Gene Coregulation and Coexpression in the Aryl Hydrocarbon Receptor-mediated Transcriptional Regulatory Network in the Mouse Liver

Navya Josyula[1], Melvin E. Andersen[2], Norbert Kaminski[3,4], Edward Dere[5,§], Timothy R. Zacharewski[5,4] and Sudin Bhattacharya[6,3,7,8,4*]

[1]Biomedical and Translational Informatics Program, Geisinger Health System, Rockville, MD 20850, USA

[2]Scitovation LLC, Research Triangle Park, NC 27709, USA

[3]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, MI, 48824, USA

[4]Institute for Integrative Toxicology, Michigan State University, East Lansing, MI, 48824, USA

[5]Department of Biochemistry & Molecular Biology, Michigan State University, East Lansing, MI, 48824, USA

[6]Department of Biomedical Engineering, Michigan State University, East Lansing, MI 48824, USA

[7]Center for Research on Ingredient Safety, Michigan State University, East Lansing, MI 48824

[8]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

[§]Currently at Genentech, South San Francisco, CA 94080

[*]Corresponding author; email: sbhattac@msu.edu

Abstract

Tissue-specific network models of chemical-induced gene perturbation can improve our mechanistic understanding of the intracellular events leading to adverse health effects resulting from chemical exposure. The aryl hydrocarbon receptor (AHR) is a ligand-inducible transcription factor (TF) that activates a battery of genes and produces a variety of species-specific adverse effects in response to the potent and persistent environmental contaminant 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). Here we assemble a global map of the AHR gene regulatory network in TCDD-treated mouse liver from a combination of previously published gene expression and genome-wide TF binding data sets. Using Kohonen self-organizing maps and subspace clustering, we show that genes co-regulated by common upstream TFs in the AHR network exhibit a pattern of co-expression. Specifically, directly-bound, indirectly-bound and non-genomic AHR target genes exhibit distinct patterns of gene expression, with the directly bound targets generally associated with highest median expression. Further, among the directly bound AHR target genes, the expression level increases with the number of AHR binding sites in the proximal promoter regions. Finally, we show that co-regulated genes in the AHR network activate distinct groups of downstream biological processes, with AHR-bound target genes enriched for metabolic processes and

enrichment of immune responses among AHR-unbound target genes, likely reflecting infiltration of immune cells into the mouse liver upon TCDD treatment. This work describes an approach to the reconstruction and analysis of transcriptional regulatory cascades underlying cellular stress response using bioinformatic and statistical tools.

Introduction

"Toxicity pathways" are normal intracellular signaling pathways that when sufficiently perturbed by exogenous chemicals can lead to an adverse outcome at the cellular level, and potentially at the level of tissues and the whole organism (NRC 2007; Whelan and Andersen 2013). Signaling, transcriptional and post-transcriptional regulatory networks underlie toxicity pathways and their dynamic behavior in response to endogenous and exogenous perturbation. It is crucial to understand the organization, structure and dynamics of these networks through mapping and modeling approaches for a quantitative assessment of the risks of chemical exposure to biological systems. Tissue-specific network models of chemical-induced perturbation can improve our understanding of the intracellular events leading to adverse effects and eventual injury from chemical exposure.

The major cellular response pathways are governed both transcriptionally and post-translationally. A core set of master regulatory transcription factors (TFs) are central actors in most molecular pathways leading to altered expression of suites of genes in response to exposure to a variety of chemical compounds (Jennings et al. 2013). These TFs, including the nuclear receptors, p53, nuclear factor erythroid 2-related factor (Nrf2), nuclear factor-κB (NF-κB), the STAT (signal transducers and activators of transcription) family and the aryl hydrocarbon receptor (AHR), typically coordinate a broad range of physiological processes like metabolism, oxidative stress response, differentiation, tumor suppression, reproduction, development and homeostasis (Tyagi et al. 2011; Ma 2013; Evans and Mangelsdorf 2014; Audet-Walsh and Giguére 2015; Wright et al. 2017). They thus act as sentinels of normal biological activity, but their inappropriate activation or inhibition can lead to adverse outcomes at the cellular or tissue level (Andersen et al. 2013).

Here we describe a network model of the AHR pathway in the mouse liver, assembled from previously published genomic data sets and newly analyzed using various computational methods. The AHR is a ligand-activated TF that belongs to the basic helix-loop-helix (bHLH)–PER-ARNT-SIM (PAS) family of proteins, which serve as sensors of developmental and environmental signals (Gu et al. 2000). The prototypical AHR ligand is TCDD (Poland et al. 1976), a persistent environmental toxicant that produces a variety of adverse effects in laboratory animals, including immune suppression, reproductive and endocrine effects, neurochemical alterations, developmental toxicity, chloracne and tumor promotion (Birnbaum 1994; Pohjanvirta and Tuomisto 1994). These effects are mediated by the transcriptional activity of the AHR, as shown by their absence or amelioration in AHR-null mice and mice with low-affinity AHR alleles (Okey et al. 1989; Gonzalez and Fernandez-Salguero 1998; Peters et al. 1999), as well as in

mice with mutations in the DNA-binding domain or nuclear localization sequence of the AHR (Bunger et al. 2003; Bunger et al. 2008). Ligand binding causes the inactive AHR in the cytosol to undergo a conformational change resulting in dissociation from its chaperone protein complex and translocation to the nucleus, where it forms a heterodimer with the related nuclear protein aryl hydrocarbon nuclear translocator (ARNT) (Hoffman et al. 1991; Whitelaw et al. 1993). The AHR-ARNT complex then binds to specific DNA sequences on target genes called dioxin response elements (DRE) containing the core sequence 5'-GCGTG-3' (Denison et al. 1988), leading to the regulation of a diverse battery of genes (Poland and Knutson 1982; Hankinson 1995). While the 5'-GCGTG-3' nucleotide core is substitution-intolerant, the flanking 5' and 3' nucleotides adjacent to the core sequence also contribute to a functional AHR binding site (Denison et al. 1988; Shen and Whitlock Jr 1992; Lusska et al. 1993; Gillesby et al. 1997). DRE-independent mechanisms of AHR binding have also been reported (Dere et al. 2011b; Huang and Elferink 2012).

While the density of AHR-bound regions in the genome of hepatic tissue from TCDD-treated mice is greatest in proximal promoter regions close to the transcription start site (TSS) of annotated genes, AHR also binds to sites distal from a TSS, e.g. in intergenic regions and 3' UTRs (Dere et al. 2011b). Moreover, only a third of the differentially expressed genes identified by microarray analysis showed AHR binding at a DRE in their proximal promoter regions, suggesting additional mechanisms of gene regulation by AHR beyond the canonical model described above (Dere et al. 2011b). These mechanisms may include target gene regulation from distal AHR-bound regions through DNA looping, or indirect regulation by AHR through tethering with a secondary TF (Farnham 2009). Such an indirect mechanism has been demonstrated in the regulation of the rat CYP1A2 gene by AHR (Sogawa et al. 2004).

Here we have mapped the TCDD-induced AHR regulatory network from a combination of gene expression and ChIP-on-chip data from the mouse liver (Dere et al. 2011b), which provides us a system-wide view of AHR-mediated gene regulation under short-term TCDD exposure (168 hr). Specifically, statistical and visualization tools were used to establish a relationship between gene co-regulation by multiple TFs and gene co-expression, and link groups of co-regulated genes to distinct downstream functional outcomes. Our focus here is on the early stages of hepatic response to TCDD exposure – longer-term exposure may lead to a different suite of adaptive responses at the cellular and tissue level.

Methods

Microarray data: Our network analysis was based on results from a previous study of gene expression profiling using whole genome oligonucleotide arrays (Agilent Technologies, Santa Clara, CA) of hepatic tissues from female C57BL/6 mice orally gavaged with 30 µg/kg of TCDD (Boverhof et al. 2005; Dere et al. 2011b). The gene expression analysis was performed in hepatic tissue from mice exposed to TCDD for 2, 4, 8, 12, 18, 24, 72 and 168 hrs. Differentially responsive genes were identified using previously described cutoffs for fold change and statistical significance (|fold change| ≥ 1.5 and posterior probabilities P1 (t) ≥ 0.999) (Eckel et al. 2004; Dere et al. 2011b).

ChIP-on-chip data: Genome-wide AHR location data were taken from the previously described ChIP-on-chip experiments (Dere et al. 2011b), where ChIP assays were performed with hepatic tissue from female C57BL/6 mice exposed to TCDD for 2 and 24 hrs. Genes were associated with AHR-enriched regions if the position of maximum fold enrichment was within 10 kb upstream of a transcriptional start site (TSS) through to the end of the 3' UTR. For the present analysis, the ChIP data for 2 and 24 hrs. were combined to obtain a unique list of ChIP enriched regions associated with annotated genes (**Supplementary Methods**; **Supplementary Code 1**).

