bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Cross-species systems analyses reveal a conserved brain transcriptional response to social challenge

Michael C. Saul[1], Charles Blatti[1,2], Wei Yang[1,2], Syed Abbas Bukhari[1,3], Hagai Y. Shpigler[1,4], Joseph M. Troy[1,3], Christopher H. Seward[1,5], Laura Sloofman[1,6], Sriram Chandrasekaran[7], Alison M. Bell[1,3,8,9], Lisa Stubbs[1,3,5,9], Gene E. Robinson[1,9,10], Sihai Dave Zhao[1,11,!], and Saurabh Sinha[1,2,10,!].

Affiliations: [1] Carl R. Woese Institute for Genomic Biology; [2] Department of Computer Science; [3] Interdisciplinary Informatics Program, University of Illinois at Urbana-Champaign, Urbana, IL USA. [4] Department of Ecology, Evolution and Behavior, Hebrew University, Jerusalem, Israel. [5] Department of Cell and Developmental Biology, University of Illinois at Urbana-Champaign, Urbana, IL USA. [6] Genetics and Genomic Sciences, Mount Sinai Health System, New York, NY USA. [7] Biomedical Engineering, University of Michigan, Ann Arbor, MI USA. [8] Department of Animal Biology; [9] Neuroscience Program; [10] Department of Entomology; [11] Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL USA.

[!] Corresponding authors.

Address correspondence to:

Sihai Dave Zhao, email: sdzhao@illinois.edu

Saurabh Sinha, email: sinhas@illinois.edu

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Behavioral responses to social challenges like territorial intrusion occur in widely diverging species. Recent work has suggested that evolutionary “toolkits” – genes and pathways with lineage-specific variations but deep conservation of function – participate in the behavioral response to social challenge. Here, we studied this toolkit at scale by probing brain transcriptomic responses to social challenge in three distantly related species: honey bees, mice, and three-spined stickleback fish. We aggregated multi-species RNA-seq expression data from specific brain regions and time points after social challenge, achieving spatio-temporal resolution substantially greater than previous work. We conducted sequencing in parallel across species, allowing fair comparison of responses. Because of our complex study design, we developed new comparative analytical methods – including a new cross-species network analysis algorithm. We identified six orthogroups of genes involved in a conserved response to social challenge, including groups represented by *Npas4* and *Nr4a1*, as well as conserved modulation of gene systems such as transcriptional regulators, ion channels, G-protein coupled receptors, and synaptic proteins. We identified two deeply conserved gene modules enriched in mitochondrial fatty acid metabolism and heat shock proteins. Our analysis of this multi-species spatio-temporal expression dataset spanning phyla describes a system wherein nuclear receptors, interacting with chaperones, induce transcriptional changes in mitochondrial activity, neural cytoarchitecture, and synaptic transmission. These data provide support for the hypothesis that core genes and gene sets conserved across animal species have been repeatedly co-opted during evolution of analogous behaviors and may therefore be considered essential toolkits of response to social challenge.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

A pivotal idea arising from evolutionary developmental biology is that across the bilateria, the same signaling and transcription factor genes, known as “toolkit” genes (Carroll, et al. 2005), underlie the patterning of basic morphological features such as the body plan and eye. This provides a conceptual framework for increasingly detailed explanations of developmental patterning in specific model organisms (Wilkins 2002). Moreover, its success has motivated researchers to ask if the toolkit idea, where ancestral genetic programs coordinating fundamental processes undergird shared phenotypes, is also applicable to studies of behavior (Toth and Robinson 2007; Rittschof and Robinson 2016).

Studying toolkits for behavior poses numerous challenges, including the relative paucity of detailed and directly comparable genetics and genomics datasets for behavioral phenotypes in most animal species, difficulties in defining correspondence between behavioral phenotypes in diverged species from different ecological contexts, and ambiguity regarding brain regions and other tissues where shared molecular mechanisms may manifest. Further, behavioral phenotypes, being transitory and directly observable only while an animal is living, cannot be readily gleaned from fossils as developmental phenotypes are (Chen, et al. 2013), giving us little evidence from the distant past that contextualizes what we observe in extant species.

In an example of the above evolutionary approach, our group recently studied whether a conserved toolkit exists for the response to a territorial intrusion by a conspecific, more generally referred to as a social challenge, in three highly diverged model social species with well-assembled genomes: the mouse, the three-spined stickleback fish, and the honey bee (Rittschof, et al. 2014). Phylogenetic analyses strongly suggest convergent evolution of relatively sophisticated social phenotypes for these species (Woodard, et al. 2011; Kapheim, et al. 2015). We generated brain transcriptomic profiles across these three species 30 minutes after exposure to the intruder and discovered several common molecular mechanisms associated with the intruder response. Though similar to evolutionary development studies in its pursuit of an evolutionary “toolkit”, our earlier study was notably different for its use of gene expression rather than direct or indirect measures of gene sequence as the primary means to identify toolkit genes. Other groups have discussed similar conservation in aggressive behavior, though such conservation was only explicitly tested within the vertebrate subphylum

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Freudenberg, et al. 2016; Malki, et al. 2016) and tacitly assumed in analyses of arthropods (Asahina, et al. 2014).

Though suggestive of shared mechanisms, it is nevertheless not possible to use Rittschof, et al. (2014) to provide a thorough description of behavioral toolkits due to factors. First, expression was measured at only a single time point after animals were exposed to the social challenge, and relatively soon after exposure (20-30 min). Such a design cannot capture longer-acting genetic programs. This simple design also limits the power of this previous study to detect responses whose spatial and temporal profiles are shaped by the unique anatomical, neuroendocrine, and metabolic properties of brains in these three species (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017). Second, evolutionarily shared mechanisms are likely to be found at various levels of organization beyond single genes – gene orthogroups, modules, groups of genes dedicated to specific biological processes, or regulatory sub-networks (Rittschof and Robinson 2016) – but analytic tools that can identify such entities across multiple species, brain regions, and time points have heretofore been lacking, and the previous dataset also did not have enough samples for accurate *de novo* gene network analysis.

We report here the results of a new detailed investigation of the shared molecular roots of social behavior, specifically response to social challenge, that remedies the above issues by using both a novel experimental design and a suite of novel computational tools developed for deep cross-species comparisons. The new experiment was explicitly designed to probe discrete brain regions in mice, sticklebacks, and honey bees for their transcriptomic responses to territorial intruder. Measurements were taken in a time series after exposure, to glean a more holistic view of a dynamic process while allowing for inter-species differences in transcriptional trajectory over time. The individual species experiments were designed and conducted in parallel – and the individual species sequencing datasets were collected in parallel – to allow for the direct comparison of datasets with minimal technical effects. We developed new computational methods that allowed us to ascertain not only individual genes and biological processes, but also coordinately expressed networks and transcriptional regulatory cascades commonly modulating behaviors across these distantly related species. Our work goes beyond existing cross-species studies of tissue-specific (Lin, et al. 2014) or developmental time-course transcriptomes (Gerstein, et al. 2014) because here, we rigorously test and quantify

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

associations between behavioral responses and evolutionarily conserved transcriptomic patterns.

While the data from our experiments have already been studied at the individual species level (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017) and are publicly available, the main goal of the present work is to integrate these cross-species data for overall comparative analysis and discovery of conserved mechanisms. Such an approach allows the elucidation and unification of conserved molecular correlates of social behavior that may not seem important in individual species, but rise to significance when looking at all three species altogether. We report below the discovery of significant shared mechanisms at varying levels of molecular organization, later discussing our conclusions from the aggregate of such evidence at all levels.

