MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3216–3241 doi: 10.1111/mec.12309

INVITED REVIEW AND META-ANALYSES

Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research

C. R. PRIMMER,* S. PAPAKOSTAS,* E. H. LEDER,* M. J. DAVIS† and M. A. RAGAN†

Department of Biology, University of Turku, 20014 Turku, Finland, †Institute for Molecular Bioscience, The University of Queensland, Brisbane, Qld 4072, Australia

Abstract

Recent advances in molecular technologies have opened up unprecedented opportunities for molecular ecologists to better understand the molecular basis of traits of ecological and evolutionary importance in almost any organism. Nevertheless, reliable and systematic inference of functionally relevant information from these masses of data remains challenging. The aim of this review is to highlight how the Gene Ontology (GO) database can be of use in resolving this challenge. The GO provides a largely species-neutral source of information on the molecular function, biological role and cellular location of tens of thousands of gene products. As it is designed to be species-neutral, the GO is well suited for cross-species use, meaning that, functional annotation derived from model organisms can be transferred to inferred orthologues in newly sequenced species. In other words, the GO can provide gene annotation information for species with nonannotated genomes. In this review, we describe the GO database, how functional information is linked with genes/gene products in model organisms, and how molecular ecologists can utilize this information to annotate their own data. Then, we outline various applications of GO for enhancing the understanding of molecular basis of traits in ecologically relevant species. We also highlight potential pitfalls, provide step-by-step recommendations for conducting a sound study in nonmodel organisms, suggest avenues for future research and outline a strategy for maximizing the benefits of a more ecological and evolutionary genomics-oriented ontology by ensuring its compatibility with the GO.

Keywords: gene annotation, genomics, ontology, proteomics, transcriptomics

Received 7 November 2012; revision received 22 February 2013; accepted 26 February 2013

General introduction

Recent rapid advances in molecular technologies have resulted in unprecedented opportunities for molecular ecologists to better understand the molecular processes regulating traits of ecological and evolutionary importance in almost any organism (Pennisi 2009). The most notable of these advances for researchers in ecology and evolution has been the advent of next-generation sequencing (NGS) technologies (Rokas & Abbot 2009). NGS enables significant proportions of the genome or transcriptome to be characterized in fine detail for essentially any organism. Indeed, studies capitalizing on the benefits of NGS technologies in (previously) genetically poorly known species are becoming more common (Vera *et al.* 2008; Hohenlohe *et al.* 2010; Bruneaux *et al.* 2013) as are reviews outlining the benefits of NGS-based approaches in ecological and evolutionary research (e.g. Rowe *et al.* 2011; De Wit *et al.* 2012). The realization that microarray and proteomic approaches can supplement more-direct methods to study gene function has also been evident recently (Forné *et al.* 2010; Leder *et al.* 2010; Weckwerth 2011; Diz *et al.* 2012; Leskinen *et al.* 2012; Papakostas *et al.* 2012). Although these technologies have reduced the challenge of generating molecular information, new challenges have arisen. One of these is inferring functionally relevant

Correspondence: Craig R Primmer, Fax: +35823336598;
E-mail: craig.primmer@utu.fi

information from these masses of data in a reliable and systematic way. The Gene Ontology (GO) database can be of use in resolving this challenge as it provides a highly structured, largely species-neutral source of information on the molecular function, biological role and cellular location of tens of thousands of gene products. The aim of this review is to highlight the potential of the GO database to assist researchers in molecular ecology to gain insights into gene function in essentially any organism. Expressed most concisely, GO can be used to provide putative functional information for the genes of species with poorly annotated genomes. To achieve this aim, we first explain the structure of the GO database and how functional information is linked to genes and gene products. We then present options for molecular ecologists to annotate their molecular data and outline various applications of GO for enhancing the understanding of the molecular basis of traits relevance in ecologically relevant species. We conclude by highlighting potential pitfalls, providing step-by-step recommendations for conducting a sound study in a non-model organism and suggesting avenues for future research. Throughout this review, unless otherwise specified, we use the term 'annotation' to refer to the assignment of *functional annotation,* in the form of GO terms, to genes and gene products, as opposed to *structural annotation* such as intron-exon boundary identification, etc. (Yandell & Ence 2012).

What is the Gene Ontology (GO) database?

An ontology is a formal structuring of knowledge (Box 1), and the GO specifically aims to provide a formal representation of (molecular) biological knowledge (Thomas *et al.* 2007). GO (http://www.geneontology.org) is built over a relational database that provides a catalogue of the biological function of genes and gene products using a standardized vocabulary (The Gene Ontology Consortium 2000). Its use of a standardized vocabulary helps to ensure that information is transferrable between studies, for example by recognizing that the terms 'translation' and 'protein synthesis' refer to the same biological process. The GO database actually encompasses three nonredundant ontologies: biological process, molecular function and cellular component. These ontologies describe aspects of the function of a gene or gene product and define the relationships between the terms. Rather than being a hierarchical tree, the GO is organized as a directed acyclic graph (DAG) in which the terms are nodes, and the relationships between them are represented as edges. This offers more flexibility than a simple hierarchy, as more specific 'child' terms can have multiple 'parents' (see Box 1 for more detail). Further, these ontologies can generally be applied at the DNA, RNA or protein levels.

From a molecular ecology perspective, one of the most important features of the GO database is its 'species-(or more generally, taxon-) neutrality', that is, it has been specifically designed to capitalize on the generally hierarchical pattern of conservation of gene and gene product structure, location and/or function in eukaryotes in particular (The Gene Ontology Consortium 2000; The Reference Genome Group of the Gene Ontology Consortium 2009). This conservation, interpreted as homology, underpins the automated transfer of information (referred to as 'evidence' below) from genetic model organisms to less well-studied species, including those important in ecological or other applied contexts. Emphasis on the transferability of information between species continues to increase within the GO consortium (Gaudet *et al.* 2011). The GO is utilized for annotating gene products, not for recording the responses of those gene products to a particular treatment or environment. Therefore, GO can be used to characterize the processes, functions and cellular locations of those gene products in any interesting scenario, whether in a drug treatment trial or in a common garden experiment. It follows that information from medically oriented experiments is highly useful for non-model organisms. As the correct identification of orthologous genes underpins the usage of the GO in ecology and evolution, the process of how this can be performed and the potential dangers of incorrect orthologue assignment are described in detail in the following sections.

What evidence underpins Gene Ontology annotation?

Before outlining how GO annotations can be assigned to nonannotated gene products, it is important to understand how annotations are assigned to model organism genome data, or indeed genome data of any species, in the first place. GO annotations are produced either manually by trained curators working within the GO Consortium member organizations, or computationally using automatic processes that exploit existing biological knowledge. Each annotation is assigned an 'evidence code' that reflects the evidence used by the curator when deciding on the correct term associations. These annotation evidence codes can be divided into three main categories: annotations based on (i) experimental evidence, (ii) curated nonexperimental evidence and (iii) (noncurated) electronic evidence. Within each of the broader categories, a number of more-specific evidence codes are defined (Škunca *et al.* 2012). An important distinction here is that while expression patterns can provide useful evidence that a gene product

Box 1

Ontology and the structure of the Gene Ontology database

An ontology is an explicit specification of concepts, including their attributes and the relationships between them, necessary to formally describe a given domain of knowledge (Gruber 1993). In the most simple case, an ontology may be a controlled vocabulary or dictionary, defining and restricting the terms (and their meanings) available for use. More-complex ontologies capture relations between concepts, commonly including a hierarchical classification presented as classes with increasingly more specific subclasses that are distinguished by attributes that differentiate sibling subclasses (*differentia*).

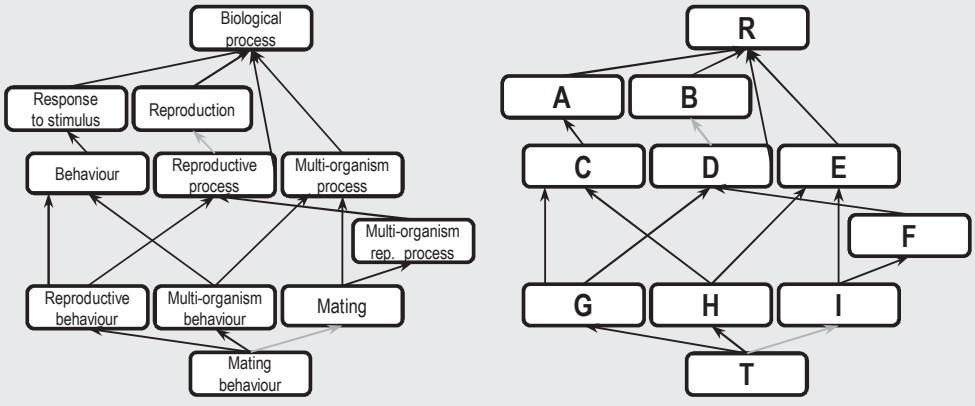
The Gene Ontology (GO) was developed by the Gene Ontology Consortium to model knowledge in the domain of molecular biology. It is structured with a moderate level of semantic complexity using the Open Biomedical Ontology (OBO) description format, a standard that has emerged from the OBO Foundry project (Smith *et al.* 2007). GO contains three distinct, well-developed conceptual hierarchies defining key concepts of molecular biology: biological process (BP), molecular function (MF) and cellular component (CC). The GO consortium describes the three domains as follows: *Biological processes are operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs and organisms; Molecular functions are the elemental activities of a gene product at the molecular level, such as binding or catalysis; and Cellular components are the parts of a cell or its extracellular environment* (see http://www.geneontology.org/GO.doc.shtml). Each hierarchy has a root class, or term, from which increasingly specific terms descend (see Graph A), largely via two distinct kinds of transitive relationship: *is_a* relations, which establish that a child term is a more-specific subclass of the parent term, and *part_of* relations, which establish that the instances associated with the child term are contained within the instances of the parent term. These relations, and others less-commonly used in the Gene Ontology, are formally defined in the OBO Relation Ontology (http://obofoundry.org/ro/).

Because the *is_a* and *part_of* relations that define its conceptual hierarchies are transitive, reflexive and antisymmetric (Bittner & Donnelly 2007), GO can be represented as a directed acyclic graph (DAG). The GO Consortium additionally requires each term should have at least one *is_a* complete path to the root, plus at least one path containing at least one *part_of* relation (see Graph A).

The Gene Ontology database

The GO database contains the specifically defined terms and relations that make up the ontology, as well as large collections of annotations, or instance data, that associate the individual *components* of the molecular biology domain (gene products, transcripts, proteins, miRNAs, etc.) with the classes that accurately describe them. These annotations are available for download in a variety of formats and, together with the terms, can be retrieved using the database search interface.

Box 1 Graph



A visualisation of the sub-graph defining the biological process "Mating behaviour" with a generic representation of the graph topology. In the generic example, terms C, D and E each have two child terms (G & H, G & F, H & I, respectively). Both C and D are parents of G, thus G has multiple parents and inherits the attributes defined in both. H and I also have multiple parents. As required by the GO specification, term T contains at least one complete *is_a* path through the ontology to the root term R (there are several such paths e.g. T→G→C→A→R), as well as at least one path to R that contains a part_of relation (indicated with grey arrows here) e.g. T→I→E→ R). Any instance of a term will also be a valid instance of every parent term along the *is_a* paths back to R; thus, any gene products annotated with the term "reproductive behaviour" can also be annotated with the processes "reproductive process" and "behaviour".

is involved in a biological process, gene expression patterns as such are not part of the domain of GO and are not described by the ontology (i.e. GO is not an expression database). Experimental evidence is generally considered as the most reliable form of annotation evidence (Škunca *et al.* 2012), but due to the time-consuming nature of such experiments, only a small proportion of annotations is supported by experimental evidence even in many model organisms (Fig. 1; Table S1, Supporting information). Curated nonexperimental annotation codes include annotations such as 'inferred from sequence or structural similarity' (ISS), 'traceable author statement' (TAS) and 'inferred by curator' (IC). The distinguishing feature of this group of evidence codes is that although no experimental evidence is available, the available evidence (computational or otherwise) has been manually reviewed by a curator. The third and largest evidence code category includes annotations that have been automatically assigned, or 'inferred from electronic annotation' (IEA). Such annotations are assigned based on some form of *in silico* analysis and have not been manually evaluated. Due to the lack of manual evaluation, IEA annotations are generally considered to be less reliable (Škunca *et al.* 2012); however, they make up the vast majority of annotation in most species (Fig. 1; Table S1, Supporting information). While evidence codes are usually ignored in applications such as GO enrichment analysis, they can be used to minimize problems of redundancy and bias (Rogers & Ben-Hur 2009; see below).

