Havrilla et al. Genome Medicine (2021) 13:91
https://doi.org/10.1186/s13073-021-00909-8

Genome Medicine

DATABASE Open Access

PhenCards: a data resource linking human phenotype information to biomedical knowledge

James M. Havrilla[1], Cong Liu[2], Xiangchen Dong[1], Chunhua Weng[2] and Kai Wang[1,3*]

Abstract

We present PhenCards (https://phencards.org), a database and web server intended as a one-stop shop for previously disconnected biomedical knowledge related to human clinical phenotypes. Users can query human phenotype terms or clinical notes. PhenCards obtains relevant disease/phenotype prevalence and co-occurrence, drug, procedural, pathway, literature, grant, and collaborator data. PhenCards recommends the most probable genetic diseases and candidate genes based on phenotype terms from clinical notes. PhenCards facilitates exploration of phenotype, e.g., which drugs cause or are prescribed for patient symptoms, which genes likely cause specific symptoms, and which comorbidities co-occur with phenotypes.

Keywords: Disease, Phenotype, Genetics, Natural Language Processing, Mendelian diseases, Rare disease, Common disease, Drug targets, Collaborative support

Background

Phenotypes describe any observable traits: common traits, such as hair color, or rarer traits like craniosynostosis (a birth defect in which the bones in a baby's skull join together too early). To understand more deeply how genetic mutations can discriminate between subclasses of phenotypes, researchers have made efforts to use phenotypic trait information to understand frequencies of phenotype occurrence in conjunction with disease and genetic mutation data [1–4]. In addition, new links between genetic pathways, expression data, and phenotype [5–7] are slowly pushing the field in a direction where researchers and clinicians can provide insights into how genetic mutations influence different pathways in the body, and how these effects are manifested in the phenotype. In recent years, several studies have shown the utility of using clinical phenotype information to facilitate gene identification from clinical genome and exome sequencing data [8–12].

Phenotype vocabularies are superb tools for facilitating the investigation and classification of genetic diseases. Copious web servers, databases, and other resources for phenotypic terms and diseases exist: HPO (Human Phenotype Ontology) [4, 13], MeSH (Medical Subject Headings) [14], OHDSI (Observational Health Data Sciences and Informatics) [15], ICD-10 (International Classification of Diseases, version 10) [16], UMLS (Unified Medical Language System) [17, 18], Disease Ontology [19, 20], OMIM (Online Mendelian Inheritance in Man) [3, 21], DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resource) [22], and Orphanet [2]. However, few of them link related biomedical information to disease and phenotypic terms, and there are numerous gaps in information. Pathway data is similarly limited to specific disease terminology [5–7]. Most phenotypic resources such as HPO or Malacards [23] are limited in their queries, as

* Correspondence: wangk@email.chop.edu
[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

their data can only access disease-specific or ontology-specific aspects of phenotype information, without comprehensive query functionality that allows one to link all phenotype terminologies to vital biomedical information including drug, gene, pathway, disease, grant, physician, literature, and clinical trial data. There are portions of this data contained in each resource, but none have access to all of it. Ideally there should be one place where clinicians and researchers can go to and find a comprehensive collection of information about a clinical phenotype. Without this singular data resource, what clinicians, researchers and genetic counselors can ascertain about clinical phenotypes is limited to their personal awareness of individual tools and databases.

To fill this void, we have created PhenCards (https://phencards.org), a data resource and a search engine for linking relevant biomedical knowledge to human clinical phenotype terms. PhenCards is among the first resources of its kind, striving to link all possible biomedical knowledge to human phenotypes. We incorporate all of the aforecited databases, drug, procedural study, surgery, pathway, funding opportunity, literature, and hospital-wide patient term co-occurrence data, as well as relevant gene information for extracted HPO terms. Providing all of this data in a single web server will allow clinicians, researchers, and genetic counselors to add new layers of information to previously limited genetic studies and make more informed clinical decisions.

Construction and content

Implementation and resources

The entire site uses HTML, Jinja2, CSS, Javascript, jQuery, and Bootstrap 4 to create the look, implement the autocompletion, and display the data on the frontend. The backend uses docker-compose [24], Elasticsearch [25], Python 3.8, Flask [26], the Apache HTTP Server [27], and Certbot with Let's Encrypt [28]. Several python packages are utilized including Elasticsearch, Ray [29], and Beautiful Soup. The site has been tested and shown to work on Windows/macOS/Linux Chrome, Opera, Edge, and Safari, iOS Safari/Chrome, and Android Chrome. The index page is easy and straightforward to use with explanations on input. Highly detailed frontend, backend, secure query, and literature search implementations, as well as how to use all site resources, are described in Additional file 1 and Additional file 2. We also possess numerous resources, some of which are real-time and some of which require regular updates (Table 1). Further information about the licenses and versions are shown in Additional file 1.

