For Peer Review

Phenopolis: an open platform for harmonisation and analysis of genetic and phenotypic data

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Bioinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Category:</td>
<td>Applications Note</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
</tbody>
</table>

| Complete List of Authors: | Pontikos, Nikolas; University College London, UCL Genetics Institute; University College London Institute of Ophthalmology; Moorfields Eye Hospital |
|                          | Yu, Jing; University of Oxford Nuffield Department of Clinical Neurosciences |
|                          | Blanco-Kelly, Fiona; University College London Institute of Ophthalmology; Moorfields Eye Hospital |
|                          | Wong, Tsz Lun Ernest; Barts and The London School of Medicine and Dentistry Blazeard Institute |
|                          | Vulliamy, Tom; Barts and The London School of Medicine and Dentistry Blazeard Institute |
|                          | Cipriani, Valentina; University College London, UCL Genetics Institute; University College London Institute of Ophthalmology |
|                          | Fiorentino, Alessia; University College London Institute of Ophthalmology |
|                          | Arno, Gavin; University College London Institute of Ophthalmology; Moorfields Eye Hospital |
|                          | Murphy, Cian; University College London, UCL Genetics Institute |
|                          | Greene, Daniel; University of Cambridge, Department of Haematology |
|                          | Jacobsen, Jules; William Harvey Research Institute |
|                          | Clarck, Tristan; University College London, Computer Science |
|                          | Gregory, David; University College London, Computer Science |
|                          | Nemeth, Andrea; University of Oxford, Nuffield Department of Clinical Neurosciences |
|                          | Halford, Stephanie; University of Oxford, Nuffield Laboratory of Ophthalmology, Nuffield Department of Clinical Neurosciences |
|                          | Downes, Susan; John Radcliffe Hospital |
|                          | Black, Graeme; Central Manchester University Hospitals NHS Foundation Trust, Manchester Royal Eye Hospital |
|                          | Webster, Andrew; University College London Institute of Ophthalmology; Moorfields Eye Hospital |
|                          | Hardcastle, Alison; University College London Institute of Ophthalmology |
|                          | Plagnol, Vincent; University College London, UCL Genetics Institute |

| Keywords: | Human Phenotype Ontology, Web services, Database |
Subject Section

Phenopolis: an open platform for harmonisation and analysis of genetic and phenotypic data

Nikolas Pontikos1,2,3, Jing Yu4+, Fiona Blanco-Kelly2,3, Tom Vulliamy5, Tsz Lun Wong5, Cian Murphy1, Valentina Cipriani1,2,3, Alessia Fiorentino2, Gavin Arno2,3, Daniel Greene6,7, Julius OB Jacobsen8, Tristan Clark9, David S Gregory9, Andrea Nemeth4, Stephanie Halford10, Susan Downes11, Graeme C Black12, Andrew R Webster2,3, Alison Hardcastle2 and Vincent Plagnol1

1 UCL Genetics Institute, University College London, London WC1E 6BT, UK, 2 Institute of Ophthalmology, University College London, London EC1V 9EL, UK, 3 Moorfields Eye Hospital, London EC1V 2PD, UK, 4 Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom, OX3 9DU, UK, 5 Blizzard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK, 6 Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK, 7 Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK, 8 William Harvey Research Institute, Barts and The London, Queen Mary’s School of Medicine and Dentistry, John Vane Building, Charterhouse Square, London, EC1M 6BQ, UK, 9 Computer Science Department, University College London, Gower Street, London, WC1E 6BT, UK, 10 Nuffield Laboratory of Ophthalmology, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK, 11 Oxford Eye Hospital, John Radcliffe Hospital, Oxford, OX3 9DU, UK and 12 Manchester Royal Eye Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK

* To whom correspondence should be addressed. † Authors contributed equally.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Phenopolis is an open-source web server providing an intuitive interface to genetic and phenotypic databases. It integrates analysis tools such as variant filtering and gene prioritisation based on phenotype. The Phenopolis platform will accelerate clinical diagnosis, gene discovery and encourage wider adoption of the Human Phenotype Ontology in the study of rare genetic diseases.

