Pu Li, Hongfang Liu and Qian Zhu

1Department of Information Systems, University of Maryland Baltimore County, USA
2Department of Health Sciences Research, Mayo Clinic, USA

Abstract
Lack of intelligent genetic testing recommendation systems that leverage a high volume of information regarding to available genetic tests, impedes widely incorporating genetic tests into regular clinical practice to further improve healthcare and reduce health cost. To fill this gap, we introduced an integrative genetic testing relevant resource called Genetic Testing Ontology (GTO), which is the first step to develop a genetic testing recommendation system, iGenetics. More specifically, we integrated multiple well-known genetic testing relevant resources, including Genetic Testing Registry (GTR), ClinVar and Human Phenotype Ontology (HPO) and formally represented it by using Web Ontology Language (OWL). A meta-ontology has been designed to capture semantics upon the nature of the selected data via Protégé. The GTO was built on top of the meta-ontology subsequently by populating specific genetic testing data. In addition, literature data from SemMedDB has been extracted and integrated into the GTO to provide scientific evidence accordingly. We performed one case study to further evaluate the GTO and demonstrate the usability of the GTO.

ABBREVIATIONS
GTO: Genetic Testing Ontology; OWL: Web Ontology Language; GTR: Genetic Testing Registry; HPO: Human Phenotype Ontology; UMLS: Unified Medical Language System; CUI: Concept Unique Identifier

INTRODUCTION
Individualized medicine, as a rapidly advancing field in healthcare, intends to make accurate predictions about a person’s susceptibility of developing disease, the course of disease, and its response to treatment based on genetic, genomic, and clinical information of individual patients. With the recent advances in genetic technology, genetic tests are available for over 2,000 rare and common conditions from over 500 laboratories [1]. Admittedly such advance introduced into the clinical practice should strongly support clinical decision-making and reduce healthcare cost, ultimately accelerate the pace of individualized medicine. However, according to two recent national surveys commissioned by UnitedHealth Group in conjunction with Harris Interactive (n=2,760; fieldwork conducted in January and February 2012) [2], physicians have not actively incorporated these tests into their regular clinical practices. About half of the physicians surveyed stated that the lack of familiarity with genetic tests is the main barrier, and over three-quarters of physicians are either somewhat or very concerned about the lack of evidence supporting the use of genetic testing. Given this barrier, this present study aims to filling the gap between a large volume of available genetic tests and insufficient usage of these genetic tests by developing an integrative genetic testing ontology, GTO. GTO will not only play a key role as a reference source to provide comprehensive information regarding to genetic tests, but also is a data foundation to support further iGenetics development.

Currently, scattered genetic testing information is presented in different formats. Most of the genetic test guidelines are published in journals that are in pdf format. GTR and Gene Reviews maintained by the National Institutes of Health (NIH) are presented in excel and html format. Such distributed storage and diverse representations lack interoperability and capabilities for data integration and automated inference of novelfindings. How to represent the information in an integrative and transformative way, which can fully support automated genetic testing information retrieval and recommendation, is a big challenge. The Web Ontology Language (OWL) [3] as a standard ontology language for the Semantic Web, has an increased degree of connectedness and increased “ability to model coherent, linked relationships” [4] compared with conventional relational
databases. To exploit these advantages to better address the challenge of genetic test recommendation, we transformed and represented genetic testing information in OWL. This will not only express genetic testing information in a formal way to enable further data integration, such as Linked Open Data, [5] but will also support automatic novel associations inference.

In this paper, we introduce materials being applied in this study in Materials section, details about the genetic test relevant data integration and the GTO generation are described in the Methods section, and followed by the Results and Discussion sections.

**MATERIALS**

**Genetic Testing Registry (GTR)** contains comprehensive information about genetic tests offered worldwide for disorders with a genetic basis. It is maintained by the NIH. Test information is voluntarily submitted by test providers [6].

