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

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

OMIM concept annotation:

Steps towards automated tagging the disease literature using PhenoMiner phenotypes

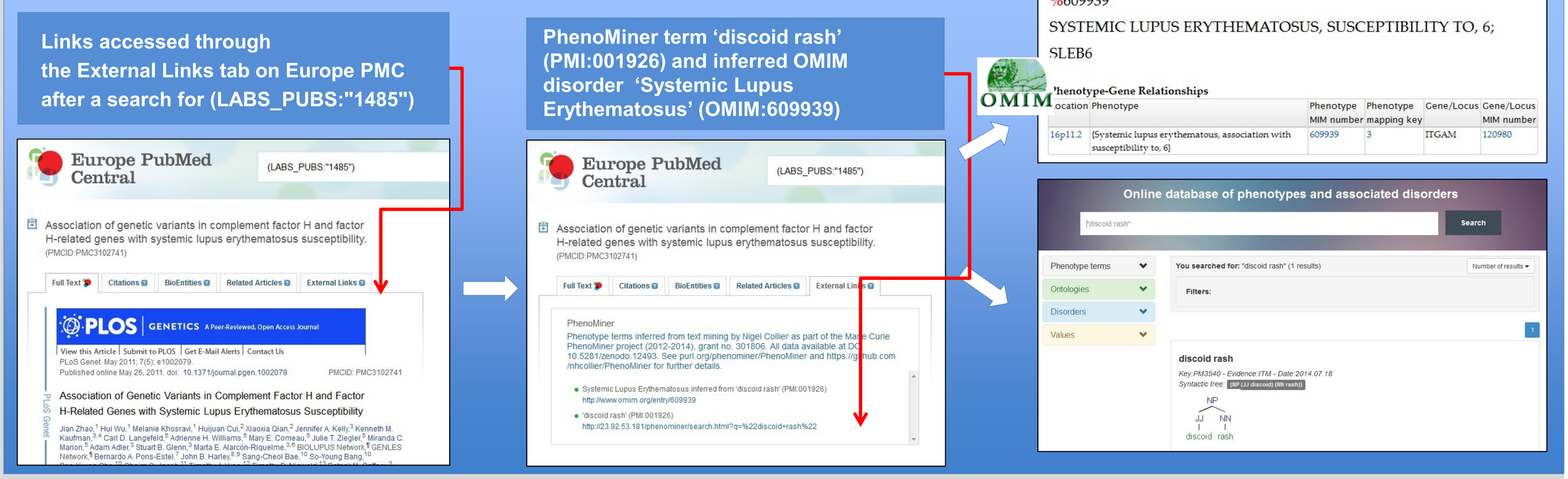
1. Introduction

Genetic dispositions play an important role in numerous human diseases. In order to understand the causes linking genes to associated disease, researchers have employed high throughput experiments to discover evidence. The Online Mendelian Inheritance of Man (http://www.ncbi.nlm.nih.gov/omim) is a manually curated database that provides bibliographic details of over 13,000 human genes and 7,800 diseases, hosted by the Johns Hopkins University. In these experiments we show our first steps towards a phenotype-based system that can automatically annotate full text scientific articles in Europe PubMed Central with OMIM concept identifiers. We compare an experimental system based on PhenoMiner phenotypes against two standard data sets: the OMIM-curated open access literature and author-mentioned OMIM accession numbers in the open access literature. Our findings are illustrated online in the EuropePMC Links service (Figure 1).

3. Gold/Silver standard annotations

As our test bed we chose to use the XML-formated subset of the open access (OA) literature published after 1989 (786k articles). From these we chose two subsets: articles that were cited by OMIM curators (3050 Articles, 5173 OMIM annotations) – considered as 'gold standard' and designated as S1 in our experiments, and articles where we used EBI's WhatIzIt [6] to text mine author citations for OMIM accession numbers (2048 articles, 5716 OMIM annotations) – considered as 'silver standard' and designated as S2 in our experiments. The 5 most common disorders were: (a) curators: breast cancer (OMIM:114480), Tumor protein P53 (OMIM:191170), Melanoma, cutaneous malignant, susceptibiltiy (OMIM:155600), Autism (OMIM:209850) and Glioma susceptibility (OMIM:137800); (b) authors: Prada-Willi syndrome (OMIM:176270), Fabry disease (OMIM:301500), Digeorge syndrome (OMIM:188400), Angelman syndrome (OMIM:105830) and Barth syndrome (OMIM:30206).

