BMC Bioinformatics

SOFTWARE

Open Access

PhenoExam: gene set analyses through integration of different phenotype databases

Alejandro Cisterna[1], Aurora González-Vidal[1], Daniel Ruiz[1], Jordi Ortiz[1], Alicia Gómez-Pascual[1], Zhongbo Chen[2], Mike Nalls[3,4,5], Faraz Faghri[3,4,5], John Hardy[2,9,10,11], Irene Díez[6], Paolo Maietta[6], Sara Álvarez[6], Mina Ryten[2,7,8] and Juan A. Botía[1,2*]

*Correspondence:
juanbot@um.es

[1] Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain
[2] Department of Neurodegenerative Disease, UCL, Institute of Neurology, London, UK
[3] Data Tecnica International LLC, Glen Echo, MD, USA
[4] Laboratory of Neurogenetics, NIA/NIH, Bethesda, MD, USA
[5] Center for Alzheimer's and Related Dememtias, NIH, Bethesda, MD, USA
[6] NIMGenetics Genómica y Medicina S.L, Madrid, Spain
[7] NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London, UK
[8] Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London WC1E 6BT, UK
[9] Reta Lila Weston Institute, UCL Queen Square Institute of Neurology, London, UK
[10] UCL Movement Disorders Centre, University College London, London, UK
[11] Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China

Abstract

Background: Gene set enrichment analysis (detecting phenotypic terms that emerge as significant in a set of genes) plays an important role in bioinformatics focused on diseases of genetic basis. To facilitate phenotype-oriented gene set analysis, we developed PhenoExam, a freely available R package for tool developers and a web interface for users, which performs: (1) phenotype and disease enrichment analysis on a gene set; (2) measures statistically significant phenotype similarities between gene sets and (3) detects significant differential phenotypes or disease terms across different databases.

Results: PhenoExam generates sensitive and accurate phenotype enrichment analyses. It is also effective in segregating gene sets or Mendelian diseases with very similar phenotypes. We tested the tool with two similar diseases (Parkinson and dystonia), to show phenotype-level similarities but also potentially interesting differences. Moreover, we used PhenoExam to validate computationally predicted new genes potentially associated with epilepsy.

Conclusions: We developed PhenoExam, a freely available R package and Web application, which performs phenotype enrichment and disease enrichment analysis on gene set G, measures statistically significant phenotype similarities between pairs of gene sets G and G′ and detects statistically significant exclusive phenotypes or disease terms, across different databases. We proved with simulations and real cases that it is useful to distinguish between gene sets or diseases with very similar phenotypes.

Github R package URL is https://github.com/alexcis95/PhenoExam.

Shiny App URL is https://alejandrocisterna.shinyapps.io/phenoexamweb/.

Keywords: Phenotype Enrichment Tool, Shiny, Epilepsy

Background

One of the main aims of clinical genetics research is to discover new gene-disease associations [1–6]. A disease is commonly diagnosed through the identification of a set of symptoms and signs associated with a particular and recognized clinical phenotype [7–10]. While some phenotypes are due to the impact of environmental factors, if a disease is inherited then the genetic variation within the individual also explains the phenotype at least partially [11]. Here, we introduce PhenoExam, a software tool to assist in the identification of new gene-phenotype associations. PhenoExam focuses on genetic diseases, harnessing all available gene-phenotype annotation resources to provide a comprehensive gene set and differential gene set annotation approach.

Over the last decade, we have seen attempts to standardize our knowledge of genetic diseases by formally linking genes to phenotypes using standard terminology, as exemplified by The Human Phenotype Ontology (HPO) [12] and The Mouse Genome Database (MGD) [13]. HPO is a standardized set of human phenotypic terms that are organized hierarchically with a directed acyclic graph and have been used to annotate all clinical entries in the Online Mendelian Inheritance in Man database (OMIM). OMIM [14] is a continuously updated catalog of human genes, genetic diseases, and traits, with a particular focus on the molecular relationship between genetic and phenotypic variation. On the other hand, MGD is the manually curated consensus representation of genotype to phenotype information including detailed information about genes and gene products. It is the authoritative source for biological reference data sets related to mouse genes, gene functions, phenotypes, and mouse models of human disease. MGD has more terms and detailed phenotypic information than HPO because scientists can perform a wider set of experiments on mice. These features increase our knowledge and can help to prioritize novel gene-phenotype relationships in humans. Beyond phenotype databases, PhenoExam also includes gene-disease association databases, namely UniProt [15], The Comparative Toxicogenomics Database (CTD) [16], Orphanet [17], The Clinical Genome Resource (ClinGen) [18], The Genomics England PanelApp [19], The Cancer Genome Interpreter (CGI) [20] and PsyGeNET [21]. It also includes CRISPRbrain [22], the first genome wide CRISPR interference and CRISPR activation screen in human neurons so we may study the potential association of phenotypic terms to specific functions of these genes in human neurons.

Apart from being a general-purpose tool for phenotype-based gene sets annotation, PhenoExam can also help in the diagnosis of genetic diseases. Currently fewer than half of patients with suspected Mendelian disorders (genetic diseases primarily resulting due to alterations in one gene) receive a molecular diagnosis [23]. Diseases with a genetic basis are usually diagnosed by looking for causal mutations in a panel of genes specifically associated with the disease. Gathering all phenotypes associated with the genes in a panel delivers a general phenotype-level description beyond the disease under study. To improve the accuracy of genetic diagnosis, we need methods to appropriately evaluate the gene level phenotypic similarity between candidate diseases. Moreover, the identification of differential phenotypes between diseases can also help towards more precise diagnostics. The identification of exclusive and/or shared phenotypes between gene panels can demonstrate common pathophysiology [24] but it can also help to create genetic links between diseases through their gene

sets [25, 26]. We can find numerous methods based on measuring disease-based phenotypic similarities by comparing sets of HPO terms e.g., Phenomizer [27], HPOSim [28], and PhenoSimWeb [29], Table 1 offers a detailed comparison amongst all tools. We also have modPhEA [30], an online resource for phenotype enrichment analysis. modPheEA helps with the gene-based phenotype enrichment analysis but just focused on one phenotype database at a time and without considering conditional analyses (two gene sets).

Phenomizer obtains the phenotype semantic similarity between sets of phenotypes based on the HPO ontology but does not rely on the use of the genes implicated in each phenotype. HPOSim is an R package that implements widely used ontology-based semantic similarity measurements to quantify phenotype similarities, and phenotype-level enrichment analysis using a hypergeometric test and the NOA method [31]. PhenoSimWeb is an online tool for measuring and visualizing phenotype similarities using HPO, uses a path-constrained Information Content-based measurement in three steps and exploits the PageRank algorithm [32]. Nevertheless, these tools did not take some important concepts into consideration. PhenoExam contributes to the field with new features. These include the ability to detect differential phenotypes between pairs of gene sets: phenotypes that are significant within one gene set only, useful for detecting featured phenotypic terms between gene sets to distinguish better between similar diseases. It also combines phenotype and disease terms. This is important to link phenotypes to specific diseases. Finally, it tries to make the interpretation of the results of the phenotypic analysis easier by using simple scores to rank significant terms as well as summary messages and interactive graphs. We also found a knowledge management platform integrating and standardizing data about disease-associated genes from multiple sources called DisGeNET [33]. While being similar to PhenoExam in finding gene-disease associations, DisGeNET does not, however, offer facilities for gene-based phenotype enrichment analysis or for detecting phenotypic conditional similarities between pairs of gene sets. PhenoExam uses as the basic substrate for gene-phenotype and gene-disease associations a number of configurable databases both in human and mouse that the user can tailor and adapt depending on the type of analysis to be performed. In PhenoExam, the phenotypic similarity between two groups of genes is performed by assessing the statistical significance of the Phenotypic Overlap Ratio (POR) between those (i.e., the number of common enriched phenotypes between the gene sets) (See methods Phenotype scores calculation).