**ClinVar** provides a publically available archive of reports of relationships among medically important variants and phenotypes. "ClinVar accessions submissions reporting human variation, interpretations of the relationship of that variation to human health and the evidence supporting each interpretation" [7].

**Human Phenotype Ontology (HPO)** captures phenotypic information to support "clinical diagnostics or as a basis for incorporating the human phoneme into large-scale computational analysis of gene expression patterns and other cellular phenomena associated with human disease [8]."

**SemMedDB** includes semantic predictions extracted from the PubMed abstracts by the NIH. Currently it contains information about approximately 70 million predications from all of PubMed citations [9].

**METHODS**

This study was designed to integrate well-known genetic testing relevant resources and formally represent it in OWL, called GTO, which will fully support data integration and automated semantic inference towards individual genetic test recommendation. Three steps have been performed accordingly, 1) data integration with genetic testing relevant data and scientific evidence data, 2) meta-ontology design to capture semantic associations presenting in the integrated genetic testing data, 3) GTO generation. More details will be described in the following sections. The general framework of the GTO is shown in Figure 1.

**Genetic testing data integration**

Data integration plays a critical role to manipulate data across multiple data resources. In this preliminary study, three genetic testing relevant data resources, namely GTR, ClinVar, HPO, and one literature data, SemMedDB have been selected and integrated to generate a comprehensive genetic testing data repository.

**Data Normalization:** To facilitate data integration, the concepts included in the aforementioned data resources have been represented in a standard form with UMLS. Two steps have been performed for UMLS mapping, 1) utilizing the existing UMLS CUIs available at the above resources, such as GTR, ClinVar and SemMedDB, 2) employing MetaMap to map those concepts from the HPO to UMLS. In addition, genes are also represented with the HUGO Gene Nomenclature Committee (HGNC) [10] gene symbol; diseases and clinical features are represented with OMIM identifiers along with their generic names; and genetic tests are represented with GTR identifiers and test names. The associated scientific evidence is presented with PubMed ID (PMID).

**Data integration:** Once data normalization accomplished, data integration is getting more straightforward. We mapped...
the relevant concepts from the four resources based on the UMLS CUIs. For those concepts without UMLS CUIs, their generic names, such as gene symbol and disease name have been used for mapping.

**Scientific evidence integration:** As mentioned in Materials section, there are total 70 million predictions included in the SemMedDB. To speed up the data retrieval process against the SemMedDB, we extracted a subset of the SemMedDB by specifying four relevant semantic types [11], ‘dsyn’ (Disease or Syndrome), ‘gngm’ (Gene or Genome), ‘lbpr’ (Laboratory Procedure), and ‘lbtr’ (Laboratory or Test Result), which covers all semantics presented in our integrative genetic testing data. The concepts included in the SemMedDB are all mapped to UMLS, thus, UMLS mapping was employed to integrate scientific evidence to the genetic testing data.

**Meta-ontology design**

The integrated data contains information about diseases, genes, clinical features, genetic tests and references. To formally represent such information in a computable fashion, we designed a meta-ontology via Protégé [12] (Version 3.5 Build 663), which is an application “plug-in for creating OWL ontologies” [12].

**Primary class definition:** We defined five primary OWL classes, namely “Disease”, “Clinical Feature”, “Gene”, “Genetic Test” and “Reference” to categorize the integrated data in terms of their own semantics.

<table>
<thead>
<tr>
<th>Object Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>associatedWithClinicalFeature</td>
<td>relationship between &quot;Disease&quot; and &quot;Clinical Feature&quot;</td>
</tr>
<tr>
<td>associatedWithGene</td>
<td>relationship between &quot;Disease&quot; and &quot;Gene&quot;; relationship between &quot;Genetic Test&quot; and &quot;Gene&quot;</td>
</tr>
<tr>
<td>testforDisease</td>
<td>relationship between &quot;Disease&quot; and &quot;Genetic Test&quot;</td>
</tr>
<tr>
<td>hasReference</td>
<td>relationship between &quot;Disease&quot; and &quot;Reference&quot;; relationship between &quot;Gene&quot; and &quot;Reference&quot;; relationship between &quot;Genetic Test&quot; and &quot;Reference&quot;; relationship between &quot;Clinical Feature&quot; and &quot;Reference&quot;</td>
</tr>
</tbody>
</table>

