OGG: a biological ontology for representing genes and genomes in specific organisms

Yongqun He*, Yue Liu, Bin Zhao
University of Michigan Medical School, Ann Arbor, MI 48109, USA: yongqunh@med.umich.edu

***Abstract* —** In this report, we present the development of the Ontology of Genes and Genomes (OGG), a biological ontology in the domain of genes and genomes. To integrate with other ontologies, OGG is aligned with the Basic Formal Ontology (BFO). OGG-specific term IDs and annotations are designed by mapping to NCBI Taxonomy IDs and NCBI Entrez Gene IDs. Each gene in OGG has over 10 annotation items, includes gene-associated Gene Ontology (GO) and PubMed article information. OGG has represented genes in human, two viruses, and four bacteria. Additionally, 7 OGG subsets are developed to represent genes and genomes of 7 model systems including mouse, fruit fly, zebrafish, yeast, *A. thaliana*, *C. elegans*, *P. falciparum*. An ontology URI dereferencing approach was designed and implemented in Ontobee to resolve the issue of dereferencing OGG terms from different OGG subset documents. OGG can be used in different cases, including SPARQL query of gene information within OGG or in combination with other ontologies, and the OGG gene term reuse in other ontologies (*e.g.*, Vaccine Ontology). The OGG project website is: https://code.google.com/p/ogg/.

Keywords—ontology; Ontology of Genes and Genomes (OGG); gene; genome; organism; vaccine; Gene Ontology (GO)

I. INTRODUCTION

Genes and genomes are fundamental to biological life and today's biological and biomedical research. In molecular biology, a gene is typically defined as the entire nucleic acids necessary for the synthesis of a functional unit including protein or RNA. A genome includes the entirety of an organism's genetic material. Depending on organism types, the genome sizes vary. For example, a human genome has a length of approximately 3.2 giga base pairs (Gb) that contains ~40,000 genes [1]. A typical *E. coli* has approximately 4.6 Mb and ~4,000 genes [2]. In contrast, a typical HIV virus has only 9.7 kb containing 10 genes [3].

Many resources of genes and genomes exist. The US National Center for Biotechnology Information (NCBI) provides several databases containing rich information about genes, genomes, and organisms. Particularly, the NCBI Taxonomy database has classified nearly one million various organisms [4]. NCBI Genome includes detailed information about genomes. The NCBI Entrez Gene (abbreviated as "NCBI Gene" later) database has accumulated over 14 million genes [5]. Other institutes and organizations also provide related information. For example, the Ensembl database includes gene and genome information for important eukaryotic organisms [6]. To facilitate data exploration, web queries and graphic visualization interfaces are also included in these resources. However, none of the gene and genome resources has been presented in an ontology.

An ontology focusing on the representation of classes of specific genes (*e.g.*, human gene *casp2*) and genomes (*e.g.*, human genome) in various organisms (*e.g.*, human or *Homo sapiens*) has not been reported. The Gene Ontology (GO) represents information about biological processes, molecular functions, and cellular components of genes or gene products [7]. Therefore, GO is not an ontology about specific genes. The GO website provides the links to gene products that are related to GO terms. For example, the web link (http://amigo.geneontology.org/amigo/gene_product/UniProtKB:C9JRR9) provides the information about a human protein (CASP2) and related GO associations. However, a gene product is not a gene itself. As a central hub of functional information on proteins, the UniProtKB is (i.e., the UniProt Knowledgebase) [8] is not an ontology. Many other gene-related ontologies also exist, for example, Sequence Ontology (SO) [9], YAMATO ontology [10], and Genetics Ontology (GXO) [10]. However, instead of representing specific genes in different organisms, these ontologies are designed to represent general top level terms of sequences, genetics, and genomics.

An ontology of specific genes and genomes for various organisms is frequently needed. For example, in the Vaccine Ontology (VO) [11] and Brucellosis Ontology (IDOBRU) [12], many genes from specific organisms (*e.g.*, bacteria and viruses) have been used for development of vaccines and generation of gene mutant. It is not optimal to generate VO and IDOBRU specific terms for these genes since these genes should come from a common ontology source for better data integration and sharing based on OBO Foundry principles [13].

To address a major bottleneck of lacking an ontology of specific genes from different organisms, we have initiated the development of a new ontology called the Ontology of Genes and Genomes (OGG). OGG is developed to incorporate existing gene and genome resources with a unique design. The OGG project (initially GGO, and later called OGG) was announced in the end of October 2013 and has received very positive feedback [14, 15]. The ontology and its namespace "OGG" have been approved by the Open Biological and Biomedical Ontologies (OBO) Foundry [14]. In this manuscript, we present the rationale, design pattern, and selected use cases of OGG.

II. METHODS

A. Ontology format and editing

OGG is generated using the W3C standard Web Ontology Language (OWL2) [16]. The Protégé-OWL editor (version 4.2) is used for manual OGG editing.