We developed PhenoExam intending to support a variety of target users, mainly clinicians, computational biologists, and geneticists. PhenoExam can help clinicians with finding phenotypes which are exclusive to diseases amongst a set of possible genetic disease candidates whose diagnosis is based on gene sequencing panels (Case 1). PhenoExam is also useful for geneticists as it can be used to improve their in-house-maintained gene panels but also to more accurately select genes involved in specific genetic studies (Case 2). Finally, computational biologists can use PhenoExam to discover new information about gene sets of interest thanks to the integration of multiple phenotype and disease databases and to compare phenotypes between known genes associated with a disease and the validation of computationally predicted disease genes (Case 2).

Table 1 Comparison of PhenoExam and other similar tools. "X" means the tool provides the function and "–" means the tool does not. "*" means the similarity scores are between phenotype terms and not between gene sets as does PhenoExam

Tool	As web	As software tool	Open source	Model Organism	Phenotype sets	Gene sets	Multiple database at once	Phenotype Enrichment Analysis	Disease Enrichment Analysis	Differential phenotypes	Diagnosis based on phenotypes	Similarity scores
PhenoExam	X	X	X	X	X	X	X	X	X	X	–	X
modPhEA	X	–	–	X	X	X	–	X	–	X	–	–
DisGeNET	X	X	X	–	X	X	X	–	X	–	–	–
Phenomizer	X	–	–	–	X	–	–	–	–	–	X	*
HPOSim	–	X	X	–	X	X	–	X	–	–	–	*
PhenoSimWeb	X	–	–	–	X	X	–	–	–	–	–	*

Design and implementation

Database integration

The set of analyses performed by PhenoExam is based on manually curated phenotypes language like HPO, gene-disease ones as OMIM but also screening-based databases like CRISPRBrain, amongst many others (see Table 2 for a complete list, description, and potential use). PhenoExam can perform a variety of analyses (Fig. 1). The integration of these different databases is possible thanks to a well-established standardization process of genes and phenotypes used by PhenoExam. Using the HUGO Gene Nomenclature Committee (HGNC) gene naming system as the common way of identifying all human genes, and the definition of a new annotation term within each annotation database to indicate the HGNC genes that do not have any phenotype term associated in the database of interest. The list of HGNC genes was obtained from [34] https://www.genenames.org/download/statistics-and-files/. The HPO gene-phenotype association list was obtained from https://archive.monarchinitiative.org/latest/tsv/gene_associations/. The new no-phenotype association (HPO:XXX No HPO phenotype) was added to HPO for all protein coding genes with no known association to phenotype. For MGD, MP terms from orthologous genes to humans were obtained from http://www.informatics.jax.org/downloads/reports/index.html#go, and the relationship between human genes—mouse genes—mouse phenotype were collected using the files (MGI_PhenoGenoMP.rpt, HMD_HumanPhenotype.rpt, VOC_MammalianPhenotype.rpt). A new no- phenotype association (MP:XXX No phenotype) was created and all the protein coding genes without a relation to phenotype were linked to this term. For CRISPRBrain, the gene-phenotype relationships were obtained from https://crisprbrain.org/simple-screen/. For the generation of this database, the phenotypes were codified in three classes for each CRISPR analysis: association to the phenotype (Positive-Hit and Negative-Hit genes in CRISPRBrain), positive association (Positive-Hit genes in CRISPRBrain) and negative association (Negative-Hit genes in CRISPRBrain). This was accomplished according to the Hit-Class label in CRISPRbrain (Positive-Hit, Negative-Hit). The non-relationship phenotype (CRB:XXX No phenotype was created and all the protein coding genes that

Table 2 Databases usable through PhenoExam and size of each in terms of genes, phenotypes and associations. Numbers reported are final, after preprocessing and unification of gene names across databases

Source	Genes	Phenotypes	Diseases	Assocs	Summary
HGCN	19,197	–	–	–	All protein coding genes
HPO	19,248	7861	–	186,290	Human gene-phenotype associations
MGD	17,900	10,243	–	242,313	Mouse gene-phenotype associations
CRISPRBrain	19,275	55	–	43,481	Cell screen gene-phenotype associations
ClinGen	19,198	–	420	19,851	Human gene-disease associations
Genomics England	19,230	–	5538	24,336	Human gene-disease associations
CTD	19,636	–	6843	58,660	Human gene-disease associations
CGI	19,198	–	177	20,361	Human gene-disease (cancer) associations
UniProt	19,204	–	3868	21,101	Human gene-disease associations
Orphanet	19,262	–	3183	2228	Human gene-disease (rare) associations
PsyGeNET	19,248	–	82	20,952	Human gene-disease associations
ALL	20,209	18,159	9348	544,022	PhenoExam tool

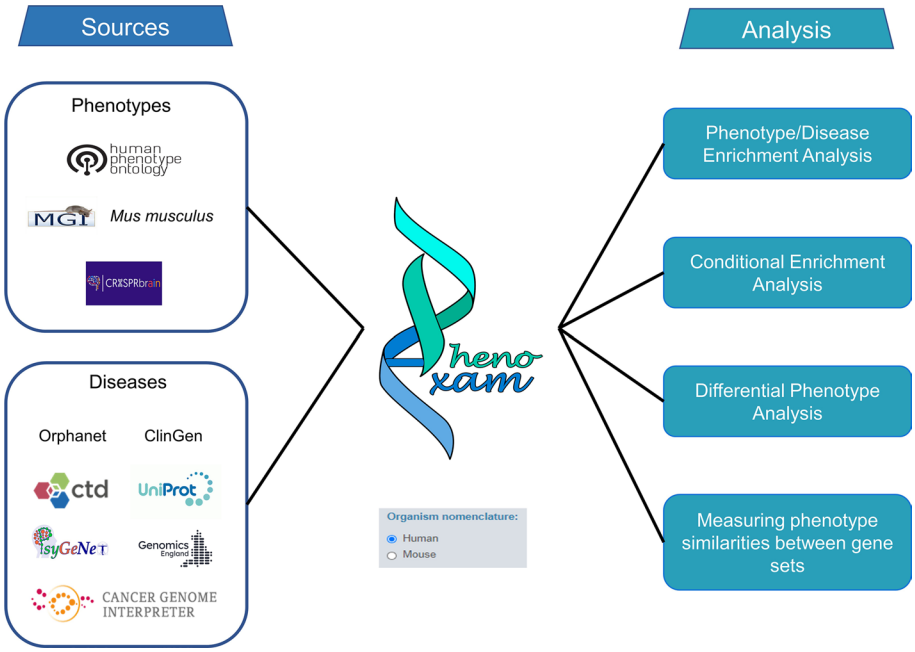

Fig. 1 Schematic representation of PhenoExam integrated databases and offered analyses. We can use PhenoExam with human or mouse genes. PhenoExam annotation databases include HPO, MGI, CRISPRBrain, CTD, ClinGen, OrphaNET, UniProt PsyGeNET, CGI and Genomics England. The tool offers a variety of analyses. Given a gene set of interest, G, the user can evaluate its enrichment for phenotypes and disease in all or a subset of the offered databases. Given two gene sets, G and G′, the user can evaluate whether the phenotype terms enriched in G are also enriched in G′ when G and G′ do not overlap e.g., G′ was predicted from G, with the Conditional Enrichment Analysis. If G and G′ show some gene overlap, the user can assess whether the gene sets show any differential phenotypes through the Differential Phenotype Analysis. We acknowledge all the sources for their contributions and we are grateful to those who permitted us to use their logos in this figure

were not related to any phenotype were related to this term. We integrate into PhenoExam only the information from curated databases (UniProt, CTD, Orphanet, ClinGen, The Genomics England PanelApp, CGI and PsyGeNET). Then the non-relationship disease term (CXXX No diseases associated) was created and all the protein coding genes that were not related to any disease were related to this term. After standardization process, the current release (v1.0) of PhenoExam contains, 659,634 gene-phenotype associations, involving 20,209 genes, 18,159 different phenotypes and 9348 different diseases (see details in Table 2).

