SYSTEMS BIOLOGY OF TRANSCRIPTION REGULATION

EDITED BY: Ekaterina Shelest, Edgar Wingender and Joerg Linde
PUBLISHED IN: Frontiers in Genetics, Frontiers in Plant Science and Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-967-9
DOI 10.3389/978-2-88919-967-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

SYSTEMS BIOLOGY OF TRANSCRIPTION REGULATION

Topic Editors:
Ekaterina Shelest, Hans-Knoell Institute, Germany
Edgar Wingender, University of Göttingen, Germany
Joerg Linde, Hans-Knoell Institute, Germany

Transcription regulation is a complex process that can be considered and investigated from different perspectives. Traditionally and due to technical reasons (including the evolution of our understanding of the underlying processes) the main focus of the research was made on the regulation of expression through transcription factors (TFs), the proteins directly binding to DNA. On the other hand, intensive research is going on in the field of chromatin structure, remodeling and its involvement in the regulation. Whatever direction we select, we can speak about several levels of regulation. For instance, concentrating on TFs, we should consider multiple regulatory layers, starting with signaling pathways and ending up with the TF binding sites in the promoters and other regulatory regions. However, it is obvious that the TF regulation, also including the upstream processes, represents a modest portion of all processes leading to gene expression. For more comprehensive description of the gene regulation, we need a systematic and holistic view, which brings us to the importance of systems biology approaches.

Advances in methodology, especially in high-throughput methods, result in an ever-growing mass of data, which in many cases is still waiting for appropriate consideration. Moreover, the accumulation of data is going faster than the development of algorithms for their systematic evaluation. Data and methods integration is indispensable for the acquiring a systematic as well as a systemic view. In addition to the huge amount of molecular or genetic components of a biological system, the even larger number of their interactions constitutes the enormous complexity of processes occurring in a living cell (organ, organism). In systems biology, these interactions are represented by networks.

Transcriptional or, more generally, gene regulatory networks are being generated from experimental ChIPseq data, by reverse engineering from transcriptomics data, or from computational predictions of transcription factor (TF) – target gene relations. While transcriptional networks are now available for many biological systems, mathematical models to simulate their dynamic behavior have been successfully developed for metabolic and, to some extent, for signaling networks, but relatively rarely for gene regulatory networks.

Systems biology approaches provide new perspectives that raise new questions. Some of them address methodological problems, others arise from the newly obtained understanding of the data. These open questions and problems are also a subject of this Research Topic.

Citation: Shelest, E., Wingender, E., Linde, J., eds. (2016). Systems Biology of Transcription Regulation. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-967-9

Table of Contents

Chapter III

frontiers in Genetics

EDITORIAL
published: 06 July 2016
doi: 10.3389/fgene.2016.00124

Editorial: Systems Biology of Transcription Regulation

Ekaterina Shelest[1]* *and Edgar Wingender*[2]

[1] *Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knoell Institute, Jena, Germany,* [2] *Institute of Bioinformatics, University Medical Center Goettingen, Goettingen, Germany*

Keywords: systems biology, transcription regulation, regulatory networks, modeling

The Editorial on the Research Topic

Systems Biology of Transcription Regulation

Systems biology (SB) is a holistic approach, an attempt to view a living system in its integrity. A system is thus considered as more than just a sum of its parts; interactions bring their flavor. Transcription regulation is in a way ideal for application of systems biology approaches, because it is complex and because it is a regulatory system. The latter puts it right in the middle of SB efforts, because regulation is central to any system: without regulation a system loses connections, its "systemic" property. Focusing on SB of transcriptional regulation, as we do in this Research Topic, is not stepping back into a reductionist approach. The complete signature of gene activities, their control, and consequences rather represents the status of a living system, for instance a single cell, in a comprehensive way. Here, we are in a good position to investigate the properties and patterns of regulatory circuits on different levels, from transcription regulation networks (TRNs) and signaling pathways to intercellular crosstalk, development, and further to physiological function on tissue and organism level—to that extent in which it depends on gene expression and its regulation.

That is more or less a perspective. Systems biology of transcription regulation, as any other systems biology, is not yet a field with a well-established set of standard methods. It is also not a field with well-defined borders and unambiguously understood content. On the one hand, the subject is too complex and simultaneously too broad, which opens a wide field of activity. On the other hand, regulation of transcription is since long in the focus of intensive research and understanding of some (usually quite narrow) parts of it is very much advanced. There is also a historical bias toward some "favorite" processes, model organisms, where we can find examples of amazing advances; however, for other, not yet well investigated processes we are often just at the stage of collecting "bricks" from which the future building of our understanding will be constructed.

This status of the SB of transcription regulation is reflected by the collection of articles in this issue. We can see the variety of views, methods, applications, and questions raised and answered: from application of state-of-the-art methods to a particular object (e.g., Wlochowitz et al.) to development of novel methods (Wachter and Beissbarth; Martignetti et al.), from discussions of critical methodological and technical issues (e.g., Madrigal) to detailed analysis of robustness mechanisms (Payne and Wagner), from first descriptions of pathways in a non-model plant (Iaria et al.) to advanced SB in well-established models (e.g., Ben-Tabou de-Leon, etc.). Let us briefly go through this collection.

For transcription regulation, at least in the part considering transcription factors (TFs), TF binding sites (TFBSs) form the basis of the pyramid. Boeva in her review leads us through the forest of existing tools for prediction of motifs and TFBSs, demonstrating in the end how application of these methods can improve the accuracy of peak-calling in CHIPSeq. TFBSs are also in the focus of the investigation of heart development regulation (Zeidler et al.). The findings suggest that TF interactions are stage-specific and support the hourglass model of heart development. Wlochowitz et al. apply the state-of-the-art tools, such as Trinity (Grabherr et al., 2011) and geneXplain (http://genexplain-platform.com/bioumlweb/),

OPEN ACCESS

Edited by:
Richard D. Emes,
University of Nottingham, UK

Reviewed by:
Ka-Chun Wong,
City University of Hong Kong, China

***Correspondence:**
Ekaterina Shelest
ekaterina.shelest@hki-jena.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 May 2016*
Accepted: *22 June 2016*
Published: *06 July 2016*

Citation:
Shelest E and Wingender E (2016) Editorial: Systems Biology of Transcription Regulation. Front. Genet. 7:124. doi: 10.3389/fgene.2016.00124

to find differences between two cancer cell lines in terms of master TFs and signaling pathways. Analyzing gene regulatory networks (GRNs) and pathway interplays, the authors come to the explanation of the invasive potential of different cancer. Transcriptome analysis is also central for the papers of Iaria et al. and Smita et al. In the former, gene expression was monitored during maturation of fruits in two olive cultivars, followed by comparative analysis and reconstruction of metabolic pathways involved in olive drupe development. This is a nice example of tissue-specific functional genomics in a non-model plant species. Smita et al. used "top-down" and "guide-gene" approaches to study transcriptome-based GRN of MYB TFs in rice. The observations of differential regulation of all 233 rice MYBs in GEO-derived microarray data along with the phylogenetic analysis demonstrated that phylogenetically close pairs of MYB TFs are involved in highly similar regulatory processes.

Bringing together different data layers is a typical SB challenge. In our Topic, we have two papers suggesting interesting approaches to it. Wachter and Beissbarth draw our attention to the fact that a lot of cellular signaling information is encoded in signaling dynamics. To take this into account, the authors suggest a novel pathway-based method for the analysis of coupled omics time-series data through inferring consensus profiles and time profile clusters. Another approach suggested by Offermann et al. is based on dynamic Boolean models inferred from time-resolved transcriptomes, protein, and phenotypic data. The models can be further optimized by fitting to experimental data and finally can describe temporal resolution of network events (regulation–transcription–feedback). Interestingly, in both papers the methods were applied to describe the same pathway, epidermal growth factor (EGF) signaling. Some new promising interactions were suggested by the first method. In the second application, EGF was confronted with NGF signaling with a very interesting outcome, suggesting that positive transcriptional feedback induces bistability in the switch between differentiation and proliferation, moreover, differentiation uses three redundant pathways.

A less typical problem is tackled by Martignetti et al.: how to estimate activity of genes based on expression data, for instance the activity of a TF from expression of its target genes? For that, the authors developed a software ROMA for quantification of the activity of gene sets with coordinated expression. Application examples demonstrate that the activity of a signaling pathway is better reflected by the set of regulated genes than by any of these genes taken individually, which is an important message for future SB applications.

The paper of Lizio et al. introduces experimental strategies to build cell-type specific TRNs. The authors use complementary approaches (CHIPseq, KD-CAGE) to identify genome-wide targets of genes of interest and warn about the problems that may arise by the usage of CHIPseq alone. This critical view is very important. Another kind of concern is expressed in the opinion paper of Madrigal, who raises a discussion of such serious issue as sequence-specific bias in chromatin assembly experiments. Indeed, this issue can be easily overlooked, and it is essential to be aware of the dangers of sequence (or any other) biases when designing an experiment or treating the results. Madrigal describes the types of bias in different analyses and the adequacy of current benchmarks.

The problem of reproducibility of individual analyses is raised by Berto et al. To extract the most confident and biologically relevant information, the authors developed a method for integration of independently derived networks into a consensus network. This approach was applied to such complex and highly variable systems as cognitive disorders.

Understanding of such properties as robustness can be only addressed from systemic perspective, making it central topic of several presented here papers. Payne and Wagner in their comprehensive review analyze the mechanisms of mutational robustness, discussing its causes and consequences. Another type of robustness—temporal control of developmental GRNs—is discussed by Ben-Tabou de-Leon. Analysis of network motifs helps us to understand how the network architecture supports the timely activation of regulatory and differentiation genes. Rigid motif combinations, such as a triple positive feedback loop conserved through bilateral, explain the robustness of the system, and suggest that this "approach" can be used in other systems as well.

Altogether, this comprehensive collection of articles provides a nice overview of the present status of SB of transcription regulation, demonstrating the advances in different areas achieved through the application of SB approaches.

AUTHOR CONTRIBUTIONS

ES and EW have read all Research Topic articles, ES drafted the review, both authors wrote the paper and approved it for publication.

ACKNOWLEDGMENTS

This work was supported by the MetastaSys project (0316173A) within the ebio initiative of the German Ministry of Education and Research (BMBF). ES was supported by CRC 1127 ChemBioSys and CRC-Transregio FungiNet by Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

OPINION
published: 22 September 2015
doi: 10.3389/fbioe.2015.00144

On accounting for sequence-specific bias in genome-wide chromatin accessibility experiments: recent advances and contradictions

Pedro Madrigal[1,2]*

[1] Wellcome Trust Sanger Institute, Cambridge, UK, [2] Department of Surgery, University of Cambridge, Cambridge, UK

Keywords: next-generation sequencing, DNase-seq, ATAC-seq, chromatin accessibility, footprinting, sequence bias, ChIP-exo

OPEN ACCESS

Edited by:
Ekaterina Shelest, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell Institute, Germany

Reviewed by:
Gaurav Sablok, Istituto Agrario San Michele, Italy
Uwe Ohler, Max Delbrueck Center, Germany

***Correspondence:**
Pedro Madrigal
pm12@sanger.ac.uk

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 June 2015*
Accepted: *07 September 2015*
Published: *22 September 2015*

Citation:
Madrigal P (2015) On accounting for sequence-specific bias in genome-wide chromatin accessibility experiments: recent advances and contradictions. Front. Bioeng. Biotechnol. 3:144. doi: 10.3389/fbioe.2015.00144

Next-Generation Sequencing for Chromatin Biology

Uncovering the protein–DNA interactions involved in cell fate, development, and disease in a time- and cell-specific manner is a fundamental goal of molecular biology. The advent of the sequencing technologies has opened a new genomic era, uncovering the information encoded in genomes, epigenomes, and transcriptomes (McPherson, 2014). For example, the popular ChIP-based techniques ChIP-seq (Johnson et al., 2007; Robertson et al., 2007) and ChIP-exo (Rhee and Pugh, 2011) are widely used to detect transcription factor (TF)-binding sites using an antibody against a single protein of interest (Mahony and Pugh, 2015). Alternative protocols assaying the chromatin landscape, such as those based on digestion by DNase I enzyme (DNase-seq), micrococcal nuclease (MNase-seq), and Tn5 transposase attack (ATAC-seq), enable the identification of DNA-binding protein footprints of many TFs in a single experiment (Tsompana and Buck, 2014). Time-series experiments might be required for the identification of those TFs cataloged as pioneer factors, allowing their effects on chromatin to be investigated (Zaret and Carroll, 2011; Pajoro et al., 2014; Sherwood et al., 2014).

Despite the initial promise of detecting the majority of TFs in one assay, DNA sequence-specific biases, together with TF-dependent binding kinetics, have been recently pinpointed as major confounding factors in DNase-seq experiments (Koohy et al., 2013; He et al., 2014; Raj and McVicker, 2014; Rusk, 2014; Sung et al., 2014). These influencing factors were not considered by any of the previous computational approaches for the analysis of next-generation sequencing chromatin accessibility data (Madrigal and Krajewski, 2012); neither those strategies based on TF-generic DNase signature nor those based on TF-specific DNase signature (Luo and Hartemink, 2013).

Alleviating Sequence-Specific Biases in DNase-seq

To partly address these challenges, four recent approaches have been published that model, predict, or explain DNase I sequence specificity in order to improve the detection of TF occupancy events at high resolution (digital genomic footprinting). The first method, FootprintMixture, uses a multinomial mixture model in which one mixture models the footprint component, and the other the background component taking into account the sequence bias (Yardimci et al., 2014). The background can be either uniform or derived from naked DNA measurements – this is the main difference with respect to the footprint component in CENTIPEDE (Pique-Regi et al., 2011), which assumes a uniform background. Alternatively, more than two components may be set to detect variability in the footprint model. Thus, the cleavage signature (number of DNase I cuts that map

to each nucleotide) is used in a multinomial mixture model to classify candidate sites as either "bound" or "unbound" aided by 6-mer DNase sequence bias cleavage frequencies (Yardimci et al., 2014). Remarkably, the authors found that sequence bias is DNase-seq protocol specific. They also found that the signature of a footprint could be formed by a mixture of DNase digestion profiles identified by unsupervised *k*-means clustering, in agreement with the observations found in an earlier study (Tewari et al., 2012). For TFs CTCF and ZNF143, variants of the consensus sequence motif associated to different footprint shapes were observed.

In the second, the DNase2TF algorithm is able to correct dinucleotide bias, detecting footprints with accuracy better or comparable to existing approaches (Sung et al., 2014). Furthermore, Sung et al. (2014) were able to predict DNase signatures using solely tetranucleotide frequency information. Although this 4-nucleotide region has the highest information content, Koohy et al. (2013) and Lazarovici et al. (2013) demonstrated information beyond a context longer than four nucleotides. Consequently, using naked (deproteinized) DNA control datasets specific to a protocol and an enzyme as well as high sequencing depth (Hesselberth et al., 2009) are now suggested recommendations for DNase-seq experiments aiming to detect footprints (Meyer and Liu, 2014).

A third approach, an improved version of HINT [HMM-based identification of TF footprints (Gusmao et al., 2014)], named as HINT-BC/HINT-BCN (Bias Correction based on hypersensitivity sites/Bias Correction based on Naked DNase-seq) includes *k*-mer based bias correction in DNase-seq data as in He et al. (2014), leading to substantial changes in the average DNase I cleavage patterns surrounding the TFs. These changes result beneficial to footprinting method accuracy (personal communication with the author).

Contradictorily, a fourth study using DNase-seq has shown that bias correction does not significantly improve the accuracy of TF binding identification (Kähärä and Lähdesmäki, 2015). In addition, this study poses a second counterintuitive idea in the field: accuracy saturates at a modest sequencing depth (30–60 million reads), and only a few TFs present improvement at deeper sequencing.

ATAC-seq Shows Sequence Cleavage Bias

It is unknown if ATAC-seq derived footprints are factor dependent or affected by Tn5 cleavage preferences (Tsompana and Buck, 2014). As expected, bioinformatic analysis of chromosome 22 in the published human datasets for 50,000 cells reveals sequence biases in ATAC-seq experiments (Buenrostro et al., 2013) (**Figure 1**), similar to those found by Koohy et al. (2013) in DNase-seq. As ATAC-seq might replace DNase-seq in the foreseeable future due to its cost and time efficiencies, and because it simultaneously allows the identification of nucleosome positions (Buenrostro et al., 2013), new computational models are necessary to evaluate intrinsic confounding factors in ATAC-seq.

A novel approach, msCentipede (Raj et al., 2014), has extended CENTIPEDE (Pique-Regi et al., 2011) from a mutinomial model to a hierarchical multiscale model. It has been evaluated on "single-hit" UW DNase-seq (Hesselberth et al., 2009) and on paired-end (PE) ATAC-seq data. Surprisingly, the "flexible model" for background DNase I cleavage rate (msCentipede-flexbg) shows very little improvement for a broad range of factors when taking into account naked DNA information from Lazarovici et al. (2013) datasets. This finding clearly contradicts those of He et al. (2014) and Sung et al. (2014). In msCentipede, the footprint signature (or cleavage profile) pattern within a factor-bound motif instance was, therefore, found to be informative when increasing the sensitivity and specificity of the TF binding site prediction. Raj et al. (2014) suggest that this might be explained by the different range of read count data between the matched consensus sequence of the candidate site/motif (10–30 bp) and the data matrix used typically by the software packages (larger sequence window, around 100–150 bp extension at each flank of the motif), which can mask the effects produced by not accounting for sequence biases within the core motif.

Are Current Benchmarks Adequate to Evaluate Bias-Corrected DNase-seq Data?

So far, a footprint of a TF, therefore, might be either detectable (and better detectable when accounting, or not, for influencing factors), or undetectable. In many studies, both problems are

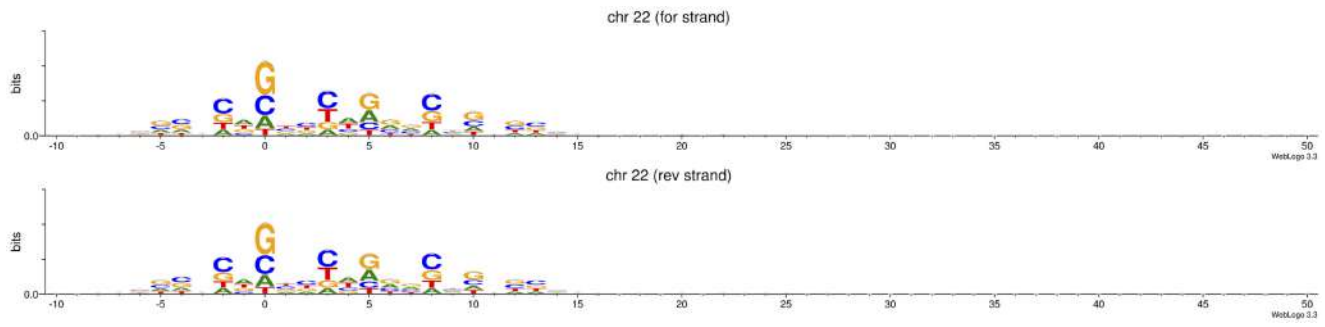


FIGURE 1 | Tn5 transposase shows sequence cleavage bias. Data represented correspond to read-start sites in reads aligned to forward and reverse strands in chromosome 22 in four ATAC-seq replicates (50 k cells per replicate) reported in Buenrostro et al. (2013). Of total, 50 bp PE reads were pre-processed with Trimmomatic v0.32 under default parameters, and then aligned to hg19 using BWA v0.7.4-r385 (Li and Durbin, 2010; Bolger et al., 2014). Sequence logos were generated using WebLogo (Crooks et al., 2004). *Y*-axis: 0.0–0.3 bits.

convoluted and addressed using the same "gold standard" datasets, such as ChIP-seq, which do not have nucleotide-level resolution. Hence, on these methods and gold standards, no reproducible improvements can be seen. This was already noted in Cuellar-Partida et al. (2012), when it was showed that simply scanning for position weight matrices in DNase I hypersensitive sites (DHSs) had the same power as CENTIPEDE. These issues also complicate data integration with TF ChIP-seq, as peaks without a footprint in DNase-seq/ATAC-seq, considered weak/indirect binding or false positives (ChIP artifacts), might instead be explained by a class of TFs with rapid kinetics. And vice versa, DNase I cleavage patterns located within "ChIP-seq unbound" sites – noted previously, e.g., in the MILLIPEDE framework, especially in yeast (Luo and Hartemink, 2013) – could support the hypothesis of footprint shape dominated by DNA sequence specificities.

Future Directions

There is room for improvement in current methodologies by making use of the sequence specificity of each enzyme/assay, including ATAC-seq, but there is no clear consensus in its importance for digital genomic footprinting. This situation is not exclusive for genome-wide chromatin accessibility experiments: modeling the sequence-specific lambda exonuclease bias in ChIP-exo did not significantly increase the identification of TF binding sites (Wang et al., 2014). Similarly, there is no clear consensus if footprint signatures at the core motif, whether they are unique or not for an individual factor, are really important for footprint identification. Establishing better benchmarks to compare performance of the algorithms across different protocols is a fundamental task. These benchmarks could be based on "differential footprints" (sites within DHSs that are bound by a factor in one condition but not the other) as a more appropriate metric to evaluate footprint identification performance instead of using ChIP-seq data (Yardimci et al., 2014). In addition, are DNase-seq software tools equally applicable to ATAC-seq without modification? If enzyme-specific biases are taken into account in a comparable experimental set-up, will DNase-seq and ATAC-seq report the same footprints for an identical sample using same algorithm parameters? This is unlikely, based on a previous comparison between open chromatin DHSs and FAIRE sites, which revealed unique regions produced in each assay (Song et al., 2011). It has been also proposed that performing, and combining, experiments with different nucleases can be an alternative to mitigate biases (He et al., 2014; Mahony and Pugh, 2015).

A greater challenge is dealing with proteins with very short residency time in the DNA as they produce mostly negligible footprints (Rusk, 2014; Sung et al., 2014). Optimizing and implementing new methods is necessary in order to enable biological insights that current methods cannot reveal.

Acknowledgments

Research in the Pedro Madrigal's laboratory is supported by ERC starting grant Relieve-IMDs and core support grant from the Wellcome Trust and MRC to the Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute.

References

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi:10.1038/nmeth.2688

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi:10.1101/gr.849004

Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* 28, 56–62. doi:10.1093/bioinformatics/btr614

Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 30, 3143–3151. doi:10.1093/bioinformatics/btu519

He, H. H., Meyer, C. A., Hu, S. S., Chen, M. W., Zang, C., Liu, Y., et al. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* 11, 73–78. doi:10.1038/nmeth.2762

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

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

Kähärä, J., and Lähdesmäki, H. (2015). BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* 31, 2852–2859. doi:10.1093/bioinformatics/btv294

Koohy, H., Down, T. A., and Hubbard, T. J. (2013). Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS ONE* 8:e69853. doi:10.1371/journal.pone.0069853

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., et al. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6376–6381. doi:10.1073/pnas.1216822110

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698

Luo, K., and Hartemink, A. J. (2013). Using DNase digestion data to accurately identify transcription factor binding sites. *Pac. Symp. Biocomput.* 80–91. doi:10.1142/9789814447973_0009

Madrigal, P., and Krajewski, P. (2012). Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front. Genet.* 3:230. doi:10.3389/fgene.2012.00230

Mahony, S., and Pugh, B. F. (2015). Protein-DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.* 1–15. doi:10.3109/10409238.2015.1051505

McPherson, J. D. (2014). A defining decade in DNA sequencing. *Nat. Methods* 11, 1003–1005. doi:10.1038/nmeth.3106

Meyer, C. A., and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* 15, 709–721. doi:10.1038/nrg3788

Pajoro, A., Madrigal, P., Muino, J. M., Matus, J. T., Jin, J., Mecchia, M. A., et al. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol.* 15, R41. doi:10.1186/gb-2014-15-3-r41

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 21, 447–455. doi:10.1101/gr.112623.110

Raj, A., and McVicker, G. (2014). The genome shows its sensitive side. *Nat. Methods* 11, 39–40. doi:10.1038/nmeth.2770

Raj, A., Shim, H., Gilad, Y., Pritchard, J. K., and Stephens, M. (2014). msCentipede: modeling heterogeneity across genomic sites improves accuracy in the inference of transcription factor binding. *bioRxiv*. doi:10.1101/012013

Rhee, H. S., and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147, 1408–1419. doi:10.1016/j.cell.2011.11.013

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

Rusk, N. (2014). Transcription factors without footprints. *Nat. Methods* 11, 988–989. doi:10.1038/nmeth.3128

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

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21, 1757–1767. doi:10.1101/gr.121541.111

Sung, M. H., Guertin, M. J., Baek, S., and Hager, G. L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* 56, 275–285. doi:10.1016/j.molcel.2014.08.016

Tewari, A. K., Yardimci, G. G., Shibata, Y., Sheffield, N. C., Song, L., Taylor, B. S., et al. (2012). Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biol.* 13, R88. doi:10.1186/gb-2012-13-10-r88

Tsompana, M., and Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 7, 33. doi:10.1186/1756-8935-7-33

Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., et al. (2014). MACE: model based analysis of ChIP-exo. *Nucleic Acids Res.* 42, e156. doi:10.1093/nar/gku846

Yardimci, G. G., Frank, C. L., Crawford, G. E., and Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* 42, 11865–11878. doi:10.1093/nar/gku810

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

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 23 February 2016
doi: 10.3389/fgene.2016.00024

Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells

Valentina Boeva [1, 2, 3, 4, 5, 6, 7, 8]*

[1] Centre de Recherche, Institut Curie, Paris, France, [2] INSERM, U900, Paris, France, [3] Mines ParisTech, Fontainebleau, France, [4] PSL Research University, Paris, France, [5] Department of Development, Reproduction and Cancer, Institut Cochin, Paris, France, [6] INSERM, U1016, Paris, France, [7] Centre National de la Recherche Scientifique UMR 8104, Paris, France, [8] Université Paris Descartes UMR-S1016, Paris, France

Eukaryotic genomes contain a variety of structured patterns: repetitive elements, binding sites of DNA and RNA associated proteins, splice sites, and so on. Often, these structured patterns can be formalized as motifs and described using a proper mathematical model such as position weight matrix and IUPAC consensus. Two key tasks are typically carried out for motifs in the context of the analysis of genomic sequences. These are: identification in a set of DNA regions of over-represented motifs from a particular motif database, and *de novo* discovery of over-represented motifs. Here we describe existing methodology to perform these two tasks for motifs characterizing transcription factor binding. When applied to the output of ChIP-seq and ChIP-exo experiments, or to promoter regions of co-modulated genes, motif analysis techniques allow for the prediction of transcription factor binding events and enable identification of transcriptional regulators and co-regulators. The usefulness of motif analysis is further exemplified in this review by how motif discovery improves peak calling in ChIP-seq and ChIP-exo experiments and, when coupled with information on gene expression, allows insights into physical mechanisms of transcriptional modulation.

Keywords: motif discovery, transcription factors, binding sites, position-specific scoring matrices, regulation of gene transcription, ChIP-seq, binding motif models

OPEN ACCESS

Edited by:
Ekaterina Shelest, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Germany

Reviewed by:
Vladimir A. Kuznetsov, Bioinformatics Institute, Singapore
Jan Grau, Martin Luther University Halle-Wittenberg, Germany

***Correspondence:**
Valentina Boeva
valentina.boeva@inserm.fr

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *30 October 2015*
Accepted: *05 February 2016*
Published: *23 February 2016*

Citation:
Boeva V (2016) Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Front. Genet. 7:24. doi: 10.3389/fgene.2016.00024

INTRODUCTION

A eukaryotic genome contains a variety of structured patterns. A far from exhaustive list of genomic patterns includes (i) tandem repeats and transposable elements, (ii) stretches of GC- or AT-rich sequences (e.g., CpG islands in mammalian genomes), (iii) binding sites of DNA associated proteins (e.g., transcription factor binding sites), (iv) splice sites, and (v) DNA and RNA binding sites of non-coding RNA molecules. Different patterns may overlap each other. Therefore, although this review is focused on motifs for transcription factor binding sites (TFBSs), we provide a short overview of other types of genomic patterns.

Transcription Factor Binding Sites (TFBSs)

Transcription factors (TFs) are proteins with DNA binding activity that are involved in the regulation of transcription. Generally, TFs modulate gene expression by binding to

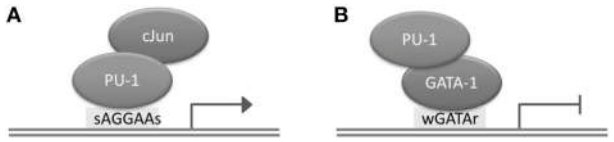


FIGURE 1 | Direct and indirect binding of TF PU.1 to DNA. (A) Direct binding of PU.1 to DNA to the consensus motif sAGGAAs, which may lead to transcriptional activation. **(B)** Indirect binding of PU.1 to DNA, which may lead to transcriptional repression.

gene promoter regions or to distal regions called enhancers. The distance between a TFBS and a transcription start site (TSS) of a gene regulated by the TF can be up to several megabases, and depends on the chromatin structure of the region (Dekker and Heard, 2015). Although TFs possess by definition DNA binding domains, they may occasionally bind DNA indirectly, by interacting with another TF. For instance, PU.1 and GATA-1 (TFs playing a critical role in the differentiation of hematopoietic lineages) interact through the ETS domain of PU.1 and the C-terminal finger region of TF GATA-1; as a result, PU.1 can bind to DNA both directly and indirectly, through the assistance of GATA-1 (**Figure 1**; Burda et al., 2010). A TF has binding preferences to a specific set of DNA sequences referred to as a "binding motif." TFs have different binding affinities for sequences forming their binding motif set. Several mathematical models have been developed to represent a binding motif and take into account its properties. One of the most commonly used models is the positional weight matrix (PWM), also called the position-specific scoring matrix (PSSM), containing the log-odds or log-probability weights for computing the binding affinity score. Construction and use of the PWM model is discussed in detail in the next section. In some cases, the same TF is able to bind quite dissimilar motifs; the motif choice may predefine the action of this TF on gene expression (Guillon et al., 2009).

TFs often interact with each other or compete for DNA binding. Consequently, their binding sites may co-localize or overlap (Wang et al., 2012). Co-localization of TFBSs can be also due to the combined action of a set of TFs: First, TFs capable of binding inactive chromatin bind to DNA and create an open chromatin environment through the recruitment of histone acetyltransferases (pioneer TFs). Then, other TFs (lacking the above capability) become able to bind DNA and activate gene transcription by interacting with the RNA polymerase machinery (Farnham, 2009). Analysis of the distance and orientation preferences between the sites of co-binding TFs helps to predict possible protein-protein interactions, and enables insights into the mechanisms of transcriptional regulation by TFs when coupled with information on gene expression modulation.

Repeats

Repeats constitute a large part of eukaryotic genomes. For instance, more than 45% of the human genome corresponds to repetitive sequences (Derrien et al., 2012). Among them, one distinguishes tandem repeats (DNA is repeated in head-to-tail fashion: microsatellites, minisatellites, and satellite sequences) and interspersed repeats (similar sequences are located throughout the genome). The latter correspond to transposable elements such as SINEs and LINEs, accounting for 12.5 and 20% of the human genome, respectively. Tandem repeats themselves account for 10–15% of the human genome. While short tandem repeats can serve as binding sites for specific transcription factors (TFs; Shi et al., 2000; Guillon et al., 2009), long satellite repeats can play a role in the 3D structure shaping of the genome. For instance, the α-satellite family of repeats (∼171 bp tandem repeats) are bound by the fundamental component of the centromere CENP-C, and are essential for centromere function by ensuring proper chromosome segregation in mitosis and meiosis (Politi et al., 2002). The TandemSWAN software (http://favorov.bioinfolab.net/swan/tool.html) allows the annotation of exact and fuzzy tandem repeats in genomic sequences (Boeva et al., 2006). It is usual to mask such repeats in order to avoid artifact discovery, for example, during analysis of next-generation sequencing data.

AT- or GC- Rich Sequences

AT- or GC- rich sequences are often located in gene promoters and play a role in transcription initiation. Approximately 24% of human genes contain an AT-rich sequence within the core promoter, with 10% containing a canonical TATA-box motif (TATAWAWR, W = A/T, R = A/G; Yang et al., 2007). The TATA-box recruits the TATA binding protein (TBP), which unwinds the DNA; also, due to weaker base-stacking interactions among A and T (than G and C), AT-rich sequences facilitate unwinding. The remaining 76% of human promoters are GC-rich and contain multiple binding sites of the transcriptional activator SP1 (Yang et al., 2007). As much as 56% of human genes, including most of the housekeeping genes, possess CpG islands, i.e., 300–3000 bp GC-rich sequences around gene TSS with a high density of CpG dinucleotides. The high methylation level of CpG sites in CpG islands has been shown to be associated with transcriptional repression. Polycomb group (PcG) repressor proteins recognize CpG islands that are unmethylated and unprotected by TFs (Klose et al., 2013). PcG proteins associate with DNA methyltransferases responsible for methylation of CpG islands (Viré et al., 2006). Also, some components of PcG proteins have histone methyltransferase activity and trimethylate histone H3 on lysine 27, which is a mark of transcriptionally silent chromatin.

Splice Site

During splicing, introns are removed from the pre-messenger RNA transcript and remaining exons are joined together to later form mature messenger RNA. Generally, in eukaryotes, the process of splicing is catalyzed by spliceosomes. These complex molecular machines recognize a donor site (almost invariably GU at the 5′ end of the intron), a branch site (adenine nucleotide followed by a pyrimidine-rich tract near the 3′ end of the intron), and an acceptor site (almost always AG at the 3′ end of the intron) on RNA transcripts. A DNA mutation in a splice site may have a wide range of functional consequences, among them exclusion of an exon from the mature mRNA, or inclusion of an intron or part of one. The latter often results in disruption of the reading frame

FIGURE 2 | Sequence logo of the PWM created by ChIPMunk (Kulakovskiy et al., 2010) using 17,781 binding site regions predicted for PU.1/Spi-1 using ChIP sequencing (ChIP-seq) data (Ridinger-Saison et al., 2012).

or a premature stop codon, and thus gives rise to a defective or truncated protein.

miRNA Binding Sites

While binding of regulatory proteins to promoter and enhancer DNA regions regulates expression of the targeted protein at the transcription level, binding of micro RNA molecules (miRNAs) to the 3′UTR region of a mRNA transcript can regulate the protein amount at the post-transcriptional level. The interaction of an miRNA as part of an active RNA-induced silencing complex (RISC) with a 3′UTR of the targeted mRNA transcript results in either inhibition of translation or increased degradation of this transcript. The miRNA complex recognizes the 6–8 nucleotides at the mRNA 3′UTR, which is complementary to the miRNA "seed" region (Bartel, 2009). In the human genome, there are more than 2000 unique miRNAs. One miRNA can target several genes, and the same 3′UTR can be targeted by multiple miRNAs. Sequence analysis of gene's 3′UTR, coupled with the analysis of evolutionary conservation of the 3′UTR region, allows the prediction of miRNA-target pairs (Yue et al., 2009). Mutations in an miRNA target site may disrupt miRNA repressive regulation, and thus result in protein overexpression (Chin et al., 2008). Alternatively, a mutation in the 3′UTR of a gene can create a new active miRNA binding site, negatively affecting gene expression (Ramsingh et al., 2010).

In this review, we present methods for *in silico* prediction of TFBSs, which can overlap any other type of genomic motif: repeats, CpG islands, splice sites, and so on. Some of the motif analysis methods discussed in this review in Section "*In silico* Detection of TFBSs" can be also applied to other types of motifs than TFBSs. In Section "Applications of Motif Analysis", we also demonstrate how motif discovery can be used to improve peak calling from chromatin immunoprecipitation (ChIP) sequencing data and obtain insights about mechanisms of transcriptional regulation by specific TFs.

IN SILICO DETECTION OF TRANSCRIPTION FACTOR BINDING SITES

We define TF binding motifs as sets of DNA sequences having high affinity for binding TFs. Each occurrence of a sequence from the binding motif in a genomic region is referred to as a motif instance. In the case of direct binding of a TF to DNA, a DNA region surrounding the binding site usually contains one or more instances of the corresponding binding motif.

There are several models for defining binding motifs. These can be used to scan a DNA sequence to predict TFBSs.

Enumeration

All sequences with the potential to be bound by a TF can be enumerated. Information about these sequences can be obtained from SELEX experiments (Oliphant et al., 1989). To allow for discrimination between sequences with strong and weak binding affinities, one can use for example the SELEX affinity score assigned to each particular k-mer.

Consensus

An alternative model for motif description is a consensus motif, constructed using the nomenclature of the International Union of Pure and Applied Chemistry (IUPAC):

A = adenine	C = cytosine
G = guanine	T = thymine
Y = T | C (pyrimidine)	R = G | A (purine)
K = G | T (keto)	M = A | C (amino)
S = G | C (strong bonds)	W = A | T (weak bonds)
B = G | T | C (all but A)	V = G | C | A (all but T)
D = G | A | T (all but C)	H = A | C | T (all but G)
N = A | G | C | T (any)	

For instance, the IUPAC consensus for the binding motif of TF PU.1/Spi-1 can be written RRVRGGAASTS (the corresponding motif logo is depicted in **Figure 2**; Ridinger-Saison et al., 2012). The shortcoming of this way of modeling binding motifs is that many functional binding sequences may not be included in the motif when using a stringent consensus, and indeed, when consensus is poor, the motif can comprise motif instances of very low binding affinity, due to the uncaptured effect of nucleotide combinations on several low-affinity positions.

Position Weight Matrix (PWM)

The PWM is the most frequently used mathematical model for binding motifs (Stormo, 2000). A PWM contains information about the position-dependent frequency or probability of each nucleotide in the motif. This information is usually represented as log-weights $\{w_{\alpha,j}\}$ of probabilities ($w_{\alpha,j} = \log(p_{\alpha,j})$) or, most frequently, odds ratios ($w_{\alpha,j} = log_2(p_{\alpha,j}/b_\alpha)$) for computing a match score. Here $p_{\alpha,j}$ is the probability of nucleotide $\alpha\alpha$ at position j, and b_α the background probability of nucleotide α. Small sample correction is usually included in $p_{\alpha,j}$ to avoid taking the logarithm of zero. A PWM match score for an arbitrary k-mer $A = a_1a_2 \ldots a_k$ is computed as $S_A = \sum_j w_{a_j,\ j}$. Recent "deep learning" techniques (Alipanahi et al., 2015) use PWMs where weights are not required to be probabilities or log-odds ratios.

PWMs can be visualized using sequence logos (Schneider and Stephens, 1990; **Figure 2**). The total height of each bin is the information content in bits of the corresponding position: $H_j = 2 - \sum_\alpha p_\alpha,\ jlog_2(p_{\alpha,j})$. The height of each nucleotide in the logo is proportional to its probability $p_{\alpha,j}$ and, for each

position, the four nucleotides are ordered by $p_{\alpha,\,j}$ with the most likely nucleotides depicted on top of the stack.

PWMs can be experimentally determined from SELEX experiments or computationally discovered from protein binding microarrays (PBMs; Berger and Bulyk, 2009), genomic-context PBM (gcPBM; Gordân et al., 2013), ChIP-seq, and ChIP-exo data.

Using the PWM motif representation, it is possible to distinguish strong binding sites (high PWM score) from weak binding sites (moderate PWM score). It may however, be a problem to discriminate weak binding sites from background (low or negative PWM score). Usually, a cutoff in the PWM score is used to decide whether a given sequence matches the motif. The choice of this cutoff is a complex statistical task that we discuss further here and in Section "Detection of TFBSs with Known PWMs".

A PWM is constructed based on single nucleotide frequencies (four letter alphabet). However, from the methodological point of view, this model can be easily extended to the 16 letter alphabet of consecutive dinucleotides. This model has been used in the *de novo* motif discovery methods Dimont (Grau et al., 2013), diChIPMunk (Kulakovskiy I. et al., 2013), and BEEML-PBM (Zhao and Stormo, 2011; Zhao et al., 2012), the latter being designed to work with PBM data.

Bayesian Networks and Other Supervised Classification Methods

Although PWM is the most widely used mathematical representation of TF specificity, it still has drawbacks. For instance, it assumes the independence of positions within the motif: each position contributes separately to the PWM score, which reflects binding affinity. Modeling position dependencies with Bayesian networks provides an elegant solution to this problem (Barash et al., 2003; Ben-Gal et al., 2005; Grau et al., 2006). However, since there is no easy way to visualize motifs defined as a Bayesian network, this approach is rarely used by the research community.

This class of models was followed by another class of graphical model approaches based on Markov models (Wasson and Hartemink, 2009; Reid et al., 2010; Mathelier and Wasserman, 2013; Eggeling et al., 2014). The approach proposed by Mathelier and Wasserman (2013) has been included in the JASPAR database. Slim probabilistic graphical models, implemented by Keilwagen and Grau (2015), can be used via a Galaxy wrapper (http://galaxy.informatik.uni-halle.de); the authors also provide an intuitive model visualization.

In addition, motifs can be modeled and searched for using k-mer frequencies via support vector machine (SVM) approaches (Holloway et al., 2005; Jiang et al., 2007; Gorkin et al., 2012; Fletez-Brant et al., 2013). This class of approaches can be successfully applied to PBM data (Agius et al., 2010; Mordelet et al., 2013).

One of the important advantages of these graphical model and SVM-based approaches is that they can account for variable spacing between half-sites of two-box TFs (examples of such motifs are shown in **Figure 6A**). The DREAM5 challenge paper provides a comparative study of different methods for modeling transcription factor sequence specificity (Weirauch et al., 2013).

Given a motif described with one of the above-listed models, one can scan a set of genomic sequences or even a whole genome in order to detect possible TF binding sites. This can be achieved by applying efficient algorithms employing deterministic and non-deterministic finite automata accepting motif instances (Navarro and Raffinot, 2002; Antoniou et al., 2006; Boeva et al., 2007; Marschall and Rahmann, 2008; Marschall, 2011; Holub, 2012). The AhoPro (http://favorov.bioinfolab.net/ahokocc/seach_motifs., html Boeva et al., 2007) and PWMTools (http://ccg.vital-it.ch/pwmtools/pwmscan.php, Iseli et al., 2007) websites allow for fast online searches of instances of motifs with several of the models described above, in a set of sequences in FASTA format or in whole genomes. More tools allowing for a fast scan of sequences in FASTA format for motif instances are listed in the next section.

In the following, we choose the PWM model to represent binding motifs. Given that a cutoff is correctly selected, we assume that a TF binds DNA sequences with PWM scores higher than the cutoff. This assumption is a very rough approximation of reality. Using a high cutoff implies rejecting most of the weak binding sites, while using a lower cutoff can result in adding too much noise to predictions and muddle biological conclusions. In practice, the cutoff can be selected in a way to predict one motif instance per 1 or 10 Kb of the genome (Kulakovskiy I. V. et al., 2013). Cutoff choice can be also based on the hypothesis that the corresponding motif is over-represented in a given set of DNA sequences; this cutoff selection strategy is discussed in the next section.

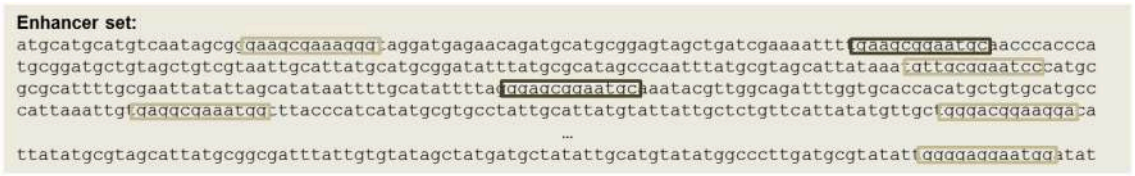
In silico detection of TFBS may be separated into two tasks: detection of binding sites of TFs with known binding motifs (PWMs), and *de novo* motif discovery. Sections "Detection of TFBSs with Known PWMs" and "*De novo* Motif Discovery" focus on these two questions.

Detection of TFBSs with Known PWMs

Detection of TF binding motif instances for known motifs has its application in promoter analysis or the analysis of more distant regulatory regions (enhancers), where the goal is to find TFs possibly regulating corresponding genes. Scanning a set of sequences with PWMs of known motifs can also be used to detect co-factor binding in ChIP-seq-derived binding site regions of a TF of interest. Alternatively, one can use known-motif discovery to assess the effect of SNPs and mutations on TF binding. With the increase in the number of sequenced genomes, the second question has recently gained in importance, and novel tools permitting annotation of variants within TF motif instances have begun to be developed (Boyle et al., 2012; Ward and Kellis, 2016).

There exist several public and commercial databases storing PWMs for known TF binding motifs.

- HOCOMOCO: a comprehensive collection of human TFBS models (Kulakovskiy I. V. et al., 2013)
- JASPAR 2016: an extensively expanded and updated open-access database of TF binding profiles that can capture



Cutoff of PWM score	# sequences in motif	# motif hits in the enhancer set	-log$_{10}$(P-value)
3.0	1020	320	5.439
3.5	880	246	7.012
4.0	690	114	2.721
4.5	530	89	2.920
5.0	256	78	3.753
5.5	216	51	4.890
6.0	198	42	5.422
6.5	120	36	6.186
7.0	80	25	2.939
7.5	44	23	3.371
8.0	16	16	3.923
8.5	2	6	3.652

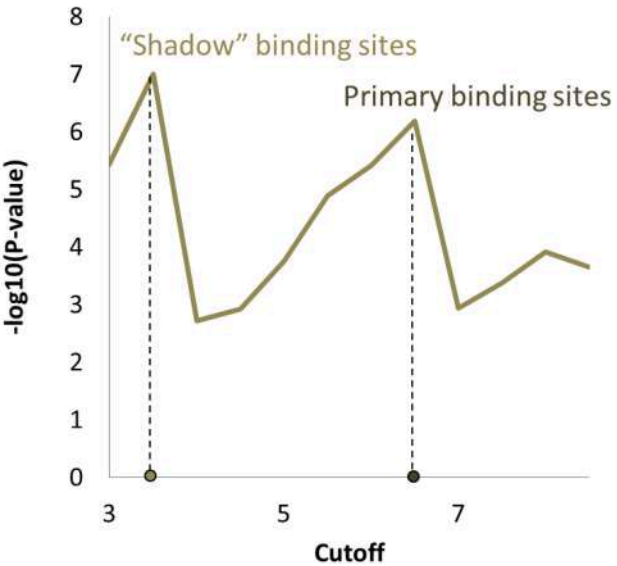


FIGURE 3 | PWM score cutoff selection for a set of enhancer regions. Two local maxima in the *P*-value graph provide two *p*-value cutoffs that correspond to primary binding sites (high cutoff) and "shadow" binding sites (low cutoff). The table shows how many potential k-mer sequences match the PWM with a given cutoff (column 2), the number of motif instances in the set of enhancers (column 3), and the corresponding *p*-value (column 4).

dinucleotide dependencies within TF binding sites (Mathelier et al., 2016)

- SwissRegulon: a database of genome-wide annotations of regulatory sites (Pachkov et al., 2007)
- TRANSFAC®: a commercial database on TFBSs, PWMs, and regulated genes in eukaryotes (Matys et al., 2006)
- footprintDB: a database summarizing motifs from HOCOMOCO, JASPAR, and other databases (Sebastian and Contreras-Moreira, 2014).

True binding sites usually score high with the corresponding PWM, while background sequences have low PWM scores. It is not sufficient to scan a DNA region to get a PWM score at each position. The main difficulty is to correctly set the cutoff on the PWM score to separate true binding sites from background. Evaluation of the statistical significance of motif instances can help solve this issue (Boeva et al., 2007).

When a PWM score cutoff c is given, it is possible to enumerate all possible sequences matching PWM with a score above the cutoff. Let us call this set $M_c = \{A_{s_1}, A_{s_2}, \ldots, A_{s_m}\}_{s_i>c}$, where each sequence A_{s_i} is a k-mer with PWM score $s_i > c$. The higher the cutoff c, the smaller the set of motif sequences M_c. Given a set of regulatory regions (enhancers or promoters) R, we can define the number $N_{R,c}$ showing how many A_{s_i} from M_c occurred in R. With a higher cutoff, fewer motif instances will be detected; corresponding binding sites are likely to have strong binding affinity. With a lower cutoff, more motif instances are detected; these may correspond to both strong and weak binding sites.

In regulatory regions, binding sites often tend to occur in clusters, and binding motifs are over-represented in the set R of regulatory sequences targeted by the transcription factor. This is not the case for random sequences. The procedure developed in Boeva et al. (2007) to specify the cutoff on the PWM score for a set R is based on this assumption.

The significance of motif instance over-representation can be measured through the p-value, i.e., the probability to observe at least the same number $N_{R,c}$ of motif instances with cutoff c in a random sequence with total length equal to the total length of sequences in R (**Figure 3**). Setting different cutoffs c, one gets different numbers of motif instances $N_{R,c}$ in R and different p-values, $P(M_c, N_{R,c})$. The minimum of $P(M_c, N_{R,c})$ over c provides a cutoff corresponding to the most significant motif over-representation in R. This approach can be equally applied to several PWM corresponding to several TF binding motifs (**Figure 4**).

The exact p-value calculation for multiple motifs with overlapping (and self-overlapping) motifs is a difficult computational task. The compound Poisson distribution formula for the p-value generally provides a good approximation, but not in the case of several highly-overlapping motifs. An exact algorithm for p-value calculation for the general case of heterotypic clusters of motifs may be based on the

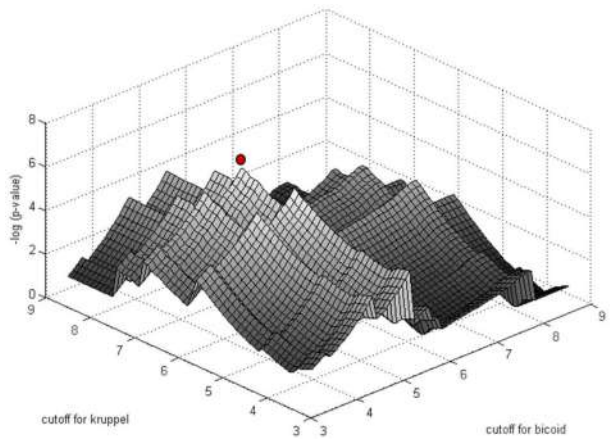


FIGURE 4 | Simultaneous PWM score cutoff selection for PWMs of two *D. melanogaster* TFs: Bicoid and Krüppel. The graph shows the distribution of $\log_{10}$(*p*-value) as a function of the cutoff for the two PWMs for the enhancer of the gene even-skipped stripe 2 (*eve2*). The red point corresponds to the most significant combination of PWM and cutoffs (from Boeva et al., 2007).

Aho-Corasick automaton, and employ a prefix tree together with a transition function (Boeva et al., 2007; Marschall and Rahmann, 2008).

The approach for automatic cutoff choice for a set of PWMs was applied to the identification of binding sites of cooperatively and anti-cooperatively functioning regulatory proteins in *D. melanogaster* (Boeva et al., 2007). By employing this method, we discovered the phenomenon of "shadow" TFBS in enhancers of the *D. melanogaster* genome. Shadow binding sites are low affinity binding sites that alone are not capable of retaining the TF long enough to ensure activation/repression, but instead are used to maintain a high concentration of TF in the vicinity of the primary binding sites. This phenomenon has been recently confirmed by other studies (Kozlov et al., 2015).

We should mention that the choice of the background model is quite important in the calculation of probabilities of motif occurrences. A Markov chain employed as a background model allows us to capture dependencies between nucleotides. This can take into account low or high frequencies of CpG nucleotides in the set of enhancer or promoter sequences.

An automatic scan of a set of DNA sequences using motifs from the databases listed above, with tool-specific cutoffs, is available through the following websites and programs:

- AME or FIMO of the MEME suite (McLeay and Bailey, 2010) http://meme-suite.org/
- SeqPos of Galaxy Cistrome (Liu et al., 2011) http://cistrome.org/ap/
- PWMScan of PWMTools (Iseli et al., 2007) http://ccg.vital-it.ch/pwmtools/pwmscan.php
- oPOSSUM-3 (Kwon et al., 2012) http://opossum.cisreg.ca/oPOSSUM3/
- HOMER (Heinz et al., 2010) http://homer.salk.edu/homer/

De novo Motif Discovery

When the PWM of a TF of interest is not known, it can be obtained using *de novo* motif discovery from a set of DNA sequences containing binding sites of this TF. The technique consists of defining the most over-represented motif in a given set of DNA sequences. The set of DNA sequences containing TFBSs of a particular protein can be obtained with SELEX, PBM or ChIP-x experiments (i.e., ChIP-seq, ChIP-exo, ORGANIC, ChIP-on-chip). ChIP-Seq (Johnson et al., 2007), ChIP-exo (Rhee and Pugh, 2011), and ORGANIC (Kasinathan et al., 2014) consist of immunoprecipitation of DNA–protein complexes and sequencing of short ends of the immunoprecipitated DNA. These techniques provide enhanced resolution of binding regions compared to ChIP-on-chip, which is based on microarrays, and have almost replaced the latter. The ChIP-exo technique provides an even better resolution of binding sites than ChIP-seq, at the expense of a more elaborate library preparation protocol, including an exonuclease step. In this section, we focus on *de novo* motif discovery in ChIP-seq datasets.

ChIP-seq yields a set of genomic regions (also called peaks) that are thought to contain TFBSs. The output of a ChIP-seq experiment can include tens of thousands of peaks, some longer than 1000 bp. Each peak position has a weight reflecting how often a given DNA fragment was cross-linked with the protein of interest during the ChIP stage (coverage profiles).

There exist a large number of methods for the *de novo* detection of over-represented motifs. The classical tool, MEME (Bailey et al., 2009), was developed for motif discovery in a small number of short DNA sequences, and scales poorly to large ChIP-seq datasets. Subsequently, several methods were newly created to analyze large sets of sequences resulting from ChIP-seq experiments: HMS (Hu et al., 2010), cERMIT (Georgiev et al., 2010), ChIPMunk (Kulakovskiy et al., 2010), diChIPMunk (Kulakovskiy I. et al., 2013), MEME-ChIP (Machanick and Bailey, 2011), POSMO (Ma et al., 2012), XXmotif (Hartmann et al., 2013), FMotif (Jia et al., 2014), Dimont (Grau et al., 2013), RSAT (Medina-Rivera et al., 2015), and DeepBind (Alipanahi et al., 2015). The latter method uses increasingly popular "deep learning" techniques; however, it has only been tested on sets of rather short input sequences (up to 101 bp).

There is a tradeoff between the user-friendliness of these tools, speed, and accuracy of predictions. For instance, the use of dinucleotide frequencies and application of read coverage profiles (.wig files) as priors for motif locations, improves the quality of resulting motifs. Both options are supported by diChIPMunk (Kulakovskiy I. et al., 2013). Dimont (Grau et al., 2013) can also use dinucleotide sequences for PWM construction and take into account peak height information, i.e., number of reads supporting each putative binding region. However, the user may find it encumbering extracting coverage information from the ChIP-seq data. Also, dinucleotide PWMs can come across as illegible in biological publications. It appears that intuitive and fast online methods based on classical PWMs are generally in higher demand by biologists than more sophisticated methods. Indeed, speed is one of the key issues in this type of analysis. In this context, k-mer enumeration methods like POSMO (Ma et al., 2012), cERMIT (Georgiev et al., 2010), and RSAT-peak-motifs

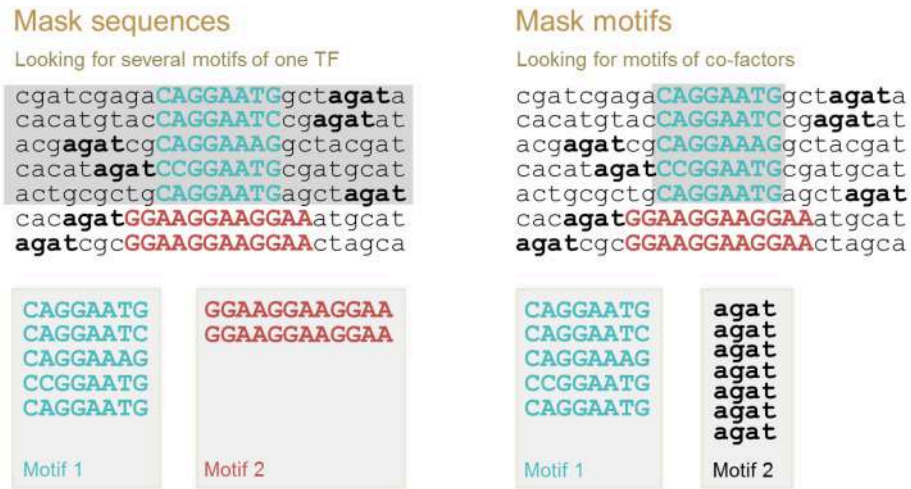


FIGURE 5 | Two modes of multiple motif detection: "Mask sequences" mode to discover binding motifs of the same TF, and "Mask motifs" mode to discover binding motifs of co-factors. After the first motif is identified, either all sequences containing this motif instance are removed from further analysis (sequences in gray, "Mask sequences" mode), or motif instances are masked (motif instances in gray, "Mask motifs" mode). The second motif is defined as the motif with the highest KDIC in the remaining nucleotide sequences.

(Medina-Rivera et al., 2015) show very competitive runtimes on large ChIP-seq datasets. However, probabilistic approaches (e.g., ChIPMunk, Dimont) may provide higher accuracy results (Grau et al., 2013). Overall, according to comparative studies, POSMO, Dimont, and ChIPMunk seem to be the most suitable methods for motif discovery among currently available ones (Ma et al., 2012; Grau et al., 2013). However, a more detailed study including more recent methods is required. More information about recently published methods is available in several reviews (Tran and Huang, 2014; Lihu and Holban, 2015). Most of the above-cited methods allow detection of *several* over-represented motifs. Below, we illustrate *de novo* multiple motif discovery with the ChIPMunk tool.

Multiple motif discovery allows us to identify (i) all possible binding motifs for the same TF and (ii) co-factor binding motifs. For these two cases, different motif discovery procedures should be applied. These two procedures are implemented in ChIPMunk as "Mask sequences" and "Mask motifs" modes. The first motif identified is always the motif with the highest Kullback discrete information content (KDIC). Then, the second motif is identified as the motif with the highest KDIC either in the sequences that do not contain the first motif ("Mask sequences" mode), or in the total set of sequences where the instances of the first motif have been masked ("Mask motifs" mode; **Figure 5**).

The underlying assumption when using the "Mask sequences" mode is that the same TF can, in some cases, bind to significantly different binding motifs; but almost every binding site region should contain at least one motif instance (Wang et al., 2012). We should mention that frequently a TF has only one binding motif; the higher the PWM score of the corresponding motif, the stronger the binding affinity (Kulakovskiy et al., 2010; Kulakovskiy I. V. et al., 2013). In this case, the "Mask sequences" mode is likely to output only one motif. This motif will be present in almost all sequences from the set. The situation where the same TF has different binding motifs, occur less frequently (Badis et al., 2009). For instance, this is the case for TFs EWS-FLI1 (Guillon et al., 2009) and NRSF (Johnson et al., 2007; **Figure 6**). Also, some proteins, such as PU.1, can bind to DNA both directly and indirectly (**Figure 1**). In these cases, the "Mask sequences" mode will provide, as a result, several motifs. This will be the motifs for the direct and indirect binding (e.g., motifs for PU.1 and GATA1 for the situation illustrated in **Figure 2**).

The underlying assumption for the use of the "Mask motifs" mode is that co-factors of the main TF bind close to the main TF in regions detected with chromatin immunoprecipitation using an antibody specific to the main TF of interest (**Figure 5**, right panel). Thus, binding motifs of co-factors can be detected as over-represented motifs after the motif instances of the main TF have been masked.

When a binding motif is identified *de novo*, it is possible to compare its PWM or IUPAC consensus with the known motif PWMs stored in the TF motif databases via:

- JASPAR (Mathelier et al., 2016)—http://jaspar.genereg.net/,
- Motif Comparison Tool of the MEME Suite (Gupta et al., 2007)—http://meme-suite.org/tools/tomtom
- MACRO-APE (Vorontsov et al., 2013)—http://autosome.ru/macroape/
- STAMP (Mahony and Benos, 2007)—http://www.benoslab.pitt.edu/stamp.

In this section, we have focused on the prediction of TFSB sites in a set of rather **short** regulatory regions provided by the user (regulatory regions obtained from ChIP-seq experiments). However, in some situations, one may be interested in analyzing much larger genomic regions (up to the whole genomes). In

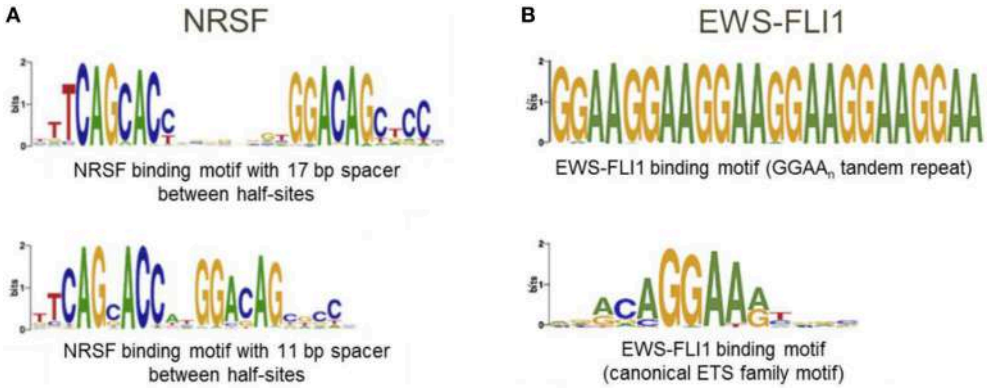


FIGURE 6 | A transcription factor can have several binding motifs. (A) Logos for binding motifs of TF NRSF with 11 and 17 bp spacer between half-sites (Johnson et al., 2007); **(B)** Logos for binding motifs of chimeric TF EWS-FLI1 (Guillon et al., 2009; Boeva et al., 2010).

this case, one can narrow down the space of possible TFBS positions by considering known open chromatin regions in a given cell type, histone marks, and by using conservation profiles between species (Zhong et al., 2013). For instance, using a PWM-based score for the promoter, together with a profile of a single histone modification (H3K4me3), can produce highly accurate predictions of TF-promoter binding (McLeay et al., 2011).

APPLICATIONS OF MOTIF ANALYSIS

Motif discovery finds its applications in the analysis of promoters of co-expressed or co-regulated genes and in the analysis of regulatory regions frequently extracted from ChIP-x experiments. In this section, we explain a frequently applied procedure for promoter analysis. Then, we provide two examples on how motif analysis can be used in the exploration of ChIP-x data. We show how motif information can be applied to get a more accurate set of TFBSs from a ChIP-x experiment, and demonstrate how motif analysis can lead to insights into mechanisms of transcriptional regulation when it is integrated with information about changes in gene expression in a TF inhibition experiment.

Promoter Analysis: Looking for Over-Represented TF Motifs

Discovery of over-represented motifs in a set of genomic regions is often used to determine TFs likely to regulate genes co-modulated following some system perturbation, e.g., knockout or knockdown of a protein or cell differentiation. This type of study is called promoter analysis; it is based on the assumption that several promoters from the gene list are regulated by the same TF via binding of this TF to the promoter area of the corresponding genes. Thus, the goal of promoter analysis is to detect known (or less frequently *de novo*) motifs for which the number of motif instances is significantly higher in the set tested compared to background. As background, one should preferably use a set of promoters of non-modulated genes. Alternatively, one can define a set of random genomic regions or simply specify a background model (e.g., a Markov model of order 1 taking into account dinucleotide frequencies in promoters). Most of the methods apply the zero-or-one occurrences per sequence (ZOOPS) model (Bailey and Elkan, 1995), which enables detection of the strongest motif in a set of sequences; under this model, the strongest motif does not necessarily have instances in every input sequence. The presence of clusters of the same motif in one sequence is not taken into account by this model. The ZOOPS model is also applied by motif discovery tools designed to analyze ChIP-seq data (described above).

There are several major caveats to this approach. First, not every motif incidence corresponds to a true binding event. Thus, the definition of promoter length affects the results of the analysis. Larger promoter regions are likely to include a certain number of false predictions of binding sites, and at the same time are likely to capture more true binding sites. The use of large regions upstream of TSS in promoter analysis is especially unjustified when looking for short or highly degenerate motifs. The second caveat is that genes can be regulated by TF binding to distant regulatory elements: enhancers. These are often tissue specific, and thus not generally included in the set of sequences in which we look for motifs. The third caveat is the selection of the cutoff on the motif strength. Some methods allow the choice of the best cutoff as that providing the lowest p-value, while other methods use predefined cutoffs (Marstrand et al., 2008). Fourth, co-factors may be required for TF binding. In this case, one should probably search for combinations of motifs within a certain distance of one another.

Several tools have been developed specifically for promoter analysis. Some tools require gene lists while others expect sequences in FASTA format as input. The latter methods can be also applied to enhancer regions.

- Web-based promoter analysis tools:
  - Amadeus (Linhart et al., 2008) http://acgt.cs.tau.ac.il/amadeus/—requires program download; can search for pairs of co-occurring motifs; accepts gene lists as input

- i-cisTarget (Herrmann et al., 2012; Imrichová et al., 2015) https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/—accepts.BED files or gene names; when gene names are provides, motif search is performed in 20 Kb window around gene TSSs overlapping with predefined candidate regularity regions
- Pscan (Zambelli et al., 2009) http://www.beaconlab.it/pscan—requires a gene list and provides a choice of 5 lengths for promoter intervals
- OTFBS (Zheng et al., 2003) http://genome.ucsf.edu/∼jiashun/OTFBS/—online version accepts no more than 200 sequences in FASTA format
- Asap (Marstrand et al., 2008) http://servers.binf.ku.dk/asap/—accepts sequences in FASTA format; PWM threshold should be selected by the user
- oPOSSUM-3 (Kwon et al., 2012) http://opossum.cisreg.ca/oPOSSUM3/—accepts both sequences in FASTA format and gene lists
- Match and P-Match (Chekmenev et al., 2005) http://www.gene-regulation.com/pub/programs.html—TRANSFAC® motif scanning algorithms
- SiTaR (Fazius et al., 2011) https://sbi.hki-jena.de/ sitar/—needs a motif in enumeration format

- Offline promoter analysis tools:
  - HOMER (Heinz et al., 2010)—command line tool to search for *de novo* motifs and compare them to known PWMs
  - Clover (Frith et al., 2004).

The motifs in the output are sorted according to the method-specific *p*-values and enrichment scores. These *p*-values may be calculated through binomial or hyper-geometric statistical tests (Frith et al., 2004; Marstrand et al., 2008; Heinz et al., 2010; Kwon et al., 2012), ranking-and-recovery analysis of predefined tracks (Imrichová et al., 2015), or using the Z-transform of scores (Linhart et al., 2008; Zambelli et al., 2009). Correction for multiple tests is optionally performed by some methods (Marstrand et al., 2008).

As mentioned earlier, complementary information about sequence conservation, regions of open chromatin, and presence of specific histone marks, helps to increase TFBS prediction accuracy (Cuellar-Partida et al., 2012; Grant et al., 2015; Imrichová et al., 2015).

Promoter analysis usually predicts binding sites independently for several TFs. However, some recent approaches propose a different strategy, where the goal is to detect combinations of binding sites of several TFs forming cis-regulatory modules (CRMs). These approaches can be based on both *de novo* discovery of motifs, or using available motifs from databases. They can be applied to a set of promoter sequences, but also on predefined sets of enhancers, which can be obtained, for example, using profiles of histone marks. Some methods such as Allegro (Halperin et al., 2009) can take into account a range of changes in gene expression to better predict CRMs.

- Online tools:
  - MatrixCatch (Deyneko et al., 2013) http://www.gene-regulation.com/cgi-bin/mcatch/MatrixCatch.pl—works with TFBS PWMs from the TRANSFAC® database; accepts a set of sequences in FASTA format
  - ModuleMiner (Loo et al., 2008) http://tomcatbackup.esat.kuleuven.be/moduleminer/—accepts Ensembl gene IDs to look for conserved CRMs upstream gene TSSs;
  - PC-TraFF (Meckbach et al., 2015) http://pctraff.bioinf.med.uni-goettingen.de/—uses TRANSFAC® PMWs on gene IDs or sequences in FASTA format
  - DistanceScan (Shelest et al., 2010) https://www.omnifung.hki-jena.de/Rpad/Distance_Scan/index.htm—requires an output from FIMO or Match
  - oPOSSUM-3 (Kwon et al., 2012) http://opossum.cisreg.ca/oPOSSUM3/—requires the name of the anchoring TF
  - MCAST (Grant et al., 2015) http://meme-suite.org/tools/mcast—a tool from the extensive MEME suite; searches for clusters of provided motifs in sequences in FASTA format
  - Cluster-Buster (Frith et al., 2003) http://zlab.bu.edu/cluster-buster/—searches for motif clusters; accepts PMWs in JASPAR or TRANSFAC® formats

- Offline tools:
  - ModuleDigger, CPModule, CORECLUST: stand-alone programs that require a set of known PWMs as input (Sun et al., 2009, 2012; Nikulova et al., 2012).

Validation of TFBSs can be carried out using a combination of chromatin immunoprecipitation with an antibody specific to the TF of interest, and real time PCR with primers specific to the predicted target region.

There are numerous illustrations of application of promoter analysis. For instance, analysis of promoters of protein coding genes and those of long non-coding RNA have shown that these two classes of genes tend to have different transcriptional regulators: motifs for 140 TFs were found to be over-represented in lncRNA gene promoters; this list of TFs includes nuclear hormone receptors and FOX family proteins (Alam et al., 2014). Dopamine-responsive genes have been shown to be regulated by the CREB protein (Frith et al., 2004). Analysis of melanocyte enhancers has predicted binding of key melanocyte TFs, including SOX10 and MITF (Gorkin et al., 2012). Motifs of 6 TFs (Hb, Foxa1, Cf2-ii, Lhx3, Mef2a, and slp1) have been found to be associated with insect bidirectional promoters (Behura and Severson, 2015). Similar analyses in the human genome have revealed 7 TFs (GABPA, MYC, E2F1, E2F4, NRF-1, CCAAT, and YY1) associated with promoter bidirectionality (Lin et al., 2007). Using promoter analysis, several ETS-domain TFs (GABPA, ELK1, and ELK4) have been discovered as likely regulators of breast cancer relevant sense-antisense gene pairs (Grinchuk et al., 2015).

The Use of Motif Information Improves the Accuracy of Binding Site Detection in ChIP-seq and ChIP-exo Data

ChIP-seq and ChIP-exo (ChIP-x) experiments have been widely used to define genomic positions of TF binding and discover TF binding motifs. The usual way to process ChIP-x data is to define TF binding regions first, then perform motif discovery to

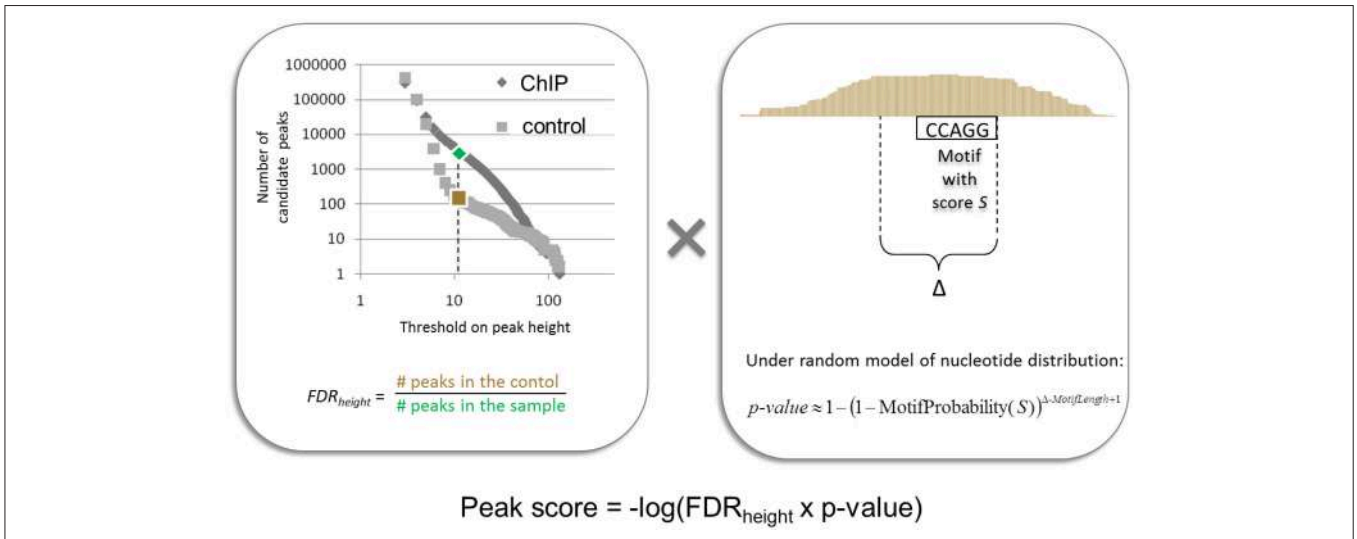


FIGURE 7 | Illustration of the procedure for peak score calculation used by the MICSA algorithm.

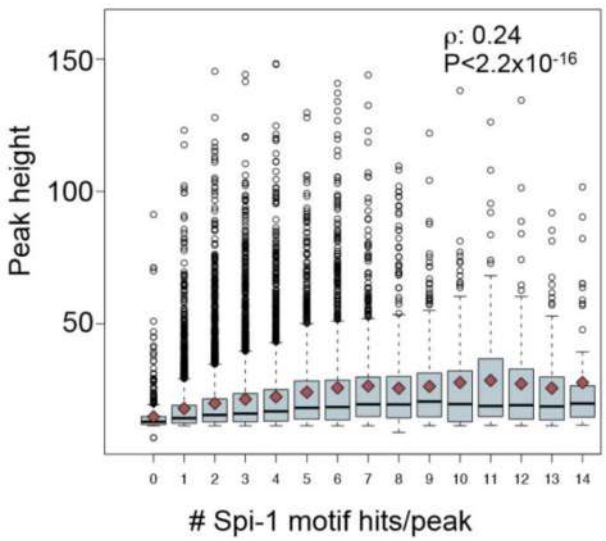


FIGURE 8 | The number of Spi-1/PU.1 motif instances correlates with the Spi-1/PU.1-binding intensity measured by the peak height. The boxplot represents the distribution of the peak heights (y-axis) for each number of Spi-1/PU.1 motif instances/peak (x-axis). The dark red squares indicate the mean values, and the black line within each box indicates the median. The Spearman coefficient correlation (ρ) and the *p*-value of correlation test are reported (from Ridinger-Saison et al., 2012).

construct PWMs of TF binding motifs. In this section, we show that simultaneous instead of successive analysis of ChIP-x signal and motif instances improves the accuracy of TFBS prediction (Boeva et al., 2010; Guo et al., 2012; Starick et al., 2015). Below, we briefly describe the main elements of ChIP-x data analysis.

In the first step of ChIP-x data analysis, by extending each read to the length of the initial immunoprecipitated DNA fragment, it is possible to identify areas of fragment overlap and locate candidate regions of TF-DNA binding. These regions with high fragment density are called candidate peaks (Fejes et al., 2008). Not every peak contains a true binding site. Low peaks (with moderate read density) can appear by chance. Thus, to characterize the read enrichment and discriminate true binding from background noise, a statistical model needs to be applied. There are more than 20 different tools that perform this task for ChIP-x TF data (Wilbanks and Facciotti, 2010; Kim et al., 2011). The background model may be based on the uniform distribution of sequenced reads along the genome. Under such a background model, a Poisson test can be applied to evaluate the significance of read over-representation in a given region (Zhang et al., 2008). Often, in the ChIP-seq protocol, a negative control experiment is performed to assess the distribution of sequenced reads in the background. Recent studies have shown that an appropriate control data set is critical for analysis of any ChIP-seq experiment, because of biases in DNA breakage during sonication (Landt et al., 2012). The ChIP-exo datasets are usually generated with negative controls.

In (Boeva et al., 2010), we presented a peak and motif calling algorithm, MICSA, based on the idea that functional binding sites of TFs should contain a consensus motif (or a set of consensus motifs). The MISCA workflow consists of four phases: (i) identification of all candidate peaks using read extension, (ii) identification of binding motif PWMs from a subset of peaks, (iii) detection of motif instances in all candidate peaks, and (iv) optimization of the peak calling output by calculating statistics taking into account information about both motif instance and depth of coverage. Importantly, MICSA identifies *several* binding motifs. The statistics calculated by MICSA allow us to retain strong binding sites (i.e., regions with high numbers of overlapping fragments) as well as weak binding sites with strong motif instances in the peak center (**Figure 7**). Weak binding sites without strong motif instances are removed from the final dataset. When applied to a ChIP-seq dataset for oncogenic TF EWS-FLI1, MICSA identified two consensus motifs (**Figure 6B**): a $(GGAA)_{\geq 6}$ microsatellite,

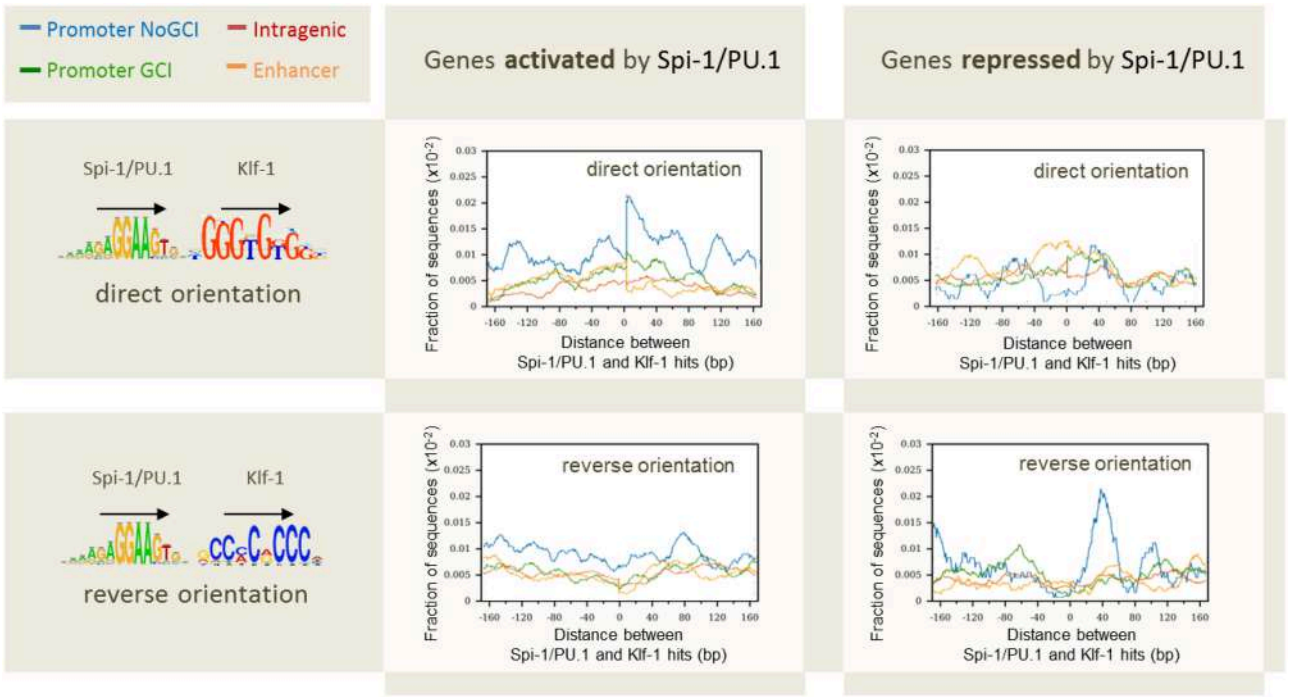


FIGURE 9 | Distribution of the distances between pairs of Spi-1/PU.1 and Klf-1 motif instances in direct or reverse orientation for genes activated and repressed by Spi-1/PU.1. The x-axis shows the length of the spacers separating pairs of Spi-1 and Klf-1 motif instances. The y-axis shows the fraction of sequences with at least one pair of motif instances separated by the selected spacer. Promoters with CpG island (GCI): green; promoter devoted to CpG island (NoGCI): blue; enhancer regions: orange; intragenic regions: red. Data from Ridinger-Saison et al. (2012).

and a motif corresponding to the consensus RCAGGAARY, further referred to as the ETS motif. Surprisingly, the ETS motif did not coincide with the FLI1 binding motif (CCGGAARY), although EWS-FLI1 and FLI1 make up the same DNA-binding domain. Further analysis revealed the tendency of sites bearing GGAA-microsatellites to activate the expression of neighboring genes (sites found from 150-kb upstream to 50-kb downstream of gene TSSs), while sites with the ETS motif do not seem to have a definite activator function. In fact, ETS-sites negatively affected gene expression when located in the 50-kb region downstream of the TSSs. When ETS sites were located further away from gene TSSs (within 1 Mb upstream or downstream), both activator and inhibitory action of EWS-FLI1 was observed. More recent research from (Riggi et al., 2014) has shown that EWS-FLI1 creates *de novo* enhancers when it binds to GGAA-microsatellites, and may disrupt existing regulatory elements of ETS family TFs when it binds to single ETS-sites.

The idea of simultaneous analysis of the ChIP-x read density signal and motif instances has been further developed by Guo et al. (2012). Their GEM algorithm consists of five main steps: (i) detect candidate binding regions, (ii) discover and cluster sets of enriched k-mers, (iii) generate a positional prior for peak calling using k-mer classes, (iv) predict binding sites with a k-mer-based positional prior, and (v) re-discover enriched k-mer clusters in peaks from (iv). On the one hand, by considering motif information, the GEM method gives a better spatial resolution of binding sites than other peak calling methods, also enabling it to resolve closely-spaced binding events. On the other hand, on 214 ENCODE ChIP-Seq experiments for 63 TFs, binding motifs discovered by GEM were overall closer to the expected ones compared to motifs discovered by other methods. In fact, in 15 cases out of 215, GEM outperformed both MEME and ChIPmunk. Using the output of GEM on ENCODE ChIP-seq data in five different cell lines, Guo et al. (2012) studied pairwise binding relationships between different TFs. As a result, 390 pairs of TFs were shown to have significant binding distance constraints within a 100 bp distance, including known interaction pairs MYC-MAX, FOS-JUN, and CTCF-YY1.

The concept of combining ChIP-exo read density with motif information has been employed in the ExoProfiler computational pipeline (Starick et al., 2015). ExoProfiler searches for both *de novo* motifs and known motifs from the JASPAR database. It then extracts regions in ChIP-seq peaks centered on motifs, and analyzes strand specific read density. By applying ExoProfiler to glucocorticoid receptor (GR) ChIP-exo data, Starick et al. (2015) discovered indirect binding of GR to DNA via cofactors (FOX proteins) and discovered a novel GR binding sequence ("combi motif"), at which a GR forms a heterodimer with other TFs (ETS or TEAD families) to activate transcription.

Getting Insights into Physical Mechanisms of Transcriptional Modulation: Co-Directional Clustered Binding of the Oncogenic TF Spi-1/PU.1 Modulates Gene Expressionin Erythroleukemia

Spi-1/PU.1 belongs to the same ETS TF family as FLI1 (the DNA-binding partner of EWS in the gene fusion causing Ewing sarcoma). Spi-1/PU.1 expression beyond physiological expression levels promotes oncogenesis in erythroid cells (Rimmelé et al., 2010). Here, we refer to our study of Spi-1/PU.1 ChIP-seq data, where motif analysis allowed us to get insights into mechanisms of how Spi-1/PU.1 physically modulates the expression of its target genes (Ridinger-Saison et al., 2012).

Analysis of the Spi-1/PU.1 ChIP-seq dataset resulted in a total of 17,781 binding site regions, which were assigned to genes using the Nebula peak-to-gene annotation module (Boeva et al., 2012). Of the 21 Spi-1/PU.1 binding sites tested, 20 were validated using real time PCR. As we detected instances of the binding motif in 88% of the Spi-1/PU.1-bound regions, we concluded that in erythroleukemia, Spi-1/PU.1 binds to DNA directly.

Interestingly, bound to a gene or even to a gene promoter, Spi-1/PU.1 rarely causes transcriptional modulation. Half of all mouse genes contained Spi-1/PU.1 binding sites, i.e., within a −30 kb region upstream of the TSS to +5 kb downstream of the transcription end, but only 8.1% (854 out of 10,560) of the Spi-1/PU.1-occupied genes were transcriptionally modulated. Therefore, we decided to study what additional factors influenced the gene modulation activity of Spi-1/PU.1.

The first factor that correlated to the modulation status of genes was the distance between gene TSS and Spi-1/PU.1 binding sites: 60% of Spi-1/PU.1-activated genes contained Spi-1/PU.1 peaks in 5 kb area around TSSs, though only 40 and 22% of repressed and non-modulated genes, respectively, had peaks within this distance around TSSs. A second factor was the binding affinity, indicated by the peak height: peaks in the promoters of activated genes were significantly higher than in the promoters of repressed and non-modulated genes (p-value $< 10^{-5}$). The binding affinity/peak height correlated with the number of motif instances per peak (**Figure 8**). In agreement with this observation, the number of Spi-1/PU.1 motif instances in Spi-1/PU.1 ChIP-seq peaks in promoters of activated genes was significantly higher than in promoters of repressed or non-modulated genes (p-values $< 10^{-6}$). The third factor was the presence of a CpG island. Our analysis also indicated that Spi-1/PU.1 binding is favored at CG-rich sequences, but the absence of CpG islands increases the potential of Spi-1/PU.1 to activate gene expression. A fourth factor was the orientation of motif instances within a regulatory region. In cases when Spi-1/PU.1 induces gene modulation (activation or repression), Spi-1/PU.1 motif instances form co-oriented clusters (head-to-tail orientation). We observed these clusters of co-oriented motifs both in promoters of up-regulated genes, and enhancers of down-regulated genes. The fifth factor was the distance and orientation of Spi-1/PU.1 binding motifs, and motifs of other TFs. To get this information, we scanned ChIP-seq peak sequences with PWMs of known TFs using PATSER (Hertz and Stormo, 1999; Transfac and Jaspar motifs libraries). The most striking pattern was observed for pairs of Spi-1/PU.1 and KLF family motifs (**Figure 9**). For instance, in promoters of Spi-1/PU.1-up-regulated genes, we observed an enrichment of Spi-1/PU.1-KLF pairs where the direct KLF motif immediately follows the direct Spi-1/PU.1 motif. The patterns observed suggest cooperative interactions between Spi-1/PU.1 and KLF family TFs. The functional significance of these observations needs to be validated by biological experiments.

CONCLUSION

Sequence analysis methods are extremely useful for decrypting the complex structure of patterns and motifs present in eukaryotic genomes. In particular, motif discovery methods applied to promoter/enhancer or ChIP-seq peak sequences enable detection of TFBSs in genomic DNA. In this review, we have presented *de novo* motif discovery techniques, and methods to find over-represented binding motifs of TFs with known motifs (PWMs). We have demonstrated that the application of these techniques improves accuracy of peak calling during ChIP-seq data analysis, and may provide novel biological insights into mechanisms of transcriptional regulation when sequence analysis is coupled with the analysis of gene expression changes. We expect that with time, motif discovery methods will become even more user-friendly, and will allow rapid processing of large datasets, while TRANSFAC®, JASPAR, and other databases will include an increasing number of TF motifs extracted from ChIP-seq experiments.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

This work has been supported by The INSERM Atip-Avenir Program and The ARC Foundation.

REFERENCES

Agius, P., Arvey, A., Chang, W., Noble, W. S., and Leslie, C. (2010). High resolution models of transcription factor-DNA affinities improve *in vitro* and *in vivo* binding predictions. *PLoS Comput. Biol.* 6:e1000916. doi: 10.1371/journal.pcbi.1000916

Alam, T., Medvedeva, Y. A., Jia, H., Brown, J. B., Lipovich, L., and Bajic, V. B. (2014). Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS ONE* 9:e109443. doi: 10.1371/journal.pone.0109443

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Antoniou, P., Holub, J., Iliopoulos, C. S., Melichar, B., and Peterlongo, P. (2006). "Finding common motifs with gaps using finite automata," in *Proceedings of the*

11th International Conference on Implementation and Application of Automata CIAA'06 (Heidelberg: Springer-Verlag), 69–77.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723. doi: 10.1126/science.1162327

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bailey, T. L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, 21–29.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). "Modeling dependencies in protein-DNA binding sites," in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology RECOMB'03* (New York, NY: ACM), 28–37.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Behura, S. K., and Severson, D. W. (2015). Bidirectional promoters of insects: genome-wide comparison, evolutionary implication and influence on gene expression. *J. Mol. Biol.* 427, 521–536. doi: 10.1016/j.jmb.2014.11.008

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

Berger, M. F., and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* 4, 393–411. doi: 10.1038/nprot.2008.195

Boeva, V., Clément, J., Régnier, M., Roytberg, M. A., and Makeev, V. J. (2007). Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms Mol. Biol.* 2:13. doi: 10.1186/1748-7188-2-13

Boeva, V., Lermine, A., Barette, C., Guillouf, C., and Barillot, E. (2012). Nebula—a web-server for advanced ChIP-seq data analysis. *Bioinformatics* 28, 2517–2519. doi: 10.1093/bioinformatics/bts463

Boeva, V., Regnier, M., Papatsenko, D., and Makeev, V. (2006). Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* 22, 676–684. doi: 10.1093/bioinformatics/btk032

Boeva, V., Surdez, D., Guillon, N., Tirode, F., Fejes, A. P., Delattre, O., et al. (2010). *De novo* motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.* 38, e126. doi: 10.1093/nar/gkq217

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Burda, P., Laslo, P., and Stopka, T. (2010). The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* 24, 1249–1257. doi: 10.1038/leu.2010.104

Chekmenev, D. S., Haid, C., and Kel, A. E. (2005). P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.* 33, W432–W437. doi: 10.1093/nar/gki441

Chin, L. J., Ratner, E., Leng, S., Zhai, R., Nallur, S., Babar, I., et al. (2008). A SNP in a let-7 microRNA complementary site in the KRAS 3′ untranslated region increases non–small cell lung cancer risk. *Cancer Res.* 68, 8535–8540. doi: 10.1158/0008-5472.CAN-08-2129

Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* 28, 56–62. doi: 10.1093/bioinformatics/btr614

Dekker, J., and Heard, E. (2015). Structural and functional diversity of topologically associating domains. *FEBS Lett.* 589, 2877–2884. doi: 10.1016/j.febslet.2015.08.044

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., et al. (2012). Fast computation and applications of genome mappability. *PLoS ONE* 7:e30377. doi: 10.1371/journal.pone.0030377

Deyneko, I. V., Kel, A. E., Kel-Margoulis, O. V., Deineko, E. V., Wingender, E., and Weiss, S. (2013). MatrixCatch - a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinformatics* 14:241. doi: 10.1186/1471-2105-14-241

Eggeling, R., Gohr, A., Keilwagen, J., Mohr, M., Posch, S., Smith, A. D., et al. (2014). On the value of intra-motif dependencies of human insulator protein CTCF. *PLoS ONE* 9:e85629. doi: 10.1371/journal.pone.0085629

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 10, 605–616. doi: 10.1038/nrg2636

Fazius, E., Shelest, V., and Shelest, E. (2011). SiTaR: a novel tool for transcription factor binding site prediction. *Bioinformatics* 27, 2806–2811. doi: 10.1093/bioinformatics/btr492

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. M. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 24, 1729–1730. doi: 10.1093/bioinformatics/btn305

Fletez-Brant, C., Lee, D., McCallion, A. S., and Beer, M. A. (2013). kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* 41, W544–W556. doi: 10.1093/nar/gkt519

Frith, M. C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* 32, 1372–1381. doi: 10.1093/nar/gkh299

Frith, M. C., Li, M. C., and Weng, Z. (2003). Cluster-buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 31, 3666–3668. doi: 10.1093/nar/gkg540

Georgiev, S., Boyle, A. P., Jayasurya, K., Ding, X., Mukherjee, S., and Ohler, U. (2010). Evidence-ranked motif identification. *Genome Biol.* 11:R19. doi: 10.1186/gb-2010-11-2-r19

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

Gorkin, D. U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S. L., Loftus, S. K., et al. (2012). Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.* 22, 2290–2301. doi: 10.1101/gr.139360.112

Grant, C. E., Johnson, J., Bailey, T. L., and Noble, W. S. (2015). MCAST: scanning for cis-regulatory motif clusters. *Bioinformatics* btv750. doi: 10.1093/bioinformatics/btv750. [Epub ahead of print].

Grau, J., Ben-Gal, I., Posch, S., and Grosse, I. (2006). VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. *Nucleic Acids Res.* 34, W529–W533. doi: 10.1093/nar/gkl212

Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative *de novo* motif discovery from high-throughput data. *Nucleic Acids Res.* 41, e197. doi: 10.C/gkt831

Grinchuk, O. V., Motakis, E., Yenamandra, S. P., Ow, G. S., Jenjaroenpun, P., Tang, Z., et al. (2015). Sense-antisense gene-pairs in breast cancer and associated pathological pathways. *Oncotarget* 6, 42197–42221. doi: 10.18632/oncotarget.6255

Guillon, N., Tirode, F., Boeva, V., Zynovyev, A., Barillot, E., and Delattre, O. (2009). The oncogenic EWS-FLI1 protein binds *in vivo* ggaa microsatellite sequences with potential transcriptional activation function. *PLoS ONE* 4:e4932. doi: 10.1371/journal.pone.0004932

Guo, Y., Mahony, S., and Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* 8:e1002638. doi: 10.1371/journal.pcbi.1002638

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8:R24. doi: 10.1186/gb-2007-8-2-r24

Halperin, Y., Linhart, C., Ulitsky, I., and Shamir, R. (2009). Allegro: analyzing expression and sequence in concert to discover regulatory programs. *Nucleic Acids Res.* 37, 1566–1579. doi: 10.1093/nar/gkn1064

Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* 23, 181–194. doi: 10.1101/gr.139881.112

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

Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (2012). i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.* 40, e114. doi: 10.1093/nar/gks543

Hertz, G. Z., and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563–577. doi: 10.1093/bioinformatics/15.7.563

Holloway, D. T., Kon, M., and DeLisi, C. (2005). Integrating genomic data to predict transcription factor binding. *Genome Inform.* 16, 83–94.

Holub, J. (2012). The finite automata approaches in stringology. *Kybernetika* 3, 386–401.

Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M., and Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.* 38, 2154–2167. doi: 10.1093/nar/gkp1180

Imrichová, H., Hulselmans, G., Kalender Atak, Z., Potier, D., and Aerts, S. (2015). i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* 43, W57–W64. doi: 10.1093/nar/gkv395

Iseli, C., Ambrosini, G., Bucher, P., and Jongeneel, C. V. (2007). Indexing Strategies for rapid searches of short words in genome sequences. *PLoS ONE* 2:e579. doi: 10.1371/journal.pone.0000579

Jia, C., Carson, M. B., Wang, Y., Lin, Y., and Lu, H. (2014). A new exhaustive method and strategy for finding motifs in ChIP-enriched regions. *PLoS ONE* 9:e86044. doi: 10.1371/journal.pone.0086044

Jiang, B., Zhang, M. Q., and Zhang, X. (2007). OSCAR: one-class SVM for accurate recognition of cis-elements. *Bioinformatics* 23, 2823–2828. doi: 10.1093/bioinformatics/btm473

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

Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* 11, 203–209. doi: 10.1038/nmeth.2766

Keilwagen, J., and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* 43, e119–e119. doi: 10.1093/nar/gkv577

Kim, H., Kim, J., Selby, H., Gao, D., Tong, T., Phang, T. L., et al. (2011). A short survey of computational analysis methods in analysing ChIP-seq data. *Hum. Genomics* 5, 117–123. doi: 10.1186/1479-7364-5-2-117

Klose, R. J., Cooper, S., Farcas, A. M., Blackledge, N. P., and Brockdorff, N. (2013). Chromatin sampling—an emerging perspective on targeting polycomb repressor proteins. *PLoS Genet.* 9:e1003717. doi: 10.1371/journal.pgen.1003717

Kozlov, K., Gursky, V. V., Kulakovskiy, I. V., Dymova, A., and Samsonova, M. (2015). Analysis of functional importance of binding sites in the Drosophila gap gene network model. *BMC Genomics* 16(Suppl. 13):S7. doi: 10.1186/1471-2164-16-S13-S7

Kulakovskiy, I., Levitsky, V., Oshchepkov, D., Bryzgalov, L., Vorontsov, I., and Makeev, V. (2013). From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.* 11, 1340004. doi: 10.1142/S0219720013400040

Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V., and Makeev, V. J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 26, 2622–2623. doi: 10.1093/bioinformatics/btq488

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., et al. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* 41, D195–D202. doi: 10.1093/nar/gks1089

Kwon, A. T., Arenillas, D. J., Hunt, R. W., and Wasserman, W. W. (2012). oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-seq datasets. *G3* 2, 987–1002. doi: 10.1534/g3.112.003202

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Lihu, A., and Holban, Ş. (2015). A review of ensemble methods for *de novo* motif discovery in ChIP-Seq data. *Brief. Bioinformatics* 16, 964–973. doi: 10.1093/bib/bbv022

Lin, J. M., Collins, P. J., Trinklein, N. D., Fu, Y., Xi, H., Myers, R. M., et al. (2007). Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res.* 17, 818–827. doi: 10.1101/gr.5623407

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

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., et al. (2011). Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* 12:R83. doi: 10.1186/gb-2011-12-8-r83

Loo, P. V., Aerts, S., Thienpont, B., Moor, B. D., Moreau, Y., and Marynen, P. (2008). ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol.* 9:R66. doi: 10.1186/gb-2008-9-4-r66

Ma, X., Kulkarni, A., Zhang, Z., Xuan, Z., Serfling, R., and Zhang, M. Q. (2012). A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res.* 40, e50–e50. doi: 10.1093/nar/gkr1135

Machanick, P., and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697. doi: 10.1093/bioinformatics/btr189

Mahony, S., and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 35, W253–W258. doi: 10.1093/nar/gkm272

Marschall, T. (2011). Construction of minimal deterministic finite automata from biological motifs. *Theor. Comput. Sci.* 412, 922–930. doi: 10.1016/j.tcs.2010.12.003

Marschall, T., and Rahmann, S. (2008). "Probabilistic arithmetic automata and their application to pattern matching statistics," in *Combinatorial Pattern Matching Lecture Notes in Computer Science*, eds. P. Ferragina and G. M. Landau (Heidelberg: Springer), 95–106. Available online at: http://link.springer.com.gate2.inist.fr/chapter/10.1007/978-3-540-69068-9_11 (Accessed December 21, 2015).

Marstrand, T. T., Frellsen, J., Moltke, I., Thiim, M., Valen, E., Retelska, D., et al. (2008). Asap: a framework for over-representation statistics for transcription factor binding sites. *PLoS ONE* 3:e1623. doi: 10.1371/journal.pone.0001623

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-Y., Denay, G., Lee, J., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 44, D110–D115. doi: 10.1093/nar/gkv1176

Mathelier, A., and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 9:e1003214. doi: 10.1371/journal.pcbi.1003214

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

McLeay, R. C., and Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11:165. doi: 10.1186/1471-2105-11-165

McLeay, R. C., Leat, C. J., and Bailey, T. L. (2011). Tissue-specific prediction of directly regulated genes. *Bioinformatics* 27, 2354–2360. doi: 10.1093/bioinformatics/btr399

Meckbach, C., Tacke, R., Hua, X., Waack, S., Wingender, E., and Gültas, M. (2015). PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformatics* 16:400. doi: 10.1186/s12859-015-0827-2

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., et al. (2015). RSAT 2015: regulatory sequence analysis tools. *Nucleic Acids Res.* 43, W50–W56. doi: 10.1093/nar/gkv362

Mordelet, F., Horton, J., Hartemink, A. J., Engelhardt, B. E., and Gordân, R. (2013). Stability selection for regression-based models of transcription factor–DNA binding specificity. *Bioinformatics* 29, i117–i125. doi: 10.1093/bioinformatics/btt221

Navarro, G., and Raffinot, M. (2002). *Flexible Pattern Matching in Strings: Practical On-line Search Algorithms for Texts and Biological Sequences*. New York, NY: Cambridge University Press.

Nikulova, A. A., Favorov, A. V., Sutormin, R. A., Makeev, V. J., and Mironov, A. A. (2012). CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation. *Nucleic Acids Res.* 40, e93. doi: 10.1093/nar/gks235

Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.* 9, 2944–2949. doi: 10.1128/MCB.9.7.2944

Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.* 35, D127–D131. doi: 10.1093/nar/gkl857

Politi, V., Perini, G., Trazzi, S., Pliss, A., Raska, I., Earnshaw, W. C., et al. (2002). CENP-C binds the alpha-satellite DNA *in vivo* at specific centromere domains. *J. Cell. Sci.* 115, 2317–2327. Available online at: http://jcs.biologists.org/content/115/11/2317.long

Ramsingh, G., Koboldt, D. C., Trissal, M., Chiappinelli, K. B., Wylie, T., Koul, S., et al. (2010). Complete characterization of the microRNAome in a patient with acute myeloid leukemia. *Blood* 116, 5316–5326. doi: 10.1182/blood-2010-05-285395

Reid, J. E., Evans, K. J., Dyer, N., Wernisch, L., and Ott, S. (2010). Variable structure motifs for transcription factor binding sites. *BMC Genomics* 11:30. doi: 10.1186/1471-2164-11-30

Rhee, H. S., and Pugh, B. F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* 147, 1408–1419. doi: 10.1016/j.cell.2011.11.013

Ridinger-Saison, M., Boeva, V., Rimmelé, P., Kulakovskiy, I., Gallais, I., Levavasseur, B., et al. (2012). Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res.* 40, 8927–8941. doi: 10.1093/nar/gks659

Riggi, N., Knoechel, B., Gillespie, S. M., Rheinbay, E., Boulay, G., Suvà, M. L., et al. (2014). EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in ewing sarcoma. *Cancer Cell* 26, 668–681. doi: 10.1016/j.ccell.2014.10.004

Rimmelé, P., Komatsu, J., Hupé, P., Roulin, C., Barillot, E., Dutreix, M., et al. (2010). Spi-1/PU.1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage. *Cancer Res.* 70, 6757–6766. doi: 10.1158/0008-5472.CAN-09-4691

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100. doi: 10.1093/nar/18.20.6097

Sebastian, A., and Contreras-Moreira, B. (2014). footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30, 258–265. doi: 10.1093/bioinformatics/btt663

Shelest, V., Albrecht, D., and Shelest, E. (2010). DistanceScan: a tool for promoter modeling. *Bioinformatics* 26, 1460–1462. doi: 10.1093/bioinformatics/btq132

Shi, X. M., Blair, H. C., Yang, X., McDonald, J. M., and Cao, X. (2000). Tandem repeat of C/EBP binding sites mediates PPARgamma2 gene transcription in glucocorticoid-induced adipocyte differentiation. *J. Cell. Biochem.* 76, 518–527. doi: 10.1002/(SICI)1097-4644(20000301)76:3%3C518::AID-JCB18%3E3.0.CO;2-M

Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., et al. (2015). ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Res.* 25, 825–835. doi: 10.1101/gr.185157.114

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

Sun, H., De Bie, T., Storms, V., Fu, Q., Dhollander, T., Lemmens, K., et al. (2009). ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC Bioinformatics* 10:S30. doi: 10.1186/1471-2105-10-S1-S30

Sun, H., Guns, T., Fierro, A. C., Thorrez, L., Nijssen, S., and Marchal, K. (2012). Unveiling combinatorial regulation through the combination of ChIP information and *in silico* cis-regulatory module detection. *Nucleic Acids Res.* 40, e90–e90. doi: 10.1093/nar/gks237

Tran, N. T. L., and Huang, C.-H. (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol. Direct* 9:4. doi: 10.1186/1745-6150-9-4

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* 439, 871–874. doi: 10.1038/nature04431

Vorontsov, I. E., Kulakovskiy, I. V., and Makeev, V. J. (2013). Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.* 8:23. doi: 10.1186/1748-7188-8-23

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

Ward, L. D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 44, D877–D881. doi: 10.1093/nar/gkv1340

Wasson, T., and Hartemink, A. J. (2009). An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* 19, 2101–2112. doi: 10.1101/gr.093450.109

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31, 126–134. doi: 10.1038/nbt.2486

Wilbanks, E. G., and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* 5:e11471. doi: 10.1371/journal.pone.0011471

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

Yue, D., Liu, H., and Huang, Y. (2009). Survey of computational algorithms for microRNA target prediction. *Curr. Genomics* 10, 478–492. doi: 10.2174/138920209789208219

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* 37, W247–W252. doi: 10.1093/nar/gkp464

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

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

Zhao, Y., and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29, 480–483. doi: 10.1038/nbt.1893

Zheng, J., Wu, J., and Sun, Z. (2003). An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res.* 31, 1995–2005. doi: 10.1093/nar/gkg287

Zhong, S., He, X., and Bar-Joseph, Z. (2013). Predicting tissue specific transcription factor binding sites. *BMC Genomics* 14:796. doi: 10.1186/1471-2164-14-796

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 March 2016
doi: 10.3389/fgene.2016.00033

Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development

Sebastian Zeidler[1, 2, 3], Cornelia Meckbach[1], Rebecca Tacke[1], Farah S. Raad[2, 3], Angelica Roa[2, 3], Shizuka Uchida[4, 5], Wolfram-Hubertus Zimmermann[2, 3], Edgar Wingender[1, 3] and Mehmet Gültas[1]*

[1] University Medical Center Göttingen, Institute of Bioinformatics, Georg-August-University Göttingen, Göttingen, Germany, [2] Heart Research Center Göttingen, University Medical Center Göttingen, Institute of Pharmacology and Toxicology, Georg-August-University Göttingen, Göttingen, Germany, [3] DZHK (German Centre for Cardiovascular Research), Göttingen, Germany, [4] Institute of Cardiovascular Regeneration, Goethe University Frankfurt, Frankfurt, Germany, [5] DZHK (German Centre for Cardiovascular Research), Frankfurt, Germany

Transcription factors (TFs) regulate gene expression in living organisms. In higher organisms, TFs often interact in non-random combinations with each other to control gene transcription. Understanding the interactions is key to decipher mechanisms underlying tissue development. The aim of this study was to analyze co-occurring transcription factor binding sites (TFBSs) in a time series dataset from a new cell-culture model of human heart muscle development in order to identify common as well as specific co-occurring TFBS pairs in the promoter regions of regulated genes which can be essential to enhance cardiac tissue developmental processes. To this end, we separated available RNAseq dataset into five temporally defined groups: (i) mesoderm induction stage; (ii) early cardiac specification stage; (iii) late cardiac specification stage; (iv) early cardiac maturation stage; (v) late cardiac maturation stage, where each of these stages is characterized by unique differentially expressed genes (DEGs). To identify TFBS pairs for each stage, we applied the MatrixCatch algorithm, which is a successful method to deduce experimentally described TFBS pairs in the promoters of the DEGs. Although DEGs in each stage are distinct, our results show that the TFBS pair networks predicted by MatrixCatch for all stages are quite similar. Thus, we extend the results of MatrixCatch utilizing a Markov clustering algorithm (MCL) to perform network analysis. Using our extended approach, we are able to separate the TFBS pair networks in several clusters to highlight stage-specific co-occurences between TFBSs. Our approach has revealed clusters that are either common (NFAT or HMGIY clusters) or specific (SMAD or AP-1 clusters) for the individual stages. Several of these clusters are likely to play an important role during the cardiomyogenesis. Further, we have shown that the related TFs of TFBSs in the clusters indicate potential synergistic or antagonistic interactions to switch between different stages. Additionally, our results suggest that cardiomyogenesis follows the hourglass model which was already proven for *Arabidopsis* and some vertebrates. This investigation helps us to get a better understanding of how each stage of cardiomyogenesis is affected by different combination of TFs. Such knowledge may help to understand basic principles of stem cell differentiation into cardiomyocytes.

Keywords: cardiomyogenesis, engineered heart muscle, MatrixCatch, Markov clustering, transcription factor collaboration

OPEN ACCESS

Edited by:
Yasset Perez-Riverol, European Bioinformatics Institute, UK

Reviewed by:
Mikhail P. Ponomarenko, Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences, Russia
Ka-Chun Wong, City University of Hong Kong, China

****Correspondence:***
Sebastian Zeidler
sebastian.zeidler@bioinf.med.uni-goettingen.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *05 November 2015*
Accepted: *23 February 2016*
Published: *23 March 2016*

Citation:
Zeidler S, Meckbach C, Tacke R, Raad FS, Roa A, Uchida S, Zimmermann WH, Wingender E and Gültas M (2016) Computational Detection of Stage-Specific Transcription Factor Clusters during Heart Development. Front. Genet. 7:33. doi: 10.3389/fgene.2016.00033

1. INTRODUCTION

Transcription factors (TFs) regulate the expression of genes and genetic programs to maintain survival and adaption to the environment in adult organisms as well as in embryo- and organogenesis. Most of them bind to recognized specific sequences in the DNA regulatory regions of genes and modify transcription, such as the assembly of the gene expression machinery. In mammalian tissues TFs often work in combinatorial interactions for precise regulation of specific programs (Boyer et al., 2005; Odom et al., 2006; Hu and Gallo, 2010; Neph et al., 2012). Such interactions can be positive, resulting in an enhanced expression of a gene or negative, resulting in reduced expression of a target gene. Thus, the identification of co-occurring transcription factor binding sites (TFBSs) in the promoter regions of regulated genes indicate potential combinatorial interactions between TFs that are important for understanding the molecular mechanisms, e.g., of tissue development during embryogenesis.

The human heart is the first organ formed during embryogenesis (Kirby, 2002; Brand, 2003; Buckingham et al., 2005; Brewer and Pizzey, 2006; Schleich et al., 2013), and it consists of different cell types, which develop simultaneously and are regulated by TFs as well as their combinatorial interactions. Until now, several groups analyzed TFs and their influence on cardiac development (Ryan and Chin, 2003; Pikkarainen et al., 2004; Peterkin et al., 2005; Brewer and Pizzey, 2006; Martin et al., 2010; Shi and Jin, 2010; Turbendian et al., 2013; Chaudhry et al., 2014; Takeuchi, 2014; Wang and Jauch, 2014). These studies mainly focus on individual TFs or their related families e.g., GATA family, TBX family, or NKX2 family (Ryan and Chin, 2003; Pikkarainen et al., 2004; Miura and Yelon, 2013; Turbendian et al., 2013). However, a detailed analysis of interactions between TFs and their role in cardiac development is limited to interactions between known cardiac TFs like NKX2-5 or MEF2 which are essential for the generation of cardiac tissues from stem cells (Martin et al., 2010; Sylva et al., 2014; Takeuchi, 2014). A complete survey of potential TF interactions by co-occurring TFBSs in the promoter regions of genes which regulate cardiac development is still missing, but needed to understand embryonic cardiac development, in particular of cardiomyocytes (CMs).

CMs comprise the most important functional cells in the human heart (Ye et al., 2013; Sylva et al., 2014). CMs show a limited potential to regenerate after myocardial infarction or other cardiovascular diseases (CVDs), which is at maximum 50% CM renewal per lifetime and less than 1% per year (Bergmann et al., 2009; Sylva et al., 2014; Takeuchi, 2014). Replacing CMs in elderly by for example enhanced cardiomyocyte proliferation may improve the quality of their life, but requires an understanding of how CMs develop and of how they can be replaced (Akhurst, 2012; Ye et al., 2013; Euler, 2015).

One approach is to apply tissue engineered myocardium to restore muscle mass and thus reintroduce contractility (Zimmermann et al., 2006). Such tissues can be generated from embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), or parthenogenetic stem cells (Soong et al., 2012; Didié et al., 2013; Ye et al., 2013; Tiburcy and Zimmermann, 2014). Controlling cardiomyogenesis *in vitro* requires insight into biological processes governing embryonic heart development. To understand cardiac development from a systems biology perspective, identification of the mechanisms controlling the expression of fate determining TFs and their regulation of transcription are of fundamental importance. Co-occurring TFBSs in the regulatory regions of genes which are specific for a particular developmental stage reveal potential TF interactions that are likely to regulate these stages. There are in fact plenty of TF-TF interactions known as implicated in organogenesis, but the specific time points when particular interactions occur, are difficult to obtain and mostly not annotated in public databases. Only intense literature surveys provide such information.

Recent studies identifying the co-occurrence of TF pairs focus either on combinatorial approaches where e.g., specific DNA-sequences bound by different TFs simultaneously were selected from a library of random sequences (Jolma et al., 2015) or approaches that focus on data integration e.g., ChIP-seq, SELEX together with Hi-C to reveal long-range chromatin interactions (Jolma et al., 2013; Wong et al., 2016). Although the selection of interacting TF pairs from a library of random sequences underpins potential interactions of TFs, it does not give any hints on the actual interactions in particular cell types or tissues. Data integration and especially Hi-C technology is very promising for the future, but currently there is a lack in publicly available data sets that cover the time dependent organogenesis of the human heart.

In this study we analyze a time series dataset obtained from RNAseq at different time points of in vitro cardiomyogenesis (Hudson et al.; in revision) to identify co-occurring TFBSs which indicate potential interacting TFs that are crucial for understanding the gene regulatory mechanisms during the heart development. The dataset consists of six different time points (day: 0, 3, 8, 13, 29, and 60) where the gene expression in the tissue culture was measured by RNAseq. The data comprises early heart development in general and can be differentiated in the following major developmental stages: (i) mesoderm induction stage (day 0–day 3); (ii) cardiac specification stage (day 3–day 13; early 3–8, late 8–13); (iii) cardiac maturation stage (day 13–day 60; early 13–29, late 29–60). For each stage we determined the set of unique differentially expressed genes (DEGs) utilizing *limma* on the FPKM-values in the dataset (Smyth, 2004). To identify specific TF interactions in individual stages, we analyzed the promoter sequences of corresponding DEGs employing the MatrixCatch approach (Deyneko et al., 2013). As a result, we observed a set of co-occurring TFBSs for each stage whose corresponding TFs are likely to represent potential core regulators of a particular developmental stage. Although the analyzed DEGs are unique in each stage, the identified TFBS pairs are highly overlapping between stages. To overcome this problem in MatrixCatch results, we further applied Markov clustering algorithm (MCL; Dongen, 2000) for the detection of clusters which contain stage specific co-occurrences between TFBSs. In recent years, MCL has gained great attention in the bioinformatics community for the detection of high-quality clusters in biological networks due to its highly effective

and successful algorithm. Especially, for the clustering of protein-protein interaction networks, several studies have shown that MCL is superior to conventional clustering approaches in terms of detection of high-quality and more accurate functional clusters (Brohée and van Helden, 2006; Vlasblom and Wodak, 2009; Shih and Parthasarathy, 2012). These articles encouraged us to utilize MCL for the elimination of negligible pairs at each stage and thus for the determination of remaining TFBS pairs, which may play crucial roles during cardiomyogenesis. To this end, we focused on clusters whose central binding site is present at almost all stages, but its partners differ stage-specifically. These clusters may regulate DEGs in each stage and are likely to be fundamentally implicated in cardiac muscle development.

2. MATERIALS AND METHODS

In this section we describe the differentially expressed genes analyzed and the methods applied and partly developed. Our analysis follows the structure of **Figure 1**.

2.1. Selection of Differentially Expressed Genes

The data, available as a FPKM normalized RNAseq time series, was mapped to corresponding gene symbols (hgnc-symbols) and further analyzed using *limma* package from the Bioconductor project for R with standard procedures (Smyth, 2004; R Core Team, 2015). The time series data describe human cardiomyogenesis in vitro at time points day 0, 3, 8, 13, 29, and 60, whereas day 0 resembles blastocyst stage development and day 60 early fetal stages (Hudson et al.; in revision). We calculated DEGs between two time points which define a particular developmental stage where: (i) day 0–3 defines the mesoderm induction stage; (ii) day 3–8 early cardiac specification; (iii) day 8–13 late cardiac specification; (iv) day 13–29 early cardiac maturation and; (v) day 29–60 the late cardiac maturation stage (this stage describes the transition from an embryonic to a fetal cardiac maturation stage). We filtered the set of all DEGs for protein coding genes (excluding TFs) and their uniqueness in a stage by comparison to all other stages with p-value ≤ 0.05 and FDR ≤ 0.01 (see **Supplementary File 1**). A heatmap of stage-specific DEGs is given in **Supplementary File 2**.

2.2. Promoter Sequences

Using UCSC genome browser (Karolchik et al., 2004), we extracted for each protein coding gene (RefSeq gene) based on its annotated transcription start site (TSS) the -1 kb putative regulatory promoter region.

It is important to note that, according to TSS annotations, a RefSeq gene can have multiple overlapping promoter regions which results in overestimation of the importance of some transcription factor binding sites (TFBSs). Thus, following the line of PC-TraFF to remove the redundancy between sequences, we filtered them regarding their TSSs (Meckbach et al., 2015). Consequently, we used in our analysis only those sequences which have no overlap.

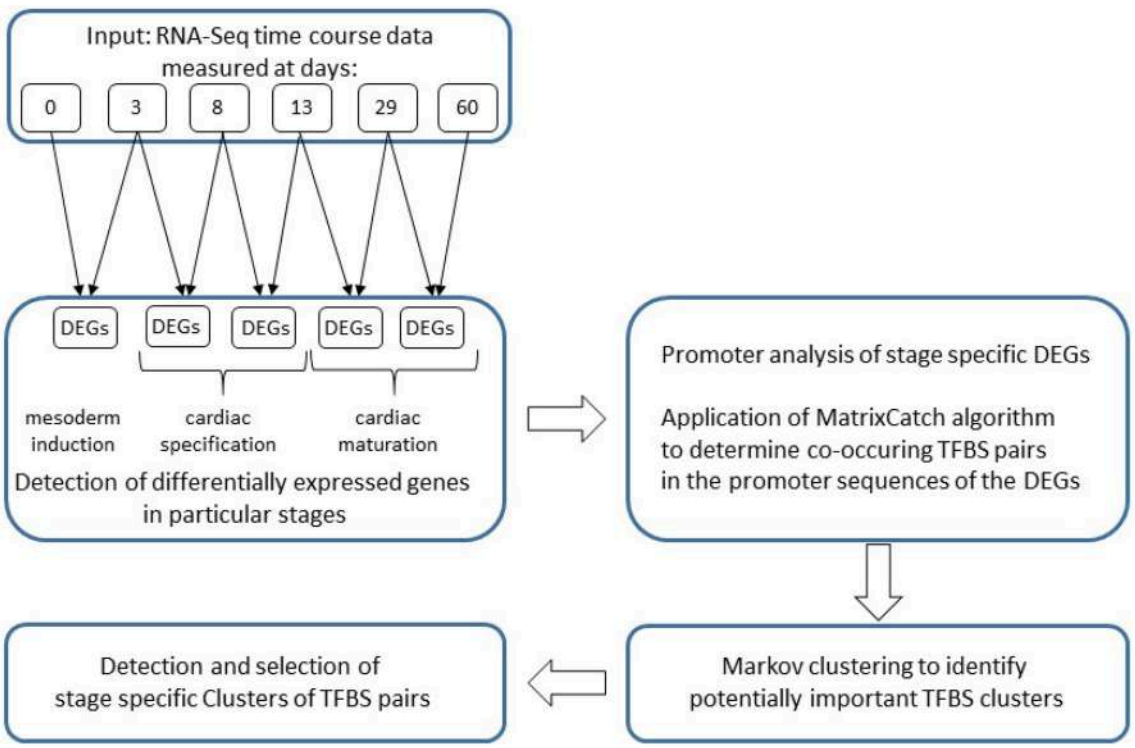


FIGURE 1 | Flowchart of the analysis applied in this study.

In this study, the assembly of the hg19 release of the human genome was used and only UCSC track refGene annotations were considered which correspond to the chromosomes chr1-chr22, chrX, and chrY.

2.3. MatrixCatch Analysis

MatrixCatch is a novel method introduced by Deyneko et al. (2013) to recognize experimentally verified TF pairs based on the co-localization of their TFBSs, known as composite regulatory modulues (CRMs), in single promoters. To detect CRMs in the individual sequences under study, MatrixCatch scans each sequence and its reverse complement using a special library of position weight matrices (PWMs). This library has been specified by considering the TF binding scores, relative orientations and distances between TFs that are experimentally known to interact, as documented in the TRANSCompel database (Kel-Margoulis et al., 2002). Consequently, the usage of MatrixCatch yields an important practical advantage since this method provides a high number of known CRMs in sequences with their biological interpretation (for details, see Deyneko et al., 2013).

In our study, we applied MatrixCatch to the promoter sequences of the filtered DEGs of the different heart developmental stages. As we have recently suggested in PC-TraFF (Meckbach et al., 2015), we prefer in this study the usage of TFBS pairs instead of CRMs, since those pairs were detected in a set of sequences. This indicates the importance of potential collaborations between corresponding TFs in the gene set of interest.

2.4. Clustering of Co-Occurring TFBSs

Since MatrixCatch provides all detected TFBS pairs of experimentally verified TF interactions in promoters, the detected pairs are highly overlapping between developmental stages. To differentiate stage specific roles of TFBS pairs, we first determined the frequency of each pair in MatrixCatch results. After that, we applied the Markov clustering algorithm (MCL; Dongen, 2000) which is able to eliminate negligible TFBS pairs based on their frequencies at each stage. To this end, we constructed an interaction network based on the TFBS pairs for each heart developmental stage, where nodes are TFBSs and edges display the co-occurrences between them.

Let $\mathcal{N} := (\mathcal{V}, \mathcal{E})$ be an undirected interaction network of TFBS pairs where any two elements $(v_i, v_j \in \mathcal{V})$ of $\mathcal{N}$ are connected by an edge $e_{(v_i,v_j)}$ belonging to $\mathcal{E}$, if and only if the corresponding TFBS pair was identified by MatrixCatch. Further, $w(v_i, v_j)$ denotes the weight of an edge $e_{(v_i,v_j)}$, which represents the observed frequency of the TFBS pair (v_i, v_j) found by MatrixCatch in the promoter sequences of genes under study.

Based on the weights of edges, an adjacency matrix $\mathcal{A}_{n\times n}$ of each network was constructed as

$$\mathcal{A}_{i,j} = \begin{cases} w(v_i, v_j) & \text{if } e_{(v_i,v_j)} \in \mathcal{E} \\ 0 & \text{else.} \end{cases}$$

$\mathcal{A}_{n\times n}$ was then converted into a row stochastic "Markov" matrix $\mathcal{M}_{n\times n}$, where $m_{i\times j}$ represents the transition probability between nodes v_i and v_j in the network under study. The most common way to construct a row stochastic transition matrix $\mathcal{M}$ is the normalization of rows in $\mathcal{A}$ to sum to 1. This process can be simply given as: $\mathcal{M} = \Delta^{-1} \cdot \mathcal{A}$, where Δ is a $n \times n$ diagonal degree matrix and defined as:

$$\Delta = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & d_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j} & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^{n} a_{2j} & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \sum_{j=1}^{n} a_{nj} \end{pmatrix}$$

Based on matrix $\mathcal{M}$, we employed MCL (Dongen, 2000) to detect densely connected TFBSs in each network. Briefly, the basic intuition of MCL was based on a simulation of stochastic flows on the underlying interaction network to separate high-flow regions from low-flow regions. To this end, Expand and Inflate operations were applied on $\mathcal{M}$ until $\mathcal{M}$ reaches its steady state. While the Expand operation corresponds to matrix multiplication ($\mathcal{M} = \mathcal{M} \times \mathcal{M}$), the Inflate operation is used to increase the contrast between higher and lower probability transitions by taking each entry $m_{i\times j}$ in $\mathcal{M}$ to the power of inflation parameter $r > 1$. Finally, $\mathcal{M}$ was re-normalized into a row stochastic matrix. The pseudo-code for MCL is given in Algorithm 1.