Table 1 Resources in PhenCards. Some resources require regular updates and downloads to stay abreast of changes, but the majority of resources are API or web-based and require no changes to stay up-to-date

Resource(s)	Method of access	Update needed?	Content
HPO [4, 13] (includes OMIM [3, 21] and Orphanet [2]), Disease Ontology [19, 20]	Elasticsearch [25] on indexed database	Yes, monthly	Standardized phenotype and disease terms
ICD-10 [16], UMLS [17, 18], OHDSI ATHENA [15], MeSH [14]	Elasticsearch on indexed database	Yes, yearly	Standardized phenotype and disease terms
Pharos (disease) [30]	API	No	Disease aliases, expression, drug, pathway, Gene Ontology data
IRS (Internal Revenue Service), Open990	Elasticsearch on indexed database	Yes, yearly	Nonprofit grants and foundations
NIH (National Institute of Health) Federal Reporter, NIH FOAs (funding opportunity announcements)	API	No	Federal grant and projects
Direct2Experts [31]	API	No	Collaborators, specialty physicians
openFDA [32], Tocris, APExBio, Pharos (target) [30], DrugCentral [33]	API	No	Federal and company drug, drug target and adverse effect data
Pathway Commons [34] and KEGG [5, 34]	API	No	Pathways: diseases, biological functions
ClinicalTrials.gov [35]	API	No	Clinical trials: studies, procedures, drugs
Columbia Open Health Data [36]	API	No	Co-occurring patient drug, procedure, and condition terms
Doc2Hpo [37]	API	No	NLP algorithm for optimally extracting terms from text
Phen2Gene [38]	API	No	Algorithm ranking candidate genes for a set of HPO terms
PubMed [39]	API	No	Biomedical literature
Google Scholar	API	No	Large-scale scholarly search engine

Basic workflow

After receiving 3 characters of user phenotype term input, the site begins to autocomplete known phenotype terms for the user (though choosing one is not required). After submitting a query, PhenCards makes multitudinous resources available to the user. We first and foremost provide aliases for matching phenotype and disease terms. There are links to each site where applicable for further investigation, and hoverable tooltips that explain all data tables on the site. Pharos [30] information for UniProt [40] disease aliases is expandable in PhenCards to obtain Gene Ontology [41] data, expression data, pathway data, and drug development stage data for the term. For further facilitating genetic studies, PhenCards utilizes Phen2Gene [38] and Pharos again to obtain relevant gene information for extracted HPO terms for user queries, and these genes link out to MedlinePlus [42] and the Pharos site. The Pharos drug target data can be further explored in PhenCards to learn ligands for the target, protein-protein interactions, its expression in certain tissues, and its novelty. We search APExBio, Tocris, DrugCentral [33], and openFDA's FAERS (FDA Adverse Event Reporting System) database [32] as well as drug indication and adverse effect data, study data, and surgery data from ClinicalTrials.gov [35] for potential treatments for the phenotype, as well as drugs that may have led to the phenotype. PhenCards integrates COHD (Columbia Open Health Data) [36] to find which drugs, conditions, and procedures significantly co-occur with phenotype terms in patients. We have consolidated pathway data for phenotype terms from several places [5, 6, 34]. PhenCards obtains active funding opportunity announcement data from the NIH, and nonprofit foundation and grant data from Open990 and the IRS, and collaborator and physician data from Direct2Experts [31]. Lastly, we have also included a complex literature search for terms using PubMed Entrez Search and Fetch [39] (Fig. 1a). Further details of the site's implementation and the API queries can be found in Additional file 1.

For users with clinical notes that may describe a de-identified patient's phenotype using free text, we are able to extract HPO terms using Doc2HPO [37] and link them to ClinicalTrials.gov data and create a custom Google Scholar query. By clicking on the HPO terms, one can investigate more information about the terms on the HPO website or use PhenCards to search for the same information in Fig. 1a. Using all of the HPO terms, we also predict a potential disease phenotype and create a candidate gene list using Phen2Gene all of which can be further investigated through links to the sites of their respective databases (Fig. 1b).

Disease prediction algorithm

We use HPO terms extracted from clinical notes by the Aho-Corasick algorithm [44] in Doc2HPO in order to find the most likely disease phenotype for a potential patient. Using an exact match in Elasticsearch for linked HPO terms to disease databases such as OMIM and Orphanet, we add the Elasticsearch scores together for each disease linked to each exact HPO term match and the disease with the highest aggregate Elasticsearch score is ranked first. The score itself is arbitrary, the scale only exists to rank the matches against one another. Elasticsearch score is calculated using an algorithm called BM25, which is similar to tf-idf (term frequency–inverse document frequency), except that it accounts for document length (greater details available in Additional file 1).

Pathway query

First, we query the KEGG Pathway database for diseases related to the phenotype term using the KEGG FIND API query. This searches names but also disease description text. Then using the official KEGG disease names we can find relevant linked pathways using KEGG's LINK API query. Pathway Commons already does all this by default for Reactome and other databases, but not for KEGG.

