DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–9
doi: 10.1093/database/baw096
Original article

Original article

Mining clinical attributes of genomic variants through assisted literature curation in Egas

Sérgio Matos,[1],* David Campos,[2] Renato Pinho,[1] Raquel M. Silva,[1,3] Matthew Mort, David N. Cooper[4] and José Luís Oliveira[1]

[1]IEETA/DETI, University of Aveiro, Aveiro, 3810-193, Portugal, [2]BMD Software, Aveiro, Portugal, [3]Department of Medical Sciences, iBiMED, University of Aveiro, Aveiro, 3810-193, Portugal and [4]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff, UK

***Corresponding Author**: Tel: +351234370510; Email: aleixomatos@ua.pt

Citation details: Matos,S., Campos,D., Pinho,R. *et al.* Mining clinical attributes of genomic variants through assisted literature curation in Egas. *Database* (2016) Vol. 2016: article ID baw096; doi:10.1093/database/baw096

Received 4 December 2015; Revised 13 May 2016; Accepted 15 May 2016

Abstract

The veritable deluge of biological data over recent years has led to the establishment of a considerable number of knowledge resources that compile curated information extracted from the literature and store it in structured form, facilitating its use and exploitation. In this article, we focus on the curation of inherited genetic variants and associated clinical attributes, such as zygosity, penetrance or inheritance mode, and describe the use of Egas for this task. Egas is a web-based platform for text-mining assisted literature curation that focuses on usability through modern design solutions and simple user interactions. Egas offers a flexible and customizable tool that allows defining the concept types and relations of interest for a given annotation task, as well as the ontologies used for normalizing each concept type. Further, annotations may be performed on raw documents or on the results of automated concept identification and relation extraction tools. Users can inspect, correct or remove automatic text-mining results, manually add new annotations, and export the results to standard formats. Egas is compatible with the most recent versions of Google Chrome, Mozilla Firefox, Internet Explorer and Safari and is available for use at https://demo.bmd-software.com/egas/.

Database URL: https://demo.bmd-software.com/egas/

Introduction

Biological and biomedical data are continually shared through the scientific literature. Structuring this vast and ever increasing amount of information into reference resources, where it can be more easily located and used, is a very challenging and expensive task. Text-mining (TM) tools are therefore becoming a normal part of the curation pipeline, helping to expedite the work of curation teams (1–3). Nonetheless, there is still a gap between the biocuration and the biomedical TM communities, in part owing to the complexity and specific requirements of each curation task. It is important to provide expert curators with tools that take advantage of automatic TM pipelines and of existing knowledge resources, but in order to effectively support users in their daily activities, these tools must be supported by interactive and highly usable interfaces that are adapted to the curation task (4).

Following this demand, several biocuration tools have been proposed, many of which have been fostered by the BioCreative Interactive Annotation Task series (http://www.biocreative.org/tasks/biocreative-v/track-5-IAT/), which has stimulated the interaction between biocurators, text miners and tool developers, validating the positive impact of assisted curation tools and promoting their adoption. Brat (5) is one of the most popular of these tools. Brat supports in-line annotation on documents and provides concept normalization features, integration of automatic annotation services, search capabilities and document comparison. However, configuring the target concepts and relations, normalization resources and automatic services is only accessible to advanced users. Additionally, document representation may become slow when displaying full-text documents with many annotations. The tagtog annotation tool (6) also offers automatic annotation services, in addition to manual editing, for a set of pre-defined concepts. This system uses the revisions made by users to update the machine-learning annotation models, thereby improving accuracy based on user inputs. MyMiner (7) is another complete web-based solution for biocuration, supporting document triage, automatic concept recognition and document comparison. However, interpreting the annotations and their textual context is less immediate, since annotations are shown on a table rather than through in-line markup. Argo (8) offers a different approach, allowing users to design their own text analysis workflow based on integrated TM components. Thus, users are able to create custom processing pipelines for concept and relation annotation, and manually curate the results of their analysis. Even though this approach is powerful, providing a high level of flexibility, creating such workflows may require advanced expertise and a steep learning curve for biocurators. Other solutions, such as BioQRator (9), PubTator (10), RLIMS-P (11, 12) and Ontogene (13) present typical web-based interfaces with tabular listings of concept and/or relation annotations with simple highlighting and sorting capabilities. Nonetheless, some of those solutions incorporate interesting features. For instance, BioQRator integrates document triage for protein–protein interactions, PubTator features a PubMed-like interface and integrates many state-of-the-art automatic solutions for concept recognition and normalization. RLIMS-P is a more specific tool, designed to extract protein phosphorylation information using carefully designed rules and patterns for identifying concepts and relations. Neves and Leser (14) present a comparative analysis of thirteen text annotation and curation tools, based on 35 criteria; these authors concluded that although comprehensive and easy-to-use solutions exist for many use cases, no tool fully satisfied all initial criteria. The major distinguishing aspects identified by the authors were the types of annotation supported by the tool and if these are configurable, import of pre-annotated documents and integration of automatic annotation methods, support for larger annotation projects and calculation of inter-annotator agreement (IAA).

