Original article

Annotating the biomedical literature for the human variome

Karin Verspoor1,2,*, Antonio Jimeno Yepes1,2, Lawrence Cavedon1,2,3, Tara McIntosh4, Asha Herten-Crabb5, Zoë Thomas5 and John-Paul Plazzer6

1National ICT Australia (NICTA), Victoria Research Laboratory, Level 2, Building 193, The University of Melbourne, Parkville VIC 3010, Australia, 2Computing and Information Systems Department, The University of Melbourne, Parkville VIC 3010, Australia, 3School of Computer Science and IT, RMIT University, GPO Box 2476, Melbourne VIC 3001, Australia, 4Wavi Inc., 2606 2nd Ave. #155, Seattle, WA 98121, USA, 5Department of Genetics, The University of Melbourne, Parkville VIC 3050, Australia and 6International Society for Gastrointestinal Hereditary Tumours (InSiGHT), Department of Colorectal Medicine and Genetics, The Royal Melbourne Hospital, Grattan St, Parkville VIC 3050, Australia

*Corresponding author: Tel: +61 3 8344 4902; Fax: +61 3 9348 1682; E-mail: karin.verspoor@nicta.com.au

Submitted 30 November 2012; Revised 15 February 2013; Accepted 12 March 2013


This article introduces the Variome Annotation Schema, a schema that aims to capture the core concepts and relations relevant to cataloguing and interpreting human genetic variation and its relationship to disease, as described in the published literature. The schema was inspired by the needs of the database curators of the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) database, but is intended to have application to genetic variation information in a range of diseases. The schema has been applied to a small corpus of full text journal publications on the subject of inherited colorectal cancer. We show that the inter-annotator agreement on annotation of this corpus ranges from 0.78 to 0.95 F-score across different entity types when exact matching is measured, and improves to a minimum F-score of 0.87 when boundary matching is relaxed. Relations show more variability in agreement, but several are reliable, with the highest, cohort-has-size, reaching 0.90 F-score. We also explore the relevance of the schema to the InSiGHT database curation process. The schema and the corpus represent an important new resource for the development of text mining solutions that address relationships among patient cohorts, disease and genetic variation, and therefore, we also discuss the role text mining might play in the curation of information related to the human variome. The corpus is available at http://opennicta.com/home/health/variome.

Introduction

The identification of associations between human genetic variation and disease phenotypes is a major thrust of current biomedical research. Such associations not only facilitate our understanding of the genetic basis for disease, but will open the door to personalized medicine, where treatment of patients can be tailored to their unique genetic characteristics. There are large-scale efforts to catalogue disease-related genetic variants in databases [e.g., OMIM (1), HGMD (2), the Human Variome Project (http://www.humanvariomeproject.org), as well as numerous databases for individual genes (3)]. Recent research has highlighted the need to automatically mine such information from the biomedical literature, and approaches for extraction of mutations and their associated genes from natural language text have been proposed (4–9). Other work extends the methods to relate such gene/mutation pairs to a specific disease (10). These approaches require annotated textual data for training and evaluation of text mining systems.

In this work, we introduce a schema for annotation of the biomedical literature that targets the core information relevant to genetic variation and lays the foundation for text mining of this information. This schema has been developed in collaboration with curators of the InSiGHT....
Background

The InSiGHT database

The InSiGHT is the peak professional body of health-care workers in the field of familial gastrointestinal (GI) cancer. InSiGHT aims to promote and coordinate efforts to improve understanding of the genetic basis, diagnosis, prevention and treatment of inherited forms of GI cancer. Lynch Syndrome and Familial Adenomatous Polyposis are the two main inherited GI cancer predisposition syndromes. InSiGHT maintains a database of genetic variants for both of these syndromes, but for this work, we focus on Lynch Syndrome, which is caused by mutations in the mismatch repair (MMR) genes. Worldwide, the annual incidence of Lynch Syndrome has been estimated to be 3% of colorectal cancer cases (11). The original database was established in the 1990s, with mutations reported by individual laboratories (12).

In 2008, InSiGHT began collaborating with the Human Variome Project (HVP) to improve systems and processes of variant sharing and interpretation. The HVP is a non-profit organization that coordinates efforts amongst individuals and groups to systematically share variants in publicly accessible databases. Around this time, two MMR gene databases were established independently (13, 14); each developed through extensive manual curation of published articles. Inspired by the vision of the HVP, InSiGHT merged the new databases with the existing database. The InSiGHT database uses the LOVD (Leiden Open Variation Database) platform (15). This is an open-source MySQL database and is commonly used for mutation database systems. Reports manually extracted from published literature comprise the majority of entries in the InSiGHT database (~75% of all 13,000 entries covering over 2500 mutations, based on input from the database curator), with the balance coming through direct submissions from clinics.

The database structure of LOVD has two main tables, one for patient information, and the other for mutation information. The fields in these tables have been configured for GI cancer data. An important use of the data is for variant interpretation, that is, the assessment of the clinical impact of a genetic variant. This is an active area of work for the InSiGHT interpretation committee, which is using information from the InSiGHT database and published literature to assign pathogenicity to each variant. Pathogenicity indicates the probability that a variant is causative for a given phenotype or disease. InSiGHT uses a five-class system proposed by the International Agency for Research on Cancer Unclassified Genetic Variants Working Group (16), with classes of neutral, probably neutral, uncertain, probably pathogenic and pathogenic. Such interpretation is important, as a significant proportion of variants are unclassified (upwards of 50% of variants) (17). Pathogenicity classifications can be calculated using a multifactorial Bayesian model, with the required supporting evidence found in published literature or other sources. The information necessary for interpretation of Lynch Syndrome-associated variants includes the following: tumour microsatellite instability (MSI) status and immunohistochemistry (IHC) results; variant frequency in cases and controls; and family history (e.g. does the variant cosegregate with disease?). Age and ethnicity of patients are also important elements of variant interpretation.

