A Comparison Between Genetics Papers Relating to Immune Disorders and Psychiatric Disorders

Mahmoud El-Haj Scott Piao Paul Rayson Jo Knight

Background

The explosion of literature in the field of genetics makes it hard to keep apace of new knowledge. Techniques developed in Natural Language Processing and Corpus Linguistics can help. Previously such techniques have been used to perform tasks such as identifying gene-gene or gene-phenotype interactions.

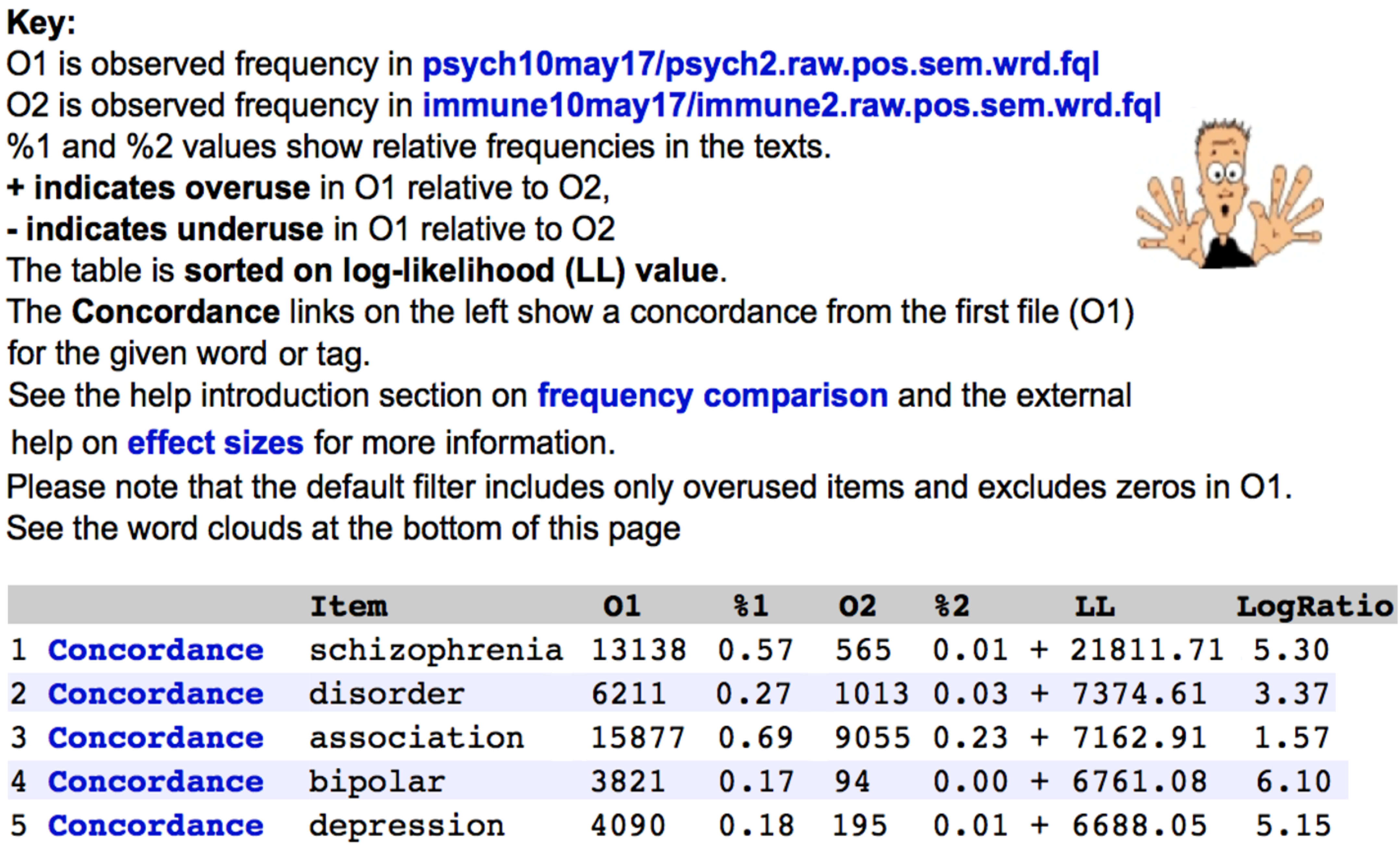
Aim

We will develop techniques to identify words that will provide new clues to disease aetiology.