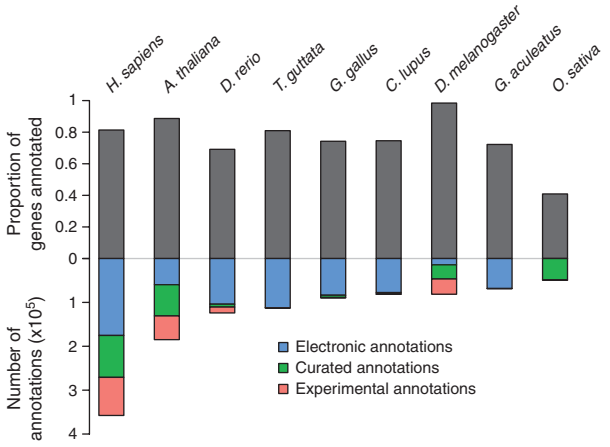


Fig. 1 The proportion of annotated genes and their types of annotations for nine sequenced genomes (as of February 2013). Humans (*Homo sapiens*) and *Arabidopsis thaliana* have the highest number of annotations for animals and plants, respectively. They also have the most experimentally derived annotations. Most other species, except *Drosophila melanogaster*, are annotated mostly electronically. Numbers of annotations available for these species and specific evidence codes are listed in Table S1 (Supporting information).

Using GO to annotate genes in nonannotated genomes

Most studies in ecology and evolution focus on nonmodel organisms (but see e.g. Landry *et al.* 2006; Arya *et al.* 2010; Lee & Mitchell-Olds 2012), which almost by definition have traditionally had little or no annotation associated with their genes and gene products. Further, given that these studies often focus on organisms in natural settings, with unknown and potentially complex population structures and lacking inbred lines, cell lines or other tools for experimental genetics, it has traditionally been difficult to gain experimental evidence for gene annotation directly in species of ecological interest. Although this is slowly changing (e.g. Edwards *et al.* 2009), the fact remains that genome annotation is heavily biased towards the traditional genetic models (Fig. 1). This is where the species neutrality of the GO database comes into play, as the annotation evidence from genetic model organisms can be transferred to less well-studied species by identifying putatively orthologous sequences and assuming they have the same function in both species. Large-scale studies support this generalization about orthologues, albeit more weakly and less consistently than is often assumed (Nehrt *et al.* 2011; Altenhoff *et al.* 2012). Thus, an inference of orthology 'lends legitimacy to the transfer of functional information' from one sequence to another (Koonin & Galperin 2003). This is, in fact, a common practice already in molecular ecological research at the single-gene level; MHC genes, for example, are frequently assumed to play an important role in the immune defence system regardless of whether experimental molecular evidence of this role is available in the study species. When scaling up from single to thousands of genes, however, there is increased potential for erroneous orthology assignment, especially when distant species are compared, or when the genome evolution of the species in question is suspected to be complex, for example involving lateral gene transfer (LGT) or genome duplication. Below, we outline the current practices used for orthologue inference in nonannotated genomes in different circumstances (see also Table 1 and Box 2).

Cross-species transfer of GO terms in practice

Genome annotation for a non-model organism starts with the assembly of sequence reads into contiguous regions (contigs), and the discovery and delineation of genes therein, that is, structural annotation (Yandell & Ence 2012). Ideally, this involves the entire research community around that organism and can be an ongoing process. Given a list of genes emerging from such a process, here we consider the next step: identification of

Box 2
Orthology and tools for inferring orthologues

What is orthology?

Orthology is a centrally important concept, although often misunderstood. Fitch (1970, 1973) recognized two subclasses of homology: orthology (resulting from speciation) and paralogy (resulting from gene duplication), and these definitions have been widely adopted (e.g. in the Instructions to Authors for *Molecular Biology and Evolution*). Thus, orthology and paralogy are best inferred on phylogenetic trees, not from function, expression, genomic location or similarity of sequence or structure. The past fifty years have seen great progress in molecular phylogenetics, but inferring high-quality trees remains a challenge, especially for the very large data sets now arising from NGS. Consequently, there has been much interest in fast surrogate approaches that, unlike rigorous tree inference, can easily be built into automated workflows. In fact, the GO Consortium establishes orthology (evidence code ISO) by 'multiple criteria generally including amino acid and/or nucleotide sequence comparisons and one or more of the following: phylogenetic analysis, coincident expression, conserved map location, functional complementation, immunological cross-reaction, similarity in subcellular localization, subunit structure, substrate specificity or response to specific inhibitors' (www.geneontology.org/GO.evidence.shtml#computational). While perhaps necessitated by the broader framework of their annotation workflow, these criteria stand well outside the established meaning of *orthology*, and best (i.e. tree-based) practice for its inference (but see Altenhoff & Dessimoz 2012). More details on orthology definition can be found in Appendix S1 (Supporting information). Tools for orthologue identification are listed in Table 1.

Orthology databases

The GO Consortium instituted the Reference Genome Annotation Project (2009) to provide direct functional annotation for human and 11 other important 'model organisms' (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus, Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe)*. Multiple points of reference offer a broader range of annotated function (e.g. photosynthesis), and (depending on the nonmodel species being annotated) potentially stronger pairwise match scores and more finely resolved trees, hence (in principle) fewer inaccuracies in orthologue assignment. Databases mapping orthology relationships between sequences in diverse taxa are also available online (http://questfororthologs.org/orthology_databases). The Clusters of Orthologous Groups (COG, http://www.ncbi.nlm.nih.gov/COG/) are constructed based on all-against-all BLAST searches of complete proteomes from several eukaryotic model organisms including *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Homo sapiens* (Tatusov *et al.* 2003). Reciprocal best hits in BLAST searches are interpreted as a pair of orthologues, and each COG represents relationships between at least three phylogenetically distant taxa. HomoloGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene) employs a similar strategy to detect putative orthologues and paralogues among the genes of 20 sequenced eukaryotic genomes, while UniGene (http://www.ncbi.nlm.nih.gov/unigene) is a system that uses BLAST to partition transcript sequences from numerous animal and plant species into nonredundant set of clusters that represent potential genes (Sayers *et al.* 2010). This vast amount of information can greatly facilitate gene annotation by orthologue identification in nonmodel organisms. However, the level of agreement between different databases is unfortunately not always high (Chen *et al.* 2007; Altenhoff & Dessimoz 2009; Shin *et al.* 2009; Boeckmann *et al.* 2011; Kristensen *et al.* 2011).

A number of easy-to-use analytical pipelines have been developed that help not only to streamline the orthologue inference process, but also to conduct downstream identification of GO annotation (Tables 1 and 2). These pipelines provide valuable heuristics for GO-based analyses, but as the default parameters may not be appropriate in specific cases, a thorough understanding of what is happening at each phase of the analysis is required. Further, key parameter settings should be reported in publications (see Box 6).

orthologues in one or more well-annotated genomes, ideally of closely related model organisms. As explained in Box 2, orthologues are defined on a phylogenetic tree, but inferring tens of thousands of trees *de novo* may not be feasible. Thus, putative orthologues are commonly identified by similarity search, usually using the Basic Local Alignment Search Tool (BLAST) or one its variants (Altschul *et al.* 1990, 1997) to find the best (highest scoring) match. By default, this best-matching sequence is taken to be the orthologue of the sequence in question, and on this basis, GO annotations are transferred from the well-annotated target to the nonannotated query sequence. Tools such as BLAST2GO (Conesa *et al.* 2005) are specifically designed as a rapid means to achieve this, combining BLAST searches and subsequent GO annotation mapping.

Table 1 Strategies for orthologue inference and subsequent GO term and functional enrichment analyses

	Orthologue inference	GO term and functional enrichment
1. *Aim*	Transfer of annotation information from a related, well annotated, species	Compare frequency of occurrence of GO terms in focal and background gene lists; further analysis of functions
2. *Standard approach(es)*	Closely related species: BLAST-based similarity search of the most closely related annotated genome; Distant species, complex genome evolution: tree-based methods (e.g. ORTHOMCL)	Implemented in BLAST2GO, GOStat, DAVID and more
3. *Potential pitfalls*	(i) Incorrect orthologue identification; (ii) Inaccurate GO annotation	(i) Incorrect choice of reference or background gene sets; (ii) Incorrect statistical approach; (iii) Inter-relationships among terms not fully captured; (iv) Lack of annotation coverage
4. *Possible Solutions*	(i) Reciprocal best hits (RBH) strategy; exclude low-complexity sequence and coiled-coil regions; increase match stringency (e-value, bit score, etc.; replace RBH with reciprocal smallest distance (Wall & Deluca 2007); replace pairwise strategy with multiple genome comparisons, for example, COGS; (ii) exclude automatically assigned GO annotations (IEA evidence code) and/or those from more distant model organisms.	(i) Genome-wide studies: use whole genome as background, otherwise use the total gene-set of the study; (ii) See Box 3; (iii) & (iv) consider options listed below
5. *Additional options*	Tree-based approach (e.g. PAINT); supplement annotation with information from protein domain signatures (e.g. InterProScan)	Further functional exploration: for example, GO hierarchy visualization (BINGO, Cytoscape), gene set or modular (network) enrichment analysis (CLUEGO)
6. *More details*	Sections "What evidence underpins Gene Ontology annotation?", "Using GO to annotate genes in nonannotated genomes", "Potential pitfalls", Box 2, Altenhoff & Dessimoz (2012)	Sections "What can GO be used for?", "Potential pitfalls", Box 3, Rivals *et al.* (2007); Rhee *et al.* (2008); Huang *et al.* (2009); Khatri *et al.* (2012)

With similarity-based approaches, the best-performing metric (e.g. similarity value e-value, or bit score) and coverage threshold depend strongly on details of the individual analysis. Factors affecting performance include the phyletic distance between query and target genomes, complexity of each gene or protein family (e.g. its size, lineage-specific gene losses, duplications, domain shuffling or LGT), quality of the genome annotation and species coverage (Trachana *et al.* 2011). Amino acid–based BLAST searches (blastp or blastx) may be more appropriate in transcriptomics or proteomics experiments where large blocks of coding sequence are available, and/or where the query and targets are phyletically distant or divergent; on the other hand, nucleotide-based searches may be more appropriate with expressed sequence tag (EST) or restriction site associated DNA (RAD) data, where non-protein-coding sequence is abundant, and/or where the query and reference species are more-closely related.

Trade-offs in orthologue inference: false positives vs. false negatives

Studies in model organisms indicate that protein-sequence similarity of at least 40–60% is required for accurate prediction of function (reviewed by Addou *et al.* 2009). Alternatively, a BLAST bit score of at least 244 has been suggested to provide accurate prediction of functional similarity (Louie *et al.* 2009). Regardless of the similarity metric(s) used, an intrinsic trade-off exists between false positives (recovering paralogues or other similar, but nonorthologous, sequences) and false negatives (missing true orthologues), at least up to a point. Chen *et al.* (2007) found that for BLASTP searches, increasing the e-value stringency beyond a certain threshold did little to reduce the proportion of false positives but increased the proportion of false negatives. Methods that combine phylogeny and similarity searches can reduce both the false-positive and false-negative rates (e.g. INPARANOID: Östlund *et al.* 2010).

Orthologue inference: best practices

Multiple metrics and criteria should be examined to help minimize false positives, and all details including threshold values should be clearly reported in the methods sections of publications where GO is applied across species. It is also important to remember that e-value is dependent on the size of the database used in the

search, so studies evaluating the performance of BLAST-based methods should also report other alignment metrics that are not database-dependent such as per cent identity and alignment coverage.

A very popular approach, which in practice provides more-robust inference of orthologues, is to use reciprocal (bidirectional) top BLAST hits between two species (Mushegian & Koonin 1996; Tatusov *et al.* 1997, 2001; Rivera *et al.* 1998; Hirsh & Fraser 2001; Kristensen *et al.* 2011). This approach, often termed reciprocal best hits (RBH), reduces the frequency at which paralogues are recovered when an orthologue is absent (Li *et al.* 2003). RBH is most effective with relatively closely related taxa and in general is less successful with non-model species for which the genome may be incompletely sequenced or annotated. The RBH approach can be extended to three or more sequences, as in methods such as COGS (Tatusov *et al.* 2001, 2003).

How closely related does a species with a well-annotated genome need to be for similarity-based approaches to be effective? Currently, the main limitation is the relatively small number of species with extensive GO annotation (The Reference Genome Group of the Gene Ontology Consortium 2009), so the choice of appropriate model species may come down to choosing between a vertebrate, a plant, an arthropod, a worm, etc. For example, in our experience with non-model vertebrates, the superior GO-term annotation for the human genome currently makes it the preferred reference genome over even zebrafish for the transfer of gene- or protein-based annotation to fish sequences. Nonetheless, considerable information is lost, especially for DNA-level comparisons, due to the lack of a more-closely related, well-annotated species. For example, of 6200 Atlantic salmon (*Salmo salar*) SNP sequences, most of which were EST-derived, GO terms could be identified using BLAST2GO for less than a half when an E-value threshold of 10^{-10} was applied (Bourret *et al.* 2013).

