: Promoter activation in the immediate early response

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[^], Stuart Aitken[1], 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, RIKEN Yokohama Campus, Yokohama, 230-0045, Japan

[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 St, Nedlands, Western Australia 6009, Australia

*Corresponding Author

Email: Colin.Semple@igmm.ed.ac.uk

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. Whereas genomic dissection on a case by case basis has not resulted in a comprehensive catalogue of immediate early genes, a rigorous meta-analysis of eight genome-wide FANTOM5 CAGE (Cap Analysis of Gene Expression) time course data sets 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 pair-wise 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. In 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.

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 minutes of a stimulus and subsides after 90 minutes [3], in contrast with its continuous overexpression in many cancers.

Immediate-early genes 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 transcription factors, 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 since 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 (FANTOM Consortium, Nature, 2014; http://fantom.gsc.riken.jp/5/). CAGE data are thus ideally suited to studying the strong burst of transcription at promoters seen in immediate-early responses. 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.

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 1a). 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. (2015) [10]. As expected, the responses of known immediate early genes 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 1b). These observations illustrate the challenges presented in IEG detection, even when studying known IEGs using a uniform experimental platform.

Optimising and refining the approach developed by Aitken et al (2015) [7] (see 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 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.

CAGE 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 7,706 Ensembl genes) and 14,376 (8,951 genes) protein-coding CAGE TSSs, depending on the dataset, and between 1,202 (692 genes) and 1,640 (858 genes) non-coding RNA 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, 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 8,785 Ensembl genes (72%) to peak in at least one of the datasets. In contrast, few genes were classified to the peak model in multiple datasets, with only 42 of such genes shared across at least seven datasets (Figure 1c), 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 1,304 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 3 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 2a) for all promoters classified as peaks in at least four datasets (the permissive set) readily demonstrated common patterns across datasets (Figure 2b). 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 2b). 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 Methods) within each class, for each dataset. The peak class was enriched for known IEGs in all datasets (electronic supplementary material, Figure S4, 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 4 or more datasets) expected to contain higher numbers of false positives than the robust set (7 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; 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).

Table 1. Enrichment of known IEGs for genes classified to the peak model in multiple datasets

Shared datasets	IEGs enrichment						# CAGE TSSs (median)
	# genes	# IEGs	# CAGE TSSs (across 8 datasets)	# IEG CAGE TSSs (across 8 datasets)	OR	p	
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

Enrichment (expressed as odds ratios) and p values for genes classified across different numbers of time series datasets.

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 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 the post transcriptional modification of many snRNPs. Kalam et al. have shown that macrophage infection with *Mycobacterium tuberculosis* results in the systematic perturbation in splicing patterns [14], 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 5 peaking TSS (Table 2; electronic supplementary material, Figure 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 [15] and HeLa [16] cells. miR-29A has

been associated with the viability and proliferation of mesenchymal stem cell and gastric cancer cells [17, 18] 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 [19-21]. 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 [22]. Although poorly characterised, 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 [23-26].

Table 2. Non-coding genes peaking in at least 7 out of 8 datasets.

Gene ID	Nº Shared Datasets	Description (PuMed ref.)
LINC00478 (MIR99AHG)	7	It has a role in cell proliferation and differentiation and it is considered a regulator of oncogenes in leukemia (PMID: 25027842)
LINC00263	7	Regulation of oligodendrocyte maturation (PMID: 25575711)
LINC-PINT	8	Putative tumor suppressor (PMID: 24070194)
LINC00963	7	Involved in the prostate cancer transition from androgen-dependent to androgen-independent and metastasis via the EGFR signaling 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)

The short descriptions of the molecular function are from the genecard database [27].

Known IEG promoters show conserved temporal order of activation across datasets

Having established common patterns of peak gene induction at similar times across datasets (Figure 2b), we hypothesised 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. Firstly, the presence of conserved gene orderings, in addition to common gene classifications, provides an additional test for functional similarity between datasets. Secondly, 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 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 < 5e-3) by comparison to the distribution of expected connections given 1,000,000 permuted datasets (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.

