Bio-ID Track Overview

Cecilia Arighi[1], Lynette Hirschman[2], Thomas Lemberger[3], Samuel Bayer[2], Robin Liechti[4], Donald Comeau[5], Cathy Wu[1]

[1]Dept. of Computer and Information Sciences, University of Delaware, Newark, DE, USA
[2]Information Technology Center, The MITRE Corporation, Bedford, M, USA
[3]EMBO, Heidelberg, Germany
[4]SIB, Lausanne, Switzerland
[5]NCBI, NIH, Bethesda, MD, USA

Abstract— **The Bio-ID track focuses on entity tagging and ID assignment to selected bioentity types, with the aim of facilitating downstream article curation both at the pre- and post-publication stages. The task is to annotate text from figure legends with the entity types and IDs for taxon (organism), gene, protein, miRNA, small molecules, cellular components, cell types and cell lines, tissues and organs. The track draws on SourceData annotated figure legends (by panel), in BioC format, and the corresponding full text articles (also BioC format) provided for context. Six teams submitted annotations: two teams submitted results for all 6 entity types; two teams submitted results for organism only; one team submitted results for miRNA and organism; and one team submitted results for small molecules. Mention level F-measures were 0.8 or better for cell type, species and gene-or-protein; however, micro-averaged normalized F-measure was significantly lower: 0.76 for species, 0.65 for cell type and below 0.6 for the other entity types. A subsequent experiment will investigate whether such systems can speed expert curation.**

Keywords—bioentity extraction; normalization; figure captions

I. Introduction

Innovations in biomedical digital curation have emerged as a critical topic to address sustainability of biological databases and research resources. Digital curation is defined as "the active management and preservation of digital resources over the lifecycle of scholarly and scientific interest, and over time for current and future generations of users" (1). In particular, there is a recognition that data curation needs to be integrated throughout the research lifecycle, without having to wait for curation by biocurators until after publication, as is the current practice for curated databases. While capturing knowledge of researchers at the time of data generation and publishing may enhance efficiency, there are significant barriers to moving curation "upstream." It is well recognized that the adoption of common database identifiers (IDs), controlled vocabularies (CVs) and ontologies facilitates data integration and re-use; however, it is nontrivial to extract IDs, CVs and ontological terms from the free texts of the scientific literature. New methods and tools need to be developed to support more effective and consistent curation at the time of paper submission.

The Bio-ID track addresses these needs for Innovations in Biomedical Digital Curation (2). Publications are one of the main vehicles for dissemination of experimental results. Researchers have new ideas, conduct experiments, write up their results summarizing those experiments, submit them to a journal and, if accepted after peer-review, the articles are disseminated in public literature databases. Publications are also the primary source of data for knowledgebase curators, who extract and summarize the relevant data in standard formats. While researchers use both the literature and knowledge bases, the latter offer efficient platforms for querying, given the linkage of data in literature to database objects. Then new ideas/hypotheses are generated to start a new cycle. Currently, there is a bottleneck in data re-use as curators spend time identifying bioentities in publications and linking these entities into their databases. We hypothesize that curation would be facilitated if articles were preprocessed to link the key bioentities to their appropriate biological knowledge bases, prior to publication (benefitting publishers) and prior to curation (speeding the downstream curation process); we refer to this as bio-ID assignment.

II. The Bio-ID Track

The Bio-ID track focuses on entity tagging and ID assignment to selected bioentity types, with the aim of facilitating downstream article curation at both the pre- and post-publication stages. This track builds on the SourceData project (3) as well as previous BioCreative experiments, including the interactive tracks (IAT), earlier gene/protein/chemical extraction tracks, the BioC interchange format and the BeCalm framework (http://www.becalm.eu/). The track is designed to foster the development of an integrated and interoperable workflow of multiple text mining tools for real-world testing in pilot publishing frameworks.

The track consists of two phases: 1) a batch phase, where the task is to annotate text from figure legends with the entity types and IDs for taxon (organism), gene, protein, miRNA, small molecules, cellular components, cell types and cell lines, tissues and organs; and 2) an interactive annotation phase, where curators can make use of the system-supplied annotations, to determine whether this speeds the curation

This work was supported under NIH grant 5R01GM080646-11S1 and NIH R13 GM109648.

process. This report describes the batch testing results. Experiments on the interactive annotation phase are now underway.