Availability and Implementation: A demo of the website is available at https://phenopolis.github.io. If you wish to install a local copy, source code and installation instruction are available at https://github.com/pontikos/phenopolis. The software is implemented using Python, MongoDB, HTML/Javascript and various bash shell scripts.

Contact: n.pontikos@ucl.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The molecular diagnosis of rare genetic diseases requires detailed clinical phenotypes and processing of large amounts of genetic data. This motivates large-scale collaborations between clinicians, geneticists and bioinformaticians across multiple sites where patient data are pooled together to increase the chances of solving rare cases, and validating novel genes. For example, the UK Inherited Retinal Dystrophy Consortium (UK-IRDC) has set up a collaboration between London, Manchester,
Oxford and Leeds to solve retinal dystrophies. A complication of multi-
37 site collaborations is that discrepancies in phenotype definitions and
38 interpretation of genetic variants can complicate the genetic diagnosis (Veu
39 et al., 2016). A solution to reduce the variability introduced by different
40 sequencing analysis pipelines is to analyse the sequence data centrally
41 and store the annotated variants in a normalised database. On the clinical
42 side, phenotype harmonisation can be improved by using nomenclatures
43 such as the Human Phenotype Ontology (HPO) Kühler et al. (2014)
44 to translate specific clinical features into a standardised, computer
45 interpretable format. We have integrated these two approaches into
46 Phenopolis, an interactive website that combines genetic and phenotypic
47 databases. With the help of HPO-encoded phenotypes, Phenopolis is
48 able to prioritise causative genes using different sources of evidence,
49 such as published disease gene associations from the Online Mendelian
50 Inheritance in Man (OMIM) (Supplementary Section 1) (Hamosh et al.,
51 2005), abstract relevance from Pubmed publications (Supplementary
52 Section 2), as well as model organism phenotype ontology analysis
53 using Ecominer (Supplementary Section 3) (Robinson et al., 2015).
54 Additionally, Phenopolis uncovers gene phenotype relationships within
55 the stored patient data through variant filtering and statistical enrichment
56 of HPO terms using and Phenogenom (Supplementary Section 4) and
57 Simfoleg (Supplementary Section 5) (Greene et al., 2016). The online
58 version, available at https://phenopolis.github.io, includes four example
59 patients with inherited retinal dystrophies and access to per gene analysis,
60 to illustrate our methods.

2 Implementation

2.1 Clinical data collection

The collection of clinical phenotype data was done retrospectively from
61 patient records and entered using the Phenotips platform (Girdia et al.,
62 2013), which provides an interface for translating detailed clinical
63 phenotypes into HPO terms. Several patient diagnoses were translated
to their closest match using HPO terminology. This included mode of
64 inheritance and modifiers such as age of onset and laterality when available.

2.2 Genetic data collection

Our internal exome database, UClex, currently comprises 4,449 patients,
collected from various research groups since 2012. Four patients solved
with genetic mutations in DRAM2 (El-Azrag et al., 2015) and TILL5
(Sergouniotis et al., 2014) are made available on the demo account.

2.3 Analysis of genetic data

The short read sequence data was aligned using novalign (version
65 3.0.08), and variants and indels were called according to GATK best
practices (joint variant calling followed by variant quality score
66 recalibration) (McKenna et al., 2013). The variants were then annotated
using the Variant Effect Predictor (McLaren et al., 2016), output to
67 JSON format, post processed by a Python script and loaded into a Mongo
68 database.