Even when no homologous sequence can be identified, several options remain for assigning a function. The most commonly encountered methods involve recognizing specific domains at the protein level, as domains typically dictate function. These domains can be valuable in annotating taxon-specific genes that may not have been characterized in any model organism, or rapidly evolving genes for which the divergence from available model organism sequences may be too large to enable identification *via* sequence similarity. As such, these methods should not be seen solely as alternatives to similarity-based searches, but rather also as extensions in some cases as an additional means of annotation assignment. InterProScan (Quevillon *et al.* 2005) performs this task, followed by INTERPRO2GO mapping that retrieves GO annotations (Burge *et al.* 2012). This approach is computationally intensive, however, particularly at full genome (or proteome) scale.

In some circumstances, similarity-based searches are inappropriate, for example, where genome evolution is suspected to have been complicated by nonhomologous gene replacement (Koonin *et al.* 1996), genome duplication (Jiao *et al.* 2011) or copy number variation (McHale *et al.* 2012), each of which undermines or complicates making the distinction between orthologues and paralogues by similarity searches alone. At larger phyletic distances or in cases where rapid evolution of new gene function may be expected (Colbourne *et al.* 2011), it is necessary to apply rules to accommodate paralogues arising from duplication after speciation. Examples of this approach can be found in the INPARANOID (Remm *et al.* 2001) and ORTHOMCL (Li *et al.* 2003) algorithms (Table 1). ORTHOMCL is similar to the INPARANOID algorithm, but clusters orthologues from multiple species and distinguishes between paralogues derived from duplications before or after a given speciation event using relative distances based on within- and between-species reciprocal BLAST hits (Li *et al.* 2003). These tools have proven invaluable for the study of taxa that have undergone repeated whole-genome duplications (e.g. Jiao *et al.* 2011).

What can GO be used for?

Experimental design

In an experimental design scenario (Fig. 2) where the function of a gene of interest is known, the GO database can be used to identify genes with similar or related functions or cellular locations in two organisms, or to identify gene products that interact in one organism, thereby guiding the expansion of a study to related or interacting genes. Alternatively, if a particular biological process, molecular function or cellular component is suspected or predicted to be of importance, GO can be queried to retrieve a list of functionally relevant candidate genes for further investigation or to test a specific hypothesis. Examples where the GO has been used for experimental design purposes in an ecological or evolutionary context are still rare, but one such study (Wenzel *et al.* 2013) is detailed in the 'Examples' section below.

Postexperiment data analyses

Postexperimental applications of the GO database (Fig. 2) are much more common, with one of the most popular applications being to make functional sense out of high-throughput molecular data. This can be carried out in a descriptive, an exploratory or a hypothesis-driven

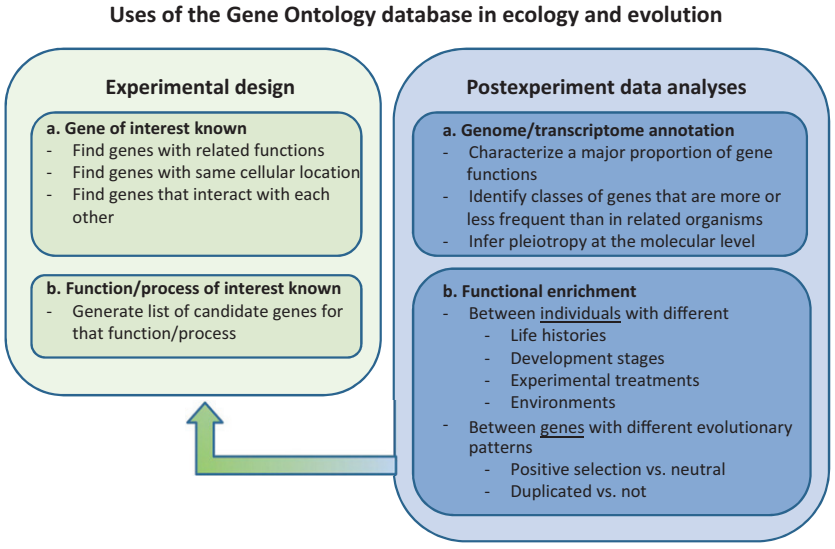


Fig. 2 Summary of the uses of the Gene Ontology database in ecology and evolution. When the function of a gene of interest is known, or a particular biological process, molecular function or cellular component is suspected or predicted to be of importance, GO can be queried to retrieve a list of functionally relevant candidate genes for further investigation or to test a specific hypothesis. Alternatively, the GO can be used to describe the functions of gene products observed in high-throughput molecular data or to identify differences in the functional categories between individuals from experimental treatments or life history stages or between genes of different categories.

fashion, or a combination of these. An example of a descriptive use of GO is its use in annotating newly sequenced genomes, or in characterizing a transcriptome or EST library. Gene and genome annotation is recognized as one of the most important phases of sequencing projects (Danchin *et al.* 2007; Yandell & Ence 2012). It involves the identification of genes and gene variants in the genome or transcriptome, including protein amino acid sequences and potential splice variants (structural annotation), followed by assigning a function to as many of the identified genes as possible (functional annotation). Presently, structural annotation is generally achieved bioinformatically and although not a trivial procedure, general guidelines are available (Yandell & Ence 2012). Determining gene functions experimentally for every organism would be a mammoth task but, as noted earlier, the GO database capitalizes on the often high level of sequence similarity among eukaryotic genes and gene products to enable functional annotation to be transferred among species. In this way, a generally robust overview of gene function can be obtained in species for which little or no direct experimental functional annotation information exists (e.g. Vera *et al.* 2008; Ji *et al.* 2012). Such information also allows comparative analyses of the distribution of gene function classification in comparison with related model organisms, thus enabling researchers to assess the completeness of genome annotation (Star *et al.* 2011), functionally compare transcriptomes of different tissues, or assess whether specific gene classes or functions are enriched or depleted, possibly indicating novel adaptations (e.g. Zhou *et al.* 2009). Further, GO can be used to annotate the probes included in genomic resources such as cDNA or SNP microarrays (Rise *et al.* 2004), facilitating GO-related analyses and making results more comparable across studies. The Atlantic salmon (*Salmo salar*) cDNA microarray provides a good example of the benefits of providing GO annotation, as numerous researchers utilize this feature of the microarray with good results (e.g. Giger *et al.* 2008; Normandeau *et al.* 2009; Renaut & Bernatchez 2011; Tadiso *et al.* 2011).

A common exploratory approach can be broadly categorized as 'gene ontology enrichment' or 'functional enrichment' tests (see Box 3). These tests enable a move from statistically testing single genes to discovering significant biological features in groups of 'interesting' genes, usually identified on the basis of a high-throughput experiment. The motivation for performing functional enrichment tests is the assumption that if a particular biological process/molecular function/cellular component plays an important role in a biological phenomenon, the gene products involved in that process should respond more significantly (in either frequency, or strength of response) than gene products unrelated to such key processes. The response being measured will depend on the study but could be the subset of gene products in a study, which are affected by positive selection, or those that have increased or decreased expression level. Enrichment tests look at the frequency

Box 3
Gene ontology enrichment tests

High-throughput experiments such as those involving proteomic or transcriptomic profiling, or sequencing, can generate very large sets of results that are typically presented as lists of *genes of interest*. Interpreting these lists is not always straightforward. Whereas a scientist may deduce the pathways or biological mechanisms underlying the results of low-throughput experiments, applying the same standard of analysis to many thousands of genes of interest is problematic. The development of the Gene Ontology and the increasing abundance of GO-annotated gene products in reference databases have enabled the development of several approaches that facilitate biological interpretation of large gene lists. Here, we focus on one of the most common approaches, the gene ontology enrichment test. Other strategies, such as pathway analysis and other knowledge-based modelling approaches, have been reviewed recently (Bauer-Mehren *et al.* 2009; Khatri *et al.* 2012).

The problem addressed by enrichment tests can be formulated as follows: given an experiment that measures the 'behaviour' of a 'population' of genes, some subset of this population will be determined to be of interest (i.e. the results of the experiment—those genes whose behaviour is influenced or changed by the treatment, condition or other variable that is the subject of the experiment). We observe that the results are associated with a set of GO terms but do not know if the frequency of these terms is significant—in other words, we are not sure if we would see the same distribution of GO terms in the results in an equivalently sized set of genes randomly sampled from the same population. Statistical enrichment tests are designed to answer this question by determining if the distribution of GO terms observed in the results is significantly different than the distribution we might expect given a random sample of equivalent size. A number of statistical tests can be applied to this general problem, and the various tests and specific problem statements are the subject of a detailed review (Rivals *et al.* 2007). Briefly, the most commonly applied statistical tests are Fisher's exact test and hypergeometric tests (which are equivalent), and the chi-squared-test and test of equality of two probabilities (also equivalent). In general, the chi-squared-test is appropriate only for large samples, whereas Fisher's exact test can be applied more generally. These tests are typically formulated to test for *enrichment*, that is, an over-representation of GO terms in the result set compared to the baseline, and are thus one-sided tests where the critical region is on the right of the distribution. However, it is possible that a researcher may be interested in *depletion*, or under-representation of terms, in which case a one-sided test with a critical region on the left is applied. If both enrichment and depletion are of interest, then a two-sided test should be applied to identify both categories of terms.

A large number of applications implementing enrichment tests are available and are reviewed elsewhere (Rivals *et al.* 2007; Khatri *et al.* 2012) and listed on the Gene Ontology Consortium website (http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools). However, before conducting an enrichment test, it is critical that researchers understand what question is in fact being addressed by the test and what the limitations of the analysis are. Most importantly, the validity of these tests depends on an accurate determination of GO-term frequency in the population of genes being measured, referred to as the *background*. For example, in a microarray experiment measuring the expression of *Arabidopsis* genes, the correct background set to use in an enrichment test is not the full set of known *Arabidopsis* genes, but rather the set of *Arabidopsis* genes present on the microarray used in the experiment (i.e. the proportion of the quantified transcriptome). While these two sets may approach equivalency in many cases, in others, they may be radically different as, for example, in the case of custom arrays designed to measure the behaviour of a specifically restricted set of genes. Tools that enable a researcher to specify a background set, as well as a list of interest, are more likely to give accurate results. Another factor that affects the validity of an enrichment test is the quality of the background set. If the experiment is based on an incomplete or poorly assembled transcriptome, this will affect the quality of the annotations being used in the test. Likewise, the extent to which GO annotations are available for the population of genes covered by the experiment is important. The term *coverage* is used to indicate the proportion of the background set for which annotations are available. Backgrounds with low GO-term coverage are not likely to produce robust results. Some tools, such as the DAVID web-tool (Table 2), will report coverage statistics for enrichment tests, providing useful insight into the extent to which researchers can rely on the resulting enrichments. The issue of coverage is expected to be most problematic with nonmodel organisms, or organisms with poorly annotated genomes (see Fig. 1).

Researchers should also keep in mind what the results of an enrichment test actually mean. The GO terms returned by such tests are not a complete categorization of the functionality present in the gene list. Many terms annotated in a list of genes may not be significant, that is, there is a significant likelihood (as measured by the *P*-value of the test) that those terms would be found at such frequency in an equivalently sized random sample of genes. Thus,

Box 3 *Continued*

these terms are not considered to be related to the experimental condition but are instead assumed to be present due to their relative frequency in the background set. If a researcher is interested in categorizing genes in a list to find out what processes are covered by (or which functions are present), then an enrichment test is not required.

Additionally, when testing a very large number of hypotheses (which for an enrichment analysis is the number of GO terms being tested, not the number of genes on the array), corrections for multiple hypothesis testing should be used to control the risk of false positives (falsely rejecting the null hypothesis, or Type I error—determining that a GO term is significant when it is not), while minimizing the chance of introducing false negatives (failing to reject a false null hypothesis, or Type II error—determining that a GO term is not significant when it is). Most tools performing rigorous statistical enrichment analysis will apply some form of correction, report both a raw *P*-value and an adjusted *P*-value and specify what correction has been applied. Approaches for correcting for multiple hypothesis testing were evaluated by Bluthgen *et al.* (2005) and reviewed by Farcomeni (2008). Some corrections, such as the Bonferroni correction, can be quite harsh and introduce unwanted false negatives, while other methods, such as the Benjamini–Hochberg method, are less conservative (Thissen *et al.* 2002). Regardless of the correction applied, a complementary approach to improving the strength of *P*-values is to test fewer hypotheses, such as by generating species-specific subsets of GO, or mapping full GO annotations into a GO slim (Davis *et al.* 2010).