Algorithm 1 : Markov Clustering Algorithm

Input: $\mathcal{M}$ and $r > 1$
Output: $\mathcal{C}$: A list of clusters
Methode:

1: $t = 0$
2: $\mathcal{M}_t = \mathcal{M}$
3: **repeat**
4: $\quad t = t + 1$
5: $\quad \mathcal{M}_t = Expand(\mathcal{M}_{t-1}) = \mathcal{M}_{t-1} \times \mathcal{M}_{t-1}$
6: $\quad \mathcal{M}_t = Inflate(\mathcal{M}_t, r) = \left\{ \frac{(m_{ij})^r}{\sum_{k=1}^{n}(m_{ik})^r} \right\}_{i,j=1}^{n}$
7: **until** $\mathcal{M}_t$ converges
8: $\mathcal{C}$: clusters($\mathcal{M}_t$)

3. RESULTS

We analyzed a time course data set which covers heart muscle development in human embryonic stem cell derived tissue cultures at days 0, 3, 8, 13, 29, and 60 (Hudson et al., in revision). These time points cover the mesoderm induction stage (day 0–day 3), the cardiac specification stage (day 3–day 13), and the cardiac maturation stage (day 13–day 29). We further defined cardiac specification and cardiac maturation into two more stages, i.e.,: (i) early cardiac specification and maturation stage from days 3–8 and days 13–29, respectively; (ii) late cardiac specification and maturation with transition from embryonic to fetal stages defined by culture days 8–13 and days 29–60, respectively. By comparison of neighboring time points, for each stage, we determined the set of DEGs and filtered them according to their uniqueness in a particular stage. Afterwards, we utilized

MatrixCatch to identify co-occurring pairs of TFBSs in the promoter regions of these DEGs. Consequently, we identified: (i) 63 TFBS pairs based on 429 DEGs for the mesoderm induction stage; (ii) 82 TFBS pairs based on 1233 DEGs for the early cardiac specification stage; (iii) 24 TFBS pairs based on 36 DEGs for the late cardiac specification stage; (iv) 52 TFBS pairs based on 205 DEGs for the early cardiac maturation stage; (v) 76 TFBS pairs based on 964 DEGs for the late cardiac maturation stage (see **Supplementary File 3**).

Due to underlying methodology of MatrixCatch, the detected TFBS pairs show a large overlap between different stages although they may play different roles in these stages. To reduce this drawback of MatrixCatch, we further applied Markov clustering algorithm that seeks to remove negligible TFBS pairs by emphasizing the roles of remaining pairs at each stage. Consequently, we obtained (i) 19 clusters for the mesoderm induction stage; (ii) 25 clusters for the early cardiac specification stage; (iii) 11 clusters for the late cardiac specification stage; (iv) 21 clusters for the early cardiac maturation stage, and (v) 24 clusters for the late cardiac maturation stage (see **Supplementary File 4**).

We focused only on clusters with V$AP1_01, V$HMGIY_Q6, V$SMAD_Q6_01, and V$NFAT_Q6 binding sites in their center (see **Figure 2**), because these clusters contain at least three interactions and the changes in their constitution provide crucial information about different cardiac developmental stages. We analyzed the TFBS pairs in these clusters according to their potential role in cardiac development. We omitted clusters, when the expression values of TF genes are below a certain threshold or their importance in heart development is currently unknown. For our analysis, we applied a FPKM threshold value of 10, which discriminates robustly between expressed TF genes and low or not expressed TF genes.

3.1. AP-1-Cluster

The AP-1-cluster is an assembly of different TFBSs with the V$AP1_01 binding site in its center (see **Figure 2A**). As described in **Table 1** and in **Figure 3**, V$AP1_01 binding site co-occurs with V$OCT_C binding site during mesoderm induction (< day 3) and early cardiac specification stage (day 3–day 8) and at late cardiac maturation stage (> day 29). Further, V$AP1_01 co-occurs with V$GATA_Q6 binding site at all stages except days 8–13. Interestingly, a co-occurring pair between V$AP1_01 and V$HNF4_Q6 binding site was detected only between day 3 and day 8. Additionally, **Figure 3** shows for these TFBSs the related TF genes which are expressed in at least one time point.

AP-1 is a family of leucine zipper transcription factors (bZIP) which forms homo- or heterodimers composed of proteins belonging to JUN or FOS protein families (Shaulian and Karin, 2002; Hess et al., 2004; Shaulian, 2010). AP-1 plays a role in the regulation of general functions like proliferation, differentiation, and apoptosis. We identified that V$AP1_01 co-occurs with V$OCT_C binding sites which are bound by AP-1 and POU-domain factors like POU5F1, respectively. POU5F1 is also known as OCT-4, which is an important pluripotency maintenance factor (Schöler et al., 1990; Nichols et al., 1998; Pesce and Schöler, 2001; Guo et al., 2002). Regarding the expression values, POU5F1 shows higher expression in early stages (< day 8) and is absent after day 13 (see **Figure 4B**). This is in contrast to AP-1, where AP-1 components (FOS as well as JUN) are not present or only present at reduced levels during early stages, but they show increased expression values after day 13 (see **Figure 4A**). This suggests that AP-1 may not be formed during early stages, where POU5F1 controls the associated genes, and that during the late cardiac maturation stage (> day 29) the analyzed genes are under control of AP-1.

Our analysis identified a co-occurrence of V$AP1_01 with V$GATA_Q6 binding sites. GATA factors form a protein family of six zinc finger transcription factors that share a highly conserved DNA-binding sequence (Orkin, 1992; Ohneda and Yamamoto, 2002; Pikkarainen et al., 2004; Brewer and Pizzey, 2006). As suggested in Brewer and Pizzey (2006), the family can be dissected into two subfamilies (GATA-1,2,3 and GATA-4,5,6), based on their expression levels in different tissues, where only GATA -4, -5 and -6 are associated with cardio- and organogenesis (Pikkarainen et al., 2004; Peterkin et al., 2005; Brewer and Pizzey, 2006; Whitfield et al., 2012; Turbendian et al., 2013). We found only GATA4 and GATA6 to be expressed. Interactions between GATA-factors and AP-1 are well known, especially co-occurrence of AP-1 together with GATA-4 in several heart cell types and in Leydig cells (Herzig et al., 1997; Suzuki et al., 1999; Schröder et al., 2006; Linnemann et al., 2011; Martin et al., 2012). In our system, GATA6 was expressed in high amounts during the mesoderm induction (< day 3) and early cardiac specification

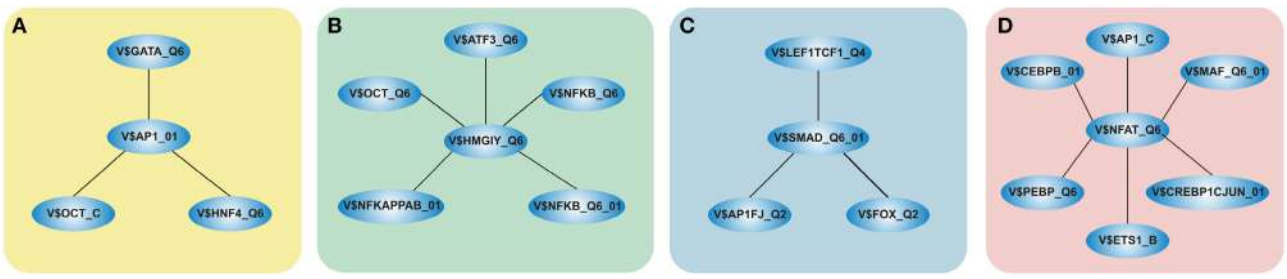


FIGURE 2 | Clusters we focus on in our analysis in the order in which they are analyzed in this study. The clusters comprise all interactions during the complete time course, identified by employing MatrixCatch and MCL. The constitution of each cluster for a particular stage is shown in the corresponding tables. **(A)** shows the AP-1-cluster, **Table 1**; **(B)** HMGIY-cluster, **Table 2**; **(C)** SMAD-cluster, **Table 4**; **(D)** NFAT-cluster, **Table 5**.

TABLE 1 | TFBS pairs within the AP-1-cluster.

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V$AP1_01 – V$OCT_C	+	+	–	–	+
V$AP1_01 – V$GATA_Q6	+	+	–	+	+
V$AP1_01 – V$HNF4_Q6	–	+	–	–	–

Constitution of co-occurring pairs in the AP-1-cluster, a "+" indicates the presence of a pair; a "–" its absence. During the late stage of cardiac specification (Day8–Day13), the cluster is completely absent.

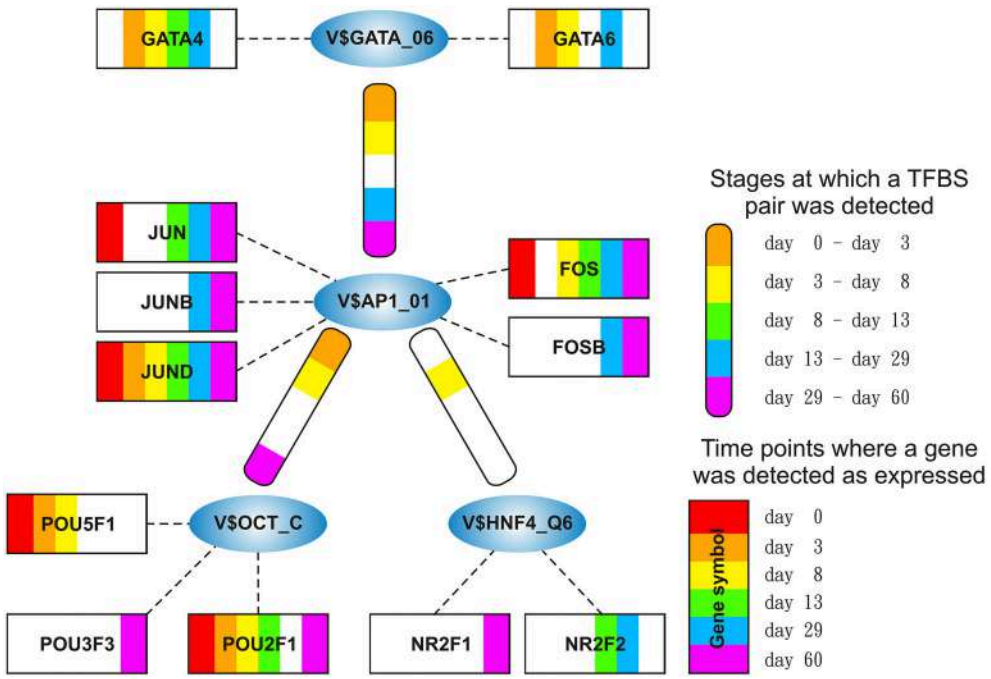


FIGURE 3 | Stage specific representation of TFBSs and the expression of associated TF genes, referring to Figure 2A. The encircled nodes represent the found TFBSs which are connected by color-coded round-edged rectangles which highlight stages where a TFBS pair was found. TF genes which are associated to TFBSs are linked by dashed lines. The TF genes are represented by color-coded rectangles representing the presence at a partiular time point. The absence of a TF gene during a particular time point or the absence of a pair during a particular stage is encoded in white. Both, the color-code for the stage specificity as well as for the gene expression of a TF gene is shown on the bottom right side. TF genes which are associated to a TFBS but are in all time points below the set threshold are omitted.

stage (day 3–day 8) but was not expressed or only at minor extent during cardiac maturation (> day 13, see **Figure 4C**). In contrast, GATA4 was expressed in high amounts during the late cardiac specification stage as well as during cardiac maturation (> day 8). The missing of AP-1 during mesoderm induction (< day 3) suggests that genes specific for mesoderm induction might be under control of GATA-6, whereas GATA-4 and AP-1 may regulate genes during cardiac maturation (> day 13), synergistically (see Pikkarainen et al., 2004 for the role of GATA-4 and GATA-6).

The role of the co-occurrence between V$AP1_01 and V$HNF4_Q6, which represents a binding site for HNF4A or HNF4G TFs, during cardiomyogenesis is uncertain. This TFBS pair was detected during early cardiac specification stage (days 3–8), but no expression of the related genes could be found. As mentioned before, the formation of AP-1 during this stage at relevant levels is uncertain (see **Figure 4A**), due to the low expression of the AP-1 components. Furthermore, the role of HNF4-genes, which where frequently reported to be associated with lipid metabolism in the liver (Watt et al., 2003; Chandra et al., 2013), during cardiac development is still unclear, but may point to changes in the metabolism at this stage.

3.2. HMGIY-Cluster

The HMGIY-cluster is assembled in a total of five TFBS pairs (see **Figures 2B**, **5**) with the V$HMGIY_Q6 binding site in its center. **Table 2** shows the co-occurring TFBS pairs of this cluster and **Figure 5** shows for these TFBSs the related TF genes which are expressed in at least one time point. The TFBS pair V$HMGIY_Q6 - V$OCT_Q6 was found during all stages and the co-occurrence between V$HMGIY_Q6 and V$ATF3_Q6 binding sites was found at days 3–8, and after day 29. Interestingly, we found in this cluster three binding sites, namely V$NFKAPPAB_01, V$NFKB_Q6_01, and

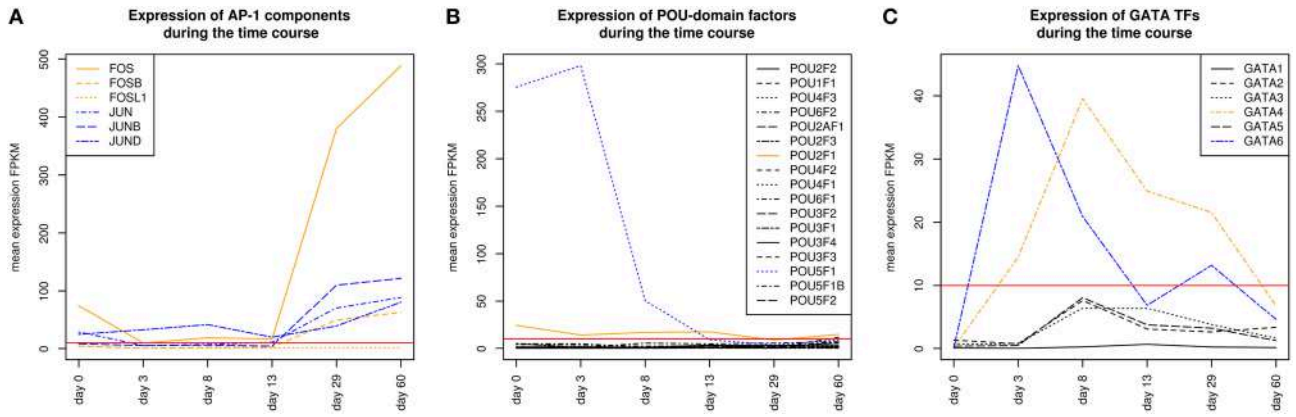


FIGURE 4 | (A) Expression of AP-1 factor genes. In orange FOS TF genes are shown, and in blue JUN TF genes. At early stages the expression levels of FOS and JUN genes which are AP-1 components is rather limited. It is likely that AP-1 cannot be formed due to the low expression of FOS genes. In later stages (> day 13), AP-1 and especially FOS increases its expression. **(B)** Expression of TF genes which contain a POU-domain, in blue (POU5F1) and orange (POU2F1) are the two genes which are above the threshold. POU5F1 which is more abundant than POU2F1 decreases during the time course and was absent after day 13. **(C)** Expression of GATA genes, in blue (GATA6) and orange (GATA4) are above the threshold. GATA6 is expressed during the mesoderm induction stage and decreases afterwards, while GATA4 becomes supreme in subsequent stages. The red lines show a FPKM value of 10 that we consider as threshold for sufficiently expressed genes which contribute to regulatory effects.

TABLE 2 | TFBS pairs within the HMGIY-cluster.

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V$HMGIY_Q6 – V$OCT_Q6	+	+	+	+	+
V$HMGIY_Q6 – V$NFKAPPAB_01	+	+	–	+	+
V$HMGIY_Q6 – V$NFKB_Q6_01	+	+	+	+	+
V$HMGIY_Q6 – V$NFKB_Q6	+	–	–	–	–
V$HMGIY_Q6 – V$ATF3_Q6	+	+	–	–	+

Constitution of co-occurring pairs within the HMGIY-cluster, a "+" indicates the presence of a matrix pair; a "–" its absence.

V$NFKB_Q6 which can be bound by the family of NF-κB-related factors. While the V$HMGIY_Q6 - V$NFKB_Q6 TFBS pair was detected only during the mesoderm induction stage (<day 3), the co-occurrence between V$HMGIY_Q6 and V$NFKB_Q6_01 binding sites was found at all stages. The TFBS pair V$HMGIY_Q6 - V$NFKAPPAB_01 was found at all stages except the late cardiac specification stage (day 8–day 13). To ensure the quality of these three NF-κB binding sites, we further investigated their position weight matrices (PWMs) as well as their binding motifs. Considering the PWMs, we observed that all PWMs have relatively high value of information content (see **Table 3**) which assess their quality. In addition, a comparison between motifs shows different binding behavior of NF-κB-related factors which could be linked to specific members of this family.

HMGA1 is a TF which is represented by the PWM V$HMGIY_Q6 and was recently described as a positive regulator of pluripotency in cellular reprogramming (Shah et al., 2012). The expression levels of HMGA1 in our system are in agreement with previous studies, which describe HMGA1 as highly abundant during embryogenesis, especially in embryonic stem cells; with intermediate expression levels in undifferentiated cancers and at low or at not detectable levels in adult differentiated cells and fibroblasts (Fusco and Fedele, 2007; Hillion et al., 2008, 2009; Resar, 2010; Chou et al., 2011; Schuldenfrei et al., 2011; Shah et al., 2012; Williams et al., 2015). The detected co-occurrence between V$HMGIY_Q6 and V$OCT_Q6 binding sites was found at all stages. The corresponding TF genes (HMGA1, HMGA2, and POU5F1) of this TFBS pair did not show such behavior (see **Figures 4B**, **6A**). HMGA1 as well as POU5F1 are expressed at high levels during early cardiac development with their maximum expression levels at day 3 and declined afterwards. However, this pair was found at later stages indicating that the detected DEGs at these stages could be potentially regulated by this pair. POU5F1 is below the threshold after day 13, whereas HMGA1 is always above the threshold but stablized at low levels. After day 13, HMGA1, which is in its expression values always more abundant than HMGA2, could regulate the detected pairs alone.

The co-occurrence of V$HMGIY_Q6 and different NF-κB binding sites was detected at all time points (see **Table 2**). Interestingly, our findings show that this interaction could occur based on different NF-κB binding sites which are bound by the same TFs. It is known that the interaction between HMGA1 and NF-κB plays a pivotal role in formation of an enhancer complex which is essential to regulate interferon-β signaling on

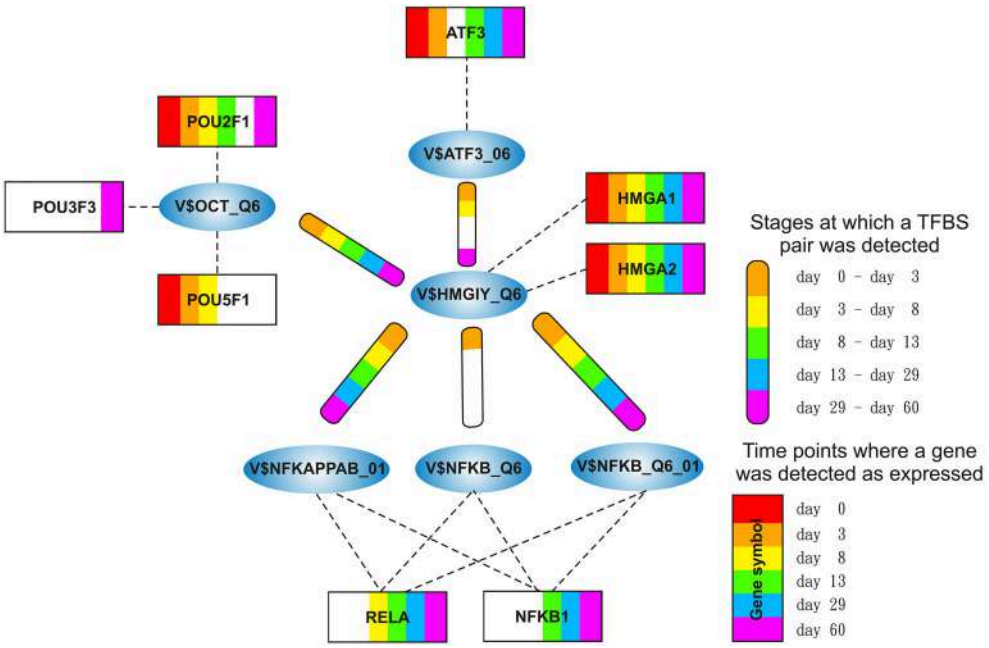


FIGURE 5 | Stage specific representation of TFBSs and the expression of associated TF genes, referring to Figure 2B. The encircled nodes represent the found TFBSs which are connected by color-coded round-edged rectangles which highlight stages where a TFBS pair was found. TF genes which are associated to TFBSs are linked by dashed lines. The TF genes are represented by color-coded rectangles representing the presence at a partiular time point. The absence of a TF gene during a particular time point or the absence of a pair during a particular stage is encoded in white. Both, the color-code for the stage specificity as well as for the gene expression of a TF gene is shown on the bottom right side. TF genes which are associated to a TFBS but are in all time points below the set threshold are omitted.

genomic level (Thanos and Maniatis, 1992; Lewis et al., 1994; Wood et al., 1995; Himes et al., 1996; Thanos and Maniatis, 1996; Mantovani et al., 1998; Perrella et al., 1999; Zhang and Verdine, 1999). Within this complex, NF-κB acts on the one hand as a key regulator in hypertrophy and, on the other hand it acts as cardioprotective factor during embryogenesis (Dewey et al., 2011; Gordon et al., 2011; Liu et al., 2012; Zhou et al., 2013). The expression levels of NF-κB genes may indicate an increasing importance of NFKB1 and especially of RELA during cardiac maturation ($>$ day 13), where it is expressed at considerable levels (see **Figure 6B**).

The co-occurrence of V$HMGIY_Q6 with the V$ATF3_Q6 binding site, which is bound by ATF3, was detected during early cardiac development until day 8 and at the latest stage after day 29. ATF-3 is a FOS-related TF, which contains a basic leucine zipper as structural motif (Chen et al., 1994). ATF-3 acts as homo- or heterodimer to activate or to repress the expression of target genes, depending on its environment. Further, it is also involved in TGF-β signaling in several cell types and in cardiac development (Ishiguro et al., 2000; Mayr and Montminy, 2001; Yan et al., 2005; Gilchrist et al., 2006; Yin et al., 2010; Lin et al., 2014). While HMGA1 is expressed at high levels during early stages (days 0–3) and is declined afterwards, the ATF3 gene is close to the threshold before day 13 and increases its expression levels during subsequent stages (see **Figure 6C**). Our results suggest that the genes regulated by this pair are under control of HMGA1 in the early stages and ATF-3 afterwards. Gilchrist et al. demonstrate the co-occurrence of ATF-3 and NF-κB binding sites in regulated target genes (Gilchrist et al., 2006). According to their binding sites, our analysis suggests that together with ATF-3 and NF-κB factor, HMGA1 may play an important role in the regulation of target genes in cardiac development.

TABLE 3 | Binding sites for different NF-κB PWMs found in the HMGIY-cluster.

PWM	Information content	Motif
V$NFKAPPAB_01	11.8	
V$NFKB_Q6_01$^{(rc)}$	13.3	
V$NFKB_Q6	14.4	

The family of NF-κB-related factors can be represented by different PWMs each of which have relatively high information content and different binding motifs. $^{(rc)}$: reverse complement

3.3. SMAD-Cluster

The SMAD-cluster is assembled in a total of three TFBS pairs with the V$SMAD_Q6_01 binding site in its center (see **Figures 2C**, **7**) . **Table 4** shows the co-occurrence of V$SMAD_Q6_01 and V$FOX_Q2 binding sites in the promoters of the regulated genes and was observed during all stages. The TFBS pair V$SMAD_Q6_01 - V$AP1FJ_Q2 was detected in our system at early stages until day 8 and at late stages after day

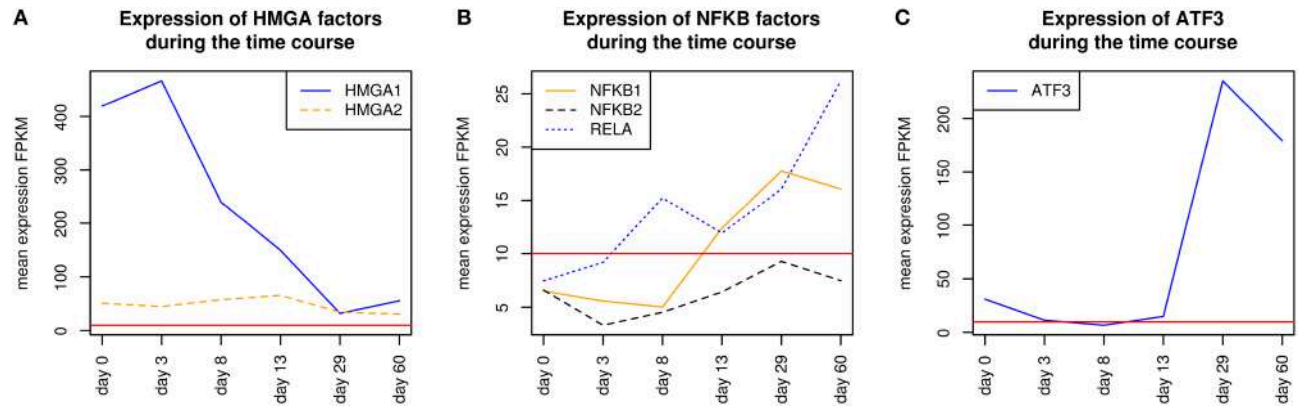


FIGURE 6 | Expression of corresponding TF genes which can be represented by the PWM V$OCT_Q6 have been shown in Figure 4B. (A) Expression of corresponding TF genes which can be represented by the PWM V$HMGIY_Q6. HMGA1 is dominant over HMGA2 but decreases during the time course. **(B)** Expression of corresponding TF genes which can be represented by the PWMs (V$NFKAPPAB_01, V$NFKB_Q6_01, V$NFKB_Q6) related to NF-$\kappa$B binding sites. While NFKB2 never reaches the threshold, RELA and NFKB1 increase their expression levels in later stages (> day 13). **(C)** Expression of the corresponding ATF3 gene to its TF which can be represented by the PWM V$ATF3_Q6. ATF3 is present in the first stage, but in subsequent stages until day 13 it is quite close to the threshold. It changes its expression levels drastically during the cardiac maturation stage (> day 13). The red lines show a FPKM value of 10 that we consider as threshold for sufficiently expressed genes which contribute to regulatory effects.

13, but not during late cardiac specification stage (days 8–13). In contrast, the co-occurrence between V$SMAD_Q6_01 and V$LEF1TCF1_Q4 was detected only during cardiac specification (days 3–13). In addition, **Figure 7** shows for these TFBSs the related TF genes which are expressed in at least one time point.

SMADs are members of a family of transcription factors that form a beta-hairpin structure which interacts with the major groove of the DNA (Burke et al., 1976; Macias et al., 2015). SMAD1-4 which can be represented by the PWM V$SMAD_Q6_01 act as TFs in the nucleus and as signaling molecules, where they are involved in numerous pathways like canonical and non-canonical SMAD-signaling pathways, TGF-β- as well as BMP- and WNT-signaling (Heldin et al., 1997; Leask and Abraham, 2004; Euler-Taimor and Heger, 2006; Pal and Khanna, 2006; Schröder et al., 2006; Leask, 2007; Ruiz-Ortega et al., 2007; Calvieri et al., 2012; Massagué, 2012; Dyer et al., 2014; Euler, 2015). **Figure 8A** shows that SMAD1, SMAD2, and SMAD4 genes are continuously expressed at all stages. The detected SMAD3 expression after day 3 exceeds the set threshold only slightly. SMAD2 and SMAD4 show the highest expression levels in our system, but the differences in their expression levels are rather small.

The co-occurrence of V$SMAD_Q6_01 and V$FOX_Q2 binding sites was detected at all stages (see **Table 4**). Recently, the cooperative regulatory interaction of FOX factors, which play an important role in cardiovascular development and in other organs (Yamagishi et al., 2003; Maeda et al., 2006; Seo and Kume, 2006; Fortin et al., 2015), with SMAD3 and SMAD4 has been shown by (Fortin et al., 2015). Although the SMAD-FOX pair can be detected during the whole time course, the expression of FOX-genes is limited to FOXH1, which seems to play a role in early heart development only (< day 13, see **Figure 8C**).

The co-occurrence between V$SMAD_Q6_01 and V$AP1FJ_Q2 binding sites were found in almost all stages except for the late cardiac specification stage (between day 8 and day 13). In adult CMs, AP-1 together with SMAD proteins modulates hypertrophic, apoptotic and fibrotic pathways. Additionally, AP-1 together with SMAD forces the shift toward apoptosis after stimulation of TGF-β-signaling (Schneiders et al., 2005; Schröder et al., 2006; Euler, 2015). In the embryonic hearts, the activation of TGF-β-pathways results in an induction of cardioprotective functions (Leask and Abraham, 2004; Pal and Khanna, 2006; Leask, 2007; Ruiz-Ortega et al., 2007; Calvieri et al., 2012; Euler, 2015). Although there is no known AP-1 SMAD interaction during cardiogenesis, Yuan et al., shows the interaction of these TFs by usage of AP-1 and SMAD decoy oligodeoxynucleotides, which reduces fibrosis in their study (Yuan et al., 2013).

The detected TFBS pair V$SMAD_Q6_01 - V$LEF1TCF1_Q4 is limited to the cardiac specification stage (day 3–day 13). TCF-7 and LEF-1 transcription factors, which are represented by V$LEF1TCF1_Q4, can be activated by β-catenin and are involved in canonical WNT-signaling (Brade et al., 2006; Chen et al., 2006; Pal and Khanna, 2006; Kwon et al., 2007; Naito et al., 2010). The measured gene expression of TCF as well as LEF genes shows that during cardiac specification both groups are quite close to or below the set threshold (see **Figure 8B**). This indicates that no TCF or LEF binding occurs, which may result in the absence of canonical WNT-signaling during cardiac specification.

3.4. NFAT-Cluster

The NFAT-cluster consists in a total of six TFBS pairs with V$NFAT_Q6 binding site in its center (see **Figures 2D**, **9**). As described in **Table 5** and **Figure 9**, V$NFAT_Q6 co-occurs with V$PEBP6_Q6 and V$ETS1_B binding sites only during the mesoderm induction stage (days 0–3). Three TFBS pairs, namely V$NFAT_Q6 - V$AP1_C, V$NFAT_Q6 - V$CREBP1CJUN_01,

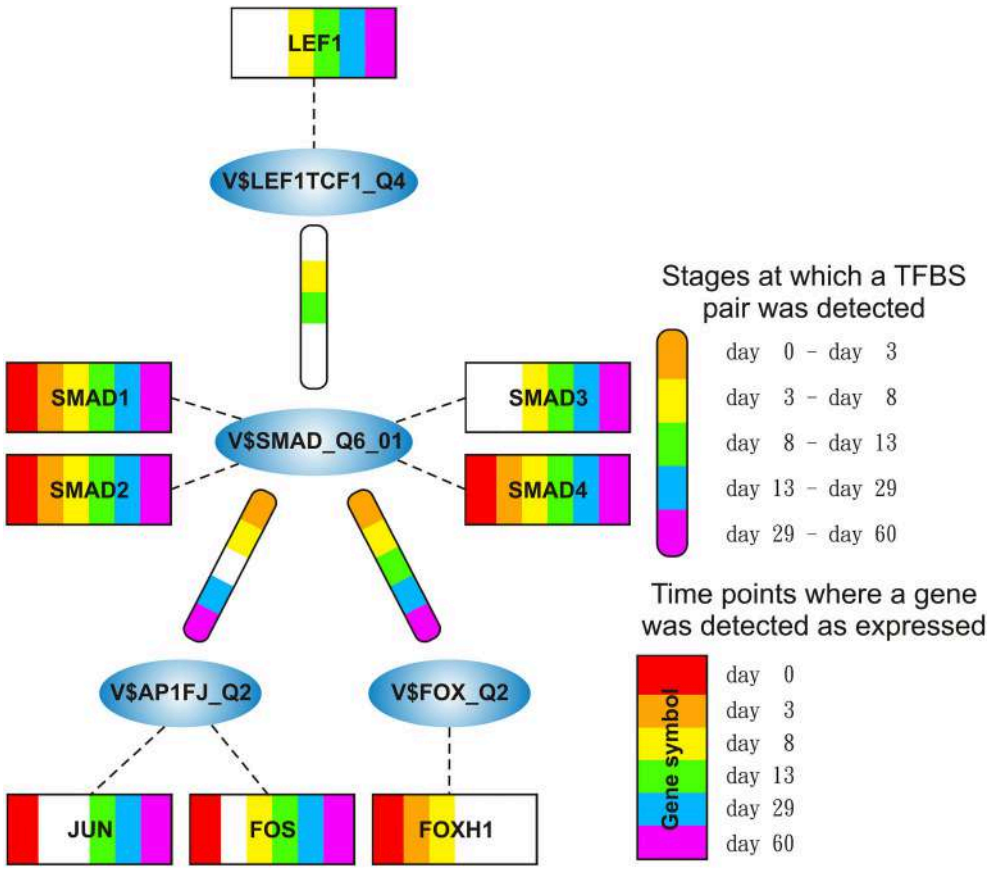


FIGURE 7 | Stage specific representation of TFBSs and the expression of associated TF genes, referring to Figure 2C. The encircled nodes represent the found TFBSs which are connected by color-coded round-edged rectangles which highlight stages where a TFBS pair was found. TF genes which are associated to TFBSs are linked by dashed lines. The TF genes are represented by color-coded rectangles representing the presence at a partiular time point. The absence of a TF gene during a particular time point or the absence of a pair during a particular stage is encoded in white. Both, the color-code for the stage specificity as well as for the gene expression of a TF gene is shown on the bottom right side. TF genes which are associated to a TFBS but are in all time points below the set threshold are omitted.

TABLE 4 | TFBS pairs within the SMAD-cluster.

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V$SMAD_Q6_01– V$FOX_Q2	+	+	+	+	+
V$SMAD_Q6_01 – V$AP1FJ_Q2	+	+	–	+	+
V$SMAD_Q6_01 – V$LEF1TCF1_Q4	–	+	+	–	–

Constitution of co-occurring pairs within the SMAD-cluster, a "+" indicates the presence of a pair; a "–" its absence.

and V$NFAT_Q6 - V$MAF_Q6_01, were found during the complete time course. The co-occurrence of V$NFAT_Q6 with V$CEBPB_01 binding sites in the promoter regions of the analyzed set of genes was found as present until day 8 and during the cardiac maturation stage after day 13. This TFBS pair was not present during the late cardiac specification stage (days 8–13). In addition, **Figure 9** shows for these TFBSs the related TF genes which are expressed in at least one time point.

Regulatory roles for NFAT factors, which can be represented by the PWM V$NFAT_Q6, have been discovered in diverse organs and cells, including the central nervous system, blood vessels, heart, skeletal muscle and haematopoietic stem cells (Macián, 2005). In general, an activation of factors of the NFAT family is calcium dependent and has been described to be of specific importance in development of the atrial myocardium and the morphogenesis of heart valves (Graef et al., 2001; Crabtree and Olson, 2002; Schubert et al., 2003; Schulz and Yutzey, 2004). In our system, only NFATC3 and NFATC4 showed expression levels above the threshold. Comparing the expression levels, NFATC4 is more abundant than NFATC3 at all time points, except for day 3, but both genes increase their expression levels at later stages and especially after day 29 (see **Figure 10A**).

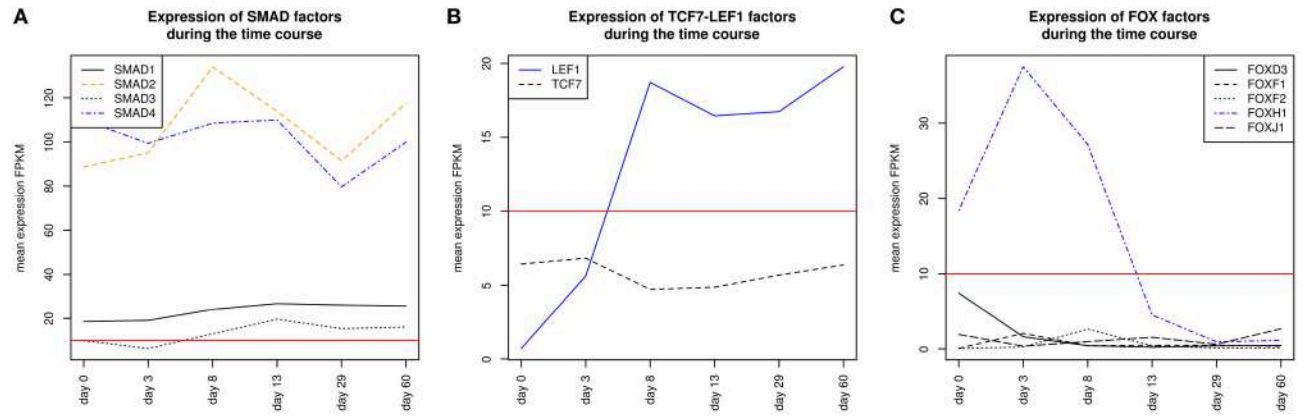


FIGURE 8 | Expression of corresponding TF genes which can be represented by the PWM V$AP1FJ_Q2 have been shown in Figure 4A. (A) Expression of corresponding genes to TFs which can be represented by the PWM V$SMAD_Q6_01. Each SMAD is expressed during the complete time course at similiar levels, while the expression levels of SMAD2/4 are higher than the expression levels of SMAD1/3. After beginning of the cardiac specification (> day 3) SMAD4 is slightly more abundant than SMAD2 and remains in this position. **(B)** Expression of corresponding TF genes which can be represented by the PWM V$LEF1TCF1_Q4, TCF7 is below the threshold set by us as a limit for robust transcription while LEF1 is clearly transcribed after the mesoderm induction stage (> day 3). The SMAD-TCFLEF-pair was found during the cardiac specification stage only (day 3–day 8). **(C)** Expression of corresponding TF genes which can be represented by the PWM V$FOX_Q2. FOXH1 is the only expressed gene and present until day 13. The red lines show a FPKM value of 10 that we consider as threshold for sufficiently expressed genes which contribute to regulatory effects.

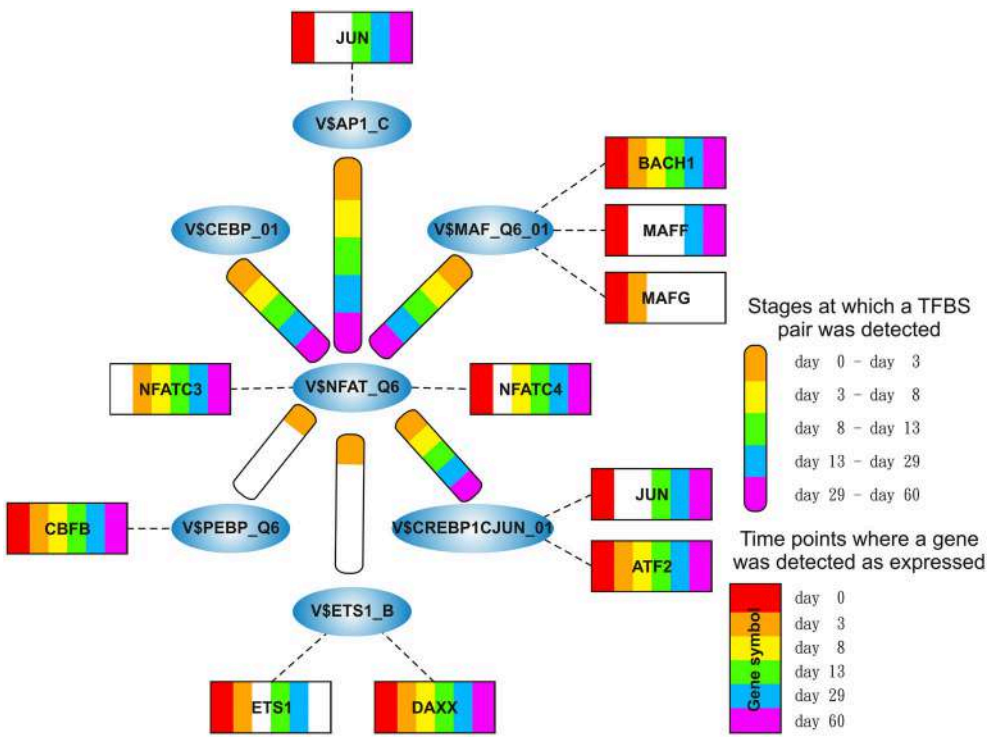


FIGURE 9 | Stage specific representation of TFBSs and the expression of associated TF genes, referring to Figure 2D. The encircled nodes represent the found TFBSs which are connected by color-coded round-edged rectangles which highlight stages where a TFBS pair was found. TF genes which are associated to TFBSs are linked by dashed lines. The TF genes are represented by color-coded rectangles representing the presence at a partiular time point. The absence of a TF gene during a particular time point or the absence of a pair during a particular stage is encoded in white. Both, the color-code for the stage specificity as well as for the gene expression of a TF gene is shown on the bottom right side. TF genes which are associated to a TFBS but are in all time points below the set threshold are omitted.

TABLE 5 | TFBS pairs within the NFAT-cluster.

	Day0–Day3	Day3–Day8	Day8–Day13	Day13–Day29	Day29–Day60
V$NFAT_Q6 – V$PEBP_Q6	+	–	–	–	–
V$NFAT_Q6 – V$AP1_C	+	+	+	+	+
V$NFAT_Q6 – V$CEBPB_01	+	+	–	+	+
V$NFAT_Q6 – V$CREBP1CJUN_01	+	+	+	+	+
V$NFAT_Q6 – V$MAF_Q6_01	+	+	+	+	+
V$NFAT_Q6 – V$ETS1_B	+	–	–	–	–

Constitution of the NFAT-cluster, a "+" indicates the presence of a matrix pair; a "–" its absence.

The detected co-occurrence of TFBS pairs V$NFAT_Q6 - V$AP1_C and V$NFAT_Q6 - V$PEBP_Q6 refers either to NFAT-AP-1 or to NFAT-RUNX interactions which have been mainly observed in the immune system (Macián, 2005). Macián et al. have demonstrated that the interaction between NFAT and AP-1 can be linked to calcineurin dependent pathways as well as to regulation of MAP kinase pathways (Macián et al., 2001). Additionally, NFAT and AP-1 cooperate in naïve T-cells with RUNX TFs as well as with NF-κB in the promoter of IL-2 during T-cell activation (see **Figures 10C,E**) (Hermann-Kleiter and Baier, 2010). In our system, the low or absent expression of RUNX indicates no relevance for these factors. However, the corresponding binding site can be also occupied by CBFB, which is associated to congenital heart anomalies and is expressed during all time points (Khan et al., 2006).

We found the co-occurring TFBS pair V$NFAT_Q6 - V$MAF_Q6_01 at all stages. For the corresponding factors it has been shown by Hogan et al. that NFAT factors and MAF were able to activate IL-4 promoters (Hogan et al., 2003). Of all TFs linked to V$MAF_Q6_01, BACH1 is expressed at all stages and is always more abundant than the other genes shown in **Figure 10B**. This suggests a synergistic interaction in gene regulation between these factors during the complete time course. Furthermore, the interaction between NFAT and MAF factors was observed simultaneously at classical NFAT-AP-1 interaction sites (Hogan et al., 2003).

The co-occurrence between V$NFAT_Q6 and V$CEBPB_01 binding sites has been described in liver cell lines by Yang and Chow (2003). The corresponding factors to this pair seem to interact in a formation of a composite enhancer complex (Yang and Chow, 2003). In our system, genes that are linked to V$CEBPB_01 binding sites are not expressed (see **Figure 10F**). The observation of this pair and its potential role in heart development remains unclear.

The role of the TFBS pair V$NFAT_Q6 - V$ETS1_B, which was detected during the mesoderm induction stage, remains unclear. ETS1, a TF gene which can be linked to the PWM V$ETS1_B, is required for the differentiation of cardiac neural crest (Gao et al., 2010). Although ETS1 was expressed during the mesoderm induction stage (days 0–3), its expression is markedly reduced afterwards. DAXX is another gene that is linked to the PWM V$ETS1_B and is at all time points more abundant than ETS1 (see **Figure 10D**). The DAXX factor inhibits apoptosis in cardiac myocytes (Zobalova et al., 2008). An interaction between NFAT and DAXX was not found in literature, and thus the role of this pair remains unclear.

4. DISCUSSION

Today, it is known that in higher organisms transcription factors have to interact with each other to regulate gene expression which leads to a proper development of tissues and organs. So far, several studies have shown that the co-occurence of TF binding sites (TFBSs) on sequences is an essential indication for the identification of interactions between TFs. In this study, we identified co-occurring TFBS pairs by applying MatrixCatch algorithm to the promoter regions of five differentially expressed gene sets, which are based on a time course dataset of developing human myocardium, modeled in a tissue engineering approach (Hudson et al., in revision). MatrixCatch is a statistically affirmed computational method for the recognition of experimentally verified interactions between TFs according to their TFBS localizations in promoters. However, MatrixCatch recognizes based on its underlying algorithm all detectable TFBS pairs of known interacting TFs in promoter regions. This results in a huge overlap between recognized pairs at different stages, although these pairs can play different roles for each stage. To eliminate this drawback of MatrixCatch to some extent, we created an interaction network based on the TFBS pairs for each stage and then applied the MCL algorithm. MCL differentiates negligible TFBS pairs from densely connected TFBS pairs within these interaction networks and thus determines clusters of TFBSs. Such clusters are important to highlight stage specific co-occurrences of TFBS pairs which provide essential knowledge in the understanding of molecular mechanism of cardiac development.

Additionally, we applied our approach to different lengths of putative promoter regions ([from −500 bp to 0], [from −500 bp to +100 bp], [from −1000 bp to 0]) to determine the influence of promoter lengths on the composition of stage-specific clusters. The results denote that there is a considerably high overlap between stage-specific clusters derived from different putative promoter regions (data not shown). Thus, we considered the -1 kb putative regulatory promoter region for our analysis, which is consistent with our experience and provides the most reliable results.

Although, we filtered MatrixCatch outputs using MCL algorithm to reduce weak co-occurrence of TFBSs in each stage, we detected in our analysis several clusters as well as TFBS pairs whose potential role during cardiac development are unclear. One possible reason for the detection of such pairs could depend on the underlying methodology of MatrixCatch. It uses a computational prediction approach which scans promoter

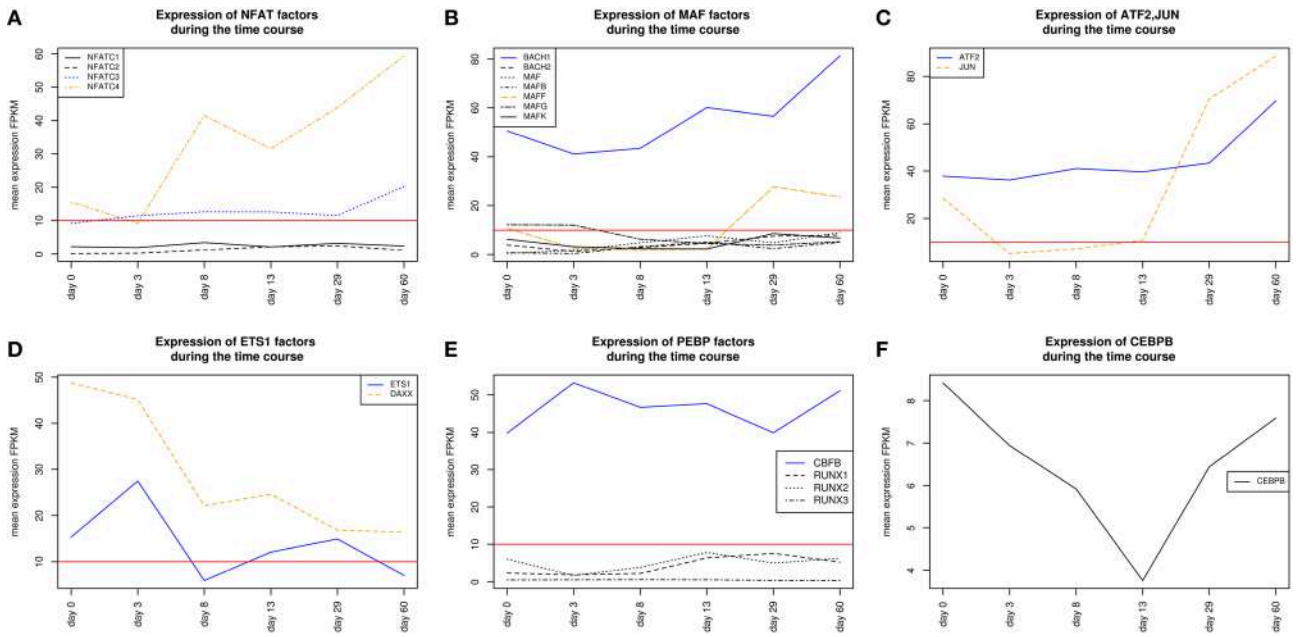


FIGURE 10 | Expression of TF genes corresponding to their factors and represented by the PWM V$AP1_C have been shown in Figure 4A. (A) Expression of TF genes which corresponds to NFAT factors represented by PWM V$NFAT_Q6. NFATC3 and NFATC4 are above the set threshold, whereas NFATC3 is more abundant than NFATC4. **(B)** Expression of TF genes which can be represented by the PWM V$MAF_Q6_01. MAFB shows expression levels slightly above the threshold set by us as a limit for robust transcription during the mesoderm induction stage while MAFF is expressed during cardiac maturation (> day 13). BACH1 is found to be expressed during the complete time course at considerable levels and is always more abundant than all other TF genes, which corresponds to V$MAF_Q6_01. Additionally, BACH1 increases its expression value after the mesoderm induction stage (> day 3). **(C)** Expression of TF genes which can be represented by the PWM V$CREBP1CJUN_01. ATF2 is expressed during the complete time course and increases its expression value in the latest stage. JUN is expressed at day 0 and after day 13 where it exceeds the expression levels of ATF2. **(D)** Expression of TF genes which can be represented by the PWM V$ETS1_B. DAXX is expressed during all time points, but its expression diminishes continuously. Nevertheless, it shows expression levels which are always above ETS1. **(E)** Expression of TF genes which can be represented by the PWM V$PEBP_Q6. Only CBFB shows expression above the threshold and was found as continuously expressed. **(F)** Expression of CEBPB which can be represented by the PWM V$CEBPB_01. CEBPB is during the complete time course below the set threshold and is considered to be low or not expressed. The red lines show a FPKM value of 10 that we consider as threshold for sufficiently expressed genes which contribute to regulatory effects.

sequences and their reverse complements to identify TFBSs using PWMs. However, computational identifications of TFBSs generally suffer from high rates of false positive predictions. Another reason for the detection of those clusters or pairs could be due to genes which are expressed at high levels but play different roles in different tissues. As a result, we could identify such clusters or pairs that might play important roles in the regulation of those genes in other tissues but not in heart. For example, we identified the TFBS pair (V$NFAT_Q6 - V$CEBPB_01) in the NFAT-cluster whose importance has been shown by Yang and Chow in liver (Yang and Chow, 2003), but the potential role of this pair during the cardiac development is unclear. In this context, we also observed the ETS cluster with the V$ETS_Q6 binding site in its center (see **Supplementary File 4**). Only some individual components, like ETS factors, in this cluster are associated with potential cardiac functionalities. However, considering TFBS pairs in the ETS cluster, we cannot verify their potential role during the cardiac development.

Our results suggest that different types of co-occurring TFBS pairs can be assigned into two main categories: (i) TFBS pairs which are present in the beginning and in later stages but absent in at least one of the subsequent stages; (ii) TFBS pairs which are present during all stages. In our clusters presented in the Result section, there are different co-occurring TFBS pairs, like V$AP1_01 - V$OCT_C and V$HMGIY_Q6 - V$ATF3_Q6, which fall into the first category. Considering the expression values of TF genes for those pairs, we observed that one TF gene was highly expressed in the beginning stages while its partner is expressed at low levels. After the re-occurrence of such a pair in later stages, the measured expression values of TF genes are exactly the opposite. Consequently, the related TFs cannot act in a synergistic manner but rather in an antagonistic manner. Very drastically, we observed this situation in the expression of AP-1 components and POU5F1, which can be linked to V$AP1_01 - V$OCT_C TFBS pair (see **Figures 4A,B**). Due to this finding we hypothesize that further TFBS pairs, which fall into the first category, could be helpful to enhance our knowledge on the combinatorial code underlying transcriptional regulation of cardiomyogenesis.

This findings could be discussed in the perspective of the "embryonic hourglass" which describes high divergence in the embryonic shape of vertebrates, insects, like *Drosophila*, and plants, in early and late developmental stages, but minor

divergence in mid-stages (Duboule, 1994; Raff and Wolpert, 1996; Kalinka et al., 2010; Quint et al., 2012). In our study, the number of DEGs as well as the number of identified clusters is high in early stages, converge to a minimum during the late cardiac specification stage (day 8–day 13) and increase afterwards again, which is consistent with the general structure of the hourglass model. Furthermore, the identified TFBS pairs, which fall into the first category, could be separated into two different subsets of genes, the one subset is up-regulated before the late cardiac specification stage, while the other subset is up-regulated afterwards and is supposed to regulated cardiac maturation processes. Our findings support the hourglass model derived by previous findings in *Arabidopsis* as well as several animals (Domazet-Lošo and Tautz, 2010; Kalinka et al., 2010; Quint et al., 2012).

In contrast to the TFBSs pairs in the first category, the co-occurrence of TFBS pairs that fall into the second category seems to indicate a synergistic cooperation between related TFs. In our presented clusters, we obtained several TFBS pairs like V$HMGIY_Q6 - V$OCT_Q6, V$SMAD_Q6_01 - V$FOX_Q2, and V$NFAT_Q6 - V$CREBP1CJUN_01 (for detail see **Tables 2**–**4**). Considering the expression values of corresponding TF genes for those pairs, we determined that these genes are regulated similarly. For instance, the TF genes HMGA1 and POU5F1, which are linked to V$HMGIY_Q6 and V$OCT_Q6, respectively, are highly expressed during first developmental stages and diminish their levels after day 3. This condition is also observed for the TFBS pair V$NFAT_Q6 - V$CREBP1CJUN_01 where the associated TF genes are expressed at low levels in the beginning and increase their expression levels in later stages.

Altogether, in our study we performed a systematic analysis of TFBS pairs to address the question of cooperation between TFs linked to TFBS pairs, which could play a crucial role through five different cardiac developmental stages. Addressing this question, our results show that some TFBS pairs can be detected at all developmental stages. Furthermore, we obtained the same TFBS pairs at very early and very late stages of the differentiation, although these stages are completely different in their functions. Especially considering expression values of related TF genes of these pairs, we determined that co-occurrence between TFBSs does not always indicate a synergistic regulation of target genes. This finding suggests that corresponding TFs of these pairs can be bound in a mutual exclusive manner, which is important during cardiac development to differentiate between stem cell programs and later embryogenic programs.

5. CONCLUSION

We identify transcription factor pairs that drive cardiac development from stem cells to mature cells in a 60 day time course dataset. Our approach is motivated by the importance of potentially interacting transcription factors represented by the co-occurrence of their TFBSs in the regulated stages specific genes and their mediated effects. We identified the relevant pairs employing MatrixCatch method with Markov clustering algorithm together to highlight stage specific clusters of co-occurring TFBS pairs. Furthermore, we analyzed the changes within these clusters to show the specificity of the gene regulation in cardiac development. Our results demonstrate that similar pairs potentially regulate different developmental stages depending on the expression values of the corresponding genes. This may define switches between embryonic and maturation programs and could contribute to a better understanding of embryonic cardiac development.

AUTHOR CONTRIBUTIONS

SZ, CM, and MG participated in the design of the study, conducted computational and statistical analyses. EW supervised the computational and statistical analyses. AR, FR prepared the time course data and the experiments. WZ supervised the experimental design and the experiments. SU prepared and processed the RNAseq data which are used in this study. SZ, CM, RT, and MG were involved in interpretation of the results and the literature survey. MG and SZ wrote the final version of the manuscript. MG conceived of and managed the project. All authors read and approved the final manuscript.

FUNDING

SZ was funded by Mediomics (Fördernummer: 01DJ13026B) of the BMBF (German Ministry of Education and Research) and the DZHK. CM was funded by ExiTox (Fördernummer: 031A269C) of the BMBF (German Ministry of Education and Research). WHZ is supported by the DZHK, the German Research Foundation (DFG ZI 708/10-1, SFB 1002 TP C04/S, and SFB 937 A18), the Foundation Leducq, the German Federal Ministry for Science and Education (BMBF FKZ 13GW0007A [BMBF/CIRM ETIII Award]), and the NIH (U01 HL099997).

ACKNOWLEDGMENTS

We would like to thank Lena Steins and Martin Haubrock for proofreading the manuscript and providing helpful advices and discussions. This work was supported by the DZHK (German Centre for Cardiovascular Research) and by the BMBF (German Ministry of Education and Research). Furthermore, we acknowledge support by the Open Access Publication Funds of the Göttingen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00033

Supplementary File 1 | DEGList.csv includes all detected stage-specific DEGs.

Supplementary File 2 | A heatmap of stage-specific DEGs.

Supplementary File 3 | TFBS pairs found by MatrixCatch for stage-specific DEGs.

Supplementary File 4 | The stage-specific networks after application of MatrixCatch and Markov clustering algorithm.

REFERENCES

Akhurst, R. J. (2012). The paradoxical TGF-β vasculopathies. *Nat. Genet.* 44, 838–839. doi: 10.1038/ng.2366

Bergmann, O., Bhardwaj, R. D., Bernard, S., Zdunek, S., Barnabé-Heider, F., Walsh, S., et al. (2009). Evidence for cardiomyocyte renewal in humans. *Science* 324, 98–102. doi: 10.1126/science.1164680

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956. doi: 10.1016/j.cell.2005.08.020

Brade, T., Männer, J., and Kühl, M. (2006). The role of Wnt signalling in cardiac development and tissue remodelling in the mature heart. *Cardiovasc. Res.* 72, 198–209. doi: 10.1016/j.cardiores.2006.06.025

Brand, T. (2003). Heart development: molecular insights into cardiac specification and early morphogenesis. *Dev. Biol.* 258, 1–19. doi: 10.1016/S0012-1606(03)00112-X

Brewer, A., and Pizzey, J. (2006). GATA factors in vertebrate heart development and disease. *Expert. Rev. Mol. Med.* 8, 1–20. doi: 10.1017/S1462399406000093

Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformat.* 7:488. doi: 10.1186/1471-2105-7-488

Buckingham, M., Meilhac, S., and Zaffran, S. (2005). Building the mammalian heart from two sources of myocardial cells. *Nat. Rev. Genet.* 6, 826–835. doi: 10.1038/nrg1710

Burke, M., Reisler, F., and Harrington, W. F. (1976). Effect of bridging the two essential thiols of myosin on its spectral and actin-binding properties. *Biochemistry* 15, 1923–1927. doi: 10.1021/bi00654a020

Calvieri, C., Rubattu, S., and Volpe, M. (2012). Molecular mechanisms underlying cardiac antihypertrophic and antifibrotic effects of natriuretic peptides. *J. Mol. Med. (Berl).* 90, 5–13. doi: 10.1007/s00109-011-0801-z

Chandra, V., Huang, P., Potluri, N., Wu, D., Kim, Y., and Rastinejad, F. (2013). Multidomain integration in the structure of the HNF-4α nuclear receptor complex. *Nature* 495, 394–398. doi: 10.1038/nature11966

Chaudhry, B., Ramsbottom, S., and Henderson, D. J. (2014). Genetics of cardiovascular development. *Prog. Mol. Biol. Transl. Sci.* 124, 19–41. doi: 10.1016/B978-0-12-386930-2.00002-1

Chen, B. P., Liang, G., Whelan, J., and Hai, T. (1994). ATF3 and ATF3Δ zip. Transcriptional repression versus activation by alternatively spliced isoforms. *J. Biol. Chem.* 269, 15819–15826.

Chen, X., Shevtsov, S. P., Hsich, E., Cui, L., Haq, S., Aronovitz, M., et al. (2006). The β-catenin/T-cell factor/lymphocyte enhancer factor signaling pathway is required for normal and stress-induced cardiac hypertrophy. *Mol. Cell. Biol.* 26, 4462–4473. doi: 10.1128/MCB.02157-05

Chou, B.-K., Mali, P., Huang, X., Ye, Z., Dowey, S. N., Resar, L. M., et al. (2011). Efficient human iPS cell derivation by a non-integrating plasmid from blood cells with unique epigenetic and gene expression signatures. *Cell. Res.* 21, 518–529. doi: 10.1038/cr.2011.12

Crabtree, G. R., and Olson, E. N. (2002). NFAT signaling: choreographing the social lives of cells. *Cell* 109 (Suppl.), S67–S79. doi: 10.1016/s0092-8674(02)00699-2

Dewey, F. E., Perez, M. V., Wheeler, M. T., Watt, C., Spin, J., Langfelder, P., et al. (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circ. Cardiovasc. Genet.* 4, 26–35. doi: 10.1161/CIRCGENETICS.110.941757

Deyneko, I. V., Kel, A. E., Kel-Margoulis, O. V., Deineko, E. V., Wingender, E., and Weiss, S. (2013). MatrixCatch–a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinformat.* 14:241. doi: 10.1186/1471-2105-14-241

Didié, M., Christalla, P., Rubart, M., Muppala, V., Döker, S., Unsöld, B., et al. (2013). Parthenogenetic stem cells for tissue-engineered heart repair. *J. Clin. Invest.* 123, 1285–1298. doi: 10.1172/JCI66854

Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468, 815–818. doi: 10.1038/nature09632

Dongen, S. (2000). *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht, Netherlands.

Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.* 135–142. Available online at: http://dev.biologists.org/content/develop/1994/Supplement/135.full.pdf

Dyer, L. A., Pi, X., and Patterson, C. (2014). The role of BMPs in endothelial cell function and dysfunction. *Trends Endocrinol. Metab.* 25, 472–480. doi: 10.1016/j.tem.2014.05.003

Euler, G. (2015). Good and bad sides of TGFβ-signaling in myocardial infarction. *Front. Physiol.* 6:66. doi: 10.3389/fphys.2015.00066

Euler-Taimor, G., and Heger, J. (2006). The complex pattern of SMAD signaling in the cardiovascular system. *Cardiovasc Res.* 69, 15–25. doi: 10.1016/j.cardiores.2005.07.007

Fortin, J., Ongaro, L., Li, Y., Tran, S., Lamba, P., Wang, Y., et al. (2015). Minireview: Activin signaling in gonadotropes: What does the FOX say to the SMAD? *Mol. Endocrinol.* 29, 963–977. doi: 10.1210/me.2015-1004

Fusco, A., and Fedele, M. (2007). Roles of HMGA proteins in cancer. *Nat. Rev. Cancer* 7, 899–910. doi: 10.1038/nrc2271

Gao, Z., Kim, G. H., Mackinnon, A. C., Flagg, A. E., Bassett, B., Earley, J. U., et al. (2010). Ets1 is required for proper migration and differentiation of the cardiac neural crest. *Development* 137, 1543–1551. doi: 10.1242/dev.047696

Gilchrist, M., Thorsson, V., Li, B., Rust, A. G., Korb, M., Roach, J. C., et al. (2006). Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* 441, 173–178. doi: 10.1038/nature04768

Gordon, J. W., Shaw, J. A., and Kirshenbaum, L. A. (2011). Multiple facets of NF-κB in the heart: to be or not to NF-κB. *Circ. Res.* 108, 1122–1132. doi: 10.1161/CIRCRESAHA.110.226928

Graef, I. A., Chen, F., and Crabtree, G. R. (2001). NFAT signaling in vertebrate development. *Curr. Opin. Genet. Dev.* 11, 505–512. doi: 10.1016/S0959-437X(00)00225-2

Guo, Y., Costa, R., Ramsey, H., Starnes, T., Vance, G., Robertson, K., et al. (2002). The embryonic stem cell transcription factors Oct-4 and FoxD3 interact to regulate endodermal-specific promoter expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3663–3667. doi: 10.1073/pnas.062041099

Heldin, C. H., Miyazono, K., and ten Dijke, P. (1997). TGF-β signalling from cell membrane to nucleus through SMAD proteins. *Nature* 390, 465–471. doi: 10.1038/37284

Hermann-Kleiter, N., and Baier, G. (2010). NFAT pulls the strings during CD4+ T helper cell effector functions. *Blood* 115, 2989–2997. doi: 10.1182/ablood-2009-10-233585

Herzig, T. C., Jobe, S. M., Aoki, H., Molkentin, J. D., Cowley, A. W. Jr, Izumo, S., et al. (1997). Angiotensin II type1a receptor gene expression in the heart: AP-1 and GATA-4 participate in the response to pressure overload. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7543–7548. doi: 10.1073/pnas.94.14.7543

Hess, J., Angel, P., and Schorpp-Kistner, M. (2004). AP-1 subunits: quarrel and harmony among siblings. *J. Cell Sci.* 117(Pt 25), 5965–5973. doi: 10.1242/jcs.01589

Hillion, J., Dhara, S., Sumter, T. F., Mukherjee, M., Di Cello, F., Belton, A., et al. (2008). The high-mobility group A1a/signal transducer and activator of transcription-3 axis: an achilles heel for hematopoietic malignancies? *Cancer Res.* 68, 10121–10127. doi: 10.1158/0008-5472.can-08-2121

Hillion, J., Wood, L. J., Mukherjee, M., Bhattacharya, R., Di Cello, F., Kowalski, J., et al. (2009). Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer. *Mol. Cancer Res.* 7, 1803–1812. doi: 10.1158/1541-7786.MCR-08-0336

Himes, S. R., Coles, L. S., Reeves, R., and Shannon, M. F. (1996). High mobility group protein I(Y) is required for function and for c-Rel binding to CD28 response elements within the GM-CSF and IL-2 promoters. *Immunity* 5, 479–489.

Hogan, P. G., Chen, L., Nardone, J., and Rao, A. (2003). Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes Dev.* 17, 2205–2232. doi: 10.1101/gad.1102703

Hu, Z., and Gallo, S. M. (2010). Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics* 11:49. doi: 10.1186/1471-2164-11-49

Ishiguro, T., Nagawa, H., Naito, M., and Tsuruo, T. (2000). Inhibitory effect of ATF3 antisense oligonucleotide on ectopic growth of HT29 human colon cancer cells. *Jpn. J. Cancer Res.* 91, 833–836. doi: 10.1111/j.1349-7006.2000.tb01021.x

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009

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

Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., et al. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468, 811–814. doi: 10.1038/nature09634

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Suppl. 1), D493–D496. doi: 10.1093/nar/gkh103

Kel-Margoulis, O. V., Kel, A. E., Reuter, I., Deineko, I. V., and Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 30, 332–334. doi: 10.1093/nar/30.1.332

Khan, A., Hyde, R. K., Dutra, A., Mohide, P., and Liu, P. (2006). Core binding factor beta (CBFB) haploinsufficiency due to an interstitial deletion at 16q21q22 resulting in delayed cranial ossification, cleft palate, congenital heart anomalies, and feeding difficulties but favorable outcome. *Am. J. Med. Genet. A* 140, 2349–2354. doi: 10.1002/ajmg.a.31479

Kirby, M. L. (2002). Molecular embryogenesis of the heart. *Pediatr. Dev. Pathol.* 5, 516–543. doi: 10.1007/s10024-002-0004-2

Kwon, C., Arnold, J., Hsiao, E. C., Taketo, M. M., Conklin, B. R., and Srivastava, D. (2007). Canonical Wnt signaling is a positive regulator of mammalian cardiac progenitors. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10894–10899. doi: 10.1073/pnas.0704044104

Leask, A. (2007). TGFβ, cardiac fibroblasts, and the fibrotic response. *Cardiovasc Res.* 74, 207–212. doi: 10.1016/j.cardiores.2006.07.012

Leask, A., and Abraham, D. J. (2004). TGF-β signaling and the fibrotic response. *FASEB J.* 18, 816–827. doi: 10.1096/fj.03-1273rev

Lewis, H., Kaszubska, W., DeLamarter, J. F., and Whelan, J. (1994). Cooperativity between two NF-κB complexes, mediated by high-mobility-group protein I(Y), is essential for cytokine-induced expression of the E-selectin promoter. *Mol. Cell Biol.* 14, 5701–5709. doi: 10.1128/MCB.14.9.5701

Lin, H., Li, H.-F., Chen, H.-H., Lai, P.-F., Juan, S.-H., Chen, J.-J., et al. (2014). Activating transcription factor 3 protects against pressure-overload heart failure via the autophagy molecule Beclin-1 pathway. *Mol. Pharmacol.* 85, 682–691. doi: 10.1124/mol.113.090092

Linnemann, A. K., O'Geen, H., Keles, S., Farnham, P. J., and Bresnick, E. H. (2011). Genetic framework for GATA factor function in vascular biology. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13641–13646. doi: 10.1073/pnas.1108440108

Liu, Q., Chen, Y., Auger-Messier, M., and Molkentin, J. D. (2012). Interaction between NFκB and NFAT coordinates cardiac hypertrophy and pathological remodeling. *Circ. Res.* 110, 1077–1086. doi: 10.1161/CIRCRESAHA.111.260729

Macián, F. (2005). NFAT proteins: key regulators of T-cell development and function. *Nat. Rev. Immunol.* 5, 472–484. doi: 10.1038/nri1632

Macián, F., López-Rodríguez, C., and Rao, A. (2001). Partners in transcription: NFAT and AP-1. *Oncogene* 20, 2476–2489. doi: 10.1038/sj.onc.1204386

Macias, M. J., Martin-Malpartida, P., and Massagué, J. (2015). Structural determinants of Smad function in TGF-β signaling. *Trends Biochem. Sci.* 40, 296–308. doi: 10.1016/j.tibs.2015.03.012

Maeda, J., Yamagishi, H., McAnally, J., Yamagishi, C., and Srivastava, D. (2006). Tbx1 is regulated by forkhead proteins in the secondary heart field. *Dev. Dyn.* 235, 701–710. doi: 10.1002/dvdy.20686

Mantovani, F., Covaceuszach, S., Rustighi, A., Sgarra, R., Heath, C., Goodwin, G. H., et al. (1998). NF-κB mediated transcriptional activation is enhanced by the architectural factor HMGI-C. *Nucleic Acids Res.* 26, 1433–1439. doi: 10.1093/nar/26.6.1433

Martin, J., Afouda, B. A., and Hoppler, S. (2010). Wnt/β-catenin signalling regulates cardiomyogenesis via GATA transcription factors. *J. Anat.* 216, 92–107. doi: 10.1111/j.1469-7580.2009.01171.x

Martin, L. J., Bergeron, F., Viger, R. S., and Tremblay, J. J. (2012). Functional cooperation between GATA factors and cJUN on the star promoter in MA-10 leydig cells. *J. Androl.* 33, 81–87. doi: 10.2164/jandrol.110.012039

Massagué, J. (2012). TGFβ signalling in context. *Nat. Rev. Mol. Cell Biol.* 13, 616–630. doi: 10.1038/nrm3434

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nat. Rev. Mol. Cell. Biol.* 2, 599–609. doi: 10.1038/35085068

Meckbach, C., Tacke, R., Hua, X., Waack, S., Wingender, E., and Gültas, M. (2015). PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformat.* 16:400. doi: 10.1186/s12859-015-0827-2

Miura, G. I., and Yelon, D. (2013). Cardiovascular biology: play it again, Gata4. *Curr. Biol.* 23, R619–R621. doi: 10.1016/j.cub.2013.06.006

Naito, A. T., Shiojima, I., and Komuro, I. (2010). Wnt signaling and aging-related heart disorders. *Circ. Res.* 107, 1295–1303. doi: 10.1161/CIRCRESAHA.110.223776

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., et al. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95, 379–391. doi: 10.1016/S0092-8674(00)81769-9

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Nekludova, L., Rolfe, P. A., Danford, T. W., et al. (2006). Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.* 2:2006.0017. doi: 10.1038/msb4100059

Ohneda, K., and Yamamoto, M. (2002). Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta. Haematol.* 108, 237–245. doi: 10.1159/000065660

Orkin, S. H. (1992). GATA-binding transcription factors in hematopoietic cells. *Blood* 80, 575–581.

Pal, R., and Khanna, A. (2006). Role of Smad- and Wnt-dependent pathways in embryonic cardiac development. *Stem Cells Dev.* 15, 29–39. doi: 10.1089/scd.2006.15.29

Perrella, M. A., Pellacani, A., Wiesel, P., Chin, M. T., Foster, L. C., Ibanez, M., et al. (1999). High mobility group-I(Y) protein facilitates nuclear factor-κB binding and transactivation of the inducible nitric-oxide synthase promoter/enhancer. *J. Biol. Chem.* 274, 9045–9052.

Pesce, M., and Schöler, H. R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells* 19, 271–278. doi: 10.1634/stemcells.19-4-271

Peterkin, T., Gibson, A., Loose, M., and Patient, R. (2005). The roles of GATA-4, -5 and -6 in vertebrate heart development. *Semin Cell Dev. Biol.* 16, 83–94. doi: 10.1016/j.semcdb.2004.10.003

Pikkarainen, S., Tokola, H., Kerkelä, R., and Ruskoaho, H. (2004). GATA transcription factors in the developing and adult heart. *Cardiovasc Res.* 63, 196–207. doi: 10.1016/j.cardiores.2004.03.025

Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature* 490, 98–101. doi: 10.1038/nature11394

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raff, R. A., and Wolpert, L. (1996). The shape of life-genes, development and evolution of animal forms. *Genet. Res.* 68:261.

Resar, L. M. S. (2010). The high mobility group A1 gene: transforming inflammatory signals into cancer? *Cancer Res.* 70, 436–439. doi: 10.1158/0008-5472.can-09-1212

Ruiz-Ortega, M., Rodríguez-Vita, J., Sanchez-Lopez, E., Carvajal, G., and Egido, J. (2007). TGF-β signaling in vascular fibrosis. *Cardiovasc. Res.* 74, 196–206. doi: 10.1016/j.cardiores.2007.02.008

Ryan, K., and Chin, A. J. (2003). T-box genes and cardiac development. *Birth Defects Res. C Embryo Today* 69, 25–37. doi: 10.1002/bdrc.10001

Schleich, J.-M., Abdulla, T., Summers, R., and Houyel, L. (2013). An overview of cardiac morphogenesis. *Arch. Cardiovasc. Dis.* 106, 612–623. doi: 10.1016/j.acvd.2013.07.001

Schneiders, D., Heger, J., Best, P., Piper, H. M., and Taimor, G. (2005). SMAD proteins are involved in apoptosis induction in ventricular cardiomyocytes. *Cardiovasc. Res.* 67, 87–96. doi: 10.1016/j.cardiores.2005.02.021

Schöler, H. R., Ruppert, S., Suzuki, N., Chowdhury, K., and Gruss, P. (1990). New type of POU domain in germ line-specific protein Oct-4. *Nature* 344, 435–439.

Schröder, D., Heger, J., Piper, H. M., and Euler, G. (2006). Angiotensin II stimulates apoptosis via TGF-β1 signaling in ventricular cardiomyocytes of rat. *J. Mol. Med. (Berl).* 84, 975–983. doi: 10.1007/s00109-006-0090-0

Schubert, W., Yang, X. Y., Yang, T. T. C., Factor, S. M., Lisanti, M. P., Molkentin, J. D., et al. (2003). Requirement of transcription factor NFAT in developing atrial myocardium. *J. Cell. Biol.* 161, 861–874. doi: 10.1083/jcb.200301058

Schuldenfrei, A., Belton, A., Kowalski, J., Talbot, C. C. Jr, Di Cello, F., Poh, W., et al. (2011). HMGA1 drives stem cell, inflammatory pathway, and cell cycle progression genes during lymphoid tumorigenesis. *BMC Genomics* 12:549. doi: 10.1186/1471-2164-12-549

Schulz, R. A., and Yutzey, K. E. (2004). Calcineurin signaling and NFAT activation in cardiovascular and skeletal muscle development. *Dev. Biol.* 266, 1–16. doi: 10.1016/j.ydbio.2003.10.008

Seo, S., and Kume, T. (2006). Forkhead transcription factors, Foxc1 and Foxc2, are required for the morphogenesis of the cardiac outflow tract. *Dev. Biol.* 296, 421–436. doi: 10.1016/j.ydbio.2006.06.012

Shah, S. N., Kerr, C., Cope, L., Zambidis, E., Liu, C., Hillion, J., et al. (2012). HMGA1 reprograms somatic cells into pluripotent stem cells by inducing stem cell transcriptional networks. *PLoS ONE* 7:e48533. doi: 10.1371/journal.pone.0048533

Shaulian, E. (2010). AP-1–The Jun proteins: Oncogenes or tumor suppressors in disguise? *Cell Signal.* 22, 894–899. doi: 10.1016/j.cellsig.2009.12.008

Shaulian, E., and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nat. Cell. Biol.* 4, E131–E136. doi: 10.1038/ncb0502-e131

Shi, G., and Jin, Y. (2010). Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res. Ther.* 1:39. doi: 10.1186/scrt39

Shih, Y.-K., and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics* 28, i473–i479. doi: 10.1093/bioinformatics/bts370

Smyth, G. K. (2004), Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25. doi: 10.2202/1544-6115.1027

Soong, P. L., Tiburcy, M., and Zimmermann, W.-H. (2012). Cardiac differentiation of human embryonic stem cells and their assembly into engineered heart muscle. *Curr. Protoc. Cell Biol.* Chapter 23:Unit23.8. doi: 10.1002/0471143030.cb2308s55

Suzuki, Y. J., Ikeda, T., Shi, S. S., Kitta, K., Kobayashi, Y. M., Morad, M., et al. (1999). Regulation of GATA-4 and AP-1 in transgenic mice overexpressing cardiac calsequestrin. *Cell Calcium* 25, 401–407.

Sylva, M., van den Hoff, M. J. B., and Moorman, A. F. M. (2014). Development of the human heart. *Am. J. Med. Genet. A* 164A, 1347–1371. doi: 10.1002/ajmg.a.35896

Takeuchi, T. (2014). Regulation of cardiomyocyte proliferation during development and regeneration. *Dev. Growth. Differ.* 56, 402–409. doi: 10.1111/dgd.12134

Thanos, D., and Maniatis, T. (1992). The high mobility group protein HMG I(Y) is required for NF-κB-dependent virus induction of the human IFN-β gene. *Cell* 71, 777–789.

Thanos, D., and Maniatis, T. (1996). In vitro assembly of enhancer complexes. *Methods Enzymol* 274, 162–173.

Tiburcy, M., and Zimmermann, W.-H. (2014). Modeling myocardial growth and hypertrophy in engineered heart muscle. *Trends Cardiovasc. Med.* 24, 7–13. doi: 10.1016/j.tcm.2013.05.003

Turbendian, H. K., Gordillo, M., Tsai, S.-Y., Lu, J., Kang, G., Liu, T.-C., et al. (2013). GATA factors efficiently direct cardiac fate from embryonic stem cells. *Development* 140, 1639–1644. doi: 10.1242/dev.093260

Vlasblom, J., and Wodak, S. J. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10:99. doi: 10.1186/1471-2105-10-99

Wang, X., and Jauch, R. (2014). OCT4: A penetrant pluripotency inducer. *Cell Regen (Lond).* 3:6. doi: 10.1186/2045-9769-3-6

Watt, A. J., Garrison, W. D., and Duncan, S. A. (2003). HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* 37, 1249–1253. doi: 10.1053/jhep.2003.50273

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13:R50. doi: 10.1186/gb-2012-13-9-r50

Williams, M. D., Zhang, X., Belton, A. S., Xian, L., Huso, T., Park, J.-J., et al. (2015). HMGA1 drives metabolic reprogramming of intestinal epithelium during hyperproliferation, polyposis, and colorectal carcinogenesis. *J. Proteome Res.* 14, 1420–1431. doi: 10.1021/pr501084s

Wong, K.-C., Li, Y., and Peng, C. (2016). Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics* 32, 321–324. doi: 10.1093/bioinformatics/btv555

Wood, L. D., Farmer, A. A., and Richmond, A. (1995). HMGI(Y) and Sp1 in addition to NF-κB regulate transcription of the MGSA/GROα gene. *Nucleic Acids Res.* 23, 4210–4219. doi: 10.1093/nar/23.20.4210

Yamagishi, H., Maeda, J., Hu, T., McAnally, J., Conway, S. J., Kume, T., et al. (2003). Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer. *Genes Dev.* 17, 269–281. doi: 10.1101/gad.1048903

Yan, C., Lu, D., Hai, T., and Boyd, D. D. (2005). Activating transcription factor 3, a stress sensor, activates p53 by blocking its ubiquitination. *EMBO J.* 24, 2425–2435. doi: 10.1038/sj.emboj.7600712

Yang, T. T. C., and Chow, C.-W. (2003). Transcription cooperation by NFAT.C/EBP composite enhancer complex. *J. Biol. Chem.* 278, 15874–15885. doi: 10.1074/jbc.M211560200

Ye, L., Zimmermann, W.-H., Garry, D. J., and Zhang, J. (2013). Patching the heart: cardiac repair from within and outside. *Circ. Res.* 113, 922–932. doi: 10.1161/CIRCRESAHA.113.300216

Yin, X., Wolford, C. C., Chang, Y.-S., McConoughey, S. J., Ramsey, S. A., Aderem, A., et al. (2010). ATF3, an adaptive-response gene, enhances TGFβ signaling and cancer-initiating cell features in breast cancer cells. *J. Cell Sci.* 123(Pt 20), 3558–3565. doi: 10.1242/jcs.064915

Yuan, H.-F., Huang, H., Li, X.-Y., Guo, W., Xing, W., Sun, Z.-Y., et al. (2013). A dual AP-1 and SMAD decoy ODN suppresses tissue fibrosis and scarring in mice. *J. Invest. Dermatol.* 133, 1080–1087. doi: 10.1038/jid.2012.443

Zhang, X. M., and Verdine, G. L. (1999). A small region in HMG I(Y) is critical for cooperation with NF-κB on DNA. *J. Biol. Chem.* 274, 20235–20243. doi: 10.1074/jbc.274.29.20235

Zhou, H., Yang, H.-X., Yuan, Y., Deng, W., Zhang, J.-Y., Bian, Z.-Y., et al. (2013). Paeoniflorin attenuates pressure overload-induced cardiac remodeling via inhibition of TGFβ/Smads and NF-κB pathways. *J. Mol. Histol.* 44, 357–367. doi: 10.1007/s10735-013-9491-x

Zimmermann, W.-H., Melnychenko, I., Wasmeier, G., Didié, M., Naito, H., Nixdorff, U., et al. (2006). Engineered heart tissue grafts improve systolic and diastolic function in infarcted rat hearts. *Nat. Med.* 12, 452–458. doi: 10.1038/nm1394

Zobalova, R., Swettenham, E., Chladova, J., Dong, L.-F., and Neuzil, J. (2008). Daxx inhibits stress-induced apoptosis in cardiac myocytes. *Redox. Rep.* 13, 263–270. doi: 10.1179/135100008X308975

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 April 2016
doi: 10.3389/fgene.2016.00042

Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines

Darius Wlochowitz[1], Martin Haubrock[1], Jetcy Arackal[2], Annalen Bleckmann[2], Alexander Wolff[3], Tim Beißbarth[3], Edgar Wingender[1] and Mehmet Gültas[1]**

[1] Institute of Bioinformatics, University Medical Center Göttingen, Göttingen, Germany, [2] Department of Hematology/Medical Oncology, University Medical Center Göttingen, Göttingen, Germany, [3] Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

OPEN ACCESS

Edited by:
Artemis Georgia Hatzigeorgiou, Biomedical Sciences Research Center Alexander Fleming, Greece

Reviewed by:
Nestoras Karathanasis, Diana Lab, Greece
Spyros Tastsoglou, University of Thessaly, Greece

***Correspondence:**
Darius Wlochowitz
darius.wlochowitz@bioinf.med.uni-goettingen.de;
Mehmet Gültas
mehmet.gueltas@bioinf.med.uni-goettingen.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 October 2015*
Accepted: *14 March 2016*
Published: *05 April 2016*

Citation:
Wlochowitz D, Haubrock M, Arackal J, Bleckmann A, Wolff A, Beißbarth T, Wingender E and Gültas M (2016) Computational Identification of Key Regulators in Two Different Colorectal Cancer Cell Lines. Front. Genet. 7:42. doi: 10.3389/fgene.2016.00042

Transcription factors (TFs) are gene regulatory proteins that are essential for an effective regulation of the transcriptional machinery. Today, it is known that their expression plays an important role in several types of cancer. Computational identification of key players in specific cancer cell lines is still an open challenge in cancer research. In this study, we present a systematic approach which combines colorectal cancer (CRC) cell lines, namely 1638N-T1 and CMT-93, and well-established computational methods in order to compare these cell lines on the level of transcriptional regulation as well as on a pathway level, i.e., the cancer cell-intrinsic pathway repertoire. For this purpose, we firstly applied the Trinity platform to detect signature genes, and then applied analyses of the geneXplain platform to these for detection of upstream transcriptional regulators and their regulatory networks. We created a CRC-specific position weight matrix (PWM) library based on the TRANSFAC database (release 2014.1) to minimize the rate of false predictions in the promoter analyses. Using our proposed workflow, we specifically focused on revealing the similarities and differences in transcriptional regulation between the two CRC cell lines, and report a number of well-known, cancer-associated TFs with significantly enriched binding sites in the promoter regions of the signature genes. We show that, although the signature genes of both cell lines show no overlap, they may still be regulated by common TFs in CRC. Based on our findings, we suggest that canonical Wnt signaling is activated in 1638N-T1, but inhibited in CMT-93 through cross-talks of Wnt signaling with the VDR signaling pathway and/or LXR-related pathways. Furthermore, our findings provide indication of several master regulators being present such as MLK3 and Mapk1 (ERK2) which might be important in cell proliferation, migration, and invasion of 1638N-T1 and CMT-93, respectively. Taken together, we provide new insights into the invasive potential of these cell lines, which can be used for development of effective cancer therapy.

Keywords: colorectal cancer, promoter analysis, pathway analysis, master regulator analysis, Wnt pathway

1. INTRODUCTION

Cancer undergoes genetic and epigenetic changes through which it acquires cellular and molecular characteristics during invasive tumor growth. These changes allow the tumor cells to evade the immune response, activate the microenvironment, invade surrounding tissues and metastasize to distant sites. The microenvironment plays an important role in this context as it may trigger anti-tumor as well as pro-tumor signals (Gao et al., 2014). Malignant tumor cells stimulate the production and secretion of growth factors, cytokines and enzymes, thereby recruiting the stroma and vasculature, which altogether results in the conversion of a normal tumor-inhibiting into a tumor-promoting microenvironment (Gao et al., 2014). In that respect, tumor aggressiveness can be linked to processes such as cell proliferation, growth, invasion, metastasis, survival as well as inflammation which are regulated by multiple signal transduction pathways. It has been suggested to summarize known signal transduction reactions into about 17 signal transduction pathways (Nebert, 2002). They are usually activated by growth factor signals from the cell surface, and further transmit the signal via transmembrane receptors to their target intracellular effectors. In tumor cells, these pathways are often dysregulated and harbor alterations in key components that can function as driver mutations, i.e., either as activation mutations (Ras, PI3K, Akt) or loss of tumor-suppressor gene function (Pten). Several cancer drivers are important integral parts of these pathways, such as receptor tyrosine kinases, and can be located upstream in signal transduction cascades. Since protein kinases propagate the signals along the cascade, they are considered attractive drug targets for therapeutic intervention using specific protein kinase inhibitors (Zwick et al., 2001; Torkamani et al., 2009; Takeuchi and Ito, 2011; Casaletto and McClatchey, 2012). To this end, many anticancer agents have been used in the context of cancer therapy to account for the number of different pathways (Casaletto and McClatchey, 2012).

The signaling pathways are interconnected and form an elaborate network of pathways that receives signals from a variety of growth factors to tightly regulate processes such as transcription, cell growth, motility, differentiation, apoptosis, and cytoskeletal organization. In addition, the outcome triggered by the integrated signaling may differ between different cell types. Therefore, knowledge on the cell type-specific pathways including their architecture and complexity provides important information on the tumor cell behavior during inhibitor therapy, i.e., the inhibitor may not achieve the desired outcome due to the utilization of alternative bypass pathways in certain tumor cells.

Signal transduction pathways converge on sets of genes with similar key functions which are regulated by upstream transcription factors (TFs). TFs occupy short and specific DNA-sequences denoted as transcription factor binding sites (TFBSs). TFs and their corresponding TFBSs recruit and regulate the transcription machinery, thereby governing selective temporal and spatial activities of their target genes. Moreover, many TFs play important roles as oncogenes and they are usually activated downstream in the signaling cascades. Consequently, their deregulated expression, aberrant activation as well as mutations contribute to tumorigenesis. For example, the TP53 gene which encodes an important transcription factor with tumor suppressor function in cancer, is known to be the most commonly mutated gene in human cancer (Kandoth et al., 2013). Unsurprisingly, TFs are central to cancer and became highly desirable points of interference in cancer gene therapy (Libermann and Zerbini, 2006). In this regard, three major transcription factor families have been considered highly desirable drug targets: (i) the NF-κB and AP-1 families of TFs; (ii) the STAT family members; (iii) the steroid receptors (Libermann and Zerbini, 2006). Although other additional TF families have been implicated in cancer to this day, there is still no comprehensive library on TFs and their specific roles in cancer and, particularly, in different cancer cell types. However, given the tumor heterogeneity and cancer cell plasticity, it can be expected that many more TFs will be associated with potentially important roles in oncogenic pathways of different cancers.

The third most common cancer in the world is colorectal cancer (CRC) which originates in the epithelial cells of the gastrointestinal track and shows a high tendency to metastasize into the liver. CRC is often caused by mutations in two well-studied signal transduction pathways, namely the Wnt and the EGFR pathways (Normanno et al., 2006; Polakis, 2012). Mouse models have been extensively used in cancer studies to directly monitor the metastatic progression in CRC. The ability to study primary tumors as well as distant metastatic sites and to manipulate the spatial and temporal expression levels of certain single genes have proven the animal model technology to be a powerful tool in cancer progression research. Such studies have often made use of APC-deficient mouse models since mutations in the adenomatous polyposis coli (APC), an important component of the Wnt signaling pathway, occur in the majority of human CRC cells (Karim and Huso, 2013). It is estimated that the canonical Wnt/β-catenin signaling pathway is abnormally activated in over 90% of CRCs (Cancer Genome Atlas Network, 2012). Briefly, the canonical Wnt pathway revolves around the intracellular levels of the transcriptional coactivator β-catenin which forms a complex with TCF/LEF, thereby controlling the expression of Wnt signaling targets, such as c-Myc and cyclin D. β-Catenin is degraded by a destruction complex that includes the tumor suppressor APC and other proteins (Stamos and Weis, 2013). Loss of APC leads to a constant activation of WNT signaling, which promotes proliferation of tumor cells.

The bottleneck in cancer research has always been a lack of effective tools to comprehensively study the complex networks of signaling pathways (Kang, 2005; Gupta and Massagué, 2006). Therefore, cancer research has largely taken advantage of the integration of animal models and bioinformatic approaches. Microarrays and nowadays RNA-sequencing techniques (RNA-Seq) are used to infer reliable gene regulatory networks based on the level of all expressed transcripts (transcriptome) (Schena et al., 1995; Mortazavi et al., 2008). The result of a transcriptome profiling experiment can be summarized in a set of expressed genes or transcription units that are meaningful for a certain experimental condition, disease state or developmental process.

These technologies have led to paradigm-shifting advances in cancer research. For example, gene expression profiles in combination with supervised clustering approaches were used in breast cancer studies which successfully discriminated between cancer patients with good prognosis from those with poor prognosis, thereby leading to the identification of prognostic cancer genes (van 't Veer et al., 2002; Weigelt et al., 2005). However, solely using genomic profiling of tumor samples only identifies individual genes of a set of signature genes, but does not provide a functional context for these genes, which is important for a mechanistic understanding of cancer-associated processes. Pathway analyses have therefore emerged as powerful tools by benefiting from the statistical power of entire gene sets using the overrepresentation in biologically defined pathways rather than interpreting meaningful functions based on the expression of individual genes.

Despite the presence of a variety of different approaches and rich literature on cancer research as mentioned above, to date, there is still need for comprehensive analyses to detect key regulators in different colorectal cancer cell lines. In this study, we made use of distinct murine cancer cell lines and system biology approaches to identify signature genes and pathways whose activation may specifically affect invasive tumor growth. In addition, we exhaustively covered a broad range of potentially important signaling pathways and focus our discussion selectively on the study of the roles of various classical and novel signaling pathways in CRC. Moreover, we aimed to highlight the meaning of specific TFs in the context of these pathways on the basis of enriched TFBSs in the promoter regions of the signature genes. We provide a comprehensive library on CRC-specific TFs and exemplarily discuss their roles in both CRC cell lines. Taken together, we identified potential discriminators between the two CRC cell lines as well as points of interference for targeted cancer therapy, thus providing further insights into the complexity of cancer.

2. MATERIALS AND METHODS

2.1. Colorectal Cancer Cell Lines

The CMT-93 cell line, a mouse colorectal polyploid carcinoma cell line, was purchased from the American Type Culture Collection, Manassas, USA (CCL223) and was cultured in DMEM High Glucose Medium (Gibco, Darmstadt, Germany) supplemented with 10% heat inactivated fetal bovine serum (FCS; Sigma, Munich, Germany). The murine colorectal cancer cell line 1638N-T1, derived from Apc1638N adenomas, was kindly provided by Ron Smits (Smits et al., 1997). Remarkably, this cell line harbors a targeted mutation at codon 1638 of the Apc gene, Apc1638T, leading to a truncated Apc protein (Smits et al., 1999). These were cultured in DMEM High Glucose Medium supplemented with 15% not heat inactivated FCS and Insulin/Transferrin/Selenium Solution (Gibco). In contrast to Smits et al., these cells were not cultured on any fibronectin/collagen/albumin-coated plates and were passaged using 0.05% (w/v) trypsin (Biochrom, Berlin, Germany), as long as they did not show any differences in their morphology, viability and proliferation.

2.2. RNA Isolation and Sequencing

Total RNA was isolated using the TRIzol Reagent (Invitrogen, Karlsruhe, Germany) including a DNase I (Roche, Mannheim, Germany) digestion. RNA integrity and quantity was assessed with the Agilent Bioanalyzer 2100 and the NanoDrop DD-1000 UV vis spectrophotometer version 3.2.1. 2 μg of total RNA were used as start material for library preparation (TruSeq Stranded mRNA Sample Prep Kit from Illumina, Cat NRS-122-2101). Accurate quantitation of cDNA libraries was performed by using the QuantiFluor dsDNA System (Promega). The size range of cDNA libraries was determined applying the DNA 1000 chip on the Bioanalyzer 2100 from Agilent (280 bp). cDNA libraries were amplified and sequenced by using the cBot and HiSeq 2000 from Illumina (SR, 1 $\times$ 51 bp, 8–9 Gb $>$ 40 M reads per sample). Sequence images were transformed with Illumina software BaseCaller to bcl-files, which were demultiplexed to FASTQ files with CASAVA (version 1.8.2). Quality check was done via FastQC (version 0.10.1, Babraham Bioinformatics).

2.3. Signature Gene Selection

We started our analyses based on 43433 gene annotations from Ensembl (mouse assembly GRCm38.p4), which were retrieved from RNA-seq samples (Section 2.2; three biological replicates for each cell line; GSE78696). Based on these samples, we obtained signature genes as follows:

Using the Trinity platform (Grabherr et al., 2011), we firstly performed a differentially expressed gene (DEG) analysis based on both cell lines. After that, employing the Trinity platform these DEGs were clustered into three main categories using a *p*-value cutoff for FDR of 0.05 and the default fold change (default: 2 (meaning 2^2 or 4-fold)): (i) genes which are most significantly upregulated in 1638N-T1 (**Supplementary Table S1**) and, at the same time, downregulated in CMT-93; (ii) genes which are most significantly upregulated in CMT-93 (**Supplementary Table S2**) and, at the same time, downregulated in 1638N-T1; (iii) the remaining DEGs which did not fall in the first and second category. In our further analysis, we only considered genes as *signature genes* which fell into the first or second category.

2.4. Data Processing

For the subsequent analyses we used the geneXplain platform (http://genexplain-platform.com/bioumlweb/), which includes the TRANSFAC and TRANSPATH databases. We used the suggested parameters from this platform if not explicitly stated otherwise.

2.4.1. Enrichment of TFBSs in Promoter Sequences

We applied a conventional enrichment analysis to the previously identified signature gene sets in order to retrieve specific TFs whose binding sites or sequence motifs are particularly enriched in their genomic regions. For the enrichment analysis, we firstly extracted for each signature gene the corresponding promoter sequence covering the -1000 to 100 bp regions relative to transcription start sites. Second, we used position weight matrices (PWMs) from the TRANSFAC database (Wingender, 2008) to predict potential TFBSs in promoters. However, computational TFBS predictions are generally considered as being flooded with high rates of false predictions. The accurate prediction of

TFBSs is still a challenging task. To minimize the rate of false predictions in our analysis, we collected a specific PWM library using literature on CRC (**Supplementary Table S3**). This library contains 229 colorectal cancer-related non-redundant matrices. In our further analysis, this library was used with the minFP profile (cut-offs minimizing false positive rate) that contains the adjusted thresholds for each PWM to minimize the prediction of false positive TFBSs. Using our library, we then employed the F-MATCH program described in Schmid et al. (2006) to determine the enriched TFBSs in promoters of the signature genes (foreground set) in comparison to a background set which contains genes with very small fold changes ($\sim$ 0) in both cell lines under study. For this purpose, F-MATCH program applies an iterative process where the initial thresholds in minFP profile are regularly altered until the best possible thresholds are defined which provide most significantly enriched TFBSs. This enrichment analysis yields important key TFs, which may not be mutated themselves, but their altered activation may potentially lead to a persistent expression of their target signature genes, thereby affecting tumorigenesis.

2.4.2. Overrepresented Pathways in Colorectal Cancer

To gain more insights into the functional properties of the signature genes and their transcriptional regulators in CRC, we investigated the overrepresented pathways. For this purpose, we observed the signal transduction and metabolic pathways from TRANSPATH (Krull et al., 2006) database which contains information about genes/molecules and reactions to build complete networks. In this study, we performed two distinct pathway analyses, of which the first one refers to the overrepresented pathways in the signature genes, and the second one is based on the enriched TFBSs found in the promoters of these signature genes.

2.4.3. Identification of Master Regulators with TRANSPATH

Master regulators (MRs) are molecules which are at the very top of regulatory hierarchy and, thus, they are not affected by any of their downstream molecules. Their identification provides important knowledge to display functional relationships of genes. In this study, using the TRANSPATH database, we employed a standard workflow with a maximum radius of 10 steps upstream of TFs to identify their potential MRs.

2.4.4. Transformation of PWMs to Their Corresponding TFs and TF Family/Subfamily Classifications

Multiple PWMs can be assigned to a TF and several TFs belong to a TF family/subfamily. To obtain the correct assignments of the PWMs to their respective TFs and TF family/subfamily, we used the annotations integrated in the geneXplain platform. TF family/subfamily classifications are curated in TFClass (http://tfclass.bioinf.med.uni-goettingen.de/tfclass) which is a classification resource with the aim to catalog TFs based on their DNA-binding characteristics (Wingender et al., 2013). TFClass incorporates a six level classification schema which consists of superclasses, classes, families, subfamilies, genera and factor species of which subfamilies and factor species are optional. At the family level, TFs are primarily grouped on basis of sequence similarities of their DNA-binding domains. The optional subfamily level comprises two more levels which represent genes and gene products, termed genera and species, respectively. TFClass uses a digit-based classification schema which is analogous to the Enzyme Commission numbering system. The schema assigns a four-digit number for the top four classification levels or a six-digit number with respect to the two optional sublevels of the subfamily level.

3. RESULTS

Classical discovery of individual markers usually involves the comparison of normal cells vs. cancer cells, which provides candidates for prognosis as well as individualized treatments. In this study, however, we focused on the *in silico* comparative analysis of two distinct cancer cell lines which serve as models to describe pathways. The cancer cell-intrinsic pathway repertoire and their activation status may differ between distinct cancer cell lines of the same cancer type, which in turn may have an impact on invasiveness and organ colonization *in vivo*. Apart from that, it still remains largely unclear as to what extent these processes are promoted or inhibited by the tumor microenvironment. Therefore, it is mandatory to first learn about the cancer cell line-specific pathway repertoire and, further, to test their functional consequences in *in vivo* models. Above all, the cell lines under study represent suitable models to investigate the molecular mechanisms by which mutations cause predisposition to the formation of multiple colorectal tumors. In addition, they can be used to screen for early disease biomarkers, and to develop therapeutic and preventive strategies.

3.1. Overview of the Analysis Workflow

Our workflow involved four major steps of which the first one was performed using the Trinity platform and all following steps using the geneXplain platform as described below (see also **Figure 1**):

1. Selection of signature genes (Section 3.2)
   a) Analysis of differentially expressed transcripts
   b) Clustering of the most differentially expressed transcripts
2. Identification of overrepresented TRANSPATH pathways based on signature genes (Section 3.3)
   a) Pathway analysis for 1638N-T1 (Section 3.3.1)
   b) Pathway analysis for CMT-93 (Section 3.3.2)
3. Identification of transcription factors (TFs) based on signature genes (Section 3.4)
   a) Prediction of enriched TFBSs in promoters using a CRC-specific PWM library
   b) Mapping of TFBSs to corresponding TFs as well as TF family/subfamily classifications
   c) Grouping of TFs as well as TF family/subfamily into three subsets: 1638NT-1- and CMT-93-intersection-specific TF set; 1638NT-1-specific TF set; CMT-93-specific TF set (Sections 3.4.1, 3.4.3, and 3.4.5)

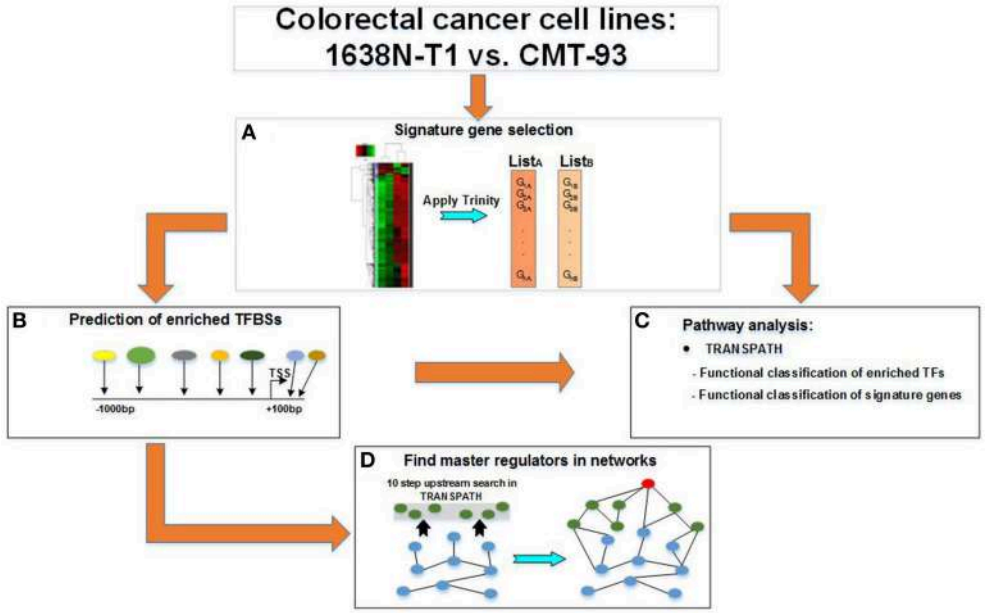


FIGURE 1 | Workflow for the study of distinct colorectal cancer cell lines. A multi-step workflow is outlined for the comparison of the 1638N-T1 and CMT-93. **(A)** The analysis begins with the identification of signature genes based on RNA-seq samples using the Trinity platform. This step generates two disjunct lists of signature genes which are further applied to different geneXplain analyses. **(B)** The signature genes are searched for overrepresented TRANSPATH pathways. Enriched transcription factor binding sites (TFBSs) are searched within the −1 kb/+100 bp promoter regions of the signature genes to obtain transcription factors (TFs). **(C)** The TFs are then searched for overrepresented TRANSPATH pathways. **(D)** A master regulatory network is generated by searching for a master regulator (red node) up to 10 steps upstream of the TFs (blue nodes) in TRANSPATH. The master regulator is connected via intermediate molecules (green nodes) with the TFs.

d) Identification of overrepresented TRANSPATH pathways based on the three TF sets (Sections 3.4.2, 3.4.4, and 3.4.6)

4. Identification of upstream master regulators in pathways based on the three TF sets (Section 3.5)

   a) Search for master regulators upstream of TRANSPATH-mapped molecules of each TF set (Sections 3.5.1, 3.5.2, and 3.5.3)
   b) Merging of master regulator pathways based on the top three master regulators found for each TF set (**Figures 2**–**4**).

3.2. Signature Genes

Tumor initiation, promotion and progression is generally driven by genes whose expression is changed in tumor cells. Comparing gene expression profiles and detection of differentially expressed transcripts between different cancer cell lines can reveal molecular characteristics of the tumor cells under study. Using the Trinity platform we identified signature genes based on their altered transcriptional regulation in the context of CRC. In total, 2296 and 2342 Ensembl gene IDs were identified for 1638N-T1 and CMT-93, respectively. **Supplementary Tables S1**, **S2** provide the full sets of signature genes for 1638N-T1 and CMT-93, respectively.

3.3. Pathway Analyses Based on Signature Gene Sets

The molecular characterization of tumor cells and the molecular mechanisms through which tumor cells acquire the capability to grow progressively, survive and metastasize are numerous and depend on genetic and environmental factors. On the other hand, tumor antigens can be recognized by host T cells, thereby triggering an immune response against the colonization of tumor cells. It is partly the activation of immune system suppressive pathways by the tumor cells which can decide whether cancer evades the anti-tumor immune responses and progresses. Moreover, the expression of various cytokines and chemokines controls the balance between anti-tumor immunity and pro-tumor inflammation. Besides cytokines and chemokines, several TFs and enzymes play critical roles in regulatory functions during tumor development. Therefore, analyzing the tumor-specific expression profiles and detection of these molecules, in particular TFs, are crucial steps in studying the molecular characteristics of tumor cells. Moreover, the knowledge about these molecules and their pathways will provide further information on the molecular mechanisms which may be linked to tumor aggressiveness. In this light, we searched for important pathways for 1638N-T1 and CMT-93 based on their signature genes and exemplarily provided references for their roles in cancer. With the previously defined signature gene sets at hand, we obtained overrepresented TRANSPATH pathways using the geneXplain platform.

3.3.1. Pathway Analysis for 1638N-T1

In total, 30 TRANSPATH pathways were found to be significantly overrepresented based on the signature genes of 1638N-T1 (**Table 1**). The top four most overrepresented pathways indicated

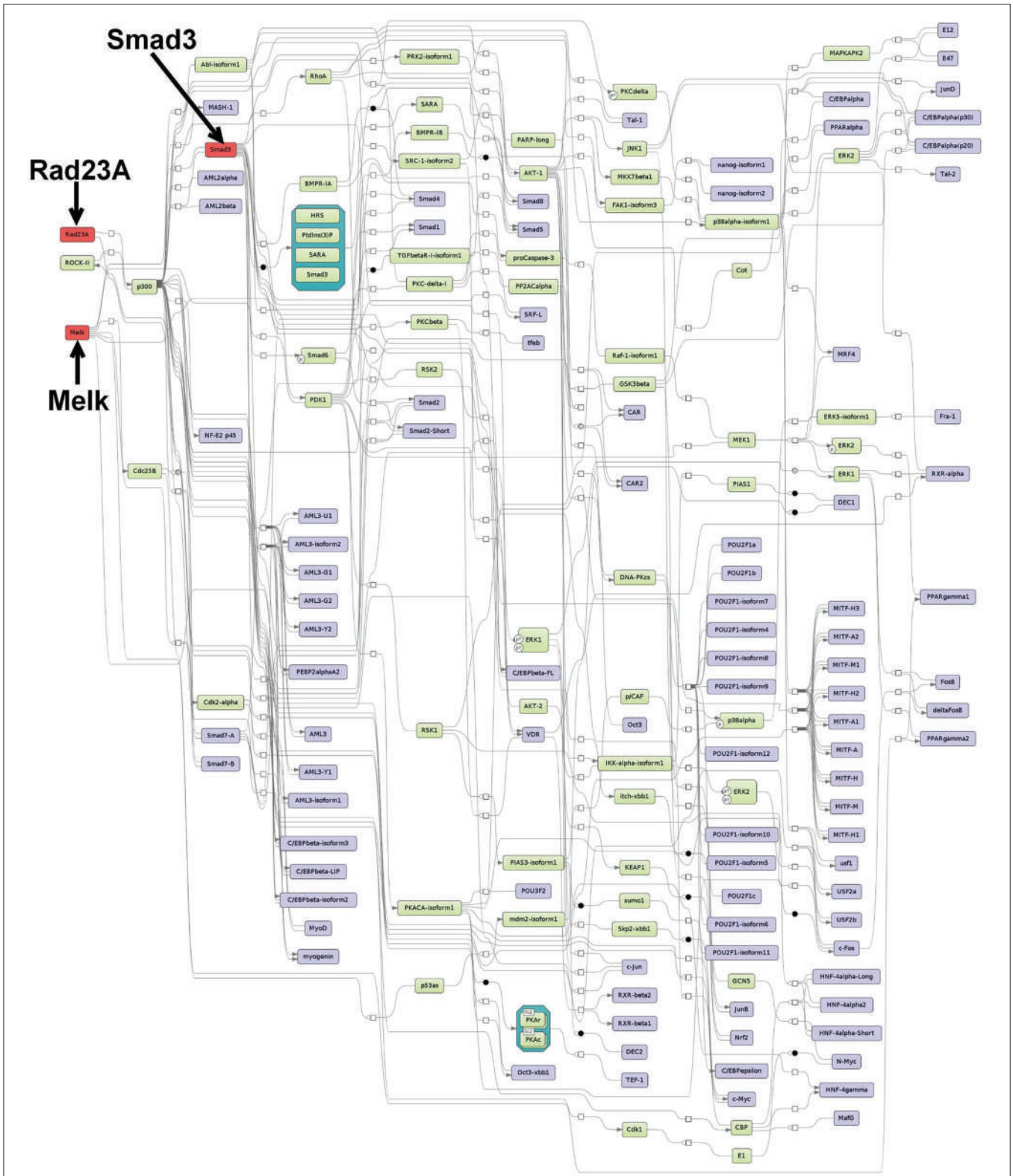


FIGURE 2 | Master regulatory network based on the intersection-specific TF set. The color coding red, blue and green represent nodes for master regulators, regulated transcription factors and connecting molecules, respectively.

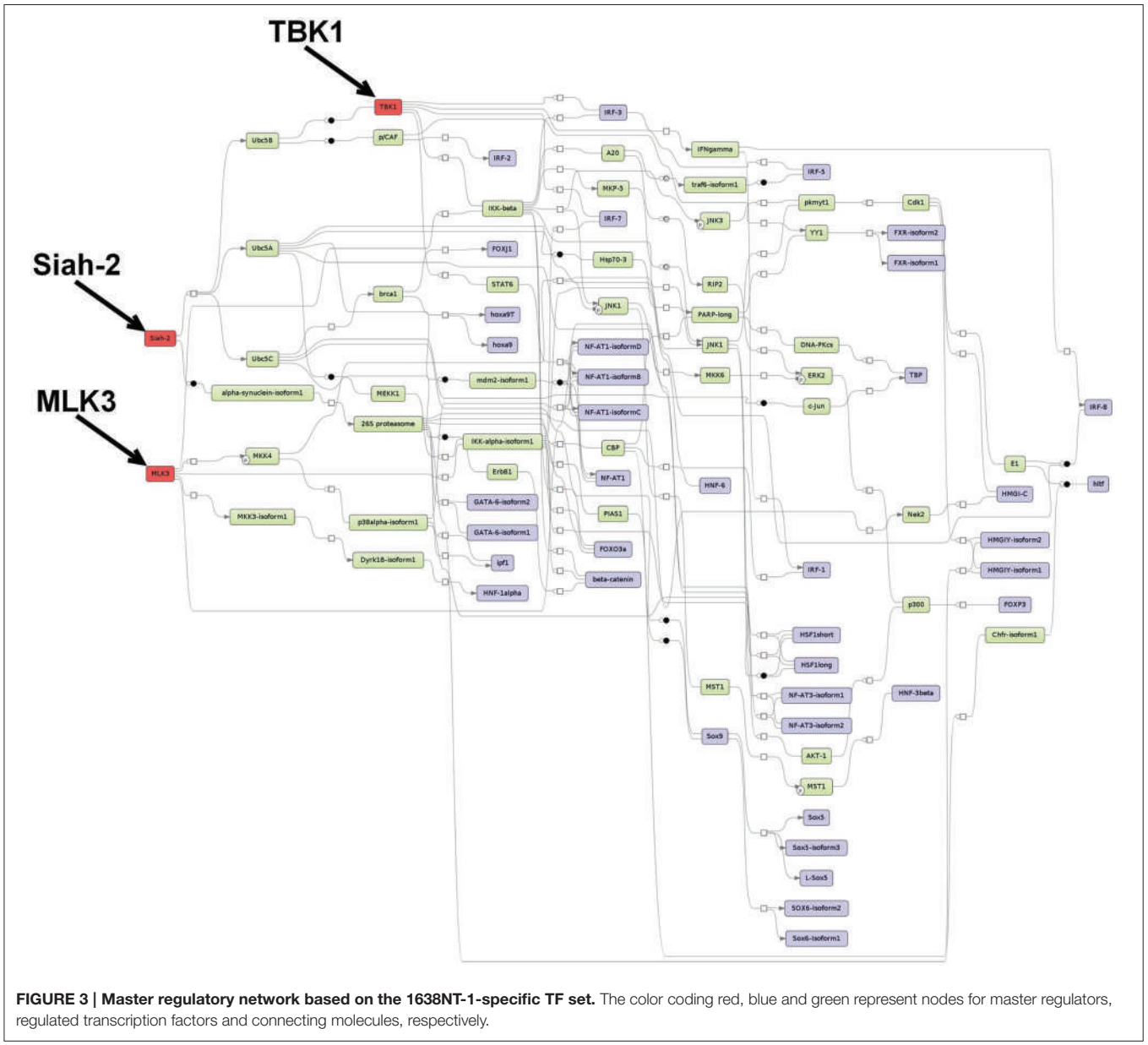


FIGURE 3 | Master regulatory network based on the 1638NT-1-specific TF set. The color coding red, blue and green represent nodes for master regulators, regulated transcription factors and connecting molecules, respectively.

a role for the signature genes Ugt1a1, Ugt1a2, Ugt1a6a, and Ugt1a7c which encode UDP-glucuronosyltransferases (UGTs). These detoxification enzymes are involved in the metabolism of endogenous and xenobiotic compounds (Cooley et al., 1982; Magnanti et al., 2000). Expression of UGTs has been implicated in human urinary bladder and colon cancer (Giuliani et al., 2005; Wang et al., 2012). Furthermore, the first two of the four pathways related to a mechanism for the detoxification of NNAL (the metabolized isoform of NNK) via UGTs-catalyzed glucuronidation pathways (Wiener et al., 2004). NNK is a tobacco agent widely known for promoting tumorigenesis and metastasis through its pro-inflammatory effects (Takahashi et al., 2010). The remaining two pathways related to glucuronidation pathways which are involved in heme degradation in response to oxidative stress. Heme ingestion leads to hyperproliferation and activation of oncogenes as well as the inhibition of the tumor suppressor p53 in response to increased cytotoxicity in the mouse colon (Ijssennagger et al., 2013). The fifth topmost overrepresented pathway corresponded to the activation of Ras-related protein Rap-1A (Rap1A) via interferon gamma (IFNγ). Rap1A is a tumor suppressor which mediates growth inhibitory responses in cancer (Alsayed et al., 2000). The cytokine IFNγ plays an important role in innate and adaptive immune responses and prevents development of primary and transplanted tumors (Ikeda et al., 2002). Further, the pathway analysis found two putative pro-inflammatory metabolic pathways which involve the molecules eicosanoid hepoxilin A3 (hepA3) and platelet

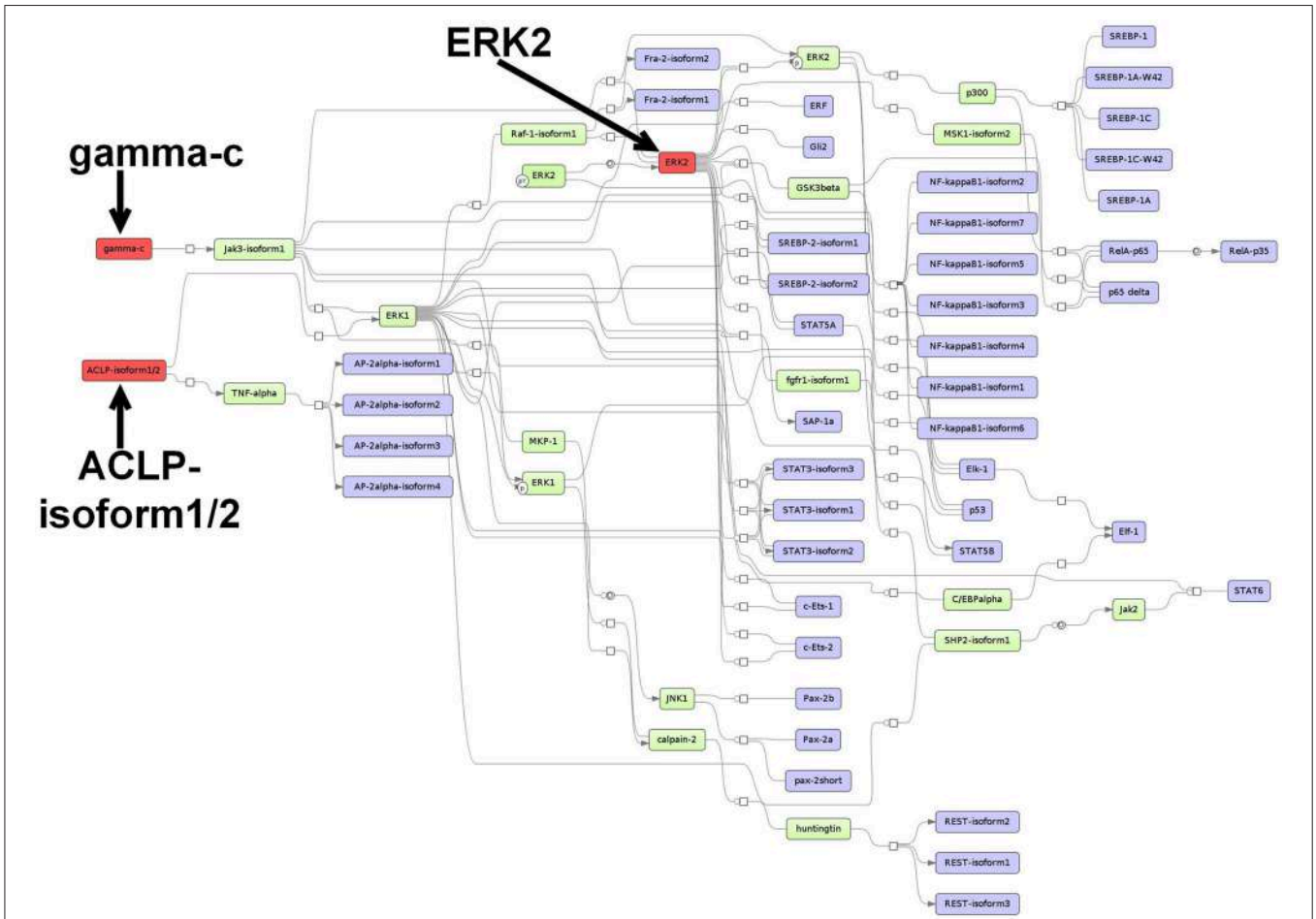


FIGURE 4 | Master regulatory network based on the CMT-93-specific TF set. The color coding red, blue and green represent nodes for master regulators, regulated transcription factors and connecting molecules, respectively.

activating factor (PAF), respectively. Both molecules have been suggested to play key roles in inflammation-associated cancer (Mrsny et al., 2004; Tsoupras et al., 2009). Furthermore, the results reported a signaling cascade which leads to the activation of mitogen-activated protein kinase 1 (Mapk1/Erk2) via interleukin-8 (IL-8). Several studies have implicated IL-8 in tumor angiogenesis, growth, and metastasis in colon, gastric and pancreatic carcinoma (Li et al., 2001, 2008; Kuai et al., 2012; Sun et al., 2014a). A recent study showed that IL-8 increases the migration in human CRC cells through the integrin alpha-V/beta-6 and chemokine receptors CXCR1/2 involving the activation of Mapk1 and Ets-1 signaling pathway (Sun et al., 2014a). Another reported pathway relates to the interleukin-3 (IL-3)-induced activation of the JAK2/STAT5 pathway. IL-3 expression via the T cell receptor signaling pathway is known to regulate growth and differentiation of hematopoietic stem cells, neutrophils, eosinophils, megakaryocytes, macrophages, lymphoid and erythroid cells (Reddy et al., 2000). Lastly, the results showed overrepresentation for the activation of Wnt signaling which is aberrantly activated in the majority of CRCs (Cancer Genome Atlas Network, 2012).

3.3.2. Pathway Analysis for CMT-93

The pathway analysis resulted in the identification of 28 overrepresented TRANSPATH pathways based on the signature genes of CMT-93 (**Table 2**). The four topmost overrepresented pathways share 13/14 hit signature genes which are all associated with the assembly of protein complexes called adherens junctions that occur in epithelial and endothelial tissues (Guo et al., 2007). One prominent signature gene amongst these hits was E-cadherin (cadherin-1/CDH1) that belongs to the cadherin superfamily and encodes a calcium-dependent cell adhesion protein. E-cadherin acts as an invasion suppressor and its loss in epithelial carcinomas permits the invasion of adjacent normal tissues. Several studies showed that the level of E-cadherin expression is inversely correlated with tumor malignancy (Vleminckx et al., 1991; Cowin et al., 2005; Junghans et al., 2005). Likewise, protein-protein interactions between E-cadherin and β-catenin result in the formation of a tumor-suppressor system (Müller et al., 1999). The regulation of β-catenin/E-cadherin has been associated with the induction of epithelial-mesenchymal transition (EMT) and metastasis (Morali et al., 2001; Kim et al., 2002; Eger et al., 2004).

TABLE 1 | Overrepresented TRANSPATH pathways for the signature genes of 1638NT-1.