Phenotype scores calculation

Phenotype enrichment analysis on a gene set G

PhenoExam obtains a list of statistically significant enriched phenotypes in a given set of gene *G* within a phenotype/disease database annotation of reference D. In order to calculate whether a gene set *G* shows enrichment in a given phenotypic term *p* belonging to D, let g be the number of genes in G associated with p. Let also gdb be the number of genes associated with p and GDB the total number of genes in the database, we model the enrichment probability with a hypergeometric distribution such that:

$$P(X = g) = \frac{\binom{gdb}{g}\binom{GDB - gdb}{|G| - g}}{\binom{GDB}{|G|}}$$

Any phenotype with $P < 0.05$ will be enriched in the G gene set. We compute this probability for each phenotypic term ph associated with 1 gene or more in G and use these probabilities as P values. PhenoExamWeb reports the raw, Bonferroni [35] and false discovery rate (FDR) [36] adjusted P values.

Phenotypic Overlap Ratio score

PhenoExam´s approach to measuring the similarity between two gene sets G and G´, within an annotation database D, is based on a score called the Phenotypic Overlap Ratio (POR). Let Gp be the number of significantly enriched terms in D for genes in G, and analogously for G′p. The POR could be computed using the widely used Jaccard index or the Forbes similarity coefficient corrected by Alroy [37] on the agreement between the subsets of significant phenotypes. PhenoExam allows users to choose between these two options accordingly to Salvatore et al. [38]. conclusions.

Jaccard index:

$$POR(G, G') = \frac{Gp \cap G'p}{Gp \cup G'p}$$

Forbes similarity coefficient corrected by Alroy:

$$N = Gp \cap G'p + Gp \backslash G'p + G'p \backslash Gp$$

$$POR(G, G') = \frac{Gp \cap G'p * (N + \sqrt{N})}{[(Gp \cap G'p + Gp \backslash G'p) * (Gp \cap G'p + G'p \backslash Gp) + Gp \cap G'p * \sqrt{N} + (Gp \backslash G'p * G'p \backslash Gp)/2]}$$

POR (G, G') takes values in [0,1], resulting in 0 when no phenotype is shared and 1 when the sets share all phenotypes (Jaccard index) or at least share all phenotypes from one set (Forbes coefficient).

Statistically significant Phenotypic Overlap Ratio

PhenoExam assess whether the POR between gene sets G and G′ is statistically significant by means of randomization. We will have two modalities of the POR, depending on whether G and G′ share genes or, on the contrary, they are disjunct (e.g., G′ was predicted from G). When G and G′ are thought to share genes, POR (G, G') is compared with POR (G, R) and with POR (G', R'), where R has the same size as G and R′ the same as G′. Genes in both R and R′ are chosen randomly within the whole set of protein coding genes. We repeat this process for m random gene sets $(R_1, R_2, \ldots, R_m)$ and $\left(R'_1, R'_2, \ldots, R'_m\right)$ to obtain an empirical P value with the proportion of random gene sets whose POR is greater than the observed one. On the other hand, when G′ is obtained by using G as input of the generation process, we say G′ is conditioned to G. Therefore, the significance test of the POR (G, G′) is reduced now to obtain an empirical P value based on the proportion of times a randomized

POR (G, R), with R any of $(R_1, R_2, \ldots, R_m)$ all with the same size of G while keeping G constant, shows higher values than the observed POR (G, G′).

Relaxed Phenotypic Overlap Ratio

The POR only considers phenotypes that were assessed as statistically significant. Sometimes, it may be of interest to relax this restriction to incorporate all phenotype/disease terms associated with G. In this case, the score is called Relaxed Phenotypic Overlap Ratio (RPOR). It is calculated in a similar way to the POR but with all phenotypes, whether these are enriched or not. In the same way, as with the POR, we can determine whether the RPOR is statistically significant by using randomization.

Phenotype relevance association analysis for gene sets

Once it has been determined that two sets of G and G' genes share some enrichment of phenotypic terms, and focusing only on the shared terms, we can measure the correlation of the number of genes of each phenotypic term as measured in G and G′ by a linear regression model and report the R^2 as the strength of this correlation together with the association *P* value. Higher values of R^2 would suggest a linear association between importance of phenotypic terms in G and importance of the same genes in G′.

Generation of the web interface

We have developed PhenoExamWeb, a web-based tool for performing phenotypic analyses using R. PhenoExamWeb shiny app is accessible at https://alejandrocisterna.shinyapps.io/phenoexamweb/. R and the shiny R package [39] were used for front-end scripting of the web interface. R scripts were used for back-end execution and analysis with the development environment of R version 3.6.3. The R package is available at https://github.com/alexcis95/PhenoExam. Note that although we offer PhenoExam through a Web application, it might be a better option to consider installing and using the R package locally for the sake of flexibility or to deploy the shiny app locally in your local workstation for computationally demanding analyses like, for example, a "comparator phenotype analysis" with more than 40 random tests. Simply download the software from https://github.com/alexcis95/PhenoExam/blob/master/PhenoExamWeb.zip and run the Rmd file locally.

Analysis with PhenoExamWeb

PhenoExamWeb requires gene symbols, human or mouse, as the input file. Then, we need to select the type of analysis: Phenotype Enrichment Analysis (One gene set) or Phenotype Comparator (Two gene sets). We also need to specify the database or databases. The workflow of PhenoExamWeb is summarized in Fig. 2. Users can follow the web tutorial on the website (https://alejandrocisterna.shinyapps.io/phenoexamweb/#section-help) and the R package tutorial on GitHub (https://raw.githack.com/alexcis95/PhenoExamWebTutorials/main/tutorial.html).

Results and discussion

PhenoExam controls type I error when used with all phenotype databases

We assessed PhenoExam for type I error given all phenotype/disease databases considered in the task of phenotypic enrichment analysis of gene sets. Firstly, we evaluated the

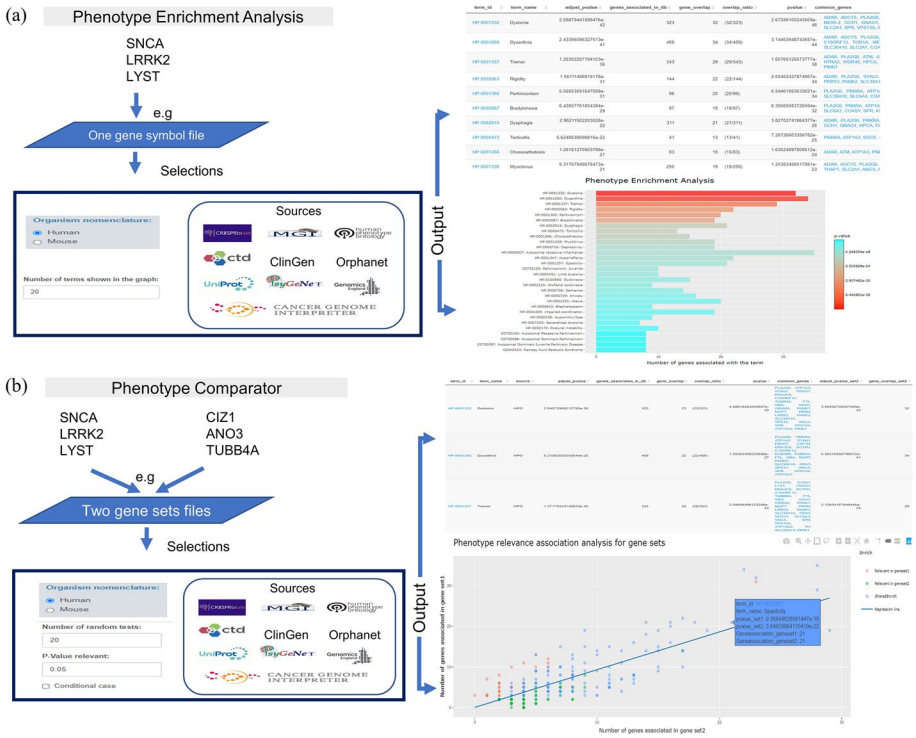

Fig. 2 PhenoExamWeb shiny app possible workflows. **a** Phenotype Enrichment Analysis: requires one gene symbol file as input file, which gene symbol nomenclature (Organism nomenclature: Human or Mouse) we use, the phenotype/disease annotation databases to be considered and the top number of terms shown in the graph. The results generate an interactive table and graph which include phenotypes, genes implicated with each term and *P* values as output. **b** The Phenotype Comparator requires two gene sets as input together with the gene symbol nomenclature (Human or Mouse) used, the annotation databases of interest for the analysis and the number of random tests to obtain empirical *P* values, the relevant *P* value threshold and whether our analysis is a conditional case (i.e., if one gene set was generated after a prediction analysis from the other and they are totally different gene sets). Finally, we obtain the summary of the analysis with the similarities phenotype scores, the differential phenotypes, interactive tables and graphs with phenotypes, genes and *P* values as output for detailed inspection and result presentation. We acknowledge all the sources for their contributions and we are grateful to those who permitted us to use their logos in this figure