The InSiGHT database curation workflow and the role of text mining

An important issue with population of biomedical databases is the on-going publication of new articles. This requires continuous effort to keep the contents of the database up to date. It has been argued that text mining is required to improve the coverage of databases (18). The role of text mining in the biocuration workflow has been carefully considered by Hirschman et al. (19). The authors conducted a survey of biological database curators and identified a ‘canonical’ workflow for biocuration, including the steps of (i) document selection, (ii) indexing of documents with biologically relevant entities and (iii) detailed curation of specific relations. The InSiGHT database curation process also follows this general paradigm, with each step tailored to the specific curation goals for the database. Articles are selected initially on the basis of a search for a mention of a key gene of interest, followed by reading of the abstract to verify the relevance of the article. The final step is reading the actual article for the relevant elements of information required in the database.
annotation. Software for locating and managing the files is available, such as Reference Manager™.

The authors of (19) further identified several insertion points for text mining technologies, including for biological entity identification and normalization, and event detection. Text mining can be applied to prioritize documents for curation, and to determine what concepts (entities, events) of interest are mentioned in those documents. The survey indicated that there was strong interest both in batch processing of articles, where the automatic processing would be followed by biocurator validation, and more interactive tools integrated into their workflow. Although the most effective integration of text mining with the InSiGHT workflow is yet to be determined, the InSiGHT curators are interested in making use of text mining. Fully automatic database population may not be realistic (20) (see Discussion section), but minimally text mining can be used to identify potentially relevant information for curation, to reduce the workload for curators and ideally to enable (semi-) automatic population of the database fields with information from published sources. A tool that can highlight relevant articles and reliably identify sections or sentences where relevant information can be found, to be manually reviewed for curatable information, would already be a great advance in reducing the workload of curators to reading a few key paragraphs or sentences.

Karamanis et al. (21) have shown that such support tools for FlyBase curation improved navigational efficiency for curators by ~58%.

This project began as an attempt to extract important types of information relating to Lynch Syndrome and its genetic underpinnings in the MMR genes. A secondary goal is to extract information to be used for the purpose of variant interpretation.

In the context of the InSiGHT database, and for genetic variant databases more broadly, there are several key pieces of information that would be highly valuable to recognize in the published literature:

(i) mentions of mutations (variation) in genes of interest;
(ii) mentions of a patient with the variant(s);
(iii) the patient’s disease status and demographic information;
(iv) for a given published study, frequency information for each genetic variant in cases/controls or the number of individuals with the variant.

Our schema therefore targets this set of information, as we will detail below. The schema has been applied to a corpus of biomedical journal articles, producing a novel resource that contains entity and relation annotations relevant for understanding genetic variation. Significantly, we have annotated many relation types that have never, to our knowledge, been included in an annotated biomedical text corpus.

Methods

We have designed the schema proposed in this work to be more broadly applicable than the specific needs of the InSiGHT database. As such, the schema—and any text mining tools that may be built based on the schema and the annotated text data—targets the goal of identifying potentially relevant information for curation of genetic variation and its relationship to disease. This includes genomic categories (e.g. gene, mutation), phenotypic categories (e.g. disease, body part) and categories related to the occurrence of mutations in disease (e.g. cohort size, age, ethnicity). In addition, the schema was designed to support eventual annotation of information for the purpose of supporting variant interpretation, captured in a broad category called characteristic. We did not explicitly target the existing structure of the InSiGHT database in designing the schema; we will consider how the schema aligns to that database in the Discussion section.

The variome annotation schema

We refer to the schema as the Variome Annotation Schema. In total, 11 entity types and 13 relation types were selected for annotation. The first version of the Variome Annotation Schema was constructed by analysing the database schema for the existing InSiGHT mutation database; further categories and relations were added based on discussions with the InSiGHT database curator, who suggested additional useful information to capture. Initial guidelines were prepared for all categories and relations, describing the intended interpretation for each of those along with examples and counter-examples.

The entity types annotated are as follows:

- **Gene**: A segment of DNA that codes for a protein.
- **Mutation**: A mutation is an alteration (deletion, insertion, substitution) of nucleotides (DNA, RNA) or amino acids (Protein).
- **Body part**: An organ or anatomical location in a person.
- **Disease**: An abnormal condition affecting the body of an organism.
- **Patient**: An individual with a disease.
- **Cohort**: A group of people; specifically any group or population of people that may be assigned a disease or characteristic. This could range from two people, e.g. two siblings, to thousands (e.g. cases or controls).
- **Size**: A number indicating the number of people in a cohort, or the number/frequency of a mutation.
- **Age**: A number or range indicating how old a person/group of people is.
- **Gender**: Terms indicating whether someone is male or female.
- **Ethnicity or Geographical Location**: Terms indicating where a person/group of people comes from, either based on ethnic origin or where they live.
- **Characteristic:** A characteristic of disease or tumour, in the sense of a property or feature that commonly occurs in or is associated with that disease or tumour. Such information is relevant to variant interpretation. For example, MSI is commonly seen in Lynch Syndrome-associated tumours.