DRE analysis in ChIP enriched regions: The ChIP enriched regions for the differentially expressed (DE) genes were computationally searched for the presence of 5'-GCGTG-3' DRE core sequences to infer the nature of AHR binding to the target genes. The putative DRE search algorithm, written in *R* (R Core Team 2016) (**Supplementary Methods**; **Supplementary Code 2**), was based on a previously described approach (Sun et al. 2004). Briefly, the genomic sequences of the enriched regions were obtained from UCSC Genome Browser (http://genome.ucsc.edu) and scanned for exact matches to the DRE core sequences on both positive and negative strands. For each matched region, the 5-bp core sequence was extended 7 bp upstream and downstream of the core. The matrix similarity (MS) scores (Quandt et al. 1995) for the 19-bp DRE sequences were calculated and compared to an MS score threshold of 0.8473 based on the lowest MS score of 13 bona fide AHR-binding sequences (Dere et al. 2011a) (i.e. sites from the literature confirmed to bind AHR). The DRE sequences with high MS scores (MS score ≥ 0.8473) were defined as putative DREs capable of binding AHR. The DE genes that were AHR-enriched and had a putative DRE in the enriched region were described as "directly bound" by AHR, while AHR-enriched genes without a putative DRE were described as "indirectly bound". The remaining DE genes that were not AHR-enriched were regarded as "unbound" / "non-genomic" targets.

Construction and visualization of the AHR transcriptional regulatory network: The DE genes from the Agilent oligonucleotide array data were searched against online databases to obtain a list of TFs that regulate these genes. The ChIP-X Enrichment Analysis (ChEA2) database (Kou et al. 2013) was used to obtain the list of regulatory TFs. To obtain the mouse-liver specific list of transcription factors, the mouse-specific TFs from ChEA2 were screened for expression in the liver using the TRANSFAC® database (Matys et al. 2003). The ensemble of DE genes including the directly and indirectly AHR-bound genes, together with their inferred transcriptional regulators, form a comprehensive network for TF-gene interactions under AHR-mediated TCDD induction. The landscape of this regulatory network was rendered using the open source network visualization tool Cytoscape (Shannon et al. 2003). The gene expression values at each time point of TCDD exposure were superposed on this network to visualize the temporal changes associated with each gene. A |fold ratio| threshold of 1.20 was used to identify the key target genes that are themselves TFs regulating other genes in the dataset (a looser fold change threshold was used for TFs than other genes as TFs tend to be more tightly regulated). To generate and annotate the network in Cytoscape, three input files describing the network topology and gene expression values were used: an AHR-gene interaction file and a TF-gene interaction file ("network files"), and a gene expression file ("attributes file"). Log_2 scaling of the fold ratios was used for visualizing gene expression. The network files were merged together to form the complete layout.

Gene expression analysis based on transcriptional groupings: A binary TF-gene interaction matrix with 43 TFs in addition to AHR was created indicating which TFs interact with which target genes. If a gene is regulated by a particular TF, then the corresponding interaction is represented as '1'; otherwise it is represented as '0'. We used this TF-gene interaction matrix to classify target genes into co-regulated groups in a transcriptional cascade, in order to examine any possible relation between co-regulation and co-expression. To generate this grouping, AHR and other key TFs that were also target genes were considered in all possible combinations to identify the expression trends for target genes in each group. The total number of genes in each co-regulated group was counted by referring to the TF-gene interaction matrix, and all groups with at least 5 genes were considered for examination of the expression patterns. A graphical analysis was performed in *R* to identify the expression patterns of target genes for each combination of regulatory TFs (**Supplementary Methods**; **Supplementary Code 3**).

Kohonen self-organizing maps to visualize gene co-expression: To further examine the relationship between the transcriptional groups and target gene expression patterns, a self-organizing map (SOM) for the AHR network was generated using the Kohonen SOM package in *R* (Wehrens and Buydens 2007). The

same TF-gene interaction matrix described above was used as input for this analysis. The SOM algorithm follows a clustering technique to group the target genes according to their TF binding patterns. Target genes with similar TF binding patterns are grouped into the same cluster or adjacent clusters, referred to as 'units' (**Supplementary Methods**; **Supplementary Code 4**).

Subspace clustering: The ORCLUS subspace clustering algorithm (Aggarwal and Yu 2000) and corresponding *R* package (Szepannek 2013) were used to cluster the differentially expressed genes into 16 non-overlapping groups. The number of clusters k = 16 and the dimensionality of each cluster l = 4 were chosen so as to minimize the cluster sparsity coefficient (Aggarwal and Yu 2000) (**Supplementary Code 5**).

Functional categorization of genes in each cluster: Gene ontology (GO) functional analysis was performed for the DE genes present in each ORCLUS cluster. Enriched GO "process" categories were identified for genes in each cluster using the *GOrilla* tool (Eden et al. 2009) with a *p*-value threshold of 10^{-3} and the list of all DE genes as background. *REViGO* (Supek et al. 2011) was used to arrange the enriched processes into a "treemap", which was then rendered as an image using the downloadable R script generated by the program (**Supplementary Code 6, Supplementary Code 7**).

Results

Differential gene expression: The raw array dataset (Dere et al. 2011b) consisted of 41,267 records with annotated genes, fold ratio and significance (P1 (t) values) at 2, 4, 8, 12, 18, 24, 72 and 168 hrs. post TCDD exposure. For genes with multiple occurrences in the dataset, the fold ratios and P1(t) values were averaged, resulting in a total of 21,307 unique gene records. After applying the statistical cutoff values for fold change and P1(t) at each expression time point, the resulting number of unique differentially expressed (DE) genes was 1,407. All 1,407 DE genes were used to generate the AHR regulatory network map.

Analysis of AHR-enriched genomic regions associated with DE genes: The ChIP-on-chip datasets for 2 and 24 hrs. time points consisted of 14,446 and 974 AHR-enriched regions respectively, with associated genes (Dere et al. 2011b). The two datasets were combined to yield a unique list of genes associated with at least one enriched region. This list of enriched genes was compared against the list of 1,407 DE genes, yielding 632 genes associated with one or more AHR-enriched regions. The AHR-enriched regions around these 632 genes were searched for putative DREs, producing three kinds of regions depending on presence and location of DREs:

(a) Regions with one or more 5-bp DRE cores centrally located such that a 7-bp upstream and downstream extension was possible for MS score calculations.

(b) Regions with DRE cores present only at the edge of the region so that the 7-bp extension in both directions was not possible.

(c) Regions with no DRE core.

A total of 144 genes were associated with AHR-enriched regions where MS score calculations were possible, and that had putative DREs, i.e., 19-bp DRE sequences with an MS score ≥ 0.8473 (see Methods). These genes were considered to be “directly bound” by AHR. For the AHR-enriched regions with (i) non-putative DRE core (i.e. MS score < 0.8473), (ii) DRE core located at edges, or (iii) DRE core not present in the enriched region, the associated genes were considered to be “indirectly bound” by AHR. In total, among the 1,407 differentially expressed genes, 632 were bound by AHR with 144 genes directly bound, 488 indirectly bound, and the remaining 775 genes unbound by AHR.

Other transcriptional regulators of the DE genes: The ChEA2 database (Kou et al. 2013) provides a comprehensive record of transcription factor/target gene interactions from genome-wide ChIP studies for

both mouse and human. The list of 1,407 DE genes from our analysis was uploaded to the ChEA2 server, which identified 104 unique mouse-specific transcriptional regulators for these genes. These 104 TFs were searched against the TRANSFAC® database to filter for expression in liver tissue, which after accounting for discrepancies in naming between ChEA2 and TRANSFAC® resulted in a list of 43 unique mouse liver-expressed TFs. Out of our 1,407 DE genes, 1,198 had interactions with at least one of these 43 transcription factors. Among the 43 TFs, seven were themselves target genes of other identified TFs differentially expressed at |fold ratio| > 1.2 in the microarray dataset. These seven TFs were NRF2, FLI1, KLF4, SOX17, CCND1, PPARG and GATA1, and in addition to AHR, form the "hubs" of the inferred mouse liver AHR network.

The AHR regulatory network: All interactions of the DE genes with AHR and the other 43 identified TFs together form the mouse liver AHR regulatory network (**Figure 1**), which consists of 44 "source" nodes interacting with 1,241 "target" nodes.

AHR and the other seven hub TFs act as both source and target nodes (AHR regulates itself). Two of these hub TFs are regulated by AHR: Nrf2 is a direct target and Fli1 an indirect target (**Figure 1**). The expression levels for up- and down-regulated genes were superposed on this network layout for each of the eight time points in the gene array study (**Supplementary Figure 1a-h**), illustrating that the gene expression levels were not monotonic in time.

To examine the transcriptional regulatory hierarchy in the network, AHR and four of the seven hub TFs (Fli1, Nrf2, Klf4 and Sox17), which were all expressed at |fold ratio|> 1.5, were grouped in all possible combinations (**Supplementary Table 1**) to assess expression of their target genes. Genes differentially expressed at least at one time point at |fold ratio| > 1.5 were chosen for this analysis, yielding 1,191 target genes regulated individually or in combination by the above five TFs (**Supplementary Table 1**). We then examined the expression pattern of groups of co-regulated genes with a count of 5 or more (**Table 1**). The time courses of genes that were up-regulated at the 168 hrs. time point (**Figure 2**) suggest that genes with the same upstream regulators have similar expression patterns.