NEW APPROACHES

Multi-scale characterization of conserved molecular basis for analogous cross-species phenotype

We probed the evolutionary toolkit of social challenge response at multiple levels of molecular organization in a uniform and systematic manner. For each level of organization – individual genes, cellular processes, co-expression modules, and TF regulons – we first identified homologous functional units in the three species as sets of genes that exhibited intra-species as well as inter-species commonality, e.g., involvement in the same cellular process, being paralogs or orthologs of each other, etc. We then tested each homologous functional unit for association with phenotype across all three species (see below). This systematic two-step approach is a novel feature of our work, and while our previous work (Rittschof, et al. 2014) reported an initial use of the approach, it is developed fully in this work.

Statistical methods for detecting simultaneous enrichment

We defined a given homologous functional unit to be associated with phenotype if all of its constituent species-specific gene sets are simultaneously enriched in phenotype-associated genes. We developed a new procedure to rigorously test for this simultaneous enrichment. We combined enrichment p-values obtained from each gene set, then tested the significance of the combined p-value by simulating a null distribution according to a precisely specified null model. Our approach gives more accurate control of false positive findings.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identifying cross-species co-expression modules with CNSRV

We developed a new method to discover homologous gene co-expression modules across divergent species. This was necessitated by our multi-scale analysis strategy but may also be of independent interest. Our approach integrates inferred gene co-expression modules in multiple highly diverged species with orthology mappings across these species (see details below and in **Materials and Methods**). Our module inference method, called "Common NetworkS ReVealed" (CNSRV), is closest in spirit to the OrthoClust method (Yan, et al. 2014), but uses a novel score for the quality of cross-species modules to avoid a bias towards large or small modules that is commonly seen with existing methods of module discovery (Langfelder and Horvath 2008). We performed systematic assessments to demonstrate that our novel methods led to less extreme module sizes and also found the resulting modules to be more statistically enriched for Gene Ontology terms.

Random walk-based approach for functional annotation of cross-species modules

Gene set enrichment tests are a popular approach to associate functions such as Gene Ontology biological processes with a given gene set. However, it is not clear how this approach can be extended to annotate homologous gene sets from multiple species in a way that also accounts for the available orthology information. We therefore adopted an alternative approach to functional annotation of a gene set in a single species, called DRaWR, which is based on Random Walk with Restarts on a network representation of genes and their annotations (Blatti and Sinha 2016). We extended this existing approach to the case of multi-species networks augmented with inter-species orthology edges. To our knowledge, the resulting "multi-species DRaWR" algorithm is the first method capable of functional annotation of gene sets in a cross-species manner.

RESULTS

Brain transcriptomic response to social challenge in three diverged species shares several orthologous gene groups

We profiled gene expression by sequencing mRNA at 30 min, 60 min, or 120 min after exposure to an intruder from discrete brain regions chosen for each species: the mushroom bodies in honey bee; the amygdala, frontal cortex, and hypothalamus in mouse; and the diencephalon

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and telencephalon in stickleback (see **Materials and Methods**). We considered only genes that were sufficiently expressed (see **Materials and Methods**) in these RNA-seq experiments for downstream analysis 10,701 in honey bee, 15,388 in mouse, and 17,435 in stickleback. Differentially expressed genes (DEGs) were obtained by comparison of intruder-exposed animals to control animals in matched conditions (see **Materials and Methods**), providing three sets of DEGs in honey bee, nine sets in mouse, and six sets in stickleback; these results have been reported elsewhere as individual species studies (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017) and are summarized in **Figure 1A**, but this is the first time that these data have been analyzed and discussed in a comparative context. These DEG sets varied in size between 36 genes (mouse amygdala, 60 min) to 1,151 (honey bee mushroom bodies, 120 min).

We were first interested in whether the same (orthologous) genes were associated with social challenge responses across these three species. However, the great evolutionary divergence of these species precludes unambiguous orthology assignments at the gene level. We instead used orthologous groups ("orthogroups") of genes as our fundamental unit of analysis. A resulting major analytical challenge is that most orthogroups contain different numbers of paralogs in the genomes of each species, and furthermore different numbers of brain regions were assessed in each species. This makes it difficult to ensure a fair comparison across species. Overcoming this issue requires carefully designed statistics, and existing approaches to this type of analysis, (such as the one we previously employed in ref. (Rittschof, et al. 2014)), cannot be applied. To address this problem, we developed a new method to identify orthogroups with the strongest evidence for activity in multiple species, where activity was measured by the proportion of DEGs, at any time point and brain region, within the orthogroup in each species (see **Materials and Methods**). Our new procedure is based on another algorithm we recently developed called Orthoverlap (Shpigler, Saul, Corona, et al. 2017) and offers stringent control of false positives.

We obtained 4,982 orthogroups common to the three species from the OrthoDB database(Kriventseva, et al. 2015), and our new method identified six orthogroups that were responsive to a social challenge in all three species at FDR ≤ 0.10 (**Figure 1B, Supplementary Table 1**). Three of the six contained at least one DEG in each of the species. Group EOG80K992 (p-value $\leq 1 \times 10^{-7}$), which contains the mouse genes *F5*, *Nrp2*, *Sned1*, and *Vwf* , is potentially involved in a deeply conserved immune response (Chang, et al. 2012), but is also

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

related to neurite outgrowth (Hey-Cunningham, et al. 2013) and axon guidance (Klagsbrun and Eichmann 2005). Group EOG8THX4X (p-value = 8.2 x 10^{-6}), which contains the mouse gene *Npas4*, is a gene that is involved in activity-dependent development of synapses(Lin, et al. 2008) and which regulates the balance between GABA and glutamate in neural circuits(Spiegel, et al. 2014). This finding is consistent with our previous work (Rittschof, et al. 2014), which also identified *Npas4* as a central gene in the conserved response to social challenge based on transcriptomic analysis (Saul, et al. 2017). Finally, group EOG8TMSCQ (p-value = 1.6 x 10^{-5}) contained subunits of the heat shock protein 70 family, which is nominally associated with stressors like heat shock that require protein refolding and that often acts in concert with co-chaperones in the heat shock protein 90 family (Mayer and Bukau 2005). Heat shock proteins from the Hsp70/Hsp90 complex have an additional documented but less discussed role, being necessary for ligand binding and subsequent signal transduction of nuclear receptors and other signaling molecules (Pratt and Toft 2003).

The remaining three statistically significant orthogroups contained DEGs in two out of the three species. We still considered these orthogroups of potential importance. For example, group EOG8M934T (p-value = 9.6 x 10^{-6}), which contains the mouse gene *Nr4a1*, only contained DEGs in honey bee and mouse. However, one of the stickleback orthologs was detected at an FDR of 0.1012 (uncorrected p-value = 0.0189) in telencephalon at 30 min, only slightly higher than the 10% FDR cutoff used for that species. This group of *Nr4a* orthologs, orphan nuclear receptors with unknown ligands, thus appears to have conserved socially regulated activity. These receptors, which are known to regulate glucose metabolism and homeostasis(Close, et al. 2013), have documented roles in memory and in object recognition (McNulty, et al. 2012) and have been documented as related to social aggression in vertebrates previously (Malki, et al. 2016). Additionally, group EOG8F4TSP (p-value = 2 x 10^{-7}), which contains “zinc finger of the cerebellum” (Zic) proteins, contained at least one DEG in both mouse and stickleback, but not in honey bee. This group of C2H2 zinc finger proteins is known for their evolutionarily conserved roles in neural development (Aruga 2004; Fujimi, et al. 2006).

Several of our findings were only made possible by the high resolution of our RNA-seq data in three species. For example, *Npas4* and *Nr4a1*, transcription factors involved in neural function and/or development, had not been identified as central molecules in the response to social challenge in each individual species (but see Rittschof, et al. 2014; Shpigler, Saul, Murdoch, et al. 2017), but our comparative analysis showed that these genes were consistently involved in

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the behavioral response in all three of our species. The multiple brain region/time point resolution of our RNA-seq data also allowed us to identify conserved genes that are transiently expressed, and/or expressed in a brain-region specific manner. For example, the heat shock orthogroup, which contains the chaperone gene *Hspa1a*, a potential cofactor with nuclear receptors like *Nr4a1*, was only active at 120 min in the mouse and in the diencephalon in the stickleback.

