Bioinformatics, 32, 2016, i620–i628
doi: 10.1093/bioinformatics/btw428
ECCB 2016

OXFORD

XGSA: A statistical method for cross-species gene set analysis

Djordje Djordjevic[1,2,*], **Kenro Kusumi**[3] **and Joshua W. K. Ho**[1,2,*]

[1]Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia, [2]St Vincent's Clinical School, University of New South Wales Australia, Darlinghurst, NSW 2010, Australia and [3]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Gene set analysis is a powerful tool for determining whether an experimentally derived set of genes is statistically significantly enriched for genes in other pre-defined gene sets, such as known pathways, gene ontology terms, or other experimentally derived gene sets. Current gene set analysis methods do not facilitate comparing gene sets across different organisms as they do not explicitly deal with homology mapping between species. There lacks a systematic investigation about the effect of complex gene homology on cross-species gene set analysis.

Results: In this study, we show that not accounting for the complex homology structure when comparing gene sets in two species can lead to false positive discoveries, especially when comparing gene sets that have complex gene homology relationships. To overcome this bias, we propose a straightforward statistical approach, called XGSA, that explicitly takes the cross-species homology mapping into consideration when doing gene set analysis. Simulation experiments confirm that XGSA can avoid false positive discoveries, while maintaining good statistical power compared to other *ad hoc* approaches for cross-species gene set analysis. We further demonstrate the effectiveness of XGSA with two real-life case studies that aim to discover conserved or species-specific molecular pathways involved in social challenge and vertebrate appendage regeneration.

Availability and Implementation: The R source code for XGSA is available under a GNU General Public License at http://github.com/VCCRI/XGSA

Contact: jho@victorchang.edu.au

1 Introduction

In the past decade, biological discovery has been transformed by the ability to profile the entire transcriptome of an organism using microarray and RNA sequencing technologies, giving us a global view into gene regulation. Often gene set analysis (GSA) is the first step in an exploratory analysis of a genome-wide expression dataset. Commonly, a set of differentially expressed (DE) genes are first identified using a statistical test, then this set of genes is compared against a database of gene sets, such as those derived from gene ontology (GO) or known molecular or signaling pathways. Many statistical methods have been adopted or developed to perform GSA (Huang *et al.*, 2009; Rivals *et al.*, 2007), including the Fisher's exact test and its variants.

One central assumption in GSA is that the gene sets being compared are subsets of the same set of genes, in practice meaning from the same species. With an increasing variety of genetic resources available in evolutionarily diverse model and non-model organisms, there is an increasing interest in utilizing these cross-species gene set resources in GSA. This is an important problem in the emerging field of comparative transcriptomics, which aim to integrate knowledge on regulation of biological pathways across the tree of life (Roux *et al.*, 2015). For example, consider the regeneration of organs and appendages, an ability present in several diverse vertebrate organisms but apparently missing from humans.

As the consistent gene set universe assumption fails when more than one species is involved, it becomes increasingly problematic as the number of many-to-many homologues increase between evolutionarily distant species. Several largely *ad hoc* methods have been proposed and used in the literature and have become the standard analysis options. The naïve cross-species mapping approach is to apply an 'at least one homologue' function to map a gene set from one species to another. This approach is the most common method for homology mapping in general and is used by the majority of existing cross-species gene set analysis web based platforms, including g:Profiler, Gene Weaver and GSGator (Baker *et al.*, 2012; Kang *et al.*, 2014; Reimand *et al.*, 2007).

When performing comparative analyses between evolutionarily distant species, many researchers remove the increased complexity from homology assignment by applying the BLAST best reciprocal hits (BRH) approach (Britto *et al.*, 2012; Gohin *et al.*, 2010; Labbé *et al.*, 2012). BRH reduces complexity by restricting homology assignments to at most one per gene, choosing the best hits for each gene (highest sequence similarity or lowest E value) and only assigning homology if the two genes are each other best hits. This implies the assumption that the best hit is the only valuable hit, which is particularly problematic when there are multiple closely scored hits in one gene family. For distantly related organisms, a large amount of non one-to-one homology information is discarded before any analysis is done, reducing the potential insight that can be gained.

Another alternative approach to reduce complexity is to perform significance testing at the level of gene family, also called an orthologous group (OG) (Kristiansson *et al.*, 2013; Rittschof *et al.*, 2014; Zheng *et al.*, 2011). In this approach, the entire OG is assigned a representative value summarizing the constituent genes (often the normalized minimum *P*-value), discarding homology information after this assignment. Traditional statistical enrichment tests are then applied at the level of the OG. The OG structure between a large selection of species can be retrieved from databases such as eggNOG, OrthoDB and InParanoid (Kriventseva *et al.*, 2015; Powell *et al.*, 2014; Sonnhammer and Ostlund, 2015). A strength of the OG framework is the ability to test gene sets from more than two species simultaneously. Nonetheless, similar to the BRH approach, one major limitation of this approach is the loss of information regarding the signal from individual genes in the same OG, and that the exact gene responsible for the final result can be unknown, making interpretation and validation challenging.

Another approach is to computationally transfer the functional annotations (based on protein domains for example) to a less studied organism from well studied ones, as facilitated by PANTHER (Mi *et al.*, 2016). This annotation transfer reduces confidence in annotation quality and relies on the assumption that the relationship between the protein domain and functional annotation is known and true, which limits its utility to molecular function annotation as opposed to more general biological pathways. Several other studies harness the strength of sequence information in microarray probes to transfer information between species (Le *et al.*, 2010; Lu *et al.*, 2009; Xie *et al.*, 2011). While these and other approaches lend more confidence and resolution than simple ID mapping, they do not create a general and principled cross-species gene set analysis framework that specifically addresses complex homology (Lu *et al.*, 2010; Yang *et al.*, 2014).

To our knowledge, there has not been any systematic investigation on the issues of cross-species GSA. An approach that utilises the full and complex homology structure between two species is not available. In this study, we discuss the statistical issues associated with cross-species gene set analyses and define an informative homology complexity score. We show that the naïve implementation of homology mapping followed by Fisher's exact test can lead to false positive discovery. To alleviate this bias, we propose a straightforward statistical test, called XGSA, to perform cross-species GSA by considering the complete homology structure between two species. Our simulations show that XGSA can indeed remove the false positive bias, while maintaining good statistical power when analyzing gene sets with complex homology structure. We apply XGSA to two real biological applications that involve comparing gene sets from distantly related organisms.

