vHOG, a multispecies vertebrate ontology of homologous organs groups

Anne Niknejad1,2, Aurélie Comte1,2, Gilles Parmentier1,2, Julien Roux1,2,†, Frederic B. Bastian1,2,* and Marc Robinson-Rechavi1,2

1Department of Ecology and Evolution, Biophore, University of Lausanne and 2Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

1INTRODUCTION

One of the main approaches to understand biological objects has long been comparative studies, from comparative anatomy in the 18th century to comparative genomics in the last decade. Comparative analysis can notably help identify adaptation, as well as functional or structural constraints (Harvey and Pagel, 1991).

Many of the data which we would like to compare, such as gene expression or phenotypes, need to be mapped to anatomy and development of organisms to be of use. To facilitate the automatic manipulation of this data, there has been an important effort to build ontologies, which describe the anatomy of human and of animal model organisms (Bard, 2008). These ontologies have tended to imply homology (Dahdul et al., 2010), which according to the principle of reality followed by the OBO Foundry should not be included within an ontology (Smith and Ceusters, 2010) (although see Merrill, 2010).

Some other efforts follow the same model as the TAO, and use general (higher level) terms from the Common Anatomy Reference Ontology (CARO) (Haendel et al., 2008), which according to the principle of reality used by the OBO Foundry should not be included within an ontology (Smith and Ceusters, 2010) (although see Merrill, 2010). The CARO was created to provide a common basis for all future anatomy ontologies and facilitate their interoperability.

Fourth, homologous structures can diverge in function or structure, to an extent that representing them together in an ontology might be difficult. Finally, there can be differences between species in the relationships among structures (Haendel et al., 2008).

There are several ongoing efforts to create multispecies ontologies for animal groups, which have chosen different answers to the challenges outlined above. The Teleost Anatomy Ontology (TAO) (Dahdul et al., 2010) is a multispecies ontology for teleost fishes. The TAO is based on the Zebrafish Anatomy Ontology (ZFA) (from the ZFIN database, Bradford et al., 2011), and uses general (higher level) terms from the Common Anatomy Reference Ontology (CARO) (Haendel et al., 2008). The CARO was created to provide a common basis for all future anatomy ontologies and facilitate their interoperability. Several other efforts follow the same model as the TAO, and include the Amphibian Anatomy Ontology and the Hymenoptera Anatomy Ontology. In each of these cases, there is an effort to describe the morphological diversity of the clade. Consequently, each ontology will include terms that are found in several species of the clade. The use of a term for several species in the TAO does not imply homology (Dahdul et al., 2010). A species-specific ontology (e.g. ZFA) is considered a subset of the multispecies ontology (e.g. TAO).
A different approach is taken by the Uberon project (Mungall et al., 2012), which maps terms from several animal anatomy ontologies, but also anatomy-related terms from the Gene Ontology or medical ontologies, and other multispecies ontologies such as the TAO or our organ groups (Bastian et al., 2008). Uberon aims to provide anatomical information in relation to the Gene Ontology and the Cell Ontology, and is neutral relative to the criterion of homology (C. Mungall, personal communication). Uberon thus groups structures based on a criterion of similarity (Dahdul et al., 2010; C. Mungall, personal communication) that is the parent concept of homology, and also of homoplasie or of functional similarity (Roux and Robinson-Rechavi, 2010).

Despite the issues raised by the use of the criterion of homology, we feel that it also presents important advantages. First, restricting to one criterion allows a clear interpretation of the ontology when used in a database; it especially allows a clear use of automatic reasoning. Second, homology is transitive, which allows us to form an ontology of ‘organisms groups’, rather than encode all pairwise relations between terms. Finally, it is the one criterion that permits correct formulation of hypotheses of adaptation or constraints in evolution (Harvey and Pagel, 1991).

Our software Homol onto to align ontologies (Parmentier et al., 2010) generates a multispecies ontology of Homologous Organs Groups (HOGs). These HOGs are used in our database of gene expression evolution, Bgee (Bastian et al., 2008), as well as in Uberon. Mappings in the HOGs are restricted to manually curated relations of homology. We use a strict definition of historical homology: ‘Homology that is defined by common descent’ (HOM:0000007, Roux and Robinson-Rechavi, 2010).

The homology in Bgee has allowed the comparison of expression patterns between species in several applications, such as the characterization of gene interactions in development (Comte et al., 2010), the study of orthologs and paralogs (Huerta-Cepas et al., 2011) (Bastian, F.B. et al., unpublished data) or the study of microRNA evolution (Roux et al., unpublished data).