Social challenge triggers conserved hormone-dependent neuronal signaling

Conserved mechanisms of the response to social challenge may emerge at higher levels of organization than that of individual genes. We asked if the same cellular processes (e.g., Gene Ontology terms) are transcriptionally active in response to social challenge, even if specific genes exhibiting differential expression are not strictly orthologs of each other. This allows us to be more sensitive to cellular mechanisms that may have evolved convergently, by repeatedly coopting the same biological pathways.

We considered gene sets defined by 341 GO terms that contained at least 5 genes in each of the species studied here. Using the new method that we developed for our analysis of conserved gene orthogroups above, we identified those GO terms that had the strongest evidence for enrichment of DEGs in each of the three species. We identified 66 GO terms at FDR ≤ 0.10 and 37 GO terms at a more stringent threshold of FWER ≤ 0.10 (**Table 1**). These centered around five major categories: hormone activity, transmembrane transport, G-protein coupled signal transduction, synaptic activity, and extracellular matrix components. This analysis has thus identified a set of processes that, though they have a slightly different complement of genes between distantly related phyla, correspond to the same general functions. Specifically, these results suggest that hormone receptors, as nuclear receptors, signaling molecules and transcription factors, are essential in the coordination of the large-scale social challenge induced transcriptional responses that potentially cause remodeling of axons and dendrites, which lead to differences in synapse-related proteins, extracellular matrix proteins, transmembrane transporters, and the modulation of GPCRs for neural signaling.

A novel method identifies conserved gene co-expression modules that respond to social challenge

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In addition to defining sets of genes using GO terms, we sought to identify coordinately expressed sets of genes, often called gene modules, *ab initio*, without the need for prior knowledge. These have become a mainstay of systems-level analysis of transcriptional programs (Langfelder and Horvath 2008). Studies in evolutionary developmental biology have noted that gene modules underlying development are deeply conserved, and are an important facet of the genetic “toolkit” concept (Toth and Robinson 2007; Peter and Davidson 2011; Rittschof and Robinson 2016). Co-expressed modules are also conserved in other biological contexts across evolutionary spans as great as humans, flies, worms, and yeast (Stuart, et al. 2003).

Using the newly developed CNSRV method for cross-species analysis, we sought to discover deeply conserved gene modules (see schematic representation of coexpressed conserved modules in **Figure 2A**) from our multi-species brain transcriptomic data, then query if any of these are regulated by social challenge commonly across the three species. Our experimental design allowed us to characterize modules with coordinated spatiotemporal expression profiles in bee, mouse, and stickleback. We then combined these results with our DEGs to identify modules that were highly responsive to social challenge (see **Materials and Methods**). These represent deeply conserved core regulatory programs, where conservation is at the level of module rather than gene.

With CNSRV, we identified 20 homologous modules (**Figure 2B**), each ranging between 140 and 523 genes in size (see **Materials and Methods** and **Supplementary Table 2**). These modules show both dense co-expression within modules in individual species (Figure 2B, central diagonal) and elevated frequency of orthology relationships between corresponding modules (Figure 2B, ancillary diagonals). Next, for each combination of brain region and time point in each species, we tested if DEGs were differentially distributed among the modules (see **Materials and Methods**), and found this to be the case (FDR $\leq$ 0.10) for all but one of the 18 species/brain region/time point combination (**Figure 2C**). Within these 17 significant combinations of region, time, and species, we then conducted *post-hoc* tests at FWER $\leq$ 0.10 to identify significantly enriched modules.

This analysis revealed that two gene co-expression modules, numbered 10 and 14, have conserved social challenge-specific activity across all three species (**Figure 2C**). Specifically, module 10 is significantly associated with DEGs in honey bee mushroom body (60 min), mouse

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

frontal cortex (120 min) and hypothalamus (30 min), as well as stickleback diencephalon (30 min) and telencephalon (30 min). Similarly, module 14 is enriched for DEGs in honey bee mushroom body (60 min), mouse hypothalamus (120 min), and stickleback diencephalon (60 min). We note that time points where the orthologous modules were observed often did not match between the species, which may have resulted from differences in the timing of behavioral responses between the species, underscoring the importance of multiple time points in the study design.

While it was instructive to observe conserved modules apparently regulated by social challenge, it was not as clear what biological functions these modules might be involved with. Functional annotation of these modules is difficult because a module in our context is not a single list of genes but a set of three different species-specific gene lists, and standard gene set enrichment tests do not take this cross-species orthology relationships into account. To solve this problem, we used our previously developed tool DRaWR (Blatti and Sinha 2016), which considers a network whose nodes are genes and annotations (e.g., Gene Ontology terms) and edges connect a gene to each of its annotations. It annotates a gene set by performing a random walk starting from nodes in the gene set and recording the annotation nodes that are visited most frequently. We extended this approach here to annotate the orthologous CNSRV modules by constructing a network using module genes, GO annotations, and orthology edges from all three of our species (see **Materials and Methods**).

Figure 2D shows the top functional annotations for modules 10 and 14, as revealed by a high DRaWR percentile score in every species, and with additional support from standard enrichment tests (hypergeometric test nominal p-value ≤ 0.05) in at least two of the three species. Module 14 comprises genes involved in cell-matrix adhesion, a process involved in neural development and plasticity (Murase and Schuman 1999); Rho GTPase binding, a process implicated in several aspects of neuronal development as well as neurological diseases (Govek, et al. 2005); and actin binding, a process associated with function and plasticity of dendritic spines and synapses (Lin and Webb 2009). Module 10 includes genes annotated for AMP deaminase activity and IMP biosynthesis, processes associated with purine balance in the brain. Purine balance and purinergic reception play well-known roles in neuronal repair and protection, acting as a bridge between neural signaling and the neural immune system in mammals (Skaper, et al. 2010; Thauerer, et al. 2012). Further, the enrichment of enoyl-CoA hydratase activity found in Module 10, as a step of fatty acid metabolism found in the Cellular

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Component and Molecular Function results, potentially bridges neural signaling and the metabolic processes previously observed in response to social challenge both across species and within individual species (Rittschof, et al. 2014; Chandrasekaran, et al. 2015). These co-expression modules bolster evidence from the conserved DEGs and from the GO results in support of a conserved transcriptomic response that includes structural proteins, heat shock proteins, and GPCR signaling proteins.

Common transcription factor regulatory activities underlie social challenge

The previous sections provide new insights into conserved biological processes and gene modules underlying the response to social challenge. We next sought to identify transcription factors (TFs) that act as master regulators of those processes and modules, using state-of-the-art tools for reconstruction of transcriptional regulatory networks in each species. In particular, we asked if the same TFs (or their paralogs) regulate brain transcriptomic response to social challenge across species. TF-gene relationships are among the best studied and most widely accepted conception of gene networks, and they have been explored in the context of genetic toolkit studies in evo-devo (Rittschof and Robinson 2016). The gene orthogroup analysis reported above (**Figure 1B**) identified multiple TF orthogroups containing social challenge DEGs; however there may be TFs which do not detectably change in transcript expression, but may for example be activated by post-transcriptional modifications. Regulatory targets for these TFs may nevertheless be socially regulated and the TFs reasonably speculated to have a role in the transcriptional response to social challenge.

To explore this idea, we constructed transcriptional regulatory networks (TRNs) for each species using the previously developed tool ASTRIX (Chandrasekaran, et al. 2011), which uses the ARACNE algorithm (Margolin, et al. 2006) to identify putative TFs for a gene, then employs Least Angle Regression (Efron, et al. 2004) to identify those TFs that best predict expression levels of that gene target in multiple experiments. In this case, each TRN had been reconstructed from different brain regions and time points within each individual species previously (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017). We used these previously reconstructed TRNs to identify TF orthogroups whose gene targets were enriched in DEGs in all three species, using the same method as described above for identifying conserved gene orthogroups and GO terms. We considered only orthogroups that contained at

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

least one TF with at least one gene target in each species. This analysis detects TFs important to social challenge even if the TFs themselves are not significantly differentially expressed.