The relation types annotated are as follows:

- **Gene has Mutation:** A mutation occurs in or near a gene, usually at a given position.
- **Patient/Cohort has Mutation:** A patient or cohort has a specific genetic variation.
- **Mutation related to Disease:** A mutation is associated with (or causes) a disease.
- **Mutation has Size:** Indicates the number or frequency of mutations.
- **Disease has Characteristic:** A characteristic of a disease/tumour.
- **Disease related to Gene:** A disease is associated with a gene—that is, a gene (when mutated) is linked to, or causes a disease.
- **Disease related to Body Part:** A disease may occur in a body part, or have a body part in its name.
- **Patient has Age:** A patient has a given age.
- **Cohort has Age:** A summary age for a cohort. Often listed as a mean or an age limit.
- **Patient/Cohort has Gender:** A patient or cohort is male or female.
- **Patient/Cohort has Ethnicity/Geographic Location:** A patient or cohort has a given ethnicity or lives in a given place.
- **Patient/Cohort has Disease:** A patient or cohort has a disease.
- **Patient/Cohort has Characteristic:** A characteristic associated with a patient or cohort.
- **Cohort has Size:** The size of a cohort group.

Here, we consider a relation to be a predicate plus its typed arguments, following the mathematical notion of a relation as a function that relates two defined classes.

The complete Variome Annotation Schema Guideline document, which includes detailed annotated examples, is available as Supplementary File S1.

Note that although mutation-relatedTo-disease and gene-relatedTo-disease are superficially similar, they reflect different granularities of the information about a gene that is associated with a disease. Accordingly, a phrase such as 'an estimate of six mutations to colorectal cancer' represents a mutation-relatedTo-disease relation in the absence of a gene mention, while 'rectal tumours have a relatively higher frequency of K-ras mutations in codons 12 and 13' contains a gene-relatedTo-disease relation connecting 'K-ras' and 'tumours' as well as a gene-has-mutation relation connecting 'K-ras' and 'mutations in codons 12 and 13'.

A mutation-relatedTo-disease relation could be inferred from those two propositions. Such similar relations are included to enable coverage of a range of linguistic patterns for expressing similar information, and for capturing as many specific propositions as possible.

**Constructing an annotated corpus**

The document annotation process consisted of three main phases:

(i) Selecting a set of documents to be annotated, and to act as the corpus;
(ii) Preparing the documents for annotation, including pre-processing, and loading them into the annotation tool, BRAT;
(iii) The actual annotation phase.

**Document selection.** Documents were firstly selected for annotation based on (some) relevance to the subject topic area. This was done using PubMed Central’s to loosely identify documents relevant to the genetics of Lynch syndrome, which covers inherited colon cancer as well as certain other cancers. This was done by using a search query consisting of the three most common Lynch syndrome genes: ‘MLH1 or MSH2 or MSH6’. This search strategy was selected to emphasize the mutation focus of the corpus, rather than a focus on the disease itself. High specificity of the query was not important: since our Schema and Guidelines are generic, we tolerated (indeed welcomed, for diversity of coverage) some documents that were outside the strict subject area. Other than the choice of the searched genes, the selection of articles was not directly targeted to the InSiGHT database, i.e. articles were not filtered for existing annotated data in InSiGHT.

Next, we downloaded only articles that were available as an open access full text publication through PubMed Central. Open access articles have been shown to be representative of the broader literature (22). Moreover, the BRAT annotation tool (23) requires articles in text form, so we retained only those articles available in HTML or XML format. As of January 2013, the PubMed query returns 4458 articles, with 1734 available in the PubMed Central Open Access collection. Articles were selected randomly from amongst the set available when the corpus was established in late 2011. For reference, there are currently 483 PubMed IDs referenced in the InSiGHT database, with only 17 available in the open access collection. Selected articles were annotated in numeric order by PubMed Central ID.

**Document preprocessing.** As mentioned above, annotation was performed using the web-based BRAT annotation tool, which supports structured annotations. Before loading the documents into BRAT, each document was split into multiple files, each major section in a different
file, to counter performance issues with BRAT over large files. Some sections were removed (i.e., those not containing relevant content, such as Author Contributions, References, etc.); those to be included were converted into plain text.

Finally, the uploaded documents were automatically pre-annotated. A number of simple regular expressions were used to identify simple clear likely occurrences of annotation schema categories. For example, expressions corresponding to the Lynch Syndrome gene names were used to annotate those items as gene, and the expression ‘[0–9]+ years? old’ (plus more like this) was used to detect likely instances of the category age. The MutationFinder tool (5) was used to detect (likely) occurrences of mutations. These pre-annotated files were then made available to the annotators, with the annotators being able to modify any auto-annotations that they considered to be incorrect.

**Annotation process.** The Annotation phase was performed by two main annotators, each a final-year undergraduate Genetics student, using the BRAT tool; Figure 1 shows a screenshot of the tool with an annotated document from our corpus. The BRAT tool supports entity annotation through selection of a span of text by keeping the left mouse button down while dragging the cursor across the span, or by double-clicking a word. A predefined set of entity types, from the Schema, are available to label the annotation. Relations are added by clicking on one entity and dragging the mouse pointer to the other entity. Again, only relation types specified in the Schema are available to label the annotation. The type constraints of each relation are checked against a configuration file; arbitrary relations are not allowed. BRAT has some limitations that placed restrictions on the Annotation Guidelines: e.g. entities must be continuous and cannot be split over multiple lines. Our Guidelines were updated to reflect such limitations.