Pathway	Hit names of signature genes	*P*-value
detoxification and bioactivation of tobacco-derived carcinogen NNK	Cbr3, Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	4.755E-4
NNK → NNAL-O-glucuronide, NNAL-N-glucuronide	Cbr3, Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	4.755E-4
heme, globin → bilirubin beta-diglucuronide	Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	0.00326
hemoglobin oxidation	Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	0.00326
IFNgamma → Rap1	Cybb, Hspa1a, Ifngr1, Ncf4	0.01253
Syk → RhoA	Syk, Vav2	0.01349
Hck → RhoA	Hck, Vav2	0.01349
hepoxilin A3 → Hepoxillin A3-D	Ggt7, Tgm2	0.01349
G-alpha-q → IP3	Cybb, Ncf4, Plcb1	0.01385
BCR → p38	C3,Cybb, Ncf4, Syk	0.01618
BCR —MLK3→ c-Jun	C3, Cybb, Ncf4, Syk	0.01618
catabolism of PAF	Enpp2, Pla2g7, Plcb1, Plcg2	0.01618
alpha IIb beta3 → Rac1	Cybb, Fyb, Ncf4, Prkg1, Syk	0.0211
alpha IIb beta3 pathway	Cybb, Fyb, Ncf4, Prkg1, Syk	0.0211
IL-8 → ERK2	Cxcl1, Cybb, Gnai1, Il8, Ncf4	0.02495
WAVE2 → Arp2/3 complex	Acta1, Actr3b, Cybb, Cyfip2, Ncf4	0.02495
Epo → Syk	Epor, Syk	0.02577
PMCA4 —/ nNOS	Dmd, Snta1	0.02577
Wnt activation of LRP5/6/frizzled/axin complex	Fzd4, Fzd8, Wnt1	0.0268
SDF-1 → G-protein	Cxcr4, Cybb, Gnai1, Ncf4, Pik3r5	0.02923
BCR → cytoskeletal reorganization	C3, Cybb, Ncf4, Syk	0.03089
BCR → c-Jun	C3, Cybb, Ncf4, Syk	0.03089
SLP-65 —/ Raf-1	Cybb, Ncf4, Plcg2	0.03503
dehydroepiandrosterone → estriol 16-glucuronide	Cyp1b1, Cyp4a12a, Cyp4b1, Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	0.03888
IL-3 → STAT5	Csf2rb, Il3ra	0.04102
beta-glucan —DECTIN1→ IP3, DAG	Plcg2, Syk	0.04102
metabolism of estrogens	Cyp1b1, Cyp4a12a, Cyp4b1, Ugt1a1, Ugt1a2, Ugt1a6a, Ugt1a7c	0.04299
Rac1 —p65PAK→ Arp2/3 complex	Acta1, Actr3b, Cybb, Ncf4	0.04396
Src → Rac1	Cybb, Ncf4, Vav2	0.0444
N-cadherin —Eplin→ actin	Acta1, Cdh2, Ctnna2	0.0444

The results included further pathways which are related to the phosphorylation and desphosphorylation of the β-catenin/E-cadherin complex. In this regard, it has been reported that phosphorylation of β-catenin, e.g., through the epidermal growth factor receptor (EGFR) or the tyrosine-protein kinase Src, leads to the dissociation of the complex and consequently to the accumulation of free β-catenin. On the contrary, dephosphorylation of β-catenin results in the formation of the complex (Müller et al., 1999). Another overrepresented pathway corresponded to nerve growth factor (NGF) signaling via the tyrosine kinase receptor TrkA. NGF has been associated with cancer cell proliferation as well as apoptosis of colon cancer cells (Molloy et al., 2011; Anagnostopoulou et al., 2013) and with angiogenesis (Romon et al., 2010). Further overrepresented pathways related to the angiopoietin-Tie signaling system which plays a role in the regulation of angiogenesis (Fagiani and Christofori, 2013). In tumors, angiopoietin-2 (Ang2) inhibits the activity of the receptor tyrosine kinase Tie2 and destabilizes blood vessels, thereby facilitating angiogenesis (Holash et al., 1999a,b; Augustin et al., 2009). Moreover, several other overrepresented pathways could be linked to anti-tumor properties. These included two p53-dependent pathways which lead to the induction of the cyclin-dependent kinase inhibitor 1 ($p21^{Cip1}$) or the p53 upregulated modulator of apoptosis (Puma), respectively. Downregulation of $p21^{Cip1}$ expression has been associated with poor prognosis and expression of Puma with a rapid apoptosis in CRC (Pasz-Walczak et al., 2001; Yu et al., 2001). Furthermore, the results also included overrepresented pathways which related to vitamin D receptor (VDR) signaling and vitamin D metabolism. VDR signaling is activated upon binding of vitamin D and plays a role in cancer progression as well as cross-talks with multiple other pathways (Slattery, 2007). For example, several studies have suggested interactions of vitamin D or its active vitamin D metabolite, calcitriol, with β-catenin (Deeb et al., 2007; Zheng et al., 2012; Klampfer, 2014). These interactions represent points of convergence between VDR and canonical Wnt signaling in CRC, which has been linked to inhibition of Wnt signaling, tumor growth inhibition, the activation of apoptotic pathways,

TABLE 2 | Overrepresented TRANSPATH pathways for the signature genes of CMT-93.

Pathway	Hit names of signature genes	*P*-value
beta-catenin:E-cadherin complex phosphorylation and dissociation	Axl, Blk, Cdh1, Epha1, Erbb3, Fes, Kit, Lck, Mertk, Ntrk1, Ret, Tek, Txk	0.00147
beta-catenin:E-cadherin complex phosphorylation and dephosphorylation	Axl, Blk, Cdh1, Epha1, Erbb3, Fes, Kit, Lck, Mertk, Ntrk1, Ret, Tek, Txk	0.00147
tyrosine dephosphorylation of plakoglobin	Axl, Blk, Cdh1, Epha1, Erbb3, Fes, Kit, Lck, Mertk, Ntrk1, Ret, Tek, Txk	0.00166
beta-catenin network	Axl, Blk, Cdh1, Epha1, Erbb3, Fes, Kit, Lck, Magi2, Mertk, Ntrk1, Ret, Tek, Txk	0.002
NGF —p75NTR→ trkA	Ngf, Ntrk1	0.00464
VDR network	Cyp27a1, Cyp2r1, Hist1h4i, Hist1h4j, Hist2h3c2, Hist2h4, Hist4h4, Vdr	0.00541
NGF → trkA	Ngf, Ntrk1	0.0133
Tie2 dephosphorylation	Ptprb,Tek	0.0133
CO2, H2O → spermine	Arg1, Car14, Car2, Car3, Car6	0.01419
Angiopoietin/Tie signaling	Dok2, Nos3, Ptprb, Sfn, Tek	0.01419
creatine biosynthesis and degradation	Car14, Car2, Car3, Car6, Gatm, Mat1a	0.01625
VDR → RXR-alpha → transcriptional activation	Hist1h4i, Hist1h4j, Hist2h3c2, Hist2h4, Hist4h4, Vdr	0.01891
sphinganine → ceramide-2,3,6,7	Cers1, Cers4, Ugcg	0.01936
urea and aspartate cycles, polyamine and creatine synthesis	Arg1, Car14, Car2, Car3, Car6, Gatm	0.02184
CO2, L-ornithine → L-arginine	Car14, Car2, Car3, Car6	0.02475
p53 → p21Cip1	Hist1h4i, Hist1h4j, Hist2h4, Hist4h4	0.02475
p53 → PUMA	Hist1h4i, Hist1h4j, Hist2h4, Hist4h4	0.02475
7-dehydrocholesterol → calcitriol	Cyp27a1, Cyp2r1	0.02542
formation of vitamin D3 and 1alpha,25-dihydroxycholecalciferol	Cyp27a1, Cyp2r1	0.02542
Nedd4 → trkA	Ngf, Ntrk1	0.02542
PKAc → NR2C	Grin1, Prkaca	0.02542
NR2A:NR2B —PKAc→ Ca	Grin1, Prkaca	0.02542
Vitamin D metabolism	Cyp27a1, Cyp2r1	0.02542
Tie2 —p56Dok-2→ PAK1	Dok2, Tek	0.02542
L-tryptophan → 5-hydroxyindoleacetate	Aldh1a7, Maoa, Tph1	0.03438
degradation of tryptophan	Acmsd, Aldh1a7, Maoa, Tph1	0.03625
Csk, CD45 → Lck	Lck, Ptprc	0.04048
NR2B:NR2C —CaMKII→ c-Fos	Camk2d, Grin1, Prkaca	0.0436

inhibition of angiogenesis and inhibition of tumor-promoting inflammation (Deeb et al., 2007; Zheng et al., 2012; Klampfer, 2014).

3.4. Promoter Analysis Based on Signature Genes

Altered gene expression is generally a result of the dysregulated activity of TFs that may play central roles as oncogenes and tumor suppressors. These proteins are often potential targets for cancer therapies due to the fact that many oncogenic signaling pathways involve TFs whose aberrant activation and inactivation contributes to tumor development and progression. We applied a promoter analysis to the previously identified signature genes in order to display which TFs are potentially important regulators in the cell lines under study. This analysis was performed using geneXplain which quantifies the enrichment of TFBSs in promoter regions of the signature genes. In total, 135 and 117 TFs were identified for 1638N-T1 and CMT-93, respectively. These numbers include 51 (**Supplementary Table S4**) and 33 TFs (**Supplementary Table S5**) that were exclusively enriched in 1638N-T1 or CMT-93, respectively, as well as 84 overlapping TFs in the intersection between both cell lines (**Supplementary Table S6**). We exemplarily highlighted several TF families/subfamilies which are present for the three TF sets. In a subsequent analysis, we additionally searched for overrepresented pathways on the basis of these sets.

3.4.1. Intersection-Specific TF Families/Subfamilies of 1638N-T1 and CMT-93

The enriched TFBSs were classified into 32 prominent TF families/subfamilies according to TFClass (**Table 3**). Our analysis detected several members of the SMAD factor family that were found to have enriched binding sites in the promoters. These factors are a major component of TGF-β signaling which is involved in the regulation of cell growth in the normal intestinal epithelium. Alterations in their expression contribute to cancer aggressiveness in CRC (Xie et al., 2003; Xu and Pasche, 2007; Korchynskyi et al., 1999; Fleming et al., 2013). Furthermore, the analysis revealed overrepresentation for members of the Jun-related factors and Fos-related factors. The protein AP-1 is composed of either Jun-Jun homodimers or Jun-Fos heterodimers and plays a role in differentiation, proliferation, and apoptosis (Ameyar et al., 2003). AP-1 is induced by c-Jun

TABLE 3 | Intersection-specific TF families/subfamilies between 1638N-T1 and CMT-93.

TF classification	TF family/subfamily
1.1.1	Jun-related factors
1.1.1.1	Jun factors
1.1.1.2	NF-E2-like factors
1.1.2	Fos-related factors
1.1.2.1	Fos factors
1.1.3	Maf-related factors
1.1.3.1	Large Maf factors
1.1.3.2	Small Maf factors
1.1.8	C/EBP-related
1.1.8.1	C/EBP
1.2.1	E2A-related factors
1.2.2	MyoD / ASC-related factors
1.2.2.1	Myogenic transcription factors
1.2.3.1	Tal / HEN-like factors
1.2.6	bHLH-ZIP factors
1.2.6.1	TFE3-like factors
1.2.6.2	USF factors
1.2.6.5	Myc / Max factors
1.2.6.7	Mad-like factors
2.1.2	Thyroid hormone receptor-related factors (NR1)
2.1.2.4	Vitamin D receptor (NR1I)
2.1.3	RXR-related receptors (NR2)
2.1.3.1	Retinoid X receptors (NR2B)
2.1.3.2	HNF-4 (NR2A)
3.1.10	POU domain factors
3.1.10.2	POU2 (Oct-1/2-like factors)
3.1.4	TALE-type homeo domain factors
3.1.4.4	PBX
6.4.1	Runt-related factors
7.1.1	SMAD factors
7.1.1.1	Regulatory Smads (R-Smad)
7.1.1.3	Repressor-Smads (I-Smad)

N-terminal protein kinases (JNK) and ERK MAPKs pathways or the canonical Wnt signaling pathway in CRC (Licato et al., 1997; Mann et al., 1999), thereby affecting CRC cell proliferation (Suto et al., 2004). Binding site enrichment was also detected for the CCAAT-enhancer binding protein (C/EBP) family of TFs whose expression has been associated with invasiveness of human colorectal cancer (Rask et al., 2000). Likewise, several members of the POU domain factor family, including Oct-4 (Pou5f1), were found in the intersection between both cell lines. It has been reported that Oct-4 promotes metastasis in CRC through EMT (Dai et al., 2013). Furthermore, Oct-4 knockdown leads to decreased Wnt pathway activity and high risk for liver metastases in CRC patients (Dai et al., 2013). Enrichment for binding sites of VDR, which belongs to the Thyroid hormone receptor-related factor (NR1) family, was also detected in the intersection. It has been suggested that vitamin D has no effect on tumor reduction in APC-deficient mice and that VDR expression is lost in the majority of the colon cancer cells (Giardina et al., 2015). Interestingly, the analysis also revealed enrichment for binding sites of β-catenin which interacts as a cofactor with members of the TCF-7-related factor family to activate Wnt target gene expression (see **Supplementary Table S6**).

3.4.2. Overrepresented TRANSPATH Pathways Based on Intersection-Specific TFs

Based on the 84 overlapping TFs in the intersection of both cell lines, the pathway analysis revealed overrepresentation for 35 TRANSPATH pathways (**Table 4**). Members of the SMAD factor family were found to be involved in many of the top overrepresented pathways. In this context, the TGF-β pathway was detected among the most overrepresented pathways. Likewise, SMADs were also found to be involved in a pathway which corresponded to the regulation of endothelin-1 (ET-1). ET-1 is a vasoconstrictor peptide, which is known to be produced by CRC cells and stimulates CRC proliferation (Asham et al., 2001; Grant et al., 2007; Knowles et al., 2012). The second most overrepresented pathway corresponded to the transcriptional regulation of ECM components. ECM sustains normal tissue homeostasis and prevents malignant transformation (Gao et al., 2014). Its anti-tumor properties are opposed by chronic inflammation, which may lead to the conversion of a tumor-inhibiting into a tumor-promoting microenvironment (Gao et al., 2014).

Furthermore, the analysis showed overrepresentation for a PPAR-related pathway which comprises the peroxisome proliferator activated receptors PPAR-α, PPAR-γ and Smads. It was shown that activation of PPAR-γ inhibits TGF-β-induced loss of E-cadherin expression, the induction of mesenchymal markers (vimentin, N-cadherin, fibronectin), MMPs and antagonizes Smad3 function, thereby preventing metastasis in lung cancer (Reka et al., 2010). This pathway has also been implicated in the induction of apoptosis as well as inhibition of cyclooxygenase-2 (COX-2) in CRC (Yang and Frucht, 2001). Activation of the PPAR pathway was shown to cause reduction in linear and clonogenic growth and, thus, it has been suggested that PPAR-γ modulates cell growth and differentiation of CRC cells (Sarraf et al., 1998). Moreover, it was shown that PPAR-γ expression is altered in APC-deficient mice, an effect which is thought to be mediated by the Wnt/β-catenin pathway (Jansson et al., 2005). In conformity with the overrepresented pathways, which were found based on the signature genes of CMT-93, a VDR network-related pathway was also found based on the intersection-specific TFs.

3.4.3. 1638N-T1-Specific TF Families/Subfamilies

The enriched TFBSs can be classified into 14 prominent TF families/subfamilies based on the 1638N-T1-specific TFs (**Table 5**). Amongst these, the factors Onecut1 and Onecut2, which belong to the HD-CUT factors family, were found to be enriched in the signature genes of 1638N-T1. Through targeting of Onecut2, the microRNA miR-429 has been reported to regulate the expression of several EMT-related markers (Sun et al., 2014b). Overall, it has been suggested that Onecut2 is involved in EMT, migration and invasion of CRC cells (Sun et al.,

TABLE 4 | Overrepresented TRANSPATH pathways based on the intersection-specific TF set of 1638N-T1 and CMT-93.

Pathway	Hit names of TFs	*P*-value
Endothelin-1 gene regulation	Fos, Jun, Smad3, Smad4	2.8851480825350153E-8
Transcriptional Regulation of ECM components	Smad2, Smad3, Smad4, Tfe3	4.2462045176114295E-7
PPAR pathway	Ppara, Pparg, Rxra, Rxrb, Smad2, Smad3	4.401281772213993E-7
BMP7 → Smad1, Smad5, Smad8	Smad1, Smad4, Smad5, Smad9	9.814052909861147E-7
TGFbeta pathway	Fos, Jun, Pparg, Smad1, Smad2, Smad3, Smad4, Smad5, Smad7, Smad9, Tfe3	1.1668767789940478E-6
SMAD7, SIK1 gene induction	Smad2, Smad3, Smad4	2.3427402430218484E-6
MIC2 signaling	Fosb, Jun, Jund, Srf	8.907929442840803E-6
Smad2/3, PPARgamma, regulation of bioavailability	Pparg, Smad2, Smad3	9.284406529601274E-6
MIC2-isoform2 —JNK, JunD→ MMP9	Fosb, Jun, Jund	4.556873521617853E-5
TGFbeta1 → Smad1, Smad2, Smad5	Smad1, Smad2, Smad5	7.900923266022097E-5
MIC2-isoform2 —FosB→ MMP9	Fosb, Jund, Srf	1.2524823045846698E-4
mammalian Hippo network	Smad2, Smad3, Smad4, Smad7, Tead1	1.609311066368507E-4
RA, 15d-PGJ2 → RXR-beta, PPAR-gamma	Pparg, Rxrb	1.830040551373434E-4
RXR-beta, VDR heterodimerization	Rxrb, Vdr	1.830040551373434E-4
Smad2/3 —TAZ→ cytoplasmic retention	Smad2, Smad3, Smad4	3.5891749789138236E-4
Sox9 —Smad3→ COL2A1	Smad2, Smad3	5.443266849267895E-4
MyoD regulation	Myod1, Tcf3	5.443266849267895E-4
MKK4 —/ PPAR-gamma	Pparg, Rxrb	0.0010793689633243411
Ctbp1 —/ Smad3	Smad3, Smad4	0.0010793689633243411
ERK1 → NQO1	Mafk, Nfe2l2	0.0010793689633243411
E2F —/ Smad4	Smad3, Smad4	0.0017836171536470523
Nrf2 → HMOX1	Mafk, Nfe2l2	0.0017836171536470523
stress-associated pathways	Jun, Mitf, Myf6, Pparg, Rxra, Rxrb	0.0023269497130444508
PRIC complex → PPAR-alpha	Ppara, Rxra	0.002652641362864685
TGFbeta1 → Smad2/3	Smad2, Smad3	0.002652641362864685
MEK → EZR	Fos, Jun	0.002652641362864685
p38 pathway	Jun, Mitf, Myf6, Pparg	0.0034193798154062926
15-Keto-PGE2 → TP63	Pparg, Smad2	0.003682094214938695
TGFbetaR-I —pak2, ERK1→ SMAD7, SERPINE1	Smad2, Smad3, Smad4	0.003904012718560197
15d-PGJ2 → PPAR-gamma	Pparg, Rxra	0.009320059910591498
Regulation of mesendoderm differentiation genes	Smad2, Smad4	0.012994431232912744
IRAK-1 —MKK3→ TNF	Fos, Jun	0.015031819490714783
JNK pathway	Jun, Pparg, Rxra, Rxrb	0.016302058148038395
VDR network	Rxra, Rxrb, Vdr	0.01758202640044028
TGFbetaR-I → ERK	Smad2, Smad3	0.04188993264127895

2014b). Onecut1 (Hnf6) expression was found to be positively correlated with the expression of p53 and E-cadherin in human lung cancer. The Onecut1-mediated induction of p53 is thought to inhibit EMT, migration and invasion (Yuan et al., 2013). Moreover, the analysis detected the HOX-related factors Cdx1 and Cdx2, which regulate intestine-specific gene expression and enterocyte differentiation (Suh et al., 1994; Suh and Traber, 1996; Taylor et al., 1997; Freund et al., 1998; Soubeyran et al., 1999; Lynch et al., 2003). In addition, it has been suggested that expression of Cdx1 reduces cancer cell proliferation by reducing cyclin D1 expression (Lynch et al., 2003). Interestingly, Cdx1 and Cdx2 also inhibit proliferation of CRC cells by blocking canonical Wnt signaling activity (Guo et al., 2004). In contrast, another study indicated that Cdx2 can promote expression of Wnt/β-catenin pathway genes (da Costa et al., 1999). Furthermore, the analysis revealed overrepresentation for several members of the interferon regulatory factor (IRF) family. Most IRFs play central roles in immune response, apoptosis and are known to exhibit tumor suppressor properties in cancer (Bouker et al., 2005). For example, anti-tumor function of IRF-1- and IRF-5-associated pathways have been suggested in CRC (Hu and Barnes, 2006; Yuan et al., 2015). The analysis also detected Sox9, a member of the SOX-related factors. Sox9 is a target as well as potential upstream regulator of Wnt signaling (Blache et al., 2004; Bastide et al., 2007).

3.4.4. Overrepresented TRANSPATH Pathways Based on 1638N-T1-Specific TFs

In total, 7 overrepresented pathways were found based on the 51 exclusive TFs for 1638NT-1 (**Table 6**). The results included

TABLE 5 | 1638N-T1-specific TF families/subfamilies.

TF classification	TF family/subfamily
3.1.1.9	CDX (Caudal type homeobox)
3.1.9	HD-CUT factors
3.1.9.1	ONECUT
3.1.10.7	HNF1-like factors
3.3.1	Forkhead box (FOX) factors
3.3.1.1	FOXA
3.3.1.6	FOXF
3.5.3	Interferon-regulatory factors
4.1.1	SOX-related factors
4.1.1.3	Sox-related factors, Group C
4.1.1.4	Sox-related factors, Group D
4.1.1.5	Sox-related factors, Group E
6.1.3	NFAT-related factors
8.2.1	HMGA factors

TABLE 6 | Overrepresented TRANSPATH pathways based on the 1638N-T1-specific TF set.

Pathway	Hit names of TFs	*P*-value
dsRNA → IRF-7:IRF-3:CBP:p300	Irf3, Irf7	3.947146928237511E-4
LPS → IRF-3:IRF-7:CBP:p300	Irf3, Irf7	0.0014260355781928803
wnt → beta-catenin	Ctnnb1, Tbp	0.005286143325503229
TLR9 pathway	Irf1, Irf7	0.0106622039857671
TLR3 pathway	Irf3, Irf7	0.01598971078797065
wnt pathway	Ctnnb1, Tbp	0.02936961872680831
TLR4 pathway	Irf3, Irf7	0.03155510304218106

overrepresented pathways which corresponded to the TLR (Toll-like receptor) pathways TLR3, TLR4, and TLR9. TLRs are pattern recognition receptors (PRRs) that play key roles in innate and adaptive immune responses. In host defence, TLRs recognize pathogens by pathogen-associated molecular patterns (PAMPs). TLRs are involved in inflammatory reponses, cell proliferation and survival, and have been associated with pro-tumor as well as anti-tumor effects in cancer (Rakoff-Nahoum and Medzhitov, 2009; Basith et al., 2012). TLR signaling pathways promote the production of cytokines and chemokines via interfering with intracellular pathways and activation of TFs, such as IRFs and NF-κB (Li et al., 2014). In particular, activation of the TLR9 pathway promotes the development of anti-tumor T-cell responses (Krieg, 2008). In contrast, it was also shown that this pathway can promote angiogenesis and cancer progression (Belmont et al., 2014; Holldack, 2014). TLR3 activation mediated by dsRNA was shown to trigger apoptosis of human breast cancer cells (Salaun et al., 2006). Additionally, signaling by IRF-3 has been implicated in TLR3-mediated apoptosis in prostate cancer (Gambara et al., 2015). Another overrepresented TRL-related pathway corresponded to the lipopolysaccharide (LPS)-induced activation of the TFs IRF-3, IRF-7, and CBP/p300 via the TLR4/MD2 complex. Moreover, it was shown that metastasis of CRC cells is increased through a signaling cascade involving

TABLE 7 | CMT-93-specific TF families/subfamilies.

TF classification	TF family/subfamily
1.1.2	Fos-related factors
1.2.6.3	SREBP factors
2.3.3.1	GLI-like factors
3.5.2	Ets-related factors
3.5.2.1	Ets-like factors
3.5.2.2	Elk-like factors
3.5.2.3	Elf-1-like factors
6.1.1	NF-kappaB-related factors
6.1.5	Early B-Cell Factor-related factors
6.2.1	STAT factors

LPS-induced TLR4 signaling as well as downstream PI3K/Akt signaling and β1 integrin activity (Hsu et al., 2011). LPS also increases phosphorylation of Mapk1 and p38, activation of NF-κB, and promotes cytokine production, such as that of IL-8, vascular endothelial growth factor (VEGF), and TGF-β in human colon cells (Tang and Zhu, 2012). Moreover, the same study has implicated TLR4 in promoting immune escape of the human colon cancer cells by inducing immunosuppressive factors and apoptosis resistance (Tang and Zhu, 2012). Strikingly, two pathways corresponded to the canonical Wnt/β-catenin signaling pathway which is of high relevance in CRC.

3.4.5. CMT-93-Specific TF Families/Subfamilies

The enriched TFBSs can be classified into 10 prominent TF families/subfamilies for the CMT-93-specific TFs (**Table 7**). The results included Ebf3 which is a member of the Early B-Cell Factor-related factors family. This family plays a role in differentiation of specific cell types such as B lymphocytes and olfactory cells (Zhao et al., 2006). Expression of Ebf3 was previously shown to promote cell cycle arrest and apoptosis in several tumor cell lines including colon carcinoma (Zhao et al., 2006).

The analysis also reported enriched TFBSs for the NF-κB-related factor family. NF-κB signaling is usually induced by inflammation and also known to be triggered by cancer progression. Many recent findings indicate that NF-κB is constitutively activated in malignant cells of various cancers including CRC (Nakshatri et al., 1997; Wang et al., 1999; Lindholm et al., 2000; Lind et al., 2001; Kojima et al., 2004), thereby promoting, cell proliferation, angiogenesis, metastasis, upregulation of chemokine secretion and other anti-apoptosis proteins (Sakamoto et al., 2009; Wang et al., 2009). Furthermore, enriched binding sites were detected for the signal transducer and activator of transcription (STAT) family which are critical regulators of immune and inflammatory responses (Yu et al., 2009). These factors play an important role in many types of cancer, including colorectal cancer, as they may promote pro-tumor inflammatory pathways such as NF-κB and JAK/STAT pathways, as well as suppress anti-tumor immunity (Wang et al., 2009; Yu et al., 2009; Slattery et al., 2013). The activation of Stat3 and Stat5 has been shown to promote cell proliferation and invasion in cancer (Yu et al., 2009), while Stat3 was also

found to be persistently activated and overexpressed in colon cancers (Klampfer, 2008). Our analysis also revealed binding site enrichment for several members of the family of Ets-related factors which are involved in diverse cellular processes, thereby often cooperatively interacting with other TFs and co-factors (Oikawa, 2004). In cancer, this family is known to regulate genes which play a role in angiogenesis, invasion and metastasis. Therefore, their altered expression has been implicated in development and progression of cancer (Bassuk and Leiden, 1997; Graves and Petersen, 1998; Oikawa and Yamada, 2003; Oikawa, 2004). Moreover, it has been suggested to use ETS-related factors as prognostic markers in cytotoxic treatment of metastatic colorectal cancer (Giessen et al., 2013).

3.4.6. Overrepresented TRANSPATH Pathways Based on CMT-93-Specific TFs

In total, 52 overrepresented pathways were found based on the 33 exclusive TFs for CMT-93 (**Table 8**). Most of these overrepresented pathways involved NF-κB family members. Further overrepresented pathways involved the tumor necrosis factor-alpha (TNF-α) of which one related to the TNF-α-mediated activation of NFκB. An increase in production of the pro-inflammatory cytokine TNF-α is linked to poor outcome in CRC (Balkwill2005, Mantovani2005, Coussens2002, Balkwill2001). Interestingly, TNF-α was shown to promote Wnt signaling through translocation of β-catenin into the nucleus in gastric tumor cells (Oguma et al., 2008).

In conformity with the results obtained for the 1638N-T1-specific TFs, TLR-related pathways for five different TLRs (TLRs 2,3,4,8,9) were also detected for the TFs of the CMT93-specific set. The results further included several overrepresented STAT factors-related pathways that included an activation of STATs by platelet-derived growth factor (PDGF)-mediated signaling. Signaling via PDGF tyrosine kinase receptors plays an important role in angiogenesis, mesenchymal cell migration, proliferation and the expression and activation of PDGF receptors is particularly associated with invasion and metastasis in CRC (Yu et al., 2003; Kitadai et al., 2006; Steller et al., 2013).

Moreover, the analysis detected overrepresentation for LXR-related pathways that implicate a role for NFκB subunits RELA/p65, NFKB1/p105, NFKB1/p50 as well as interleukin-1 beta (IL-1β). Interestingly, the signature gene set for CMT-93 included the factors Nr1h2 and Nr1h3, two members of the thyroid hormone receptor-related factor (NR1) family. These genes encode liver X receptors (LXRs), of which the oxysterol receptor LXRα (Nr1h3) is thought to increase caspase-dependent apoptosis, slow growth of xenograft tumors in CRC mouse models and may negatively interfere with Wnt signaling through direct binding to β-catenin in CRC (Uno et al., 2009; Sasso et al., 2013). Hence, LXRs have been considered important potential targets in cancer therapeutics on account of their tumor suppressor activities (Sasso et al., 2013; Vedin et al., 2013; Lin and Åke Gustafsson, 2015). With respect to IL-1β, this pro-inflammatory cytokine has been associated with angiogenesis, invasiveness of different tumor cells and increased risk of CRC (Voronov et al., 2003; Andersen et al., 2013).

3.5. Identification of Upstream Master Regulators in Pathways Based on TF Sets

In the previous step, we reported potentially important TFs for the sets of signature genes, on the basis of which we defined sets of TFs for the intersection between the two cell lines as well as for the 1638N-T1-specific and CMT-93-specific TFs. Since signal transduction pathways can modulate the activity of nuclear TFs, activation mutations in these pathways can lead to the altered expression of the TFs and their target genes. These pathways are diverse in both their complexity and the mechanism of signal transduction, and even more complexity is added through cross-talks or transactivation signals between different pathways. Therefore, we were interested in the detection of upstream regulators, called master regulators, for the previously defined TF sets. We additionally aimed to construct the upstream pathways which may regulate activity or inhibition of the TFs.

We applied the master regulator analysis from geneXplain to each of the three TF sets, namely the intersection with overlapping TFs between 1638N-T1 and CMT-93, the 1638N-T1-specific and the CMT-93-specific TFs. This workflow will first map the set-specific TFs to TRANSPATH molecules and then search based on the TRANSPATH knowledge for upstream master regulators. We report the top three master regulators for each TF set (**Table 9**) and provide references for their roles in cancer. Noteworthy, we only proposed distinct master regulators for each gene set, i.e., different splice variants or isoforms of a master regulator reported by the analysis were counted as the same master regulator.

The master regulators and their pathways, denoted as master regulator pathways, constitute the set-specific TFs which are either connected to other set-specific TFs or intermediate molecules. These intermediate molecules are not contained within the respective TF sets but function as a bridge between the set-specific TFs and the master regulator(s) in the pathways. Since the pathways of the top ranked master regulators share many of the interacting nodes and, thus, are very similar to each other, we merged the top 3 master regulator pathways for each set into one network.

3.5.1. Prediction of Master Regulators and Construction of a Master Regulatory Network Based on the Intersection-Specific TF Set

For the intersection-specific TF set, we obtained the three master regulators Rad23A, Smad3, and Melk that reach 91, 74, and 93 TFs from the set, respectively. The master regulator Rad23A is involved in DNA damage recognition and nucleotide-excision repair. A recent study has implicated Rad23A in nuclear translocation of the apoptosis-inducing factor (AIF) during induction of cell death (Sudhakar and Chow, 2014). However, not much is known about its specific function in CRC.

As a major component of the TGF-β signaling pathway, the Smad3 master regulator plays a pivotal role in survival, invasion, and metastasis of CRC cells (Xu and Pasche, 2007; Fleming et al., 2013). However, despite the fact that not much is known about the pathogenic role of Smad3, mutations in the gene occur rather rarely in human CRC (Ku et al., 2007). Loss of Smad3 has been

TABLE 8 | Overrepresented TRANSPATH pathways based on the CMT-93-specific TF set.

Pathway	Hit names of TFs	*P*-value
PDGF B → STATs	Stat3, Stat5a, Stat5b	6.272149884041184E-7
STAT5 → Ccnd1	Stat5a, Stat5b	3.1953088992325076E-5
STAT5 → CISH	Stat5a, Stat5b	3.1953088992325076E-5
STAT5 → CSN2	Stat5a, Stat5b	3.1953088992325076E-5
PDGF B → STAT1alpha, STAT5	Stat5a, Stat5b	9.554461322021479E-5
importin-alpha3 → NFkappaB	Nfkb1, Rela	9.554461322021479E-5
Pin1 → p50:RelA-p65	Nfkb1, Rela	9.554461322021479E-5
Epo —Jak2→ STAT5	Stat5a, Stat5b	1.9046201145220444E-4
Epo → STAT5	Stat5a, Stat5b	1.9046201145220444E-4
IL-3 → STAT5	Stat5a, Stat5b	3.1639480397985514E-4
LXR —/ IL1B	Nfkb1, Rela	3.1639480397985514E-4
SOCS-1 → p50:RelA-p65	Nfkb1, Rela	4.730345816689199E-4
TLR8 —Btk→ NF-kappaB	Nfkb1, Rela	4.730345816689199E-4
TLR9 —Btk→NF-kappaB	Nfkb1, Rela	4.730345816689199E-4
p50:RelA-p65 → SELE	Nfkb1, Rela	4.730345816689199E-4
IFNalpha/beta pathway	Stat3, Stat5a, Stat5b	6.440779144960161E-4
fMLP → NFkappaB	Nfkb1, Rela	6.600749950470129E-4
IL-2 → STAT5	Stat5a, Stat5b	6.600749950470129E-4
IFNalpha, IFNbeta → STAT5	Stat5a, Stat5b	6.600749950470129E-4
LXR network	Nfkb1, Rela	6.600749950470129E-4
IL-2 - STAT5 pathway	Stat5a, Stat5b	8.772117434506635E-4
cPKC —CARD9→TRAF6	Nfkb1, Rela	8.772117434506635E-4
mannan, Dectin2	Nfkb1, Rela	8.772117434506635E-4
EDA-A2 —TRAF3→ p50:RelA-p65	Nfkb1, Rela	0.0011241425642107344
EDA-A1 → p50:RelA-p65	Nfkb1,Rela	0.0011241425642107344
IL-1 pathway	Elk1, Nfkb1, Rela	0.0012748952830245175
neurotrophic signaling	Elk1, Nfkb1, Rela, Trp53	0.0012979014272467505
NGF —p75NTR→ p50:RelA-p65	Nfkb1, Rela	0.001400567221887963
CH000000333	Nfkb1, Rela	0.0017061874975544628
EDAR → NF-kappaB	Nfkb1, Rela	0.0017061874975544628
TNF-alpha → p50:RelA-p65	Nfkb1, Rela	0.0024038320457071337
PDGF pathway	Stat3, Stat5a, Stat5b	0.004038573581262634
TBK1:TRIF:IKK-i → p50:RelA	Nfkb1, Rela	0.004136568270131099
dsRNA → p50:RelA	Nfkb1, Rela	0.004136568270131099
RANKL → p38	Nfkb1, Rela	0.004638374899213325
LAT → p50:RelA	Nfkb1, Rela	0.005722351182439321
EDAR pathway	Nfkb1, Rela	0.009597224599851443
T-cell antigen receptor pathway	Elk1, Nfkb1, Rela	0.009985935589865625
LPS → NF-kappaB	Nfkb1, Rela	0.011871624770568048
NF-kappaB → genes encoding endothelial adhesion molecules	Nfkb1,Rela	0.011871624770568048
Epo pathway	Stat5a, Stat5b	0.012677905293747096
TLR9 pathway	Nfkb1, Rela	0.012677905293747096
IL-1beta → p50:RelA	Nfkb1, Rela	0.01436112389208374
TLR3 pathway	Nfkb1, Rela	0.018969411877391994
TNFR1 signaling	Nfkb1, Rela	0.019957669453188952
diacyl lipopeptide, TLR2	Nfkb1, Rela	0.019957669453188952
p38 pathway	Stat3, Trp53	0.029796036656231952
PRL pathway	Stat5a, Stat5b	0.03343465884776058
p50:RelA-p65 → IL8	Nfkb1, Rela	0.03343465884776058
IL-3 signaling	Stat5a, Stat5b	0.03343465884776058
TLR4 pathway	Nfkb1, Rela	0.037242517103675384
TLR2-mediated signaling	Nfkb1, Rela	0.04394876054564345

TABLE 9 | Top three master regulators for three TF sets: Intersection-specific TFs of the two cell lines, 1638N-T1-specific TFs and CMT-93-specific TFs.

Rank	Intersection set	1638N-T1-specific set	CMT-93-specific set
1	Rad23A	MLK3	Aebp1 (ACLP)
2	Smad3	TBK1	Il2rg (gamma-c)
3	Melk	Siah2	Mapk1 (ERK2)

associated with metastasis in CRC, an outcome that is thought to be dependent on chronic inflammation, e.g., triggered by bacterial infection (Zhu et al., 1998; Maggio-Price et al., 2006).

The third master regulator maternal embryonic leucine zipper kinase (Melk) is a known embryonic and neural stem cell marker and belongs to the family of serine/threonine kinases (Choi and Ku, 2011). Melk is normally expressed in cells that undergo proliferation during embryonic development, however, elevated expression has been particularly observed in variety of different cancer cell types including colorectal cancer (Gray et al., 2005; Badouel et al., 2010; Ganguly et al., 2015). Moreover, it has been shown that Melk knockdown decreases proliferation and tumor growth in CRC and, thus, it has been proposed to use Melk as a therapeutic target for cancer (Gray et al., 2005).

The merged master regulatory network consisted of 155 nodes (**Figure 2**, **Supplementary Table S7** and **Supplementary Figure S7**). The master regulators Rad23A and Smad3 were found most upstream in the hierarchy of the network. Rad23A was connected via the nodes p300 and CBP to the other nodes in the network, whereas Smad3 was connected to a variety of nodes which also included important cancer-associated TFs such as c-Myc, Runt-related factors, and Smad factors. Likewise, the master regulator Melk featured cascades through several molecules including Smad factors and p53 (see **Figure 2** for more details).

3.5.2. Prediction of Master Regulators and Construction of a Master Regulator Network Based on the 1638N-T1-Specific TF Set

The master regulator analysis detected Mlk3, Tbk1 and Siah2, which reach 28, 22, and 37 TFs from the 1638N-T1-specific set, respectively. The first master regulator MLK3 is a serine/threonine kinase that activates p38 MAP kinase, ERK, and JNK signaling pathways (Velho et al., 2014). MLK3-mediated activation has been shown to promote invasion and metastasis in several cancer types, including breast and gastric cancers (Chen et al., 2010; Mishra et al., 2010; Chen and Gallo, 2012; Cronan et al., 2012). Moreover, it has been proposed that mutant MLK3 is involved in the deregulation of several important CRC-associated signaling pathways such as WNT, MAPK, NOTCH, TGF-β, and P53 (Velho et al., 2014). Concerning Wnt signaling pathways in MLK3 mutant cells, it has been shown that components of the canonical Wnt pathway were found to be downregulated, while components of the non-canonical planar cell polarity (PCP) pathway were found to be upregulated.

The proposed master regulator TBK1 is a member of the non-canonical IκB protein kinases which is involved in the activation of IRF3 and c-Rel and NF-κB in cancer. The role of TBK1 is poorly investigated in CRC. However, several studies associated TBK1 with malignant transformation, cell growth and proliferation (Chien et al., 2006; Kim et al., 2013a,b).

The third master regulator Siah2 is an E3 ubiquitin ligase that regulates the degradation of a variety of substrates such as the nuclear corepressor (N-CoR), TRAF2, 2-oxoglutarate dehydrogenase-complex protein E2 (OGDC-E2), TIEG, and β-catenin (Zhang et al., 1998; Matsuzawa and Reed, 2001; Habelhah et al., 2002, 2004; Johnsen et al., 2002). Siah2 has been implicated in MAPK signaling, mitochondrial dynamics and cell survival (Nakayama et al., 2009; Kim et al., 2011). In addition, several studies have indicated that Siah2 functions as a proto-oncogene, while the Siah1 isoform has been associated with tumor suppressor activity (Wong and Möller, 2013; Gopalsamy et al., 2014). Although its role in CRC remains unclear, Siah2 has been suggested to promote invasion and metastasis in a variety of other cancers, including prostate, breast and liver (Qi et al., 2010, 2013; Behling et al., 2011; Malz et al., 2012; Sarkar et al., 2012; Wong et al., 2012; Gopalsamy et al., 2014).

The merged master regulatory network consisted of 52 nodes (**Figure 3**, **Supplementary Table S8** and **Supplementary Figure S8**). MLK3 and Siah2 were found most upstream in the hierarchy of the network, whereas TBK1 was found downstream of the network branch which is regulated by Siah2. MLK3 featured cascades through MKK3-isoform1, 4, and 6, and IKK-alpha-isoform1, and -beta. Siah2 was connected via the molecule alpha-synuclein-isoform1, Ubc5A, B, and C. TBK1 was connected via IRF3, 5, and 7, STAT6, and IKK-beta to its downstream nodes.

3.5.3. Prediction of Master Regulators and Construction of a Master Regulator Network Based on the CMT-93-Specific TF Set

For the CMT-93-specific TFs, the analysis reported the master regulators Aebp1 (ACLP), Il2rg (gamma-c) and Mapk1 (ERK2), which reach 43, 36, and 31 TFs from the set, respectively. The first proposed master regulator, Aebp1, is known to act as a transcriptional repressor in adipogenesis (Ladha et al., 2012). Aebp1 is upregulated in the majority of the primary glioblastoma multiforme (GBM) and loss of Aebp1 function was shown to result in apoptosis (Ladha et al., 2012). Moreover, Aebp1 induces NF-κB activity which leads to macrophage inflammatory responsiveness and affects tumor cell growth and survival (Majdalawieh et al., 2007). In the context of breast cancer tumorigenesis, Aebp1 has been suggested to be involved in the regulation of the cross-talk between mammary epithelium and stroma (Holloway et al., 2012). To this date, the role of Aebp1 remains largely unclear in CRC.

The second master regulator corresponded to the interleukin 2 receptor subunit gamma (Il2rg/gamma-c) which heterodimerizes with several interleukin receptors, including receptors for the interleukins −2, −4, −7, −9, −15, and −21 (Nata et al., 2015). Interleukins receptor signaling pathways are known to play crucial roles in inflammation-dependent progression and anti-tumor responses in CRC (West et al., 2015).

The last master regulator Mapk1 (ERK2) belongs to the MAP-kinases, which regulate cell growth, differentiation, proliferation, migration, and apoptosis (Santarpia et al., 2012). MAPKs act

downstream of several growth-factor receptors such as Egfr, which are often found overexpressed and activated in CRC (Fang and Richardson, 2005). Thus, it has been stated that the ERK MAPK pathway plays a central role in the progression of CRC (Fang and Richardson, 2005). In addition, it has been proposed that this pathway but not the JNK pathway or the p38 MAPK pathway is the key regulator of cell proliferation in CRC (Fang and Richardson, 2005).

The merged master regulatory network was composed of 65 nodes (**Figure 4**, **Supplementary Table S9** and **Supplementary Figure S9**). ACLP (Aebp1) and Il2rg (gamma-c) were found to be the regulators most upstream in the network. ACLP (Aebp1) was connected via the nodes ERK1 and TNF-alpha to the other nodes in the network. The master regulator Il2rg (gamma-c) featured a cascade through Jak3-isoform1, whereas the master regulator Mapk1 (ERK2) was connected to several molecules and TF families, including SREBP factors, STAT factors and Ets-like factors (see **Figure 4** for more details).

3.5.4. A Comparison with Randomly Selected Gene Sets

To test the prediction quality of our results and, whether they are specific for CRC, we performed a comparison between our results and those found for randomly drawn gene sets. Thus, we first randomly selected 10 gene sets, each of which had the same sample size as the signature genes analyzed in this study. After that, each random gene set was analyzed in the same way as both signature gene sets. In this regard, we started with TFBS enrichment analyses (see Section 2.4.1) for the detection of enriched TFBSs in the promoter regions of each random gene set. After retrieving the corresponding TFs, we observed that 17 TFs were common to each of the 10 random gene sets. Interestingly, 13 out of these 17 TFs were also detected based on both CRC signature gene sets (see Section 3.4.1). To determine their potential role in the context of our results, we further searched for overrepresented TRANSPATH pathways and master regulators based on these 13 TFs (see Section 2.4.2 and 2.4.3). The results of these analyses showed that there were no overrepresented pathways and, beyond that, the master regulators were completely different from those presented in Section 3.5.1, 3.5.2, and 3.5.3. Finally, we searched for overrepresented TRANSPATH pathways based on each random gene set (see Section 2.3.2). As expected, the overrepresented pathways found for each random gene set were completely different among themselves and, thus, they have no overlap with the pathways presented in the Section 3.3.1 and 3.3.2.

4. DISCUSSION

In this study, we specifically focused on revealing the similarities and differences with respect to the transcriptional regulation as well as the pathway repertoire of two distinct colorectal cancer (CRC) cell lines, namely 1638N-T1 and CMT-93, in a direct comparison. Based on signature genes that are most significantly upregulated in cancer cell type I and cancer cell type II, respectively, our approach aimed to identify the upstream transcriptional regulators and their regulatory networks.

Our results indicated that many of the pathways, which were identified based on the signature genes, can be linked to both pro-tumor as well as anti-tumor properties. In particular, we found pathways for 1638N-T1 which play a role in the detoxification of carcinogens, immune response, and apoptosis. Additionally, we found pathways which can be linked to oxidative stress, inflammation, cell migration, proliferation and survival. Oxidative stress is one important environmental factor in cancer as it is genotoxic and contributes to mutations (Beckman and Ames, 1997). During tumor progression, cells harbor mutations that reduce growth-limiting effects in pathways such as TGF-β signaling which becomes a tumor-promoting pathway due to mutations in later stages of CRC (Jakowlew, 2006; Bellam and Pasche, 2010; Calon et al., 2012). Therefore, it is likely that the results include many putative anti-tumor pathways that contain mutations in the cell lines, which is an important aspect to be addressed in future investigations.

On the level of transcriptional regulation, we identified a number of well-known, cancer-associated TFs with significantly enriched binding sites in the promoter regions of the signature genes. These TFs belong to a variety of TF families/subfamilies and are known to form protein-protein interactions with each other such as Jun factors and Fos factors which form the heterodimeric AP-1 protein (Chen et al., 1996; Shaulian and Karin, 2002; Eferl and Wagner, 2003). Likewise, nuclear receptors (NRs) of the subfamilies vitamin D receptors (NR1I) and retinoid X receptors form the VDR-RXR heterodimer complex (Orlov et al., 2012) that has been implicated in anticancer therapeutics (Friedrich et al., 2002; Sepulveda et al., 2006; Deeb et al., 2007; Matsuda and Kitagishi, 2013). In this light, it is known that TFs do not regulate their target genes in solitude, but interact with other TFs and cofactors in specific combinations for a fine-tuned control of gene expression (Gerstein et al., 2012). In addition, we identified different TF families/subfamilies that have overlapping binding sites and may act in a synergistic, additive, or antagonistic fashion in cancer. Kittler et al. revealed binding redundancy for NRs and their putative cooperating TFs in breast cancer on the basis of 39 factors, whereas non-overlapping binding sites were found to occur rarely (Kittler et al., 2013). Taken together, although the signature genes of both cell lines show no overlap, they may still be regulated by common factors in CRC.

We revealed that 62 and 72% of the TFs for 1638N-T1 and CMT-93, respectively, were found in the intersection of both cell lines. Consequently, only 38 and 28% of the TFs were exclusive for 1638N-T1 and CMT-93, respectively, whose implications in signal transduction pathways might explain phenotypic differences between the two cell lines with regard to tumor growth and metastasis. We deduced cross-talks between several pathways that might have an impact on tumor progression in the cell lines. For the APC-deficient 1638N-T1 cell line, we found overrepresented pathways which related to the activation of the canonical Wnt signaling pathway (**Tables 1**, **6**). Wnt signaling activity is known to contribute to tumor aggressiveness; therefore, it is often targeted in cancer therapy (Anastas and Moon, 2013; Loh et al., 2013). It has also been stated that enhancement of canonical Wnt signaling activity is required

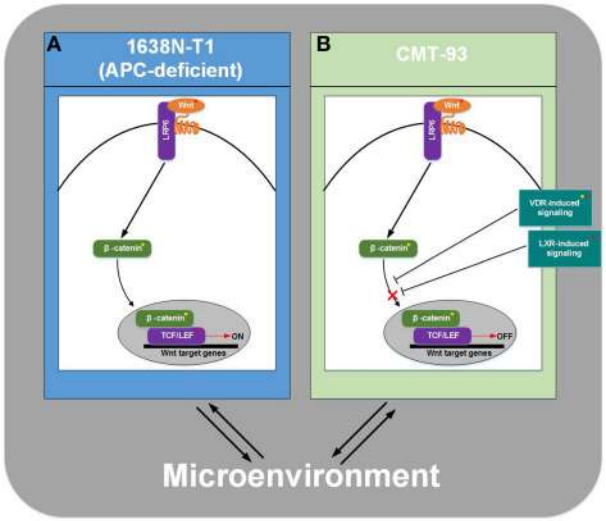


FIGURE 5 | Schema for potential state of canonical Wnt signaling pathway in mouse models. (A) Wnt signaling is activated in 1638N-T1. **(B)** Wnt signaling is inhibited in CMT-93 through cross-talks with VDR- and/or LXR-induced pathways. Interaction of tumor cells with the microenvironment has an impact on cell proliferation, invasion, and metastasis in mouse models. Signature genes and transcription factors/cofactors, whose binding sites were found to be enriched in promoters, are indicated by a red asterisk or a yellow asterisk, respectively.

for tumor progression and metastasis (Oguma et al., 2008). On the other hand, we showed several pathways for CMT-93 which have been previously associated with an inhibition of Wnt signaling. Two of these pathways related to VDR signaling and LXR-induced signaling (**Tables 4**, **8**). Strikingly, VDR and LXRα (Nr1h3) were included in the signature genes for CMT-93 (see **Supplementary Table S2**), and VDR also showed significantly enriched binding sites (see **Supplementary Table S5**). Previous studies have investigated the activation of VDR as well as LXR in APC-deficient mice and observed that the activity of both factors decreased tumor growth (Zheng et al., 2012; Sasso et al., 2013). In addition, LXR expression was found to be downregulated in colon tumors of APC-deficient mice compared with adjacent normal mucosa (Su et al., 1992; Sasso et al., 2013). We also found that CTNNB1, which encodes β-catenin, was not included in the signature genes of any of the two cell lines, but showed significant binding site enrichment (see **Supplementary Table S6**). With respect to TCF-7-related factors, the genes Tcf7l1 and Lef1, however, were included in the signature genes of 1638N-T1. Interestingly, VDR and LXR can both directly bind to β-catenin, thereby preventing β-catenin from binding to its target sites (Uno et al., 2009; Makoukji et al., 2011; Zheng et al., 2012; Larriba et al., 2013; Shackleford et al., 2013; Lim et al., 2014).

All things considered, supported by the knowledge that 1638N-T1 cells harbor a mutation in the APC gene, which leads to aberrant Wnt pathway activation: we suggest that Wnt signaling is activated in 1638N-T1, but inhibited in CMT-93 through cross-talks of canonical Wnt signaling with VDR signaling pathway and/or LXR-related pathways. Consequently, we suggest that Wnt signaling-driven tumor formation and growth should be increased in mouse models involving 1638N-T1 compared to ones involving CMT-93. Though, many additional factors have to be taken into account when monitoring cell proliferation, invasion, and metastasis in mouse models. Several previous studies indicated synergistic effects between K-Ras and canonical Wnt signaling harboring APC mutations in CRC (Janssen et al., 2006; Luo et al., 2009; Lemieux et al., 2015). Furthermore, during development of effective cancer therapies, tumor cells grown *in vitro* are transplanted into ectopic sites of immunocompromized mice that do not reject tumor cells (Sharpless and Depinho, 2006; Richmond and Su, 2008; Hung et al., 2010). It has been stated that these xenograft models may fail to recapitulate the heterogeneity of cancer and the microenvironment, i.e., the interaction between tumor cells and supporting stroma (Hung et al., 2010). In the end, regardless of the fact that Wnt signaling may be aberrantly activated in 1638N-T1, a variety of different factors have an impact on the capacity of tumor cells to grow, proliferate, and metastasize in mouse models. We summarized our observations concerning the potential state of canonical Wnt signaling in the cell lines (**Figure 5**).

The master regulator analyses revealed several potential candidates which might be useful as therapeutic targets for cancer therapy. Master regulators were inferred from a network model that explicitly displayed the regulatory cascades between TFs. Beside several master regulators with yet unknown roles in CRC, we found MLK3 and Mapk1 (ERK2) which might be important in cancer cell proliferation, invasion, and metastasis of 1638N-T1 and CMT-93, respectively. Above all, our master regulatory networks can be used as models to generate testable hypotheses for studying the phenotypic differences between 1638N-T1 and CMT-93.

5. CONCLUSION

In this study, we have presented a systematic approach which combines colorectal cancer (CRC) cell lines, namely 1638N-T1 and CMT-93, and well-established computational methods in order to compare these cell lines on the level of transcriptional regulation as well as on a pathway level, i.e., the cancer cell-intrinsic pathway repertoire. We used the Trinity platform and the geneXplain platform to identify significantly upregulated genes in each of the cell lines as well as their upstream transcriptional regulators, on the basis of which we generated regulatory networks. Our findings suggested that the Wnt signaling pathway is activated in 1638N-T1, but inhibited in CMT-93 cells through cross-talks with other pathways. Moreover, we identified a number of well-known, cancer-associated TFs for both cell lines and provided indication of several master regulators being present such as MLK3 and Mapk1 (ERK2) which might be important in cell proliferation, migration, and invasion of 1638N-T1 and CMT-93, respectively. Using our systematic approach, we have provided new insights into the invasive potential of individual CRC cell lines, which can be used for development of effective cancer therapy.

AUTHOR CONTRIBUTIONS

DW and MG participated in the design of the study, conducted computational, and statistical analyses. EW supervised the computational and statistical analyses. MH and TB interpreted the results with DW. JA prepared the colorectal cancer cell lines for this study. AW processed the RNA-Seq data (FASTQ files) and prepared the RNA-Seq count data. AB was involved in the preparation of the cell lines and interpretation of the results in perspective of colorectal cancer biology. DW and MG conceived of and managed the project and wrote the final version of the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Ron Smits for providing the murine colorectal cancer cell line 1638N-T1 as well as the Transcriptome and Genome Analysis Laboratory (TAL) in Göttingen for the preparation of the RNA-Seq data (FASTQ files). We would also like to thank Cornelia Meckbach for proofreading the manuscript. This work was supported by the ebio initiative of the German Ministry of Education and Research (BMBF) and DW was funded by the MetastaSys project (0316173A) within the ebio initiative. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00042

Supplementary Table S1 | Signature genes for colorectal cell line 1638N-T1.

Supplementary Table S2 | Signature genes for colorectal cell line CMT-93.

Supplementary Table S3 | Colorectal cancer-related non-redundant PWM library.

Supplementary Table S4 | 1638N-T1-specific TF set.

Supplementary Table S5 | CMT-93-specific TF set.

Supplementary Table S6 | Intersection-specific TF set between 1638N-T1 and CMT-93.

Supplementary Table S7 | Master regulatory network based on the intersection-specific TF set in Pair Graph File format.

Supplementary Figure S7 | Master regulatory network based on the intersection-specific TF set as Scalable Vector Graphics (SVG).

Supplementary Table S8 | Master regulatory network based on the 1638NT-1-specific TF set in Pair Graph File format.

Supplementary Figure S8 | Master regulatory network based on the 1638NT-1-specific TF set as Scalable Vector Graphics (SVG).

Supplementary Table S9 | Master regulatory network based on the CMT-93-specific TF set in Pair Graph File format.

Supplementary Figure S9 | Master regulatory network based on the CMT-93-specific TF set as Scalable Vector Graphics (SVG).

REFERENCES

Alsayed, Y., Uddin, S., Ahmad, S., Majchrzak, B., Druker, B. J., Fish, E. N., et al. (2000). IFN-γ activates the C3G/Rap1 signaling pathway. *J. Immunol.* 164, 1800–1806. doi: 10.4049/jimmunol.164.4.1800

Ameyar, M., Wisniewska, M., and Weitzman, J. B. (2003). A role for AP-1 in apoptosis: the case for and against. *Biochimie* 85, 747–752. doi: 10.1016/j.biochi.2003.09.006

Anagnostopoulou, V., Pediaditakis, I., Alkahtani, S., Alarifi, S. A., Schmidt, E.-M., Lang, F., et al. (2013). Differential effects of dehydroepiandrosterone and testosterone in prostate and colon cancer cell apoptosis: the role of nerve growth factor (NGF) receptors. *Endocrinology* 154, 2446–2456. doi: 10.1210/en.2012-2249

Anastas, J. N., and Moon, R. T. (2013). WNT signalling pathways as therapeutic targets in cancer. *Nat. Rev. Cancer* 13, 11–26. doi: 10.1038/nrc3419

Andersen, V., Holst, R., Kopp, T. I., Tjønneland, A., and Vogel, U. (2013). Interactions between diet, lifestyle and IL10, IL1B, and PTGS2/COX-2 gene polymorphisms in relation to risk of colorectal cancer in a prospective Danish case-cohort study. *PLoS ONE* 8:e78366. doi: 10.1371/journal.pone.0078366

Asham, E., Shankar, A., Loizidou, M., Fredericks, S., Miller, K., Boulos, P. B., et al. (2001). Increased endothelin-1 in colorectal cancer and reduction of tumour growth by ET(A) receptor antagonism. *Br. J. Cancer* 85, 1759–1763. doi: 10.1054/bjoc.2001.2193

Augustin, H. G., Koh, G. Y., Thurston, G., and Alitalo, K. (2009). Control of vascular morphogenesis and homeostasis through the angiopoietin-Tie system. *Nat. Rev. Mol. Cell Biol.* 10, 165–177. doi: 10.1038/nrm2639

Badouel, C., Chartrain, I., Blot, J., and Tassan, J.-P. (2010). Maternal embryonic leucine zipper kinase is stabilized in mitosis by phosphorylation and is partially degraded upon mitotic exit. *Exp. Cell Res.* 316, 2166–2173. doi: 10.1016/j.yexcr.2010.04.019

Basith, S., Manavalan, B., Yoo, T. H., Kim, S. G., and Choi, S. (2012). Roles of toll-like receptors in cancer: a double-edged sword for defense and offense. *Arch. Pharm. Res.* 35, 1297–1316. doi: 10.1007/s12272-012-0802-7

Bassuk, A. G., and Leiden, J. M. (1997). The role of Ets transcription factors in the development and function of the mammalian immune system. *Adv. Immunol.* 64, 65–104. doi: 10.1016/S0065-2776(08)60887-1

Bastide, P., Darido, C., Pannequin, J., Kist, R., Robine, S., Marty-Double, C., et al. (2007). Sox9 regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. *J. Cell Biol.* 178, 635–648. doi: 10.1083/jcb.200704152

Beckman, K. B., and Ames, B. N. (1997). Oxidative decay of DNA. *J. Biol. Chem.* 272, 19633–19636. doi: 10.1074/jbc.272.32.19633

Behling, K. C., Tang, A., Freydin, B., Chervoneva, I., Kadakia, S., Schwartz, G. F., et al. (2011). Increased SIAH expression predicts ductal carcinoma *in situ* (DCIS) progression to invasive carcinoma. *Breast Cancer Res. Treat.* 129, 717–724. doi: 10.1007/s10549-010-1254-8

Bellam, N., and Pasche, B. (2010). TGF-β signaling alterations and colon cancer. *Cancer Treat. Res.* 155, 85–103. doi: 10.1007/978-1-4419-6033-7_5

Belmont, L., Rabbe, N., Antoine, M., Cathelin, D., Guignabert, C., Kurie, J., et al. (2014). Expression of TLR9 in tumor-infiltrating mononuclear cells enhances angiogenesis and is associated with a worse survival in lung cancer. *Int. J. Cancer* 134, 765–777. doi: 10.1002/ijc.28413

Blache, P., van de Wetering, M., Duluc, I., Domon, C., Berta, P., Freund, J.-N., et al. (2004). SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. *J. Cell Biol.* 166, 37–47. doi: 10.1083/jcb.200311021

Bouker, K. B., Skaar, T. C., Riggins, R. B., Harburger, D. S., Fernandez, D. R., et al. (2005). Interferon regulatory factor-1 (IRF-1) exhibits tumor suppressor activities in breast cancer associated with caspase activation and induction of apoptosis. *Carcinogenesis* 26, 1527–1535. doi: 10.1093/carcin/bgi113

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V. F., Iglesias, M., Céspedes, M. V., et al. (2012). Dependency of colorectal cancer on a TGF-β-driven

program in stromal cells for metastasis initiation. *Cancer Cell* 22, 571–584. doi: 10.1016/j.ccr.2012.08.013

Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Casaletto, J. B., and McClatchey, A. I. (2012). Spatial regulation of receptor tyrosine kinases in development and cancer. *Nat. Rev. Cancer* 12, 387–400. doi: 10.1038/nrc3277

Chen, J., and Gallo, K. A. (2012). MLK3 regulates paxillin phosphorylation in chemokine-mediated breast cancer cell migration and invasion to drive metastasis. *Cancer Res.* 72, 4130–4140. doi: 10.1158/0008-5472.CAN-12-0655

Chen, J., Miller, E. M., and Gallo, K. A. (2010). MLK3 is critical for breast cancer cell migration and promotes a malignant phenotype in mammary epithelial cells. *Oncogene* 29, 4399–4411. doi: 10.1038/onc.2010.198

Chen, T. K., Smith, L. M., Gebhardt, D. K., Birrer, M. J., and Brown, P. H. (1996). Activation and inhibition of the AP-1 complex in human breast cancer cells. *Mol. Carcinog.* 15, 215–226.

Chien, Y., Kim, S., Bumeister, R., Loo, Y.-M., Kwon, S. W., Johnson, C. L., et al. (2006). RalB GTPase-mediated activation of the IκB family kinase TBK1 couples innate immune signaling to tumor cell survival. *Cell* 127, 157–170. doi: 10.1016/j.cell.2006.08.034

Choi, S., and Ku, J.-L. (2011). Resistance of colorectal cancer cells to radiation and 5-FU is associated with MELK expression. *Biochem. Biophys. Res. Commun.* 412, 207–213. doi: 10.1016/j.bbrc.2011.07.060

Cooley, D. A., Frazier, O. H., and Kahan, B. D. (1982). Cardiac transplantation with the use of cyclosporin a for immunologic suppression. *Tex Heart Inst. J.* 9, 247–251.

Cowin, P., Rowlands, T. M., and Hatsell, S. J. (2005). Cadherins and catenins in breast cancer. *Curr. Opin. Cell Biol.* 17, 499–508. doi: 10.1016/j.ceb.2005.08.014

Cronan, M. R., Nakamura, K., Johnson, N. L., Granger, D. A., Cuevas, B. D., Wang, J.-G., et al. (2012). Defining MAP3 kinases required for MDA-MB-231 cell tumor growth and metastasis. *Oncogene* 31, 3889–3900. doi: 10.1038/onc.2011.544

da Costa, L. T., He, T. C., Yu, J., Sparks, A. B., Morin, P. J., Polyak, K., et al. (1999). CDX2 is mutated in a colorectal cancer with normal APC/β-catenin signaling. *Oncogene* 18, 5010–5014. doi: 10.1038/sj.onc.1202872

Dai, X., Ge, J., Wang, X., Qian, X., Zhang, C., and Li, X. (2013). OCT4 regulates epithelial-mesenchymal transition and its knockdown inhibits colorectal cancer cell migration and invasion. *Oncol. Rep.* 29, 155–160. doi: 10.3892/or.2012.2086

Deeb, K. K., Trump, D. L., and Johnson, C. S. (2007). Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat. Rev. Cancer* 7, 684–700. doi: 10.1038/nrc2196

Eferl, R., and Wagner, E. F. (2003). AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer* 3, 859–868. doi: 10.1038/nrc1209

Eger, A., Stockinger, A., Park, J., Langkopf, E., Mikula, M., Gotzmann, J., et al. (2004). β-Catenin and TGFβ signalling cooperate to maintain a mesenchymal phenotype after FosER-induced epithelial to mesenchymal transition. *Oncogene* 23, 2672–2680. doi: 10.1038/sj.onc.1207416

Fagiani, E., and Christofori, G. (2013). Angiopoietins in angiogenesis. *Cancer Lett.* 328, 18–26. doi: 10.1016/j.canlet.2012.08.018

Fang, J. Y., and Richardson, B. C. (2005). The MAPK signalling pathways and colorectal cancer. *Lancet Oncol.* 6, 322–327. doi: 10.1016/S1470-2045(05)70168-6

Fleming, N. I., Jorissen, R. N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., et al. (2013). SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res.* 73, 725–735. doi: 10.1158/0008-5472.CAN-12-2706

Freund, J. N., Domon-Dell, C., Kedinger, M., and Duluc, I. (1998). The Cdx-1 and Cdx-2 homeobox genes in the intestine. *Biochem. Cell Biol.* 76, 957–969. doi: 10.1139/bcb-76-6-957

Friedrich, M., Axt-Fliedner, R., Villena-Heinsen, C., Tilgen, W., Schmidt, W., and Reichrath, J. (2002). Analysis of vitamin D-receptor (VDR) and retinoid X-receptor alpha in breast cancer. *Histochem. J.* 34, 35–40. doi: 10.1023/A:1021343825552

Gambara, G., Desideri, M., Stoppacciaro, A., Padula, F., Cesaris, P. D., Starace, D., et al. (2015). TLR3 engagement induces IRF-3-dependent apoptosis in androgen-sensitive prostate cancer cells and inhibits tumour growth *in vivo*. *J. Cell Mol. Med.* 19, 327–339. doi: 10.1111/jcmm.12379

Ganguly, R., Mohyeldin, A., Thiel, J., Kornblum, H. I., Beullens, M., and Nakano, I. (2015). MELK-a conserved kinase: functions, signaling, cancer, and controversy. *Clin. Transl. Med.* 4:11. doi: 10.1186/s40169-014-0045-y

Gao, F., Liang, B., Reddy, S. T., Farias-Eisner, R., and Su, X. (2014). Role of inflammation-associated microenvironment in tumorigenesis and metastasis. *Curr. Cancer Drug Targets* 14, 30–45. doi: 10.2174/1568009613136660107

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from encode data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Giardina, C., Nakanishi, M., Khan, A., Kuratnik, A., Xu, W., Brenner, B., et al. (2015). Regulation of VDR Expression in Apc-Mutant Mice, Human Colon Cancers and Adenomas. *Cancer Prev. Res.* 8, 387–399. doi: 10.1158/1940-6207.CAPR-14-0371

Giessen, C., Laubender, R. P., von Weikersthal, L. F., Schalhorn, A., Modest, D. P., Stintzing, S., et al. (2013). Early tumor shrinkage in metastatic colorectal cancer: retrospective analysis from an irinotecan-based randomized first-line trial. *Cancer Sci.* 104, 718–724. doi: 10.1111/cas.12148

Giuliani, L., Ciotti, M., Stoppacciaro, A., Pasquini, A., Silvestri, I., Matteis, A. D., et al. (2005). Udp-glucuronosyltransferases 1A expression in human urinary bladder and colon cancer by immunohistochemistry. *Oncol. Rep.* 13, 185–191. doi: 10.3892/or.13.2.185

Gopalsamy, A., Hagen, T., and Swaminathan, K. (2014). Investigating the molecular basis of Siah1 and Siah2 E3 ubiquitin ligase substrate specificity. *PLoS ONE* 9:e106547. doi: 10.1371/journal.pone.0106547

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grant, K., Knowles, J., Dawas, K., Burnstock, G., Taylor, I., and Loizidou, M. (2007). Mechanisms of endothelin 1-stimulated proliferation in colorectal cancer cell lines. *Br. J. Surg.* 94, 106–112. doi: 10.1002/bjs.5536

Graves, B. J., and Petersen, J. M. (1998). Specificity within the ets family of transcription factors. *Adv. Cancer Res.* 75, 1–55. doi: 10.1016/S0065-230X(08)60738-1

Gray, D., Jubb, A. M., Hogue, D., Dowd, P., Kljavin, N., Yi, S., et al. (2005). Maternal embryonic leucine zipper kinase/murine protein serine-threonine kinase 38 is a promising therapeutic target for multiple cancers. *Cancer Res.* 65, 9751–9761. doi: 10.1158/0008-5472.CAN-04-4531

Guo, R., Sakamoto, H., Sugiura, S., and Ogawa, M. (2007). Endothelial cell motility is compatible with junctional integrity. *J. Cell Physiol.* 211, 327–335. doi: 10.1002/jcp.20937

Guo, R.-J., Huang, E., Ezaki, T., Patel, N., Sinclair, K., Wu, J., et al. (2004). Cdx1 inhibits human colon cancer cell proliferation by reducing β-catenin/T-cell factor transcriptional activity. *J. Biol. Chem.* 279, 36865–36875. doi: 10.1074/jbc.M405213200

Gupta, G. P., and Massagué, J. (2006). Cancer metastasis: building a framework. *Cell* 127, 679–695. doi: 10.1016/j.cell.2006.11.001

Habelhah, H., Frew, I. J., Laine, A., Janes, P. W., Relaix, F., Sassoon, D., et al. (2002). Stress-induced decrease in TRAF2 stability is mediated by Siah2. *EMBO J.* 21, 5756–5765. doi: 10.1093/emboj/cdf576

Habelhah, H., Laine, A., Erdjument-Bromage, H., Tempst, P., Gershwin, M. E., Bowtell, D. D. L., et al. (2004). Regulation of 2-oxoglutarate (α-ketoglutarate) dehydrogenase stability by the RING finger ubiquitin ligase Siah. *J. Biol. Chem.* 279, 53782–53788. doi: 10.1074/jbc.M410315200

Holash, J., Maisonpierre, P. C., Compton, D., Boland, P., Alexander, C. R., Zagzag, D., et al. (1999a). Vessel cooption, regression, and growth in tumors mediated by angiopoietins and VEGF. *Science* 284, 1994–1998. doi: 10.1126/science.284.5422.1994

Holash, J., Wiegand, S. J., and Yancopoulos, G. D. (1999b). New model of tumor angiogenesis: dynamic balance between vessel regression and growth mediated by angiopoietins and VEGF. *Oncogene* 18, 5356–5362. doi: 10.1038/sj.onc.1203035

Holldack, J. (2014). Toll-like receptors as therapeutic targets for cancer. *Drug Discov. Today* 19, 379–382. doi: 10.1016/j.drudis.2013.08.020

Holloway, R. W., Bogachev, O., Bharadwaj, A. G., McCluskey, G. D., Majdalawieh, A. F., Zhang, L., et al. (2012). Stromal adipocyte enhancer-binding protein (AEBP1) promotes mammary epithelial cell hyperplasia via

proinflammatory and hedgehog signaling. *J. Biol. Chem.* 287, 39171–39181. doi: 10.1074/jbc.M112.404293

Hsu, R. Y. C., Chan, C. H. F., Spicer, J. D., Rousseau, M. C., Giannias, B., Rousseau, S., et al. (2011). LPS-induced TLR4 signaling in human colorectal cancer cells increases β1 integrin-mediated cell adhesion and liver metastasis. *Cancer Res.* 71, 1989–1998. doi: 10.1158/0008-5472.CAN-10-2833

Hu, G., and Barnes, B. J. (2006). Interferon regulatory factor-5-regulated pathways as a target for colorectal cancer therapeutics. *Exp. Rev. Anticancer. Ther.* 6, 775–784. doi: 10.1586/14737140.6.5.775

Hung, K. E., Maricevich, M. A., Richard, L. G., Chen, W. Y., Richardson, M. P., Kunin, A., et al. (2010). Development of a mouse model for sporadic and metastatic colon tumors and its use in assessing drug treatment. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1565–1570. doi: 10.1073/pnas.0908682107

Ijssennagger, N., Rijnierse, A., de Wit, N. J. W., Boekschoten, M. V., Dekker, J., Schonewille, A., et al. (2013). Dietary heme induces acute oxidative stress, but delayed cytotoxicity and compensatory hyperproliferation in mouse colon. *Carcinogenesis* 34, 1628–1635. doi: 10.1093/carcin/bgt084

Ikeda, H., Old, L. J., and Schreiber, R. D. (2002). The roles of IFNγ in protection against tumor development and cancer immunoediting. *Cytokine Growth Factor Rev.* 13, 95–109. doi: 10.1016/S1359-6101(01)00038-7

Jakowlew, S. B. (2006). Transforming growth factor-β in cancer and metastasis. *Cancer Metastasis Rev.* 25, 435–457. doi: 10.1007/s10555-006-9006-2

Janssen, K.-P., Alberici, P., Fsihi, H., Gaspar, C., Breukel, C., Franken, P., et al. (2006). APC and oncogenic KRAS are synergistic in enhancing Wnt signaling in intestinal tumor formation and progression. *Gastroenterology* 131, 1096–1109. doi: 10.1053/j.gastro.2006.08.011

Jansson, E. A., Are, A., Greicius, G., Kuo, I.-C., Kelly, D., Arulampalam, V., and et al. (2005). The Wnt/β-catenin signaling pathway targets PPARγ activity in colon cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1460–1465. doi: 10.1073/pnas.0405928102

Johnsen, S. A., Subramaniam, M., Monroe, D. G., Janknecht, R., and Spelsberg, T. C. (2002). Modulation of transforming growth factor β (TGFβ)/Smad transcriptional responses through targeted degradation of TGFβ-inducible early gene-1 by human seven in absentia homologue. *J. Biol. Chem.* 277, 30754–30759. doi: 10.1074/jbc.M204812200

Junghans, D., Haas, I. G., and Kemler, R. (2005). Mammalian cadherins and protocadherins: about cell death, synapses and processing. *Curr. Opin. Cell Biol.* 17, 446–452. doi: 10.1016/j.ceb.2005.08.008

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Kang, Y. (2005). Functional genomic analysis of cancer metastasis: biologic insights and clinical implications. *Exp. Rev. Mol. Diagn.* 5, 385–395. doi: 10.1586/14737159.5.3.385

Karim, B. O., and Huso, D. L. (2013). Mouse models for colorectal cancer. *Am. J. Cancer Res.* 3, 240–250.

Kim, H., Scimia, M. C., Wilkinson, D., Trelles, R. D., Wood, M. R., Bowtell, D., et al. (2011). Fine-tuning of Drp1/Fis1 availability by AKAP121/Siah2 regulates mitochondrial adaptation to hypoxia. *Mol. Cell* 44, 532–544. doi: 10.1016/j.molcel.2011.08.045

Kim, J.-Y., Beg, A. A., and Haura, E. B. (2013a). Non-canonical IKKs, IKKϵ and TBK1, as novel therapeutic targets in the treatment of non-small cell lung cancer. *Exp. Opin. Ther. Targets* 17, 1109–1112. doi: 10.1517/14728222.2013.833188

Kim, J.-Y., Welsh, E. A., Oguz, U., Fang, B., Bai, Y., Kinose, F., et al. (2013b). Dissection of TBK1 signaling via phosphoproteomics in lung cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12414–12419. doi: 10.1073/pnas.1220674110

Kim, K., Lu, Z., and Hay, E. D. (2002). Direct evidence for a role of β-catenin/LEF-1 signaling pathway in induction of EMT. *Cell Biol. Int.* 26, 463–476. doi: 10.1006/cbir.2002.0901

Kitadai, Y., Sasaki, T., Kuwai, T., Nakamura, T., Bucana, C. D., Hamilton, S. R., et al. (2006). Expression of activated platelet-derived growth factor receptor in stromal cells of human colon carcinomas is associated with metastatic potential. *Int. J. Cancer* 119, 2567–2574. doi: 10.1002/ijc.22229

Kittler, R., Zhou, J., Hua, S., Ma, L., Liu, Y., Pendleton, E., et al. (2013). A comprehensive nuclear receptor network for breast cancer cells. *Cell Rep.* 3, 538–551. doi: 10.1016/j.celrep.2013.01.004

Klampfer, L. (2008). The role of signal transducers and activators of transcription in colon cancer. *Front. Biosci.* 13, 2888–2899. doi: 10.2741/2893

Klampfer, L. (2014). Vitamin D and colon cancer. *World J. Gastrointest. Oncol.* 6, 430–437. doi: 10.4251/wjgo.v6.i11.430

Knowles, J. P., Shi-Wen, X., ul Haque, S., Bhalla, A., Dashwood, M. R., Yang, S., et al. (2012). Endothelin-1 stimulates colon cancer adjacent fibroblasts. *Int. J. Cancer* 130, 1264–1272. doi: 10.1002/ijc.26090

Kojima, M., Morisaki, T., Sasaki, N., Nakano, K., Mibu, R., Tanaka, M., et al. (2004). Increased nuclear factor-kB activation in human colorectal carcinoma and its correlation with tumor progression. *Anticancer Res.* 24, 675–681.

Korchynskyi, O., Landström, M., Stoika, R., Funa, K., Heldin, C. H., ten Dijke, P., et al. (1999). Expression of Smad proteins in human colorectal cancer. *Int. J. Cancer* 82, 197–202.

Krieg, A. M. (2008). Toll-like receptor 9 (TLR9) agonists in the treatment of cancer. *Oncogene* 27, 161–167. doi: 10.1038/sj.onc.1210911

Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., et al. (2006). TRANSPATH®: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* 34, D546–D551. doi: 10.1093/nar/gkj107

Ku, J.-L., Park, S.-H., Yoon, K.-A., Shin, Y.-K., Kim, K.-H., Choi, J.-S., et al. (2007). Genetic alterations of the TGF-β signaling pathway in colorectal cancer cell lines: A novel mutation in Smad3 associated with the inactivation of TGF-β-induced transcriptional activation. *Cancer Lett.* 247, 283–292. doi: 10.1016/j.canlet.2006.05.008

Kuai, W.-X., Wang, Q., Yang, X.-Z., Zhao, Y., Yu, R., and Tang, X.-J. (2012). Interleukin-8 associates with adhesion, migration, invasion and chemosensitivity of human gastric cancer cells. *World J. Gastroenterol.* 18, 979–985. doi: 10.3748/wjg.v18.i9.979

Ladha, J., Sinha, S., Bhat, V., Donakonda, S., and Rao, S. M. R. (2012). Identification of genomic targets of transcription factor AEBP1 and its role in survival of glioma cells. *Mol. Cancer Res.* 10, 1039–1051. doi: 10.1158/1541-7786.mcr-11-0488

Larriba, M. J., González-Sancho, J. M., Barbáchano, A., Niell, N., Ferrer-Mayorga, G., and Muñoz, A. (2013). Vitamin D is a multilevel repressor of Wnt/β-catenin signaling in cancer cells. *Cancers* 5, 1242–1260. doi: 10.3390/cancers5041242

Lemieux, E., Cagnol, S., Beaudry, K., Carrier, J., and Rivard, N. (2015). Oncogenic KRAS signalling promotes the Wnt/β-catenin pathway through LRP6 in colorectal cancer. *Oncogene* 34, 4914–4927. doi: 10.1038/onc.2014.416

Li, A., Varney, M. L., and Singh, R. K. (2001). Expression of interleukin 8 and its receptors in human colon carcinoma cells with different metastatic potentials. *Clin. Cancer Res.* 7, 3298–3304.

Li, M., Zhang, Y., Feurino, L. W., Wang, H., Fisher, W. E., Brunicardi, F. C., et al. (2008). Interleukin-8 increases vascular endothelial growth factor and neuropilin expression and stimulates ERK activation in human pancreatic cancer. *Cancer Sci.* 99, 733–737. doi: 10.1111/j.1349-7006.2008.00740.x

Li, T.-T., Ogino, S., and Qian, Z. R. (2014). Toll-like receptor signaling in colorectal cancer: Carcinogenesis to cancer therapy. *World J. Gastroenterol.* 20, 17699–17708. doi: 10.3748/wjg.v20.i47.17699

Libermann, T. A., and Zerbini, L. F. (2006). Targeting transcription factors for cancer gene therapy. *Curr. Gene Ther.* 6, 17–33. doi: 10.2174/156652306775515501

Licato, L. L., Keku, T. O., Wurzelmann, J. I., Murray, S. C., Woosley, J. T., Sandler, R. S., et al. (1997). *In vivo* activation of mitogen-activated protein kinases in rat intestinal neoplasia. *Gastroenterology* 113, 1589–1598. doi: 10.1053/gast.1997.v113.pm9352861

Lim, Y. Y., Kim, S. Y., Kim, H. M., Li, K. S., Kim, M. N., Park, K.-C., et al. (2014). Potential relationship between the canonical Wnt signalling pathway and expression of the vitamin D receptor in alopecia. *Clin. Exp. Dermatol.* 39, 368–375. doi: 10.1111/ced.12241

Lin, C.-Y., and Åke Gustafsson, J. (2015). Targeting liver x receptors in cancer therapeutics. *Nat. Rev. Cancer* 15, 216–224. doi: 10.1038/nrc3912

Lind, D. S., Hochwald, S. N., Malaty, J., Rekkas, S., Hebig, P., Mishra, G., et al. (2001). Nuclear factor-κB is upregulated in colorectal cancer. *Surgery* 130, 363–369. doi: 10.1067/msy.2001.116672

Lindholm, P. F., Bub, J., Kaul, S., Shidham, V. B., and Kajdacsy-Balla, A. (2000). The role of constitutive NF-κB activity in PC-3 human prostate cancer cell invasive behavior. *Clin. Exp. Metastasis* 18, 471–479. doi: 10.1023/A:1011845725394

Loh, Y. N., Hedditch, E. L., Baker, L. A., Jary, E., Ward, R. L., and Ford, C. E. (2013). The Wnt signalling pathway is upregulated in an *in vitro* model of acquired tamoxifen resistant breast cancer. *BMC Cancer* 13:174. doi: 10.1186/1471-2407-13-174

Luo, F., Brooks, D. G., Ye, H., Hamoudi, R., Poulogiannis, G., Patek, C. E., et al. (2009). Mutated K-ras(Asp12) promotes tumourigenesis in Apc(Min) mice more in the large than the small intestines, with synergistic effects between K-ras and Wnt pathways. *Int. J. Exp. Pathol.* 90, 558–574. doi: 10.1111/j.1365-2613.2009.00667.x

Lynch, J., Keller, M., Guo, R.-J., Yang, D., and Traber, P. (2003). Cdx1 inhibits the proliferation of human colon cancer cells by reducing cyclin D1 gene expression. *Oncogene* 22, 6395–6407. doi: 10.1038/sj.onc.1206770

Maggio-Price, L., Treuting, P., Zeng, W., Tsang, M., Bielefeldt-Ohmann, H., and Iritani, B. M. (2006). Helicobacter infection is required for inflammation and colon cancer in SMAD3-deficient mice. *Cancer Res.* 66, 828–838. doi: 10.1158/0008-5472.CAN-05-2448

Magnanti, M., Giuliani, L., Gandini, O., Gazzaniga, P., Santiemma, V., Ciotti, M., et al. (2000). Follicle-stimulating hormone, testosterone, and hypoxia differentially regulate UDP-glucuronosyltransferase 1 isoforms expression in rat sertoli and peritubular myoid cells. *J. Steroid. Biochem. Mol. Biol.* 74, 149–155. doi: 10.1016/S0960-0760(00)00095-9

Majdalawieh, A., Zhang, L., and Ro, H.-S. (2007). Adipocyte Enhancer-binding Protein-1 Promotes Macrophage Inflammatory Responsiveness by Up-Regulating NF-κB via IκBa Negative Regulation. *Mol. Biol. Cell* 18, 930–942. doi: 10.1091/mbc.E06-03-0217

Makoukji, J., Shackleford, G., Meffre, D., Grenier, J., Liere, P., Lobaccaro, J.-M. A., et al. (2011). Interplay between LXR and Wnt/β-catenin signaling in the negative regulation of peripheral myelin genes by oxysterols. *J. Neurosci.* 31, 9620–9629. doi: 10.1523/JNEUROSCI.0761-11.2011

Malz, M., Aulmann, A., Samarin, J., Bissinger, M., Longerich, T., Schmitt, S., et al. (2012). Nuclear accumulation of seven in absentia homologue-2 supports motility and proliferation of liver cancer cells. *Int. J. Cancer* 131, 2016–2026. doi: 10.1002/ijc.27473

Mann, B., Gelos, M., Siedow, A., Hanski, M. L., Gratchev, A., Ilyas, M., et al. (1999). Target genes of β-catenin-T cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. *Proc. Natl. Acad. Sci. U.S.A.* 96, 1603–1608. doi: 10.1073/pnas.96.4.1603

Matsuda, S., and Kitagishi, Y. (2013). Peroxisome proliferator-activated receptor and vitamin D receptor signaling pathways in cancer cells. *Cancers* 5, 1261–1270. doi: 10.3390/cancers5041261

Matsuzawa, S. I., and Reed, J. C. (2001). Siah-1, SIP, and Ebi collaborate in a novel pathway for β-catenin degradation linked to p53 responses. *Mol. Cell* 7, 915–926. doi: 10.1016/S1097-2765(01)00242-8

Mishra, P., Senthivinayagam, S., Rangasamy, V., Sondarva, G., and Rana, B. (2010). Mixed lineage kinase-3/JNK1 axis promotes migration of human gastric cancer cells following gastrin stimulation. *Mol. Endocrinol.* 24, 598–607. doi: 10.1210/me.2009-0387

Müller, T., Choidas, A., Reichmann, E., and Ullrich, A. (1999). Phosphorylation and free pool of β-catenin are regulated by tyrosine kinases and tyrosine phosphatases during epithelial cell migration. *J. Biol. Chem.* 274, 10173–10183. doi: 10.1074/jbc.274.15.10173

Molloy, N. H., Read, D. E., and Gorman, A. M. (2011). Nerve growth factor in cancer cell death and survival. *Cancers* 3, 510–530. doi: 10.3390/cancers3010510

Morali, O. G., Delmas, V., Moore, R., Jeanney, C., Thiery, J. P., and Larue, L. (2001). IGF-II induces rapid β-catenin relocation to the nucleus during epithelium to mesenchyme transition. *Oncogene* 20, 4942–4950. doi: 10.1038/sj.onc.1204660

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

Mrsny, R. J., Gewirtz, A. T., Siccardi, D., Savidge, T., Hurley, B. P., Madara, J. L., et al. (2004). Identification of hepoxilin A3 in inflammatory events: a required role in neutrophil migration across intestinal epithelia. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7421–7426. doi: 10.1073/pnas.0400832101

Nakayama, K., Qi, J., and Ronai, Z. (2009). The ubiquitin ligase Siah2 and the hypoxia response. *Mol. Cancer Res.* 7, 443–451. doi: 10.1158/1541-7786.MCR-08-0458

Nakshatri, H., Bhat-Nakshatri, P., Martin, D. A., Goulet, R. J., and Sledge, G. W. (1997). Constitutive activation of NF-κB during progression of breast cancer to hormone-independent growth. *Mol. Cell Biol.* 17, 3629–3639. doi: 10.1128/MCB.17.7.3629

Nata, T., Basheer, A., Cocchi, F., van Besien, R., Massoud, R., Jacobson, S., et al. (2015). Targeting the binding interface on a shared receptor subunit of a cytokine family enables the inhibition of multiple member cytokines with selectable target spectrum. *J. Biol. Chem.* 290, 22338–22351. doi: 10.1074/jbc.M115.661074

Nebert, D. W. (2002). Transcription factors and cancer: an overview. *Toxicology* 181–182, 131–141. doi: 10.1016/S0300-483X(02)00269-X

Normanno, N., Luca, A. D., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., et al. (2006). Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* 366, 2–16. doi: 10.1016/j.gene.2005.10.018

Oguma, K., Oshima, H., Aoki, M., Uchio, R., Naka, K., Nakamura, S., et al. (2008). Activated macrophages promote Wnt signalling through tumour necrosis factor-α in gastric tumour cells. *EMBO J.* 27, 1671–1681. doi: 10.1038/emboj.2008.105

Oikawa, T. (2004). ETS transcription factors: possible targets for cancer therapy. *Cancer Sci.* 95, 626–633. doi: 10.1111/j.1349-7006.2004.tb03320.x

Oikawa, T., and Yamada, T. (2003). Molecular biology of the Ets family of transcription factors. *Gene* 303, 11–34. doi: 10.1016/S0378-1119(02)01156-3

Orlov, I., Rochel, N., Moras, D., and Klaholz, B. P. (2012). Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *EMBO J.* 31, 291–300. doi: 10.1038/emboj.2011.445

Pasz-Walczak, G., Kordek, R., and Faflik, M. (2001). P21 (WAF1) expression in colorectal cancer: correlation with P53 and cyclin D1 expression, clinicopathological parameters and prognosis. *Pathol. Res. Pract.* 197, 683–689. doi: 10.1078/0344-0338-00146

Polakis, P. (2012). Wnt signaling in cancer. *Cold Spring Harb. Perspect. Biol.* 4:a008052. doi: 10.1101/cshperspect.a008052

Qi, J., Nakayama, K., Cardiff, R. D., Borowsky, A. D., Kaul, K., Williams, R., et al. (2010). Siah2-dependent concerted activity of HIF and FoxA2 regulates formation of neuroendocrine phenotype and neuroendocrine prostate tumors. *Cancer Cell* 18, 23–38. doi: 10.1016/j.ccr.2010.05.024

Qi, J., Tripathi, M., Mishra, R., Sahgal, N., Fazli, L., Fazil, L., et al. (2013). The E3 ubiquitin ligase Siah2 contributes to castration-resistant prostate cancer by regulation of androgen receptor transcriptional activity. *Cancer Cell* 23, 332–346. doi: 10.1016/j.ccr.2013.02.016

Rakoff-Nahoum, S., and Medzhitov, R. (2009). Toll-like receptors and cancer. *Nat. Rev. Cancer* 9, 57–63. doi: 10.1038/nrc2541

Rask, K., Thörn, M., Pontén, F., Kraaz, W., Sundfeldt, K., Hedin, L., et al. (2000). Increased expression of the transcription factors CCAAT-enhancer binding protein-β (C/EBβ) and C/EBPζ (CHOP) correlate with invasiveness of human colorectal cancer. *Int. J. Cancer* 86, 337–343. doi: 10.1002/(SICI)1097-0215(20000501)86:3<337::AID-IJC6>3.0.CO;2-3

Reddy, E. P., Korapati, A., Chaturvedi, P., and Rane, S. (2000). IL-3 signaling and the role of Src kinases, JAKs and STATs: a covert liaison unveiled. *Oncogene* 19, 2532–2547. doi: 10.1038/sj.onc.1203594

Reka, A. K., Kurapati, H., Narala, V. R., Bommer, G., Chen, J., Standiford, T. J., et al. (2010). Peroxisome proliferator-activated receptor-γ activation inhibits tumor metastasis by antagonizing Smad3-mediated epithelial-mesenchymal transition. *Mol. Cancer Ther.* 9, 3221–3232. doi: 10.1158/1535-7163.MCT-10-0570

Richmond, A., and Su, Y. (2008). Mouse xenograft models vs gem models for human cancer therapeutics. *Dis. Model. Mech.* 1, 78–82. doi: 10.1242/dmm.000976

Romon, R., Adriaenssens, E., Lagadec, C., Germain, E., Hondermarck, H., and Bourhis, X. L. (2010). Nerve growth factor promotes breast cancer angiogenesis by activating multiple pathways. *Mol. Cancer* 9:157. doi: 10.1186/1476-4598-9-157

Sakamoto, K., Maeda, S., Hikiba, Y., Nakagawa, H., Hayakawa, Y., Shibata, W., et al. (2009). Constitutive NF-κB activation in colorectal carcinoma plays a key role in angiogenesis, promoting tumor growth. *Clin. Cancer Res.* 15, 2248–2258. doi: 10.1158/1078-0432.CCR-08-1383

Salaun, B., Coste, I., Rissoan, M.-C., Lebecque, S. J., and Renno, T. (2006). TLR3 can directly trigger apoptosis in human cancer cells. *J. Immunol.* 176, 4894–4901. doi: 10.4049/jimmunol.176.8.4894

Santarpia, L., Lippman, S. M., and El-Naggar, A. K. (2012). Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opin. Ther. Targets* 16, 103–119. doi: 10.1517/14728222.2011.645805

Sarkar, T. R., Sharan, S., Wang, J., Pawar, S. A., Cantwell, C. A., Johnson, P. F., et al. (2012). Identification of a Src tyrosine kinase/SIAH2 E3 ubiquitin ligase pathway that regulates C/EBPδ expression and contributes to transformation of breast tumor cells. *Mol. Cell Biol.* 32, 320–332. doi: 10.1128/MCB.05790-11

Sarraf, P., Mueller, E., Jones, D., King, F. J., DeAngelo, D. J., Partridge, J. B., et al. (1998). Differentiation and reversal of malignant changes in colon cancer through PPARγ. *Nat. Med.* 4, 1046–1052. doi: 10.1038/2030

Sasso, G. L., Bovenga, F., Murzilli, S., Salvatore, L., Tullio, G. D., Martelli, N., et al. (2013). Liver X receptors inhibit proliferation of human colorectal cancer cells and growth of intestinal tumors in mice. *Gastroenterology* 144, 1497–1507, 1507.e1–13. doi: 10.1053/j.gastro.2013.02.005

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470. doi: 10.1126/science.270.5235.467

Schmid, C. D., Perier, R., Praz, V., and Bucher, P. (2006). EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* 34, D82–D85. doi: 10.1093/nar/gkj146

Sepulveda, V. A. T., Weigel, N. L., and Falzon, M. (2006). Prostate cancer cell type-specific involvement of the VDR and RXR in regulation of the human PTHrP gene via a negative VDRE. *Steroids* 71, 102–115. doi: 10.1016/j.steroids.2005.08.009

Shackleford, G., Makoukji, J., Grenier, J., Liere, P., Meffre, D., and Massaad, C. (2013). Differential regulation of Wnt/β-catenin signaling by Liver X Receptors in Schwann cells and oligodendrocytes. *Biochem. Pharmacol.* 86, 106–114. doi: 10.1016/j.bcp.2013.02.036

Sharpless, N. E., and Depinho, R. A. (2006). The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov.* 5, 741–754. doi: 10.1038/nrd2110

Shaulian, E., and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nat. Cell Biol.* 4, E131–E136. doi: 10.1038/ncb0502-e131

Slattery, M. L. (2007). Vitamin D receptor gene (VDR) associations with cancer. *Nutr. Rev.* 65, S102–S104. doi: 10.1301/nr.2007.aug.S102-S104

Slattery, M. L., Lundgreen, A., Kadlubar, S. A., Bondurant, K. L., and Wolff, R. K. (2013). JAK/STAT/SOCS-signaling pathway and colon and rectal cancer. *Mol. Carcinog.* 52, 155–166. doi: 10.1002/mc.21841

Smits, R., Kartheuser, A., Jagmohan-Changur, S., Leblanc, V., Breukel, C., de Vries, A., et al. (1997). Loss of Apc and the entire chromosome 18 but absence of mutations at the Ras and Tp53 genes in intestinal tumors from Apc1638N, a mouse model for Apc-driven carcinogenesis. *Carcinogenesis* 18, 321–327. doi: 10.1093/carcin/18.2.321

Smits, R., Kielman, M. F., Breukel, C., Zurcher, C., Neufeld, K., Jagmohan-Changur, S., et al. (1999). Apc1638T: a mouse model delineating critical domains of the adenomatous polyposis coli protein involved in tumorigenesis and development. *Genes Dev.* 13, 1309–1321. doi: 10.1101/gad.13.10.1309

Soubeyran, P., Andre, F., Lissitzky, J. C., Mallo, G. V., Moucadel, V., Roccabianca, M., et al. (1999). Cdx1 promotes differentiation in a rat intestinal epithelial cell line. *Gastroenterology* 117, 1326–1338. doi: 10.1016/S0016-5085(99)70283-0

Stamos, J. L., and Weis, W. I. (2013). The β-catenin destruction complex. *Cold Spring Harb. Perspect. Biol.* 5:a007898. doi: 10.1101/cshperspect.a007898

Steller, E. J. A., Raats, D. A., Koster, J., Rutten, B., Govaert, K. M., Emmink, B. L., et al. (2013). PDGFRB promotes liver metastasis formation of mesenchymal-like colorectal tumor cells. *Neoplasia* 15, 204–217. doi: 10.1593/neo.121726

Su, L. K., Kinzler, K. W., Vogelstein, B., Preisinger, A. C., Moser, A. R., Luongo, C., et al. (1992). Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science* 256, 668–670. doi: 10.1126/science.1350108

Sudhakar, J. N., and Chow, K.-C. (2014). Human RAD23 homolog A is required for the nuclear translocation of apoptosis-inducing factor during induction of cell death. *Biol. Cell* 106, 359–376. doi: 10.1111/boc.201400013

Suh, E., Chen, L., Taylor, J., and Traber, P. G. (1994). A homeodomain protein related to caudal regulates intestine-specific gene transcription. *Mol. Cell. Biol.* 14, 7340–7351. doi: 10.1128/MCB.14.11.7340

Suh, E., and Traber, P. G. (1996). An intestine-specific homeobox gene regulates proliferation and differentiation. *Mol. Cell. Biol.* 16, 619–625. doi: 10.1128/MCB.16.2.619

Sun, Q., Sun, F., Wang, B., Liu, S., Niu, W., Liu, E., et al. (2014a). Interleukin-8 promotes cell migration through integrin $\alpha v\beta 6$ upregulation in colorectal cancer. *Cancer Lett.* 354, 245–253. doi: 10.1016/j.canlet.2014.08.021

Sun, Y., Shen, S., Liu, X., Tang, H., Wang, Z., Yu, Z., et al. (2014b). MiR-429 inhibits cells growth and invasion and regulates EMT-related marker genes by targeting Onecut2 in colorectal carcinoma. *Mol. Cell Biochem.* 390, 19–30. doi: 10.1007/s11010-013-1950-x

Suto, R., Tominaga, K., Mizuguchi, H., Sasaki, E., Higuchi, K., Kim, S., et al. (2004). Dominant-negative mutant of c-Jun gene transfer: a novel therapeutic strategy for colorectal cancer. *Gene Ther.* 11, 187–193. doi: 10.1038/sj.gt.3302158

Takahashi, H., Ogata, H., Nishigaki, R., Broide, D. H., and Karin, M. (2010). Tobacco smoke promotes lung tumorigenesis by triggering IKKβ- and JNK1-dependent inflammation. *Cancer Cell* 17, 89–97. doi: 10.1016/j.ccr.2009.12.008

Takeuchi, K., and Ito, F. (2011). Receptor tyrosine kinases and targeted cancer therapeutics. *Biol. Pharm. Bull.* 34, 1774–1780. doi: 10.1248/bpb.34.1774

Tang, X., and Zhu, Y. (2012). TLR4 signaling promotes immune escape of human colon cancer cells by inducing immunosuppressive cytokines and apoptosis resistance. *Oncol. Res.* 20, 15–24. doi: 10.3727/096504012X13425470196092

Taylor, J. K., Boll, W., Levy, T., Suh, E., Siang, S., Mantei, N., et al. (1997). Comparison of intestinal phospholipase A/lysophospholipase and sucrase-isomaltase genes suggest a common structure for enterocyte-specific promoters. *DNA Cell Biol.* 16, 1419–1428. doi: 10.1089/dna.1997.16.1419

Torkamani, A., Verkhivker, G., and Schork, N. J. (2009). Cancer driver mutations in protein kinase genes. *Cancer Lett.* 281, 117–127. doi: 10.1016/j.canlet.2008.11.008

Tsoupras, A. B., Iatrou, C., Frangia, C., and Demopoulos, C. A. (2009). The implication of platelet activating factor in cancer growth and metastasis: potent beneficial role of PAF-inhibitors and antioxidants. *Infect. Disord. Drug Targets* 9, 390–399. doi: 10.2174/187152609788922555

Uno, S., Endo, K., Jeong, Y., Kawana, K., Miyachi, H., Hashimoto, Y., et al. (2009). Suppression of β-catenin signaling by liver X receptor ligands. *Biochem. Pharmacol.* 77, 186–195. doi: 10.1016/j.bcp.2008.10.007

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536. doi: 10.1038/415530a

Vedin, L.-L., Åke Gustafsson, J., and Steffensen, K. R. (2013). The oxysterol receptors LXRα and LXRβ suppress proliferation in the colon. *Mol. Carcinog.* 52, 835–844. doi: 10.1002/mc.21924

Velho, S., Pinto, A., Licastro, D., Oliveira, M. J., Sousa, F., Stupka, E., et al. (2014). Dissecting the signaling pathways associated with the oncogenic activity of MLK3 P252H mutation. *BMC Cancer* 14:182. doi: 10.1186/1471-2407-14-182

Vleminckx, K., Vakaet, L., Mareel, M., Fiers, W., and van Roy, F. (1991). Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell* 66, 107–119. doi: 10.1016/0092-8674(91)90143-M

Voronov, E., Shouval, D. S., Krelin, Y., Cagnano, E., Benharroch, D., Iwakura, Y., et al. (2003). IL-1 is required for tumor invasiveness and angiogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 2645–2650. doi: 10.1073/pnas.0437939100

Wang, S., Liu, Z., Wang, L., and Zhang, X. (2009). NF-κB signaling pathway, inflammation and colorectal cancer. *Cell Mol. Immunol.* 6, 327–334. doi: 10.1038/cmi.2009.43

Wang, W., Abbruzzese, J. L., Evans, D. B., Larry, L., Cleary, K. R., and Chiao, P. J. (1999). The nuclear factor-κB RelA transcription factor is constitutively activated in human pancreatic adenocarcinoma cells. *Clin. Cancer Res.* 5, 119–127.

Wang, Y., Shen, L., Xu, N., Wang, J.-W., Jiao, S.-C., Liu, Z.-Y., et al. (2012). UGT1A1 predicts outcome in colorectal cancer treated with irinotecan and fluorouracil. *World J. Gastroenterol.* 18, 6635–6644. doi: 10.3748/wjg.v18.i45.6635

Weigelt, B., Peterse, J. L., and van 't Veer, L. J. (2005). Breast cancer metastasis: markers and models. *Nat. Rev. Cancer* 5, 591–602. doi: 10.1038/nrc1670

West, N. R., McCuaig, S., Franchini, F., and Powrie, F. (2015). Emerging cytokine networks in colorectal cancer. *Nat. Rev. Immunol.* 15, 615–629. doi: 10.1038/nri3896

Wiener, D., Doerge, D. R., Fang, J.-L., Upadhyaya, P., and Lazarus, P. (2004). Characterization of N-glucuronidation of the lung carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) in human liver:

importance of UDP-glucuronosyltransferase 1A4. *Drug Metab. Dispos.* 32, 72–79. doi: 10.1124/dmd.32.1.72

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016

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

Wong, C. S. F., and Möller, A. (2013). Siah: a promising anticancer target. *Cancer Res.* 73, 2400–2406. doi: 10.1158/0008-5472.CAN-12-4348

Wong, C. S. F., Sceneay, J., House, C. M., Halse, H. M., Liu, M. C. P., George, J., et al. (2012). Vascular normalization by loss of Siah2 results in increased chemotherapeutic efficacy. *Cancer Res.* 72, 1694–1704. doi: 10.1158/0008-5472.CAN-11-3310

Xie, W., Rimm, D. L., Lin, Y., Shih, W. J., and Reiss, M. (2003). Loss of Smad signaling in human colorectal cancer is associated with advanced disease and poor prognosis. *Cancer J.* 9, 302–312. doi: 10.1097/00130404-200307000-00013

Xu, Y., and Pasche, B. (2007). TGF-β signaling alterations and susceptibility to colorectal cancer. *Hum. Mol. Genet.* 16 , R14–R20. doi: 10.1093/hmg/ddl486

Yang, W. L., and Frucht, H. (2001). Activation of the PPAR pathway induces apoptosis and COX-2 inhibition in HT-29 human colon cancer cells. *Carcinogenesis* 22, 1379–1383. doi: 10.1093/carcin/22.9.1379

Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. *Nat. Rev. Cancer* 9, 798–809. doi: 10.1038/nrc2734

Yu, J., Ustach, C., and Kim, H.-R. C. (2003). Platelet-derived growth factor signaling and human cancer. *J. Biochem. Mol. Biol.* 36, 49–59. doi: 10.5483/BMBRep.2003.36.1.049

Yu, J., Zhang, L., Hwang, P. M., Kinzler, K. W., and Vogelstein, B. (2001). PUMA induces the rapid apoptosis of colorectal cancer cells. *Mol. Cell* 7, 673–682. doi: 10.1016/S1097-2765(01)00213-1

Yuan, L., Zhou, C., Lu, Y., Hong, M., Zhang, Z., Zhang, Z., et al. (2015). IFN-γ-mediated IRF1/miR-29b feedback loop suppresses colorectal cancer cell growth and metastasis by repressing IGF1. *Cancer Lett.* 359, 136–147. doi: 10.1016/j.canlet.2015.01.003

Yuan, X.-W., Wang, D.-M., Hu, Y., Tang, Y.-N., Shi, W.-W., Guo, X.-J., et al. (2013). Hepatocyte nuclear factor 6 suppresses the migration and invasive growth of lung cancer cells through p53 and the inhibition of epithelial-mesenchymal transition. *J. Biol. Chem.* 288, 31206–31216. doi: 10.1074/jbc.M113.480285

Zhang, J., Guenther, M. G., Carthew, R. W., and Lazar, M. A. (1998). Proteasomal regulation of nuclear receptor corepressor-mediated repression. *Genes Dev.* 12, 1775–1780. doi: 10.1101/gad.12.12.1775

Zhao, L. Y., Niu, Y., Santiago, A., Liu, J., Albert, S. H., Robertson, K. D., et al. (2006). An EBF3-mediated transcriptional program that induces cell cycle arrest and apoptosis. *Cancer Res.* 66, 9445–9452. doi: 10.1158/0008-5472.CAN-06-1713

Zheng, W., Wong, K. E., Zhang, Z., Dougherty, U., Mustafi, R., Kong, J., et al. (2012). Inactivation of the vitamin D receptor in APC(min/+) mice reveals a critical role for the vitamin D receptor in intestinal tumor growth. *Int. J. Cancer* 130, 10–19. doi: 10.1002/ijc.25992

Zhu, Y., Richardson, J. A., Parada, L. F., and Graff, J. M. (1998). Smad3 mutant mice develop metastatic colorectal cancer. *Cell* 94, 703–714. doi: 10.1016/S0092-8674(00)81730-4

Zwick, E., Bange, J., and Ullrich, A. (2001). Receptor tyrosine kinase signalling as a target for cancer intervention strategies. *Endocr. Relat. Cancer* 8, 161–173. doi: 10.1677/erc.0.0080161

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 19 January 2016
doi: 10.3389/fpls.2015.01246

A *De novo* Transcriptomic Approach to Identify Flavonoids and Anthocyanins "Switch-Off" in Olive (*Olea europaea* L.) Drupes at Different Stages of Maturation

Domenico L. Iaria[1], *Adriana Chiappetta*[2] *and Innocenzo Muzzalupo*[1,3]*

[1] *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria, Centro di Ricerca per l'Olivicoltura e l'Industria Olearia, Cosenza, Italy,* [2] *Dipartimento di Biologia, Ecologia e Scienze della Terra, Università della Calabria, Cosenza, Italy,* [3] *Dipartimento di Farmacia, Scienze della Salute e della Nutrizione, Università della Calabria, Cosenza, Italy*

Highlights

- A *de novo* transcriptome reconstruction of olive drupes was performed in two genotypes
- Gene expression was monitored during drupe development in two olive cultivars
- Transcripts involved in flavonoid and anthocyanin pathways were analyzed in Cassanese and Leucocarpa cultivars
- Both cultivar and developmental stage impact gene expression in *Olea europaea* fruits.

During ripening, the fruits of the olive tree (*Olea europaea* L.) undergo a progressive chromatic change characterized by the formation of a red-brown "spot" which gradually extends on the epidermis and in the innermost part of the mesocarp. This event finds an exception in the Leucocarpa cultivar, in which we observe a destabilized equilibrium between the metabolisms of chlorophyll and other pigments, particularly the anthocyanins whose switch-off during maturation promotes the white coloration of fruits. Despite its importance, genomic information on the olive tree is still lacking. Different RNA-seq libraries were generated from drupes of "Leucocarpa" and "Cassanese" olive genotypes, sampled at 100 and 130 days after flowering (DAF), and were used in order to identify transcripts involved in the main phenotypic changes of fruits during maturation and their corresponding expression patterns. A total of 103,359 transcripts were obtained and 3792 and 3064 were differentially expressed in "Leucocarpa" and "Cassanese" genotypes, respectively, during 100–130 DAF transition. Among them flavonoid and anthocyanin related transcripts such as phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate-CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonol 3′-hydrogenase (F3′H), flavonol 3′5 ′-hydrogenase (F3′5′H), flavonol synthase (FLS), dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS), UDP-glucose:anthocianidin: flavonoid glucosyltransferase (UFGT) were identified. These results contribute to reducing the current gap in information regarding metabolic processes, including those linked to fruit pigmentation in the olive.

Keywords: *Olea europaea*, flavonoid and anthocyanin pathway, RNA-seq, *de novo* assembly, gene expression

Edited by:
Jörg Linde, Leibniz-Institute for Natural Product Research and Infection Biology -Hans-Knoell-Institute, Germany

Reviewed by:
Thiruvarangan Ramaraj, National Center for Genome Resources, USA
Gabriele Bucci, San Raffaele Scientific Institute, Italy

***Correspondence:**
Innocenzo Muzzalupo
innocenzo.muzzalupo@entecra.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: *07 October 2015*
Accepted: *21 December 2015*
Published: *19 January 2016*

Citation:
Iaria DL, Chiappetta A and Muzzalupo I (2016) A De novo Transcriptomic Approach to Identify Flavonoids and Anthocyanins "Switch-Off" in Olive (Olea europaea L.) Drupes at Different Stages of Maturation. Front. Plant Sci. 6:1246. doi: 10.3389/fpls.2015.01246

INTRODUCTION

The olive tree (*Olea europaea* L. subsp. *europaea* var. *europaea*) is one of the most important and widespread fruit trees in the Mediterranean area. It belongs to the *Oleaceae* family, which includes 600 species within 25 genera. It is widely distributed on all continents, from temperate areas in the north to sub-tropical regions and from low to high altitudes. Native to Mediterranean regions, *Olea europaea* is the only species within the genus *Olea* that produces edible fruits (Green and Wickens, 1989; Wallander and Albert, 2000; Green, 2002; FAOSTAT, 2008[1]). The quality of its products, olive oil and table olives, is highly dependent on the agronomic and organoleptic characteristics of its drupes. These characteristics vary in relation to the genetic traits, varieties, the stage of ripeness, as well as in relation to the different susceptibility to environmental growth conditions (Loumou and Giourga, 2003; Conde et al., 2008).

The genuineness of olive oil is important within the "Mediterranean diet." Several research and epidemiological studies link healthy aspects of its components; in particular, olive oil is known to exert protective effects against vascular disease and the onset of cancer (Vauzour et al., 2010). These features are correlated to the high percentage of monounsaturated fats as well as to the high content of antioxidant compounds such as phenols and tocopherols, which, together with other components, characterize the nutraceutical profile of olive products (Pérez-Jiménez et al., 2007; Bruno et al., 2009; Muzzalupo et al., 2011). Phenolic compounds represent a complex mixture in olive derived products responsible for the anti-atherogenic and anti-cancerogenic effects, and for antioxidant properties (Hashim et al., 2008; Llorente-Cortes et al., 2010; Martinelli and Tonutti, 2012). Despite the importance and uniqueness of olive products, the long juvenile developmental phase and its intrinsic self-incompatibility mechanisms slow down current olive breeding programs, which are still very long. Although the current breeding strategies can now benefit from the availability of new polymorphic genetic markers, characterization of the olive germplasm is still far from complete (Baldoni et al., 2009; Muzzalupo, 2012; Muzzalupo et al., 2014).

Therefore, it is of prime importance to focus research programs toward innovative improvement strategies to support conventional programs. In particular, a wider characterization of genes related to plant product quality and to adaptive mechanisms, could provide new information and tools to support both Marker Aided Selection (MAS) strategies and biotechnological approaches. This would aid the development of new growing techniques to increase productivity and quality of this unique species.

Anthocyanins are the most widely distributed group of pigments in plants. They are synthesized via the phenylpropanoid pathway and are mainly responsible for the mauve, red, blue, and purple colors in flowers, fruits, leaves, seeds, and other organs in most flowering plants. As one of the most ubiquitous class of flavonoids, anthocyanins possess a multitude of biological roles, including protection against solar exposure and ultraviolet radiation, free radical scavenging and anti-oxidative capacity, defense against many different pathogens, and attraction of predators for seed dispersal. Anthocyanins also play a role in consumer preference for flower and fruit quality, potential food health properties, and related horticultural attributes. As a result, classical breeding, as well as transgene technologies, have been used to enhance or create novel colors in ornamental and food crops (Chalker-Scott, 1999; Schaefer et al., 2004; Takahama, 2004; Stommel et al., 2009).

The enzymes involved in the anthocyanin biosynthetic pathway are well characterized. Many of the genes encoding these enzymes have been cloned and share high sequence similarity across species and exhibit tissue- or development-specific expression. Chalcone synthase (CHS) is the first enzymatic step of the biosynthetic pathway (Coe et al., 1981; Dooner, 1983; Koes et al., 1989, 2005). Subsequently chalcone isomerase (CHI) catalyzes the isomerization of chalcone to naringenin (van Tunen et al., 1988, 1989; Grotewold and Peterson, 1994; Griesbach and Beck, 2005). Flavanone 3-hydroxylase (F3H) converts naringen into dihydrokaempferol, which is converted to anthocyanins by the action of three enzymes. Dihydroflavonol is first converted to a colorless leucoanthocyanidin by dihydroflavonol 4-reductase (DFR). Leucoanthocyanidins are subsequently converted to colored anthocyanidins by anthocyanidin synthase (ANS) finally, the UDP-glucose-flavonoid 3-O-glucosyltransferase (UFGT) creates the anthocyanin-3-glucoside. Within the path, the CHS is the first and key regulatory enzyme of flavonoid biosynthesis and the DFR is the first committed enzyme of anthocyanin biosynthesis in the flavonoid pathway (Holton and Cornish, 1995; Ramsay and Glover, 2005; Martinelli and Tonutti, 2012).

Despite having been recently studied in different olive cultivars (Alagna et al., 2009; Galla et al., 2009; Martinelli and Tonutti, 2012) the molecular mechanisms involved in the regulation of biosynthesis are still unknown.

Tissue- or developmental-specific expression exhibited by anthocyanin structural genes is controlled by a set of regulatory genes. It is known that MYB, bHLH MYC, and WD40 repeat proteins, interacting together to form a regulatory complex that controls anthocyanin structural genes at the transcriptional level (Dixon et al., 2005; Ramsay and Glover, 2005; He et al., 2008; Tian et al., 2008; Alagna et al., 2009; Galla et al., 2009; Stommel et al., 2009; Martinelli and Tonutti, 2012; Ravaglia et al., 2013; Chiappetta et al., 2015).

It has been suggested that a functional MYB-MYC-WD complex directly binds the cis-element of structural gene through MYB transcription factor, while R-like MYC might bind indirectly via a hypothetical R interaction protein (RIP) (Ramsay and Glover, 2005). R-like MYC is centered in the complex that interacts with a MYB factor with WD proteins on its sides. Together, they activate the entire set of anthocyanin biosynthesis genes (Stommel et al., 2009).

[1]FAOSTAT 2008 home page Columbia URL: http://www.columbia.edu/cgi-bin/cul/resolve?ASL9609

The aim of this work was to define the main transcriptomic profile differences during olive drupe development and to identify the transcripts involved in flavonoid and anthocyanin metabolism.

We have chosen to analyze the transcriptome profile at 100 and 130 days after flowering (DAF), through an Illumina RNA-seq approach, to identify the transcripts along flavonoids and anthocyanins biosynthetic pathways and to monitor their expression levels during ripening. A *de novo* transcriptome reconstruction of olive fruits was performed together with a full expression analysis between samples from "Leucocarpa," an olive variety characterized by a switch-off in skin color at full ripeness, and "Cassanese," used as control plant. Significant differences in flavonoid and anthocyanin transcript expression profiles emerged, both during fruit maturation and in relation to genotypes. Consequently, from the wide array of information obtained, our attention was focused on the identified candidate genes set, the expression of which was confirmed by quantitative PCR. In addition, the expression patterns of different MYB, MYC, and WDR transcriptional activators was compared to CHS, DFR, and ANS genes during fruit ripening (Matus et al., 2009; Ravaglia et al., 2013).

MATERIALS AND METHODS

Plant Materials

Olive drupes, of *Olea europaea* L. Leucocarpa and Cassanese cv were used. Drupes were collected from 20-year-old plants, clonally propagated and belonging to the olive germplasm collection of the Agricultural Research Council—Olive Growing and Oil Industry Research Centre, CREA-OLI in Mirto-Crosia (Cosenza, Calabria, Italy). Olive trees were grown using the same field conditions and were located at latitude 39°37′04.57″N, longitude 16°45′42.00″E and altitude 8 m asl).

Fruit sampling was performed as previously described (Matus et al., 2009): for each cultivar, drupes ($n = 30$,) were randomly collected at 100 and 130 DAF (**Figure S1**). In order to minimize the effects related to asynchronous fruits maturation within the same tree, drupes with similar pigmentation were picked from all around the external parts of the tree canopy. Concerning drupe pigmentation, the epi-mesocarp tissues, was totally green in color at 100 DAF whereas at 130 DAF the pulp pigmentation was 50% brown in "Cassanese" and totally unpigmented in "Leucocarpa" drupes (**Figure S1**).

All samples were fixed in liquid nitrogen and stored at −80°C for both RNA-seq and qRT-PCR experiments.

RNA-Seq Library Preparation and Sequencing

In order to obtain a general overview of the transcripts and metabolic pathways involved in fruit maturation and to avoid cross contamination from non-homogeneous tissue separation, sample pooling strategy has been here used (Peng et al., 2003). Pooling reduces variability by minimizing individual variation and represents an alternative approach to biological replicates in experiments where the interest is not on the individual but rather on characteristics of the population (e.g., common changes in expression patterns; Karp and Lilley, 2007, 2009).

Total RNA was extracted from the epi-mesocarp tissues of drupes collected together, using the RNeasy Plant Mini kit (Qiagen) according to the manufacturer's instructions. Each RNA sample was subjected to DNase digestion (DNase I, Roche) to remove any DNA contamination and pooled equally, as previously described (Muzzalupo et al., 2012). RNA was quantified by the NanoDrop Spectrophotometer ND-2000 and quality was checked by electrophoresis (28S rRNA/18S rRNA ratios). Samples with a concentration of ≥400 ng/μl, OD260/280=1.8∼2.2, RNA 28S:18S ≥ 1.0, and RNA Integrity Number (RIN) ≥ 7.0 were used for cDNA library preparation.

Standard RNA-seq library preparation and sequencing via Illumina HiSeq TM 2000 was carried out by Technology Services of the Institute of Applied Genomics (IGA, Udine, Italy). For each sample a single-end (SE) sequencing cDNA library was constructed with a fragment length range of 50 bp. Each library was created using two replicates, consisting of a separate pool of 30 homogeneous fruits.

RNA-Seq Data Filter and *De novo* Assembly by Trinity

The raw Fastq "reads" (NCBI PDA/SRAaccession numbers: SRR1574719, SRR1574772, SRR1573503, SRR1574328, **Table 1**) were analyzed and filtered, respectively with FastQC and Fastx Toolkit softwares to obtain high quality *de novo* transcriptome sequence data. Each sequence set was filtered using the following criteria: (*i*) reads containing the sequencing adaptor were removed; (*ii*) reads with unknown nucleotides comprising more than 5% were removed; (*iii*) low-quality reads with ambiguous sequence "N" were trimmed and discarded.

Since the olive tree does not have a reference genome, the *de novo* assembly of the clean reads into transcripts was performed using the Trinity program (Grabherr et al., 2011; Haas et al., 2013), a useful method for the efficient and robust *de novo* reconstruction of transcriptomes from RNA-seq data (Ward et al., 2012; Gutierrez-Gonzalez et al., 2013; Liang et al., 2013; Liu et al., 2013; Pallavicini et al., 2013; Tulin et al., 2013).

Trinity was run via script using 128 GB of ram, 12 cpu thread and a minimum assembled contig length to report set to 300 bp.

Trinity sequentially combines Inchworm, Chrysalis and Butterfly modules to process large RNA-seq reads data, partitioning the sequence data into many individual de Bruijn graphs, representing transcriptional complexity at a given gene or locus (Grabherr et al., 2011; Haas et al., 2013).

Analysis of Transcript Assembly

For non-model organisms, one metric for evaluating the transcript assembly quality is to examine the number of transcripts that appear to be full-length or nearly full-length

TABLE 1 | Assembled transcripts for each sample.

Sample	Raw reads	Used reads	Assembled transcripts	Contig N50	Mapped reads
Leucocarpa 100 DAF	28,700,100	23,687,921	22,959	754	84.07%
Leucocarpa 130 DAF	28,121,963	23,122,308	26,203	829	84.15%
Cassanese 100 DAF	28,550,901	23,394,526	22,709	767	83.82%
Cassanese 130 DAF	57,106,631	48,153,012	31,485	972	85.49%

if compared to a closely related organism to examine full-length coverage. In this context, a more general analysis was performed aligning the assembled transcripts against all known plant proteins determining the number of unique top matching proteins that are aligned in 70–100% range of its length by full-length transcript analysis (Haas et al., 2013). Therefore, a blastable database has been created to perform a local blastx search where only the single best matching Trinity transcript was outputted for each top matching entry.

To validate our de novo assembly read remapping has been realized using bowtie2 (Langmead and Salzberg, 2012); for each data set a bowtie2 index was created, and then the number of reads that map to our transcriptome have been counted.

Abundance Estimation and Differentially Expressed Trinity Transcripts

For abundance estimation of transcriptome assemblies RSEM software was used (Li and Dewey, 2011). RSEM is a package for estimating gene and isoform expression levels from RNA-seq data. The current version of RSEM, was bundled with the Trinity software package.

Moreover, Trinity currently supports the use of Bioconductor tools (edgeR and DESeq) to compute differential expression analysis in the assembled transcriptome (Anders and Huber, 2010; Robinson et al., 2010; Grabherr et al., 2011; Haas et al., 2013). In order to identify statistically significant differences in transcript expression between samples, the number of reads/transcripts, the depth of sequencing, the transcripts length (longer transcripts generate more fragment reads) and the expression level of the transcripts were considered. Expression values, normalized for each of these factors were measured in FPKM (fragments per feature kilo base per million reads mapped) (Trapnell et al., 2010; Robinson and Oshlack, 2010) and used to make a comparison across multiple samples and replicates. Trinity supports the use of TMM (trimmed mean of *M*-values) normalization (Lekanne Deprez et al., 2002; Dillies et al., 2012), to account for differences in the mass composition of the RNA-seq samples, which does not change the fragment count data, but provides a scaling parameter that yields an effective library size (total map able reads) for each sample. This effective library size is then used in the FPKM calculations.