III. DATA

Training materials were drawn from annotated figure panel legends from SourceData (3). SourceData is working with publishers and authors to create machine-readable descriptions of underlying data in figures and figure legends. By referring to established public databases of biological terms, the specific biological entities, their roles as target, intervention or outcome measure in each paper can be consistently identified. SourceData (TL, RL) made available a large set of curated figure panel legends for the BioID track; the SourceData team made these data sets available "as is", to provide a "public" data set for training and a "private" (not previously released) reference data set for testing system performance.

The track organizers (DC, SB) converted these annotated panel captions into BioC format, which were provided to participants, along with the corresponding full text articles (also BioC format) for context. Participants participated by submitting annotations for one or more of the bioentity types, in BioC format for the set of captions.

A. *Bio-ID Training and Test Data Sets*

The training data set consisted of 13,573 annotated figure panel captions corresponding to 3,658 figures from 570 full length articles from some 22 journals, for a total of 102,717 annotations. Table 1 shows the distribution of entity types across the training corpus. The test data set consisted of 4,310 annotated figure panel captions from 1,154 figures taken from 196 full length journal articles, with 30,286 annotations in total. This corresponds to roughly 6+ figures/document and 3-4 panels per figure with some 7 annotations per panel caption.

TABLE 1 DESCRIPTION OF TRAINING DATA BY ENTITY TYPE

Entity Type	# articles w entity	Total # by entity type	# unique by entity type
Protein/Gene (561 articles)	UniProt (548)	30211	2833
	NCBI gene (537)	21766	2451
miRNA	9	167	13
Small molecules (513 articles)	ChEBI (506)	9869	786
	PubChem (134)	700	140
Cellular component	482	7310	376
Cell types and cell lines (482 articles)	Cellosaurus (351)	5783	230
	Cell Ontology (300)	4638	217
Tissues & organs	316	5870	459
Organisms & species	454	7952	147

B. *Data Set Content*

The training data contains a set of files in BioC format. Each file consists of a collection of figure captions for a given article annotated by SourceData curators. The name of the file corresponds to its PubMed Central ID and the corresponding full-text article is provided separately. Within a BioC file, each <document> element contains the annotation for each individual figure panel within the article; the <id> given is made up of the PMCID followed by the Figure_number-panel (e.g., <id>5048346 Figure_1-A</id>), as shown in the example in Fig. 1.

Note that the text for the panel may consist of discontinuous text derived from the Figure caption; it may also include legends from several panels.

C. *Annotation Guidelines*

Below are important considerations of the curation practice (extracted from the SourceData definitions for curators by Thomas Lemberger):

- The SourceData description of the data presented in scientific figures specifies the entities that are relevant to the scientific meaning of the data. To be tagged by curators, a figure panel must report experimental data. Panels that present schematics, computational simulation results, overviews and workflows are not tagged.
- In the text of a relevant panel legend, all mentions of terms that correspond to entities types on Table 2 are tagged.
- Entities are assigned to only one entity type of those listed in Table 2.
- In general, generic terms referring to broad classes of biological components (e.g. 'proteins', 'cells', 'animals') are not be tagged unless they refer to the object of an assay.

```
[<document>
  <id>5048346 Figure_1-A</id>
  <infon key="sourcedata_document">2225</infon>
  <infon key="doi">10.15252/embj.201694885</infon>
  <infon key="pmc_id">5048346</infon>
  <infon key="figure">Figure 1-A</infon>
  <infon
key="sourcedata_figure_dir">Figure_1-A</infon>
  <passage>
   <offset>0</offset>
   <text>A. The localization of NSUN3 was analysed in
HEK293 cells stably expressing NSUN3-GFP (green).
NSUN3-GFP and staining with a Mitotracker (red) are
shown separately and in an overlay with DAPI to indicate
nuclei. The scale bar represents 5 m.</text>
   <annotation id="1">
    <infon key="type">Uniprot:Q9H649</infon>
    <infon key="sourcedata_figure_annot_id">1</infon>
    <infon
key="sourcedata_article_annot_id">1</infon>
    <location offset="23" length="5"/>
    <text>NSUN3</text>
   </annotation>
   …
```

Fig. 1. Example of BioC training data file

TABLE 2 ENTITY TYPES WITH CORRESPONDING RESOURCES FPR LINKAGE (NORMALIZATION)