Co-regulation and co-expression in the AHR network: To take a closer look at whether co-regulation in the AHR network is associated with co-expression, we clustered the 1191 target genes into self-organizing maps (SOMs) based on the factors that regulate them (**Figure 3A**). Each circular unit in the SOM represents a grouping of genes (individual dots within a unit). The SOM algorithm groups genes into units such that genes in a single unit or adjacent units have a similar combination of TFs regulating them (no gene

expression values were used for clustering), while genes in distant units have more dissimilar regulators. The median expression level (log_2 fold change) of the genes in each unit was then superposed on the SOM as a continuous color scale with blue indicating suppression and red activation (panels in **Figure 3B**). A distinct pattern emerges over the time course, with the units with high median expression at 168 hrs. localized at the lower right corner of the SOM (**Figure 3B**). The median time courses of the genes in adjacent units are also quite similar (**Supplementary Figure 2**). This analysis shows a strong association between gene co-regulation and co-expression in the AHR network.

Localized clustering of co-regulated genes: We further attempted to cluster the 1,191 target genes considered in the SOM analysis above based on regulation by the 44 TFs. Fundamentally, the clustering problem may be stated as: "Given a set of data points, partition them into a set of groups which are as similar as possible" (Aggarwal 2014). If we consider the binary TF-gene connectivity matrix, with genes in rows (observations), TFs in columns (features) and each matrix element equaling 1 or 0 depending on whether a TF binds a gene, we have a high-dimensional clustering problem with feature localization, i.e. different groups of genes are regulated by different subsets of TFs. Global clustering methods like k-means, or dimensionality reduction approaches like principal components analysis do not perform well in this situation, which motivated the development of high-dimensional subspace clustering methods (Aggarwal 2014). These methods include "projected clustering" or "subspace clustering" approaches like PROCLUS (Aggarwal et al. 1999), CLIQUE (Agrawal et al. 2005) and ORCLUS (Aggarwal and Yu 2000), where feature selection or transformation is performed specific to different localities of the data (Aggarwal 2014). ORCLUS in particular is suited for data sets like ours where relevant subspaces may be arbitrarily oriented due to inter-feature correlations (Aggarwal and Yu 2000), i.e. many TFs are correlated in term of which genes they regulate.

We used the ORCLUS algorithm to group the DE genes by TF connectivity into 16 clusters (**Figure 4A**), illustrating both the sparsity of the TF-gene connectivity matrix and the fact that different clusters of genes are regulated by different subsets of TFs. In particular, there is a marked contrast between Cluster 2, where none of 157 genes is bound by AHR, and Cluster 6, where all 123 genes are. Cluster 2 genes can thus be said to comprise a "non-genomic pathway" and Cluster 6 genes a "genomic pathway" with respect to regulation by AHR. The most frequent regulators in Cluster 2 are PPARG, regulating 46 genes, STAT3 (33 genes), CEBPB, NANOG (30 genes each), CREB1, GATA2 and SUZ12 (27 genes each). In contrast, the most frequent regulators in Cluster 6 are AHR (all 123 genes), followed by SUZ12 (55 genes), PPARG (42 genes), CREB1 (29 genes), MYC (25 genes), NANOG (24 genes), E2F1 and TAL1 (22 genes each).

AHR binding and gene expression: We plotted the time courses of $\log_2$ fold change values for all genes in Clusters 2 and 6 (**Figures 5A** and **5B**). Genes in the "non-genomic" Cluster 2 are downregulated or moderately upregulated at earlier times (**Figure 5A**), with about two-thirds of the genes showing upregulation at the later time points. On the contrary, a majority of genes in the "genomic" Cluster 6 (**Figure 5B**) are moderately to strongly upregulated at all time points, with a smaller subset showing consistent downregulation. This led us to suspect that there may be a link between binding of a gene by AHR and its expression level, and we separately plotted the time courses of genes that are (a) directly bound, (b) indirectly bound, and (c) unbound by AHR (**Figures 6A-C**). Nearly all genes in the directly bound group (**Figure 6A**) are upregulated moderately or strongly at multiple time points, whereas in the indirectly bound group (**Figure 6B**), about half of the genes are consistently downregulated. Finally, the unbound group shows an unusual pattern of gene expression (**Figure 6C**), where most genes are downregulated at the first two time points, but then about two-thirds of the genes show progressive upregulation up to the 168 h time point. This observation suggests a cascade structure in the AHR network, where genes not proximally bound by AHR are bound by other TFs activated at intermediate to later points in the time course, or are targets of long range interaction with distally-bound AHR, leading to their upregulation at later times.

These differences between direct, indirect and unbound AHR target genes are also highlighted in overlaid box and violin plots (**Figures 7A-D**), showing the respective distributions of expression level of the three groups of genes at multiple time points. At each time point shown, the middle 50% (first to third quartile) of the directly bound genes are all upregulated, while the indirectly bound group is symmetrically distributed with about half of the genes upregulated. In the unbound group, most genes are downregulated at earlier time points, but at 168 hrs., the distribution is considerably right-skewed with many genes upregulated. Overall, the directly regulated group has the highest median expression (except at 168 hrs.), and also has the most outliers on the high expression end, the furthest outlier being the Cyp1a1 gene. Given that the Cyp1a1 gene has a high number of DREs in its proximal promoter region (Li et al. 2014), we examined the relationship among expression level and number of proximal promoter DREs for the direct target genes at various time points (**Figures 8A-D**). There is an increase in mean expression level with increasing number of DREs, a trend that gets stronger at later time points. This is likely due to a larger number of AHR molecules binding to DREs in the promoter regions, leading to a higher degree of activation of proximal genes. This hypothesis is supported by previous findings that in the human liver, genes with more transcriptional regulators bound in their promoter regions were more highly expressed (Odom et al. 2006).

Distinct gene clusters activate distinct biological processes: We carried out gene ontology (GO) analysis on the six major clusters of genes labeled in **Figure 4A**. The genes in the six clusters enrich for different groups of biological processes (**Figure 4B**). In particular, the genomic cluster (Cluster 6) is enriched for genes associated with metabolic processes and ribosome biogenesis, whereas the major GO categories associated with the non-genomic cluster (Cluster 2) are immune regulatory processes. Interestingly, Cluster 5 is enriched for cell migration and activation of cellular defense mechanisms. Presumably this reflects immune cell infiltration into the mouse liver under exposure to TCDD (Fader et al. 2015). Cluster 16 is also enriched for immune system response. Thus, co-regulated genes in the AHR network in the mouse liver show patterns of co-expression, and lead to differential downstream activation of biological processes.

Discussion

Ligand-activated transcription factors underlie most major cellular response pathways. These TF-governed molecular pathways tend to have a similar organizational structure with key functional components that act as signal sensors (co-binding proteins) and transducers (protein kinases) to complement the central role of the TF (Simmons et al. 2009). The inactivated TF is typically sequestered in the cytoplasm or nucleus. Upon activation by its ligand (endogenous or exogenous molecule), the TF is able to bind specific response elements in the promoter regions of target genes and activate or inhibit expression of suites of genes in a coordinated manner. Beyond these "direct target" genes, there are additional genes that bind the master regulatory TF indirectly through tethering interactions with secondary TFs (George et al. 2011; Shen et al. 2011; McMullen et al. 2014). In fact, combinatorial control of gene expression by TFs is a common feature of cellular pathways, since binding sites are often clustered in the genome, allowing multiple TFs to act in a coordinated fashion to induce or suppress groups of genes in specific cell types under particular conditions (George et al. 2011). In addition, a surprisingly large number of genes are activated or inhibited in a "non-genomic" manner, showing no evidence of binding by the master regulatory TF of the stimulated pathway in their promoter regions (van der Meer et al. 2010; Dere et al. 2011b; Shen et al. 2011; McMullen et al. 2014). These observations collectively suggest that combining gene expression data from transcriptome profiling with high-throughput genome-wide analysis of TF binding can provide an integrated, systems-level view of the structure and function of transcription factor-governed molecular pathways (Blais and Dynlacht 2005; Walhout 2006; Dere et al. 2011b; Limonciel et al. 2015).