Using the initial Guidelines document, all project team members (both annotators, the database curator and InSiGHT project member, and the Language Technology researchers) jointly annotated the abstract of a single article: the abstract was selected to be dense with

---

**Figure 1.** A screenshot of the BRAT tool (23) being used to annotate a document in the InSiGHT corpus.
annotation categories. This exercise was designed to immediately identify any problematic or unclear guidelines, which were then corrected or clarified. The initial annotation phase then involved the two annotators annotating five full articles, according to the Guidelines; the resulting annotated documents were examined for agreement between the annotators, and particularly for any differences in the way categories were filled. Such disagreements were resolved via meetings involving all team members; any disputes were resolved by the curator of the existing database. The articles were re-annotated, and the Guidelines document was updated to reflect the resolutions and clarifications to differences in interpretation between the annotators.

Following this initial phase, the annotators were given five further articles to double-annotate to verify agreed interpretation of all annotation categories and relations. Each annotator had some further questions during this second phase—these were quickly resolved and the Guidelines clarified where appropriate.

Having verified acceptable inter-annotator agreement on this set, the remaining articles were divided amongst the two main annotators, with each article being assigned one annotator. Other minor modifications to the Guidelines and interpretation of Schema categories were made during the formal annotation phase, whether raised by one or both annotators; these were again resolved by discussion, with final resolution left to the curator. After such a clarification, one or both annotators would revisit any articles they had already annotated to ensure their use of that category reflected the updated Guidelines.

Results

To date, 10 journal articles (listed in Table 1) have been (doubly) annotated following the Variome Annotation Schema, and 21 additional (singly annotated) articles will be ready soon. The entity and relation annotations are stored using the file format representation of the BioNLP Shared Task (http://2011.bionlp-st.org/home/file-formats). The current corpus of 10 journal articles is split into 120 units defined by article sections and contains 42921 words. Corpus annotation, after the annotator training phase and with no more revisions to the Guidelines, requires ~4 h per article. The corpus consisting of the 10 doubly annotated articles, with the rest of the corpus to follow, is available at http://opennicta.com/home/health/variome.

To evaluate the consistency of the corpus annotations, we measure inter-annotator agreement (IAA) over the articles annotated by both annotators. Note that we measured IAA after the annotators reviewed all their annotations after any modifications to the Guidelines—i.e. the reported IAA measurements reflect final document annotations consistent with the final agreed Guidelines document.

Although the kappa statistic (24) is typically used to measure IAA, it cannot be applied in our case, as it requires estimating the ‘random distribution’ based on a negative set of annotations that is not available (25). Therefore, we use F-measure (F1 score), using the standard formulas (TP = True Positives, FP = False Positives, FN = False Negatives, Precision = TP/(TP+FP), Recall = TP/(TP + FN), F1 = (2 * Precision * Recall)/(Precision + Recall)). Since F-measure is symmetric, it captures the results of comparing the annotations from one annotator with the other.

When comparing the annotation of entities between the two annotators, there is agreement if each annotator annotates the same entity: i.e., both the textual span (begin/end boundaries) of the annotated entity and the entity type match. Statistics on the agreement of entity annotations is available in Table 2. We find that there is broad agreement in the annotation of entities. Many of the differences are due to boundary mismatches that have been automatically resolved. Boundary mismatches were typically related to more specific annotation by one of the annotators—e.g., one annotator selected the phrase ‘FAP cancers’ while the other only annotated the substring ‘cancers’. Table 2 also shows agreement for the case where the annotation boundaries are relaxed, i.e., where two entity annotations of the same type overlap a given span of text but do not have exactly matching begin/end points, and this shows even higher agreement. We discuss boundary differences further in the Discussion section.

The relations have three components: the type of relation (has or relatedTo) and two arguments filled with annotated entities. We consider there to be agreement if there is an agreement on the relation type itself, as well as agreement on the arguments. The direction of the relation is not relevant for the comparison (e.g. gene-has-mutation is the same relation as mutation-has-gene). The entity types cohort and patient have been merged, as they refer, in

<table>
<thead>
<tr>
<th>Table 1. The articles included in the doubly annotated portion of the human variome corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed ID</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>16202134</td>
</tr>
<tr>
<td>16356174</td>
</tr>
<tr>
<td>16403224</td>
</tr>
<tr>
<td>16426447</td>
</tr>
<tr>
<td>16879751</td>
</tr>
<tr>
<td>16982006</td>
</tr>
<tr>
<td>16879389</td>
</tr>
<tr>
<td>18257912</td>
</tr>
<tr>
<td>18433509</td>
</tr>
<tr>
<td>21247423</td>
</tr>
</tbody>
</table>
practice, to the same entity type (a cohort of size 1 is a patient). This reduces the number of candidate relations to be checked.

Since relation agreement relies on entity agreement, the relation agreement numbers shown in Table 3 are lower than for entity annotation. Many of the disagreements are due to boundary disagreements and to different interpretations of the guidelines. The disagreements can in many cases be automatically resolved, first by resolving the entity annotation disagreements and then by adding the missing relations that are based on those entities. We therefore developed a set of rules to produce a merged set of annotations. These rules follow the annotation guidelines and the advice of the InSiGHT database curator. For most disagreements, entities annotated by

