USING TEXTPRESSO FOR INFORMATION RETRIEVAL, FACT EXTRACTION AND DATABASE ENTRY
WormBase: a collaborative effort for genetics, genomics and biology of *C. elegans* and some related nematodes
How does experimental data get into WormBase?

WormBase literature curation pipeline

Daily, automated search using keyword ‘elegans’

Manually select papers for inclusion into WormBase

Read all papers and ‘flag’ various data types (first-pass)

email notification to individual data-type curators

paper-entity association experiment curation
Why incorporate text mining into our curation pipeline?

WormBase curators curate >25 different types of data, but there are still data types we don’t yet curate.

For some data types we do curate, we have a backlog

- older papers <2001
- newer curation tasks

More genomes = more curation

If text mining tools can improve curation efficiency, they’ll help with all of this!
**Textpresso** : Text Mining for Literature Curation

http://www.textpresso.org

Key features:

- Search full text of articles, returns sentences
- Keyword (+synonyms) and/or category searches

  ‘regulation’: attenuate, downregulate, enhancing, inhibiting, misregulated, etc.

- Sort by score, year, journal, etc.

  sort by score = efficient prioritization

- Open source, with 19 different implementations

"...all wild-type embryos have foci of GFP:SAS-4 associated with both sperm centrioles..."
for Paper-Entity Association:
What objects are in a paper?

Gene names

Alleles

Transgenes

Life stages

Anatomy terms

Chemicals

Human disease
Paper-entity association: alleles and transgenes

Principle: standardized nomenclature allows for pattern matching in full text using regular expressions

Alleles: q71, lg6001, e2141ts, oz12oz75

\[ [a-z]{1-3}\d+\w*\d* \]

98% recall

Transgenes: syls1, nls2, bln1

80% recall, >95% precision
for Data Type Identification:
What kinds of experiments?

- Cloning
- Mutant Phenotypes
- RNAi
- Alleles
- Expression Patterns
- Antibodies
- Gene Regulation
- Interactions
- Genes
- Gene Products
- Site of Action
We performed RNAi feeding using lim-7 cDNA (see Section 2.2) on homozygous wild-type and heterozygous (tm674/hT2 [let GFP]) animals...

...we exposed ain-1 (ku322) animals to ain-2 (RNAi), starting with L1 larvae and using a published RNAi feeding construct and protocol...
Data type identification: Antibodies and antibody production

Principle: pattern matching and combinatorial category search to find evidence for antibodies in full text

“The immunoprecipitated protein complexes were subsequently immunoblotted using anti-HCF-1, anti-DAF-16, or anti-GFP antibodies.”

“To determine the expression pattern of sepa-1, we raised antibodies against the KIX domain of SEPA-1.”

Antibody pattern
- anti-DAF-16
- anti-HCF-1

Antibody verbs
- produced
- generated
- raised

Antibody terms
- antibody
- antiserum
- antisera
- antibodies

85% recall
> 80% precision
Finding Curatable Information

- Loss of MAD1/MDF-1 suppresses the meiotic defect of mat-3(or180)
- NXF-1 binds RNA directly
- GFP:: RAB-7 is visualized on the membranes of late endosomes
- srsx-3 expression was present in young odr-1 larvae but absent in odr-1 adults
Fact Extraction: GO Cellular Component Curation

Principle: combinatorial category searches to retrieve experimental results from the full text of papers

Why GO Cellular Component curation?

- Subcellular localization is often expressed in a stereotypical way within a single sentence

  GFP:: RAB-7 is visualized on the membranes of late endosomes

- Limited number of experimental strategies or assays
  - Microscopy
  - Fractionation

- Typically annotate using only one GO evidence code, IDA
Fact Extraction: GO Cellular Component Curation

Needed to make new categories:

What words and phrases are diagnostic of subcellular localization experiments?

1 - Curators read ~240 papers containing localization data

2 – Collected ~1,700 sentences that report experimental results

3 – Examined word usage and frequency

This approach took advantage of the work that curators already do and also allowed us to add annotations to WormBase while we created the new categories.
Three new categories:

**Cellular Components**

**Verbs**

**Assay Terms**

*IDA*-1 and *IDA*-1::GFP were not restricted to presynaptic sites but were abundantly localized throughout cell bodies, axons and dendrites of expressing neurons.
Can we annotate from the search results?

Test searches:
search with names of previously uncurated *C. elegans* proteins and three new categories

Criteria:
Returned sentences must contain a *C. elegans* protein plus one word from each of the three categories

Results:

Papers: 79.1% recall, 61.8% precision

Sentences: 30.3% recall, 80.1% precision

*Annotations: 66.2% recall, 97.3% precision
Fact Extraction: GO Cellular Component Curation

Is Textpresso-based curation more efficient?

YES!

Tests comparing manual vs. Textpresso-based curation:

Textpresso yields 8- to 15-fold improvement in curation efficiency
Incorporating Textpresso into our Curation Pipeline

Weekly searches:

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Suggested GO terms based on previous annotation</th>
<th>Textpresso Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**WDR-23**: WDR-23: GFP is expressed in nuclei of the hypodermis, intestine and head neurons.

**SKN-1**: Exposure to stressors or inhibition of the proteasome causes SKN-1::GFP to accumulate in the nuclei of intestinal cells (1, 27).

**SKN-1**: As shown in Figures 6C-D, RNAi of wdr-23, ddb-1 and cul4 caused accumulation of SKN-1::GFP in intestinal nuclei.

**WDR-23**: As shown in Figure 7A, skn-1 mRNA levels were not significantly (P > 0.05) increased by knockdown of these genes, demonstrating that WDR-23 and the CUL-4/DDB-1 complex regulate nuclear accumulation of SKN-1 protein independently of skn-1 mRNA levels.
Onward! Future Developments and Plans

Improve Search Results

False Positives
- Textpresso searches by paper section
- Word exclusion

False Negatives
- What do we miss and why?
  - Missing category terms
  - Non-standard nomenclature
  - Information in multiple sentences

More Category Development
- Use Textpresso searches to help make new categories
- Category Editor – interactive, iterative Textpresso searches

Incorporate Textpresso Searches into Existing Curation Tools
- Phenote for Phenotype and GO Curation
## Acknowledgements

**Textpresso – Caltech**

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juancarlos Chan</td>
</tr>
<tr>
<td>Ruihua Fang</td>
</tr>
<tr>
<td>Joshua Jaffery*</td>
</tr>
<tr>
<td>Eimear Kenny*</td>
</tr>
<tr>
<td>Hans-Michael Müller</td>
</tr>
<tr>
<td>Arun Rangarajan</td>
</tr>
<tr>
<td>Tracy Teal*</td>
</tr>
</tbody>
</table>

**WormBase – Caltech**

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juancarlos Chan</td>
</tr>
<tr>
<td>Wen Chen</td>
</tr>
<tr>
<td>Jolene Fernandes</td>
</tr>
<tr>
<td>Ranjana Kishore</td>
</tr>
<tr>
<td>Raymond Lee</td>
</tr>
<tr>
<td>Cecilia Nakamura</td>
</tr>
<tr>
<td>Andrei Petcherski*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slava Petcherski*</td>
</tr>
<tr>
<td>Gary Schindleman</td>
</tr>
<tr>
<td>Erich Schwarz</td>
</tr>
<tr>
<td>Kimberly Van Auken</td>
</tr>
<tr>
<td>Daniel Wang</td>
</tr>
<tr>
<td>Xiaodong Wang</td>
</tr>
<tr>
<td>Karen Yook</td>
</tr>
</tbody>
</table>

*alumni

**Paul Sternberg, PI**

**Funding:**

- National Human Genome Research Institute, NIH
- WormBase, Textpresso, Gene Ontology

**Bioinformatics Group, Lewis Sigler Institute for Integrative Genomics, Princeton University**