Accordingly, we have integrated TCDD-induced gene expression and multiple genome-wide TF binding data sets for a global view of the AHR regulatory pathway in the mouse liver. Using a combination of self-organizing maps and subspace clustering, we show that there is a pattern of co-regulated genes in the AHR pathway being co-expressed, as previously observed in *Saccharomyces cerevisiae* (Yu et al. 2003; Allocco et al. 2004). In particular, directly-bound, indirectly-bound and unbound AHR target genes have distinct patterns of gene expression, with the directly-bound group showing higher median expression. Further, among the direct AHR target genes, the expression level increases with the number of AHR-binding DRE sites in the proximal promoter regions. Finally, we found that co-regulated gene clusters activated distinct groups of downstream biological processes, with the AHR-bound genomic cluster enriched for metabolic processes and the AHR-unbound non-genomic cluster primarily activating immune processes. This work, together with other recent studies of the PPARα and estrogen receptor pathways

(McMullen et al. 2014; Pendse et al. 2016), illustrates the application of bioinformatic and statistical tools for reconstruction and analysis of the transcriptional regulatory cascades underlying cellular stress response.

Acknowledgments

The authors would like to thank Agnes (Forgacs) Karmaus, Arindam Banerjee, Rory Conolly and Qiang Zhang for helpful discussions. This work was supported by the US EPA STAR Program (EPA Grant Number: R835000) and the Superfund Research Program of the National Institute of Environmental Health Sciences (P42ES04911).

References

Aggarwal C, Yu P. 2000. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 70-81.

Aggarwal CC. 2014. An Introduction to Cluster Analysis. In *Data Clustering: Algorithms and Applications*, (ed. CC Aggarwal, CK Reddy). Chapman and Hall/CRC Boca Raton, FL.

Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS. 1999. Fast algorithms for projected clustering. *ACM SIGMOD Record* **28**: 61.

Agrawal R, Gehrke J, Gunopulos D, Raghavan P. 2005. Automatic Subspace Clustering of High Dimensional Data. *Data Mining and Knowledge Discovery* **11**: 5.

Allocco DJ, Kohane IS, Butte AJ. 2004. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**.

Andersen ME, McMullen PD, Bhattacharya S. 2013. Toxicogenomics for transcription factor-governed molecular pathways: Moving on to roles beyond classification and prediction. *Archives of Toxicology* **87**: 7-11.

Audet-Walsh É, Giguére V. 2015. The multiple universes of estrogen-related receptor α and γ in metabolic control and related diseases. *Acta Pharmacologica Sinica* **36**: 51-61.

Birnbaum LS. 1994. The mechanism of dioxin toxicity: Relationship to risk assessment. *Environmental Health Perspectives* **102**: 157-167.

Blais A, Dynlacht BD. 2005. Constructing transcriptional regulatory networks. *Genes and Development* **19**: 1499-1511.

Boverhof DR, Burgoon LD, Tashiro C, Chittim B, Harkema JR, Jump DB, Zacharewski TR. 2005. Temporal and dose-dependent hepatic gene expression patterns in mice provide new insights into TCDD-mediated hepatotoxicity. *Toxicological Sciences* **85**: 1048-1063.

Bunger MK, Glover E, Moran SM, Walisser JA, Lahvis GP, Hsu EI, Bradfield CA. 2008. Abnormal liver development and resistance to 2,3,7,8-tetrachlorodibenzo-p-dioxin toxicity in mice carrying a mutation in the DNA-Binding domain of the aryl hydrocarbon receptor. *Toxicological Sciences* **106**: 83-92.

Bunger MK, Moran SM, Glover E, Thomae TL, Lahvis GP, Lin BC, Bradfield CA. 2003. Resistance to 2,3,7,8-tetrachlorodibenzo-p-dioxin toxicity and abnormal liver development in mice carrying a mutation in the nuclear localization sequence of the aryl hydrocarbon receptor. *Journal of Biological Chemistry* **278**: 17767-17774.

Denison MS, Fisher JM, Whitlock Jr JP. 1988. The DNA recognition site for the dioxin-Ah receptor complex. Nucleotide sequence and functional analysis. *Journal of Biological Chemistry* **263**: 17221-17224.

Dere E, Forgacs AL, Zacharewski TR, Burgoon LD. 2011a. Genome-wide computational analysis of dioxin response element location and distribution in the human, mouse, and rat genomes. *Chemical Research in Toxicology* **24**: 494-504.

Dere E, Lo R, Celius T, Matthews J, Zacharewski TR. 2011b. Integration of Genome-Wide Computation DRE Search, AhR ChIP-chip and Gene Expression Analyses of TCDD-Elicited Responses in the Mouse Liver. *BMC Genomics* **12**.

Eckel JE, Gennings C, Chinchilli VM, Burgoon LD, Zacharewski TR. 2004. Empirical bayes gene screening tool for time-course or dose-response microarray data. *Journal of Biopharmaceutical Statistics* **14**: 647-670.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**.

Evans RM, Mangelsdorf DJ. 2014. Nuclear receptors, RXR, and the big bang. *Cell* **157**: 255-266.

Fader KA, Nault R, Ammendolia DA, Harkema JR, Williams KJ, Crawford RB, Kaminski NE, Potter D, Sharratt B, Zacharewski TR. 2015. 2,3,7,8-tetrachlorodibenzo-p-dioxin alters lipid metabolism and depletes immune cell populations in the Jejunum of C57BL/6 mice. *Toxicological Sciences* **148**: 567-580.

Farnham PJ. 2009. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics* **10**: 605-616.

George CL, Lightman SL, Biddie SC. 2011. Transcription factor interactions in genomic nuclear receptor function. *Epigenomics* **3**: 471-485.

Gillesby BE, Stanostefano M, Porter W, Safe S, Wu ZF, Zacharewski TR. 1997. Identification of a motif within the 5' regulatory region of pS2 which is responsible for AP-1 binding and TCDD-mediated suppression. *Biochemistry* **36**: 6080-6089.

Gonzalez FJ, Fernandez-Salguero P. 1998. The aryl hydrocarbon receptor. Studies using the AHR-null mice. *Drug Metabolism and Disposition* **26**: 1194-l1198.

Gu YZ, Hogenesch JB, Bradfield CA. 2000. The PAS superfamily: Sensors of environmental and developmental signals. In *Annual Review of Pharmacology and Toxicology*, Vol 40, pp. 519-561.

Hankinson O. 1995. The aryl hydrocarbon receptor complex. *Annual Review of Pharmacology and Toxicology* **35**: 307-340.

Hoffman EC, Reyes H, Chu FF, Sander F, Conley LH, Brooks BA, Hankinson O. 1991. Cloning of a factor required for activity of the Ah (dioxin) receptor. *Science* **252**: 954-958.

Huang G, Elferink CJ. 2012. A novel nonconsensus xenobiotic response element capable of mediating aryl hydrocarbon receptor-dependent gene expression. *Molecular Pharmacology* **81**: 338-347.

Jennings P, Limonciel A, Felice L, Leonard MO. 2013. An overview of transcriptional regulation in response to toxicological insult. *Archives of Toxicology* **87**: 49-72.

Kou Y, Chen EY, Clark NR, Duan Q, Tan CM, Ma'ayan A. 2013. ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome. In *Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 84, 89, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013 Proceedings*, doi:10.1007/978-3-642-40511-2_30 (ed. A Cuzzocrea, et al.), pp. 416-430. Springer Berlin Heidelberg, Berlin, Heidelberg.

Li S, Pei X, Zhang W, Xie H, Zhao B. 2014. Functional Analysis of the Dioxin Response Elements (DREs) of the Murine CYP1A1 Gene Promoter: Beyond the Core DRE Sequence. *International Journal of Molecular Sciences* **15**: 6475.

Limonciel A, Moenks K, Stanzel S, Truisi GL, Parmentier C, Aschauer L, Wilmes A, Richert L, Hewitt P, Mueller SO et al. 2015. Transcriptomics hit the target: Monitoring of ligand-activated and stress response pathways for chemical testing. *Toxicology in Vitro* doi:10.1016/j.tiv.2014.12.011.

Lusska A, Shen E, Whitlock Jr JP. 1993. Protein-DNA interactions at a dioxin-responsive enhancer: Analysis of six bona fide DNA-binding sites for the liganded Ah receptor. *Journal of Biological Chemistry* **268**: 6575-6580.

Ma Q. 2013. Role of Nrf2 in oxidative stress and toxicity. In *Annual Review of Pharmacology and Toxicology*, Vol 53, pp. 401-426.

Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV et al. 2003. TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**: 374-378.

McMullen PD, Bhattacharya S, Woods CG, Sun B, Yarborough K, Ross SM, Miller ME, McBride MT, Lecluyse EL, Clewell RA et al. 2014. A map of the PPARα transcription regulatory network for primary human hepatocytes. *Chemico-Biological Interactions* **209**: 14-24.

NRC. 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. The National Academies Press, Washington, DC.

Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, Gifford DK, Fraenkel E, Bell GI, Young RA. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Molecular systems biology* **2**: 2006 0017.

Okey AB, Vella LM, Harper PA. 1989. Detection and characterization of a low affinity form of cytosolic Ah receptor in livers of mice nonresponsive to induction of cytochrome P1-450 by 3-methylcholanthrene. *Molecular Pharmacology* **35**: 823-830.