<table>
<thead>
<tr>
<th>Table 2. Entity annotation statistics</th>
<th>Annotator 1</th>
<th>Annotator 2</th>
<th>Strict boundary match</th>
<th>Relaxed boundary match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agreed</td>
<td>F-measure</td>
<td>Agreed</td>
<td>F-measure</td>
</tr>
<tr>
<td>Age</td>
<td>86</td>
<td>85</td>
<td>67</td>
<td>0.7836</td>
</tr>
<tr>
<td>Body-part</td>
<td>407</td>
<td>432</td>
<td>394</td>
<td>0.9392</td>
</tr>
<tr>
<td>Characteristic</td>
<td>1037</td>
<td>1035</td>
<td>849</td>
<td>0.8195</td>
</tr>
<tr>
<td>Cohort-patient</td>
<td>1189</td>
<td>1096</td>
<td>944</td>
<td>0.8263</td>
</tr>
<tr>
<td>Disease</td>
<td>1475</td>
<td>1497</td>
<td>1365</td>
<td>0.9186</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>62</td>
<td>56</td>
<td>56</td>
<td>0.9492</td>
</tr>
<tr>
<td>Gender</td>
<td>60</td>
<td>57</td>
<td>49</td>
<td>0.8376</td>
</tr>
<tr>
<td>Gene</td>
<td>918</td>
<td>1078</td>
<td>902</td>
<td>0.9038</td>
</tr>
<tr>
<td>Mutation</td>
<td>544</td>
<td>528</td>
<td>440</td>
<td>0.8209</td>
</tr>
<tr>
<td>Size</td>
<td>606</td>
<td>669</td>
<td>584</td>
<td>0.9161</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Relation inter-annotator agreement</th>
<th>Entity 1</th>
<th>Entity 2</th>
<th>Annotator 1</th>
<th>Annotator 2</th>
<th>Agreed</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>has Age Cohort/Patient</td>
<td>78</td>
<td>71</td>
<td>57</td>
<td>0.7651</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Characteristic Cohort/Patient</td>
<td>0</td>
<td>231</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Characteristic Disease</td>
<td>925</td>
<td>661</td>
<td>557</td>
<td>0.7024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Cohort/Patient Disease</td>
<td>612</td>
<td>549</td>
<td>446</td>
<td>0.7683</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Cohort/Patient Ethnicity</td>
<td>42</td>
<td>32</td>
<td>28</td>
<td>0.7568</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Cohort/Patient Gender</td>
<td>66</td>
<td>46</td>
<td>35</td>
<td>0.6250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Cohort/Patient Mutation</td>
<td>245</td>
<td>207</td>
<td>147</td>
<td>0.6504</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Cohort/Patient Size</td>
<td>599</td>
<td>617</td>
<td>545</td>
<td>0.9016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Gene Mutation</td>
<td>491</td>
<td>457</td>
<td>410</td>
<td>0.8650</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Mutation Size</td>
<td>0</td>
<td>37</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relatedTo Body-part Disease</td>
<td>392</td>
<td>390</td>
<td>337</td>
<td>0.8619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relatedTo Disease Gene</td>
<td>31</td>
<td>45</td>
<td>4</td>
<td>0.1053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relatedTo Disease Mutation</td>
<td>104</td>
<td>50</td>
<td>28</td>
<td>0.3636</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4. Merged entity type statistics</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>85</td>
</tr>
<tr>
<td>Body-part</td>
<td>465</td>
</tr>
<tr>
<td>Characteristic</td>
<td>986</td>
</tr>
<tr>
<td>Cohort-patient</td>
<td>1272</td>
</tr>
<tr>
<td>Disease</td>
<td>1700</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>62</td>
</tr>
<tr>
<td>Gender</td>
<td>62</td>
</tr>
<tr>
<td>Gene</td>
<td>1086</td>
</tr>
<tr>
<td>Mutation</td>
<td>598</td>
</tr>
<tr>
<td>Size</td>
<td>675</td>
</tr>
</tbody>
</table>
just one annotator were added to the merged set, as they generally were valid mentions missed by the other annotator. If both annotators had annotated the same entity, the largest span is preferred in most cases. Instances of the characteristic entity type, as they are modifiers, were removed if they did not take part in any relation, i.e. characteristics cannot stand alone, but rather only have meaning as an argument of a has-characteristic relation. The same is true for size annotations, which do not have meaning outside of a cohort-has-size or mutation-has-size relation. Some missing annotations were added to comply with the annotation of diseases: e.g., the occurrence of the body part ‘colon’ within the disease annotation ‘colon cancer’, which were not consistently annotated according to the guidelines.

Annotations from both annotators have been merged into a single corpus. The rules for merging the annotations are based on the analysis previously mentioned in the Results section. Table 4 shows the entity statistics for the merged set. With this merged set of entities, we have reviewed the relations. Once the entity disagreements are resolved, many relation disagreements are also resolved. We manually reviewed the disagreements and merged the relation annotations by adding the relations annotated by each annotator.

**Discussion**

**Alignment of variome annotation schema to InSiGHT**

To assess the Variome Annotation Schema for use in the InSiGHT database curation process, the database curator reviewed several of the articles in our corpus for information relevant to the database. The articles selected for the corpus had not been previously included in the database. The curator read unannotated versions of the articles and identified the core information he would typically include in the database. This information was then compared with the annotations for those same articles created by the annotators.

We find that the information about genes and mutations was in general properly identified and linked to the patient or cohort. This includes not only the identification of the cohort but also its size, thereby providing the basic curatable information about different cohort groups.