Quantitative PCR

Gene expression analysis was performed by quantitative real-time PCR on a 7500 fast real time PCR system (Applied Biosystems) with SYBR® Select Master Mix. The oligonucleotide primer sets (**Table 1**) used for qRT-PCR analysis were designed using Primer3 (http://primer3.ut.ee/).

Each primer pair (Supplementary data, **Table S1**) generated a single specific amplicon on the 3′-end of target sequence. PCR products were about 150–200 bp long and primer pair average efficiency ranged between 0.95 and 1.0. The housekeeping olive ELONGATION FACTOR 1 (EF1) gene (CAQ17046.1) was used to normalize the expression levels (Galla et al., 2009; Trapnell et al., 2010). Amplification reactions were prepared in a final volume of 20 µl according to the manufacturer's instructions.

All reactions were run in triplicate in 96-well reaction plates, and negative controls were set. The cycling parameters were as follows: one cycle at 95°C for 3 min to activate the Taq enzyme, followed by 40 cycles of denaturation at 95°C for 10 s and annealing-extension at 58°C for 30 s. To confirm the occurrence of a unique PCR product, the "melting curve" (Lekanne Deprez et al., 2002) was evaluated by an increase of 0.5°C every 10 s within a 60–95°C range and a unique "melting peak" in every reaction was observed. The comparisons of cycle threshold (CT) values were obtained analysing data with the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). The means of gene expression levels were calculated from two biological repeats, obtained from two independent experiments.

Blast2GO

To assign gene ontology (GO) terms in our DE data sets, we used BLASTx 2.2.26+, BLOSUM62 similarity matrix, and Blast2GO database version August 2011 programs (Conesa et al., 2005; Morgulis, 2008). The definition of each GO term was determined by the GO Consortium: http://www.geneontology.org and can be found using the EMBL European Bioinformatics Institute QuickGO: http://www.ebi.ac.uk/QuickGO or the Gene Ontology Normal Usage Tracking System, GONUTS: http://gowiki.tamu.edu/wiki/index.php/Main_Page.

Pathway assignments were determined following the Kyoto Encyclopedia of Genes and Genomes pathway database (Kanehisa et al., 2008) using BLASTX with an *E*-value threshold of 1.0E-5.

MapMan (http://mapman.gabipd.org/) analysis was done using our DE transcripts rearranged as input experimental dataset. Using the Mercator web application we can assign MapMan "Bins" to DNA sequences (Thimm et al., 2004; Lohse et al., 2014). The output was used as a mapping file for data visualization in MapMan. The Mercator tool generates functional predictions by searching a variety of reference

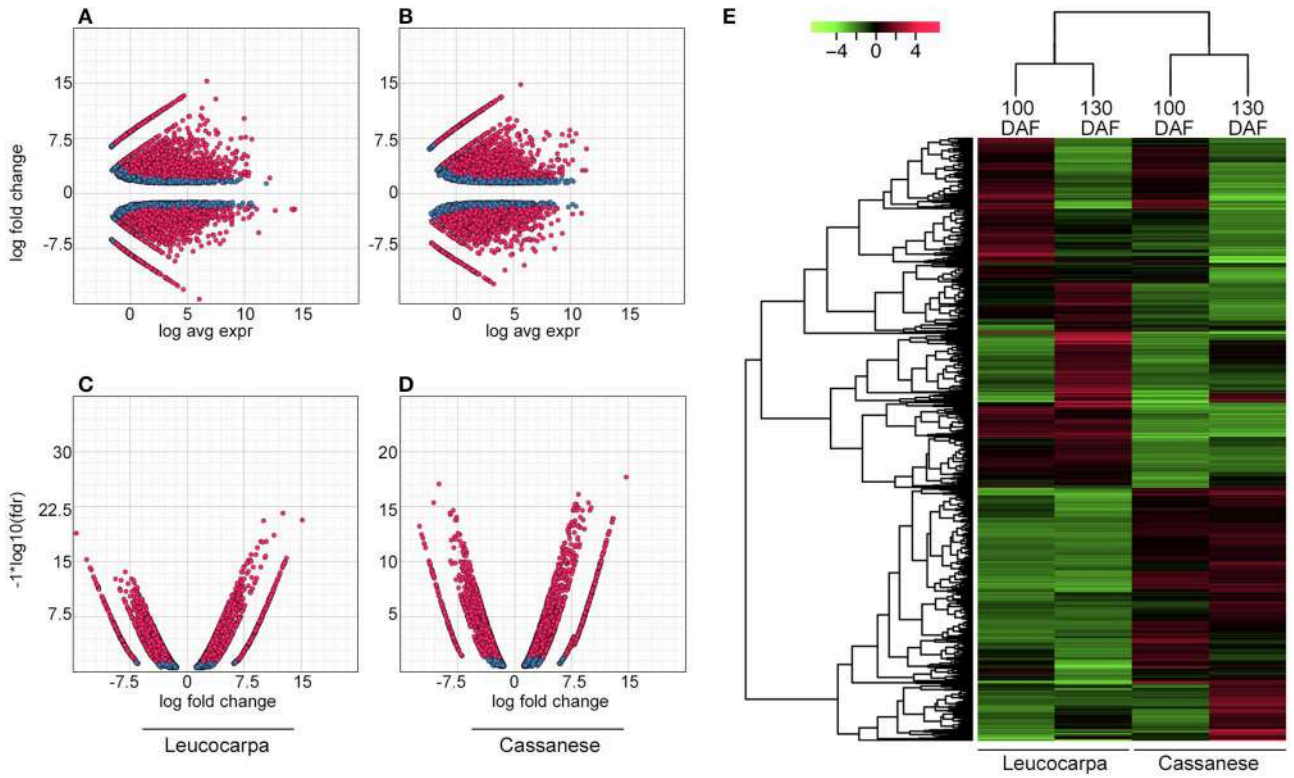


FIGURE 1 | Whole transcripts expression during fruit ripening. In **(A,B)**, MA plot for differential expression analysis generated by EdgeR: for each gene, the $\log_2$ (fold change) ($\log_2$(100DAF/130DAF)) between the 100 and 130 DAF samples is plotted (A, y axis) against the gene's $\log_2$(average expression) (M, x axis). In **(C,D)**, the Volcano plot reports a FDR ($-\log_{10}$FDR, y axis) as a function of $\log_2$ (fold change) between the 100 and 130 DAF samples (logFC, x axis). Transcripts that are identified as significantly differentially expressed at most 0.1% FDR are colored in red. In **(E)**, the heat map show the relative transcript/sample expression levels. Green and red colors are used to indicate the transcripts up to four-fold up- and down- regulated, respectively. Expression values (FPKM) are $\log_2$ transformed and then median-centered by transcript. The dendrogram, on the left, orders whole transcripts set in relation to their level of expression.

databases (BLAST-based, RPSBLAST based and InterProScan) and subsequently evaluating and compiling the search results for each input gene to propose a functional Bin.

RESULTS

RNA-Seq Library Sequencing and De novo Transcriptome Assembly by Trinity

Starting from four RNA-seq libraries, corresponding to two fruit developmental stages (100 and 130 DAF) of *Olea europaea* "Leucocarpa" and "Cassanese," 147,789,544 raw reads were generated from 50 bp insert library. A total of 142,479,595 high-quality SE reads were identified and used for *Olea europaea* trascriptome assembly, through the Trinity software. Using the 25-mer in Trinity and a minimum assembled contig length set to 300 bp, we found 103,359 transcripts. The total used reads, the total assembled transcripts, N50 statistics for each sample and remapping results are indicated in **Table 1**.

A total of 93,623 likely coding sequences were extracted with the Transdecoder utility, to identify the longest ORF (Open Reading Frame) from the transcript assembly, reporting that ORF scored according to the Markov model. In all, 9597 of the olive transcripts had a BLAST hit with an *E*-value of less than 1e-20, and 19,708 of the extracted reference coding sequences are considered to be approximately "full length," with the Trinity contigs aligning the matching UniProt reference transcript's length by more than 70%.

TABLE 2 | Number of differential expressed transcripts during 100–130 DAF transition for each cultivar.

Sample	Total transcripts	DE trascripts
Leucocarpa 100–130 DAF	49,162	3792
Cassanese 100–130 DAF	54,194	3064

Differential Expression Analysis

To estimate the differential gene expression between fruits of both considered cultivars at each developmental stage, a single assembly, based on combining all reads across all samples as inputs was generated. A single assembly was chosen to avoid difficulty in comparing the results across the different samples, due to differences in assembled transcript lengths and contiguity. Then, reads were aligned separately back to the single assembly, in order to identify the number of DE transcripts with a False

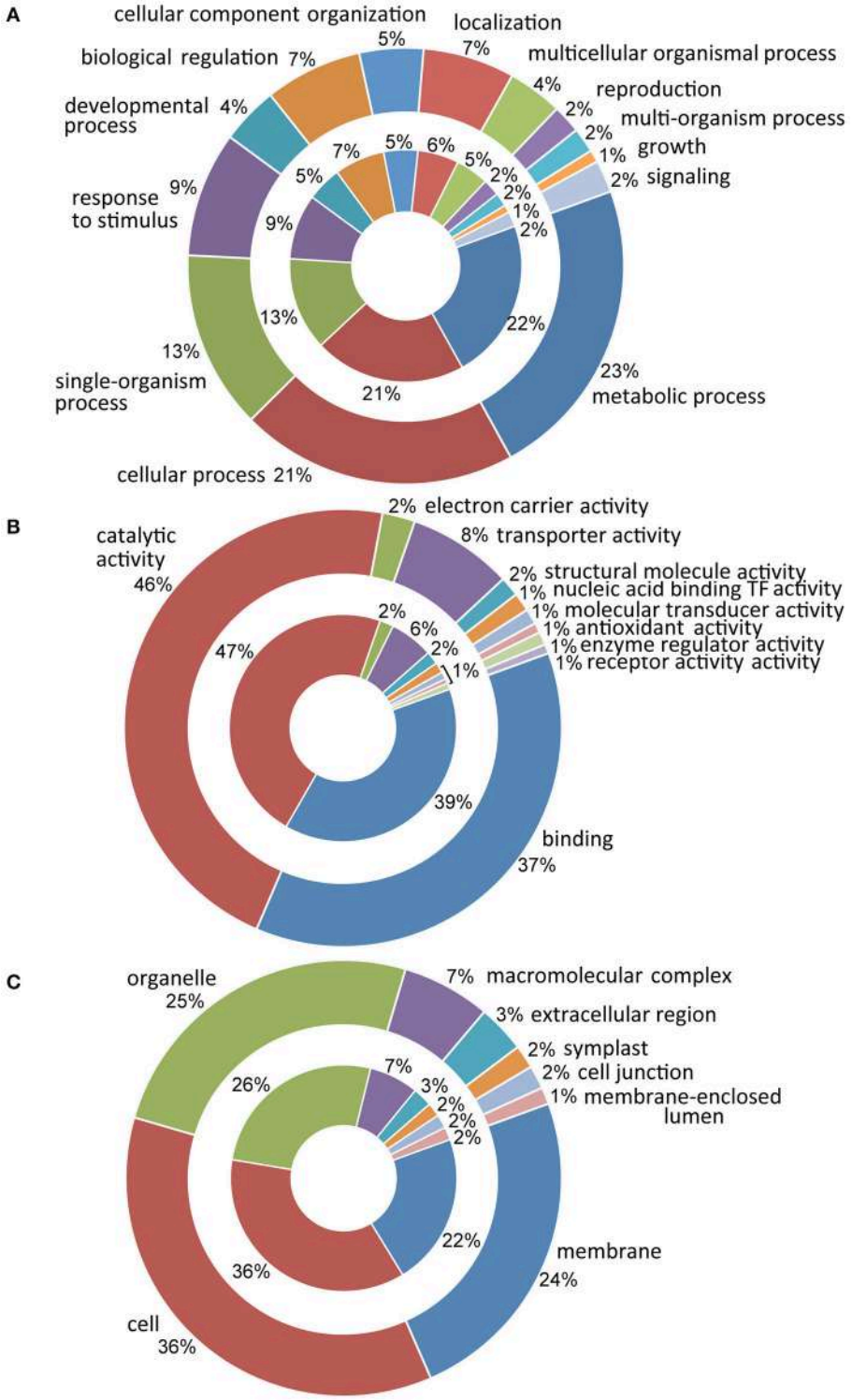


FIGURE 2 | Distribution of ontological categories (level 2 GO terms) in Leucocarpa (inner chart) and Cassanese cvs (outer chart) DE transcripts according to: biological process (A), molecular function (B) and cellular component (C). In A metabolic process and cellular process are the most represented groupings; the divisions relating to catalytic activity and binding are strongly represented in **(B)**, while in C cell, organelle and membrane categories are represented. At the side of anthological categories, the percentage of the transcripts within each class is reported.

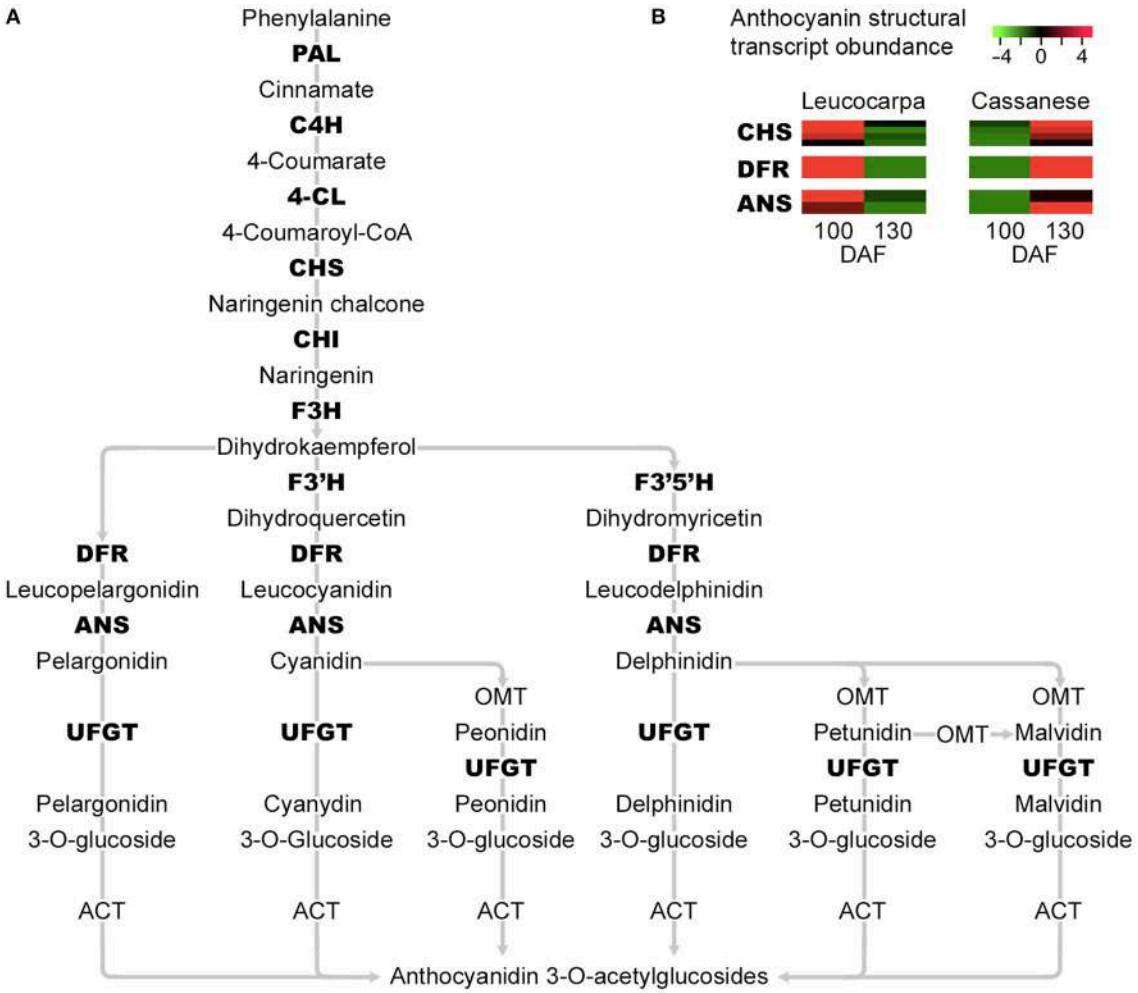


FIGURE 3 | Simplified representation of the main steps in the flavonoid and anthocyanin enzymatic pathways. The transcripts identified as differentially expressed in the 100–130 DAF transition and which show a reduced expression in the "Leucocarpa" epi-mesocarp at 130 DAF and an opposite expression pattern in "Cassanese" epi-mesocarp, that leads at 130 DAF to a normal veraison stage are indicated in bold (A). *Phenylalanine ammonia lyase* (*PAL*), *cinnamate 4-hydroxylase* (*C4H*), *4-coumarate-CoA Ligase* (*4CL*), *chalcone synthase* (*CHS*), *chalcone isomerase* (*CHI*), *Flavonol 3-hydrogenase* (*F3H*), *Flavonol 3′-hydrogenase* (*F3′H*), *Flavonol 3′5′-hydrogenase* (*F3′5′H*), *dihydroflavonol 4-reductase* (*DFR*), *anthocyanidin synthase* (*ANS*), and *UDP-glucose: anthocianidin:flavonoid glucosyltransferase* (UFGT). In **(B)**, the expression abundance of anthocyanin structural genes *CHS*, *DFR,* and *ANS* identified in our whole transcript expressions analysis are highlighted. Each row show the relative expression abundance of transcript clusters; green and red colors are used to indicate the transcript levels four-fold up- and down- regulated, respectively.

Discovery Rate (FDR) value of at most 0.001 and at least four-fold difference in expression values according to the Trinity protocol.

For this purpose, it was possible to identify the DE transcripts sets of each cultivar, during the 100–130 DAF transition from Trinity scripts that leverage the R software. In this context, 3792 and 3064 DE transcripts (of 49,162 and 54,194 total transcripts, respectively) were identified in "Leucocarpa" and "Cassanese." The fold change and the statistical significance values between different developmental stage and cultivar were also estimated.

Trinity facilitates analysis of RNA-seq data, including scripts for extracting transcripts that are above some statistical significance (FDR threshold) and fold-change in expression. To examine expression across multiple samples, the FPKM expression values across samples have been normalize, which will account for differences in RNA composition, afterwards TMM normalization generate a matrix of normalized FPKM values across all samples.

These adjusted library sizes are used to recompute the FPKM expression values. Although the raw fragment counts are used for differential expression analysis, the normalized FPKM values are used below in examining profiles of expression across different samples, each DE set of transcripts was displayed as MA plots (where M = log ratios and A = mean values) (**Figures 1A,B**), volcano plots (**Figures 1C,D**) and clustered heat maps (**Figure 1E**). A correlation matrix (**Figure S2**) for the different developmental stages across cultivars, reveals that samples are more highly correlated within cultivar than between cultivar.

Functional Annotation of Differentially Expressed Transcript Sets

The *in silico* analysis of the entire sets of DE transcripts, performed by querying databases of genes and proteins (NCBI, ExPASy, InterProScan) and the functional annotation software Blast2GO, have allowed for each sequence to be traced back to the gene family and to be annotated according to the terms of the three main Gene Ontology vocabularies (**Figure 2**). Since analyses were conducted on the same organ and developmental stages, in both analyzed cultivars a fairly overlapped distribution of GO terms was observed during the developmental transition. In particular, the most represented ontological categories were membrane (GO:0016020), cell (GO:0005623) and organelle (GO:0043226). Molecular functional categories were strongly represented by terms related to catalytic activity (GO:0003824) with 47 and 46% in Leucocarpa and Cassanese cvs, respectively, followed by binding (GO:0005488) and transporter activity (GO:0005215). Finally, more than 10 categories were identified at the biological process level with metabolic and cellular processes (GO:0008152, GO:0009987), among the groups most represented, highlighting the intense and complex metabolic and regulatory activities during fruit maturation.

In order to trace back to the pathways, such as flavonoids and anthocyanin, (map 00941 and 00942, **Figures S3A,B,** respectively), which were more closely involved in the transition between 100 and 130 DAF, the whole DE transcripts set was examined through the Kyoto Encyclopedia of Genes and Genomes (KEGG). Functional analysis was implemented in Mapman, to focus gene expression changes via Image Annotator. All obtained results are consistent with a down regulation of flavonoid and anthocyanins metabolism in Leucocarpa cv, while an opposite trend was observed in "Cassanese" (**Figure S4**).

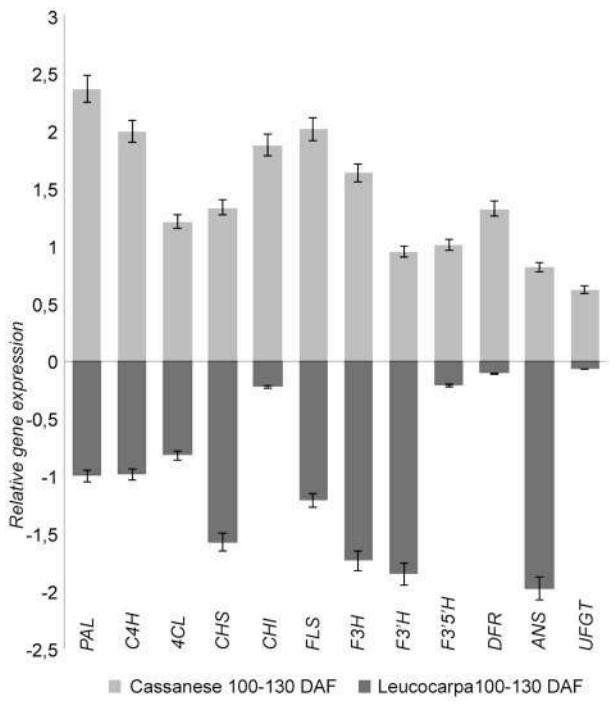


FIGURE 4 | Relative transcripts expression during fruit ripening in Leucocarpa (dark gray) and Cassanese (light gray) cvs. The qRT-PCR results (log fold change) are presented as a proportion of the highest value after normalization with the EF1 house-keeping gene; for each cv 100 DAF samples are used as calibrator. The means ± s.e. of two independent biological replicates are reported.

Gene Expression during Olive Fruits Ripening

We performed a quantitative RNA-seq analisys in a cultivar of *Olea europaea* species, whose fruits are characterized by a switch-off in skin color at full ripeness, to identify the genes involved in flavonoid andanthocyanin biosynthesis.

The transcripts set in flavonoid and anthocyanin pathways were identified in our Illumina datasets. It includes 11 transcripts: phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate-CoA Ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonol 3′-hydrogenase (F3′H), flavonol 3′5′-hydrogenase (F3′5′H), flavonol synthase (FLS), dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS), UDP-glucose: anthocianidin:flavonoid glucosyltransferase (UFGT) (Supplementary data, **Table S1**). Moreover, it was possible to identify different member of MYB, MYC and WD transcription factors related to the regulatory complex that controls anthocyanin structural genes at the transcriptional level (Takahama, 2004).

Interestingly, the quantitative gene expression analysis does not seem to show significant differences during olive fruit development in Leucocarpa and Cassanese cvs (**Table 2**). Indeed, focusing attention on the paths that control the biosynthesis of pigments and the natural reduction of photosynthetic pigments during the veraison stage (Pua and Davey, 2010), the "Leucocarpa" was characterized by a broad down-regulation of CHS, DFR, and ANS transcripts (**Figure 3**), during the 100–130 DAF transition compared to Cassanese cv.

The estimated fold change of the selected genes was also confirmed by quantitative PCR experiments (**Figure 4**). In particular, the expression of transcripts putatively involved in the selected pathway were more highly expressed in "Cassanese" genotype than in "Leucocarpa."

This genome-wide overview on flavonoid and antocyanidin genes also allowed us to select different members of MYB, MYC, and WD transcription factors (TF), within the differentially expressed gene set, linkable to anthocyanins regulatory circuit (Dixon et al., 2005; He et al., 2008; Tian et al., 2008; Stommel et al., 2009; Jaakola, 2013). The abundance estimation analysis made it possible to compare the identified TFs in all analyzed samples. In the "Cassanese" plant, despite a slight decline, the amount of transcripts during 100–130 DAF transition was consistent with the increased anthocyanin structural gene expressions and metabolite accumulation during growth of fruits; whereas in the Leucocarpa cv the identified TFs are primarily characterized by lower expression levels and a general reduction in expression abundance during ripening transition. The differences were most evident when the comparison was carried out at the same stage

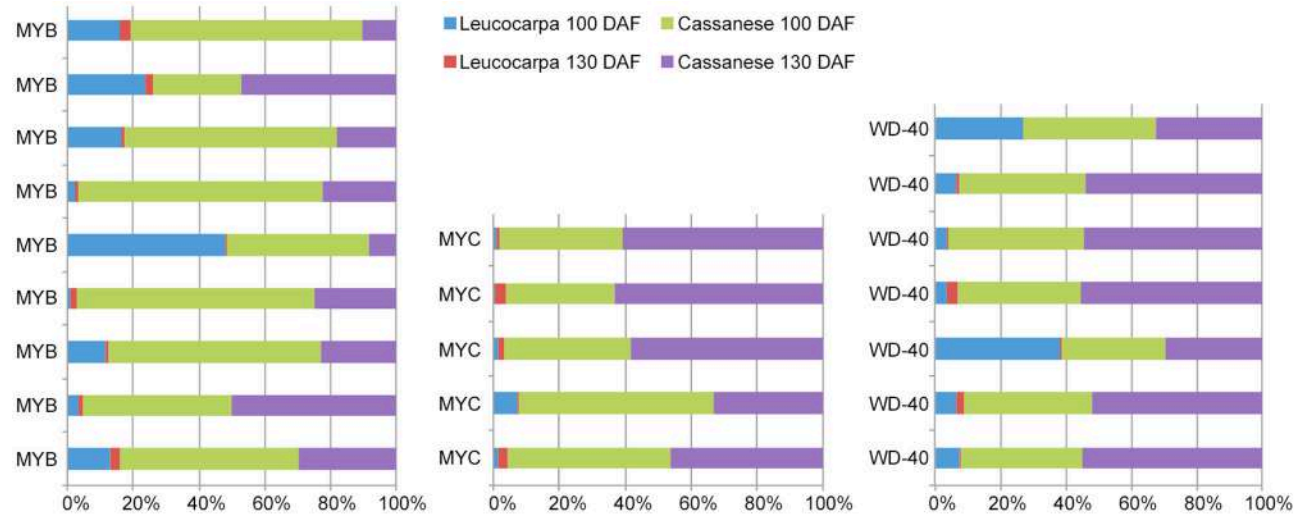


FIGURE 5 | Comparisons in *MYB*, *MYC*, and *WDR* transcripts abundance between samples. Each data is displayed as a stacked bar. Transcripts expression levels were taken from the complete FPKM normalized matrix that were identified as differentially expressed.

(100 or 130 DAF) of maturation. 9 MYB, 5 MYC, and 7 WD TF undergo a decrease in expression during transition, in contrast to Cassanese cv where they appear to participate in the activation pathway (**Figure 5**).

DISCUSSIONS

In the present work we used the Illumina RNA-seq technology to identify the transcripts along flavonoids and anthocyanins biosynthetic pathways and to monitor their expression levels during ripening, by comparing two olive cultivars characterized by different phenological behavior at ripening in terms of anthocyanin accumulation and general pigmentation. We also used a *de novo* transcriptome assembly strategies performed in many plants, including rice, maize, sesame, bamboo, poplar, sweet potato, *Eucalyptus* tree, chickpea, and orchid (Mizrachi et al., 2010; Wang et al., 2010; Fu et al., 2011; Garg et al., 2011; Wei et al., 2011; Zhang et al., 2011a).

The characterization of the genetic entity of olive cultivars has benefited from new molecular biology and high-throughput sequencing methods (Alagna et al., 2009; Galla et al., 2009; Bazakos et al., 2012; Muñoz-Mérida et al., 2013). Through the analysis of massive data it is possible to identify/investigate the genetic pathways that underlie specific, or more general, agronomic traits in the physiological performance of the plants belonging to the *Olea europaea* species.

Between different high-throughput methods, the Illumina sequencing is the best next generation technology, both less costly and more efficient, for transcriptome analysis, if compared with 454 platform, in particular when used in non-model organisms, where genomic sequences are unknown.

Even though this technology has been previously restricted to the re-sequencing of organisms with available reference genomes (Nagalakshmi et al., 2008), its recent improvement has enabled the development of *de novo* strategies for robust trascriptome reconstruction for non-model plants from short reads and their assembly into unigenes.

Through this approach we identified anthocyanin genes, including PAL, C4H, 4CL, CHS, CHI, F3H, F3′H, F3′5′H, FLS, DFR, ANS, from two olive cultivars.

In addition, different transcription factor members with similarity to MYB, MYC, and WD40 family and involved in anthocyanin biosynthesis were also found. Furthermore, the transcripts abundance of identified genes was correlated to the accumulation rate of anthocyanin metabolites.

The anthocyanin biosynthesis pathway has been extensively studied in several plant species, such as petunia, pears, goji berry, bilberry and black raspberry (Jaakola et al., 2002; Zeng et al., 2014). During the ripening progression, many species including the olive tree accumulate anthocyanin in their fruits (Jaakola et al., 2002; Sweetman et al., 2009; Zhang et al., 2011b). In this context, anthocyanins are considered potent marker to monitor ripening stages and organoleptic quality of fruits.

In apple, the regulatory circuit in anthocyanin biosynthesis is tuned by the MYB-MYC-WD40 protein complexes (Ramsay and Glover, 2005; Schaart et al., 2012). Moreover the R2R3-MYB and bHLH TFs are able to activate structural genes, including CHS, DFR and ANS, and ultimately promote anthocyanin accumulation in fruits (Chagné et al., 2013; Umemura et al., 2013; Zeng et al., 2014). In our case the transcripts abundance of MYB, MYC, and WD40-*type* TFs was higher in Cassanese cultivar than in Leucocarpa and was also directly related to anthocyanin accumulation.

In conclusion, the comparative approach performed provide an invaluable resource to identify genes involved in fruit maturation and to define the metabolic pathway and tissue specific functional genomics in non-model plant species. The characterization of transcripts from flavonoid and anthocyanin biosynthetic pathways and the analysis of their expression level in olive fruits is an important goal to understand the veraison

event of fruits and to increase the knowledge on these antioxidant molecules, important for human health.

AUTHOR CONTRIBUTIONS

DI performed research and discussed results. AC designed research analyzed data and discussed result. IM designed research analyzed data and discussed results. All authors contributed to improving the papers and approved the final manuscript.

ACKNOWLEDGMENTS

The authors are very grateful to Dr. Sabrina Micali (CREA-FRU of Roma, Italy) for excellent technical and scientific assistance. This research was supported by the "Certificazione della composizione varietale, dell'origine geografica e dell'assenza di prodotti di sintesi negli oli extravergini di oliva—CERTOLIO 2012-2014" projects and by grant from University of Calabria (MIUR ex 60) to Prof. AC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2015.01246

Figure S1 | "Cassanese" (A–A″) and "Leucocarpa" (B–B″) fruits, in the top and bottom respectively, sampled during 100 (A,B) and 130 (A′,A″,B′,B″) DAF transition.

Figure S2 | Correlation matrix for sample across cultivar: samples are more highly correlated within cultivar than between cultivar.

Figure S3 | KEGG pathway for flavonoid and anthocyanins biosynthesis, map: 00941 (A) and 00942 (B). The isoforms of differentially expressed transcripts controlling flavonoid as well as anthocyanin biosynthesis were mapped.

Figure S4 | MapMan visualization of changes in expression levels of genes associated with secondary metabolism. Green denotes down-regulation and red up-regulation. Changes in **(A,B)** gene expression after 130 DAF compared to 100 DAF as calibrator, in Cassanese and Leucocarpa cvs respectively.

Table S1 | Sequences list and qRT-PCR primers set to validate selected targets.

REFERENCES

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., et al. (2009). Comparative 454 pyrosequencing of transcripts from olive genotypes during fruit development. *BMC Genomics* 10:399. doi: 10.1186/1471-2164-10-399

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

Baldoni, L., Cultrera, N. G., Mariotti, R., Ricciolini, C., Arcioni, S., Vendramin, G. G., et al. (2009). A consensus list of microsatellite markers for olive genotyping. *Mol. Breeding* 24, 213–231. doi: 10.1007/s11032-009-9285-8

Bazakos, C., Manioudaki, M. E., Therios, I., Voyiatzis, D., Kafetzopoulos, D., Awada, T., et al. (2012). Comparative transcriptome analysis of two olive cultivars in response to NaCl-Stress. *PLoS ONE* 7:e42931. doi: 10.1371/journal.pone.0042931

Bruno, L., Chiappetta, A., Muzzalupo, I., Gagliardi, C., Iaria, D., Bruno, A., et al. (2009). Role of geranylgeranyl reductase gene in organ development and stress response in olive (*Olea europaea*). plants. *Funct. Plant Biol.* 36, 370–381. doi: 10.1071/FP08219

Chagné, D., Lin-Wang, K., Espley, R. V., Volz, R. K., How, N. M., Rouse, S., et al. (2013). An ancient duplication of apple MYB transcription factors is responsible for novel red fruit-flesh phenotypes. *Plant Physiol.* 161, 225–239. doi: 10.1104/pp.112.206771

Chalker-Scott, L. (1999). Environmental significance of anthocyanins in plant stress responses. *Photochem. Photobiol.* 70, 1–9. doi: 10.1111/j.1751-1097.1999.tb01944.x

Chiappetta, A., Muto, A., Bruno, L., Woloszynska, M, Van Lijsebettens, M., and Bitonti, M. B. (2015). A dehydrin gene isolated from feral olive enhances drought tolerance in *Arabidopsis* transgenic plants. *Front. Plant Sci.* 6:392. doi: 10.3389/fpls.2015.00392

Coe, E. H., McCormick, S., and Modena, S. A. (1981). White pollen in maize. *J. Hered.* 72, 318–320.

Conde, C., Delrot, S., and Gerós, H. (2008). Physiological, biochemical and molecular changes occurring during olive development and ripening. *J. Plant Physiol.* 165, 1545–1562. doi: 10.1016/j.jplph.2008.04.018

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high- throughput RNA sequencing data analysis. *Brief. Bioinformatics* 14, 671–683. doi: 10.1093/bib/bbs046

Dixon, R. A., Xie, D. Y., and Sharma, S. B. (2005). Proanthocyanidins - a final frontier in flavonoid research? *New Phytol.* 165, 9–28. doi: 10.1111/j.1469-8137.2004.01217.x

Dooner, H. K. (1983). Coordinate genetic regulation of flavonoid biosynthetic enzymes in maize. *Mol. Gen. Genet.* 189, 136–141. doi: 10.1007/BF00326066

Fu, C. H., Chen, Y. W., Hsiao, Y. Y., Pan, Z. J., Liu, Z. J., Huang, Y. M., et al. (2011). OrchidBase: a collection of sequences of the transcriptome derived from orchids. *Plant Cell Physiol.* 52, 238–243. doi: 10.1093/pcp/pcq201

Galla, G., Barcaccia, G., Ramina, A., Collani, S., Alagna, F., Baldoni, L., et al. (2009). Computational annotation of genes differentially expressed along olive fruit development. *BMC Plant Biol.* 9:128. doi: 10.1186/1471-2229-9-128

Garg, R., Patel, R. K., Tyagi, A. K., and Jain, M. (2011). *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* 18, 53–63. doi: 10.1093/dnares/dsq028

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Green, P. S. (2002). A revision of *Olea* L (Oleaceae). *Kew Bull.* 140, 57–91. doi: 10.2307/4110824

Green, P. S., and Wickens, G. E. (1989). *The Olea Europaea Complex*. Edinburgh: Univ Press.

Griesbach, R. J., and Beck, R. M. (2005). Sequence analysis of the chalcone synthase gene in four *Petunia* taxa. *J. Am. Soc. Hort. Sci.* 130, 360–365.

Grotewold, E., and Peterson, T. (1994). Isolation and characterization of a maize gene encoding chalcone flavonone isomerase. *Mol. Gen. Genet.* 242, 1–8.

Gutierrez-Gonzalez, J. J., Tu, Z. J., and Garvin, D. F. (2013). Analysis and annotation of the hexaploid oat seed transcriptome. *BMC Genomics* 14:471. doi: 10.1186/1471-2164-14-471

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hashim, Y. Z., Rowland, I. R., McGlynn, H., Servili, M., Selvaggini, R., Taticchi, A., et al. (2008). Inhibitory effects of olive oil phenolics on invasion in human colon adenocarcinoma cells *in vitro*. *Int. J. Cancer* 122, 495–500. doi: 10.1002/ijc.23148

He, F., Pan, Q-H, Shi, Y., and Duan, C. Q. (2008). Biosynthesis and genetic regulation of proanthocyanidins in plants. *Molecules* 13, 2674–2703. doi: 10.3390/molecules13102674

Holton, T. A., and Cornish, E. C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* 7, 1071–1083. doi: 10.1105/tpc.7.7.1071

Jaakola, L. (2013). New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant Sci.* 18, 477–483. doi: 10.1016/j.tplants.2013.06.003

Jaakola, L., Määttä K, Pirttilä AM, Törrönen, R., Kärenlampi, S., and Hohtola, A. (2002). Expression of genes involved in anthocyanin biosynthesis in relation to anthocyanin, proanthocyanidin, and flavonol levels during bilberry fruit development. *Plant Physiol.* 130, 729–739. doi: 10.1104/pp.006957

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Karp, N. A., and Lilley, K. S. (2007). Design and analysis issues in quantitative proteomics studies. *Proteomics* 7, 42–50. doi: 10.1002/pmic.200700683

Karp, N. A., and Lilley, K. S. (2009). Investigating sample pooling strategies for DIGE experiments to address biological variability. *Proteomics* 9, 388–397. doi: 10.1002/pmic.200800485

Koes, R. E., Spelt, C. E., and Mol, J. N. (1989). The chalcone synthase multigene family of *Petunia hybrida* (V30): differential, light regulated expression during flower development and UV light induction. *Plant Mol. Biol.* 12, 213–225. doi: 10.1007/BF00020506

Koes, R., Verweij, W., and Quattrocchio, F. (2005). Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci.* 10, 236–242. doi: 10.1016/j.tplants.2005.03.002

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lekanne Deprez, R. H., Fijnvandraat, A. C., Ruijter, J. M., and Moorman, A. F. (2002). Sensitivity and accuracy of quantitative real-time polymerase chain reaction using SYBR green I depends on cDNA synthesis conditions. *Anal. Biochem.* 307, 63–69. doi: 10.1016/S0003-2697(02)00021-0

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Liang, C., Liu, X., Yiu, S. M., and Lim, B. L. (2013). *De novo* assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing. *BMC Genomics* 14:146. doi: 10.1186/1471-2164-14-146

Liu, S., Li, W., Wu, Y., Chen, C., and Lei, J. (2013). *De novo* transcriptome assembly in chili pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids. *PLoS ONE* 8:e48156. doi: 10.1371/journal.pone.0048156

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Llorente-Cortes, V., Estruch, R., Mena, M. P., Ros, E., González, M. A., Fitó M, et al. (2010). Effect of Mediterranean diet on the expression of pro-atherogenic genes in a population at high cardiovascular risk. *Atherosclerosis* 208, 442–450. doi: 10.1016/j.atherosclerosis.2009.08.004

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231

Loumou, A., and Giourga, C. (2003). Olive groves: the life and identity of the Mediterranean. *Agric. Hum. Values* 20, 87–95. doi: 10.1023/A:1022444005336

Martinelli, F., and Tonutti, P. (2012). Flavonoid metabolism and gene expression in developing olive (*Olea europaea* L.) fruit. *Plant Biosyst.* 146, 164–170. doi: 10.1080/11263504.2012.681320

Matus, J. T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., Arce-Johnson, P., et al. (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifera*. *J. Exp. Bot.* 60, 853–867. doi: 10.1093/jxb/ern336

Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F., and Myburg, A. A. (2010). *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 11:681. doi: 10.1186/1471-2164-11-681

Morgulis, A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics* 24, 1757–1764. doi: 10.1093/bioinformatics/btn322

Muñoz-Mérida, A., González-Plaza, J. J., Cañada, A., Blanco, A. M., García-López Mdel, C., Rodríguez, J. M., et al. (2013). *De novo* assembly and functional annotation of the olive (*Olea europaea*) transcriptome. *DNA Res.* 20, 93–108. doi: 10.1093/dnares/dss036

Muzzalupo, I. (2012). *Olive Germplasm – Italian Catalogue of Olive Varieties*. Rijeka: InTech

Muzzalupo, I., Macchione, B., Bucci, C., Stefanizzi, F., Perri, E., Chiappetta, A., et al. (2012). LOX gene transcript accumulation in olive (*Olea europaea* L.) fruits at different stages of maturation: relationship between volatile compounds, environmental factors, and technological treatments for oil extraction. *Sci. World J.* 2012, 1–9. doi: 10.1100/2012/532179

Muzzalupo, I., Stefanizzi, F., Perri, E., and Chiappetta, A. (2011). Transcript levels of CHL P gene, antioxidants and chlorophylls contents in olive (*Olea europaea* L.) pericarps: a comparative study on eleven olive cultivars harvested in two ripening stages. *Plant Foods Hum. Nutr.* 66, 1–10. doi: 10.1007/s11130-011-0208-6

Muzzalupo, I., Vendramin, G. G., and Chiappetta, A. (2014). Genetic biodiversity of Italian olives (*Olea europaea*) germplasm analyzed by SSR markers. *Sci. World J.* 2014, 1–12. doi: 10.1155/2014/296590

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

Pallavicini, A., Canapa, A., Barucca, M., Alf Ldi, J., Biscotti, M. A., Buonocore, F., et al. (2013). Analysis of the transcriptome of the Indonesian coelacanth Latimeria menadoensis. *BMC Genomics* 14:538. doi: 10.1186/1471-2164-14-538

Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W., and Stromberg, A. (2003). Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* 4:26. doi: 10.1186/1471-2105-4-26

Pérez-Jiménez, F., Ruano, J., Perez-Martinez, P., Lopez-Segura, F., and Lopez-Miranda, J. (2007). The influence of olive oil on human health: not a question of fat alone. *Mol. Nutr. Food Res.* 51, 1199–1208. doi: 10.1002/mnfr.200600273

Pua, E. C., and Davey, M. R. (2010). *Plant Developmental Biology - Biotechnological Perspectives*, Vol. 1. Heidelberg; Dordrecht; London; New York, NY: Springer.

Ramsay, N. A., and Glover, B. J. (2005). MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends Plant Sci.* 10, 63–70. doi: 10.1016/j.tplants.2004.12.011

Ravaglia, D., Espley, R. V., Henry-Kirk, R. A., Andreotti, C., Ziosi, V., Hellens, R. P., et al. (2013). Transcriptional regulation of flavonoid biosynthesis in nectarine (*Prunus persica*) by a set of R2R3 MYB transcription factors. *BMC Plant Biol.* 13:68. doi: 10.1186/1471-2229-13-68

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Schaart, J. G., Dubos, C., Romero De La Fuente, I., van Houwelingen, A. M. M. L., de Vos, R. C. H., Jonker, H. H., et al. (2012). Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (*Fragaria x ananassa*) fruits. *New Phytol.* 197, 454–467. doi: 10.1111/nph.12017

Schaefer, H. M., Schaefer, V., and Levey, D. J. (2004). How plant-animal interactions signal new insights in communication. *Trends Ecol. Evol.* 19, 577–584. doi: 10.1016/j.tree.2004.08.003

Stommel, J. R., Lightbourn, G. J., Winkel, B. S., and Griesbach, R. J. (2009). Transcription factor families regulate the anthocyanin biosynthetic pathway in *Capsicum annuum*. *J. Am. Soc. Hort. Sci.* 134, 244–251.

Sweetman, C., Deluc, L. G., Cramer, G. R., Ford, C. M., and Soole, K. L. (2009). Regulation of malate metabolism in grape berry and other developing fruits. *Phytochemistry* 70, 1329–1344. doi: 10.1016/j.phytochem.2009.08.006

Takahama, U. (2004). Oxidation of vacuolar and apoplastic phenolic substrates by peroxidases: physiological significance of the oxidation reactions. *Phytochem. Rev.* 3, 207–219. doi: 10.1023/B:PHYT.0000047805.08470.e3

Thimm, O., Bläesing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939. doi: 10.1111/j.1365-313X.2004.02016.x

Tian, L., Pang, Y., and Dixon, R. A. (2008). Biosynthesis and genetic engineering of proanthocyanidins and (iso)flavonoids. *Phytochem. Rev.* 7, 445–465. doi: 10.1007/s11101-007-9076-y

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals

unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tulin, S., Aguiar, D., Istrail, S., and Smith, J. (2013). A quantitative reference transcriptome for *Nematostella vectensis* early embryonic development: a pipeline for de novo assembly in emerging model systems. *Evodevo* 4:16. doi: 10.1186/2041-9139-4-16

Umemura, H., Otagaki, S., Wada, M., Kondo, S., and Matsumoto, S. (2013). Expression and functional analysis of a novel MYB gene, MdMYB110a_JP, responsible for red flesh, not skin color in apple fruit. *Planta* 238, 65–76. doi: 10.1007/s00425-013-1875-3

van Tunen, A. J., Hartman, S. A., Mur, L. A., and Mol, J. N. (1989). Regulation of chalcone flavone isomerase (CHI) gene expression in *Petunia hybrida*: the use of alternative promoters in corolla, anthers and pollen. *Plant Mol. Biol.* 12, 539–551. doi: 10.1007/BF00036968

van Tunen, A. J., Koes, R. E., Spelt, C. E., van der Krol, A. R., Stuitje, A. R., Mol, J. N., et al. (1988). Cloning of the two chalcone flavanone isomerase genes from *Petunia hybrida*: coordinate light-regulated and differential expression of flavonoid genes. *EMBOJ* 7, 1257–1263.

Vauzour, D., Rodriguez-Mateos, A., Corona, G., Oruna-Concha, M. J., and Spencer, J. P. E. (2010). Polyphenols and human health: prevention of disease and mechanisms of action. *Nutrients* 2, 1106–1131. doi: 10.3390/nu2111106

Wallander, E., and Albert, V. A. (2000). Phylogeny and classification of Oleaceae based on rps16 and trnL-F sequence data. *Am. J. Bot.* 12, 1827–1841. doi: 10.2307/2656836

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., et al. (2010). *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genomics* 11:726. doi: 10.1186/1471-2164-11-726

Ward, J. A., Ponnala, L., and Weber, C. A. (2012). Strategies for transcriptome analysis in non model plants. *Am. J. Bot.* 99, 267–276. doi: 10.3732/ajb.1100334

Wei, W., Qi, X., Wang, L., Zhang, Y., Hua, W., Li, D., et al. (2011). Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12:451. doi: 10.1186/1471-2164-12-451

Zeng, S., Wu, M., Zou, C., Liu, X., Shen, X., Hayward, A., et al. (2014). Comparative analysis of anthocyanin biosynthesis during fruit development in two *Lycium* species. *Physiol. Plant.* 150, 505–516. doi: 10.1111/ppl.12131

Zhang, J., Wang, X., Yu, O., Tang, J., Gu, X., Wan, X., et al. (2011b). Metabolic profiling of strawberry (*Fragaria x ananassa* Duch) during fruit development and maturation. *J. Exp. Bot.* 62, 1103–1118. doi: 10.1093/jxb/erq343

Zhang, Y. J., Ma, P. F., and Li, D. Z. (2011a). High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (*Poaceae: Bambusoideae*). *PLoS ONE* 6:e20596. doi: 10.1371/journal.pone.0020596

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 24 December 2015
doi: 10.3389/fpls.2015.01157

Transcriptional Regulatory Network Analysis of MYB Transcription Factor Family Genes in Rice

Shuchi Smita[1,2†], *Amit Katiyar*[1,2†], *Viswanathan Chinnusamy*[3], *Dev M. Pandey*[2] and *Kailash C. Bansal*[1*]

[1] ICAR-National Bureau of Plant Genetic Resources, Indian Agricultural Research Institute, New Delhi, India, [2] Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi, India, [3] Division of Plant Physiology, ICAR-Indian Agricultural Research Institute, New Delhi, India

MYB transcription factor (TF) is one of the largest TF families and regulates defense responses to various stresses, hormone signaling as well as many metabolic and developmental processes in plants. Understanding these regulatory hierarchies of gene expression networks in response to developmental and environmental cues is a major challenge due to the complex interactions between the genetic elements. Correlation analyses are useful to unravel co-regulated gene pairs governing biological process as well as identification of new candidate hub genes in response to these complex processes. High throughput expression profiling data are highly useful for construction of co-expression networks. In the present study, we utilized transcriptome data for comprehensive regulatory network studies of MYB TFs by "top-down" and "guide-gene" approaches. More than 50% of *OsMYBs* were strongly correlated under 50 experimental conditions with 51 hub genes via "top-down" approach. Further, clusters were identified using Markov Clustering (MCL). To maximize the clustering performance, parameter evaluation of the MCL inflation score (I) was performed in terms of enriched GO categories by measuring F-score. Comparison of co-expressed cluster and clads analyzed from phylogenetic analysis signifies their evolutionarily conserved co-regulatory role. We utilized compendium of known interaction and biological role with Gene Ontology enrichment analysis to hypothesize function of coexpressed *OsMYBs*. In the other part, the transcriptional regulatory network analysis by "guide-gene" approach revealed 40 putative targets of 26 OsMYB TF hubs with high correlation value utilizing 815 microarray data. The putative targets with MYB-binding cis-elements enrichment in their promoter region, functional co-occurrence as well as nuclear localization supports our finding. Specially, enrichment of MYB binding regions involved in drought-inducibility implying their regulatory role in drought response in rice. Thus, the co-regulatory network analysis facilitated the identification of complex *OsMYB* regulatory networks, and candidate target regulon genes of selected guide MYB genes. The results contribute to the candidate gene screening, and experimentally testable hypotheses for potential regulatory MYB TFs, and their targets under stress conditions.

Keywords: MYB TF, co-expression, co-regulatory, abiotic stress, rice, network analysis

OPEN ACCESS

Edited by:
Jörg Linde, Leibniz-Institute for Natural Product Research and Infection Biology-Hans-Knoell-Institute, Germany

Reviewed by:
Hemant Ritturaj Kushwaha, International Centre for Genetic Engineering and Biotechnology, New Delhi, India
Mika Gustafsson, Linköping University, Sweden

****Correspondence:***
Kailash C. Bansal
kailashbansal@hotmail.com

†Present Address:
Shuchi Smita, The McFadden Northern Plains Biostress Laboratory, Plant Science Department, College of Agriculture and Biological Sciences, South Dakota State University, Brookings, SD, USA
Amit Katiyar, Department of Biophysics, All India Institute of Medical Science, Ansari Nagar, New Delhi, India

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: *30 September 2015*
Accepted: *07 December 2015*
Published: *24 December 2015*

Citation:
Smita S, Katiyar A, Chinnusamy V, Pandey DM and Bansal KC (2015) Transcriptional Regulatory Network Analysis of MYB Transcription Factor Family Genes in Rice. Front. Plant Sci. 6:1157. doi: 10.3389/fpls.2015.01157

INTRODUCTION

Plants are exposed to several environmental factors and accordingly modulate their growth and development. Excess or deficit of these environmental factors from their optimum levels adversely affect the plant growth and thus crop yield (Gao et al., 2007; Shinozaki and Yamaguchi-Shinozaki, 2007; Bansal et al., 2012). Plants respond and adapt to these cues, through various molecular, biochemical and physiological processes. These processes are regulated by transcriptional regulators which mediate the transcriptional regulation of several effector genes required for stress tolerance. Hence, understanding the regulatory hierarchy of gene expression in response to diverse environmental cues is important to improve the plant processes for enhancing agricultural production.

Systematic analysis of transcriptome data decipher regulatory networks, that helps in identification of candidate genes with certain degree of coordinated expression (Xue et al., 2012; Zhang et al., 2012; Smita et al., 2013). Correlation analyses are useful to identify co-regulated gene pairs in a signal transduction pathway as well as in identifying new candidate genes for specific processes (Gigolashvili et al., 2009; Mounet et al., 2009; Vandepoele et al., 2009). Proteins encoded by highly co-regulated genes are co-localized within the cell and often physically interact with each other. Several gene clustering methods are used to identify functionally coupled genes based on expression similarity (co-expression) levels in a given set of conditions. To study the functional association among genes "guide-gene" and "top-down" approaches are generally used in system biology study. In the guide-gene approach, genes with known functions are utilized to retrieve the correlated genes in the co-expression network, while top-down approach (non-targeted) is used to identify the local module from the large network based on network topology (Patnala et al., 2013). Further, relating these modules to functional enrichment analysis leads to the identification of gene function.

Network approach have been successfully applied in order to analyze correlated genes and hub genes using high throughput expression profiling data (Aoki et al., 2007; Yuan et al., 2008; Cramer et al., 2011; Movahedi et al., 2012). The major progress in molecular genetic analyses led to the identification of several genes and TFs that directly and/or indirectly (i.e., regulated by other pathway product) regulate the plant responses to abiotic stresses (Chinnusamy et al., 2004; Nakashima et al., 2009; Xu et al., 2011). TF genes encompass a considerable portion in plant genome, and can be grouped into different, often large, gene families on the basis of their specific DNA-binding domain. This specific DNA binding domain of TF interacts with target *cis*-elements in the promoter sequence, thereby controlling the expression of the target gene. The MYB domain containing TFs constitute one of the largest TF families in plant kingdom (Qu and Zhu, 2006). The first MYB (myeloblastosis) family of transcription factor identified was the "Oncogene" v-MYB identified in avian myeloblastosis virus (Klempnauer et al., 1982). Three v-MYB-related genes namely c-MYB, A-MYB, and B-MYB were subsequently identified in many vertebrates (Martin and Paz-Ares, 1997; Weston, 1998). *MYB* genes code for TFs with a characteristic 52 amino acid MYB motifs. These TFs contain one to four MYB domain direct repeats termed as R1, R2, R3, and R4 (Du et al., 2009). As their name implies, one R-*MYB* (*MYB*-related), two R-*MYB, three* R-*MYB*, four R-*MYB* have one, two, three, and four repeats, respectively. Each MYB domain has three regularly spaced tryptophan residues that are separated by 18 or 19 amino acid residues, and each domain form helix-turn-helix fold that is crucial for MYB TF–DNA interaction (Saikumar et al., 1990). Among these, two R-MYB (R2R3) are the richest class of *MYB* TF super-family genes in plants (Dubos et al., 2010). The MYB TFs play important role in wide range of biological processes such as cell cycle regulation (Cominelli and Tonelli, 2009), cell proliferation (Xie et al., 2010), developmental processes (Komaki and Sugimoto, 2012), hormone signal transduction (Zhao et al., 2014), and abiotic stress responses (Dai et al., 2007; Liu et al., 2011; Seo et al., 2011; Katiyar et al., 2012) in plants. Several researches have demonstrated the regulatory role especially of R2R3-*MYB* genes in various abiotic stresses responses (Pattanaik et al., 2010; Yun et al., 2010; Du et al., 2012; Zhang et al., 2012).

Advances in high throughput *omics* technologies complemented with comprehensive system biology approaches offers many ways to identify gene networks that operate in a given time or a biological processes. Several TF families have been explored for regulatory network study (Meier et al., 2008; Berri et al., 2009; Lim et al., 2010; Ouyang et al., 2012), while the MYB family network has not been explored in spite of its important roles in several biological processes. In the present study, we applied co-expression network based analysis, to dissect *MYB* transcriptional regulatory networks and their correlated links in rice. Taking into account the role of *MYBs* in diverse biological processes, we selected transcriptome data for five major processes such as developmental stages, abiotic stress response, biotic stress response, hormone signaling, and phosphorus deficiency stress response. Comprehensive correlation approach was employed to answer: (i) how *OsMYBs* network connectivity relates to the significant level of co-expression between *OsMYBs* by top-down approach; and (ii) how transcriptional regulatory network based analysis complementing with *cis*-regulatory elements relates to the putative target genes by guide-gene approach. Thus, the study revealed insight into the discovery of new links and usefulness of characterizing the interacting target genes that lead to the formation of complex transcriptional regulatory network (TRN) in plants.

METHODS

OsMYB Identification and Their Genome-Wide Expression Profiling for Top-Down Approach

MYB domain was retrieved by searching for PFAM-ID PF00249 (MYB domain) as a query in rice genome at TIGR (http://rice.plantbiology.msu.edu/). The non-redundant dataset of *MYB* genes identified in rice genome MSU (release 7) was used as input for further validation by domain search at the Pfam database. Only the longest splice form was selected when

more than one alternative splicing sequence was found for the same locus. These analyses led to the identification of 237 non-redundant *OsMYBs* genes in our study. Further, we discarded the loci lacking MYB-DNA binding domain but annotated as MYB protein family in MSU. Finally, we identified 233 *OsMYBs* genes in rice genome and named these MYBs following the nomenclature scheme suggested for TF genes in grasses (Gray et al., 2009). Affymetrix rice arrays were downloaded from NCBI Gene Expression Omnibus (GEO) (platform: GPL2025). Total fifty Affymetrix rice arrays representing five different conditions abiotic (drought, cold, salt), biotic (Magnaporthe oryzae strain Guy11), developmental stages (embryo, endosperm, root, leaf, and seedling), phosphorus deficiency, and hormone treatment (auxin; indole-3-acetic acid, and benzyl aminopurine) with minimum of two biological replicates were retrieved. The microarray data have been retrieved from NCBI GEO under the accession number of GSE6901, GSE18361, GSE11966, GSE35984, and GSE5167 (**Table S1**). Original.CEL files for were normalized using RMA (Bolstad et al., 2003) a package of the statistical software R-version 2.6.1, part of Bioconductor http://www.bioconductor.org/ (R Development Core, Gentleman et al., 2004). Normalization on total signal was performed using the "Robust Multi-array Average-RMA" method. In brief, gene expression raw data analysis was done using the robust multichip analysis algorithm (RMA) and *t*-test was used to calculate the *P*-value of the expression change of each probe set in each biological perturbation. Differentially expressed genes (DEGs) were identified based on normalized signal intensities of biological replicates for each samples using the limma package (Diboun et al., 2006). Fold change of gene expression was calculated using average signal intensities of biological replicates for each sample. *OsMYBs* were considered to be significantly up/down regulated when the log of expression value is $\geq$1.5 with adjusted $P < 0.05$.

Mapping of probes to gene models were done by searching in the MSU Rice Genome Annotation Project release—7 (based on a new pseudomolecule assembly, Os-Nipponbare-Reference-IRGSP-1.0). Microarray data used in the study were from Affymetrix platform (GPL2025) chip containing 57,381 probe sets, each consisting of 11 pairs of 25-mers probes. The 123 probes designed for bacterial/phage control were not included in further analysis. Particularly, when we searched probes matching for *OsMYBs*—264 probe sets matched for 223 *OsMYB* loci (more than one probes matched with one loci). Out of 223 *OsMYBs,* 219 were mapped to 262 probe sets, while no probe sets for 14 *OsMYBs*. Among 219 *OsMYBs*, 183 *MYB* genes had single probe, while the remaining 36 *OsMYBs* were represented by more than one probe. To avoid ambiguity during analysis, the average expression was calculated for the genes having multiple probes.

Expression Correlation Network Construction

The expression correlations assembled in matrix of all-versus-all *OsMYB* genes were calculated by Pearson correlation coefficient (PCC; *r*-value) that capture the linear relationships between any two given components. Expression correlation data were used for correlation network, where nodes represent genes and edges are correlation coefficient value among gene pair. The network was further visualized and analyzed using Cytoscape version 2.8.3 (Shannon et al., 2003).

Module Detection, Assessment and GO Enrichment Analysis

Highly interconnected genes were identified by best graph partitioning algorithms called Markov Clustering algorithm (MCL) (Van Dongen, 2008). The MCL algorithm is designed specifically for clustering of simple or weighted graphs. The MCL algorithm finds cluster structure in graphs by a mathematical bootstrapping procedure. Since the results of MCL depend heavily on the choice of an inflation parameter (I), we applied MCL to the networks constructed with varied I between 1.1 and 3.0 to identify the functional clusters. Clusters with less than three probesets are often biologically meaningless and were removed.

Further, the evaluation of functionally enriched were done by assessment of gene ontology (GO) term overrepresentation within a cluster, as discussed by Wong et al. (2014). Gene Ontology enrichment analysis was done by "g:Profiler" Gene Ontology enrichment analysis tool (http://biit.cs.ut.ee/gprofiler/) using the hypergeometric distribution adjusted by set count sizes (SCS) for multiple hypothesis correction (Reimand et al., 2011). SCS threshold remove enriched false positive GO terms and prioritizes truly significant results. Each probe IDs were assigned GO term, if it crossed the threshold adjusted *P*-values (SCS) < 0.05. The evaluation of cluster performance using MCL at various *I*-values was determined by calculating the fraction of modules enriched with one annotation at FDR <0.05 (expressed as specificity) and the fraction of annotations enriched in at least one module at FDR < 0.05 (expressed as sensitivity), having at least two genes associated with the enriched annotation (Wong et al., 2013). The specificity and sensitivity values were then summarized as a functional enrichment score, the F-measure, calculated as the harmonic mean between specificity and sensitivity [(2 $\times$ Specificity $\times$ Sensitivity)/(Specificity + Sensitivity)].

Phylogenetic Analysis

Multiple sequence alignment of full OsMYB amino acid sequences was performed by Clustal X 2.0.11 using default parameters. Rooted phylogenetic tree topologies were constructed by the Neighbor-Joining (NJ) method and the distances were obtained using a PAM-like distance matrix. The pairwise deletion and p-distance model parameters were used. Bootstrap test (1000 replicates) was performed to validate the phylogenetic tree. The phylogenetic tree image was displayed with the iTOL programme (http://itol.embl.de/; Letunic and Bork, 2011). In tree view, the branches with >1000 bootstrap were shown as green nodes, while red nodes had >80 but <1000 bootstrap value. Most of the genes with high Bootstrap values shown the evolutionary relatedness of genes with high confidence.

Transcriptional Co-regulatory Network Construction and Inference Using guide-Gene Approach

The transcriptional co-regulatory network was built by RiceFREND database (http://ricefrend.dna.affrc.go.jp/) with hierarchy equal to two and mutual rank was set as five (Sato et al., 2013). The database contains 815 microarray data from various tissues at different developmental stages and plant hormone treatment conditions with the access of single and multiple guide-gene searches. In order to exclude the expression correlation due to the constitutive expression pattern, the correlated genes with weighted PCCs higher than the optimal (0.6) thresholds were only extracted from the database and considered as the putative co-expressed genes.

Cis-Element Enrichment Analysis

PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) was used to predict *cis*-regulatory elements in the promoter region (1 kb upstream from the translational start codon (Lescot et al., 2002). Over representation of *cis*-regulatory elements in promoter region (−1000 bp) were performed by de novo motif finder Multiple EM for motif elicitation tool (MEME; Bailey et al., 2006) with maximum number of motif set to five, $E = 0.01$, minimum motif width 6 and maximum motif width 10.

Subcellular Localization Prediction

Subcellular localization was predicted using consensus results of four localization predictor; Plant-PLoc (version 2) http://www.csbio.sjtu.edu.cn/bioinf/plant/ (Chou and Shen, 2008), (ii) WoLF PSORT http://wolfpsort.org/ (Horton et al., 2007), (iii) CELLO (version 2.5) http://cello.life.nctu.edu.tw/ (Yu et al., 2006), and (iv) GO slim from TIGR-MSU database.

RESULTS

OsMYB Co-regulatory Network Using Top-Down Approach

Retrieval of *OsMYBs* and Transcriptome Data Pre-Processing

By a reiterative database exploration with Pfam-ID PF00249 as a query at TIGR, a total of 237 nucleotide sequences were retrieved from rice genome as putative *OsMYB* genes with at least one MYB domain. These candidate genes were further examined by searching for MYB domain at Pfam database. Based on this, we identified 233 MYB genes and named them following the nomenclature scheme suggested earlier (Gray et al., 2009; **Table S2**). Computational domain analysis of final non-redundant set of 233 *MYB* genes showed the presence of several other functional domains including WD domain, G-beta repeat, response regulator receiver domain, BTB/POZ domain, SWIRM/Zinc finger domain, and MYB-CC type transfactor (LHEQLE motif). In total, 113 *MYB*, 70 *MYB* related, 44 *G2*-like *MYB,* and 6 *ARR-B MYB* genes were identified and mapped on rice chromosomes. We observed the variant density distribution of *MYB* genes on rice chromosomes. It reflects the genome/ tandem duplication and gene amplification of *MYB* over evolutionary time.

Gene regulation in response to a physiological perturbation and those triggered by developmental stages can be inferred by appending one dataset with the other. As MYB has diverse role in stresses as well as developmental stages, we have mined and append genome wide expression data of *OsMYBs* from a total of 50 Affymetrix rice arrays for different conditions *viz.* abiotic (GSE6901), biotic (GSE18361), developmental stages (GSE11966), phosphorus deficiency (GSE35984), and hormone treatment (GSE5167; **Table S1**). Differentially expressed *OsMYBs* were identified based on normalized signal intensities of biological replicates for each sample. About 20% *OsMYBs* showed significant expression change (log fold ≥ 1.5; adjusted $P = 0.05$) in at least one of the experiment (**Table S3**). Gene Ontology enrichment analysis showed that *OsMYBs* differentially expressed were associated with genes involved in the regulation of biological process such as response to freezing, abiotic stress, endogenous stimulus, environmental stimulus, regulation of two-component signal transduction system (phosphorelay), etc., (**Table S4**). The transcriptional responses of *MYB* TFs to several cues clearly indicated the existence of a complex regulatory circuit comprising transcriptional activator as well as repressors. Hence, we utilized and correlated these data for understanding of regulatory network in further analysis.

OsMYB Co-expression Network Construction with Cross-Validated Expression Correlations

The complete expression data of 219 *OsMYBs* (mapped to the probsets; see Section *OsMYB* Identification and Their Genome-wide Expression Profiling for top-down Approach) was further recruited for co-regulatory network analysis. The correlations were measured using log transformed (logarithmic) expression values and co-expression network was built as well as analyzed with Cytoscape (**Table S5A**). The topology for networks was examined at different threshold of PCC. This showed that increasing PCC cutoff value leads to decrease in number of both nodes and edges (**Figure 1A**). It was observed that with increasing the PCC value from 0.85 to 0.90, the number of nodes was reduced by 37.67%, while the number of edges was dropped drastically by 69.46%. This drastic reduction in the number of edges may drop important biological interaction. Hence, to possess relatively large number of nodes and their correlation in the network, we opted 0.85 as stringent PCC cutoff value. For the topology, selecting PCC cutoff 0.85 was confirmed by plotting the number of edges, nodes, and network density as a function of the threshold values. The network density at the governed cutoff was ∼0.027 in co-expression network, and increased thereafter (**Figure 1B**). The network created in this study satisfied the scale free topology (**Figure 1C**; Albert and Barabasi, 2000).

The preliminary co-expression network was constructed by connecting genes with PCC magnitude >0.85 and said to be strongly coexpressed genes (PCC > 0.85; positively co-expressed and < −0.85; negatively coexpressed) (**Figure S1**). Total of 146 (66.67%) *OsMYBs* and 298 correlations in network at 0.85 PCC cutoffs were obtained. Among all correlation, a total of 95.30% paired genes had positive correlation; while 4.69% paired

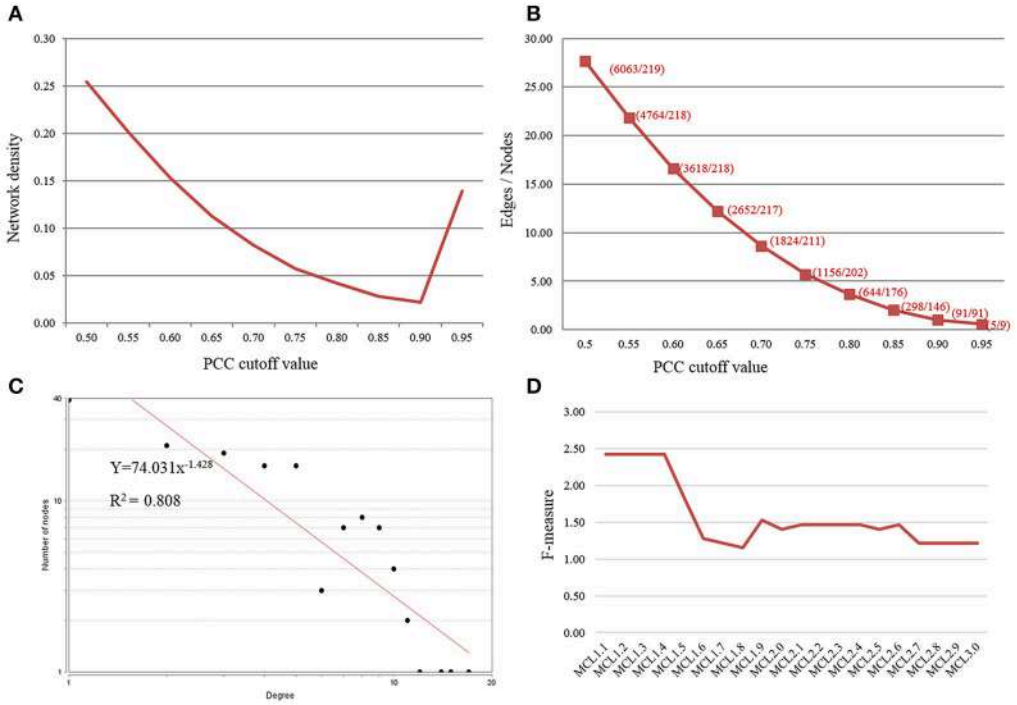


FIGURE 1 | Selection of Pearson correlation coefficient (PCC) threshold value. (A) Plot of number of edges and nodes vs. PCC threshold value. **(B)** Plot of Network density as a function of PCC threshold value. **(C)** Network satisfying scale-free topology showed the node degree distribution following power law ($R^2 > 0.8$). **(D)** Parameter evaluation and optimization of the MCL inflation score (*I*) for cluster performance by F-measure.

genes had negative correlation (**Table S5B**). Genes with positive correlation depict the role of interacting partner in a coordinated manner in similar biological pathway, while genes showing negative correlation might be effective in opposite regulation of genes for a physiological response. This analysis revealed the existence of three major co-regulatory sub-networks with nodes having greater than 3°, in networks (**Figure S1**). Network analyses revealed that 151 out of 219 (68.95%) of the rice *MYB* genes analyzed in this study are coexpressed with diverse degree of connectivity with other *OsMYBs*.

Specificity of Module With GO Enrichment

Grouping of the cluster of coexpressed genes into "modules" also reflects regulatory relationships found in biological systems. One can conclude the function of unknown genes through "guilt by association" with well-characterized genes. We grouped the biologically related coexpressed genes by modular analysis to unravel the underlying functional processes. Several graph clustering methods based on sharing of common functional and expression relatedness are being used in biological science. We subjected the whole *OsMYB* network for module analysis by MCL (Markov Cluster) algorithm (Van Dongen, 2008). This algorithm has an important Inflation parameter (*I*). Higher value for *I* tends to produce a large number of modules but smaller in size. Parameter evaluation and optimization of the MCL inflation score (*I*) is often necessary to maximize clustering performance (the quality of derived GO predictions based on specificity, sensitivity and F-measure; Wong et al., 2013). We examined different inflation values between 1.1 and 3.0. At inflation value 1.1–1.3, no modules were obtained. At *I* value of 1.4 onwards diverse number of modules were obtained in network. Further, relating the largest module to diverse functional categories gives clue to opt the inflation cutoff value. We observed that an MCL *I* parameter of 1.4 produced the best clustering solution in terms of enrichment significance for GO biological process (BP) of most of the cluster and highest *F*-score (see the details in Methods Section; **Table S6**, **Figure 1D**). Therefore, with the inflation value set at 1.4, MCL detected 11 modules in the network with modularity (0.256; **Figure 2**). As node degree distribution, the module size distribution was also observed highly skewed. The largest module had 103 nodes; whereas smallest module had two nodes with one correlated edge in the network. Distribution of hub nodes was observed to be restricted to module 1 only.

We took the modules having more than three correlated edges (i.e., six modules) for modular GO enrichment analysis. The network possesses more number of edges and confers co-regulation of genes even with large differences in expression level. We examined the significant modular GO functional enrichment analysis for six modules using g:profiler tool with cut-off using the hypergeometric distribution adjusted by set count sizes (SCS) $p \leq 0.05$ (**Figure 2**). The module genes were significantly enriched in response to gibberellin stimulus (GO:0009739; g:scs < 6.94E-06), jasmonic acid stimulus (GO:0009753; g:scs < 5.54E-06), hormone stimulus (GO:0009725; g:scs < 1.17E-02), auxin stimulus (GO:0009733; g:scs < 6.27E-03), temperature homeostasis (GO:0001659; g:scs < 2.66E-04), abiotic stimulus

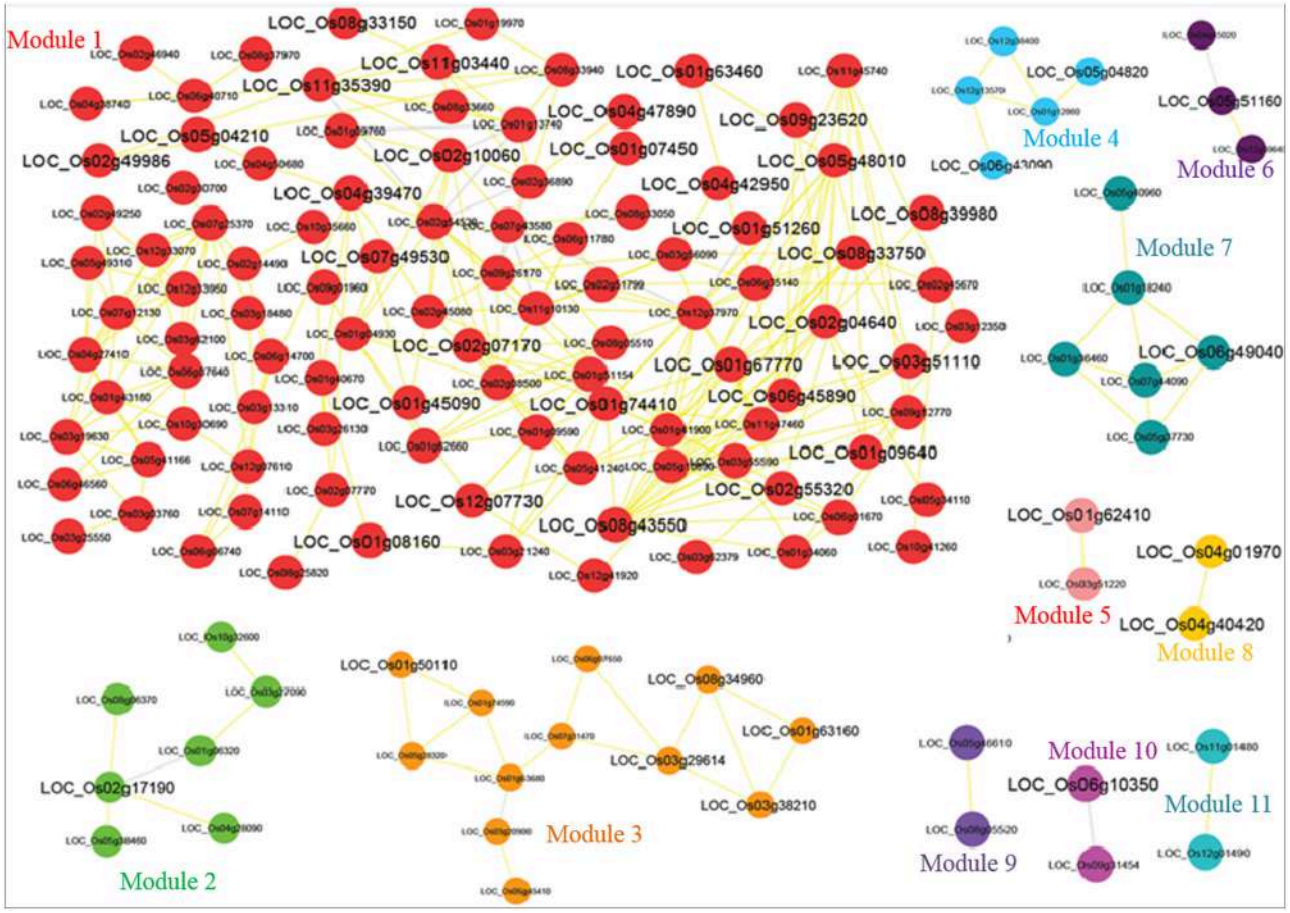


FIGURE 2 | Co-regulatory rice *OsMYB* network identified by top-down approach. The 11 modules are shown in different color. Fonts in larger size indicate differentially expressed *OsMYBs*. The positive and negative correlation value is shown by yellow and gray color edges, respectively.

(GO:0009628; g:scs < 5.10E-04), cold (GO:0009409; g:scs < 2.20E-03), response to freezing (GO:0050826; g:scs < 2.96E-04) etc. with highest significance. *OsMYBs* of module 3 were found to be significantly enriched with GO term positive regulation of response to stimulus (GO:0048584; g:scs < 1.04E-02). Besides, the molecular functions related to DNA binding and nucleic acid binding were significantly enriched. More detailed knowledge about the significant and unique biological processes, molecular functions, and cellular component where the *OsMYBs* act are given in **Table S4**.

Evaluating the Relationship Between Differential Expression and Functional Coherence of a Modular *OsMYBs*

The correlation analysis gave a hint to correlate the significant relationship between regulatory modular *OsMYB* genes and the differentially expressed *OsMYBs*. To investigate this relationship between differentially expressed genes in the network, we assessed topological properties of network and function of *OsMYB* nodes and hubs (labeled in red color in **Figure S1**, **Figure 2**). We observed this kind of relationship especially in 1st, 2nd, and 7th modules. Analysis showed that more than 50% of the genes of module 1 were found to be upregulated under drought conditions. Among them, one pair of *OsMYB*; *LOC_Os09g23620* and *LOC_Os02g04640* was positively correlated (0.80) with each other. We observed that *LOC_Os02g55320* and *LOC_Os01g67770* were positively correlated (0.90) with each other and were found to be up regulated in leaf by more than two-fold with significant enrichment of two-component signal transduction system.

First module gene *LOC_Os03g51110* was found to be upregulated in leaf and down regulated in phosphorous deficiency and significantly enriched with response to organic substance. This gene positively correlated with other upregulated genes in the leaf *viz. LOC_Os08g43550, LOC_Os06g45890,* and *LOC_Os08g33750*. Most of the genes in second modules are induced in leaves, which imply that this module may serve as a tissue specific regulator in rice leaves, whereas some of them were found to be down regulated in root. *LOC_Os11g03440* showed positive correlation with *LOC_Os11g35390*. Interestingly, module 3 contained 12 *OsMYBs* that were found to be negative regulator of leaf and all these genes were found to be correlated with each other. We observed correlation of *LOC_Os01g63160* with two other *OsMYBs* viz. *LOC_Os08g34960*

and *LOC_Os03g38210* genes, while *LOC_Os03g38210* correlates with *LOC_Os03g29614* and *LOC_Os08g34960*.

Assessment of Phylogenetic Conserved Modules

Considering the fact that the knowledge of sequence conservation is additive in identification of coexpressed gene clusters (Elnitski et al., 2006), phylogenetic analysis was performed with the Maximum Likelihood method using all OsMYB protein sequences to infer diverse conserved cluster. The tree revealed six main phylogenetic groups, which were further sub-grouped in to smaller clades based upon the bootstrap values. We then mapped the selected six functionally enriched modules (see Section Specificity of Module with GO Enrichment) on the phylogenetic tree (**Figure 3**). Particularly, genes lie in module 1, 2, and 3 were found to be in different clade with high bootstrap values. This illustration was signifying the sequence conservation of these modules as well as their co-regulatory roles. Majority of the network modules clearly grouped in to different phylogenetic groups suggesting that evolutionarily diverse *OsMYB*s contributing to orchestrate a specific common signal transduction pathway in a network.

All clades identified based on evolutionary relatedness showed the existence of co-expressed *MYB* genes in clusters. Moreover, some of the *OsMYBs* of module 1, 2, 3, and 4 showed strong positive correlation within the whole network module as well as sequence conservation. For example, module 1 gene *LOC_Os12g37970* had significant positive correlation (0.90) with *LOC_Os11g47460* and observed to be evolutionarily conserved in largest phylogenetic group. *LOC_Os07g44090* of module 4 had strong positive correlation (0.90) with *LOC_Os01g18240* and occupied in thirrd phylogenetic cluster. We observed that OsMYB2P-1 (LOC_Os05g04820) protein was close to LOC_Os01g65370, LOC_Os05g3550, and OsMYB4 (LOC_Os04g43680) in 3rd phylogenetic cluster. Specificity of the genes lies in one module as well as together in one phylogenetic clad suggested its evolutionary role in co-regulatory manner.

Hub *OsMYBs* in Regulatory Network Exhibit Biological Significance

Genes with high degree of connectivity either positive/negative correlation was defined as hub genes. In this study, we defined "hubs" as nodes having five and more than five connectivity in the whole network (Patil and Nakamura, 2006; Lu et al., 2007). We found 51 *OsMYBs* as hub genes which were present in network (**Table S5C**). Additionally, candidate hub nodes that were significantly enriched in higher level of biological processes such as signaling were adopted as a factor for potential hub genes in the network. We observed high correlation (positive/negative) among hub nodes themselves. Among 51 hubs, 48 hub *OsMYBs* were significantly enriched with GO term, while three hub genes were not found to be enriched with any GO term. Among 48 hub *OsMYB*s, 17 were significantly enriched with response to salicylic acid stimulus, stimulus, hormone stimulus, jasmonic acid stimulus, gibberellin stimulus, and abscisic acid stimulus related GO biological processes (**Table 1**). Results revealed that nodes pertaining to molecular functions such as DNA binding (GO:0003677; g:scs < 4.29e-32), nucleic acid binding (GO:0003676; g:scs < 1.13e-21), two-component response regulator activity (GO:0000156; g:scs < 2.91E-02), organic cyclic compound binding (GO:0097159; g:scs < 7.12e-14), etc. The details of all 48 hub nodes and significantly enriched GO biological processes were summarized in the **Table S4**.

TABLE 1 | Hub *OsMYB* genes that were significantly enriched with abiotic stress and hormone related Gene Ontology (biological process).

Hub node MSU_ID	Degree
LOC_Os01g13740	8
LOC_Os01g62660	5
LOC_Os01g67770	10
LOC_Os02g08500	12
LOC_Os02g10060	7
LOC_Os02g36890	8
LOC_Os02g54520	14
LOC_Os02g55320	5
LOC_Os03g51110	6
LOC_Os03g55590	5
LOC_Os04g39470	5
LOC_Os05g48010	9
LOC_Os06g01670	5
LOC_Os06g11780	6
LOC_Os07g43580	5
LOC_Os08g43550	15
LOC_Os11g35390	5
LOC_Os12g37970	17

Nodes in bold are differentially expressed in at least one of the condition.

The hub node *LOC_Os12g37970* with highest degree had 17 coexpressed neighbors; 15 positive and 2 negative, with an average correlation 0.88 and 0.86, respectively (**Figure 4**). GO analysis of sub-network of this highest degree node revealed that five nodes are significantly enriched with GO biological processes in response to stimulus and response to hormone stimulus. Among 17 coexpressed *OsMYBs*, six were found to be differentially expressed in at least one of the conditions taken in the present study. Where, three (*LOC_Os01g74410*, *LOC_Os11g47460,* and *LOC_Os07g43580*) were differentially expressed in our previous study under drought condition with more than 1.5-fold change (Katiyar et al., 2012). The function of individual genes was explored on the basis of GO annotation and found to be involved in endogenous stimulus, stress, abiotic, signal transduction pathways for all positively correlated genes. While two pair of genes with negative correlation; first the *LOC_Os07g43580* has role in cell death, lipid metabolic process, biotic stimulus and other one *LOC_Os01g51260* has role in flower development. These data clearly showed that the hub genes and their interacting genes as putative nodes to function in several stresses and hormones signaling pathway.

Abiotic Stress Responsive *OsMYB* Transcriptional Regulatory Network (TRN) by Guide-Gene Approach

Identifying directly co-regulated genes (i.e., genes that are both co-expressed and share conserved upstream regulatory

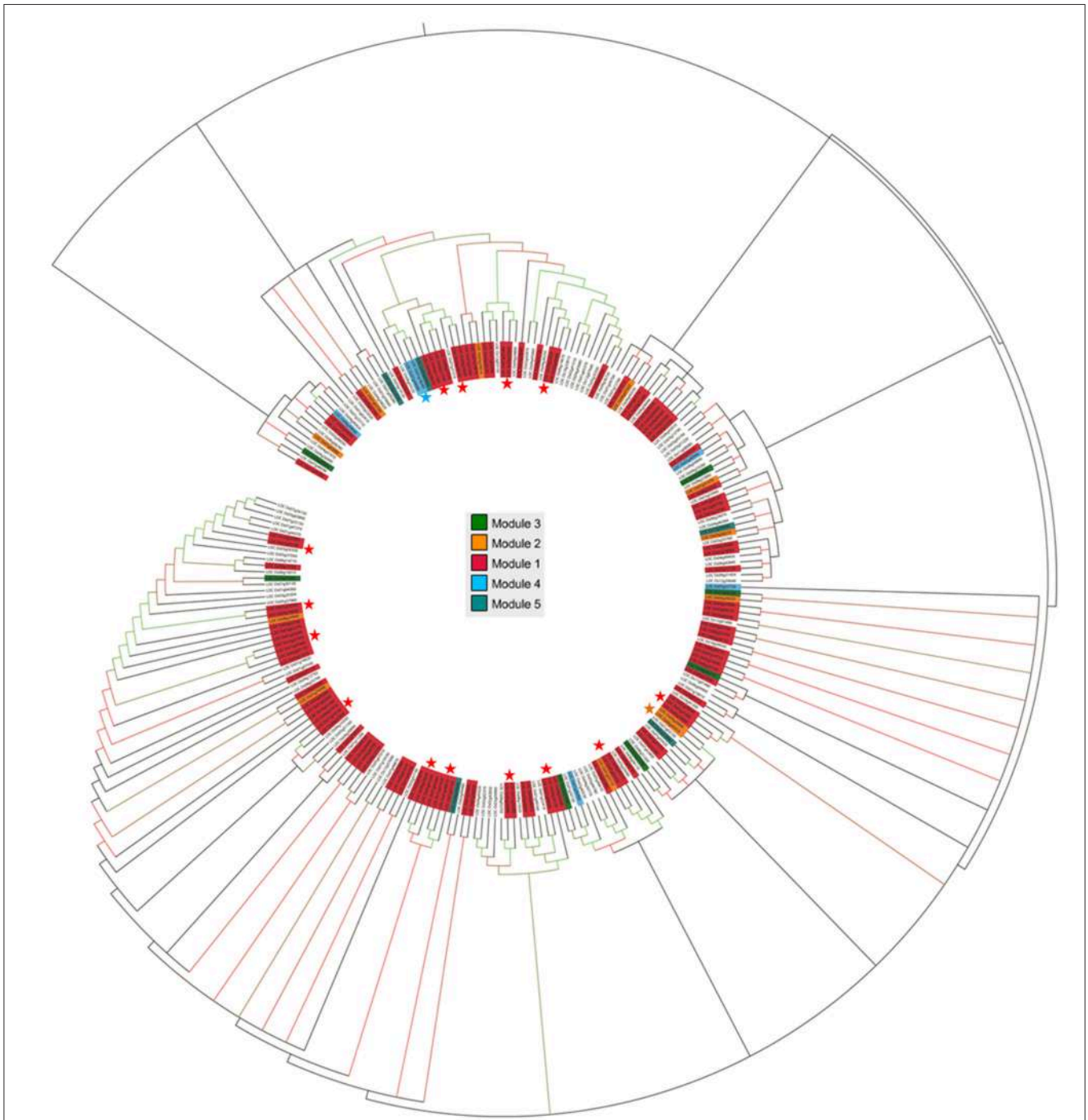


FIGURE 3 | Phylogenetic tree of OsMYB proteins. *OsMYBs* lie in module 1 (highlighted in red), module 2 (in orange), module 3 (in green), and module 4 (in blue), were found to be evolutionary related with high bootstrap values. Gene pair marked with star (*) showed their sequence conservation with high boot strap value as well as coexpression which lies in same module. Bootstrap values higher than 80 are indicated by colored nodes (green nodes with >1000 bootstrap value; red nodes with >80 but <1000 bootstrap value).

sequences) is important for exploring the underlying transcriptional regulatory network and putative target genes (Imam et al., 2015). For this purpose, based on the available biological knowledge, certain *OsMYBs* were selected as guide genes that are known to play key role in a specific biological process. Total of 35 *OsMYB*s were chosen as guide genes to build global co-expression network that included 17 *OsMYBs* with previously known functions and 18 *OsMYB*s with more than two

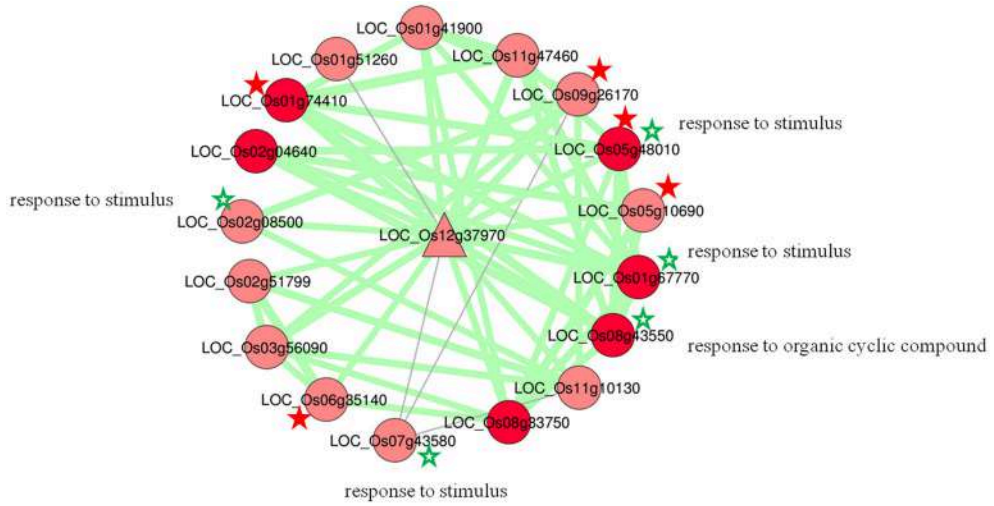


FIGURE 4 | Hub *OsMYB*LOC_Os12g37970 with the highest connection in co-regulatory network. Differentially expressed nodes are in dark red color. Genes with abiotic stress related GO term are marked with green star (*). Characterized genes having role in abiotic stress response are highlighted with red star (*).

fold up-regulation under drought conditions in our previous study (Katiyar et al., 2012; **Table S7A**). The transcriptional regulatory networks have two types of nodes namely "TFs hub" and putative target genes. We employed recently published RiceFREND co-expression tool that contains microarray data for abscisic acid, gibberellins, jasmonic acid, developmental stages, etc., for co-expressed gene identification based on mutual ranking. Since hormones play significant role in adaptive response of plants to abiotic and biotic stresses, we opted RiceFREND database with multiple guide genes search option to understand the underlying transcriptional regulatory network. The resulting regulatory networks derived from this analysis contained a total of 163 correlated nodes (TFs and putative target genes) with 158 correlations that include 24 guide genes with cutoff of weighted PCC $>$ 0.6 and mutual rank $<$5 (**Figure 5**; **Table S7B**).

The GO enrichment analysis of target genes showed that significant enrichment of biological processes such as response to abiotic stimulus (GO:0009628; g:scs $<$ 1.25E-02), response to salicylic acid stimulus (GO:000975; g:scs $<$ 5.36E-04), response to ethylene stimulus (GO:0009723; g:scs $<$ 2.24E-02) response to gibberellin stimulus (GO:0009739; g:scs $<$ 1.12E-03), etc. Interestingly as expected, the molecular function enrichment showed the term DNA binding (GO:0003677; g:scs $<$ 1.04E-03) with highest enrichment. The cellular component showed the nucleus (GO:0005634; g:scs $<$ 6.04E-08), intracellular organelle (GO:0043229; g:scs $<$ 2.10E-03) with highest enrichment (**Table S7C**).

Co-regulated Drought Responsive Putative Target Genes Of *OsMYBs*

Most of the guide *OsMYB* genes in the network were found to be involved in drought response and hence, the coexpressed genes were analyzed for the presence of drought response (or abiotic stress related) regulatory elements in their promoters. As shown in **Figure 5**, transcriptional regulators based on coordinated expression and over representation of the *cis*-elements associated with the *OsMYB* in putative target genes may support our finding. For this purpose, *OsMYB* co-regulatory network was further analyzed for similar promoter *cis*-elements. A total of 53 genes as a direct neighbor of 26 guide *OsMYBs* were found. Localization prediction showed that the majority of the co-regulated MYB TF-target pairs have nuclear localization. The presence of nuclear localization signal and GO cellular location in MYB TFs and their target genes suggest that these pairs are not only co-expressed but also localized in the same cellular (nucleus) location. Further, this suggests their putative physical interactions and function in the same signaling/gene expression pathway.

The results encouraged us to identify putative targets of guide *OsMYB* genes having MYB binding *cis*-elements in their promoter region. Interestingly, we observed around 40 (75%) putative target genes with at least one MYB binding region in their promoter region (**Table 2**). Remarkably, among all 40 putative targets, 27 (∼67%) were found to be enriched with 44 MYB binding regions involved in drought-inducibility (MBS; CAACTG, and TAACTG), implying their regulatory role in drought response. Among 27, nine were annotated as unknown proteins having MYB binding *cis*-element in their promoter. Furthermore, MYB binding site involved in light responsiveness (MRE; AACCTAA) and flavonoid biosynthetic gene regulation (MBSII; AAAAGTTAGTTA) were also found to be enriched in the putative target genes. The results suggested the multiple functionality of *MYB* targeting genes which have association with abiotic stress, function in light signaling, flavonoid biosynthesis and circadian control (Kuno et al., 2003; Dubos et al., 2010).

Along the MYB binding site involved in these processes, several other *cis*-elements were also found in good frequency.

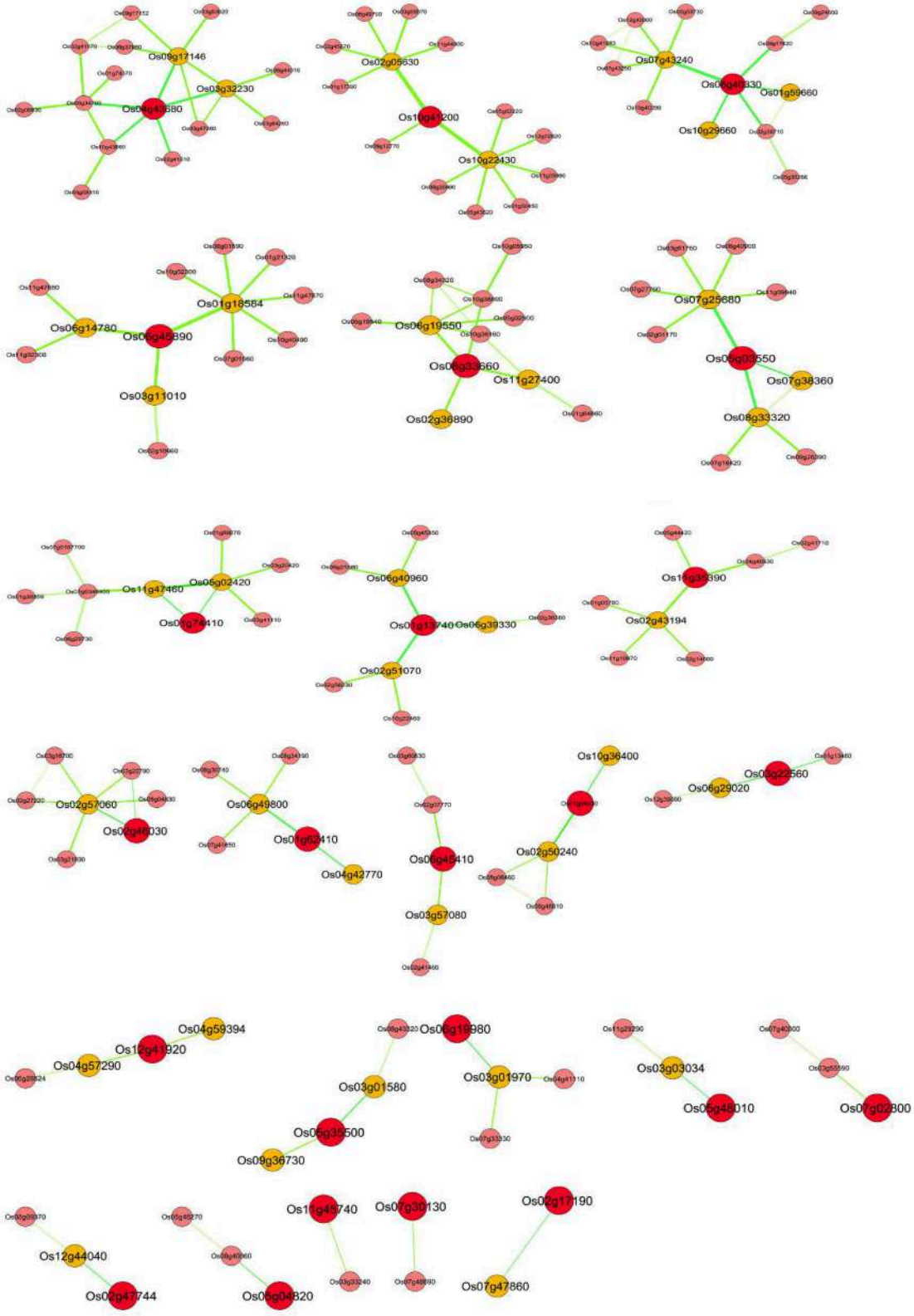


FIGURE 5 | *OsMYB* transcriptional co-regulatory network constructed using guide-gene approach. The co-expression network of 24 reported drought responsive genes as guide *OsMYBs* (RAP_ID, enlarged red circle); their putative first neighbor target gene (RAP_ID, orange circle) based on integrative analysis of coexpressed gene and over representation of target promoter motif enrichment with that class of transcription factor. A link between two nodes indicates direct interaction with PCC > 0.64 and MR <10. The thickness and brightness of the edges represents the confidence of the interaction.

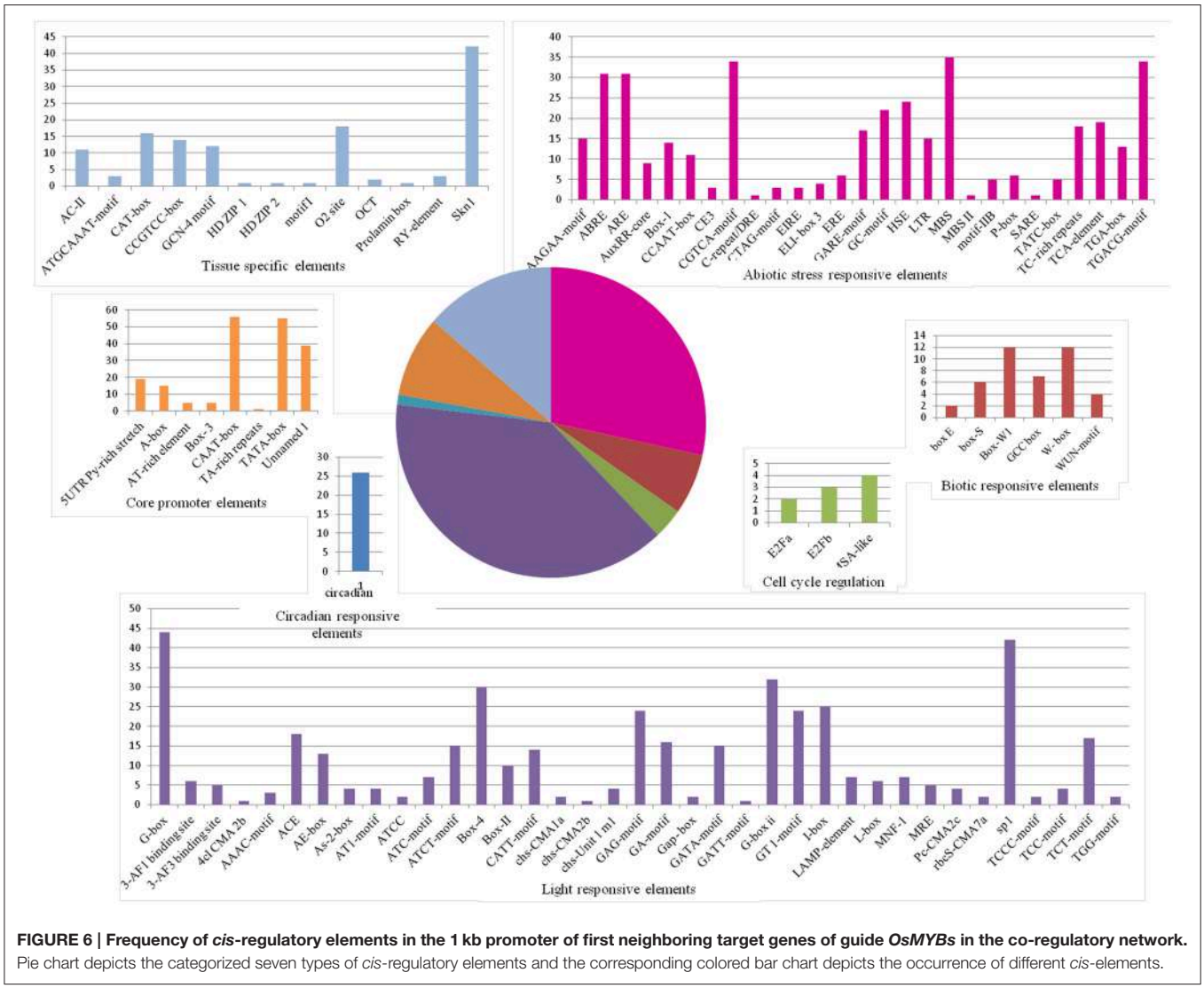


FIGURE 6 | Frequency of *cis*-regulatory elements in the 1 kb promoter of first neighboring target genes of guide *OsMYBs* in the co-regulatory network. Pie chart depicts the categorized seven types of *cis*-regulatory elements and the corresponding colored bar chart depicts the occurrence of different *cis*-elements.

We categorized all the *cis*-elements in the seven broad categories on the basis of responsiveness for any perturbation (**Figure 6**). We observed the enrichment of light, abiotic stress and tissue specific *cis*-elements in the promoter region of first neighbor target of guide *OsMYBs*. Detailed promoter content has been summarized in **Table S8A**. Furthermore, the position of 44 MYB binding region involved in drought-inducibility revealed distinct patterns of sites related to proximal/distal location with respect to transcription start site (TSS). Majority of them (up to 75%) are far from TSS (∼200 bp) indicating their distal type of gene expression regulation. Furthermore, the enrichment analysis of motif in 1000 bp promoter region performed by using MEME with minimum motif width 8 and maximum motif width 10 with *E*-value set to 0.01 (**Table S8B**). Results showed that four motifs were highly conserved in 186 sites in maximum of the target promoter sequences (**Figure 7**). Interestingly, we found CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) motif which has been reported to be binding region of CCA1 MYB-related transcription factor (Wang et al., 1997). It supports our findings that these target genes identified in global co-regulatory network are putative and a researchable area in future.

Consideration of the phylogenetic conservation of binding sites of the promoter elements can enhance the accuracy and have a higher likelihood of being functional *in vivo* (Elnitski et al., 2006). This approach relies on the principle that biologically important TF-binding sites are more likely to be conserved during evolution (Harbison et al., 2004; Dieterich et al., 2005). Therefore, relationship between phylogenetically conserved 1 kb promoter region of all correlated gene pair in the global network and modules were investigated (**Figure S2**). Results showed the evolutionary conservation of several pair of correlated genes. Co-regulated genes with MYB binding regions were examined for evolutionary conservation. Results showed the presence of putative target genes having MYB binding *cis*-element from module 2; 6–10 were evolutionarily

TABLE 2 | The guide *OsMYB* genes and their first neighbor as putative target with MYB binding *cis*-elements within 1 kb upstream promoter region.

Guide gene	First neighbor gene; PCC	MYB binding related *Cis*-elements*	Strand	Position
Os04g43680 (MYB family transcription factor, OsMYB4)	Os03g32230 (ZOS3-12—C2H2 zinc finger protein); 0.7	TAACTG	+	521
		CAACTG	+	566
	Os09g17146 (unknown protein); 0.7	TAACTG	+	582
		CGGTCA	–	941
		TAACTG	–	688
		AACCTAA	–	497
Os10g41200 (Transcription factor MYBS3, OsMYBS3)	Os02g05630 (protein phosphatase 2C, putative); 0.7	TAACTG	+	729
	Os10g22430 (gibberellin response modulator protein); 0.7	CAACTG	–	641
Os06g45890 (MYB family transcription factor)	Os01g18584 (WRKY9); 0.8	CAACTG	+	27
	Os03g11010 (natural resistance-associated macrophage protein); 0.7	TAACTG	+	51
		CAACTG	+	516
		TAACTG	+	128
	Os06g14780 (unknown protein); 0.7	TAACTG	–	754
Os06g40330 (GAMYB-like1)	Os01g59660 (GAMyb); 0.7	CGGTCA	+	222
		CGGTCA	–	479
	Os10g29660 (TFIID, TATA-binding protein); 0.7	CAACTG	–	115
	Os07g43240 (SKP1-like protein 1B); 0.7	TAACTG	–	191
		CAACTG	–	286
Os05g03550 (MYB family transcription factor)	Os07g25680 (protein kinase domain containing protein); 0.7	CAACTG	+	905
		AACCTAA	+	757
	Os07g38360(unknown protein); 0.7	CAACGG	+	691
	Os08g33320 (unknown protein); 0.7	AACCTAA	–	210
Os08g33660 (MYB family transcription factor)	Os02g36890 (MYB family transcription factor); 0.6	CGGTCA	–	377
	OS10g38800 (leucine-rich repeat transmembrane protein kinase); 0.7	CGGTCA	–	196
		CGGTCA	+	360
	Os11g27400 (Glycoside hydrolase); 0.7	CAACTG	+	16
		TAACTG	–	358
		TAACTG	–	278
	Os06g19550 (Short-chain dehydrogenase/reductase SDR domain containing protein); 0.7	CGGTCA	–	377
Os01g74410 (MYB59)	Os11g47460 (MYB family transcription factor); 0.8	CAACGG	–	364
		TAACTG	–	61
		CAACTG	+	790
	Os05g02420 (unknown protein); 0.8	CAACGG	–	314
Os01g13740 (MYB family transcription factor)	Os06g39330 (UDP-glucuronosyl/UDP-glucosyltransferase family protein); 0.7	AACCTAA	+	257
	Os06g40960 (ZOS6-05 - C2H2 zinc finger protein); 0.7	TAACTG	+	77
	Os02g51070 (Starch synthase isoform zSTSII-2); 0.7	CAACGG	+	525
		CAACTG	–	288

(Continued)

TABLE 2 | Continued

Guide gene	First neighbor gene; PCC	MYB binding related *Cis*-elements*	Strand	Position
Os11g35390 (MYB family transcription factor)	Os02g43194 (Aldehyde dehydrogenase); 0.7	CGGTCA	+	124
		CAACTG	+	555
		CGGTCA	–	267
		TAACTG	–	909
Os02g46030 (OsMyb1R)	Os02g57060 (OsCttP2 - Putative C-terminal processing peptidase homolog); 0.8	CAACGG	+	792
Os01g62410 (OsMYB3R–2)	Os04g42770 (unknown protein); 0.6	CAACGG	–	326
		CAACGG	+	345
		TAACTG	+	413
		CAACTG	–	491
		CGGTCA	+	458
		TAACTG	–	823
	Os06g49800 (ubiquitin interaction motif family protein); 0.6	CAACTG	–	744
		CAACTG	+	755
Os06g45410 (MYB family transcription factor)	Os03g57080 (PLA IIIA/PLP7, Patatin-like phospholipase family protein); 0.6	CGGTCA	–	141
		CAACTG	+	921
Os01g04930 (OsMYB2)	Os10g36400 (GIL1); 0.6	TAACTG	+	808
	Os02g50240 (glutamine synthetase, catalytic domain containing protein); 0.7	TAACTG	–	471
Os03g22560 (MYB family transcription factor)	Os06g29020 (retrotransposon protein); 0.6	CAACGG	+	515
		CGGTCA	+	189
Os06g19980 (MYB family transcription factor)	Os03g01970 (THO complex subunit 1); 0.8	TAACTG	–	555
		CAACTG	+	689
Os05g35500 (MYB family transcription factor)	Os09g36730 (P-type R2R3 Myb protein); 0.6	CAACTG	–	75
	Os03g01580 (unknown protein); 0.6	CAACTG	–	75
Os12g41920 (Similar to Single myb histone 6)	Os04g59394 (unknown protein); 0.7	TAACTG	+	67
	Os04g57290 (OsFBX153 - F-box domain containing protein); 0.6	TAACTG	–	700
		CAACTG	+	925
Os02g47744 (MYB family transcription factor)	Os12g44040 (transposon protein); 0.7	TAACTG	–	797
		AAAAGTTAGTTA	+	786
Os05g48010 (OsMYB55)	Os03g03034 (flavonol synthase/flavanone 3-hydroxylase); 0.6	TAACTG	+	533
Os07g30130 (Myb, DNA-binding domain containing protein)	OS07g48690 (DUF630/DUF632 domains containing protein); 0.7	TAACTG	–	328
Os02g17190 (Myb, DNA-binding domain containing protein)	Os07g47860 (tRNA synthetase); 0.7	CAACTG	+	286

**Seven types of MYB binding cis-elements were present—CAACGG, (CCAAT-box; MYBHv1 binding site); AACCTAA, (MRE; MYB binding site involved in light responsiveness); MBSII, (AAAAGTTAGTTA; MYB binding site involved in flavonoid biosynthetic genes regulation); TAACTG, (MBS; MYB binding site involved in drought-inducibility); CAACTG, (MBS; MYB binding site involved in drought-inducibility); CGGTCA, (MBS; MYB Binding Site).*

conserved. Thus, the analysis performed via top down and guide gene approaches in this study identified the highly correlated hub *OsMYBs* and drought responsive putative target genes of *OsMYBs*. Several uncharacterized hub genes as well as co-expressed genes with guide genes annotated as unknown proteins in co-expression network represent high confidence candidate regulator awaiting further examination and validation *in vitro*.

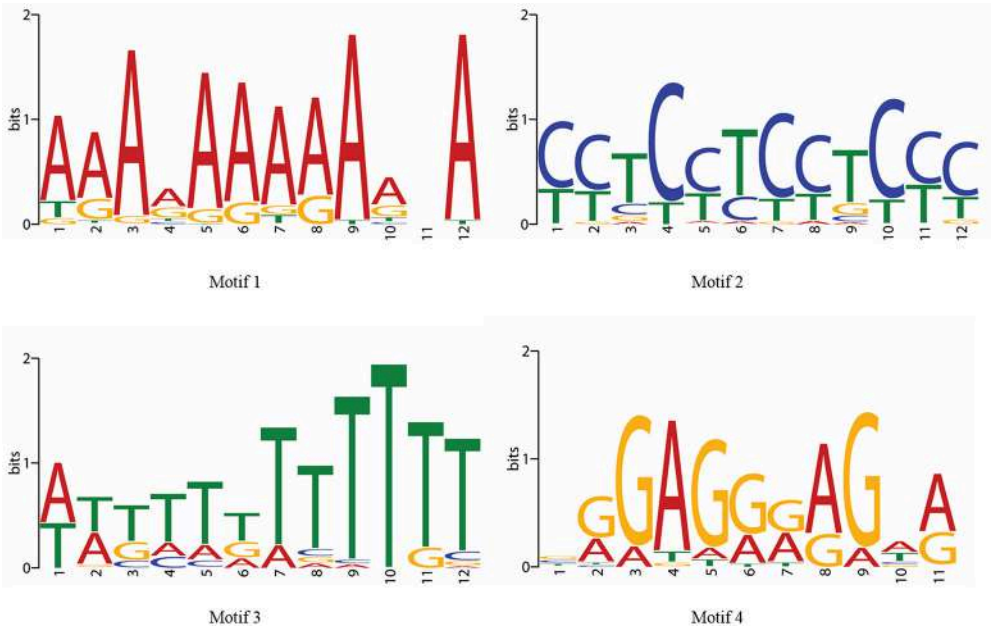


FIGURE 7 | Four enriched motifs logo in the 1 kb promoter region of first neighboring target genes of guide *OsMYBs* in the co-regulatory network.

DISCUSSION

Inferring Function of Candidate *OsMYBs* in Co-expressed Modules

In this study, we carried out transcriptome analysis of *OsMYB* gene family in different abiotic, biotic, hormone stress and developmental stages to identify underlying regulatory network. The *OsMYBs* were first analyzed for their differential expression and putative functions. We found, *OsMYBs* differentially expressed were associated with genes involved in the regulation of biological process such as response to freezing, abiotic stress, endogenous stimulus, environmental stimulus, regulation of two-component signal transduction system (phosphorelay), The two-component system has been shown to play an important role in response to environmental stimuli and growth regulation (Hwang and Sheen, 2001; Du et al., 2007).

The subset of genes that are differentially expressed in particular sample are also observed to be correlated with each other in a co-expression network (Cho et al., 2012). In the *OsMYBs* network of co-expressed genes identified, from the function of known gene in the network, the potential function the co-expressed genes may be inferred and could be selected as candidates for functional verification by *in vivo* approaches. The preliminary gene network of *OsMYBs* was constructed with the relative stringent thresholds to reduce false connections. Module identification and comparison with DEGs showed, correlated *OsMYB* pair in 1st, 2nd, and 7th modules was also differentially regulated under any stress conditions taken in consideration (**Figure 2**). GO enrichment assessment of the modules revealed the significant enrichment of term related to abiotic stress related responses. Some of the candidate genes correlating with already characterized genes for a particular condition showed their role in similar biological pathways as extracted by GO analysis also. Taken together, the coexpression results largely confirm results from previous studies and provided additional clues into the complex molecular mechanism of *OsMYBs*. *OsMYB3R-2* (*LOC_Os01g62410*) was found to be differentially expressed in drought and had positive correlation with *LOC_Os03g51220* which was found to be involved in biosynthetic process. *OsMYB3R-2* is known to confer tolerance to freezing, drought, and salt stresses in transgenic Arabidopsis (Ma et al., 2009). Several predicted *OsMYBs* were activated at early response mechanism in chilling stress (Yun et al., 2010).

LOC_Os06g45410 positively correlated with *LOC_Os03g20900* and has role in biosynthetic processes (**Table S5B**). In a previous study, it was shown that *LOC_Os03g20900* has a positive correlation (0.80) with *OsATG6a* which is involved in abiotic stress (heat, cold, and drought) and abscisic acid responses (Rana et al., 2012). The *MYB* genes have been studied for their cross talk in abiotic stress and hormone regulated gene expression (Peleg and Blumwald, 2011). ABA and auxin responses were regulated by ABI5-like1 (ABL1), a bZIP transcription factor, and the expression of *LOC_Os05g04820* was changed in *abl1* mutant (Yang et al., 2011). In our study, we observed its positive correlation with *LOC_Os01g12860*. A large number of TFs interact with calmodulin (CaMs) to mediate both biotic and abiotic stress responses (Laluk et al., 2012). Recently, several putative *OsMYBs* have been reported to interact with calmodulin (Chantarachot et al., 2012). In our study, we found correlation of CaM binding MYBs i.e.,

LOC_Os05g04210, LOC_Os11g45740 and *LOC_Os01g45090* with other *OsMYBs*. GO slim analysis revealed that the participation of first two genes (*LOC_Os05g04210* and *LOC_Os11g45740*) in response to abiotic stimulus and all trios in response to endogenous stimulus. In consistent with previous study, several *OsMYBs* of module were previously shown to play significant role in activation of immune response, regulation of response to stress as well as in defense response signaling pathway (Glazebrook, 2001). Module 1 genes pair were upregulated in leaf and significant enrichment of two-component signal transduction system. The two-component signal transduction system plays central role in cytokinin signaling and growth (Skerker et al., 2008; Schaller et al., 2011). Recently, it has been reported that the substantial difference in hormone signaling in several response regulators due to variation within their MYB-like DNA binding motif (Tsai et al., 2012). Hence, the correlated *OsMYB* genes may be good candidates for functional characterization of their role in abiotic stress and hormone responses.

Further identifying the hub nodes showed 51 hubs *OsMYB* in our study. These hub genes might have important roles in organizing the functional modules (Barabási and Oltvai, 2004). Some of the high degree functionally characterized hub genes such as *OsMYBS1* (*LOC_Os01g34060*), *OsMYBS2* (*LOC_Os10g41260*), and *OsMYBS3* (*LOC_Os10g41200*) have been studied previously and found to mediate sugar and hormone regulation of α-amylase gene expression (Lu et al., 2002). Moreover, *OsMYB3* is known to be essential for conferring cold tolerance to rice plants (Su et al., 2010). Another *OsMYB55* (*LOC_Os05g48010*) with 9° has been shown to confer high temperature stress tolerance and modulation of amino acid metabolism (Wahid et al., 2007). A highest hub node *LOC_Os12g37970* with 15 positively coexpressed MYB genes with their enriched GO terms "response to stimulus" and "hormone stimulus" as well as differential expression pattern suggest their function in stress and hormone signaling pathway (**Figure 4**). Where two negatively coexpressed OsMYBs with the hub genes showed their function in flower development, cell death and lipid metabolic process. That shows, environmental stress lead to the modulation in flower development and cell death might be due to (reactive oxygen species) ROS formation (Petrov et al., 2015).

Interestingly, we look at numerous scientific reports demonstrated the characterized genes in stress signal pathways from this highest hub cluster (**Figure 4**). Some of the correlated OsMYBs with this highest hub genes such as *LOC_Os01g74410* has been characterized for significant improvement in tolerance to drought and salinity stresses in rice (Xiong et al., 2014). The ortholog of *LOC_Os01g74410* i.e., TaMYB13-1 was also evidenced as transcriptional activator for fructan synthesis that known as protecting agent for drought and cold stress (Xue et al., 2011). The other coexpresssed *LOC_Os01g51260* corresponds to the *Arabidopsis* MYB TF *AT3G13890* that known to be activator of secondary wall thickening (Yang et al., 2007) and *LOC_Os08g33750* ortholog in maize for ethylene-induced lysigenous aerenchyma formation under aerobic conditions (Takahashi et al., 2015). Another positive correlated *OsMYB LOC_Os09g26170* was recently study for their significant role in MG-response and stress-responsive signal transduction pathways. (Kaur et al., 2015). Remarkably, two of the correlated 561 genes *LOC_Os05g10690* and *LOC_Os05g48010* were patented for enhancing yield-related traits in plants by modulating expression in a plant (Molinero, 2013). Hence, we hypothesize this high hub gene cluster have specific role in regulation of stress tolerance, in particular in defense mechanism as well as in crop yield improvement. And thus characterization of some uncharacterized MYB TF from this cluster can be a promising future direction.

Phylogenetically Preserved *OsMYBs* Reveals Strong Associations Between Genes Co-expression, Function and Evolution

The phylogenetic footprinting might be additive to coexpressed cluster and successfully being applied to determine expression association of genes (Elnitski et al., 2006). Exploring the co-expression and phylogenetic analysis suggested that the highly co-expressed genes with known role in specific regulatory processes were preserved in the network. We found such type of relation in module 1, 2, 3, and 4 (**Figure 3**). Two of the *OsMYB2* (*LOC_Os01g18240*, *LOC_Os05g04820*) genes were found to be upregulated in phase-I of chilling stress, where *OsMYB2* (*LOC_Os01g18240*) positively correlated with *LOC_Os07g44090* (phylogenetically also closely related), *LOC_Os05g40960, LOC_Os01g36460* and *LOC_Os06g49040*. The phylogenetically close pair was found to be involved in highly similar type of processes such as response to biosynthetic process, endogenous stimulus, reproduction, post-embryonic development. Two of the genes with high degree *viz. LOC_Os01g74410 (MYB59)* and *LOC_Os01g51154 (R1-MYB)* were found to be highly correlated with several other *MYB* genes in the network (**Table S5B**). It is in agreement with the study that the expression of these genes are modulated both by cold independent conditions (Park et al., 2010). We observed that OsMYB2P-1 (LOC_Os05g04820) protein was close to LOC_Os01g65370, LOC_Os05g3550, and OsMYB4 (LOC_Os04g43680) in 3rd phylogenetic cluster. OsMYB2P-1 is known to regulate phosphate starvation, cold, salt and osmotic stress responses, and also found to be up-regulated in phosphorus starvation in this study. This is in agreement with the results by Dai et al. (2012). A system biology approach has identified R2R3 motif *MYB28* and two homologs, *MYB29* and *MYB76* genes that form a single clade with distinct and overlapping functions in regulation of aliphatic glucosinolates (Sønderby et al., 2007). These evidences showed the important regulatory roles of MYBs in several biological processes. Moreover, *OsMYB4* is known to express in cold-mediated and cold-independent transcriptional network (Park et al., 2010). Evaluation of data revealed that the cluster of genes that are co-expressed lie in distinct phylogenetic clade,

suggesting functional redundancy and their evolution by recent duplication.

Deciphering Transcriptional Regulatory Network for Putative Target Gene Identification

The first step in gene regulation is transcriptional regulation which is governed by the recognition of *cis*-element by the DNA binding domain of TFs. The assembly of TFs on the promoter *cis*-element region and their interaction in regulatory network profoundly influence the target gene expression. It is known that genes with similar expression pattern in the same biological function are likely to be regulated by same TF(s) (i.e., co-regulated) having similar *cis*-regulatory elements for the TFs were liable for putative target gene identification (Wang and Stormo, 2003; Walhout, 2006; Wang et al., 2009; Imam et al., 2015). Hence, we created another *OsMYB* network by guide gene approach to identify the putative target *OsMYB* genes on the basis of functional co-occurrence as well as MYB recognition *cis*-elements in their promoter region.

Among TFs, we observed ten guide *OsMYBs* were in correlation with other *OsMYB* genes forming a more complex feedback network. We also observe the presence of feedback motif in the target OsMYBs Comparing the results from both top-down and guide-gene approach showed the conservation of one correlated pair of *OsMYB* (*LOC_Os11g47460*, *LOC_Os01g74410*; PCC 0.98). Among correlated TFs such as WRKY, ZOS6-05—C2H2 zinc finger protein and helix–loop–helix (bHLH) protein were found. This suggests that the function of OsMYB proteins might require participation of various members of these transcription factors (**Table S7B**). It is in partial agreement with the recent study that showed transcriptional regulation by MYB–bHLH–WD40 (MBW) complex in the late step of flavonoid biosynthetic pathway (Hichri et al., 2011), GL2 expression and the non-hair or trichome fate (Schiefelbein, 2003).