Entity Type	Resources	Example infon key="type"	Generic infon key="type"
Protein	UniProt	Uniprot:Q9H649	Uniprot:<Accession>
Gene	NCBI gene	NCBI gene:4137	NCBI gene:<ID>
miRNA	Rfam	Rfam:RF00076	Rfam:<ID>
Small molecules	ChEBI (primary)	CHEBI:15996	CHEBI:<ID>
	PubChem (secondary)	PubChem:5717066	PubChem:<id>
Cellular component	GO cellular component	GO:0005886	<GOID>
Cell types and cell lines	Cellosaurus (primary)	CVCL_U985	<accession>
	Cell Ontology (secondary)	CL:0000540	<CLID>
Tissues & organs	Uberon	Uberon:UBERON:0002048	Uberon:<UBERONID>
Organisms & species	NCBI Taxonomy	NCBI taxon:10090	NCBI taxon:<ID>

Sources: All ID sources are publicly available:
UniProt: www.uniprot.org
Rfam: http://rfam.xfam.org/
PubChem: https://pubchem.ncbi.nlm.nih.gov/
Cellosaurus: http://web.expasy.org/cellosaurus/
Uberon: http://uberon.github.io/
NCBI gene: https://www.ncbi.nlm.nih.gov/gene
ChEBI: https://www.ebi.ac.uk/chebi/
Gene ontology: http://geneontology.org/
Cell Ontology: http://obofoundry.org/ontology/cl.html
NCBI Taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy

• Some terms, such as those referring to proteins or genes, can be appended with prefixes or suffixes that indicate a post-translational modification, a mutation or other variations of the actual base term. In such cases, prefixes or suffixes are left out and only the base term is tagged (e.g., in text describing a mutant of B-RAF 'B-RAF(V600E)', only B-RAF is tagged; similarly with p-AKT1 that designates the phosphorylated form of AKT1, only AKT1 is tagged).

• In other cases, a prefix is added to an entity to denote a species origin, in which case the prefix should be kept (e.g., dMyc to denote the homolog in Drosophila of Myc)

• Some components are engineered by assembling or fusing multiple sub-components; these are tagged individually. For example, the term 'RAS-GFP' referring to a fusion protein between GFP and RAS is annotated with two tags: 'RAS' and 'GFP'.

D. Special considerations for the BIO-ID track

• Participating teams can provide annotation for one or more types of those described in Table 1.

• For this track, the gene/protein type can be treated interchangeably (i.e., all proteins and gene mentions can be linked to NCBI gene and/or UniProt identifiers).

• Some entity types, like small molecules, have more than one resource listed; however, one of the resources is considered primary. To help with training and comparison, we provide mappings, to the extent possible, between identifiers of the same entity type. However, the best practice would be to look up the entity in the preferred resource, and if it cannot be found, then look it up in the secondary resource.

• In particular, for UniProtKB, UniProtKB/Swiss-Prot (reviewed) entries should be linked whenever possible; TrEMBL (unreviewed) entries should only be linked when no corresponding Swiss-Prot entry is available.

IV. SCORING

The scorer reports scores at three levels: the individual caption level, the document level, and the corpus level. At the last two levels, scores are aggregated per bioentity type from the previous level. Scoring is done both at the mention level (every occurrence of a bioentity is tagged for its type) and at the normalized identifier level, where the set of unique identifiers in a caption are compared to the reference set of identifiers from SourceData.

Mention level annotations are divided into two classes: the first class consists of annotations which are linked to a biological resource (the "normalized annotations", e.g., in Fig. 1, "NSUN3" is annotated with "Uniprot:Q9H649"). The second class consists of annotations that are associated with an entity type, but not linked to a biological resource, for example, in the text "GFP is green and Calbindin staining is red.", "GFP" is tagged as "protein:GFP" which labels GFP as a protein, but does not link it to an identifier in a specific resource.

The scorer reports mention level scores in 4 conditions:

Score Types	All Entities	Normalizable Only
Exact Match	Any Exact	Norm Exact
Overlapping Match	Any Overlap	Norm Overlap

Exact span match requires that the span (in byte offset) match exactly against the SourceData reference standard; span overlap relaxes this condition to allow a match if the span overlaps with the reference annotation at all. This means that the "overlap" score will be greater than or equal to the "exact" score. The scorer computes mention-level recall/precision/f1-measure for each of these four conditions.