Known IEGs and candidate IEGs 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 Ca2+-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 S8A, Wilcoxon $p < 2.2e\text{-}16$), however, some candidate IEGs promoters show notably similar timing (electronic supplementary material, Figure S8C, 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 material, Figure S8B, Wilcoxon $p < 0.05$) but they

16

occur earlier than known IEGs (electronic supplementary material, Figure S8D, Wilcoxon p < 0.05 for all datasets).

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 [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 nonclassical 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 endoplasmic reticulum 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 Methods) of the robust set of genes (electronic supplementary material, Figure S9).

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 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 [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.

Materials and methods

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. (2015) [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 >10 TPM (tags per million), while CAGE clusters corresponding to ncRNA were thresholded to >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/).

Model-based classification of TSS 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, logZ. 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

19

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, $t_p = t_d + t_s$.

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

$$(1)$$

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

$$(2)$$

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

$$(3)$$

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

$$(4)$$

However, an alternative rate $\delta = \frac{\log(0.1)}{t_d}$ was also used to model the slower dynamics of transcripts peaking later in time, and the best fitting model selected during the decision step. Normalising 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 categorisation decision.

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 ≤ 0.05) were sought in flanking 400bp windows centred on the middle of each CAGE TSS analysed), using FIMO [41] from the MEME package (version 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 ≤ 0.05).

Pathway and Gene Ontology enrichment

Functional and pathway enrichments were assessed using GOrilla [13] and InnateDB [28] respectively (FDR ≤ 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 was assigned using Fisher's exact test.

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 ($t_{p\ gene-A} <$ $t_{p\ gene-B}$), and this order is observed in 6 or more of the other 7 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 4,516 out of 1,000,000 permutations indicating that the number of temporal connections was statistically significant ($p < 5e-3$).

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, N-7006 Trondheim, Norway) curated the list of IEGs within the FANTOM5 consortium.

Authors Contributions

AV carried out the computational analysis and drafted the manuscript; CAS and SA designed the study and drafted the manuscript; MI, HK, EA, TL, COD, PC and ARRF generated and processed the CAGE data coordinated by YH. All authors gave final approval for publication.

References

1.   Fowler T, Sen R, Roy AL. Regulation of primary response genes. Molecular cell. 2011;44(3):348-60.
2.   Healy S, Khan P, Davie JR. Immediate early response genes and cell transformation. Pharmacology & therapeutics. 2013;137(1):64-77.
3.   Greenberg ME, Ziff EB. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. Nature. 1984;311:433-8.
4.   Bahrami S, Drabløs F. Gene regulation in the immediate-early response process. Advances in Biological Regulation. 2016.
5.   O'Donnell A, Odrowaz Z, Sharrocks AD. Immediate-early gene activation by the MAPK pathways: what do and don't we know? : Portland Press Limited; 2012.
6.   Fowler T, Suh H, Buratowski S, Roy AL. Regulation of primary response genes in B cells. Journal of Biological Chemistry. 2013;288(21):14906-16.
7.   Aitken S, Magi S, Alhendi AM, Itoh M, Kawaji H, Lassmann T, et al. Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response. PLoS Comput Biol. 2015;11(4):e1004217.
8.   Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.
9.   Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nature Reviews Genetics. 2012;13(8):552-64.
10.   Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347(6225):1010-4.
11.   Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The Ensembl genome database project. Nucleic acids research. 2002;30(1):38-41.
12.   Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one. 2011;6(7):e21800.
13.   Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics. 2009;10(1):1.
14.   Kalam H, Fontana MF, Kumar D. Alternate splicing of transcripts shape macrophage response to Mycobacterium tuberculosis infection. PLoS pathogens. 2017;13(3):e1006236.
15.   Avraham R, Sas-Chen A, Manor O, Steinfeld I, Shalgi R, Tarcic G, et al. EGF decreases the abundance of microRNAs that restrain oncogenic transcription factors. Sci Signal. 2010;3(124):ra43.
16.   Llorens F, Hummel M, Pantano L, Pastor X, Vivancos A, Castillo E, et al. Microarray and deep sequencing cross-platform analysis of the mirRNome and isomiR variation in response to epidermal growth factor. BMC genomics. 2013;14(1):371.
17.   Liu X, Cai J, Sun Y, Gong R, Sun D, Zhong X, et al. MicroRNA-29a inhibits cell migration and invasion via targeting Roundabout homolog 1 in gastric cancer cells. Molecular medicine reports. 2015;12(3):3944-50.