The workflow for performing an enrichment test is outlined in Table 1, and examples of available analysis tools are listed in Table 2. While many applications exist for performing these tests, researchers are encouraged to work through a formulation of the problem that they want to solve, clearly determine the null hypothesis being evaluated and carefully identify the appropriate background set for testing, to obtain accurate and informative results. Finally, enrichment tests answer a very specific question regarding the probability of seeing GO terms in a gene list given the frequency of those terms in a background set, and the validity of the result depends on factors such as those described previously.

of GO terms in experimental results and compare that frequency to the observed background frequency in the set of genes measured in the experiment. If the frequency of a term is significantly different from what is expected based on the background frequency, then the term is said to be *enriched* (or potentially *depleted:* Box 3).

Since GO represents the largest repository of functional roles of gene products, several methodologies for functional enrichment analyses utilize GO annotations for defining gene product function (Tables 1 and 2). In this way, GO annotations offer a basis for ascertaining whether genes of a certain function are over- or under-represented in two contrasting experimental groups via a likelihood ratio test (Box 3). Such groups could represent different kinds of individuals, for example, with different life history or developmental stages, control vs. treatment groups, or individuals from contrasting environments; or they could contrast groups of genes or gene products within a species (e.g. those evolving under positive selection vs. neutrally or those retained following genome duplication). Table 3 includes a non-exhaustive list of such studies. The same approach can be used in a hypothesis-testing framework by specifically asking whether certain GO terms are significantly over- or under-represented in the results from a particular experiment (Kim & Caetano-Anollés 2010).

Examples of the use of GO in ecology and evolution

Phylogenomics and GO analysis reveals significance of genome duplication in plant diversification

The success of angiosperm plants has been recognized to be due, at least in part, to the evolution of innovations following whole-genome duplication (e.g. De Bodt *et al.* 2005). Recently, it was unclear if this duplication pre-dated the divergence of angiosperms and what kind of genes may have subsequently aided their spread. Jiao *et al.* (2011) addressed these questions using a phylogenomics approach. They used genome sequences that were available for nine plant and moss species and sequenced a further 12.6 million ESTs from key plant lineages; from these, they identified a set of almost 800 'orthogroups' (clusters of inferred homologous genes originating from a common ancestral gene in a defined organismal ancestor: Wapinski *et al.* 2007), phylogenetic analyses of which provided convincing evidence for two distinct genome duplication events: one in the common ancestor of all angiosperms, the other in the common ancestor of all seed plants. The authors then conducted a functional enrichment analysis based on GO annotations to shed light on the particular biological processes, and the genes behind them, that may

Table 2 Examples of tools available for GO term browsing, annotation and downstream analyses

Tool	Purpose	Address	References*	Comments
GO Browsers				
AMIGO	An interface to search and browse GO annotation data	http://amigo.geneontology.org/	–	The 'official' tool of the Gene Ontology
QUICKGO	An interface to search and browse GO annotation data	http://www.ebi.ac.uk/QuickGO/	1	
GO annotation via orthologue identification				
BLAST2GO	Putative orthologue identification via a BLAST search	http://www.blast2go.com	2	Also conducts downstream analyses
ARGOT2	Putative orthologue identification via a BLAST search	http://www.medcomp.medicina.unipd.it/Argot2	3	Also conducts downstream analyses
INPARANOID	Putative orthologue identification via pairwise species comparisons	http://inparanoid.sbc.su.se/cgi-bin/index.cgi	4	
ORTHOMCL	Putative orthologue identification using reciprocal BLAST	http://orthomcl.org	5	Best option when complex genome evolution is suspected
INTERPRO2GO	Putative orthologue identification via protein domain identification	http://www.ebi.ac.uk/GOA/InterPro2GO.html	6	Also conducts downstream analyses
TREEFAM	Tree-based method for putative orthologue identification	http://www.treefam.org/	7	
PHYLOMEDB	Tree-based method for putative orthologue identification	http://phylomedb.org/	8	
Orthologue databases				
COGS	Maintains clusters of orthologous groups (COGS) of proteins	http://www.ncbi.nlm.nih.gov/COG/	9	Delineated by comparing protein sequences encoded in 66 complete genomes, representing 38 major phylogenetic lineages
ORTHOMCL BD	Houses orthologue group predictions for 150 genomes	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi	10	
GO term enrichment analysis				
GORILLA	GO enrichment and visualization tool	http://cbl-gorilla.cs.technion.ac.il/	11	
BINGO	GO enrichment and visualization tool	http://apps.cytoscape.org/apps/bingo	12	Cytoscape plugin† Highly customizable
CLUEGO	GO enrichment and visualization tool	http://apps.cytoscape.org/apps/cluego	13	Cytoscape plugin† Infers groups of functionally similar GO terms
GO term redundancy estimation				
REVIGO	Finds representative subsets of related GO terms using semantic similarity measures	http://revigo.irb.hr/	14	

Table 2 *Continued*

Tool	Purpose	Address	References*	Comments
G-SESAME	Measures the semantic similarities of GO terms	http://bioinformatics.clemson.edu/G-SESAME/	15	
Taxon specific GO resources				
G:PROFILER	Web server for functional interpretation of gene lists in >80 species	http://biit.cs.ut.ee/gprofiler/index.cgi	16	
AGRIGO	GO analysis toolkit and database for agriculturally relevant species	http://bioinfo.cau.edu.cn/agriGO/	17	
Other				
DAVID	A comprehensive set of functional annotation tools for any given gene list	http://david.abcc.ncifcrf.gov/home.jsp	18	Recognizes identifiers from various model organisms (including *Arabidopsis thaliana*, *Danio rerio* and *Gallus gallus*)
GOTOOLS	Contains various tools developed by the GO Consortium and by third parties	http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools	–	
BIOCONDUCTOR	R-based package with >400 GO-related modules	http://www.bioconductor.org	19	
PINA	Network analysis platform that integrates protein–protein interaction information	http://cbg.garvan.unsw.edu.au/pina/	20	

*1—Binns *et al.* 2009; 2—Conesa *et al.* 2005; 3—Falda *et al.* 2012; 4—Östlund *et al.* 2010; 5—Li *et al.* 2003; 6—Burge *et al.* 2012; 7—Ruan *et al.* 2008; 8—Huerta-Cepas *et al.* 2011; 9—Tatusov *et al.* 2003; 10—Chen *et al.* 2006; 11—Eden *et al.* 2009; 12—Maere *et al.* 2005; 13—Bindea *et al.* 2009; 14—Supek *et al.* 2011; 15—Du *et al.* 2009; 16—Reimand *et al.* 2007; 17—Du *et al.* 2010; 18—Dennis *et al.* 2003; 19—Gentleman *et al.* 2004; 20—Wu *et al.* 2009.

†Cytoscape is a an open-source platform for complex network analysis and visualization (http://www.cytoscape.org/, Shannon *et al.* 2003).

have been important in the origin and rapid diversification of angiosperms. To do this, they first reduced the size of the GO database, retaining terms likely to be of relevance, using the *Arabadopsis* GO slim (Box 4) as a starting point. They custom annotated the orthogroups that lacked GO annotation and added them to the GO slim, thereby enabling analysis of a more-complete data set. These analyses revealed several functional categories that were enriched in the orthogroups of both genome duplication events, including regulatory functions important to seed and flower development such as tranferases and binding proteins, transcription factors and protein kinases. The authors were able to conclude that retention of certain types of genes following genome duplication has been common during the re-diploidization process during the evolutionary history of plants.

Use of the GO for testing sexual selection handicap hypotheses

An innovative use of the GO in experimental design was recently reported by Wenzel *et al.* (2013). They used GO annotations to identify genes for a targeted expression analysis in red grouse (*Lagopus lagopus scoticus*) aimed at testing two sexual selection handicap hypotheses: the immunocompetence handicap hypothesis (ICHH) and the oxidative stress handicap hypothesis (OSHH). Firstly, they identified GO terms for biological process and molecular function predicted to be important in the respective handicap hypotheses: processes of immune function were proposed to be involved in the ICHH and were identified via the GO biological process term *immune system process* (GO: 0002376), while processes that respond to generation of reactive oxygen species were proposed to be involved in the OSHH and were identified via the GO biological process term *response to oxidative stress* (GO: 0006979) and the GO molecular function term *antioxidant activity* (GO: 0016209). To do this, the 5925 unique transcripts present on their custom cDNA microarray were annotated with GO terms employing a hierarchical search strategy using BLAST2GO as follows: first, the entire SWISSPROT database was queried with a BLAST e-value threshold of 10^{-10}, after which sequences with no match were

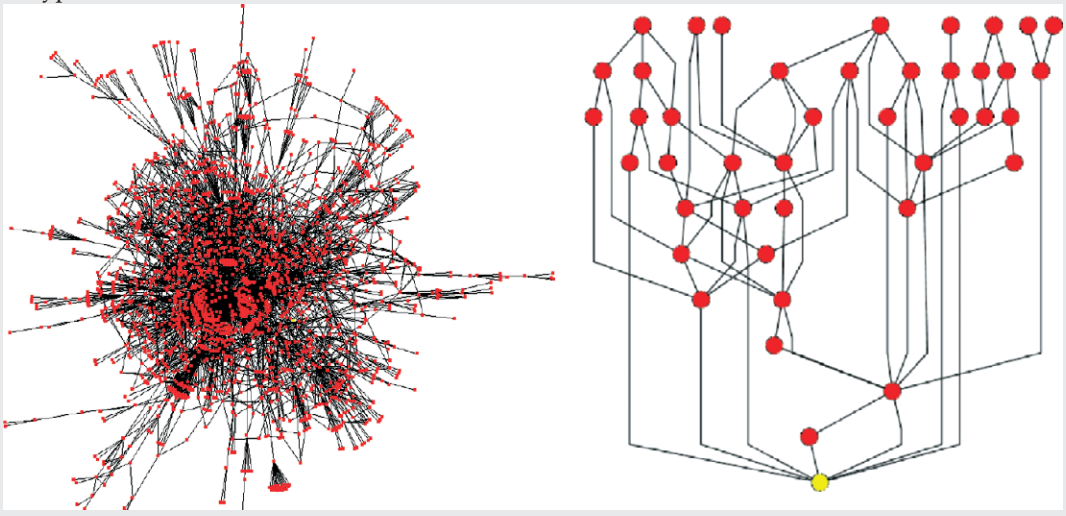
Box 4
GO slims

The Gene Ontology provides terms to cover the gene products of all known organisms, and thus now contains in excess of 35 000 terms. While all terms are required for gene product associations or to support the ontology structure (see Box 1), many terms will not be required to annotate a specific set of gene products, for example, the genes represented in a microarray experiment, or the genes of a particular species. For a number of reasons, some explored below, researchers may find it useful to work with a subset of GO terms. The Gene Ontology Consortium has defined such subsets as 'GO slims' (http://www.geneontology.org/GO.slims.shtml). There are many contexts in which a GO slim may be useful. These include supporting annotation for an experiment, categorizing the genes of a particular species, or performing conceptual mapping between the literature and the GO. In the following discussion, we will assume that the purpose of specifying a GO slim is to create a species-specific subset of the GO.

A GO slim is much smaller in size than the full Gene Ontology (e.g. Box 4 graph). More-detailed terms tend to be sacrificed, and gene annotation associated with terms targeted for removal is transferred to more-general terms higher up the tree, exploiting the transitive nature of relationships in the GO (Box 1). Additionally, entire branches of terms that are irrelevant for the organism of interest can be removed.

GO slims have been created for a variety of organisms and are listed on the GO Consortium website, but few of these are actively maintained by the consortium and annotated within the GO flat file that specifies the terms for the ontology. The consortium maintains a generic GO slim, as well as a Plant slim, a Yeast slim and slims for several other species. A large number of archived slims are available but are not actively maintained. One example of a recently created GO slim in an ecologically relevant species is that created for the eel *Anguilla anguilla* (Coppe *et al.* 2010).

Typically, slims are created either by a community resource (for example, the *Candida albicans* slim was developed by the *Candida* Genome Database), or as part of a targeted research effort (e.g. the Honeybee ESTs slim: Whitfield *et al.* 2002). Most slims are created in a manual process, using tools such as OBO Edit for support; however, automated and semi-automated methods are also available (Kusnierczyk 2008; Davis *et al.* 2010). Slims can be used to provide high-level summaries of the functions present in a given gene set and may be applied to reduce the number of hypotheses tested in enrichment tests (see Box 3).



Box 4 graph

The graph of the GO Cellular Component hierarchy and a cellular component GO slim for human proteins produced using the method of Davis *et al.* (2010), demonstrating how a GO slim reduces the number of terms and the complexity of the ontology, with a corresponding loss of more detailed terms.

queried against the chicken (*Gallus gallus domesticus*) and then the zebrafinch (*Taeniopygia guttata*) GenBank protein databases. GO terms were then assigned based on the closest match and further augmented using a procedure that infers biological processes from commonly associated molecular functions, thus increasing

Table 3 A nonexhaustive list of applications of the Gene Ontology in ecological and evolutionary contexts. Studies marked with an asterisk indicate those elaborated as examples in the text

Approach	Species	Molecular level	Description	Reference
Functional enrichment				
Life history/ development stage comparison	Ant (*Camponotus festinatus)*	RNA	Detected differences in gene expression between larval and adult ants	1
	Brown trout (*Salmo trutta*)	RNA	Identified enriched functional categories in trout with different life history strategies	2
	Grayling (*Thymallus thymallus*)	Protein	Identified enriched functional categories in two early life history stages (eyed egg and post hatch)	3
	Fire ant (*Solenopsis invicta*)	RNA	Identified enriched functional categories in virgin queens following orphaning	4
	Whitefish (*Coregonus spp.*)	RNA	Functional enrichment comparison of two whitefish ecotypes and their hybrids	5
	Brook charr (*Salvelinus fontinalis*)	RNA	Compared expression patterns of anadromous and resident charr in a common garden setting in muscle and gill tissue.	6
'Control vs. treatment'	Coral reef fish (*Pomacentrus moluccensis*)	RNA	Compared expression in thermal stress and control groups using a zebrafish microarray	7
	Native grass (*Andropogon gerardii*)	RNA	Stress response study (temperature and drought) using maize genomic resources	8
	Soil arthropod (*Folsomia candida*)	RNA	Sampled soils from different locations (dairy, forest, agriculture, natural grassland) and exposed laboratory-reared animals to the soil and assessed RNA expression	9
	Whitefish* (*Coregonus lavaretus*)	Protein	Compared proteomic expression of two whitefish ecotypes in salt- and freshwater	10
Postgenome duplication	Various fishes	DNA	Used GO to identify the functions of genes retained in duplicate following whole-genome duplication	11
	Rice (*Oryza sativa*)	DNA	Used GO to identify functions of genes retained following whole-genome duplication	12
	Yeast (*Saccharomyces* spp.)	DNA	Identified functional differences in retained gene duplicates in yeast strains from different environments	13
	Various plants*	DNA	Assessed whole-genome sequences in multiple plant species and used GO analysis to identify the functional classes of retained duplicated genes.	14
Positive selection	Mouse (*Mus musculus*)	Protein	GO used for functional categorization of rapidly evolving proteins in the mouse sperm proteome	15
	Seagrasses (*Posidonia oceanica* & *Zostera marina*)	RNA (ESTs)	GO used to identify functional enrichment in positively selected genes between sea-grasses and land plants	16
	Fungal pathogens (*Botrytis* spp. & *Sclerotinia sclerotiorum*)	DNA	GO used to identify the functions of positively selected genes	17
	Sheep (*Ovis aries*)	DNA	GO used to identify the functions of positively selected SNPs	18
Other	Eukaryotes	DNA	Examined whether horizontally transferred genes were enriched for specific functions	19
	Three-spined sticklebacks (*Gasterosteus aculeatus*)	RNA	Identified functionally enriched categories in genes expressed differentially in males and females	20
	Various plants	DNA	Identified functionally enriched genes common to broad phylogenetic plant lineages,	21

Table 3 *Continued*

Approach	Species	Molecular level	Description	Reference
			thus identifying key processes during plant evolution	
	Vertebrates	DNA	Identified functionally enriched genes in vertebrate lineages with common gains in regulatory elements	22
	Mussels (*Mytilus californianus*)	RNA	Compared RNA expression in mussels sampled from differing vertical shore locations in several populations	23
Genome/ transcriptome sequencing	Glanville fritillary (*Melitaea cinxia*)	RNA	GO used for functional characterization of the transcriptome	24
	Parasite *(Schistosoma japonicum)*	DNA	GO used for gene function classification	25
	Eel (*Anguilla anguilla*)	DNA	Also created a GO slim	26
	Sumatran and Bornean orangutans (*Pongo* spp.)	DNA	GO analysis indicated genes exhibiting positive selection enriched for vision genes	27
	Atlantic cod (*Gadus morhua*)	DNA	GO categories used to confirm gene content was similar to other fish species	28
	Fire ants *(S. invicta)*	DNA	GO analysis indicated that methylated genes of the newly sequenced genome are enriched for certain functions	29
Candidate identification	Honeybee (*Apis mellifera*)	DNA	Used GO to identify potential candidate genes in QTL regions identified in the study (based on predicted important functions)	30
Hypothesis testing	Red grouse* (*Lagopus lagopus scoticus*)	RNA	Used GO to identify and then examine genes with functions predicted to be important for two alternative sexual selection handicap model hypotheses.	31

1—Goodisman *et al*. (2005); 2—Giger *et al*. (2008); 3—Papakostas *et al*. (2010); 4—Wurm *et al*. (2010); 5—Renaut & Bernatchez (2011); 6—Boulet *et al*. (2012); 7—Kassahn *et al*. (2007); 8—Travers *et al*. (2010); 9—De Boer *et al*. (2011); 10—Papakostas *et al*. (2012); 11—Brunet *et al*. (2006); 12—Wu *et al*. (2008); 13—Ames *et al*.(2010); 14—Jiao *et al*. (2011); 15—Dorus *et al*. (2010); 16—Wissler *et al*. (2011); 17—Aguileta *et al*. (2012); 18—Kijas *et al*. (2012); 19—Kim & Caetano-Anolles (2010); 20—Leder *et al*. (2010); 21—Lee *et al*. (2011); 22—Lowe *et al*. (2011); 23—Place *et al*. (2012); 24—Vera *et al*. (2008); 25—Zhou *et al*. (2009); 26—Coppe *et al*. (2010); 27—Locke *et al*. (2011); 28—Star *et al*. (2011); 29—Wurm *et al*. (2011); 30—Oxley *et al*. (2010); 31—Wenzel *et al*. (2013).

the coverage of biological process annotations by up to 15% (Myhre *et al*. 2006). GO annotations could be assigned to just under a third (1864) of the transcripts. Of these, 282 were associated with the *immune system process* GO term, and 65 with the *oxidative stress/antioxidant activity* terms. Secondly, the response of transcript expression in these focal genes to experimentally increased testosterone levels was assessed in three different tissues and three experimental parasite treatment groups (anthelmintic treatment, natural chronic parasite infection and parasite challenge). The relative difference in transcriptomic response between natural and increased testosterone levels and the associated *P*-value for the null hypothesis of no differential response was assessed for both focal gene groups, and the false discovery rate (Benjamini & Hochberg 1995) was used to account for multiple testing and also to estimate the power of the microarray data to detect significant differences in expression (expected false discovery rate: eFDR). Possibly due to the large number of treatment comparisons and inclusion of individuals from several natural populations (which can result in unexplained environmental variation confounding interpretations), relatively few significant changes in transcript expression in the focal genes were observed. Given the low proportion of GO terms identified based on sequence similarity, an alternative approach would be to use a nonorthology-based method such as one based on the prediction of functional domains as implemented in, for example InterProScan (Tables 1 and 2). Nevertheless, some support for the ICHH was reported based on the results of one of the three tissues (caecum). The authors highlighted the issues of tissue choice and environmental context in their case study but emphasized the util-

ity of GO to shed light onto the physiological mode of action of handicap mechanisms.

GO enrichment and protein–protein interaction network analysis reveals divergent responses to salinity in two whitefish populations

Efficient osmoregulation is a vital physiological function in aquatic organisms, as it enables survival in environments with different salinity levels. Papakostas *et al.* (2012) studied the molecular basis of salinity tolerance in European whitefish (*Coregonus lavaretus*) by conducting a common garden experiment in which the fertilized eggs of two whitefish ecotypes, one freshwater spawning and one brackish-water spawning, were raised in salinities ranging from 0 ppt (freshwater) to 10 ppt (brackish-water). The molecular responses of hatchlings from both populations raised in the highest and lowest salinity levels were studied using a proteomics approach. About 1500 proteins were quantified using the Atlantic salmon proteins in the UniProt database as a search database for the sequenced peptides. To overcome the current poor annotation of Atlantic salmon proteins, GO terms for human orthologues were employed for functional analyses. A remarkable difference in molecular response to salinity change was observed between the freshwater and brackish-water populations. The brackish- and freshwater populations shared only six of the 115 proteins that changed expression levels in response to salinity; functional enrichment analysis based on GO annotations using BINGO (Maere *et al.* 2005) shed more light on the specific molecular mechanisms involved. Proteins with modified expression in freshwater whitefish were annotated with functions related to osmotic stress response, specifically cell volume regulation associated with calcium ion imbalance. In contrast, proteins with modified expression in brackish-water whitefish were annotated with functions known to be involved in routine salinity acclimation/adaptation, such as sodium ion transport. This analysis, combined with one that enables proteins to be placed into interaction networks (again, based on human orthologues), suggests that these whitefish ecotypes have adapted to their respective salinity environment despite background genome divergence being relatively low (microsatellite F_{ST} 0.049).

Potential pitfalls

Despite the popularity and wide use of GO-based analysis, there are a number of factors about which researchers should be wary, especially in comparisons with distantly related species. Some of the factors are specific to use of the GO in non-model organisms, while others are also relevant to the use of GO in general.

Incorrect orthologue identification

As highlighted previously, annotation using GO relies heavily on correct orthologue inference, so errors during this procedure can obviously lead to erroneous inference. Broadly speaking, three sorts of errors can arise in this process: (i) the sequence of interest may be incorrectly identified as the homologue of a database sequence during the search; (ii) the sequence may be correctly identified as a homologue but incorrectly as an orthologue, with the likely result that too-precise a GO term will be transferred from the target organism; and (iii) the sequence may be correctly identified as an orthologue, but its function has nonetheless diverged during the separate evolution of the two species. Although each of these risks can be mitigated, albeit not necessarily avoided, by limiting similarity searches to more closely related species, sometimes this may in fact do more harm than good due to the far superior annotation of the main genetic model species. So although fewer incorrect orthology assignments would be made if the genes of unstudied teleost fish were matched against zebrafish, quite probably this benefit would be outweighed by the (currently) lower-quality GO annotation of the zebrafish genome compared with that of human. Thus, for example, human *PRMT1*, protein arginine methyltransferase 1, in the NCBI Gene database is annotated with 12 GO terms for biological process (as of March 2013), whereas the same gene in zebrafish has only three biological process terms associated with it. Such discrepancies can influence downstream analyses. To illustrate this, we re-examined a set of 40 genes found to be up-regulated in female three spine stickleback using data from the study described by Leder *et al.* (2010). The zebrafish and human sequences matching the transcripts of these 40 stickleback genes were then identified using BLASTX. The highest (i.e. least-significant) *e*-value for human was 8.00E−11, while for zebrafish, it was 3.00E−30. The majority of the zebrafish and human orthologues (72.5%) had the same gene symbol, and most others had the same gene description linked to their Entrez identifiers. Using the species-specific Entrez identifiers associated with the best BLAST match (annotations as of 9/2012), typical downstream analyses were conducted using both DAVID Functional Annotation Clustering and ClueGo enrichment clustering in Cytotoscape (see Table 2 for tool details). In ClueGo, using the zebrafish Entrez identifiers, 26 functionally enriched GO terms for biological process were identified (all evidence codes), with these identifiers clustering into three groups; however, for the same 40 genes using human Entrez identifiers, 100 GO terms, almost four times as many, were identified and formed eight

functional groups. In DAVID, all three GO categories were used but a similar result was observed. Using the human Entrez identifiers, 17 functionally enriched clusters were observed, whereas only nine were observed using zebrafish identifiers. In both cases, the significance of the clusters and of the individual terms was, in general, higher using the human annotation. Due to the larger number of terms and the greater depth of the functional terms, more-specific functions were recovered using the human annotation. This is not to say that one data set is more correct, but rather that in this case, use of human annotations provided finer detail. As a caveat, the tissue examined in this case was liver, and many of the metabolic processes are conserved across taxa, and hence, gene function with humans may be more likely to be similar in more distant species in this example than if one were studying gene expression in gill tissue, or osmoregulation or other taxon-specific processes. This example does, however, highlight the importance of considering such factors when inferring orthologues and conducting enrichment tests.