2.4 Website implementation

The Phenopolis website was implemented using the Python Flask web
69 framework by extending the ExAC codebase [11](runnning on top of a Mongo
70 database (Figure 1.A). Javascript was used for visualisations (mostly using
71 D3.js) and to provide interactive features. The website provides five main
entry points:

- The home page: summary statistics of genetic and phenotypic data, as well as auto-completing search bar to search by phenotype, gene name or patient id.
- The all patients page: summary data of all patients and their candidate genes for which the user has access permission.
- The patient page: the patient phenotypes and a table of filtered variants per patient prioritised based on gene. The causal variants are expected to be in this list, ranked at the top of the table.
- The gene page: the variants and the patients in which they occur, as well as the gene-HPO analysis.
- The phenotype page: a prioritised list of genes per phenotype, based on known association and gene enrichment analysis.

3 Applications

3.1 Clinical application: gene prioritization by patient

Given a list of genetic variants and the phenotype of a patient, the first task
5 towards a molecular diagnosis is to prioritise potentially causative genes.
For each case, variants are first filtered based on user-defined thresholds:

- Allele count less than 5 in our internal database and in ExAC (Lek et al., 2015).
- K-scores frequency less than 0.05
- Exclude non-exonic variants or variants on non-coding transcripts. Splicing variants are kept.

Next, gene panels from the gene to HPO/OMIM mapping available on the
72 HPO website [2], and more specialist gene panels, such as Retnet [3]
for retinal genes, are used to highlight candidate genes which match the
phenotypic description and inheritance pattern. We have also developed a
Venn diagram visualisation to highlight genes which are associated to
more than one phenotype (Supplementary Section 1) (Figure 1.B). We also
provide a filterable variant table in which genes are ranked based on their

![Fig. 1. A. Overview of the pipeline. HPO-encoded phenotypes are entered using Phenotips. The Variant Call Format file is annotated by the Variant Effect Predictor and translated to JSON for import into MongoDB. OMIM, Pubmed and Ensembl data are also imported into the Mongo database, on which we run the Phenopipe, Exomizer, Simfoleg and Phenogenom to cense the genes. A Python Flask server is used as the front end and to display the front end points to the website. B. Venn diagram visualisation of HPO gene overlaps highlighting AR12BP. C. Phenopipe visualisation of gene AR12BP (crescent mode). The size of the circle is inversely proportional to the p-value. Clicking on the nodes brings up information about the individuals and variants. "Red cone dystrophy" and "Nyctalopia" are significantly enriched for AR12BP with respective p-values of 0.000172 and 0.00001.](image-url)
Phenopolis

3

“Pontikos_main” — 2016/11/17 — 13:04 — page 3 — #3

Phenopols

Pubmed, Exomiser or Phenogenon gene scores (Supplementary section 6).

3.2 Research application: HPO signature per gene

Given a sufficiently large and phenotypically diverse collection of cases, gene to phenotype patterns start emerging. In order to assign phenotype associations per genes based on our patient database, we have developed a gene-based HPO enrichment and visualisation tool, Phenogenon. (Figure 1.C). We have also integrated the SimReg tool, which suggests a characteristic phenotype per gene (Greene et al., 2016). Both methods work on a filtered list of variants and are explained in detail in the Supplementary sections 4 and 5.

3.3 Research application: genes ranked per HPO term

Individuals with the specified HPO term and their solved gene are listed on this page. We retrieve the list of known disease genes from the gene-HPO/OMIM mapping [2] and we score these genes with Phenogenon to assess their support in our dataset. Furthermore, we rank all genes according to their Phenogenon score for this HPO term to enable gene discovery in our dataset.