18.   Zhang Y, Zhou S. MicroRNA-29a inhibits mesenchymal stem cell viability and proliferation by targeting Roundabout 1. Molecular Medicine Reports. 2015;12(4):6178-84.

19.   Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(39):13944-9.

20.   Lerner M, Harada M, Lovén J, Castro J, Davis Z, Oscier D, et al. 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. 2009;315(17):2941-52.

21.   Gao S-m, Xing C-y, Chen C-q, Lin S-s, Dong P-h, Yu F-j. miR-15a and miR-16-1 inhibit the proliferation of leukemic cells by down-regulating WT1 protein level. Journal of Experimental & Clinical Cancer Research. 2011;30(1):1.

22.   Gökmen-Polar Y, Zavodszky M, Chen X, Gu X, Kodira C, Badve S. Abstract P2-06-05: LINC00478: A novel tumor suppressor in breast cancer. Cancer Research. 2016;76(4 Supplement):P2-06-5-P2--5.

23.   Mills JD, Kavanagh T, Kim WS, Chen BJ, Waters PD, Halliday GM, et al. High expression of long intervening non-coding RNA OLMALINC in the human cortical white matter is associated with regulation of oligodendrocyte maturation. Molecular brain. 2015;8(1):1.

24.   Müller S, Raulefs S, Bruns P, Afonso-Grunz F, Plötner A, Thermann R, et al. Next-generation sequencing reveals novel differentially regulated mRNAs, lncRNAs, miRNAs, sdRNAs and a piRNA in pancreatic cancer. Molecular cancer. 2015;14(1):1.

25.   Marín-Béjar O, Marchese FP, Athie A, Sánchez Y, González J, Segura V, et al. Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2. Genome biology. 2013;14(9):1.

26.   Wang L, Han S, Jin G, Zhou X, Li M, Ying X, et al. Linc00963: A novel, long non-coding RNA involved in the transition of prostate cancer from androgen-dependence to androgen-independence. International journal of oncology. 2014;44(6):2041-9.

27.   Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from Gene data mining to disease genome sequence analyses. Current Protocols in Bioinformatics. 2016:1.30. 1-1.. 3.

28.   Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic acids research. 2012:gks1147.

29.   Schratt G, Weinhold B, Lundberg AS, Schuck S, Berger J, Schwarz H, et al. Serum response factor is required for immediate-early gene activation yet is dispensable for proliferation of embryonic stem cells. Molecular and cellular biology. 2001;21(8):2933-43.

30.   Treisman R. Regulation of transcription by MAP kinase cascades. Current opinion in cell biology. 1996;8(2):205-15.

31.   Andruska N, Zheng X, Yang X, Helferich WG, Shapiro DJ. Anticipatory Estrogen Activation of the Unfolded Protein Response is Linked to Cell Proliferation and Poor Survival in Estrogen Receptor α Positive Breast Cancer. Oncogene. 2015;34(29):3760.

32.   Shapiro DJ, Livezey M, Yu L, Zheng X, Andruska N. Anticipatory UPR activation: A protective pathway and target in cancer. Trends in Endocrinology & Metabolism. 2016;27(10):731-41.

33.   Skalet AH, Isler JA, King LB, Harding HP, Ron D, Monroe JG. Rapid B cell receptor-induced unfolded protein response in nonsecretory B cells correlates with pro-versus antiapoptotic cell fate. Journal of Biological Chemistry. 2005;280(48):39762-71.

34.   He Y, Sun S, Sha H, Liu Z, Yang L, Xue Z, et al. Emerging roles for XBP1, a sUPeR transcription factor. Gene expression. 2010;15(1):13-25.