Table 5 presents a basic analysis of how the information in the Variome Annotation Schema corresponds to fields in the current InSiGHT database. While several of the annotated concepts and relations map directly to existing fields, several others do not. The mutation concept, for instance, as annotated according to the guidelines, in some cases refers to strings that contain constituents that in turn map to the distinct database fields of exon/intron number, variant name and protein change. For example, the annotation of the sentence ‘a c.1864C>A transversion in exon 12 of hMSH2 gene at the heterozygous state…leading to a proline 622 to threonine (p.Pro622Thr) amino acid substitution’, with two mutation annotations indicated with underlining, would correspond to values in database fields for the gene (hMSH2), exon (12), variant (c.1864C>A) and protein change (p.Pro622Thr). This example also shows that the annotation schema does not distinguish between DNA and protein mutations, whereas the database does. Body part maps to the Disease field of the database, though it does not have good conceptual alignment to that field, because it is the primary place in the database where disease localization is recorded. As indicated in the table, body part can also appear in the Additional Phenotype field of the database. Concept annotations such as age, gender and ethnicity can be assumed to correspond to a specific patient; this information is more

<table>
<thead>
<tr>
<th>Annotation type</th>
<th>Primary database fields</th>
<th>Other database fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>Gene</td>
<td>Additional phenotype</td>
</tr>
<tr>
<td>Mutation</td>
<td>(Exon/intron number, variant name, protein change)</td>
<td>Additional phenotype</td>
</tr>
<tr>
<td>Disease</td>
<td>Disease</td>
<td></td>
</tr>
<tr>
<td>Body part</td>
<td>Disease</td>
<td></td>
</tr>
<tr>
<td>Mutation-has-size</td>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>Age, patient-has-age</td>
<td>Patient age</td>
<td></td>
</tr>
<tr>
<td>Gender, patient-has-gender</td>
<td>Patient gender</td>
<td></td>
</tr>
<tr>
<td>Ethnicity, patient-has-ethnicity</td>
<td>Ethnicity</td>
<td>Geographic location</td>
</tr>
<tr>
<td>Cohort + cohort-has-size</td>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>Characteristic</td>
<td>MSI (microsatellite instability)</td>
<td>IHC (immunohistochemistry)</td>
</tr>
<tr>
<td>N/A</td>
<td>Functional assay</td>
<td>Functional assay result</td>
</tr>
<tr>
<td>N/A</td>
<td><em>In silico</em> prediction</td>
<td><em>In silico</em> result</td>
</tr>
</tbody>
</table>
reliable if a specific relation involving a patient is identified. Ethnicity in the Variome Annotation Schema is ambiguous; we do not discriminate between Ethnicity and Geographic Location, though this distinction exists in the database schema, and therefore that concept may map to either field. Some characteristics correspond to the database fields of MSI and IHC. Cells labelled ‘N/A’ correspond to concepts that are unique to the database. In silico predictions and in vitro assay results are not included in the annotation schema due to their complexity, though they form an important part of InSiGHT’s variant interpretation process.

Several additional difficulties were identified in relating the information relevant for curation to the corpus annotation. First, all entities and relations in the article are annotated according to the schema, although they may not always be relevant to the scope of the database. For instance, in the InSiGHT database, only germline mutations are relevant due to the focus on inherited cancers. The annotation schema specifies that all mutations should be annotated; this includes somatic mutations that would not be included according to the database criteria. This suggests that an additional discrimination task to differentiate the two types of mutations might be required. Second, another relevancy issue arises in relation to the specific diseases discussed in the articles. While the articles were initially selected on the basis of genes known to be relevant to Lynch Syndrome, these genes are also discussed in the context of sporadic or other cancers or indeed cancer cell lines. Some filtering would be required to specifically meet the needs of the database curators by only highlighting genetic variants specifically relevant to the focus disease of the database.

The LOVD schema used in the InSiGHT database is designed to handle individual patient- and mutation-level information. Therefore, the database uses mutation and patient identifiers as key fields, with all other information anchored to those fields. Published articles, on the other hand, often report on multiple patients in a summary, rather than specific cases. This summary information cannot be directly mapped to a database record in the current database structure. Furthermore, published articles may discuss e.g., Lynch Syndrome patients in general, without highlighting a specific mutation. Again, without a concrete mutation to tie the information to, it is not possible to record this information in the database. On the other hand, the generic information about those patient groups that the Variome Annotation Schema targets is potentially useful for understanding Lynch Syndrome even without a specific variant mention.

Some of these difficulties could be overcome by a post-annotation filtering step to exclude unwanted data on the basis of a relevancy assessment. Others can be addressed through an alteration to the database schema to increase the type of information allowed. For example, summary information could be included in addition to individual patient data.

A specific challenge to text mining that arises from this analysis is that several of the key mutations in one of the articles (PubMed ID 18257912/PubMed Central ID 2275286) appear (only) in a table. The information in tables was not in scope for the annotators; the annotation was limited to information appearing in the main text ("prose" sentences of natural language) of the article. Therefore, this information was missed entirely in the annotation. Text mining of this information will require analysis of the content of tables in articles; semantic interpretation of tables is a difficult problem (26, 27).

Despite the discrepancies and challenges we have identified, we remain convinced that tools developed on the basis of the corpus can be deployed in the context of InSiGHT database curation. As suggested above, tools that can highlight relevant articles and reliably identify relevant information in those articles, to be manually reviewed for curatable information, would help greatly to reduce curator workload. Fully automated database population is not required in order for the tools to be useful; computationally assisted curation would already make a large difference. Our analysis suggests that the data annotated with the Variome Annotation Schema would facilitate progress towards such useful tools.