Method

We have performed two searches in PubMed. One to extract articles relating to genetic association studies of immune disorders, another to extract articles relating to genetic association studies of psychiatric studies.

We use Wmatrix (http://ucrel.lancs.ac.uk/wmatrix/) to compare the frequency of words between the two corpora. Frequency differences are ranked on the basis of log likelihood results. Those words more frequent in the psychiatric corpora are presented in the wordles. (Generated using http://www.wordle.net/compose). Figure 1 shows the raw results. Figure 2 shows the results once expected words are removed (e.g. names of psychiatric disorders)



Key:
O1 is observed frequency in **psych10may17/psych2.raw.pos.sem.wrd.fql**
O2 is observed frequency in **immune10may17/immune2.raw.pos.sem.wrd.fql**
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2
The table is **sorted on log-likelihood (LL) value**.
The **Concordance** links on the left show a concordance from the first file (O1) for the given word or tag.
See the help introduction section on **frequency comparison** and the external help on **effect sizes** for more information.
Please note that the default filter includes only overused items and excludes zeros in O1.
See the word clouds at the bottom of this page

		Item	O1	%1	O2	%2	LL	LogRatio
1	Concordance	schizophrenia	13138	0.57	565	0.01	+ 21811.71	5.30
2	Concordance	disorder	6211	0.27	1013	0.03	+ 7374.61	3.37
3	Concordance	association	15877	0.69	9055	0.23	+ 7162.91	1.57
4	Concordance	bipolar	3821	0.17	94	0.00	+ 6761.08	6.10
5	Concordance	depression	4090	0.18	195	0.01	+ 6688.05	5.15

Results - Data

We have a corpus based on genetic association studies of immune-related diseases (21,422 papers, 4,815,641 words) and one based on psychiatric diseases (15,151 papers, 2,817,417 words)

Results - findings

- As expected subject specific words have a much higher proportional representation (Figure 1).
- Once expected words are removed other less predictable words such as "hydroxylase" are also found to be more frequent in psychiatric literature (Figure 2).
- Preliminary finding suggest genetic psychiatric literature is much more focused on family based studies.

Figure 1

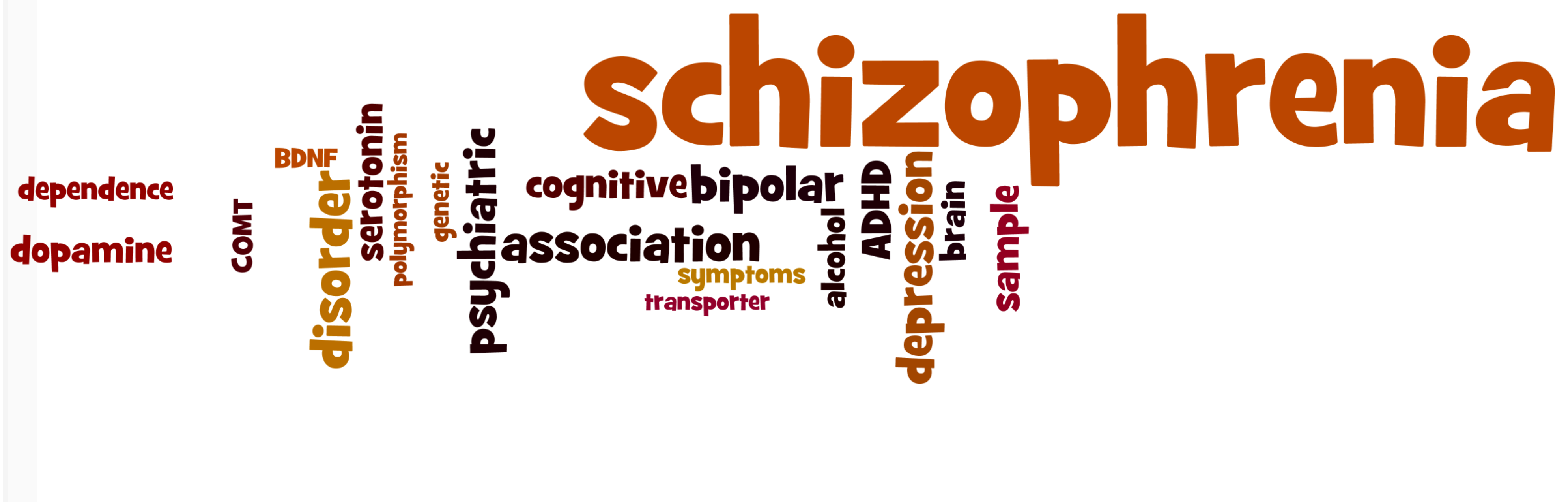
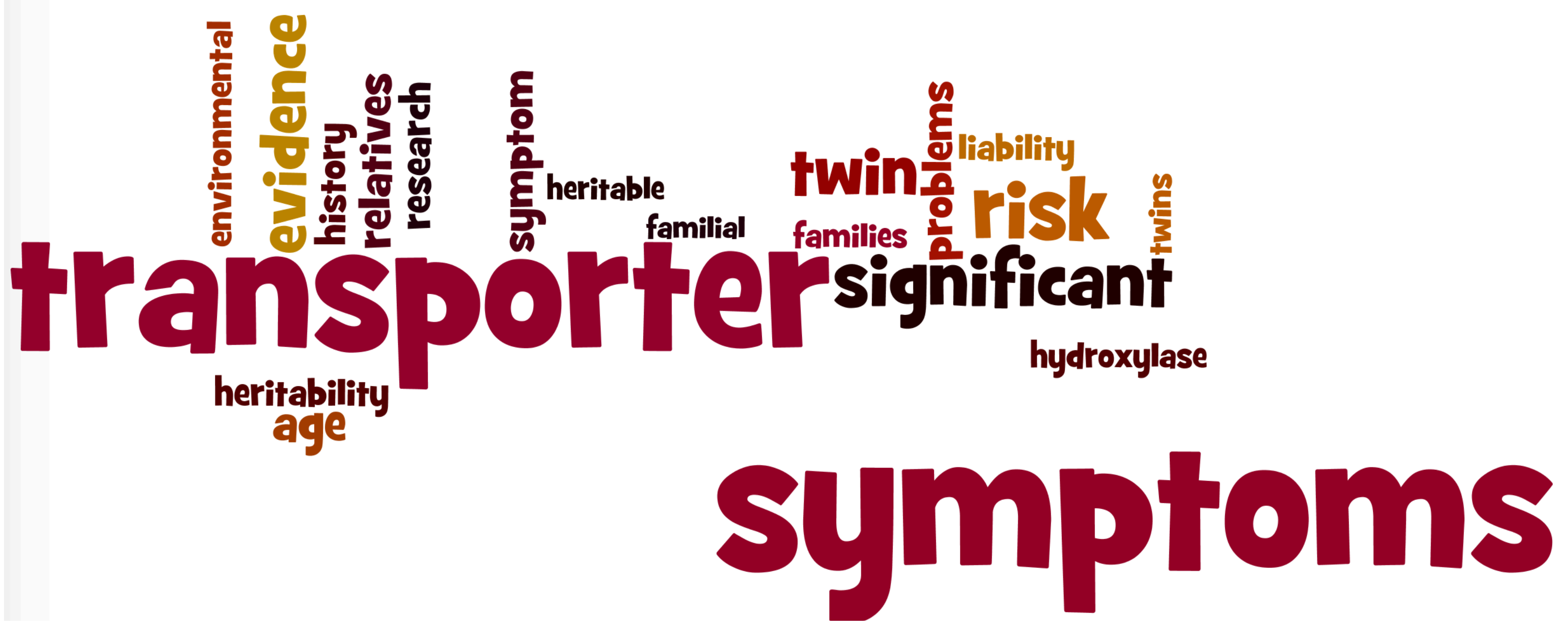


Figure 2



Conclusion and Future Work

This approach identifies known differences in the literature and has the potential to identify unknown differences in the literature.

We are currently working on:

- reproducible methods to filter expected words to allow us to focus on new discovery.
- further refinement of the searches.
- Detailed exploration of the results using techniques such as semantic tagging based on known ontologies.

Project Website: http://wp.lancs.ac.uk/btm/
Data Science Institute – Lancaster University http://www.Lancaster.ac.uk/dsi/