**RESULTS**

**Data normalization**

As mentioned in Methods section, all concepts extracted from the four selected resources have been mapped to UMLS.
and represented by UMLS CUI. There are total 3,172 concepts with UMLS CUI originally. An extra step was performed to retrieve UMLS CUI for the concepts from the HPO via MetaMap, additionally 3,118 concepts have been retrieved their UMLS CUIs.

**Genetic testing data integration**

Four primary genetic testing data has been integrated. Table 3 shows the details about the integrated information.

**GTO generation**

Once the meta-ontology has been defined, we populated the instance data from the integrated genetic testing data repository into this ontology and generated the GTO. There are 3,062 unique diseases, 5,375 unique clinical features, 13,738 unique genetic tests, 2,214 unique genes, and 378,556 PMIDs included in the GTO. GTO can be accessed from the below link:

http://bioportal.bioontology.org/ontologies/GTO_TESTING

**“Arts Syndrome”**

Arts syndrome is a disorder that causes serious neurological problems in males. Females can also be affected by this condition, but they typically have much milder symptoms. Boys with Arts syndrome have profound sensorineural hearing loss. Other features of the disorder include weak muscle tone (hypotonia), impaired muscle coordination (ataxia), developmental delay, and intellectual disability. In early childhood, affected boys develop vision loss caused by degeneration of nerves that carry information from the eyes to the brain (optic nerve atrophy). They also experience loss of sensation and weakness in the limbs (peripheral neuropathy).

There are so few publications about this rare disease to provide necessary support for clinical diagnosis. Thus, we
Table 3: Data integration results.

<table>
<thead>
<tr>
<th>Data resources</th>
<th>Data integration results</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinVar</td>
<td>2,214 unique genes along with their gene id and gene symbol</td>
</tr>
<tr>
<td>GTR</td>
<td>13,738 unique genetic tests with test ID; 3,171 unique GTR identifiers; 3,058 unique object names; 771 SNOMET-CT ID and Gene ID</td>
</tr>
<tr>
<td>HPO</td>
<td>3,062 unique diseases; 5,421 unique HPO identifier; 5,375 unique clinical features; 3,632 unique definitions; 2,488 unique subclasses; 3,261 unique external references; 1,521 unique alternative IDs; 1,024 unique comments</td>
</tr>
<tr>
<td>Semantic Medline</td>
<td>1,378,566 PubMed citations (PMIDs)</td>
</tr>
</tbody>
</table>

Figure 4: Logic workflow for genetic test recommendation.
attempted to explore the GTO and evaluate the capability of the GTO for clinical decision support. Figure 4 is a logic workflow by leveraging domain knowledge offered by physicians and critical information retrieved from the GTO to recommend the appropriate genetic tests for an eight-year-old boy with Arts Syndrome.

Based on the clinical observations (clinical features), two diseases, "Arts Syndrome" and "Combined Oxidative Phosphorylation Deficiency" have been identified by the GTO. To further support clinical decision-making, relevant scientific references retrieved from the GTO play a key role. "X-linked ataxia, weakness, deafness, and loss of vision in early childhood with a fatal course" [13] provides more information about "Arts Syndrome", for example, a fatal course is in early childhood, and the high risk population is male. "Mitochondrial translation and beyond: processes implicated in combined oxidative phosphorylation deficiencies" [14] describes more details about "Combined Oxidative Phosphorylation Deficiency", for example, it applies to both genders and patients typically have severe liver dysfunction during the course of the disease. Therefore, more confident decision can be made that the boy may have "Arts Syndrome" as the patient is a boy and no abnormal liver function based on the reference review. Prior to making the final diagnosis, the physician would like to order one genetic test to the patient for further confirmation.