Analysis of annotation agreement

In general, entity annotation agreement on the corpus is quite high and therefore will serve as reliable example data. We have reviewed entity types for which the agreement is lower than 0.9 by F-measure. Many disagreements are boundary disagreements or entities overlooked by one of the two annotators. Examples of disagreement have been extracted and examined by the InSiGHT database curator to understand and resolve them. Disagreements in the age entity type are due to terms being annotated that do not directly denote age, such as ‘at older age’, ‘earlier in life’, ‘very early in life’. For the cohort-patient entity type, many disagreements are due to disagreements in the boundary of the entity annotation (e.g. ‘Chinese’ versus ‘Chinese population’, or ‘MSI-H CRC’ versus ‘CRC’ or ‘seven cases’ versus ‘cases’ alone). Another disagreement example involves the annotation of relatives of a patient (e.g., a patient’s mother or father), which in some cases carry a relevant mutation and should be annotated. Examples of boundary disagreement for the gender entity annotations include an annotation of the phrase ‘proband’s father’ rather than ‘father’ alone; in this case ‘father’ is the only word denoting the gender and so the shorter annotation is preferred. Finally, examples of boundary disagreements for the mutation entity type are related to specificity of the annotation. In the following mutation examples, the largest span should be annotated to better

As mentioned previously, the agreement on annotation of the characteristic entity type is lower than for other entities. Examination of these annotations revealed that this is due to a lack of a fully coherent semantic definition in the Guidelines. That is, the notion of a ‘characteristic’ or ‘property’ of something could apply to nearly anything that is associated to the entity. To obtain a clearer idea of the kinds of terms that in practice have been annotated as characteristics, we manually mapped each characteristic annotation to a UMLS\textsuperscript{R} Semantic Group (28) (using judgment to select the closest group). The statistics of the resulting mapping are shown in Table 6; all characteristic annotations map to one of four Semantic Groups, with most belonging to either ‘Concepts & Ideas’, ‘Disorders’ or ‘Physiology’. In Table 7, we show the relation statistics of the merged set with the characteristic category split into the UMLS Semantic Groups. We see, for instance, that cohorts/patients tend to be associated with ‘Disorder’ characteristics more often than other kinds of characteristics. These semantic groups can be used to guide the selection of appropriate information for inclusion in a characteristic annotation. That is, the semantic groups could be used to refine the definition of characteristic to an entity from one of the four groups. If an annotation attempts to label some piece of information that falls outside of one of the four semantic groups as a characteristic, it can be flagged as not satisfying the semantic constraints, or at least requiring review. Furthermore, these semantic groups could provide a way to recognize characteristics more generically: a term in an article that is recognized as belonging to one of these groups can be highlighted as potentially relevant for describing a cohort or disease. This analysis is an attempt to ground the notion of a characteristic to concepts from an existing semantic resource.

Examination of the relation agreement in Table 3 reveals that there are some relations which have no agreement (indicated as “−” F-measure). These are relations that were only annotated by one of the annotators. This could be due to misinterpretation of the Guidelines. For instance, one of the annotators may not have understood that characteristics can be associated with cohorts or patients. The relation mutation-has-size was added late in the annotation process, and one of the annotators (annotator 1) was unable to review the files to add it; therefore, we cannot assess agreement on this relation. This has additional implications for the annotation of size: since size has to be related to either a mutation, cohort or patient entity, annotator 2 produced a larger set of size annotations.

In addition, there are relations with very low agreement. Examples of these relation types are disease-relatedTo-gene and disease-relatedTo-mutation. The main reason

<table>
<thead>
<tr>
<th>Table 6. Mapping of ‘characteristic’ to UMLS semantic groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic group</td>
</tr>
<tr>
<td>Concepts and ideas</td>
</tr>
<tr>
<td>Disorders</td>
</tr>
<tr>
<td>Phenomena</td>
</tr>
<tr>
<td>Physiology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7. Frequency of relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Concepts and ideas</td>
</tr>
<tr>
<td>has Disorders</td>
</tr>
<tr>
<td>has Disorders</td>
</tr>
<tr>
<td>has Disorders</td>
</tr>
<tr>
<td>has Disorders</td>
</tr>
<tr>
<td>has Disorders</td>
</tr>
<tr>
<td>has Phenomena</td>
</tr>
<tr>
<td>has Phenomena</td>
</tr>
<tr>
<td>has Phenomena</td>
</tr>
<tr>
<td>has Phenomena</td>
</tr>
<tr>
<td>has Physiology</td>
</tr>
<tr>
<td>has Physiology</td>
</tr>
<tr>
<td>has Physiology</td>
</tr>
<tr>
<td>has Physiology</td>
</tr>
<tr>
<td>has Physiology</td>
</tr>
<tr>
<td>has Age</td>
</tr>
<tr>
<td>has Body-part</td>
</tr>
<tr>
<td>has Body-part</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Cohort-patient</td>
</tr>
<tr>
<td>has Disease</td>
</tr>
<tr>
<td>has Gene</td>
</tr>
<tr>
<td>has Mutation</td>
</tr>
<tr>
<td>relatedTo</td>
</tr>
<tr>
<td>relatedTo</td>
</tr>
<tr>
<td>relatedTo</td>
</tr>
</tbody>
</table>
for these discrepancies is that instances of the relation in
text are simply missed by one of the annotators. There are a
large number of annotations that have been correctly iden-
tified and this has been resolved by merging the relations
from both annotators into the final annotation set. The data in Table 7 shows the total number of relations after
merging the work of both annotators, coupled with the
breakdown of the characteristic type into more specific
UMLS Semantic Group categories.

Text mining variant analysis
The corpus we have annotated following the Variome
Annotation Schema introduced in this article will serve as
an important resource for training and evaluating text
mining tools that target information extraction of genetic
variation and its relationship to disease. The use of the
corpus for this purpose will be explored in detail in future
work. While the articles selected for inclusion in our corpus
are derived on the basis of some association to Lynch
Syndrome, the entity and relation types we have targeted
for annotation are also generally applicable to genetic vari-
ation in other disease contexts.