possibility of finding a phenotypic term erroneously enriched, due to random chance, amongst all the terms at the database, for gene sets of varying sizes. For such purpose we performed simulations of phenotype enrichment analysis for different random gene sets with a variable number of protein coding gene sizes (5, 10, 20, 40, 80, 160, 320, 640) tested in all annotation databases. Each combination of gene set size and database was simulated 1000 times, yielding a total of 80,000 simulations. A graphical representation of the summary of results appears in Fig. 3. PhenoExam maintains type I error under control, see Fig. 3, (a) plot, with a significance level of 0.05 as the number of significant tests is always under 0.05 ratio. We observed a negative correlation between gene set size and proportion of false positive tests, $r = -0.453$, $P = 0.026$. Type I error is harder to control with Genomics England Panel App (GEL) and Orphanet gene sets. PhenoExam only controls type I error when the gene set size is greater than 80 for Orphanet and 180 for Genomics England. We believe that the difficulties in keeping under control type I error are due to the number of average disease terms associated with each gene, i.e. 4.39 for GEL and 7 for Orphanet when for the rest of the disease databases is, on average,

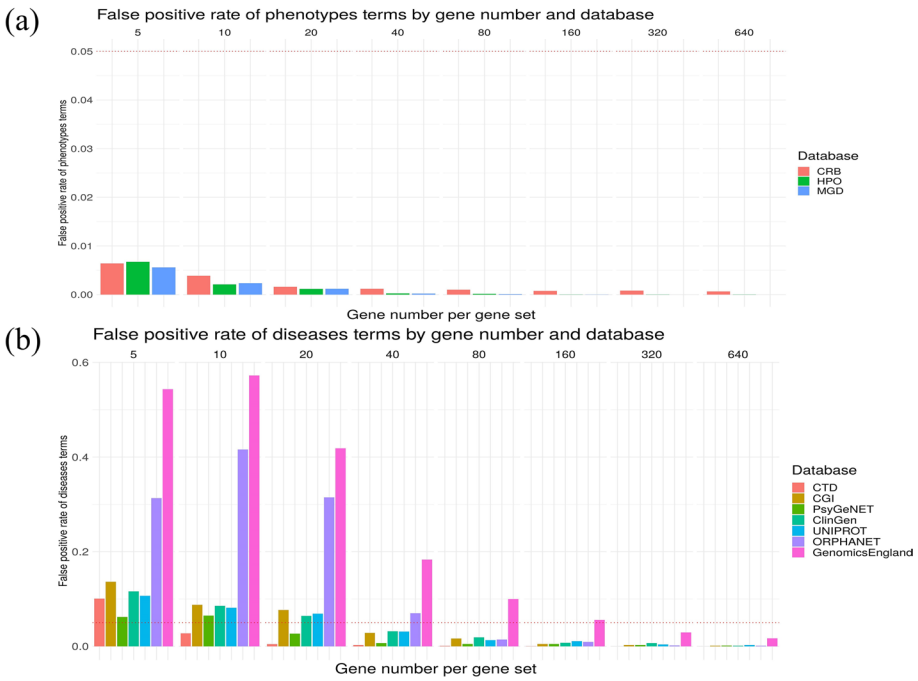

Fig. 3 False positive rate of phenotype (**a**) and disease (**b**) terms enrichment across varying gene set sizes (5, 10, 20, 40, 80, 160, 320, 640) per phenotype/disease database. As the simulation points out, CRB, HPO, MGD, are perfectly usable for any gene set size, CTD is recommended for gene set sizes over 10, PsyGeNET for 20, CGI, ClinGen and Uniprot for 40, Orphanet for 80 and GEL for gene set sizes over 180

17.7. Moreover, there is a negative correlation between the number of genes per random gene set and the type I error, $r = -0.381$, $P = 0.0038$. Therefore, both the number of terms associated with each gene and the size of the gene sets used as input are crucial to obtain enough gene-phenotype relationships to maintain in this way, type I error under control. For these reasons, we recommend using CTD, HPO, MGD or CRB for analyses implying gene sets of size 10. These are, roughly, less than the number of genes we can find in many biological pathway. We recommend using PsyGeNET, ClinGen, UNIPROT or CGI with 40 genes or more. These usually are less than the number of genes detected at most genome-wide association studies. We only recommend the inclusion of the Orphanet and GEL when we have at least 80 and 180 genes respectively. Users can find more information about what database they need to use at https://alejandrocisterna.shinyapps.io/phenoexamweb/#section-help

PhenoExam differentiates between gene sets with very similar phenotypes

We evaluated how accurate PhenoExam is when computing the POR (detecting phenotype similarities) between gene sets by comparing genetic forms of epilepsy (261 genes from NIMGenetics epilepsy panel) and "artificial" gene sets constructed with variable POR with the original epilepsy gene set and additional genes with similar phenotypic connectivity not associated to epilepsy. In these additional genes we injected a 5% of noise with genes associated with epilepsy phenotypic terms. We performed 1000 simulations for the artificial genes sets (261 genes) constructed with different proportions of epilepsy genes between (0–100%) and different proportions of other genes (0–100%). We

calculated the POR significance test between the real and the artificial gene sets (Fig. 4). PhenoExam is sensitive in detecting differences between gene composition changes ($\cong 1\%$) in different gene sets, which in this case are 3 genes. We observed a positive linear relationship between POR and the proportions of epilepsy genes in the artificial gene sets, 0.9674 R^2 ($P < 2.2 \times 10^{-16}$) (Fig. 4a). We assessed that PhenoExam can distinguish well amongst the epilepsy real genes and the artificial gene sets constructed with high proportions of epilepsy genes (94–99% epilepsy genes) that gather very similar phenotypes with a t-test in all cases ($P < 2.2 \times 10^{-16}$) (Fig. 4b).

Case 1: The analysis between juvenile-onset Parkinson's disease (PD) and early onset dystonia (EOD) reveals they hold phenotype-level similarities but also potentially interesting differential phenotypes

We applied PhenoExam to the detection of differential phenotypes between gene sets by comparing two genetic diseases with similar symptoms: juvenile-onset Parkinson's disease (PD) and early-onset dystonia (EOD). PD and EOD both are movement disorders, PD is caused by a degeneration in the basal ganglia, and it has predominant symptoms consisting of tremor, rigidity, bradykinesia, postural instability and progressive dementia

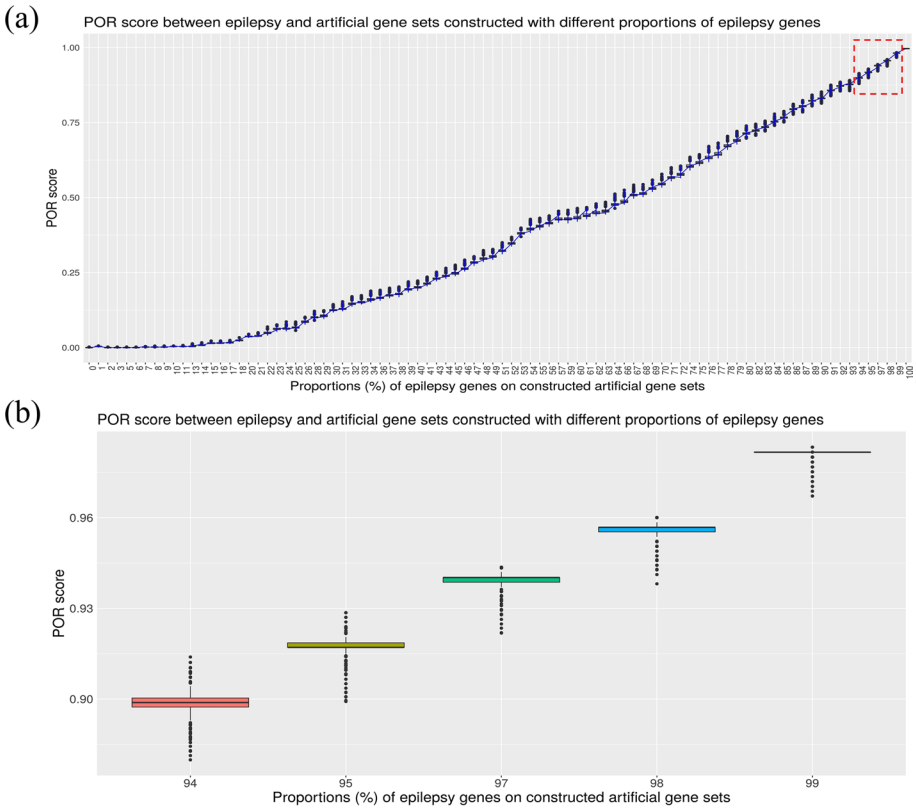

Fig. 4 POR significance test between the real and the artificial gene sets constructed with different proportions of epilepsy genes (**a**) and detailed zoom of POR score between the real and the artificial gene sets constructed with different proportions of epilepsy genes (94–99% epilepsy genes). **a** We observed a positive linear relationship between POR and the proportions of epilepsy genes in the artificial gene sets, 0.9674 R^2 ($P < 2.2 \times 10^{-16}$). **b** PhenoExam can distinguish well amongst the epilepsy real gene set and the artificial gene sets constructed with high proportions of epilepsy genes (94–99% epilepsy genes) that gather very similar phenotypes with a t-test in all cases ($P < 2.2 \times 10^{-16}$)

[40]. EOD is a disease characterized by involuntary muscle contractions leading to abnormal posturing and movements and postures, occurring with or without other neurological symptoms [41]. In our case we compared 35 PD genes and 50 EOD genes from Genomics England PanelApp (Additional file 1), with 19 genes in the overlapping set (54.3% of genes on PD gene set). We ran a separate phenotype enrichment analysis for PD and EOD, using HPO, MGD, CTD and CRISPRBrain databases simultaneously (given the simulation analyses performed above, these are the databases recommended by PhenoExam) (Fig. 5). We obtained a table for PD (Additional file 2: Table S1) and EOD (Additional file 3: Table S2). The top two most enriched phenotypes, in each input database, for PD genes were Bradykinesia (HP: 0002067; $P = 2.16 \times 10^{-60}$) and Parkinsonism (HP: 0001300; $P = 2.62 \times 10^{-51}$) for HPO, Abnormal gait (MP: 0001406; $P = 3.78 \times 10^{-13}$) and Neuron degeneration (MP: 0003224; $P = 9.98 \times 10^{-13}$) for MGD, Parkinsonism, Juvenile (C0752105; $P = 7.49 \times 10^{-28}$) and Ramsay Hunt Paralysis Syndrome (C0242423; $P = 7.49 \times 10^{-28}$) for CTD, and no enrichment found for CRISPRBrain. All the enrichment terms found are supported by the literature [42–45]. At the EOD analysis, we found Dystonia (HP: 0001332; $P = 3.51 \times 10^{-42}$) and Dysarthria (HP: 0001260; $P = 5.38 \times 10^{-41}$) for HPO, impaired coordination (MP: 0001405; $P = 7.4 \times 10^{-14}$) and Abnormal gait (MP: 0001406; $P = 3.17 \times 10^{-10}$) for MGD, Parkinsonism, Juvenile (C0752105; $P = 7.4 \times 10^{-13}$) and Ramsay Hunt Paralysis Syndrome (C0242423; $P = 7.4 \times 10^{-13}$) for CTD, and again no enriched term for CRISPRBrain. Above mentioned phenotype terms are associated with dystonia according to several articles [46–50].

We wanted to compare PD and EOD gene sets, through the Phenotype Comparator analysis in PhenoExamWeb (see Fig. 6) using HPO, MGD, CTD and CRISPRBrain as the databases selected, and a randomization based on 1000 null tests. This comparison yielded 139 shared significant phenotypic terms (out of 273 unique significant phenotypic terms in both, POR = 0.509 ($P < 0.001$). Phenotype relevance association analysis for PD and EOD (i.e., whether the shared phenotypes are similar in relevance, i.e., in the number of genes associated with them, within each gene set) results in an adjusted R squared of 0.643 ($P < 9.23 \times 10^{-63}$) which suggests that an important portion

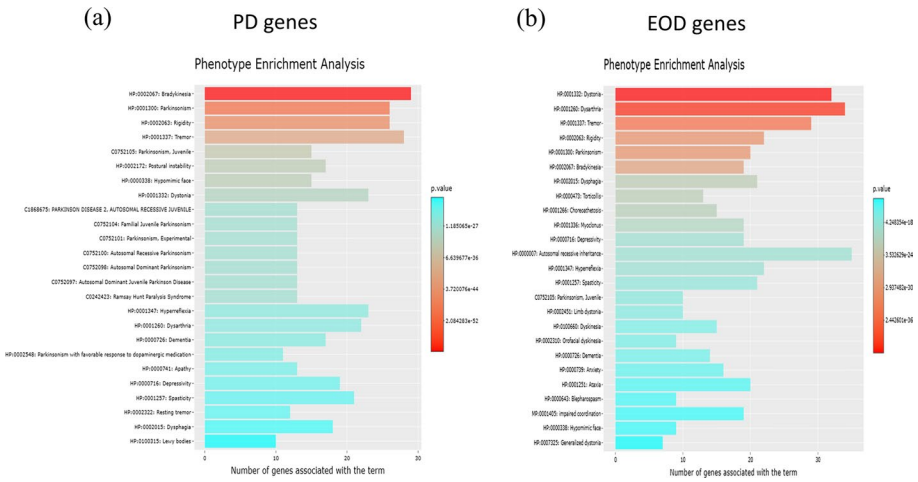

Fig. 5 Phenotype Enrichment Analysis in PhenoExam for each gene set. The graph shows the 25 most enriched terms for PD genes (**a**) and for EOD genes (**b**)

of the common phenotypes are similar in relevance. We actually see they share phenotypic terms such as Tremor (HP: 0001337), Bradykinesia (HP: 0002067), Rigidity (HP: 0002063), Dystonia (HP: 0001332), Abnormal gait (MP: 0001406) or Neuron degeneration (MP: 0003224) (Additional file 4: Table S3). But we also detect differential phenotypes that can be displayed by interactive graphs and tables on the web. For example, significant terms exclusive from the PD gene set phenotypes include Astrocytosis (MP: 0003354; $P < 5.17 \times 10^{-12}$), Substantia nigra gliosis (HP: 0011960; $P < 4.15 \times 10^{-11}$), Neuronal loss in central nervous system (HP: 0002529; $P < 3.74 \times 10^{-6}$), Orthostatic hypotension due to autonomic dysfunction (HP: 0004926; $P < 9.96 \times 10^{-6}$) and Lewy Body Disease (C0752347; $P < 1.11 \times 10^{-3}$) (Additional file 5: Table S4). Above mentioned phenotype terms are associated more or only with PD according to several articles [51–56]. The same analysis identified Writer's cramp (HP: 0002356; $P < 1.37 \times 10^{-9}$) as exclusive to EOD and this refers to a type of focal dystonia [57]. We also found Hypoplasia of the corpus callosum (HP: 0002079; $P < 3.56 \times 10^{-5}$), a controversial and not widely studied phenotype in dystonia [58, 59] and Acanthocytosis (HP: 0001927; $P < 2.76 \times 10^{-3}$) a term normally associated with chorea-acanthocytosis, other disease with dystonia's similar symptoms [60]. Microcephaly (HP: 0000252; $P < 4.17 \times 10^{-4}$) is associated with dystonia and several genes such as KMT2B [61, 62]. We also found Intellectual disability, mild (HP: 0001256; $P < 4.68 \times 10^{-3}$), Dystonia, Primary (C0752203; $P < 3.26 \times 10^{-7}$) and Hyperactive deep tendon reflexes (HP: 0006801; $P < 4.31 \times 10^{-2}$) that is associated with Paroxysmal dyskinesia (PxD) [63] (Additional file 6: Table S5).

Case 2: New likely epilepsy genes predicted by G2PML recapitulate phenotype terms of known epilepsy genes

Let us suppose it is possible to discover new Mendelian genes associated with a specific disease (congenital epilepsy in this case) by finding non-linear patterns of the genes in

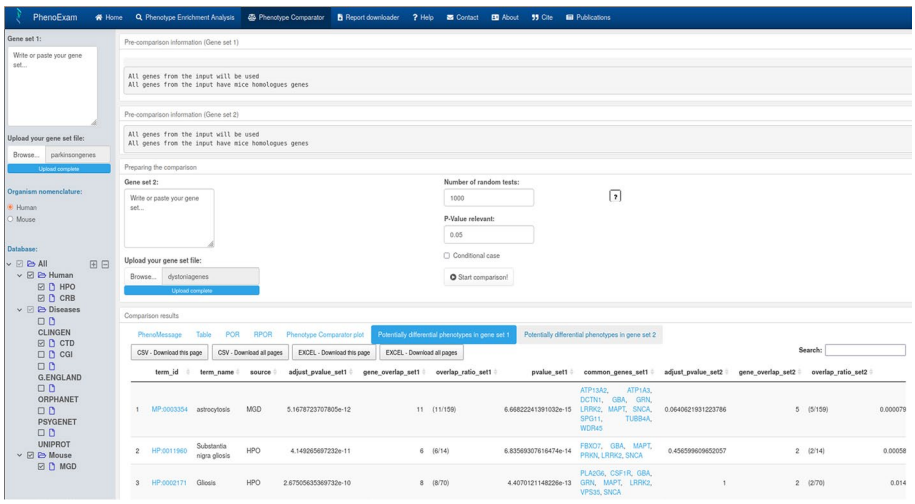

Fig. 6 Phenotype Comparator analysis view. We selected PD genes as gene set 1, EOD genes as gene set 2, HPO, MGD, CRISPRBrain and CTD databases and 1000 random tests. We obtained as output interactive tables with the shared phenotypes and the differential phenotypes, plots, PhenoExam phenotype similarities scores and information

that panel based on their description through properties based on genomic, transcriptomic and genetics of each gene with machine learning techniques. Therefore, in order to discover new genes, we aim at finding very similar genes in terms of those properties (see G2PML paper at biorxiv [64]). The question we face is: do those genes predicted to be linked to congenital genetic forms of epilepsy recapitulate similar phenotypes to the genes in the panel of origin? The more supportive the answer points to a phenotype recapitulation, the better the predictions made by G2PML. This is an example of what we call a conditional case, comparing phenotypes in gene sets G and G′ when they are disjunct and G′ was generated using G as seeds. More specifically, G refers to epilepsy genes from an in-house maintained epilepsy panel (261 genes) at NIMGenetics. Moreover, G′ is a set of 209 new genes as predicted by G2PML.

We carried out the Phenotype Comparator analysis in PhenoExamWeb with the conditional case option marked, gene set 1 was the epilepsy genes, gene set 2 was the new likely epilepsy genes predicted by G2PML, HPO, MGD, CRISPRBrain and CTD databases selected at the same time and we chose 1000 random tests. We obtained the Pheno Message from PhenoExamWeb that they shared 106 significant phenotypic terms (out of 734 unique significant phenotypic terms in both), which yields a POR of 0.144 ($P<0.001$). Phenotype relevance association analysis for epilepsy associated genes and epilepsy predicted genes (i.e., whether the shared phenotypes are similar in relevance, i.e., in the number of genes associated with them, within each gene set) results in an adjusted R squared of 0.331 ($P<4.35\times10^{-66}$) which suggests that an important portion of the common phenotypes are similar in relevance. The P values were obtained through the randomization of 1000 random gene sets. We also obtained a table with the phenotypes shared between gene sets (Additional file 7: Table S6). New likely epilepsy genes predicted by G2PML, e.g., DDX3X, KCNH1, TBL1XR1, DLG4 or PDE2A, recapitulate phenotype terms of known epilepsy genes, we check they share epilepsy significant phenotypic terms such as Seizures (HP: 0001250), Global developmental delay (HP: 0001263), Microcephaly (HP: 0000252), abnormal brain morphology (MP: 0002152), hyperactivity (MP: 0001399) and diseases terms without Bonferroni adjust Epilepsy (C0014544) and Autistic Disorder (C0004352). We also found they recapitulate interesting CRISPRBrain terms such as Association with Labile Iron (FeRhoNox Intensity) in Glutamatergic Neuron (CRB: 0000004) and Positive hit with Peroxidized Lipids (Liperfluo Intensity) in Glutamatergic Neuron (CRB: 0000008). Above mentioned phenotype terms are associated with epilepsy according to several articles [65–73]. We also provided the number of genetic variants from the Epi25 whole-exome sequencing (WES) case–control study of each epilepsy gene predicted, we obtained 665 genetic variants in cases and 446 in controls (OR = 1.49) (Additional file 8: Table S7) [74].

Conclusions

We developed PhenoExam, a freely available R package and Web application, which performs phenotype enrichment and disease enrichment analysis on gene set G, measures statistically significant phenotype similarities between pairs of gene sets G and G′ and detects statistically significant exclusive phenotypes or disease terms, across different databases. PhenoExam just required the names of genes in the gene sets as input and which databases to test for enrichment. It allows us to switch from the gene space and the phenotype

space. PhenoExam integrates phenotype data from different databases. And each database is focused on specific diseases and organisms. Therefore, choosing a database for the analyses requires of a basic knowledge of the user about the diseases used there to appropriately understand the analysis outcome. PhenoExam can identify the statistically significant and differential phenotypes of a gene set as we showed with PD, EOD, epilepsy, and likely epilepsy predicted genes. We proved with simulations that it is useful to distinguish between gene sets or diseases with very similar phenotypes through projecting genes into their annotation based phenotypical spaces. With the PD and EOD example above, we clearly see they hold phenotype-level similarities but also potentially interesting differential phenotypes. The conditional case studied between epilepsy associated and epilepsy predicted genes show they hold epilepsy phenotype terms in common, which is useful for the validation of computationally epilepsy predicted disease genes. Therefore, PhenoExam effectively discovers links between phenotypic terms across annotation databases by integrating different annotation databases. All these findings are supported with interactive plots (see tutorials at GitHub project) to foster the visualization and interpretation of findings.

Availability and requirements

Project name: PhenoExam.

Project home page: https://alejandrocisterna.shinyapps.io/phenoexamweb/

Source code is available at https://github.com/alexcis95/PhenoExam

Operating system(s): Windows, Linux, Mac OS.

Programming language: R language.

License: GPL-2|GPL-3 Any restrictions to use by non-academics: none.

Abbreviations

CGI	Cancer Genome Interpreter
ClinGen	Clinical Genome Resource
CTD	Comparative Toxicogenomics Database
EOD	Early-Onset Dystonia
GEL	Genomics England Panel App
HGNC	HUGO Gene Nomenclature Committee
HPO	Human Phenotype Ontology
MGD	Mouse Genome Database
RPOR	Relaxed Phenotypic Overlap Ratio
OMIM	Online Mendelian Inheritance in Man
PD	Parkinson's disease
POR	Phenotypic Overlap Ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-05122-x.

Additional file 1. PD and EOD genes. PD and EOD genes selected from Genomics England PanelApp.

Additional file 2: Table S1. PhenoExam enrichment analysis using PD genes. PhenoExam output using PD genes.

Additional file 3: Table S2. PhenoExam enrichment analysis using EOD genes. PhenoExam output using EOD genes.

Additional file 4: Table S3. PhenoExam comparator analyses between PD and EOD genes. Shared phenotypic and diseases terms between PD and EOD genes.

Additional file 5: Table S4. Significant terms exclusive from the PD genes. Significant phenotypic terms exclusive from the PD genes and the results obtained in EOD genes.

Additional file 6: Table S5. Significant terms exclusive from the EOD genes. Significant phenotypic terms exclusive from the EOD genes and the results obtained in PD genes.

Additional file 7: Table S6. PhenoExam comparator analyses between epilepsy and epilepsy predicted genes. Table with data from the phenotypes shared between gene sets.

Additional file 8: Table S7. Genetic variants detected from Epi25 whole-exome sequencing in epilepsy predicted genes. Data and number of genetic variants from the Epi25 whole-exome sequencing (WES) case-control study of each epilepsy gene predicted, we obtained 665 genetic variants in cases and 446 in controls.

Acknowledgements
A prototypical version of the software tool was presented at the European Bioconductor Meeting 2020 (EBM2020).

Author contributions
AC and DR wrote the code; JO and AGV tested the code. AGP and ZC tested the tool and provided insights regarding the biological interpretation of the analysis. MN and FF collected CRIPRBrain data and provided insights regarding CRISPRBrain terms in our analysis. ID, PM and SA provided an epilepsy gene panel and analysis regarding epilepsy genes. AC prepared the diagrams and figures. JAB, MR, JH and AC wrote the manuscript with the help of the rest of the authors. AGV, AGP, ZC, JH, MN, FF, ID, PM and SA helped check and improve the manuscript. JAB designed the study. All authors read and approved the final manuscript.

Authors' information
Alejandro Cisterna is a PhD candidate at Universidad de Murcia and his research interests include machine learning in bioinformatics and neurological diseases.

Aurora González-Vidal is postdoc at Universidad de Murcia and her research interests include data analysis and algorithm design and implementation.

Daniel Ruiz is a MSc. student at Universidad de Mrcia and his research interests include Web development and development of protocols for 5G Networks.

Jordi Ortiz is Assistant Professor at Universidad de Murcia and his research interests include Virtualization of Software services and cloud computing.

Alicia Gómez-Pascual is PhD candidate at Universidad de Murcia and her research interests include single-cell transcriptomics and machine learning algorithm development.

Zhongbo Chen is PhD candidate at the Institute of Child Health, University College London and her research interests include genomics of rare diseases and genetic diagnostic.

Mike Nalls is PhD at the Data Tecnica International, his research interests include genetics of neurodegeneration and machine learning algorithms for the diagnosis of Parkinson's disease.

Faraz Faghri is PhD at the Data Tecnica International, his research interests include genetics of neurodegeneration and machine learning algorithms for the diagnosis of Parkinson's disease.

John Hardy is Full Professor at the Institute of Neurology, University College London, and his research interests include the genetics of neurodegeneration.

Irene Díez is research assistant at NIMGenetics and her research interests include the genetic diagnostic of congenital diseases.

Paolo Maietta is PhD at NIMGenetics and her research interests include the genetic diagnostic of congenital diseases.

Sara Álvarez is PhD and Scientific Officer at NIMGenetics and her research interests include the genetic diagnostic of congenital diseases.

Mina Ryten is PhD, MD and Group Leader at the Ryten Lab, Institute of Child Health and her research interests include human transcriptomics and its role in brain related diseases.

Juan A. Botía is PhD, Full Professor at Universidad de Murcia and his research interests include Artificial Intelligence and AI based algorithm development for the study of neurodegeneration.

Funding
This work was supported by the Science and Technology Agency, Séneca Foundation, Comunidad Autónoma Región de Murcia, Spain. AC was supported by the Science and Technology Agency, Séneca Foundation, Comunidad Autónoma Región de Murcia, Spain through the Grant [20762/FPI/18]. JB was supported by the same foundation through the research project [00007/COVI/20]. AG-M was funded by the same foundation through [21230/PD/19]. JH and MR were supported by the United Kingdom Medical Research Council (MRC), with JH supported by a Grant [MR/N026004/] and MR through the award of a Tenure Track Clinician Scientist Fellowship [MR/N008324/1]. JH was also supported by the United Kingdom Dementia Research Institute, The Wellcome Trust [202903/Z/16/Z], the Dolby Family Fund, the BRCNIHR Biomedical Research Centre, and the NIHR. None of these funding agencies played any role in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials
Project home page: https://alejandrocisterna.shinyapps.io/phenoexamweb/
Source code is available at https://github.com/alexcis95/PhenoExam
Documentation: https://raw.githack.com/alexcis95/PhenoExamWebTutorials/main/tutorial.html
Data:
HGNC: https://www.genenames.org/download/statistics-and-files/ (protein-coding gene).
HPO: https://hpo.jax.org/app/data/annotations (phenotypes_to_genes.txt) https://archive.monarchinitiative.org/latest/tsv/gene_associations/
MGI: http://www.informatics.jax.org/downloads/reports/index.html#go (MGI_PhenoGenoMP.rpt, HMD_HumanPhenotype.rpt, VOC_MammalianPhenotype.rpt).
CRISPRBrain: https://crisprbrain.org/simple-screen/ (Exporting each screen name).
UniProt: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/humsavar.txt
CTD: http://ctdbase.org/downloads/#gd (CTD_genes_diseases.csv.gz).
Orphanet: https://github.com/Orphanet/Orphadata_aggregated/tree/master/Genes%20associated%20with%20rare%20diseases
ClinGen: https://search.clinicalgenome.org/kb/downloads (Gene-Disease).
The Genomics England PanelApp: https://panelapp.genomicsengland.co.uk/panels/ (Each panel).
CGI: https://www.cancergenomeinterpreter.org/2018/data/catalog_of_validated_oncogenic_mutations_latest.zip?ts=20180216
PsyGeNET: http://www.psygenet.org/ds/PsyGeNET/results/all_GeneDiseaseAssociations.tar.gz

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interest
The authors declare that they have no competing interests.

Received: 2 May 2022 Accepted: 22 December 2022
Published online: 31 December 2022

References

1. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature. 2001;409(6822):853–5.
2. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.
3. Schaub MA, Boyle AP, Kundaje A, et al. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22(9):1748–59.
4. Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet. 2004;13(Suppl_1):R57–64.
5. Osborne JD, Flatow J, Holko M, et al. Annotating the human genome with Disease Ontology. BMC Genomics. 2009;10(S1):S6.
6. Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.
7. Dorland WAN. Dorland's illustrated medical dictionary. 32nd ed. Philadelphia: Elsevier/Saunders; 2012.
8. Temple LK, McLeod RS, Gallinger S, et al. Defining disease in the genomics era. Science. 2001;293(5531):807–8.
9. Scully JL. What is a disease? EMBO Rep. 2004;5(7):650–3. https://doi.org/10.1038/sj.embor.7400195.
10. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010;86(4):560–72.
11. Hunter DJ. Gene–environment interactions in human diseases. Nat Rev Genet. 2005;6(4):287–98.
12. Robinson PN, Köhler S, Bauer S, et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83(5):610–5.
13. Bult CJ, Blake JA, Smith CL, et al. The Mouse Genome Database Group, 2019. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2019;47(D1):D801–6.
14. Amberger JS, Bocchini CA, Schiettecatte F, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789–98. https://doi.org/10.1093/nar/gku1205.
15. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15. https://doi.org/10.1093/nar/gky1049.
16. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res. 2021;49(D1):D1138–43. https://doi.org/10.1093/nar/gkaa891.
17. Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. http://www.orpha.net Accessed (date of access).
18. Rehm HL, Berg JS, Brooks LD, et al. ClinGen—The Clinical Genome Resource. N Engl J Med. 2015;372:2235–42. https://doi.org/10.1056/NEJMsr1406261.

19. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–5. https://doi.org/10.1038/s41588-019-0528-2.
20. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25. https://doi.org/10.1186/s13073-018-0531-8.
21. Gutiérrez-Sacristán A, Grosdidier S, Valverde O, et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. Bioinformatics. 2015. https://doi.org/10.1093/bioinformatics/btv301.
22. Tian R, Abarientos A, Hong J, et al. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nat Neurosci. 2021. https://doi.org/10.1038/s41593-021-00862-0.
23. Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Science Translational Medicine. 2014;6(252):252ra123.
24. Kalaria R. Similarities between Alzheimer's disease and vascular dementia. J Neurol Sci. 2002;203–204:29–34.
25. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236–41.
26. Lohi H, Turnbull J, Zhao XC, et al. Genetic diagnosis in Lafora disease Genotype–phenotype correlations and diagnostic pitfalls. Neurology. 2007;68(13):996–1001. https://doi.org/10.1212/01.wnl.0000258561.02248.2f.
27. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457–64.
28. Deng Y, Gao L, Wang B, et al. Hposim: an r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS ONE. 2015;10(2):0115692.
29. Peng J, Xue H, Hui W, et al. An online tool for measuring and visualizing phenotype similarities using HPO. BMC Genomics. 2018;19(6):89–97.
30. Weng MP, Liao BY. modPhEA: model organism Phenotype Enrichment Analysis on eukaryotic gene sets. Bioinformatics. 2017;33(21):3505–7.
31. Wang J, Huang Q, Liu ZP, et al. NOA: a novel Network Ontology Analysis method. Nucleic Acids Res. 2011;39(13):e87–e87.
32. Page L, Motwani R, Brin S, et al. The pagerank citation ranking: bringing order to the web. Stanford Digital Libraries Working Paper, 1999. 2009; 9(1):1–14.
33. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–55. https://doi.org/10.1093/nar/gkz1021.
34. Braschi B, Denny P, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019;47(D1):D786–92.
35. Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze. 1936;8:3–62.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. https://www.jstor.org/stable/2346101.
37. Alroy J. A new twist on a very old binary similarity coefficient. Ecology. 2015;96:575–86.
38. Salvatore S, et al. Beware the Jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis. Brief Bioinform. 2020;21:1523–30.
39. Winston Chang, Joe Cheng, JJ Allaire, et al. (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny
40. Niemann N, Jankovic J. Juvenile Parkinsonism: differential diagnosis, genetics, and treatment. Parkinsonism Relat Disord. 2019;67:74–89.
41. Breakefield XO, Blood AJ, Li Y, et al. The pathophysiological basis of dystonias. Nat Rev Neurosci. 2008;9(3):222–34. https://doi.org/10.1038/nrn2337.
42. Berardelli A, Rothwell JC, Thompson PD, et al. Pathophysiology of bradykinesia in Parkinson's disease. Brain. 2001;124(11):2131–46. https://doi.org/10.1093/brain/124.11.2131.
43. Chen PH, Wang RL, Liou DJ, et al. Gait disorders in Parkinson's disease: assessment and management. Int J Gerontol. 2013;7(4):189–93.
44. Hunot S, Hirsch EC. Neuroinflammatory processes in Parkinson's disease. Ann Neurol. 2003;53(S3):S49–60.
45. O'Keeffe GW, Sullivan AM. Evidence for dopaminergic axonal degeneration as an early pathological process in Parkinson's disease. Parkinsonism Relat Disord. 2018;56:9–15.
46. Albanese A, Di Giovanni M, Lalli S. Dystonia: diagnosis and management. Eur J Neurol. 2019;26(1):5–17.
47. Brashear A, Farlow MR, Butler IJ, et al. Variable phenotype of rapid-onset dystonia-parkinsonism. Mov Disord. 1996;11(2):151–6.
48. Romano R, Bertolino A, Gigante A, et al. Impaired cognitive functions in adult-onset primary cranial cervical dystonia. Parkinsonism Relat Disord. 2014;20(2):162–5.
49. Furuya S, Tominaga K, Miyazaki F, et al. Losing dexterity: patterns of impaired coordination of finger movements in musician's dystonia. Sci Rep. 2015;5(1):1–14.
50. Castagna A, Frittoli S, Ferrarin M, et al. Quantitative gait analysis in parkin disease: possible role of dystonia. Mov Disord. 2016;31(11):1720–8.
51. Booth H, Hirst WD, Wade-Martins R. The role of astrocyte dysfunction in Parkinson's disease pathogenesis. Trends Neurosci. 2017;40(6):358–70. https://doi.org/10.1016/j.tins.2017.04.001.
52. Kim CY, Wirth T, Hubsch C, et al. Early-onset parkinsonism is a manifestation of the PPP2R5D p. E200K mutation. Ann Neurol. 2020;88(5):1028–33.
53. Van Muiswinkel FL, De Vos RAI, Bol JGJM, et al. Expression of NAD (P) H: quinone oxidoreductase in the normal and Parkinsonian substantia nigra. Neurobiol Aging. 2004;25(9):1253–62.
54. Zarow C, Lyness SA, Mortimer JA, et al. Neuronal loss is greater in the locus coeruleus than nucleus basalis and substantia nigra in Alzheimer and Parkinson diseases. Arch Neurol. 2003;60(3):337–41.
55. Ziemssen T, Reichmann H. Cardiovascular autonomic dysfunction in Parkinson's disease. J Neurol Sci. 2010;289(1–2):74–80.

56. Aarsland D, Kurz MW. The epidemiology of dementia associated with Parkinson disease. J Neurol Sci. 2010;289(1–2):18–22.
57. Rosenkranz K, Williamon A, Butler K, et al. Pathophysiological differences between musician's dystonia and writer's cramp. Brain. 2005;128(4):918–31.
58. Ibrahim MH, Fadhil A, Ali SS, et al. Could dystonia be initial presentation of corpus callosum infarction in young age patients? A case report study. Neurosci Med. 2015;6(02):62.
59. Colosimo C, Pantano P, Calistri V, et al. Diffusion tensor imaging in primary cervical dystoniaJournal of Neurology. Neurosurg Psychiatry. 2005;76:1591–3.
60. Schneider SA, Lang AE, Moro E, et al. Characteristic head drops and axial extension in advanced chorea-acanthocytosis. Mov Disord. 2010;25(10):1487–91.
61. Gorman KM, Meyer E, Kurian MA. Review of the phenotype of early-onset generalised progressive dystonia due to mutations in KMT2B. Eur J Paediatr Neurol. 2018;22(2):245–56.
62. Lohmann K, Klein C. Update on the genetics of dystonia. Curr Neurol Neurosci Rep. 2017;17(3):26.
63. Groffen AJ, Klapwijk T, van Rootselaar AF, et al. Genetic and phenotypic heterogeneity in sporadic and familial forms of paroxysmal dyskinesia. J Neurol. 2013;260(1):93–9. https://doi.org/10.1007/s00415-012-6592-5.
64. Botía, J. A., Guelfi, S., Zhang, D., et al. (2018). G2P: Using machine learning to understand and predict genes causing rare neurological disorders. bioRxiv, 288845.
65. Stafstrom CE, Carmant L. Seizures and epilepsy: an overview for neuroscientists. Cold Spring Harb Perspect Med. 2015;5(6): a022426.
66. Ishiura H, Doi K, Mitsui J, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat Genet. 2018;50(4):581–90.
67. Trinka E, Höfler J, Zerbs A. Causes of status epilepticus. Epilepsia. 2012;53:127–38.
68. Abdel-Salam GM, Halász AA, Czeizel AE. Association of epilepsy with different groups of microcephaly. Dev Med Child Neurol. 2000;42(11):760–7.
69. Carvalho MDC, Ximenes RA, Montarroyos UR, et al. Early epilepsy in children with Zika-related microcephaly in a cohort in Recife, Brazil: Characteristics, electroencephalographic findings, and treatment response. Epilepsia. 2020;61(3):509–18.
70. Ricobaraza A, Mora-Jimenez L, Puerta E, et al. Epilepsy and neuropsychiatric comorbidities in mice carrying a recurrent Dravet syndrome SCN1A missense mutation. Sci Rep. 2019;9:14172. https://doi.org/10.1038/s41598-019-50627-w.
71. Parisi P, Moavero R, Verrotti A, et al. Attention deficit hyperactivity disorder in children with epilepsy. Brain Develop. 2010;32(1):10–6.
72. Lee BH, Smith T, Paciorkowski AR. Autism spectrum disorder and epilepsy: disorders with a shared biology. Epilepsy Behav. 2015;47:191–201.
73. Tuchman R, Rapin I. Epilepsy in autism. Lancet Neurol. 2002;1(6):352–8.
74. Epi25 Collaborative. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. Am J Hum Genet. 2019;105(2):267–282. https://doi.org/10.1016/j.ajhg.2019.05.020

Publisher's Note