Conclusively, in the present study we identified co-regulatory network and functional co-occurrence of modules of *OsMYB* genes in rice. This will contribute to illustrate the functions of gene cooperation pathways that have not yet been identified by classical genetic analyses. In the first part of the study, we adopted the top-down approach to decipher the *OsMYBs* with correlated expression pattern in different development and stress conditions. We defined the existence of *OsMYBs* gene clusters comprising both phylogenetically related and unrelated genes that were strongly coexpressed, signifying their evolutionary role in co-regulatory manner. A sum of 51 most highly connected hub *OsMYB*s were identified, some of them were expected to play the significant regulatory roles in abiotic stress tolerance. As the hubs have high correlation value, they may play crucial role in stress tolerance as well as development.

More importantly, our analyses revealed the existence of *OsMYBs* transcriptionally co-regulatory networks by taking guide *OsMYB* genes with known function under abiotic stress condition. This provided insight into the functional association of several uncharacterized genes and coexpressed putative target genes possessing MYB binding *cis*-elements in their promoter region. The presence of drought responsive MYB binding *cis*-elements in the putative target genes and guide genes with known drought stress response identified the co-regulatory network in response to drought stress. In several instances, these rationales for candidate gene screening and functional validation allowed us to generate hypotheses, which are experimentally testable and their relevance in a specific process involved in plant response to stress or hormone signals. Functional testing of *in vivo* interaction or action of the candidate co-expressed gene network modules and hubs will significantly enhance our knowledge on the function of MYB family and help develop improved rice genotypes. Therefore, the network modules predicted in the present study were of high biological relevance and revealed putative role for uncharacterized genes. Further, the outcome of the study offers new biological insights into the transcriptional regulatory networks that await experimental validation.

AUTHOR CONTRIBUTIONS

SS, VC, KB conceived and designed the experiments. SS performed the experiments. SS analyzed the data. AK performed computational analysis. SS, VC, DP, KB wrote the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the Indian Council of Agricultural Research (ICAR) for supporting this work through the ICAR-sponsored project on the National Initiative on Climate Resilient Agriculture (NICRA) project. VC was supported by NASF (ICAR) project (grant No. NFBSFARA/Phen 2015).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2015.01157

Figure S1 | *OsMYB* correlation network viewed in Cytoscape. Nodes in red color are differentially expressed at least in one condition.

Figure S2 | Phylogenetic tree of 1 kb promoter sequences of guide *OsMYB* genes and their putative target genes. Putative target genes having MYB binding *cis*- elements in their promoter region are shown by italic fonts.

Table S1 | Detailed list of affymetrix rice genome microarray datasets used for *OsMYB* genes correlation network analysis via top-down approach.

Table S2 | List of retrieved *OsMYB* genes with their putative function.

Table S3 | Differentially expressed *OsMYB* genes under diverse microarray experiments.

Table S4 | Modular gene ontology enrichment analysis.

Table S5 | Average logarithmic (A) signal values of 219 *OsMYB* protein encoding genes expressed under different microarray experiments. (**B**) Pearson correlation coefficient (PCC) among *OsMYB* genes in top down approach. (**C**) Simple and complex topological properties of correlation network of *OsMYB* genes. Red highlighted is hub nodes.

Table S6 | Parameter evaluation and optimization of the MCL inflation score (*I*).

Table S7 | (A) List of guide genes used to create global co-regulatory network via guide-gene approach. (**B**) Global co-regulatory network of guide *OsMYB* genes and their correlated allies with their description. (**C**) Gene ontology enrichment analysis of target genes.

Table S8 | (A)List of *cis*-elements in 1 kb upstream promoter region of direct first neighbor of guide *OsMYB* genes in global co-regulatory network. (B) Motif enrichment analyses by MEME of direct first neighbor of guide *OsMYB* genes in global co-regulatory network.

REFERENCES

Albert, R., and Barabasi, A.-L. (2000). Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* 85, 5234–5237. doi: 10.1103/PhysRevLett.85.5234

Aoki, K., Ogata, Y., and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.* 48, 381–390. doi: 10.1093/pcp/pcm013

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Bansal, K. C., Katiyar, A., Smita, S., and Chinnusamy, V. (2012). Functional genomics and computational biology tools for gene discovery for abiotic stress tolerance. *Improv. Crop Resist. Abiotic Stress Vols 1,* 2, 321–335. doi: 10.1002/9783527632930.ch14

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Berri, S., Abbruscato, P., Faivre-Rampant, O., Brasileiro, A. C., Fumasoni, I., Satoh, K., et al. (2009). Characterization of WRKY co-regulatory networks in rice and *Arabidopsis*. *BMC Plant Biol.* 9:120. doi: 10.1186/1471-2229-9-120

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma. Oxf. Engl.* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Chantarachot, T., Buaboocha, T., Gu, H., and Chadchawan, S. (2012). Putative calmodulin-binding R2R3-MYB transcription factors in rice (*Oryza sativa* L.). *Thai. J. Bot.* 4, 101–112.

Chinnusamy, V., Schumaker, K., and Zhu, J.-K. (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *J. Exp. Bot.* 55, 225–236. doi: 10.1093/jxb/erh005

Cho, D.-Y., Kim, Y.-A., and Przytycka, T. M. (2012). Network biology approach to complex diseases. *PLoS Comput. Biol.* 8:e1002820. doi: 10.1371/journal.pcbi.1002820

Chou, K.-C., and Shen, H.-B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Cominelli, E., and Tonelli, C. (2009). A new role for plant R2R3-MYB transcription factors in cell cycle regulation. *Cell Res.* 19, 1231–1232. doi: 10.1038/cr.2009.123

Cramer, G. R., Urano, K., Delrot, S., Pezzotti, M., and Shinozaki, K. (2011). Effects of abiotic stress on plants: a systems biology perspective. *BMC Plant Biol.* 11:163. doi: 10.1186/1471-2229-11-163

Dai, X., Wang, Y., Yang, A., and Zhang, W.-H. (2012). OsMYB2P-1, an R2R3 MYB transcription factor, is involved in the regulation of phosphate-starvation responses and root architecture in rice. *Plant Physiol.* 159, 169–183. doi: 10.1104/pp.112.194217

Dai, X., Xu, Y., Ma, Q., Xu, W., Wang, T., Xue, Y., et al. (2007). Overexpression of an R1R2R3 MYB gene, OsMYB3R-2, increases tolerance to freezing, drought, and salt stress in transgenic Arabidopsis. *Plant Physiol.* 143, 1739–1751. doi: 10.1104/pp.106.094532

Diboun, I., Wernisch, L., Orengo, C. A., and Koltzenburg, M. (2006). Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics* 7:252. doi: 10.1186/1471-2164-7-252

Dieterich, C., Grossmann, S., Tanzer, A., Röpcke, S., Arndt, P. F., Stadler, P. F., et al. (2005). Comparative promoter region analysis powered by CORG. *BMC Genomics* 6:24. doi: 10.1186/1471-2164-6-24

Du, H., Feng, B.-R., Yang, S.-S., Huang, Y.-B., and Tang, Y.-X. (2012). The R2R3-MYB transcription factor gene family in maize. *PLoS ONE* 7:e37463. doi: 10.1371/journal.pone.0037463

Du, H., Zhang, L., Liu, L., Tang, X.-F., Yang, W.-J., Wu, Y.-M., et al. (2009). Biochemical and molecular characterization of plant MYB transcription factor family. *Biochem. Biokhimiia* 74, 1–11. doi: 10.1134/s0006297909010015

Du, L., Jiao, F., Chu, J., Jin, G., Chen, M., and Wu, P. (2007). The two-component signal system in rice (*Oryza sativa* L.): a genome-wide study of cytokinin signal perception and transduction. *Genomics* 89, 697–707. doi: 10.1016/j.ygeno.2007.02.001

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* 15, 573–581. doi: 10.1016/j.tplants.2010.06.005

Elnitski, L., Jin, V. X., Farnham, P. J., and Jones, S. J. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.* 16, 1455–1464. doi: 10.1101/gr.4140006

Gao, J.-P., Chao, D.-Y., and Lin, H.-X. (2007). Understanding abiotic stress tolerance mechanisms: recent studies on stress response in rice. *J. Integr. Plant Biol.* 49, 742–750. doi: 10.1111/j.1744-7909.2007.00495.x

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Gigolashvili, T., Yatusevich, R., Rollwitz, I., Humphry, M., Gershenzon, J., and Flügge, U.-I. (2009). The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in *Arabidopsis thaliana*. *Plant Cell Online* 21, 1813–1829. doi: 10.1105/tpc.109.066399

Glazebrook, J. (2001). Genes controlling expression of defense responses in Arabidopsis–2001 status. *Curr. Opin. Plant Biol.* 4, 301–308. doi: 10.1016/S1369-5266(00)00177-1

Gray, J., Bevan, M., Brutnell, T., Buell, C. R., Cone, K., Hake, S., et al. (2009). A recommendation for naming transcription factor proteins in the grasses. *Plant Physiol.* 149, 4–6. doi: 10.1104/pp.108.128504

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104. doi: 10.1038/nature02800

Hichri, I., Barrieu, F., Bogs, J., Kappel, C., Delrot, S., and Lauvergeat, V. (2011). Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *J. Exp. Bot.* 62, 2465–2483. doi: 10.1093/jxb/erq442

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Hwang, I., and Sheen, J. (2001). Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature* 413, 383–389. doi: 10.1038/35096500

Imam, S., Noguera, D. R., and Donohue, T. J. (2015). An integrated approach to reconstructing genome-scale transcriptional regulatory networks. *PLoS Comput. Biol.* 11:e1004103. doi: 10.1371/journal.pcbi.1004103

Katiyar, A., Smita, S., Lenka, S. K., Rajwanshi, R., Chinnusamy, V., and Bansal, K. C. (2012). Genome-wide classification and expression analysis of MYB transcription factor families in rice and *Arabidopsis*. *BMC Genomics* 13:544. doi: 10.1186/1471-2164-13-544

Kaur, C., Kushwaha, H. R., Mustafiz, A., Pareek, A., Sopory, S. K., and Singla-Pareek, S. L. (2015). Analysis of global gene expression profile of rice in response to methylglyoxal indicates its possible role as a stress signal molecule. *Front. Plant Sci.* 6:682. doi: 10.3389/fpls.2015.00682

Klempnauer, K. H., Gonda, T. J., and Bishop, J. M. (1982). Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. *Cell* 31, 453–463. doi: 10.1016/0092-8674(82)90138-6

Komaki, S., and Sugimoto, K. (2012). Control of the plant cell cycle by developmental and environmental cues. *Plant Cell Physiol.* 53, 953–964. doi: 10.1093/pcp/pcs070

Kuno, N., Møller, S. G., Shinomura, T., Xu, X., Chua, N.-H., and Furuya, M. (2003). The novel MYB protein EARLY-PHYTOCHROME-RESPONSIVE1

is a component of a slave circadian oscillator in *Arabidopsis*. *Plant Cell* 15, 2476–2488. doi: 10.1105/tpc.014217

Laluk, K., Prasad, K. V. S. K., Savchenko, T., Celesnik, H., Dehesh, K., Levy, M., et al. (2012). The calmodulin-binding transcription factor SIGNAL RESPONSIVE1 is a novel regulator of glucosinolate metabolism and herbivory tolerance in *Arabidopsis*. *Plant Cell Physiol.* 53, 2008–2015. doi: 10.1093/pcp/pcs143

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39, W475–W478. doi: 10.1093/nar/gkr201

Lim, S. D., Yim, W. C., Moon, J.-C., Kim, D. S., Lee, B.-M., and Jang, C. S. (2010). A gene family encoding RING finger proteins in rice: their expansion, expression diversity, and co-expressed genes. *Plant Mol. Biol.* 72, 369–380. doi: 10.1007/s11103-009-9576-9

Liu, H., Zhou, X., Dong, N., Liu, X., Zhang, H., and Zhang, Z. (2011). Expression of a wheat MYB gene in transgenic tobacco enhances resistance to *Ralstonia solanacearum*, and to drought and salt stresses. *Funct. Integr. Genomics* 11, 431–443. doi: 10.1007/s10142-011-0228-1

Lu, C.-A., Ho, T. D., Ho, S.-L., and Yu, S.-M. (2002). Three novel MYB proteins with one DNA binding repeat mediate sugar and hormone regulation of alpha-amylase gene expression. *Plant Cell* 14, 1963–1980. doi: 10.1105/tpc.001735

Lu, X., Jain, V. V., Finn, P. W., and Perkins, D. L. (2007). Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol.* 3:98. doi: 10.1038/msb4100138

Ma, Q., Dai, X., Xu, Y., Guo, J., Liu, Y., Chen, N., et al. (2009). Enhanced tolerance to chilling stress in OsMYB3R-2 transgenic rice is mediated by alteration in cell cycle and ectopic expression of stress genes. *Plant Physiol.* 150, 244–256. doi: 10.1104/pp.108.133454

Martin, C., and Paz-Ares, J. (1997). MYB transcription factors in plants. *Trends Genet. TIG* 13, 67–73. doi: 10.1016/S0168-9525(96)10049-4

Meier, S., Gehring, C., MacPherson, C. R., Kaur, M., Maqungo, M., Reuben, S., et al. (2008). The Promoter signatures in rice LEA genes can be used to build a co-expressing LEA gene network. *Rice* 1, 177–187. doi: 10.1007/s12284-008-9017-4

Molinero, A. I. S. (2013). *Plants having Enhanced Yield-Related Traits and a Method for Making the Same*. Available online at: http://www.google.com/patents/US8569575 [Accessed November 16, 2015].

Mounet, F., Moing, A., Garcia, V., Petit, J., Maucourt, M., Deborde, C., et al. (2009). Gene and metabolite regulatory network analysis of early developing fruit tissues highlights new candidate genes for the control of tomato fruit composition and development. *Plant Physiol.* 149, 1505–1528. doi: 10.1104/pp.108.133967

Movahedi, S., van Bel, M., Heyndrickx, K. S., and Vandepoele, K. (2012). Comparative co-expression analysis in plant biology. *Plant Cell Environ.* 35, 1787–1798. doi: 10.1111/j.1365-3040.2012.02517.x

Nakashima, K., Ito, Y., and Yamaguchi-Shinozaki, K. (2009). Transcriptional regulatory networks in response to abiotic stresses in *Arabidopsis* and grasses. *Plant Physiol.* 149, 88–95. doi: 10.1104/pp.108.129791

Ouyang, Y., Huang, X., Lu, Z., and Yao, J. (2012). Genomic survey, expression profile and co-expression network analysis of OsWD40 family in rice. *BMC Genomics* 13:100. doi: 10.1186/1471-2164-13-100

Park, M.-R., Yun, K.-Y., Mohanty, B., Herath, V., Xu, F., Wijaya, E., et al. (2010). Supra-optimal expression of the cold-regulated OsMyb4 transcription factor in transgenic rice changes the complexity of transcriptional network with major effects on stress tolerance and panicle development. *Plant Cell Environ.* 33, 2209–2230. doi: 10.1111/j.1365-3040.2010.02221.x

Patil, A., and Nakamura, H. (2006). Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* 580, 2041–2045. doi: 10.1016/j.febslet.2006.03.003

Patnala, R., Clements, J., and Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful *in silico* tools. *BMC Genet.* 14:39. doi: 10.1186/1471-2156-14-39

Pattanaik, S., Kong, Q., Zaitlin, D., Werkman, J. R., Xie, C. H., Patra, B., et al. (2010). Isolation and functional characterization of a floral tissue-specific R2R3 MYB regulator from tobacco. *Planta* 231, 1061–1076. doi: 10.1007/s00425-010-1108-y

Peleg, Z., and Blumwald, E. (2011). Hormone balance and abiotic stress tolerance in crop plants. *Curr. Opin. Plant Biol.* 14, 290–295. doi: 10.1016/j.pbi.2011.02.001

Petrov, V., Hille, J., Mueller-Roeber, B., and Gechev, T. S. (2015). ROS-mediated abiotic stress-induced programmed cell death in plants. *Front. Plant Sci.* 6:69. doi: 10.3389/fpls.2015.00069

Qu, L.-J., and Zhu, Y.-X. (2006). Transcription factor families in *Arabidopsis*: major progress and outstanding issues for future research. *Curr. Opin. Plant Biol.* 9, 544–549. doi: 10.1016/j.pbi.2006.07.005

Rana, R. M., Dong, S., Ali, Z., Huang, J., and Zhang, H. S. (2012). Regulation of ATG6/Beclin-1 homologs by abiotic stresses and hormones in rice (*Oryza sativa* L.). *Genet. Mol. Res. GMR* 11, 3676–3687. doi: 10.4238/2012.August.17.3

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* 39, W307–W315. doi: 10.1093/nar/gkr378

Saikumar, P., Murali, R., and Reddy, E. P. (1990). Role of tryptophan repeats and flanking amino acids in Myb-DNA interactions. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8452–8456. doi: 10.1073/pnas.87.21.8452

Sato, Y., Namiki, N., Takehisa, H., Kamatsuki, K., Minami, H., Ikawa, H., et al. (2013). RiceFREND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.* 41, D1214–D1221. doi: 10.1093/nar/gks1122

Schaller, G. E., Shiu, S.-H., and Armitage, J. P. (2011). Two-component systems and their co-option for eukaryotic signal transduction. *Curr. Biol.* 21, R320–R330. doi: 10.1016/j.cub.2011.02.045

Schiefelbein, J. (2003). Cell-fate specification in the epidermis: a common patterning mechanism in the root and shoot. *Curr. Opin. Plant Biol.* 6, 74–78. doi: 10.1016/S136952660200002X

Seo, P. J., Lee, S. B., Suh, M. C., Park, M.-J., Go, Y. S., and Park, C.-M. (2011). The MYB96 transcription factor regulates cuticular wax biosynthesis under drought conditions in *Arabidopsis*. *Plant Cell* 23, 1138–1152. doi: 10.1105/tpc.111.083485

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shinozaki, K., and Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. *J. Exp. Bot.* 58, 221–227. doi: 10.1093/jxb/erl164

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., et al. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell* 133, 1043–1054. doi: 10.1016/j.cell.2008.04.040

Smita, S., Katiyar, A., Pandey, D. M., Chinnusamy, V., Archak, S., and Bansal, K. C. (2013). Identification of conserved drought stress responsive gene-network across tissues and developmental stages in rice. *Bioinformation* 9:72. doi: 10.6026/97320630009072

Sønderby, I. E., Hansen, B. G., Bjarnholt, N., Ticconi, C., Halkier, B. A., and Kliebenstein, D. J. (2007). A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PLoS ONE* 2:e1322. doi: 10.1371/journal.pone.0001322

Su, C.-F., Wang, Y.-C., Hsieh, T.-H., Lu, C.-A., Tseng, T.-H., and Yu, S.-M. (2010). A novel MYBS3-dependent pathway confers cold tolerance in rice. *Plant Physiol.* 153, 145–158. doi: 10.1104/pp.110.153015

Takahashi, H., Yamauchi, T., Rajhi, I., Nishizawa, N. K., and Nakazono, M. (2015). Transcript profiles in cortical cells of maize primary root during ethylene-induced lysigenous aerenchyma formation under aerobic conditions. *Ann. Bot.* 115, 879–894. doi: 10.1093/aob/mcv018

Tsai, Y.-C., Weir, N. R., Hill, K., Zhang, W., Kim, H. J., Shiu, S.-H., et al. (2012). Characterization of genes involved in cytokinin signaling and metabolism from rice. *Plant Physiol.* 158, 1666–1684. doi: 10.1104/pp.111.192765

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. *Plant Physiol.* 150, 535–546. doi: 10.1104/pp.109.136028

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi: 10.1137/040608635

Wahid, A., Gelani, S., Ashraf, M., and Foolad, M. R. (2007). Heat tolerance in plants: an overview. *Environ. Exp. Bot.* 61, 199–223. doi: 10.1016/j.envexpbot.2007.05.011

Walhout, A. J. M. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* 16, 1445–1454. doi: 10.1101/gr.5321506

Wang, T., and Stormo, G. D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19, 2369–2380. doi: 10.1093/bioinformatics/btg329

Wang, X., Haberer, G., and Mayer, K. F. X. (2009). Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics* 10:284. doi: 10.1186/1471-2164-10-284

Wang, Z. Y., Kenigsbuch, D., Sun, L., Harel, E., Ong, M. S., and Tobin, E. M. (1997). A Myb-related transcription factor is involved in the phytochrome regulation of an *Arabidopsis* Lhcb gene. *Plant Cell* 9, 491–507. doi: 10.1105/tpc.9.4.491

Weston, K. (1998). Myb proteins in life, death and differentiation. *Curr. Opin. Genet. Dev.* 8, 76–81. doi: 10.1016/S0959-437X(98)80065-8

Wong, D. C., Sweetman, C., Drew, D. P., and Ford, C. M. (2013). VTCdb: a gene co-expression database for the crop species *Vitis vinifera* (grapevine). *BMC Genomics* 14:882. doi: 10.1186/1471-2164-14-882

Wong, D. C., Sweetman, C., and Ford, C. M. (2014). Annotation of gene function in citrus using gene expression information and co-expression networks. *BMC Plant Biol.* 14:186. doi: 10.1186/1471-2229-14-186

Xie, Z., Lee, E., Lucas, J. R., Morohashi, K., Li, D., Murray, J. A. H., et al. (2010). Regulation of cell proliferation in the stomatal lineage by the *Arabidopsis* MYB FOUR LIPS via direct targeting of core cell cycle genes. *Plant Cell* 22, 2306–2321. doi: 10.1105/tpc.110.074609

Xiong, H., Li, J., Liu, P., Duan, J., Zhao, Y., Guo, X., et al. (2014). Overexpression of OsMYB48-1, a Novel MYB-related transcription factor, enhances drought and salinity tolerance in rice. *PLoS ONE* 9:e92913. doi: 10.1371/journal.pone.0092913

Xu, Z.-S., Chen, M., Li, L.-C., and Ma, Y.-Z. (2011). Functions and application of the AP2/ERF transcription factor family in crop improvementf. *J. Integr. Plant Biol.* 53, 570–585. doi: 10.1111/j.1744-7909.2011.01062.x

Xue, G.-P., Kooiker, M., Drenth, J., and McIntyre, C. L. (2011). TaMYB13 is a transcriptional activator of fructosyltransferase genes involved in β-2,6-linked fructan synthesis in wheat. *Plant J.* 68, 857–870. doi: 10.1111/j.1365-313X.2011.04737.x

Xue, L.-J., Zhang, J.-J., and Xue, H.-W. (2012). Genome-wide analysis of the complex transcriptional networks of rice developing seeds. *PLoS ONE* 7:e31081. doi: 10.1371/journal.pone.0031081

Yang, C., Xu, Z., Song, J., Conner, K., Vizcay Barrena, G., and Wilson, Z. A. (2007). *Arabidopsis* MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence. *Plant Cell* 19, 534–548. doi: 10.1105/tpc.106.046391

Yang, X., Yang, Y.-N., Xue, L.-J., Zou, M.-J., Liu, J.-Y., Chen, F., et al. (2011). Rice ABI5-Like1 regulates abscisic acid and auxin responses by affecting the expression of ABRE-containing genes. *Plant Physiol.* 156, 1397–1409. doi: 10.1104/pp.111.173427

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins* 64, 643–651. doi: 10.1002/prot.21018

Yuan, J. S., Galbraith, D. W., Dai, S. Y., Griffin, P., and Stewart, C. N. Jr. (2008). Plant systems biology comes of age. *Trends Plant Sci.* 13, 165–171. doi: 10.1016/j.tplants.2008.02.003

Yun, K.-Y., Park, M. R., Mohanty, B., Herath, V., Xu, F., Mauleon, R., et al. (2010). Transcriptional regulatory network triggered by oxidative signals configures the early response mechanisms of japonica rice to chilling stress. *BMC Plant Biol.* 10:16. doi: 10.1186/1471-2229-10-16

Zhang, L., Yu, S., Zuo, K., Luo, L., and Tang, K. (2012). Identification of gene modules associated with drought response in rice by network-based Analysis. *PLoS ONE* 7:e33748. doi: 10.1371/journal.pone.0033748

Zhao, Y., Xing, L., Wang, X., Hou, Y.-J., Gao, J., Wang, P., et al. (2014). The ABA receptor PYL8 promotes lateral root growth by enhancing MYB77-dependent transcription of auxin-responsive genes. *Sci. Signal.* 7:ra53. doi: 10.1126/scisignal.2005051

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 January 2016
doi: 10.3389/fgene.2015.00351

Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data

Astrid Wachter and Tim Beißbarth*

Department of Medical Statistics, University Medical Center, Göttingen, Germany

OPEN ACCESS

Edited by:
Ekaterina Shelest,
Hans-Knoell Institute, Germany

Reviewed by:
Frank Emmert-Streib,
Tampere University of Technology, Finland
Lorenz Adlung,
German Cancer Research Center (DKFZ), Germany

****Correspondence:***
Astrid Wachter
astrid.wachter@med.uni-goettingen.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 October 2015*
Accepted: *03 December 2015*
Published: *07 January 2016*

Citation:
Wachter A and Beißbarth T (2016) Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data. Front. Genet. 6:351. doi: 10.3389/fgene.2015.00351

Identification of dynamic signaling mechanisms on different cellular layers is now facilitated as the increased usage of various high-throughput techniques goes along with decreasing costs for individual experiments. A lot of these signaling mechanisms are known to be coordinated by their dynamics, turning time-course data sets into valuable information sources for inference of regulatory mechanisms. However, the combined analysis of parallel time-course measurements from different high-throughput platforms still constitutes a major challenge requiring sophisticated bioinformatic tools in order to ease biological interpretation. We developed a new pathway-based integration approach for the analysis of coupled omics time-series data, which we implemented in the R package *pwOmics*. Unlike many other approaches, our approach acknowledges the role of the different cellular layers of measurement and infers consensus profiles and time profile clusters for further biological interpretation. We investigated a time-course data set on epidermal growth factor stimulation of human mammary epithelial cells generated on the two layers of RNA and proteins. The data was analyzed using our new approach with a focus on feedback signaling and pathway crosstalk. We could confirm known regulatory patterns relevant in the physiological cellular response to epidermal growth factor stimulation as well as identify interesting new interactions in this signaling context, such as the regulatory influence of the connective tissue growth factor on transferrin receptor or the influence of growth arrest and DNA-damage-inducible alpha on the connective tissue growth factor. Thus, we show that integrated cross-platform analysis provides a deeper understanding of regulatory signaling mechanisms. Combined with time-course information it enables the characterization of dynamic signaling processes and leads to the identification of important regulatory interactions which might be dysregulated in disease with adverse effects.

Keywords: omics, data integration, high-throughput, time-series, EGF signaling

INTRODUCTION

Omics data integration is a conclusive concept for a systemic understanding of biological signaling mechanisms, both in healthy conditions and disease (Kristensen et al., 2014; Ritchie et al., 2015). The combination of different types of omics data can provide a more comprehensive and complete picture of individual cellular mechanisms. Furthermore, a cross-platform analysis represents a measure to overcome individual platform biases and technical limitations (Yeger-Lotem et al., 2009).

An even more informative approach is to analyze time-course data sets from different omics levels, as a lot of cellular signaling information is encoded in signaling dynamics (Purvis and Lahav, 2013). This type of data provides more than only a single "snapshot" of the underlying biological processes, thus it can augment the knowledge we have about cellular signaling events considerably. With these data feedback signaling loops, molecular interactions and pathway crosstalk can be tracked over time. Thus, combining different types of omics data with time course information enables a comprehensive characterization of cellular responses upon stimulation and also a detection of regulatory mechanisms initiated by specific perturbations. In **Figure 1** a selection of dynamic regulatory signaling mechanisms on protein and gene layer is depicted. These effects become directly apparent in such omics data sets, so the "dynamic knowledge" we can collect may also provide us with an idea of modifications responsible for pathologic signaling and signaling dynamics, thus forming a basis for an improvement of treatment strategies.

Of course, such parallel time-course data sets are even more challenging to analyze and interpret as they include an additional dimension and require a meaningful cross-platform integration method. Hence, there is a demand for bioinformatic tools that can deal with the diverse data types and combine them in such a way that their output enables a straightforward biological interpretation of the data. Although a lot of individual data integration methods have been developed so far, they mostly address very specific integration questions (Balbin et al., 2013; Hamon et al., 2014), are not implemented as tools which can be freely used by other biologists and bioinformaticians [e.g., QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City[1])] or do not acknowledge the different nature of different omics data types (Ding et al., 2012; Sun et al., 2014). Very few tools also include the biologically very interesting aspect of time-course data analysis (Rogers et al., 2008), although these types of data sets are expected to be generated more often in the near future (Bar-Joseph et al., 2012) in order to address systems biology questions.

We developed a pathway-based data integration approach for the analysis of coupled high-throughput time-course measurements on the cellular layers of proteins, transcripts and genes. We implemented this approach as R package *pwOmics,* that we presented earlier (Wachter and Beissbarth, 2015). In brief, *pwOmics* joins the tools of network analysis: It uses public signaling pathway knowledge to map molecular network interactions, thereby identifying activated and inactivated genes and proteins in cellular signaling upon perturbation. Thus, the cellular layers on which the data is collected are acknowledged during data analysis while simultaneously considering the

[1]www.quiagen.com/ingenuity.

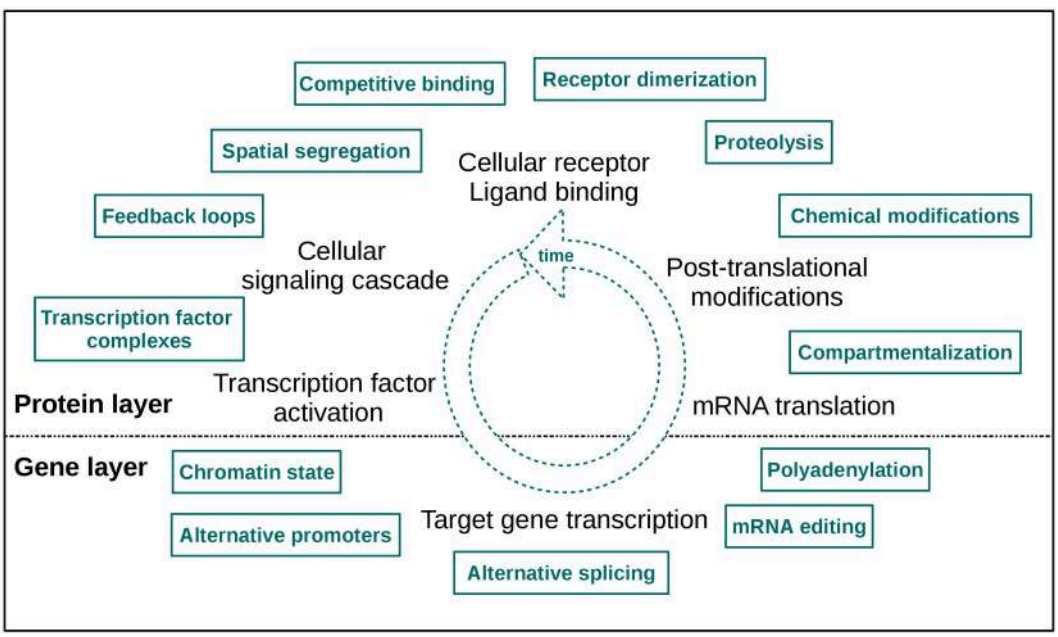


FIGURE 1 | A selection of cellular layer specific regulatory signaling mechanisms. The two layers of measurement are indicated as "protein" and "gene layer." The high number of effectors illustrates the mechanistic fine-tuning of signaling. Note that this fine-tuning also takes place in the dimension of time.

dynamics. Here we describe and test the utility of our method in more detail.

Epidermal growth factor (EGF) signaling has already been studied comprehensively in comparison to other signaling pathways as dysregulation is associated with poor prognosis in many human malignancies (Lurje and Lenz, 2009). As various high-throughput and low-throughput omics data sets are available and a lot of knowledge is already acquired on the basis of which methodical evaluation can be performed, it constitutes an adequate example for investigation of new approaches. The data set analyzed here measures the mitogenic response of human mammary epithelial cells (HMEC) to EGF on the proteomic and the transcriptomic layer over time (Waters et al., 2012), thereby representing physiological signaling conditions. **Figure 2** depicts the experimental design used in the study. EGF stimulation is associated with cellular proliferation, differentiation and survival (Herbst, 2004) and directly affects signaling pathways such as the MAPK signaling pathway, the ERBB signaling pathway and the RAS signaling pathway.

We chose the comparably well characterized example of EGF signaling in order to map the results of our new pathway-based integration approach to known experimental results for methodical evaluation and to reveal new dynamically relevant mechanisms in EGF signaling on the different functional layers. We focus on feedback signaling and pathway crosstalk, both complex regulatory mechanisms that have been under intensive biological investigation in individual experiments in physiological and pathological conditions (Avraham and Yarden, 2011; Wang et al., 2011).

METHODS

Data Set

The data set investigated with the new pathway-based integration approach was generated in a study on network analysis of EGF signaling. The experimental design used is illustrated in **Figure 2**, the measurements included transcriptomic, proteomic and phosphoproteomic data generation. Further details as well as the preprocessing steps performed on both microarray raw data and proteomic raw data are described in Waters et al. (2012). The raw microarray data files are available via the Gene Expression Omnibus database, GSE15668 (Waters et al., 2012). The corresponding proteomic data is also publicly available[2].

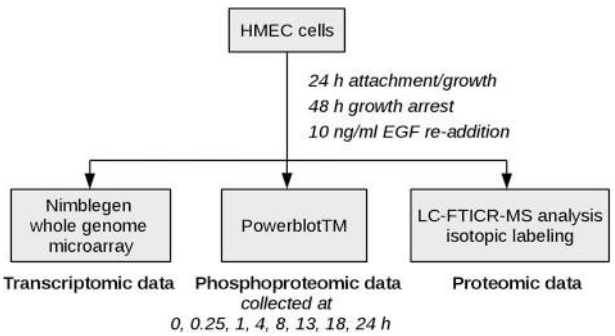


FIGURE 2 | Experimental design. HMEC cells were seeded and allowed to attach and grow for 24 h. After 48 h of growth arrest with medium lacking serum, EGF and other growth factors, EGF was added again to monitor the mitogenic response of the cells. Samples for high-throughput genomic and proteomic measurements were taken at time points 0, 0.25, 1, 4, 8, 13, 18, 24 h after EGF stimulation. The 0.25 h time point was excluded from the microarray data set due to quality issues, therefore the coupled data set on which our analysis is based includes time points 0, 1, 4, 8, 13, 18, and 24 h after EGF stimulation.

Shortly, biological samples were hybridized against NimbleGen microarrays. A quality check revealed that time point 0.25 h failed to hybridize, therefore the coupled data set analyzed here includes only time points 0, 1, 4, 8, 13, 18, and 24 h after EGF stimulation. Proteome analysis was performed MS-based, while phosphoproteome data were collected as part of a parallel western blot analysis. For each time point differentially expressed transcripts or differentially abundant phosphoproteins/proteins compared to time point 0 h were determined. Raw microarray data was quantile normalized before performing a pairwise analysis of variance with a 5% false discovery rate to determine differentially expressed transcripts. Proteome and phosphoproteome levels were considered significant when passing specific quality checks and showing a fold change $\geq$1.5.

Databases

Pathway information used for the pathway-based integration approach were taken from KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2014), Reactome (Croft et al., 2014), Pathway Interaction Database (Schaefer et al., 2009), and Biocarta (Nishimura, 2001). This information was used as gene sets in the analysis of the phosphoproteome data and combined with its topological information in the transcriptome data analysis. It was downloaded via the AnnotationHub R package[3] from Bioconductor (Huber et al., 2015) as BioPAX level 2 files and then processed further with the rBiopaxParser R package (Kramer et al., 2013). The transcription factor (TF)—target gene interaction information from the TRANSFAC® database (Biobase version 2014.4; Matys et al., 2006) was used. Network reconstruction was based on the connected protein-protein interaction (PPI) network of the STRING database (Franceschini et al., 2013).

Analyses

All analysis steps described here are based on pre-processed transcriptome, proteome and phosphoproteome data, as described in Waters et al. (2012). Main analyses steps were performed with the R package *pwOmics* (Wachter and Beissbarth, 2015). Our methodical framework is depicted in **Figures 3**, **4**.

Data Processing

First, individual analyses of the omics data sets were performed during phosphoprotein data based downstream and transcript based upstream analysis (**Figure 3**). For the downstream analysis an identification of the pathways, which include differentially

[2]http://omics.pnl.gov.

[3]Morgan, M., Carlson, M., Tenenbaum, D., and Arora, S. *AnnotationHub: Client to Access AnnotationHub Resources.* R package version 2.0.0.

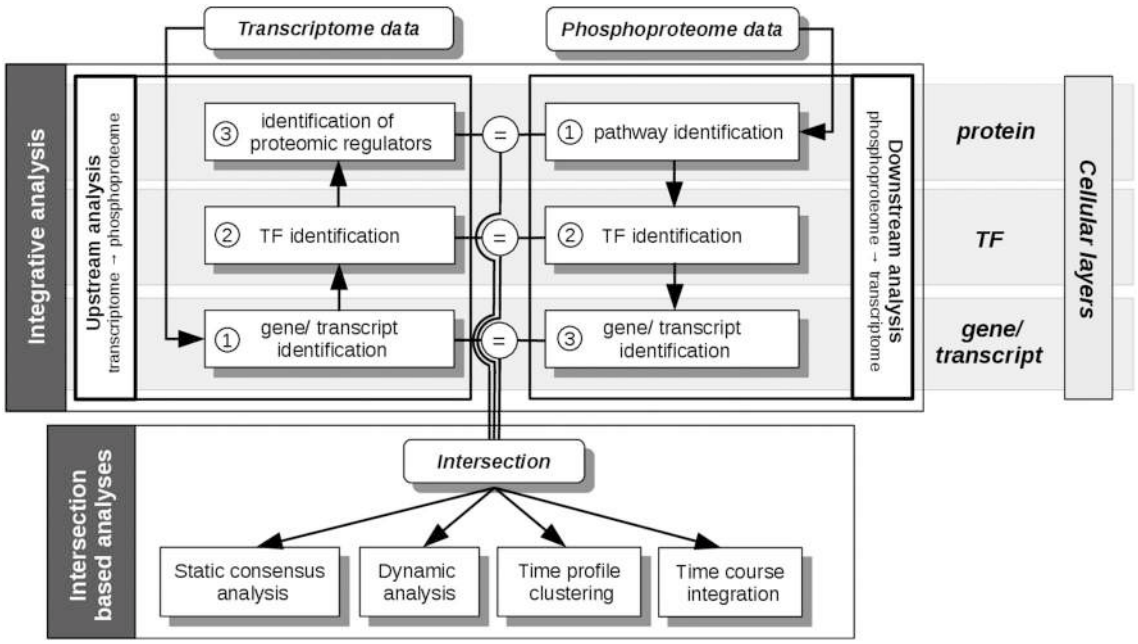


FIGURE 3 | pwOmics analyses steps. In the initial integrative analysis a downstream analysis of the phosphoproteome data and an upstream analysis of the transcriptome data is performed. The former includes the identification of pathways that include differentially abundant phosphoproteins, the identification of the TFs in these pathways and the determination of downstream target genes. In the upstream analysis the differentially expressed transcripts are identified, as well as their upstream TFs. By determining the pathways of these TFs also potential proteomic regulators can be identified. The intersection of the molecules on each cellular layer (protein, TF and gene/transcript) is determined before the intersection based analyses are performed. These include a static consensus analysis that can be performed for each measured time point, the consensus-based dynamic analysis that enables the generation of a probabilistic network exploiting the time-course information of those molecules that are part of the consensus analysis result. Furthermore, in a time profile clustering co-regulation patterns can be identified. Eventually, the time course integration allows to map downstream consensus transcripts with differentially abundant proteins. The "=" sign depicts the molecular overlap on each cellular layer, corresponding to the layer-specific consensus molecules.

abundant phosphoproteins, was performed. The transcription factors of these pathways were then found by matching the gene sets of the pathways against the transcription factors listed in the transcription factor—target gene database. Downstream target genes were identified, equivalently. The downstream analysis is based in general on the assumption of downstream regulation upon protein phosphorylation. Upstream analysis identified the upstream TFs of significantly differentially regulated transcripts. Subsequently, pathways including these TFs were identified in order to find possible upstream proteomic regulators of differentially expressed transcripts. The parameters chosen here corresponded to at least one TF per pathway for pathway identification and 10 orders of neighbors identified upstream of the TF for potential proteomic regulators. The results of each functional layer of signaling (pathway layer, TF layer, and gene/transcript layer) of downstream and upstream analysis were compared. These analyses steps were performed for each time point. Gene and protein ID matching was done by conversion of all IDs to HUGO gene symbols.

Static Consensus Analysis

In the static consensus analysis integrated signaling networks were constructed based on intersecting proteins, TFs, genes and transcripts on each functional layer (**Figure 4A**). The consensus proteins and TFs were mapped to the PPI STRING database and Steiner trees were generated via a shortest paths based approximation algorithm (Sadeghi and Fröhlich, 2013). The graphs were then completed by adding the corresponding TF—target interactions using TRANSFAC information. In case both consensus gene and consensus protein were part of the static consensus graph feedback loops were added.

Dynamic Consensus Analysis

In order to leverage the complete dynamic information from the data sets dynamic analysis was performed on basis of all consensus molecules (**Figure 4B**). The data associated with these nodes was used to fit cubic smoothing splines in order to generate a sufficiently dense data set for network inference via empirical Bayes estimation of a dynamic bayesian network with the R package ebdbNet (Rau et al., 2010). The generation of data points was based on the simplifying assumption of a gradual change of signaling over time. For further parameters default values were chosen. For visualization of the dynamic bayesian network a probability threshold was chosen which reflects a moderate number of regulatory interactions with a high probability in the network. The resulting threshold for plotting of the edges corresponded to a probability of an edge to be present by chance of 0.15.

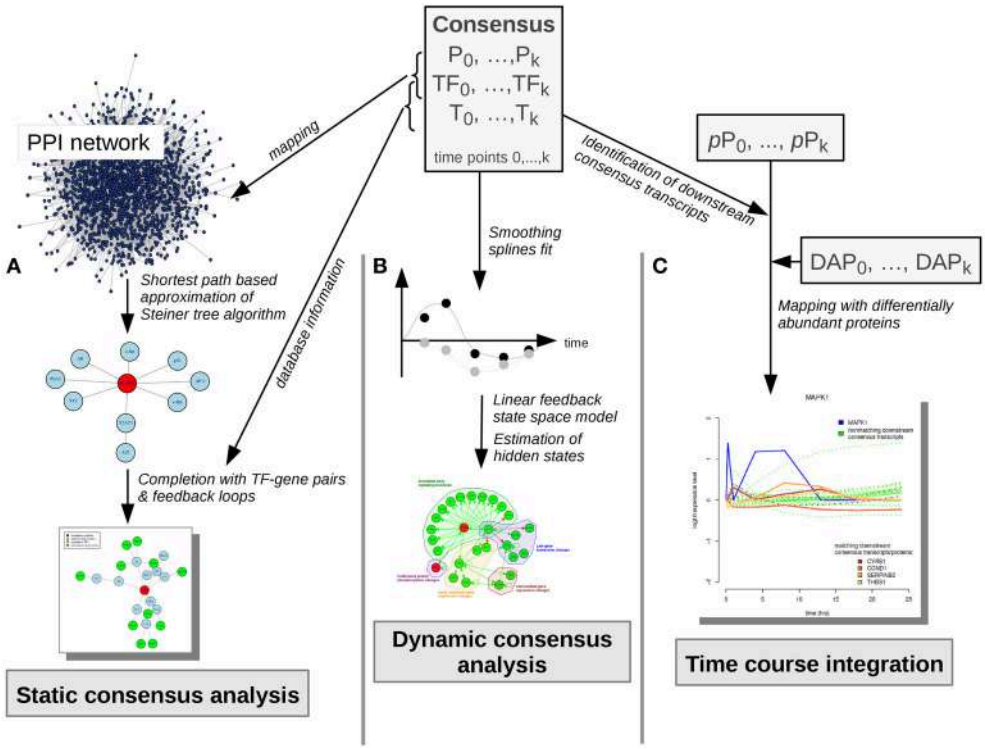


FIGURE 4 | Consensus molecule analyses. Consensus molecules on each cellular layer are used for the static consensus analysis, the dynamic consensus analysis and the time-course integration. **(A)** In the static consensus analysis static graphs are generated based on the PPI-mapped consensus proteins and transcription factors, an approximation of the Steiner tree algorithm is applied and the connected networks are complemented with TF-target interactions from TRANSFAC database. In case both consensus gene and corresponding consensus protein are part of the network, feedback loops are added. **(B)** In the dynamic consensus analysis smoothing splines are fitted to the time courses for all consensus molecules. Based on the higher density data set a linear feedback state space model is generated, hidden states are estimated and a probabilistic network is generated with dynamic Bayesian network inference (ebdbNet R package). **(C)** In the time course integration downstream consensus transcripts of the differentially abundant phosphoproteins are identified. These are mapped to the differentially abundant proteins. Time-courses of the downstream signaling players are visualized, subsequently. P, consensus proteins; TF, consensus transcription factors; T, consensus transcripts; pP, phosphoproteins; DAP, differentially abundant proteins.

Time Profile Clustering

Additionally, time profile clustering was performed in order to identify co-regulation patterns: Combining the described integration approach with a soft clustering implemented as fuzzy c-means algorithm (Kumar and Futschik, 2007) yielded an integrated time profile clustering based on the log-fold changes of consensus proteins and transcripts.

Time Course Integration

For further time course based integration with the proteome data set downstream consensus transcripts of the measured phosphoproteins were determined (**Figure 4C**). In a next step these were mapped to proteins, that were significantly differentially abundant at any time point (**Figure 2**, proteomic data).

RESULTS

Individual Downstream and Upstream Analyses

We performed individual downstream and upstream analyses of the phosphoproteome and microarray data sets taking into account the different functional layers of the cell the data originates from. The used pathway information exploits the signaling knowledge stored in public databases. **Figure 3** illustrates the steps of the individual analyses and further analysis steps explained in the next sections. **Table 1** shows the corresponding numbers of identified molecules and pathways on the different functional cellular layers in downstream and upstream analysis.

The data set for the phosphoproteome based downstream analysis is very small with only five phosphoprotein abundances investigated. However, as these were chosen thoroughly in the experiment we observe a considerable number of pathways that are influenced in downstream signaling. Altogether 121 pathways were identified when querying the four pathway databases used for the analysis. However, this set might include partly redundant pathways when originating from different databases, but describing the same signaling pathway. Pathways that are identified in every time point include e.g., the Biocarta "egf signaling" pathway, the NCI "EGF receptor (ErbB1) signaling pathway," the NCI pathway "EGFR-dependent Endothelin signaling events" or the NCI pathway "ErbB1 downstream signaling." Furthermore, a number of pathways are identified

TABLE 1 | Individual analysis.

Time after EGF stimulation [h]	0.25	1	4	8	13	18	24
DOWNSTREAM ANALYSIS							
No. of differentially abundant phosphoproteins	5	3	3	2	3	2	2
No. of pathways	121	68	98	90	81	79	79
No. of TFs	64	61	62	62	62	62	62
No. of potential target genes	1296	1293	1294	1294	1295	1295	1295
UPSTREAM ANALYSIS							
No. of differentially expressed transcripts	–	35	87	66	85	134	1551
No. of TFs	–	140	111	146	199	212	480
No. of pathways	–	163	154	169	200	200	230
No. of potential upstream proteomic regulators	–	871	950	897	920	976	1023

Downstream and upstream analyses characteristics over time. The expected bottleneck on the transcription factor layer can be observed. In the downstream analysis most pathways are overlapping, so we observe no large difference in the target gene numbers. The pre-processed proteomic data set comprises one time point of measurement more than the transcriptomic data set (0.25 h after EGF stimulation).

that are involved in cellular adhesion, STAT3 dependent signaling and PI3K signaling. Differential abundance of phopho-MAPK14 was only identified at time point 0.25 h after EGF stimulation. Corresponding pathways identified for that time point included e.g., the Biocarta "p38 mapk signaling pathway" and the Biocarta "mapkinase signaling pathway." According to the TF—target gene database the identified TFs activate the expression of a high number of genes as shown in **Table 1**.

In the transcriptome based upstream analysis an identification of upstream TFs was performed based on the differentially expressed transcripts. Corresponding numbers at each time point after EGF stimulation are displayed in **Table 1**. Identified upstream pathways included e.g., the "MAPK signaling pathway," the "EGF receptor (ErbB1) signaling pathway" and the "ErbB1 downstream signaling" pathway. The higher numbers of differentially expressed transcripts resulted likewise in the identification of more pathways. In those pathway sets the topological information enabled the identification of possible upstream proteomic regulators, subsequently.

The pathways identified in the downstream and upstream analyses at each measured time point after EGF stimulation are part of the Supplementary Material (**Tables S2**, **S3**).

Consensus Analysis

In the static consensus analysis we integrated the results of the different platforms for each time point on each functional layer. The aim was to reduce the individual downstream and upstream analyses results to molecule sets which include those molecules identified from both platforms and to reduce at the same time false positive molecules on the different functional layers. Exemplary, the consensus network of 1 h after EGF stimulation is shown in **Figure 5A**, later time point static consensus networks are part of the Supplementary Material (**Figures S2**–**S7**). These networks provide interaction and regulatory information on the consensus molecules. Yet, in our further analyses we focus on the static consensus profiles reflecting the presence of specific molecules in the consensus networks at each time point, as illustrated in **Figure 5B**. The static consensus profiles were used to explore the static consensus characteristics of certain molecules in order to evaluate the integration method. As dynamic signaling is especially interesting with regard to feedback signaling mechanisms and pathway crosstalk, we focus on these two signaling patterns in the following. **Figure 5B** shows the static consensus profiles of the members of the static consensus graph 1 h after EGF stimulation. A considerable number of genes being part of this consensus graph are exclusively found at this early time point. The profiles additionally show that both *PLAU,* the urokinase-type plasminogen activator, and *CTGF*, the connective tissue growth factor, comprise late regulatory changes. A figure with all static consensus profiles is part of the Supplementary Material (**Figure S1**). In these, 13 of 19 genes that are at least identified at two time points not including the 1 h time point after stimulation show a sustained pattern, indicative of a secondary cellular response. The genes without such a sustained pattern are *PLAU*, *CTGF* and *IL1A*, being already active 1 h after EGF stimulation or genes showing an intermediate activation.

Next, we investigated the pattern of proteins in the static consensus networks as well as the identified steiner nodes. The first group comprises the intersection of differentially abundant phosphoproteins in the proteomic data set and the potential upstream proteomic regulators of the differentially expressed genes. The second group is derived by generating Steiner trees after mapping the consensus molecules to the PPI network and might be functionally interesting, as its nodes are candidates for the regulation of the unconnected, mapped proteins. The static consensus profiles of the included proteins and the steiner node identified in this analysis are shown in **Figure 5C**. Transcription factor STAT3 is identified on the transcription factor layer at all-time points. MAPK1 is identified 4–8 h after EGF stimulation. PRKAR2B is identified later on (18–24 h after stimulation) on the protein layer. VAV1 is identified as a Steiner node in the static consensus graph 24 h after stimulation.

Additionally, we wanted to test in how far our integratory pathway-based approach is able to trace pathway crosstalk in the given data sets. In order to do so we chose a crosstalk mechanism which we expected to be reflected in the data set as it is not exclusively based on phosphorylation or ubiquitylation

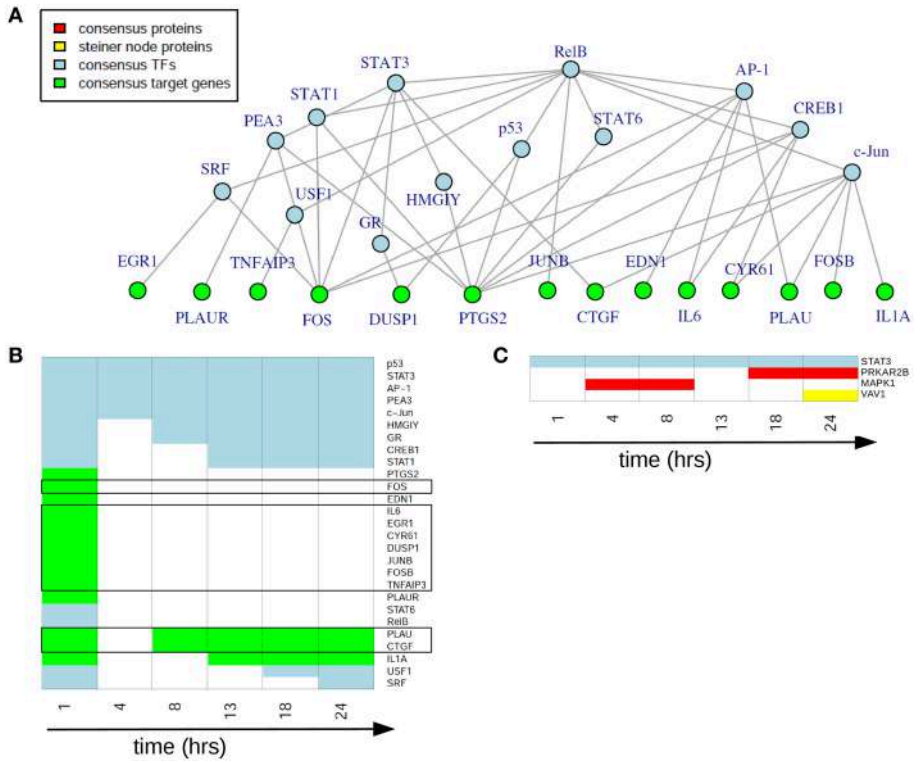


FIGURE 5 | Static consensus analysis results. (A) Static consensus graph for time point 1 h after EGF stimulation. **(B)** Static consensus profiles for members of the static consensus graph 1 h after EGF stimulation. Colors in the heatmap correspond to colors used in the consensus graphs, "white" boxes represent no membership in the consensus graph at that time point after EGF stimulation. Genes known to be IEGs (according to Tullai et al., 2007) are framed in black. **(C)** Static consensus profiles for selected proteins.

events. This mechanism is characterized by the activation of metalloproteinases (MMPs) by G-protein-coupled-receptors (GPCRs; Yarden and Sliwkowski, 2001). Upon activation MMPs cleave membrane-tethered ErbB ligands, which enables their binding to ErbB receptors, thereby positively regulating the ErbB signaling pathway. With EGFR being a receptor of the ErbB family our approach could identify a considerable number of the mentioned regulatory molecules in the consensus molecules (**Table 2**). Expression of different MMPs is observed starting at time point 4 h after EGF stimulation. Differentially expressed ErbB ligands for the different time points after EGF stimulation could be identified (such as self-induced EGF and AREG).

TABLE 2 | Consensus analysis.

Time after EGF stimulation [h]	1	4	8	13	18	24
MMPs	–	MMP1	MMP1	MMP1	MMP1	MMP1
			MMP1	MMP1	MMP1	MMP2
			MMP1	MMP1	MMP1	MMP10
ErbB ligands	–	–	–	EGF	AREG	AREG
					EGF	EGF

Regulatory molecules identified on the gene layer that are hypothesized to be involved in the signaling crosstalk via GPCRs and MMPs. GPCRs activate MMPs which then cleave the membrane-bound ErbB ligands leading to activated ErbB signaling (Yarden and Sliwkowski, 2001). Although differential expression is not direct evidence for the activity of these molecules, such regulatory mechanism can be hypothesized here.

Exploiting Dynamic Information of Coupled Time Course Data Sets

Our pathway-based approach additionally enables the utilization of the complete time-series for each molecule in order to generate a probabilistic network displaying those nodes of the network with a high posterior probability of interaction. The dynamic analysis is based on the simplifying assumption of a gradual change in signaling over time, as existing high-frequency components are not considered due to the small sampling rate. Each consensus molecule at any time point after EGF stimulation was taken into account. With this approach we obtained the probabilistic network displayed in **Figure 6**. This network is a reduced way to look at activating or inhibiting relationships between consensus proteins and genes. Here, we observe mainly activating relationships corresponding to an activation of the regulatory effect of EGF stimulation and not to upregulation directly. Likewise an inhibiting relationship in the network does not imply a downregulation, but the inhibition of the effects induced by EGF stimulation.

In total, we could identify five subgroups in the consensus-based dynamic network by mapping them to the times in which they are part of the consensus graphs (**Figure 6**): (1) immediate

early signaling processes, (2) early, but sustained gene expression changes, (3) intermediate gene expression changes, (4) late gene expression changes, and (5) continuous protein phosphorylation changes. In the group of the "immediate early signaling processes" most early response genes that were identified in the static consensus profiles are activated by the protein MAPK1 and the gene *IL1A*. This group reflects early phosphorylation induced transcriptional changes. The next group, consisting of five genes, is the group of "early, but sustained gene expression changes" upon EGF stimulation. It includes *CTGF*, a connective growth tissue factor. Its regulation is activated by MAPK1, *FKBP5*, *GADD45A* and also self-activation is observed. *CTGF* itself has activatory influence on gene members of its own group (*IGFBP3*, *FKBP5*), but also on members of the "intermediate gene expression changes" group and the "late gene expression changes" group. Two further members (*PLAU* and *ODC1*) are influenced by *IL1A*, a hub gene in the network, which we assigned to the "immediate early signaling processes" group and to the "late gene expression changes" group, as it shows immediate membership in the static consensus graphs, but also a late response profile. A small group showing intermediate gene expression changes comprises *TFRC* and *GADD45A*. We observe in the graph that *GADD45A* activates itself, but also *PCNA*, a gene of the "late gene expression changes" group. *PCNA* is additionally self-activated, as well as externally activated by the ErbB ligand *AREG* and *ASPH*, the aspartate beta-hydroxylase. *AREG* and *ASPH* are upregulated late after EGF stimulation. *IL1A* also activates *SLC3A2*, the solute carrier family 3 member 2, and inhibits *LAMA3*, a proliferating cell nuclear antigen, laminin alpha 3. The second protein being part of the network is the transcription factor STAT3. The changes in STAT3 phosphorylation are found in the consensus graphs over all time points, thus we assign it to the group of "continuous protein phosphorylation changes." Beside the activating influence of MAPK1 also autoregulation of STAT3 can be detected.

Time Profile Clustering

In order to identify co-regulation patterns in the signaling response after EGF stimulation we performed time profile clustering. We obtained four dynamic co-regulation patterns of which two exhibit positive regulation and two exhibit negative regulation. Both positive and negative clusters each comprise one cluster of immediate regulation and one of delayed regulation. The clusters are depicted in **Figure 7**. Corresponding molecule membership in the four different clusters is listed in the Supplementary Material (**Table S1**). Cluster 1 is immediately activated and thus contains various immediate early genes, but also the proteins MAPK1 and STAT3, which are part of the consensus-based dynamic analysis. Compared to the groups identified in the latter analysis this cluster constitutes the immediate early signaling processes together with early, but sustained gene expression changes. Cluster 2 is the biggest cluster with 52 members and is the delayed positively regulated cluster. Cluster 3 only comprises two members (RARRES3 and SLC3A2), both of which are showing a delayed negative dynamic co-regulation. Cluster 4 is the early negatively regulated cluster.

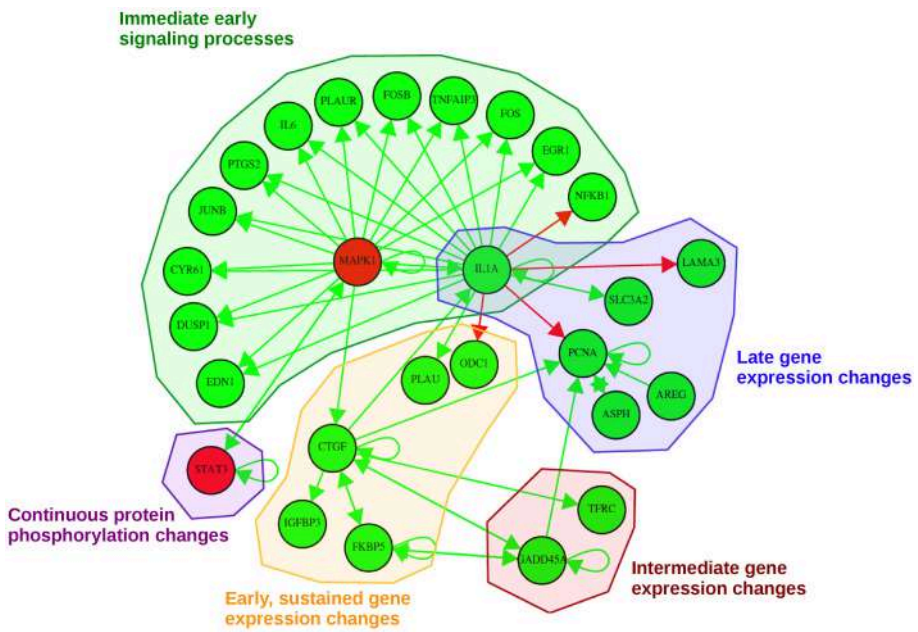


FIGURE 6 | Probabilistic network displaying result of the consensus-based dynamic analysis. For network inference all consensus genes and proteins at any time point were considered. For visualization of the dynamic bayesian network a probability threshold was chosen corresponding to a probability of an edge to be present by chance of 0.15. Five groups could be identified comprising direct immediate early signaling processes, continuous protein phosphorylation changes, late gene expression changes, intermediate gene expression changes and early, but sustained gene expression changes upon stimulation. Activating regulatory effects are depicted with green edges whereas inhibiting regulatory effects are depicted as red edges. Consensus protein nodes are colored in red, consensus transcript nodes in green. Activation/inhibition refers to changes in the regulatory effects initiated by EGF stimulation, not to activated or inhibited expression.

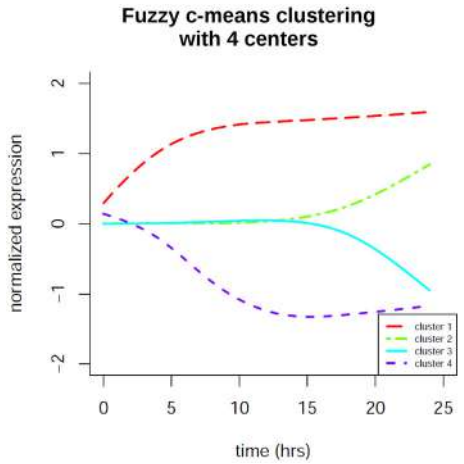


FIGURE 7 | Fuzzy c-means time profile clustering revealed 4 co-regulation clusters with distinct cluster sizes. Two of the clusters exhibit positive regulation and two exhibit negative regulation. Both positive and negative clusters each comprise one cluster of immediate regulation and one of delayed regulation. The clusters contain both protein and gene expression changes. Cluster membership is listed in the Supplementary Material.

Time Course Integration

The results of the time-course integration based on the consensus analysis results are displayed in **Figure 8** and in the Supplementary Material (**Figure S8**). Of the five phosphoproteins that were measured over time in the coupled data set we could identify four phosphoproteins with their downstream transcripts being part of our consensus analysis and mapping to differentially abundant proteins (MAPK1, STAT3, MAPK14, and PRKAR2B). MAPK1 downstream analysis revealed four transcripts (**Figure 8A**), which mapped to significantly differential proteins, CYR61—cysteine-rich angiogenic inducer 61, CCND1—cyclin D1, SERPINB2—serpin peptidase inhibitor, clade B, member 2, and THBS1—thrombospondin 1. MAPK1 itself shows increased phosphorylation levels in the very beginning after EGF stimulation and again between 1 and 13 h after EGF stimulation. In regard to temporal coordination CYR61 shows correlating temporal expression on the transcript and protein layer up to time point 4 h after EGF stimulation, but then a rather opposed pattern. CCND1 belongs to the group of cyclins and thus exhibits a specific expression and degradation pattern over the cell cycle, in this way contributing to the temporal coordination of mitotic events. Here we can observe an opposed temporal pattern of transcripts and proteins over the whole timespan measured: While on the mRNA layer, CCND1 shows higher expression levels after EGF stimulation, the corresponding proteins are found at lower levels over the whole time course. High mRNA-to-protein levels have already been reported by Waters et al. (2012). In the time-course SERPINB2 shows slowly rising levels of transcripts after EGF stimulation, whereas on the protein layer there is a direct decrease, an intermediate increase, and a second decrease again to the 0-level at 18 h after EGF stimulation. THBS1 protein levels are similar to that of SERPINB2, however, here we observe rather correlating transcript levels in the beginning and deviating ones after the 18 h time point.

STAT3 is the phosphoprotein showing the most downstream transcripts that match to significantly regulated proteins (**Figure 8B**). STAT3 itself shows sustained high expression levels over the whole time-course. All MAPK1 downstream transcripts that are part of the consensus analysis also belong to the downstream transcripts of STAT3. Further ones are *SLC3A2*, *FKBP5*, *PPP2CA*, *CD44,* and *ODC1*. All of these except for *ODC1* show anti-correlating patterns between transcripts and proteins until 4 h after EGF stimulation. For later time points most pairs exhibit correlating behavior. MAPK14 also has *CYR61*, *CCND1,* and *SERPINB2* as downstream targets with corresponding proteins being significantly differentially abundant, whereas for PRKAR2B only *CYR61* could be identified.

DISCUSSION

Pathway Layer Based Integration

In the downstream and upstream analyses the results indicate that pathway identification based on differentially abundant phosphoproteins and differentially expressed transcripts is effective. In both pathway sets those pathways known to be activated by EGF stimulation were identified reliably in the different databases, expectedly the "EGF signaling pathway" itself. This shows, that the two data sets are in concordance on the pathway layer even if they are measured on different cellular layers and analyzed individually. Based on these initial results a pathway-based integration was considered to be constructive. However, downstream and upstream analyses might also introduce false positive findings, which we aimed to reduce from further analysis steps by the subsequent intersection analysis. The small set of phosphoproteins measured over time gives a strong basis for the pathway layer based integration as they were selected carefully for the experiment and belong to key pathways in EGF signaling. However, a larger set of phosphoprotein data as obtained now e.g., from mass-spectrometry approaches could lead to more robust results.

Consensus Analysis Enables Identification of Regulatory Dynamics

In order to evaluate our methods it is important to first classify the data according to their temporal transcriptional domains. According to Avraham and Yarden (2011) feedback mechanisms in EGFR signaling can be assigned to two temporal domains, one of them being the immediate group which includes receptor endocytosis, secondary phosphorylation and further protein modifications, the other constituting the late group which includes newly synthesized adaptors, transcriptional repressors, RNA-binding proteins and phosphatases of the mitogen-activated protein kinase (MAPK) pathway. Especially the integrated data with parallel time points between 1 and 24 h after EGF stimulation thus reflects the late group capturing the transcriptional regulation with a wave-like regulation of

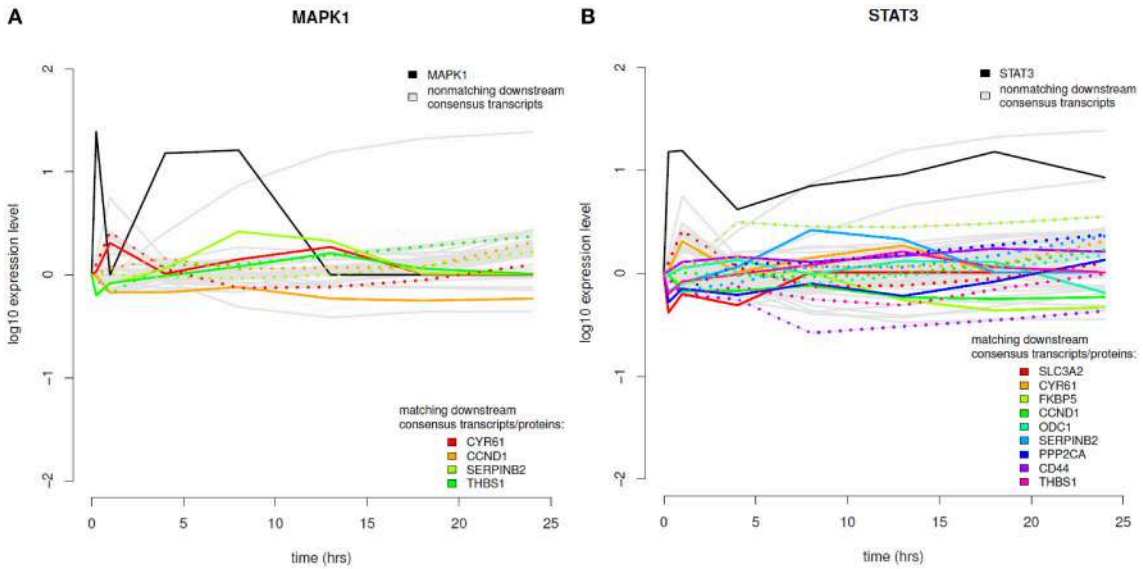


FIGURE 8 | Time course integration. (A) MAPK1 downstream consensus transcripts identified were mapped to differentially abundant proteins. **(B)** Time course integration for downstream consensus transcripts of STAT3. Note that the measurement range of the expression profiles across platforms can vary. Phosphoprotein time course data is shown in solid, black lines, non-matching transcript data in solid, gray lines and matching transcript and proteome data in rainbow color palette with proteins depicted as solid lines and transcripts depicted as dotted lines.

immediate early genes (IEGs), delayed early genes (DEGs), secondary response genes (SRGs; Avraham and Yarden, 2011) and their corresponding subsequent protein expression. IEGs are known to induce transcriptional changes of DEGs which then reduce the regulation of IEGs in a feedback subsequently, but initiate regulation of SRG expression. Based on this transcriptional regulation scheme the measured time points in the investigated data sets capture stimulation of both IEGs and DEGs 1 h after EGF stimulation while in subsequent time points we expect only regulation of SRGs, conferring the stable cellular phenotype.

We used the static consensus analysis in order to generate a static view on the integrated networks at each time point. Via static consensus profiles we can identify transcription factors with regulatory effects and their regulated consensus molecules on the gene layer at the 1 h time point. A large number of those genes were already reported to be IEGs in the cellular response to growth factor stimulation according to Tullai et al. (2007). PLAU and CTGF, regulated as well at later time points, apparently have an additional function in the definition of the phenotype. The two-phase regulation pattern indicates 2-fold tasks and can be interpreted to underly direct or indirect auto-feedback regulation.

The static consensus profiles of most SRGs, in contrast, are supposed to show a sustained activity. This is exactly what we find in our consensus graph analysis.

Due to the low number of differentially abundant phosphoproteins as a starting point the number of intersecting proteins from downstream and upstream analyses are low, as well. MAPK1 is involved in a variety of cellular growth processes such as proliferation and differentiation, thus its presence in the consensus graph corresponds well to the expected cellular response after EGF stimulation. As a regulatory subunit of the cAMP-dependent protein kinases PRKAR2B is involved in various cellular functions. With its late activity we suspect an involvement in the cellular reconstruction processes taking place for the final phenotype definition. The VAV proteins are guanine nucleotide exchange factors that activate pathways leading to cytoskeletal actin rearrangements and transcriptional alterations (Han et al., 1998). Thus, its functional association can be linked to cellular restructuring during proliferation.

In EGF signaling several pathways are involved which do not only process signals in a linear way but also enable cross-pathway regulatory influence on transcription. Oda et al. (2005) tried to compress all known signaling interactions into a comprehensive pathway map, resulting in a bow-tie architecture signaling pathway. As this network has to convey fine-tuned messages, it is deducable that slight dysregulation results in pathological transcriptional responses. Many crosstalk mechanisms have been investigated in more detail, most of them under pathological conditions. However, in order to understand the consequences of such dysregulation it is essential to also have a detailed understanding of physiological pathway crosstalk mechanisms. This is why we reviewed the consensus molecules in terms of their possible role in the crosstalk described by Yarden and Sliwkowski (2001). The large number of identified consensus molecules implicated in this crosstalk on the gene layer supports our hypothesis, that they are part of this signaling crosstalk mechanism.

As the described regulatory dynamic patterns are based on two independent data sets from different platforms we suppose that

this pattern is not identified due to measurement bias and thus has a biologically relevant function in the cellular response.

Identification of Regulatory Mechanisms by Exploiting Dynamic Information of Coupled Time Course Data Sets

In order to fully exploit the dynamic information of the time course data sets, we inferred a probabilistic network based on all consensus molecules. This network enables an identification of important players in the cellular response to EGF as well as the determination of inhibitory or activating regulation patterns.

The consensus proteins which are part of the dynamic network are MAPK1 and STAT3, both being part of the starting phosphoprotein data set. This indicates, that their important role in EGF signaling can be confirmed as such via the transcriptomic data set. STAT3 is a transcription factor, which is phosphorylated upon growth factor stimulation of the cell and builds homo- or heterodimers, which can then translocate to the nucleus and activate transcription (Park et al., 1996). It has multiple target genes with its protein products being involved in proliferative processes. MAPK1 is associated with cellular processes such as proliferation, differentiation and transcriptional regulation. Both show a self-activation as well as a mutual activation, which illustrates their functional relevance in EGF signaling. This regulatory interaction between MAPK1, also known as ERK2, and STAT3 is triggered via the activation of the MAPK/ERK cascade upon EGF stimulation, leading to MAPK1 phosphorylation by upstream kinases. STAT3 transcriptional activation by phosphorylation of STAT3 pS727 is then performed by the serine/threonine kinase ERK (Zhang and Liu, 2002), leading to activation of STAT3, which then acts as transcription factor and initiates the expression of downstream target genes. Target genes of STAT3 that might lead to further activation of MAPK1 are e.g., downstream transcription factors, multiplying indirectly the effective activation, or EGFR allowing for binding of more EGF. Furthermore, JAK2 is a target gene of STAT3, which can contribute to positive auto-feedback of STAT3 via the JAK-STAT pathway (Dauer et al., 2005).

Beside the already discussed early regulation processes and the protein phosphorylation changes of STAT3, the other identified groups are particularly interesting for further interpretation: The regulation of *CTGF*, the connective growth tissue factor, is activated by MAPK1, *FKBP5*, *GADD45A* and by itself. Interestingly, we observe auto-feedback regulation here, as already suspected from the static consensus profiles. *CTGF* is a hub gene in the consensus-based dynamic network, so the activation of its downregulation upon EGF stimulation is associated with downregulation of other genes in this cluster, such as *FKBP5*, or genes of the "intermediate gene expression changes" group. One of these is *GADD45A*, the growth arrest and DNA-damage-inducible alpha, which activates the regulation of PCNA. It is known to comprise increased transcript levels when cells are subjected to arrest conditions, treatment with DNA-damaging agents and environmental stresses (Hollander et al., 1993), thus we suspect the experimental design of the experiment with the chosen growth arrest time to be of no direct harm to the cells. PCNA, the proliferating cell nuclear antigen, is a cofactor of DNA polymerase delta and plays a central role during DNA replication. In DNA damage response it is positioned at the replication fork coordinating replication with DNA repair and DNA damage tolerance pathways (Cazzalini et al., 2014). Thus, its function is intensely needed in the phase of cellular remodeling and proliferation. The link between *GADD45A* and *PCNA*, that we determined with our integrative analysis, was previously reported (Chen et al., 1995).

AREG is upregulated in the "late gene expression changes" group as part of the regulatory pathway crosstalk loop via metalloproteinases described above and presumably provides an additional amplifying cellular way of an activation cascade after initial EGF stimulation. Also ASPH, which is thought to play an important role in calcium homeostasis (Treves et al., 2000), is part of this group. With its diverse roles e.g., as a messenger between cellular compartments calcium regulation is essential for proliferating cells.

IL1A, as another hub in the network, has immediate and late regulatory influence. In the "late gene expression changes" group it activates *SLC3A2*, solute carrier family 3 member 2, and inhibits *LAMA3*, proliferating cell nuclear antigen, laminin alpha 3. With their functions in regulating intracellular calcium levels, amino acid transport, formation and function of the basement membrane, cell migration and mechanical signal transduction and DNA replication, this part of the network rather shows the expression changes which represent the secondary (late) response of the cells.

In summary, we identified MAPK1, *IL1A* and *CTGF* as main players driving EGF stimulation response in the cell. Interestingly, we could detect the link between *GADD45A* and *PCNA* in two independent high-throughput time course data sets measured on different platforms using our pathway-based integration approach. As a matter of course, with a higher temporal resolution of the coupled time course measurements more accurate results can be identified by our approach, as less intermediate time points need to be estimated. To gain insight into the biological response after an external stimulation at least four time points after the stimulation time point are necessary, though there is a high information content in such coupled data sets on the different cellular layers. The chosen time points and the temporal resolution, however, need to be adjusted specifically to the cellular signaling dynamics and the stimulation of choice in order to reflect the crucial time points of regulation.

Time Profile Clustering Identifies Four Dynamic Co-Regulation Patterns Ruling EGF Signaling

With our time profile clustering approach we could identify four co-regulation patterns with distinct functions in the cellular response to EGF signaling. Cluster 1 contains many of the directly upregulated immediate early genes. Most of these are in fact downregulated again after their early response, which is not reflected by this cluster, as it contains also a considerable number of genes that are secondary response genes and are only upregulated at later time points (such as MMP1 or MMP10)

or immediate early genes which are upregulated again at later time points (PLAU or IL1A). Our hypothesis, that cluster 2 includes mainly genes upregulated as secondary response genes, responsible for the phenotype definition, holds true, when having a closer look to the members: We observe *CCND1*, the cyclin family protein, *ANXA1* and *ASPH*, *LAMA3* and *AREG*, which were identified in the consensus-based dynamic analysis in the group of late gene expression changes, *VEGFC*, a vascular endothelial growth factor promoting angiogenesis, *CCND2*—cyclin D2, *NME1*—nucleoside diphosphate kinase 1, which has been associated with high tumor metastatic potential based on different studies (MacDonald et al., 1996) and many more genes which act during cellular proliferation and migration. As cell cycle inhibitory protein coding genes we can observe the membership of *CDKN1A*, the cyclin-dependent kinase inhibitor 1A, which is tightly controlled by transcription factor p53 (He et al., 2005). Its membership in cluster 2 might be due to the high importance of balancing proliferation processes against growth stimulating processes in physiological tissue. Further we observe *PTHLH*, the parathyroid hormone-like hormone, to be part of this cluster, which regulates the epithelial-mesenchymal interactions during formation of mammary glands and teeth (Wysolmerski, 2012). Additionally the protein PRKAR2B is part of this cluster, indicating its late activation, which we already observe in the phosphoproteome data individually. However, here we see the confirmation that it is part of the consensus data from the two independent data sets generated on different platforms. Also *MMP2* is part of cluster 2 as well its regulatory counterpart, *TIMP1*, a metallopeptidase inhibitor. As the other metalloproteinases identified in the static consensus graphs (*MMP1* and *MMP10*) are not members of cluster 2, but of the immediately positively regulated cluster 1, it can be assumed, that *TIMP1* activation might also have a negative regulatory impact on these late after EGF stimulation. In the delayed downregulated cluster 3 we observe RARRES3, the retinoic acid receptor responder 3, which is known for its growth inhibitory effects (Hsu and Chang, 2015). A late downregulation thus can have the function of preventing contrasting growth signals. SLC3A2, the solute carrier family 3 member 2, encodes a subunit of a cell surface transmembrane protein complex responsible for regulation of L-type amino acid transport, which is essential for cellular growth and proliferation (Yanagida et al., 2001). Cluster 4, the early negatively regulated cluster, comprises *CTGF*, the connective tissue growth factor, whose downregulation might enhance proliferation of cells upon EGF stimulation. A further member is *IGFBP3*, the insulin-like growth factor binding protein 3, which potentiates insulin-like growth factor action and thereby also stimulates growth promoting effects (Cubbage et al., 1990). Supposedly, the cells do need less proliferating activation via IGF, when there is the growth-promoting stimulation of EGF. This underlines again that signaling patterns are tightly regulated in regard to their dynamics.

Time Course Integration of Consensus Graphs with Proteome Data

We were interested in how far our approach reveals the dynamics of elements in the regulatory cascade of a stimulation induced phosphorylation cascade triggering a specific gene expression, which then leads to the generation of proteins needed in the cellular response to that particular stimulation. Therefore, after integrating the phosphoproteome data in the first pathway layer based integration, we integrated in a second step also the proteome data with the results of our pathway-based integrative analysis dynamically. The delay between consensus transcript generation and their corresponding protein generation reflects the time the cell needs for the complete translational and post-translational process. However, it is known that differences in protein abundance are only attributable to mRNA levels by about 20–40% (Brockmann et al., 2007). This underlines the importance of post-translational modification and is the reason why we assumed the correlation between increasing and decreasing transcript expression and corresponding protein generation to be rather marginal.

For the interpretation of these results we need to be aware of the different ranges of the expression ratios in the data sets of different platforms. Thus, a direct comparison of the expression levels between transcripts and proteins is not possible, however, a dynamic interpretation is feasible.

Dynamically, we observe both correlating and non-correlating expression level patterns between transcripts and corresponding proteins. Based on the time resolution of the measurements we assume the time delay reflecting the translational and post-translational processes to be not necessarily observable in the data, as they can lie in a wide time range. Indeed, correlating behavior seems not to be shifted in time in our analysis for certain transcripts (e.g., for CYR61 up to 4 h after EGF stimulation or THBS1 up to 13 h after EGF stimulation), however, when performed on a time-series data set with higher resolution, such time shifts might be observable. Non-correlating expression level patterns indicate post-translational modifications or a possibly very rapid degradation of mRNA or the protein product, which is not captured in the low resolution time measurements. Of the identified pairs CYR61 is a growth factor inducible protein which promotes the adhesion of endothelial cells (Brigstock, 2002), CCND1 is a protein contributing to coordination of mitosis. High levels of SERPINB2 have been observed to exhibit an anti-proliferative effect (Croucher et al., 2008). In the time courses we see an intermediate increase of its protein levels, but an overall anti-correlating pattern between protein and transcript levels. THBS1, thrombospondin 1, is known as angiogenesis regulator (Chandrasekaran et al., 2000). Its protein levels are similar to that of SERPINB2, however, here we observe rather correlating expression levels, indicating less post-transcriptional modification. Also changes in the correlation behavior can be observed, indicative for a secondary regulatory influence. This could be induced by variations in mRNA degradation, protein degradation rates or post-translational modifications.

From the transcript/protein pairs that are observed as part of the regulatory loops CYR61, THBS1, and CCND1 clearly have a high influence on EGF stimulated cells during cellular proliferation, differentiation and survival, while the detection of SERPINB2 is more intriguing. It is known to inhibit urokinase plasminogen activators (PLAUs), but its physiological function has not been characterized comprehensively, although activity

in the adaptive immune response has been reported (Schroder et al., 2011). As we based the time-course integration on the consensus analysis the discussed time-courses are supported by both transcriptome and proteome data set. Thus, we hypothesize the interaction of SERPINB2 and PLAU, its inhibition target, to be of high relevance for proliferative processes. Our hypothesis is supported also by literature in the context of cancer: SERPINB2 has been associated with increased survival in breast cancer patients (Duffy, 2004).

With the integrated time-courses of phosphoproteins, downstream consensus-graph transcripts and their corresponding proteins the data implies an extensive post-translational modification of a number of proteins. This we see in the transcript/protein pairs investigated in detail here, but also in the downstream transcripts depicted in gray in **Figure 8**, with no corresponding proteins in the list of significantly differentially abundant proteins. Therefore, our results correspond to what is known about the low percentage of protein concentration variations that are affected by mRNA abundances directly (Vogel and Marcotte, 2012). However, our approach not only enables a general overall classification of correlating or anti-correlating transcript/protein pairs, but in addition a time-resolved interpretation of consensus-based regulatory processes.

Comparison of Separate Data Set Analysis with Integrated Consensus-Based Analysis

To comprehensively assess the advantage of our data integration approach based on public pathway knowledge we compared its results with the ones gained by a separate analysis of the individual proteomic and transcriptomic data sets. Waters et al. (2012) performed a separate pathway analysis and reported network statistics, such as the number of nodes in the largest cluster, the number of edges in the network and the two primary hub nodes, however, this analysis was limited to data measured 0–4 h after EGF stimulation. Interestingly, the hub genes identified in the microarray based network were the transcription factors *FOS* and *EGR1*, while the hub proteins identified in the proteome data were EGFR and ITGB1. Comparing these results to our results from the pathway-based integrative analysis, we likewise observe FOS and EGR1 to be highly important regarding regulatory mechanisms during the initial cellular response. Yet, we additionally derived further information than what is given by the separate analysis: We evaluated these genes to play a significant role in the immediate early cellular reaction based on static consensus profiles. Furthermore, we saw that these are mainly influenced by *IL1A* and the phosphorylation of MAPK1 directly as well as indirectly. Based on the time profile clustering we saw on top that they belong to the early positively regulated cluster. The protein hubs that are identified via the separate analysis, however, cannot be found in our consensus analysis, as the consensus is confined to the small set of measured phosphoproteins.

In a second separate analysis of the proteomic and transcriptomic data sets Waters et al. (2012) performed separate gene set enrichment on the basis of differentially expressed proteins and transcripts. The three most significant biological processes identified for the transcriptomic data set were "cell cycle," "mitosis," and "protein folding," while for the proteomic data set the most significant process was "protein synthesis." In a comparison the authors found considerable differences in the gene set enrichment results. Although this type of analysis is widely used for gene expression data it is arguable in how far "gene set" and "protein set" enrichment should be compared directly due to the different biological layers the data and possibly also network knowledge originates from. Thus, we see an inherent problem in the simplified layer-unspecific comparison with subsequent interpretation. Additionally, the results allow no conclusions or hypothesis generation on the molecular level.

In summary, we conclude that the integrated analysis of the two data sets moves the focus to the dynamic interplay of regulatory mechanisms and enables a layer specific and detailed regulatory analysis of the cellular response to external stimulation.

Comparison of Data Integration Approaches in Coupled High-Throughput Data Sets

The data integration approaches applied by Waters et al. (2012) were based on RNA/protein pairs cross-referenced between the platforms. However, no layer-specific analysis was performed. In a canonical correlation analysis the 199 RNA/protein pairs comprising all measurement time points were investigated with the result of intense post-transcriptional regulation on the protein layer. The benefit compared to a simple correlation analysis is that it captures also concordance or disconcordance of pairs when a temporal delay is observed. With our time-course integration we could also observe this effect, individually for specific phosphoprotein initiated signaling cascades. With our approach it is additionally possible to analyze transcriptional and translational dynamics of each cascade individually.

In the integrative analysis of Waters et al. (2012) major cell processes of the combined data were then ranked to early (0–4 h), intermediate (8–13 h) and late (18–24 h) time domains after EGF stimulation. A general shift from categories "cytoskeletal organization" and "regulation of cell cycle" (0–4 h) toward anti-apoptotic and cell adhesion pathways (8–13 h) was observed. An increased representation of the "mitosis" category between 18 and 24 h after stimulation corresponded to an increase of mitotic cells monitored by flow cytometry in parallel. A direct comparison of the analyses results is not possible here, though the results we found in the consensus-based dynamic analysis of the data agree roughly with the results of Waters et al. (2012), when comparing the function of individual consensus molecules with the GO biological process category names. Although having category names enables in general a better overview of the data, it does not allow individual identification of regulatory interactions. Therefore, we consider our approach as valuable additional method in order to get a better understanding of the dynamic biological processes.

Furthermore, integrated signaling networks from all data sets were investigated in Waters et al. (2012). Not surprisingly, the microarray data set contributed the highest number of nodes in the merged network. Compared to the signaling networks from

single data sets, the integrated network comprised increasingly linked nodes, reflected in the number of edges and the degree of the largest cluster reported. The two primary hub nodes of the integrated network were *FOS* and *SRC*, while the hub nodes in the network generated from exclusively microarray data were *FOS* and *EGR1*, generated exclusively from proteome data EGFR and ITGB1 and exclusively from phosphoproteome data STAT3 and MAPK1. Interestingly, we also found *FOS* and *EGR1*, as well as STAT3 and MAPK1 as consensus molecules in our consensus-based dynamic analysis with considerable regulatory influence during the cellular response after EGF stimulation. The proteome hub nodes EGFR and ITGB1, as well as the hub node *SRC* from the integrated network were not part of our results due to the low number of phosphoproteins measured in the study. However, we found already considerable amount of regulatory mechanisms when including only the phosphoproteome data set as initial data set in our analysis. The MMP cascades identified in the integrated analysis from Waters et al. (2012) as most robust response to EGF stimulation were identified as consensus molecule based process by our approach as well.

Unfortunately, in the integrated analysis of Waters et al. (2012) only time domains were considered in contrast to our individual time point analysis. This enables a rough summarized view on the signaling process, yet it does not fully exploit the information encoded in the dynamics. Likewise, the GO term analysis performed is based on a subset of RNA/protein pairs and results in a summarized interpretation, but it does not enable an individual regulatory mechanistic interpretation. Thus, we consider our approach as valuable complement in the analysis of coupled high-throughput data sets.

CONCLUSION

The presented data integration approach shows a way to gain a much deeper understanding of biological processes if time-course measurements and data from different high-throughput platforms representing the different functional layers of the cell are combined. Our approach enables a functional linking of regulatory processes over the transcriptional and translational cycle, even if the temporal resolution of the example data set is quite low, data has only been measured on two functional cellular layers and the phosphoproteome data set is very limited. This sets the basis for the integration of further cellular layers, as following regulation upon external perturbation in a detailed way provides a much deeper understanding of biological processing.

Bioinformatic tools like the R package *pwOmics* promote the generation of coupled data sets as they offer the possibility of an integrated analysis and help to sort the vast data sets in a biologically interpretable manner. By applying the different analysis steps implemented in *pwOmics* we showed that biological interpretation is facilitated and the results correspond to current biological knowledge about EGF stimulation generated in low and high-throughput experiments. Furthermore, we identified interesting regulatory relationships that were not observed yet in physiological EGF signaling. As our approach considers data from the different functional cellular layers individually, it enables to identify the regulatory interplay between these layers. We have demonstrated this in the consensus analysis, which is able to identify the molecular response minutes to hours after stimulation as feedback mechanism with a wave-like regulatory pattern generated by IEGs, DEGs, and SRGs and their corresponding proteins. We could also identify previously published pathway crosstalk via activation of MMPs (Yarden and Sliwkowski, 2001). Furthermore, we could ascertain the link in EGF signaling between the two molecules GADD45A and PCNA, in the investigated data sets, which was previously reported (Chen et al., 1995). Interestingly, we also found PTHLH in the consensus molecules as part of the secondary cellular response, which is involved in the formation of mammary glands (Wysolmerski, 2012). Furthermore, we could identify the regulatory interaction of PLAU and SERPINB2 to be also of high relevance in physiological EGF signaling. Compared with the previously performed integrative analysis on the coupled data set we gain a complementary, and much more detailed view on cellular signaling processes, enabling the generation of biological hypothesis about individual regulatory mechanisms involved in the dynamic interplay of signaling pathways and feedback responses. With the examples stated above we could show, that our integrative approach is able to identify regulatory patterns, molecular interactions and dynamically orchestrated cellular response mechanisms.

In order to link the different functional cellular layers it is beneficial and necessary to integrate knowledge from public databases which builds a frame for placing and linking the individual analysis results. This has the advantage of utilizing a vast amount of collected and curated information, which stays unused otherwise and can add an additional information layer for interpretation of the data. On the other hand this prior knowledge also directs the results in a certain extent, thus the quality of the databases used has to be taken into consideration when interpreting the overall results. A further caveat is that the public database knowledge available in most databases is not cell type or tissue specific resulting in a generalized analysis. However, as more cell type or tissue specific knowledge is collected such databases can be build up and integrated in the presented analysis workflow.

In the consensus-based dynamic analysis we make the simplifying assumption of a gradual change of signaling over time. Clearly, this does not hold true for individual cells and still is a rough assumption for a set of cells as there have been found oscillatory mechanisms which work at high frequencies (Avraham and Yarden, 2011), for example, and which are purely not identifiable via such a time resolution. However, we can still gain a lot of knowledge about the regulatory processes that are encoded in the comparably slow dynamic processes. Of course, there can be even more biologically functional layers measured in high-throughput experiments in a parallel manner over time, such as siRNA, epigenetic influences etc. At the moment such data sets are still rare, but we expect them to be generated increasingly. It will be interesting for future projects to include such additional layers into an integrative analysis.

We showed that the hypotheses on regulatory mechanisms generated via our integrative approach could be confirmed with

independent low-throughput data sets. Although such time-course data sets measured in parallel enable a detailed analysis, it is not yet possible to infer from these data sets every regulatory aspect in detail. Nevertheless, our approach is a step toward portraying the whole picture of regulatory influences on the molecular level.

AVAILABILITY

Main analysis steps of the pathway-based integration approach of coupled time-series omics data described in this manuscript are implemented in the R package *pwOmics* (Wachter and Beissbarth, 2015).

AUTHOR CONTRIBUTIONS

AW developed the method, performed data analysis and wrote the manuscript. TB conceived the design, envisioned the project and revised the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support by BMBF e:Bio program grant MetastaSys [0316173A] and by BMBF e:Med grant MMML-Demonstrators [031A428B]. We additionally acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00351

Figure S1 | Static consensus profiles of all members of the static consensus graphs. Color coding corresponds to the one used in the static consensus graphs (red, consensus proteins; yellow, steiner node proteins; lightblue, consensus transcription factors; green, consensus genes).

Figure S2 | Static consensus graphs for time points 1 h after EGF stimulation.

Figure S3 | Static consensus graphs for time points 4 h after EGF stimulation.

Figure S4 | Static consensus graphs for time points 8 h after EGF stimulation.

Figure S5 | Static consensus graphs for time points 13 h after EGF stimulation.

Figure S6 | Static consensus graphs for time points 18 h after EGF stimulation.

Figure S7 | Static consensus graphs for time points 24 h after EGF stimulation.

Figure S8 | Time course integration for phosphoproteins MAPK14 and PRKAR2B. Downstream consensus transcripts identified for MAPK14 and PRKAR2B were mapped to differentially abundant proteins. Note that the measurement range of the expression profiles across platforms can vary. Phosphoprotein time course data is shown in solid, black lines, non-matching transcript data in solid, gray lines and matching transcript and proteome data in rainbow color palette with proteins depicted as solid lines and transcripts depicted as dotted lines.

Table S1 | List of molecule cluster membership in the time profile analysis. Data origin is encoded in the abbreviation after each protein/gene name (_g, microarray data; _p, proteome data).

Table S2 | Lists of pathways identified in the downstream analysis based on the phosphoprotein data for time points 0.25, 1, 4, 8, 13, 18, and 24 h after EGF stimulation. Table includes information about the pathway database used for pathway identification (as part of their ID) and the corresponding pathway names.

Table S3 | Lists of pathways identified in the upstream analysis based on the differentially expressed transcripts for time points 1, 4, 8, 13, 18, and 24 h after EGF stimulation.

REFERENCES

Avraham, R., and Yarden, Y. (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.* 12, 104–117. doi: 10.1038/nrm3048

Balbin, O. A., Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., et al. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun.* 4:2617. doi: 10.1038/ncomms3617

Bar-Joseph, Z., Gitter, A., and 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

Brigstock, D. R. (2002). Regulation of angiogenesis and endothelial cell function by connective tissue growth factor (CTGF) and cysteine-rich 61 (CYR61). *Angiogenesis* 5, 153–165. doi: 10.1023/A:1023823803510

Brockmann, R., Beyer, A., Heinisch, J. J., and Wilhelm, T. (2007). Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput. Biol.* 3:e57. doi: 10.1371/journal.pcbi.0030057

Cazzalini, O., Sommatis, S., Tillhon, M., Dutto, I., Bachi, A., Rapp, A., et al. (2014). CBP and p300 acetylate PCNA to link its degradation with nucleotide excision repair synthesis. *Nucleic Acids Res.* 42, 8433–8448. doi: 10.1093/nar/gku533

Chandrasekaran, L., He, C.-Z., Al-Barazi, H., Krutzsch, H. C., Iruela-Arispe, M. L., and Roberts, D. D. (2000). Cell contact–dependent activation of α3β1 integrin modulates endothelial cell responses to thrombospondin-1. *Mol. Biol. Cell.* 11, 2885–2900. doi: 10.1091/mbc.11.9.2885

Chen, I. T., Smith, M. L., O'Connor, P. M., and Fornace, A. J. (1995). Direct interaction of Gadd45 with PCNA and evidence for competitive interaction of Gadd45 and p21Waf1/Cip1 with PCNA. *Oncogene* 11, 1931–1937.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weise, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Croucher, D. R., Saunders, D. N., Lobov, S., and Ranson, M. (2008). Revisiting the biological roles of PAI2 (SERPINB2) in cancer. *Nat. Rev. Cancer* 8, 535–545. doi: 10.1038/nrc2400

Cubbage, M. L., Suwanichkul, A., and Powell, D. R. (1990). Insulin-like growth factor binding protein-3. Organization of the human chromosomal gene and demonstration of promoter activity. *J. Biol. Chem.* 265, 12642–12649.

Dauer, D. J., Ferraro, B., Song, L., Yu, B., Mora, L., Buettner, R., et al. (2005). Stat3 regulates genes common to both wound healing and cancer. *Oncogene* 24, 3397–3408. doi: 10.1038/sj.onc.1208469

Ding, Y., Chen, M., Liu, Z., Ding, D., Ye, Y., Zhang, M., et al. (2012). atBioNet– an integrated network analysis tool for genomics and biomarker discovery. *BMC Genomics* 13:325. doi: 10.1186/1471-2164-13-325

Duffy, M. J. (2004). The urokinase plasminogen activator system: role in malignancy. *Curr. Pharm. Des.* 10, 39–49. doi: 10.2174/1381612043453559

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

Hamon, J., Jennings, P., and Bois, F. Y. (2014). Systems biology modeling of omics data: effect of cyclosporine a on the Nrf2 pathway in human renal cells. *BMC Syst. Biol.* 8:76. doi: 10.1186/1752-0509-8-76

Han, J., Luby-Phelps, K., Das, B., Shu, X., Xia, Y., Mosteller, R. D., et al. (1998). Role of substrates and products of PI 3-kinase in regulating activation of rac-related guanosine triphosphatases by Vav. *Science* 279, 558–560. doi: 10.1126/science.279.5350.558

He, G., Siddik, Z. H., Huang, Z., Wang, R., Koomen, J., Kobayashi, R., et al. (2005). Induction of p21 by p53 following DNA damage inhibits both Cdk4 and Cdk2 activities. *Oncogene* 24, 2929–2943. doi: 10.1038/sj.onc.1208474

Herbst, R. S. (2004). Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys.* 59, 21–26. doi: 10.1016/j.ijrobp.2003.11.041

Hollander, M. C., Alamo, I., Jackman, J., Wang, M. G., McBride, O. W., and Fornace, A. J. (1993). Analysis of the mammalian gadd45 gene and its response to DNA damage. *J. Biol. Chem.* 268, 24385–24393.

Hsu, T.-H., and Chang, T.-C. (2015). RARRES3 regulates signal transduction through post-translational protein modifications. *Mol. Cell. Oncol.* 2:e999512. doi: 10.1080/23723556.2014.999512

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Meth.* 12, 115–121. doi: 10.1038/nmeth.3252

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A., and Beissbarth, T. (2013). rBiopaxParser - an R package to parse, modify and visualize BioPAX data. *Bioinformatics* 29, 520–522. doi: 10.1093/bioinformatics/bts710

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A. L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Kumar, L., and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* 2, 5–7. doi: 10.6026/97320630002005

Lurje, G., and Lenz, H. J. (2009). EGFR signaling and drug discovery. *Oncology* 77, 400–410. doi: 10.1159/000279388

MacDonald, N. J., Freije, J. M. P., Stracke, M. L., Manrow, R. E., and Steeg, P. S. (1996). Site-directed Mutagenesis of nm23-H1 mutation of proline 96 or serine 120 abrogates its motility inhibitory activity upon transfection into human breast carcinoma cells. *J. Biol. Chem.* 271, 25107–25116. doi: 10.1074/jbc.271.41.25107

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

Nishimura, D. (2001). BioCarta. *Biotechnol. Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* 1, 2005.0010. doi: 10.1038/msb4100014

Park, O. K., Schaefer, T. S., and Nathans, D. (1996). *In vitro* activation of Stat3 by epidermal growth factor receptor kinase. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13704–13708. doi: 10.1073/pnas.93.24.13704

Purvis, J. E., and Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. *Cell* 152, 945–956. doi: 10.1016/j.cell.2013.02.005

Rau, A., Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). An empirical bayesian method for estimating biological networks from temporal microarray data. *Stat. Appl. Genet. Mol. Biol.* 9:9. doi: 10.2202/1544-6115.1513

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16, 85–97. doi: 10.1038/nrg3868

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., et al. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* 24, 2894–2900. doi: 10.1093/bioinformatics/btn553

Sadeghi, A., and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics* 14:144. doi: 10.1186/1471-2105-14-144

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653

Schroder, W. A., Major, L., and Suhrbier, A. (2011). The role of SerpinB2 in immunity. *Crit. Rev. Immunol.* 31, 15–30. doi: 10.1615/CritRevImmunol.v31.i1.20

Sun, H., Wang, H., Zhu, R., Tang, K., Gong, Q., Cui, J., et al. (2014). iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* 30, 737–739. doi: 10.1093/bioinformatics/btt576

Treves, S., Feriotto, G., Moccagatta, L., Gambari, R., and Zorzato, F. (2000). Molecular cloning, expression, functional characterization, chromosomal localization, and gene structure of junctate, a novel integral calcium binding protein of Sarco(endo)plasmic reticulum membrane. *J. Biol. Chem.* 275, 39555–39568. doi: 10.1074/jbc.M005473200

Tullai, J. W., Schaffer, M. E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G. M. (2007). Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *J. Biol. Chem.* 282, 23981–23995. doi: 10.1074/jbc.M702044200

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi: 10.1038/nrg3185

Wachter, A., and Beissbarth, T. (2015). pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics* 31, 3072–3074. doi: 10.1093/bioinformatics/btv323

Wang, D., Xia, D., and Dubois, R. N. (2011). The crosstalk of PTGS2 and EGF signaling pathways in colorectal cancer. *Cancers* 3, 3894–3908. doi: 10.3390/cancers3043894

Waters, K. M., Liu, T., Quesenberry, R. D., Willse, A. R., Bandyopadhyay, S., Kathmann, L. E., et al. (2012). Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PLoS ONE* 7:e34515. doi: 10.1371/journal.pone.0034515

Wysolmerski, J. J. (2012). Parathyroid hormone-related protein: an update. *J. Clin. Endocrinol. Metab.* 97, 2947–2956. doi: 10.1210/jc.2012-2142

Yanagida, O., Kanai, Y., Chairoungdua, A., Kim, D. K., Segawa, H., Nii, T., et al. (2001). Human L-type amino acid transporter 1 (LAT1): characterization of function and expression in tumor cell lines. *Biochim. Biophys. Acta* 1514, 291–302. doi: 10.1016/s0005-2736(01)00384-4

Yarden, Y., and Sliwkowski, M. X. (2001). Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell Biol.* 2, 127–137. doi: 10.1038/35052073

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* 41, 316–323. doi: 10.1038/ng.337

Zhang, W., and Liu, H. T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* 12, 9–18. doi: 10.1038/sj.cr.7290105

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 April 2016
doi: 10.3389/fgene.2016.00044

Boolean Modeling Reveals the Necessity of Transcriptional Regulation for Bistability in PC12 Cell Differentiation

Barbara Offermann[1‡], Steffen Knauer[1†‡], Amit Singh[‡], María L. Fernández-Cachón[1], Martin Klose[1], Silke Kowar[1], Hauke Busch[1,2,3§] and Melanie Boerries[1,2,3*§]*

[1] Systems Biology of the Cellular Microenvironment Group, Institute of Molecular Medicine and Cell Research, Albert-Ludwigs-University Freiburg, Freiburg, Germany, [2] German Cancer Consortium, Freiburg, Germany, [3] German Cancer Research Center, Heidelberg, Germany

OPEN ACCESS

Edited by:
Ekaterina Shelest,
Hans-Knöll-Institute, Germany

Reviewed by:
Julio Vera González,
University Hospital Erlangen, Germany
Nils Blüthgen,
Charité-Universitätsmedizin Berlin, Germany

***Correspondence:**
Hauke Busch
h.busch@dkfz.de;
Melanie Boerries
m.boerries@dkfz.de

†Present Address:
Steffen Knauer,
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

‡These authors are co-first authors.
§These authors are co-last authors.

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 20 November 2015
Accepted: 14 March 2016
Published: 14 April 2016

Citation:
Offermann B, Knauer S, Singh A, Fernández-Cachón ML, Klose M, Kowar S, Busch H and Boerries M (2016) Boolean Modeling Reveals the Necessity of Transcriptional Regulation for Bistability in PC12 Cell Differentiation. Front. Genet. 7:44. doi: 10.3389/fgene.2016.00044

The nerve growth factor NGF has been shown to cause cell fate decisions toward either differentiation or proliferation depending on the relative activity of downstream pERK, pAKT, or pJNK signaling. However, how these protein signals are translated into and fed back from transcriptional activity to complete cellular differentiation over a time span of hours to days is still an open question. Comparing the time-resolved transcriptome response of NGF- or EGF-stimulated PC12 cells over 24 h in combination with protein and phenotype data we inferred a dynamic Boolean model capturing the temporal sequence of protein signaling, transcriptional response and subsequent autocrine feedback. Network topology was optimized by fitting the model to time-resolved transcriptome data under MEK, PI3K, or JNK inhibition. The integrated model confirmed the parallel use of MAPK/ERK, PI3K/AKT, and JNK/JUN for PC12 cell differentiation. Redundancy of cell signaling is demonstrated from the inhibition of the different MAPK pathways. As suggested *in silico* and confirmed *in vitro*, differentiation was substantially suppressed under JNK inhibition, yet delayed only under MEK/ERK inhibition. Most importantly, we found that positive transcriptional feedback induces bistability in the cell fate switch. *De novo* gene expression was necessary to activate autocrine feedback that caused Urokinase-Type Plasminogen Activator (uPA) Receptor signaling to perpetuate the MAPK activity, finally resulting in the expression of late, differentiation related genes. Thus, the cellular decision toward differentiation depends on the establishment of a transcriptome-induced positive feedback between protein signaling and gene expression thereby constituting a robust control between proliferation and differentiation.

Keywords: PC12 cells, Boolean modeling, NGF signaling, EGF signaling, bistability

1. INTRODUCTION

The rat pheochromocytoma cells PC12 are a long established *in vitro* model to study neuronal differentiation, proliferation and survival (Greene and Tischler, 1976; Burstein et al., 1982; Cowley et al., 1994). After stimulation with the nerve growth factor (NGF), a small, secreted protein from the neurotrophin family, PC12 cells differentiate into sympathetic neuron-like cells, which is

morphologically marked by neurite outgrowth over a time course of up to 6 days (Levi-Montalcini, 1987; Chao, 1992; Fiore et al., 2009; Weber et al., 2013). NGF binds with high affinity to the TrkA receptor (tyrosine kinase receptor A), thereby activating several downstream protein signaling pathways including primarily the protein kinase C/phospholipase C (PKC/PLC), the phosphoinositide 3-kinase/protein kinase B (PI3K/AKT) and the mitogen-activated protein kinase/extracellular signal-regulated kinase (MAPK/ERK) pathways (Kaplan et al., 1991; Jing et al., 1992; Vaudry et al., 2002). Beyond these immediate downstream pathways, further studies showed the involvement of Interleukin 6 (IL6), Urokinase plasminogen activator (uPA) and Tumor Necrosis Factor Receptor Superfamily Member 12A (TNFRSF12A) in PC12 cell differentiation (Marshall, 1995; Wu and Bradshaw, 1996; Leppä et al., 1998; Xing et al., 1998; Farias-Eisner et al., 2000, 2001; Vaudry et al., 2002; Tanabe et al., 2003). Sustained ERK activation is seen as necessary and sufficient for the successful PC12 cell differentiation under NGF stimulation (Avraham and Yarden, 2011; Chen et al., 2012), whereas transient ERK activation upon epidermal growth factor (EGF) stimulation results in proliferation (Gotoh et al., 1990; Qui and Green, 1992; Marshall, 1995; Vaudry et al., 2002). In fact, selective pathway inhibition or other external stimuli that modulate the duration of ERK activation likewise determine the cellular decision between proliferation and differentiation (Dikic et al., 1994; Vaudry et al., 2002; Santos et al., 2007). Consequently, the MAPK signaling network, as the key pathway in the cellular response, has been studied thoroughly *in vitro* and *in silico* (Sasagawa et al., 2005; von Kriegsheim et al., 2009; Saito et al., 2013). Interestingly, both EGF and NGF provoke a similar transcriptional program within the first hour. Therefore, differences in cellular signaling must be due (i) to differential regulation of multiple downstream pathways and (ii) late gene response programs (>1 h) that feed back into the protein signaling cascade. As an example for pathway crosstalk, both, the MAPK/ERK and c-Jun N-terminal kinase (JNK) pathways regulate c-Jun activity and are necessary for PC12 cell differentiation (Leppä et al., 1998; Waetzig and Herdegen, 2003; Marek et al., 2004), while uPA receptor (uPAR) signaling, as a result of transcriptional AP1 (Activator Protein-1) regulation, is necessary for differentiation of unprimed PC12 cells (Farias-Eisner et al., 2000; Mullenbrock et al., 2011).

In the present study, we combined time-resolved transcriptome analysis of EGF and NGF stimulated PC12 cells up to 24 h with inhibition of MAPK/ERK, JNK/JUN, and PI3K/AKT signaling, to develop a Boolean Model of PC12 cell differentiation that combines protein signaling, gene regulation and autocrine feedback. The Boolean approach allows to derive important predictions without detailed quantitative kinetic data and parameters over different time scales (Singh et al., 2012). Protein signaling comprised MAPK/ERK, JNK/JUN, and PI3K/AKT pathways. Based on the upstream transcription factor analysis and transcriptional regulation of *Mmp10* (Matrix Metallopeptidase 10), *Serpine1* (Serpin Peptidase Inhibitor, Clade E, Member) and *Itga1* (Integrin, Alpha 1), we further included an autocrine feedback via uPAR signaling. The model topology was trained on the transcriptional response after pathway inhibition. Inhibition of JNK completely blocked PC12 cell differentiation and long-term expression of target transcription factors (TFs), such as various Kruppel-like factors (*Klf2*, *4*, *6* and *10*), *Maff* (V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog F) and AP1. Interestingly, inhibition of MEK (mitogen-activated protein kinase kinase), blocking the phosphorylation of ERK, slowed down, but not completely abolished cell differentiation. Neurite quantification over 6 days confirmed a late and reduced, but significant PC12 differentiation, which hinted at alternative pathway usage through JNK. Inhibition of the PI3K/AKT pathway, which is involved in cell proliferation (Chen et al., 2012), even increased the neuronal morphology and neurite outgrowth.

In conclusion, our Boolean modeling approach shows the complex interplay of protein signaling, transcription factor activity and gene regulatory feedback in the decision and perpetuation of PC12 cell differentiation after NGF stimulation.

2. MATERIALS AND METHODS

2.1. Cell Culture and Stimulation

PC12 cells were obtained from ATCC (American Type Culture Collection, UK) and were cultured at 37°C at 5% CO_2 in RPMI 1640 medium, supplemented with 10% Horse Serum, 5% Fetal Bovine Serum, 1% penicillin/streptomycin (PAN Biotech, Germany) and 1% glutamine (PAN Biotech, Germany). For cell stimulation, 500,000 cells/well were seeded on collagen coated 6 well plates (Corning, NY, USA). The following day, cells were stimulated with 50 ng/ml rat nerve growth factor (NGF; Promega, Madison, WI, USA) or 75 ng/ml epidermal growth factor (EGF; R&D Systems; Wiesbaden, Germany) for the corresponding times. For the pathway inhibition experiments, the following inhibitors were used and added 60 min before NGF was added, mitogen-activated protein inhibitor at a concentration of 20 μM (MEKi; U0126 from Promega, Madison, WI, USA), phosphoinositide 3-kinase inhibitor at a concentration of 40 μM (PI3Ki; LY-294002 from Enzo Life Sciences, New York, USA) and c-Jun N-terminal kinase inhibitor at a concentration of 20 μM (JNKi; SP600125 from Sigma-Aldrich, St. Louis, USA). The inhibitors were dissolved in DMSO and were further diluted in cell culture medium at their working concentration. Control cells were treated with DMSO at the same concentration that was present in the cells with inhibitor treatment.

2.2. RNA Isolation and Quantitative Real Time PCR (qRT-PCR)

Total RNA was isolated from 500,000 cells per timepoint according to the manufacturer's protocol (Universal RNA Purification Kit, Roboklon, Germany). RNA integrity was measured using an Agilent Bioanalyzer-2000 (Agilent Technologies GmbH, Waldbronn, Germany), and its content quantified by NanoDrop ND-1000 (Thermo Fisher Scientific, Wilmington, USA). For RT-qPCR, double strand cDNA was synthesized from 1 μg of total RNA using the iScript™ cDNA Synthesis kit (Quanta Biosciences, Gaithersburg, USA) according to the manufacturer instructions. RT-qPCR was performed in a CFX96 instrument (BioRad, Hercules, CA, USA) using a

SYBR Green master mix. Relative gene expression levels were calculated with the 2-ΔΔCt method, using HPRT1 and 18S ribosomal RNA as reference genes. Post-run analyses were performed using Bio-Rad CFX Manager version 2.0 and the threshold cycles (Cts) were calculated from a baseline subtracted curve fit. See Supplementary Table 1 for primer pair sequences.

2.3. Microscopy and Quantification

Live phase contrast images from PC12 cells under the different conditions were acquired using a Nikon Eclipse Ti Inverted Microscope (Nikon; Düsseldorf, Germany) equipped with a Perfect Focus System (PFS) and a Digital cooled Sight Camera (DS-QiMc; Nikon, Germany) as described in (Weber et al., 2013). Briefly, PC12 cells were cultured in collagen coated 6-well plates (500,000 cells/well) and treated as described in "Cell culture and stimulation" and 150 images per well, every second day were recorded with the same spatial pattern. Cell differentiation is calculated by the ratio of the two described imaging features (Weber et al., 2013) convex hull (CH) to cell area (CA) for 150 images per well over 6 days (Weber et al., unpublished data).

2.4. Western Blot

For each timepoint and condition 3×10^6 PC12 cells (for inhibition experiments) or 5×10^6 PC12 cells (for EGF vs. NGF comparison) were seeded in 10cm collagen coated Cell BIND dishes (Corning; Germany). Cells were collected after 5, 10, 30 min, 1, 2, 4, 6, 8, 12, 24, and 48 h in 200 μl RIPA buffer (containing 0.5% SDS), supplemented with proteinase inhibitor (complete mini EDTA free tablets, Roche, Basel, Switzerland) and Benzonase (Merck), and lysed for 20 min under agitation. A total of 30 μg protein was loaded per lane and run in 10% SDS- polyacrylamide gels, transferred to polyvinylidene difluoride membranes. Membranes were cut horizontally into fragments according to the expected sizes of the protein of interest and immunoblotted with antibodies against total p44/42 (ERK1/2, 1:2000, #9102S, Cell Signaling Technology [CST]), phospho p44/42 (pERK1/2, 1:2000, #9101S, CST), total JNK (JNK1/2, 1:1000, #9258S, CST), phospho JNK (Thr183/Tyr185, 1:1000, #4668S, CST), total AKT (1:1000, #4691S, CST), phospho AKT (1:1000, Ser473, #9271S, CST) or GAPDH (1:2000,# MAB374, Millipore) overnight at 4°C. Proteins were visualized with chemiluminescence on SuperSignal West Pico Chemiluminiscent Substrate imager (Thermo-Fischer, Massachusetts, USA) after 1h of incubation with appropriate horseradish peroxidase-linked secondary antibody (Sigma-Aldrich). Immunoblots were quantified using ImageJ (image analyzer camera LAS4000, Fujifilm, Tokyo, Japan). Blots were normalized to total GAPDH and an internal standard (IS) was used for normalization between membranes.

2.5. Microarray Analysis and Data Pre-processing

Time-resolved gene expression data of stimulated PC12 were recorded at $t = [1, 2, 3, 4, 5, 6, 8, 12, 24]$ h and $t = [1, 2, 3, 4, 6, 8, 12, 24]$ h for NGF and EGF stimulation, respectively. Control timepoints were measured at 0, 2, 4, 6, 8, 12, 24 h. Total RNA was isolated, labeled and hybridized to an Illumina RatRef-12 BeadChip (Illumina, San Diego, CA, USA) according to the manufacturers protocol. Raw microarray data were processed and quantile normalized using the Bioconductor R package beadarray (Ritchie et al., 2011). Illumina Probes were mapped to reannotated Entrez IDs using the Illumina Ratv1 annotation data (v. 1.26) from Bioconductor. If several probes mapped to the same Entrez ID, the one having the largest interquartile range was retained. This resulted in 15,348 annotated genes, whose expression was further batch corrected according to their chip identity (Johnson et al., 2007). Finally, gene expression time series were smoothed by a 5*th* order polynomial to take advantage of the high sampling rate and replicates at 0, 12, and 24 h. Microarray data have been deposited at Gene Expression Omnibus (GEO) under the accession number GSE74327.

2.6. Multi-Dimensional Scaling

To determine significantly regulated genes over time we performed a multi-dimensional scaling (MDS) using the HiT-MDS algorithm (Strickert et al., 2005). The algorithm projects the 15348×15348 distance matrix D of the pairwise Euclidean distances between all genes onto a two dimensional space, while preserving distances in D as best as possible. Genes varying strongly and uniquely over time will appear as outliers in the MDS point distribution. The uniqueness of a gene expression profile was quantified by fitting a two-dimensional skewed Gaussian distribution (Azzalini, 2015) to the MDS point density function.

2.7. Clustering Gene Expression Patterns

To cluster the gene timeseries, we applied the Cluster Affinity Search Technique (CAST), which considers the genes and their similarity over times as nodes and weighted edges of graph, respectively (Ben-Dor et al., 1999). All clusters are considered as unrelated entities and there is no pre-defined number of clusters. Instead a threshold parameter, here $t = 0.8$, determines the affinity between genes and this the final number of gene clusters. Inverse or anti-correlative behavior of genes after NGF or EGF stimulation was determined by fitting a linear model to the smoothed gene expression. Genes having a significant slope with opposite sign and an $r^2 > 0.7$ were taken as anti-correlated.

2.8. Enrichment Analysis of Transcription Factor Target Gene Sets

Upstream analysis for putative transcription factors regulating the EGF and NGF transcriptome responses over time were assessed by a Gene Set Enrichment analysis (Luo et al., 2009) using paired control to treatment samples for each timepoint with an overall cutoff q-value < 0.01. As gene sets we used the transcription factor target lists from the Molecular Signatures Database (MSigDB, version 5.0) (Subramanian et al., 2005), for which we mapped the human genes to the rat orthologs using BiomaRt (Huang et al., 2014).

2.9. Boolean Model

We used a Boolean model framework for dynamic analysis of PC12 cell differentiation. Based on our microarray data and literature knowledge we constructed a highly connected prior knowledge network (PKN) consisting of 63 nodes and 109 edges (cf. Supplementary Table 2). The R/Bioconductor package CellNetOptimizer (CNO) (Saez-Rodriguez et al., 2009) was used to optimize the PKN by reducing redundant nodes, unobservable states and edges. For this we rescaled the qRT-PCR fold change values between 0 and 1 and then transformed with a Hill function $f(x) = \frac{x^n}{x^n + k^n}$ as suggested in Saez-Rodriguez et al. (2009), where $n = 2$ and $k = 0.5$ denote the Hill coefficient and the threshold, above which a node is considered "on," respectively. Changing the Hill coefficient between $1 \leq n \leq 6$ did not change the results qualitatively. Model topology optimization was performed via the CellNORdt, which allows fitting with time course data. (See Supplementary Table 3 for stimulus, inhibition and time course data). We set the maximal CPU run time for the underlying genetic algorithm (GA) to 100 s and the relative tolerance to 0.01, using default parameters from the CNO otherwise. A representative evolution of the average and best residual error in a GA run is depicted in Supplementary Image 1A. The solutions quickly converge to a quasi steady state within the time window of simulation of 100 s. The following edges were fixed to prior to optimization based on literature knowledge: NGF → PI3K, NGF → RAS, NGF → PLC, AP1 → NPY, MEK/ERK & JNK → *Jund*, MEK/ERK & JNK → *Junb*, *Fosl1* & *Jund* → *AP1*, *Mmp10* → RAS, RAS → MEK, PLC → MEK. Model optimization was performed 100 times and edges were retained, if they appeared in 70% of the runs. This cutoff was chosen to generate a sparse network with robust edges. Performing more runs did not change the results qualitatively (cf. Supplementary Image 1B). Model simulations were performed using the R/Bioconductor package BoolNet (Müssel et al., 2010). The reference publications from which the interactions have been inferred as well as their Boolean transition functions are listed in Supplementary Table 4.

3. RESULTS

3.1. Gene Response of PC12 Cells Diverges for NGF and EGF on Long Time Scales

To elucidate the dynamic gene response of NGF and EGF, we measured the transcriptome dynamics using Illumina RatRef-12 Expression BeadChips. PC12 cells were either stimulated with NGF or EGF, and collected at the following timepoints: 1, 2, 3, 4, 5, 6, 8, 12, and 24 h. The unstimulated control samples (ctrl) were collected in parallel. Gene expression time series were smoothed by a 5*th* order polynomial to take advantage of the high sampling rate. Finally, we mapped array probes to their respective Entrez IDs, resulting in 15,348 annotated genes.

A bi plot of the principal component analysis (PCA) for the 1000 most varying genes depicted a clear separation of the control, NGF and EGF samples. The PCA scores, representing the NGF and EGF treated samples, showed a qualitatively similar behavior up to 4 h after stimulation, yet differed markedly beyond that time (**Figure 1A**, left). The absolute length and direction of the PCA loadings (**Figure 1A**, right) indicate the contribution of individual genes to the position of the scores. Correspondingly, several immediate early genes, such as *Junb* (Jun B Proto-Oncogene), *Fos* (FBJ Murine Osteosarcoma Viral Oncogene Homolog), *Ier2* (Immediate Early Response 2), and *Egr1* (Early Growth Response 1) contributed to the early gene response under both EGF and NGF stimulation, while members of the uPAR/Integrin signaling complex, such as *Mmp13/10/3* (Matrix Metallopeptidase 13/10/3), *Plat* (Plasminogen Activator, Tissue) and *Serpine1* (Serpin Peptidase Inhibitor, Clade E, Member 1) determined, among others, the separation of the NGF from the EGF trajectory. Loadings that point toward the control and late EGF response samples, like *Cdca7* (Cell Division Cycle Associated 7) and *G0s2* (G0/G1 Switch 2), are clearly related to cell cycle progression and additionally highlight the difference in proliferation vs. differentiation. In conclusion, the NGF gene response, and thus PC12 cell differentiation, must be determined by late transcriptional feedback events, that trigger and sustain MAPK/ERK signaling.

Next, we sought to functionally analyze the transcriptional differences in early and late gene regulation after EGF and NGF stimulation. For this we selected genes that are (i) strongly regulated (log_2 fold change of < -1.7 or > 1.7 in two consecutive timepoints) and (ii) have a unique temporal expression profile according to a multi-dimensional scaling (MDS) analysis (p-value < 0.01) (cf. Supplementary Image 2). We found 152 and 402 genes, meeting both criteria, in the EGF and NGF data, respectively, among which 126 genes are shared by both conditions. **Figure 1B** depicts a clustering of these differentially i.e., top-regulated genes. A cluster affinity search technique (Ben-Dor et al., 1999) identified five EGF (E1-E3b) and seven NGF (N1-N4B) gene response clusters (cf. Supplementary Table 5 and Supplementary Image 3). Interestingly, the EGF stimulus induced a short pulse-like response with rapid return to original gene expression levels, while the NGF stimulus induced a combination of short-impulse like (N1 - N2b) and long sustained gene expression patterns with several clusters (N3a-N4b) sustaining their expression over time (cf. circled insets in **Figure 1B**).