In this article, we focus on the task of identifying mentions of human genomic variants in the biomedical literature, and associating these mentions to corresponding genes, phenotypes and clinical attributes such as mode of inheritance, penetrance and zygosity. To support this task, we propose Egas, a project-oriented and customizable web platform for TM assisted literature curation that offers manual and automatic annotation of concepts and concept relations, with simple in-line representations and straightforward user interactions. Egas provides ‘annotation-as-a-service’ through a centrally managed pipeline, including document collections, users, configurations, annotations, back-end data storage and document processing and TM tools (15). Its flexible architecture, allowing the definition of concept and relation types to use on each annotation project, integration of automatic annotation web-services and the ability to import pre-annotated documents in different formats (e.g. the A1 standoff annotation (16) and BioC inline XML format (17)) facilitates adaptation to different annotation requirements. Additionally, several ontologies from NCBO BioPortal (http://bioportal.bioontology.org/) are integrated and can be configured as the normalization ontology to use for each concept type. These features give curation teams the practicality and simplicity of using an annotation service configured according to each project’s annotation guidelines.

Curation of Genomic Variant Clinical Attributes

The study of genetic variations and their association with diseases is a major focus of biomedical research (18). Several databases have been set up to curate and store these associations and related information in structured form, but automated methods to extract this information from text are essential in order to keep track of the burgeoning biomedical literature. This has prompted the development of various TM tools for identifying genetic variants in biomedical texts. Wei *et al.* (19) described tmVar, a conditional random field (CRF)-based variant extraction tool together with a corpus of 500 Medline abstracts used for training and testing the models. The authors compared their model to the rule-based system, MutationFinder (20), and showed significant improvements in terms of recognition performance when tested in their corpus as well as on the MutationFinder corpus. tmVar uses an extensive feature set and a specially crafted set of labels that allows the CRF model to identify different variation types. Additionally, a small set of post-processing rules is applied to further improve the model predictions, leading to a state-of-the-art F-score of 91.4% (87.7% when no post-processing was applied), an increase of over 13 percentage points when compared with MutationFinder. Other studies have considered not only the identification of variant mentions but also the association with disease. Doughty *et al.* (21) used their own rule-based variant extraction tool, EMU, to identify variant mentions in two sets of abstracts related to prostate cancer and breast cancer, respectively. They manually validated 51 mutations related to prostate cancer and 128 mutations related to breast cancer that had not yet been annotated in large reference databases. Ravikumar *et al.* (22) used MutationFinder to identify variants in Medline abstracts, and associated these with automatically annotated disease and protein mentions by using the dependency graph of sentences containing these three entities. They obtained an *F*-score of 64.3% when comparing to gold-standard data from the UniProt database.

Our curation task involved annotating mentions of human inherited pathogenic gene variations in Medline abstracts, as well as association to genes, diseases and clinical attributes such as inheritance mode and penetrance. The task was organized in concert with the Human Gene Mutation Database (HGMD), a comprehensive collection of germline mutations in nuclear genes that underlie, or are associated with, human inherited disease (23). By March 2016, this database contained over 183 000 different lesions detected in over 7000 different genes, with new mutation entries currently accumulating at a rate exceeding 12000 per annum. HGMD is used as a central unified disease-oriented mutation repository by human molecular geneticists, genome scientists, molecular biologists, clinicians and genetic counsellors as well as by those specializing in biopharmaceuticals, bioinformatics and personalized genomics. The public version of HGMD (http://www.hgmd.org) is freely available to registered users from academic institutions/non-profit organizations whilst the subscription version (HGMD Professional) is available to academic, clinical and commercial users under license via Qiagen Inc (http://www.biobase-international.com/product/hgmd).

The annotation task was performed on a corpus of 100 Medline abstracts selected after prioritizing, using a classifier trained with information from the documents previously used to curate information in HGMD, the 28000 results obtained from the PubMed search:

Table 1. Concepts for curation task

Group	Concept
Gene	Gene
Disease	Disease
Variation	SNV
	InDel
	Rearrangement
Clinical attributes	Mode of inheritance
	Penetrance
	Zygosity
	Age of patient (years)
	Age of onset (years)

```
genetic disease, inborn[MeSH Terms] AND (polymorphism, genetic[MeSH Terms] OR deletion[Title/Abstract] OR substitution [Title/Abstract] OR insertion[Title/Abstract] OR duplication[Title/Abstract] OR indel[Title/Abstract] OR delin[Title/Abstract] OR conversion[Title/Abstract] OR translocation[Title/Abstract] OR inversion [Title/Abstract]) AND hasabstract[text] AND humans[MeSH Terms] AND English[lang] NOT Review[ptyp] NOT "genome wide"
```

Table 1 lists the concepts defined for the task, where variations were subdivided into single nucleotide variants (SNVs), insertion/deletions (InDel) and rearrangements. Automatic entity recognition tools were used to identify mentions of these concepts, in order to facilitate and accelerate the curator's work, but the automatic annotation of relations (e.g. between a variant and a gene or disease mention) was not considered. In order to ground the annotations, concepts were assigned a concept identifier from an established ontology. The following terminologies were selected for grounding the concept annotations in this corpus:

i. Diseases—Online Mendelian Inheritance in Man (OMIM)
ii. Genes (or proteins)—HUGO Gene Nomenclature Committee (HGNC)
iii. SNVs—dbSNP

OMIM and HUGO are part of the various vocabularies available in Egas for concept normalisation. The dbSNP data, on the other hand, were specially collected and integrated for this annotation task. This required indexing all variants available in the dbSNP database, together with all Human Genome Variation Society names associated with each variant as well as the corresponding gene names, to facilitate matching the textual mention to an existing

Table 2. List of clinical attributes considered for annotating human variants

Concept	Clinical metadata	UMLS CUI
Mode of inheritance	Autosomal dominant	C0265385
	Autosomal recessive	C0441748
Zygosity	Homozygous	C0019904
	Heterozygous	C0019425
	Hemizygous	C1881036
Penetrance	Complete penetrance	C1840470
	Reduced penetrance	C1867989
	Variable penetrance and expressivity	C3276568
	Incomplete penetrance of some features	C2750454
	Incomplete, age-associated penetrance	C3280136
Age	Age of patient (years)	C0001779
	Age of onset (years)	C0206132

dbSNP identifier. Normalization of insertion-deletions and rearrangements was not considered.