Erroneous GO annotations

Problems can also arise if the sequence of interest and the database sequence are indeed orthologues, but the GO annotations for the latter are incorrect. Such problems are normally associated with the processes used to assign annotations in the GO database rather than with a particular study. Because of this, detecting such errors remains very challenging and is indeed an active field of study in genomics (Jones *et al.* 2007; Škunca *et al.* 2012). For molecular ecologists, the best way to get an idea of the reliability or confidence of the annotation for a particular gene is to examine the evidence code for each annotation. At a more-general level, it can be expected that a higher proportion of annotation errors will be observed in automatically assigned annotations (IEA evidence codes: Jones *et al.* 2007; Deegan *et al.* 2010). This is an important point to keep in mind, as the vast majority of GO annotations have been assigned using IEA evidence (Fig 1, Table S1, Supporting information). On the other hand, more-general IEA terms tend to be better-predicted (du Plessis *et al.* 2011). One solution to limit the effects of potentially incorrect annotations would be to exclude all automatically generated annotations (i.e. all annotations with IEA codes). The problem with this strategy is that in many cases, the number of annotated sequences will be significantly reduced. Therefore, researchers will be required to make a choice between retaining a higher number of genes with potentially lower confidence or using a reduced gene set.

To compare the consequences of these alternative strategies, we re-analysed a proteomic data set aimed at characterizing the proteome of a salmonid fish, European grayling (*Thymallus thymallus*), at the eyed egg and hatching stages of embryonic development (Papakostas *et al.* 2010). More specifically, we compared the results of an enrichment analysis obtained using all GO annotations (as was reported in the original article) to those obtained if only curated annotations were used (i.e. we excluded GO annotations with the evidence code IEA). Details of the results can be found in Appendix S2 (Supporting information). Briefly, at both developmental stages, the number of significantly enriched terms detected was more than double when all GO annotations were included in the analysis (141 vs. 55 and 158 vs. 72 for eyed-egg and hatch stages, respectively). This is probably the result of several factors including the greater number of terms available for analysis, as well as higher statistical power. However, while the majority (45 and 54, respectively) of the significantly enriched terms identified when IEA annotations were excluded were also identified when IEA annotations were included, 17 and 18 new significantly enriched terms were revealed by excluding IEA annotations. Analysis of the functional similarity of the terms identified as significant in the alternative analyses, as estimated by semantic similarity index (Du *et al.* 2009), indicated that both analyses identified terms with similar biological functions (semantic similarity 0.714–0.727), and therefore, the overall biological conclusions would be similar regardless of which data set was used. The same conclusion could be drawn when considering the lists of most significant GO terms, with the same GO terms commonly being found in the top five terms of both analyses. Therefore, in this example, it appears that similar biological conclusions would have been drawn regardless of whether IEA annotations were included or not.

Redundancy in lists of enriched terms

Due to term interdependency and multiple parent–child relationships in the GO DAG, several instances of parent terms may appear significant in enrichment tests simply because they include genes from multiple child terms (Masseroli & Pinciroli 2005; Supek *et al.* 2011). This kind of redundancy inflates enrichment lists and typically hampers summation of biological meaning (e.g. Fig. 3). There are several ways to deal with this problem. Tools like BINGO (Maere *et al.* 2005) or GORILLA (Eden *et al.* 2009) rely on visualization of the DAG (Fig. 3), while GO trimming (Jantzen *et al.* 2011) uses the parent–child relationships from the GO DAG to identify redundant terms. Other recent tools like REVIGO

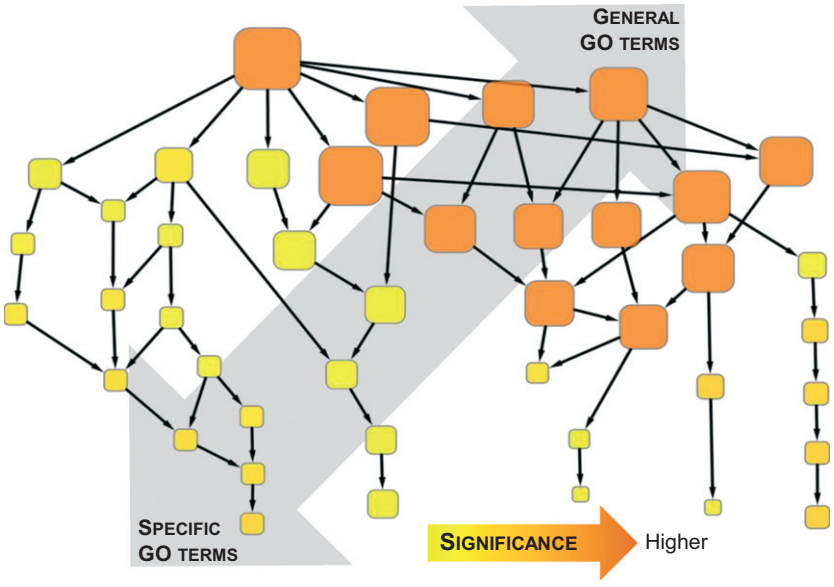


Fig. 3 An empirical example depicting how GO term interdependency influences the significance of enrichment. All terms were significantly enriched following a Benjamini–Hochberg FDR correction of $P < 0.05$, but higher level (also called 'parent', more general) terms were more significantly enriched because they include the genes from multiple lower level (also called 'child', more-specific) terms. Size of the nodes indicates more genes under the specific GO term. GO term names have been omitted for the sake of simplicity. The figure has been generated based on data from the study described by Papakostas *et al.* (2010).

(Supek *et al.* 2011) and RedundancyMiner (Zeeberg *et al.* 2011) use semantic similarity measures, that is, numerical values reflecting the closeness in meaning between GO terms (Pesquita *et al.* 2009). Another approach is the use of GO Slims (Davis *et al.* 2010), which are cut-down versions of GO, although these limit analysis to more-general terms (Box 4).

Terms with taxon restrictions

Under this description, GO includes any class assigned to gene functions specific for certain taxa. As of 5 October 2012, 463 GO terms were found in this category [http://www.geneontology.org/GO.doc.sensu.shtml, Table S1 (Supporting information)]. For instance, 'lactation' (GO: 0007595) and 'mammary gland development' (GO: 0030879) are specific to mammals, and 'CAM photosynthesis' (GO: 0009761) and 'root development' (GO: 0048364) to green plants. Specificity is defined with *only_in_taxon* or *never_in_taxon* arguments followed by the identifier of the taxonomic unit. Different collections of organisms have been assigned different taxon-restricted terms. These can be as general as 'cellular organisms' or 'Eukaryota', or as specific as 'Insecta' or 'Teleostei'. We note that GO class definitions remain more or less species-neutral, so one can be sure about the specificity of a particular class only by accessing this information.

Overlooking these restrictions can lead to errors and inconsistencies, especially by automated annotation pipelines. For example, GO terms related to photosynthesis have been detected in electronically annotated *Drosophila* data (Deegan *et al.* 2010). Could such discrepancies affect conclusions drawn from the cross-species use of GOs in high-throughput experiments? With minimum care, it is unlikely for taxon-restricted terms to be significantly enriched in the wrong species. In addition, we anticipate public databases to have taken this problem into consideration. For example, when searching the 90 901 annotations in the publically available zebrafish ontology file (*Danio rerio*.goa_zebrafish file as of 9 January 2013), we found only two cases of mis-annotation to the 16 mammalian *only_in_taxon* classes GO terms: mammary gland development (GO:0030879) and secondary neural tube formation (GO:0014021) were assigned to the genes *lef1* (UniProt Accession: Q9W7C0) and *scrib* (Q4H4B6) with IBA and IMP evidence codes, respectively. Perhaps a more-important issue is the information missed when transferring GO terms across taxa. Taxon-specific functions cannot be inferred from phylogenetically distant taxa, potentially resulting in a loss of important information. For example, there are 505 *only_in_taxon* Insecta and 1160 *only_in_taxon* Arthropoda annotations in *Drosophila* (as of 28 September 2012); these classes will be missed when non-model insect species are functionally annotated using data from orthologues of noninsect species.

Missing or incompletely annotated gene products

No ontology can ever be complete, so it is important to remember that the absence of an annotation does not mean the absence of function. Further, incompleteness of functional annotations may bias interpretation or enrichment tests, for example, by missing taxon-specific functions. For example, despite human having by far the most annotations, 17% of human genes still have no GO annotation at all (Fig. 1). For this reason, not only

Box 5

Best-practice guidelines for annotating a non-model organism in EEG research

The procedures listed here are aimed at enhancing the quality of GO annotation practices in non-model organisms. They are considerably more detailed than are currently used in ecological and evolutionary research. The benefits of adhering to these guidelines are, however, several fold. Firstly, they enhance the accuracy of downstream analyses of data sets such as functional enrichment tests to identify the processes, functions or locations of gene products, or exploring functions for the detection of expanded or missing gene families. Secondly, carefully generated and reported annotation becomes a valuable resource for the entire research community as well as enhancing the possibilities for future use of the data set by other researchers. The considerations relevant to the annotation of sequences from ecologically relevant non-model organisms are similar to those faced by all annotation groups. However, certain issues will be faced more frequently. The following guidelines are based on the GO Annotation Guidelines, Standard Operating Procedures (SOP), SOP for the GO reference genome project, and manuscript review guidelines provided to GO consortium (GOC) members. GO documentation is referenced where appropriate.

1 *Understand the Gene Ontology structure, conceptual coverage and application*: a good starting point is to review annotation guidelines used by consortium members (see http://www.geneontology.org/GO.annotation.shtml). Recognizing that individual research groups do not have the resources of large organismal databases, GO has established standard operating procedures (SOP) for small groups (http://www.geneontology.org/GO.annotation.SOP.shtml#small_lab) for a variety of annotation tasks, including annotating ESTs, genome sequence, microarray data sets or peptide sequences. Where possible, annotation should adhere to these standards and practices.

2 *Develop an annotation methodology*: consider the lines of evidence acceptable for GO annotation (see discussion of evidence codes) and annotation workflows in the SOP provided by GO (http://www.geneontology.org/GO.annotation.SOP.shtml) to select methods appropriate for the data and resources available:

   a Identify the type of gene products and associated sequence information that is to be used in annotation and review methods suitable for use with these data; for example, with proteins consider a method such as InterProScan (http://www.geneontology.org/GO.annotation.interproscan.shtml), which uses an InterProtoGO mapping (http://www.geneontology.org/GO.indices.shtml) to assign annotation, or another well-described method that has been used successfully elsewhere.

   b If using sequence-similarity-based functional transfer methods, adhere to the standards used by GOC members in annotation, which include:

      i Functional transfer should be made only between orthologues (see Box 2 for methods used to identify orthologues) except in exceptional circumstance (protein family information may sometimes be used);

      ii BLAST hits alone are not routinely used for functional transfer and should be complemented by other lines of evidence or replaced with stronger orthologue inference methods (see Box 2); and

      iii When transferring annotations, it may be necessary to assign a higher-level (i.e. less-specific) term, particularly when the annotation species has duplicate members of the gene product in question and functions may be partitioned (subfunctionalization) or subtly changed (neo-functionalization).

   c Consider annotation tools recommended for use by GOC members, for example, PAINT (Phylogenetic Annotation and Inference Tool: http://wiki.geneontology.org/index.php/PAINT), which is made available with curation guidelines (http://wiki.geneontology.org/index.php/PAINT_User_Guide#Curation_Guidelines) and a SOP (http://wiki.geneontology.org/index.php/PAINT_SOP).

3 *Generate annotation for all known/predicted gene products in the new species of interest*

   a Precisely record all relevant tools, parameters, data sources, data versions and availability; annotate this set as comprehensively as possible because the validity and usefulness of downstream analysis, for example, GO enrichment tests, depends on annotation coverage. Annotations of *no information* as described below are nonetheless useful, that is, about the lack of knowledge.

   b Assign the correct evidence codes; if annotation is electronically generated, and not reviewed, IEA should be used.

Box 5 *Continued*

c Where no annotation can be established, annotate the gene product with the relevant root terms (molecular function GO:0003674, biological process GO:0008150, or cellular component GO:0005575) which indicate that no knowledge is available about a gene product in that part of GO; assign the ND (No biological Data available) evidence code.

d Format the annotations using a GOC standard format (http://www.geneontology.org/GO.format.annotation.shtml) which includes:

i Evidence codes;

ii A reference to the experimental method used to generate the annotation;

iii Annotation provenance: with certain kinds of evidence, the WITH/FROM field should be populated to indicate the species of origin of the inferred annotation (see http://www.geneontology.org/GO.evidence.shtml#withUsage). This may be a particularly important issue in ecologically relevant species.