Figure 1C depicts a network representation of the enrichment analysis using a hypergeometric test on Gene Ontologies (GO). Enriched upregulated biological functions were identified in gene lists E1, E2a, N1, N2a, N3a, N4a and in both groups of inversely regulated genes (cf. Supplementary Table 6). Nodes correspond to GO terms, with numbers indicating the joint enrichment scores. Nodes sharing at least 20 percent of their genes are connected by solid or dotted edges, if the connected nodes lie within a stimulus or across NGF and EGF treatment. Early transcription factor activity is common to both, NGF and EGF signaling, (clusters E1 and N1) as well as MAPK signaling genes (clusters E2a and N3a). The latter, however, is more prominent and enriched at later points in time after NGF stimulation (N3a) compared to the EGF induced response (E2a). Here, a less and earlier enrichment of MAPK signaling genes was seen. Moreover, a second network of transcription factor activity could be identified after NGF stimulation (cluster N2a) that does not have any equivalent after EGF stimulation. It seems, that the initial response (first hour) is

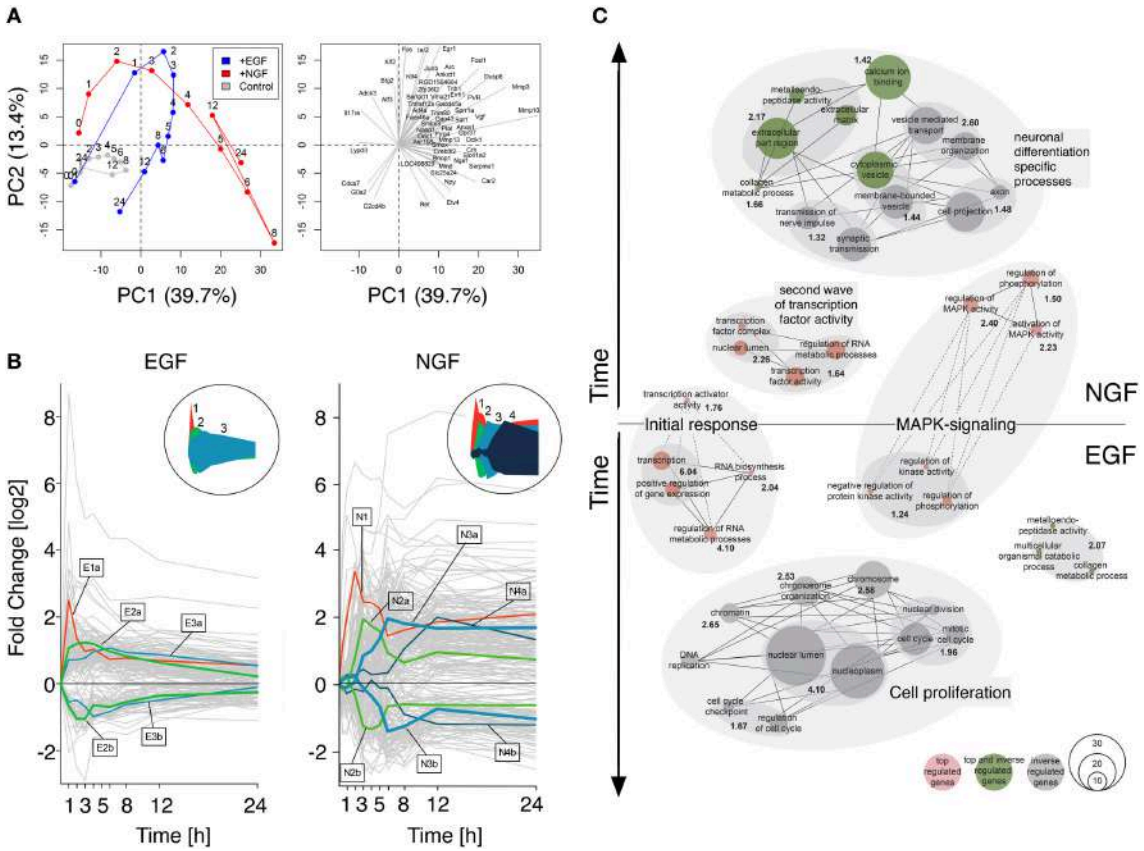


FIGURE 1 | Gene response dynamics after NGF or EGF stimulation. (A) Principal component analysis (PCA) of the PC12 cell transcriptomes after NGF (red), EGF (blue) and control treatment (gray). The PCA scores (left panel) and loadings (right panel) correspond to the samples and genes, respectively. Samples in the left panel have been connected to guide the eye. Clearly, EGF and NGF samples remain close in the first 3 h and separate at later timepoints, indicating a different cellular phenotype. Right panel: 50 largest loading vectors indicating the impact and time of action of individual genes. Immediate early genes, like *Fos* or *Ier2* point toward early timepoints, while loadings pointing toward the right, like *Vgf* or *Npy*, correspond to late timepoints and are most likely involved in differentiation. **(B)** Expression clusters of top regulated genes. The left and right panels depict the response of individual genes to EGF and NGF stimulation, respectively (gray lines). Cluster centroids are marked by lines with the cluster size encoded by line thickness. The circular inserts depict the cluster centroid envelopes for EGF and NGF, respectively. **(C)** Network representation of functional enrichment of NGF and EGF response genes. The network is comprised of GO-term clusters having a significant enrichment ($-log_{10}$(*p*-value) > 1.3) as shown in bold black numbers. Red, gray and green nodes contain in this order top-regulated genes, inversely-regulated genes between EGF and NGF or both. The vertical node location corresponds to the peak regulation of their genes, while node size is proportional to the number of genes in a functional category. Edges indicate a gene overlap of > 20% between nodes, being drawn as dashed lines, if they are shared between EGF and NGF.

controlled by a shared set of top-regulated genes (cf. **Figure 1C**, dashed lines). The cell-fate specific processes, however, seem to be orchestrated by different set of genes (cf. **Figure 1C**, separate networks). Many of the genes executing proliferation or differentiation specific processes fall into the category of inversely regulated genes and are not amongst the set of top-regulated genes identified earlier (cf. **Figure 1C**, green and gray nodes, cf. Material and Methods, cf. Supplementary Table 7). The genes involved in the procession of extracellular matrix and cytoplasmic vesicles, however, constitute an exception: these genes are both top and inverse-regulated (cf. **Figure 1C**, green nodes).

In summary, functional analysis of the gene clusters revealed an initiation of the differentiation and proliferation process by a shared set of differentially regulated genes. Specific functions, such as transmission of nerve impulse or DNA replication, however, seemed to be executed by two distinct gene groups that are when comparing the EGF to the NGF stimulus inversely regulated over time. Additionally, a second network of genes involved in transcription factor activity was identified in the NGF data set, which lacked a corresponding network in the EGF data set.

3.2. Simulation of a Boolean Network

Based on the above gene response analyses we sought to identify the mechanisms that sustain MAPK signaling activity after NGF stimulation. Our transcriptome timeseries analysis revealed that the decision process between proliferation and differentiation was spread out over several hours during which transcriptional feedback through an additional set of transcription factors was present after NGF stimulation, only (cf. **Figure 1C**). To further elucidate the transcription factors upstream of the gene

response after EGF or NGF stimulation we performed a gene set enrichment analysis (GSEA) (Luo et al., 2009) on the paired NGF to control and EGF to control transcriptome timeseries. As gene sets we used the motif gene sets from the Molecular Signatures Database (MSigDB v5.0) (Subramanian et al., 2005) and mapped the human genes onto the rat orthologs using BiomaRt (Huang et al., 2014).

Figure 2A compares the temporal significance of transcription factors for EGF and NGF stimulation. EGF elicited an early, yet transient significance of all transcription factors, while the time-resolved transcription factor significances for NGF showed early, transient and late activity. **Figure 2B** depicts the differences in TF significance between NGF and EGF. The most down-regulated TFs relative to EGF are E2F1, EBF1, SOX9 and SP1, all of which are linked to cell proliferation (Bastide et al., 2007; Hallstrom et al., 2008; György et al., 2012; Zhang et al., 2014).

Mullenbrock et al. (2011) showed late NGF-induced genes up to 4 h were preferentially regulated by AP1 and CREB (cAMP response element-binding protein). While AP1 was among the most persistently up-regulated transcription factors, we found a transient significance for CREB1, only, peaking at 3 and 6 h, under EGF or NGF stimulation, respectively, which indicated the importance of further TFs beyond that time window. In fact, we found the highest positive differences in the transcription factors BACH2, AP1, as well as ELF2 and ETV4. The latter two belong to the ETS transcription factor family. In particular ETV4, a member of the PEA3 subfamily of ETS, has been shown to promote neurite outgrowth (Fontanet et al., 2013; Kandemir et al., 2014). BACH2, member of the BTB-basic region leucine zipper transcription factor family, is known to down-regulate proliferation and is involved in neuronal differentiation of neoblastoma cells via p21 expression (Shim et al., 2006) and it interacts with the transcription factor MAFF (V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog F) (Kannan et al., 2012) that is necessary for differentiation.

To analyze the early cellular response upon treatment, we additionally compared the phosphorylation levels of pERK, pAKT and pJNK under NGF and EGF stimulation over time (**Figure 2C**). As expected, pERK increased after NGF and EGF stimulation, showing a persistent up-regulation for 8 h or pulse-like response, respectively. pJNK was continuously up-regulated under NGF relative to EGF stimulation, whereas pAKT responded similar to both stimuli, yet showed a consistently higher phosphorylation under EGF beyond 2 h. Taken together, this corroborates the roles of both pERK and pJNK as well as pAKT in PC12 cell differentiation and proliferation, respectively (Waetzig and Herdegen, 2003; Chen et al., 2012).

Based on the combined transcriptome, upstream transcription factor and protein analyses we next developed a comprehensive prior knowledge interaction network (PKN) for NGF induced PC12 cell differentiation. The PKN comprises key players of known pathways involved in PC12 cell differentiation, such as ERK/PLC/PI3K/JNK/P38/uPAR/NPY and integrin signaling, as well as "linker nodes" to obtain a minimal, yet fully connected network, consisting of 63 nodes and 109 reactions (cf. Supplementary Table 4 for reference publications). The network is depicted in Supplementary Image 4 with differentially regulated genes obtained from our timeseries marked in red and points of inhibition indicated by orange. A Cytoscape readable network format is provided in Supplementary Table 2. Albeit the included PKN pathways are much more complex, our focus was on simulating a biologically plausible signaling flow, including protein signaling, gene response and autocrine signaling as follows: stimulated TrkA receptor activates the downstream pathways PLC/PKC, MAPK/ERK, PI3K/AKT, and JNK/P38. Phosphorylated ERK, PI3K and P38/JNK together activate different transcription factors such as *Fosl1*, *Fos*, *Junb*, *Btg2*, *Klf2/5/6/10*, *Cited2*, *Maff*, and *Egr1*, which are important for PC12 cell differentiation according to our analysis and literature (Cao et al., 1990; Ito et al., 1990; Levkovitz and Baraban, 2002; Gil et al., 2004; Eriksson et al., 2007).

Junb and *Fos* initiate the AP1 system, which in turn induces uPA/uPAR signaling, triggering the formation of plasmin (Avraham and Yarden, 2011). The latter is a major factor for the induction of *Mmp3/Mmp10*, linking degradation of the extracellular matrix (ECM) with integrin signaling. The integrins transmit extracellular signaling back via the focal adhesion kinase (FAK) (Singh et al., 2012). FAK activates again the SHC protein, which closes the autocrine signaling. Previous studies reported that uPAR expression is necessary for NGF-induced PC12 cell differentiation (Farias-Eisner et al., 2000; Mullenbrock et al., 2011). A second autocrine signaling loop in the initial PKN putatively acts via the AP1 system, which in turn activates the Neuropeptide Y (NPY/NPYY1 pathway). NPY is a sympathetic co-transmitter that acts via G protein-coupled receptors through interactions with its NPYY1 receptors (Selbie and Hill, 1998; Pons et al., 2008). NPYY1 receptor further activates Ca^{2+} dependent PKC /PLCgamma and subsequently convergences to ERK signaling.

To optimize the highly connected PKN we used CellNetOptimizer (CNO) (Saez-Rodriguez et al., 2009). The CNO first compresses the network, i.e., it deletes unobservable nodes and then optimizes the network topology using a genetic algorithm. We trained the PKN using gene expression of selected differentially regulated genes under NGF stimulus and inhibition of either MEK, JNK, or PI3K (**Figure 3A**, MEKi, JNKi and PI3Ki). The overall gene response showed a gradual decline in fold change from NGF via MEK to JNK inhibition, while inhibition of PI3K only moderately impacted the gene expression (**Figure 3A**). The most affected genes under MEK and JNK inhibition were members of the uPAR signaling pathway, *Mmp10*, *Mmp3*, and *Plaur* as well as the transcription factors *Fosl1* and *Egr1*, *Plaur*, *Dusp6* (Dual Specificity Phosphatase 6) and lastly *Npy*.

Topology optimization using the above perturbations led to a greatly reduced network. Optimization lumped linear pathways into one node, such as the autocrine feedback via uPA/PLAT to *Itga1* and FAK or MEK to ERK transition. The reduced network revealed both MAPK/ERK and JNK as the central network hubs, distributing the upstream signals to downstream genes. It includes two positive feedback via AP1 and uPAR signaling back to FAK and MAPK as well as AP1 to Npy and PKC/PLC back to MAPK. To comply with prior knowledge, we re-expanded linear

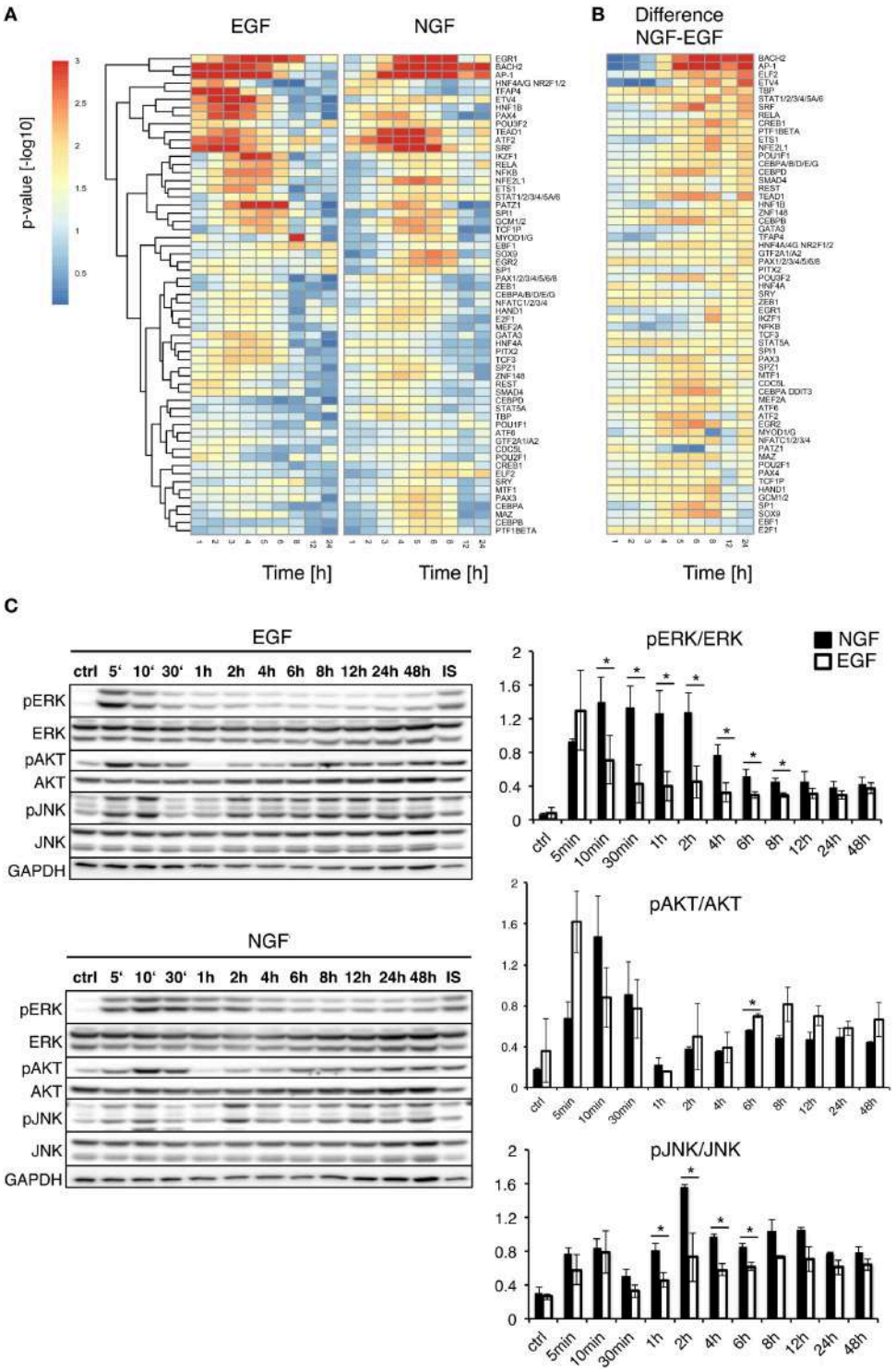


FIGURE 2 | Upstream analysis of gene expression timeseries. (A) Upstream Gene Set Enrichment Analysis for transcription factors. The heatmaps depict the significance of transcription factors putatively controlling the gene response after EGF (left) or NGF stimulation (right). All TFs are significantly regulated (FDR corrected p-value < 0.01) after NGF treatment. TFs have been clustered by their Euclidean distance across all conditions using a complete linkage method. **(B)** Difference in TF p-value significance (NGF-EGF). Rows were ordered from the most positive to the most negative difference at $t = 12$ and 24 h. **(C)** Time-resolved quantification of pERK, pAKT and pJNK after EGF and NGF treatment. Original western blots from PC12 cells treated with 75 ng/ml EGF and 50 ng/ml NGF over time. GAPDH is shown as loading control, IS: Internal Standard. Statistical analysis of the pERK/ERK, pAKT/AKT and pJNK/JNK levels are shown on the right panel. An increased and significant higher pERK/ERK level is shown in NGF stimulated (shown as black bars) cells compared to EGF (shown as white bars). A similar trend is visible for pJNK/JNK. A * denotes a p-value < 0.05, data points obtained in duplicates and triplicates.

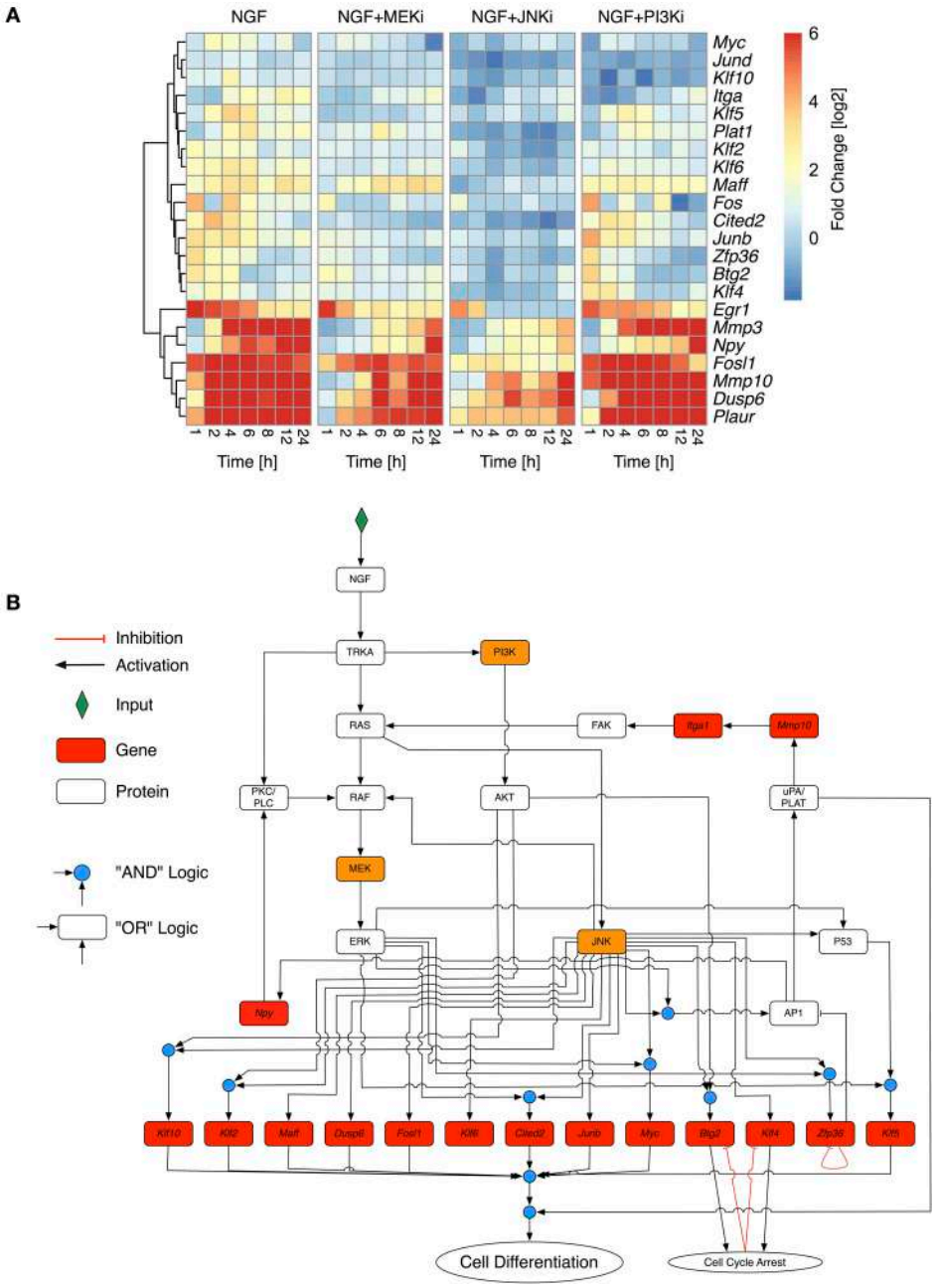


FIGURE 3 | Selective inhibition of NGF-induced PC12 differentiation. (A) Fold change values of selected response genes in PC12 cells after NGF stimulation under additional inhibition of MEK (NGF+MEKi), JNK (NGF+JNKi), or PI3K (NGF+PI3Ki). Fold change values have been calculated from biological triplicates relative to the unstimulated control per timepoint. To retain the contrast of less variable genes the maximal fold change has been restrained to +6. Genes have been clustered by their Euclidean distance across all conditions using a complete linkage method. **(B)** Optimized Boolean Network based on the training data in **(A)**. Nodes in red have been measured on the transcript level. Orange nodes indicate inhibited proteins.

pathways and added known down-stream target genes, such that the final network, shown in **Figure 3B**, comprised 32 nodes and 52 edges. We assumed that PC12 differentiation occurs, if the majority of these genes is activated together with uPAR signaling. Due to the inherent difficulty of Boolean networks to incorporate negative feedback loops, we revised the network topology of the reduced network to include transient gene activity of several moderately responding genes. *Klf4* and *Btg2* have been previously been indicated as immediate early genes in PC12 cell differentiation (Dijkmans et al., 2009) and are involved in growth arrest (Tirone, 2001; Yoon et al., 2003), which is a necessary prerequisite for differentiation and degradation of

mRNA, respectively. While the explicit mechanism of how *Klf4* and *Btg2* are regulated remains unclear, we assumed an auto-inhibition once they mediated their growth arrest effect. *Zfp36* belongs to the TTP (Tristetraprolin) family of proteins and has been shown to degrade AU-rich mRNAs, particularly of early response genes (Amit et al., 2007). It negatively regulates its own expression (Tiedje et al., 2012) and therefore in the model effectively delays the activity of AP1 before switching itself off. Of note, another member of the TTP protein family, Zfp36l2 (zinc finger protein 36, C3H type-like 2) is constitutively expressed at long times after NGF stimulation (data not shown) and might act as another long-term negative feedback regulator and causing downregulation of *Egr1*, *Fos*, and *Junb*. Indeed, our experimental data revealed a reduction on gene expression of *Egr1*, *Fos* and *Junb* over time (**Figure 3A**).

We simulated the optimized and re-expanded Boolean network (cf. Supplementary Table 8) using the BoolNet R/Bioconductor package (Müssel et al., 2010), performing two types of simulations. First, we tested the robustness and alternative attractors by setting NGF to "on" and randomly initializing all other network nodes. The nodes were then synchronously updated until a steady state was reached. Within $n = 10^7$ different simulations, the same final network state with "cell differentiation" set to "on" was always reached. Although this was not an exhaustive search given the number of possible initial network states, it still demonstrated the robustness of the network output. Next, to show the information flow from the NGF receptor to the downstream nodes under different inhibitory conditions, we initialized all nodes except NGF to "off" and performed synchronous updates until a steady state was reached (**Figure 4A**). Without inhibition, NGF sequentially switches on MAPK, AKT and JNK pathways as well as uPAR signaling. *Klf4*, *Btg2*, and *Zfp36* become transiently active, with the latter delaying AP1 activity. Blocking MEK (NGF+MEKi) inhibited ERK and thus several downstream targets, including the uPAR feedback. As the latter is assumed indispensable for PC12 cell differentiation, (Farias-Eisner et al., 2000, 2001), the model predicted inhibition of PC12 cell differentiation. The same phenotype is found, when blocking JNK (NGF+JNKi). In comparison to NGF+MEKi it even abrogated the activity of downstream targets altogether. Inhibition of PI3K (NGF+PI3Ki) solely affected PI3K and its downstream target protein AKT and target genes *Maff* and *Klf10*, yet cell differentiation persisted.

Taken together, we developed a core network from the downstream interactome of PC12 cell pathways involved in differentiation. The model captured the dynamic pathway activation after NGF stimulation and various inhibitions. It assigned central and synergistic roles for ERK and JNK in PC12 differentiation with JNK having the largest impact on the network activity.

3.3. Model Analysis and Experimental Confirmation

Network simulations were confirmed by live phase-contrast imaging (**Figure 4B**) and western blot analyses (**Figure 5**). We measured the convex hull (CH) to cell area (CA) ratio of PC12 cells on days 2, 4, and 6. A large convex hull due to extended neurite (marked as red arrow heads in **Figure 4B**) and small overall cell area is indicative of differentiation (**Figure 4B**, right panel). Clearly, the continuous CH/CA ratio at day 2 was largest for NGF stimulation and NGF stimulation with additional PI3K inhibition, which corresponded well with the cell differentiation set to "on" in the network simulations under these condition. One can speculate whether inhibition of the pro-proliferative PI3K pathway amplifies cell differentiation, possibly relieving a negative feedback. Indeed, a Western blot of the pERK/ERK ratio depicted a trend to higher ERK phosphorylation relative to NGF stimulation under PI3K inhibition (**Figure 5**) and phase-contrast images of PC12 cells show more and longer neurites in comparison to cells treated only with NGF or in combination to MEKi and JNKi (**Figure 4B**, NGF+PI3Ki). Interestingly, image analysis suggested not a stop, but rather a delay of cell differentiation under MEK inhibition. In detail, PC12 cells show no neurites under MEKi after 2 days of combined NGF treatment compared to NGF alone or NGF-PI3Ki. After 4 and 6 days of NGF+MEKi treatment, less cells have neurites in comparison to cells that were only treated with NGF (**Figure 4B**, NGF+MEKi). In line with literature, pERK levels were reduced, yet pJNK levels were likewise increased, indicating a redirection of protein activity under MEK inhibition (**Figure 5**, right panel). Likewise, the gene expression showed a reduced, but not completely abolished fold change for *Mmp10* (**Figure 3A**) and also an up-regulation of Dusp6. Although the discrete Boolean model could not simulate gradual responses, MEK inhibition still resulted in the activation of several downstream target genes necessary for PC12 cell differentiation, while none of these were active under JNK inhibition. In summary, modeling and simulation suggested that PC12 differentiation involved the activity of both JNK/JUN, MAPK/ERK and PI3K/AKT signaling pathways. The establishment of a positive, autocrine feedback loop was indispensable to active late and persistent gene expression.

4. DISCUSSION

PC12 cells are a well established model to study the cellular decisions toward proliferation or differentiation. Nevertheless, there is still a lack of understanding on how protein signaling and gene regulation interact on different time scales to decide on a long-term, sustained phenotype. Given the fact that PC12 cell cycle and differentiation last up to 4 and 6 days, respectively (Greene and Tischler, 1976; Luo et al., 1999; Adamski et al., 2007), late events occurring beyond the first hours are most likely to be important for sustaining the cellular decision. However, few studies that have compared the long-term effect of EGF and NGF in PC12 cells. They focused either on NGF alone (Dijkmans et al., 2008, 2009), on individual (Angelastro et al., 2000; Marek et al., 2004; Lee et al., 2005; Chung et al., 2010), or early time-points (Mullenbrock et al., 2011).

Previous studies have identified expression of immediate early genes (IEG), such as *Egr1*, *Junb*, and *Fos* together with delayed early genes (DEG), like *Dusp6*, *Mmp3/10*, *Fosl1*, and *Atf3* as necessary for PC12 cell differentiation (Vician et al.,

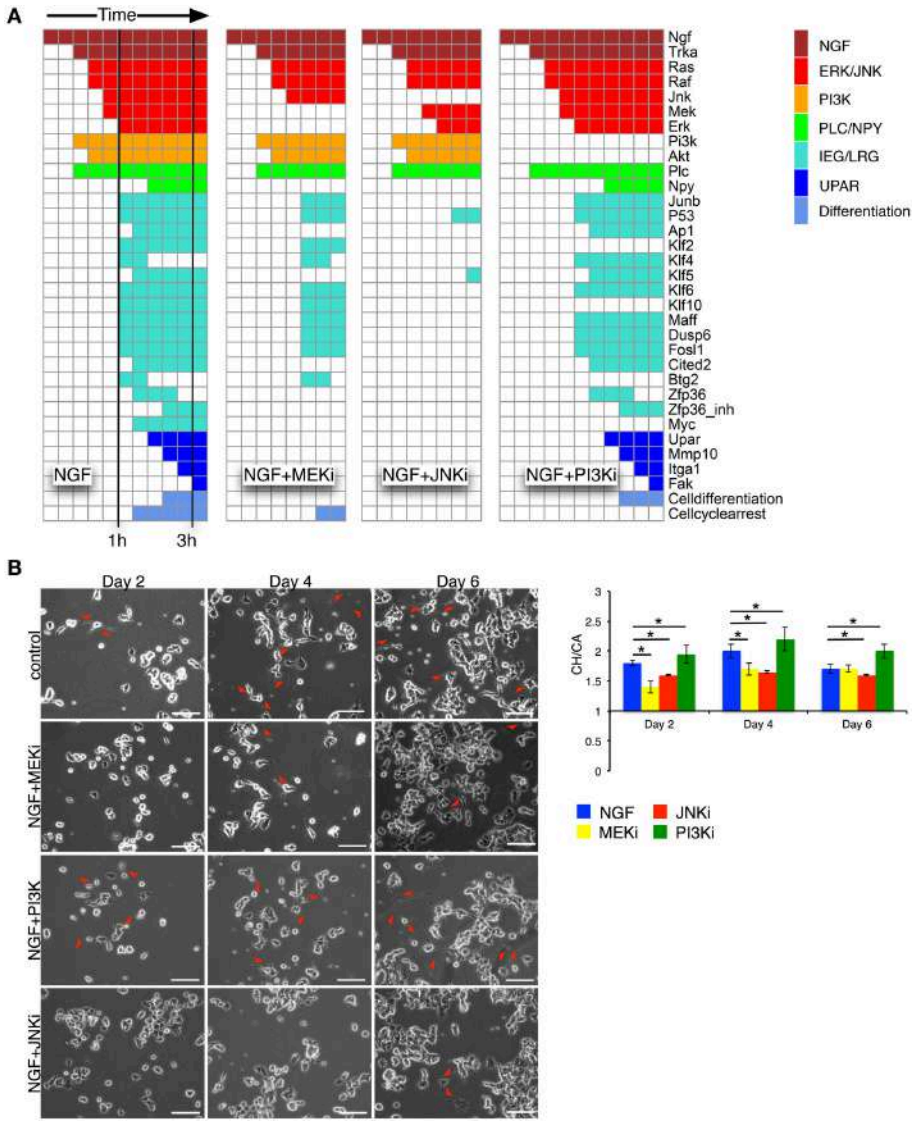


FIGURE 4 | Network simulation of time sequential pathway activation and experimental validation. (A) The heatmaps depict the path to attractor upon NGF stimulation. Columns correspond to synchronous update steps of the Boolean network. Time progresses from left to right until a steady state is reached. Initially all nodes, except NGF, are set to zero. Colored boxes correspond to activated nodes with the color denoting individual pathways/node categories. Cells are predicted to differentiate, if the node "Cell differentiation" is active, as in the case for NGF, or NGF+PI3Ki treatment. **(B)** Left: phase contrast images for days 2, 4, and 6 are shown for the 4 different conditions: NGF (control), NGF+MEKi, NGF+PI3Ki and NGF+JNKi. Red arrows depict sites of neurite outgrowth in differentiating PC12 cells. Bar: 100 μm. Right: statistical analysis of PC12 cell differentiation from phase contrast imaging for the different conditions are shown as convex hull (CH) to cell area (CA) ratio. Bars show Mean ± SEM, $n = 2$, (*t-test p-value < 0.05).

1997; Levkovitz et al., 2001; Dijkmans et al., 2008; Mullenbrock et al., 2011). However, we found all these genes strongly regulated by both EGF and NGF stimulation (Supplementary Table 5), however, showing differences in their expression kinetics (**Figure 1**). Akin to differences in the pERK dynamics, these results suggest that cellular decisions toward differentiation or proliferation are driven by the differences in the gene expression kinetics.

It has been suggested before that distinct cellular stimuli activate similar sets of response genes, whose expression dynamics, rather than their composition, determine cellular decisions (Murphy and Blenis, 2006; Amit et al., 2007; Yosef and Regev, 2011). Single expression bursts are likely to stimulate proliferation, while complex, wave-like expression patterns induce differentiation (Bar-Joseph et al., 2012). Accordingly, EGF elicited a pulse-like gene response, while NGF induced a complex, wave-like gene response (**Figure 1B**). After EGF stimulation the expression of IEGs, *Egr1*, *Fos*, and *Junb* was quickly attenuated through the rapid up-regulation of their negative regulators, namely *Fosl1*, *Atf3*, *Maff*, *Klf2*, and *Zfp36l2* and contributing to

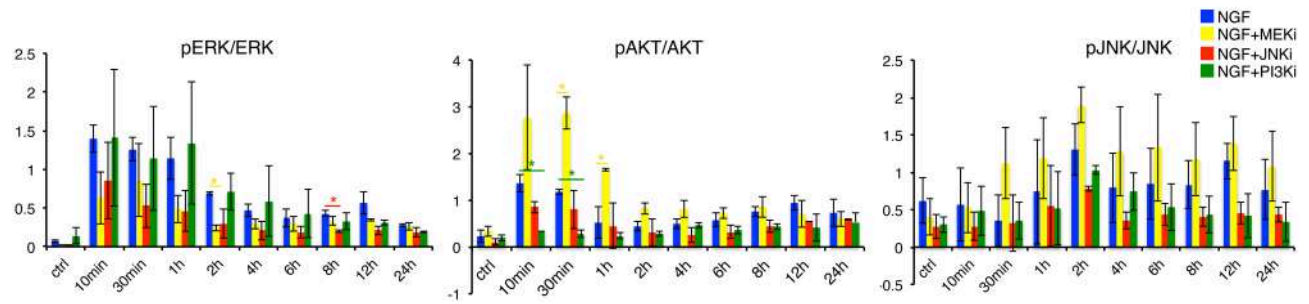


FIGURE 5 | Quantification of pERK, pAKT and pJNK levels under NGF and individual inhibitor treatments. Determination of pERK/ERK, pAKT/AKT and pJNK/JNK under NGF, NGF+MEKi, NGF+JNKi and NGF+PI3Ki treatment. Left panel: pERK/ERK levels decrease over time under NGF plus MEK and JNK inhibition. In contrast, PI3K inhibition shows a similar increase and sustained pERK/ERK levels over time compared to NGF treated PC12 cells alone. Interestingly, pAKT/AKT is increased under NGF+MEKi treatment, which is particularly significant in the early timepoints (30 min and 1 h) compared to NGF alone or the other two inhibitors. The latter two show decreased pAKT/AKT levels over time (middle panel). A * denotes a p-value < 0.05, data points obtained in duplicates and triplicates.

a pulse-like gene expression. Furthermore, *Fosl1* counteracts *Fos* and AP1 (Hoffmann et al., 2005) and *Atf3* has been shown to modulate *Egr1* activity (Giraldo et al., 2012), while *Maff* and *Klf2* negatively regulate serum response and STAT-responsive promoter elements (Amit et al., 2007). The same genes respond after NGF stimulation, however with a delayed response and might be one of the reasons for the stronger and longer gene and pERK response under NGF stimulation (Murphy et al., 2002, 2004; Murphy and Blenis, 2006; Saito et al., 2013).

A recent study by Mullenbrock et al. (2011) compared the transcriptome response of PC12 cells to EGF and NGF stimulation up to 4 h. Using chromatin immunoprecipitation they found a preferential regulation of late genes through AP1 and CREB TFs after NGF stimulation, which is in line with our findings (**Figure 2A**). However, we predicted a constitutive significance for AP1 up to 24 h, while CREB1 displayed a transient importance, being most abundant at 6 h after stimulation. Furthermore, we found a switch in the composition of transcriptional master regulators between 4 and 12 h. During this time, late TFs, such as BACH2, ETS1 and ELF2 become active.

Supplementary Image 5 depicts a Volcano plot of their target genes. Beyond the early gene targets, such as *Fosl1* or *Junb*, the late TFs additionally target related to cytoskeleton, morphogenesis and apoptosis, such as Tumor Necrosis Factor Receptor Superfamily, Member 12A (*Tnfrsf12a*), Doublecortin-Like Kinase 1 (*Dclk1*), Nerve Growth Factor Inducible *Vgf*, Coronin, Actin Binding Protein, 1A (*Coro1a*, Growth Arrest And DNA-Damage-Inducible, Alpha (*Gadd45a*) and *Npy*. Of note, we found *Rasa2* among the targets, which has recently been identified as a driver for differentiation through a negative feedback between PI3K and RAS (Chen et al., 2012).

A recent study by Aoki et al. (2013) investigated the downstream gene response upon light-induced intermittent and continuous ERK activation in normal rat kidney epithelial cells. Similar to the TF activity after EGF and NGF stimulation in PC12 cells, intermittent pERK activity caused up-regulation of *Fos*, *Egfr*, *Ier2*, and *Fgf21*, which were putatively controlled through serum response factor (SRF) and CREB binding sites, while sustained pERK activity caused gene regulation controlled by AP1 and BACH1. One can speculate that it is more the temporal dynamics of pERK and less the upstream ligands, such as EGF or NGF, that eventually encode the transcriptional program deciding on the cell fate.

To elucidate the various pathways and downstream target genes under NGF stimulation we constructed a Boolean model based on our transcriptome and additional literature data. A prior knowledge network revealed a highly interconnected pathway map transmitting NGF-induced signals. Training the network via inhibition of MEK, JNK or PI3K reduced the number of edges and nodes by about 80% and revealed the MAPK/JNK pathway as second signaling hub next to MAPK/ERK. Moreover, blocking the JNK pathway had a more drastic effect on cell differentiation than blocking MAPK/ERK via inhibition of MEK through UO126. Indeed, studies on the effect of MEK inhibition for PC12 cell differentiation are inconclusive. Early studies report how MEK inhibition completely averted PC12 cell differentiation (Pang et al., 1995; Klesse et al., 1999), while recent experiments suggest a decrease, rather than full inhibition of differentiation (Levkovitz et al., 2001; Chung et al., 2014). Our results were in line with the latter. Despite a significant reduction in pERK (**Figure 5**), our cell morphology measurements detected merely a decrease in the formation of neurites, rather than full inhibition of differentiation. The reason for this discrepancy could lie in the time scale of observation. MEK inhibition delayed differentiation and it took 6 days to eventually overcome this delay (**Figure 4B**). This confirmed the modeling results, which established JNK as key regulator that is closely interlinked with MAPK/ERK signaling. In concert with pERK, also pJNK becomes constitutively active upon NGF stimulation (**Figure 2C**). Moreover, blocking pERK through MEK even increased pJNK (and pAKT) levels, while pERK decreased after JNK inhibition, verifying a crosstalk between JNK and ERK pathways. Previous reports suggested such a crosstalk due to dual-phosphatase interaction (Fey et al., 2012), while other studies proposed that JNK phosphorylates RAF (Adler et al., 2005; Chen et al., 2012) and thereby contributing to MAPK/ERK activity. However, the mechanistic details governing

the crosstalk remain unclear so far. In conclusion, while previous studies assigned parallel, non-redundant roles to MAPK/ERK and MAPK/JNK (Waetzig and Herdegen, 2003), our results show that JNK signaling might be even the main driver for PC12 cell differentiation.

Next to the negative feedback loops through *Klf4*, *Zfp36*, and *Btg2*, arresting cell cycle and attenuating mRNA abundance, we included also two positive feedback loops via uPAR and integrin signaling as well as through Neuropeptide Y and PKC/PLC signaling. Positive feedback loops are a common regulatory pattern in molecular biology to induce bistability switch-like behavior, particularly in cell fate decisions and differentiation (Xiong and Ferrell, 2003; Mitrophanov and Groisman, 2008; Kueh et al., 2013). In fact, multiple feedbacks deciding between PC12 cell differentiation and proliferation, have been studied on the level of MAPK signaling (Santos et al., 2007; von Kriegsheim et al., 2009). Recently, Ryu et al. (2015) used a FRET construct to quantify pERK dynamics on a single cell level after growth factor stimulation. While the cell population average still resembled the hitherto described transient and sustained pERK activity after respective EGF and NGF stimulation, the authors found a highly heterogenous response on the single cell level. Pulsed stimulation, however, not only synchronized MAPK activity between cells, but also triggered PC12 differentiation upon EGF stimulation, if the integrated pERK signal was large enough. The authors concluded that thus not only MAPK signaling, but also further pathways are responsible for the cell fate decision. Sparta et al. (2015) used a similar experimental approach to single cell response of human MCF10A-5e cells to show that EGFR activity induced a frequency modulation response, while TrkA activity caused amplitude modulation of pERK levels. The authors explained these finding by additional receptor-dependent signaling networks beyond the core Ras-Raf-MEK-ERK pathway. Extending on this idea, our data and model suggest autocrine signaling as further feedbacks that sustain the expression of differentiation inducing TFs. Indeed, uPAR and also *Npy* activity were strongly correlated with differentiation (**Figure 3A**) and neither Npy nor uPAR signaling were activated upon EGF stimulation (data not shown). In line with this finding previous studies reported that uPAR expression is necessary for NGF-induced PC12 cell differentiation (Farias-Eisner et al., 2000; Mullenbrock et al., 2011). SERPINE1 regulating the plasminogen activator-plasmin proteolysis was shown to promote neurite outgrowth and phosphorylation of the TrkA receptor and ERK (Soeda et al., 2006, 2008). In our model we included the necessity of uPAR signaling though the activation of late genes, such as *Klf5*, yet the causal relationship between uPAR signaling and late gene expression remains unclear. However, uPAR signaling could constitute the additional positive feedbacks beyond MAPK signaling that were predicted by Ryu et al. (2015), which would be interesting to test on the single cell level. Reporters for uPAR and/or JNK activity should likewise show a heterogenous activity and correlate with the per-cell differentiation status, which could potentially be modeled within a stochastic differential equation framework.

In conclusion, our approach has identified the short and long-term transcriptional activity in PC12 cells after NGF and EGF stimulation. Modeling the pathway orchestration using a Boolean model we identified feedback regulations beyond MAPK signaling that attenuate and sustain the cellular decision toward differentiation. Extending on previous studies we established JNK as a key player in PC12 cell differentiation that might have equal, if not even more importance than ERK during this process. Over time AP1 was accompanied by a variety of transcription factors serving signal attenuation, signal maintenance and morphological change of the cell, which demonstrates that the decision toward differentiation is a time sequential process over at least 12 h.

AUTHOR CONTRIBUTIONS

MF, SKn, MK, SK, and BO performed the experiments. MF, SKn, AS, and BO performed the data analysis. HB And MB conceived the project, performed the data analysis and wrote the manuscript with BO. All authors approved the final manuscript.

ACKNOWLEDGMENTS

This work was was supported by the Deutsche Forschungsgemeinschaft grant InKoMBio: SPP 1395. The authors greatly acknowledge the Genomics and Proteomics Core Facility, German Cancer Research Center/DKFZ, Heidelberg, Germany for their microarray service.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00044

REFERENCES

Adamski, D., Mayol, J.-F., Platet, N., Berger, F., Hérodin, F., and Wion, D. (2007). Effects of Hoechst 33342 on C2c12 and PC12 cell differentiation. *FEBS Lett.* 581, 3076–3080. doi: 10.1016/j.febslet.2007.05.073

Adler, V., Qu, Y., Smith, S. J., Izotova, L., Pestka, S., Kung, H. F., et al. (2005). Functional interactions of Raf and MEK with Jun-N-terminal kinase (JNK) result in a positive feedback loop on the oncogenic Ras signaling pathway. *Biochemistry* 44, 10784–10795. doi: 10.1021/bi050619j

Amit, I., Citri, A., Shay, T., Lu, Y., Katz, M., Zhang, F., et al. (2007). A module of negative feedback regulators defines growth factor signaling. *Nat. Genet.* 39, 503–512. doi: 10.1038/ng1987

Angelastro, J. M., Klimaschewski, L., Tang, S., Vitolo, O. V., Weissman, T. A., Donlin, L. T., et al. (2000). Identification of diverse nerve growth factor-regulated genes by serial analysis of gene expression (SAGE) profiling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10424–10429. doi: 10.1073/pnas.97.19.10424

Aoki, K., Kumagai, Y., Sakurai, A., Komatsu, N., Fujita, Y., Shionyu, C., et al. (2013). Stochastic ERK activation induced by noise and cell-to-cell propagation regulates cell density-dependent proliferation. *Mol. Cell* 52, 529–540. doi: 10.1016/j.molcel.2013.09.015

Avraham, R., and Yarden, Y. (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.* 12, 104–117. doi: 10.1038/nrm3048

Azzalini, A. (2015). *The R package sn: The skew-normal and skew-t distributions (version 1.2-4)*. Universitate di Padova.

Bar-Joseph, Z., Gitter, A., and 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

Bastide, P., Darido, C., Pannequin, J., Kist, R., Robine, S., Marty-Double, C., et al. (2007). Sox9 regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. *J. Cell Biol.* 178, 635–648. doi: 10.1083/jcb.200704152

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* 6, 281–297. doi: 10.1089/106652799318274

Burstein, D. E., Blumberg, P. M., and Greene, L. A. (1982). Nerve growth factor-induced neuronal differentiation of PC12 pheochromocytoma cells: lack of inhibition by a tumor promoter. *Brain Res.* 247, 115–119. doi: 10.1016/0006-8993(82)91033-2

Cao, X. M., Koski, R. A., Gashler, A., McKiernan, M., Morris, C. F., Gaffney, R., et al. (1990). Identification and characterization of the Egr-1 gene product, a DNA-binding zinc finger protein induced by differentiation and growth signals. *Mol. Cell. Biol.* 10, 1931–1939. doi: 10.1128/MCB.10.5.1931

Chao, M. V. (1992). Neurotrophin receptors: a window into neuronal differentiation. *Neuron* 9, 583–593. doi: 10.1016/0896-6273(92)90023-7

Chen, J.-Y., Lin, J.-R., Cimprich, K. A., and Meyer, T. (2012). A two-dimensional ERK-AKT signaling code for an NGF-triggered cell-fate decision. *Mol. Cell* 45, 196–209. doi: 10.1016/j.molcel.2011.11.023

Chung, J., Kubota, H., Ozaki, Y.-I., Uda, S., and Kuroda, S. (2010). Timing-dependent actions of NGF required for cell differentiation. *PLoS ONE* 5:e9011. doi: 10.1371/journal.pone.0009011

Chung, J., Miura, N., Ito, A., Sawada, M., Nishikawa, S., Kuroda, K., et al. (2014). Single-cell heterogeneity in suppression of PC12 differentiation by direct microinjection of a differentiation inhibitor, U0126. *Cell Biol. Int.* 38, 1215–1220. doi: 10.1002/cbin.10296

Cowley, S., Paterson, H., Kemp, P., and Marshall, C. J. (1994). Activation of MAP kinase kinase is necessary and sufficient for PC12 differentiation and for transformation of NIH 3T3 cells. *Cell* 77, 841–852. doi: 10.1016/0092-8674(94)90133-3

Dijkmans, T. F., van Hooijdonk, L. W. A., Schouten, T. G., Kamphorst, J. T., Fitzsimons, C. P., and Vreugdenhil, E. (2009). Identification of new Nerve Growth Factor-responsive immediate-early genes. *Brain Res.* 1249, 19–33. doi: 10.1016/j.brainres.2008.10.050

Dijkmans, T. F., van Hooijdonk, L. W. A., Schouten, T. G., Kamphorst, J. T., Vellinga, A. C. A., Meerman, J. H. N., et al. (2008). Temporal and functional dynamics of the transcriptome during nerve growth factor-induced differentiation. *J. Neurochem.* 105, 2388–2403. doi: 10.1111/j.1471-4159.2008.05338.x

Dikic, I., Schlessinger, J., and Lax, I. (1994). PC12 cells overexpressing the insulin receptor undergo insulin-dependent neuronal differentiation. *Curr. Biol.* 4, 702–708. doi: 10.1016/S0960-9822(00)00155-X

Eriksson, M., Taskinen, M., and Leppä, S. (2007). Mitogen activated protein kinase-dependent activation of c-Jun and c-Fos is required for neuronal differentiation but not for growth and stress response in PC12 cells. *J. Cell. Physiol.* 210, 538–548. doi: 10.1002/jcp.20907

Farias-Eisner, R., Vician, L., Reddy, S., Basconcillo, R., Rabbani, S. A., Wu, Y. Y., et al. (2001). Expression of the urokinase plasminogen activator receptor is transiently required during "priming" of PC12 cells in nerve growth factor-directed cellular differentiation. *J. Neurosci. Res.* 63, 341–346.

Farias-Eisner, R., Vician, L., Silver, A., Reddy, S., Rabbani, S. A., and Herschman, H. R. (2000). The urokinase plasminogen activator receptor (UPAR) is preferentially induced by nerve growth factor in PC12 pheochromocytoma cells and is required for NGF-driven differentiation. *J. Neurosci.* 20, 230–239.

Fey, D., Croucher, D. R., Kolch, W., and Kholodenko, B. N. (2012). Crosstalk and signaling switches in mitogen-activated protein kinase cascades. *Front. Physiol.* 3:355. doi: 10.3389/fphys.2012.00355

Fiore, M., Chaldakov, G. N., and Aloe, L. (2009). Nerve growth factor as a signaling molecule for nerve cells and also for the neuroendocrine-immune systems. *Rev. Neurosci.* 20, 133–145. doi: 10.1515/REVNEURO.2009.20.2.133

Fontanet, P., Irala, D., Alsina, F. C., Paratcha, G., and Ledda, F. (2013). Pea3 transcription factor family members Etv4 and Etv5 mediate retrograde signaling and axonal growth of DRG sensory neurons in response to NGF. *J. Neurosci.* 33, 15940–15951. doi: 10.1523/JNEUROSCI.0928-13.2013

Gil, G. A., Bussolino, D. F., Portal, M. M., Alfonso Pecchio, A., Renner, M. L., Borioli, G. A., et al. (2004). c-Fos activated phospholipid synthesis is required for neurite elongation in differentiating PC12 cells. *Mol. Biol. Cell* 15, 1881–1894. doi: 10.1091/mbc.E03-09-0705

Giraldo, A., Barrett, O. P. T., Tindall, M. J., Fuller, S. J., Amirak, E., Bhattacharya, B. S., et al. (2012). Feedback regulation by Atf3 in the endothelin-1-responsive transcriptome of cardiomyocytes: Egr1 is a principal Atf3 target. *Biochem. J.* 444, 343–355. doi: 10.1042/BJ20120125

Gotoh, Y., Nishida, E., Yamashita, T., Hoshi, M., Kawakami, M., and Sakai, H. (1990). Microtubule-associated-protein (MAP) kinase activated by nerve growth factor and epidermal growth factor in PC12 cells. Identity with the mitogen-activated MAP kinase of fibroblastic cells. *Eur. J. Biochem.* 193, 661–669. doi: 10.1111/j.1432-1033.1990.tb19384.x

Greene, L. A., and Tischler, A. S. (1976). Establishment of a noradrenergic clonal line of rat adrenal pheochromocytoma cells which respond to nerve growth factor. *Proc. Natl. Acad. Sci. U.S.A* 73, 2424–2428. doi: 10.1073/pnas.73.7.2424

György, I., Boller, S., Nechanitzky, R., Mandel, E., Pott, S., Liu, E., et al. (2012). Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells. *Genes Dev.* 26, 668–682. doi: 10.1101/gad.187328.112

Hallstrom, T. C., Mori, S., and Nevins, J. R. (2008). An E2f1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer Cell* 13, 11–22. doi: 10.1016/j.ccr.2007.11.031

Hoffmann, E., Thiefes, A., Buhrow, D., Dittrich-Breiholz, O., Schneider, H., Resch, K., et al. (2005). MEK1-dependent delayed expression of Fos-related antigen-1 counteracts c-Fos and p65 NF-kappaB-mediated interleukin-8 transcription in response to cytokines or growth factors. *J. Biol. Chem.* 280, 9706–9718. doi: 10.1074/jbc.M407071200

Huang, L., Feng, G., Du, P., Xia, T., Wang, X., Jing, W., et al. (2014). *GeneAnswers: Integrated Interpretation of Genes*. R package version 2.10.0.

Ito, E., Sweterlitsch, L. A., Tran, P. B., Rauscher, F. J. III, and Narayanan, R. (1990). Inhibition of PC-12 cell differentiation by the immediate early gene fra-1. *Oncogene* 5, 1755–1760.

Jing, S., Tapley, P., and Barbacid, M. (1992). Nerve growth factor mediates signal transduction through trk homodimer receptors. *Neuron* 9, 1067–1079. doi: 10.1016/0896-6273(92)90066-M

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Kandemir, B., Caglayan, B., Hausott, B., Erdogan, B., Dag, U., Demir, O., et al. (2014). Pea3 transcription factor promotes neurite outgrowth. *Front. Mol. Neurosci.* 7:59. doi: 10.3389/fnmol.2014.00059

Kannan, M. B., Solovieva, V., and Blank, V. (2012). The small MAF transcription factors MAFF, MAFG and MAFK: current knowledge and perspectives. *Biochim. Biophys. Acta* 1823, 1841–1846. doi: 10.1016/j.bbamcr.2012.06.012

Kaplan, D. R., Martin-Zanca, D., and Parada, L. F. (1991). Tyrosine phosphorylation and tyrosine kinase activity of the trk proto-oncogene product induced by NGF. *Nature* 350, 158–160. doi: 10.1038/350158a0

Klesse, L. J., Meyers, K. A., Marshall, C. J., and Parada, L. F. (1999). Nerve growth factor induces survival and differentiation through two distinct signaling cascades in PC12 cells. *Oncogene* 18, 2055–2068. doi: 10.1038/sj.onc.1202524

Kueh, H. Y., Champhekar, A., Champhekhar, A., Nutt, S. L., Elowitz, M. B., and Rothenberg, E. V. (2013). Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. *Science* 341, 670–673. doi: 10.1126/science.1240831

Lee, K.-H., Ryu, C. J., Hong, H. J., Kim, J., and Lee, E. H. (2005). CDNA microarray analysis of nerve growth factor-regulated gene expression profile in rat PC12 cells. *Neurochem. Res.* 30, 533–540. doi: 10.1007/s11064-005-2688-y

Leppä, S., Saffrich, R., Ansorge, W., and Bohmann, D. (1998). Differential regulation of c-Jun by ERK and JNK during PC12 cell differentiation. *EMBO J.* 17, 4404–4413. doi: 10.1093/emboj/17.15.4404

Levi-Montalcini, R. (1987). The nerve growth factor 35 years later. *Science* 237, 1154–1162. doi: 10.1126/science.3306916

Levkovitz, Y., and Baraban, J. M. (2002). A dominant negative Egr inhibitor blocks nerve growth factor-induced neurite outgrowth by suppressing c-Jun activation: role of an Egr/c-Jun complex. *J. Neurosci.* 22, 3845–3854.

Levkovitz, Y., O'Donovan, K. J., and Baraban, J. M. (2001). Blockade of NGF-induced neurite outgrowth by a dominant-negative inhibitor of the egr family of transcription regulatory factors. *J. Neurosci.* 21, 45–52.

Luo, J., West, J. R., Cook, R. T., and Pantazis, N. J. (1999). Ethanol induces cell death and cell cycle delay in cultures of pheochromocytoma PC12 cells. *Alcohol. Clin. Exp. Res.* 23, 644–656. doi: 10.1111/j.1530-0277.1999.tb04166.x

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10:161. doi: 10.1186/1471-2105-10-161

Marek, L., Levresse, V., Amura, C., Zentrich, E., Van Putten, V., Nemenoff, R. A., et al. (2004). Multiple signaling conduits regulate global differentiation-specific gene expression in PC12 cells. *J. Cell. Physiol.* 201, 459–469. doi: 10.1002/jcp.20087

Marshall, C. J. (1995). Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* 80, 179–185. doi: 10.1016/0092-8674(95)90401-8

Mitrophanov, A. Y., and Groisman, E. A. (2008). Positive feedback in cellular control systems. *Bioessays* 30, 542–555. doi: 10.1002/bies.20769

Mullenbrock, S., Shah, J., and Cooper, G. M. (2011). Global expression analysis identified a preferentially nerve growth factor-induced transcriptional program regulated by sustained mitogen-activated protein kinase/extracellular signal-regulated kinase (ERK) and AP-1 protein activation during PC12 cell differentiation. *J. Biol. Chem.* 286, 45131–45145. doi: 10.1074/jbc.M111.274076

Murphy, L. O., and Blenis, J. (2006). MAPK signal specificity: the right place at the right time. *Trends Biochem. Sci.* 31, 268–275. doi: 10.1016/j.tibs.2006.03.009

Murphy, L. O., MacKeigan, J. P., and Blenis, J. (2004). A network of immediate early gene products propagates subtle differences in mitogen-activated protein kinase signal amplitude and duration. *Mol. Cell. Biol.* 24, 144–153. doi: 10.1128/MCB.24.1.144-153.2004

Murphy, L. O., Smith, S., Chen, R.-H., Fingar, D. C., and Blenis, J. (2002). Molecular interpretation of ERK signal duration by immediate early gene products. *Nat. Cell Biol.* 4, 556–564. doi: 10.1038/ncb822

Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010). BoolNet–an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* 26, 1378–1380. doi: 10.1093/bioinformatics/btq124

Pang, L., Sawada, T., Decker, S. J., and Saltiel, A. R. (1995). Inhibition of MAP kinase kinase blocks the differentiation of PC-12 cells induced by nerve growth factor. *J. Biol. Chem.* 270, 13585–13588. doi: 10.1074/jbc.270.23.13585

Pons, J., Kitlinska, J., Jacques, D., Perreault, C., Nader, M., Everhart, L., et al. (2008). Interactions of multiple signaling pathways in neuropeptide Y-mediated bimodal vascular smooth muscle cell growth. *Can. J. Physiol. Pharmacol.* 86, 438–448. doi: 10.1139/Y08-054

Qui, M. S., and Green, S. H. (1992). PC12 cell neuronal differentiation is associated with prolonged p21ras activity and consequent prolonged ERK activity. *Neuron* 9, 705–717. doi: 10.1016/0896-6273(92)90033-A

Ritchie, M. E., Dunning, M. J., Smith, M. L., Shi, W., and Lynch, A. G. (2011). BeadArray expression analysis using bioconductor. *PLoS Comput. Biol.* 7:e1002276. doi: 10.1371/journal.pcbi.1002276

Ryu, H., Chung, M., Dobrzynski, M., Fey, D., Blum, Y., Lee, S. S., et al. (2015). Frequency modulation of ERK activation dynamics rewires cell fate. *Mol. Syst. Biol.* 11, 838–838. doi: 10.15252/msb.20156458

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., et al. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.* 5:331. doi: 10.1038/msb.2009.87

Saito, T. H., Uda, S., Tsuchiya, T., Ozaki, Y.-I., and Kuroda, S. (2013). Temporal decoding of MAP kinase and CREB phosphorylation by selective immediate early gene expression. *PLoS ONE* 8:e57037. doi: 10.1371/journal.pone.0057037

Santos, S. D., Verveer, P. J., and Bastiaens, P. I. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat. Cell Biol.* 9, 324–330. doi: 10.1038/ncb1543

Sasagawa, S., Ozaki, Y., Fujita, K., and Kuroda, S. (2005). Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nat. Cell Biol.* 7, 365–373. doi: 10.1038/ncb1233

Selbie, L. A., and Hill, S. J. (1998). G protein-coupled-receptor cross-talk: the fine-tuning of multiple receptor-signalling pathways. *Trends Pharmacol. Sci.* 19, 87–93. doi: 10.1016/S0165-6147(97)01166-8

Shim, K. S., Rosner, M., Freilinger, A., Lubec, G., and Hengstschläger, M. (2006). Bach2 is involved in neuronal differentiation of N1E-115 neuroblastoma cells. *Exp. Cell Res.* 312, 2264–2278. doi: 10.1016/j.yexcr.2006.03.018

Singh, A., Nascimento, J. M., Kowar, S., Busch, H., and Boerries, M. (2012). Boolean approach to signalling pathway modelling in HGF-induced keratinocyte migration. *Bioinformatics* 28, i495–i501. doi: 10.1093/bioinformatics/bts410

Soeda, S., Koyanagi, S., Kuramoto, Y., Kimura, M., Oda, M., Kozako, T., et al. (2008). Anti-apoptotic roles of plasminogen activator inhibitor-1 as a neurotrophic factor in the central nervous system. *Thromb. Haemost.* 100, 1014–1020. doi: 10.1160/th08-04-0259

Soeda, S., Shinomiya, K., Ochiai, T., Koyanagi, S., Toda, A., Eyanagi, R., et al. (2006). Plasminogen activator inhibitor-1 aids nerve growth factor-induced differentiation and survival of pheochromocytoma cells by activating both the extracellular signal-regulated kinase and c-Jun pathways. *Neuroscience* 141, 101–108. doi: 10.1016/j.neuroscience.2006.03.026

Sparta, B., Pargett, M., Minguet, M., Distor, K., Bell, G., and Albeck, J. G. (2015). Receptor level mechanisms are required for epidermal growth factor (EGF)-stimulated extracellular signal-regulated kinase (ERK) activity pulses. *J. Biol. Chem.* 290, 24784–24792. doi: 10.1074/jbc.M115.662247

Strickert, M., Teichmann, S., Sreenivasulu, N., and Seiffert, U. (2005). "High-throughput multi-dimensional scaling (HiT-MDS) for cDNA-array expression data," in *Artificial Neural Networks: Biological Inspirations ICANN 2005, number 3696 in Lecture Notes in Computer Science*, eds W. Duch, J. Kacprzyk, E. Oja, and S. A. Zadrrony (Berlin; Heidelberg: Springer), 625–633.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tanabe, K., Bonilla, I., Winkles, J. A., and Strittmatter, S. M. (2003). Fibroblast growth factor-inducible-14 is induced in axotomized neurons and promotes neurite outgrowth. *J. Neurosci.* 23, 9675–9686.

Tiedje, C., Ronkina, N., Tehrani, M., Dhamija, S., Laass, K., Holtmann, H., et al. (2012). The p38/MK2-driven exchange between tristetraprolin and HuR regulates AURich elementdependent translation. *PLoS Genet.* 8:e1002977. doi: 10.1371/journal.pgen.1002977

Tirone, F. (2001). The gene PC3(TIS21/BTG2), prototype member of the PC3/BTG/TOB family: regulator in control of cell growth, differentiation, and DNA repair? *J. Cell. Physiol.* 187, 155–165. doi: 10.1002/jcp.1062

Vaudry, D., Stork, P. J., Lazarovici, P., and Eiden, L. E. (2002). Signaling pathways for PC12 cell differentiation: making the right connections. *Science* 296, 1648–1649. doi: 10.1126/science.1071552

Vician, L., Basconcillo, R., and Herschman, H. R. (1997). Identification of genes preferentially induced by nerve growth factor versus epidermal growth factor in PC12 pheochromocytoma cells by means of representational difference analysis. *J. Neurosci. Res.* 50, 32–43.

von Kriegsheim, A., Baiocchi, D., Birtwistle, M., Sumpton, D., Bienvenut, W., Morrice, N., et al. (2009). Cell fate decisions are specified by the dynamic ERK interactome. *Nat. Cell Biol.* 11, 1458–1464. doi: 10.1038/ncb1994

Waetzig, V., and Herdegen, T. (2003). The concerted signaling of ERK1/2 and JNKs is essential for PC12 cell neuritogenesis and converges at the level of target proteins. *Mol. Cell. Neurosci.* 24, 238–249. doi: 10.1016/S1044-7431(03)00126-X

Weber, S., Fernandez-Cachon, M. L., Nascimento, J. M., Knauer, S., Offermann, B., Murphy, R. F., et al. (2013). Label-free detection of neuronal differentiation in cell populations using high-throughput live-cell imaging of PC12 cells. *PLoS ONE* 8:e56690. doi: 10.1371/journal.pone.0056690

Wu, Y. Y., and Bradshaw, R. A. (1996). Synergistic induction of neurite outgrowth by nerve growth factor or epidermal growth factor and interleukin-6 in PC12 cells. *J. Biol. Chem.* 271, 13033–13039. doi: 10.1074/jbc.271.22.13033

Xing, J., Kornhauser, J. M., Xia, Z., Thiele, E. A., and Greenberg, M. E. (1998). Nerve growth factor activates extracellular signal-regulated kinase and

p38 mitogen-activated protein kinase pathways to stimulate CREB serine 133 phosphorylation. *Mol. Cell. Biol.* 18, 1946–1955. doi: 10.1128/MCB.18.4.1946

Xiong, W., and Ferrell, J. E. (2003). A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* 426, 460–465. doi: 10.1038/nature02089

Yoon, H. S., Chen, X., and Yang, V. W. (2003). Krüppel-like factor 4 mediates p53-dependent G1/S cell cycle arrest in response to DNA damage. *J. Biol. Chem.* 278, 2101–2105. doi: 10.1074/jbc.M211027200

Yosef, N., and Regev, A. (2011). Impulse control: temporal dynamics in gene transcription. *Cell* 144, 886–896. doi: 10.1016/j.cell.2011.02.015

Zhang, J.-P., Zhang, H., Wang, H.-B., Li, Y.-X., Liu, G.-H., Xing, S., et al. (2014). Down-regulation of Sp1 suppresses cell proliferation, clonogenicity and the expressions of stem cell markers in nasopharyngeal carcinoma. *J. Trans. Med.* 12, 222. doi: 10.1186/s12967-014-0222-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 19 February 2016
doi: 10.3389/fgene.2016.00018

ROMA: Representation and Quantification of Module Activity from Target Expression Data

Loredana Martignetti[1, 2, 3, 4], *Laurence Calzone*[1, 2, 3, 4], *Eric Bonnet*[1, 2, 3, 4], *Emmanuel Barillot*[1, 2, 3, 4] *and Andrei Zinovyev*[1, 2, 3, 4]*

[1] Computational and Systems Biology of Cancer, Institut Curie, Paris, France, [2] PSL Research University, Paris, France, [3] Institut National de la Santé et de la Recherche Médicale U900, Paris, France, [4] Mines ParisTech, Paris, France

In many analyses of high-throughput data in systems biology, there is a need to quantify the activity of a set of genes in individual samples. A typical example is the case where it is necessary to estimate the activity of a transcription factor (which is often not directly measurable) from the expression of its target genes. We present here ROMA (Representation and quantification Of Module Activities) Java software, designed for fast and robust computation of the activity of gene sets (or modules) with coordinated expression. ROMA activity quantification is based on the simplest uni-factor linear model of gene regulation that approximates the expression data of a gene set by its first principal component. The proposed algorithm implements novel functionalities: it provides several method modifications for principal components computation, including weighted, robust and centered methods; it distinguishes overdispersed modules (based on the variance explained by the first principal component) and coordinated modules (based on the significance of the spectral gap); finally, it computes statistical significance of the estimated module overdispersion or coordination. ROMA can be applied in many contexts, from estimating differential activities of transcriptional factors to finding overdispersed pathways in single-cell transcriptomics data. We describe here the principles of ROMA providing several practical examples of its use. ROMA source code is available at https://github.com/sysbio-curie/Roma.

Keywords: module activity, gene set, overdispersed pathway, coordinated pathway, gene expression, proteomics, transcription factors

OPEN ACCESS

Edited by:
Ekaterina Shelest, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knoell Institute, Germany

Reviewed by:
Ka-Chun Wong, City University of Hong Kong, China
Dirk Fey, University College Dublin, Ireland

***Correspondence:**
Andrei Zinovyev
andrei.zinovyev@curie.fr

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 November 2015*
Accepted: *29 January 2106*
Published: *19 February 2016*

Citation:
Martignetti L, Calzone L, Bonnet E, Barillot E and Zinovyev A (2016) ROMA: Representation and Quantification of Module Activity from Target Expression Data. Front. Genet. 7:18. doi: 10.3389/fgene.2016.00018

1. INTRODUCTION

The current availability of high-throughput genomics techniques such as transcriptomics makes it possible to accurately measure molecular profiles of a biological system at multiple levels (Hawkins et al., 2010). Given the large amounts of quantitative data produced by these system-wide experiments, the interpretation of results in terms of cellular processes and pathways becomes a crucial issue. Dedicated integrative analyses are needed to synthesize and transform data into valuable biological insight (Hawkins et al., 2010).

Many biological and clinical applications require the comparison of samples from different conditions. The objective of the analysis often requires highlighting signaling pathways and transcriptional programs that distinguish between the compared conditions. A widely used approach in cancer genomics consists in comparing measurements at the single gene or protein

level to identify potential indicators of a particular disease state (biomarkers) or driver genes causally linked to the tumor initiation and progression (Barillot et al., 2012). In recent years, it has become clear that in cancer and other systemic diseases the same pathways can be affected by defects in different individual genes and that molecular profiles of tumor samples are more similar at the pathway level than at the gene level (Wang et al., 2010). Application of pathway-based approaches in the analysis of genomic data can help capturing biological information that is otherwise undetectable by focusing on individual genes. The idea of pathway quantification is widely exploited to extract biological information from high-throughput data (Levine et al., 2006; Ramos-Rodriguez et al., 2012; Borisov et al., 2014).

Here we propose an algorithm, released as a software, Representation Of Module Activity (ROMA), that was designed to address the issue of quantifying the activity of gene sets (further referred to as modules) characterized by coordinated gene expression. These modules can correspond to genes sharing the same functional annotations or regulatory motifs, genes belonging to the same pathway or genes forming a group of frequently coexpressed genes. The idea behind ROMA consists in quantifying module activity by computing the largest amount of one-dimensional variance across samples explained by the genes in the module (property of the first principal component or PC1). This is interpreted as a result of the action of a hidden factor on the expression of target module genes and variability in the activity of this factor in the studied collection of samples. This setting corresponds to the simplest linear model of gene expression regulation (for example, see Schreiber and Baumann, 2007; **Figure 1**).

ROMA implements several novel functionalities compared to existing related approaches. It allows determining genes within a group of genes contributing the most to the PC1 definition; it provides several alternative methods for PC1 computation, including weighted, robust and centered versions of principal component analysis; it estimates the statistical significance of the amount of variance explained by PC1 in two different ways; it distinguishes *overdispersed* and *coordinated* modules.

Here overdispersion of a gene set signifies that the amount of variance explained by PC1 computed for a dataset restricted to the genes from the set is significantly larger than for a random gene set of the same size. Coordinated gene set means that the spectral gap between the first and the second eigenvalues of the co-variance matrix computed for the restricted dataset is significantly larger than for a random gene set of the same size. Overdispersion signifies higher variability of a gene set even without increased correlations between genes. Coordination signifies relatively high degree of expression level correlation between genes in a gene set. Overdispersed set might be not coordinated: this is interpreted as simultaneous strong influence of several factors on the expression of the genes in the set. Coordinated set might be not overdispersed: this corresponds to a relatively weak but detectable activity of one single transcription or other factor on gene set expression. The most interesting and interpretable case is the case of simultaneous overdispersion and coordination of a gene set.

Naive quantification of the module activity frequently consists in computing the average or the median expression of the genes in the module in a given sample or, in opposite, relies on a single gene marker of module activity. ROMA is particularly suitable to model cases in which the different genes do not contribute similarly to the activity of the module, like the case in which some genes may be more important than others to define the activity of the module, or the case in which some genes are expected to negatively correlate with the activity of the module (e.g., p21, an inhibitor of the cyclin-dependent kinase complexes, may belong to a module of genes involved in the G1/S transition).

Several pathway quantification methods have been already proposed to recapitulate the activity of a module by computing the first metagene in the singular value decomposition (SVD) of the expression matrix restricted to the genes of the module (Tomfohr et al., 2005). In Bild et al. (2006) similar strategy was exploited in order to define the activity of several cancer-related pathways [MYC, RASA1 (RAS), SRC, Wnt/β-catenin and loss of RB function] on a large collection of human cancer transcriptomes. In Fan et al. (2016) the authors suggested the notion of "overdispersed pathway" in single-cell transcriptomic analysis framework such that the measure of activity in a set of genes is quantified by the statistical significance of the overdispersion explained by the first (weighted) principal component (PC1), computed for a set of single-cell transcriptomic profiles. Other methods have been developed for estimating module activity scores in individual samples such as single-sample extension of GSEA (ssGSEA) (Barbie et al., 2009) or OncoFinder (Borisov et al., 2014).

We illustrate the use of ROMA with four examples. In the first example, we quantify activities of several transcription factors (TFs) in metastatic and non-metastatic human colon tumor samples. In the second example, ROMA explores the transcriptional activity of modules in a comprehensive map of molecular interactions involving RB/E2F pathway in bladder cancer. The third application exploits ROMA to quantify transcriptional activity of targets for the oncogenic chimeric transcription factor EWS–FLI1 responsible of Ewing sarcoma initiation. Finally, we show an application of ROMA in the context of single-cell transcriptomic temporal profiling of myoblast differentiation (Trapnell et al., 2014).

2. MATERIALS AND METHODS

2.1. First Principal Component as the Simplest Uni-Factor Linear Model of Gene Expression Regulation

Let us consider the simplest model of gene regulation in which it is assumed that the expression of a gene g in sample s is proportional to the activity of one factor F (which can be a transcription or any other endogenous or exogenous factor affecting gene expression) in sample s with positive or negative (response) coefficient (**Figure 1A**):

$$Expression(gene\, g, sample\, s) \approx \alpha_g^F Activity_s^F + B_s, \qquad (a)$$

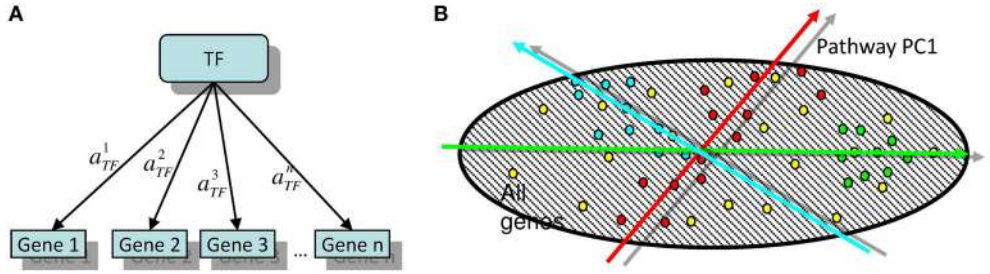


FIGURE 1 | (A) The simplest linear model of gene regulation: expression of a gene is proportional to the activity of a transcription factor TF. **(B)** Illustration of two possible configurations of target genes in the global gene expression space. Here, the points of different color signify the genes participating in different modules. Red points are symmetrically overdispersed in both directions from the global data distribution center, while the green points are displaced with respect to the global data distribution center. Computing the principal component passing through the center of the global (represented by dashed area) data distribution allows quantifying both types of pattern in a similar and comparable way. Moreover, the position of the center of the global data distribution defines the reference point with respect to which the sign of the projection of a gene onto the principal component can be defined.

where α_g^F is the coefficient of response of a gene g to the factor F, $Activity_s^F$ is the activity of the factor F in sample s, and B_s represents any sample-specific bias in measuring gene expression, affecting expression of all genes in sample s (B_s is analogous of the regression intercept in this linear model). In all further computations, we will assume that $\sum_s Expression\,(g, s) = 0$ for all genes. Without this normalization, there is a possibility that the computed PC1 will only explain the variations in the basal gene expression (which is frequently the case). By applying double-centering of the gene expression matrix, containing genes in a gene set G_i, i.e., making both $\sum_s Expression\,(g, s) = 0$ and $\sum_{g\in G_i} Expression\,(g, s) = 0$, one can achieve also $B_s = 0$. We do not suppose this normalization in the rest of this manuscript, because different gene sets can have different shift with respect to the center of the global distribution, hence, $B_s = 0$ can not be achieved for all gene sets at the same time.

Typically neither $Activity_s^F$ (activities of the factor in individual samples) nor α_g^F (the strength with which the factor F affects individual genes) are directly measurable. However, the simplest model fitting problem

$$\sum_s \sum_g \left(Expression(gene\,g, sample\,s) - \alpha_g^F Activity_s^F - B_s\right)^2 \to \min, \tag{b}$$

with constraints

$$\sum_g (\alpha_g)^2 = 1, \sum_g \alpha_g = 0 \tag{c}$$

is solved by finding the PC1 of the expression dataset $Expression\,(g, s), g \in G_i, s \in S$ restricted to the genes from a selected gene set G_i over all sample set S. If the data set does not contain missing values, then $B_s = \frac{1}{|G_i|}\sum_g Expression\,(g, s)$. To find both $Activity_s^F$ and α_g^F, one can apply the standard iterative SVD (Singular Value Decomposition) algorithm (e.g., see Gorban and Zinovyev, 2009), by starting with a random vector $Activity_s^F$ and computing $\alpha_g^F = \frac{\sum_s (Expression\,(g,s) - B_s) Activity_s^F}{\sum_s (Activity_s^F)^2}$. Then, the computed α_g^F are normalized to satisfy (c), and the new vector of factor activities is computed: $Activity_s^F = \sum_g \alpha_g^F Expression\,(g, s)$. The iterations are repeated until convergence. The constraints (c) are needed to guarantee convergence of this simple algorithm avoiding possible stretching or systematic drift of the α_g^F values.

Throughout the article, we will refer to a gene set G_i as "module" (accompanied by proper gene weights and signs if possible, as described below), where the biological interpretation of a "module" can be any functionally related list of genes, such as a set of direct targets of a transcription factor or other regulatory molecule, genes participating in the same signaling pathway as it is described in pathway databases, set of genomically co-localized genes, a set of genes containing the same motif for a transcription binding site, a set of co-expressed genes as a response to a particular perturbation, etc.

2.2. Principal Component Computation with Weights or Fixed Center

Computation of the PC1 can take into account the a priori estimated relative importance of a gene g in the module G_i. In order to achieve this, ROMA takes as an input the module descriptions which consist of a list of genes with a signed weight $w_g^{(G_i)}$ specified when possible (positive for "activators" and negative for "inhibitors" and undefined sign if the role of the gene is not known). The weights can be assigned only for some of the module genes with others being assigned the default 1.0 weight and undefined sign.

The computation of the principal components in ROMA is performed by the standard weighted SVD iterative algorithm as described in Gorban and Zinovyev (2009), where the weights for SVD are taken as the absolute values of the weights $|w_g^{(G_i)}|$ of the genes in the module. Introducing weights corresponds to generalizing the model fitting problem (d) to

$$\sum_s \sum_g |w_g^{(G_i)}| \left(Expression\,(gene\,g, sample\,s) - \alpha_g^F\, Activity_s^F - B_s\right)^2 \to \min. \tag{d}$$

Furthermore, in many cases, the activity of a module does not correspond to overdispersion of the module in the global gene expression space but to a shift of the genes in a particular direction (see **Figure 1B**). It is possible to quantify simultaneously this configuration of points and the overdispersed pattern using a simple modification of principal component computation such that the principal component would always pass through the center of the global distribution of points. This corresponds to the following modification of the initial linear model of gene regulation:

$$\sum_s \sum_g |w_g^{(G_i)}| \left(Expression\,(g,s) - \alpha_g^F Activity_s^F - C_s^{fixed}\right)^2 \rightarrow \min, \quad \text{(e)}$$

where C_s^{fixed} is the global central point of the data distribution. In this case, we do not assume (c) and it might be that all α_gs will possess the same sign (e.g., all targets being activated by a transcription factor).

We call this way of computing principal components as "PCA with fixed center." It is used by default in ROMA, though standard PCA (d) can be also used.

2.3. Orienting Principal Components

In the standard principal component analysis, all components are computed with undefined orientation sign: there is an inherent mirror symmetry in the optimization problem (d) because the optimized function is symmetric with respect to $\alpha_g \rightarrow -\alpha_g$, $Activity_s^F \rightarrow -Activity_s^F$ transformation. In ROMA we use the *a priori* information about the signs of genes in the module G_i to prefer one of two possible orientations of the PC1. We choose the orientation of PC1 for which

$$\sum_{g \in W^{(G_i)}} w_g^{(G_i)} \alpha_g^{(G_i)} > 0, \quad \text{(f)}$$

where $W^{(G_i)}$ is the set of genes in G_i for which both sign and weight are defined in the module description.

2.4. Computing Robust First Principal Component

The computation of the PC1 can be affected even by a single outlier in the data set. In order to increase robustness of the PC1 computation, we apply here the "leave-one-out" cross-validation approach (Hastie et al., 2001). We compute the distribution of L_1^i values where L_1^i is the variance explained by the PC1 with the point i removed. The distribution L_1^i is converted into a set of z-values, and all points with the absolute z-value bigger than z_{max} are removed from the dataset, where z_{max} is specified as a parameter (3.0 by default).

2.5. Estimating Statistical Significance of the Variance Explained for a Module

The PC1 can be computed for any random set of genes, and it will assign the hidden factor activity in the samples for any randomly chosen gene set. In order to avoid overfitting, we perform an empirical statistical test estimating the probability of a module to be *overdispersed* (i.e., to explain in the PC1 more variance than expected for a random set of genes) or *coordinated* (i.e., to explain in the PC1 more variance compared to the second principal component than expected for a random set of genes). Let us denote by L_1 the amount of variance explained by the PC1 and by L_2 the amount of variance explained by the second principal component. It is important to notice that the randomly expected values of both L_1 and L_2 strongly depend on the size of the module for which it is computed. Therefore, we compute the empirical null distributions for values L_1 and $\frac{L_1}{L_2}$ for K randomly chosen modules of the same size as the tested gene set.

In practice, there is frequently a need to test many module definitions. Estimating the null distribution for each tested gene set might lead to very expensive computations in terms of time. In ROMA, we do not compute the overdispersion significance scores for all possible module sizes, but instead we approximate them on predefined grid of size values. In order to rapidly estimate the significance of both overdispersion score (L_1) and the coordinatedness score ($\frac{L_1}{L_2}$), we construct the null distributions for a selected representative list of module sizes. The representative module sizes are chosen to be uniformly distributed in the log scale between the minimal size of the module in the collection and the maximal module size. For computing the empirical p-value, the null distribution which is the closest one in terms of size in the log scale is chosen.

2.6. Data Preprocessing for ROMA

The input format for gene or protein expression for ROMA is a tab-delimited text file with columns corresponding to biological samples and rows corresponding to genes or proteins. The first line is assumed to contain the sample identifiers while the first column is assumed to contain the non-redundant names of genes or proteins. In addition, ROMA can use description of samples also in tab-delimited text file format, in which the first row is assumed to contain the names of the features with which the samples are annotated and the first column will contain the names of the samples, in the same format as they are defined in the first row of the expression data table.

Optionally the input expression data can be centered or double-centered. If the data table contains missing values, they can be imputed using the approximation of the data matrix with missing values by a complete lower-rank matrix. For this, the user has to specify the rank k_{rank} of the approximative complete matrix. After this, k_{rank} principal components are calculated using the PCA algorithm able to work with missing data values (Gorban and Zinovyev, 2009). This PCA decomposition is used to construct the lower rank complete approximative matrix, from which the missing values in the initial data are imputed. For further computations, the completed initial data matrix of full rank is used.

2.7. ROMA Implementation and Workflow Description

ROMA is implemented as a Java library which can be launched in command line. For computation of weighted PCA, and PCA with fixed center, ROMA exploits *vdaoengine* library. ROMA

source code with instructions to build and run the application are available at http://github.com/sysbio-curie/Roma.

The analysis workflow is schematized in **Figure 2**. The algorithm requires as an input a genome-wide expression data matrix and a gmt file with predefined modules. The analysis comprises a multistep procedure for (i) extracting expression submatrices corresponding to each module, (ii) quantifying robust PC1 based module activities and (iii) assessing the statistical significance of the L_1 and $\frac{L_1}{L_2}$ values. ROMA provides as outputs different text files and tables including: a module score table with the overdispersion scores (L_1) and the coordinatedness scores ($\frac{L_1}{L_2}$) with corresponding p-values for each module, a matrix file with rows containing the activity scores of each module across samples, a table for each module reporting the projections of genes in the PC1-PC2 space computed for a given module.

3. RESULTS

As previously mentioned, typical scenarios for applying ROMA is to measure the activity of a transcription factor. It can also be applied in other cases, such as finding the activity of a kinase from phosphoproteomic data, or finding an abstract aggregated "activity" of a set of functionally related genes (such as genes belonging to the same pathway), assuming that overdispersed or coordinated behavior of the genes in the pathway is an indicator of its active state. We describe the application of ROMA to multiple case studies. In three of them, the biological information about the activity of the modules under study was a priori available and confirmed by ROMA results. The last case study shows an exploratory analysis by ROMA applied to single-cell RNA-seq data.

3.1. Notch, Wnt, and p53 Pathways Activity in Human Colon Cancer

As a first case study, we applied ROMA to quantify the activity of Notch, Wnt and p53 pathways in invasive and non-invasive human colon tumors. In a previous study on a mouse model, p53 loss of function and Notch gain of function have been predicted to have synergistic effect in the induction of the epithelial to mesenchymal (EMT)-like phenotype (Chanrion et al., 2014). To investigate in human data the involvement of Wnt, p53, and Notch pathways in EMT induction, we used a publicly available gene expression dataset of human colon cancer samples from The Cancer Genome Atlas (TCGA) project (Muzny et al., 2012) and compared the activity scores of Notch, Wnt and p53 pathways in metastatic and non-metastatic samples. Genome-scale expression profiles of 121 tumor samples were used in our analysis.

Differential expression analysis of single genes involved in Wnt and Notch signaling pathways did not show significant changes between metastatic and non-metastatic tumors (see File S1). Thus, we investigated the involvement of these pathways by computing with ROMA the activity scores of their downstream target sets. Levels of pathway activity across tumor samples revealed that Notch and Wnt pathways were significantly activated, whereas the p53 pathway was downregulated in the metastatic compared to non-metastatic tumors (**Figure 3**). Molecular Signature Database (Subramanian et al., 2005) was used to select the sets of target genes for Notch and Wnt pathways (see File S3). Among several available modules, we chose the ones having the best differential activity score between metastatic and non-metastatic samples for computing Notch and Wnt pathway activities. For p53 pathway activity, we used a set of known p53 primary targets (Kannan et al., 2001).

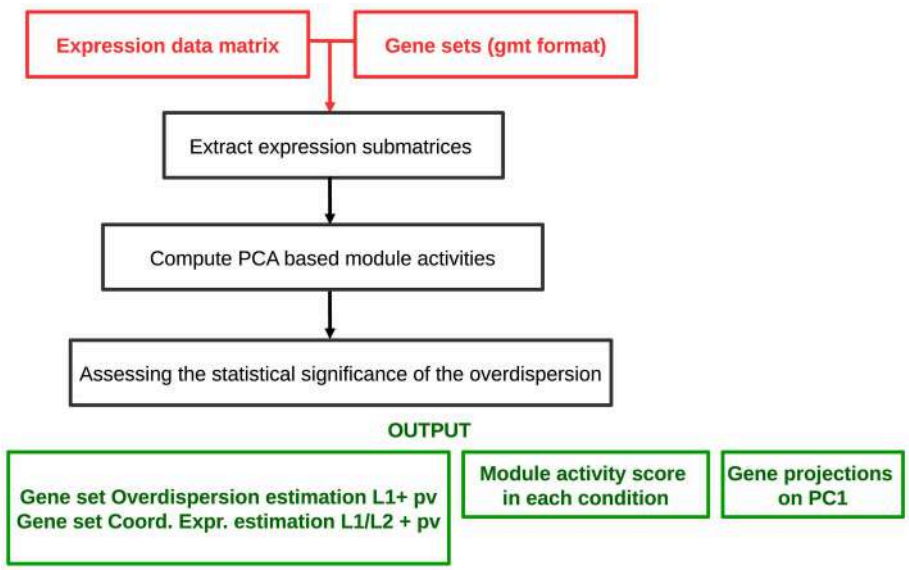


FIGURE 2 | Schematized workflow of the ROMA algorithm.

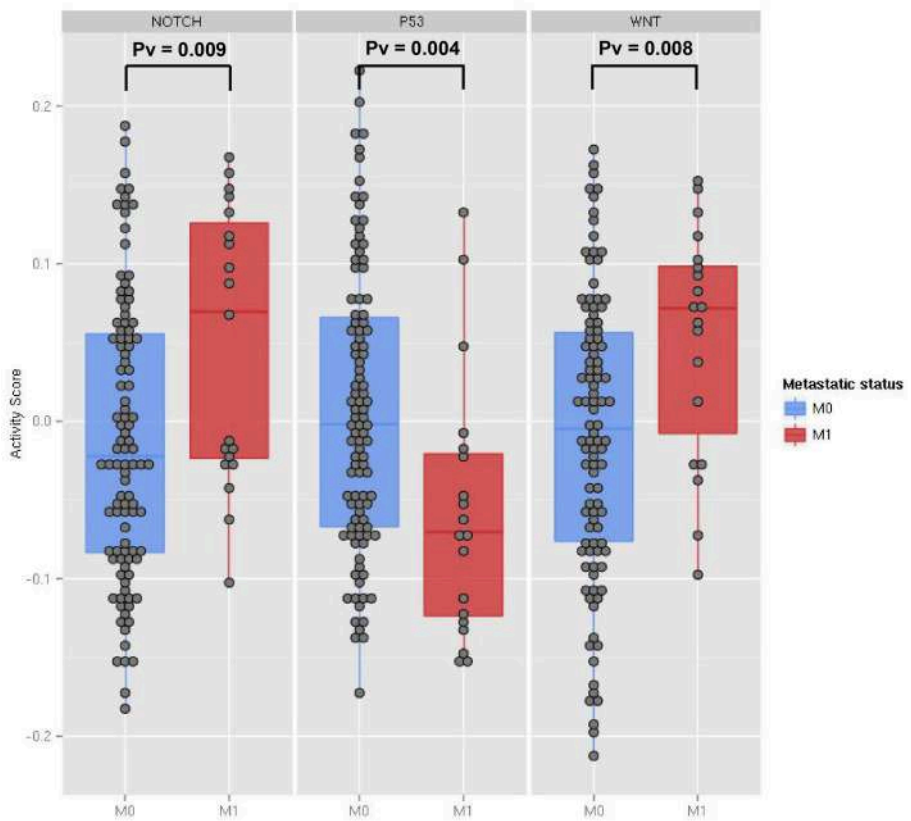


FIGURE 3 | The activity scores computed for the Notch, p53 and Wnt pathways in human transcriptome data from TCGA colon cancer samples. The data points represent primary tumor samples grouped as non-metastatic (blue) and metastatic (red) according to the observation of distant metastases. *P*-values are calculated using the two-sample Kolmogorov-Smirnov test between the two groups.

3.2. Dysregulated Signaling Pathways in Bladder Cancer

We performed the ROMA analysis on a transcriptome dataset of bladder tumors with clinical information about the stage of the tumors (Lindgren et al., 2010). Two groups of samples were selected for comparison, invasive and superficial. Normal samples are also provided (details can be found in File S1). The modules of genes chosen for this analysis are those that are known to be frequently dysregulated in this cancer and that include, among others, cell cycle and apoptotic pathways (see File S2). Inside each module, the genes that are known to be representative of the activity of the module are specified as positive contributors of the module, e.g., E2F1, E2F2, and E2F3 are assigned a positive sign in the module E2F, whereas RB1 is assigned a negative weight. The modules that appear in the analysis are the ones for which at least 8 genes are found in the dataset. We plotted the module activity scores for which the L1 *p*-value was lower than 0.05 onto an influence network (**Figure 4**) for the three cases: normal samples, superficial tumors, and invasive tumors. The influence network was drawn using CellDesigner software with connections extracted by manual literature mining. We also plotted the module NF-KB signaling that has a *p*-value of 0.12, knowing that the activity of this module cannot be as trusted as the others.

We find that in normal samples and superficial tumor samples, the activity for the modules of the E2F1, E2F2 and E2F3 target genes is lower than in invasive tumors, as opposed to the target genes of the inhibitory transcription factors E2F4 and E2F6. This is in accordance with what is expected. Indeed, in the invasive group, tumors show a higher proliferation rate. Also, TGFb activity is lower in the invasive group than in the superficial one. Interestingly, the activity of the death signaling pathway (DDR signaling) is high in normal samples, lower in superficial tumors and start to be higher again in invasive tumors. RTK signaling activity, representing growth factors, is low in normal samples but is found high in both tumor groups. Indeed, genetic alterations in the EGFR, FGFR3, and RAS pathways are typical of tumor initiation and progression in bladder.

3.3. Estimating Activity of EWS/FLI-1 Chimeric Transcription Factor in Ewing Sarcoma

We tested ROMA algorithm on transcriptome time-course measurements performed on Ewing sarcoma inducible cell lines after EWS-FLI1 silencing and re-expression (Tirode

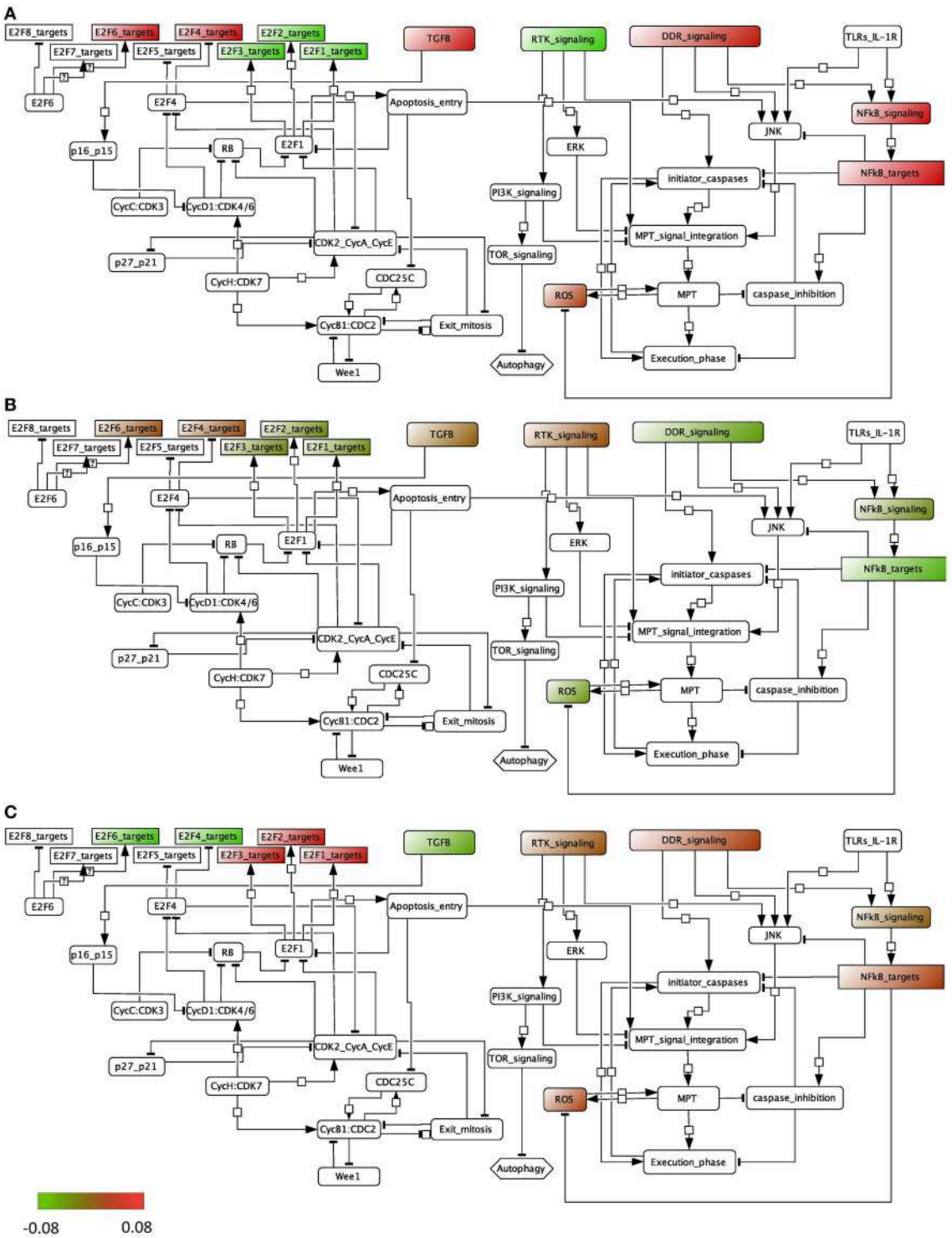


FIGURE 4 | Representation of the module activity of bladder dataset (Lindgren) onto a signaling network that is drawn from literature known facts and that illustrates the module activity for (A) normal samples, (B) superficial tumors, and (C) invasive tumors.

et al., 2007; Stoll et al., 2013). EWS-FLI1 is a chimeric transcription factor specific to Ewing sarcoma disease and responsible for a tumorigenic phenotype. Different studies have reported opposing transcriptional activity of EWS-FLI1 whether it binds to transcriptional co-activators (Fuchs et al., 2003) or transcriptional co-repressors (Sankar et al., 2013). Since

EWS-FLI1 functions as both an activator and an inhibitor, the simple average expression of its target genes does not reflect its active/inactive state (see boxplot in File S1). Instead, weights obtained when applying ROMA to the expression matrix of target genes provide an appropriate measure of EWS-FLI1 activity (see File S4).

We studied the effect of EWS-FLI1 on a predefined signature of dysregulated genes (Hancock and Lessnick, 2008) by computing the activity score of this set of targets over time. First, ROMA analysis was performed for the whole set of genes. In this case, the sign of the weights for some target genes was specified according to a priori biological knowledge about the regulation of up and down targets. Secondly, the same analysis was performed by splitting the initial signature in two separated modules for the predicted up and down-regulated targets. Among the three tested modules, the whole signature target set showed the most significant overdispersion pattern across time points, with $L_1 = 0.52$ (p-value $= 0.001$). ROMA analysis using down-regulated targets gave a better overdispersion signal compared to up-regulated targets (see detailed results in File S1). We expected the activity scores of the EWS-FLI1 set of targets to show modulation of the expression of targets of EWS-FLI1 over time. Results confirmed that the activity scores of both up and down-regulated target sets properly reflected the dynamics of EWS-FLI1 expression during the inhibitory ($t = 0 - 10$ days) and rescue ($t = 10 - 27$ days) time series experiments (**Figure 5A**). Instead, the average expression of the same set of targets did not show modulation across the time points.

We tested whether the expression of other modules than EWS-FLI1 targets showed a significantly overdispersed pattern upon EWS-FLI1 inhibition and reactivation. This could reveal relevant biological functions affected by EWS-FLI1 expression. ROMA analysis was performed on the EWS-FLI1 transcriptome time-series using a large collection of predefined signaling pathways from Molecular Signature Database (MSigDB Liberzon et al., 2011). In this example, we used a subset of MSigDB limited to the pathway definitions imported from KEGG (Ogata et al., 1999), REACTOME (Croft et al., 2014), BIOCARTA (Nishimura, 2001) pathway databases. To these sets, we added 59 definitions of modules from Atlas of Cancer Signaling Network (ACSN) (Kuperstein et al., 2015) and the set of potential transcriptional targets of EWS/FLI-1 chimeric oncogene (Hancock and Lessnick, 2008). In total, this resulted in 1121 modules. Out of all these modules, 23 had significant overdispersion in time series measurements with p-value < 0.05 (see File S5). For these modules, we distinguished two different kinetics in their response to EWS-FLI1 expression reflected by their activity score, one having switch-like response similar to EWS-FLI1 signature targets and a second one similar to a pulse-like response (**Figure 5B**).

3.4. Detecting Overdispersed Pathways in Single-Cell RNASeq Data

Application of module activity estimation is particularly interesting to determine molecular pathways contributing to the non-genetic heterogeneity of cell populations in the context of single cell transcriptomics data analysis (Fan et al., 2016). In order to demonstrate that ROMA can be used to detect overdispersed pathways in single cell transcriptomics data, we applied it to a set of 372 individual cell transcriptomic profiles measured in several time points after induction of differentiation in a skeletal myoblast cell culture (Trapnell et al., 2014).

The collection of gene sets used for this example was taken as in the previous section. ROMA has detected a number of overdispersed pathways (many more than in the previous examples) revealing major biological functions contributing to the cell-to-cell transcriptome variation. As expected, clustering overdispersed pathways according to their module activity

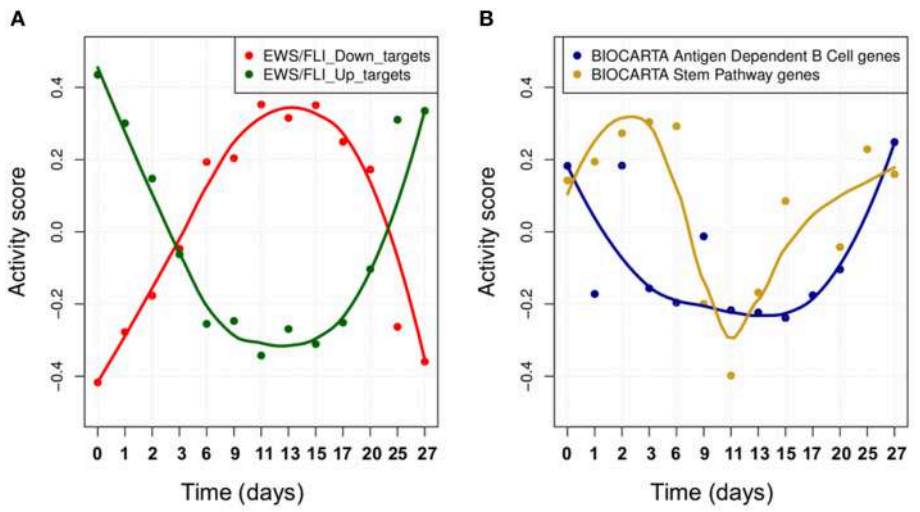


FIGURE 5 | (A) The activity scores of both up and down-regulated target sets during EWS-FLI1 inhibition (days 0–10) and rescue (days 10–27) time series experiments. EWS-FL1 time-course related to the dataset was measured and reported in **Figure 3A** of (33). **(B)** Smoothed temporal activity profile for two overdispersed pathways found by ROMA in the analysis of time series expression profile after inhibition of EWS-FLI1.

score profiles (see Supplementary Materials) distinguished a large cluster of signatures related to cell cycle and closely related DNA replication and DNA repair. A large cluster of 50 signatures mixed modules related to apoptosis, respiratory electron transport, TCA cycle and various metabolism and catabolism-related modules. A cluster of 10 signatures was related to translation. Another cluster of 16 signatures contained modules related to transcription, mRNA splicing and mRNA processing. Relatively small cluster contained six signatures related to glucose transport and, surprisingly, metabolism of non-coding RNA. Two smaller clusters included five gene signatures related to extracellular matrix organization, and muscle contraction together with cardiomyopathy (which is probably more specific to the cellular function of myoblasts).

In **Figure 6A** we show several examples of overdispersion pattern observed in the single-cell RNASeq dataset. We observed that most overdispersed modules obtained high score due to a systematic shift with respect to the global gene distribution, such as the leftmost E2F3_TARGETS signature in **Figure 6A**. In **Figure 6B** we show the profiles of module activity scores across all cells, ordered in time. E2F3_TARGETS signature from ACSN pathway database probably marks the cells in the active proliferation state. One can see that the number of proliferating cells drops at the time point T24 when compared to the time point T0. However, there remains a significant number of proliferating cells after T24. Interestingly, the modules can be classified into those showing clear bimodal distribution of activity scores and those having unimodal distribution (e.g., see the KEGG dilated cardiomyophathy profile in **Figure 6B**). One can observe also that the variance of module activity scores might vary significantly from one time point to another (see the same profile on **Figure 6B**).

Note that in all of the four analyses presented above, we have found a large set REACTOME_OLFACTORY_SIGNALING_PATHWAY overdispersed. Olfactory receptors are known to be a common confounding signal in many mutation profiling analyses (Lawrence et al., 2013). It seems that this is also reflected in pathway overdispersion analysis, based on transcriptomic data of normal or cancer cells. We are not aware that this phenomenon was described before.

4. DISCUSSION

Quantifying the activity of biologically related modules is a widely exploited strategy to extract biological information from high-throughput data. In the analysis of genomic data, using gene sets as aggregated variables can help to capture biological information that is otherwise undetectable by focusing only on individual genes. We introduced the ROMA algorithm which deals with this problem of quantifying the activity of modules by fast and robust computation of the simplest linear model of gene regulation based on computing the PC1 of the expression data matrix and estimating the statistical significance of such approximation.

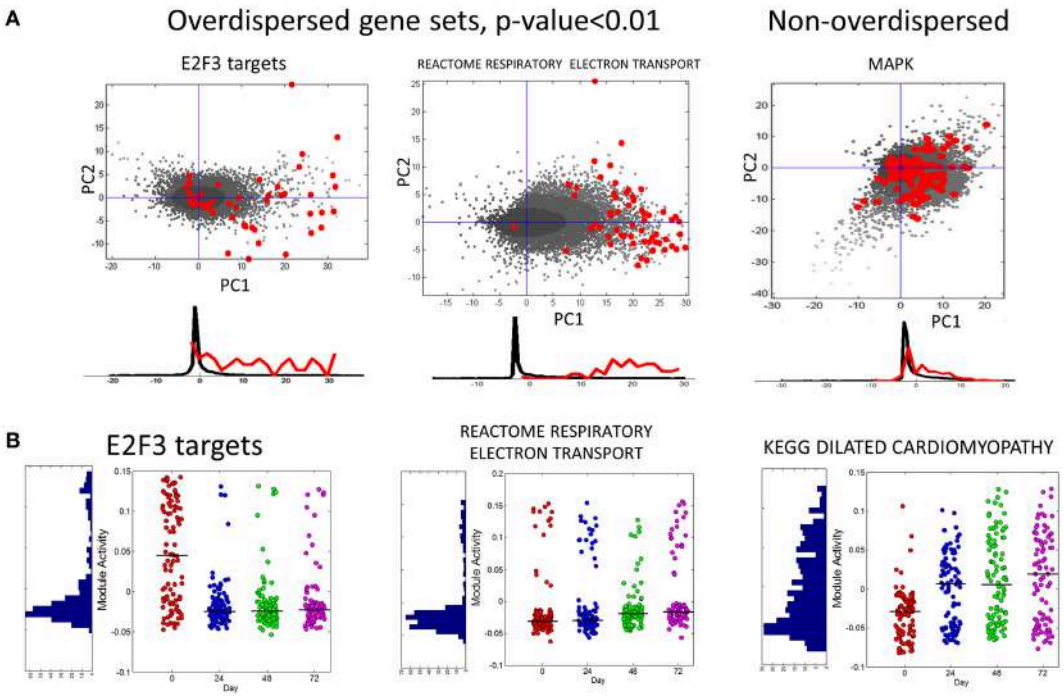


FIGURE 6 | (A) Examples of overdispersed and non-overdispersed pathways in single-cell RNA-Seq data. Red points are the genes of the pathways, shown in the projection on the first two principal components computed for these points. Black points show the global distribution projected in the first two principal components of the pathway. Below the scatterplot, the histogram of gene projections on the PC1 is shown separately for the genes in the pathway (red) and for the global distribution (black). **(B)** Module activation score in single cells. The x-axis corresponds to four time points (T0-T72). The black line shows the median module activation score within the same time point. On the left of the graph the histogram of module activation scores for all cells in all time points is shown.

We tested ROMA on a first case study to quantify the activity of Notch, Wnt and p53 pathways in metastatic and non-metastatic tumors from human colon cancer transcriptome data. Unlike single gene expression analysis, the ROMA algorithm has effectively shown the involvement of these pathways in the metastastic process by detecting their differential activity. In this study, the sets of downstream transcriptional targets reflect the activity of the associated pathways better than any individual gene involved in the signaling cascades. In similar gene set analysis, ROMA can be considered as a powerful algorithm to detect coordinated but small changes of several genes in a pathway.

In our second example ROMA was used to map the expression profiles of bladder patients on an influence graph that recapitulates the molecular interactions between different pathways. The information extracted from the data correlates to what is known about the tumor progression in bladder cancer. To complete the analysis, it would be possible to translate the influence network into a logical model. This would consist in associating to each module (equivalent to a variable of the model) a logical rule linking all of his inputs with the logical operators AND, OR, and NOT. For instance, ROS would be written as follows: ROS = MPT AND NOT NFkB_targets. Thus, if the influence network was to be translated into a logical model and simulated for each patient profile (set of mutations or genetic alterations known for the genes included in the model) with accompanying clinical information (stage of the tumor), we would expect to see the solutions of the simulation, referred to as stable states, of an invasive patient with active E2F1, E2F2, and EF3 target variables (equal to 1) whereas the stable states for patients with superficial tumors with these variables equal to 0. The data analysis performed with ROMA is also one way to assess that the logical rules are in accordance with the dataset that is studied and thus that the model represents correctly the dynamics of bladder tumorigenesis. Another possible use of ROMA in the context of network modeling can be in the selection of the pathways of interest to include in the model. Constructing a structural model of a specific complex molecular process can be based on literature information combined with an exploratory analysis of pathway databases to identify those pathways that are active or inactive in a particular cellular condition.

In the third example, we described the application of ROMA in quantifying transcriptional activity of targets of EWS-FLI1 from time-course measurements. Since this oncogenic TF can have both inhibitory and activating properties, ROMA analysis was performed first for the whole set of known target genes and secondly by splitting the set in two separated modules for the up and down regulated targets. The whole signature target set was the most significantly overdispersed. This is consistent with the fact that a larger set of co-regulated genes, regardless of the regulation sign, is expected to generate a stronger overdispersion signal. This is an advantageous property of ROMA compared to other gene set testing methods, such as GSEA, that estimate the significance of enrichment score by considering separately the positively and negatively scoring gene sets. Also, several TFs can have both inhibitory and activating function; ROMA can be applied without information about the sign of the TF effect on its targets. In time series data, the scores calculated on the sets of targets can give information on the kinetics of the transcriptional response. The activity scores of targets reflect the dynamics of EWS-FLI1 expression during the inhibitory and rescue experiments.

In the fourth example, ROMA is applied to detect overdispersed pathways in single cell transcriptomics data. This is particulary interesting application of unsupervised ROMA approach, because it potentially allows quantifying the non-genetic heterogeneity of a cell population on pathway level. Multiple gene sets have been shown to be overdispersed in this case: therefore, clustering them based on the activity profiles over the cell population helps identifying the major functional aspects contributing to cell-to-cell variance.

In many studies ROMA can be applied to unravel the effective status of a TF protein from the expression of its target genes. The predicted activity values can be validated experimentally. if the active form of a transcription factor or other factor is known and can be measured (i.e., by mass spectrometry measurements), or the factor represents a measurable phenotypic read-out (such as cell growth or age).

Oncogenes and tumor suppressor regulatory genes, such as p53, often carry mutations in their DNA sequences. However, such DNA changes do not always have a clear effect at the phenotypic level. On the other hand, the function of oncogenes or tumor suppressors can be compromised by other mechanisms than DNA mutations, like for example alterations in DNA methylation. Computing activity score of transcriptional target sets is a useful method to assess the active or inactive status of regulatory oncogenes or tumor suppressors. We can also imagine to label tumor samples in a more reliable manner by relying both on the targets activity score and on DNA mutations. Our previous study shows that the estimated activity of p53 in tumor samples is better associated to the clinical outcome than expression or mutation status of p53 alone (unpublished data). Recent advances in chromatin immunoprecipitation with next-generation DNA sequencing (ChIP-Seq) have provided large collections of detected TFBSs with high sensitivity that facilitate the comprehensive annotation of TF targets sets.

The idea of applying ROMA in order to investigate the effect of regulatory molecules can be generalized in order to study other classes of regulators, such as kinases, phosphatases, microRNAs, etc. The availability of large-scale proteomics and phosphoproteomics data gives unprecedented knowledge about post-transcriptional and post-translational regulation happening in the cell. The ROMA algorithm can be applied to analyze quantitative phosphoproteomics profiles and identify overdispersed patterns of predefined sets of proteins sharing common phosphorylation sites. By exploiting this information it would be possible to infer active or inactive kinases/phosphatases.

Multiple types of analyses using ROMA can be performed in order to explore microRNA regulation. First, microRNA genes appear often organized in genomic clusters that are not randomly composed, meaning that this clustered structure is evolutionary conserved and is likely to be related to miRNAs coordinated regulatory action. Comparing expression level of clustered miRNAs in different conditions, the variation in the

abundance of each individual miRNA of the cluster can be weak and not detectable by standard statistical hypotheses testing applied to individual miRNA expression levels, while the overdispersed expression pattern of the entire cluster can produce a statistically significant signal and reveal its differential activity.

ROMA can also be useful for the identification of microRNA regulation by expression analysis of target genes. The module approach is particularly suitable to infer miRNA regulatory effect from target expression profiles, since miRNA effect is subtle at the level of individual target but affects a large number of genes (Martignetti et al., 2015).

ROMA can be used in combination with unsupervised methods for metagene extraction from omics data such as Independent Component Analysis (ICA) for helping component interpretation (Zinovyev et al., 2013; Biton et al., 2014).

In the future it would be interesting to generalize the linear model of ROMA method onto a non-linear case, through application of non-linear versions of principal component analysis such as principal curves (Gorban and Zinovyev, 2001; Gorban et al., 2008) or principal trees (Gorban and Zinovyev, 2009). Indeed, distributions of gene expression profiles are demonstrated to contain non-linearities (Drier et al., 2013) and branching points. For example, a variant of principal curve approach was suggested in Trapnell et al. (2014) in order to recapitulate the non-linear dynamics of myoblast differentiation. Non-linearity leads to the situation when there exists no one single set of genes contributing the most to the definition of module activity: this set will depend on a particular region of the gene expression space. This will complicate the interpretation of the module activity: however, many ideas introduced in ROMA (estimating statistical significance of overdispersion, robust modification of non-linear PCA, etc.) will remain applicable.

To conclude, we prove that ROMA is useful when applied to different biological case studies. ROMA will contribute to the set of tools routinely applied in systems biology according to the application examples outlined before. In the future, we will provide a Graphical User Interface to facilitate the use of the ROMA algorithm, in the form of a Cytoscape app (Smoot et al., 2011; Saito et al., 2012).

AUTHOR CONTRIBUTIONS

LM, LC, EBa, and AZ designed and implemented the methodology. EB packaged the code and worked on improving the methodology. LM, LC, and AZ has provided the examples of methodology use. All authors have read and worked on the manuscript.

ACKNOWLEDGMENTS

The research leading to these results of this article have received funding from the European Union Seventh Framework Programme (FP72007-2013) ASSET project under grant agreement number FP7-HEALTH-2010-259348. The work was supported by ITMO Cancer SysBio program (MOSAIC project) and by Institut National de la Santé et de la Recherche Médicale (U900 and U830 budget). This work has received support under the program "Investissements d'Avenir" launched by the French Government and implemented by ANR with the references ANR-11-BINF-0001 (Project ABS4NGS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00018

SUPPLEMENTAL DATA

File S1 | Supplemental Material about the described examples and results.

File S2 | Modules gmt file used in bladder cancer data analysis.

File S3 | Sets of target genes for p53, Notch and Wnt pathways from Molecular Signature Database.

File S4 | Weights for up-regulated genes, down-regulated genes and full signature obtained by ROMA to the expression matrix of target genes.

File S5 | Modules gmt file used for the two last examples. The pathway definitions come from KEGG, Reactome, Biocarta and ACSN.

REFERENCES

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic kras-driven cancers require tbk1. *Nature* 462, 108–112. doi: 10.1038/nature08460

Barillot, E., Calzone, L., Hupe, P., Vert, J.-P., and Zinovyev, A. (2012). *Computational Systems Biology of Cancer*. Boca Raton, FL: CRC Press.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., et al. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* 9, 1235–1245. doi: 10.1016/j.celrep.2014.10.035

Borisov, N. M., Terekhanova, N. V., Aliper, A. M., Venkova, L. S., Smirnov, P. Y., Roumiantsev, S., et al. (2014). Signaling pathways activation profiles make better markers of cancer than expression of individual genes. *Oncotarget* 5, 10198–10205. doi: 10.18632/oncotarget.2548

Chanrion, M., Kuperstein, I., Barrière, C., El Marjou, F., Cohen, D., Vignjevic, D., et al. (2014). Concomitant notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. *Nat. Commun.* 5:5005. doi: 10.1038/ncomms6005

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The reactome pathway knowledgebase. *Nucl. Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* doi: 10.1038/nmeth.3734. [Epub ahead of print].

Fuchs, B., Inwards, C. Y., and Janknecht, R. (2003). Upregulation of the matrix metalloproteinase-1 gene by the ewings sarcoma associated ews-er81 and ews-fli-1 oncoproteins, c-jun and p300. *FEBS Lett.* 553, 104–108. doi: 10.1016/S0014-5793(03)00984-0

Gorban, A., Kegl, B., Wunsch, D., and Zinovyev, A., (eds.). (2008). *Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58*. Berlin: Springer. doi: 10.1007/978-3-540-73750-6

Gorban, A., and Zinovyev, A. (2001). *Visualization of Data by Method of Elastic Maps and its Applications in Genomics, Economics and Sociology*. IHES Preprints (IHES/M/01/36). Available online at: http://preprints.ihes.fr/M01/Resu/resu-M01-36.html

Gorban, A. N., and Zinovyev, A. (2009). "Principal graphs and manifolds," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, eds E. S. Olivas, J. D. M. Guererro, M. M. Sober, J. R. M. Benedito, and A. J. S. Lopes (Hershey, PA: IGI Global).

Hancock, J. D., and Lessnick, S. L. (2008). A transcriptional profiling meta-analysis reveals a core ews-fli gene expression signature. *Cell Cycle* 7, 250–256. doi: 10.4161/cc.7.2.5229

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning, Springer Series in Statistics*. New York, NY: Springer New York Inc. doi: 10.1007/978-0-387-21606-5

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486. doi: 10.1038/nrg2795

Kannan, K., Amariglio, N., Rechavi, G., Jakob-Hirsch, J., Kela, I., Kaminski, N., et al. (2001). Dna microarrays identification of primary and secondary target genes regulated by p53. *Oncogene* 20, 2225–2234. doi: 10.1038/sj.onc.1204319

Kuperstein, I., Bonnet, E., Nguyen, H.-A., Cohen, D., Viara, E., Grieco, L., et al. (2015). Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis* 4:e160. doi: 10.1038/oncsis.2015.19

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Levine, D. M., Haynor, D. R., Castle, J. C., Stepaniants, S. B., Pellegrini, M., Mao, M., et al. (2006). Pathway and gene-set activation measurement from mrna expression data: the tissue distribution of human pathways. *Genome Biol.* 7:R93. doi: 10.1186/gb-2006-7-10-r93

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lindgren, D., Frigyesi, A., Gudjonsson, S., Sjödahl, G., Hallden, C., Chebil, G., et al. (2010). Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res.* 70, 3463–3472. doi: 10.1158/0008-5472.CAN-09-4213

Martignetti, L., Tesson, B., Almeida, A., Zinovyev, A., Tucker, G. C., Dubois, T., et al. (2015). Detection of mirna regulatory effect on triple negative breast cancer transcriptome. *BMC Genomics* 16(Suppl. 6):S4. doi: 10.1186/1471-2164-16-S6-S4

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Nishimura, D. (2001). Biocarta. *Biotech. Softw. Int. Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Ramos-Rodriguez, R.-R., Cuevas-Diaz-Duran, R., Falciani, F., Tamez-Peña, J.-G., and Trevino, V. (2012). Compadre: an r and web resource for pathway activity analysis by component decompositions. *Bioinformatics* 28, 2701–2702. doi: 10.1093/bioinformatics/bts513

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., et al. (2012). A travel guide to cytoscape plugins. *Nat. Methods* 9, 1069–1076. doi: 10.1038/nmeth.2212

Sankar, S., Bell, R., Stephens, B., Zhuo, R., Sharma, S., Bearss, D. J., et al. (2013). Mechanism and relevance of ews/fli-mediated transcriptional repression in ewing sarcoma. *Oncogene* 32, 5089–5100. doi: 10.1038/onc.2012.525

Schreiber, A. W., and Baumann, U. (2007). A framework for gene expression analysis. *Bioinformatics* 23, 191–197. doi: 10.1093/bioinformatics/btl591

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Stoll, G., Surdez, D., Tirode, F., Laud, K., Barillot, E., Zinovyev, A., et al. (2013). Systems biology of ewing sarcoma: a network model of ews-fli1 effect on proliferation and apoptosis. *Nucl. Acids Res.* 41, 8853–8871. doi: 10.1093/nar/gkt678

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tirode, F., Laud-Duval, K., Prieur, A., Delorme, B., Charbord, P., and Delattre, O. (2007). Mesenchymal stem cell features of ewing tumors. *Cancer Cell* 11, 421–429. doi: 10.1016/j.ccr.2007.02.027

Tomfohr, J., Lu, J., and Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6:225. doi: 10.1186/1471-2105-6-225

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843–854. doi: 10.1038/nrg2884

Zinovyev, A., Kairov, U., Karpenyuk, T., and Ramanculov, E. (2013). Blind source separation methods for deconvolution of complex signals in cancer biology. *Biochem. Biophys. Res. Commun.* 430, 1182–1187. doi: 10.1016/j.bbrc.2012.12.043

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 November 2015
doi: 10.3389/fgene.2015.00331

Mapping Mammalian Cell-type-specific Transcriptional Regulatory Networks Using KD-CAGE and ChIP-seq Data in the TC-YIK Cell Line

*Marina Lizio[1,2], Yuri Ishizu[1,2], Masayoshi Itoh[1,2,3], Timo Lassmann[1,2,4], Akira Hasegawa[1,2], Atsutaka Kubosaki[1], Jessica Severin[1,2], Hideya Kawaji[1,2,3], Yukio Nakamura[5], the FANTOM consortium[1], Harukazu Suzuki[1,2], Yoshihide Hayashizaki[1,3], Piero Carninci[1,2] and Alistair R. R. Forrest[1,2,6]**

[1] RIKEN Center for Life Science Technologies, Yokohama, Japan, [2] Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan, [3] RIKEN Preventive Medicine and Diagnosis Innovation Program, Yokohama, Japan, [4] Telethon Kids Institute, The University of Western Australia, Subiaco, WA, Australia, [5] Cell Engineering Division, RIKEN BioResource Center, Ibaraki, Japan, [6] QEII Medical Centre and Centre for Medical Research, Harry Perkins Institute of Medical Research, The University of Western Australia, Nedlands, WA, Australia

OPEN ACCESS

Edited by:
Edgar Wingender,
The University Medical Center Göttingen, Germany

Reviewed by:
Mikael Boden,
The University of Queensland, Australia
Ka-Chun Wong,
City University of Hong Kong, China

***Correspondence:**
Alistair R. R. Forrest
alistair.forrest@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 21 September 2015
Accepted: 30 October 2015
Published: 18 November 2015

Citation:
Lizio M, Ishizu Y, Itoh M, Lassmann T, Hasegawa A, Kubosaki A, Severin J, Kawaji H, Nakamura Y, FANTOM consortium, Suzuki H, Hayashizaki Y, Carninci P and Forrest ARR (2015) Mapping Mammalian Cell-type-specific Transcriptional Regulatory Networks Using KD-CAGE and ChIP-seq Data in the TC-YIK Cell Line. Front. Genet. 6:331. doi: 10.3389/fgene.2015.00331

Mammals are composed of hundreds of different cell types with specialized functions. Each of these cellular phenotypes are controlled by different combinations of transcription factors. Using a human non islet cell insulinoma cell line (TC-YIK) which expresses insulin and the majority of known pancreatic beta cell specific genes as an example, we describe a general approach to identify key cell-type-specific transcription factors (TFs) and their direct and indirect targets. By ranking all human TFs by their level of enriched expression in TC-YIK relative to a broad collection of samples (FANTOM5), we confirmed known key regulators of pancreatic function and development. Systematic siRNA mediated perturbation of these TFs followed by qRT-PCR revealed their interconnections with *NEUROD1* at the top of the regulation hierarchy and its depletion drastically reducing insulin levels. For 15 of the TF knock-downs (KD), we then used Cap Analysis of Gene Expression (CAGE) to identify thousands of their targets genome-wide (KD-CAGE). The data confirm *NEUROD1* as a key positive regulator in the transcriptional regulatory network (TRN), and *ISL1,* and *PROX1* as antagonists. As a complimentary approach we used ChIP-seq on four of these factors to identify *NEUROD1*, *LMX1A*, *PAX6,* and *RFX6* binding sites in the human genome. Examining the overlap between genes perturbed in the KD-CAGE experiments and genes with a ChIP-seq peak within 50 kb of their promoter, we identified direct transcriptional targets of these TFs. Integration of KD-CAGE and ChIP-seq data shows that both *NEUROD1* and *LMX1A* work as the main transcriptional activators. In the core TRN (i.e., TF-TF only), *NEUROD1* directly transcriptionally activates the pancreatic TFs *HSF4, INSM1, MLXIPL, MYT1, NKX6-3, ONECUT2, PAX4, PROX1, RFX6, ST18, DACH1,* and *SHOX2*, while *LMX1A* directly transcriptionally

activates *DACH1, SHOX2, PAX6,* and *PDX1*. Analysis of these complementary datasets suggests the need for caution in interpreting ChIP-seq datasets. (1) A large fraction of binding sites are at distal enhancer sites and cannot be directly associated to their targets, without chromatin conformation data. (2) Many peaks may be non-functional: even when there is a peak at a promoter, the expression of the gene may not be affected in the matching perturbation experiment.