Pendse SN, Maertens A, Rosenberg M, Roy D, Fasani RA, Vantangoli MM, Madnick SJ, Boekelheide K, Fornace AJ, Odwin SA et al. 2016. Information-dependent enrichment analysis reveals time-dependent transcriptional regulation of the estrogen pathway of toxicity. *Archives of Toxicology* doi:10.1007/s00204-016-1824-6: 1-14.

Peters JM, Narotsky MG, Elizondo G, Fernandez-Salguero PM, Gonzalez FJ, Abbott BD. 1999. Amelioration of TCDD-induced teratogenesis in aryl hydrocarbon receptor (AhR)-null mice. *Toxicological Sciences* **47**: 86-92.

Pohjanvirta R, Tuomisto J. 1994. Short-term toxicity of 2,3,7,8-tetrachlorodibenzo-p-dioxin in laboratory animals: Effects, mechanisms, and animal models. *Pharmacological Reviews* **46**: 483-549.

Poland A, Glover E, Kende AS. 1976. Stereospecific, high affinity binding of 2,3,7,8 tetrachlorodibenzo p dioxin by hepatic cytosol. Evidence that the binding species is receptor for induction of aryl hydrocarbon hydroxylase. *Journal of Biological Chemistry* **251**: 4936-4946.

Poland A, Knutson JC. 1982. 2,3,7,8-tetrachlorodibenzo-p-dioxin and related halogenated aromatic hydrocarbons: examination of the mechanism of toxicity. *Annual Review of Pharmacology and Toxicology* **22**: 517-554.

Quandt K, Frech K, Karas H, Wingender E, Werner T. 1995. MatInd and matinspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* **23**: 4878-4884.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research* **13**: 2498-2504.

Shen C, Huang Y, Liu Y, Wang G, Zhao Y, Wang Z, Teng M, Wang Y, Flockhart DA, Skaar TC et al. 2011. A modulated empirical Bayes model for identifying topological and temporal estrogen receptor alpha regulatory networks in breast cancer. *BMC Syst Biol* **5**: 67.

Shen ES, Whitlock Jr JP. 1992. Protein-DNA interactions at a dioxin-responsive enhancer: Mutational analysis of the DNA-binding site for the liganded Ah receptor. *Journal of Biological Chemistry* **267**: 6815-6819.

Simmons SO, Fan CY, Ramabhadran R. 2009. Cellular stress response pathway system as a sentinel ensemble in toxicological screening. *Toxicological Sciences* **111**: 202-225.

Sogawa K, Numayama-Tsuruta K, Takahashi T, Matsushita N, Miura C, Nikawa JI, Gotoh O, Kikuchi Y, Fujii-Kuriyama Y. 2004. A novel induction mechanism of the rat CYP1A2 gene mediated by Ah receptor-Arnt heterodimer. *Biochemical and Biophysical Research Communications* **318**: 746-755.

Sun YV, Boverhof DR, Burgoon LD, Fielden MR, Zacharewski TR. 2004. Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. *Nucleic Acids Research* **32**: 4512-4523.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**.

Szepannek G. 2013. orclus: ORCLUS subspace clustering. R package version 0.2-5.

Tyagi S, Gupta P, Saini AS, Kaushal C, Sharma S. 2011. The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. *J Adv Pharm Technol Res* **2**: 236-240.

van der Meer DLM, Degenhardt T, Väisänen S, de Groot PJ, Heinäniemi M, de Vries SC, Müller M, Carlberg C, Kersten S. 2010. Profiling of promoter occupancy by PPARα in human hepatoma cells via ChIP-chip analysis. *Nucleic Acids Research* **38**: 2839-2850.

Walhout AJM. 2006. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Research* **16**: 1445-1454.

Wehrens R, Buydens LMC. 2007. Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software* **21**: 1-19.

Whelan M, Andersen ME. 2013. Toxicity Pathways – from concepts to application in chemical safety assessment. JRC, Luxembourg: Publications Office of the European Union.

Whitelaw M, Pongratz I, Wilhelmsson A, Gustafsson JÅ, Poellinger L. 1993. Ligand-dependent recruitment of the arnt coregulator determines DNA recognition by the dioxin receptor. *Molecular and Cellular Biology* **13**: 2504-2514.

Wright EJ, Pereira De Castro K, Joshi AD, Elferink CJ. 2017. Canonical and non-canonical aryl hydrocarbon receptor signaling pathways. *Current Opinion in Toxicology* **2**: 87-92.

Yu H, Luscombe NM, Qian J, Gerstein M. 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in Genetics* **19**: 422-427.

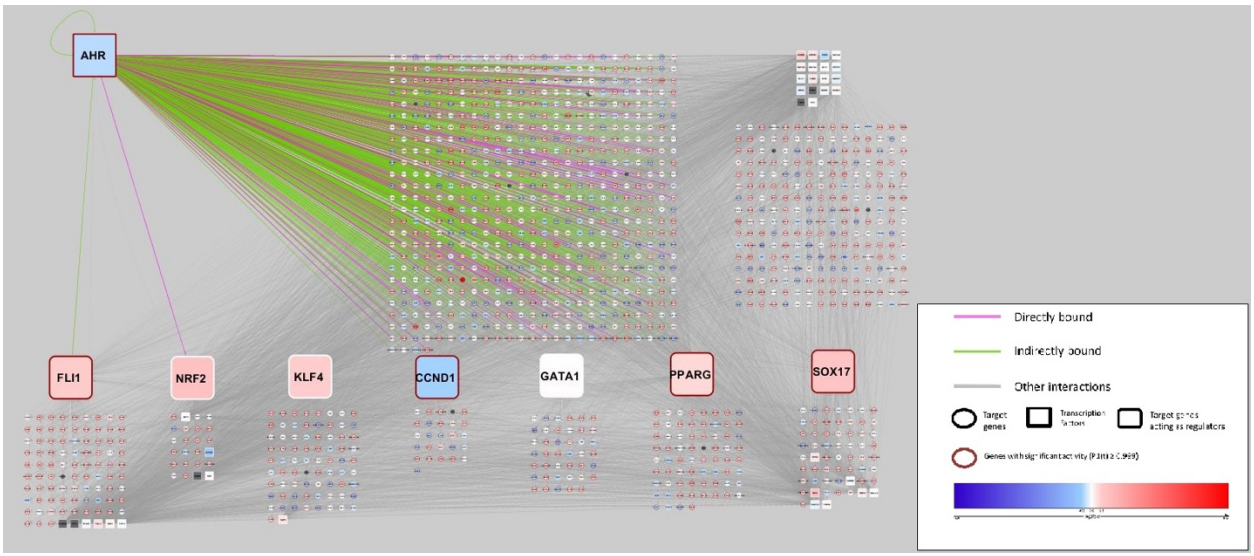


Figure 1. A comprehensive AHR transcriptional regulatory landscape in the mouse liver, viewed at time t = 168h and TCDD dose = 30 µg/kg. The AHR network in the mouse liver reveals a hierarchical structure. Transcription factor (TF) nodes are shown as rectangles, target genes as circles. Edges represent TF-target gene binding. AHR is at the top left of the figure, with AHR interactions shown with pink and green edges indicating direct and indirect binding respectively. The seven key non-AHR TFs that are also target genes of other TFs are shown in the bottom half of the figure along with their target genes. Other TFs in the network and their target genes are shown in the top right of the figure. Each node is colored according to their $\log_2$ fold change ratio at 168hrs (nodes in grey were not differentially expressed at 168hrs). Non-AHR TF-gene interactions are shown as grey edges.