The scorer also computes recall/precision/F1-measure on the normalized IDs which are found, both micro-averaged over the corpus and macro-averaged by document. For the

TABLE 3 TOP SCORING RUNS BY ENITYT TYPE FOR MENTION F1 SCORE ("OVERLAP ALL") AND MICRO-AVERAGED NORMALIZATION F1

	Mention - All Overlap				Normalization Micro Avg					
Entity Type	Team	P	R	F1	Team	P	R	F1	# Teams	# Runs
cell_type_or_line	407	0.84	0.76	0.80	422_2	0.78	0.56	0.65	2	3
cellular_component	407	0.73	0.55	0.63	422_2	0.55	0.45	0.49	2	3
gene_or_protein	407	0.83	0.84	0.83	407	0.47	0.34	0.40	2	3
organism_or_species	407	0.88	0.83	0.85	393	0.66	0.88	0.76	5	9
small_molecule	407	0.80	0.60	0.69	422_2	0.59	0.47	0.52	3	5
tissue_or_organ	407	0.79	0.63	0.70	407	0.53	0.49	0.51	2	3
miRNA	386	not scored					not scored		1	1

comparison of results, we report here on exact and overlap matching against all annotations for mention level scoring; and micro-averaged scoring for the normalized identifiers.

V. RESULTS

A total of six teams participated. Two teams provided annotations for all six entity types; two teams provided annotations for species only; one team did small molecules only, and one team did miRNA (not scored) and species. The results for the top scoring runs are shown in Table 3, along with the number of teams and runs for each bioentity type. The top scores were selected based on the top F1 score for the "Overlap All" computation for the mention level score. The micro-averaged F1 was used to rank the normalization scores. The set of scores (all types of mention level scores plus micro-averaged normalization scores) for all runs is shown in the Appendix; the high score in each category is highlighted with a green background.

The highest performance was achieved for organism/species, which also had the most participating teams (see Table 3 and Figs. 2 and 3). Fig. 4 shows the results (at both mention level and normalization) for gene/protein as well as organism. Determining the species is a necessary step to determine a correct gene or protein identifier, since these are dependent on species. However, the species may not be explicitly mentioned in the figure caption, making it necessary to use the full text to determine the species or organism for each experiment. We see that the species scores are quite high (both for mention level and normalized micro-average – grey and gold stars). For gene/protein mention, the "Any Overlap" scores (black triangles) are also quite good, but the "Norm Overlap" scores (green triangles) are less good, and the micro-averaged gene/protein normalization scores (red triangles) are much worse – indeed Fig. 3 shows that results for gene/protein normalization are the lowest among the bioentities. One issue is that correct normalization requires stripping of complex prefixes and suffixes from gene names – a challenging task. A second issue is that only two teams tackled the gene/protein task – and these were not the high-scoring systems for species. The results suggest that progress is being made at the mention level, but that gene/protein normalization remains challenging.

ACKNOWLEDGMENTS

The BioID Track has been supported by NIH grants 5R01GM080646-11S1 and NIH R13 GM109648. LH and SB have been supported under MITRE's Internal Research and Development Program.

REFERENCES

1. Lee, C., and Tibbo, H. (2007) Digital Curation and Trusted Repositories: Steps Toward Success. Journal of Digital Information 8, 2

2. https://datascience.nih.gov/

3. Liechti, R., George, N., El-Gebali, S., Götz,,L., Crespo, I., Xenarios, I., and Lemberger, T. (2016) SourceData - a semantic platform for curating and searching figures. preprint in BioRxiv doi: https://doi.org/10.1101/058529