To annotate the clinical attributes, we collected terms (and their synonyms) from OMIM, Human Phenotype Ontology, NCBI Metathesaurus, and NCI Thesaurus. A small ontology was created using the selected terms, and integrated in Egas. Table 2 lists the set of clinical attributes considered for this task and the Unified Medical Language System (UMLS) concept unique identifiers (CUI) used to represent the categorical values of these attributes. All synonyms for these terms were compiled in a dictionary for automatic matching in the texts.

Curation Tool

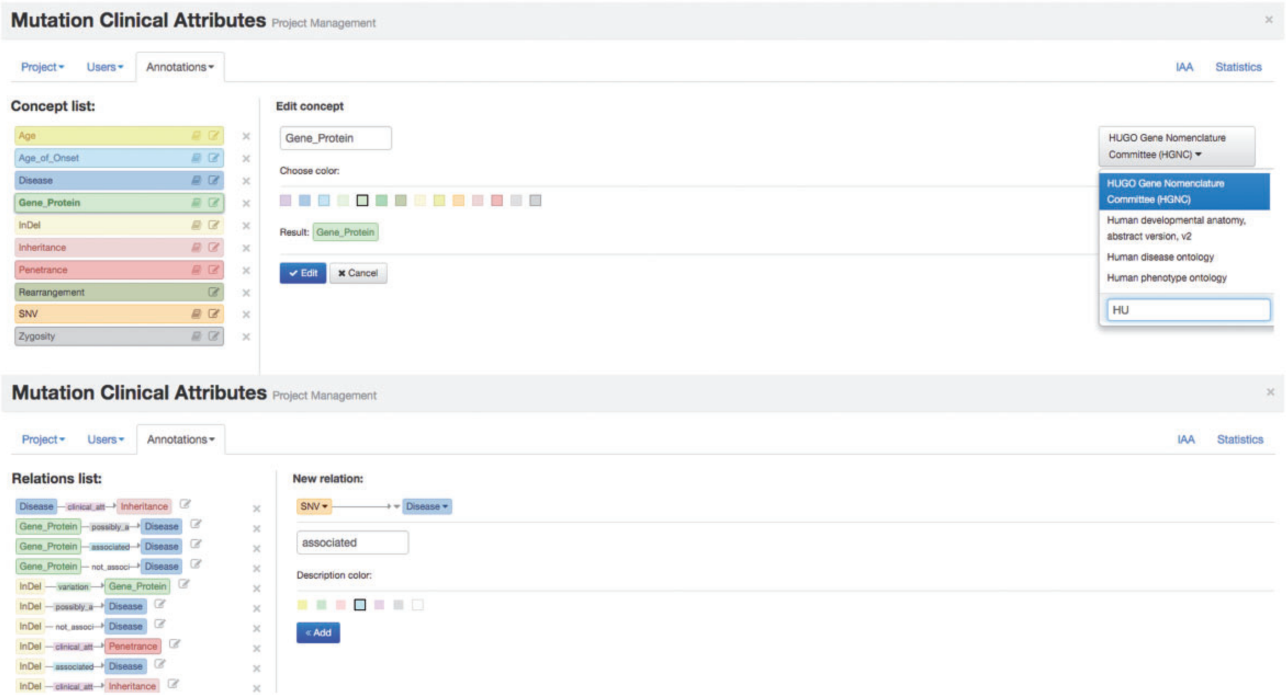
Egas organizes curation tasks in projects. Each project created in Egas has its own workspace, and comprises a curation or document annotation task, performed on a collection of documents, by a team of (one or more) curators, and considering a pre-defined set of concept and relation types defined by the curation guidelines. Project administrators have access to the administration panel that allows them to add and manage users (curators) associated with the project, import documents, and define project characteristics and annotation guidelines. The project administrator can freely define the relevant concept and relation types and the normalization ontology to use for each concept type, according to the requirements of the task. Furthermore, to facilitate the annotation work, each different concept and relation type may be associated with a markup colour (Figure 1) .

Egas supports collaborative and blind annotation projects. In the case of collaborative projects users can work on the same documents, see the changes introduced by other users, and use the project chat to discuss details of the annotation task, contributing to more consistent results. Blind annotation projects allow project administrators to assign different portions of the corpus to different users, with a configurable overlap. Administrators may evaluate the annotation consistency by visually comparing the annotations or by calculating the IAA (see 'Results' section). Figure 2 shows the 'Users' tab in the user management panel, where the project administrator can partition the corpus and assign partitions to curators. Partitions may be created manually, by defining their size in terms of proportion of the corpus, or a simple overlapping schema can be used. In this case, the administrator only needs to define the percentage of overlap between curators, and the tool generates and assigns the partitions according to the number of curators in the project. Any remaining documents are equally distributed among curators. In the example shown, an overlap of 33% is defined and each of the three curators is assigned a shared partition comprising 33% of the documents in the corpus, plus an individual partition with 22% of the corpus. To facilitate assigning or removing a partition previously assigned to a curator, the coloured partition boxes may be dragged over to (or from) a curator on the right-hand side of the panel.