Phen2Gene query

Using the "Phenotype search" query, which is the main function of the site, leads users to the results page. There, the number one HPO term result from the phenotype term query is used to search for the top 1000 candidate genes in Phen2Gene [38]. If using the "Clinical notes" query function of the site, Doc2HPO [37] extracts the non-negated HPO terms and PhenCards uses all of them to query Phen2Gene for the top 1000 candidate genes.

COHD concept co-occurrence ranking

Patient concepts (conditions, drugs, and procedures) from COHD are derived from the Columbia University Irving Medical Center's Observational Health Data Sciences and Informatics (OHDSI) database and are in structured OMOP (Observational Medical Outcomes Partnership) format. The p-values for each respective concept pairing's co-occurrence are based on chi-square tests of their frequencies in the COHD database (further details of calculation in Additional file 1).

Utility and discussion

Performing a patient-centered query

PhenCards provides a great default example of what a user can investigate using clinical notes in free text format drafted by clinicians or researchers. In this example, these clinical notes come courtesy of Griffin et al., from the section "Clinical Presentation and Family History" [43]. Using these very thorough de-identified clinical

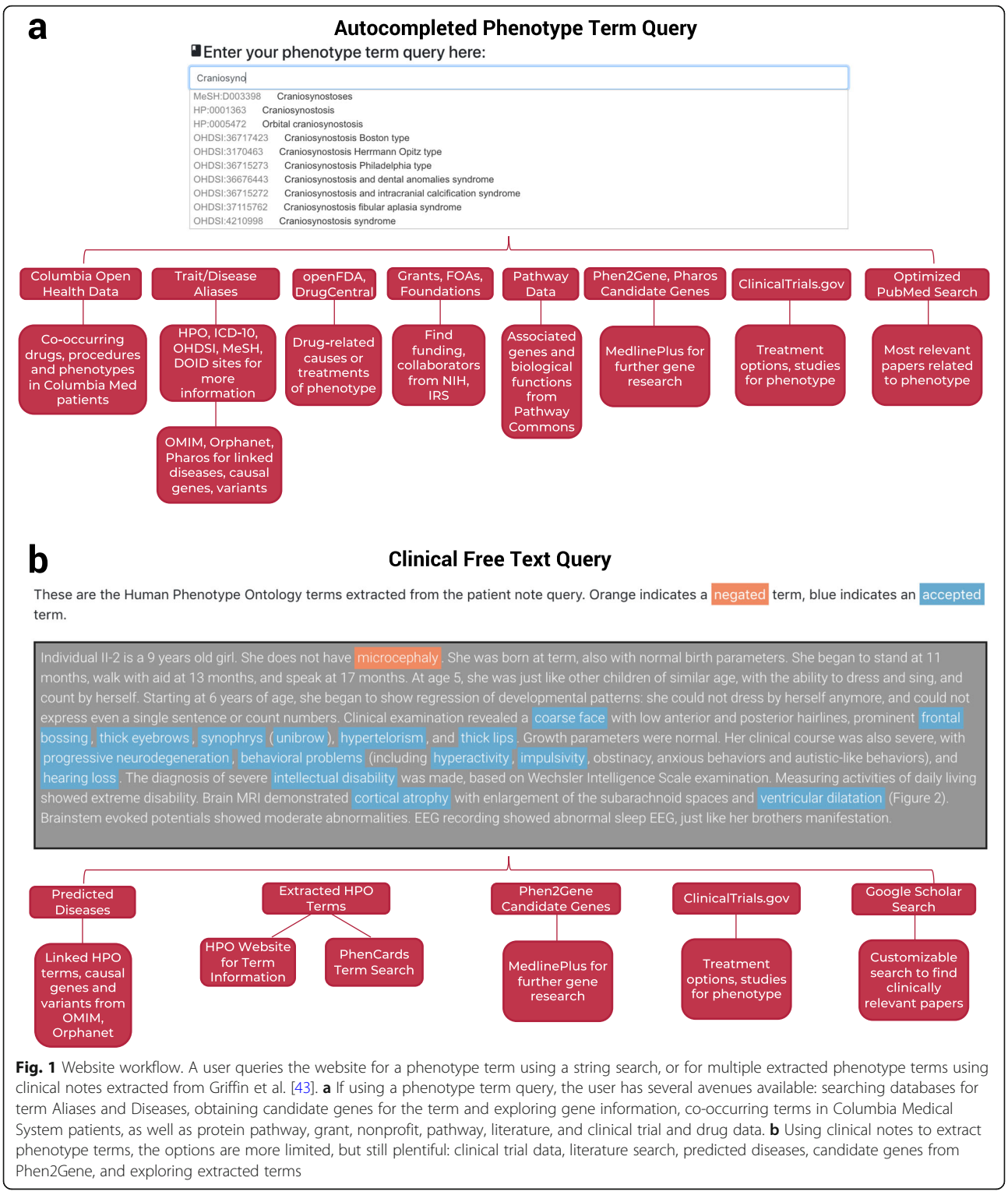

Fig. 1 Website workflow. A user queries the website for a phenotype term using a string search, or for multiple extracted phenotype terms using clinical notes extracted from Griffin et al. [43]. **a** If using a phenotype term query, the user has several avenues available: searching databases for term Aliases and Diseases, obtaining candidate genes for the term and exploring gene information, co-occurring terms in Columbia Medical System patients, as well as protein pathway, grant, nonprofit, pathway, literature, and clinical trial and drug data. **b** Using clinical notes to extract phenotype terms, the options are more limited, but still plentiful: clinical trial data, literature search, predicted diseases, candidate genes from Phen2Gene, and exploring extracted terms