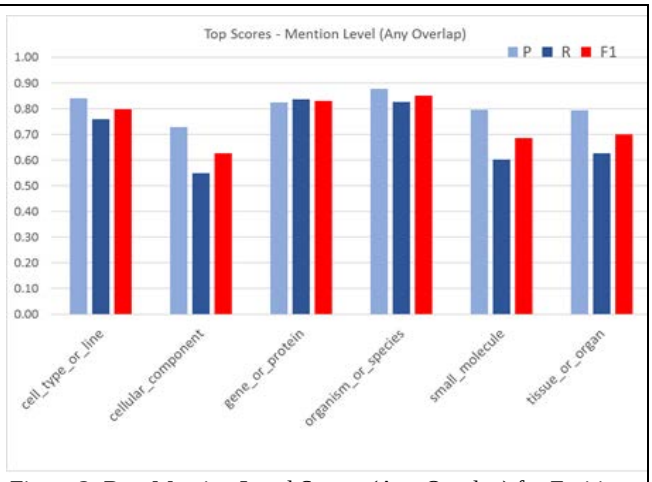

Figure 2: Best Mention Level Scores (Any Overlap) for Entities

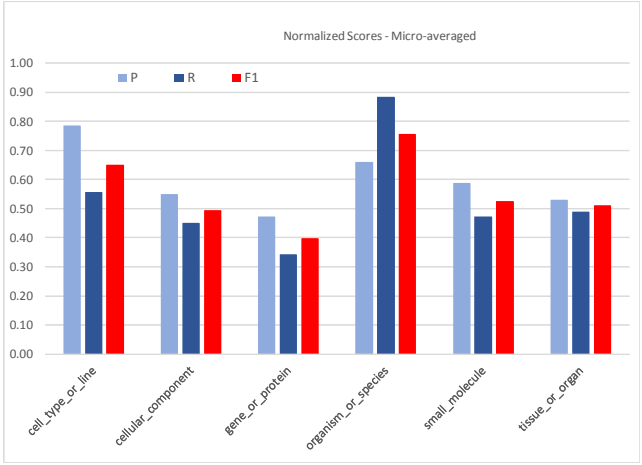

Figure 3: Best Normalized Scores (Micro-Averaged) for Entities

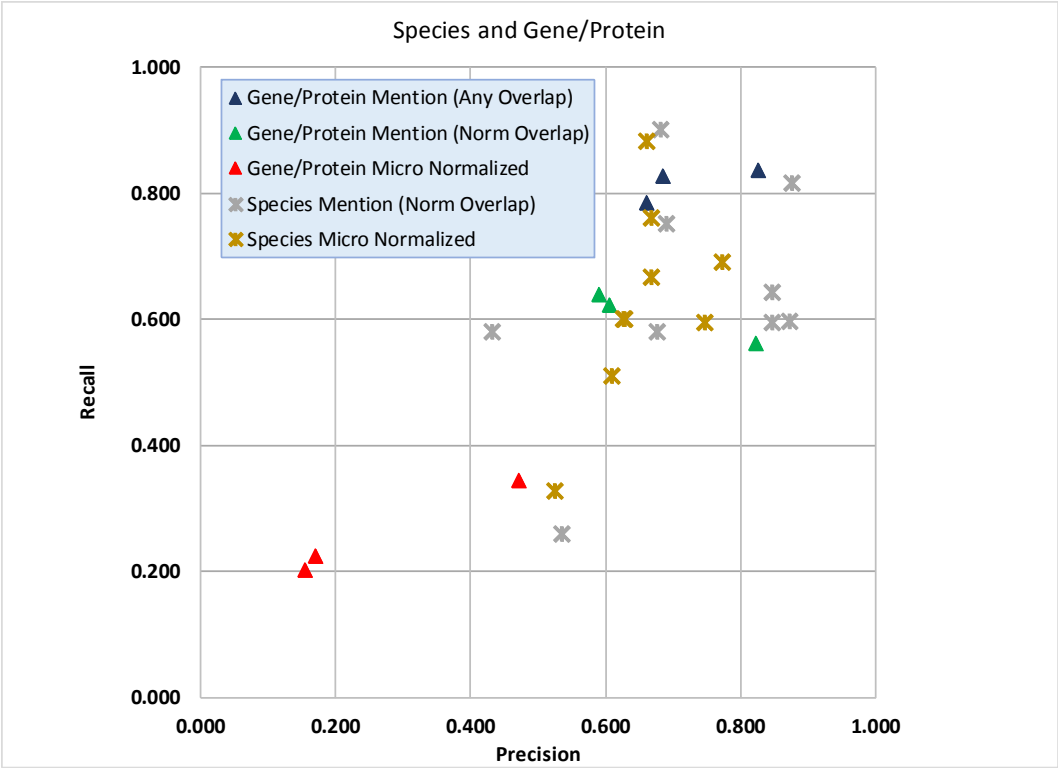

Figure 4: Gene/Protein and Species Results

Appendix: Scores for Mention Level and Micro-Averaged Normalized Scores for all Runs