We detected six TF orthogroups (FDR $\leq$ 0.10) that are likely to be conserved regulators of the transcriptomic response to social challenge (**Table 2**). For instance, the orthogroup EOG8KWM99 comprises the mouse TF genes *Pbx1* and *Pbx3*, for which the ASTRIX-derived TRN included 2 target genes in mouse, both of which are social challenge DEGs, 45 targets (including 3 DEGs) in stickleback and 25 targets (including 9 DEGs) in honey bee. Further, one orthogroup containing the mouse neural development transcription factor genes *Rax* and *Pax6* may be involved in the conserved regulation of the formation of new neurons from a neural stem cell lineage (Davis, et al. 2003; Pak, et al. 2014). *Rax* was identified as a transcriptional regulator in our earlier work (Rittschof, et al. 2014). One particularly interesting TF, the orphaned nuclear receptor mouse gene *Nr2e1*, has been implicated in our previous cross-species work (Rittschof, et al. 2014), in aggression in mice (Abrahams, et al. 2005), and in aggression in flies (Davis, et al. 2014). These results identify specific transcriptional regulators that appear to be important central regulators of the processes described in the above sections and therefore constitute potential key conserved master regulators of the transcriptional response to challenge.

DISCUSSION

The evolution of gene regulatory programs is a subject of long-standing interest (Halfon and Michelson 2002) and has been studied by cross-species comparisons of cis-regulatory sequences (Sinha, et al. 2004), TF-DNA binding (Consortium 2012), as well as gene expression measurements in matched tissues and organs (Gerstein, et al. 2014; Lin, et al. 2014; Breschi, et al. 2017). An important achievement of our study was its explicit coupling of gene and gene network comparisons with objectively defined and analogous phenotypic states measured experimentally. This approach utilizes an array of novel tools with a common theme of studying different tiers of organization for evidence of a shared genetic program: each test assays if groups of related genes – orthogroups, functional systems, co-expressed modules, or transcription factor regulons – have a non-random association with socially responsive genes expressed in the brain in multiple species. The technical novelty of our methods allowing for comparative associations between genes and phenotypes does not just elucidate new clues to the mechanisms of response to social challenge; it will allow for future work to identify similar

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

deeply conserved molecular systems in association with other phenotypes of interest. Moreover, the scope of our transcriptome-wide comparisons distinguishes this work from more directed studies of regulatory evolution where expression and cis-regulatory divergence of individual genes was linked to morphological differences between species (Wray 2007). Our goal is similar to the work of Malki, et al. (2016), who identified compared aggression-related DEGs in prefrontal cortex of mouse and zebrafish, but our study pursues the goal through an experiment design whereby data were collected from multiple species in a parallel manner, addressing several key technical and statistical challenges in the process.

One technical challenge observed in our data is the difficulty in matching gene expression sets across such long evolutionary distances. We note that time points where the orthologous modules were observed often did not match between the species, which may have resulted from differences in metabolic rates between the species. This observation underscores the importance of temporal series in our study design. Furthermore, it demonstrates that experimental design in future studies must proceed carefully to identify matching expression sets across species. We were able to go beyond identification of differentially expressed genes and rigorously analyze co-expression relationships only because our experimental design included multiple brain regions and time points. Thus, the design gave us access to the higher order biological mentioned above, significantly elaborating upon our earlier work (Rittschof, et al. 2014).

Using results derived from these novel methods developed for our data, we propose a system of genes acting commonly in the adult brain of these diverged species to transduce social challenge stimuli into transcriptional and epigenomic responses. This is graphically summarized in **Figure 3**. It involves the integration of nuclear receptor signaling to drive the transcriptional regulatory events that result in changes in neural signaling observed after a social challenge. We speculate that because nuclear receptors are both liganded receptors and transcription factors, they act as key drivers of the large-scale transcriptional changes seen across all of these species. We further speculate that these transcriptional changes happen in concert with transcription factors commonly associated with neural development to drive neural signaling modulation, which likely take place through alterations in dendritic architecture, axon architecture, signaling molecules like GPCRs and ion channels, mitochondrial metabolism, or all of these processes simultaneously.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this pathway, we call specific attention to the signaling molecules, transcription factors, and nuclear receptors that can act as both. Specifically, the various homologs of *Npas4*, *Nr2e1*, and *Nr4a1* are transcription factor genes well-known in neural response to stimuli (Abrahams, et al. 2005; Maxwell and Muscat 2006; Kim, et al. 2010). We speculate that the ancestral versions of these genes, which were likely present in the most recent common ancestor of all living bilaterians, were potentially already active in the response to social challenge stimuli that was exhibited by their contemporaries around the time of the Cambrian explosion (Carbone and Narbonne 2014). Translating these gene expression patterns into knowledge about how the cellular systems inside the brain change in response to social challenge is an important next step. Such research will require careful work across species to identify important points of similarity as well as how these systems diverge.

Though we discussed the role and neurobiological relevance of some of the above-mentioned systems in detail in our previous work – we described hormone receptors in sticklebacks (Bukhari, et al. 2017), developmental transcription factors in mice (Saul, et al. 2017), dendritic architecture in honey bees (Shpigler, Saul, Murdoch, et al. 2017), and GPCRs in all three species (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017) – the present work unifies these systems in their role in social responsiveness into a whole. The genes and systems we propose as drivers of the response to social challenge constitute real, testable connections for a conserved genetic program for the response to a social challenge, something that was lacking before this analysis. However, we note that these genes may not be specific to social contexts, but may instead coordinate information from multiple contexts, and thus, the specificity of these gene sets for social challenge response also needs rigorous testing.

MATERIALS AND METHODS

DEGs

RNA-seq data were collected as described previously (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017). The data for these three sets are deposited in the GEO under accession numbers: GSE85876 (honey bee), GSE80346 (mouse), and GSE96673 (threespined stickleback). In each species, we used a 1 CPM cutoff for an equivalent of the smallest group size for expression, as proposed in the edgeR documentation (Robinson, et al. 2010). FDR thresholds from each individual species paper – 5% for bee, 10% for mouse, and

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10% for stickleback – were used to compile the DEG lists. We chose a lower FDR threshold for the honey bee because its experimental design was more powerful.

OrthoDB

Using the raw data from OrthoDB v8 (Kriventseva, et al. 2015), we first filtered for the three species of interest. We then identified the orthogroups present within all of the three individual species used in this experiment, a total of 4,982 orthogroups. We found all paralogs inside of each orthogroup for the individual species, which brought us to a total of 10,158 genes in mouse, 6,725 genes in bee, and 10,869 genes in stickleback.

Identifying orthogroups with a conserved response to social challenge

In each species, for each orthogroup we only considered genes in the orthogroup and in the corresponding species' gene "universe", that is, the full complement of genes expressed above a threshold in each species. Under the null hypothesis of no orthogroup activity in response to social challenge, we modeled the number of DEGs contained in an orthogroup as a hypergeometric random variable. We tested if each orthogroup contained more DEGs from that species than expected by chance, using a one-tailed hypergeometric test. We did not separate brain region- and time point-specific DEGs within each species at this stage of analysis. This resulted in three p-values for each orthogroup, p_{bee}, p_{mouse}, and p_{fish}, which we then aggregated using Fisher's combination test statistic $T = -2\ln p_{bee} - 2\ln p_{mouse} - 2\ln p_{fish}$.

For each orthogroup triplet, we calculated the p-value of the test statistic T under the reasonable assumption that the p_{bee}, p_{mouse}, and p_{fish} were statistically independent. If they were uniformly distributed, classical theory gives that T would be χ_6^2-distributed under the null hypothesis that none of the three orthogroups was responsive to social challenge. However, due to the discrete nature of the hypergeometric variables from which the p_{bee}, p_{mouse}, and p_{fish} were calculated, we resorted to simulations to calculate the true p-value of T. We simulated 5 million instances of the hypergeometric variables for each orthogroup in each species under the null hypothesis and calculated the p-value of each othogroup triplet's T using the simulated distribution.

Technically, the alternative hypothesis of this test is that there is at least one species in which the corresponding orthogroup is enriched in social challenge DEGs. This does not exactly match the conservation hypothesis, which should state that *all* orthogroups in all three species are enriched in DEGs. However, this latter hypothesis corresponds to a composite null

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

hypothesis, which is difficult to formally test without sacrificing a great deal of statistical power. Here we instead test the simpler sharp null hypothesis where all orthogroups are inactive, but it is well-known that the test statistic T that we have chosen is most powerful when all three orthogroups are enriched in DEGs. Thus our tests are oriented toward the desired conservation hypothesis, and our results indicate that we are indeed able to capture deeply conserved orthogroups.

Identifying GO terms and TF orthogroups with a conserved response to social challenge

We downloaded GO annotations for each species from Ensembl Biomart (Ensembl v83, Kinsella, et al. 2011) and considered only the 341 terms that contained at least 5 genes. We used our newly developed orthogroup analysis method, described above, to identify terms that were significantly enriched in DEGs in multiple species.

TRNs were reconstructed for each individual species individually as previously described (Bukhari, et al. 2017; Saul, et al. 2017; Shpigler, Saul, Murdoch, et al. 2017). In each species, for each orthogroup of TFs, we collected the gene targets of all TFs in the orthogroup into a single set. We then used our analysis method to identify TF orthogroups whose target sets were enriched in DEGs in multiple species.

CNSRV

Construction of co-expression networks: For each species, we first calculated coexpression of gene pairs as the Pearson correlation of their expression values in a specific brain region at different time points after exposure (including intruder-exposed as well as control animals), and retained pairs that had correlation coefficient above 0.7 in all brain regions considered for that species.

Cross-species co-expression module detection: The algorithm partitions the genes in each species' co-expression network into $K = 20$ non-overlapping clusters, referred to be identifiers 1, 2, … K, such that cluster i in one species "corresponds to" clusters labeled i in the other species. The algorithm seeks to find partitions such that (1) clusters in each species exhibit "modularity" (Newman 2006) – high density of within cluster coexpression edges compared to cross-cluster density of such edges, and (2) corresponding clusters in a pair of species exhibit high density of orthology edges. (An orthology edge is created for any pair of genes in the same

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

orthogroup from the two species). To meet these two goals, the CNSRV method attempts to maximize the following objective function:

$$\boldsymbol{Q} = (\boldsymbol{1}-\boldsymbol{\lambda})\sum_{\boldsymbol{s}=\boldsymbol{1}}^{\boldsymbol{S}}\sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{K}} \widehat{\boldsymbol{w}}_{\boldsymbol{ks}}\boldsymbol{log}_{\boldsymbol{2}}(\widehat{\boldsymbol{v}}_{\boldsymbol{ks}}) + \boldsymbol{\lambda}\sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{K}}\sum_{\boldsymbol{i},\boldsymbol{j}\in[\boldsymbol{1}..\boldsymbol{S}],\boldsymbol{i}\neq\boldsymbol{j}}\ \sum_{(\boldsymbol{a},\boldsymbol{b})\in\boldsymbol{Orth}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k})} \widehat{\boldsymbol{\omega}}_{\boldsymbol{ab}}$$

Here *S* is the number of species, *K* is the desired number of clusters. $\widehat{w}_k$ is the normalized count of co-expression edges in cluster *k* of species *s*, defined as $\widehat{w}_{ks} = w_{ks}/E_s$, where w_{ks} is the number of co-expression edges in cluster *k* of that species and E_s is the total number of edges in that species. Similarly, $\hat{v}_{ks}$ is the normalized count of co-expression edges connected to nodes in cluster *k* of species *s*, defined as $\hat{v}_{ks} = v_{ks}/E_s$, where v_{ks} is the count of co-expression edges incident to nodes in cluster *k* in that species. (a, b) refers to any pair of orthologous genes from species *i* and *j* such that both genes are in cluster *k* of their respective species. To normalize the number of orthologous edges from many-to-many gene mappings, $\widehat{\omega}_{ab} = 1/2\,(1/d_a + 1/d_b)$ where d_a is the number of orthologous edges from gene a in species i to genes in species j. The two terms in this formula represent the “modularity” and “orthology” goals respectively, and are weighted by factors of λ and (1 - λ) respectively. We chose a value of λ = 0.05 to provide a suitable balance between the co-expression modularity and cross-species sharing aspects of our desired gene modules (**Supplementary Figure SF1**).

The objective function is maximized with a Simulated Annealing algorithm. Initially, genes are assigned random cluster labels from 1 to *K* and the “temperature” variable is set to 10. In each proposed move, a gene is selected at random and assigned a different cluster label. The objective function is re-evaluated, “good” moves that generate a better score are accepted, whereas “bad” moves are rejected with probability that depends on the score of the proposed reassignment and the temperature variable. Specifically, the probability of accepting a proposed move that generates a new clustering with score Q_{new}, assuming the current score is Q_{cur}, is given by min(1, $(Q_{new}/Q_{old})^T$), where temperature *T* changes across iterations according to the cooling schedule $T_{k+1} = \alpha T_k$, where $\alpha = 0.9$, and *k* is the iteration index. This results in bad moves being rejected with low probability in earlier iterations (when the “temperature” is higher), and with higher probability in later iterations. The iterative procedure stops once no good move can be found after certain amount of attempts, or a pre-determined number of iterations have been performed.

Identifying gene co-expression modules with a conserved response to social challenge

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We used 19-df chi-square tests of independence to test if the DEGs in each species/brain region/time point combination were distributed randomly across the 20 modules. A non-random distribution indicates that exposure to social challenge results in certain modules being more activated than others. To identify the active ones, we used *post-hoc* hypergeometric enrichment tests in the species/brain region/time point combinations with significant chi-square tests.

To annotate these active modules, we extended our previously reported DRaWR tool (Blatti and Sinha 2016, see **Supplemenatary Methods** for details). DRaWR takes a heterogeneous biological network and ranks all annotation nodes in the network for their proximity to a set of gene nodes of interest by using random walks. We constructed a network containing "gene nodes" representing genes from all three species and "annotation nodes" that represent Gene Ontology annotations (obtained from Biomart for Ensembl v83, Kinsella, et al. 2011) and Pfam domains (whose presence was predicted using HMMER, Finn, et al. 2011). Edges connected genes with their properties (GO annotations and Pfam domains), and also connected homologous pairs of genes from the same or different species. For a given module, we executed the DRaWR random walk now with restarts from module genes from all three species, so that the method is also able to "walk" from a gene to its ortholog(s) in other species. Separately, we also executed the random walk with restarts from module genes of each species individually, and selected annotation nodes that were ranked highest across all four restart configurations. As such, an annotation that is highly ranked by our new multi-species DRaWR technique is either enriched in module genes from multiple species or enriched in orthologs of those genes (even if it is not enriched in the module genes themselves), or both. We also required that the reported annotations be significantly enriched (p-value < 0.05 using one-sided Fisher exact test) in at least one of the three species.

ACKNOWLEDGEMENTS

We would like to thank Yoshitsugu Oono and Jian Ma for their critical comments during the compilation of this manuscript. This research was primarily funded by grant #SFLife 291812 from the Simons Foundation. In addition, this research was supported in part by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative and by NSF grant DMS-1613005.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Abrahams BS, Kwok MC, Trinh E, Budaghzadeh S, Hossain SM, Simpson EM. 2005. Pathological aggression in "fierce" mice corrected by human nuclear receptor 2E1. J Neurosci 25:6263-6270.