B. *Ontology term reuse*

OGG imports the whole set of the Basic Formal Ontology (BFO) as its upper level ontology [17]. BFO has been used as an upper level ontology used by over 100 biological and biomedical ontologies. The alignment of OGG with BFO makes it possible to integrate OGG with other ontologies. To support ontology interoperability, many terms from reliable ontologies are reused. To facilitate the reusing process, OntoFox [18] was applied for automatically extracting individual terms from existing ontologies, including NCBITaxon (*i.e.*, a taxonomy ontology based on the NCBI Taxonomy database) [19], the Ontology for Biomedical Investigations (OBI) [20], and Information Artifact Ontology (IAO) [21].

C. *New OGG term generation*

New OGG-specific terms were generated using new OGG IDs with the prefix of "OGG_" followed by 10 digits. An OGG-base OWL file was first generated to include basic OGG hierarchy and key terms. The data of the NCBI Gene database was downloaded from the NCBI Gene FTP (ftp://ftp.ncbi.nih.gov/gene/). A MongoDB database (http://www.mongodb.org/) was generated to parse and store the downloaded NCBI Gene contents. To avoid name conflicts, a specific scheme is designed to assign non-redundant OGG IDs. Based on the pre-defined scheme and using the OGG-base and MongoDB data, a Java program was developed to generate new OGG IDs, hierarchies, and annotations.

D. *OGG URI dereferencing:*

A URI "dereferencing" is defined as an act of retrieving a representation of a resource identified by a uniform resource identifier (URI) [22]. Following the default OBO Foundry domain dereferencing policy, OGG URIs are directed to be resolved in Ontobee [23]. However, since different OGG OWL files (*e.g.*, ogg.owl and ogg-mm.owl) exist and all OGG subsets use the same OGG namespace, for a given OGG term URI, Ontobee was not be able to identify which OGG OWL file to use for the URI dereferencing. This issue was solved with a special design and updated Ontobee program as described in the Results section.

E. *OGG use cases:*

Three OGG use cases are introduced. First, OGG was used as a knowledge base for SPARQL query of various gene and genome information. Second, since OGG includes gene-associated GO IDs, SPARQL queries were developed to query both OGG and GO for useful gene-related information. Third, the OGG terms of genes and genomes were reused in existing ontologies such as the Vaccine Ontology (VO) [11].

III. RESULTS

A. *OGG ontology design and development*

(1) OGG is aligned with BFO and OBO Foundry ontologies

The OGG was developed by first identifying the relations among gene, genome, and organism. Specifically, an organism has a genome, and a genome has many genes. OGG represents both genes and genomes as BFO:*material entity* (Fig. 1).

The OGG:*gene* (OGG_0000000002) is defined as "a material entity that represents the entire DNA sequence required for synthesis of a functional protein or RNA molecule" [24]. In addition to the coding regions (exons), a gene includes transcription-control regions and sometimes introns. Although the majority of genes encode proteins, some encode tRNAs, rRNAs, and other types of RNA. It is noted that the OGG 'gene' is an ontology class or type [25]. Although OGG focuses on the representation of specific genes in different species, these specific genes are subclasses of the OGG:gene, and they are not ontology individuals or tokens (*i.e.*, spatio-temporal particulars) [25].

The default OGG covers 7 model organisms, including Homo sapiens (*e.g.*, human), two viruses, and four bacteria (Fig. 2). The two viruses are HIV and influenza virus. The four bacteria include *Escherichia coli*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* (a common opportunistic and nosocomial pathogen), and *Brucella melitensis* (cause of a common zoonotic disease brucellosis). The organism information including their hierarchy was extracted from the NCBITaxon ontology using the OntoFox program [18]. Corresponding to a specific "organism X" (*e.g.*, human), the terms 'genome of organism X' and 'gene of organism X' were generated in OGG. The hierarchical structures of the genomes and genes of all the organisms maintain the same as the hierarchy of these organisms shown in the NCBITaxon taxonomy ontology (Fig. 2). As shown in Fig. 2, a large number of OGG terms are generated using the strategy of ontology cross-product generation [26]. For example, the OGG term 'gene of Eukaryota' (OGG_2000009606) is a cross-product term generated using the OGG term 'gene' and the NCBITaxon term 'Eukaryota'. Particularly, '*gene of Eukaryota*' *'is gene of organism' some Eukaryota*.

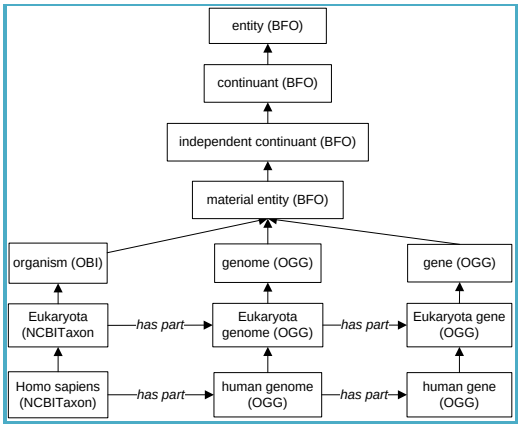

Fig. 1. Basic OGG hierarchy of gene and genome representation. OGG is aligned with BFO. Like organism, genes and genomes are material entities. The relations among an organism, a genome, and a gene are that an organism *has part* a genome, and a genome *has part* a gene. For example, a human organism has a human genome, and a human genome has genes. It is noted that other organisms are not included in this figure. The term '*has part*' is a regular OWL object property. All the arrows without the '*has part*' label represent the *rdf:subClassOf* (or called *is a*) relation.