The HOGs used in Bgee are part of the database schema, are constrained according to the database optimization, and are not formatted for easy external use. Yet, it is also desirable to provide an ontology which is optimized for inter-operability and reuse by the community. Thus, we present here a CARO-compliant version of the HOG ontology, with all terms and relations carefully curated. The homology in Bgee has allowed the comparison of expression patterns between species in several applications, such as the characterization of gene interactions in development (Comte et al., 2010), the study of orthologs and paralogs (Huerta-Cepas et al., 2011) (Bastian, F.B. et al., unpublished data) or the study of microRNA evolution (Roux et al., unpublished data).

The HOGs used in Bgee are part of the database schema, are constrained according to the database optimization, and are not formatted for easy external use. Yet, it is also desirable to provide an ontology which is optimized for inter-operability and reuse by the community. Thus, we present here a CARO-compliant version of the HOG ontology, with all terms and relations carefully curated. The homology in Bgee has allowed the comparison of expression patterns between species in several applications, such as the characterization of gene interactions in development (Comte et al., 2010), the study of orthologs and paralogs (Huerta-Cepas et al., 2011) (Bastian, F.B. et al., unpublished data) or the study of microRNA evolution (Roux et al., unpublished data).

The homology in Bgee has allowed the comparison of expression patterns between species in several applications, such as the characterization of gene interactions in development (Comte et al., 2010), the study of orthologs and paralogs (Huerta-Cepas et al., 2011) (Bastian, F.B. et al., unpublished data) or the study of microRNA evolution (Roux et al., unpublished data).

The vHOG ontology version 1.4 (December 2011) has 1184 terms, 306 is_a relations and 1181 part_of relations. There are 664 synonyms: 547 terms have definitions; 67% of terms have only one parent, 25% have two parents, and the others have three or more. There are 493 cross-references (xref) to other multispecies OBO ontologies. These do not include the mappings to species-specific ontologies, as we consider that mappings from multispecies ontologies to species-specific ontologies should not be treated as xrefs, but should be encoded separately in an association file, similar to the Gene Ontology Annotation mapping (Barrell et al., 2009). This differs from the approach of Uberon, which includes cross-references to the source ontologies in the multispecies ontology itself (Dahdul et al., 2010).

All mappings of terms from the species-specific ontologies to the vHOG terms were manually curated. There are 5129 terms from species-specific ontologies mapped to 1189 vHOG terms, which represent 2259 hypotheses of homology between vertebrates (i.e. 2259 pair-wise relations between structures in different species). There are 15 vHOG terms with no mappings, which correspond to higher level CARO terms (e.g. ‘material anatomical entity’). The semantics of the mapping is ‘treat-xrefs-as-equivalent’ for CARO, and equivalent to ‘treat-xref-as-genus-differentia’ for the mapping to species-specific ontologies.

The mappings were annotated with ‘support codes’ to provide confidence information (Table 1). Means to provide confidence metadata information in support of annotations are currently being discussed in the framework of the Evidence Code Ontology (ECO) (Gene Ontology Consortium, 2010) and of the International Society for Biocuration. Our objective is to rely on the use of the ECO, once a standard to address the issue of confidence has emerged in the biocuration and ontology communities.

In practice, the support codes ‘obvious’ and ‘inferred’ are essentially used for homology between the two mammals considered in our ontology: 97% of all mappings with the code ‘inferred’ are for human and mouse. All mappings with the code ‘obvious’ are for human and mouse, except ‘whole organism’ which has the code ‘obvious’ for all ontologies.

Among the vHOG groups noted ‘debated’, several concern bones and the skeletal system, due to citations such as the following: ‘Whether this “biomineralization toolkit” of genes reflects a parallel co-option of a common suite of genes or the inheritance of a skeletogenic gene regulatory network from a biomineralizing common ancestor remains an open debate’ (Murdock and Donoghue, 2011) (more references in the association file). Another case of debate is the ‘vomeronasal organ’ (vHOG:0000665). None of the three bibliographical references cited in the association file is conclusive concerning the existence and homology of this term, according to OBO Foundry guidelines. As noted in the TAO publication (Dahdul et al., 2010), it is at present difficult to implement this guideline in practice for anatomical ontologies. In future versions of the vHOG, we will continue working toward this aim.

The terms and relations thus generated were incorporated into the framework of the CARO (Haendel et al., 2008). We were careful to implement this in a manner consistent with other anatomical ontologies using CARO: TAO, XAO and ZFA. This especially concerns the relations between vHOG terms derived from species-specific ontology alignments and higher level CARO terms, e.g. ‘digit’ is_a ‘anatomical cluster’.