Keywords: ChIP-seq, transcriptional regulatory network, perturbation, pancreas, CAGE, FANTOM5

INTRODUCTION

Regulation of gene expression by combinations of transcription factors (TFs) is a fundamental process that determines cellular identity and functions. TFs have the ability to recognize and bind short sequence motifs throughout the genome, and, either alone or in combination with other TFs, modulate mRNA levels in a cell until it acquires the predetermined phenotype (Mitchell and Tjian, 1989; Wray et al., 2003). In humans it has been estimated that there are at least 411 different cell types (Vickaryous and Hall, 2006) and 1500–2000 different transcription factors (Roach et al., 2007; Vaquerizas et al., 2009; Wingender et al., 2015), with ∼430 TFs expressed at appreciable levels in any given primary cell type (Forrest et al., 2014). Identifying key cell type specific transcription factors and their targets is fundamental to understanding cellular states, and is important for regenerative medicine where efforts are made to direct differentiation of stem cells toward a medically relevant cell type (Cahan et al., 2014).

Over the years, multiple approaches to map the targets of TFs have been developed. Computational approaches that predict TF targets based upon their co-expression with a given TF and/or the presence of a transcription factor binding site motif (TFBS) in their promoter regions have helped to identify direct targets (Wasserman and Sandelin, 2004; Tompa et al., 2005; Valouev et al., 2008; FANTOM Consortium et al., 2009); however, these are purely predictive methods and the validation rate, when experimental validations are carried out, is low. Motif prediction methods are limited as the vast majority of our TFs have no well-defined TFBS, and TFs from the same family bind very similar motifs. Even for those cases where a motif is known, the information content is so low that the majority of binding site predictions will likely be false positives (Wasserman and Sandelin, 2004). Lastly, unless the expression levels of the TFs themselves are taken into consideration, inaccurate predictions can be made where a binding event may be predicted as important despite the fact that the corresponding TF is not even present in the cell.

Alternatively, TF targets can be identified experimentally. Experimental perturbation of TFs (Hilger-Eversheim et al., 2000) followed by expression profiling can identify global sets of genes affected by the given TF. This is a powerful approach, but does not discriminate direct from indirect targets (genes regulated by TFs which are regulated by the perturbed TF). Another experimental approach directly determines physical binding sites in the genome using protocols such as ChIP-CHIP, DamID or ChIP-seq (van Steensel and Henikoff, 2000; Horak et al., 2002; Robertson et al., 2007). The caveat with these methods lies in that they do not distinguish functional from non-functional binding. By combining the perturbation and physical interaction approaches we can overcome the limitations of each.

The remaining issue, however, is the scale of the problem. TF-target interactions vary between cell types as there are different combinations of transcription factors expressed and different chromatin configurations in each cell type. Thus, ultimately, what we need is a compendium of cell type specific regulatory networks for every cell type that makes up the human body. Given its scale, the problem necessitates prioritization of the cell type to be studied and the sets of TFs considered. We need ways to identify which TFs are most important to a given cell type.

Recently, the FANTOM5 project used single molecule sequencing to generate CAGE (Kanamori-Katayama et al., 2011) across a large collection of human and mouse primary cells, cell lines and tissue samples, providing a nearly comprehensive set of human and mouse, promoter and enhancer regions and their expression profiles (Andersson et al., 2014; Forrest et al., 2014). Importantly, for the prioritization of key TFs, the FANTOM5 CAGE data boasts expression profiles for 94% (1665/1762) of human TFs; this can be used to generate cell-type-specific ranked lists (expression relative to median across almost 1000 samples). What emerged from those lists is that the TFs with the most enriched expression in a given primary cell type often had phenotypes relevant to that cell type [e.g., mutations of osteoblast enriched TFs resulted in bone phenotypes, hematopoietic stem cell enriched TFs in blood phenotypes and inner ear hair cell enriched TFs in deafness (Forrest et al., 2014)]. These enriched TFs are therefore likely key components of cell-type-specific transcriptional regulatory networks (TRNs). To probe cell type enriched TFs in more detail, we explored an integrated approach for dissecting TRNs using siRNA knock-down, qRT-PCR, CAGE (Shiraki et al., 2003), and ChIP-seq (Robertson et al., 2007).

The large numbers of cells required for our systematic studies made it necessary to find an easily expandable cell line. Reviewing the FANTOM5 expression profiles, we chose an interesting cell line, TC-YIK (Ichimura et al., 1991), derived from an argyrophilic small cell carcinoma (ASCC) of the uterine cervix, which expresses insulin and showed enriched expression for dozens of pancreatic transcription factors. We show that TC-YIK cells express 75% of a set of genes previously reported as islet cell specific and 85% of a set of genes previously reported as beta cell specific. Given the difficulty in obtaining primary human beta cells for research, our results may be of interest to studying pancreatic transcriptional regulation, with the caveat that we

are only using TC-YIK as an experimentally tractable cell line model to examine the prediction of key TFs; it is a non-islet-cell insulinoma and therefore the regulatory edges inferred here may not generalize to primary islet cells.

Using newly created genome-wide datasets on TC-YIK enriched TFs, and a comparative set of non-enriched TFs, we sought to determine the importance of each factor in maintaining the TC-YIK cell state. Knock-down followed by CAGE profiling allowed us to identify, genome-wide, the set of genes affected by each TF, while integration with ChIP-seq data on the same factors allowed us to further discriminate direct from indirect TF targets. We present the results of the TC-YIK analysis and show that the combination of CAGE and ChIP-seq on key TFs is a powerful approach for studying mammalian transcriptional networks and necessary for dissection of direct and indirect edges. An overview of the datasets used, our analysis and the main findings are summarized in the workflow shown in **Figure 1**.

This work is part of the FANTOM5 project. Data download, genomic tools and co-published manuscripts have been summarized at http://fantom.gsc.riken.jp/5/.

RESULTS

The TC-YIK Cell Line Expresses Pancreatic Islet Cell Transcripts

Previously, TC-YIK cells were shown to generate neurosecretory granules and express chromogranin A (*CHGA*) and gastrin (*GAST*; Ichimura et al., 1991). A systematic review of endocrine hormones and peptides detected in TC-YIK confirmed *CHGA* and *GAST* were expressed at high levels and revealed also expression of insulin (*INS*), ghrelin (*GHRL*), and transthyretin (*TTR*; **Table 1**). All of these proteins [insulin, gastrin (*GAST*; Wang et al., 1993; Rooman et al., 2002; Téllez et al., 2011),

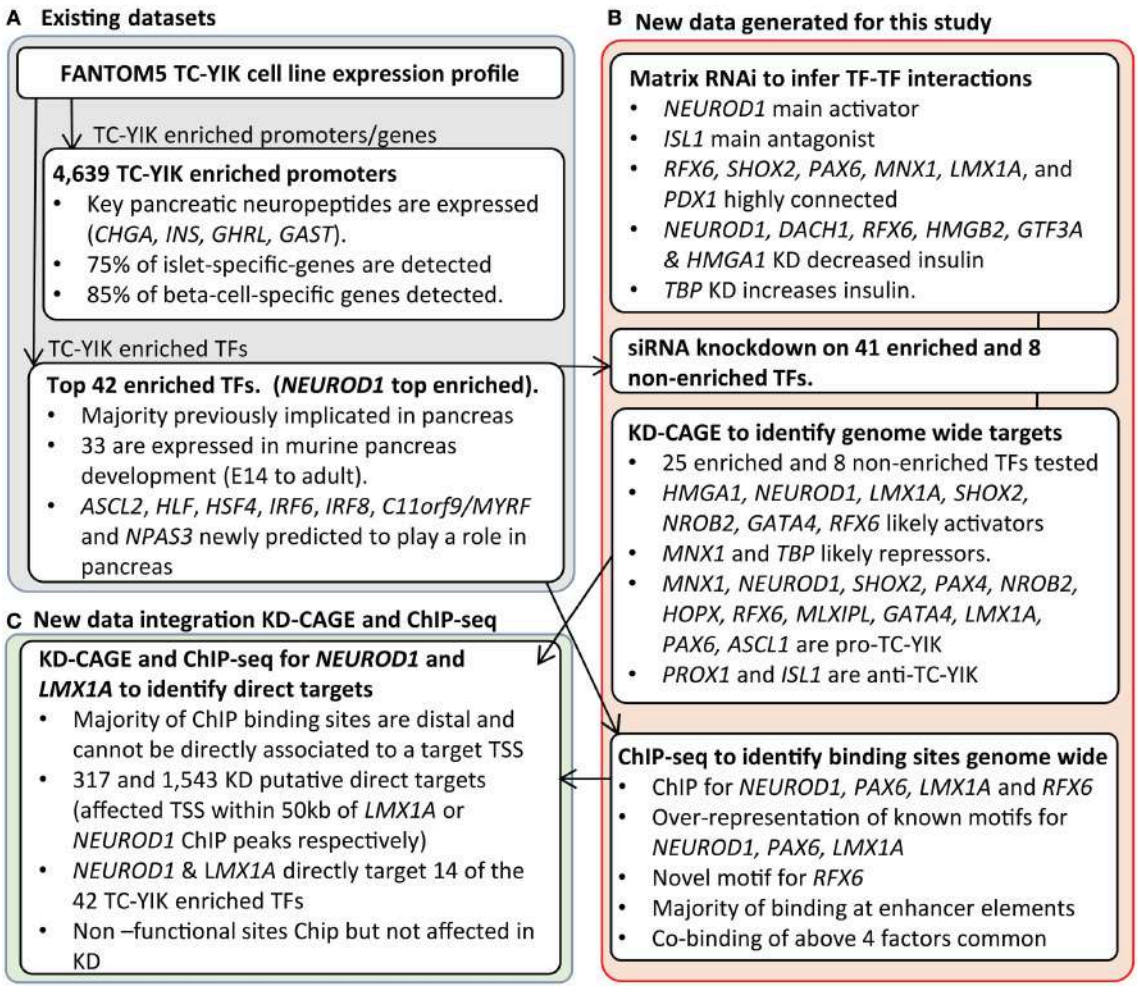


FIGURE 1 | Diagram showing the workflow of the analyses. (A) The FANTOM5 data reveal the TC-YIK cell line expresses key pancreatic neuropeptides and pancreatic transcription factors. **(B)** New data is generated for this study including siRNA perturbation of TC-YIK enriched and non-enriched TFs. siRNA perturbed samples are profiled by qRTPCR in a matrix RNAi design and by CAGE to globally identify target promoters. ChIP-seq on 4 key TFs is also carried out to identify genome wide binding sites. **(C)** KD-CAGE and ChIP-seq data are integrated to identify directly regulated targets of *NEUROD1* and *LMX1A*.

TABLE 1 | Neurosecretory peptide expression in TC-YIK.

Gene	Expression in FANTOM5 (TPM)			
	TC-YIK	Rank (out of 988 samples)	Max	Sample expressing highest level of peptide
CHGA	6062.51	1	6062.51	TC-YIK
TTR	1202.73	21	60441.3	medulla oblongata, adult
GAST	1096.66	1	1096.66	TC-YIK
INS	50.13	4	5119.98	Duodenum, fetal
GHRL	15.37	5	54.13	Eosinophils
SST	7.81	93	3612.79	Duodenum, fetal
IAPP	0	NA	26.58	Pancreas, adult
GCG	0	NA	3534.95	Gastric cancer cell line AZ521

ghrelin [(*GHRL*; Date et al., 2002; Wang et al., 2008; Arnes et al., 2012), transthyretin (*TTR*; Refai et al., 2005; Su et al., 2012), and chromogranin A (*CHGA*, a precursor of pancreatic chromostatin; Cetin et al., 1993)] play key roles in the pancreas (**Table 1**). In contrast to insulin, which is a biomarker for pancreatic beta cells, somatostatin (*SST*), glucagon (*GCG*), and islet amyloid polypeptide (*IAPP*), the biomarkers for pancreatic delta, alpha, and gamma cells, respectively, were lowly expressed or absent in TC-YIK cells. We next examined the expression of genes described in the beta cell gene atlas (Kutlu et al., 2009) as being specifically expressed in human islets. We find that 75% of the 938 human islet tissue specific genes reported by the authors are detected in TC-YIK [**Supplementary Table 1**, $\geq$ 5 tags per million (TPM)]. The authors provide a further subset of 445 genes that are enriched in alpha and/or beta cells and overlap the islet specific list (76 are expressed $>$ 2-fold higher in alpha cells and 153 are expressed $>$ 2-fold higher in beta cells). In TC-YIK, we find that 65% of these alpha cell enriched genes and 85% of the beta cell enriched genes are detected (**Supplementary Table 2**, $\geq$ 5 TPM). From this review we conclude that, although TC-YIK does not completely recapitulate the beta cell transcriptome, it shares significant similarity to islet cells. For this reason TC-YIK is sufficiently interesting for the purposes of an investigative study integrating CAGE and ChIP-seq data. Lastly, although there are rare reports of non-islet-cell insulinomas that ectopically express insulin [e.g., kidney (Ramkumar et al., 2014), liver (Furrer et al., 2001), brain (Nakamura et al., 2001)] and additional cases of argyrophilic small cell carcinoma (ASCC) of cervix (Kiang et al., 1973; Seckl et al., 1999), ours is the first report to our knowledge that identifies a non-islet-cell line (TC-YIK) where the majority of the beta cell program is active.

Pancreatic Transcription Factors are Enriched in TC-YIK cells

To identify TC-YIK-enriched-transcription factors, we ranked all 1665 human TFs according to their expression in TC-YIK cells relative to the median expression across the 988 human samples in the FANTOM5 phase 1 collection (Forrest et al., 2014). The highest ranked TF was *NEUROD1*, a factor known to be key in the differentiation of beta cells and insulin production (Itkin-Ansari et al., 2005; Guo et al., 2012). Furthermore, of the 42 most TC-YIK enriched TFs (enrichment score $>$ 1.25, $\sim$18-fold enrichment over median expression levels), 33 were previously implicated in pancreatic biology, including direct regulators of insulin (Sander and German, 1997), key factors for islet cell development (Wang et al., 2005; Guo et al., 2011), genes associated with diabetes (Foti et al., 2005) and with pancreatic endocrine tumors (Johansson et al., 2008; **Table 2**, **Supplementary Table 3**).

CAGE profiling of the mouse orthologs throughout pancreatic development (also profiled in FANTOM5) detected 33 of the 42 TFs in at least one stage with most changing expression levels over time (**Supplementary Figure 1**). This added support for a further seven of the remaining nine TFs enriched in TC-YIK (*ASCL2, HLF, HSF4, IRF6, IRF8, MYRF,* and *NPAS3*) as likely important factors in pancreatic development.

Assessing the Interconnection of Key TFs

A key question is whether the cell type enriched TFs identified in FANTOM5 are key regulators of the cellular state and whether these enriched factors are more (or less) important than housekeeping TFs that are more broadly expressed. Logic would suggest that those TFs expressed in an enriched manner are more likely to be regulated by other enriched TFs, and that their targets are also more likely to be enriched. To test our assumption, we first carried out siRNA perturbation of a set of enriched and non-enriched (but expressed) TFs in TC-YIK cells and assessed their effect on expression of enriched and non-enriched targets by qRT-PCR.

Multiple siRNAs were tested for each enriched factor and the one with the best efficiency was kept; siRNAs for 26 TFs reduced expression below 50%, a further 7 suboptimal siRNAs reduced expression to 51–77% of that of the scrambled control, while for the remaining TFs we were unable to find an efficient siRNA (**Supplementary Table 4**). An additional 8 non-enriched TFs were also perturbed below 50% (**Table 2**). After perturbation, RNA was extracted and qRT-PCR was used to measure the knock-down response in a 41 $\times$ 52 matrix of expression changes, where 41 columns represent the TFs that were perturbed and 52 rows represent the measured qRT-PCR values of target genes after perturbation (**Supplementary Table 5**). Experiments were carried out in triplicate and knock-down was assessed relative to a scrambled siRNA sequence. Of the $\sim$2000 potential (TF-target) edges tested, 551 were up- or down-regulated 1.5-fold or more [threshold as used in our previous studies (Tomaru et al., 2009)].

Looking at the number of affected targets for each TF knock-down (out degree) and the number of knock-downs that affected each TF (in degree; summarized in **Supplementary Table 6**) we identified *NEUROD1* as a key activator at the top of the hierarchy. *NEUROD1* knock-down caused down-regulation of 21 of the 52 tested targets (the most influenced being *PAX4*, followed by *GHRL, INS, GAST, CHGA, GCK, RFX6,* and *PAX6*). In an analogous way, *ISL1* was the main antagonist in the network, where its knock-down affected 11 targets, all of which were up-regulated (among those *CHGA, LMX1A, PAX4*, and *NEUROD1*). Other likely key TFs, *RFX6, SHOX2, PAX6, MNX1, LMX1A*, and *PDX1* also strongly affected several targets.

TABLE 2 | TFs enriched in TC-YIK and their putative function in pancreas.

TF_symbol	Expression TPM	Enrichment log10 (TC-YIK+1/median+1)	Insulin or pancreatic biology?	Detected in mouse developing pancreas	Experiments
TRANSCRIPTION FACTORS WITH ENRICHED EXPRESSION IN TC-YIK CELLS					
NEUROD1	593	2.77	Yes	Yes	Si, CA, CS
INSM1	519	2.72	Yes	Yes	–
PAX6	296	2.47	Yes	Yes	Si, CA, CS
NKX6-3	239	2.38	Yes	No	–
ARX	237	2.38	Yes	Yes	Si
MLXIPL	218	2.34	Yes	Yes	Si, CA
RFX6	146	2.17	Yes	Yes	Si, CA, CS
ONECUT2	151	2.14	Yes	Yes	Si, CA
PAX4	133	2.13	Yes	Yes	Si, CA
PDX1	127	2.11	Yes	Yes	Si
DACH1	269	2.05	Yes	Yes	Si, CA
ISL1	102	2.01	Yes	Yes	Si, CA, CS
FEV	94	1.98	Yes	No	Si
HOPX	168	1.95	Yes	Yes	Si, CA
FOXA2	88	1.95	Yes	Yes	Si
ST18	78	1.90	Yes	Yes	–
HNF4G	75	1.88	Yes	Yes	–
PROX1	106	1.84	Yes	Yes	Si, CA
HNF4A	69	1.84	Yes	Yes	Si
ELF3	51	1.71	Yes	Yes	Si
SHOX2	62	1.70	Yes	No	Si, CA
NPAS3	55	1.63	No	Yes	–
CDX2	41	1.63	Yes	Yes	–
HOXA10	40	1.61	Yes	No	Si
MNX1	38	1.59	Yes	Yes	Si, CA
ASCL2	34	1.54	No	Yes	–
TFAP2A	97	1.53	Yes	No	–
IRF8	31	1.51	No	Yes	Si
CASZ1	70	1.51	Yes	Yes	–
SIX3	30	1.49	No	No	Si
C11orf9/MYRF	62	1.49	No	Yes	–
MYT1	26	1.43	Yes	Yes	Si
HOXB13	26	1.43	Yes	No	Si
ASCL1	25	1.42	Yes	Yes	Si, CA
NR0B2	24	1.41	Yes	Yes	Si
LMX1A	24	1.40	Yes	No	Si, CA, CS
HSF4	27	1.33	No	Yes	–
HES6	71	1.32	Yes	Yes	–
HLF	23	1.31	No	Yes	Si
IRF6	23	1.30	No	Yes	–
DLX6	19	1.29	No	No	Si
GATA4	18	1.28	Yes	Yes	Si, CA
UBIQUITOUS TRANSCRIPTION FACTORS EXPRESSED IN TC-YIK BUT NOT ENRICHED					
ATF5	290	0.73	No	Yes	Si, CA
HMGB2	243	0.37	No	Yes	Si, CA
GTF3A	213	0.36	No	Yes	Si, CA
HMGA1	672	0.34	Yes	Yes	Si, CA

(Continued)

TABLE 2 | Continued

TBP	29	0.15	No	Yes	Si, CA
TAF9	80	0.09	No	Yes	Si, CA
TCF25	90	−0.10	No	Yes	Si, CA
TAF10	75	−0.33	No	Yes	Si, CA

*An extended version of the table is provided as **Supplementary Table 3** with references to pancreatic biology. Experiments used in this paper (Si, siRNA perturbation; CA, cap analysis of gene expression; CS, ChIP-seq). TC-YIK enriched factors that were not tested by siRNA were excluded due to oligo design or knock-down efficiency problems.*

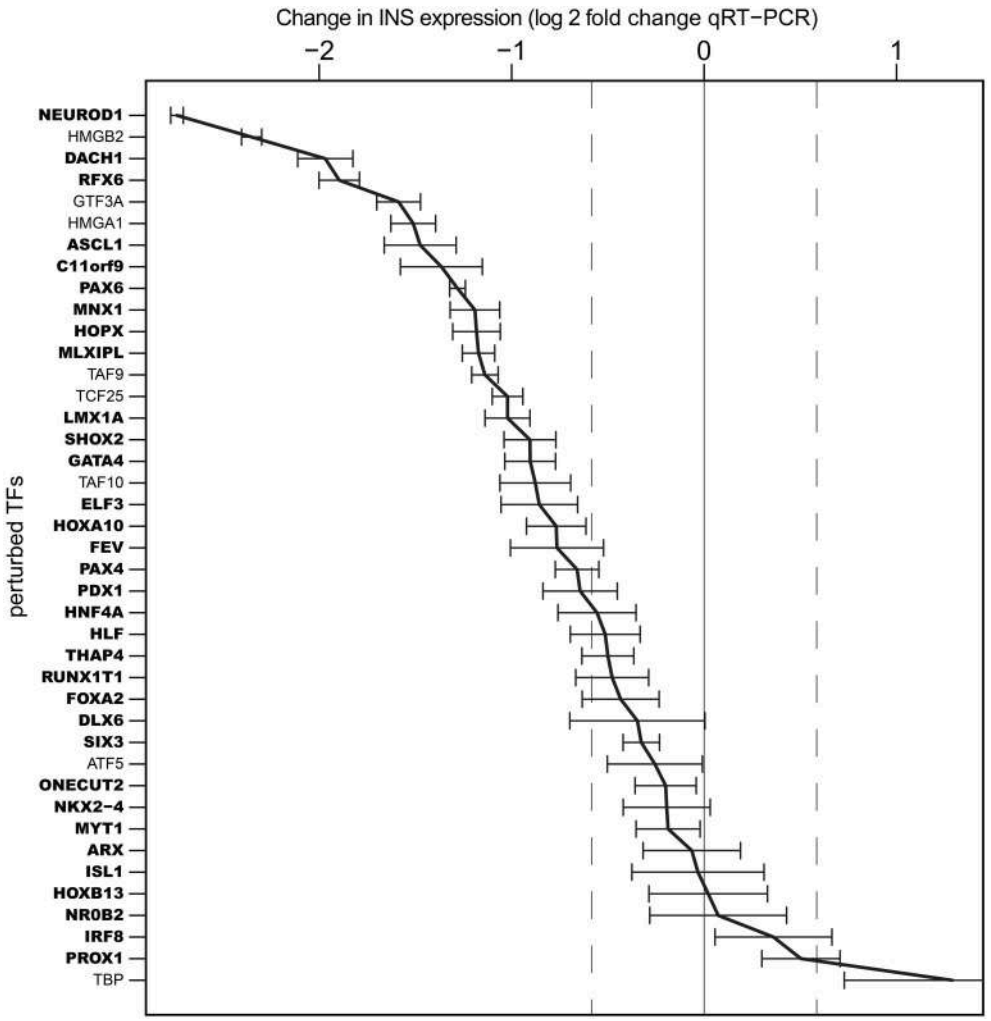


FIGURE 2 | Influence of transcription factor knock-down on *INS* expression. Log2 expression fold changes for *INS* gene upon siRNA perturbation of 41 TFs. *NEUROD1* knock-down caused the most down–regulation of insulin expression, while highest up-regulation was observed in *TBP* knock-down. Error bars indicate standard deviation over triplicate measurements. TFs in bold indicate those that were TC-YIK-enriched rather than ubiquitous.

Of note, knock-down of 28 of the 33 TFs enriched in TC-YIK and 7 of the 8 non-enriched TFs affected insulin expression levels, with the enriched factors *NEUROD1, DACH1, RFX6*, and the non-enriched TFs *HMGB2, GTF3A*, and *HMGA1* knock-down causing the greatest decreases in insulin transcript levels (**Figure 2**). Interestingly, knock-down of the non-enriched TF TATA binding protein (*TBP*) led to the highest increase in insulin transcript, which may indicate a shift in the balance between TATA dependent and TATA independent transcription.

Identifying Genome-wide TF Targets using Knock-down and Cage

The above section focused on a limited and biased set of 52 target transcripts. We next applied CAGE [KD-CAGE; (Vitezic et al., 2010)] to identify genome-wide the sets of promoters that were perturbed after knock-down of 15 of the enriched TFs and all 8 non-enriched TFs using the same RNA samples as used in the qRT-PCR. Notably the fold changes observed by CAGE and qRT-PCR were highly correlated

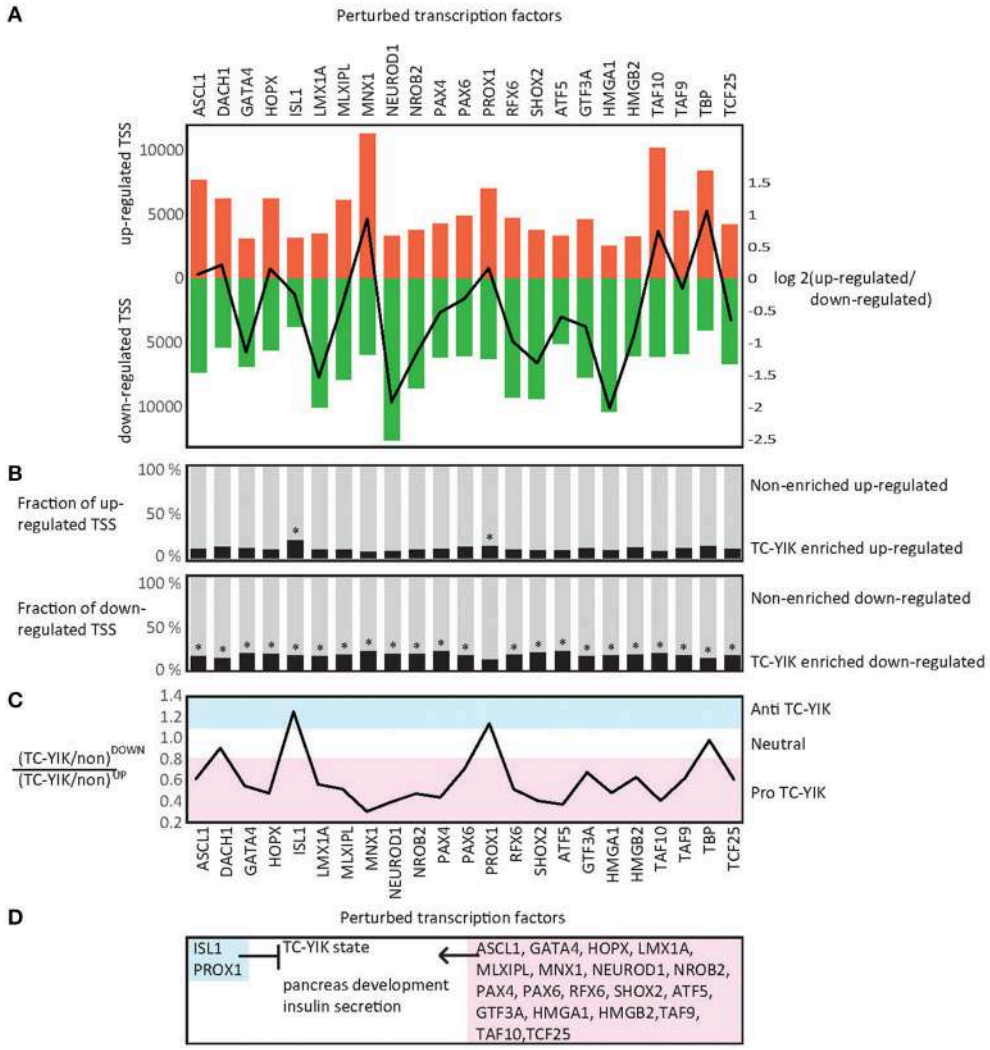


FIGURE 3 | KD-CAGE analysis. (A) Up-regulated and down-regulated TSSs in KD-CAGE experiments. Bars indicate, the numbers of up-regulated, and down-regulated TSSs detected by edgeR ($p < 0.05$) after siRNA knockdown of each factor. Line indicates the log transformed ratio of up-regulated to down-regulated TSS (e.g., note *NEUROD1* causes a much larger number of down-regulated TSS than up-regulated ones, while *MNX1* shows the reverse). **(B)** Fractions of up or down-regulated promoters that are TC-YIK-enriched or non-enriched. **(C)** Comparison of the ratios of TC-YIK-enriched to non-enriched promoters for up and down-regulated TSS sets. Note, *ISL1* and *PROX1* appear antagonistic to the TC-YIK state. **(D)** Diagram summarizing the results of the state enrichment and gene ontology enrichment analyses. *Indicates at least 15% of the up or down-regulated promoters were TC-YIK enriched.

(**Supplementary Figure 2**), indicating the suitability of CAGE for this experiment.

Promoters specifically affected by the TF knock-downs in comparison to scrambled siRNA control samples were then identified using edgeR (Robinson et al., 2010; **Supplementary Table 7**). Similar numbers of affected promoters were detected for enriched and non-enriched TFs; between 8229 and 19,467 and between 9922 and 18,362 promoters respectively (**Supplementary Table 8**). For six of the TF knock-downs (*HMGA1, NEUROD1, LMX1A, SHOX2, NROB2, GATA4, RFX6*), there were at least twice as many down-regulated promoters as up-regulated ones, suggesting that these factors work as activators. Conversely, for knock-down of *MNX1* and *TBP* we observed at least twice as many up-regulated promoters as down-regulated ones, suggesting they work as repressors (**Figure 3A**).

Identifying TFs Important for Maintaining Cell State

To understand which TFs are responsible for maintaining the TC-YIK cell state, we next identified a set of 4639 promoters with enriched expression (>3-fold) in TC-YIK compared to median expression in FANTOM5. We refer to this set as TC-YIK-enriched-promoters, and to the remainder as non-enriched-promoters. We then used these sets to separate TFs into synergists or antagonists to the cell fate: if perturbation of

a TF causes down-regulation of a significantly larger fraction of TC-YIK-enriched-promoters than non-enriched-promoters, then this would suggest that the factor in question is important for maintaining the TC-YIK state (pro-TC-YIK); similarly, if the perturbation led to up-regulation of a significantly larger fraction of TC-YIK-enriched-promoters than non-enriched-promoters, this would suggest that the factor antagonizes the TC-YIK state (anti-TC-YIK).

Starting from the assumption that TC-YIK state is maintained by regulation of TC-YIK-enriched-promoters, we checked, for each TF knock-down, whether TC-YIK-enriched-promoters were more likely to be affected (either up- or down- regulated) compared to a random event. Knock-down of all factors resulted in significantly more TC-YIK-enriched-promoters being perturbed (in either direction) than expected (hypergeometric probability test, **Supplementary Table 8**), and testing the up- and down-regulated sets separately also showed that for all perturbations significantly more TC-YIK-enriched-promoters were up-regulated and significantly more TC-YIK-enriched-promoters were down-regulated than expected by chance. This suggests that all tested TFs contribute to some extent to the maintenance of the TC-YIK state (**Supplementary Table 8**, **Figure 3B**).

Of particular note, *NEUROD1* knock-down led to down-regulation of 50% of the TC-YIK-enriched-promoters, and *ISL1* knock-down led to up-regulation of the most TC-YIK-enriched-promoters compared to the other factors, suggesting that they are pro- and anti-TC-YIK factors respectively (**Figure 3B**). To examine this in more detail we calculated the ratios of TC-YIK-enriched-promoters to non-enriched-promoters in the up-regulated sets over the down-regulated sets. High ratios correspond to anti-TC-YIK TFs and low ratios correspond to pro-TC-YIK TFs (**Figure 3C**). To compare these ratios systematically we used Chi-square with Yates correction to test for significant differences (**Supplementary Table 8**).

Using the above mentioned metric the TC-YIK-enriched factors *MNX1, NEUROD1, SHOX2, PAX4, NROB2, HOPX, RFX6, MLXIPL, GATA4, LMX1A, PAX6, ASCL1* and the non-enriched factors *ATF5, TAF10, HMGA1, TCF25, TAF9, HMGB2, GTF3A* all appear to be pro-TC-YIK (**Figure 3C**). In the case of *ISL1* and *PROX1* the ratios are shifted in the opposite direction with a higher fraction of up-regulated TC-YIK-enriched-promoters compared to non-enriched-promoters, indicating they act as antagonists to the TC-YIK state (**Figure 3C**). Interestingly, *MNX1* knock-down led to up-regulation of many non-enriched-promoters (10,483 up vs. 4426 down, ratio = 2.37), and relatively few TC-YIK-enriched-promoters (821 up vs. 1453 down, ratio = 0.57). Thus, *MNX1* is pro-TC-YIK but appears to do this by actively repressing non-enriched-promoters.

TC-YIK TFs Regulate Pancreatic Genes

Many GO terms were significantly enriched in the up- and down-regulated gene sets, including terms related to pancreatic development and function (**Supplementary Table 9**). In particular, the following down-regulated gene sets were enriched for the terms "pancreas development" (*ATF5, MNX1, NEUROD1, PAX4, RFX6, SHOX2, TAF9*), "insulin secretion" (*ATF5, GATA4, HOPX, LMX1A, MLXIPL, MNX1, NEUROD1, NROB2, PAX6, RFX6, SHOX2, TAF10, TAF9, TBP*), "cellular response to insulin stimulus" (*ATF5, GATA4, LMX1A, MLXIPL, NEUROD1, NROB2, PAX4, PAX6, RFX6, TAF9, TCF25*), "glycogen biosynthetic process" (*ATF5, HOPX, LMX1A, MNX1, NEUROD1, NROB2*), glycogen catabolic process (*GTF3A, NROB2, SHOX2*), and "glycogen metabolic process" (*HOPX, NEUROD1, NROB2*). While, for the upregulated gene lists, *ISL1* appears to be an antagonist to the pancreatic program with its knockdown leading to up-regulation of a gene set enriched for the terms "glucose homeostasis," "pancreas development," "regulation of glucose metabolic process," "insulin secretion," "endocrine pancreas development," "endocrine system development," and "peptide hormone secretion" (**Supplementary Table 9**).

In summary, it appears that both enriched and non-enriched factors contribute to the TC-YIK TRN and that, intriguingly, despite *ISL1* and *PROX1* both being enriched in TC-YIK, they seem to be antagonists to the system (**Figure 3D**).

Protein-DNA Edge Mapping by ChIP-seq of *NEUROD1*, *LMX1A*, *RFX6*, and *PAX6*

As the perturbation edges identified above could be either direct or indirect, we next used ChIP-seq data for four of the TC-YIK enriched factors to generate a paired complimentary dataset which would identify the genomic binding sites of the same factor. Integration of these two edge types (KD-CAGE and ChIP-seq) should allow us to discriminate direct from indirect edges. Biological duplicates for each factor were generated and ChIP-seq binding peaks were called relative to input chromatin using MACS (Zhang et al., 2008). We note that the number of peaks called for the same target in different biological replicates varied (NEUROD1: 7195 and 14,949 peaks, LMX1A: 7622 and 7361 peaks, PAX6: 587 and 7866 peaks, RFX6: 960 and 1659 peaks). To be conservative we only used peaks that were called as reproducible with 90% likelihood using the irreproducible discovery rate (Li et al., 2011) method (IDR $\leq$ 0.1) which yielded 144 RFX6 peaks, 190 PAX6 peaks, 4506 NEUROD1 peaks and 2166 LMX1A peaks. Scanning these peaks for known TFBS motifs using HOMER (Heinz et al., 2010) found significant enrichment for the relevant motifs (NeuroD1/Homer motif was found in 46% of NEUROD1 peaks, 7.4% of background; Lmx1a-mouse/Jaspar–9% of LMX1A peaks, 4.7% of background; PAX6/SwissRegulon–11% of PAX6 peaks, 2.2% of background, **Supplementary Figure 3**). For *RFX6* there is no known motif; however, the motifs of other *RFX* family members, and in particular *RFX5,* were enriched (37% of RFX6 peaks and 3% of background). *De-novo* motif finding on the *RFX6* ChIP-seq data identified a novel motif that is found in 58% of RFX6 peaks and 4% of background sequences. This motif closely resembles, but is different from, other RFX family motifs (**Figure 4A**).

Examining the distribution of binding in the genome, we observed that the four factors often bound in combination at the same sites, and seldom bound at promoters. For example in the *RERE* locus we observed co-binding of *NEUROD1* and *LMX1A*, and *NEUROD1* and *RFX6,* respectively, at distinct sites (see boxes in **Figure 4B**). Genome wide, co-binding of two or more of these enriched factors was common, with more than half

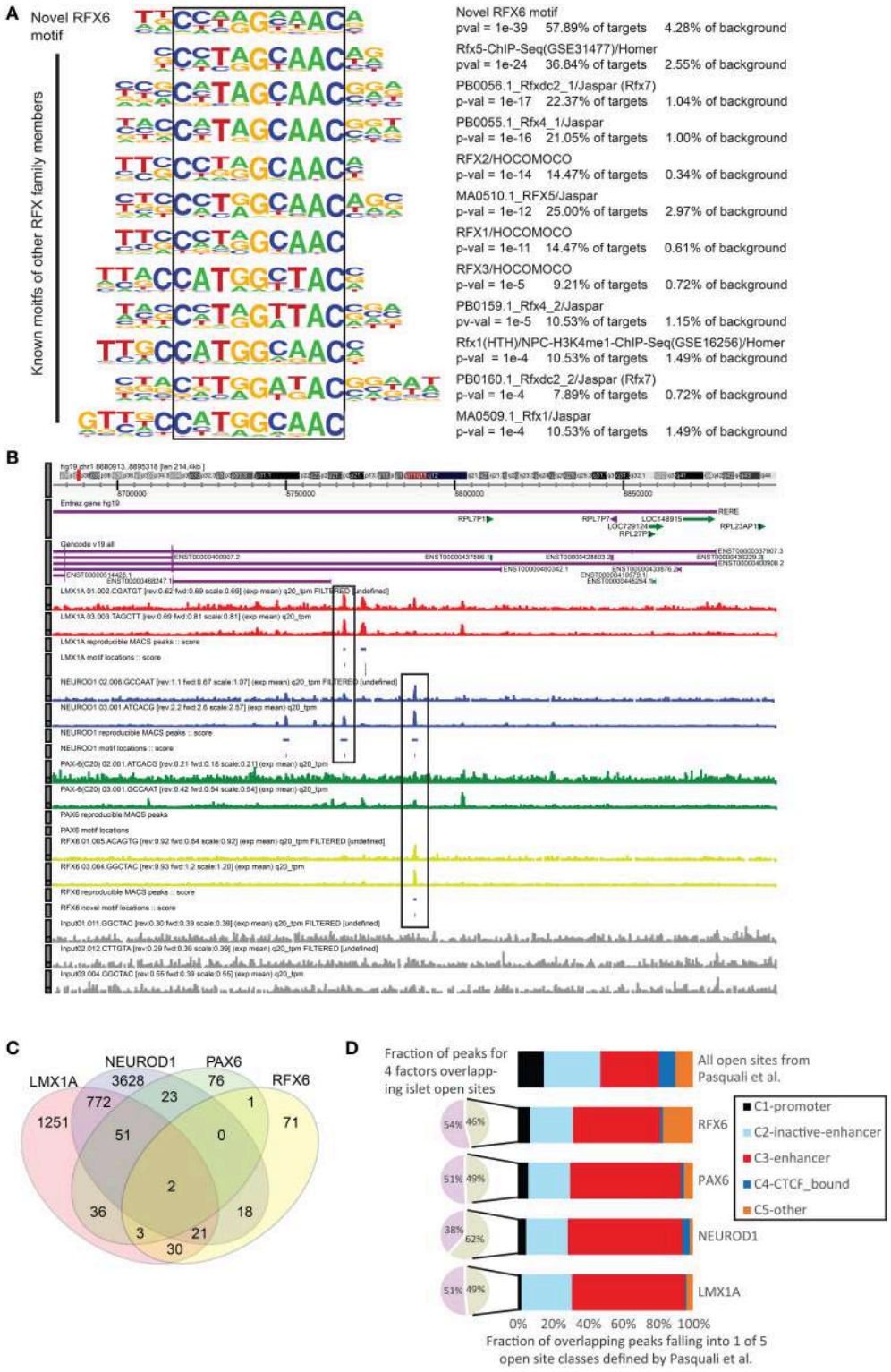


FIGURE 4 | ChIP-seq analysis of *NEUROD1*, *LMX1A*, *PAX6*, and *RFX6* in TC-YIK cells. (A) Comparison of the novel *RFX6* motif to that of other *RFX* members. Note that it is unlikely that the antibody used (S-15, Santa Cruz) would cross react with any other *RFX* family member as the antibody is raised against a peptide in the unique extended C-terminus of the protein which is not present in any of the other RFX family members. **(B)** ZENBU genome browser (Severin et al., 2014) view showing combinatorial binding of *LMX1A-NEUROD1* and *NEUROD1-RFX6* in the first intron of the *RERE* locus. Red, *LMX1A*; Blue, *NEUROD1*; Green, *PAX6*; Yellow, *RFX6*; Gray, input chromatin. **(C)** Venn diagram showing the degree of overlap between the peaks called for the four factors, numbers correspond to count of peaks overlapping by at least 1 base. **(D)** Comparison of the TF ChIP-seq peaks to open chromatin sites identified in human islet cell material by Pasquali et al. (2014).

of the *RFX6* and *PAX6* sites overlapping a *LMX1A* or *NEUROD1* site (**Figure 4C**).

Given (1) the paucity of promoter proximal binding of these factors and (2) the ample similarity between TC-YIK cellular program and endocrine program, we compared the binding sites to a map of open chromatin sites in human islet cells. Pasquali et al. (2014) integrated FAIRE-seq, and ChIP-seq of H2A.Z, H3K4me1, H3K4me3, H3K27ac, and CTCF to classify open sites in the genome of human islets as promoters (C1), poised/inactive enhancers (C2), active enhancers (C3), CTCF-bound sites (C4), and other open sites (C5). In our ChIP-seq data, we found that between 46 and 62% of peaks overlapped at least one of these open chromatin sites (this was comparable to the overlap seen by the authors for their own TF ChIP-seq experiments; 48 to 81% for *NKX2.2, PDX1, FOXA2, NKX6.1*, and *MAFB*). For those peaks overlapping the islet cell open sites, we observed enriched binding at active enhancer sites and depletion of promoter sites for all four factors (**Figure 4D**, **Supplementary Table 10**), suggesting that these factors primarily work at enhancers.

In support of this observation, both *NEUROD1* and *PAX6* have been reported previously to bind enhancer regions (Andersen et al., 1999; Aota et al., 2003; Scardigli et al., 2003; Inoue et al., 2007; Babu et al., 2008), and a recent *PAX6* ChIP-seq dataset in neuroectoderm cells identified multiple *PAX6* regulated enhancers, and reported that less than 2% of 16,000 *PAX6* peaks are near TSS of coding genes (Bhinge et al., 2014). In the case of *RFX6* there is still little known about its functional targets. Other *RFX* family members have been reported to be bound at enhancers (Reith et al., 1994; Maijgren et al., 2004; Creyghton et al., 2010; Watts et al., 2011), and in the Pasquali *et al.* study an *RFX* motif was over-represented at islet cell enhancer clusters (Pasquali et al., 2014). Intriguingly, *RFX6* had twice as many peaks overlapping class C5 than expected, suggesting that *RFX* binding may be one of the earliest events at opening of sites (Niesen et al., 2005). For *LMX1A,* ours is the first report of its involvement at enhancers.

Integration of ChIP-seq and KD-CAGE Data to Identify Direct Transcriptional Targets of TFs

By combining KD-CAGE with ChIP-seq data for *LMX1A, NEUROD1, PAX6*, and *RFX6*, we hoped to identify directly regulated promoters (that is, promoters perturbed in the knock-down experiments that also had matching nearby ChIP-seq signal). In the case of *NEUROD1* and *LMX1A*, we observed that promoters closest to a matching ChIP-seq peak were indeed affected. In particular for *NEUROD1*, almost 80% of promoters within 1 kb of a NeuroD1 ChIP-seq peak were down-regulated and for *LMX1A* almost 70% of promoters within 1 kb of an Lmx1a ChIP-seq peak were down-regulated (**Figure 5A**). Both cases indicate that these factors work primarily as transcriptional activators. As one moves further away from a ChIP-seq peak the fraction of down-regulated promoters drops, however, even at distances greater than 5 kb (up to 100 kb) from a TSS we observed a higher proportion of down-regulated TSS compared to that seen for those >100 kb away, suggesting that both factors can affect gene expression in *cis* from neighboring enhancer elements (the closer the element, the higher the probability of being affected). Repeating the analysis only using peaks with or without a TFBS motif showed no significant differences in the fractions of TSS likely to be affected. In fact, for the case of *LMX1A* and *NEUROD1* the fraction of perturbed TSS increased at shorter distances relative to a ChIP-seq peak, regardless of whether the ChIP-seq peak overlapped a motif or not (**Supplementary Table 11**). In the case of *RFX6* and *PAX6*, we observed no such distance-dependent effect, suggesting that either these factors work predominantly via distal sites or that the small number of ChIP-seq peaks observed for these two factors confounded the analysis.

Finally it is worth noting that not all proximal sites appear to be functional. For *NEUROD1* and *LMX1A* respectively, 17 and 18% of the TSSs within 1 kb of a ChIP-seq peak for the same factor were unaffected in the knock-down. An example is shown for the *EYS* locus. ChIP-seq and TFBS predictions support binding of *LMX1A* and *NEUROD1* at the *EYS* promoter, but only *NEUROD1* perturbation affected *EYS* expression levels (**Figure 5B**; other examples are shown in **Supplementary Figure 4**).

Role of *NEUROD1* and *LMX1A* in the TC-YIK TRN

Our original objective had been to integrate KD-CAGE and ChIP-seq to identify directly regulated targets (in this case of *NEUROD1*, *LMX1A*, *PAX6*, and *RFX6*). However, based on the results above, we conclude that the majority of binding events happen at enhancers, and only in the case of *NEUROD1* and *LMX1A* where we observed enrichment for perturbed TSS at shorter distances to the TSS can we infer direct promoter mediated edges. For these two factors, we considered TSS that are down-regulated at least 1.5-fold and with a ChIP-seq peak at a distance of less than 50 kb as likely direct targets. This identified 317 and 1543 directly regulated promoters for *LMX1A* and *NEUROD1* respectively (**Supplementary Table 12**). Finally, to understand the hierarchy of these factors we checked whether they directly regulate any of the other TC-YIK enriched TFs identified in the beginning of the paper. Focusing on the core network (TF-TF) we find that both NEUROD1 and LMX1A directly target 12 and 4 TC-YIK enriched TFs, respectively, but do not directly regulate each other (**Figure 5C**).

CONCLUSION

In this paper we have introduced an experimental strategy to elucidate cell type specific transcriptional regulatory networks. We start by identifying cell type enriched transcription factors (pre-computed lists for all primary cell types available online from the FANTOM web resource (Lizio et al., 2015) http://fantom.gsc.riken.jp/5/) and then use a combination of siRNA perturbation, CAGE and ChIP-seq to identify their direct and indirect targets. This strategy leverages the strengths of both approaches. Application of CAGE to siRNA perturbed samples identifies affected genes and ChIP-seq identifies directly bound targets. We show that ChIP-seq alone is insufficient

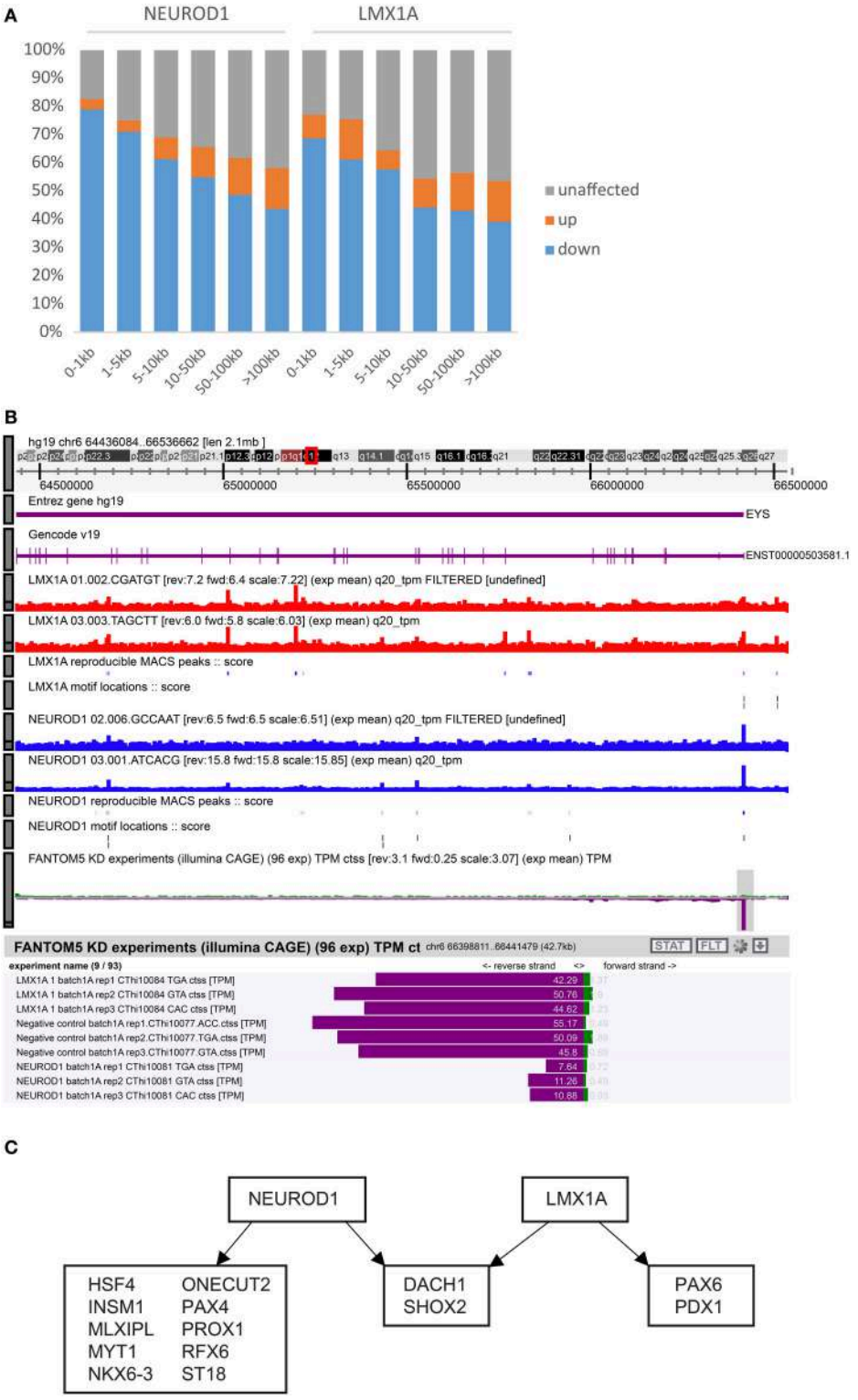


FIGURE 5 | Integration of KD-CAGE and ChIP-seq to identify direct edges. (A) Bar graph showing the fractions of up-regulated (orange), down-regulated (blue), and unaffected (gray) TSS in the knock-down of *NEUROD1* or *LMX1A*. Bars correspond to different distance bins from a ChIP-seq peak for the same factor. **(B)** Example of putative non-functional binding of *LMX1A* at the *EYS* locus. Note the presence of multiple *NEUROD1* and *LMX1A* ChIP-seq peaks and relevant motifs, but only the *NEUROD1* knock-down affected *EYS* expression (more examples shown in **Supplementary Figure 4**). **(C)** Diagram showing TC-YIK enriched transcription factors (from **Supplementary Table 4**) that are directly regulated by *NEUROD1* or *LMX1A*. To be called a direct target, we require at least one TSS of the target gene to be down-regulated 1.5-fold with a *p*-value of 0.05 and it must be within 50 kb of a ChIP-seq peak for the same factor.

to discriminate functional from non-functional bound sites, while perturbation approaches alone cannot unequivocally discriminate direct from indirect targets. It is important to precise that we are not questioning the power of ChIP methods in identifying direct and indirect *binding* (Gordan et al., 2009); the novelty of our approach lies in demonstrating that even in the presence of a TF-DNA interaction, *regulation* of target genes can happen only if the site of interaction is functional. This work highlights an important and yet undervalued matter, as in many previous publications researchers have assumed the nearest gene to, or any gene within a fixed distance of, a ChIP-seq peak, is a direct target (Shin et al., 2009; Bottomly et al., 2010; Tallack et al., 2010; Schodel et al., 2011). This is clearly an oversimplification. We have shown that almost a fifth of TSS within 1 kb of a *NEUROD1* or *LMX1A* ChIP-seq peak are unaffected in matching siRNA knock-down. This could mean that these sites are non-functional or that they are cell-context dependent (Osmanbeyoglu et al., 2012; Whitfield et al., 2012).

Aside from exploring this strategy to build TRNs, we have introduced TC-YIK as a model to study transcriptional regulation of pancreatic genes. There is a need for such cell line models, as the majority of viable post mortem islet cell material is used for transplants into diabetic patients, thus pancreatic beta cells for research are difficult to obtain. Moreover, the isolation of pure beta cell populations, the lack of protocols to expand them in culture and the number of cells required to carry out extensive perturbation and chromatin immuno-precipitation experiments are prohibitive. We have shown by CAGE profiling that 85% of the beta cell genes identified by the beta cell gene atlas (Kutlu et al., 2009) are expressed in TC-YIK and that *NEUROD1*, *LMX1A*, *PAX6,* and *RFX6* binding sites in TC-YIK are enriched at islet cell active enhancer sites. Furthermore, TC-YIK cells express key transcription factors known to be involved in pancreatic cell development and differentiation, including *NEUROD1, PDX1,* and *FOXA2* (Wang et al., 2002; Itkin-Ansari et al., 2005; Guo et al., 2012). In fact, 33 of the top 42 most TC-YIK enriched TFs are implicated in pancreatic biology. In addition, 33 homolog TFs are expressed in developing mouse pancreas. On this account, we, for the first time, find evidence of *ASCL2, HLF, HSF4, IRF6, IRF8, C11orf9/MYRF,* and *NPAS3* playing a role in pancreatic neuroendocrine gene expression and development. The only two TFs without prior references in the literature or detectable expression in the FANTOM5 mouse pancreatic samples were *SIX3* and *DLX6*, respectively. Despite this, *DLX6* expression has previously been reported in earlier pancreatic stages (E12.5 and E13.5; Gasa et al., 2004). This thorough review shows that the majority of transcription factors with enriched expression in TC-YIK have a role in pancreatic development and thus, TC-YIK is an important cell line model for studying transcriptional regulation of pancreatic gene expression.

Genome-wide expression profiling of the perturbed samples by CAGE revealed multiple insights. The majority of TF knock-downs led to more down-regulated genes than up-regulated ones, suggesting these TFs primarily work as activators, in agreement with the arguments of Hurst et al. (2014). From this logic, we predict *HMGA1, NEUROD1, LMX1A, SHOX2, NROB2, GATA4, RFX6* as likely activators and *MNX1* and *TBP* as likely repressors. Although there is the possibility that a predicted activator is in fact a repressor of an activator and a predicted repressor is an activator of a repressor, we find that both *GATA4* (Rojas et al., 2008) and *LMX1A* (Andersson et al., 2006) have direct evidence as transcriptional activators and *MNX1* (William et al., 2003) has been confirmed as a transcriptional repressor. By incorporating ChIP-seq data we can verify the roles of TFs directly. For both *NEUROD1* and *LMX1A* we show that they work as direct transcriptional activators. This clarifies the role of *NEUROD1* as a previous work reported it as both a transcriptional repressor and activator (Itkin-Ansari et al., 2005). Integration of the CAGE and ChIP-seq data clearly shows that >75% of TSS proximal to *NEUROD1* are down-regulated in *NEUROD1* knock-down (**Figure 5A**). In the previous work by Itkin-Ansari et al. the authors used perturbation (over-expression) alone and assumed *SST* down-regulation upon *NEUROD1* over-expression indicated it was a target that was directly transcriptionally repressed; we think it is more likely that *NEUROD1* indirectly antagonizes *SST* expression via other pancreatic TFs. This highlights the value of using both perturbation and ChIP-seq approaches.

In terms of what the application of our strategy to TC-YIK has told us about pancreatic gene expression, and the hierarchy of TFs, firstly we have shown that not only enriched (*MNX1, NEUROD1, SHOX2, PAX4, NROB2, HOPX, RFX6, MLXIPL, GATA4, LMX1A, PAX6, ASCL1*) but also non-enriched factors (*ATF5, TAF10, HMGA1, TCF25, TAF9, HMGB2, GTF3A*) contribute to the maintenance of the TC-YIK state. It is thus important to consider housekeeping TFs, too, when building cell-specific TRNs since they often work cooperatively with state specific factors (Ravasi et al., 2010). Our analysis also identified *ISL1* and *PROX1* as likely antagonists to the state. It may be that these antagonists help maintain a stem/progenitor like state (Wang et al., 2005; Eberhardt et al., 2006). We show that *NEUROD1* and *LMX1A* are both directly activating multiple other pancreatic TFs, and that based on our data they do not directly regulate each other (**Figure 5C**).

Finally, building cell-type-specific TRNs will require further work and integration of newer data types. In the case of *RFX6* and *PAX6* we made no predictions of their direct targets as there were few peaks bound at promoter regions and there was no enrichment for perturbed TSS near these peaks. This could be due to lower quality or less efficient antibodies used for the two factors, or could reflect lower expression levels compared to the other factors. Despite this, for all four factors (including the higher quality *NEUROD1* and *LMX1A* experiments) the majority of peaks were at putative enhancer regions. In conclusion, mammalian TRN models will need to incorporate distal regulatory elements as well, as proximal elements. To address this issue in the future we will need to use protocols such as ChIA-PET (Fullwood et al., 2009) and HiC (Dixon et al., 2012) to link distal elements with the TSS that they regulate. We believe that such chromatin conformation methods combined with KD-CAGE and ChIP-seq have the potential to identify gold standard regulatory events at both promoters and enhancers, and are key to understanding how each cell type is wired.

METHODS

Selection of Transcription Factors Significantly Enriched in TC-YIK for siRNA Knock Down

A pre-computed list of TFs with enriched expression in TC-YIK was downloaded from FANTOM5's sample browser SSTAR [direct link: http://fantom.gsc.riken.jp/5/sstar/FF:10589-108D4, see FANTOM web resource (Lizio et al., 2015)]. Enrichment is based on expression in the sample compared to the median expression across all samples in the FANTOM5 collection. The enrichment score is defined as log10[(expression in TC-YIK + 1)/(median expression in FANTOM5 + 1)]. The top 33 genes with enriched expression in TC-YIK were targeted for siRNA knock-down using stealth siRNAs from Invitrogen. As a comparison we also targeted a set of 8 non enriched TFs (*TAF9*, *TAF10*, *ATF5*, *GTF3A*, *TCF25*, *TBP*, *HMGA1*, *HMGB2*) that were expressed in TC-YIK at similar levels. In addition to these TFs, six target genes (*INS, CHGA, GHRL, GCK, GAST, TTR*) and five additional target TF genes where we were unable to find effective siRNAs (*ASCL2*, *CBFA2T2*, *CDX2*, *INSM1*, *TFAP2A*) were also added to the set. The combined set was used for systematic siRNA KD in triplicate of one factor at a time followed by qRT-PCR measurements of the perturbed genes in a Matrix RNAi design as described in Tomaru et al. (2009). siRNA sequences, knock-down efficiency and primers used in qRT-PCR are provided in **Supplementary Table 9**.

Cell Culture

TC-YIK (Ichimura et al., 1991; Human cervical cancer) cells were provided by RIKEN BRC (Cell no: RCB0443). Cells were grown in RPMI1640 (GIBCO), 10% fetal bovine serum (CCB), 1% penicillin/streptomycin (Wako). TC-YIK cells were incubated at 37°C in a humidified 5% CO2 incubator.

Genome-wide KD-CAGE

KD experiments followed by CAGE were profiled (see below) to obtain genome-wide promoter activities. Of the 41 most enriched TFs that were selected for Matrix RNAi, 15 among the most perturbed and all 8 non-enriched genes were chosen for siRNA transfection followed by CAGE. The 15 enriched TFs targeted for CAGE analysis were selected in a semi-random fashion that favored TFs that affected insulin expression in the qRT-PCR results (**Figure 2**). *NEUROD1, DACH1, RFX6, ASCL1, PAX6, MNX1, HOPX, MLXIPL, LMX1A, SHOX2, GATA4,* and *PAX4* knock-down significantly reduced *INS* transcript levels. *PROX1, NR0B2,* and *ISL1* were selected based on their reported roles in pancreatic biology as putative repressors, rather than their effect on *INS* levels. Experiments were carried out in biological triplicate, and scrambled siRNA samples were prepared as negative control. While the KD method has been previously described (Vitezic et al., 2010), we used a new variant of CAGE developed for the Illumina Hiseq 2500 called nAnT-iCAGE (Murata et al., 2014). Briefly, 5 μg of RNA was used for each sample and libraries were combined in 8-plex using different barcodes. Tags were de-multiplexed and mapped to the human genome (hg19) using BWA (Li and Durbin, 2010), yielding an average of 8.9 M mapped counts per sample (map quality > 20). Expression tables were made by counting the numbers of mapped tags falling under the 184,827 robust CAGE peaks regions identified in FANTOM5 (Forrest et al., 2014). Differential expressed promoters in TF knock-downs vs scrambled controls were identified using edgeR (Robinson et al., 2010) with a significance threshold of 0.05.

Chromatin Immunoprecipitation Assay

Chromatin was prepared and immunoprecipitation carried out as described previously (Kubosaki et al., 2009).

List of antibodies used in the ChIP-seq experiments: *LMX1A* [LMX1A (C-17), sc-54273X Santa Cruz], *NEUROD1* [Neuro D (G-20), sc-1086X Santa Cruz], *RFX6* [RFX6 (S-15), sc-169145X Santa Cruz], and *PAX6* [Anti Pax-6 (C-20), Human (Goat), sc-7750 X Santa Cruz]. Note to readers, the following antibodies were also tried but failed in ChIP-seq: [Santa Cruz: Anti ISL1 (K-20) sc-23590X; Anti PAX6 (AD2.38) sc-32766X; Anti Dlx-6 (G-20) sc-18154; Anti HB9 (H-20) sc-22542; Anti DLX6 (C-20) sc-18155; Anti PDX-1 (A-17) sc-14664 X; and Abnova: Anti ISL1 (H00003670-M05)].

All experiments were carried out as biological duplicates. Immunoprecipitated and input chromatin samples were incorporated into 4-plex ChIP-seq libraries using the NEBnext kit (New England Biolabs). Libraries were labeled with a 6 bp barcode and then pooled to be sequenced on Illumina HiSeq2000.

Sequencing results were mapped to the human genome (hg19) using BWA software (Li and Durbin, 2010) providing an average of ∼180 M mapped tags per lane (or, alternatively, ∼45 M per sample), with a mapping rate of >96%. After mapping we performed peak calling using MACS software (Zhang et al., 2008) with the recommended default parameter settings for point binding type of events [mfold=(Refai et al., 2005; Tompa et al., 2005), bandwidth=300]. We additionally used Irreproducible Discovery Rate analysis (Li et al., 2011), to identify reproducible peaks which were used for downstream analysis.

Motif Enrichment Analysis

We used HOMER software for de-novo motif discovery (Heinz et al., 2010), as well as to calculate over-representation of known motifs. Known motifs provided with HOMER (v4.6, 3-29-2014) were expanded by importing all known *NEUROD1*, *LMX1A*, *PAX6,* and *RFX* motifs from SwissRegulon (Pachkov et al., 2007), JASPAR (Bryne et al., 2008), UniPROBE (Newburger and Bulyk, 2009), and HOCOMOCO (Kulakovskiy et al., 2013), into HOMER before carrying out the scan. We used the function *findMotifsGenome.pl* to discover motifs in all reproducible peaks for each factor (genomic regions from hg19) with the option "–mask" to filter out bindings on repeats. The target sequences are the regions under the peaks and the background regions are randomly sampled sequences from the genome (Hg19) with similar GC content as the target sequences.

Gene Ontology Enrichment Analysis

The R Bioconductor GOstats package (Falcon and Gentleman, 2007) was used to obtain gene ontology enrichment scores. For the ChIP-seq GO analysis was performed on bound TSSs, while for the CAGE KD experiments, the up- and down-regulated genes were analyzed separately. For both analyses, all genes expressed in TC-YIK (>1 TPM) were used as the background.

Data Access

This work is part of the FANTOM5 project. Data download, genomic tools and co-published manuscripts have been summarized at http://fantom.gsc.riken.jp/5/. A ZENBU genome browser view displaying TC-YIK related expression data can be accessed at this URL: http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=e3YeqamiJBWhbPgPq59ubD;loc=hg19::chr14:93349815..93441266 [Reviewer username: lizio2014-review@riken.jp, password: lizio2014 (note: if problems after logging in, re-enter the URL and try again. Password will be removed at publication)]. All sequencing data used in this study has been deposited to DDBJ Read Archive (http://www.ddbj.nig.ac.jp/) with accession number DRA002420 (CAGE data) and DRA002468 (ChIP-seq data). CAGE expression profiles and enrichment of TFs for TC-YIK cell line are part of the FANTOM5 main data set. siRNA perturbations, CAGE-KD, and ChIP-seq experiments were generated separately for this study. Additional material can be found at the following URL (http://fantom.gsc.riken.jp/5/suppl/Lizio_et_al_2014/?cultureKey=&q=5/suppl/Lizio_et_al_2014 Reviewer username: m.lizio, password: m.lizio).

AUTHOR CONTRIBUTIONS

AF designed the study and wrote the manuscript; ML carried out all bioinformatics analyses and wrote the manuscript; YI carried out the siRNA perturbations, qRT-PCR and chromatin immunoprecipitation experiments with help from AK; MI provided the CAGE libraries; TL and AH mapped the CAGE data; YN provided the TC-YIK cell line; JS helped with visualization in ZENBU; HK contributed to the ChIP-seq analysis and provided the set of CAGE peaks; HS, HK, PC, YH, and AF supervised the project.

FUNDING

FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Yoshihide Hayashizaki; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to YH. We thank Michiel de Hoon for proofreading the manuscript. We would also like to thank RIKEN BRC for providing the TC-YIK cell line samples and thank GeNAS for data production. ARRF is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust and funds raised by the MACA Ride to Conquer Cancer.

ACKNOWLEDGMENTS

FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to YH; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to YH. We thank Michiel de Hoon for proofreading the manuscript. We would also like to thank RIKEN BRC for providing the TC-YIK cell line samples and thank GeNAS for data production. AF is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust and funds raised by the MACA Ride to Conquer Cancer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00331

Supplementary Figure 1 | Homolog TF genes expressed in mouse pancreas development series. CAGE expression profiles for 33 of the 42 human homolog TC-YIK-enriched TFs. Only TFs with expression above 1TPM for at least one developmental stage are shown. On the x-axis are developmental stages, from E14 until adult state. The y-axis shows expression levels (normalized TPM).

Supplementary Figure 2 | CAGE KD and qRT-PCR KD comparison. Plots for 23 transcription factors matched in both CAGE and qRT-PCR. Fold changes largely agree between technologies. Each dot represents the fold change value of a target gene among the pool of 52 perturbed genes in the matrix RNAi pilot study.

Supplementary Figure 3 | HOMER Motif scan summary. Enrichment of relevant known motif and top novel motif is shown for *NEUROD1, LMX1A, PAX6*, and *RFX6*. Expanded results are available online at (http://fantom.gsc.riken.jp/5/suppl/Lizio_et_al_2014).

Supplementary Figure 4 | ZENBU genome browser views showing integration of CAGE and ChIP-seq profiles for LMX1A and NEUROD1. (A) SYT4 and PLK4 loci have proximal binding of both factors and are affected in both of the knock-downs. **(B)** GPD2 and RSRC1 loci have proximal binding of both factors but are affected in both the knock-downs. **(C)** PROX1 and ID4 have proximal binding of both factors but only the knock-down of NEUROD1 affects expression.

Supplementary Table 1 | Human islet cell enriched transcripts. Detection of human islet cell enriched transcripts from the beta cell gene atlas (Kutlu et al., 2009) in TC-YIK.

Supplementary Table 2 | Rat alpha and beta cell enriched transcripts. Detection of human orthologs of rat alpha and beta cell enriched transcripts from the beta cell gene atlas (Kutlu et al., 2009) in TC-YIK.

Supplementary Table 3 | Extended main Table 2. TFs enriched in TC-YIK and their putative function in pancreas.

Supplementary Table 4 | siRNAs and primers used in this study.

Supplementary Table 5 | Matrix RNAi results. Pilot study of systematic knock-down and qRT-PCR expression measurements for TC-YIK enriched transcription factors.

Supplementary Table 6 | Affected targets and in/out degree. Summary of the matrix RNAi study: numbers of affected targets, in- and out-degree and effects on *INS* gene.

Supplementary Table 7 | Promoters perturbed by TF knockdown. List of promoters detected by edgeR in KD-CAGE sets (*p*-value of 0.05, 1.5FC).

Supplementary Table 8 | Summary of affected promoters in CAGE KD. Numbers of differentially expressed promoters in CAGE KD and ratios of affected TC-YIK enriched promoters.

Supplementary Table 9 | Gene ontology enrichment of perturbed genes. GO enrichment analysis for CAGE KD differentially expressed promoters (split in up- and down- regulated).

Supplementary Table 10 | Overlap with open chromatin regions. Overlap of TC-YIC ChIP-seq peaks and C1-C5 open chromatin regions as defined in Pasquali et al. (2014).

Supplementary Table 11 | ChIP-seq- CAGE integration. Relationship between distance from ChIP-seq peak and perturbation in CAGE, for peaks (all, +motif, −motif).

Supplementary Table 12 | Direct targets of NEUROD1 and LMX1A. TSS that are down-regulated 1.5-fold, *p*-value of 0.05 and within 50 kb of a ChIP-seq peak for the same factor.

REFERENCES

Andersen, F. G., Jensen, J., Heller, R. S., Petersen, H. V., Larsson, L. I., Madsen, O. D., et al. (1999). Pax6 and Pdx1 form a functional complex on the rat somatostatin gene upstream enhancer. *FEBS Lett.* 445, 315–320. doi: 10.1016/S0014-5793(99)00144-1

Andersson, E., Tryggvason, U., Deng, Q., Friling, S., Alekseenko, Z., Robert, B., et al. (2006). Identification of intrinsic determinants of midbrain dopamine neurons. *Cell* 124, 393–405. doi: 10.1016/j.cell.2005.10.037

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787

Aota, S., Nakajima, N., Sakamoto, R., Watanabe, S., Ibaraki, N., and Okazaki, K. (2003). Pax6 autoregulation mediated by direct interaction of Pax6 protein with the head surface ectoderm-specific enhancer of the mouse Pax6 gene. *Dev. Biol.* 257, 1–13. doi: 10.1016/S0012-1606(03)00058-7

Arnes, L., Hill, J. T., Gross, S., Magnuson, M. A., and Sussel, L. (2012). Ghrelin expression in the mouse pancreas defines a unique multipotent progenitor population. *PLoS ONE*7:e52026. doi: 10.1371/journal.pone.0052026

Babu, D. A., Chakrabarti, S. K., Garmey, J. C., and Mirmira, R. G. (2008). Pdx1 and BETA2/NeuroD1 participate in a transcriptional complex that mediates short-range DNA looping at the insulin gene. *J. Biol. Chem.* 283, 8164–8172. doi: 10.1074/jbc.M800336200

Bhinge, A., Poschmann, J., Namboori, S. C., Tian, X., Jia Hui Loh, S., Traczyk, A., et al. (2014). MiR-135b is a direct PAX6 target and specifies human neuroectoderm by inhibiting TGF-beta/BMP signaling. *EMBO J.* 33, 1271–1283. doi: 10.1002/embj.201387215

Bottomly, D., Kyler, S. L., McWeeney, S. K., and Yochum, G. S. (2010). Identification of {beta}-catenin binding regions in colon cancer cells using ChIP-Seq. *Nucleic Acids Res.* 38, 5735–5745. doi: 10.1093/nar/gkq363

Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., et al. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 36, D102–D16. doi: 10.1093/nar/gkm955

Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q., and Collins, J. J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903–915. doi: 10.1016/j.cell.2014.07.020

Cetin, Y., Aunis, D., Bader, M. F., Galindo, E., Jörns, A., Bargsten, G., et al. (1993). Chromostatin, a chromogranin A-derived bioactive peptide, is present in human pancreatic insulin (beta) cells. *Proc. Natl. Acad. Sci. U.S.A.* 90, 2360–2364. doi: 10.1073/pnas.90.6.2360

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

Date, Y., Nakazato, M., Hashiguchi, S., Dezaki, K., Mondal, M. S., Hosoda, H., et al. (2002). Ghrelin is present in pancreatic alpha-cells of humans and rats and stimulates insulin secretion. *Diabetes* 51, 124–129. doi: 10.2337/diabetes.51.1.124

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

Eberhardt, M., Salmon, P., von Mach, M. A., Hengstler, J. G., Brulport, M., Linscheid, P., et al. (2006). Multipotential nestin and Isl-1 positive mesenchymal stem cells isolated from human pancreatic islets. *Biochem. Biophys. Res. Commun.* 345, 1167–1176. doi: 10.1016/j.bbrc.2006.05.016

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257–258. doi: 10.1093/bioinformatics/btl567

FANTOM Consortium, Suzuki, H., Forrest, A. R., van Nimwegen, E., Daub, C. O., Balwierz, P. J., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* 41, 553–562. doi: 10.1038/ng.375

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Lassmann, T., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Foti, D., Chiefari, E., Fedele, M., Iuliano, R., Brunetti, L., Paonessa, F., et al. (2005). Lack of the architectural factor HMGA1 causes insulin resistance and diabetes in humans and mice. *Nat. Med.* 11, 765–773. doi: 10.1038/nm1254

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64. doi: 10.1038/nature08497

Furrer, J., Hättenschwiler, A., Komminoth, P., Pfammatter, T., and Wiesli, P. (2001). Carcinoid syndrome, acromegaly, and hypoglycemia due to an insulin-secreting neuroendocrine tumor of the liver. *J. Clin. Endocrinol. Metab.* 86, 2227–2230. doi: 10.1210/jcem.86.5.7461

Gasa, R., Mrejen, C., Leachman, N., Otten, M., Barnes, M., Wang, J., et al. (2004). Proendocrine genes coordinate the pancreatic islet differentiation program *in vitro*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13245–13250. doi: 10.1073/pnas.0405301101

Gordân, R., Hartemink, A. J., and Bulyk, M. L. (2009). Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.* 19, 2090–2100. doi: 10.1101/gr.094144.109

Guo, Q. S., Zhu, M. Y., Wang, L., Fan, X. J., Lu, Y. H., Wang, Z. W., et al. (2012). Combined transfection of the three transcriptional factors, PDX-1, NeuroD1, and MafA, causes differentiation of bone marrow mesenchymal stem cells into insulin-producing cells. *Exp. Diabetes Res.* 2012:672013. doi: 10.1155/2012/672013

Guo, T., Wang, W., Zhang, H., Liu, Y., Chen, P., Ma, K., et al. (2011). ISL1 promotes pancreatic islet cell proliferation. *PLoS ONE*6:e22387. doi: 10.1371/journal.pone.0022387

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

Hilger-Eversheim, K., Moser, M., Schorle, H., and Buettner, R. (2000). Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control. *Gene* 260, 1–12. doi: 10.1016/S0378-1119(00)00454-6

Horak, C. E., Mahajan, M. C., Luscombe, N. M., Gerstein, M., Weissman, S. M., and Snyder, M. (2002). GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc. Natl. Acad. Sci. U.S.A.* 99, 2924–2929. doi: 10.1073/pnas.052706999

Hurst, L. D., Sachenkova, O., Daub, C., Forrest, A. R., the FANTOM consortium, and Huminiecki, L. (2014). A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol.* 15, 413. doi: 10.1186/s13059-014-0413-3

Ichimura, H., Yamasaki, M., Tamura, I., Katsumoto, T., Sawada, M., Kurimura, O., et al. (1991). Establishment and characterization of a new cell line TC-YIK originating from argyrophil small cell carcinoma of the uterine cervix integrating HPV16 DNA. *Cancer* 67, 2327–2332.

Inoue, M., Kamachi, Y., Matsunami, H., Imada, K., Uchikawa, M., and Kondoh, H. (2007). PAX6 and SOX2-dependent regulation of the Sox2 enhancer N-3

involved in embryonic visual system development. *Genes Cells* 12, 1049–1061. doi: 10.1111/j.13652-443.2007.01114.x

Itkin-Ansari, P., Marcora, E., Geron, I., Tyrberg, B., Demeterco, C., Hao, E., et al. (2005). NeuroD1 in the endocrine pancreas: localization and dual function as an activator and repressor. *Dev. Dyn.* 233, 946–953. doi: 10.1002/dvdy.20443

Johansson, T., Lejonklou, M. H., Ekeblad, S., Stålberg, P., and Skogseid, B. (2008). Lack of nuclear expression of hairy and enhancer of split-1 (HES1) in pancreatic endocrine tumors. *Horm. Metab. Res.* 40, 354–359. doi: 10.1055/s-2008-1076695

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., et al. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* 21, 1150–1159. doi: 10.1101/gr.115469.110

Kiang, D. T., Bauer, G. E., and Kennedy, B. J. (1973). Immunoassayable insulin in carcinoma of the cervix associated with hypoglycemia. *Cancer* 31, 801–805.

Kubosaki, A., Tomaru, Y., Tagami, M., Arner, E., Miura, H., Suzuki, T., et al. (2009). Genome-wide investigation of *in vivo* EGR-1 binding sites in monocytic differentiation. *Genome Biol.* 10:R41. doi: 10.1186/gb-2009-10-4-r41

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., et al. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* 41, D195–D202. doi: 10.1093/nar/gks1089

Kutlu, B., Burdick, D., Baxter, D., Rasschaert, J., Flamez, D., Eizirik, D. L., et al. (2009). Detailed transcriptome atlas of the pancreatic beta cell. *BMC Med. Genomics* 2:3. doi: 10.1186/1755-8794-2-3

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). IDR, Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* 5, 24. doi: 10.1214/11-AOAS466

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 16, 22. doi: 10.1186/s13059-014-0560-6

Maijgren, S., Sur, I., Nilsson, M., and Toftgård, R. (2004). Involvement of RFX proteins in transcriptional activation from a Ras-responsive enhancer element. *Arch. Dermatol. Res.* 295, 482–489. doi: 10.1007/s00403-004-0456-5

Mitchell, P. J., and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* 245, 371–378. doi: 10.1126/science.2667136

Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., and Itoh, M. (2014). Detecting Expressed Genes Using CAGE. *Methods Mol. Biol.* 1164, 67–85. doi: 10.1007/978-1-4939-0805-9_7

Nakamura, T., Kishi, A., Nishio, Y., Maegawa, H., Egawa, K., Wong, N. C., et al. (2001). Insulin production in a neuroectodermal tumor that expresses islet factor-1, but not pancreatic-duodenal homeobox 1. *J. Clin. Endocrinol. Metab.* 86, 1795–1800. doi: 10.1210/jcem.86.4.7429

Newburger, D. E., and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 37, D77–D82. doi: 10.1093/nar/gkn660

Niesen, M. I., Osborne, A. R., Yang, H., Rastogi, S., Chellappan, S., Cheng, J. Q., et al. (2005). Activation of a methylated promoter mediated by a sequence-specific DNA-binding protein, RFX. *J. Biol. Chem.* 280, 38914–38922. doi: 10.1074/jbc.M504633200

Osmanbeyoglu, H. U., Hartmaier, R. J., Oesterreich, S., and Lu, X. (2012). Improving ChIP-seq peak-calling for functional co-regulator binding by integrating multiple sources of biological information. *BMC Genomics* 13(Suppl. 1), S1. doi: 10.1186/1471-2164-13-S1-S1

Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.* 35, D127–D131. doi: 10.1093/nar/gks1145

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* 46, 136–143. doi: 10.1038/ng.2870

Ramkumar, S., Dhingra, A., Jyotsna, V., Ganie, M. A., Das, C. J., Seth, A., et al. (2014). Ectopic insulin secreting neuroendocrine tumor of kidney with recurrent hypoglycemia: a diagnostic dilemma. *BMC Endocr. Disord.* 14:36. doi: 10.1186/1472-6823-14-36

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044

Refai, E., Dekki, N., Yang, S. N., Imreh, G., Cabrera, O., Yu, L., et al. (2005). Transthyretin constitutes a functional component in pancreatic beta-cell stimulus-secretion coupling. *Proc. Natl. Acad. Sci. U.S.A.* 102, 17020–17025. doi: 10.1073/pnas.0503219102

Reith, W., Ucla, C., Barras, E., Gaud, A., Durand, B., Herrero-Sanchez, C., et al. (1994). RFX1, a transactivator of hepatitis B virus enhancer I, belongs to a novel family of homodimeric and heterodimeric DNA-binding proteins. *Mol. Cell. Biol.* 14, 1230–1244. doi: 10.1128/MCB.14.2.1230

Roach, J. C., Smith, K. D., Strobe, K. L., Nissen, S. M., Haudenschild, C. D., Zhou, D., et al. (2007). Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16245–16250. doi: 10.1073/pnas.0707757104

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

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rojas, A., Kong, S. W., Agarwal, P., Gilliss, B., Pu, W. T., and Black, B. L. (2008). GATA4 is a direct transcriptional activator of cyclin D2 and Cdk4 and is required for cardiomyocyte proliferation in anterior heart field-derived myocardium. *Mol. Cell. Biol.* 28, 5420–5431. doi: 10.1128/MCB.00717-08

Rooman, I., Lardon, J., and Bouwens, L. (2002). Gastrin stimulates beta-cell neogenesis and increases islet mass from transdifferentiated but not from normal exocrine pancreas tissue. *Diabetes* 51, 686–690. doi: 10.2337/diabetes.51.3.686

Sander, M., and German, M. S. (1997). The beta cell transcription factors and development of the pancreas. *J. Mol. Med.* 75, 327–340. doi: 10.1007/s001090050118

Scardigli, R., Bäumer, N., Gruss, P., Guillemot, F., and Le Roux, I. (2003). Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6. *Development* 130, 3269–3281. doi: 10.1242/dev.00539

Schödel, J., Oikonomopoulos, S., Ragoussis, J., Pugh, C. W., Ratcliffe, P. J., and Mole, D. R. (2011). High-resolution genome-wide mapping of HIF-binding sites by ChIP-seq. *Blood* 117, e207–e217. doi: 10.1182/blood-2010-10-314427

Seckl, M. J., Mulholland, P. J., Bishop, A. E., Teale, J. D., Hales, C. N., Glaser, M., et al. (1999). Hypoglycemia due to an insulin-secreting small-cell carcinoma of the cervix. *N. Engl. J. Med.* 341, 733–736. doi: 10.1056/NEJM199909023411004

Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C. O., Hayashizaki, Y., et al. (2014). Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* 32, 217–219. doi: 10.1038/nbt.2840

Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. (2009). CEAS: cis-regulatory element annotation system. *Bioinformatics* 25, 2605–2606. doi: 10.1093/bioinformatics/btp479

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15776–15781. doi: 10.1073/pnas.2136655100

Su, Y., Jono, H., Misumi, Y., Senokuchi, T., Guo, J., Ueda, M., et al. (2012). Novel function of transthyretin in pancreatic alpha cells. *FEBS Lett.* 586, 4215–4222. doi: 10.1016/j.febslet.2012.10.025

Tallack, M. R., Whitington, T., Yuen, W. S., Wainwright, E. N., Keys, J. R., Gardiner, B. B., et al. (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* 20, 1052–1063. doi: 10.1101/gr.106575.110

Téllez, N., Joanny, G., Escoriza, J., Vilaseca, M., and Montanya, E. (2011). Gastrin treatment stimulates beta-cell regeneration and improves glucose tolerance in 95% pancreatectomized rats. *Endocrinology* 152, 2580–2588. doi: 10.1210/en.2011-0066

Tomaru, Y., Simon, C., Forrest, A. R., Miura, H., Kubosaki, A., Hayashizaki, Y., et al. (2009). Regulatory interdependence of myeloid transcription factors revealed by Matrix RNAi analysis. *Genome Biol.* 10:R121. doi: 10.1186/gb-2009-10-11-r121

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

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., et al. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* 5, 829–834. doi: 10.1038/nmeth.1246

van Steensel, B., and Henikoff, S. (2000). Identification of *in vivo* DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* 18, 424–428. doi: 10.1038/74487

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263. doi: 10.1038/nrg2538

Vickaryous, M. K., and Hall, B. K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* 81, 425–455. doi: 10.1017/S1464793106007068

Vitezic, M., Lassmann, T., Forrest, A. R., Suzuki, M., Tomaru, Y., Kawai, J., et al. (2010). Building promoter aware transcriptional regulatory networks using siRNA perturbation and deepCAGE. *Nucleic Acids Res.* 38, 8141–8148. doi: 10.1093/nar/gkq729

Wang, H., Gauthier, B. R., Hagenfeldt-Johansson, K. A., Iezzi, M., and Wollheim, C. B. (2002). Foxa2 (HNF3beta) controls multiple genes implicated in metabolism-secretion coupling of glucose-induced insulin release. *J. Biol. Chem.* 277, 17564–17570. doi: 10.1074/jbc.M111037200

Wang, J., Kilic, G., Aydin, M., Burke, Z., Oliver, G., and Sosa-Pineda, B. (2005). Prox1 activity controls pancreas morphogenesis and participates in the production of "secondary transition" pancreatic endocrine cells. *Dev. Biol.* 286, 182–194. doi: 10.1016/j.ydbio.2005.07.021

Wang, Q., Elghazi, L., Martin, S., Martins, I., Srinivasan, R. S., Geng, X., et al. (2008). Ghrelin is a novel target of Pax4 in endocrine progenitors of the pancreas and duodenum. *Dev. Dyn.* 237, 51–61. doi: 10.1002/dvdy.21379

Wang, T. C., Bonner-Weir, S., Oates, P. S., Chulak, M., Simon, B., Merlino, G. T., et al. (1993). Pancreatic gastrin stimulates islet differentiation of transforming growth factor alpha-induced ductular precursor cells. *J. Clin. Invest.* 92, 1349–1356. doi: 10.1172/JCI116708

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5, 276–287. doi: 10.1038/nrg1315

Watts, J. A., Zhang, C., Klein-Szanto, A. J., Kormish, J. D., Fu, J., Zhang, M. Q., et al. (2011). Study of FoxA pioneer factor at silent genes reveals Rfx-repressed enhancer at Cdx2 and a potential indicator of esophageal adenocarcinoma development. *PLoS Genet.* 7:e1002277. doi: 10.1371/journal.pgen.1002277

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13, R50. doi: 10.1186/1471-2164-13-S1-S1

William, C. M., Tanabe, Y., and Jessell, T. M. (2003). Regulation of motor neuron subtype identity by repressor activity of Mnx class homeodomain proteins. *Development* 130, 1523–1536. doi: 10.1242/dev.00358

Wingender, E., Schoeps, T., Haubrock, M., and Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* 43, D97–D102. doi: 10.1093/nar/gku1064

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., et al. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20, 1377–1419. doi: 10.1093/molbev/msg140

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 27 October 2015
doi: 10.3389/fgene.2015.00322

Mechanisms of mutational robustness in transcriptional regulation

Joshua L. Payne[1, 2]* *and Andreas Wagner*[1, 2, 3]

[1] Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3] The Santa Fe Institute, Santa Fe, NM, USA

Robustness is the invariance of a phenotype in the face of environmental or genetic change. The phenotypes produced by transcriptional regulatory circuits are gene expression patterns that are to some extent robust to mutations. Here we review several causes of this robustness. They include robustness of individual transcription factor binding sites, homotypic clusters of such sites, redundant enhancers, transcription factors, redundant transcription factors, and the wiring of transcriptional regulatory circuits. Such robustness can either be an adaptation by itself, a byproduct of other adaptations, or the result of biophysical principles and non-adaptive forces of genome evolution. The potential consequences of such robustness include complex regulatory network topologies that arise through neutral evolution, as well as cryptic variation, i.e., genotypic divergence without phenotypic divergence. On the longest evolutionary timescales, the robustness of transcriptional regulation has helped shape life as we know it, by facilitating evolutionary innovations that helped organisms such as flowering plants and vertebrates diversify.

Keywords: homotypic clusters, redundancy, regulatory networks, shadow enhancers, transcription factor binding sites

OPEN ACCESS

Edited by:
Ekaterina Shelest, Hans-Knoell Institute, Germany

Reviewed by:
Bartek Wilczynski, University of Warsaw, Poland
Ka-Chun Wong, The Chinese University of Hong Kong, China

***Correspondence:**
Joshua L. Payne
joshua.payne@ieu.uzh.ch

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *11 September 2015*
Accepted: *10 October 2015*
Published: *27 October 2015*

Citation:
Payne JL and Wagner A (2015) Mechanisms of mutational robustness in transcriptional regulation. Front. Genet. 6:322. doi: 10.3389/fgene.2015.00322

1. INTRODUCTION

Robustness is the invariance of a phenotype in the face of environmental or genetic change. The phenotypes of living systems exhibit robustness at multiple scales of organization, ranging from the structural properties of macromolecules (Bloom et al., 2005; Wagner, 2008) to the preferred carbon sources of entire metabolisms (Samal et al., 2010). An immense body of work has focused on elucidating the mechanisms of robustness in living systems (reviewed in de Visser et al., 2003; Kitano, 2004; Stelling et al., 2004; Wagner, 2005; Masel and Siegal, 2009). Here we highlight a subset of this work, specifically those studies that have addressed the mechanisms of mutational robustness in transcriptional regulation.

Transcriptional regulation is fundamental to the control of gene expression. It allows cells to respond to environmental signals (Ptashne and Gann, 2002), such as hormones or sugars, and it drives fundamental behavioral and developmental processes, such as mating in yeast (Tsong et al., 2006) and embryonic patterning in fruit flies (Lawrence, 1992). Transcriptional regulation is largely carried out by transcription factors (TFs), proteins that bind short DNA sequences—TF binding sites—in the promoters or enhancers of genes. Such binding may induce or repress gene expression by promoting or inhibiting the recruitment of RNA polymerase. Given the fundamental

importance of when and where genes are expressed, it is crucial that transcriptional regulation is robust to perturbation.

Genetic perturbations that may affect transcriptional regulation occur in both *cis* and in *trans*. They include point mutations in TF binding sites, which may impact transcriptional regulation by changing the affinity of a binding site for its cognate TF. They also include the insertion or deletion of large segments of DNA within promoters or enhancers, which may add or remove one or more regulatory interactions from a regulatory circuit. And they include changes to the amino acid sequence of the activation or DNA binding domains of a TF, which may alter the entire binding repertoire of the TF. Such perturbations can be deleterious, as shown by the numerous disease-associated mutations within gene regulatory regions and within genes that encode TFs (Vaquerizas et al., 2009; Maurano et al., 2012; Lee and Young, 2013).

Transcriptional regulation is not only subject to a litany of genetic insults, it is also remarkably robust to these insults (Weirauch and Hughes, 2010). Gene expression phenotypes are often insensitive to mutations in TF binding sites (Kasowski et al., 2010; Kwasnieski et al., 2012), to the turnover of regulatory control from one TF to another (Ludwig et al., 2000; Odom et al., 2007), to variation in gene expression levels (Garfield et al., 2013), and even to the rewiring of entire transcriptional regulatory circuits (Tsong et al., 2006; Isalan et al., 2008; Swanson et al., 2011). Here, we review the mechanisms that underlie this mutational robustness (**Figure 1**). Reviews of the equally important topic of robustness to environmental perturbations can be found elsewhere (Eldar et al., 2004; Alon, 2007; Macneil and Walhout, 2011; Silva-Rocha and de Lorenzo, 2010), as can primary literature on the contribution of post-transcriptional regulation to robust gene expression (McManus et al., 2014).

2. MECHANISMS OF ROBUSTNESS

2.1. Transcription Factor Binding Sites

TF binding sites are short DNA sequences (6–12 base pairs) that bind TFs to regulate gene expression. On the one hand, mutations in TF binding sites can be deleterious, as shown by their involvement in human disease (Pomerantz et al., 2009; Musunuru et al., 2010; Harismendy et al., 2011), including cancer (Khurana et al., 2013; Weinhold et al., 2014; Katainen et al., 2015; Melton et al., 2015). For instance, of 2931 disease-associated single nucleotide polymorphisms located within regulatory DNA, 93.2% fall within TF binding sites (Maurano et al., 2012). On the other hand, cross-species comparisons of regulatory regions often uncover variation in TF binding sites without obvious differences in the gene expression patterns that are driven by these sites (Ludwig et al., 2000; Odom et al., 2007). In addition, within-species variation in TF binding sites is common (Garfield et al., 2012; Spivakov et al., 2012; Arbiza et al., 2013; Khurana et al., 2013; Zheng et al., 2011), and such inter-individual differences often do not affect the expression level of target genes (Kasowski et al., 2010; Zheng et al., 2010).

The simplest cause of such mutational robustness is that individual binding sites are themselves robust to mutation. That is, they can often tolerate mutations without losing the ability to bind their cognate TFs. This results from two properties of TFs: (1) They typically bind dozens, if not hundreds of distinct DNA sequences (Sengupta et al., 2002; Berger et al., 2006; Badis et al., 2009; Wong et al., 2013) and (2) these sequences are almost always organized as large *genotype networks* in the space of all possible binding sites (Payne and Wagner, 2014). In such a genotype network, nodes represent DNA sequences that bind a particular TF and edges connect nodes if their corresponding sequences differ by a single small DNA mutation. Genotype networks confer robustness, because a mutation to any site in a TF's binding site repertoire is likely to yield another site that is also in the repertoire, thus preserving binding. Moreover, the binding affinities of neighboring sites in a genotype network are strongly correlated, indicating that a site's affinity for a TF is also robust to mutation. This is important, because mutations that affect binding affinity may impact the expression of a TF's target genes (Kasowski et al., 2010; Shultzaberger et al., 2010; Sharon et al., 2012). In addition, it is worth highlighting that the very short length of TF binding sites itself confers mutational robustness: Even though longer sites may offer greater specificity, they are also more susceptible to mutational disruption (Stewart et al., 2012).

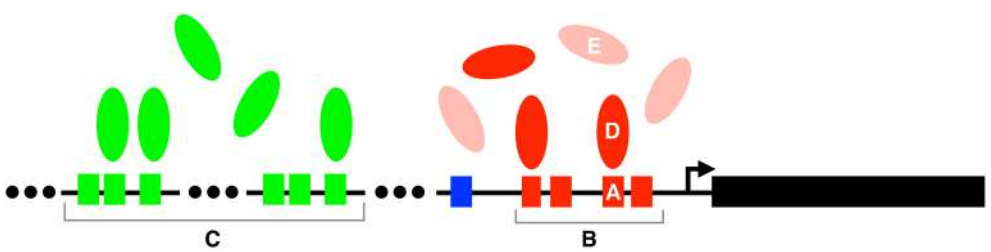


FIGURE 1 | Mechanisms of mutational robustness in transcriptional regulation. Robustness can be conferred by (A) individual transcription factor binding sites, (B) homotypic clusters of such sites, (C) redundant enhancers, (D) individual transcription factors, and (E) redundant transcription factors. Small colored boxes represent transcription factor binding sites, ellipsoids represent transcription factors, and the arrow represents the transcription start site of the gene indicated by the large black rectangle. The lightly shaded ellipses in (E) represent paralogs of the transcription factors (red ellipses) in (D). Both the red and green transcription factors regulate the expression of the black gene. These regulatory interactions are part of a larger regulatory network, whose structural properties can also influence the robustness of transcriptional regulation.

2.2. Homotypic Clusters of Transcription Factor Binding Sites

Regulatory regions often contain multiple binding sites for the same TF (Johnson et al., 1979; Giniger and Ptashne, 1988; Carey et al., 1990; Thanos and Maniatis, 1995; Wasserman and Fickett, 1998; Krivan and Wasserman, 2001; Berman et al., 2002; Ezer et al., 2014). Such *homotypic clusters* of binding sites are common in both prokaryotic and eukaryotic organisms, including bacteria (Gama-Castro et al., 2011), fruit flies (Lifanov et al., 2003), and humans (Gotea et al., 2010). For example, in humans, 62% of promoters and roughly 40% of 487 experimentally-validated developmental enhancers contain such clusters (Gotea et al., 2010). The benefits of homotypic clusters include threshold-dependent (Lebrecht et al., 2005) and graded (Giogetti et al., 2010) transcriptional responses to input signals.

An additional benefit of homotypic clusters is mutational robustness. Experiments with high-throughput promoter screens show that increasing the number of binding sites within a homotypic cluster has a saturating effect on gene expression, such that increasing the number of sites beyond a threshold results in no further impact on gene expression (Sharon et al., 2012; Smith et al., 2013). This apparent redundancy of a subset of a cluster's binding sites can provide robustness to mutation. For example, the promoter of the mouse HTF9 genes contains a homotypic cluster of binding sites for the TF Sp1, and deletion of all but one of these sites has no effect on the expression of HTF9 genes (Somma et al., 1991). Similarly, mutations in a binding site of the human TF PU.1 are less likely to impact gene expression if a second, non-mutated site is nearby (Kilpinen et al., 2013). This finding echoes earlier observations made in an analysis of polymorphic TF binding sites in *Drosophila melanogaster*, which found that sites were more likely to tolerate deleterious mutations if they were located nearby other sites for the same TF (Spivakov et al., 2012).

2.3. Redundant Enhancers

Enhancers are DNA sequences (50–1500 base pair) that bind one or more TFs to activate the transcription of genes, often in a cell-specific manner (Banerji et al., 1981; de Villiers et al., 1982; Gillies et al., 1983; Small et al., 1996; Levine et al., 2014; Shlyueva et al., 2014). Enhancers often target genes across long chromosomal distances, but typically within well-defined structural units called topologically associating domains (Dixon et al., 2012). Many genes are regulated by more than one enhancer, as exemplified by the gap genes in *Drosophila*, which control anterior-posterior patterning in the developing embryo. For example, the gap genes *hunchback*, *Kruppel*, and *knirps* are each regulated by two distinct enhancers that work together to produce bands of gene expression in the presumptive head, thorax, and abdomen (Perry et al., 2011). More generally, a genome-wide analysis of enhancer activity in *Drosophila* S2 cells found that 434 genes are regulated by at least two enhancers, and 203 of these genes are regulated by more than five enhancers (Arnold et al., 2013). For many genes, all of the gene's enhancers are necessary to drive appropriate expression. For example, both of the enhancers that regulate the gap gene *hunchback* are necessary to ensure the gene's correct expression in the developing embryo (Perry et al., 2011). In some genes, however, enhancers appear to be functionally redundant: Under normal growth conditions, only one of a gene's multiple enhancers are necessary to drive correct expression (Frankel et al., 2010; Perry et al., 2010).

Redundant enhancers—sometimes referred to as *shadow enhancers* (Hong et al., 2008)—provide not only robustness to environmental perturbations (Frankel et al., 2010; Perry et al., 2010), but also robustness to mutations. This is because deletion of one enhancer is often insufficient to disrupt normal gene expression, even if the enhancers are only partially redundant. For example, the *Drosophila* gene *snail*—a key determinant of dorsal-ventral patterning—is regulated by two enhancers, and deletion of either of these enhancers does not alter the gene's expression pattern in the presumptive mesoderm under normal growth conditions (Perry et al., 2010). Redundant enhancers can also provide robustness to mutations that affect the expression level of their cognate TFs (Frankel et al., 2010; Perry et al., 2010). For example, the two enhancers of *snail* drive a normal pattern of expression upon reduction of the expression level of Dorsal, an activator of *snail*, whereas deletion of one of these enhancers yields erratic patterns of *snail* expression in response to this genetic perturbation (Perry et al., 2010).

We note that shadow enhancers do not always provide mutational robustness. For example, the *Drosophila* gene *shavenbaby* is regulated by three primary enhancers and two shadow enhancers (Frankel et al., 2010). While the shadow enhancers are not necessary to drive the gene's epidermal expression pattern under normal growth conditions, their presence does not compensate for the inactivation of any one of the three primary enhancers (McGregor et al., 2007).

2.4. Transcription Factors

Transcription factors are also to some extent robust to mutations, including those that change the amino acid sequence of the protein's DNA binding domain. There are at least two causes of this robustness. First, amino acid substitutions in a TF's DNA binding domain may have little or no effect on the TF's binding specificity. For example, the human helix-loop-helix transcription factor Max contacts DNA at five residues, and amino acid substitutions in three of these residues have no effect on binding specificity (Maerkl and Quake, 2009). Second, transcription factors often bind DNA cooperatively, and the presence of cofactors may ameliorate the effects of amino acid substitutions that impair binding specificity. For example, the binding specificity of Matα1, a regulator of cell-type specification in ascomycete fungi, has diverged so extensively among *S. cerevisiae* and *C. albicans* that the sequences recognized by these proteins appear unrelated by bioinformatic criteria (Baker et al., 2011). Nonetheless, Matα1 controls the same set of core genes in these two species, because its recognition sequences evolved along with it. This was most likely facilitated by a protein-protein interaction with Mcm1, which is conserved among *S. cerevisiae* and *C. albicans*, and may have helped stabilize Matα1 while its interaction with DNA slowly changed.

Despite these examples, it should be emphasized that mutations in a transcription factor's DNA binding domain often do affect binding specificity and that cofactors cannot

always compensate for such changes. Because transcription factors typically regulate the expression of multiple genes, such mutations are often deleterious. This is demonstrated both by the common implication of such mutations in disease (Lee and Young, 2013) and by the high level of conservation of one-to-one transcription factor orthologs across highly diverse species (Nitta et al., 2015).

2.5. Redundant Transcription Factors

Gene duplication, which creates paralogous genes within the same genome, is a driving force in evolution. In eukaryotes, for instance, gene duplicates are estimated to arise at a rate of 0.01 per gene per million years (Lynch and Conery, 2000), and between 30 and 65 percent of a typical eukaryote's genes have paralogs (Zhang, 2003). Because gene duplicates are often functionally redundant at their time of origin, it is possible that they play compensatory roles, acting as a backup if one of the paralogs is functionally compromised. This possibility has led to a large body of research on redundant genes as a source of mutational robustness (e.g., Conant and Wagner, 2003; Gu et al., 2003).

Gene duplication has played an important role in the evolution of transcriptional regulatory systems. For example, an estimated 68% of TFs in yeast (Teichmann and Babu, 2003) and 73% of TFs in *Escherichia coli* (Madan Babu and Teichmann, 2003) are the result of gene duplication. Many of these paralogous transcription factors appear fully or partially redundant in function, because they recognize the same sets of binding sites *in vitro* (Weirauch et al., 2014) and bind to some of the same genomic regions *in vivo*. For example, genome-wide binding profiles of three ETS TFs in human T cells revealed that nearly 10% of 17,000 promoters bound more than two of the three TFs, and probably at the same binding site (Hollenhorst et al., 2007). A broader view of redundant TFs is provided by enhanced yeast one-hybrid assays (Reece-Hoyes et al., 2011), which have facilitated a test of nearly 400,000 putative binding events among 1086 human TFs and 360 enhancers (Fuxman Bass et al., 2015). This analysis found that human enhancers often bind multiple TFs that typically belong to the same TF family. Moreover, the greater the number of enhancers that a pair of TFs shares, the more likely it is that these factors are coexpressed, and the less likely it is that each factor is essential for viability (Fuxman Bass et al., 2015), providing additional support for their compensatory roles. Indeed, even distant paralogs may compensate for one another, at least in part (Kafri et al., 2005; He and Zhang, 2006; Tischler et al., 2006).

2.6. Global Topological Properties of Transcriptional Regulatory Networks

The transcriptional regulatory networks of organisms as different as bacteria and humans exhibit strikingly similar structural properties, including a heavy-tailed degree distribution, a modular organization, and non-random assortativity (Barabási and Oltvai, 2004; Boyle et al., 2014; Sorrells and Johnson, 2015). Each of these properties may confer mutational robustness in transcriptional regulation.

Many biological networks, including transcriptional regulatory networks, exhibit a heavy-tailed degree distribution (Aldana et al., 2007). Such networks are characterized by a preponderance of nodes with few connections and a small number of nodes with many connections. This topological property can endow a network with robustness to random gene deletion, because such deletions are more likely to affect low-degree nodes than high-degree nodes, and are therefore unlikely to disrupt the structure of a network (Albert et al., 2000). Simulations of model regulatory networks with heavy-tailed degree distributions show that such networks exhibit stable dynamical behavior over a broader range of parameter values than networks with a homogeneous degree distribution (Aldana and Cluzel, 2003). They are also more robust to both gene duplication (Aldana et al., 2007) and edge rewiring (Greenbury et al., 2010).

Transcriptional regulatory networks are modular. They can be decomposed into subnetworks of genes that are coregulated in response to different conditions and that are involved in distinct functions (Ihmels et al., 2002; Segal et al., 2003; Peter and Davidson, 2009). For example, an analysis of gene expression data in yeast uncovered 85 partially overlapping modules that participate in distinct cellular processes, including sporulation and rRNA processing (Ihmels et al., 2002). Similarly, the regulatory network controlling embryogenesis in the sea urchin has been decomposed into several modules that each perform distinct functions in patterning the pre-gastrular embryo, such as restricting gene expression to specific subdomains (Peter and Davidson, 2009). Such modularity may serve to contain damage, limiting the propagation of a mutation's effects to those genes that are also part of the module. For example, the yeast TF Ypl230w drives the expression of a module of hundreds of genes during entry to stationary phase. Analysis of differential gene expression upon deletion of Ypl230w found that differentially expressed genes were enriched within the module, indicating that the effect of the perturbation was largely contained (Segal et al., 2003). Similar observations have been made in simulations of model regulatory networks (Poblanno-Balp and Gershenson, 2011). It is therefore conceivable that modularity confers mutational robustness (Wagner et al., 2007), although in the context of transcriptional regulation, we currently have very little empirical evidence to support this possibility.

Assortativity is the propensity of nodes in a network to connect to other nodes with similar properties (Newman, 2002). For instance, in a network that is assortative with respect to the number of neighbors that a node (TF) has, nodes with many neighbors tend to connect to other nodes with many neighbors, and nodes with few neighbors tend to connect to nodes with few neighbors. Simulations of model transcriptional regulatory networks suggest that degree assortativity can confer robustness to mutations in regulatory regions (Pechenick et al., 2012) and to gene duplications (Pechenick et al., 2013). The transcriptional regulatory networks of 41 distinct human cell and tissue types exhibit such an assortativity signature (Pechenick et al., 2014), raising the possibility that this structural property confers robustness to transcriptional regulation in humans.

3. ORIGINS OF ROBUSTNESS

There are at least three possible origins of mutational robustness (de Visser et al., 2003): (1) Mutational robustness may itself

be an adaptation to mutations, i.e., it may exist because it provides a selective advantage; (2) It may be a byproduct of other adaptations, such as environmental robustness; or, (3) It may be neither a direct adaptation nor an indirect by-product of an adaptation, and thus a non-adaptive result of biophysical principles or non-adaptive evolutionary forces.

The first, adaptive view can be traced to at least the early 1990s, when genetic studies first showed that many genes, including genes encoding TFs, are duplicated (Thomas, 1993). This observation raised the question whether such gene redundancy exists to protect genes against otherwise deleterious mutations, and lead to modeling work addressing this question (Clark, 1994; Nowak et al., 1997; Wagner, 1999, 2000; Lynch et al., 2001; O'Hely, 2006). Such models apply in principle not only to redundant genes, but also to binding site clusters with redundant sites and to redundant enhancers.

Redundancy is not the only route to adaptive robustness. In the context of transcriptional regulation, this became clear once it became possible to analyze the structure of genotype spaces of model transcriptional regulatory circuits. In such spaces, one finds that circuits with a given gene expression pattern usually form large and connected genotype networks, where differences between neighboring genotypes (circuits) can be caused by small genetic changes, such as alterations of single regulatory interactions (Ciliberti et al., 2007; Cotterell and Sharpe, 2010; Payne et al., 2014). Individual circuits in such a network can change their regulatory interactions without changing their expression pattern. Because these circuits also vary considerably in their mutational robustness, they can evolve increased robustness via a series of small mutations that maintain their expression phenotype. Empirical data on TF binding sites demonstrate that such sites show a similar organization in the space of DNA sequences (Payne and Wagner, 2014). In consequence, their mutational robustness could in principle increase through gradual genetic change (e.g., point mutations) that preserve transcription factor binding.

Despite these observations, robustness is unlikely to confer a sufficiently strong advantage in a binding site, regulatory circuit, or a redundant regulatory element to be maintained by natural selection in most evolving populations. The reason is that its selective advantage is small, i.e., on the order of the mutation rate μ, because selection of increased robustness is effective only when a population of organisms (binding sites, circuits, etc.) are polymorphic for robustness. Elementary population genetics dictates that this will be the case only when the product of the effective population size N and the mutation rate μ is much greater than one ($N\mu \gg 1$) (van Nimwegen et al., 1999; Wagner, 2000). Especially for small mutational targets, this requires huge population sizes and very large mutation rates. Therefore, although robustness may sometimes be an adaptation, this is likely the exception rather than the rule.

Mutational robustness may also arise as a byproduct of selection for other traits, most notably robustness to environmental change (Wagner, 1997; Meiklejohn and Hartl, 2002). This is particularly relevant for transcriptional regulation, which is frought with noise, including stochastic fluctuations in signaling molecules and variable temperatures (Macneil and Walhout, 2011). Such noise can be viewed as incessant change in the molecular environment where transcriptional regulation operates. Shadow enhancers provide a useful example. As we mentioned in Section 2.3, the regulatory region of the *Drosophila* gene *snail* comprises two enhancers. Either of them is sufficient to drive wild-type gene expression patterns under normal growth conditions (Perry et al., 2010), which provides a source of mutational robustness. Under extreme temperatures, however, deletion of either of the enhancers results in aberant gene expression patterns, suggesting that the primary function of the shadow enhancer is to provide robustness to the destabilizing effects of sub-optimal temperatures, as is also the case for the two shadow enhancers associated with the *Drosophila* gene *shavenbaby* (Frankel et al., 2010). Additional support for the origin of mutational robustness as a byproduct of environmental robustness is found in model transcriptional regulatory circuits, which exhibit a positive correlation between mutational and environmental robustness (Ciliberti et al., 2007), such that selection for environmental robustness facilitates mutational robustness.

Finally, mutational robustness may also be a consequence of biophysical principles underlying transcriptional regulation, or of non-adaptive forces of genome evolution, i.e., genetic drift, mutation, and recombination.

For example, homotypic clusters of TF binding sites may evolve simply because there are more ways to build a regulatory region using many low-affinity sites than there are with few high-affinity sites (He et al., 2012). The reason is that there are many more distinct DNA sequences that bind TFs with low affinity than with high affinity (Badis et al., 2009). In addition, such clusters could simply result from the inefficiency of selection at removing insertions, such that insertions containing TF binding sites accumulate over time (Lynch, 2007), or they may be a byproduct of recombination within regulatory regions (Lynch, 2007; Paixao and Azevedo, 2010). Moreover, the spatial organization of homotypic clusters may reflect a mutational bias toward deletions, as such mutations are more likely to bring different sites closer together than farther apart (Lusk and Eisen, 2000).

Similarly, robustness-conferring topological properties, such as heavy-tailed degree distributions, can originate as a by-product of biophysical principles. For example, a biophysical model of protein-protein interactions shows that this distribution can emerge if the number of surface-exposed hydrophobic amino acids on a protein follows a simple random distribution (Deeds et al., 2006). In addition, evolutionary forces other than natural selection can enhance the robustness of regulatory networks. For instance, heavy-tailed degree distributions (Lynch, 2007), a modular organization (Wagner et al., 2007), and the enrichment of particular circuit motifs (Artzy-Randrup et al., 2004; Cordero and Hogeweg, 2006; Sorrells and Johnson, 2015) can all emerge through random genetic drift.

4. CONSEQUENCES OF ROBUSTNESS

Mutational robustness in transcriptional regulation has several consequences that emerge on evolutionary timescales. First,

the mutational robustness of regulatory regions permits their evolutionary divergence without a corresponding divergence in the gene expression patterns they control. This phenomenon is often observed among closely-related species (Weirauch and Hughes, 2010). During such divergence, substantial binding site turnover may occur, such that different sets of TFs may regulate orthologous genes in different species (Moses et al., 2006; Borneman et al., 2007; Schmidt et al., 2010). Binding site turnover can even occur among activating and repressing TFs and can alter the architecture of a regulatory circuit, all without altering its gene expression phenotype (Tanay et al., 2005; Tsong et al., 2006; Swanson et al., 2011). A well-known practical consequence of this divergence is that regulatory regions are exceptionally difficult to align.

A related consequence of mutational robustness is that regulatory regions can accumulate genetic diversity within a population. Such diversity is often referred to as cryptic, because it does not generate phenotypic variation (Gibson and Dworkin, 2004; McGuigan and Sgro, 2009). However, cryptic diversity may generate phenotypic variation upon environmental or genetic perturbation (Rutherford and Lindquist, 1998; Queitsch et al., 2002). Cryptic diversity is commonly observed in DNA sequences regulating transcription (Rockman and Wray, 2002), including TF binding sites (Balhoff and Wray, 2005; Kasowski et al., 2010; Spivakov et al., 2012; Arbiza et al., 2013). Computational models of transcriptional regulatory circuits hint that such diversity may generate phenotypic variation in response to genetic or environmental perturbations (Siegal and Bergman, 2002; Bergman and Siegal, 2003). However, we currently have no experimental evidence that standing cryptic diversity in gene regulatory regions contributes to adaptation in transcriptional regulation.

Yet another consequence of mutational robustness is that it permits regulatory interactions to originate that do not contribute to gene regulation at the time of their origin. Over time, the accumulation of such non-functional interactions can give rise to dense, highly-interconnected transcriptional regulatory networks (Sorrells and Johnson, 2015). This is especially true if binding sites are short, regulatory regions are long, and TF binding specificities are low. Evidence exists that each of these conditions are met, especially in eukaryotes, where binding sites are on average merely ten nucleotides long (Stewart et al., 2012), regulatory regions comprise promoters and enhancers that span thousands of nucleotides (The ENCODE Project Consortium, 2012), and the average information content per nucleotide of binding sites is roughly 65% of the maximum, indicating modest specificity (Stewart et al., 2012). Taken together with evidence that synthetically-added regulatory interactions rarely impact phenotype (Isalan et al., 2008), these observations suggest that mutational robustness may contribute to the apparent complexity of transcriptional regulatory networks. What is more, non-functional regulatory interactions may form the substrate of subsequent adaptations (Isalan et al., 2008), implicating mutational robustness in the evolution of novel transcriptional regulatory programs.

A final consequence of robustness emerges from the duplication of transcription factor genes. By providing a back-up gene for any one essential molecular function, gene duplication facilitates the evolution of genes with novel functions (Ohno, 1970; Hahn, 2009; Innan and Kondrashov, 2010; Rensing, 2014), such as TFs with altered binding site repertoires that can take on novel regulatory roles (Pérez et al., 2014). Over long evolutionary time scales, this ability can have profound consequences. For example, gene and genome duplications that created novel homeobox TF genes have been implicated in the diversification of the vertebrate body plan (Carroll et al., 2001), and duplication of genes encoding MADS box TFs has played an important role in the diversification of flowering plants (De Bodt et al., 2003; Irish, 2003). In other words, robust transcriptional regulation has helped shape life as we know it.

AUTHOR CONTRIBUTIONS

JP and AW conceived of and wrote the paper.

FUNDING

JP acknowledges support through the Ambizione program of the Swiss National Science Foundation. AW acknowledges support through Swiss National Science Foundation grant 31003A_146137, as well as through the University Priority Research Program in Evolutionary Biology at the University of Zurich.

REFERENCES

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382. doi: 10.1038/35019019

Aldana, M., Balleza, E., Kauffman, S., and Resendiz, O. (2007). Robustness and evolvability in genetic regulatory networks. *J. Theor. Biol.* 245, 433–448. doi: 10.1016/j.jtbi.2006.10.027

Aldana, M., and Cluzel, P. (2003). A natural class of robust networks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8710–8714. doi: 10.1073/pnas.1536783100

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461. doi: 10.1038/nrg2102

Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., et al. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* 45, 723–729. doi: 10.1038/ng.2658

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, L. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077. doi: 10.1126/science.1232542

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on "network motifs: simple building blocks of complex networks" and "superfamilies of evolved and designed networks." *Science* 305, 1107c. doi: 10.1126/science.1099334

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723. doi: 10.1126/science.1162327

Baker, C. R., Tuch, B. B., and Johnson, A. D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7493–7498. doi: 10.1073/pnas.1019177108

Balhoff, J. P., and Wray, G. A. (2005). Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8591–8596. doi: 10.1073/pnas.0409638102

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 199–308. doi: 10.1016/0092-8674(81)90413-X

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W. III., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* 24, 1429–1435. doi: 10.1038/nbt1246

Bergman, A., and Siegal, M. L. (2003). Evolutionary capacitance as a general feature of complex gene networks. *Nature* 424, 549–552. doi: 10.1038/nature01765

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *drosophila* genome. *Proc. Natl. Acad. Sci. U.S.A.* 99, 757–762. doi: 10.1073/pnas.231608898

Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., and Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. U.S.A.* 102, 606–611. doi: 10.1073/pnas.0406744102

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Serringhaus, M. R., et al. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* 317, 815–819. doi: 10.1126/science.1140748

Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., et al. (2014). Comparative analysis of regulatory information and circuits across distant species. *Nature* 512, 453–456. doi: 10.1038/nature13668

Carey, M., Lin, Y.-S., Green, M. R., and Ptashne, M. (1990). A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. *Nature* 345, 361–364. doi: 10.1038/345361a0

Carroll, S. B., Grenier, J. K., and Weatherbee, S. D. (2001). *From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design*. Malden, MA: Blackwell.