Existing text mining tools
There has been some prior effort relevant to text mining
for genetic variation which we review briefly here. Several
systems have addressed identification of mutations in text,
as reviewed in (9), including the system MutationFinder
that we used for pre-annotation of mutations (5). These
tools typically ignore splice-site mutations, insertions, dele-
tions, stop codons and frame shifts; they focus on single
point mutations. More recent work attempts to identify
the functional impact of such mutations, e.g., the effect
of a protein mutation on kinetic properties or protein sta-
bility (8, 9). The Extractor of Mutations (EMU) tool identifies
mutations and their associated genes related to Breast and
Prostate Cancers (10). The Mutator tool (7) uses regular
expressions to recognize mutations and was tested on muta-
tions related to Fabry disease. The LEAP-FS system aims to
recognize all protein amino acid mentions in text, including
mutations but also bare mentions (29), and subsequent
work with that tool addresses identifying relations
between residues and their associated proteins in text
(30) as well as functional classification of those residues as
catalytic (31).

Other information extraction work has addressed recog-
nition of some of the other entity categories annotated in
our schema. Existing methods, usually using machine learn-
ing techniques such as conditional random fields, address
recognition of diseases (32, 33) and genes, e.g. the GENIA
(34) or ABNER (35) systems. The remaining entity types have
not been studied as thoroughly but could be annotated
using terminologies like the UMLS Metathesaurus8, for which MetaMap
(36, 37) would be a first choice. The Metathesaurus concepts are grouped into meaningful
categories like ‘Age Group’, ‘Family Group’ or ‘Population
Group’, which are relevant to some of the entity types. In
addition, we have shown how the characteristic entity
type can be mapped to UMLS Semantic Groups. Regular
expressions could be considered as well to identify the
age of cohorts and the size entity type. Other work (e.g.
38, 39) has addressed annotation of PICO (Population,
Intervention, Comparison, Outcome) or similar criteria
used in Evidence-Based Medicine. While these categories
are superficially related to our aims, that work does not
address specific patient/cohort information and relation-
ships involving these categories.

Existing work on annotation of relations addresses only
a limited number of relation types in the biomedical
domain. In addition to the work on gene/protein–mutation
relationships mentioned above, protein–protein inter-
actions and other specific events such as gene expression
and transcription have been studied in community challenges (40, 41). For many of the relation types in our
work (e.g. cohort-has-size), there is no existing work that
we are aware of. Pattern matching-based systems (42) or
machine learning (43) approaches are suitable for consider-
ation for such relation annotation.

The feasibility of automatic genetic variant
database population
Our analysis indicates that fully automated population of
a genetic variant database is not likely to be possible,
given subtle database-specific relevancy judgments that
are required. However, for some specific entity and relation
types, text mining may be suitable for initial population of
a database record.

A key feature of the EMU, Mutator and LEAP-FS systems
is that they exploit known sequence information about
genes to validate identified gene or protein–mutation
relationships; in (7, 10), this external knowledge is applied
as a filter after a putative gene–mutation relationship is
identified while in (30), it is used to build reliable training
data for inferring linguistic relational patterns. Such work
has highlighted the importance of this physical information
for reliable extraction of information relevant to variants.
However, in general, such information is not always
straightforward to apply, due to inconsistencies in refer-
ences to genomic coordinates and gene nomenclature
(44). These inconsistencies will need to be resolved in text
mining solutions that depend on using this information to
improve accuracy.

We note that, given nomenclature variation for muta-
tions and other relevant categories of information—
notably phenotypic and characteristic information—even
organization of disease-related mutations into a database
does not provide the final solution to easy access to
comprehensive mutation data (45). Text mining can provide
value in this context by providing tools that target clearly specified annotation schema for well-defined information types and by mapping natural language descriptions to standard nomenclature (46) or to controlled vocabulary or ontology terms, as we have done with the UMLS Semantic Groups. This provides the semantic glue that enables relating disparate information on genetic variation, enabling standardization and improved querying (20).

Conclusion

We have introduced the Variome Annotation Schema. This schema aims to capture the core information relevant to genetic variant databases, and discussed the application of that schema to a small corpus of full text publications. We found there was good inter-annotator agreement on the basic entity annotations, in particular when some relaxation of annotation boundaries is permitted, and good agreement on most relation annotations. We showed that the somewhat imprecise entity type of characteristic can be broken down into four UMLS Semantic Groups; the use of these groups will improve the consistency of annotation with the schema.

The corpus we have built will provide an important resource for building text mining systems that can support the curation of genetic variation and associated phenotypic data, for the InSiGHT database as well as other gene- and disease-specific databases. There are currently text mining tools that target some of the aspects of the Variome Annotation Schema, but several of the concepts and most of the relation types we introduce have not been previously considered for text mining. The corpus provides an opportunity to develop new tools more targeted to the needs of the context of the human variome. We will do this in future work, and make the resource available to the community when the remaining annotation is completed.

Supplementary Data

Supplementary data are available at Database Online.

Acknowledgements

Installation and configuration of the BRAT annotation tool, along with most of the document pre-processing, was performed by Lars Yencken. We also thank Sampo Pyysalo and Pontus Stenetorp and other (former) members of the Tsujii Lab at the University of Tokyo for their support of BRAT.

Funding

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The InSiGHT database curator John-Paul Plazzer was generously supported by the George Hicks Foundation in Melbourne, Australia (to September 2012) and is currently supported through The Royal Melbourne Hospital Foundation. Funding for open access charge: NICTA.

Conflict of interest. None declared.

References