The vHOG ontology version 1.4 (December 2011) has 1184 terms, 306 is_a relations and 1181 part_of relations. There are 664 synonyms: 547 terms have definitions; 67% of terms have only one parent, 25% have two parents, and the others have three or more. There are 493 cross-references (xref) to other multispecies OBO ontologies. These do not include the mappings to species-specific ontologies, as we consider that mappings from multispecies ontologies to species-specific ontologies should not be treated as xrefs, but should be encoded separately in an association file, similar to the Gene Ontology Annotation mapping (Barrell et al., 2009). This differs from the approach of Uberon, which includes cross-references to the source ontologies in the multispecies ontology itself (Dahdul et al., 2010).

All mappings of terms from the species-specific ontologies to the vHOG terms were manually curated. There are 5129 terms from species-specific ontologies mapped to 1189 vHOG terms, which represent 2259 hypotheses of homology between vertebrates (i.e. 2259 pair-wise relations between structures in different species). There are 15 vHOG terms with no mappings, which correspond to higher level CARO terms (e.g. ‘material anatomical entity’). The semantics of the mapping is ‘treat-xrefs-as-equivalent’ for CARO, and equivalent to ‘treat-xref-as-genus-differentia’ for the mapping to species-specific ontologies.

The mappings were annotated with ‘support codes’ to provide confidence information (Table 1). Means to provide confidence metadata information in support of annotations are currently being discussed in the framework of the Evidence Code Ontology (ECO) (Gene Ontology Consortium, 2010) and of the International Society for Biocuration. Our objective is to rely on the use of the ECO, once a standard to address the issue of confidence has emerged in the biocuration and ontology communities.

In practice, the support codes ‘obvious’ and ‘inferred’ are essentially used for homology between the two mammals considered in our ontology: 97% of all mappings with the code ‘inferred’ are for human and mouse. All mappings with the code ‘obvious’ are for human and mouse, except ‘whole organism’ which has the code ‘obvious’ for all ontologies.

Among the vHOG groups noted ‘debated’, several concern bones and the skeletal system, due to citations such as the following: ‘Whether this “biomineralization toolkit” of genes reflects a parallel co-option of a common suite of genes or the inheritance of a skeletogenic gene regulatory network from a biomineralizing common ancestor remains an open debate’ (Murdock and Donoghue, 2011) (more references in the association file). Another case of debate is the ‘vomeronasal organ’ (vHOG:0000665). None of the three bibliographical references cited in the association file is conclusive concerning the existence and homology of this term.
To answer these needs, we propose the first ontology of vertebrate homology. The vHOG aims at following OBO Foundry rules, and homologous organs, the vHOG. We use a strict definition of historical homology in an ontology. Since homology is always a hypothesis, the mappings of species-specific structures to vHOG terms are kept in a separate association file (see also Dahdul et al., 2010; Haendel et al., 2008). Structures that have no homolog are not included in the vHOG ontology, but can still be found in each of the ontologies which are mapped to it. Thus, inclusion in the multispecies ontology carries a clear biological meaning, without hampering the fine description of each species. Finally, divergent homologous structures can be mapped to the same vHOG term, while keeping their individual definitions.

At present, the vHOG is limited to homologies between those model species for which anatomical ontologies are publicly available. We plan to extend it to more diverse species, while maintaining the restriction to terms describing organs or tissues with evidence of homology.

The advantages and drawbacks of our strict homology approach are clear when comparing the vHOG with the Uberon. The Uberon contains many more terms (6806 as of October 2011). It includes homology mappings from our project, since the Bgee HOGs are one of Uberon’s source ontologies. Since it is not limited to homology, Uberon includes for example a term ‘eye’ (UBERON:0000970), which has the children ‘compound eye’ (UBERON:000018) and ‘simple eye’ (UBERON:000047). An automatic reasoner cannot distinguish the case of compound eyes, which are all homologous, from the case of the parent ‘eye’, which includes homologs and analogs. And in less obvious cases, it can be difficult to recover such information even for non-automated reasoning, i.e. by a biologist user. For example, auditory ossicles are all mapped to UBERON:0001686, whereas the amphibian ‘auditory ossicle’ (XAO:0000214) is not homologous to the mammalian ‘auditory ossicles’; in vHOG it is mapped to ‘hyomandibula – stapes’ (VHOG:0000688). On the other hand, Uberon provides information that is not included in vHOG, such as develops_from relations.

Thus, Uberon and vHOG are complementary projects, the one focused on function and on integrating as many terms as possible, the other focused on a more restrictive set of terms, with strict homology definitions.