Ciliberti, S., Martin, O. C., and Wagner, A. (2007). Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput. Biol.* 3:e15. doi: 10.1371/journal.pcbi.0030015

Clark, A. G. (1994). Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. U.S.A.* 91, 2950–2954. doi: 10.1073/pnas.91.8.2950

Conant, G. C., and Wagner, A. (2003). Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. R. Soc. Lond. B*, 271, 89–96. doi: 10.1098/rspb.2003.2560

Cordero, O. X., and Hogeweg, P. (2006). Feed-forward loop circuits as a side effect of genome evolution. *Mol. Biol. Evol.* 23, 1931–1936. doi: 10.1093/molbev/msl060

Cotterell, J., and Sharpe, J. (2010). An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol. Syst. Biol.* 6, 425. doi: 10.1038/msb.2010.74

De Bodt, S., Raes, J., Van de Peer, Y., and Theissen, G. (2003). And then there were many: MADS goes genomic. *Trends Plant Sci.* 8, 475–483. doi: 10.1016/j.tplants.2003.09.006

de Villiers, J., Olson, L., Tyndall, C., and Schaffner, W. (1982). Transcriptional 'enhancers' from sv40 and polyoma virus show a cell type preference. *Nucleic Acids Res.* 10, 7965–7976. doi: 10.1093/nar/10.24.7965

de Visser, J. A. G. M., Hermisson, J., Wagner, G. P., Ancel Meyers, L., Bagheri-Chaichian, H., Blankchard, J. L., et al. (2003). Evolution and detection of genetic robustness. *Evolution* 57, 1959–1972. doi: 10.1111/j.0014-3820.2003.tb00377.x

Deeds, E. J., Ashenberg, O., and Shakhnovich, E. I. (2006). A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 311–316. doi: 10.1073/pnas.0509715102

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

Eldar, A., Shilo, B.-Z., and Barkai, N. (2004). Elucidating mechanisms underlying robustness of morphogen gradients. *Curr. Opin. Genet. Dev.* 14, 435–439. doi: 10.1016/j.gde.2004.06.009

Ezer, D., Zabet, N. R., and Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: a model system for understanding the physical mechanics of gene expression. *Comput. Struct. Biotechnol. J.* 10, 63–69. doi: 10.1016/j.csbj.2014.07.005

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* 466, 490–493. doi: 10.1038/nature09158

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

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz, R. L., Solano-Lira, H., et al. (2011). RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated wihin genetic sensory response units (gensor units). *Nucleic Acids Res.* 39, D98–D105. doi: 10.1093/nar/gkq1110

Garfield, D., Haygood, R., Nielsen, W. J., and Wray, G. A. (2012). Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin *Stronglyocentrotus purpuratus*. *Evol. Dev.* 14, 152–167. doi: 10.1111/j.1525-142X.2012.00532.x

Garfield, D. D., Runcie, D. E., Babbitt, C. C., Haygood, R., Nielsen, W. J., and Wray, G. A. (2013). The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network. *PLoS Biol.* 11:e1001696. doi: 10.1371/journal.pbio.1001696

Gibson, G., and Dworkin, I. (2004). Uncovering cryptic genetic variation. *Nat. Rev. Genet.* 5, 681–690. doi: 10.1038/nrg1426

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* 33, 717–728. doi: 10.1016/0092-8674(83)90014-4

Giniger, E., and Ptashne, M. (1988). Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc. Natl. Acad. Sci. U.S.A.* 85, 382–386. doi: 10.1073/pnas.85.2.382

Giogetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., et al. (2010). Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded trancriptional responses to environmental inputs. *Mol. Cell* 37, 418–428. doi: 10.1016/j.molcel.2010.01.016

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* 20, 565–577. doi: 10.1101/gr.104471.109

Greenbury, S. F., Johnson, I. G., Smith, M. A., Doye, J. P. K., and Louis, A. A. (2010). The effect of scale-free topology on the robustness and evolvability of genetic regulatory networks. *J. Theor. Biol.* 267, 48–61. doi: 10.1016/j.jtbi.2010.08.006

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66. doi: 10.1038/nature01198

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* 100, 605–617. doi: 10.1093/jhered/esp047

Harismendy, O., Notani, D., Song, X., Rahim, N. G., Tanasa, B., Heintzman, N., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* 470, 264–268. doi: 10.1038/nature09753

He, X., Duque, T. S. P. C., and Sinha, S. (2012). Evolutionary origins of transcription factor binding sites. *Mol. Biol. Evol.* 29, 1059–1070. doi: 10.1093/molbev/msr277

He, X., and Zhang, J. (2006). Transcriptional reprogramming and backup between duplicate genes: is it a genomewide phenomenon? *Genetics* 172, 1363–1367. doi: 10.1534/genetics.105.049890

Hollenhorst, P. C., Shah, A. A., Hopkins, C., and Graves, B. J. (2007). Genome-wide analysis reveal properties of redundant and specific promoter occupancy within the *ETS* gene family. *Genes Dev.* 21, 1882–1894. doi: 10.1101/gad.1561707

Hong, J. W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science* 321, 1314. doi: 10.1126/science.1160631

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31, 370–377. doi: 10.1038/ng941

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689

Irish, V. F. (2003). The evolution of floral homeotic gene function. *BioEssays* 25, 637–646. doi: 10.1002/bies.10292

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., et al. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840–846. doi: 10.1038/nature06847

Johnson, A. D., Meyer, B. J., and Ptashne, M. (1979). Interactions between DNA-bound repressors govern regulation by the λ phage repressor. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5061–5065. doi: 10.1073/pnas.76.10.5061

Kafri, R., Bar-Even, A., and Pilpel, Y. (2005). Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* 37, 295–299. doi: 10.1038/ng1523

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., et al. (2010). Variation in transcription factor binding among humans. *Science* 328, 232–235. doi: 10.1126/science.1183621

Katainen, R., Dave, K., Pitknen, E., Palin, K., Kivioja, T., Vlimki, N., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* 47, 818–821. doi: 10.1038/ng.3335

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., et al. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342, 84. doi: 10.1126/science.1235587

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., et al. (2013). Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science* 342, 744–747. doi: 10.1126/science.1242463

Kitano, H. (2004). Biological robustness. *Nat. Rev. Genet.* 5, 826–837. doi: 10.1038/nrg1471

Krivan, W., and Wasserman, W. W. (2001). A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* 11, 1559–1566. doi: 10.1101/gr.180601

Kwasnieski, J. C., Mogno, I., Meyers, C. A., Corbo, J. C., and Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19498–19503. doi: 10.1073/pnas.1210678109

Lawrence, P. A. (1992). *The Making of a Fly: The Genetics of Animal Design*. Oxford: Blackwell Science Ltd.

Lebrecht, D., Foehr, M., Smith, E., Lopes, F. J. P., Vanario-Alonso, C. E., Reinitz, J., et al. (2005). Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13176–13181. doi: 10.1073/pnas.0506462102

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251. doi: 10.1016/j.cell.2013.02.014

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

Lifanov, A. P., Makeev, V. J., Nazina, A. G., and Papatsenko, D. A. (2003). Homotypic regulatory clusters in *Drosophila*. *Genome Res.* 13, 579–588. doi: 10.1101/gr.668403

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567. doi: 10.1038/35000615

Lusk, R. W., and Eisen, M. B. (2000). Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. *PLoS Genet.* 6:e1000829. doi: 10.1371/journal.pgen.1000829

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Gen.* 8, 803–813. doi: 10.1038/nrg2192

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Lynch, M., O'Hely, M., Walsh, B., and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics* 159, 1789–1804.

Macneil, L. T., and Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.* 21, 645–657. doi: 10.1101/gr.097378.109

Madan Babu, M., and Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network of *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244. doi: 10.1093/nar/gkg210

Maerkl, S. J., and Quake, S. R. (2009). Experimental determination of the evolvability of a transcription factor. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18650–18655. doi: 10.1073/pnas.0907688106

Masel, J., and Siegal, M. L. (2009). Robustness: mechanisms and consequences. *Trends Genet.* 25, 395–403. doi: 10.1016/j.tig.2009.07.005

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

McGregor, A. P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D. G., Payre, F., et al. (2007). Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature* 448, 587–590. doi: 10.1038/nature05988

McGuigan, K., and Sgrò, C. M. (2009). Evolutionary consequences of cryptic genetic variation. *Trends Ecol. Evol.* 24, 305–311. doi: 10.1016/j.tree.2009.02.001

McManus, C. J., May, G. E., Spealman, P., and Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* 24, 422–430. doi: 10.1101/gr.164996.113

Meiklejohn, C. D., and Hartl, D. L. (2002). A single mode of canalization. *Trends Ecol. Evol.* 17, 468–473. doi: 10.1016/S0169-5347(02)02596-X

Melton, C., Reuter, J. A., Spacek, D. V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* 47, 710–716. doi: 10.1038/ng.3332

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X., Biggin, M. D., et al. (2006). Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput. Biol.* 2:e130. doi: 10.1371/journal.pcbi.0020130

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K., et al. (2010). From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature* 466, 714–719. doi: 10.1038/nature09266

Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.* 89:208701. doi: 10.1103/PhysRevLett.89.208701

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., et al. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* 4:e04837. doi: 10.7554/eLife.04837

Nowak, M. A., Boerlijst, M. C., Cooke, J., and Maynard Smith, J. M. (1997). Evolution of genetic redundancy. *Nature* 388, 167–171. doi: 10.1038/40618

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., et al. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* 39, 730–732. doi: 10.1038/ng2047

O'Hely, M. (2006). A diffusion approach to approximating preservation probabilities for gene duplicates. *J. Math. Biol.* 53, 215–230. doi: 10.1007/s00285-006-0001-6

Ohno, S. (1970). *Evolution by Gene Duplication*. New York, NY: Blackwell Science Ltd; Springer. doi: 10.1007/978-3-642-86659-3

Paixão, T., and Azevedo, R. B. R. (2010). Redundancy and the evolution of cis-regulatory element multiplicity. *PLoS Comput. Biol.* 6:e10000848. doi: 10.1371/journal.pcbi.1000848

Payne, J. L., Moore, J. H., and Wagner, A. (2014). Robustness, evolvability, and the logic of genetic regulation. *Artif. Life* 20, 111–126. doi: 10.1162/ARTL_a_00099

Payne, J. L., and Wagner, A. (2014). The robustness and evolvability of transcription factor binding sites. *Science* 466, 714–719. doi: 10.1126/science.1249046

Pechenick, D. A., Moore, J. H., and Payne, J. L. (2013). The influence of assortativity on the robustness and evolvability of gene regulatory networks upon gene birth. *J. Theor. Biol.* 330, 26–36. doi: 10.1016/j.jtbi.2013.03.019

Pechenick, D. A., Payne, J. L., and Moore, J. H. (2012). The influence of assortativity on the robustness of signal-integration logic in gene regulatory networks. *J. Theor. Biol.* 296, 21–32. doi: 10.1016/j.jtbi.2011.11.029

Pechenick, D. A., Payne, J. L., and Moore, J. H. (2014). Phenotypic robustness and the assortativity signature of human transcription factor networks. *PLoS Comput. Biol.* 10:e1003780. doi: 10.1371/journal.pcbi.1003780

Pérez, J. C., Fordyce, P. M., Lohse, M. B., Hanson-Smith, V., DeRisi, J. L., and Johnson, A. D. (2014). How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. *Genes Dev.* 28, 1272–1277. doi: 10.1101/gad.242271.114

Perry, M. W., Boettiger, A. N., Bothma, J. P., and Levine, M. (2010). Shadow enhancers foster robustness of *Drosophila* gastrulation. *Curr. Biol.* 20, 1562–1567. doi: 10.1016/j.cub.2010.07.043

Perry, M. W., Boettiger, A. N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the *drosophila* embryo. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13570–13575. doi: 10.1073/pnas.1109873108

Peter, I. S., and Davidson, E. H. (2009). Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS Lett.* 583, 3948–3958. doi: 10.1016/j.febslet.2009.11.060

Poblanno-Balp, R., and Gershenson, C. (2011). Modular random boolean networks. *Artif. Life* 17, 331–351. doi: 10.1162/artl_a_00042

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. *Nat. Genet.* 41, 882–884. doi: 10.1038/ng.403

Ptashne, M., and Gann, A. (2002). *Genes & Signals*. New York, NY: Cold Spring Harbor Laboratory Press.

Queitsch, C., Sanger, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor for phenotypic variation. *Nature* 417, 618–624. doi: 10.1038/nature749

Reece-Hoyes, J. S., Diallo, A., Lajoie, B., Kent, A., Shrestha, S., Kadreppa, S., et al. (2011). Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat. Methods* 8, 1059–1064. doi: 10.1038/nmeth.1748

Rensing, S. A. (2014). Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.* 17, 43–48. doi: 10.1016/j.pbi.2013.11.002

Rockman, M. V., and Wray, G. A. (2002). Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 19, 1991–2004. doi: 10.1093/oxfordjournals.molbev.a004023

Rutherford, S. L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342. doi: 10.1038/24550

Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C., and Wagner, A. (2010). Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* 4:30. doi: 10.1186/1752-0509-4-30

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040. doi: 10.1126/science.1186176

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176. doi: 10.1038/ng1165

Sengupta, A. M., Djordjevic, M., and Shraiman, B. I. (2002). Specificity and robustness in transcription control networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 2072–2077. doi: 10.1073/pnas.022388499

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., et al. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* 30, 521–530. doi: 10.1038/nbt.2205

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286. doi: 10.1038/nrg3682

Shultzaberger, R. K., Malashock, D. S., Kirsch, J. F., and Eisen, M. B. (2010). The fitness landscape fo *cis*-acting binding sites in different promoter and environmental contexts. *PLoS Genet.* 6:e1001042. doi: 10.1371/journal.pgen.1001042

Siegal, M. L., and Bergman, A. (2002). Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10528–10532. doi: 10.1073/pnas.102303999

Silva-Rocha, R., and de Lorenzo, V. (2010). Noise and robustness in prokaryotic regulatory networks. *Annu. Rev. Microbiol.* 64, 257–275. doi: 10.1146/annurev.micro.091208.073229

Small, S., Blair, A., and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev. Biol.* 175, 314–324. doi: 10.1006/dbio.1996.0117

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., et al. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* 45, 1021–1028. doi: 10.1038/ng.2713

Somma, M. P., Pisano, C., and Lavia, P. (1991). The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. *Nucleic Acids Res.* 19, 2817–2824. doi: 10.1093/nar/19.11.2817

Sorrells, T. R., and Johnson, A. D. (2015). Making sense of transcription networks. *Cell* 161, 714–723. doi: 10.1016/j.cell.2015.04.014

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., et al. (2012). Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.* 13, R49. doi: 10.1186/gb-2012-13-9-r49

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. III., and Doyle, J. (2004). Robustness of cellular functions. *Cell* 118, 675–685. doi: 10.1016/j.cell.2004.09.008

Stewart, A. J., Hannenhalli, S., and Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics* 192, 973–985. doi: 10.1534/genetics.112.143370

Swanson, C. I., Schwimmer, D. B., and Barolo, S. (2011). Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr. Biol.* 21, 1186–1196. doi: 10.1016/j.cub.2011.05.056

Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal evolution in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7203–7208. doi: 10.1073/pnas.0502521102

Teichmann, S. A., and Babu, M. M. (2003). Gene regulatory network growth by duplication. *Nat. Genet.* 36, 492–496. doi: 10.1038/ng1340

Thanos, D., and Maniatis, T. (1995). Virus induction of human ifnβ gene expression requires the assembly of an enhanceosome. *Cell* 83, 1091–1100. doi: 10.1016/0092-8674(95)90136-1

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Thomas, J. H. (1993). Thinking about genetic redundancy. *Trends Genet.* 9, 395–399. doi: 10.1016/0168-9525(93)90140-D

Tischler, J., Lehner, B., Chen, N., and Fraser, A. G. (2006). Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol.* 7:R69. doi: 10.1186/gb-2006-7-8-r69

Tsong, A. E., Tuch, B. B., Li, H., and Johnson, A. D. (2006). Evolution of alternative transcriptional circuits with identical logic. *Nature* 443, 415–420. doi: 10.1038/nature05099

van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9716–9720. doi: 10.1073/pnas.96.17.9716

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263. doi: 10.1038/nrg2538

Wagner, A. (1997). A population genetic theory of canalization. *Evolution* 51, 329–347. doi: 10.2307/2411105

Wagner, A. (1999). Redundant gene functions and natural selection. *J. Evol. Biol.* 12, 1–16. doi: 10.1046/j.1420-9101.1999.00008.x

Wagner, A. (2000). Robustness against mutations in genetic networks of yeast. *Nat. Genet.* 24, 355–361. doi: 10.1038/74174

Wagner, A. (2005). *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.

Wagner, A. (2008). Robustness and evolvability: a paradox resolved. *Proc. Biol. Sci.* 275, 91–100. doi: 10.1098/rspb.2007.1137

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet.* 8, 921–931. doi: 10.1038/nrg2267

Wasserman, W. W., and Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167–181. doi: 10.1006/jmbi.1998.1700

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 46, 1160–1165. doi: 10.1038/ng.3101

Weirauch, M. T., and Hughes, T. R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more

they stay the same. *Trends Genet.* 26, 66–74. doi: 10.1016/j.tig.2009.12.002

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

Wong, K.-C., Chan, T.-M., Peng, C., Li, Y., and Zhang, Z. (2013). DNA motif elucidation using belief propagation. *Nucleic Acids Res.* 41, e153. doi: 10.1093/nar/gkt574

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298. doi: 10.1016/S0169-5347(03)00033-8

Zheng, W., Gianoulis, T. A., Karczewski, K. J., Zhao, H., and Snyder, M. (2011). Regulatory variation within and between species. *Annu. Rev. Genomics Hum. Genet.* 12, 327–346. doi: 10.1146/annurev-genom-082908-150139

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M., and Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature* 464, 1187–1191. doi: 10.1038/nature08934

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 15 February 2016
doi: 10.3389/fgene.2016.00016

Robustness and Accuracy in Sea Urchin Developmental Gene Regulatory Networks

Smadar Ben-Tabou de-Leon *

The Department of Marine Biology, The University of Haifa, Haifa, Israel

Developmental gene regulatory networks robustly control the timely activation of regulatory and differentiation genes. The structure of these networks underlies their capacity to buffer intrinsic and extrinsic noise and maintain embryonic morphology. Here I illustrate how the use of specific architectures by the sea urchin developmental regulatory networks enables the robust control of cell fate decisions. The Wnt-βcatenin signaling pathway patterns the primary embryonic axis while the BMP signaling pathway patterns the secondary embryonic axis in the sea urchin embryo and across bilateria. Interestingly, in the sea urchin in both cases, the signaling pathway that defines the axis controls directly the expression of a set of downstream regulatory genes. I propose that this direct activation of a set of regulatory genes enables a uniform regulatory response and a clear cut cell fate decision in the endoderm and in the dorsal ectoderm. The specification of the mesodermal pigment cell lineage is activated by Delta signaling that initiates a triple positive feedback loop that locks down the pigment specification state. I propose that the use of compound positive feedback circuitry provides the endodermal cells enough time to turn off mesodermal genes and ensures correct mesoderm vs. endoderm fate decision. Thus, I argue that understanding the control properties of repeatedly used regulatory architectures illuminates their role in embryogenesis and provides possible explanations to their resistance to evolutionary change.

Keywords: developmental gene regulatory network, development and evolution, compound network motifs, sea urchins, Wnt signaling pathway, BMP signaling, Delta-Notch signaling

OPEN ACCESS

Edited by:
Ekaterina Shelest, Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knoell Institute, Germany

Reviewed by:
Pavel Loskot, Swansea University, UK
Tom Thorne, University of Edinburgh, UK

***Correspondence:**
Smadar Ben-Tabou de-Leon
sben-tab@univ.haifa.ac.il

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *02 December 2015*
Accepted: *28 January 2016*
Published: *15 February 2016*

Citation:
Ben-Tabou de-Leon S (2016) Robustness and Accuracy in Sea Urchin Developmental Gene Regulatory Networks. Front. Genet. 7:16. doi: 10.3389/fgene.2016.00016

INTRODUCTION

Robustness, the perseverance of phenotype through genetic and environmental changes (de Visser et al., 2003), is a prominent property of embryo development. Thus, embryos can maintain their morphologies through a wide range of temperatures and pH (Runcie et al., 2012; Pespeni et al., 2013; Kuntz and Eisen, 2014) and within substantial genetic variation (Garfield et al., 2013). This robustness of the developmental program relays on various levels of molecular control, among them, transcription factor binding to the DNA, enhancer structure and the architecture of developmental gene regulatory networks (reviewed in de Visser et al., 2003; Kitano, 2007; Payne and Wagner, 2015). Here I describe the repeated use of specific network architectures in the sea urchin developmental gene regulatory networks, and illustrate how they contribute to robust cell fate decision.

The current model of the sea urchin developmental regulatory networks encompasses all the embryonic territories up to gastrulation and is one of the most elaborate of its kind

(Saudemont et al., 2010; Peter and Davidson, 2011; Materna and Davidson, 2012; Ben-Tabou de-Leon et al., 2013). A major strength of this network model is the extensive *cis*-regulatory analyses conducted for many nodes (e.g., Nam et al., 2007; Ben-Tabou de Leon and Davidson, 2010; Ransick and Davidson, 2012). Thus, the direct connectivity of this network is highly reliable and can provide a systems level view of how network architecture contributes to the precise control of embryonic axes formation and germ layer specification.

Within the sea urchin regulatory network, specific network architectures are repeatedly used to control various patterning events at different embryonic territories (Ben-Tabou de-Leon and Davidson, 2006; Peter and Davidson, 2009). These network architectures are composed of multiple interconnected common network motifs: switches, feedforward and feedback loops (Ben-Tabou de-Leon and Davidson, 2006; Peter and Davidson, 2009). The concept of "common network motifs" originated more than a decade ago by Alon and colleagues that identified typical three-node network circuitries overrepresented in bacterial transcriptional regulatory networks (Shen-Orr et al., 2002). Since then, similar and other network motifs were identified in other biological systems and their intensive study illuminates the relationship between motif structure and its control function (Hornung and Barkai, 2008; Shoval and Alon, 2010). Here I illustrate how compound interconnected network motifs are used by the sea urchin developmental gene regulatory networks and propose that their control properties are utilized to ensure robustness and accuracy of cell fate decisions.

WNT-βCATENIN REGULATION OF PRIMARY AXIS FORMATION AND ENDODERM SPECIFICATION

Extensive research had shown the extreme conservation of the role of the Wnt-βcatenin signaling pathway in primary axis formation and endoderm specification across metazoan (Petersen and Reddien, 2009). The model of the sea urchin developmental regulatory networks reveal how Wnt-βcatenin spatial information is transformed into specific cell fate decisions. The primary axis in the sea urchin embryo, the animal-vegetal axis, is initiated by nuclear localization of βcatenin in all the cells of the vegetal half of the embryo [**Figure 1A**, endomesodermal lineages, B, βcatenin nuclearization pattern (Logan et al., 1999)]. When βcatenin enters the nucleus it forms an activating complex with the transcription factor Tcf that otherwise forms a repressor complex with Groucho. The βcatenin-Tcf switch initiates the specification of both mesoderm and endoderm in the vegetal half of the sea urchin embryo (**Figures 1A–E**).

βcatenin-Tcf switch directly activates the expression of a set of endodermal regulatory genes, *hox11/13*, *blimp1*, *foxa,* and *bra*, in a staggered manner [**Figure 1C** (Cui et al., 2014)]. That is, the expression of each of these gene is turned on at a different time, but their spatial expression overlap, at least at the earlier stages of their expression (Minokawa et al., 2005; Livi and Davidson, 2006; Peter and Davidson, 2010, 2011). Each of these genes has functional Tcf sites in its enhancers, indicating direct control of Wnt signaling through βcatenin /Groucho-Tcf switch (**Figure 1F**, Minokawa et al., 2005; Smith et al., 2007, 2008; Ben-Tabou de Leon and Davidson, 2010).

At Mesenchyme blastula stage, βcatenin clears from the mesodermal nuclei, first from the skeletogenic lineage and then from the non-skeletogenic mesoderm [**Figure 1B** (Logan et al., 1999)]. When βcatenin is cleared from the mesodermal nuclei the Tcf sites on the enhancers of the endodermal genes control their clearance from the mesoderm territories through Tcf-Groucho mediated repression (Ben-Tabou de Leon and Davidson, 2010) and thus regulate the endoderm—mesoderm cell fate decision (**Figure 1F**, Peter and Davidson, 2011). Apparently, βcatenin-Tcf acts as a permissive switch and restricts the expression of these genes spatially, while their differential activation time is defined by their specific activators (**Figure 1F**). I suggest that this mode of regulatory circuitry decouples the spatial from the temporal regulation and promotes a uniform spatial response of all the endodermal genes. Thus, βcatenin-Tcf/Groucho-Tcf switch ensures that the endodermal genes will be cleared from the mesodermal domain at the right developmental stage and guarantees a clear-cut cell fate decision.

DELTA-NOTCH ACTIVATION OF A TRIPLE POSITIVE FEEDBACK CIRCUIT AND MESODERM CELL FATE SPECIFICATION

The Delta-Notch signaling pathway is highly conserved in metazoan and controls glial vs. neural differentiation (Gaiano and Fishell, 2002). Early in sea urchin embryogenesis, the gene that encodes the ligand Delta is activated indirectly by the βcatenin-Tcf input in the skeletogenic mesoderm (**Figure 1D**, Oliveri et al., 2008). The reception of Delta in the neighboring tier of cells, Veg2, activates the gene that encodes the transcription factor *glial cells missing* [GCM, **Figures 1E,G** (Ransick and Davidson, 2006; Croce and McClay, 2010)]. GCM then establishes a triple positive feedback loop by directly activating the expression of the transcription factor GataE, that activates the expression of the transcription factor Six1/2, that feeds back to activate GCM expression (**Figures 1E,G**, Ransick and Davidson, 2012). GCM-GataE-Six1/2 triple positive feedback loop maintains the expression of these three genes in the pigment cell lineage after Delta signal stops being received in these cells (20 hpf in *S. purpuratus*, **Figures 1D,E,G**).

The tier of cells where GCM is first activated, Veg2, give rise to both endoderm and non-skeletogenic mesoderm lineages (**Figure 1A**, 12 hpf). When Veg2 cells divide, only the future pigment cells remain in direct contact with the Delta secreting SM cells, while the future endodermal cells lose this contact and therefore lose the Delta signal (**Figure 1A**, 15 hpf). Hence, in the endodermal cells the Delta signal is not received long enough to establish the triple positive feedback loop so GCM expression turns off there (**Figure 1E**, 15 and 20 hpf, Ransick and Davidson, 2006, 2012; Croce and McClay, 2010). The transient Delta signal is practically filtered in the endodermal cells by the mesodermal positive feedback loop to allow correct endodermal fate decision.

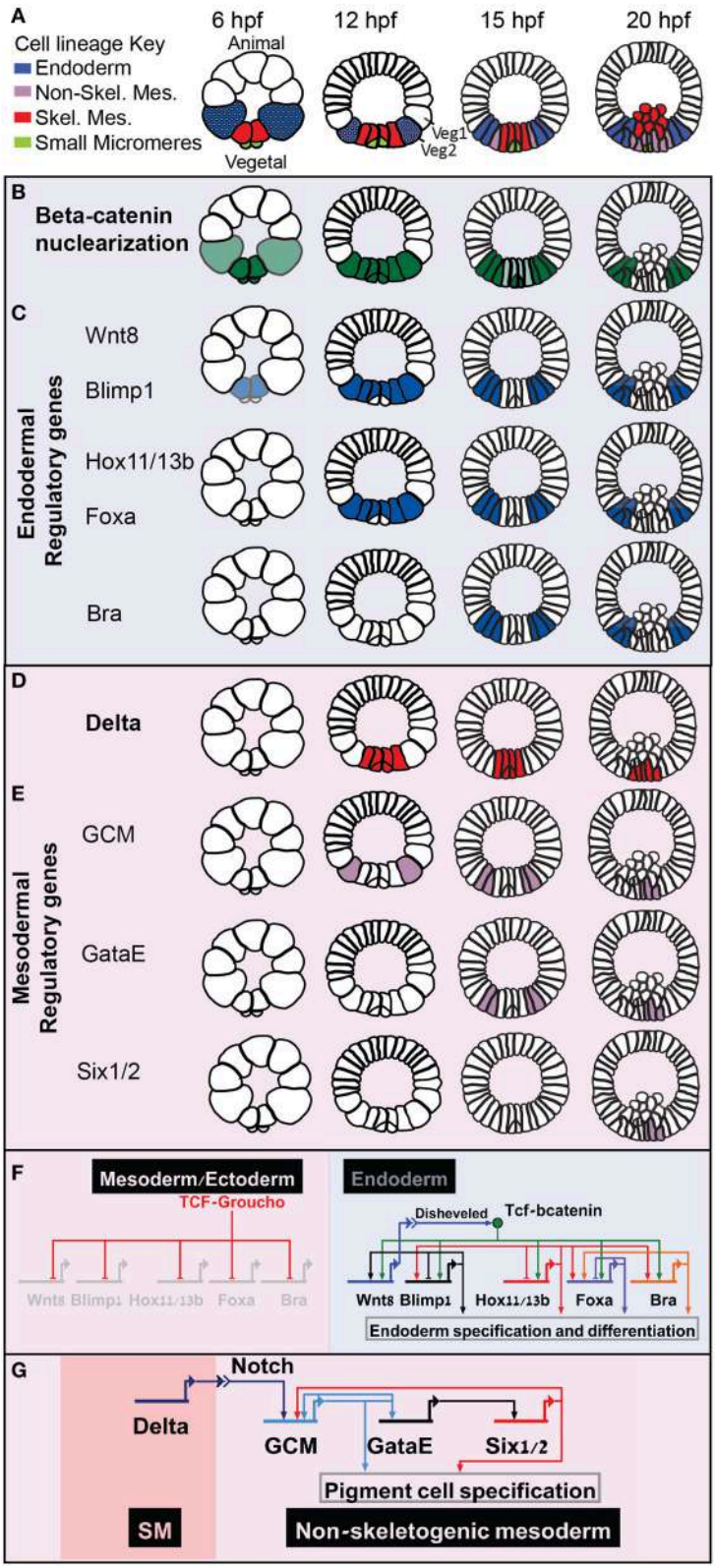


FIGURE 1 | Sea urchin embryonic development and endoderm specification. Developmental time is described in hours post fertilization according the developmental rate of the purple sea urchin, *S. purpuratus*. **(A)** Sea urchin endomesoderm cell lineage diagram. Color key is described in the figure. **(B)** βcatenin nuclearization pattern, dark green indicates high concentration, light green low. **(C)** Spatio-temporal expression profiles of endodermal control genes. **(D)** Partial

(Continued)

FIGURE 1 | Continued
endodermal GRN model depicting Tcf/βcatenin-Tcf/Groucho switch and regulatory interactions within the endodermal genes. **(E)** Spatio-temporal expression of the Delta ligand. **(F)** spatio-temporal expression of non-skeletogenic mesodermal genes. **(G)** GRN model of the triple positive feedback loop that Delta reception activates in the non-skeletogenic cells.

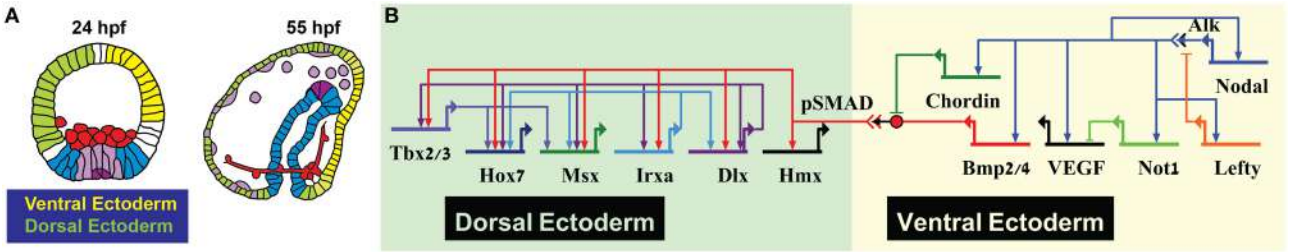


FIGURE 2 | Sea urchin dorsal-ventral patterning. (A) Sea urchin lineage diagram showing ventral (yellow) and dorsal (light green) ectoderm. **(B)** Partial model on Dosrsal-Ventral patterning in the sea urchin depicting key regulatory processes in the ectoderm.

Previous theoretical studies of three component circuits show that feedback circuitry is more efficient than other architectures in buffering noise in the inducing signal while keeping high responsivity to the level of the signal (Hornung and Barkai, 2008). According to these studies, noise reduction in positive feedback circuits results from effectively slowing the response dynamics and allowing for better averaging of the induction signal over time. Additionally, mathematical modeling of the kinetics of positive feedback loops shows that compound positive feedback circuitry is less responsive than single positive feedback loop to low levels of activating signals (Ben-Tabou de-Leon, 2010). These studies suggests that compound positive feedback circuitry filters better low and transient signals compared to single positive feedback loops and thus are a more reliable mechanism for regulatory state lock down. This could be the reason for the common use of compound positive feedback circuits by developmental networks instead of single gene positive feedback loop.

TGFB PATHWAYS CONTROL OF SECONDARY AXIS AND ECTODERM SPECIFICATION

The gene regulatory networks that pattern the secondary embryonic axis, the dorsal-ventral axis of the sea urchin embryo, use similar circuit architectures to those discussed above. Nodal signaling directly activates the ventral ectoderm regulatory genes that then interact with each other to form subdomains within the ventral ectoderm [**Figure 2** (Saudemont et al., 2010; Li et al., 2014)]. Two of Nodal targets at the ventral ectoderm are the ligand BMP2/4 and its inhibitor Chordin. Chordin inhibits BMP reception at the ventral side so the mediator of BMP signaling, the transcription factor SMAD1/5/8, is phosphorylated and activates transcription only in the dorsal side of the embryo (**Figure 2B**, Saudemont et al., 2010; Ben-Tabou de-Leon et al., 2013). BMP operates in a feed-forward structure, directly activating the expression of dorsal transcription factors that then regulate one another forming compound positive feedback loop (**Figure 2B**, Ben-Tabou de-Leon et al., 2013). Thus, BMP provides a temporal cue that uniformly boosts the expression of the aboral transcription factors at the exact time when the first genes that specify the neighboring territory, the ciliated band, are turning on (Ben-Tabou de-Leon et al., 2013).

CONCLUSIONS: PRECISE AND HIGHLY CONSERVED CONTROL OF EXPRESSION DYNAMICS

As we gain more information on the structure and function of gene regulatory networks we can start asking why are specific architectures used more than others and why are they so deeply conserved? A recent paper revealed remarkable conservation of regulatory gene expression dynamics between two sea urchin species after 40 million years of independent evolution (Gildor and Ben-Tabou de-Leon, 2015). The use of direct activation by signaling pathways and compound positive feedback circuitry described above could underlie this strong conservation of expression dynamics and the observed robustness within genotypic variance and different environmental conditions.

Direct activation by a signaling pathway might be a general strategy used by developmental gene regulatory networks to guarantees a uniform timely response of a set of key regulatory genes. This strategy could also explain the deep conservation of the role of Wnt and BMP pathways in primary embryonic axes specification. If the activation of the downstream gene regulatory network was in a cascade of regulatory interactions, there were only a few regulatory changes required to replace Wnt or BMP with alternative signaling input. It is much less likely to replace Wnt or BMP signaling when they activate the entire set of genes that define the endoderm or dorsal ectoderm specification, respectively. Thus, the direct activation of large sets of regulatory genes by signaling pathways might be important for clear cut cell fate decision on one hand, and on the other hand imposes

a strong constraint on the use of these signaling pathways in developing embryos.

Similar argument could explain the extreme conservation of well-studied compound positive feedback circuits. Specifically, the compound positive feedback circuit that controls the lock down of endoderm cell fate specification was conserved across 500 years of echinoderm evolution (Hinman et al., 2003); the compound positive feedback circuit that controls heart development is conserved between human and fly (Olson, 2006). It seems that any regulatory change within these critical control circuits must have reduced the circuit precision and therefore had been selected against. Thus, understanding the control properties of repeatedly used regulatory architectures illuminates their function in developing embryos and provides possible explanation to their resistance to evolutionary change.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the Marie Curie Carrier Integration Grant FP7-PEOPLE-2012-CIG, grant number 321758.

REFERENCES

Ben-Tabou de-Leon, S. (2010). Perturbation analysis analyzed - mathematical modeling of intact and perturbed gene regulatory circuits for animal development. *Dev. Biol.* 344, 1110–1118. doi: 10.1016/j.ydbio.2010.06.020

Ben-Tabou de Leon, S., and Davidson, E. H. (2010). Information processing at the foxa node of the sea urchin endomesoderm specification network. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10103–10108. doi: 10.1073/pnas.1004824107

Ben-Tabou de-Leon, S., and Davidson, E. H. (2006). Deciphering the underlying mechanism of specification and differentiation: the sea urchin gene regulatory network. *Sci. STKE* 2006:pe47. doi: 10.1126/stke.3612006pe47

Ben-Tabou de-Leon, S., Su, Y. H., Lin, K. T., Li, E., and Davidson, E. H. (2013). Gene regulatory control in the sea urchin aboral ectoderm: spatial initiation, signaling inputs, and cell fate lockdown. *Dev. Biol.* 374, 245–254. doi: 10.1016/j.ydbio.2012.11.013

Croce, J. C., and McClay, D. R. (2010). Dynamics of Delta/Notch signaling on endomesoderm segregation in the sea urchin embryo. *Development* 137, 83–91. doi: 10.1242/dev.044149

Cui, M., Siriwon, N., Li, E., Davidson, E. H., and Peter, I. S. (2014). Specific functions of the Wnt signaling system in gene regulatory networks throughout the early sea urchin embryo. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5029–E5038. doi: 10.1073/pnas.1419141111

de Visser, J. A., Hermisson, J., Wagner, G. P., Ancel Meyers, L., Bagheri-Chaichian, H., Blanchard, J. L., et al. (2003). Perspective: evolution and detection of genetic robustness. *Evolution* 57, 1959–1972. doi: 10.1111/j.0014-3820.2003.tb00377.x

Gaiano, N., and Fishell, G. (2002). The role of notch in promoting glial and neural stem cell fates. *Annu. Rev. Neurosci.* 25, 471–490. doi: 10.1146/annurev.neuro.25.030702.130823

Garfield, D. A., Runcie, D. E., Babbitt, C. C., Haygood, R., Nielsen, W. J., and Wray, G. A. (2013). The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network. *PLoS Biol.* 11:e1001696. doi: 10.1371/journal.pbio.1001696

Gildor, T., and Ben-Tabou de-Leon, S. (2015). Comparative study of regulatory circuits in two sea urchin species reveals tight control of timing and high conservation of expression dynamics. *PLoS Genet.* 11:e1005435. doi: 10.1371/journal.pgen.1005435

Hinman, V. F., Nguyen, A. T., Cameron, R. A., and Davidson, E. H. (2003). Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13356–13361. doi: 10.1073/pnas.2235868100

Hornung, G., and Barkai, N. (2008). Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput. Biol.* 4:e8. doi: 10.1371/journal.pcbi.0040008

Kitano, H. (2007). Towards a theory of biological robustness. *Mol. Syst. Biol.* 3:137. doi: 10.1038/msb4100179

Kuntz, S. G., and Eisen, M. B. (2014). Drosophila embryogenesis scales uniformly across temperature in developmentally diverse species. *PLoS Genet.* 10:e1004293. doi: 10.1371/journal.pgen.1004293

Li, E., Cui, M., Peter, I. S., and Davidson, E. H. (2014). Encoding regulatory state boundaries in the pregastrular oral ectoderm of the sea urchin embryo. *Proc. Natl. Acad. Sci. U.S.A.* 111, E906–E913. doi: 10.1073/pnas.1323105111

Livi, C. B., and Davidson, E. H. (2006). Expression and function of blimp1/krox, an alternatively transcribed regulatory gene of the sea urchin endomesoderm network. *Dev. Biol.* 293, 513–525. doi: 10.1016/j.ydbio.2006.02.021

Logan, C. Y., Miller, J. R., Ferkowicz, M. J., and McClay, D. R. (1999). Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development* 126, 345–357.

Materna, S. C., and Davidson, E. H. (2012). A comprehensive analysis of Delta signaling in pre-gastrular sea urchin embryos. *Dev. Biol.* 364, 77–87. doi: 10.1016/j.ydbio.2012.01.017

Minokawa, T., Wikramanayake, A. H., and Davidson, E. H. (2005). cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Dev. Biol.* 288, 545–558. doi: 10.1016/j.ydbio.2005.09.047

Nam, J., Su, Y. H., Lee, P. Y., Robertson, A. J., Coffman, J. A., and Davidson, E. H. (2007). Cis-regulatory control of the nodal gene, initiator of the sea urchin oral ectoderm gene network. *Dev. Biol.* 306, 860–869. doi: 10.1016/j.ydbio.2007.03.033

Oliveri, P., Tu, Q., and Davidson, E. H. (2008). Global regulatory logic for specification of an embryonic cell lineage. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5955–5962. doi: 10.1073/pnas.0711220105

Olson, E. N. (2006). Gene regulatory networks in the evolution and development of the heart. *Science* 313, 1922–1927. doi: 10.1126/science.1132292

Payne, J. L., and Wagner, A. (2015). Mechanisms of mutational robustness in transcriptional regulation. *Front. Genet.* 6:322. doi: 10.3389/fgene.2015.00322

Pespeni, M. H., Chan, F., Menge, B. A., and Palumbi, S. R. (2013). Signs of adaptation to local pH conditions across an environmental mosaic in the California Current Ecosystem. *Integr. Comp. Biol.* 53, 857–870. doi: 10.1093/icb/ict094

Peter, I., and Davidson, E. H. (2010). Genomic programs for endoderm specification in sea urchin embryos. *Dev. Biol.* 344:469. doi: 10.1016/j.ydbio.2010.05.225

Peter, I. S., and Davidson, E. H. (2009). Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS Lett.* 583, 3948–3958. doi: 10.1016/j.febslet.2009.11.060

Peter, I. S., and Davidson, E. H. (2011). A gene regulatory network controlling the embryonic specification of endoderm. *Nature* 474, 635–639. doi: 10.1038/nature10100

Petersen, C. P., and Reddien, P. W. (2009). Wnt signaling and the polarity of the primary body axis. *Cell* 139, 1056–1068. doi: 10.1016/j.cell.2009.11.035

Ransick, A., and Davidson, E. H. (2006). cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Dev. Biol.* 297, 587–602. doi: 10.1016/j.ydbio.2006.05.037

Ransick, A., and Davidson, E. H. (2012). Cis-regulatory logic driving glial cells missing: self-sustaining circuitry in later embryogenesis. *Dev. Biol.* 364, 259–267. doi: 10.1016/j.ydbio.2012.02.003

Runcie, D. E., Garfield, D. A., Babbitt, C. C., Wygoda, J. A., Mukherjee, S., and Wray, G. A. (2012). Genetics of gene expression responses to temperature stress in a sea urchin gene network. *Mol. Ecol.* 21, 4547–4562. doi: 10.1111/j.1365-294x.2012.05717.x

Saudemont, A., Haillot, E., Mekpoh, F., Bessodes, N., Quirin, M., Lapraz, F., et al. (2010). Ancestral regulatory circuits governing ectoderm patterning downstream of Nodal and BMP2/4 revealed by gene regulatory network analysis in an echinoderm. *PLoS Genet.* 6:e1001259. doi: 10.1371/journal.pgen.1001259

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68. doi: 10.1038/ng881

Shoval, O., and Alon, U. (2010). SnapShot: network motifs. *Cell* 143:326–e1. doi: 10.1016/j.cell.2010.09.050

Smith, J., Kraemer, E., Liu, H., Theodoris, C., and Davidson, E. (2008). A spatially dynamic cohort of regulatory genes in the endomesodermal gene network of the sea urchin embryo. *Dev. Biol.* 313, 863–875. doi: 10.1016/j.ydbio.2007.10.042

Smith, J., Theodoris, C., and Davidson, E. H. (2007). A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science* 318, 794–797. doi: 10.1126/science.1146524

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2016
doi: 10.3389/fgene.2016.00031

A Consensus Network of Gene Regulatory Factors in the Human Frontal Lobe

Stefano Berto[1,2,3]*, *Alvaro Perdomo-Sabogal*[1], *Daniel Gerighausen*[1], *Jing Qin*[4,5] and *Katja Nowick*[1,2]*

[1] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Leipzig, Germany, [2] Paul-Flechsig Institute for Brain Research, University of Leipzig, Leipzig, Germany, [3] Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, TX, USA, [4] Department of Mathematics and Computer Sciences, University of Southern Denmark, Odense, Denmark, [5] Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria

OPEN ACCESS

Edited by:
Edgar Wingender,
University Medical Center Goettingen,
Germany

Reviewed by:
Beisi Xu,
St. Jude Children's Research Hospital,
USA
Vsevolod Jurievich Makeev,
Vavilov Institute of General Genetics,
Russia

***Correspondence:**
Stefano Berto
stefano.berto@utsouthwestern.edu;
stefano@bioinf.uni-leipzig.de;
Katja Nowick
nowick@bioinf.uni-leipzig.de

Specialty section:
This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics

Received: 30 October 2015
Accepted: 18 February 2016
Published: 08 March 2016

Citation:
Berto S, Perdomo-Sabogal A,
Gerighausen D, Qin J and Nowick K
(2016) A Consensus Network of Gene
Regulatory Factors in the Human
Frontal Lobe. Front. Genet. 7:31.
doi: 10.3389/fgene.2016.00031

Cognitive abilities, such as memory, learning, language, problem solving, and planning, involve the frontal lobe and other brain areas. Not much is known yet about the molecular basis of cognitive abilities, but it seems clear that cognitive abilities are determined by the interplay of many genes. One approach for analyzing the genetic networks involved in cognitive functions is to study the coexpression networks of genes with known importance for proper cognitive functions, such as genes that have been associated with cognitive disorders like intellectual disability (ID) or autism spectrum disorders (ASD). Because many of these genes are gene regulatory factors (GRFs) we aimed to provide insights into the gene regulatory networks active in the human frontal lobe. Using genome wide human frontal lobe expression data from 10 independent data sets, we first derived 10 individual coexpression networks for all GRFs including their potential target genes. We observed a high level of variability among these 10 independently derived networks, pointing out that relying on results from a single study can only provide limited biological insights. To instead focus on the most confident information from these 10 networks we developed a method for integrating such independently derived networks into a consensus network. This consensus network revealed robust GRF interactions that are conserved across the frontal lobes of different healthy human individuals. Within this network, we detected a strong central module that is enriched for 166 GRFs known to be involved in brain development and/or cognitive disorders. Interestingly, several hubs of the consensus network encode for GRFs that have not yet been associated with brain functions. Their central role in the network suggests them as excellent new candidates for playing an essential role in the regulatory network of the human frontal lobe, which should be investigated in future studies.

Keywords: transcription factor, coexpression network, weighted topological overlap network, consensus network, cognitive abilities, cognitive disorders, prefrontal cortex (PFC)

INTRODUCTION

Broadly defined, cognition refers to the biological mechanisms through which animals perceive, learn and memorize information from the environment and decide to act upon them (Shettleworth, 2009). In humans, cognitive processes such as use of language, social behavior, and decision-making have been attributed to the frontal lobe (Duncan et al., 1996; Chayer and Freedman, 2001). However, the actual molecular mechanisms that underlie these morphological changes are still not well understood.

Candidate genes that are involved in the molecular mechanisms of cognition can be identified through biomedical studies on cognitive disorders. For example, causative mutations point to the genes that should in their wild-type variants be important for providing for healthy cognitive abilities. Research on cognitive disorders such as Alzheimer's disease (AD; Bullido et al., 1998), intellectual disability (ID; Kaufman et al., 2010), autism spectrum disorder (ASD; Bailey et al., 1996; Voineagu et al., 2011; Berg and Geschwind, 2012; Ecker et al., 2012), schizophrenia (SZ; Andreasen, 1995), circadian rhythm and bipolar disorder (BD; Akula et al., 2014, 2016; Takahashi, 2015), Parkinson's disease (PD; Polymeropoulos, 2000), and several syndromes or disorders associated with ID or cognitive impairment (SY; Greydanus and Pratt, 2005) has thus already identified several candidate genes involved in cognition. Importantly, these studies also revealed that most cognitive disorders are complex and phenotypically and genetically heterogeneous (Sebat et al., 2007; Tsankova et al., 2007; Voineagu et al., 2011; Weyn-Vanhentenryck et al., 2014), thus creating challenges for studying these disorders.

Transcriptome and network analyses bear great potential for overcoming some of these challenges and uncovering the genetic interactions and molecular mechanisms causing such complex disorders. For example, recent studies have used network approaches to identify coexpressed ASD and ID modules implicated in synaptic development, chromatin remodeling and early transcriptional regulation (Parikshak et al., 2013; Willsey et al., 2013; De Rubeis et al., 2014). However, coexpression networks can have many false positive inferences. One way to reduce the effect of false positives is to calculate weighted topological overlap (wTO) networks (Zhang and Horvath, 2005; Nowick et al., 2009). Another drawback is that most network studies so far have only analyzed data from one dataset. However, it is unclear how variable independently derived networks are and depend, for instance, on the technical platform or on the particular samples/individuals that were used to produce the dataset. We thus analyzed and compared here 10 different transcriptome datasets from individual human frontal lobe samples, which have been produced with different platforms (microarrays and RNA-Seq), and developed a method for integrating the coexpression wTO networks calculated from them into one consensus network of high confidence level.

Several reasons prompted us to especially focus on the role of gene regulatory factors (GRFs) in the consensus network of the frontal lobe. First, because GRFs regulate the expression of many genes, they are expected to be among the most important players in these networks and might provide important insights about the molecular mechanisms taking place in this tissue. Second, primate specific zinc finger genes with a *Krüppel*-associated box (KRAB-ZNFs) are also enriched among the genes expressed during frontal lobe development (Zhang et al., 2011), which leads to the hypothesis that at least some GRFs might contribute to human specific cognitive abilities. Third, we show here that GRFs are enriched among the candidate genes for ID and ASD, thus suggesting an important role of GRFs in the gene regulatory processes and circuitry of such cognitive disorders. Taken together, GRFs are thus good candidates for providing essential information about the molecular mechanisms that set the stage for cognition.

To identify and analyze GRF proteins with potential implications in cognition in more detail, we used our in-house list of all 3315 human GRFs (Perdomo-Sabogal et al., under preparation). This catalog includes information from the most relevant studies in the area of human GRF inventories (see Section Materials and Methods), and includes information about proteins involved in different regulatory mechanisms such as DNA-binding proteins, cofactors that associate with transcription factors, histone and chromatin modifiers, among others. We also performed a comprehensive literature survey and compiled a list of 676 GRFs that are known to be important during human brain development or that have been associated with cognitive disorders. We will refer to this set of 676 GRFs as "Brain-GRFs" (Table S1). Using our high-confidence consensus network we identified here several GRFs, including 166 "Brain-GRFs" that are hubs and thus seem to be important for the gene regulatory processes in the human frontal lobe.

MATERIALS AND METHODS

Data Sets

The raw and processed data from microarrays and RNA-Seq were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Microarrays were analyzed using the R programming language and Bioconductor packages (Ihaka and Gentleman, 1996). For the microarrays, we determined gene expression levels (RMA values) and MAS5 detection p-value from the probes using the "affy" and "oligo" package, respectively of the platform used (Gautier et al., 2004; Carvalho and Irizarry, 2010). We considered only the probesets significantly detected in at least one individual ($p < 0.05$). Furthermore, for genes represented by more than one expressed probeset, we calculated the mean of the expression values of all its probesets. For the RNA-Seq data, we used published RPKM values when available (BrainSpan). Otherwise, we processed and analyzed the raw data by mapping of the reads using segemehl (Hoffmann et al., 2009) and calculating RPKM values using R programming language and R libraries such as GenomicRanges, GenomicFeatures, and Rsamtools (Lawrence et al., 2013). All the raw data were mapped to the hg19 genome. All expression values were then filtered for RPKM values > 0.5 for 90% of the samples. All samples were used from the following datasets: FrontalVal

[GSE25219] (Kang et al., 2011), NeoVal [GSE11512] (Somel et al., 2009), KhatVal [SRA028456] (Somel et al., 2011), and GexVal [GSE22521] (Liu et al., 2012). Only the data from the control individuals were selected from the DisVal [GSE53987], BipRval [GSE53239] (Akula et al., 2014), and BipVal [GSE5388] (Ryan et al., 2006) datasets. From the BrainSpan dataset we selected the samples from the frontal lobe regions and subset them such that individuals with same ages (13 total individuals per dataset) were used.

Catalog of Gene Regulatory Factor Proteins

The GRF catalog we used for building our GRFs consensus network of the human frontal lobe was initially built by Perdomo-Sabogal et al. (under preparation). For this catalog the information for 3315 GRF proteins sourced from the most seminal studies in the area of human GRF inventories (Messina et al., 2004; Vaquerizas et al., 2009; Ravasi et al., 2010; Nowick et al., 2011; Corsinotti et al., 2013; Tripathi et al., 2013; Wingender et al., 2013, 2015) that are associated with gene ontology terms for regulation of transcription, DNA-depending transcription, RNA polymerase II transcription cofactor and co-repressor activity, chromatin binding, modification, remodeling, or silencing, among others, were manually curated.

Gene Sets

The ASD gene list was compiled using the SFARI gene database (09/20/2015, 740 genes; Basu et al., 2009; Banerjee-Basu and Packer, 2010). In the analysis, we included all the 740 genes. In addition, we also calculated the overlap between GRFs and ASD genes with strong association with S category (syndromic) and strong evidence (levels 1–4). ASD modules (asdM12 and asdM16) were obtained from an independent genome-wide expression study that compared ASD with healthy post-mortem brain tissues (Voineagu et al., 2011).

GRFs with association with Parkinson's disease, Alzheimer's disease, and Schizophrenia where filtered according to their significant evidence in more than two GWAS studies (Allen et al., 2008; Bertram, 2009; Jia et al., 2010; Lill et al., 2012). Additional schizophrenia GRFs were derived from independent publication with 108 loci implicated in schizophrenia (Consortium SWGotPG, 2014). ID and FMRP targets genes were collected from independent publications (Inlow and Restifo, 2004; Ropers, 2008; Darnell et al., 2011; van Bokhoven, 2011; Lubs et al., 2012; Consortium SWGotPG, 2014).

Other brain related GRFs were manually selected using web sources such as OMIM and independent databases such as SGZR (Hamosh et al., 2005; Jia et al., 2010). We prioritize GRFs that have evidence on brain functions, synaptic transmission, and brain development.

wTO Calculation

Spearman rank correlations were used to correlate the expression values of the GRF genes with the expression values of all genes, separately in each of the 10 datasets. Note that only expressed genes were considered in each dataset and that the number of expressed GRFs and other genes differs between the datasets. We extracted all significant correlations ($p < 0.05$) for calculating the weighted topological overlap values ($\omega = wTO$) between all pairs of expressed GRF genes for each dataset as previously described (Nowick et al., 2009). The calculation is based on a real symmetric matrix $A = [a_{ij}]$, in which a_{ij} is a real number ranging between −1 to 1 that indicates the correlation coefficient between the i -th and j -th GRF in the dataset. In particular, we have $a_{ii} = 0$. Comparing with the previous method (Zhang and Horvath, 2005), our method incorporates both significant (Spearman rank correlation; $p < 0.05$) positive and negative correlations of two GRFs' correlated gene sets (u) described as follow: $a_{ij} \in [0, 1]$ when $a_{ij} \geq 0 \rightarrow \ a_{iu}a_{ju} \geq 0$ for all u and $a_{ij} \in [-1, 0]$ when $a_{ij} \leq 0 \rightarrow \ a_{iu}a_{ju} \leq 0$ for all u. This condition results in a positive wTO value for the GRFs i and j if they are both correlated in the same direction with u, while in a negative wTO value if i and j are correlated with u in the opposite direction.

Inserting the weighted connectivity of a node i as:

$$K_i = \sum\nolimits_i a_{ij},$$

and the connectivity between i and j as:

$C = A * A^T$, the weighted topological overlap is calculated as:

$$\omega_{ij} = \frac{c_{ij} + a_{ij}}{min\left(K_i, K_j\right) + 1 - \left|a_{ij}\right|}$$

To evaluate the reliability of each wTO network, we performed 100 permutations by randomizing the expression values of each individual. This effectively assigned a random expression value to each gene of a particular individual out of all the available gene expression values for that individual. The permutation was done separately for each individual. We then calculated 100 permuted wTO networks for each dataset. We determined the number of links in the empirically derived ("real") network for multiple wTO cutoffs [0.1:0.6] and compared it to the number of links with the same wTO cutoff in the 100 permuted networks. This method allowed us to determine a p-value for how different the empirical networks are from random expectation and to calculate a false positive rate for the links in each network. All empirically derived networks had more links at all tested wTO values compared to the permuted networks, demonstrating that the empirically derived networks are different from random expectation (Table S2D).

Consensus Network Construction

To construct the consensus network, we first analyzed the distributions of the wTO values of all GRF-GRF pairs across all datasets using the boxplot.stats function in R (Williamson et al., 1989) to have an overall view of the data sets. Our results show that the distributions of wTO values of the datasets BipRVal, DisVal, and FrontalVal are different from the other datasets (**Figure 2**). Based on these observations, we chose the Wilcoxon rank sum test for our subsequent analysis, since it is a non-parametric test and hence robust against outliers. Thus, we are able to construct the consensus network by taking all the wTO values from all the datasets into consideration. Furthermore, to identify significant GRF-GRF pairs, we performed another Wilcoxon rank sum test with alternative hypothesis greater than

|wTO|> 0.3. By applying this test, we avoided potential false positive links due to high variation of wTO values across the datasets. If the result was significant ($p < 0.05$), we considered this GRF-GRF pair as a significant pair. For each of these detected significant GRF-GRF pairs, we then calculated its consensus wTO value as the median of all 10 individual wTO values. Note here, we opted for |wTO|>0.3 as cutoff in the hypothesis, because this was the mean of the cutoffs at which the 10 networks differed from random expectation with $p < 0.01$.

Network Visualization

For network visualization, we used Cytoscape 3.0. Node attributes were used according to our manually curated Brain-GRF list, the Human Proteome map (Kim et al., 2014), and the FMRP targets (Darnell et al., 2011). We included the Cytoscape session (the file is publically available on http://www.nowicklab.info/?page_id=470) for manual visualization of GRF-GRF interactions as additional file.

Statistics

For gene set enrichments, *p*-values were calculated with a one-sided Fisher's exact test function in R (alternative = "g," confidence level = 0.99, simulated *p*-value with 1000 replicates). A one-sided Wilcoxon ranked test was implemented to evaluate the enrichment of the connectivity between species (alternative = "g," confidence level = 0.99, paired = FALSE). *P*-values for overlaps were calculated with hypergeometric tests using a custom made R script. We retained an independent background (BrainSpan expressed gene = 15585 genes). *P*-values were subsequently adjusted for multiple comparisons using Benjamini-Hochberg FDR procedure. Two-way permutation test of 1000 was performed to validate the overlaps. First we randomized the external gene sets (e.g., ASD genes) by randomly selecting the same number of genes from an independent brain expressed genes list (e.g., BrainSpan gene set) and subsequently calculating the overlap *p*-values with the GRF gene set. The second approach randomized the internal gene sets (e.g., GRF gene set) by randomly selecting the same number of genes as GRFs that were expressed and subsequently calculating the overlap *p*-values. Analysis for RNA-seq, microarray, and correlation filtering were performed using custom made R and SQL scripts. To calculate the correlation and wTO, we developed a Java-based program.

Enrichment for Transcription Factor Binding Sites (TFBS)

For the TFBS enrichment, we focused on the 5421 genes that are expressed in all datasets and correlated with at least one GRF in each of the 10 different datasets. To test whether correlated genes might be target genes of the respective GRF, we performed a ChIP Enrichment Analysis (ChEA) using the ENCODE database and data from Chip-Seq, Chip-Chip, Chip-PET, and DamID experiments (Lachmann et al., 2010). We also performed a TFBS enrichment analysis using the Jolma and JASPAR databases (Jolma et al., 2013; Mathelier et al., 2014). We tested for enrichment of TFBSs included in those databases within the 2 kb upstream region of the 5421 genes using

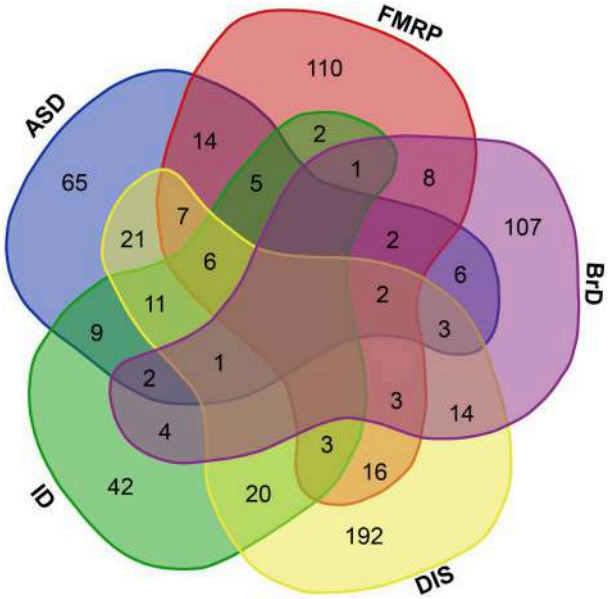


FIGURE 1 | Brain-GRFs association. Overlap between GRFs implicated in autism (ASD) or intellectual disability (ID), GRFs that are FMRP targets (FMRP), GRFs involved in brain development and functions (BrD), and GRFs implicated in syndromes or disorders (DIS). Empty space represents no overlap between sets. The overlap shows the commonalities of GRFs implicated in multiple disorders and syndromes.

CentriMo (default parameters) implemented in the MEME suite (Bailey et al., 2009; Bailey and Machanick, 2012). As background, we used the 2 kb upstream regions of the remaining protein coding genes and CpG islands.

Protein–Protein-Interactions Enrichment

Protein–Protein-Interactions (PPIs) were compiled from BioGRID and InWeb using the method described in Parikshak et al. (2013). We used the set of 5421 genes commonly expressed in all 10 datasets. Then we determined the GRF-gene pairs that were called to interact as proteins according to BioGRID and InWeb (Rossin et al., 2011; Chatr-Aryamontri et al., 2013). GRF-gene pairs that were present in each of the 10 datasets and were indicated to interact as proteins were then combined to a consensus PPI network. Fisher's exact test was used for testing the enrichment of PPI in Brain-GRFs and other GRFs.

GO Enrichment

For the GO enrichment analysis in the consensus network, we first ranked the genes of each dataset according to the number of GRFs they were correlated with. Then we summed up the ranks across the 10 datasets. The ranked list of the sums of the ranks was used as input for the Wilcoxon test implemented in FUNC (Prüfer et al., 2007) for the GO enrichment analysis. This method allowed us to understand the relative importance of a gene in each dataset according to the rank position. We next summarized the ranks across the 10 datasets, thus obtaining a general rank (rank-sum). The GO enrichment test was performed using FUNC

TABLE 1 | Platforms description.

Dataset	Names	Assession number	Sample	Type	Permutation (|wTO|)
BipRVal	*Bipolar RNA-seq Values*	GSE53239	11	Adult	>0.39
BipVal	*Bipolar Microarray Values*	GSE5388	31	Adult	>0.24
DfcVal	*DFC RNA-seq Values*	BrainSpan	13	Developmental	>0.36
DisVal	*Disorder Microarray Values*	GSE53987	19	Adult	>0.30
FrontVal	*Frontal Pole Microarray Values*	GSE25219	348	Developmental	>0.10
GexVal	*Gene Expression Microarray Values*	GSE22521	25	Developmental	>0.25
KhatVal	*RNA-seq Values from a Khaitovich study*	SRA028456	12	Developmental	>0.36
NeoVal	*Neoteny Microarray Values*	GSE11512	44	Developmental	>0.19
OfcVal	*OFC RNA-seq Values*	BrainSpan	13	Developmental	>0.37
VfcVal	*VFC RNA-seq Values*	BrainSpan	13	Developmental	>0.37

We used multiple platforms to uncover the GRF—GRF interactions.
The first column represents the chosen name of each dataset.
The second column showed complete name of the dataset, the used platform, and Values as indication for the wTO calculation.
The third column contains the accession numbers of each dataset.
The fourth column indicates the number of samples used for the analysis.
The fifth column indicates the type of dataset.
The sixth column shows the |wTO| cutoff at which the dataset has significantly more links in the empirical network than in the permuted ones.
For BipRval, BipVal, DisVal, we used only the control samples consisting of healthy individuals (see Materials and methods).

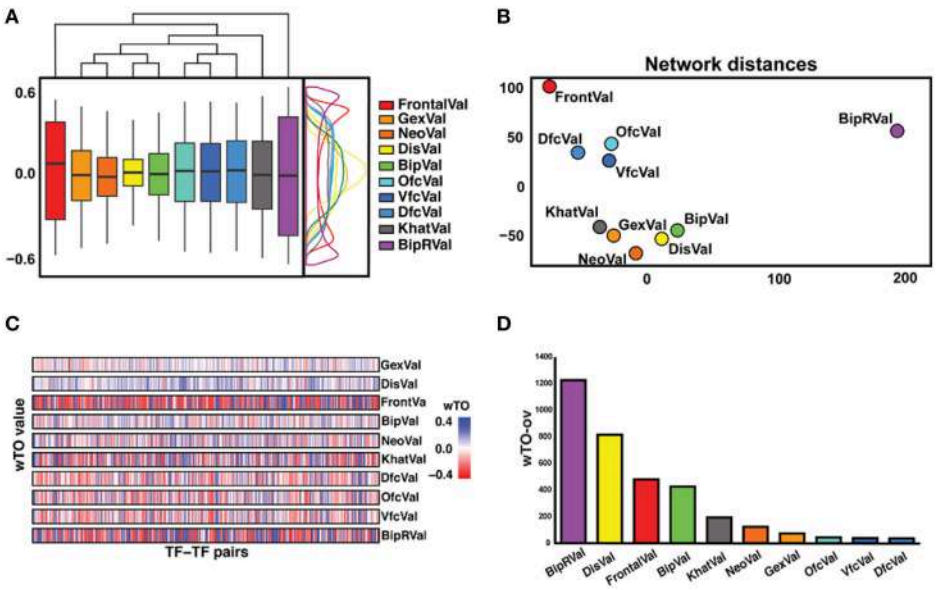


FIGURE 2 | Overview of differences and similarities between datasets. (A) Representation of the distribution of the wTO values of the 10 datasets. On the right side, a wTO density plot. On the top, a clustering map of the datasets showing FrontalVal and BipRVal as outliers compared with the remaining datasets. **(B)** Two-dimensional scaling plot in which the circles represent the datasets used in this study. The BipRVal dataset is the most different dataset compared to the other datasets. The three BrainSpan datasets (DfcVal, OfcVal, VfcVal) cluster together. The microarray datasets (GexVal, NeoVal, DisVal, BipVal) showed a consistent clustering with one additional RNA-seq dataset (KhatVal). FrontalVal is not clustering with any of the other microarray or RNA-Seq datasets. This clustering suggests that the wTO networks do not simply cluster according to experimental platforms. **(C)** Overall stripe chart of the wTO values across the 10 datasets. Red represents negative wTO values whereas blue represents positive wTO values. As also seen in **Figure 2A**, FrontalVal and BipRVal wTO values differ most from the other datasets. **(D)** Barplot representing the numbers of detected wTO outlier values (wTO-ov) per dataset. BipRVal contained the highest number of outliers underlining it as being the most distant dataset.

(Prüfer et al., 2007). We used a Wilcoxon rank-based test for GO enrichment among the genes with highest rank-sums. For the GO analyses we only analyzed GO groups with at least 20 genes per group. We report GO groups with enrichment with $p < 0.01$ before and after refinement.

For the analysis of GO enrichment within each individual network among genes correlated with the selected Brain-GRF hubs we collected for each hub its correlated genes in all the 10 datasets. The remaining set of expressed genes was used as background set. We used the hypergeometric test implemented

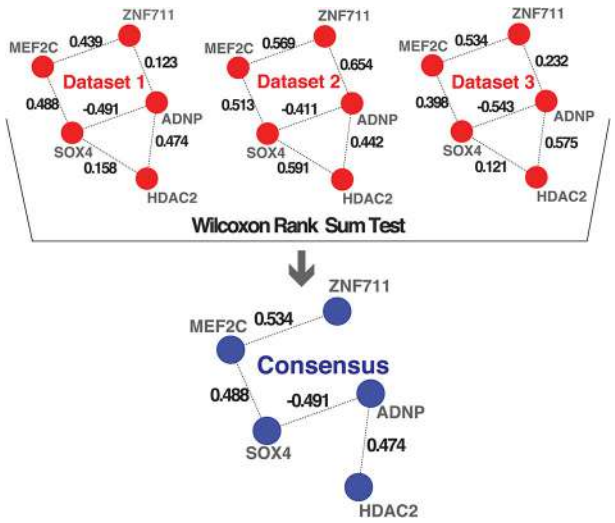


FIGURE 3 | Consensus method. Schematic representation of the method we implemented for combining multiple networks into a consensus network. The examples shown in the first part highlight hypothetical interactions present in three independent datasets. The numbers on the links represent the wTO values calculated using our method. We performed a Wilcoxon rank sum test to statistically determine which links had wTO values that were significantly higher than a chosen cutoff (|wTO|> 0.3) across all datasets. The blue network represents the consensus network containing only these significant links. The numbers shown at the links of the consensus network are the median wTO values calculated from the respective links in the 10 datasets. The links that not full-filled our statistical criteria due to high variation between dataset and cutoff trimming were consequently excluded.

in FUNC for the GO enrichment analysis considering only GO groups with at least 20 genes per group. We report GO groups with enrichment with $p < 0.01$ before and after refinement. Finally, we summarized the 10 lists of significant GO categories into one single list, thus removing duplicated GO categories. We also parsed the analyzed GO categories into a list of developmental categories using CateGOrizer (Hu et al., 2008).

RESULTS

Gene Regulatory Factors Involved in Brain Development and Cognitive Disorders

Within this list of human GRFs we identified 676 GRFs that are involved in cognitive functions, brain development, and disorders by using different sources (see Materials and Methods; **Figure 1** and Table S1). A prevalence of genes coding for GRFs among genes associated with some cognitive disorders has been observed before (Hong et al., 2005; West and Greenberg, 2011; Parikshak et al., 2013; De Rubeis et al., 2014; Nord et al., 2015). We here tested if this observation represents a significant overrepresentation of GRF genes among genes implicated in cognitive disorders. Among the 401 genes implicated in ID, we identified 106 genes coding for GRFs, which represents a highly significant enrichment of GRFs among all ID genes (hypergeometric test, $p = 2.03 \times 10^{-07}$). The SFARI database (Basu et al., 2009; Banerjee-Basu and Packer, 2010) currently contains 740 genes implicated in autism. Among those, 297 genes show strong evidence of ASD association. We identified 154 GRFs among the 740 genes (78 among the 297 ASD genes with strong association), which demonstrates that there is also a highly significant overrepresentation of GRFs among genes associated with autism (hypergeometric test, $p = 0.0001$). We further investigated whether GRFs are enriched among the target genes of the Fragile-X Mental Retardation Protein (FMRP). This protein was previously shown to play an important role in ASD-pathways by exerting translational regulation during human brain development (Darnell et al., 2011). Among the set of 842 FMRP target genes predicted by HITs-CLIP, we identified 179 GRF genes revealing a significant overrepresentation of GRF genes (hypergeometric test, $p = 0.0001$). In addition, GRFs are also significantly enriched for genes highly expressed in neurons (hypergeometric test, $p < 0.001$) and astrocytes (hypergeometric test, $p < 0.05$) compared with other brain cell-type expressed genes (Zhang et al., 2014).

Taken together, these findings show that GRF genes are enriched among candidate genes for cognitive disorders and cell important for brain functions, metabolism, and structure. Therefore, they are likely to be good candidates for providing essential information about the molecular processes involved in the organization and functioning of neural circuits that support healthy cognitive abilities.

A Consensus Network of High Confidence

To investigate the roles of all GRFs in the frontal lobe, we analyzed 10 genome-wide expression datasets comprised of frontal lobe samples from individuals of different ages and obtained with different techniques (**Table 1**). We first analyzed each dataset independently to investigate the consistency of the coexpression networks derived from these independent datasets.

Specifically, from each dataset, we constructed a weighted topological overlap (wTO) network taking into account all expressed GRFs and their coexpressed genes (Nowick et al., 2009). For constructing this wTO network, we first identified all genes that are significantly correlated in expression (i.e., coexpressed) with a particular GRF. These genes include putative target genes and genes coding for interaction partners of that GRF. The wTO of a pair of GRFs then represents the commonality of these two GRFs in their sets of coexpressed genes. Because GRFs can function as activators or repressors of gene expression, we take into account the sign of the correlation when calculating the wTO. Pairs of GRFs with |wTO|values above a certain cutoff are connected by a link in the wTO network visualization (see Materials and Methods).

Even though each network is supported to significantly differ from random expectation, we noted differences between the 10 networks, for instance, in the distribution of the wTO values and when comparing the wTO values for particular links between the datasets (**Figures 2A,B**). The differences between the networks can probably be explained by biological variation between individuals, but also by technical variations such as in RNA extraction methods, RIN values, and RNA library preparation procedures. We observed that the dataset BipRVal differs the most from the other datasets by having the highest number of wTO outliers, followed by datasets DisVal and FrontalVal

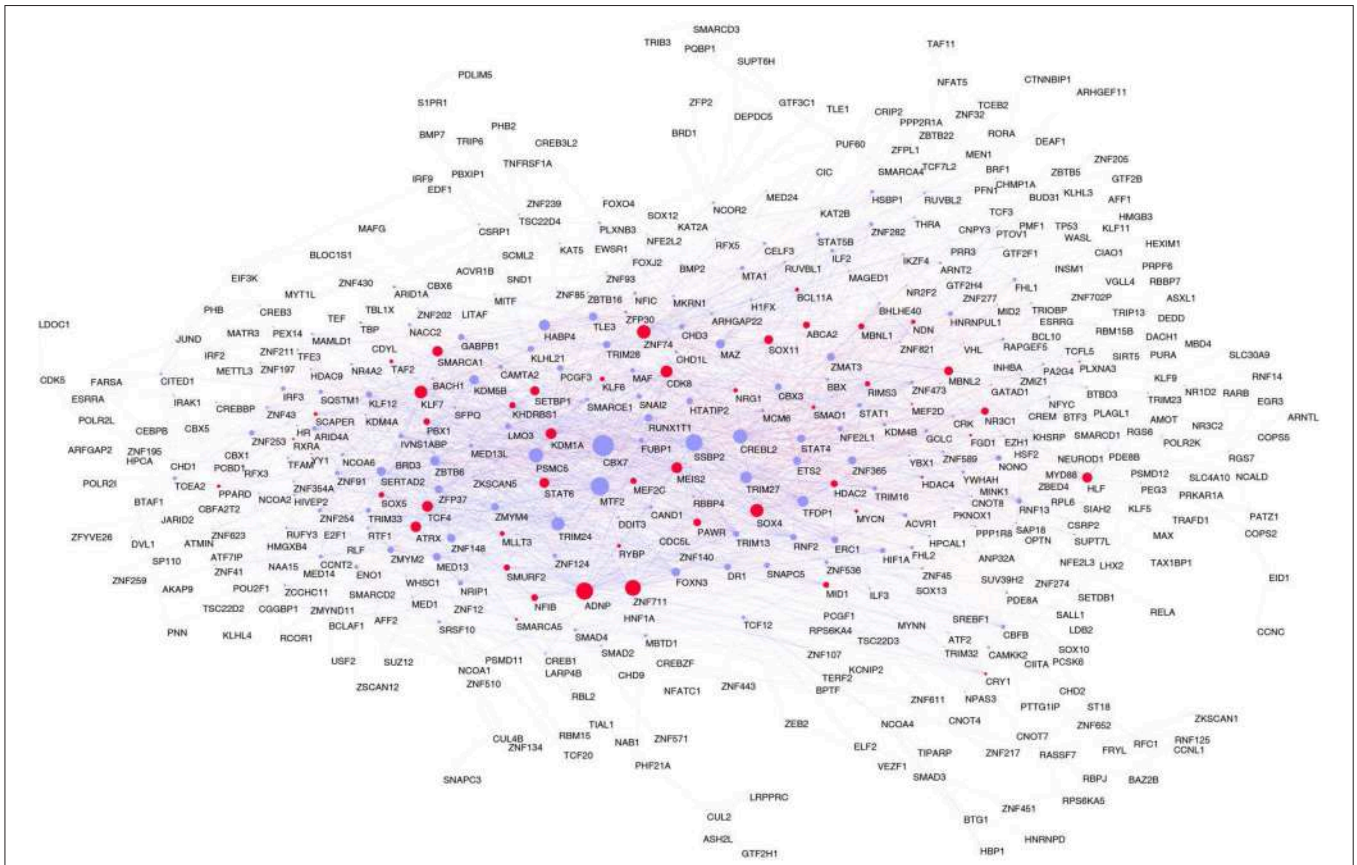


FIGURE 4 | Consensus network. In red Brain-GRFs; in blue all other GRFs. Node size is proportional to the number of links. Links with positive wTO values are in blue and links negative wTO values are shown in red.

(**Figures 2B,C**). All in all, we found that merely 19% (287930) of all links between GRFs are present in all 10 wTO networks. Given such variation between the networks, we think it is dangerous to rely on only one dataset when making inferences about biological relationships. Instead, multiple datasets should be combined to alleviate the dependence of the results on a particular set of individuals, developmental time points, different RNA library preparations, and gene expression measurement platforms and to focus on the most consistently observed links.

To combine the 10 independently derived networks into a consensus network with higher confidence, we considered them as biological replicates. We evaluated for each GRF—GRF pair, whether the distribution of strengths of their links across the 0 datasets is significantly higher than a chosen cutoff (Wilcoxon rank sum test, $p < 0.05$; **Figure 3** and see Materials and Methods). If so, the link was included into our consensus network. The resulting consensus network for $|wTO|>0.3$ consists of 2516 links (**Figure 4** and Tables S2A,B). This method allowed us to pinpoint the links with the strongest consistency across multiple networks. To determine the final weight of the links in the consensus network, we calculated the median of all wTO values for the respective GRF—GRF pair.

Brain-GRF Genes Are Often Hubs and Highly Interconnected in the Frontal Lobe Consensus Network

Focusing on the most consistent links as determined by our consensus network, we next analyzed how the known Brain-GRFs are integrated into this consensus network. Of the total of 676 Brain-GRFs, 166 are present in the consensus network. Interestingly, this represents a significant enrichment of Brain-GRFs among the 498 GRFs of the consensus network (Fisher exact test, $p = 1.79 \times 10^{-11}$, Odd Ratio = 2.2). Remarkably, the group of Brain-GRFs has a higher connectivity (number of links) compared to other GRFs in the consensus network (Wilcoxon rank sum test, $p = 0.015$). Those finding suggests that known Brain-GRFs have stronger and more consistent functional relationships amongst each other than other GRFs in the frontal lobe.

To confirm the transcriptional pathways suggested by our consensus network, we examined whether there is enrichment of the GRF binding sites in the regulatory sequences of the 5421 genes that are correlated with at least one of the 498 GRFs of the consensus network (Table S2C). To this end, we first performed a ChIP enrichment analysis (ChEA) using the updated ENCODE database and a manually curated list of

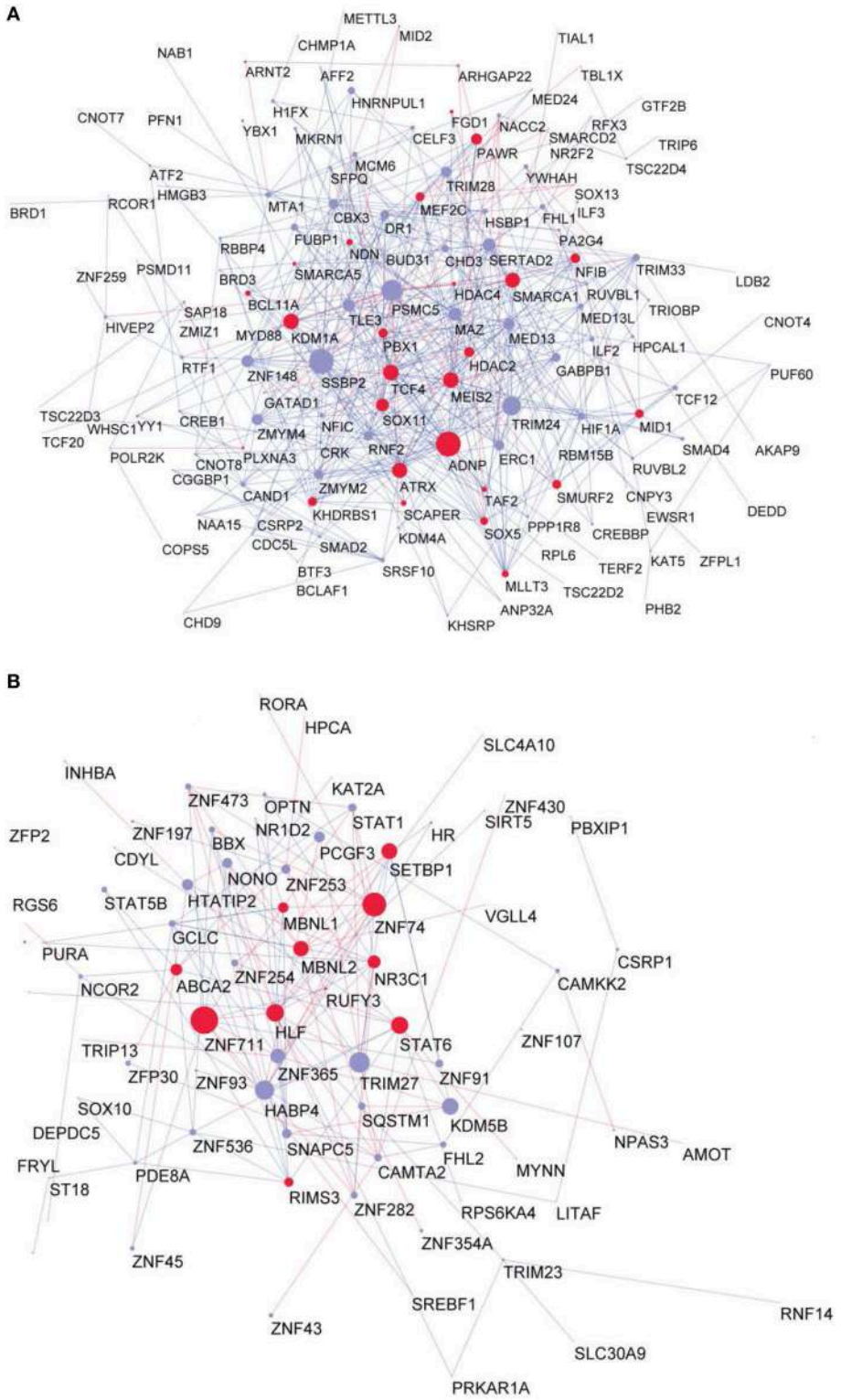


FIGURE 5 | Proteome GRF modules with red nodes representing the Brain-GRFs whereas in blue the other GRFs. Links with positive wTO values are in blue and links negative wTO values are shown in red. **(A)** Fetal module. **(B)** Adult module. Brain-GRFs are significantly enriched in the fetal module showing higher connectivity compared with the other GRFs.

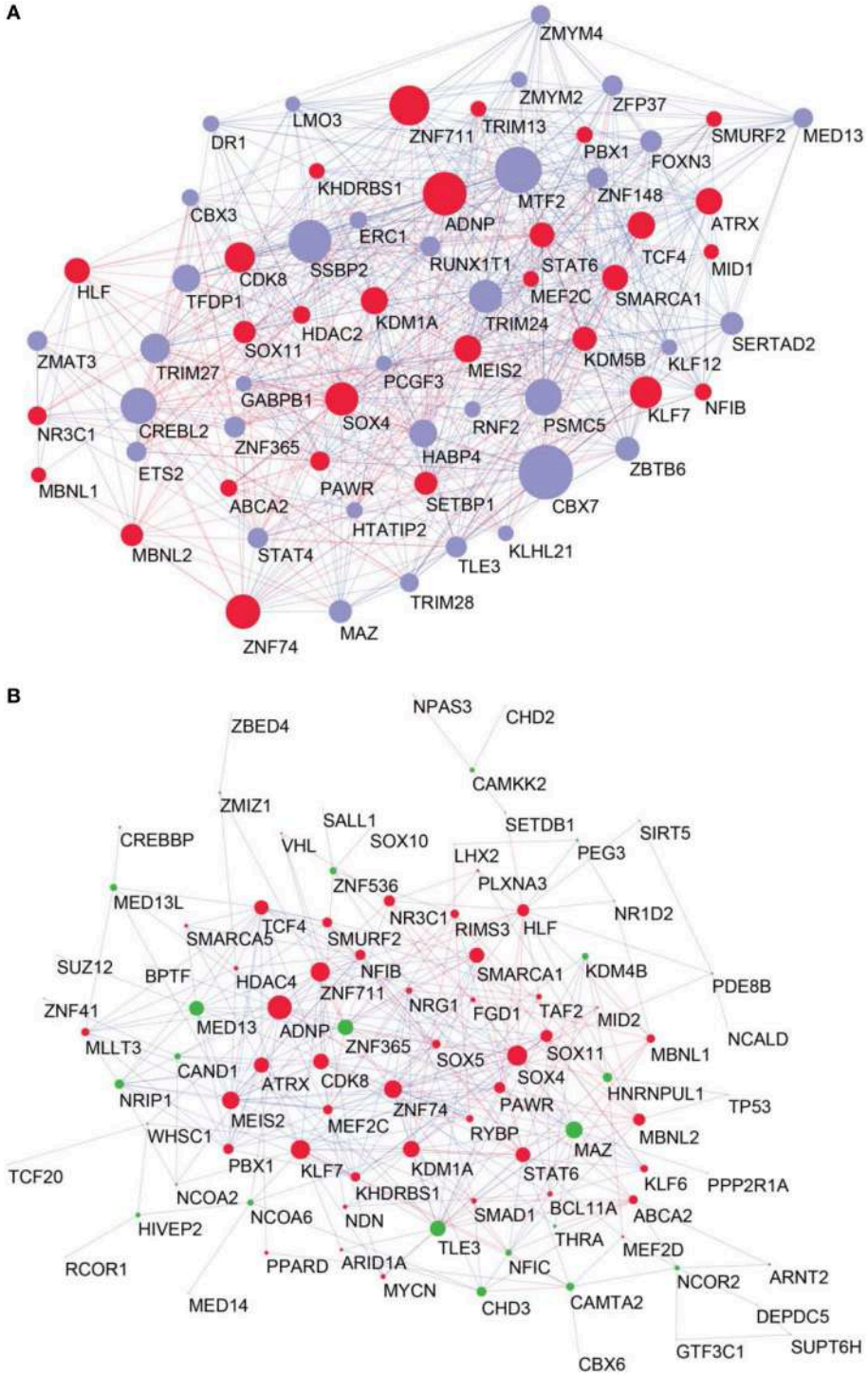


FIGURE 6 | High confident consensus network and proteomics networks. (A) Representation of the frontal lobe consensus network. Shown are the most highly connected hubs (degree > 25). Red nodes highlight Brain-GRFs, while blue nodes represent all other GRFs. The size of a node is proportional to its number of links: bigger nodes represent hubs in the network. Links with positive wTO values are in blue and links with negative wTO values are shown in red. **(B)** Brain-GRFs and FMRP targets module. Red nodes highlight the Brain-GRFs, while the green nodes highlight GRFs that are FMRP targets. The size of the nodes is proportional to their number of links.

target genes uncovered by ChIP-Seq, Chip-chip, ChIP-PET, and DamID from multiple studies (Lachmann et al., 2010). We found that the TFBS of 55 GRFs in the consensus network are significantly enriched among the regulatory sequences of the 5421 genes ($p < 0.05$ after Benjamini-Hochberg correction). Among those 55 GRFs, we found, for instance, *HDAC2* involved in synaptic plasticity and neural circuits (Guan et al., 2009), *ATF2* linked to neuronal apoptosis and cell migration (Yuan et al., 2009), and *CHD2* implicated in ASD and epilepsy (Rauch et al., 2012; Table S3A). Secondly, using the Jaspar and Jolma databases (Jolma et al., 2013; Mathelier et al., 2014), we found an enrichment of binding sites for 34 additional GRFs of the consensus network within the 2 kb region upstream of the transcription start site of the 5421 genes (Fisher exact test, $p < 0.05$ after Benjamini-Hochberg correction; Jolma et al., 2013; Mathelier et al., 2014). Here, we found enrichment for binding sites of *ARNTL*, important for circadian rhythm associated with BD (Nievergelt et al., 2006), *MEF2D*, involved in neuronal differentiation and PD (Yang et al., 2009), and *MEF2C*, involved in ASD, ID, and epilepsy (Novara et al., 2010) among others (Table S3B).

Coexpressed genes can also indicate protein interaction partners. Thus, we next examined protein—protein interactions (PPI) among the 498 GRFs and the 5421 correlated genes utilizing the annotations from BioGRID (Stark et al., 2006) and InWeb. We found that correlated GRF-gene pairs were significantly enriched within the PPI interactions (Fisher exact test, $p = 2.2 \times 10^{-6}$, Odd Ratio > 3), thus providing an additional confirmation of the potential functional interactions between GRFs and their correlated genes (Table S4).

In addition to the Brain-GRF enrichment, we examined the overlap between our consensus network with two coexpression modules, asdM12 and asdM16, that have been implicated in ASD previously (Voineagu et al., 2011). Remarkably, the consensus network overlaps significantly with the asdM12 module that is associated with synaptic development and dysregulated in ASD brains (hypergeometric test, $p = 0.045$). This result suggests that functional relationship of the GRFs in our consensus network plays a role in ASD.

To investigate whether the GRFs are also highly expressed at protein level in a fetal or adult brain, we superimposed our consensus network with a proteome map of the human brain at different stages, which was derived using mass-spectrometry proteomics (Kim et al., 2014). This strategy allowed us to understand the potential roles of the GRFs in the period of brain development and circuitry formation compared with an adult brain. Interestingly, overall the GRFs of our consensus network have higher expression and significantly more links in the fetal module compared to the adult module (Wilcoxon rank sum test, $p = 0.006$). The known Brain-GRFs are specifically enriched in the fetal module (Fisher exact test, $p = 0.03$, OR = 1.5) with generally higher number of links in comparison to other GRFs (Wilcoxon rank sum test, $p = 0.002$; **Figures 5A,B**).

To determine the most important GRFs in the consensus network of the human frontal pole, we determined the GRFs with the highest numbers of links (**Figures 6A,B**). Examples of such hubs include *ADNP*, *ZFN711*, *ZNF74*, and *SOX4*, which are all Brain-GRFs. Interestingly, those Brain-GRFs are also strongly interconnected with other Brain-GRFs (e.g., *MEF2C*, *PBX1*, *SMARCA1*, an *SOX11*) and GRFs that are FMRP-targets (e.g., *KDM4B*, *MED13*, *NRIP1*, and *ZNF365*), suggesting a high functional interrelationship between various Brain-GRFs (**Figure 7**). Of note, in addition to the Brain-GRFs, the consensus network also contained hubs that yet are not implicated in brain functions or disorders. For example we detected GRFs important for embryogenesis (e.g., *CBX7*, *TFDP1,* and *TLE3*; Dehni et al., 1995; Morey et al., 2013; Laing et al., 2015) and energy metabolism (e.g., *PSMC5* and *SERTAD2*; Hoyle et al., 1997; Liew et al., 2013). Due to their strong connectivity to known Brain-GRFs in the consensus network, it seems likely that also these GRFs play an important role in the human frontal lobe circuitries. Taken together, our results suggest GRFs that are important for shaping the transcriptional circuitry of the human frontal lobe, including novel candidates for experimental validation of their roles at brain level and potential association with cognitive disorders.

To infer more about the functions of the GRFs in the consensus network, we performed a Gene Ontology (GO) enrichment analysis among the genes correlated with the GRFs (see Materials and Methods). We found significant enrichment for genes involved in metabolism, signaling, transport, translation, and RNA splicing (**Figure 8A**). We also specifically tested for GO enrichment of the genes correlated with three Brain-GRFs that are the strongest hubs in the consensus network: *ADNP*, *ZNF711,* and *ZNF74* (see Materials and Methods). Overall, we found similar GO groups enriched for these hubs like we did for the consensus network as a whole. However, there were also hub-specifically enriched GO categories such as brain development, methylation, and regulation of synaptic transmission, which suggests a specific role of these three GRFs in the regulation of genes important for these particular brain functions (**Figures 8B–D**; Table S5).

DISCUSSION

Comprehending the characteristic complexity of cognitive disorders, such as ASD and ID, still represents a challenge in neurosciences. An important step toward understanding this complexity is to elucidate the molecular networks of healthy human brains. In this study, we specifically compiled a set of 676 "Brain-GRF" genes implicated in brain development and cognitive disorders and analyzed their co-expression networks to gain first insights into which gene regulatory pathways these genes may be involved in in the frontal lobe of healthy individuals. Importantly, we discovered that networks derived from independent studies differ considerably from each other, highlighting a potential danger of relying on just one dataset. After combining these independent networks into a consensus network containing the links that are the most conserved across them, we were able to identify robust relationships between GRFs in the coexpression network of the frontal lobe of healthy human individuals. We further discovered that, while

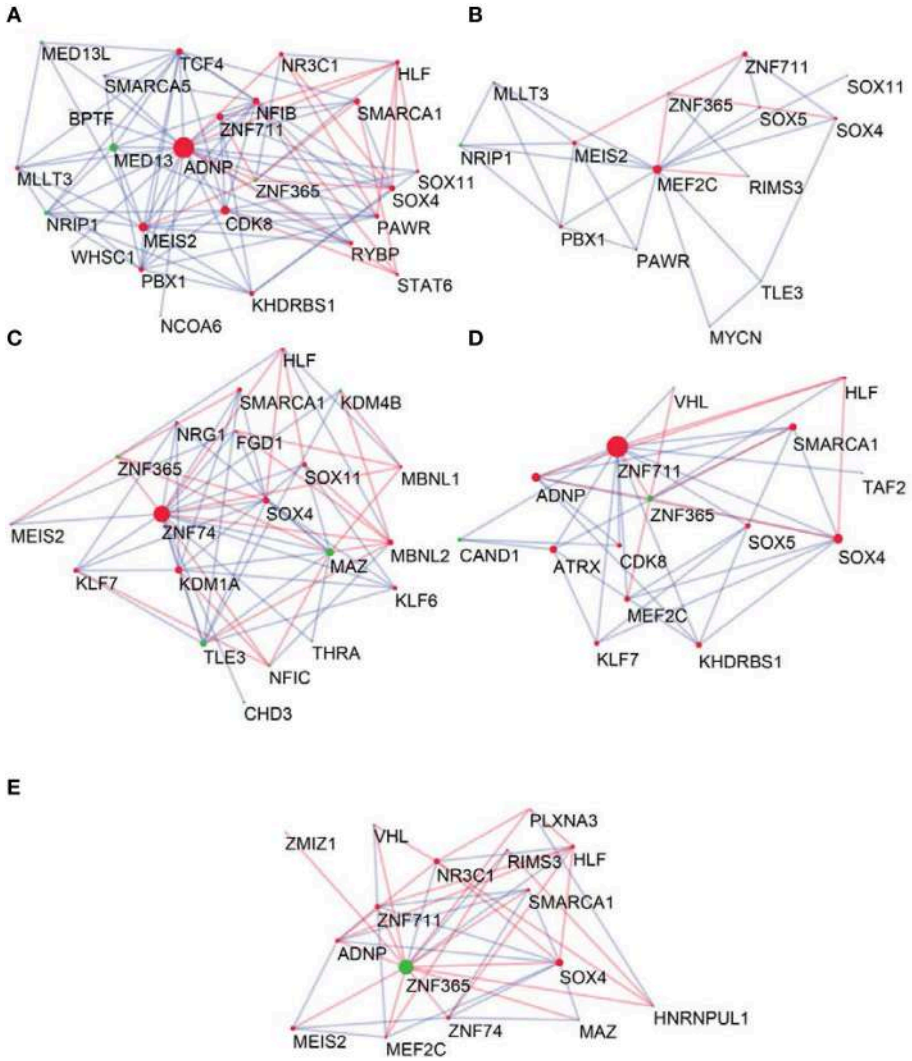


FIGURE 7 | Neighbors of hub Brain-GRFs and their strongly connected partners. (A) ADNP module, **(B)** MEF2C module, **(C)** ZNF74 module, **(D)** ZNF711 module, and **(E)** ZNF365 module. Red nodes highlight Brain-GRFs whereas green nodes represent FMRP targets. Links with positive wTO values are in blue and links negative wTO values are shown in red. Each hub Brain-GRFs is interestingly associated with other known Brain-GRFs highlighting potential interactions and common pathways.

some hubs in the consensus network are known "Brain-GRF" genes, others have not been linked to functions in the brain before.

The function of most GRFs is still only insufficiently characterized. However, insights into the functions and interactions of our human frontal lobe consensus network can be gained from the expression patterns of the GRFs, the GO enrichment of the genes correlated with the GRFs, and disorders the GRFs have been associated with. Many hubs of the consensus network are also expressed in tissues other than brain. However, we observed that a considerable number of them (115 in total), for example *ZNF711*, *ADNP, MEF2C, SOX11*, and *CBX7*, have higher expression in mouse neurons than in other brain cells, such as glia, astrocytes, oligodendrocytes, myelinating oligodendrocytes, and endothelial cells (Zhang et al., 2014), suggesting that they have an essential role in neurons. In addition, we also discovered that the GRFs of our network play dominant roles in the fetal proteome module, (Kim et al., 2014) supporting the reasoning that these GRFs might regulate important processes during brain development such as forming the necessary brain structures for proper brain functions, including cognitive functions. Despite being ubiquitously expressed, it is plausible that some GRFs might only be hubs in the frontal lobe, a possibility that needs to be investigated further when data becomes available.

Our GO analysis revealed that the hub GRFs of the frontal lobe consensus network are likely to regulate genes involved in splicing, translation, metabolism, signaling, and synaptic transmission in the frontal lobe. Interestingly, these GO categories seem to be important for several brain functions.

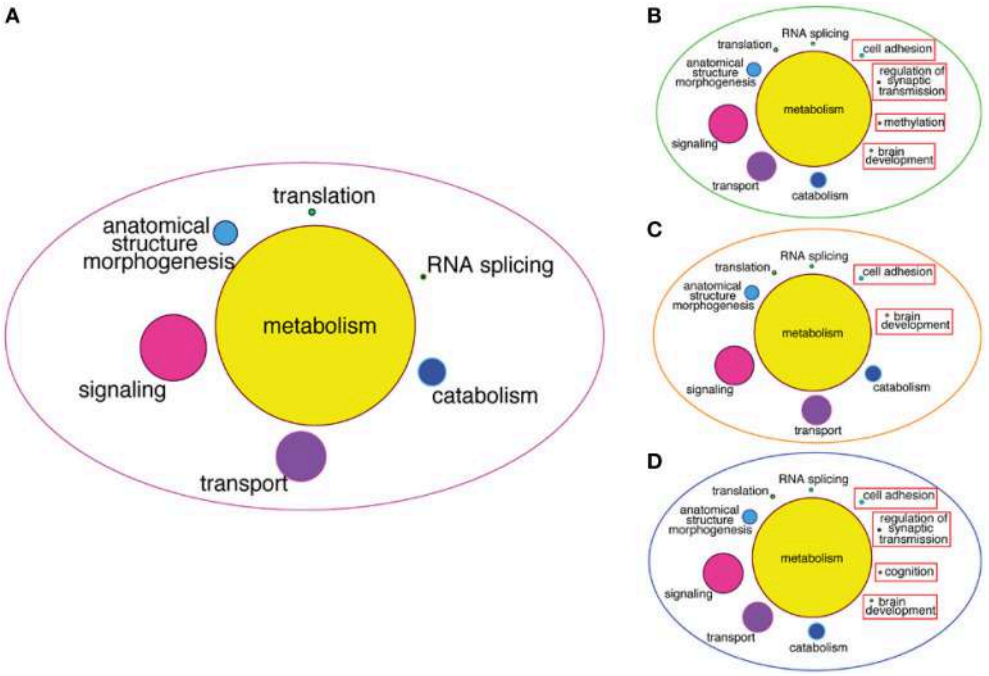


FIGURE 8 | GO enrichment among correlated genes of the consensus network and of Brain-GRF hubs. (A) GO categories that are enriched among the correlated genes of the GRFs of the consensus network. The categories for metabolism represent 46% of the enrichment. **(B–D)** GO categories enriched among the correlated genes of three selected Brain-GRF hubs of the consensus network (ADNP, ZNF711, and ZNF74, respectively). Interestingly, Brain-GRFs showed specific enrichment for categories involved in cognition and brain development.

For instance, translational mechanisms have been shown to play a role in the mechanisms of memory formation and synaptic plasticity (Richter and Klann, 2009) and RNA splicing mechanisms have been implicated in neuronal development (Li et al., 2007; Weyn-Vanhentenryck et al., 2014). Genes involved in metabolism might be important to provide the brain with the necessary energy for its functions. Signaling and synaptic transmission are important for the communication between neurons and relevant to allow for cognitive abilities. We thus suggest the interactions of the GRFs in the frontal lobe network are critically underlying the regulatory processes that allow for these vital brain functions.

We found a significant enrichment of known Brain-GRFs, including GRFs implicated in ASD, ID, or SY in our consensus network, indicating that it forms the basis for setting the stage for healthy cognitive abilities. For instance, the three strongest hubs are *ZNF711*, associated with ID (Tarpey et al., 2009), *ADNP*, involved in ID and ASD (Helsmoortel et al., 2014; Iossifov et al., 2014), and *ZNF74*, involved in ID and SY (Ravassard et al., 1999). Being in these central network positions presumably renders them to risk genes that increase the likelihood for developing brain disorders. We speculate that interaction between *ZNF711* and *ZNF74* reflect biological pathways that might be important for intellectual abilities. In line with this potential, genes correlated with *ZNF711* and *ZNF74* are enriched for functions such as axon development, brain development and regulation of synaptic transmission, which are likely important for the development and maintenance of healthy cognitive skills. Another hub in our GRF consensus network is *MEF2C*, a GRF that is important for synaptic plasticity and has been implicated in ASD (Ebert and Greenberg, 2013). *MEF2C* is also strongly associated with other Brain-GRFs such as *ZNF711*, *SOX11*, and *SOX5*, defining a strongly interconnected module of GRFs involved in regulatory pathways that might control cognitive functions (Uwanogho et al., 1995; Jankowski et al., 2006; Tarpey et al., 2009; Schanze et al., 2013). Our analysis highlighted also hubs that are targeted by FMRP, pointing to pathways that might be (dys)regulated at the post-transcriptional level. For instance, *CREBBP*, a GRF associated with ASD and ID (Barnby et al., 2005), *HDAC4*, implicated in ID and ASD (Pinto et al., 2014), *ZNF365*, which has been discovered in a module strongly associated with ASD in a brain expression study (Voineagu et al., 2011), and *KDM5B* and *KDM4B*, recently implicated in ASD using another weighted network approach (TADA; De Rubeis et al., 2014; Iossifov et al., 2014). *CREB* transcription factors and *HDAC4* are further known to regulate synaptic plasticity and memory formation (Silva et al., 1998; Hardingham et al., 2001; Vecsey et al., 2007; Thomson et al., 2008; Kim et al., 2012; Sando et al., 2012). These observations lead us to speculate that Brain-GRFs are strongly dependent on each other by sharing functional pathways and target genes. Further experimental studies are needed to identify shared targets of these and other GRFs to confirm their role in human frontal lobe functions and disorders.

Supporting our speculation that Brain-GRFs depend on each other, we found that Brain-GRFs have significantly more links than other GRFs and are strongly interconnected in the human frontal lobe network. Importantly, in addition to 30 known Brain-GRFs that are hubs, we identified further 36 GRF genes that are hubs in the frontal lobe consensus network but were not included in our Brain-GRFs list. Interestingly, one of these hubs, *GABPB1* encodes for a subunit of the hetero-tetrameric GABP consisting of two GABPA and two GABPB subunits (Batchelor et al., 1998). GABPA was recently found to bind human-specific binding sites and regulate gene expression of at least four genes (*ALDOA*, *HSPA8*, *TP73,* and *TMBIM6*) that have been associated with cognitive diseases such as autism, AZ, PD and other brain disorders (Perdomo-Sabogal et al., 2016). To 0 explore if more of these hubs might be associated with brain functions, we mined the (non-curated) data from DisGeNET (Piñero et al., 2015). We found that at least 12 of these hub GRFs may be connected with mental diseases and other neurological pathologies such as AZ (*DR1*, *ETS2*, *TFDP1*, *and TRIM13*), PD (*RUNX1T1*), SZ (*ZNF365*), developmental verbal dyspraxia (*ERC1*) and central neuroblastoma (*LMO3*, *PSMC5*, *TRIM13*, *TRIM24, ZMAT3*), among others. This suggests that with our method we have potentially identified novel candidates for being associated with important, if not essential, functions in the brain. We speculate that sequence and regulatory changes altering the regulatory activity or expression of these 36 hub GRFs could have medical relevance. It would thus be highly interesting to experimentally investigate their functions at brain level.

The structure and organization of the consensus network we are presenting here provides insights into regulatory circuits of the human frontal lobe. However, a yet unanswered question is how the network that we described for the human frontal lobe differs from the network of other brain regions, tissues or species. We expect that the relevant data for addressing this question will become available soon. We also expect that more GRFs will be discovered to be involved in brain functions. In future studies similar strategies as we presented here can then be implemented to enrich our knowledge about the molecular basis and regulatory networks underlying cognitive abilities.

AUTHOR CONTRIBUTIONS

SB designed and executed research; AP contributed material; DG contributed analysis programs and visualizations; JQ contributed methods and designed research; KN designed research; All authors wrote and discussed the manuscript.

FUNDING

This work was supported by a grant from Volkswagen Foundation within the initiative "Evolutionary Biology" awarded to KN and by a fellowship from the Departamento Administrativo de Ciencia, Tecnologia e Innovacion Colciencias from Colombia, calls Francisco Jose de Caldas 497/2009 awarded to AP. This work was funded by the Austrian Science Fund (FWF): M1619-N28 awarded to JQ.

ACKNOWLEDGMENTS

The authors would like to thank Genevieve Konopka for helpful comments and discussions. The authors would also like to thank Neelroop Parikshak for sharing with us the protein–protein-interaction manually curated database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00031

REFERENCES

Akula, N., Barb, J., Jiang, X., Wendland, J. R., Choi, K. H., and Sen, S. K. (2014). RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol. Psychiatry* 19, 1179–1185. doi: 10.1038/mp.2013.170

Akula, N., Wendland, J. R., Choi, K. H., and McMahon, F. J. (2016). An integrative genomic study implicates the postsynaptic density in the pathogenesis of bipolar disorder. *Neuropsychopharmacology* 41, 886–895. doi: 10.1038/npp.2015.218

Allen, N. C., Bagade, S., McQueen, M. B., Ioannidis, J. P., Kavvoura, F. K., Khoury, M. J., et al. (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.* 40, 827–834. doi: 10.1038/ng.171

Andreasen, N. C. (1995). Symptoms, signs, and diagnosis of schizophrenia. *Lancet* 346, 477–481. doi: 10.1016/S0140-6736(95)91325-4

Bailey, A., Phillips, W., and Rutter, M. (1996). Autism: towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *J. Child Psychol. Psychiatry* 37, 89–126. doi: 10.1111/j.1469-7610.1996.tb01381.x

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bailey, T. L., and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* 40, e128–e128. doi: 10.1093/nar/gks433

Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. *Dis. Model. Mech.* 3, 133–135. doi: 10.1242/dmm.005439

Barnby, G., Abbott, A., Sykes, N., Morris, A., Weeks, D. E., Mott, R., et al. (2005). Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. *Am. J. Hum. Genet.* 76, 950–966. doi: 10.1086/430454

Basu, S. N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Res.* 37, D832–D836. doi: 10.1093/nar/gkn835

Batchelor, A. H., Piper, D. E., de la Brousse, F. C., McKnight, S. L., and Wolberger, C. (1998). The structure of GABPα/β: an ETS domain-ankyrin repeat heterodimer bound to DNA. *Science* 279, 1037–1041. doi: 10.1126/science.279.5353.1037

Berg, J. M., and Geschwind, D. H. (2012). Autism genetics: searching for specificity and convergence. *Genome Biol.* 13, 247. doi: 10.1186/gb-2012-13-7-247

Bertram, L. (2009). Alzheimers disease genetics current status and future perspectives. *Int. Rev. Neurobiol.* 84, 167–184. doi: 10.1016/S0074-7742(09)00409-7

Bullido, M. J., Artiga, M. J., Recuero, M., Sastre, I., Garcia, M. A., Aldudo, J., et al. (1998). A polymorphism in the regulatory region of APOE associated

with risk for Alzheimers dementia. *Nat. Genet.* 18, 69–71. doi: 10.1038/ng0198-69

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi: 10.1093/bioinformatics/btq431

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–D823. doi: 10.1093/nar/gks1158

Chayer, C., and Freedman, M. (2001). Frontal lobe functions. *Curr. Neurol. Neurosci. Rep.* 1, 547–552. doi: 10.1007/s11910-001-0060-4

Consortium SWGotPG (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi: 10.1038/nature13595

Corsinotti, A., Kapopoulou, A., Gubelmann, C., Imbeault, M., Santoni de Sio, F. R., Rowe, H. M., et al. (2013). Global and stage specific patterns of Kruppel-associated-box zinc finger protein gene expression in murine early embryonic cells. *PLoS ONE* 8:e56721. doi: 10.1371/journal.pone.0056721

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y., Mele, A., Fraser, C. E., et al. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146, 247–261. doi: 10.1016/j.cell.2011.06.013

Dehni, G., Liu, Y., Husain, J., and Stifani, S. (1995). TLE expression correlates with mouse embryonic segmentation, neurogenesis, and epithelial determination. *Mech. Dev.* 53, 369–381. doi: 10.1016/0925-4773(95)00452-1

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209–215. doi: 10.1038/nature13772

Duncan, J., Emslie, H., Williams, P., Johnson, R., and Freer, C. (1996). Intelligence and the frontal lobe: the organization of goal-directed behavior. *Cogn. Psychol.* 30, 257–303. doi: 10.1006/cogp.1996.0008

Ebert, D. H., and Greenberg, M. E. (2013). Activity-dependent neuronal signalling and autism spectrum disorder. *Nature* 493, 327–337. doi: 10.1038/nature11860

Ecker, C., Suckling, J., Deoni, S. C., Lombardo, M. V., Bullmore, E. T., Baron-Cohen, S., et al. (2012). Brain anatomy and its relationship to behavior in adults with autism spectrum disorder: a multicenter magnetic resonance imaging study. *Arch. Gen. Psychiatry* 69, 195–209. doi: 10.1001/archgenpsychiatry.2011.1251

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Greydanus, D. E., and Pratt, H. D. (2005). Syndromes and disorders associated with mental retardation. *Indian J. Pediatr.* 72, 859–864. doi: 10.1007/BF02731116

Guan, J. S., Haggarty, S. J., Giacometti, E., Dannenberg, J. H., Joseph, N., Gao, J., et al. (2009). HDAC2 negatively regulates memory formation and synaptic plasticity. *Nature* 459, 55–60. doi: 10.1038/nature07925

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Hardingham, G. E., Arnold, F. J., and Bading, H. (2001). Nuclear calcium signaling controls CREB-mediated gene expression triggered by synaptic activity. *Nat. Neurosci.* 4, 261–267. doi: 10.1038/85109

Helsmoortel, C., Vulto-van Silfhout, A. T., Coe, B. P., Vandeweyer, G., Rooms, L., van den Ende, J., et al. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nat. Genet.* 46, 380–384. doi: 10.1038/ng.2899

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., et al. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* 5:e1000502. doi: 10.1371/journal.pcbi.1000502

Hong, E. J., West, A. E., and Greenberg, M. E. (2005). Transcriptional control of cognitive development. *Curr. Opin. Neurobiol.* 15, 21–28. doi: 10.1016/j.conb.2005.01.002

Hoyle, J., Tan, K. H., and Fisher, E. M. (1997). Localization of genes encoding two human one-domain members of the AAA family: PSMC5 (the thyroid hormone receptor-interacting protein, TRIP1) and PSMC3 (the Tat-binding protein, TBP1). *Hum. Genet.* 99, 285–288. doi: 10.1007/s004390050356

Hu, Z.-L., Bao, J., and Reecy, J. M. (2008). CateGOrizer: a web-based program to batch analyze gene ontology classification categories. *Online J. Bioinformatics* 9, 108–112. Available online at: http://onljvetres.com/geneontologyabs2008.htm; http://www.animalgenome.org/tools/catego/index.html

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* 5, 299–314.

Inlow, J. K., and Restifo, L. L. (2004). Molecular and comparative genetics of mental retardation. *Genetics* 166, 835–881. doi: 10.1534/genetics.166.2.835

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221. doi: 10.1038/nature13908

Jankowski, M. P., Cornuet, P. K., McIlwrath, S., Koerber, H. R., and Albers, K. M. (2006). SRY-box containing gene 11 (Sox11) transcription factor is required for neuron survival and neurite growth. *Neuroscience* 143, 501–514. doi: 10.1016/j.neuroscience.2006.09.010

Jia, P., Sun, J., Guo, A., and Zhao, Z. (2010). SZGR: a comprehensive schizophrenia gene resource. *Mol. Psychiatry* 15, 453–462. doi: 10.1038/mp.2009.93

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

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* 478, 483–489. doi: 10.1038/nature10523

Kaufman, L., Ayub, M., and Vincent, J. B. (2010). The genetic basis of non-syndromic intellectual disability: a review. *J. Neurodev. Disord.* 2, 182–209. doi: 10.1007/s11689-010-9055-2

Kim, M.S., Akhtar, M. W., Adachi, M., Mahgoub, M., Bassel-Duby, R., Kavalali, E. T., et al. (2012). An essential role for histone deacetylase 4 in synaptic plasticity and memory formation. *J. Neurosci.* 32, 10879–10886. doi: 10.1523/JNEUROSCI.2089-12.2012

Kim, M.S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., et al. (2014). A draft map of the human proteome. *Nature* 509, 575–581. doi: 10.1038/nature13302

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 26, 2438–2444. doi: 10.1093/bioinformatics/btq466

Laing, A. F., Lowell, S., and Brickman, J. M. (2015). Gro/TLE enables embryonic stem cell differentiation by repressing pluripotent gene expression. *Dev. Biol.* 397, 56–66. doi: 10.1016/j.ydbio.2014.10.007

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Li, Q., Lee, J.-A., and Black, D. L. (2007). Neuronal regulation of alternative pre-mRNA splicing. *Nat. Rev. Neurosci.* 8:819–831. doi: 10.1038/nrn2237

Liew, C. W., Boucher, J., Cheong, J. K., Vernochet, C., Koh, H.-J., Mallol, C., et al. (2013). Ablation of TRIP-Br2, a regulator of fat lipolysis, thermogenesis and oxidative metabolism, prevents diet-induced obesity and insulin resistance. *Nat. Med.* 19, 217–226. doi: 10.1038/nm.3056

Lill, C. M., Roehr, J. T., McQueen, M. B., Kavvoura, F. K., Bagade, S., Schjeide, B. M., et al. (2012). Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. *PLoS Genet.* 8:e1002548. doi: 10.1371/journal.pgen.1002548

Liu, X., Somel, M., Tang, L., Yan, Z., Jiang, X., Guo, S., et al. (2012). Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res.* 22, 611–622. doi: 10.1101/gr.127324.111

Lubs, H. A., Stevenson, R. E., and Schwartz, C. E. (2012). Fragile X and X-linked intellectual disability: four decades of discovery. *Am. J. Hum. Genet.* 90, 579–590. doi: 10.1016/j.ajhg.2012.02.018

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., et al. (2014). 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

Messina, D. N., Glasscock, J., Gish, W., and Lovett, M. (2004). An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* 14, 2041–2047. doi: 10.1101/gr.2584104

Morey, L., Aloia, L., Cozzuto, L., Benitah, S. A., and Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. *Cell Rep.* 3, 60–69. doi: 10.1016/j.celrep.2012.11.026

Nievergelt, C. M., Kripke, D. F., Barrett, T. B., Burg, E., Remick, R. A., Sadovnick, A. D., et al. (2006). Suggestive evidence for association of the circadian genes

PERIOD3 and ARNTL with bipolar disorder. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 141B, 234–241. doi: 10.1002/ajmg.b.30252

Nord, A. S., Pattabiraman, K., Visel, A., and Rubenstein, J. L. (2015). Genomic perspectives of transcriptional regulation in forebrain development. *Neuron* 85, 27–47. doi: 10.1016/j.neuron.2014.11.011

Novara, F., Beri, S., Giorda, R., Ortibus, E., Nageshappa, S., Darra, F., et al. (2010). Refining the phenotype associated with MEF2C haploinsufficiency. *Clin. Genet.* 78, 471–477. doi: 10.1111/j.1399-0004.2010.01413.x

Nowick, K., Fields, C., Gernat, T., Caetano-Anolles, D., Kholina, N., and Stubbs, L. (2011). Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS ONE* 6:e21553. doi: 10.1371/journal.pone.0021553

Nowick, K., Gernat, T., Almaas, E., and Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc. Natl. Acad. Sci. U.S.A.* 106, 22358–22363. doi: 10.1073/pnas.0911376106

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., et al. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 1008–1021. doi: 10.1016/j.cell.2013.10.031

Perdomo-Sabogal, A., Nowick, K., Piccini, I., Sudbrak, R., Lehrach, H., Yaspo, M. L., et al. (2016). Human lineage-specific transcriptional regulation through GA-binding protein transcription factor alpha (GABPa). *Mol. Biol. Evol.* doi: 10.1093/molbev/msw007. [Epub ahead of print].

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015:bav028. doi: 10.1093/database/bav028

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* 94, 677–694. doi: 10.1016/j.ajhg.2014.03.018

Polymeropoulos, M. H. (2000). Genetics of Parkinson's disease. *Ann. N.Y. Acad. Sci.* 920, 28–32. doi: 10.1111/j.1749-6632.2000.tb06901.x

Prüfer, K., Muetzel, B., Do, H. H., Weiss, G., Khaitovich, P., Rahm, E., et al. (2007). FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinform.* 8:41. doi: 10.1186/1471-2105-8-41

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674–1682. doi: 10.1016/S0140-6736(12)61480-9

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044

Ravassard, P., Côté, F., Grondin, B., Bazinet, M., Mallet, J., and Aubry, M. (1999). ZNF74, a gene deleted in DiGeorge syndrome, is expressed in human neural crest-derived tissues and foregut endoderm epithelia. *Genomics* 62, 82–85. doi: 10.1006/geno.1999.5982

Richter, J. D., and Klann, E. (2009). Making synaptic plasticity and memory last: mechanisms of translational regulation. *Genes Dev.* 23, 1–11. doi: 10.1101/gad.1735809

Ropers, H. H. (2008). Genetics of intellectual disability. *Curr. Opin. Genet. Dev.* 18, 241–250. doi: 10.1016/j.gde.2008.07.008

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., et al. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* 7:e1001273. doi: 10.1371/journal.pgen.1001273

Ryan, M. M., Lockstone, H. E., Huffaker, S. J., Wayland, M. T., Webster, M. J., and Bahn, S. (2006). Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol. Psychiatry* 11, 965–978. doi: 10.1038/sj.mp.4001875

Sando, R.IIIrd, Gounko, N., Pieraut, S., Liao, L., Yates, J., and Maximov, A. (2012). HDAC4 governs a transcriptional program essential for synaptic plasticity and memory. *Cell* 151, 821–834. doi: 10.1016/j.cell.2012.09.037

Schanze, I., Schanze, D., Bacino, C. A., Douzgou, S., Kerr, B., and Zenker, M. (2013). Haploinsufficiency of SOX5, a member of the SOX (SRY-related HMG-box) family of transcription factors is a cause of intellectual disability. *Eur. J. Med. Genet.* 56, 108–113. doi: 10.1016/j.ejmg.2012.11.001

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of *de novo* copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659

Shettleworth, S. J. (2009). *Cognition, Evolution, and Behavior*. New York, NY: Oxford University Press.

Silva, A. J., Kogan, J. H., Frankland, P. W., and Kida, S. (1998). CREB and memory. *Annu. Rev. Neurosci.* 21, 127–148. doi: 10.1146/annurev.neuro.21.1.127

Somel, M., Franz, H., Yan, Z., Lorenc, A., Guo, S., Giger, T., et al. (2009). Transcriptional neoteny in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5743–5748. doi: 10.1073/pnas.0900544106

Somel, M., Liu, X., Tang, L., Yan, Z., Hu, H., Guo, S., et al. (2011). MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol.* 9:e1001214. doi: 10.1371/journal.pbio.1001214

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi: 10.1093/nar/gkj109

Takahashi, J. S. (2015). Molecular components of the circadian clock in mammals. *Diabetes Obes. Metab.* 17(Suppl. 1), 6–11.

Tarpey, P. S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., et al. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat. Genet.* 41, 535–543. doi: 10.1038/ng.367

Thomson, D. M., Herway, S. T., Fillmore, N., Kim, H., Brown, J. D., Barrow, J. R., et al. (2008). AMP-activated protein kinase phosphorylates transcription factors of the CREB family. *J. Appl. Physiol.* 104, 429–438. doi: 10.1152/japplphysiol.00900.2007

Tripathi, S., Christie, K. R., Balakrishnan, R., Huntley, R., Hill, D. P., Thommesen, L., et al. (2013). Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database* 2013:bat062. doi: 10.1093/database/bat062

Tsankova, N., Renthal, W., Kumar, A., and Nestler, E. J. (2007). Epigenetic regulation in psychiatric disorders. *Nat. Rev. Neurosci.* 8, 355–367. doi: 10.1038/nrn2132

Uwanogho, D., Rex, M., Cartwright, E. J., Pearl, G., Healy, C., Scotting, P. J., et al. (1995). Embryonic expression of the chicken Sox2, Sox3 and Sox11 genes suggests an interactive role in neuronal development. *Mech. Dev.* 49, 23–36. doi: 10.1016/0925-4773(94)00299-3

van Bokhoven, H. (2011). Genetic and epigenetic networks in intellectual disabilities. *Annu. Rev. Genet.* 45, 81–104. doi: 10.1146/annurev-genet-110410-132512

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263. doi: 10.1038/nrg2538

Vecsey, C. G., Hawk, J. D., Lattal, K. M., Stein, J. M., Fabian, S. A., Attner, M. A., et al. (2007). Histone deacetylase inhibitors enhance memory and synaptic plasticity via CREB: CBP-dependent transcriptional activation. *J. Neurosci.* 27, 6128–6140. doi: 10.1523/JNEUROSCI.0296-07.2007

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384. doi: 10.1038/nature10110

West, A. E., and Greenberg, M. E. (2011). Neuronal activity–regulated gene transcription in synapse development and cognitive function. *Cold Spring Harb. Perspect. Biol.* 3:a005744. doi: 10.1101/cshperspect.a005744

Weyn-Vanhentenryck, S. M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* 6, 1139–1152. doi: 10.1016/j.celrep.2014.02.005

Williamson, D. F., Parker, R. A., and Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. *Ann. Intern. Med.* 110, 916–921. doi: 10.7326/0003-4819-110-11-916

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155, 997–1007. doi: 10.1016/j.cell.2013.10.020

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

Wingender, E., Schoeps, T., Haubrock, M., and Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* 43, D97–D102. doi: 10.1093/nar/gku1064

Yang, Q., She, H., Gearing, M., Colla, E., Lee, M., Shacka, J. J., et al. (2009). Regulation of neuronal survival factor MEF2D by chaperone-mediated autophagy. *Science* 323, 124–127. doi: 10.1126/science.1166088

Yuan, Z., Gong, S., Luo, J., Zheng, Z., Song, B., Ma, S., et al. (2009). Opposing roles for ATF2 and c-Fos in c-Jun-mediated neuronal apoptosis. *Mol. Cell. Biol.* 29, 2431–2442. doi: 10.1128/MCB.01344-08

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., et al. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* 34, 11929–11947. doi: 10.1523/JNEUROSCI.1860-14.2014

Zhang, Y. E., Landback, P., Vibranovski, M. D., and Long, M. (2011). Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* 9:e1001179. doi: 10.1371/journal.pbio.1001179

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on