Entity Type	Run	Any Exact P	Any Exact R	Any Exact F1	Any Overlap P	Any Overlap R	Any Overlap F1	Norm Exact P	Norm Exact R	Norm Exact F1	Norm Overlap P	Norm Overlap R	Norm Overlap F1	Micro Norm P	Micro Norm R	Micro Norm F1
cell type or line	407	**0.783**	**0.708**	**0.744**	**0.841**	0.761	**0.799**	0.774	**0.699**	**0.735**	0.830	**0.750**	**0.788**	0.600	**0.576**	0.588
	422_1	0.563	0.493	0.526	0.764	0.670	0.714	0.638	0.430	0.513	0.827	0.557	0.666	0.686	0.505	0.582
	422_2	0.653	0.650	0.652	0.766	**0.763**	0.765	**0.822**	0.551	0.660	**0.866**	0.581	0.695	**0.784**	0.557	**0.651**
cellular compo-nent	407	**0.673**	**0.508**	**0.579**	**0.728**	**0.550**	**0.627**	**0.685**	**0.482**	**0.566**	**0.737**	**0.519**	**0.609**	0.456	0.371	0.410
	422_1	0.400	0.353	0.375	0.600	0.528	0.562	0.424	0.345	0.380	0.629	0.511	0.563	0.472	**0.467**	0.469
	422_2	0.548	0.439	0.488	0.629	0.505	0.560	0.456	0.319	0.376	0.659	0.461	0.543	**0.550**	0.450	**0.495**
gene or protein	407	**0.729**	**0.739**	**0.734**	**0.825**	**0.836**	**0.831**	0.795	0.543	**0.645**	**0.823**	0.561	**0.667**	**0.472**	**0.343**	**0.397**
	422_1	0.412	0.490	0.447	0.660	0.785	0.717	0.437	0.474	0.455	0.590	**0.640**	0.614	0.154	0.202	0.175
	422_2	0.509	0.613	0.556	0.686	0.826	0.749	0.456	0.469	0.463	0.605	0.622	0.614	0.170	0.224	0.193
organism or species	386_1	0.668	0.572	0.616	0.677	0.580	0.625	0.667	0.572	0.616	0.677	0.580	0.625	0.628	0.601	0.614
	386_2	0.426	0.572	0.489	0.433	0.580	0.496	0.426	0.572	0.488	0.432	0.580	0.495	0.626	0.601	0.613
	393_1	0.663	0.874	0.754	0.683	**0.900**	0.776	0.662	0.874	0.754	0.682	0.900	0.776	0.660	**0.883**	**0.756**
	393_2	0.671	0.731	0.699	0.690	0.752	0.720	0.670	0.731	0.699	0.689	0.752	0.719	0.668	0.760	0.711
	393_3	0.516	0.251	0.337	0.536	0.260	0.350	0.516	0.251	0.338	0.536	0.260	0.350	0.526	0.327	0.403
	407	**0.860**	**0.809**	**0.834**	**0.878**	0.826	**0.852**	**0.857**	**0.797**	**0.826**	**0.877**	**0.815**	**0.845**	0.668	0.667	0.667
	408	0.813	0.616	0.701	0.848	0.642	0.731	0.812	0.616	0.701	0.847	0.642	0.730	0.609	0.510	0.555
	422_1	0.649	0.496	0.563	0.837	0.640	0.726	0.686	0.483	0.567	0.846	0.595	0.699	0.747	0.594	0.662
	422_2	0.746	0.715	0.730	0.814	0.780	0.796	0.723	0.495	0.588	0.872	0.597	0.709	0.772	0.691	0.729
small molecule	407	**0.775**	**0.587**	**0.668**	0.796	**0.603**	**0.686**	**0.787**	**0.581**	**0.668**	0.803	**0.593**	**0.682**	0.244	0.240	0.242
	422_1	0.503	0.359	0.419	0.697	0.496	0.579	0.504	0.326	0.396	0.696	0.451	0.547	0.587	**0.473**	**0.524**
	422_2	0.562	0.451	0.500	0.683	0.548	0.608	0.490	0.331	0.395	0.686	0.463	0.553	0.654	0.394	0.492
	425_1	0.584	0.065	0.116	**0.915**	0.101	0.182	0.582	0.069	0.123	**0.910**	0.107	0.192	**0.901**	0.150	0.257
	425_2	0.563	0.050	0.091	0.902	0.079	0.146	0.538	0.051	0.092	0.856	0.080	0.147	0.791	0.102	0.180
tissue or organ	407	**0.727**	**0.575**	**0.643**	**0.793**	**0.627**	**0.701**	**0.604**	**0.542**	**0.572**	0.669	**0.600**	**0.632**	0.531	**0.490**	**0.510**
	422_1	0.547	0.405	0.465	0.741	0.548	0.630	0.519	0.394	0.448	**0.678**	0.515	0.586	0.547	0.434	0.484
	422_2	0.572	0.559	0.565	0.671	0.654	0.662	0.497	0.417	0.453	0.654	0.548	0.596	**0.584**	0.442	0.503