Şenay Kafkas[1], Jee-Hyub Kim[1], Jo McEntyre[1] and Nigel Collier[1]

(1) European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

# Analysis of PhenoMiner

## phenotypes in the open access full text literature

## 1. Introduction

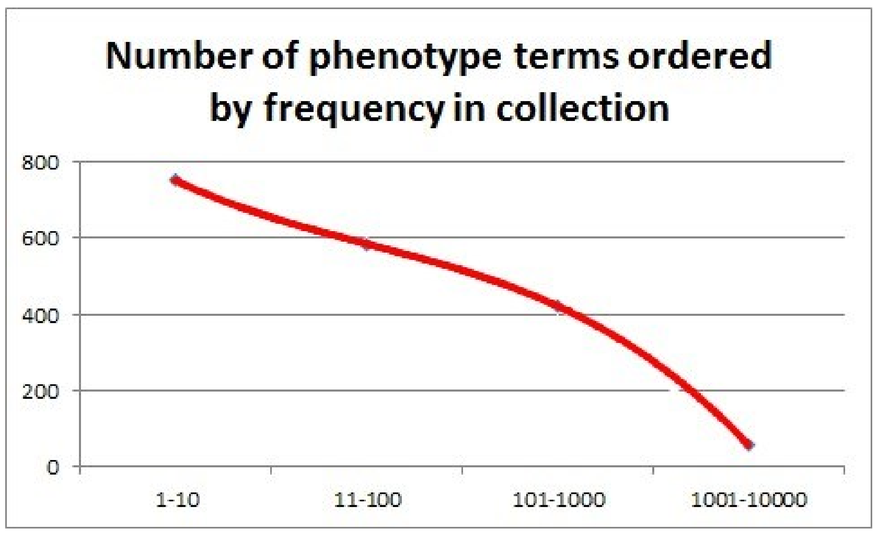
The free text scientific literature has enormous potential to support knowledge discovery in biomedical pipelines. Although substantial work has taken place to identify and link biomedical terms such as genes and diseases, until now phenotypic descriptions such as '*parasitized red blood cells*', '*bifid uvula*' and '*discoid rash*' have not received much attention. Phenotype terms denote important clinical concepts through variation from normal morphology, physiology or behaviour [1] and can be used as evidence to help identify the disease under consideration. In recent years ontologies [2,3] have been constructed by hand using expert intuition but we still do know little about the actual terms used by authors or their distribution in the literature. In this paper we begin to make progress in this area by using a database of phenotypes (DOI 10.5281/zenodo.12493) mined from the BMC open access corpus (http://www.biomedcentral.com/about/ datamining) and looking at the distributional properties across the whole of the open access literature.

## 2. Phenotype identification

In order to capture phenotype descriptions and harmonise them to external ontologies a natural language processing pipeline was constructed that exploited named entity recognition [3], full parsing and conceptual analysis using the NCBO Annotator to derive an Entity Quality representation [3]. Candidate phenotypes were associated with human heritable disorders from OMIM using association analysis on PubMed literature citations which were then filtered using association rule mining [4]. The output of 4,898 phenotypes from this pipeline has been used in these experiments to provide the basis for indexing the open access full text collection at Europe PMC (http://europepmc.org/). The resulting phenotypes and their associations are available from Zenodo (DOI 10.5281/zenodo.12493). Using the 866k full text articles in the open access collection, we filtered to keep only those that had been published after 1989 with XML formatting. This provided 786k OA articles which we then formed the basis for our analysis below.

## 3.Frequency distribution of phenotypes across the OA corpus

Figure 1 below shows a plot of the frequency ranges of the PhenoMiner (PM) terms (1-10, 11-100, 101-1000 and 1001-10000) against the number of terms in those frequency ranges. Note that the values on the curve do not represent a continuous sample. Whilst the frequency of terms is inversely proportional to their rank frequency, they do not appear to display the classic Zipf's distribution.