notes, we extract several useful HPO terms, which rank the actual disease diagnosis of Aarskog-Scott syndrome as the first result in our disease prediction algorithm (Fig. 2). As the disease name is identical in both OMIM and Orphanet, by clicking the disease name link out to HPO for external navigation, we can see the causal gene identified in Griffin et al., FGD1, is listed as the causal gene for Aarskog-Scott syndrome on both sites. Additionally, FGD1 is ranked 3rd by Phen2Gene. If a user had these notes for this patient as well as variant data

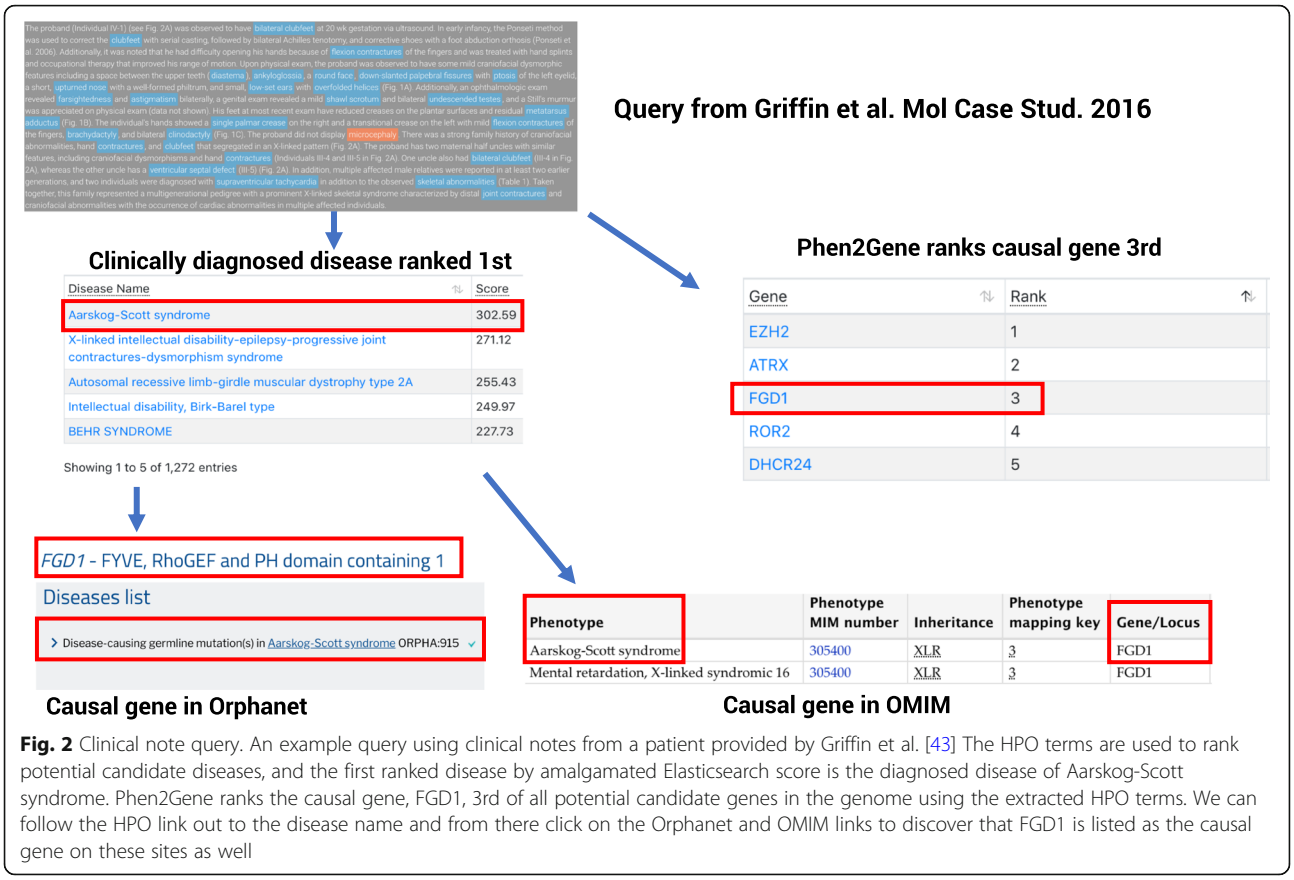

Fig. 2 Clinical note query. An example query using clinical notes from a patient provided by Griffin et al. [43] The HPO terms are used to rank potential candidate diseases, and the first ranked disease by amalgamated Elasticsearch score is the diagnosed disease of Aarskog-Scott syndrome. Phen2Gene ranks the causal gene, FGD1, 3rd of all potential candidate genes in the genome using the extracted HPO terms. We can follow the HPO link out to the disease name and from there click on the Orphanet and OMIM links to discover that FGD1 is listed as the causal gene on these sites as well

for the patient, this could rank the gene even higher by only accepting the genes with overlapping variant data as potential candidates. The algorithms are fairly robust for a large number of HPO terms, as 5 terms were arbitrarily removed in 5 different combinations for this example and we still returned Aarskog-Scott syndrome and FGD1 at the same or higher ranks.