2 Methods

2.1 Problem definition

Let $A = \{a_1, a_2, \ldots, a_l\}$ and $B = \{b_1, b_2, \ldots, b_k\}$ denote the set of all genes (the gene universe) in two species A and B, respectively. We further define subsets A' and B' as the *gene set of interest* in species A and B respectively, where $A' \subseteq A$ and $B' \subseteq B$.

Let there be a *homology mapping function*, $m(a, b)$, that describes the sequence homology relationship between any gene a in species A and any gene b in species B:

$$m(a,b) = \begin{cases} 1, & \text{if } a \text{ and } b \text{ are homologous} \\ 0, & \text{otherwise.} \end{cases}$$

Given gene sets of interest A' and B', we can further define their homologous partners in the other species as $B'' = \{b \in B : m(a,b) = 1, \exists a \in A'\}$ and $A'' = \{a \in A : m(a,b) = 1, \exists b \in B'\}$

The cross-species gene set analysis problem can be defined as a hypothesis test where the null hypothesis H_μ is that the membership of A' and A'' are independent and B' and B'' are independent (Fig. 1).

2.2 XGSA

We calculate the probability p_A of co-membership of A' and A'' equal to or greater than the observed co-membership if H_μ is true, using the hypergeometric distribution,

$$p_A = \sum_{k=|A'\cap A''|}^{\min(|A'|,|A''|)} \frac{\binom{|A'|}{k}\binom{|A_u|-|A'|}{|A''|-k}}{\binom{|A_u|}{|A''|}}$$

where A_u is the gene universe in A that has homology to the gene universe in species B, $A_u = \{a \in A : m(a,b) = 1, \exists b \in B\}$. This is equivalent to an upper tail Fisher's exact test. Similarly, we compute the probability p_B for observing the co-membership of B' and B'' if H_μ is true,

$$p_B = \sum_{k=|B'\cap B''|}^{\min(|B'|,|B''|)} \frac{\binom{|B'|}{k}\binom{|B_u|-|B'|}{|B''|-k}}{\binom{|B_u|}{|B''|}}.$$

We calculate a statistic p to estimate the probability of H_μ being true, as the maximum of p_A and p_B,

$$p = \max(p_A, p_B).$$

We take the maximum in order to reduce the false positive rate caused by complex homology, as illustrated in Figure 1.

2.3 Naïve approach

We compared the performance of XGSA with other *ad hoc* approaches for cross-species GSA. The naïve approach is equivalent to doing the above test in only one of the species, e.g. species A. In this case, the P-value is the same as p_A.

2.4 Best reciprocal hits

The best reciprocal hits approach only differs from the naïve approach in that it reduces the complexity first. We created a subset of

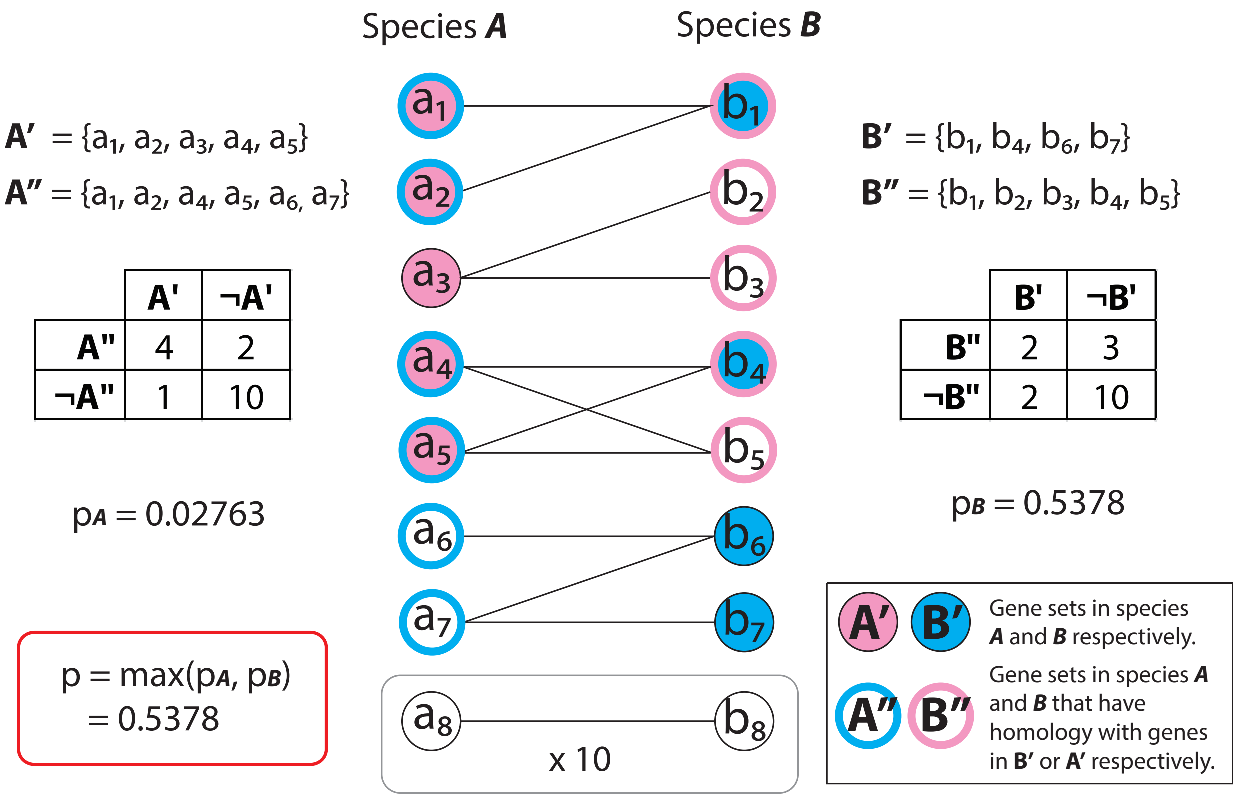
Species A
Species B
A′ = {a1, a2, a3, a4, a5}
A″ = {a1, a2, a4, a5, a6, a7}
A' | ¬A'
A" | 4 | 2
¬A" | 1 | 10
pA = 0.02763
p = max(pA, pB)
= 0.5378
a1
a2
a3
a4
a5
a6
a7
a8
b1
b2
b3
b4
b5
b6
b7
b8
x 10
B′ = {b1, b4, b6, b7}
B″ = {b1, b2, b3, b4, b5}
B' | ¬B'
B" | 2 | 3
¬B" | 2 | 10
pB = 0.5378
A′
B′
Gene sets in species A and B respectively.
A″
B″
Gene sets in species A and B that have homology with genes in B′ or A′ respectively.

Fig. 1. A schematic diagram illustrating the XGSA method. Nodes represent genes in species *A* and *B*, with edges representing homology, and shading and outlines representing gene set membership. The gray box represents the remainder of the homologous universe of one-to-one relationships not assigned to either gene set. The tables are contingency tables describing the observed overlap of the homologous gene sets in the two species. p_A and p_B show the different *P*-values derived from performing Fisher's exact test in each species. The red box indicates the final value *p* produced by XGSA

homology mappings for which the retained human and zebrafish homology mappings were each other highest scoring partners, based on sequence similarity percentages from Ensembl.

2.5 Orthologous group

We downloaded OG annotations from OrthoDB with no filtering applied. We mapped genes to OGs and calculated p_A at the OG level.

2.6 Automatically identifying homology between species using Ensembl BioMart

Following standard practice (Reimand *et al.*, 2007; Yates *et al.*, 2016), we accessed Ensembl BioMart programmatically through the R package *biomaRt* (Durinck *et al.*, 2009) and retrieved homology mapping between Ensembl gene ids in two species. We turned this mapping into a sparse matrix using the R package *Matrix*.

2.7 Homology complexity score

We define a measure of complexity for a gene set in one species with respect to its homology mapping to another species (*A* and *B*), as the fraction of genes in the gene set GS_A in species *A* which have more than one homologue in species *B*,

$$\text{Complexity}(GS_A, B) = \frac{|a \in GS_A : \sum_{b \in B} m(a, b) > 1|}{|GS_A|}.$$

2.8 Statistical power analysis

For each human gene ontology (GO) term, we find all of the zebrafish homologues for that GO term. Intuitively, when gene sets devoid of any homologous genes are tested in a cross-species gene set enrichment test, the *P*-value for that test should be 1 (no match). Alternatively, when the entire set of homologues is tested, the *P*-value should be close to zero (perfect match). Based on this logic, if we incrementally add homologous genes to the gene set enrichment test, the *P*-value should decrease. We can then interpret the rate at which several methods reach significance as an indicator of their relative power for that cross-species gene set enrichment test.

We start with a zebrafish gene set consisting of the same number of non-homologous genes as there are zebrafish homologues to the chosen human GO gene set. We incrementally substitute each non-homologous gene in the zebrafish set with a homologous gene, and perform enrichment testing after each substitution.

2.9 Data preprocessing for the vertebrate regeneration case study

We downloaded four spinal cord regeneration datasets from three species, zebrafish (*Danio rerio*), lizard (*Anolis carolinensis*) and Western clawed frog (*Xenopus tropicalis*). We reprocessed the zebrafish and frog results from the raw microarray data because the lists of DE genes were not available in the original papers. All processing was done in R using the *limma* package and custom scripts unless otherwise noted. Benjamini–Hochberg multiple hypothesis testing correction was applied in each case.

2.9.1 Zebrafish 1

Raw agilent microarray data were downloaded from GEO (accession GSE39295), corrected for background effects (offset = 16), log-transformed and quantile normalized (Hui *et al.*, 2014). Probes with an average expression less than 8 were removed as 'not present' probes after visual inspection of probe intensity distribution. Differential expression at each post-injury timepoint compared to time zero control was computed to match the published study design. We applied an absolute T-statistic threshold of 7 resulting in 404 significantly differentially expressed genes with Ensembl gene IDs across the five timepoints.

2.9.2 Zebrafish 2

Preprocessing of raw data (accession GSE20460) as above (Guo *et al.*, 2011). Differential expression at 4 and 12 h post-injury compared to matched sham timepoints was computed to match the published study design, although we omitted the 264 h timepoint due to poor data quality. We applied an absolute T-statistic threshold of 4 resulting in 62 significantly differentially expressed genes across the two timepoints.

2.9.3 Frog

Raw Affymetrix CEL files were downloaded from Array Express (accession E-MEXP-2420) corrected for background effects, log-transformed and quantile normalized using the RMA method (Love *et al.*, 2011). Probes with an average expression less than 6 were removed as 'not present' probes after visual inspection of probe intensity distribution. Differential expression followed a time series design with 6 h post-amputation (PA) versus 0 h control, 24 h PA versus 6 h PA and 60 h PA versus 24 h PA, to match the published study design. We applied an absolute T-statistic threshold of 4 resulting in 666 significantly differentially expressed genes across the three timepoints.

2.9.4 Lizard

We retrieved the differentially expressed gene lists from the supplementary files of the published study (Hutchins *et al.*, 2014).

3 Results

3.1 Human and zebrafish gene sets exhibit a broad range of complex homology

We chose two model organisms with well annotated genomes, *Homo sapiens* (human) and *Danio rerio* (zebrafish), retrieving homology mappings between 15 908 human Ensembl gene IDs and 18 777 zebrafish Ensembl gene IDs. Henceforth the term 'genes' refers to Ensembl gene IDs. 4179 human genes map to more than one zebrafish gene, and 2218 zebrafish genes map to more than one human gene, corresponding to 26.3% and 11.8% of the respective homologous genomes.

We constructed a BRH subset of homology mappings (see Section 2), henceforth referred to as the BRH set. The BRH set has 1807 fewer human genes and 4493 fewer zebrafish genes than the complete set, corresponding to a reduction of 11.4% and 23.9% of the respective homologous genomes.

We calculated human–zebrafish complexity scores (see Section 2) for each gene set in the gene ontology (GO). We observe a wide range of complexity occurs in GO.

3.2 Naïve cross-species GSA approach results in a systematic bias

When a random selection of 1000 human genes is tested against the human GO using the Fisher's exact test as implemented in TopGO (Alexa and Rahnenfuhrer, 2010), there are no significant results passing the significance thresholds after multiple testing correction, and a relatively uniform distribution of *P*-values is observed as expected (Fig. 2A, red bars). The same is true when 1000 zebrafish genes are tested against the zebrafish GO, and these results were consistent for 100 different random selections of genes.

In contrast, when all the human homologues of 1000 randomly selected zebrafish genes is tested against the human GO, a very strong enrichment of small *P*-values is observed (Fig. 2A, blue bars). An enrichment of small *P*-values passing multiple hypothesis testing thresholds can be interpreted as evidence for a strong signal in the data. Considering that the original selection of genes was entirely random, this indicates that these significant *P*-values are false positives.

Repeating this virtual assay 100 times reveals that some GO terms appear repeatedly in the list of enriched gene sets, with the most recurrent gene set 'flavonoid glucoronidation' enriched in 38% of trials (Fig. 2B). This shows a systematic bias leading to false positive results when using Fisher's exact test, introduced by naïve homology mapping from zebrafish to human genes. Importantly this bias is species-dependent; as different pairs of species show different biased GO terms. For example, mapping 1000 random genes from *Xenopus tropicalis* to *Mus musculus* and performing GSA results in a different set of biased GO terms, including' sensory perception of chemical stimulus' in 65% of virtual assays. This indicates that the bias may result from the complex homology mapping between two species. When we look at the genetic homology between genes annotated with the GO term 'flavonoid glucoronidation' we see several striking examples of complex homology (Fig. 2C). Comparing the complexity scores of the repeated bias GO terms versus all other GO terms shows that the bias GO terms have a significantly higher gene set complexity on average (two-sided *t*-test, *P*-value = 1.156e−07) (Fig. 2D). When we use the same sets of randomly selected zebrafish genes but map them to human genes using the BRH homology mapping, the bias disappears (data not shown). Taken together, these findings provide evidence that the cause of the bias is the introduction of complex homology mapping into the testing framework without compensation.

3.3 XGSA alleviates the bias in the naïve method

Using a toy example of the cross-species testing problem, we observe that the directionality of the complex homology mapping creates the bias (Fig. 1). Our solution involves performing testing in both species / directions, and combining the results. This means that both species act as the host for a Fisher's exact test, with the test set being naïvely mapped from the gene set in the other species. We then return the maximum *P*-value of the pair of tests. We call this approach XGSA (see Section 2). Intuitively, this means that the gene set overlap must be significant in both species—that is, in both directions of testing (Fig. 1). In this way, we reduce the effect of complex homology on the resulting *P*-value. When we applied our method to the same 100 repetitions of 1000 randomly selected zebrafish genes we saw that the systematic bias disappears—zero out of 100 repetitions had any significantly enriched human GO terms. By accounting for the effects of complex homology in our statistical testing framework, we can remove the bias while still utilizing the full complex homology structure.

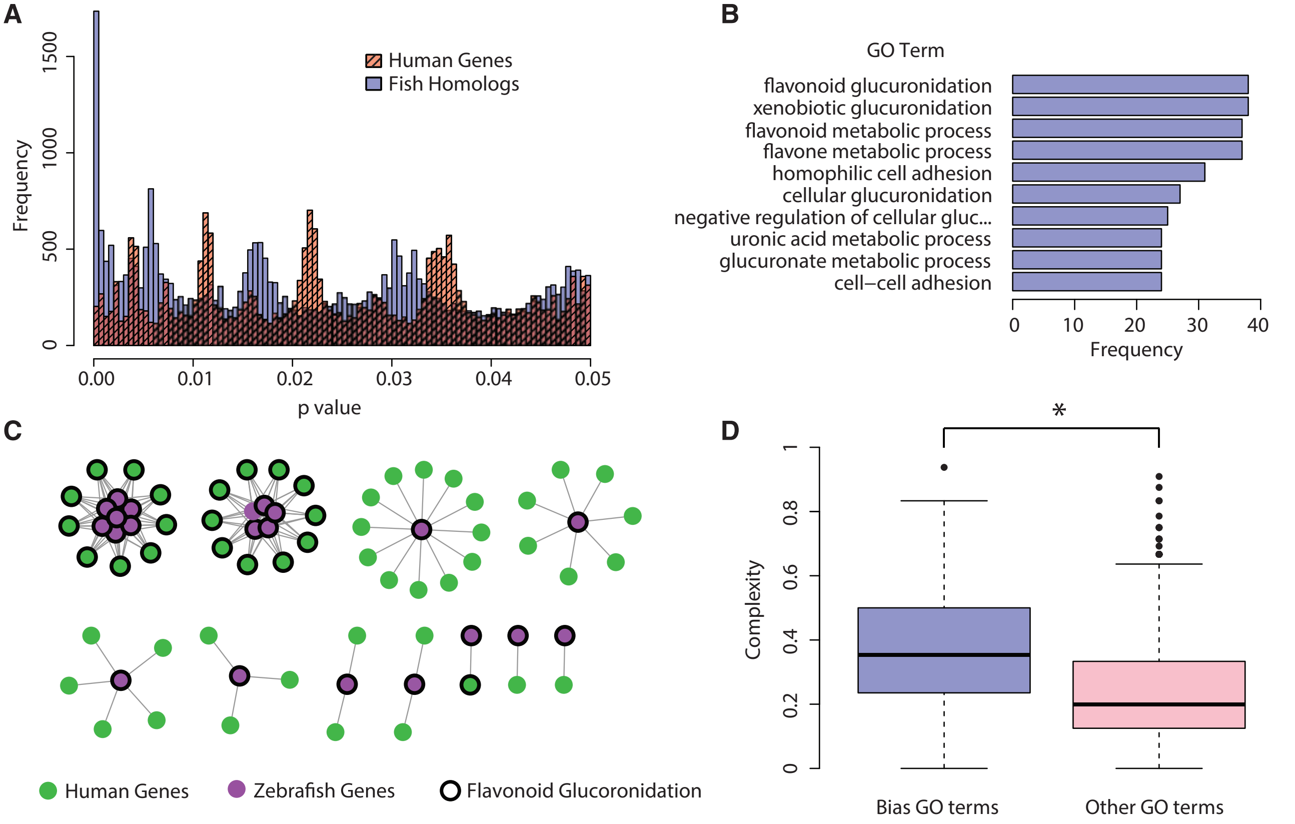

Fig. 2. Identification of bias in naïve cross-species gene set analysis. (A) *P*-values from human GO analysis of 1000 randomly selected human genes and 1000 randomly selected zebrafish genes naïvely mapped to their human homologs. (B) Frequency of the 10 most common false-positive GO terms from 100 repeats of the experiment. (C) Human and zebrafish homology relationships for genes assigned to the GO term 'flavonoid glucoronidation'. (D) Complexity scores of false positive GO terms versus all other GO terms

3.4 Simulation studies show XGSA maintains good statistical power even when analyzing gene sets with complex homology

As this problem has not been studied in depth before and no gold standard exists against which we can evaluate our method, we devised a novel testing approach to compare the power of different methods (see Section 2). Briefly, after choosing a human GO gene set, we incrementally replace zebrafish genes that are not homologous to the GO set with genes that are homologous and calculate significance using different cross-species gene set enrichment testing approaches. Based on the assumption that zero homologous genes should return a *P*-value of 1 and all homologous genes a *P*-value close to zero, we can compare the rate at which the *P*-value decreases as genes are replaced (Fig. 3).

We found that for low complexity gene sets, the four methods of naïve mapping, BRH, OG and XGSA perform comparably with no practical difference at commonly used thresholds (Fig. 3A). However, when testing higher complexity gene sets the power of XGSA becomes clearer (Fig. 3B). By retaining the full homology structure XGSA continues to gain power from genes assigned to complex gene families, as opposed to BRH and OG in which the power curve plateaus when complex genes are added. The over-sensitivity of the naïve method to high complexity gene sets can be observed as abrupt rises in the curve above the diagonal. In contrast, XGSA maintains a near linear diagonal power curve as with low complexity gene sets.

We can summarize these curves by measuring the relative area under them (Fig. 4). We find that for zero complexity gene sets all methods perform similarly with small differences due to various gene universe sizes—XGSA receives a lower score because it uses the most extensive gene universe. ODB then drops in detection power as complexity increases in the tested gene sets and plateaus are introduced to the power curve. As gene set complexity increases the advantage of XGSA over both ODB and BRH becomes clear. While it seems that XGSA may be too sensitive as complexity increases, this is because the zero *p*-value saturates earlier in large and complex gene sets, causing the power curve to change shape and the AUC to increase.

3.5 Case study 1: discovering conserved pathways in social challenge in evolutionarily distant organisms

Rittschof *et al.* (2014) studied the transcriptomic changes associated with social challenge in three species: stickleback fish (*Gasterosteus aculeatus*), mouse (*Mus musculus*) and honey bee (*Apis mellifera*). They performed a ranked GSA (Sartor *et al.*, 2009) on their DE genes for each species by assigning GO membership based on protein domain (sequence) information using PANTHER. They also performed a cross-species analysis using the homologous triplet OG approach by harnessing the OrthoDB database, and used the mouse GO as the reference gene sets. We downloaded their lists of DE genes from each species and sought to recreate their analysis using XGSA. Because the honey bee *Apis mellifera* is not yet included in the Ensembl BioMart homology database, we mapped the 182 honey bee genes to 153 fly (*Drosophila melanogaster*) genes using OrthoDB as suggested by FlyBase (Attrill *et al.*, 2016), and used the fly genes to continue the analysis.

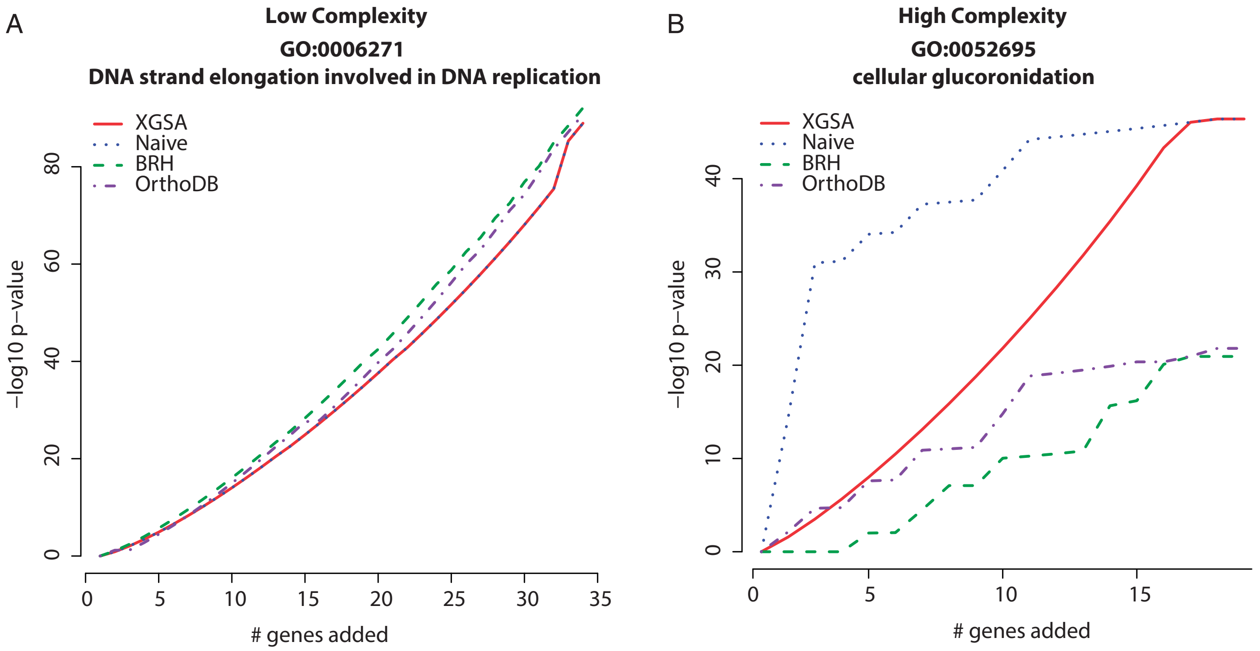

Fig. 3. Simulations show an increased power of XGSA for high complexity gene sets. (A) Low complexity gene sets show comparable performance, with all methods demonstrating a smooth near linear power curve as expected. (B) With high complexity gene sets the difference in the methods becomes obvious. OrthoDB represents the orthologous group implementation using the OrthoDB database

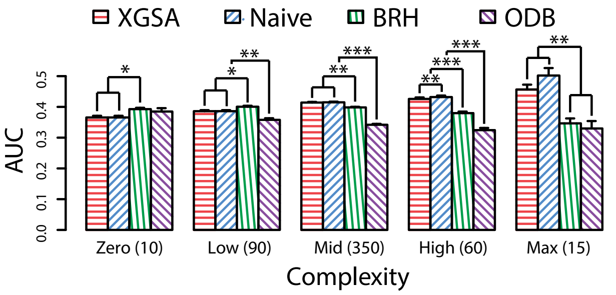

Fig. 4. Area under the curve of different methods performance during the GO simulation study. Error bars indicate the standard deviation around the mean. In parentheses on the X-axis labels is indicated the number of GO terms in each complexity bracket. The apparent improvement sensitivity of the naïve method for maximum complexity gene sets is in fact the detection of false positives as shown in Figure 2. $*P < 0.01$, $**P < 1 \times 10^{-4}$ and $***$ $P < 1 \times 10^{-8}$ by two sided *t*-test

We visualize our results as a molecular concept map (MCM) (Rhodes *et al.*, 2007)—a network diagram where each node represents a gene set and each edge represents a significant overlap between gene sets (Fig. 5). Unlike Rittschof *et al.*, we are directly comparing the experimental gene sets from all three species against the standard mouse GO terms which allows us to interpret them in a single MCM. This approach is different from the approach used by Rittschof *et al.* where they used computationally inferred GO membership for each species.

As with the original study we found very little in the way of shared significant gene sets between two or more species when comparing the tests performed for each individual species. However, gene set similarity clustering shows that some gene set categories span multiple species, including ion transport and regulation of neuronal and muscle activity, particularly between fly and mouse. We also see a mouse specific viral and ribosomal response, as well as a fish specific phototransduction response. Furthermore we included KEGG pathways into the MCM analysis, allowing us to identify interesting and relevant pathways for social challenge such as Long Term Depression and Long Term Potentiation, and the only gene set significant in two species (mouse and fly), Dilated Cardiomyopathy.

We found that 39%, 33% and 12% of our significant GO gene sets overlapped with Rittschof *et al.* results in mouse, stickleback and honey bee, respectively. Furthermore, 21% of our total significant GO gene sets were significant in the Rittschof *et al.* homologous triplet OG analysis, including representative gene sets spanning their major result categories.

Our comparison with the study by Rittschof *et al.* raises several issues. As a limitation, our analysis used the DE gene sets as opposed to the ranked list used with GSEA in the original study. As we also do not know the GO terms universe or term—gene assignment used in that study, we cannot declare how closely our results matched theirs. That our results, although retrieving fewer and different GO terms, spanned their species-specific and homologous triplet categories indicate that we recreated many of their key findings. Mapping from honey bee to fly was clearly not ideal and so it is not surprising the fairly low correspondence of significant gene sets for that species. An alternative is to create homology mappings from honey bee to stickleback fish and mouse using BLAST, to enable direct comparisons.

3.6 Case study 2: XGSA reveals conserved molecular pathways in vertebrate organ regeneration

Many vertebrates display the ability to regenerate entire appendages, but humans or other common mammalian animal models have very limited capacity to regenerate. With the availability of whole genome sequences and functional genetic technologies for reptilian and amphibian species with significant regenerative capacity, genome-wide comparative studies of gene expression dynamics during organ regeneration are now possible. Lizards, which are amniote vertebrates like humans, are able to lose and regenerate a functional tail with regrowth and patterning of cartilage, muscle, vasculature, spinal cord and skin (Hutchins *et al.*, 2014). In addition to the

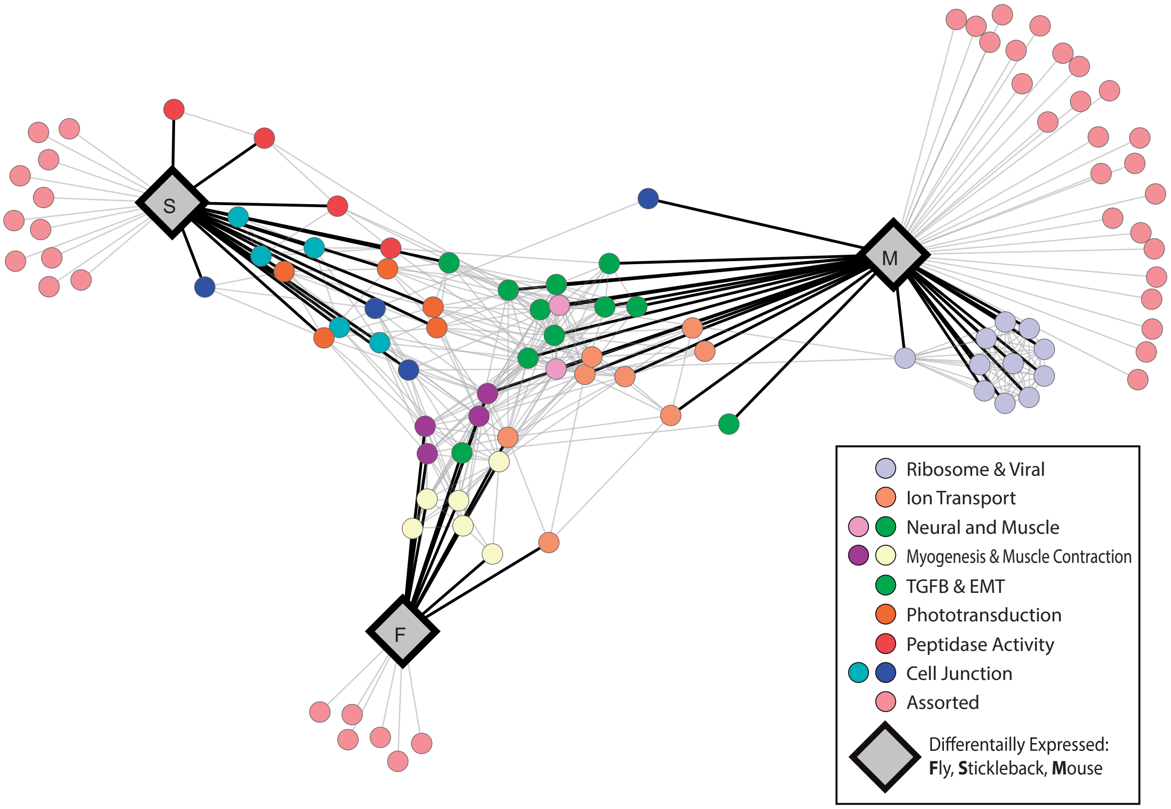

Fig. 5. Cross-species gene set analysis of transcriptional response to social challenge. (A) Molecular concept map showing the results of the XGSA pipeline, nodes represent gene sets and edges represent a significant overlap between gene sets. B) Differences in complexity between gene sets recovered by XGSA versus Rittschof *et al.* (2014)

lizard, tadpoles of the African clawed frog, *Xenopus laevis*, are also capable of regenerating their tails and fins, and there are extensive genomic resources available for this model. One important task in regenerative biology is to identify molecular pathways that are conserved in multiple regenerative vertebrates during organ regeneration.

Here, we performed a case study to investigate spinal cord regeneration across three species for which transcriptomic profiling of regenerating spinal cord tissues was available; zebrafish (Guo *et al.*, 2011; Hui *et al.*, 2014), lizard (Hutchins *et al.*, 2014) and frog (Love *et al.*, 2011). We sought to explore the biology captured in these datasets by leveraging the extensive gene sets available for human and zebrafish in GO, MSigDB and SPEED (Subramanian *et al.*, 2005). In total, our analysis included 97 079 XGSA tests between 2804 gene sets which took 30 min on a single core and resulted in 175 significant overlaps. We analyzed the results using an MCM (Fig. 6).

Zebrafish and tadpole regeneration gene sets show a direct overlap between them as well as many shared enriched gene sets, including *TNFa* and *E2F* signaling, cell cycle, DNA repair and oocyte maturation signals. The lizard gene sets are more isolated from the other two, which is itself not surprising due to their different experimental designs and tissues being profiled. We found that Lizard and tadpole share endothelial to mesenchymal transition and extra cellular matrix assembly related signals. In the base of the lizards regenerating tail, we see a very strong enrichment of muscle-related gene sets, likely due to the dominance of this tissue in this regenerative region.

When we compared our gene sets against the MSigDB perturbation gene sets, we find a gene set related to human carious teeth that overlap significantly (adjusted P-value < 0.05) with all three species. The pulpal tissue of human carious teeth has been reported to be a source of active multipotent mesenchymal stem cells and may represent a tissue with limited regenerative capacity in human (Rajendran *et al.*, 2013). When focusing on the TF targets and immune gene sets, we found the motif for *SRF* (a known regeneration stimulant (Stern *et al.*, 2013)) is enriched in DE genes from both lizard and zebrafish, and that there is a conserved immune response in zebrafish and tadpole.

We further looked at which genes are commonly DE between multiple species. Aurora Kinase A (*AURKA*), which plays a crucial role in spindle assembly, was DE in lizard and zebrafish regeneration experiments, and is known to be required for regeneration in mouse (Pérez de Castro *et al.*, 2013). Furthermore, *AURKB* was DE in the tadpole regeneration experiment, suggesting an evolutionarily conserved role for Aurora kinases in regeneration. Another gene of interest is Keratin 19 (*KRT19*), a marker of hepatic stem cells, endothelial mesenchymal transition and TGFβ signaling. *KRT19* was DE in both lizard and tadpole regeneration experiments, with other keratins being DE in zebrafish. Thirteen more DE genes were conserved between zebrafish and tadpole regeneration, including *PLK1* which is required for cardiac regeneration in zebrafish (Jopling *et al.*, 2010), *KIF23* which controls G2/M arrest and is also DE in axolotl limb regeneration, *SOCS3* which has been shown to suppress optic nerve regeneration in mice (Smith *et al.*, 2009), and several hepatocyte regeneration markers (*KIF20a*, *MCM4* and *LIG1*).

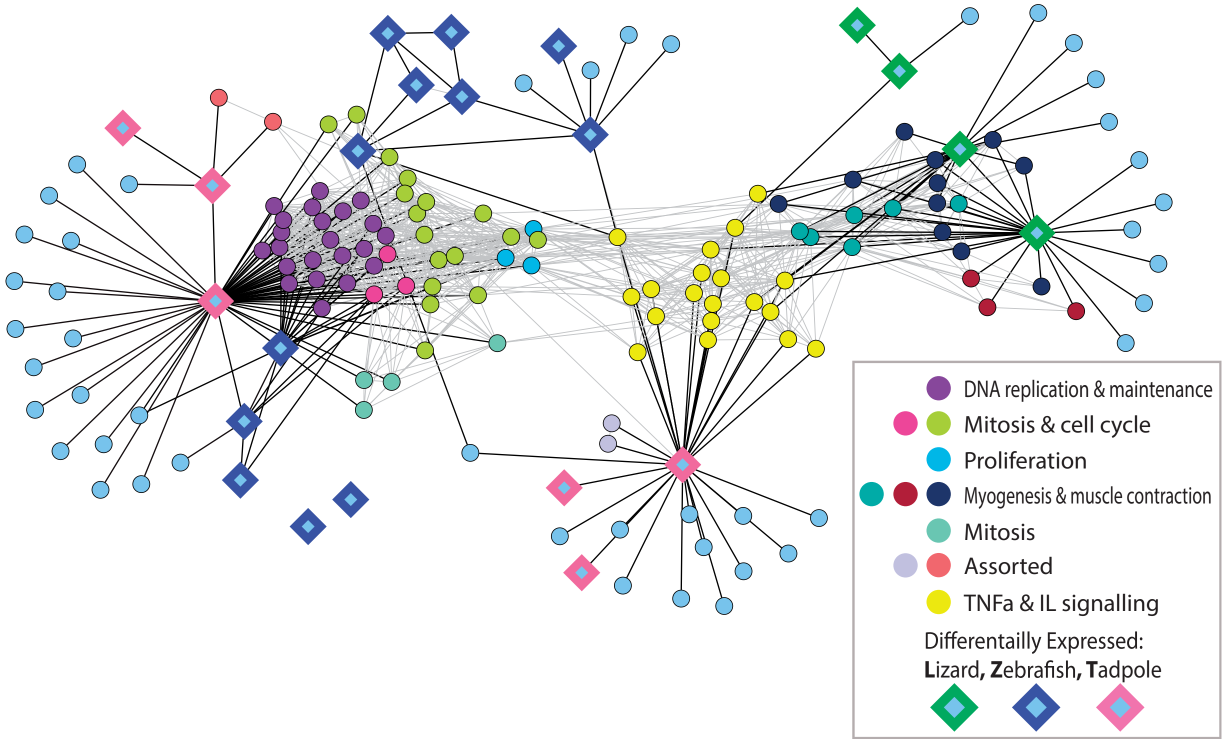
DNA replication & maintenance
Mitosis & cell cycle
Proliferation
Myogenesis & muscle contraction
Mitosis
Assorted
TNFa & IL signalling
Differentailly Expressed:
Lizard, Zebrafish, Tadpole

Fig. 6. Molecular concept map (MCM) showing the overview of the cross-species spinal cord regeneration gene set analysis

4 Discussion

The main contributions of this work are: (1) formulation of the cross-species gene set analysis problem, (2) investigation of the statistical bias that may arise when comparing gene sets with complex homology relationships, (3) development of a statistical hypothesis testing approach called XGSA and (4) demonstration of how XGSA can be used in conjunction with MCM to identify evolutionarily conserved and species-specific molecular pathways using two real datasets.