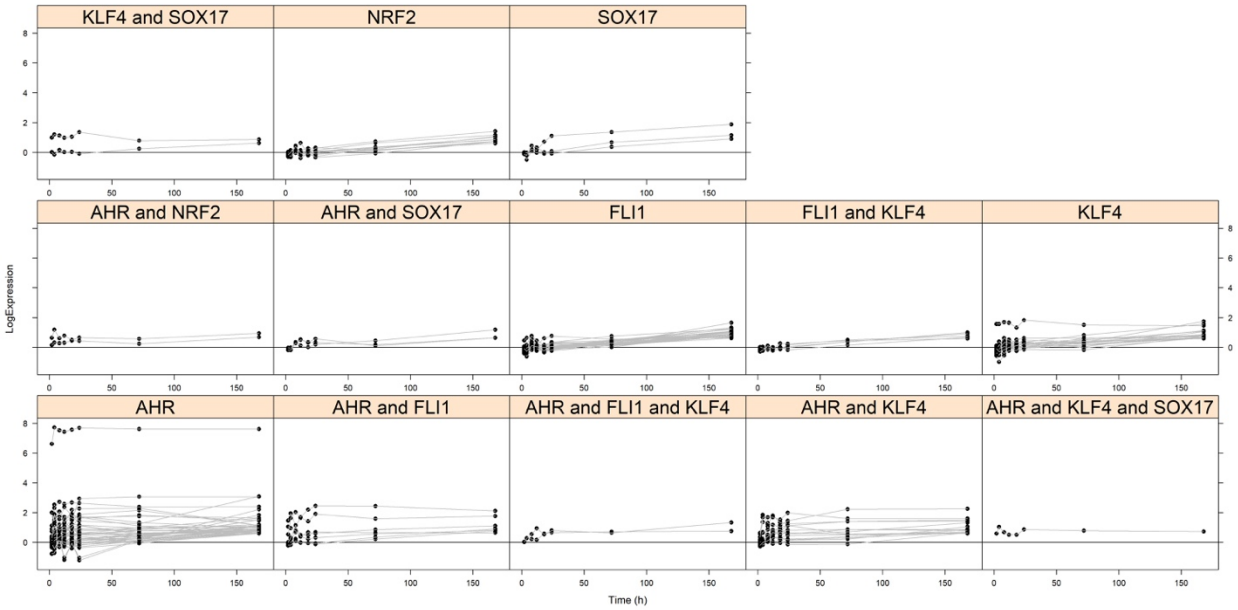
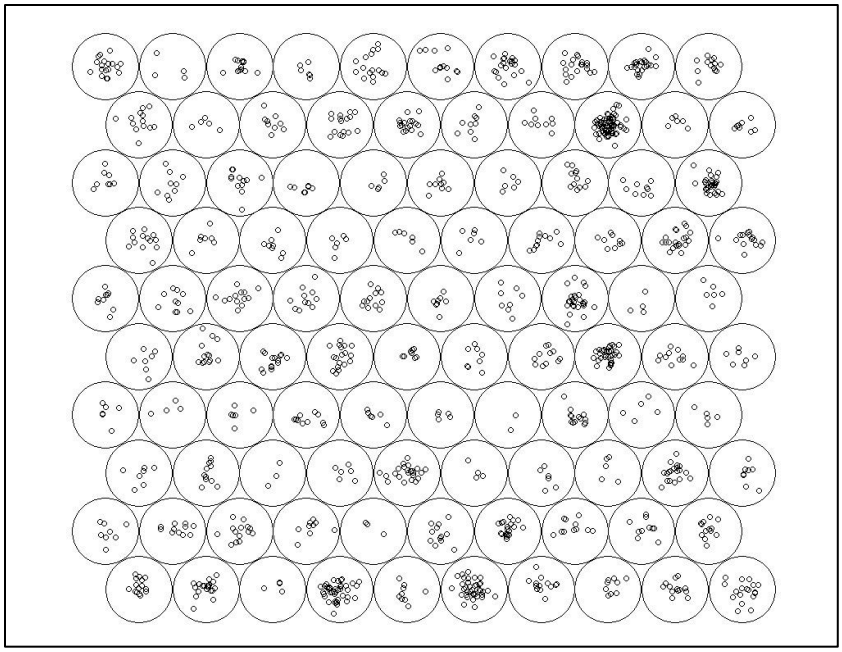
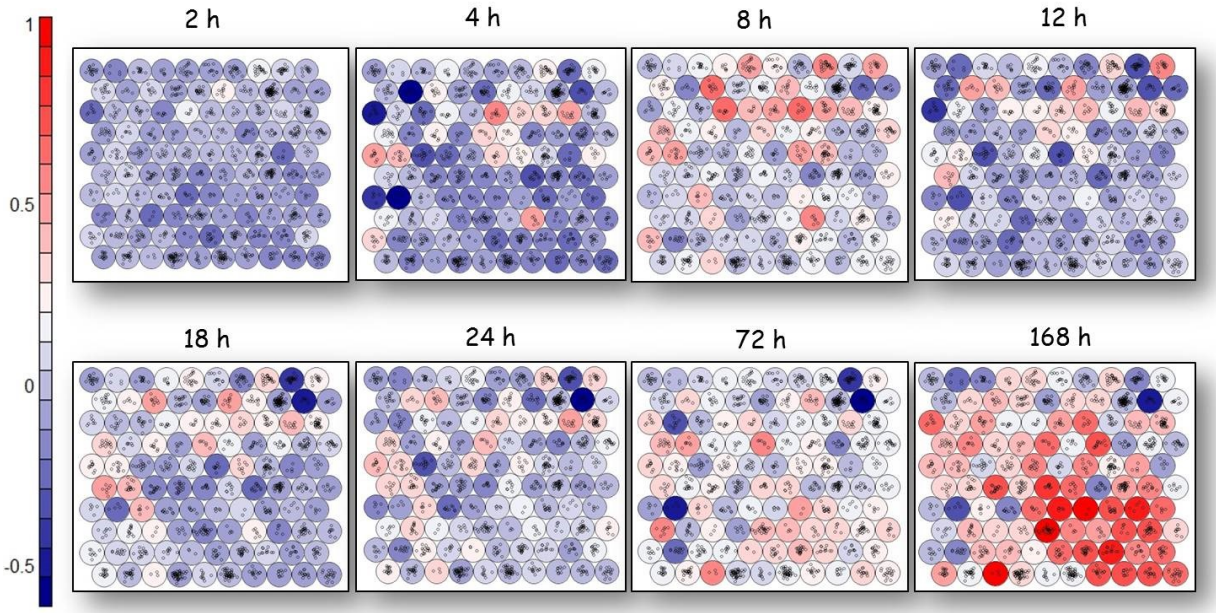


Figure 2: Time courses of genes grouped by transcriptional regulators (only genes up-regulated at 168hrs. shown). Genes grouped by transcriptional regulators show similar expression patterns. The vertical axis denotes $\log_2$ fold change.



(A)



(B)

Figure 3: Median expression of genes organized into Kohonen Self-Organizing Map (SOM). (A) The mapping of genes clustered in each unit according to TF binding patterns. (B) The temporal gene expression patterns of the SOM units, confirming the coregulation and coexpression patterns of the genes. The continuous color scale shows the median $\log_2$ fold change expression values for the genes in each unit, with blue indicating suppression and red activation.

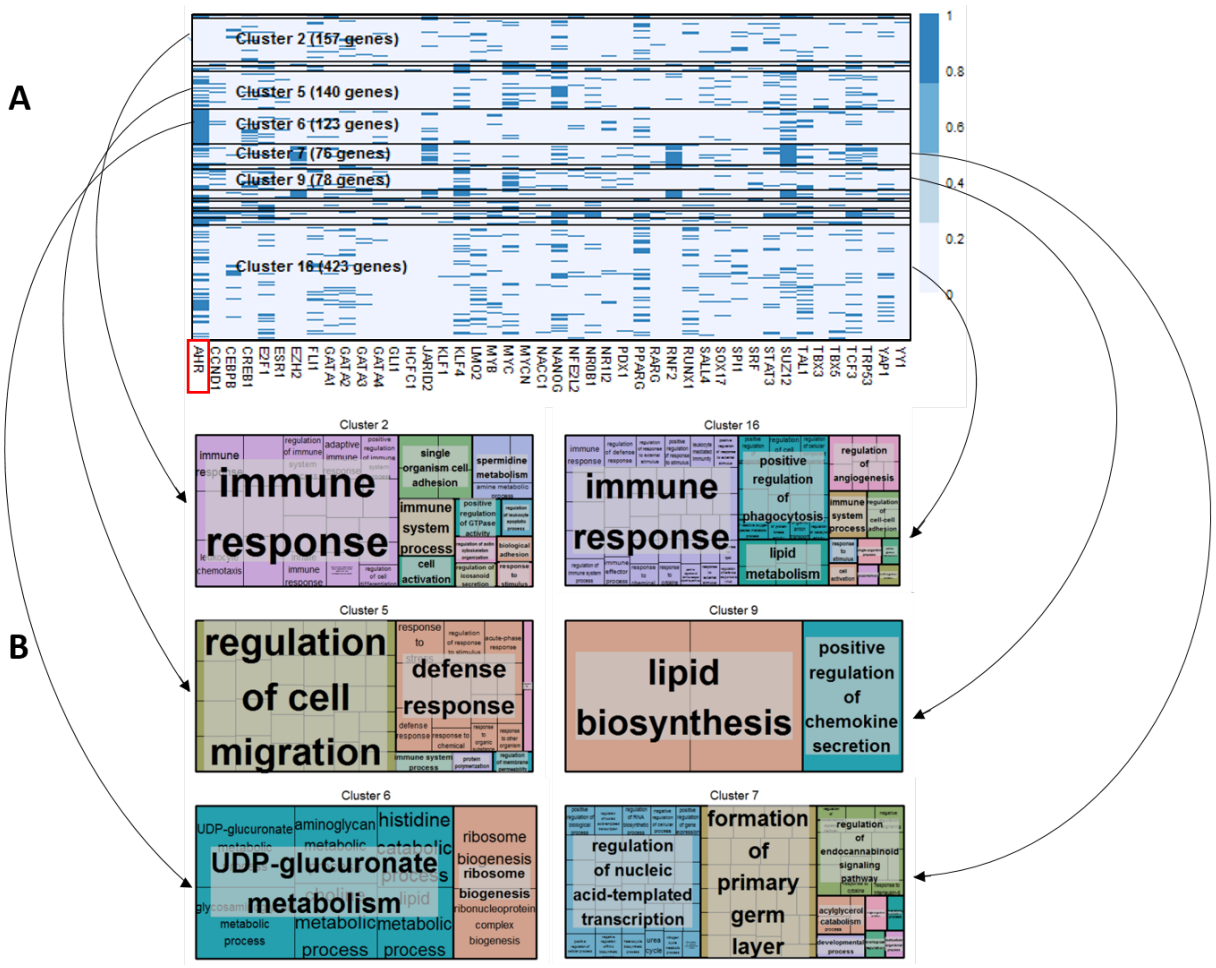
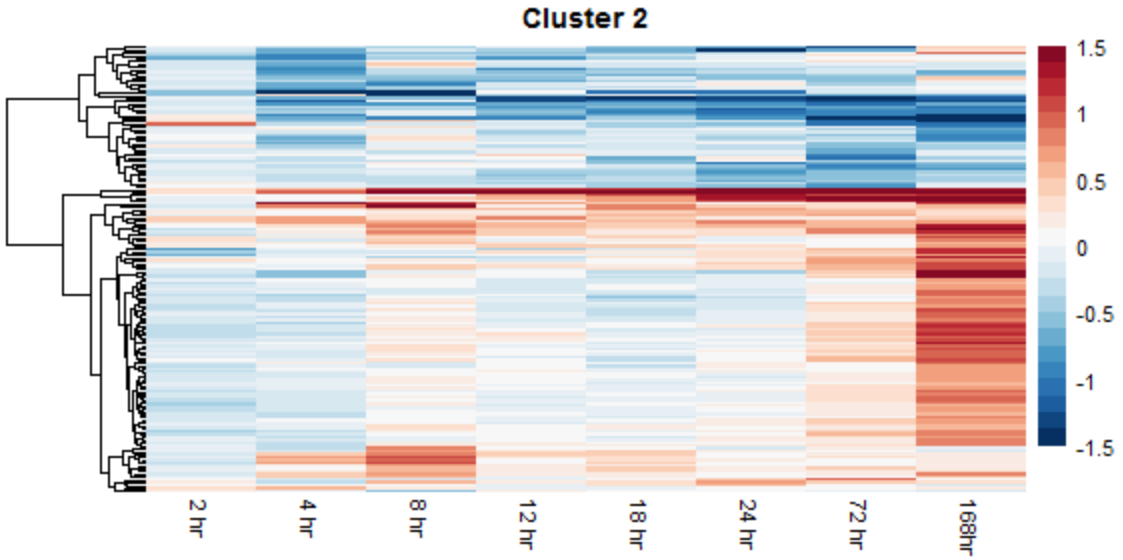
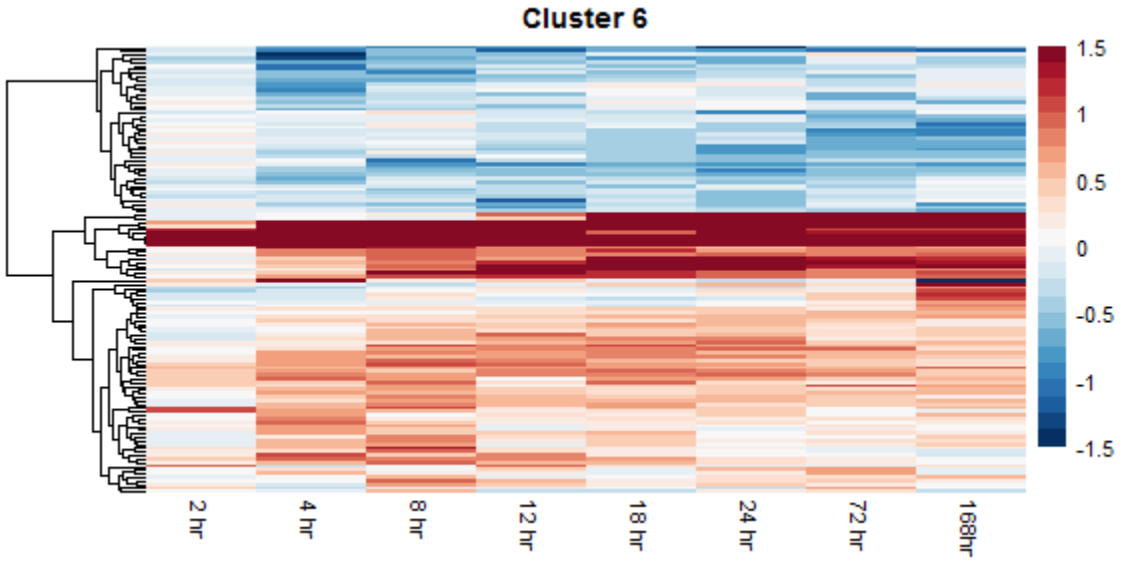


Figure 4: Distinct co-regulated clusters of genes in the AHR network activate distinct downstream pathways. (A) Clustered TF-gene connectivity matrix with 1191 genes in rows and 44 TFs in columns. Elements of the matrix have the value 1 (dark blue) if the corresponding TF and gene are bound; and 0 (pale blue) otherwise. Clusters of more than 50 genes are labeled. (B) Gene Ontology (GO) process categories associated with clusters of more than 50 genes. Sizes of individual boxes representing GO processes are proportional to negative $\log_{10}$ p-values of enrichment.



(A)



(B)

Figure 5: Heatmaps showing time courses of log_2 fold change for all 157 genes in Cluster 2 (A) and all 123 genes in Cluster 6 (B). For visualization of the heatmap, log_2 fold change values > 1.5 were set to 1.5 and values < -1.5 to -1.5. Blue indicates downregulation and red upregulation.

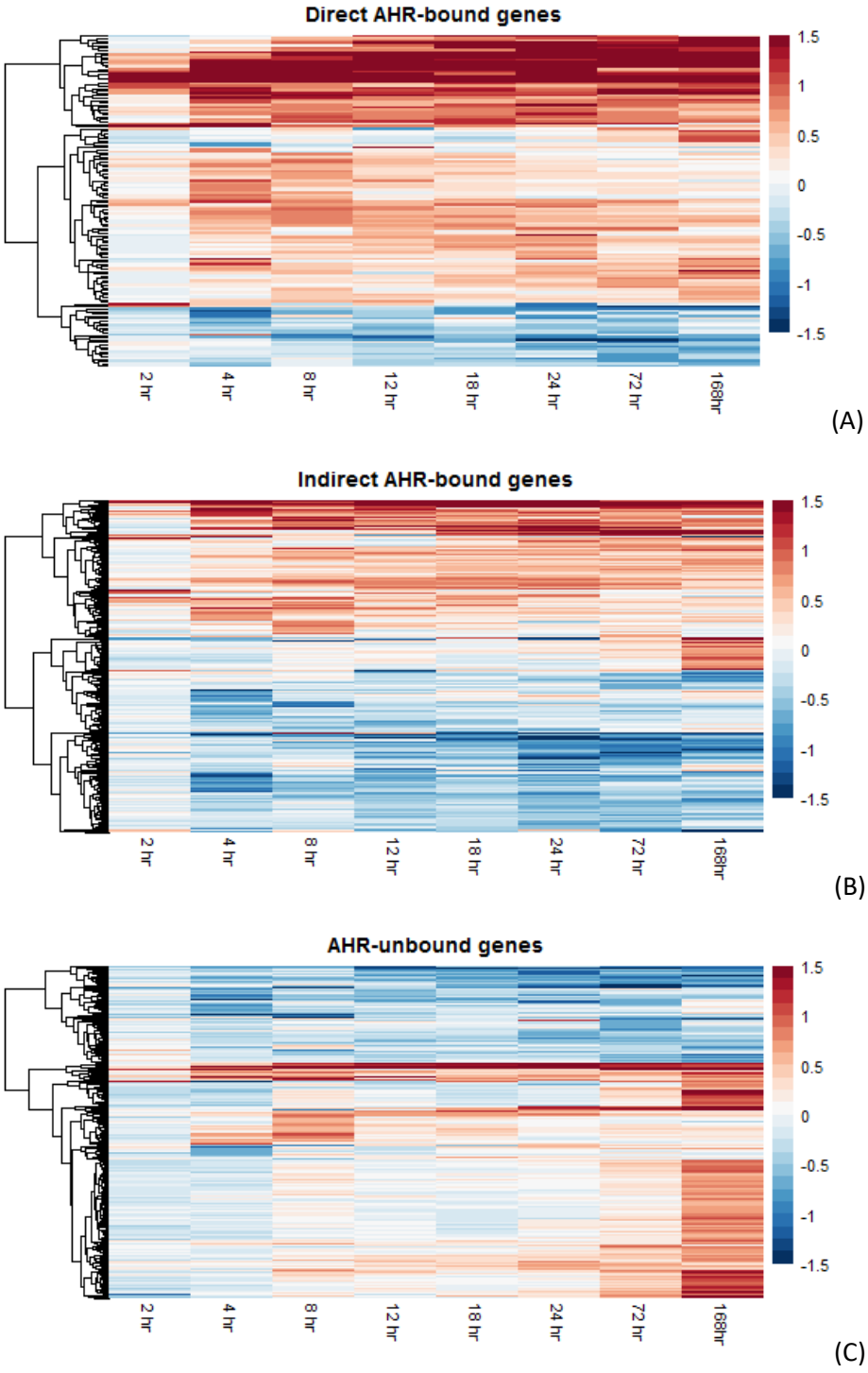


Figure 6: Heatmaps showing time courses of log_2 fold change for 140 genes directly bound (A), 477 genes indirectly bound (B), and 574 genes unbound (C) by AHR. For visualization of the heatmap, log_2 fold change values > 1.5 were set to 1.5 and values < -1.5 to -1.5. There are proportionately more upregulated genes in (A) compared to (B), and in (B) compared to (C). Blue indicates downregulation and red upregulation.