FGD1 provides a potential drug target for the patient. If FGD1 has been used previously and proven unsuccessful, the ClinicalTrials.gov search may provide other treatment alternatives, and the customizable literature search in Google Scholar is preloaded with the HPO terms and can be used to search for relevant literature. In fact, we used it to search for several test case studies with clinical free text. The example in Fig. 2 only requires 3 of the terms to be used in our Google Scholar search to obtain the Griffin et al. paper.

Using phenotype search to investigate a rare symptom

The phenotype term search provides access to a myriad of data sources, but its primary use is to aid researchers and genetic counselors in further investigating a patient's phenotype when they do not have access to clinical notes or curated terms. Using a rare symptom, "craniosynostosis" which only occurs in as few as 0.04% of births [45], we can show how a researcher might narrow down useful information (Fig. 3a). Using the openFDA and DrugCentral FAERS data, we can see that SSRI antidepressants like paroxetine are a common cause of infant craniosynostosis as a birth defect (and from drinking breast milk with the drug) [46]; this birth defect has also been associated with fluoxetine [47] and sertraline [48]. The most common genetic cause of craniosynostosis is mutations in FGFR2, which is the first ranked gene in both the Phen2Gene candidate gene and Pharos drug target results. Beare-Stevenson cutis gyrata syndrome is commonly associated with this symptom and mutations in FGFR2 [49] on OMIM and Orphanet, as is "Craniosynostosis, nonspecific" from OMIM, both of which appear in the Diseases results for this term. Pathway Commons further supports this with the top Reactome hit "Activated point mutants of FGFR2," where the first mentioned disease entry is Beare-Stevenson cutis gyrata syndrome; this pathway can also be used for alternative potential drug targets. A basic literature search of fluoxetine shows the drug affects one of the FGFR2 sub-pathways, explaining a likely cause of this potential genetic mutation [50].

Relevant surgical procedures and investigational study data can also be found for this condition thanks to COHD and ClinicalTrials.gov data collected by PhenCards. Finally, PhenCards provides several non-profit