35.   Piperi C, Adamopoulos C, Papavassiliou AG. XBP1: a pivotal transcriptional regulator of glucose and lipid metabolism. Trends in Endocrinology & Metabolism. 2016;27(3):119-22.

36.   Hu E, Mueller E, Oliviero S, Papaioannou V, Johnson R, Spiegelman B. 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. 1994;13(13):3094.

37.   Fei J, Viedt C, Soto U, Elsing C, Jahn L, Kreuzer J. Endothelin-1 and Smooth Muscle Cells. Arteriosclerosis, thrombosis, and vascular biology. 2000;20(5):1244-9.

38.   Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome biology. 2015;16(1):22.

39.   Aitken S, Akman OE. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. BMC systems biology. 2013;7(1):72.

40.   Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic acids research. 2013;42(D1):D142-D7.

41.   Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-8.
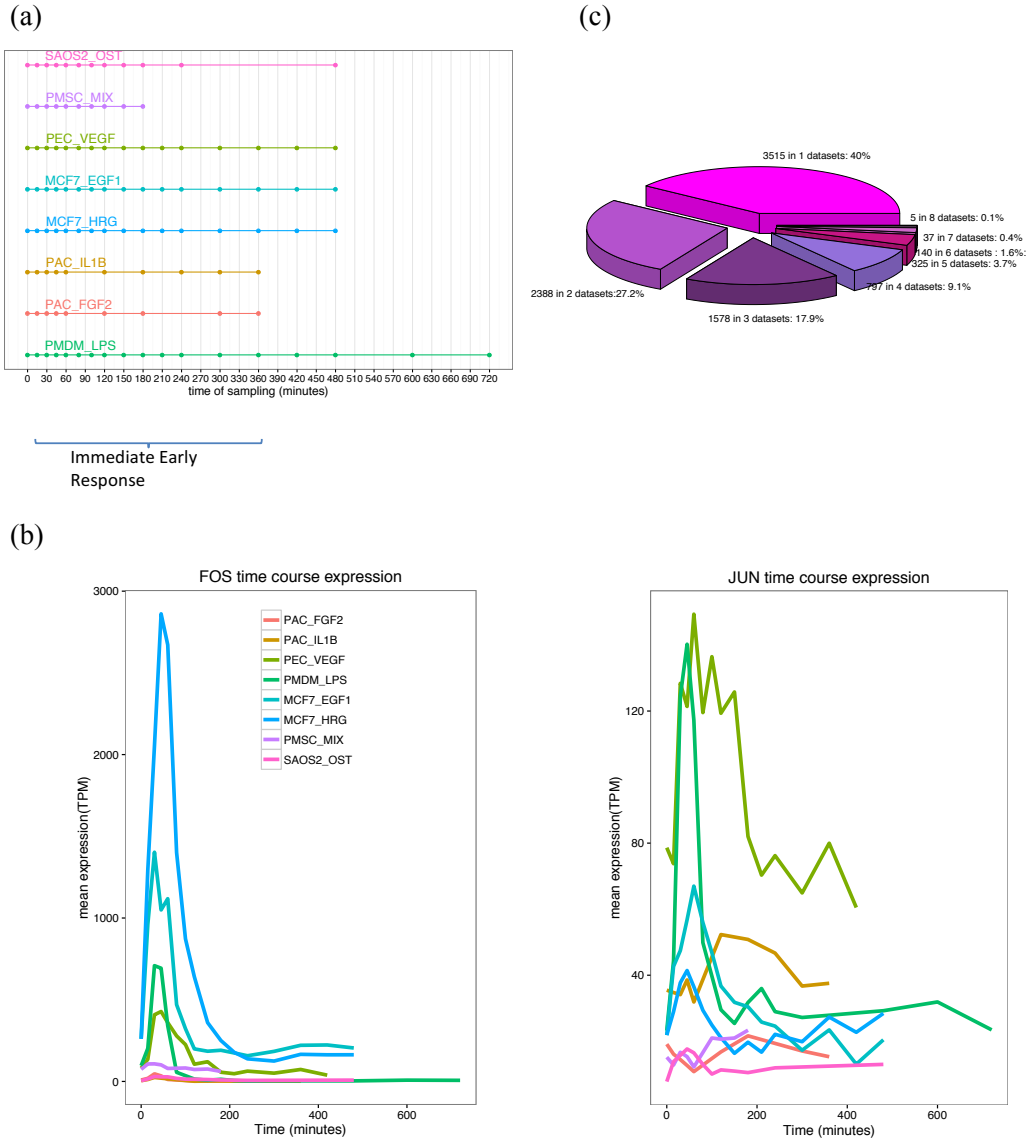
Figure 1. Time course datasets demonstrating the immediate early response. (a) Schematic of the eight time course datasets considered, horizontal lines indicate the time span and symbols show the sampling times. Time zero corresponds to inactivated or quiescent cells in all cases. (b) The time course expression profile of FOS (left) and JUN (right) 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 2 or more datasets.
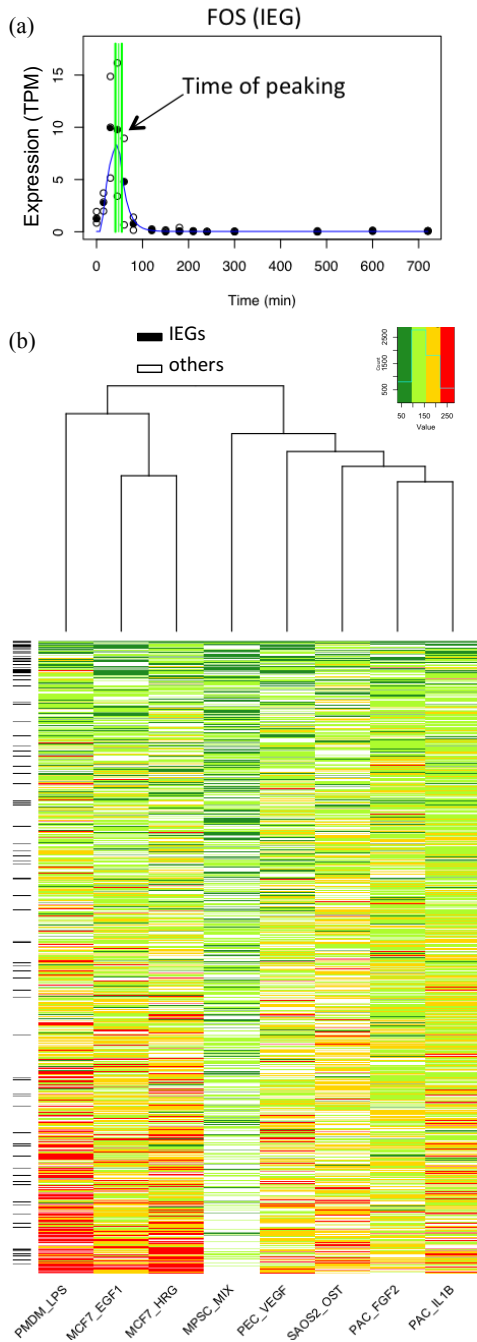
Figure 2. Broad trends in peak expression times across datasets. (a) Identification of the peak time parameter (t_p) of FOS estimated from the PMDM_LPS time series (filled symbols indicate the median TPM; unfilled symbols