4 Discussion

There are currently several closed-source commercial online alternatives that provide variant filtering and prioritisation, for example Saphir [4], Congenica [5] and Omicia [6]. However their costs limit broad usage and they are not readily extensible. There are also open-source alternatives such as Seq2 [7] and Gemini (Paia et al., 2013) but currently neither provides full integration with HPO. As it stands, Phenopolis is an ideal platform for studying pleiotropic genes (Supplementary Figure 3) and how variation in different parts of the same gene could lead to different seemingly unrelated phenotypes. In the next iteration of our software, we plan to integrate tissue expression databases, allowing for genes and transcripts to be prioritised by cell type when the disease affects a specific tissue type. Furthermore, we are working on including copy number variation data, inferred from exomes using ExomeDepth (Plagnol et al., 2013). We also plan on interfacing with the Genomics England GenePanel app to retrieve relevant genes and contribute novel disease genes. Collection of phenotypes and prioritisation of genes can help elucidate which features are informative for a particular gene and warrant close inspection in clinic. The systematic chronological ordering of patient features obtained from clinical history can be informative in discerning between conditions which might appear similar, for example rod-cone and cone-rod dystrophy. Currently, a limitation to obtaining detailed phenotypes for our retrospective cases is the manual input of HPO terms and we are investigating data mining of health records to pull detailed phenotypes for our retrospective cases is the manual input of Rod-cone and Cone-rod dystrophy. Currently, a limitation to obtaining patient features obtained from clinical history can be informative in warrant close inspection in clinic. The systematic chronological ordering of which features are informative for a particular gene and disease genes. Collection of phenotypes and prioritisation of genes can help elucidate the GenePanel app to retrieve relevant genes and contribute novel disease genes. Collection of phenotypes and prioritisation of genes can help elucidate which features are informative for a particular gene and warrant close inspection in clinic. The systematic chronological ordering of patient features obtained from clinical history can be informative in discerning between conditions which might appear similar, for example rod-cone and cone-rod dystrophy. Currently, a limitation to obtaining detailed phenotypes for our retrospective cases is the manual input of HPO terms and we are investigating data mining of health records to pull detailed phenotypes for our retrospective cases is the manual input of Rod-cone and Cone-rod dystrophy. Currently, a limitation to obtaining patient features obtained from clinical history can be informative in warrant close inspection in clinic. The systematic chronological ordering of which features are informative for a particular gene and disease genes. Collection of phenotypes and prioritisation of genes can help elucidate the

5 URLs

1. https://github.com/konradjk/exac_browser
3. https://sph.uth.edu/Retnet
4. www.saphitor.com
5. www.congenica.com
6. www.omicia.com
7. https://seqr.broadinstitute.org/

6 Funding

This work was supported by the UK-IRDC, funded by Retinitis Pigmentosa Fighting Blindness and Fight for Sight.

7 Acknowledgements

We wish to acknowledge all patients that contributed their DNA to this project. We also acknowledge the Computer Science High Performance Cluster for providing us with the computing platform on which to analyse our data and host our webserver. We wish to thank Michael Morgan for his advice on Phenogenon. NP and JY wrote the manuscript and VC, ARW and DG critically reviewed it. CM performed the per HPO term gene burden test. DG carried out the SimReg analysis. JJ and VC helped with the Exomiser analysis. FBK, TV, VC, JY and GA provided the clinical patient data in HPO format. TC and DSG provided the computing infrastructure and the server configuration.

8 References

Fig. 1. A. Overview of the pipeline. HPO-encoded phenotypes are entered using Phenotips. The Variant Call Format files are annotated by the Variant Effect Predictor and translated to JSON for import into MongoDB. OMIM, Pubmed and ExAC data are also imported into the Mongo database, on which we run the PubmedScore, Exomiser, SimReg and Phenogenon to score the genes. A Python Flask server is used as the front-end to display the four entry points to the website. B. Venn diagram visualisation of HPO-gene overlap highlighting ARL2BP. C. Phenogenon visualisation of gene ARL2BP (recessive mode). The size of the circles is inversely proportional to the p-value. Clicking on the nodes brings up information about the individuals and variants. “Rod-cone dystrophy” and “Nyctalopia” are significantly enriched for ARL2BP with respective p-values of 0.00172 and 0.00051.

270x203mm (300 x 300 DPI)