In OGG, a '*gene disposition*' is defined as a BFO:*disposition* where a gene has a tendency of being

* Corresponding author of the paper.

expressed to different gene products such as protein and RNA. Corresponding to various gene dispositions [27], OGG includes a hierarchy of different types of organism genes under the branch of '*material entity*'. For example, OGG includes a term called 'protein-coding gene' that has the disposition of '*protein-coding gene disposition*'. For each specific species, there are also different specific types of genes in each organism, such as '*protein-coding gene of Homo sapiens*' (Fig. 2). Indeed, the type of genes with the highest number of genes is usually the protein-coding gene. There are many different RNA gene types including ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and non-coding RNA (ncRNA) (Fig. 2).

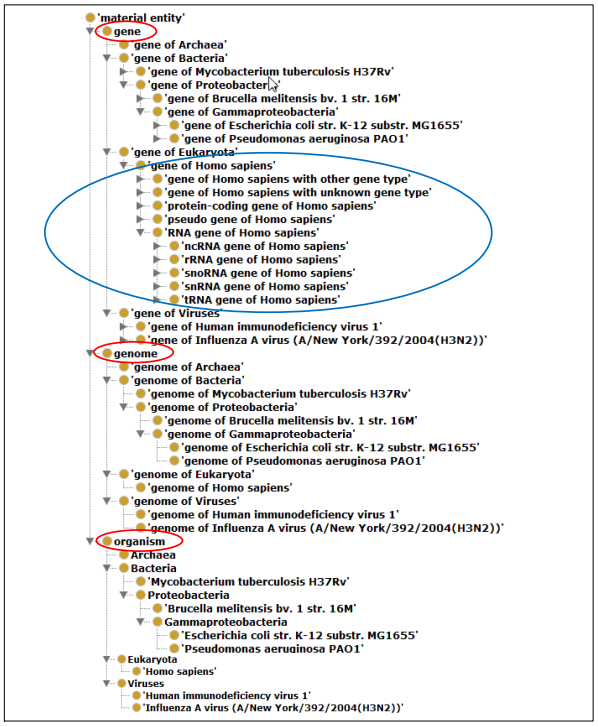

Fig. 2. The hierarchies of 7 model organisms and related genes and genomes in OGG. The ontology terms of 7 organisms and their hierarchy were retrieved using OntoFox from the source ontology NCBITaxon (with the OntoFox option of "includeComputedIntermediates"). The genes and genomes (labeled with red circles) of organisms have the same hierarchy structure as the organism hierarchy. The hierarchy under the gene of a specific organism (*e.g.*, Homo sapiens or human) is showed in the blue circle. The Protégé-OWL editor was used for visualization.

(2) Automatic OGG gene and genome ID assignments

With millions of genes sequenced and annotated, it is a challenge to assign OGG gene IDs without redundancy. We have thus generated a special scheme (or called algorithm) for new OGG ID assignments (Fig. 3).

The key part of this scheme is ontology ID mapping with NCBITaxon IDs and NCBI Gene IDs, the two sets of reliable and non-redundant identifiers from the NCBI resources. The resource of the NCBI organism taxonomy database has been transformed to the NCBITaxon organism taxonomy ontology [19]. Making OGG genome and gene IDs map to NCBITaxon IDs and NCBI Gene IDs allow us to design and develop computer programs to automatically generate reliable and non-redundant OGG genome and gene URIs (Fig. 3). A gene can be expressed into different types of gene productions. NCBI summarizes 12 gene types (*e.g.*, protein-coding and tRNA gene types) based on the gene products [27]. Correspondingly, OGG includes 12 *gene dispositions* mapping to these 12 gene types. Based on a specific gene disposition associated with a gene, our program classifies the gene type. The BFO object property (*i.e.*, relation) '*has disposition at all times*' (BFO_0000162) has been generated to represent a relation between a gene and a gene disposition. For example, the '*protein-coding gene of Homo sapiens*' '*has disposition at all times*' *some 'protein-coding gene disposition'*.

As an example, Fig. 3B illustrates how OGG is used to assign IDs and annotations for a human gene CASP2 (*i.e.*, *casp2*) that encodes a human protein Caspase-2. The same design pattern is applied to all other genes in other organisms.