Aruga J. 2004. The role of Zic genes in neural development. Mol Cell Neurosci 26:205-221.

Asahina K, Watanabe K, Duistermars BJ, Hoopfer E, Gonzalez CR, Eyjolfsdottir EA, Perona P, Anderson DJ. 2014. Tachykinin-expressing neurons control male-specific aggressive arousal in Drosophila. Cell 156:221-235.

Blatti C, Sinha S. 2016. Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks. Bioinformatics 32:2167-2175.

Breschi A, Gingeras TR, Guigo R. 2017. Comparative transcriptomics in human and mouse. Nat Rev Genet 18:425-440.

Bukhari SA, Saul MC, Seward CH, Zhang H, Bensky M, James N, Zhao SD, Chandrasekaran S, Stubbs L, Bell AM. 2017. Temporal Dynamics of Neurogenomic Plasticity in Response to Social Interactions in Male Threespined Sticklebacks. PLOS Genetics 13:e1006840.

Carbone C, Narbonne GM. 2014. When life got smart: the evolution of behavioral complexity through the Ediacaran and early Cambrian of NW Canada. Journal of Paleontology 88:309-330.

Carroll SB, Grenier JK, Weatherbee SD. 2005. From DNA to diversity : molecular genetics and the evolution of animal design. Malden, MA: Blackwell Pub.

Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE. 2011. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proceedings of the National Academy of Sciences of the United States of America 108:18020-18025.

Chandrasekaran S, Rittschof C, Djukovic D, Gu H, Raftery D, Price N, Robinson G. 2015. Aggression is associated with aerobic glycolysis in the honey bee brain1. Genes, Brain and Behavior 14:158-166.

Chang HJ, Dhanasingh I, Gou X, Rice AM, Dushay MS. 2012. Loss of Hemolectin reduces the survival of Drosophila larvae after wounding. Dev Comp Immunol 36:274-278.

Chen Z, Zhou C, Meyer M, Xiang K, Schiffbauer JD, Yuan X, Xiao S. 2013. Trace fossil evidence for Ediacaran bilaterian animals with complex behaviors. Precambrian Research 224:690-701.

Close AF, Rouillard C, Buteau J. 2013. NR4A orphan nuclear receptors in glucose homeostasis: a minireview. Diabetes Metab 39:478-484.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Davis RJ, Tavsanli BC, Dittrich C, Walldorf U, Mardon G. 2003. Drosophila retinal homeobox (drx) is not required for establishment of the visual system, but is required for brain and clypeus development. Dev Biol 259:272-287.

Davis SM, Thomas AL, Nomie KJ, Huang L, Dierick HA. 2014. Tailless and Atrophin control Drosophila aggression by regulating neuropeptide signalling in the pars intercerebralis. Nat Commun 5:3177.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression.407-499.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39:W29-37.
Freudenberg F, Carreno Gutierrez H, Post AM, Reif A, Norton WH. 2016. Aggression in non-human vertebrates: Genetic mechanisms and molecular pathways. Am J Med Genet B Neuropsychiatr Genet 171:603-640.
Fujimi TJ, Mikoshiba K, Aruga J. 2006. Xenopus Zic4: conservation and diversification of expression profiles and protein function among the Xenopus Zic family. Dev Dyn 235:3379-3386.
Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, et al. 2014. Comparative analysis of the transcriptome across distant species. Nature 512:445-448.
Govek EE, Newey SE, Van Aelst L. 2005. The role of the Rho GTPases in neuronal development. Genes Dev 19:1-49.
Halfon MS, Michelson AM. 2002. Exploring genetic regulatory networks in metazoan development: methods and models. Physiol Genomics 10:131-143.
Hey-Cunningham AJ, Markham R, Fraser IS, Berbic M. 2013. Dysregulation of vascular endothelial growth factors and their neuropilin receptors in the eutopic endometrium of women with endometriosis. Reprod Sci 20:1382-1389.
Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, et al. 2015. Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. Science 348:1139-1143.
Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182-187.
Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.
Klagsbrun M, Eichmann A. 2005. A role for axon guidance receptors and ligands in blood vessel development and tumor angiogenesis. Cytokine Growth Factor Rev 16:535-548.
Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Research 43:D250-256.
Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.
Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A 111:17224-17229.
Lin WH, Webb DJ. 2009. Actin and Actin-Binding Proteins: Masters of Dendritic Spine Formation, Morphology, and Function. Open Neurosci J 3:54-66.
Lin Y, Bloodgood BL, Hauser JL, Lapan AD, Koon AC, Kim TK, Hu LS, Malik AN, Greenberg ME. 2008. Activity-dependent regulation of inhibitory synapse development by Npas4. Nature 455:1198-1204.
Malki K, Du Rietz E, Crusio WE, Pain O, Paya-Cano J, Karadaghi RL, Sluyter F, de Boer SF, Sandnabba K, Schalkwyk LC, et al. 2016. Transcriptome analysis of genes

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and gene networks involved in aggressive behavior in mouse and zebrafish. Am J Med Genet B Neuropsychiatr Genet 171:827-838.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7.

Maxwell MA, Muscat GE. 2006. The NR4A subgroup: immediate early response genes with pleiotropic physiological roles. Nucl Recept Signal 4:e002.

Mayer MP, Bukau B. 2005. Hsp70 chaperones: cellular functions and molecular mechanism. Cell Mol Life Sci 62:670-684.

McNulty SE, Barrett RM, Vogel-Ciernia A, Malvaez M, Hernandez N, Davatolhagh MF, Matheos DP, Schiffman A, Wood MA. 2012. Differential roles for Nr4a1 and Nr4a2 in object location vs. object recognition long-term memory. Learning & Memory 19:588-592.

Murase S, Schuman EM. 1999. The role of cell adhesion molecules in synaptic plasticity and memory. Curr Opin Cell Biol 11:549-553.

Newman ME. 2006. Modularity and community structure in networks. Proc Natl Acad Sci U S A 103:8577-8582.

Pak T, Yoo S, Miranda-Angulo AL, Wang H, Blackshaw S. 2014. Rax-CreERT2 knock-in mice: a tool for selective and conditional gene deletion in progenitor cells and radial glia of the retina and hypothalamus. PLoS ONE 9:e90381.

Peter IS, Davidson EH. 2011. Evolution of gene regulatory networks controlling body plan development. Cell 144:970-985.

Pratt WB, Toft DO. 2003. Regulation of signaling protein function and trafficking by the hsp90/hsp70-based chaperone machinery. Exp Biol Med (Maywood) 228:111-133.

Rittschof CC, Bukhari SA, Sloofman LG, Troy JM, Caetano-Anollés D, Cash-Ahmed A, Kent M, Lu X, Sanogo YO, Weisner PA. 2014. Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. Proceedings of the National Academy of Sciences 111:17929-17934.

Rittschof CC, Robinson GE. 2016. Behavioral Genetic Toolkits: Toward the Evolutionary Origins of Complex Phenotypes. Curr Top Dev Biol 119:157-204.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.

Saul MC, Seward CH, Troy JM, Zhang H, Sloofman LG, Lu X, Weisner PA, Caetano-Anolles D, Sun H, Zhao SD, et al. 2017. Transcriptional regulatory dynamics drive coordinated metabolic and neural response to social challenge in mice. Genome Res.

Shpigler HY, Saul MC, Corona F, Block L, Cash Ahmed A, Zhao SD, Robinson GE. 2017. Deep evolutionary conservation of autism-related genes. Proc Natl Acad Sci U S A.

Shpigler HY, Saul MC, Murdoch EE, Cash-Ahmed AC, Seward CH, Sloofman L, Chandrasekaran S, Sinha S, Stubbs LJ, Robinson GE. 2017. Behavioral, transcriptomic and epigenetic responses to social challenge in honey bees. Genes Brain Behav.

Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. 2004. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics 5:129.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Skaper SD, Debetto P, Giusti P. 2010. The P2X7 purinergic receptor: from physiology to neurological disorders. The FASEB Journal 24:337-345.

Spiegel I, Mardinly AR, Gabel HW, Bazinet JE, Couch CH, Tzeng CP, Harmin DA, Greenberg ME. 2014. Npas4 regulates excitatory-inhibitory balance within neural circuits through cell-type-specific gene programs. Cell 157:1216-1229.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302:249-255.

Thauerer B, zur Nedden S, Baier-Bitterlich G. 2012. Purine nucleosides: endogenous neuroprotectants in hypoxic brain. Journal of Neurochemistry 121:329-342.

Toth AL, Robinson GE. 2007. Evo-devo and the evolution of social behavior. Trends in Genetics 23:334-341.

Wilkins AS. 2002. The evolution of developmental pathways. Sunderland, Mass.: Sinauer Associates.

Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. Proc Natl Acad Sci U S A 108:7472-7477.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206-216.

Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. 2014. OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol 15:R100.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Multiple cross-species-mapped GO terms show conserved activity in response to social challenge.

Biological Process GO ID – Term	**P Sim.**	**Ratio (Hits / Total Genes in GO Term)**		
		Honey Bee	Mouse	Stickleback
GO:0007186 – G-protein coupled receptor signaling pathway	$< 1 \times 10^{-7}$	30/139	52/344	58/399
GO:0007218 – neuropeptide signaling pathway	$< 1 \times 10^{-7}$	5/17	20/70	2/6
GO:0007601 – visual perception	$< 1 \times 10^{-7}$	1/6	7/71	12/16
GO:0055085 – transmembrane transport	$< 1 \times 10^{-7}$	36/246	42/338	67/425
GO:0007155 – cell adhesion	2.0×10^{-7}	8/54	40/308	20/120
GO:0006836 – neurotransmitter transport	4.0×10^{-7}	4/16	11/28	3/28
GO:0043401 – steroid hormone mediated signaling pathway	5.4×10^{-7}	11/21	9/52	5/61
GO:0007169 – transmembrane receptor protein tyrosine kinase signaling pathway	5.8×10^{-6}	1/7	14/81	11/45
GO:0006811 – ion transport	6.6×10^{-6}	14/108	17/176	39/190
GO:0006366 – transcription from RNA polymerase II promoter	7.2×10^{-6}	2/22	41/314	0/6
GO:0007165 – signal transduction	1.8×10^{-4}	50/285	60/633	50/506
GO:0007166 – cell surface receptor signaling pathway	2.9×10^{-4}	3/16	17/132	11/58

Cellular Component GO ID – Term	**P Sim.**	**Ratio (Hits / Total Genes in GO Term)**		
		Honey Bee	Mouse	Stickleback
GO:0005576 – extracellular region	$< 1 \times 10^{-7}$	27/113	85/481	31/176
GO:0005578 – proteinaceous extracellular matrix	$< 1 \times 10^{-7}$	2/20	40/193	3/14
GO:0005615 – extracellular space	$< 1 \times 10^{-7}$	6/17	107/712	5/18
GO:0005886 – plasma membrane	$< 1 \times 10^{-7}$	8/58	203/2146	23/109
GO:0005887 – integral component of plasma membrane	$< 1 \times 10^{-7}$	3/7	83/562	1/9
GO:0016021 – integral component of membrane	$< 1 \times 10^{-7}$	119/785	245/3266	153/1212
GO:0016459 – myosin complex	$< 1 \times 10^{-7}$	8/23	2/39	21/53
GO:0031012 – extracellular matrix	2.0×10^{-7}	2/11	28/158	6/40
GO:0016020 – membrane	4.8×10^{-6}	123/813	147/2521	174/1229
GO:0045202 – synapse	6.7×10^{-5}	1/25	31/236	1/6

Molecular Function GO ID – Term	**P Sim.**	**Ratio (Hits / Total Genes in GO Term)**		
		Honey Bee	Mouse	Stickleback
GO:0003774 – motor activity	$< 1 \times 10^{-7}$	8/23	2/51	21/53
GO:0004930 – G-protein coupled receptor activity	$< 1 \times 10^{-7}$	26/124	37/247	51/360
GO:0005179 – hormone activity	$< 1 \times 10^{-7}$	3/8	12/41	10/40
GO:0005509 – calcium ion binding	$< 1 \times 10^{-7}$	29/169	57/488	90/535
GO:0043565 – sequence-specific DNA binding	2.0×10^{-7}	32/133	44/385	48/365
GO:0005515 – protein binding	1.2×10^{-6}	279/1635	534/8766	417/3710
GO:0003707 – steroid hormone receptor activity	1.4×10^{-6}	10/18	9/45	5/62
GO:0005216 – ion channel activity	2.4×10^{-6}	5/65	11/109	30/119
GO:0005198 – structural molecule activity	4.8×10^{-6}	2/24	12/86	20/79

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GO:0005201 – extracellular matrix structural constituent	5.6 x 10^{-6}	4/7	6/26	7/25
GO:0020037 – heme binding	3.3 x 10^{-5}	10/63	15/77	11/72
GO:0004714 – transmembrane receptor protein tyrosine kinase activity	3.4 x 10^{-5}	1/6	8/36	8/28
GO:0005215 – transporter activity	4.6 x 10^{-5}	13/95	19/110	16/122
GO:0005506 – iron ion binding	2.6 x 10^{-4}	8/60	16/95	11/82
GO:0003700 – transcription factor activity, sequence-specific DNA binding	2.7 x 10^{-4}	40/168	51/637	38/346

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Conserved transcription factor expressed in the brain implicated as regulators of response to social challenge

	Ratio (Hits / Total Targets)				
Orthogroup	**Honey Bee**	**Mouse**	**Stickleback**	**p**	**Mouse Gene Names**
EOG8JT1HM	1/3	12/84	30/128	0	*Sp3, Klf16, Klf13, Klf10, Sp4, Klf4, Klf9, Sp1, Sp7, Klf5, Klf12, Klf7*
EOG873R3N	0/1	8/42	11/34	0.000004	*Barx2, Nkx2-1, Hmx2, Hhex*
EOG8KWM99	9/25	2/2	3/45	0.0003072	*Pbx3, Pbx1*
EOG86DNH2	5/9	0/1	1/1	0.0012816	*Nr2e1*
EOG8JWWWP	4/5	1/6	0/6	0.0055148	*Gsx1*
EOG81RRB5	5/105	1/2	9/42	0.0171374	*Arx, Pax6, Rax*

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

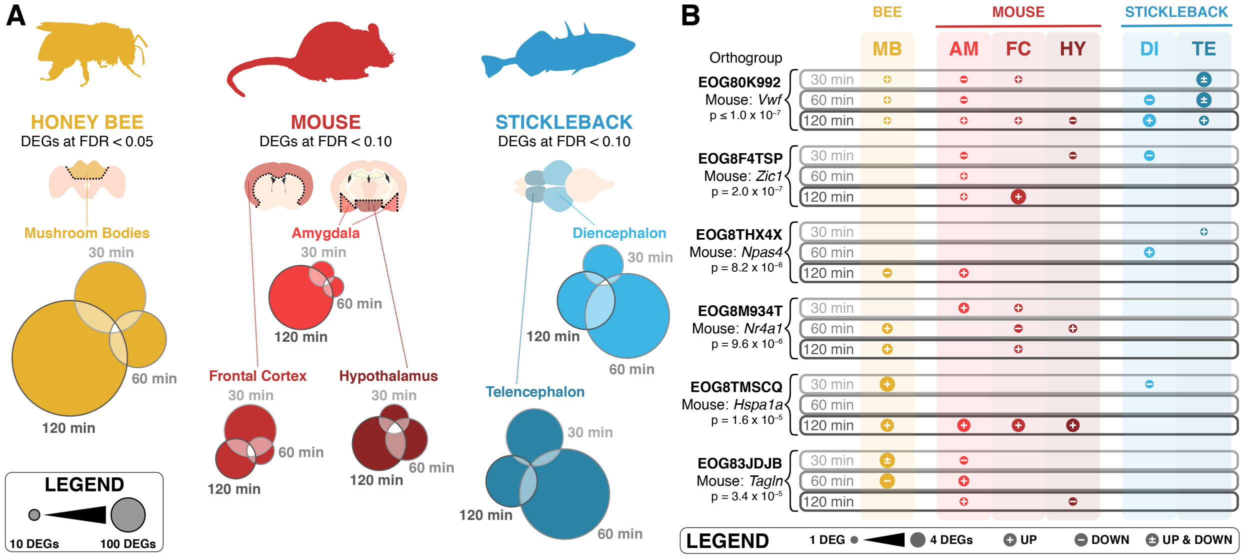

Figure 1: Differential gene expression in response to social challenge across species. A) Description of differentially expressed genes (DEGs) within each brain region assayed for each species. Honey bee DEGs are called at FDR < 0.05 while mouse and stickleback DEGs are called at FDR < 0.10. B) Orthogroups with significant conservation of differential expression across all three species.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