Fig 1: High specificity phenotypes and OMIM associations shown in Europe PMC Links



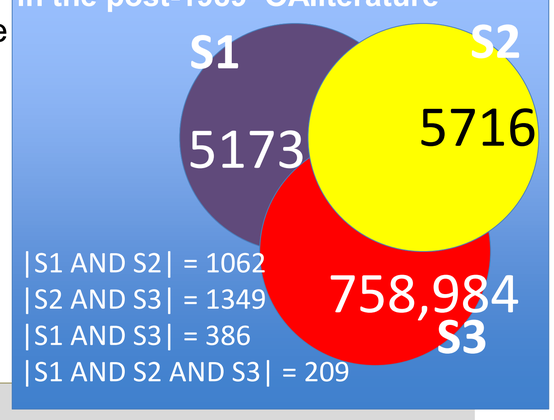
2. Phenotype identification

Descriptions such as '*parasitized red blood cells*', '*bifid uvula*' and '*discoid rash*' denote clinical phenotypes that show a deviation from normal morphology, physiology or behaviour [1]. In order to capture such descriptions and harmonise them to external ontologies a natural language processing pipeline was constructed that exploited named entity recognition [2], full parsing and conceptual analysis using the NCBO Annotator (http://bioportal.bioontology.org/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 our automated OMIM annotation system. The resulting phenotypes and their associations are available from Zenodo (DOI 10.5281/zenodo.12493).

4. Inferring OMIM disorders

We filtered the output of the PhenoMiner associations to keep only the 10 most strongly associated phenotypes with the top 10 ranked phenotype associations. Using this list of high specificity associations we tagged the full open access collection of articles, assigning an OMIM identifier to the document if the phenotype appeared anywhere in the full text. As shown in Figure 2 this resulted in 758,984 OMIM terms being assigned to the articles – an order of magnitude greater than either the curation or author assignment methods. On a simple measure of overlap we found that the PhenoMiner inferred OMIM terms (S3) agreed most strongly with those from authors (S2: 1349 agreements out of 5716 author-assigned terms) with less overlap to curators (S1: 386 agreements out of 5173 curator-assigned terms). It is interesting to note that the author-derived OMIM annotations (S2) and the OMIM annotations inferred from text mining (S3) have a much higher agreement than (S1) and S3. This implies that surface level phenotype mentions might be more valuable for learning author OMIM classifications. In this pilot study we have shown potential for automated enrichment of the full text literature with OMIM identifiers and highlighted important differences in author and curator assignment s.

Fig 2: Number of OMIM annotations in the post-1989 OAliterature



References

[1] Robinson, P and Webber, C (2014), "Phenotype ontologies and cross-species analysis for translational research", PLoS genetics, 10(4). [2] Collier, N. et al. (2013), "Learning to recognize phenotype candidates in the auto-immune literature using SVM re-ranking", *PloS one*, 8(10). [3] Washington, N. et al. (2009), "Linking human diseases to animal models using ontology-based phenotype annotation", *PLoS biology*, *7*(11). [4] Collier et al. (2014), "PhenoMiner: from text to a database of phenotypes associated with OMIM diseases", under review. [5] Kafkas, Ş. et al. (2013). Database citation in full text biomedical articles. *PloS one*, *8*(5). [6] Rebholz-Schuhmann, D., et al. (2008). Text processing through Web services: calling Whatizit. *Bioinformatics*, *24*(2), 296-298.

5. Acknowledgements

NC gratefully acknowledges a Marie Curie grant for the PhenoMiner project (grant no. 301806). All data available at DOI 10.5281/zenodo.12493.

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