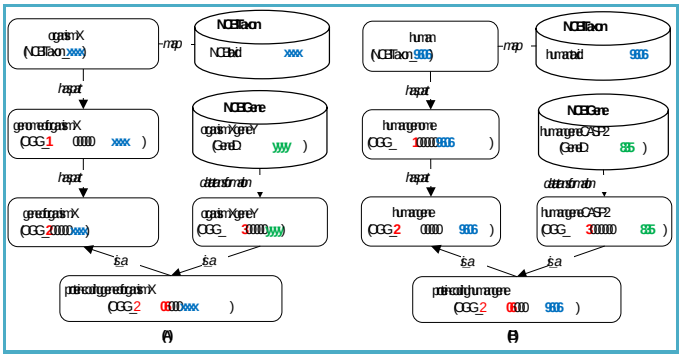

Fig. 3. OGG ID assignment strategy. (A) General design. The 10-digits of OGG IDs for genome and gene of an organism map to the corresponding NCBITaxon ID (*e.g.*, 9606) with a pre-defined first digit "1" or "2", respectively. To further label the gene type of an organism gene, the 2nd and 3rd digits of the 10-digit number are used. A specific OGG gene ID maps to its corresponding NCBI Gene ID with an additional pre-defined first digit "3". The gene type information is used to generate gene hierarchies in OGG. The second and third digits are used to represent specific gene type, *e.g.*, "06" representing protein-coding gene type. The relations among these terms are indicated with italicized relation terms. **(B) An example: human gene CASP2** (*i.e.*, *casp2*). The example illustrates how the OGG ID assignment strategy works. Based on the design, the OGG term URI for this human gene is: http://purl.obolibrary.org/obo/OGG_3000000835.

Since both NCBI Taxonomy IDs and NCBI Gene IDs are unique (non-redundant) and stable among all organisms, our OGG naming design can be reused to efficiently generate new OGG subsets for other organisms without a naming conflict.

(3) OGG gene annotations use the NCBI Gene resource

The gene annotation information from the NCBI Gene database was extracted and used to annotate genes using OGG-predefined annotation or object properties. In total up to 17 annotation items are provided for each gene. Examples of the annotations include gene symbol, alternative terms, NCBI Gene ID, description, and associated GO and PubMed IDs (Fig. 4).

One of the gene annotations is the GO IDs associated with a specific gene. For example, CASP2 is associated with *GO_004197 (EC: IDA; PMID: 10980123)* (Fig. 4), where EC: IDA means "Evidence Code" (EC) "Inferred from Direct

Assay" (IDA). PMID is the PubMeD unique identifier. Some genes are associated with a large number of GO IDs. For example, human TP53 gene is associated with over 6,000 GO IDs. To show all these IDs in a single HTML page is neither necessary nor user-friendly. Therefore, we have chosen to show up to 20 GO IDs in the Ontobee page (See red-highlighted text in Fig. 2). All the other GO IDs associated with the gene can be retrieved by viewing the page source (Fig. 4B). Instead of HTML source code, the source of an ontology term URI in Ontobee is generated as the easy-to-parse RDF/OWL format [23].

Fig. 4. Example of OGG gene term annotations using Ontobee. The human gene CASP2 is used as an example here. In total 14 different types of annotations are included for this gene. **(A) HTML display of the gene information**. Only up to 20 GO IDs and 50 PMIDs are displayed in the HTML web page. **(B) Page source of the OGG term URI**. The complete list of the GO associations is provided in the web page source. Google Chrome was used as the web browser. Note that only parts of the HTML and page source contents are viewed here.

(4) Statistics of OGG and released OGG subsets

At current stage, OGG has been developed to represent the information of all genes and genomes of 14 organisms (Table 1). Due to the large number of genes in these 14 organisms, it is not feasible to put all the genes of all sequenced organism genomes into single OWL document. Therefore, in addition to the 7 organisms covered in the default OGG, we have generated OGG subsets targeting for different model organisms. For example, OGG-Mm represents the OGG subset for *Mus musculus* (*i.e.*, mouse). The development of OGG subsets follows the same strategy as shown in Fig. 1-4. The statistical numbers of the OGG and different OGG subsets are included in Table 1.

Table 1: Statistics of OGG as of May 14, 2014

Species/strain (Common name)	strain	NCBI Taxon id	Subset name (#terms)
H. sapiens (human)	-	9606	OGG (69,800)
Bacteria			
B. melitensis	16M	224914	
E. coli	MG1655	511145	
M. tuberculosis	H37Rv	83332	
P. aeruginosa	PAO1	208964	
Viruses			
human immunodeficiency virus (HIV)	-		
Influenza virus	392/2004 (A/H3N2)	335341	
A. thaliana	-	3702	OGG-At (33,774)
C. elegans (roundworm)	-	6239	OGG-Ce (45,912)
D. melanogaster (fruit fly)	-	7227	OGG-Dm (23,574)
D. rerio (zebrafish)	-	7955	OGG-Dr (36,792)
M. musculus (mouse)	-	10090	OGG-Mm (69,539)
P. falciparum	3D7	36329	OGG-Pf (5,694)
S. cerevisiae (yeast)	S288c	559292	OGG-Sc (6,535)

B. OGG term URI dereferencing and query in Ontobee

An ontology term URI denoting a thing is referred to and looked up ("dereferenced") by people and user agents. An OGG URI includes an HTTP domain name and an OGG ID. As an approved OBO Foundry candidate ontology, OGG uses the domain name http://purl.obolibrary.org/obo/, where http://purl.obolibrary.org/ is a CNAME (*i.e.*, a canonical name or alias) that redirects to http://purl.org. To have an OGG URI, the domain name is followed by an OGG term ID.

According to the OBO Foundry PURL domain dereferencing policy [28], an OGG term URI is by default dereferenced in Ontobee (Fig. 5). For example, based on this policy, the OGG term URI:

http://purl.obolibrary.org/obo/OGG_3000000835 (mouse CASP2 gene) should be directed to:

http://www.ontobee.org/browser/rdf.php?o=OGG&iri=http://purl.obolibrary.org/obo/OGG_3000000835

However, by our design, the mouse gene is located in the OGG-Mm subset file instead of the default OGG file. Since all OGG-specific terms in OGG and different OGG subsets use the same OGG prefix "OGG_", an OGG term in an OGG subset may be mistakenly dereferenced using the default OGG instead of its corresponding OGG subset (*e.g.*, OGG-Mm).

To solve this issue, we have developed and implemented a new strategy in Ontobee as illustrated in Fig. 5. Basically, once Ontobee detects an OGG term for dereferencing, it will act based on different conditions. For example, when the OGG term ID starts with the number "3", Ontobee will know that this is an OGG gene term. The Ontobee program will then identify the NCBI Gene ID based on the OGG ID assignment strategy (Fig. 3). Using a web NCBI E-utility program embedded in Ontobee, the NCBI Taxonomy ID associated with this gene will be identified. The Ontobee database maintains a predefined mapping table between NCBI Taxonomy IDs and OGG subset names. Based on the mapping result, Ontobee will know which OGG subset stored in the Ontobee RDF triple store should be used for retrieving the term information and

displaying the information. Fig. 5 illustrates how an OGG term (*i.e.*, human gene CASP2) is dereferenced in Ontobee.

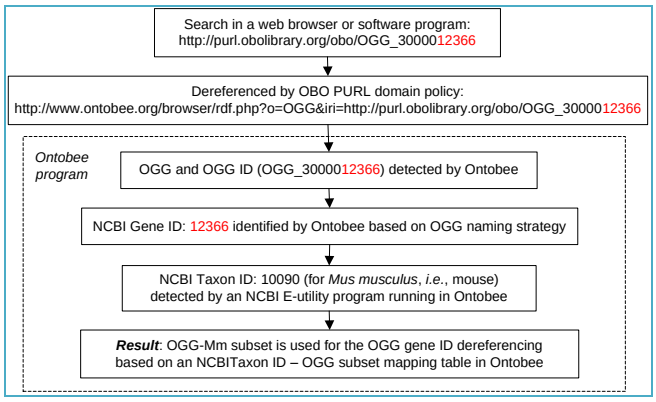

Fig. 5. Illustration of Ontobee dereferencing OGG term URI. of OGG. An example OGG term URI of representing mouse gene CASP2 is dereferenced in Ontobee using the OGG-Mm subset. This pipeline shows all the steps where those steps inside the dashed box occur inside Ontobee. Note that the NCBI Taxonomy ID associated with an OGG gene is already stored as an annotation content of the OGG gene record (see Fig. 4). See text for more detail.

It is noted that an alternative solution for the dereferencing problem is to provide a direct mapping between an OGG gene term ID and an OGG subset. Before the mapping, all we have is an ontology name (*i.e.*, OGG) and an ontology term IRIs. If we store a mapping from OGG terms to OGG subset names directly, we will have to store a huge number of mappings due to the availability of a huge number of OGG gene terms. Since there is no specified range of gene IDs available for easy mapping, each individual OGG term will need a specific mapping. This is very space-consuming. Furthermore, if new gene terms are added, we will have to add new mappings. It will be much more challenging to maintain. In comparison, since the mapping before a NCBI Gene ID and its Taxonomy ID is available already recorded, our design of "Gene ID – Taxonomy ID – OGG subset" is more robust and maintainable.

In addition to OGG, some other ontologies, such as the Infectious Disease Ontology (IDO) [29], also have different ontology subsets (*e.g.*, the IDO-core and IDOBRU [12]) but use the same namespace. In such cases, appropriate dereferencing of ontology terms can be very challenging. The solution designed and implemented in this OGG study provides a novel and feasible example on how to address this situation. Indeed, we have recently used a similar mapping approach to solve the issue of IDOBRU ontology term dereferencing. In the IDOBRU dereferencing case, since a specific range of IDO IDs were pre-assigned to IDOBRU, an examination of an IDO ID allows Ontobee to determine which subset (IDO-core or IDOBRU) to use for term dereferencing.

C. OGG use cases:

OGG can be used for different applications. Three use cases are introduced as follows:

Use Case 1: Query OGG for gene information

OWL-formatted OGG is stored in the Ontobee RDF triple store, a database system based on the Resource Description Framework (RDF) [23]. SPARQL is an RDF query language able to retrieve data stored in the triple store. Therefore, SPARQL queries can be developed to query the rich gene and genome information represented in OGG and OGG subsets. For example, Fig. 6 provides an example of SPARQL querying the number of human tRNA genes (OGG_2010009606). With only a few lines of code, this query shows that 579 tRNA genes exist in the human organism.