4 *Make annotations available to the research community*, either via publications, or preferably by submission to the Gene Ontology; where no database has been established to manage GO annotations for a species (as is likely to be the case in EEG), groups can contribute annotations to the central repository via the UniProtKB GO Annotation (UniProtKB-GOA) multispecies annotation group (see http://www.geneontology.org/GO.annotation.shtml#single). Currently, submissions to the repository must be agreed with the annotation group by contacting them directly in advance for instructions (see http://www.ebi.ac.uk/GOA/contribute.html).

5 *Adhere to minimum reporting standards* in the preparation of manuscripts describing GO-based work (see Box 6).

ecologists and evolutionary biologists, but also human genetics researchers could benefit from an increased effort in the ecological and evolutionary genomics (EEG) field to report functional annotation information in the GO.

Recommendations for conducting a sound study using GO and minimum reporting standards

The variety of research outlined in Table 3 clearly indicates the versatility of approaches by which functional annotations can be assigned to nonannotated genomes using the GO database, and the information subsequently applied to infer gene functions in ecologically relevant species. However, for the time being at least, molecular ecologists will be faced with important decisions regarding trade-offs between quantity and the potential quality of the assigned annotations. Unfortunately, perhaps not surprisingly, no solution will be suitable in all cases. For this reason, we recommend that researchers familiarize themselves with the methodologies behind 'all-in-one' tools such as BLAST2GO, as careful adjustment of parameters and/or use of

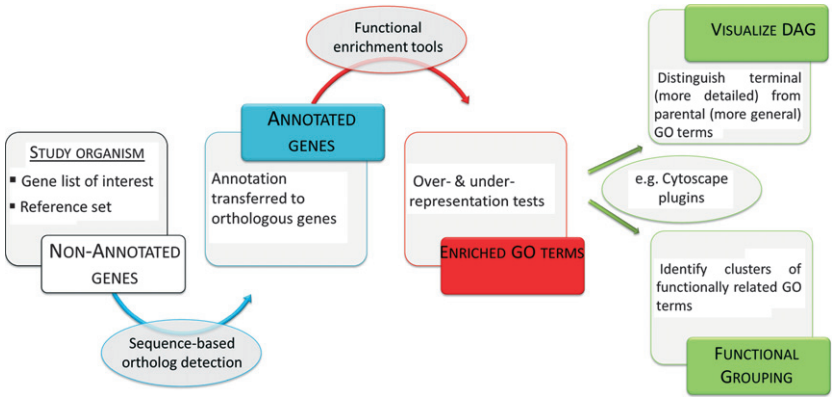


Fig. 4 Workflow for performing functional (GO term) enrichment analysis of a gene list of interest in a nonannotated organism. The first step is to retrieve annotations from putative orthologues of a well-annotated genome (blue). Then, enrichment analysis is conducted as in any organism (red). Finally, functional inference based on the enriched GO terms largely depends on the biological questions been asked. Typically, visualization or clustering approaches can greatly reduce redundancy and help describe large lists of enriched GO terms in a concise manner (green). Best practices for each of these phases are outlined in the review.

Box 6
Minimal information reporting guidelines in Gene Ontology experiments

1 *Generating new annotations for a non-model organism*

a Clearly describe the technique (i.e. annotation transfer from annotated orthologues, annotation generated from functional analysis tool (such as InterPro), direct experimentation) used to assign GO terms to the un-annotated organism;
b For sequence-similarity-based methods, the thresholds applied to e-value, identity, bit score and alignment coverage must be reported;
c For nonsequence similarity–based methods, describe the annotation process and report all tools and parameter settings;
d In the case of transferring annotation, the source of the existing annotations used for the study should be clearly described, and the description should include the database version and access date, and version of the Gene Ontology present in the annotation data;
e The source of the non-model organism data should be clearly described, including the database version and access date if applicable;
f Where data sets are not archived by online sources, provide relevant data sets as supplementary data so future researchers can reproduce analyses and verify results, and if possible assign a persistent identifier for the data.

2 *Reporting inferred GO terms*

a Make the newly inferred annotations available to the wider research community in an accepted standard format (i.e. GO annotation format http://www.geneontology.org/GO.format.annotation.shtml);
b Assign accurate evidence codes (see Box 5);
c Use unique, searchable and appropriate identifiers for all molecules and do not use gene names as unique identifiers;

3 *Enrichment tests: Statistical methods and multiple test corrections*

a Provide a clear statement of the hypothesis and null hypothesis;
b Describe and reference the statistical test used for the enrichment tests and clearly describe and reference the type of multiple test correction applied;
c Clearly describe the background annotation data used in the enrichment test and provide these data if they are not already publicly available;
d Provide query lists;
e Report all significant results.

complementary tools can provide more robust results. In Box 5, we provide some general guidelines for functionally annotating non-model organism gene products and in Fig. 4, we summarize the workflow for performing a functional enrichment analysis of a gene list of interest in a nonannotated organism.

In addition to the technical requirements for a sound study listed previously, certain minimum information reporting standards should be followed in manuscripts using GO with largely un-annotated genomes. Adherence to these reporting criteria will enable research groups to critically assess the experimental procedures used by others, facilitate progress in the field as researchers become more aware of detailed computational parameters used in studies and ensure reproducibility of experiments. In Box 6, we outline a set of minimum reporting guidelines for the three important phases of GO experiments illustrated in Fig. 4.

Recommendations for the EEG field in general

Given the amazing amounts of data that are being generated by NGS and high-throughput proteomics, some effort is required by the molecular ecology community to improve the knowledge base and make this information reliable and accessible. For example, a curated database of non-model organism biological processes, molecular function and cellular components that incor-

porates information from multiple experiments across disciplines would greatly improve the reliability of orthology and function. Having information from microarrays, RNA-Seq, shotgun proteomics, and Western blots on specific tissue expression, changes due to specific treatments or other experimental evidence concerning a specific transcript or protein would greatly enhance the prediction of function. In fact, many of these data already exist in public databases such as Array Express (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) and in the underlying literature. For example, in the study of whitefish salinity response described previously, significant responses were identified from electronically inferred GO terms in the human database. However, as reported in the original article, considerable experimental support for the inferred responses was already available (reviewed by Hwang *et al.* 2011). Other examples of convincing experimental evidence being available for ecologically relevant species (threespine sticklebacks, *Gasterosteus aculeatus*), but this information being absent from the GO, include the roles of the *EDA* and *pit-1* genes in armory and pelvic girdle development (Colosimo *et al.* 2005; Chan *et al.* 2010) as well as the role of the *spiggin* protein in nest building (Jones *et al.* 2001). To date, such existing information is rarely included in the GO, and thus, an active effort to further categorize and curate available experimental gene function information in a broader range of species is required. We strongly encourage EEG researchers to make the effort to report the results of relevant experiments, in the form of GO annotations, to the GO (as described in Box 5).

There is no doubt that the GO has changed the way in which researchers interpret the results of high-throughput experiments in molecular genetics. However, an ontology designed to describe the processes, functions and locations of gene products will not capture many domain-specific concepts that are of interest in disciplines adopting new genomic technologies. The argument has been made elsewhere that the EEG community needs an ontology that can capture important domain-specific concepts (Pavey *et al.* 2012). Although we feel that to some extent, Pavey *et al.* undervalued the important role that the GO can play in EEG, we agree wholeheartedly that an ecology and evolution ontology that could be used alongside the GO would be highly valuable. Here, we offer some recommendations for such an effort so that the community can extract maximum benefit from the valuable development and annotation effort already present in the GO.

To ensure compatibility with GO, a new EEG ontology should be generated in adherence with the principles set out by the Open Biological and Biomedical Ontology (OBO) Foundry, including use of the OBO Format. Other ontology languages and formats exist, but the value of the GO in genomic science is a strong argument for using OBO Foundry-compliant design principles and formats. Other ontologies are available in this format and could be considered for integration with an EEG ontology. For example, the Environmental Ontology (EVO: http://www.environmentontology.org/home/about-envo) and the ontologies maintained by Gramene Ontologies (http://www.gramene.org/plant_ontology/index.html#eo) may contain elements relevant to the definition of terms in an ontology for EEG.

There is no need for an EEG ontology to be completely separate from, or work in parallel with, the GO or other ontologies. Compatible ontologies can be integrated using *cross products* to establish formal definitions for terms and add power and specificity to the concepts so defined. The GO website provides an example of a term that is defined as the cross product of two terms, one from the GO and one from the Cell Ontology (Bard *et al.*, 2005) (see GO website http://geneontology.org/GO.ontology.structure.shtml#xp for details).

Finally, the success of any ontology is measured by its adoption by the research community. An ontology must therefore be useful, logical, easy to use and critically supported during on-going phases of extension and development. Further, it should represent a commitment by a research community to adopt specific, formal terms and definitions for the concepts critical to that domain. As such, community involvement in the design and development of an ontology is vital to its eventual success, as is the participation of ontological engineers and members of related ontology development projects, and in this case, especially the GO. We therefore echo the call of Pavey *et al.* (2012) and strongly urge members of the EEG community to get organized and participate in the development of such an ontology.

Acknowledgements

This review was initiated while CRP was a visiting researcher at the University of Queensland (funded by the Finnish Academy, grants 137710, 141231). MAR's research is supported by ARC, NHMRC and the J.S. McDonnell Foundation and SP and EHL are supported by the Finnish Academy. We thank Matthieu Bruneaux and Shihab Hasan for their help with bioinformatics analyses and three anonymous reviewers and the review editor for extremely constructive comments on an earlier version of the manuscript.

References

Addou S, Rentzsch R, Lee D, Orengo CA (2009) Domain-based and family-specific sequence identity thresholds increase the

levels of reliable protein function transfer. *Journal of Molecular Biology*, **387**, 416–430.

Aguileta G, Lengelle J, Chiapello H *et al.* (2012) Genes under positive selection in a model plant pathogenic fungus, Botrytis. *Infection, Genetics and Evolution*, **12**, 987–996.

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, **5**, e1000262.

Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. *Methods in Molecular Biology*, **855**, 259–279.

Altenhoff AM, Studer RA, Robinson-Rochavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, **8**, e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Ames RM, Rash BM, Hentges KE *et al.* (2010) Gene duplication and environmental adaptation within yeast populations. *Genome Biology and Evolution*, **2**, 591–601.

Arya GH, Weber AL, Wang P *et al.* (2010) Natural variation, functional pleiotropy and transcriptional contexts of odorant binding protein genes in *Drosophila melanogaster*. *Genetics*, **186**, 1475–1485.

Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. *Genome Biology*, **6**, R21.

Bauer-Mehren A, Furlong LI, Sanz F (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, **5**, 290.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bindea G, Mlecnik B, Hackl H *et al.* (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

Binns D, Dimmer E, Huntley R *et al.* (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.

Bittner T, Donnelly M (2007) Logical properties of foundational relations in bio-ontologies. *Artificial Intelligence in Medicine*, **39**, 197–216.

Bluthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Informatics*, **16**, 106–115.

Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessmoz C (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Briefings in Bioinformatics*, **12**, 423–435.

Boulet M, Normandeu E, Bougas B, Audet C, Bernatcez L (2012) Comparative transcriptomics of anadromous and resident brook charr *Salvelinus fontinalis* before their first salt water transition *Current*. *Zoology*, **58**, 155–167.

Bourret V, Kent MP, Primmer CR *et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology*, **22**, 532–551.

Bruneaux M, Johnston S, Herczeg G, Merilä J, Primmer CR, Vasemägi A (2013) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified RAD tag approach. *Molecular Ecology*, **22**, 565–582.

Brunet FG, Roest Crollius H, Paris M *et al.* (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, **23**, 1808–1816.

Burge S, Kelly E, Lonsdale D *et al.* (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database-The Journal of Biological Databases and Curation*, **2012**, bar068.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*, **327**, 302–305.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**, D363–D368.

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, e383.

Colbourne JK, Pfrender ME, Gilbert D *et al.* (2011) The ecoresponsive genome of Daphnia pulex. *Science*, **331**, 555–561.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science*, **307**, 1928–1933.

Conesa A, Götz S, Garcia-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Coppe A, Pujolar JM, Maes GE *et al.* (2010) Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*, **11**, 635.

Danchin EGJ, Levasseur A, Rascol VL, Gouret P, Pontarotti P (2007) The use of evolutionary biology concepts for genome annotation. *Journal of Experimental Zoology (Molecular Development and Evolution)*, **308B**, 26–36.

Davis MJ, Sehgal MSB, Ragan MA (2010) Automatic, context-specific generation of Gene Ontology slims. *BMC Bioinformatics*, **11**, 498.