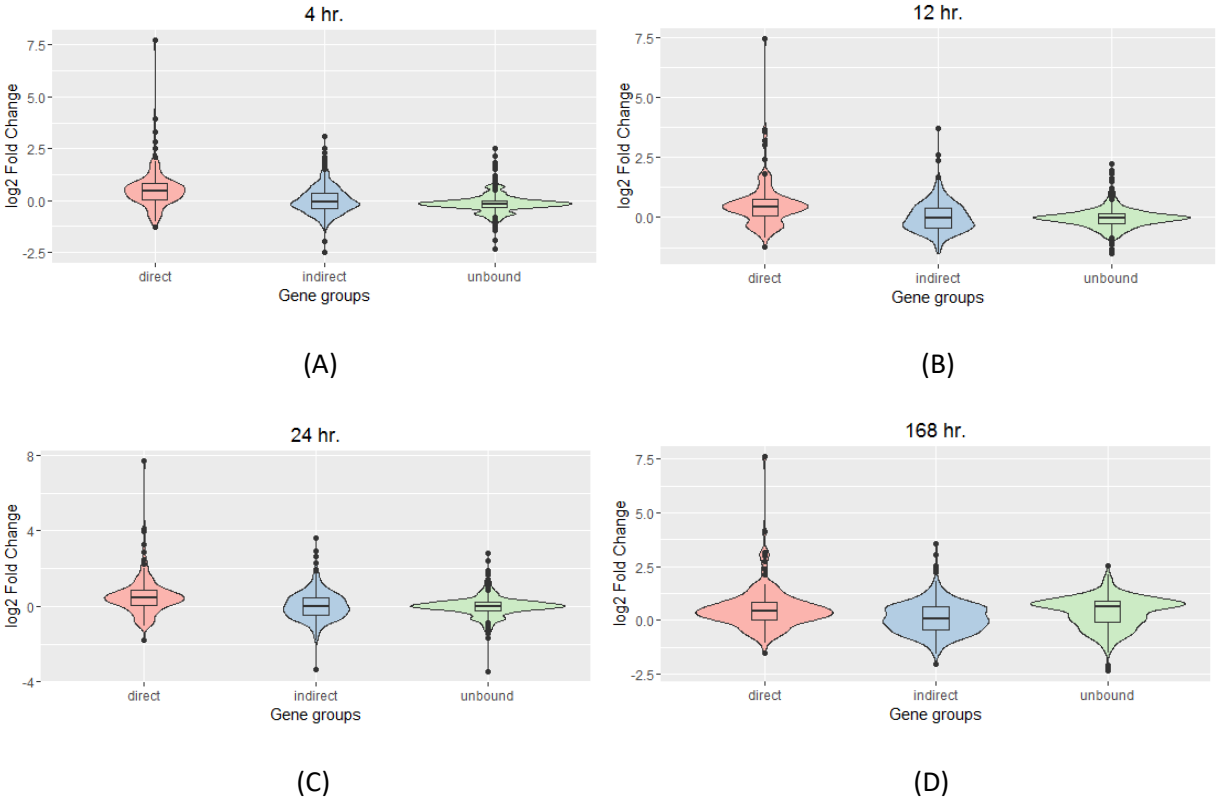


Figure 7: Overlaid box and violin plots showing the distribution in differential expression of direct (n = 140), indirect (n = 477) and unbound (n = 574) AHR target genes at 4 hr. (A), 12 hr. (B), 24 hr. (C) and 168 hr. (D). These plots illustrate the respective distributions of expression level of the three groups of genes at multiple time points, with the box plots illustrating the median, first and third quartile, and outliers; and the overlaid violin plots showing a rotated histogram of the distribution of gene expression.

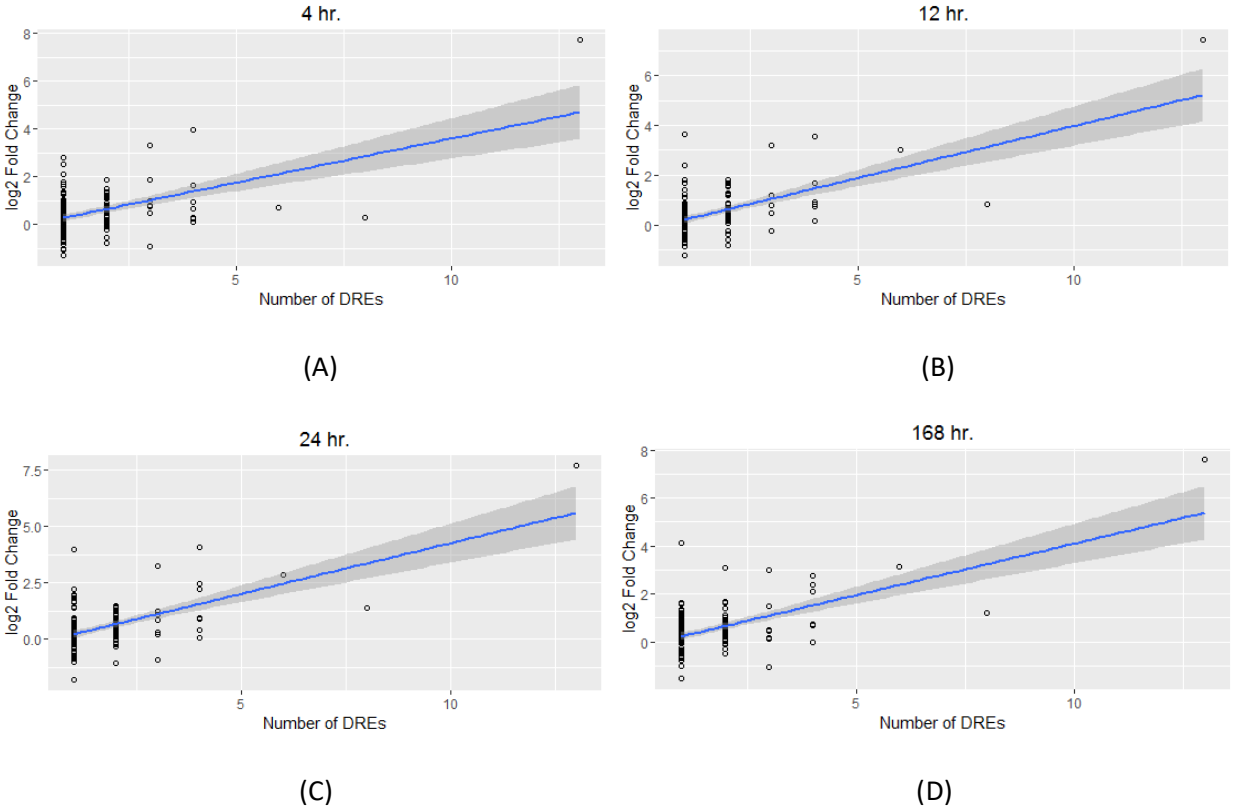


Figure 8: Increase in expression level of direct AHR target genes with number of DREs in proximal promoter regions at 4 hr. (A), 12 hr. (B), 24 hr. (C) and 168 hr. (D). Circles denote individual genes; linear regression fit shown in blue line with shaded region showing 95% confidence interval.

	Combinations	TF.Groups	Gene.Count
1	AHR	G1	210
2	FLI1	G2	36
3	NFE2L2	G3	13
4	KLF4	G4	55
5	SOX17	G5	10
6	AHR|FLI1	G6	15
7	AHR|NFE2L2	G7	12
8	AHR|KLF4	G8	54
9	AHR|SOX17	G9	7
10	FLI1|KLF4	G11	9
11	KLF4|SOX17	G15	6
12	AHR|FLI1|KLF4	G17	7
13	AHR|KLF4|SOX17	G21	5

Table 1: Count of genes in each transcriptional grouping. All possible groupings among the “hub” TFs in the network, AHR, Fli1, Nrf2 (NFe2L2), Klf4 and Sox17 were considered. Groups with gene counts > 5 are shown in the table.