are individual replicates; green lines represent t_p and one standard deviation 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.
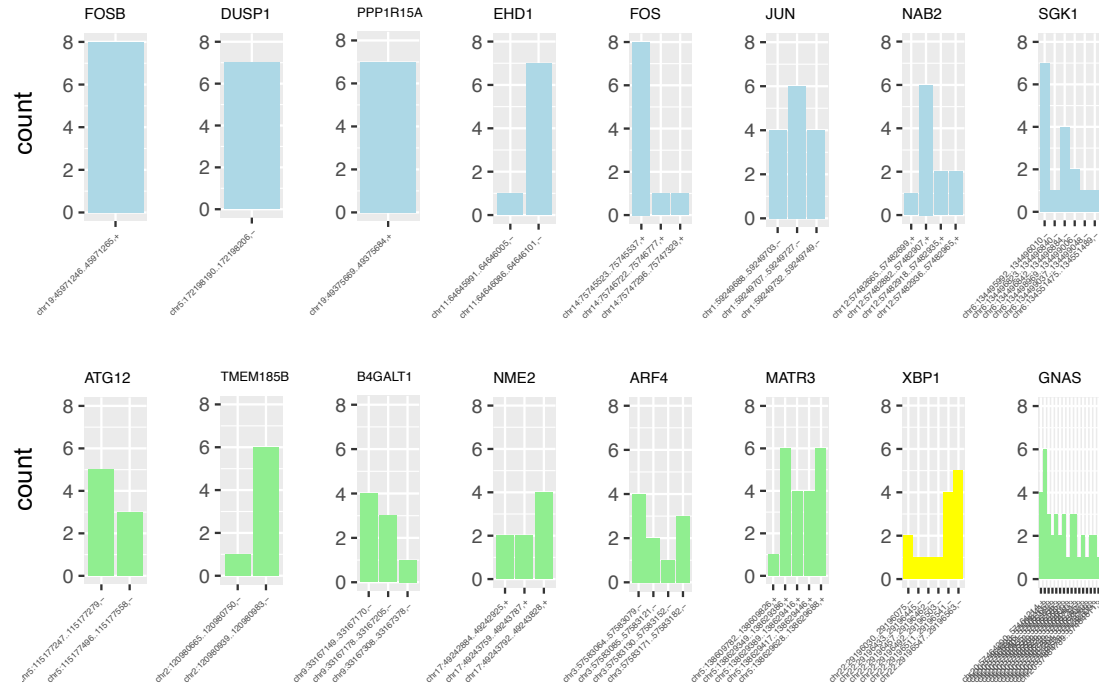
Figure 3. Promoter usage across time series datasets. For representative genes, barcharts show the number of datasets where each TSS peaks to illustrate the diversity of TSS usage and commonality of the peaking response. Known IEGS are shown in blue, transcription factors in yellow and other genes in green. FOSB has a single TSS that peaks in 8 datasets, JUN has three TSS each peaking in 4 or more datasets, and XBP1 has 6 TSS that peak in between 1 and 6 datasets.
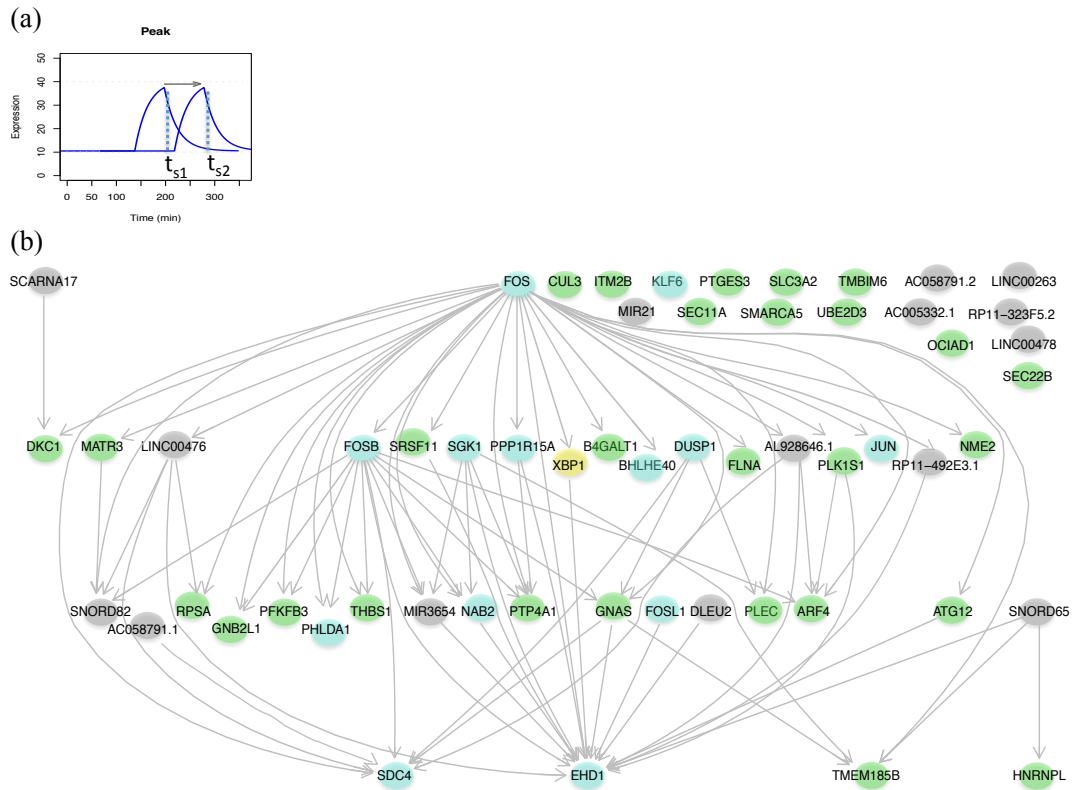