De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, **20**, 591–597.

De Boer TE, Birlutiu A, Bochdanovits Z *et al.* (2011) Transcriptional plasticity of a soil arthropod across different ecological conditions. *Molecular Ecology*, **20**, 1144–1154.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Deegan JI, Dimmer EC, Mungall CJ (2010) Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics*, **11**, 530.

Dennis G, Sherman BT, Hosack DA *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, **4**, R60.

Diz AP, Martinez-Fernandez M, Rolan-Alvarez E (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology*, **21**, 1060–1080.

Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL (2010) Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Molecular Biology and Evolution*, **27**, 1235–1246.

Du Z, Li L, Chen C-C, Yu PS, Wang JZ (2009) G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, **37**, W345–W347.

Du Z, Zhou X, Ling Y, Zhang ZH, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, **38**, W64–W70.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Edwards AC, Ayroles JF, Stone EA *et al.* (2009) A transcriptional network associated with natural variation in Drosophila aggressive behavior. *Genome Biology*, **10**, R76.

Falda M, Toppo S, Pescarolo A *et al.* (2012) Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics*, **13** (Suppl 4), S14.

Farcomeni A (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347–388.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**, 99–113.

Fitch WM (1973) Aspects of molecular evolution. *Annual Reviews of Genetics*, **7**, 343–380.

Forné I, Abián J, Cerdà J (2010) Fish proteome analysis: Model organisms and non-sequenced species. *Proteomics*, **10**, 858–872.

Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, **12**, 449–462.

Gentleman RC, Carey VJ, Bates DM *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Giger T, Excoffier L, Amstutz U *et al.* (2008) Population transcriptomics of life-history variation in the genus Salmo. *Molecular Ecology*, **17**, 3095–3108.

Goodisman MAD, Isoe J, Wheeler DE, Wells MA (2005) Evolution of insect metamorphosis : a microarray-based study of larval and adult gene expression in the ant Camponotus festinatus. *Evolution*, **59**, 858–870.

Gruber T (1993) Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, **43**, 907–928.

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in three-spined stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP *et al.* (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, **39**, D556–D560.

Hwang P-P, Lee T-H, Lin L-Y (2011) Ion regulation in fish gills: recent progress in the cellular and molecular mechanisms. *American Journal of Physiology*, **301**, R28–R47.

Jantzen SG, Sutherland BJ, Minkley DR, Koop BF (2011) GO Trimming: systematically reducing redundancy in large Gene Ontology datasets. *BMC Research Notes*, **4**, 267.

Ji P, Liu G, Xu J, Wang X, Li J *et al.* (2012) Characterization of common carp transcriptome: sequencing, *de novo* assembly, annotation and comparative genomics. *PLoS ONE*, **7**, e35152.

Jiao Y, Wickett NJ, Ayyampalayam S *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.

Jones I, Lindberg C, Jakobsson S *et al.* (2001) Molecular cloning and characterization of spiggin. An androgen-regulated extraorganismal adhesive with structural similarities to von Willebrand Factor-related proteins. *The Journal of Biological Chemistry*, **276**, 17857–17863.

Jones CE, Brown AL, Baumann U (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, **8**, 170.

Kassahn KS, Caley MJ, Ward AC *et al.* (2007) Heterologous microarray experiments used to identify the early gene response to heat stress in a coral reef fish. *Molecular Ecology*, **16**, 1749–1763.

Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, **8**, e1002375.

Kijas JW, Lenstra JA, Hayes B *et al.* (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology*, **10**, e1001258.

Kim KM, Caetano-Anollés G (2010) Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Molecular Biology and Evolution*, **27**, 1710–1733.

Koonin EV, Galperin MY (2003) *Sequence—Evolution—Function: Computational Approaches in Comparative Genomics*. Kluwer Academic, Boston.

Koonin EV, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. *Trends in Genetics*, **12**, 334–336.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for gene orthology inference. *Briefings in Bioinformatics*, **12**, 379–391.

Kusnierczyk W (2008) Taxonomy-based partitioning of the Gene Ontology. *Journal of Biomedical Informatics*, **41**, 282–292.

Landry CR, Townsend JP, Hartl DL, Cavalieri D (2006) Ecological and evolutionary genomics of *Saccharomyces cerevisiae*. *Molecular Ecology*, **15**, 575–591.

Leder EH, Cano JM, Leinonen T *et al.* (2010) Female-biased expression on the X chromosome as a key step in sex chromosome evolution in three spine sticklebacks. *Molecular Biology and Evolution*, **27**, 1495–1503.

Lee C-R, Mitchell-Olds T (2012) Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome. *Molecular Biology and Evolution*, **29**, 3721–3728.

Lee EK, Cibrian-Jaramillo A, Kolokotronis S-O *et al.* (2011) A functional phylogenomic view of the seed plants. *PLoS Genetics*, **7**, e1002411.

Leskinen PK, Laaksonen T, Ruuskanen S, Primmer CR, Leder EH (2012) The proteomics of feather development in pied

flycatchers (*Ficedula hypoleuca*) with different plumage coloration. *Molecular Ecology*, **21**, 5762–5777.

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.

Locke DP, Hillier LW, Warren WC *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.

Louie B, Higdon R, Kolker E (2009) A statistical model of protein sequence similarity and function similarity reveals overly-specific function predictions. *PLoS ONE*, **4**, e7546.

Lowe CB, Kellis M, Siepel A *et al.* (2011) Three periods of regulatory innovation during vertebrate evolution. *Science*, **333**, 1019–1024.

Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Masseroli M, Pinciroli F (2005) Using Gene Ontology and genomic controlled vocabularies to analyze high-throughput gene lists: three tool comparison. *Computers in Biology and Medicine*, **36**, 731–747.

McHale LK, Haun WJ, Xu WW *et al.* (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, **159**, 1295–1308.

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences USA*, **93**, 10268–10273.

Myhre S, Tveit H, Mollestad T, Laegreid A (2006) Additional gene ontology structure for improved biological reasoning. *Bioinformatics*, **22**, 2020–2027.

Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, **7**, e1002073.

Normandeau E, Hutchings JA, Fraser DJ, Bernatchez L (2009) Population-specific gene expression responses to hybridization between farm and wild Atlantic salmon. *Evolutionary Applications*, **2**, 489–503.

Östlund G, Schmitt T, Forslund K *et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, **38**, D196–D203.

Oxley PR, Spivak M, Oldroyd BP (2010) Six quantitative trait loci influence task thresholds for hygienic behaviour in honeybees (*Apis mellifera*). *Molecular Ecology*, **19**, 1452–1461.

Papakostas S, Vollestad LA, Primmer CR, Leder EH (2010) Proteomic profiling of early life stages of European grayling (*Thymallus thymallus*). *Journal of Proteome Research*, **9**, 4790–4800.

Papakostas S, Vasemägi A, Vähä J-P, Himberg M, Peil L, Primmer CR (2012) A proteomics approach reveals divergent molecular responses to salinity in populations of European whitefish (*Coregonus lavaretus*). *Molecular Ecology*, **21**, 3516–3530.

Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology & Evolution*, **27**, 673–678.

Pennisi E (2009) Ecological genomics gets down to genes–and function. *Science*, **326**, 1620–1621.

Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, **5**, e1000443.

Place SP, Menge B, Hofmann GE (2012) Transcriptome profiles link environmental variation and physiological response of *Mytilus californianus* between Pacific tides. *Functional Ecology*, **26**, 144–155.

du Plessis L, Škunca N, Dessimoz C (2011) The what, where, how and why of gene ontology-a primer for bioinformaticians. *Briefings in Bioinformatics*, **12**, 723–735.

Quevillon E, Silventoinen V, Pillai S *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Research*, **33**, W116–W120.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, **35**, W193–W200.

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**, 1041–1052.

Renaut S, Bernatchez L (2011) Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Heredity*, **106**, 1003–1011.

Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, **9**, 509–515.

Rise ML, von Schalburg KR, Brown GD *et al.* (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*, **14**, 478–490.

Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6239–6244.

Rogers MF, Ben-Hur A (2009) The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics*, **25**, 1173–1177.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Genetics*, **24**, 192–200.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

Ruan J, Li H, Chen Z *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Research*, **36**, D735–D740.

Sayers EW, Barrett T, Benson DA *et al.* (2010) Database resources of the National Centre for Biotechnology Information. *Nucleic Acids Research*, **38**, D5–D16.

Shannon P, Markiel A, Ozier O *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

Shin CJ, Davis MJ, Ragan MA (2009) Towards the mammalian interactome: inference of a core mammalian interaction set in mouse. *Proteomics*, **9**, 5256–5266.

Škunca N, Altenhoff A, Dessimoz C (2012) Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, **8**, e1002533.

Smith B, Ashburner M, Rosse C *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25**, 1251–1255.

Star B, Nederbragt AJ, Jentoft S *et al.* (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**, 207–210.

Supek F, Bosnjak M, Škunca N, Smuc T (2011) REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE*, **6**, e21800.

Tadiso T, Krasnov A, Skugor S *et al.* (2011) Gene expression analyses of immune responses in Atlantic salmon during early stages of infection by salmon louse (*Lepeophtheirus salmonis*) revealed bi-phasic responses coinciding with the copepod-chalimus transition. *BMC Genomics*, **12**, 141.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov RL, Natale DA, Garkavtsev IV *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, **29**, 22–28.

Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

The Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's reference genome project: a unified framework for functional annotation across species. *PLoS Computational Biology*, **5**, e1000431.

Thissen D, Steinberg L, Kuang D (2002) Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in mutiple comparisons. *Journal of Educational and Behavioral Statistics*, **27**, 77–83.

Thomas PD, Mi H, Lewis S (2007) Ontology annotation: mapping genomic regions to biological function. *Current Opinion in Chemical Biology*, **11**, 4–11.

Trachana K, Larsson TA, Powell S *et al.* (2011) Orthology prediction methods: a quality assessment using curated protein families. *BioEssays*, **33**, 769–780.

Travers SE, Tang Z, Caragea D *et al.* (2010) Variation in gene expression of *Andropogon gerardii* in response to altered environmental conditions associated with climate change. *Journal of Ecology*, **98**, 374–383.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wall DP, Deluca T (2007) Ortholog detection using the reciprocal smallest distance algorithm. *Methods in Molecular Biology*, **396**, 95–110.

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i558.

Weckwerth W (2011) Green systems biology—from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of Proteomics*, **75**, 284–305.

Wenzel MA, Webster LMI, Paterson S *et al.* (2013) A transcriptomic investigation of handicap models in sexual selection. *Behavioral Ecology and Sociobiology*, **67**, 221–234.

Whitfield CW, Band MR, Bonaldo MF *et al.* (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behaviour in the honey bee. *Genome Research*, **12**, 555–566.

Wissler L, Codoñer FM, Gu J *et al.* (2011) Back to the sea twice: identifying candidate plant genes for molecular evolution to marine life. *BMC Evolutionary Biology*, **11**, 8.

Wu Y, Zhu Z, Ma L, Chen M (2008) The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. *Molecular Biology and Evolution*, **25**, 1003–1006.

Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP, Hautaniemi S (2009) Integrated network analysis platform for protein-protein interactions. *Nature Methods*, **6**, 75–77.

Wurm Y, Wang J, Keller L (2010) Changes in reproductive roles are associated with changes in gene expression in fire ant queens. *Molecular Ecology*, **19**, 1200–1211.

Wurm Y, Wang J, Riba-Grognuz O *et al.* . (2011) The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 5679–5684.

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, **13**, 329–342.

Zeeberg B, Liu H, Kahn A, Ehler M *et al.* (2011) RedundancyMiner: de-replication of redundant GO categories in microarray and proteomics analysis. *BMC Bioinformatics*, **12**, 52.

Zhou Y, Zheng H, Chen Y *et al.* (2009) The *Schistosoma japonicum* genome reveals features of host–parasite interplay. *Nature*, **460**, 345–351.

C.R.P., S.P. and E.L. share a common interest in understanding the molecular basis of traits of ecological and evolutionary importance in aquatic organisms. M.A.R. and M.J.D. are interested in development and application of bioinformatic approaches for studying topics including phylogenomics, lateral genetic transfer and cancer biomolecular networks as well as the use of ontology in biological systems modelling.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Gene ontology evidence codes and their frequency (as of 1.02.2013) in some 'traditional model' and 'emerging genomic model' organisms.

Table S2 Summary of taxon-specific GO terms.

Table S3 Results and data associated with Appendix S2.

Table S4 Results and data associated with Appendix S2.

Appendix S1 Orthology and related GO evidence codes.

Appendix S2 Enrichment test differences when using ALL vs. non-IEA evidence codes: a case study using a previously published data set.