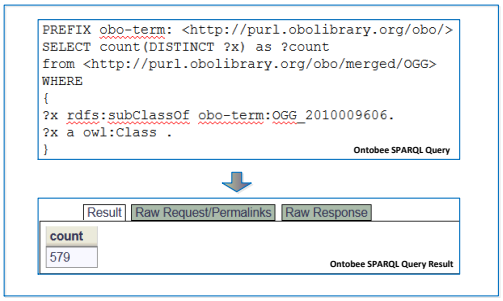

Fig. 6. SPARQL query of RNA genes in human. The OGG term OGG_2010009606 is 'tRNA gene of Homo sapiens'. The query was performed using the Ontobee SPARQL query interface: http://www.ontobee.org/sparql/.

Use Case 2: Query OGG & GO for the gene-GO associations

Besides querying OGG class hierarchy as shown above, the rich annotation contents of OGG genes can also be queried. As shown in Fig. 4, an OGG gene is usually associated with many GO terms that represent the biological processes, cellular components, or molecular functions of the gene product [7]. To identify what or how many genes are associated with a GO term, we can use SPARQL query again. Fig. 7 provides a SPARQL query example of identifying how many mouse genes are associated with GO 'leukocyte apoptotic process' (GO_0071887) and the subclasses of the GO term. Based on GO, GO_0071887 has 18 subclasses in 5 layers. The SPARQL query shown in Fig. 7 is able to identify all the OGG genes that are associated with GO_0071887 or any of its subclasses.