Figure 3 illustrates the annotation panel in Egas. The central box displays the content of the text being curated, showing the concepts and relations that have been identified. Concepts are shown as coloured boxes, using the colours defined in the project configuration. Hovering the mouse over an annotation reveals the corresponding semantic type and normalization information. Relations are shown as lines, tagged with the relation type. Coloured boxes connected by the relation markup are placed under the concepts that participate in the relation, making it easy to identify the entire relation.

Assisted curation

The curation work in Egas may start with raw texts or with pre-processed texts, containing automatically identified concepts and relations that will be revised by the curators. This can be achieved by importing a previously processed document collection in either A1 or BioC formats, or by using concept and/or relation extraction webservices to annotate a set of documents in the collection. Likewise, annotated documents can be exported in A1 or BioC formats, allowing users to store the generated information locally in order, for instance, to add it to a local knowledgebase or for use in TM pipelines.



Mutation Clinical Attributes Project Management

Concept list: Age, Age_of_Onset, Disease, Gene_Protein, InDel, Inheritance, Penetrance, Rearrangement, SNV, Zygosity

Edit concept: Gene_Protein

Choose color:

Result: Gene_Protein

HUGO Gene Nomenclature Committee (HGNC)

Human developmental anatomy, abstract version, v2

Human disease ontology

Human phenotype ontology

HU

Mutation Clinical Attributes Project Management

Relations list:

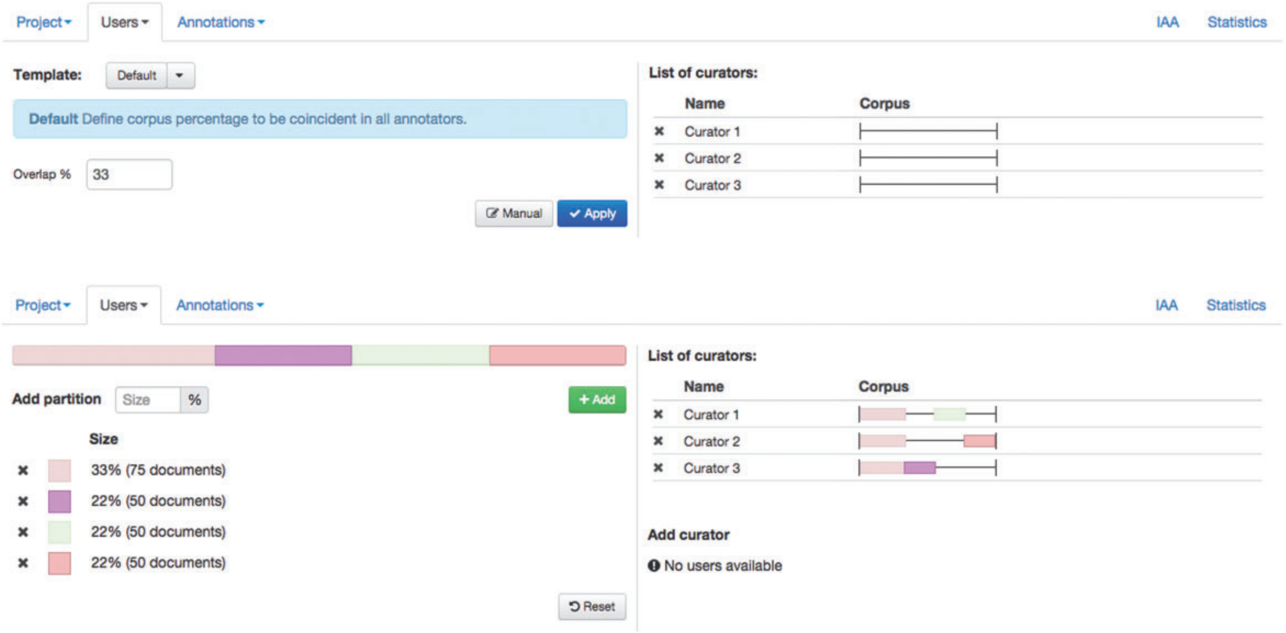
- Disease —clinical_att→ Inheritance
- Gene_Protein —possibly_a→ Disease
- Gene_Protein —associated→ Disease
- Gene_Protein —not_associ→ Disease
- InDel —variation→ Gene_Protein
- InDel —possibly_a→ Disease
- InDel —not_associ→ Disease
- InDel —clinical_att→ Penetrance
- InDel —associated→ Disease
- InDel —clinical_att→ Inheritance

New relation: SNV → Disease

associated

Description color:

Figure 1. Egas administration panel illustrating the definition of concept type ‘Gene_Protein’ linked to HGNC for normalization, and of the ‘associated’ relation between ‘SNV’ and ‘Disease’ concepts.



Template: Default

Default Define corpus percentage to be coincident in all annotators.

Overlap % 33

List of curators:

	Name	Corpus
✖	Curator 1	
✖	Curator 2	
✖	Curator 3	

Add partition Size %

		Size
✖	▨	33% (75 documents)
✖	▨	22% (50 documents)
✖	▨	22% (50 documents)
✖	▨	22% (50 documents)

List of curators:

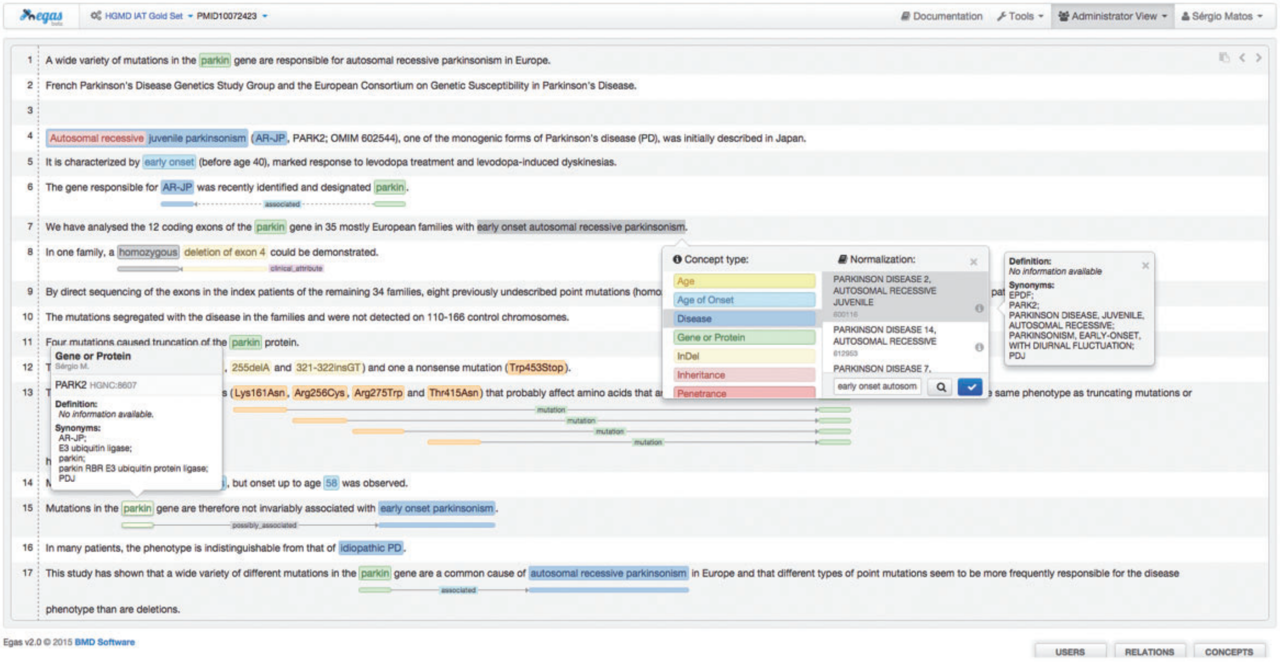
	Name	Corpus
✖	Curator 1	
✖	Curator 2	
✖	Curator 3	

Add curator

❶ No users available

Figure 2. Administration interface showing the definition and assignment of corpus partition to curators.

For this task, documents were pre-annotated with Neji (24), a biomedical concept annotation framework, and imported to Egas in A1 format. Genes and SNVs were identified through CRF machine-learning models, whereas mentions of diseases and variant’s clinical attributes were identified through dictionary-matching using terms from the UMLS ‘Disorders’ semantic group and from the clinical attributes ontology described earlier. The dictionary-based annotation of diseases achieved an *F*-score of 85.0% (24) when evaluated on the NCBI Disease corpus (25). The gene recognition model was trained and evaluated on the BioCreative II Gene Mention corpus (26), and achieved an *F*-score of 87.5% on the test set of that corpus (27). For SNVs, we used



1 A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe.

2 French Parkinson's Disease Genetics Study Group and the European Consortium on Genetic Susceptibility in Parkinson's Disease.

3

4 Autosomal recessive juvenile parkinsonism (AR-JP, PARK2; OMIM 602544), one of the monogenic forms of Parkinson's disease (PD), was initially described in Japan.

5 It is characterized by early onset (before age 40), marked response to levodopa treatment and levodopa-induced dyskinesias.

6 The gene responsible for AR-JP was recently identified and designated parkin.

7 We have analysed the 12 coding exons of the parkin gene in 35 mostly European families with early onset autosomal recessive parkinsonism.

8 In one family, a homozygous deletion of exon 4 could be demonstrated.

9 By direct sequencing of the exons in the index patients of the remaining 34 families, eight previously undescribed point mutations (homo pat

10 The mutations segregated with the disease in the families and were not detected on 110-166 control chromosomes.

11 Four mutations caused truncation of the parkin protein.

12 T , 255delA and 321-322insGT) and one a nonsense mutation (Trp453Stop).

13 T s (Lys161Asn, Arg256Cys, Arg275Trp and Thr415Asn) that probably affect amino acids that ar e same phenotype as truncating mutations or

h

14 M , but onset up to age 58 was observed.

15 Mutations in the parkin gene are therefore not invariably associated with early onset parkinsonism.

16 In many patients, the phenotype is indistinguishable from that of idiopathic PD.

17 This study has shown that a wide variety of different mutations in the parkin gene are a common cause of autosomal recessive parkinsonism in Europe and that different types of point mutations seem to be more frequently responsible for the disease phenotype than are deletions.

Gene or Protein
Sérgio M.
PARK2 HGNC:8607
Definition:
No information available.
Synonyms:
AR-JP;
E3 ubiquitin ligase;
parkin;
parkin RBR E3 ubiquitin protein ligase;
PDJ

PARKINSON DISEASE 2, AUTOSOMAL RECESSIVE JUVENILE
600116
PARKINSON DISEASE 14, AUTOSOMAL RECESSIVE
612953
PARKINSON DISEASE 7,

early onset autosom

Definition:
No information available
Synonyms:
EPDF;
PARK2;
PARKINSON DISEASE, JUVENILE, AUTOSOMAL RECESSIVE;
PARKINSONISM, EARLY-ONSET, WITH DIURNAL FLUCTUATION;
PDJ

Egas v2.0 © 2015 BMD Software

Figure 3. Egas annotation interface illustrating the addition of a new concept annotation of type ‘Disease’, and its normalization to an OMIM concept. A concept information tooltip is shown when hovering the mouse pointer over an existing annotation.