The GTO provides a list of genetic tests along with related scientific reference for "Arts Syndrome" accordingly, including "PRPS1 Related Disorders: PRPS1 Full Gene Sequencing" (Genetic Test ID 502471), and other tests (Genetic Test ID 305645, 319358 and 331864). After reviewing test descriptions through link under "Test Details" and leveraging his/her professional expertise, the physician may determine one of the best tests for this patient.

**DISCUSSION**

In this study, we successfully developed a genetic testing ontology, GTO by integrating public genetic testing relevant data resources. The aim of the GTO development is not only gathering information about the available genetic tests for supporting regular clinical practice, but also eventually improve health information communication to meet clinical needs. We discuss the benefit of the GTO gained from this study, the limitations of this study observed and the future plan proposed as below.

In this study, we were focusing on a centralized genetic testing ontology development. Three major public genetic testing resources, GTR, ClinVar and HPO that provide critical information about genetic tests have been integrated in the GTO. Meanwhile, to strengthen the associations presented in the GTO, scientific evidence extracted from the SemMedDB has been integrated into the GTO. The capability of the GTO for assisting in clinical decision making has been demonstrated in the case study "Arts Syndrome" section. We realized that in order to increase the accuracy of the test recommendation and automate the recommendation process, more genetic testing relevant data should be included to enlarge the coverage of the GTO. To be specific, we will integrate more genetic testing relevant data from three aspects, 1) integrating more public available genetic testing relevant data, such as OMIM (Online Mendelian Inheritance in Man), MedGen; 2) integrating patient data extracted Electronic Medical Records (EMRs). In our previous study [15], we have successfully applied EMRs to identify common clinical features to support Wilson disease screening test recommendation. We will extend that study by identifying clinical features regarding to more genetic diseases, and then to be able to recommend more genetic tests; 3) integrating information extracted from the authorized reviews of genetic disorder, GeneReviews that provides professional review for each disease, including rich genetic testing relevant information. An initial survey has been done to extract genetic testing relevant information from GeneReviews [16], which has identified the gaps between the current NLP (Natural Language Processing) annotation tools and incomplete annotation results. For the next step, a concrete experiment will be performed to fill the identified gap and extract information from the entire set of GeneReviews and integrated into the GTO.

A list of genetic tests can be identified by querying the GTO, however, no ranking rules have been integrated into the current version of the GTO to recommend the best appropriate genetic test accordingly. In this study, we stored a list of PubMed IDs for each associations presented in the GTO as one instance of the Reference class. It will be easy to wrap such list into one ranking rule by calculating the number of co-occurrences based on PubMed reference corresponding to each relevant association. A larger number of co-occurrence indicates the association (e.g., disease-genetic test) is supported by more scientific evidence, and then such test will be ranked in the top list. Furthermore, we will design more robust ranking algorithm, e.g., PageRank [17] to identify the best genetic test.

In addition to the genetic testing data formal representation, the mission of the GTO development is to recommend the right genetic test to the right patient. The tests can be recommended via the workflow introduced in the case study section by leveraging domain knowledge and information provided by the GTO. However, automatic prediction plays a key role to identify the best tests according to the clinical observations, especially for those observations can not be found in the GTO directly. Two means we will pursue accordingly in the next step, 1) we will define appropriate inference rules and integrate into the GTO for automated semantic inference. 2) We will design prediction models on top of the GTO, consequently a genetic test recommendation system called iGenetics [18] will be designed and released.

**CONCLUSION**

In this study, we successfully developed a genetic testing ontology (GTO) to integrate major public genetic testing resources in OWL. We have demonstrated the capability of the GTO for genetic testing data representation and genetic test recommendation. In the future study, we will extend the GTO with more diverse genetic testing data including EMRs and GeneReviews and build a genetic test recommendation system (iGenetics) for genetic test recommendation on the basis of the GTO.

**REFERENCES**

2. UnitedHealth Group CfHRM. Personalized Medicine: Trends and prospects for the new science of genetic testing and molecular diagnostics.