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
SELECT DISTINCT ?s ?labelogg ?annotation
from <http://purl.obolibrary.org/obo/merged/GO>
from <http://purl.obolibrary.org/obo/merged/OGG-Mm>
WHERE
{
 { #Note: Get OGG genes associated with GO_0071887
   ?s a owl:Class .
   ?s rdfs:label ?labelogg .
   ?s obo:OGG_0000000029 ?annotation .
   FILTER regex(?annotation, "GO_0071887") .
 }
 union
 { #Note: Get OGG genes with descendants of GO_0071887
   ?s a owl:Class .
   ?s rdfs:label ?labelogg .
   ?s obo:OGG_0000000029 ?annotation .
   FILTER regex(?annotation, bif:substring(?x, 32, 10)) .
   ?x rdfs:subClassOf obo:GO_0071887 option (transitive) .
   ?x rdfs:label  ?labelgo .
   ?x a owl:Class .
 }
}
```

Fig. 7. Query OGG and GO for genes associated with GO_0071887 ("leukocyte apoptotic process"). The term OGG_0000000029 is an object property 'has GO association'. In total 28 genes were found. See text for detail.

Use Case 3: OGG term reuse in other ontology development

One driving biological project for the OGG development is the usage of the same OGG gene representation across different ontologies, such as the Vaccine Ontology (VO) [30]

and Brucellosis Ontology [12]. An example is shown in Fig. 8. Using OntoFox, we imported 10 *M. tuberculosis* gene terms from OGG (more OGG terms will later be imported to VO). These OGG gene terms were used to logically represent many live attenuated *M. tuberculosis* vaccines. For example, the OGG term for *M. tuberculosis* gene *drrC* (OGG_3000888491) is now used in VO to define a vaccine '*Mycobacterium tuberculosis drrC mutant vaccine*' (VO_0002780) as:

'has part' some ('Mycobacterium tuberculosis' and ('has gene mutation' some drrC))

In this case, 'has gene mutation' represents a shortcut relation between an organism and a gene where the organism has a mutation of the gene. After the OGG term is imported to VO, it is also possible to add additional annotation to the OGG term inside VO. For example, a comment is added to annotate *M. tuberculosis* gene *drrC* in the content of VO (Fig. 8).

Fig. 8. Usage of OGG gene terms in VO. Ten *M. tuberculosis* gene terms were imported to VO by OntoFox [18]. These ten genes were mutated from wild type *M. tuberculosis* for generating live attenuated vaccines. Note that this this is a screenshot of an Ontobee web page dereferencing the OGG term: http://purl.obolibrary.org/obo/OGG_3000888491.

IV. Discussion

In this paper, we have introduced the Ontology of Genes and Genomes (OGG). OGG is aligned with the BFO, making it possible for OGG to integrate with over 100 other BFO-aligned biological and biomedical ontologies.

The rationale and methods of the OGG development has been well discussed and vetted among ontology developers in the OBO Foundry discussion email list (obo-discuss). One major session of discussions occurred in October 2013. Another major session of discussions occurred in the end of March and early April 2014. In here, we want to summarize a few most important issues we have discussed.

Currently, OGG defines the term "gene" inside OGG. The reason why OGG does not use the "gene" definition in the Sequence Ontology (SO) is that current SO version still treats the "gene" as a sequence feature instead of a material entity as defined in OGG. Instead of being a material entity, the SO:*gene* (SO_0000704) is classified under the branch of SO:*sequence_feature* (SO_0000110), which is aligned with the BFO term 'generically dependent continuant' [9]. Therefore, SO describes the gene sequences that inhere in genes rather than the genes themselves. However, SO developers have realized the gap between the gene as a material entity (a BFO 'independent continuant') and the gene sequence as a 'generically dependent continuant', and proposed to fill the gap by Sequence Ontology:Molecules (SOM), an ontology of molecules with genomic origin [9]. Based on the discussion between OGG and SO developers, once the SO improvements are made, OGG will discuss with SO and align its definition with SO [31]. Meanwhile, other ontologies, including the Genetics Ontology (GXO) [10] and the Ontology for Genetic Interval (OGI) [32], have represented gene-related entities with different details and emphases. There are also many unresolved issues in how to represent and analyze many gene/genome-related entities such as different types of genomic segments, and relations between genes and alleles [10]. Ontology terms with the same label in natural language may have different meanings in different ontologies. A collaborative and integrative work among these different ontologies would support shared and community-based ontological representation of gene-related entities.

The Protein Ontology (PR) [33] has initially been developed to primarily represent protein groups. The recent versions of PR have also included specific proteins from different organisms. Both PR and OGG developers realize that the representations of specific prokaryotic and eukaryotic proteins are critical for different applications such as the study of host-microbe interactions and vaccine design [34]. Proteins are the main type of gene products. PR and OGG developers have been communicating and collaborating in the development of these two important ontologies.

Another recent discussion in OBO-discuss email list is on the usage of NCBI Gene or Genome namespace or the usage of OGG namespace to represent the genes and genomes [15]. In general, it has been agreed that commonly referenced public resources such as NCBI Gene and Ensembl databases store the data about the entities (*e.g.*, gene). They are different from the gene entities represented in the ontology. Therefore, it is not recommended by OBO Foundry to use resource names (*e.g.*, NCBI Gene) as the namespace of an ontology. However, the data resource is required to be cited as a definition source. A linking to the resource page mechanism is also being discussed inside the ontology community.

Since current OGG design relies on the existence of a gene and organism in the NCBI Gene and Taxonomy resources, the design does not cover the scenario when a gene or an organism is not recorded in these NCBI resources. For example, African

swine fever virus (ASFV) isolate Zi UK gene (GenBank accession number: AF015681; GenBank GI: 2905984) is a virulence determinant [35]. This ASFV isolate is not classified in the NCBI Taxonomy database and thus does not have an NCBI Taxonomy ID (or an NCBITaxon ontology term ID). The NCBI GenBank record of this gene (http://www.ncbi.nlm.nih.gov/nuccore/AF015681) uses the NCBI Taxonomy ID of 10497, which is the ASFV species taxonomy ID instead of the ID for the ASFV isolate. Although this gene from the ASFV isolate Zi exists in the GenBank database, the gene is not listed in the NCBI Gene database. One major difference between the NCBI GenBank and Gene resources is that the GenBank sequences are obtained primarily through public submissions [36], but the NCBI Gene database includes non-redundant curated gene data representing our current knowledge of known genes in different organisms [5]. In such a case when a gene record is in GenBank (or a non-NCBI resource) but not in NCBI Gene, different ways may be used to represent this gene in OGG. For example, we may generate an OGG gene ID "OGG_AF015681", where the "AF015681" is the accession number of the gene in GenBank. This strategy of ontology ID generation is similar to how the Protein Ontology (PR) reuses the UniProtKB protein accession numbers [37]. The usage of such a strategy should be cautious since it might potentially cause duplications between different gene records in OGG.

The OGG representation of specific genes in different organisms supports gene-related data integration and ontology reuse. Three use cases are demonstrated in this manuscript. More use cases can be identified. For example, OGG can be used to represent genes whose expression levels are measured using different DNA microarray technologies. The usage of OGG genes makes it possible to compare gene expression levels with the same gene representation. In the Big Data era, OGG provides a standard gene representation to be used in the field of Semantic Web.

ACKNOWLEDGMENT

We thank Drs. Chris Mungall, Alan Ruttenberg, Barry Smith, Jie Zheng, Yu Lin, Richard H. Scheuermann, Erick Antezana, and Darren Natale for their valuable discussions and feedback. This research is supported by NIH grant R01AI081062.

REFERENCES

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, *et al.*, "The sequence of the human genome," *Science*, vol. 291, pp. 1304-51, Feb 16 2001.

[2] F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, *et al.*, "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, pp. 1453-74, Sep 5 1997.

[3] S. Wain-Hobson, P. Sonigo, O. Danos, S. Cole, and M. Alizon, "Nucleotide sequence of the AIDS virus, LAV," *Cell*, vol. 40, pp. 9-17, Jan 1985.

[4] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Res*, vol. 40, pp. D136-43, Jan 2012.

[5] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res*, vol. 39, pp. D52-7, Jan 2011.

[6] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, *et al.*, "Ensembl 2014," *Nucleic Acids Res*, vol. 42, pp. D749-55, Jan 2014.

[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, May 2000.

[8] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, "UniProtKB/Swiss-Prot," *Methods Mol Biol*, vol. 406, pp. 89-112, 2007.

[9] C. J. Mungall, C. Batchelor, and K. Eilbeck, "Evolution of the Sequence Ontology terms and relationships," *J Biomed Inform*, vol. 44, pp. 87-93, Feb 2011.

[10] H. Masuya and R. Mizoguchi, "An Ontology of Gene," in *Proc. of the 3rd International Conference on Biomedical Ontology (ICBO 2012)*, Graz, Austria, 2012, pp. 1-5.

[11] Y. He, L. Cowell, A. D. Diehl, H. L. Mobley, B. Peters, A. Ruttenberg, *et al.*, "VO: Vaccine Ontology," in *The 1st International Conference on Biomedical Ontology (ICBO-2009)*, Buffalo, NY, USA, 2009, p. http://precedings.nature.com/documents/3552/version/1.

[12] Y. Lin, Z. Xiang, and Y. He, "Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology," *J Biomed Semantics*, vol. 2, p. 9, 2011.

[13] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, pp. 1251-5, Nov 2007.

[14] Y. He. (2013). *Announcement of the Ontology of Genes and Genomes (OGG)*. Available: https://groups.google.com/forum/#!topic/ogg-discuss/wy0132CCdNA

[15] OBO-discuss. (2014). *OGG Updates*. Available: https://groups.google.com/forum/#!msg/obo-discuss/Ls2BhZIzMu4/3ShybVtK5j8J

[16] W3C, "OWL 2 Web Ontology Language document overview," pp. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/. Accessed on March 1, 2014, 2009.

[17] P. Grenon and B. Smith, "SNAP and SPAN: Towards Dynamic Spatial Ontology," *Spatial Cognition and Computation*, vol. 4, pp. 69-103, 2004.

[18] Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, and Y. He, "OntoFox: web-based support for ontology reuse," *BMC Res Notes*, vol. 3, p. 175, 2010.

[19] *OBO Foundry wiki. Introduction of the NCBITaxon ontology*. Available: http://www.obofoundry.org/wiki/index.php/NCBITaxon:Main_Page

[20] R. R. Brinkman, M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord, *et al.*, "Modeling biomedical experimental processes with OBI," *J Biomed Semantics*, vol. 1 Suppl 1, p. S7, 2010.

[21] IAO. *Information Artifact Ontology*. Available: http://code.google.com/p/information-artifact-ontology/

[22] R. Lewis. (2007, Nov 13). *Dereferencing HTTP URIs*. Available: http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14

[23] Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He, "Ontobee: A linked data server and browser for ontology terms," in *The 2nd International Conference on Biomedical Ontologies (ICBO)*, Buffalo, NY, USA, 2011, pp. Pages 279-281 [http://ceur-ws.org/Vol-833/paper48.pdf].

[24] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*. New York: W. H. Freeman and Company, 2000.

[25] L. Wetzel, "Types and tokens," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Spring 2014 Edition ed, 2014.

[26] C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, *et al.*, "Cross-product extensions of the Gene Ontology," *J Biomed Inform*, vol. 44, pp. 80-6, Feb 2011.

[27] J. Ostell. (2011). *NCBI Entrezgene definitions*. Available: http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn

[28] M. Courtot and O. F. O. Committee. (2014, OBO PURL Domain configuration of the OBO PURL domain. Available: https://code.google.com/p/obo-foundry-operations-committee/wiki/OBOPURLDomain

[29] L. G. Cowell and B. Smith, "Infectious Disease Ontology," in *Infectious Disease Informatics*, V. Sintchenko, Ed., ed New York Dordrecht Heidelberg London: Springer, 2010, pp. 373-395.

[30] Y. Lin and Y. He, "Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses," *J Biomed Semantics*, vol. 3, p. 17, Dec 20 2012.

[31] Y. He and C. Mungall. (2013). *OGG vs SO*. Available: https://groups.google.com/forum/#!topic/ogg-discuss/Woi05g0nf0c

[32] Y. Lin and P. Simons, "DNA sequence from below: a nominalist approach," in *Interdisciplinary Ontology Vol.3 - Proceedings of the Third Interdisciplinary Meeting*, Tokyo, Japan, 2010, pp. 79-88.

[33] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, *et al.*, "The Protein Ontology: a structured representation of protein forms and complexes," *Nucleic Acids Res*, vol. 39, pp. D539-45, Jan 2011.

[34] Y. He, R. Rappuoli, A. S. De Groot, and R. T. Chen, "Emerging vaccine informatics," *J Biomed Biotechnol*, vol. 2010, p. 218590, 2010.

[35] L. Zsak, E. Caler, Z. Lu, G. F. Kutish, J. G. Neilan, and D. L. Rock, "A nonessential African swine fever virus gene UK is a significant virulence determinant in domestic swine," *J Virol*, vol. 72, pp. 1028-35, Feb 1998.

[36] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, *et al.*, "GenBank," *Nucleic Acids Res*, vol. 41, pp. D36-42, Jan 2013.

[37] D. A. Natale, C. N. Arighi, J. A. Blake, C. J. Bult, K. R. Christie, J. Cowart, *et al.*, "Protein Ontology: a controlled structured network of protein entities," *Nucleic Acids Res*, vol. 42, pp. D415-21, Jan 2014.