Neji and the tmVar corpus to train our own CRF model, achieving an *F*-score of 86.0%.

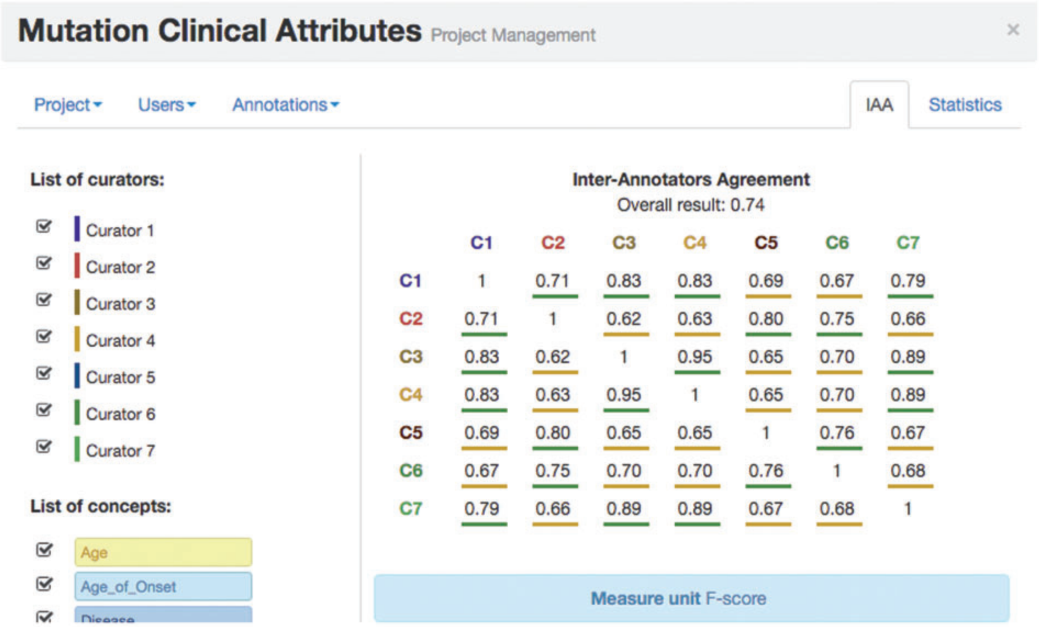
Results

Seven curators were asked to annotate documents that were pre-analyzed by the automatic concept recognition tool (half of the corpus), and raw documents (the remaining corpus), in order to evaluate the added benefit of TM-assisted curation. Curators had to revise the automatically generated annotations, correct any erroneous concept annotations and add missing ones, normalize the concept mentions, and add associations between the identified concepts. For the raw documents, curators had to add all the concept, normalization and relation annotations. The tool recorded the time taken by each curator to curate each document, as well as the number of annotated concepts and relations.

Three curators annotated the complete corpus while two other curators followed a time-limited work plan, that is, they worked on the curation task for a total of four hours, half of the time curating pre-annotated documents and the other half working on raw (non-annotated) documents. The remaining two curators annotated a small portion of the corpus: 13 and 9 documents.

Figure 4 illustrates the IAA panel in Egas. The IAA is calculated as the average of f-scores between each pair of curators, taking into account the documents shared by both. For this task, the complete corpus was assigned to all curators (100% overlap) in order to maximize the number of shared documents. An overall IAA of 0.74 was obtained, with paired agreements varying between 0.62 and 0.95. The IAA panel allows the project manager to obtain the agreement for each pair of curators, for all the concept types or by selecting the concept types to consider.

To evaluate the impact of the TM-assisted workflow, we compared the IAA obtained on the pre-annotated documents against the agreement obtained on documents without automatic annotations and on all documents in the corpus. IAA was significantly higher when performing the curation task on pre-annotated documents than on documents without automatic annotations (*P*-value $\ll$ 0.01, one-tailed paired *t*-test). The results shown in Figure 5 indicate that annotation consistency can be enhanced through the use of automated TM services. This improvement may possibly be explained by the added simplicity of checking and correcting existing annotations when compared with manually adding all annotations, and by the fact that the pre-annotation promotes completeness, that is, curators are encouraged to perform a more comprehensive annotation of the document. Annotations of ‘Age’, ‘Age of Onset’, ‘InDel’ and ‘Rearrangement’ were not considered for these results, as these were only present on a very small number of documents. The accuracy of the automatic annotations could be evaluated by comparison to the results provided by the curators. Figure 6 shows average evaluation metrics of automatic annotations against all curators, with precision ranging from 0.77 for gene to 1.0 for



Mutation Clinical Attributes Project Management ×

List of curators:

- ☑ Curator 1
- ☑ Curator 2
- ☑ Curator 3
- ☑ Curator 4
- ☑ Curator 5
- ☑ Curator 6
- ☑ Curator 7

List of concepts:

- ☑ Age
- ☑ Age_of_Onset
- ☑ Disease

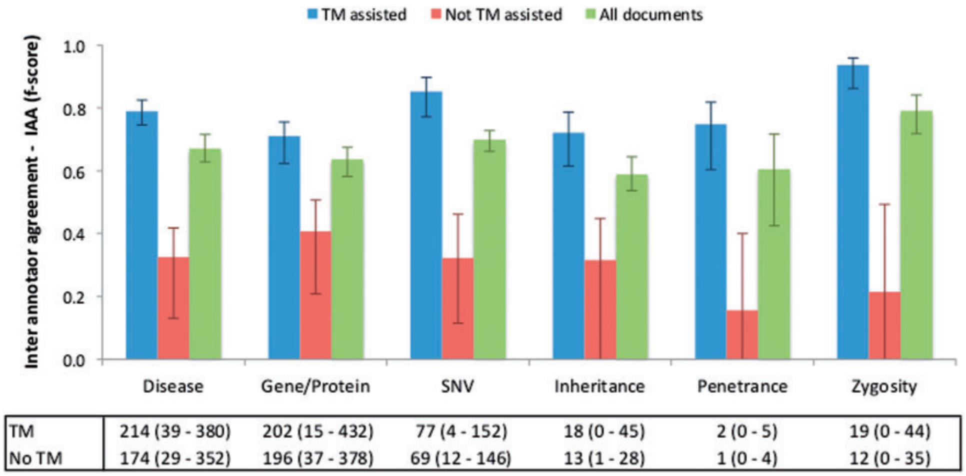
Inter-Annotators Agreement

Overall result: 0.74

	C1	C2	C3	C4	C5	C6	C7
C1	1	0.71	0.83	0.83	0.69	0.67	0.79
C2	0.71	1	0.62	0.63	0.80	0.75	0.66
C3	0.83	0.62	1	0.95	0.65	0.70	0.89
C4	0.83	0.63	0.95	1	0.65	0.70	0.89
C5	0.69	0.80	0.65	0.65	1	0.76	0.67
C6	0.67	0.75	0.70	0.70	0.76	1	0.68
C7	0.79	0.66	0.89	0.89	0.67	0.68	1

Measure unit F-score

Figure 4. IAA panel in Egas. The IAA is calculated as the *F*-score between each pair of curators, and the average of these values is taken as the overall result. Annotations for each concept type and each curator can be included or removed from the calculation by using the checkboxes on the left.



	Disease	Gene/Protein	SNV	Inheritance	Penetrance	Zygosity
TM	214 (39 - 380)	202 (15 - 432)	77 (4 - 152)	18 (0 - 45)	2 (0 - 5)	19 (0 - 44)
No TM	174 (29 - 352)	196 (37 - 378)	69 (12 - 146)	13 (1 - 28)	1 (0 - 4)	12 (0 - 35)

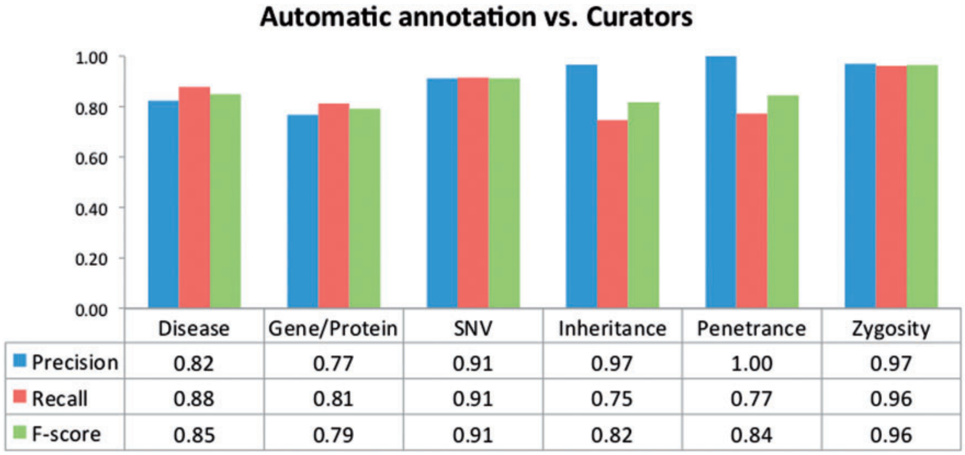
Figure 5. IAA obtained on documents pre-annotated with automated TM services, without pre-annotation, and on all documents. The error lines show the minimum and maximum IAA values for all pairs of annotators. The data table shows the average (range) number of annotations for each concept type using TM-assisted vs. not TM-assisted curation.

penetrance and recall ranging from 0.75 for inheritance to 0.96 for zygosity. These results indicate the quality of annotations provided and further support the effectiveness of TM-assisted literature curation.

We also evaluated the number of annotations and the time taken to perform the task on each set of documents. As can be seen from the results in Table 3, there are in general more concept annotations in documents that had been previously annotated by the concept recognition tool and, on average, less time is required to add annotations to these documents than to documents that had not been automatically annotated. Although no relation extraction tool was used to pre-annotate the corpus, the results indicate that automatic concept recognition may also help increasing the number of relations annotated. Two of the seven curators were not considered for these results, as they only curated 9 and 13 documents, respectively.

Conclusion

We describe the use of Egas, a configurable web-based document annotation and curation tool, on a literature curation task focused on the annotation of mentions of



	Disease	Gene/Protein	SNV	Inheritance	Penetrance	Zygosity
🟦 Precision	0.82	0.77	0.91	0.97	1.00	0.97
🟥 Recall	0.88	0.81	0.91	0.75	0.77	0.96
🟩 F-score	0.85	0.79	0.91	0.82	0.84	0.96

Figure 6. Evaluation of the automatic annotations provided in the corpus. The data table shows the average precision, recall and *F*-score against all curators.

Table 3. Comparison of TM-assisted versus non-assisted curation

	TM assisted	Non-assisted	*P*-value
Concepts	744	664	0.12
Relations	217	157	0.10
Concepts/article	23.5	17.4	0.13
Relations/article	6.8	4.2	0.12
Time/article	219	245	0.25
Time/concept	10.8	13.1	0.17
Time/relation	44.3	56.1	0.13

An automatic tool was used to annotate concepts but not relations. Values shown are averages for five curators. *P*-value obtained using a one-tailed paired *t*-test.

human inherited pathogenic gene variations and their association with genes, diseases and clinical attributes. The tool allows teams of curators to work on a shared curation project, following a set of configurable concept and relation types. The curation task can be performed over a collection of raw text documents or by reviewing automatic concept and relation annotations, obtained either with the included concept and relation identification service or through external annotation tools.

To validate Egas and its assisted curation features, we participated in the BioCreative Interactive Annotation task, obtaining positive evaluation in terms of usability, learnability and design on a user survey conducted by the task organizers (28). Additionally, we analysed the IAA and time spent when annotating raw documents and when working on documents previously annotated by an automatic concept recognition tool. The survey results and our comparative analysis show that a TM-assisted curation pipeline brings benefits in terms of efficiency and consistency of the curation results.

Acknowledgements

The authors would like to thank the organizers of the BioCreative User Interactive Task, and the curators that participated in the testing and evaluation of Egas for their valuable feedback.

Funding

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant no. 115372). iBiMED is supported by COMPETE/FEDER UID/BIM/04501/2013. SM is funded by Fundação para a Ciência e a Tecnologia—FCT under the FCT Investigator programme. D.C. has received support from the HemoSpec European project (EC contract number 611682). D.N.C. and M.M. receive financial support from Qiagen Inc through a License Agreement with Cardiff University. Funding for open access charge: FCT [IF/01694/2013/CP1162/CT0018].

Conflict of interest. None declared.

References

1. Hirschman,L., Burns,G.A., Krallinger,M. *et al.* (2012) Text mining for the biocuration workflow. *Database*, 2012, bas020.
2. Alex,B., Grover,C., Haddow,B. *et al.* (2008) Assisted curation: does text mining really help? *Pac. Symp. Biocomput*., 2008, 556–567.
3. Karamanis,N., Seal,R., Lewin,I. *et al.* (2008) Natural language processing in aid of FlyBase curators. *BMC Bioinformatics*, 9, 193.
4. Bolchini,D., Finkelstein,A., Perrone,V. *et al.* (2009) Better bioinformatics through usability analysis. *Bioinformatics*, 25, 406–412.
5. Stenetorp,P., Pyysalo,S., Topic,G. *et al.* (2012) BRAT: a web-based tool for NLP-assisted text annotation.In: *European Chapter of the Association for Computational Linguistics*. Avignon, France, p. 102.
6. Cejuela,J.M., McQuilton,P., Ponting,L. *et al.* (2012) tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database*, 2014, bau033.
7. Salgado,D., Krallinger,M., Depaule,M. *et al.* (2012) MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, 28, 2285–2287.

8. Rak,R., Rowley,A., Black,W. and Ananiadou,S. (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, 2012, bas010.
9. Kwon,D., Kim,S., Shin,S.Y. *et al*. (2014) Assisting manual literature curation for protein–protein interactions using BioQRator. *Database*, 2014, bau067.
10. Wei,C.H., Kao,H.Y. and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*., 41, W518–W522.
11. Torii,M., Li,G., Li,Z. *et al*. (2014) RLIMS-P: an online text-mining tool for literature-based extraction of protein phosphorylation information. *Database*, 2014, bau081.
12. Torii,M., Arighi,C.N., Li,G. *et al*. (2015) RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. *IEEE/ACM Trans. Comput. Biol. Bioinform*, 12, 17–29.
13. Rinaldi,F., Clematide,S., Schneider,G. *et al*. (2010) ODIN: an advanced interface for the curation of biomedical literature. *Nature Precedings*, 10.1038/npre.2010.5169.1.
14. Neves,M. and Leser,U. (2012) A survey on annotation tools for the biomedical literature. *Brief. Bioinform*., 15, 327–340.
15. Campos,D., Lourenço,J., Matos,S., and Oliveira,J.L. (2014) Egas: a collaborative and interactive document curation platform. *Database*, 2014, bau048.
16. Standoff format - brat rapid annotation tool. Available online at: http://brat.nlplab.org/standoff.html.
17. Comeau,D.C., Doğan,R.I., Ciccarese,P. *et al*. (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013, bat064.
18. Verspoor,K., Yepes,A.J., Cavedon,L. *et al*. (2013) Annotating the biomedical literature for the human variome. *Database*, 2013, bat019.
19. Wei,C.H., Harris,B.R., Kao,H.Y. and Lu,Z. (2013) tmVar: A text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29, 1433–1439.
20. Caporaso,J.G., Baumgartner,W.A. Jr, Randolph,D.A. *et al*. (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23, 1862–1865.
21. Doughty,E., Kertesz-Farkas,A., Bodenreider,O. *et al*. (2011) Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27, 408–415.
22. Ravikumar,K.E., Wagholikar,K.B., Li,D. *et al*. (2015) Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature. *BMC Bioinformatics*, 16, 185.
23. Stenson,P.D., Mort,M., Ball,E.V. *et al*. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet*., 133, 1–9.
24. Campos,D., Matos,S. and Oliveira,J.L. (2013) A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14, 281.
25. Doğan,R.I., Leaman,R. and Lu,Z. (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform*., 47, 1–10.
26. Tanabe,L., Xie,N., Thom,L. *et al*. (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6, S3.
27. Campos,D., Matos, S., and Oliveira, J. Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14, 54.
28. Wang,Q., Abdul,S.S., Almeida,L. *et al*. (to appear) Overview of the Interactive Task in BioCreative V.