Effectively, current GSA approaches deal with the complex homology mapping issue by reducing the complexity of the homology mapping (i.e. by removing non one-to-one homologous gene pairs or abstracting the test to a higher level). In contrast, XGSA takes into account the entire homology structure when performing GSA. The benefit of XGSA is increased power to detect enrichment of gene sets with complex homology. XGSA also alleviates the false positive bias introduced by the naïve testing framework by ensuring gene set enrichment is significant in both species, overcoming the main limitation of ‘at least one homologue’ mapping. When compared to existing cross-species GSA approaches, XGSA balances both sensitivity and specificity for all gene sets.

Our work does not deal with the issues of comparing ranked lists, like the GSEA method (Subramanian *et al.*, 2005). Also, our method currently treats each homology relationship equally (absent or present), whereas the information about the extent of sequence homology was not used. Nonetheless, based on the formulation of the problem statement, we aim to extend our method to incorporate these features.

We have implemented the source code for XGSA in R. By harnessing the Ensembl BioMart portal our framework utilizes the latest homology structure on a growing number of species supported in Ensembl (currently 69). Due to the flexibility and simplicity of our R framework such that users can include custom homology matrices for unsupported species, the potential for XGSA to unlock cross-species gene set analyses is widespread. The typical use case is when investigating gene sets from an organism without a comprehensive gene set database. If the organism is supported by Ensembl the XGSA workflow is straightforward, otherwise the user needs to compute homology to a genomic model organism to unleash XGSA. A second use case is when cross-species analysis is central to the biological questions being studied, such as in our case study of spinal cord regeneration. The ability to integrate gene sets from different species together into a unified network-based visualization such as a MCM improves speed and confidence when interpreting insights from traditionally problematic cross-species gene set analyses. This improved workflow is expected to be valuable for researchers in practice (Huang *et al.*, 2009).

Funding

This study was supported in part by funds from the New South Wales Ministry of Health, a National Health and Medical Research Council / National Heart Foundation Career Development Fellowship (1105271), and a Burroughs Wellcome Fund Collaborative Research Travel Grant.

Conflict of Interest: none declared.

References

Alexa,A. and Rahnenfuhrer,J. (2010). *topGO: topGO: Enrichment analysis for Gene Ontology.* R package version 2.18.0. Bioconductor, Seattle, WA.

Attrill,H. *et al.* (2016) FlyBase: establishing a Gene Group resource for *Drosophila melanogaster. Nucleic Acids Res.*, **44**, D786–D792.

Baker,E.J. *et al.* (2012) GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Res.*, **40**, D1067–D1076.

Britto,R. *et al.* (2012) GPSy: a cross-species gene prioritization system for conserved biological processes–application in male gamete development. *Nucleic Acids Res.*, **40**, W458–W465.

Durinck,S. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.

Gohin,M. *et al.* (2010) Comparative transcriptomic analysis of follicle-enclosed oocyte maturational and developmental competence acquisition in two non-mammalian vertebrates. *BMC Genomics*, **11**, 18.

Guo,Y. *et al.* (2011) Transcription factor Sox11b is involved in spinal cord regeneration in adult zebrafish. *Neuroscience*, **172**, 329–341.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Hui,S.P. *et al.* (2014) Genome wide expression profiling during spinal cord regeneration identifies comprehensive cellular responses in zebrafish. *PLoS One*, **9**, e84212.

Hutchins,E.D. *et al.* (2014) Transcriptomic analysis of tail regeneration in the lizard *Anolis carolinensis* reveals activation of conserved vertebrate developmental and repair mechanisms. *PLoS One*, **9**, e105004.

Jopling,C. *et al.* (2010) Zebrafish heart regeneration occurs by cardiomyocyte dedifferentiation and proliferation. *Nature*, **464**, 606–609.

Kang,H. *et al.* (2014) gsGator: an integrated web platform for cross-species gene set analysis. *BMC Bioinformatics*, **15**, 13.

Kristiansson,E. *et al.* (2013) A novel method for cross-species gene expression analysis. *BMC Bioinformatics*, **14**, 70.

Kriventseva,E.V. *et al.* (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, **43**, D250–D256.

Labbé,R.M. *et al.* (2012) A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells (Dayton, Ohio)*, **30**, 1734–1745.

Le,H.S. *et al.* (2010) Cross-species queries of large gene expression databases. *Bioinformatics*, **26**, 2416–2423.

Love,N.R. *et al.* (2011) Genome-wide analysis of gene expression during *Xenopus tropicalis* tadpole tail regeneration. *BMC Dev. Biol.*, **11**, 70.

Lu,Y. *et al.* (2009) Cross species analysis of microarray expression data. *Bioinformatics*, **25**, 1476–1483.

Lu,Y. *et al.* (2010) Cross species expression analysis of innate immune response. *J. Comput. Biol.*, **17**, 253–268.

Mi,H. *et al.* (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.

Pérez de Castro,I. *et al.* (2013) Requirements for Aurora-A in tissue regeneration and tumor development in adult mammals. *Cancer Res.*, **73**, 6804–6815.

Powell,S. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.

Rajendran,R. *et al.* (2013) Regenerative potential of dental pulp mesenchymal stem cells harvested from high caries patient's teeth. *J. Stem Cells*, **8**, 25–41.

Reimand,J. *et al.* (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.

Rhodes,D.R. *et al.* (2007) Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia (New York, N.Y.)*, **9**, 443–454.

Rittschof,C.C. *et al.* (2014) Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proc. Natl. Acad. Sci. USA*, **111**, 17929–17934.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Roux,J. *et al.* (2015) What to compare and how: comparative transcriptomics for Evo-Devo. *J. Exp. Zool. B Mol. Dev. Evol.*, **324**, 372–382.

Sartor,M.A. *et al.* (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics (Oxford, England)*, **25**, 211–217.

Smith,P.D. *et al.* (2009) SOCS3 deletion promotes optic nerve regeneration in vivo. *Neuron*, **64**, 617–623.

Sonnhammer,E.L.L. and Ostlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.

Stern,S. *et al.* (2013) The transcription factor serum response factor stimulates axon regeneration through cytoplasmic localization and cofilin interaction. *J. Neurosci.*, **33**, 18836–18848.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Xie,C. *et al.* (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.*, **39**, W316–W322.

Yang,Z. *et al.* (2014). Multi-Scale Gaussian Mixtures for Cross-species Study. In *Proceedings of the International Multiconference of Engineers and Computer Scientists*, volume 1. Hong Kong, IAENG.

Yates,A. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.

Zheng,W. *et al.* (2011) Comparative transcriptome analyses indicate molecular homology of zebrafish swimbladder and mammalian lung. *PLoS One*, **6**, e24019.