As an example of application of an ontology based on homology, we have queried gene expression available in Bgee for human, mouse and zebrafish, for HoxA5 orthologs (http://tinyurl.com/bgee10-boxa5). In human, there are 63 organs

### Table 1. Support for mapping of species-specific anatomical ontologies to vHOG

<table>
<thead>
<tr>
<th>Support code</th>
<th>Meaning</th>
<th>References(^a)</th>
<th>Terms mapped</th>
<th>vHOGs with mapping(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obvious</td>
<td>General knowledge, no need for reference</td>
<td>No 44</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Well established</td>
<td>No debate in the literature</td>
<td>Yes 3754</td>
<td>815</td>
<td></td>
</tr>
<tr>
<td>Debated</td>
<td>Debate in the literature(^c)</td>
<td>Yes 27</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Uncertain</td>
<td>Not clearly established</td>
<td>Variable(^d)</td>
<td>378 103</td>
<td></td>
</tr>
<tr>
<td>Inferred</td>
<td>Deduced from references which do not discuss this mapping explicitly, or personal communication from experts</td>
<td>No 926</td>
<td>236</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Yes if at least one bibliographic reference is provided for each mapping.

\(^b\)The total is more than the number of vHOG terms, because different mappings to a same vHOG term can have different support.

\(^c\)A consensus is chosen and presented, but the debate is documented.

\(^d\)Either there is a reference in which the homology is discussed as uncertain, and it is provided, or this code is used when well-established or obvious relations between closely related species (e.g. human and mouse) are extended to other species (e.g. zebrafish).

organ in different tetrapodes (Doving and Trotier, 1998; Kardong, 2006 p. 669; Smith et al., 2001), and Doving and Trotier (1998) specify: ‘The opinions concerning the presence and functioning of the vomeronasal organ in humans are controversial’. Conversely, 87% of all mappings for zebrafish, the most divergent species included in vHOG, are ‘well established’ in the literature. Still, 7% are ‘uncertain’, as for example the ovary (ZFA:0000403 mapped to VHOG:0000251: ‘(...) while it is likely that Urbilateria lacked a complex somatic reproductive system, it is at present impossible to speculate on whether or not it possessed a true gonad’ (Extavour, 2007) (full citation in the association file) (see also DeFalco and Capel, 2008).

Finally, it should be noted that terms which represent different developmental stages of the same organ are mapped together. For example, ‘presumptive midbrain’ (ZFA:0000148) and ‘midbrain’ (ZFA:0000128) are both mapped to ‘midbrain’ (VHOG:0000069).

This leads to some loss of information concerning developmental relations, and notably the relation develops_from is not implemented in the vHOG. But it provides a simpler description of vertebrate anatomy, which has proven relevant and useful to studying gene expression patterns (e.g. it is used in the Prosite database) (Sigrist et al., 2010). This has proven especially useful for human and mouse, for which different ontologies describe anatomy during development, and in the adult; vHOG (and the HOGs in Bgee) provide a high-quality mapping between these ontologies. Of note, for queries on gene expression in Bgee, a developmental stage can be specified, which then recovers only data from the correct structure.
or tissues with expression, in zebrafish 12 and in mouse 201. Unsurprisingly, given that the data are from targeted in situ hybridizations, most of the expression detected in zebrafish is shared with mammals. But there is also evidence from three high-quality in situ hybridization experiments of zebrafish-specific expression in the pharyngeal arch. Importantly, the homology of this structure is defined, and has been studied, in mouse and human (the branchial arch), confirming that the expression pattern of HoxA5 is probably zebrafish-specific. Conversely, the abundance of large-scale reports of expression in many organs leads to an uninformative mouse expression pattern, i.e. HoxA5 is detected to some degree in many structures where no biological role has been reported. Here homology information allows us to filter the data to recover the signal. Restricting to structures with homologous expression in human, for example, highlights expression in structures in which HoxA5 has been shown to play a functional role (Bouchet et al., 2009; Chen et al., 2005), such as the reproductive system (ovary, testis, uterus), forelimb, gut, bone and components of the respiratory system. Thus, the homology information in the ontology allowed both the identification of a species-specific patterns and of functionally important conserved expression.

4 CONCLUSION
Fine-grained yet large-scale comparisons between model organisms, especially vertebrates such as mouse or zebrafish and humans, is increasingly important. In addition to providing a framework for evolutionary studies, the vHOG provides a unique tool for relating humans and model organisms. Additionally, the association files of vHOG and HOGr are unique resources in providing detailed judgments of homology between anatomical structures, with supporting evidence from the literature.

ACKNOWLEDGEMENTS
Emilie Person and Balazs Laurenczy helped with annotations in earlier versions of the HOG ontology. We thank Linda Z. Holland for helpful discussions concerning anatomical homology of chordates. We thank Chris Mungall for helpful discussions concerning Uberon. We thank all members of the Robinson-Rechavi lab for helpful discussions. We thank three anonymous reviewers for their very helpful comments.

Funding: Etat de Vaud; the Décrypton program of Association Française contre les Myopathies; the European program Crescendo; and the Swiss National Science Foundation (grant 31003A_133011/1) and the Swiss Institute of Bioinformatics.

Conflict of Interest: none declared.

REFERENCES