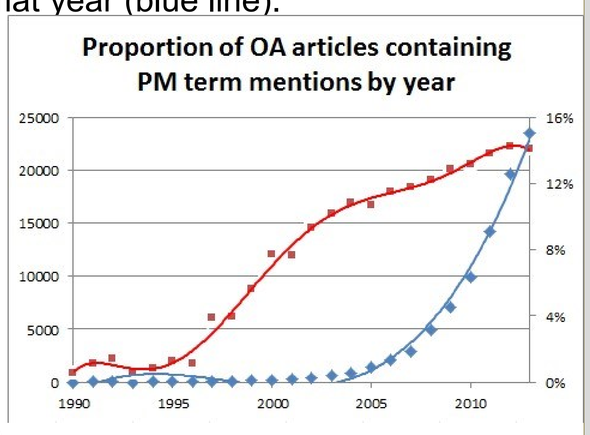



The top 5 most frequent PM phenotypes in the corpus are '' severe hypoglycemia" (PMI:004267), "white matter lesions" (PMI:004885), "peripheral blood smear" (PMI:003572) and "phosphorylated tau" (PMI:003635). Other highly ranked PM terms contain entity quality patterns that conform to phenotypes but are actually anatomical expressions (e.g. "superior temporal gyrus", PMI:006460), disorders (e.g. "peripheral artery disease", PMI:003563) or procedures (e.g. "solid organ transplantation", PMI:004497).

## 4. Phenotype mentions over time

Figure 2 below shows the number of PM term mentions in OA articles by year of publication. We can see that in 2013 this reaches a peak of 19,637 articles mentioning at least one phenotype (red line) – or about 14% of the articles published in that year (blue line).

The proportion of articles showing relevance to phenotypes in the PM database appears to be growing at a polynomial rate (blue line). This might be due to a shift in focus towards phenotype-oriented studies or It might simply reflect the overlap between the OA corpus and the source BMC collection.




## 5. Phenotype mentions by Journal

We examined the distribution of PM terms for each OA journal and ranked each journal by the proportion of articles mentioning at least 200 PM terms. The top ranked journals indicate topical relevance for the phenotypes in our database: (1) Orphanet Journal of Rare Diseases (52.3%), (2) Annals of the Indian Academy of Neurology (45.7%), (3) Journal of Medical Case Reports (39.7%), (4) Cardiovascular Ultrasound (38.9%) and (5) the Journal of Headache and Pain (38.8%). The least topically relevant journals applying the same criteria were (1) Nucleic Acids Research (2.7%), (2) Scientific Reports (4.8%), (3) BMC Genomics (5.4%), (4) The Scientific World Journal (6.2%) and the EMBO Journal (6.2%). This might indicate that the phenotypes in our collection are both more applicable and more likely to appear in journals that are specific to affected systems.

## 6. Implications and future work

The present study provides a useful insight into the distribution of phenotypes in the open access full text literature. We have shown how this distribution varies across time and across journals. The data appears to be showing increased mentioning of phenotypes over time. At the same time it indicates an extremely small set of very high frequency phenotypes which could be focussed on a narrow set of disorders such as breast cancer, skin cancer and pain. In future work we intend to cluster and categorise the phenotypes in more detail to try and understand these patterns more concisely.

## References

[1] Robinson, P and Webber, C (2014), "Phenotype ontologies and cross-species analysis for translational research", PLoS genetics, 10(4). [2] Robinson, P. N. et al. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, *83*(5), 610-615, [3] Smith, C. L., et al. (2004). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, *6*(1), R7, [4] Collier et al. (2014), "PhenoMiner: from text to a database of phenotypes associated with OMIM diseases", under review.

## 5. Acknowledgements

Nigel Collier gratefully acknowledges a Marie Curie grant for the PhenoMiner project (grant no. 301806). All data available at DOI 10.5281/zenodo.12493 and collaboration with Tudor Groza, Anika Oellrich, Damian Smedley, Peter Robinson and Dietrich Rebholz-Schuhmann during development and evaluation of PhenoMiner.

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK

Tel. +44 (0) 1223 494 444
comms@ebi.ac.uk
www.ebi.ac.uk