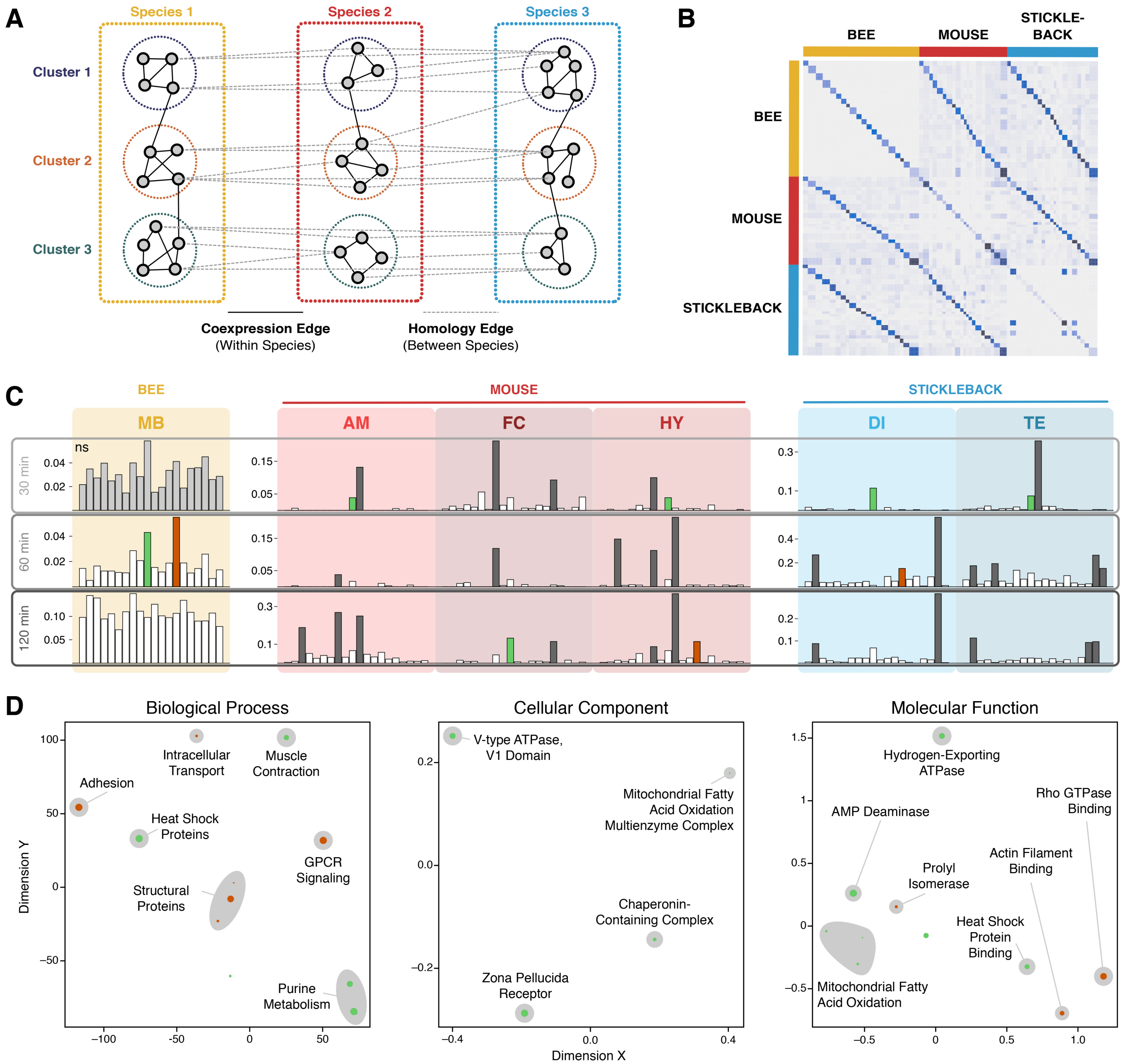

Figure 2: Cross-species coexpression module algorithm conceptual schematic and results. A) Schematic of CNSRV, the cross-species clustering algorithm used to find conserved gene modules, which uses evidence derived from both coexpression and conservation to find gene modules enriched in conservation. B) Clustering results from CNSRV show that conserved modules, shown by the ancillary diagonals off the main diagonal, cluster better between species than do unmatched modules. C) Enrichment results for DEGs for CNSRV modules within each species reveal significant differences among clusters in all but honey bee mushroom body at 30 min (light gray). Multiple CNSRV clusters were enriched in individual species (dark gray), but two modules – 10 and 14, shown in green and dark orange respectively – show enrichment for differential expression across all 3 species. D) Multidimensional scaling on semantic distances

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for GO terms enriched in the cross-species DRAWR results show clusters of GO terms commonly related to clusters 10 and 14 across all 3 species. Larger points associated with each GO term correspond to stronger p-values. Gray clouds correspond to a high-level biological description of the GO terms within each cluster.

bioRxiv preprint doi: https://doi.org/10.1101/219444; this version posted November 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

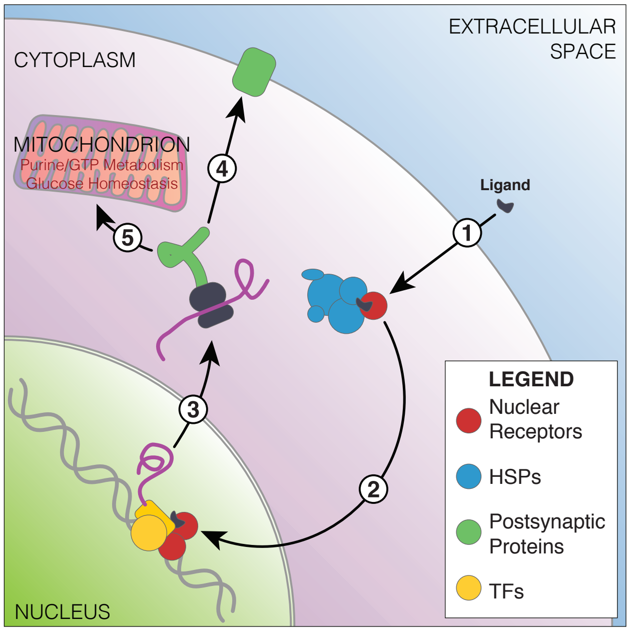

Figure 3: Schematic representation of genes and gene sets found enriched in the brain's response to social challenge across honey bees, stickleback fish and mice. Hypothesized pathway includes 1) nuclear receptor signaling interacting with heat shock/chaperones. 2) These nuclear receptors translocate across the membrane, interacting with well known neurally active transcription factors to cause 3) alterations in transcription. These induce changes in 4) postsynaptic proteins and 5) mitochondrial function.