Figure 4. Conserved activation network. (a) Schematic profiles of two peaking genes, with temporal precedence indicated by the arrow. (b) Conserved temporal precedence between IEGs (light blue nodes), transcription factors (yellow nodes) non-coding RNA (grey nodes) and other protein-coding genes (green nodes) is shown by directed edges. A subset of IEGs in this network are also transcription factors (FOS, KLF6, FOSB, BHLHE40, JUN, and FOSL1).
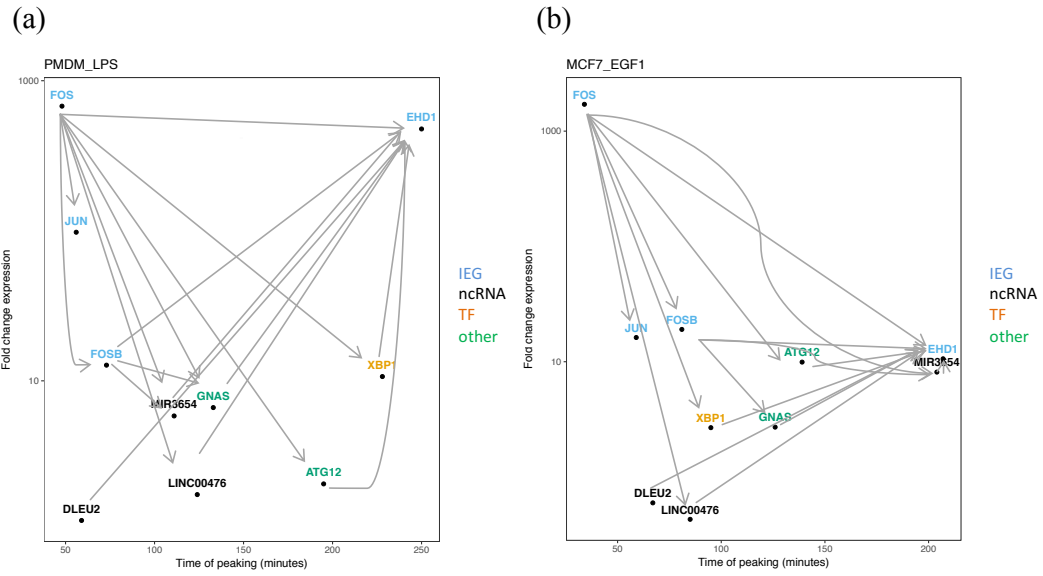
Figure 5. Transcriptional dynamics of genes classified to the peak model. Scatterplots of log fold change against the time of peaking for selected genes, with conserved temporal precedence indicated by arrows for PMDM_LPS (a) and MCF7_EGF1 (b). FOS peaks earliest and has many conserved temporal relations to later peaking genes, while EHD1 peaks late and has many conserved temporal orderings with earlier peaking genes.