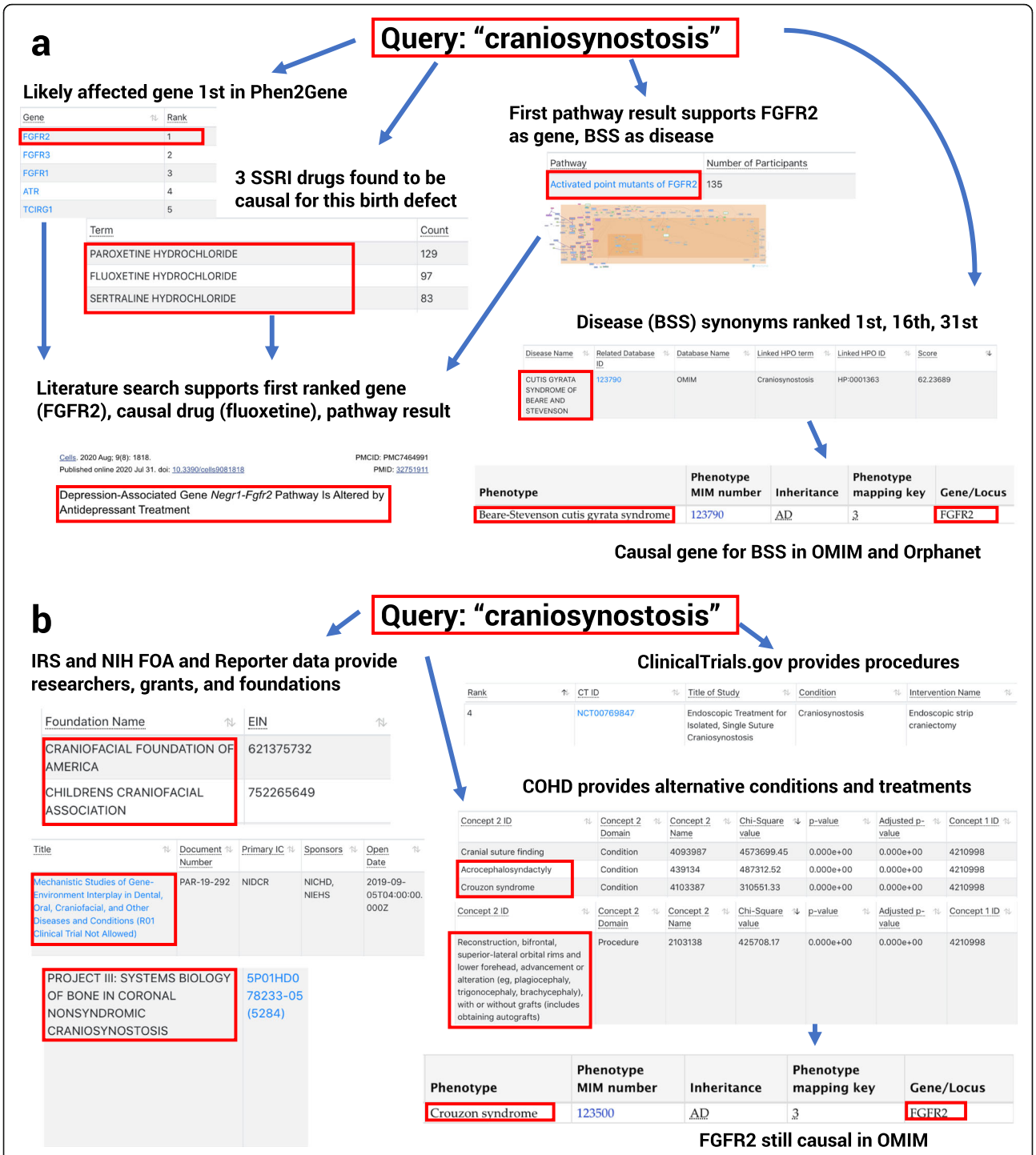

Fig. 3 Phenotype term query. An example query using "craniosynostosis" as the searched term. **a** A researcher may wish to know what genes are likely causal for a disease or phenotypic trait. FGFR2 is shown to be the most likely causal gene on OMIM, Orphanet, and Phen2Gene for craniosynostosis, particularly for "Beare-Stevenson Syndrome" (BSS). Pathway information from Reactome for "Activated point mutants of FGFR2" demonstrates further evidence that FGFR2 is the most likely causal gene and provides several alternative drug target candidates. Literature and openFDA data also support fluoxetine, which has significant effects on FGFR2 and its pathways, as a potential cause of the condition. **b** Alternative ways to look at this symptom include finding alternative co-occurring conditions in COHD, in addition to finding past treatments for patients in the Columbia University Irving Medical Center in conjunction with clinical trial data. FGFR2 is still shown to be a causal gene for many alternative syndromes with this condition. Finally, a user can find current funded research, its principal investigators (PIs), new sources of NIH funding, and non-profit foundations that support research and treatment for the condition

foundations where one could seek financial help for treating or researching this ailment, including the Craniofacial Foundation of America. There are several relevant active FOAs displayed from the NIH, and funded grants and their PIs' information from the NIH Federal Reporter service (Fig. 3b). In addition, we provide access to collaborators and physicians via Direct2Experts. Our goal is to arm researchers, genetic counselors, and clinicians with enough knowledge to assist their patients in relieving their ailments, whatever path they choose.

Discussion and future directions

The PhenCards web server can facilitate biomedical researchers in formulating new hypotheses on specific clinical phenotypes, locating potential funding and financial support, and help clinicians in analyzing clinical texts and derive possible diagnoses. If a clinician has trouble discerning their patient's phenotype from their notes, our phenotype prediction algorithm can assist in this feat, the extracted terms can be used to rank candidate causal genes, and provide potential sources of treatment by linking clinical trials data. If the disease is too new to have this information in existing databases, we provide the means to supplement this with a literature search involving the extracted terms. For a more specific individual condition, the phenotype term bestows the researcher with a wellspring of knowledge, such as potentially related diseases, prescribed drugs or drugs that caused it, clinical trials specific to the phenotype, pathways involving the condition, causal genes for the condition, and procedures and conditions co-occurring with the phenotype in other patient data. Finally, if the varied information fails to supply the researcher with new hypotheses, it can put them in touch with other researchers, physicians, grants, and foundations that may lend support in the investigation or treatment of the condition.

The current version of the PhenCards server has some limitations. Currently, only the English language is supported, though moving to other languages in the future for some resources is not difficult: UMLS and OHDSI, for example, support several other languages, and Elasticsearch can be configured to allow for multilingual query decomposition. Language translations are already underway by the HPO team who has gathered volunteer groups worldwide to translate and parse the HPO terms. We plan to support the multi-language HPO upgrade, as well as incorporate terminologies that already include more than one language. To further linguistic inclusivity, we are also making efforts to parse clinical notes from other languages using NLP algorithms specialized to those languages and map them to these multi-language terminology sets. In the near future, we aspire to continue adding resources, more drug databases, pathway databases, incorporate complex searching into the patient page, and link more disease information to the search terms. Furthermore, our data on physicians, foundations, grants, and research data is also currently limited to the USA only, but we would appreciate incorporating this data from other countries. For patient co-occurrence data, COHD only reflects the population demographics in one medical institution. PhenCards would benefit greatly from NLP-based question and answer services like "What diseases have seizures?" or "What drugs are linked to conditions with palmar creases?" Lastly, we would relish the idea of incorporating pathology, physical or facial images related to phenotype queries; there are some site scraping and Twitter-based AI bot builders we have begun collaborating with to accomplish this.

Conclusions

While databases such as HPO have been extremely useful for researchers in genomic medicine, PhenCards (https://phencards.org) adds a new layer of investigational ability for phenotype terminology that did not exist previously, by integrating multiple sources of biomedical knowledge and linking them to presentations of phenotype. When clinical researchers discover a new undiagnosed case, characterizing it by phenotypic traits is always the first step. PhenCards can not only aid in finding similar diseases, but link all relevant information to give researchers the best possible chance of identifying it by comparing it to similar diseases or even supplying novel candidate genes. We sincerely hope that with researcher and community involvement we can add even more useful knowledge to our web server. Our goal is to provide both a one-stop shop and a lasting, continuously updated resource that will allow for novel insight into research of human phenotypes to further our understanding of human health and both rare and common diseases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00909-8.

Additional file 1. Contains Supplementary Methods and Results.

Additional file 2: Supplementary Video on how to use the website. A tutorial that is also available on YouTube with captions for those hard of hearing at https://www.youtube.com/watch?v=9FT4pFgeA08.

Acknowledgements

We thank the developers of HPO, OMIM, OHDSI, UMLS, Pharos, NIH RePORTER, MeSH, Orphanet, Direct2Experts, Disease Ontology, ICD, Pathway Commons, Columbia Open Health Data, PubMed, the IRS, openFDA, ClinicalTrials.gov, and DrugCentral for continuous development of their respective databases over the past several years, which have greatly facilitated and standardized clinical diagnosis of affected individuals with suspected genetic disorders, discovering drug targets for disease, and

furthered the development of our site. We would also like to acknowledge Jacqueline Peng for her contributions to the initial template of the site.

Authors' contributions

JH and KW conceived and designed the project. XD provided help with the initial site design. CL helped JH integrate Doc2HPO and Elasticsearch. JH wrote the draft manuscript and created all the figures and case studies. JH designed the site and algorithms and gathered resources. All authors read and approved the final manuscript.

Funding

This study is supported by NIH/NLM/NHGRI grant LM012895 (JH, CL, CW, KW), NIH/NIGMS grant GM132713 (JH, KW), and the CHOP Research Institute (XD, KW).

Availability of data and materials

All data can be accessed through https://phencards.org and its API. The code for creating the site is on https://github.com/WGLab/PhenCards [51] and the data used for constructing the Lucene index with Elasticsearch is on Zenodo at: https://zenodo.org/record/4755959 [52].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [2]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA. [3]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

Received: 3 February 2021 Accepted: 13 May 2021
Published online: 25 May 2021

References

1. Papatheodorou I, Oellrich A, Smedley D. Linking gene expression to phenotypes via pathway information. J Biomed Semantics. 2015;6(1):17. https://doi.org/10.1186/s13326-015-0013-5.
2. Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC. Orphanet: a European database for rare diseases. Ned Tijdschr Geneeskd. 2008;152(9):518–9.
3. McKusick VA. Mendelian inheritance in man and its online version, OMIM. Am J Hum Genet. 2007;80(4):588–604. https://doi.org/10.1086/514346.
4. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018–D27. https://doi.org/10.1093/nar/gky1105.
5. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
6. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(Database issue):D428–32. https://doi.org/10.1093/nar/gki072.
7. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. Plos Biol. 2008;6(7):e184. https://doi.org/10.1371/journal.pbio.0060184.
8. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med. 2018;379(15):1452–62. https://doi.org/10.1056/NEJMra1615014.
9. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12(9):841–3. https://doi.org/10.1038/nmeth.3484.
10. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics P, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8. https://doi.org/10.1101/gr.160325.113.
11. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. Am J Hum Genet. 2018;103(1):58–73. https://doi.org/10.1016/j.ajhg.2018.05.010.
12. Birgmeier J, Haeussler M, Deisseroth CA, Steinberg EH, Jagadeesh KA, Ratner AJ, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020;12(544). https://doi.org/10.1126/scitranslmed.aau9113.
13. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–D17. https://doi.org/10.1093/nar/gkaa1043.
14. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265–6.
15. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–8.
16. World HO. International Statistical Classification of Diseases and Related Health Problems: Tabular list: World Health Organization; 2004.
17. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267–70. https://doi.org/10.1093/nar/gkh061.
18. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. J Am Med Inform Assoc. 2020;27(10):1606–11. https://doi.org/10.1093/jamia/ocaa084.
19. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(Database issue):D940–6. https://doi.org/10.1093/nar/gkr972.
20. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019;47(D1):D955–D62. https://doi.org/10.1093/nar/gky1032.
21. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47(D1):D1038–D43. https://doi.org/10.1093/nar/gky1151.
22. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84(4):524–33. https://doi.org/10.1016/j.ajhg.2009.03.010.
23. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, et al. MalaCards: an integrated compendium for diseases and their annotation. Database. 2013;2013:bat018. https://doi.org/10.1093/database/bat018.
24. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.
25. Gormley C, Tong Z Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. 1st ed. Sebastopol: O'Reilly Media; 2015.
26. Grinberg M. Flask web development: developing web applications with Python. 2nd ed. Sebastopol: O'Reilly Media; 2018.
27. Fielding RT, Kaiser G. The Apache HTTP Server Project. IEEE Internet Comput. 1997;1(4):88–90. https://doi.org/10.1109/4236.612229.
28. Aas J, Barnes R, Case B, Durumeric Z, Eckersley P, Flores-López A, et al. Let's Encrypt: an automated certificate authority to encrypt the entire web. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security: Association for Computing Machinery; 2019. p. 2473-87.
29. Moritz P, Nishihara R, Wang S, Tumanov A, Liaw R, Liang E, et al. Ray: a distributed framework for emerging AI applications. arXiv [csDC]. 2017.
30. Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, et al. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017;45(D1):D995–D1002. https://doi.org/10.1093/nar/gkw1072.
31. Weber GM, Barnett W, Conlon M, Eichmann D, Kibbe W, Falk-Krzesinski H, et al. Direct2Experts Collaboration Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. J Am Med Inform Assoc. 2011;18(Suppl 1):i157–60. https://doi.org/10.1136/amiajnl-2011-000200.
32. Kass-Hout TA, Xu Z, Mohebbi M, Nelsen H, Baker A, Levine J, et al. OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. J Am Med Inform Assoc. 2016;23(3):596–600. https://doi.org/10.1093/jamia/ocv153.
33. Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen D-T, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res. 2021;49(D1):D1160–D9. https://doi.org/10.1093/nar/gkaa997.

34. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39(Database issue):D685–90. https://doi.org/10.1093/nar/gkq1039.
35. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials. gov results database—update and key issues. N Engl J Med. 2011;364(9):852–60. https://doi.org/10.1056/NEJMsa1012065.
36. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. Sci Data. 2018;5(1):180273. https://doi.org/10.1038/sdata.2018.273.
37. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. 2019;47(W1):W566–W70. https://doi.org/10.1093/nar/gkz386.
38. Zhao M, Havrilla JM, Fang L, Chen Y, Peng J, Liu C, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genom Bioinform. 2020;2(2):lqaa032. https://doi.org/10.1093/nargab/lqaa032.
39. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7–19. https://doi.org/10.1093/nar/gkv1290.
40. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–12. https://doi.org/10.1093/nar/gku989.
41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
42. Miller N, Lacroix EM, Backus JE. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc. 2000;88(1):11–7.
43. Griffin LB, Farley FA, Antonellis A, Keegan CE. A novel FGD1 mutation in a family with Aarskog-Scott syndrome and predominant features of congenital joint contractures. Cold Spring Harb Mol Case Stud. 2016;2(4):a000943. https://doi.org/10.1101/mcs.a000943.
44. Aho AV, Corasick MJ. Efficient string matching. Communications of the ACM. 1975;18(6):333–40. https://doi.org/10.1145/360825.360855.
45. Johnson D, Wilkie AOM. Craniosynostosis. Eur J Hum Genet. 2011;19(4):369–76. https://doi.org/10.1038/ejhg.2010.235.
46. Nevels RM, Gontkovsky ST, Williams BE. Paroxetine-the antidepressant from hell? Probably not,but caution required. Psychopharmacol Bull. 2016;46(1):77–104.
47. Reefhuis J, Devine O, Friedman JM, Louik C, Honein MA, National Birth Defects Prevention S. Specific SSRIs and birth defects: Bayesian analysis to interpret new data in the context of previous reports. BMJ. 2015;351:h3190. https://doi.org/10.1136/bmj.h3190.
48. Bérard A, Zhao J-P, Sheehy O. Sertraline use during pregnancy and the risk of major malformations. Am J Obstet Gynecol. 2015;212(6):795.e1-.e12. https://doi.org/10.1016/j.ajog.2015.01.034.
49. Wenger T, Miller D, Evans K. FGFR craniosynostosis syndromes overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, LJH B, Stephens K, et al., editors. GeneReviews®. Seattle: University of Washington; 1998.
50. Carboni L, Pischedda F, Piccoli G, Lauria M, Musazzi L, Popoli M, et al. Depression-associated gene Negr1-Fgfr2 pathway is altered by antidepressant treatment. Cells. 2020;9(8). https://doi.org/10.3390/cells9081818.
51. Havrilla J, Dong S, Liu C, Wang K. WGLab/PhenCards: First public release of PhenCards; 2021. https://doi.org/10.5281/zenodo.4759231.
52. Havrilla J. PhenCards v.1.0.0 database; 2021. https://doi.org/10.5281/zenodo.4755959.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions