Published online 18 November 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1353–D1359***
https://doi.org/10.1093/nar/gkac1046

The next-generation Open Targets Platform: reimagined, redesigned, rebuilt

David Ochoa[1,2,*], Andrew Hercules[1,2], Miguel Carmona[1,2], Daniel Suveges[1,2], Jarrod Baker[1,2], Cinzia Malangone[1,2], Irene Lopez[1,2], Alfredo Miranda[1,2], Carlos Cruz-Castillo[1,2], Luca Fumis[1,2], Manuel Bernal-Llinares[1,2], Kirill Tsukanov[1,2], Helena Cornu[1,2], Konstantinos Tsirigos[1,2], Olesya Razuvayevskaya[1,2], Annalisa Buniello[1,2], Jeremy Schwartzentruber[1,3], Mohd Karim[1,3], Bruno Ariano[1,3], Ricardo Esteban Martinez Osorio[1,2], Javier Ferrer[1,2], Xiangyu Ge[1,3], Sandra Machlitt-Northen[4], Asier Gonzalez-Uriarte[1,2], Shyamasree Saha[2], Santosh Tirunagari[2], Chintan Mehta[1,2], Juan María Roldán-Romero[1,2], Stuart Horswell[1,3], Sarah Young[1,3], Maya Ghoussaini[1,3], David G. Hulcoop[1,4], Ian Dunham[1,2,3] and Ellen M. McDonagh[1,2,*]

[1]Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK and [4]GlaxoSmithKline plc, GSK Medicines Research Centre, Gunnels Wood Road, Stevenage, SG1 2NY, UK

Received September 15, 2022; Revised October 14, 2022; Editorial Decision October 20, 2022; Accepted October 27, 2022

ABSTRACT

The Open Targets Platform (https://platform.opentargets.org/) is an open source resource to systematically assist drug target identification and prioritisation using publicly available data. Since our last update, we have reimagined, redesigned, and rebuilt the Platform in order to streamline data integration and harmonisation, expand the ways in which users can explore the data, and improve the user experience. The gene–disease causal evidence has been enhanced and expanded to better capture disease causality across rare, common, and somatic diseases. For target and drug annotations, we have incorporated new features that help assess target safety and tractability, including genetic constraint, PROTACtability assessments, and AlphaFold structure predictions. We have also introduced new machine learning applications for knowledge extraction from the published literature, clinical trial information, and drug labels. The new technologies and frameworks introduced since the last update will ease the introduction of new features and the creation of separate instances of the Platform adapted to user requirements. Our new Community forum, expanded training materials, and outreach programme support our users in a range of use cases.

INTRODUCTION

The Open Targets consortium brings together expertise and capabilities from academic and pharmaceutical industry partners with the vision to systematically identify targets that will ultimately lead to more effective and safer drugs for disease treatment. To facilitate therapeutic hypothesis building, the consortium experimentally generates novel evidence and contextualises it with available knowledge in the public domain. In this task, genetics provides an unprecedented source of causal evidence, to the extent that two-

[*]To whom correspondence should be addressed. Tel: +44 1223 494330; Email: emcdonagh@ebi.ac.uk
Correspondence may also be addressed to David Ochoa. Tel: +44 1223 494330; Email: ochoa@ebi.ac.uk
Present addresses:
Ellen M. McDonagh, Open Targets and European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.
David Ochoa, Open Targets and European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.

thirds of the drugs approved in 2021 were directly or indirectly supported by genetic evidence (1–3).

The Open Targets Platform (https://platform.opentargets.org/) provides an open source informatic solution for the identification and prioritisation of targets using publicly available data (4). The Platform provides the necessary knowledgebase to characterise targets, diseases and drugs in the context of drug discovery, as well as the relationships between the entities, with particular focus on target–disease associations (Figure 1). Powered with an in-house target identification scoring framework, evidence is aggregated across sources to provide ranked lists of gene–disease associations. To maintain up-to-date-evidence, ensure regular updates from external data sources and integrate user feedback and new features, the Platform is released five times a year.

Within the last 2 years, the Platform has been widely used by the community as a source of truth in various contexts, including; supporting Crohn's disease associations or approved kinase inhibitors, as a prioritisation tool in the search for drug repurposing opportunities, or more generally as a data source to build corporate knowledge graphs to assist drug development (5–7). Moreover, our open-source code has been re-used by the NIH National Cancer Institute Childhood Cancer Data Initiative to develop and launch a customised Molecular Targets Platform that integrates patient-specific data, aiming to facilitate therapeutic hypotheses building and target discovery in paediatric cancer (https://moleculartargets.ccdi.cancer.gov/). This is just the first example of the many potential applications of the Platform open-source code to create separate instances in order to integrate experimental or pre-publication results.

Herein we provide an update on the significant enhancements made to the Platform since our last publication that contribute to supporting the aims of the consortium as well as the world-wide community (Figure 2).

Enhanced target–disease evidence

The Platform assists target identification by integrating multiple layers of information pinpointing likely causal targets with biological plausibility. A fundamental element of this effort is the target–disease evidence, which we collate from 22 different sources, covering evidence derived from genetic associations, somatic mutations, known drugs, differential expression, animal models, pathways and systems biology. Since the last update, we have expanded the Platform to include several new data sources which we describe below, as well as providing regular updates and expanding the available information on each piece of evidence. Some of the data streams stem directly from public resources in our partner institutions (8–14). An important aspect of the data integration is that all the phenotypic traits are mapped using the Experimental Factor Ontology (EFO), which we have recently expanded to better cover quantitative traits and medical procedures (15).

Common and complex diseases. For common variation, the Platform uniquely relies on the closely developed post-GWAS analysis available in Open Targets Genetics (https://genetics.opentargets.org/) (16). Since the last update, the Genetics portal has integrated new studies deposited in the GWAS Catalog—with or without summary statistics—as well as a series of FinnGen public data releases (17,18). To enhance the causal inference, the Genetics portal enriched the L2G machine learning (ML) regression with additional eQTLs from GTEx v8 and pQTLs from 6 new studies (19,20).

Gene burden analysis. The increasing number of sequenced individuals has led to a proliferation of studies aimed at identifying rare or ultra-rare variants in coding regions that contribute to common diseases. To respond to this new source of target–disease evidence, we included a new Gene Burden dataset capturing the results from collapsing analysis on exome or whole-genome sequencing studies (Figure 3). Among other publications, we ingested three studies analysing the 450 000 exomes sequenced by the UK BioBank to a total of 5459 unique gene–disease associations, 57% of them not previously reported on any other genetic resource included in the Platform (21–23). The aggregation of likely deleterious variants in coding regions also provides a putative mechanistic explanation that can be leveraged to build a therapeutic hypothesis. Moreover, gene burden constitutes the first ancestry-specific evidence in the Platform, a feature that will be expanded to other data sources where possible.

Rare diseases. The Platform continues to integrate several resources providing variant- and gene-centric evidence derived from the clinical assessment of families or individuals with rare diseases. Since the last update, we included new versions of these resources mapped to EFO and two new datasets: Orphanet and the ClinGen Gene-Disease Validity assessments which provide an evaluation of the strength of the evidence supporting or refuting a claim that variation in a particular gene causes a particular disease (24,25). Each respective resource currently brings 6003 and 1655 unique associations, with 19.3% and 28.7% unique target–disease pairs not captured by any other genetic resource in the Platform. To align with the Gene Curation Coalition guidelines, we have adopted the standardised terminologies for confidence level and display allelic requirements when available (26). In collaboration with the European Variation Archive, we have also included complex structural variants in ClinVar such as insertions, deletions, and small tandem repeats, aiming to cover a broader range of disease-causing mechanisms (27).

Somatic variation. A new data widget was created to bring in cancer biomarker annotations sourced from the Cancer Genome Interpreter, which provide evidence for target–disease associations (and importantly the level of clinical evidence from pre-clinical to approved drug standard-of-care guidelines) (28). These also contribute to biomarker-drug response and drug resistance information. We also brought in cancer driver predictions from IntOgen's analysis of the Pan-Cancer Analysis of Whole Genomes study of the International Cancer Genome Consortium, and developed the visualisation to provide clear information of the different driver methods used to predict the target as a driver gene in different cohorts (29).

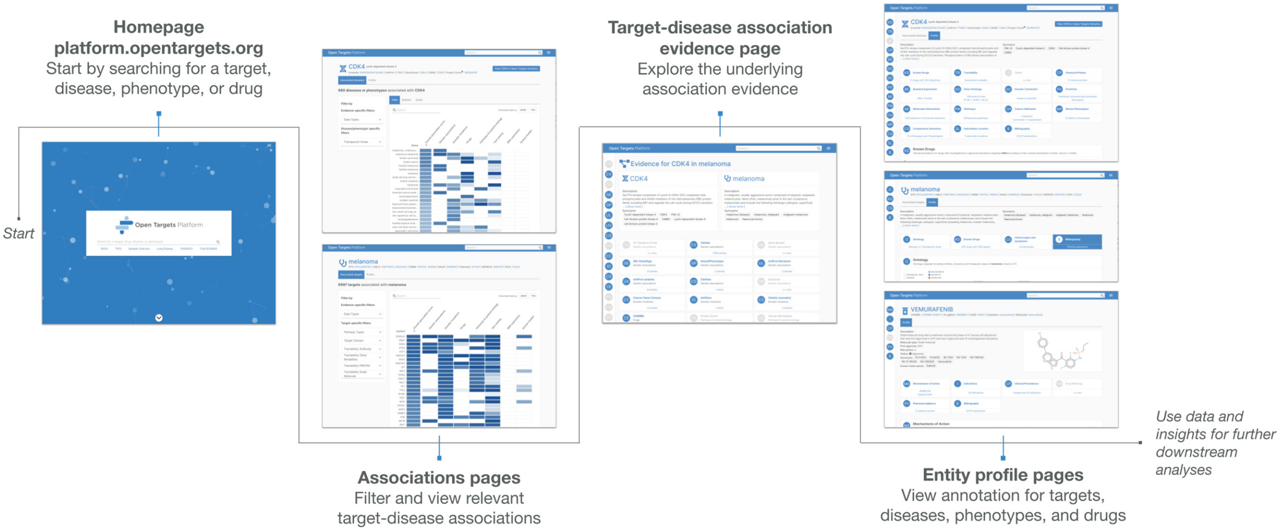
Homepage
platform.opentargets.org
Start by searching for a target, disease, phenotype, or drug
Start
Open Targets Platform
Associations pages
Filter and view relevant target-disease associations
Target-disease association evidence page
Explore the underlying association evidence
Evidence for CDK4 in melanoma
CDK4
melanoma
VEMURAFENIB
Entity profile pages
View annotation for targets, diseases, phenotypes, and drugs
Use data and insights for further downstream analyses

Figure 1. A journey through the Platform web interface. The Open Targets Platform web interface is the first point of access for most users, and was completely redesigned to create the Next Generation Platform. The unified search box is connected to a series of tools allowing users to query different therapeutic hypotheses. From the homepage, users can navigate to association pages, with prioritised lists of target–disease associations. From there, users can access target–disease evidence pages, detailing the available evidence for an association. Once the evidence for a target–disease association has been assessed, users can explore entity profile pages, containing annotation information for each target, disease/phenotype and drug in the Platform to further build their hypothesis. For targets, this includes investigating whether it is expressed in a suitable tissue, what type of modality may be suitable and whether modulation is likely to be safe, whether there are already known drugs or available chemical probes for validation experiments, and whether interacting proteins may be more suitable targets. For disease/phenotype, the user can investigate known drugs and their targets, or explore targets associated with disease phenotypes through ontology expansion. Drug annotation pages provide a user with the mechanism of action and safety information related to modulating a target.

New literature mining strategy

The Platform relies on knowledge extraction from the biomedical literature as an additional source of target–disease evidence and annotation for the three main entities: targets, diseases, and drugs. In collaboration with Europe PubMedCentral, we refactored the pipeline to perform named-entity recognition on all abstracts and open access full-text documents using a BioBERT model fine-tuned for the task (9,30). To normalise the identified words to the Platform entities, we leverage identifiers, names, symbols and synonyms from diverse sources. In line with the previous implementation, once the entity occurrences are associated with up-to-date identifiers the new pipeline generates evidence based on sentences in which target and disease co-occurred (31). The bibliography widget allows users to interrogate the literature by entity, displaying a list of publications mentioning the selected term. The user can also expand the selection by including other suggested entities in the search based on the similarity of their BioBERT-driven descriptors to the previously selected terms.

The development and integration of natural language processing (NLP) models into the Platform pipelines have enabled additional applications. We now provide a classification of the reasons why a clinical trial stopped—more significantly, separating studies stopped due to efficacy or side effects from those halted due to other reasons independent of the therapeutic hypothesis. In parallel, a ML model is used to expand the list of drug indications by mining drug labels (11). Overall, the expansion of NLP applications allows the Platform to provide insights into semi-structured data relevant to building a more robust therapeutic hypothesis.

Improving target characterisation

Prioritising likely causal targets in a drug discovery program requires a comprehensive understanding of the target biology that could inform risks and strategies for modulating the target. Since the last update, several enhancements aim to expand the Platform's ability to inform about the target tractability for different modalities. For example, the PROTAC tractability assessment includes a set of target properties that can help determine whether protein degradation would be a suitable modulation strategy (32). Similarly, we expanded the list of tool molecules that are selective and specific against the target by including the data integration from Probes and Drugs (33). The recent inclusion of the AlphaFold protein structure predictions in the target profile page allows the identification of potential druggable pockets in previously unresolved 3D models (34). This information, combined with the enhanced subcellular localisation widget, better informs about the most likely viable target modulation strategy (35).

Identifying target properties that might raise safety issues can also influence the prioritisation strategy. A new 'Genetic Constraint' data widget aims to capture how evolutionarily protected a gene is from loss-of-function variation, a proxy to inform whether modulating the target is likely to be tolerated. This section includes the observed and expected frequency of synonymous, nonsynonymous, and predicted

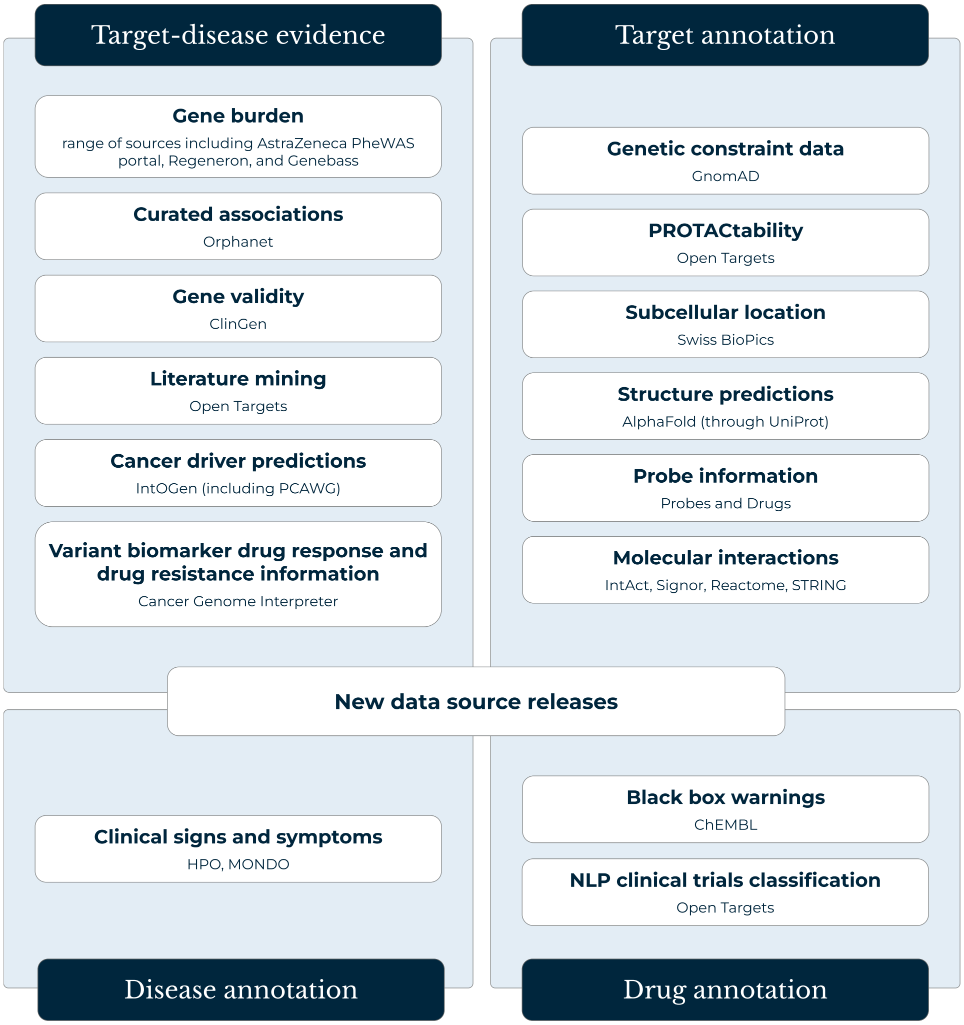

Figure 2. New data in the Open Targets Platform. Additional sources of target–disease evidence and new types of annotation for targets diseases and drugs during the last 2 years.

loss-of-function mutations together with the pLOEUF genetic constraint assessment provided by GnomAD (36). In addition, the safety widget information has been expanded to incorporate further safety events associated with targets. On drug profile pages, BlackBox warnings and drug withdrawal information have also been expanded to increase the drug-associated safety events (37).

To contextualise the target in the context of its community, we incorporated the new molecular interactions section (38). This feature aggregates and integrates the network topology, as well as the supporting evidence from 4 different sources capturing different types of interactions; IntAct (physical interactions), STRING (functional interactions), Signor (directional signalling interactions), and Reactome (enzymatic reaction pathway-based interactions) (14,39–41). The data widget for molecular interactions allows the user to expand upon their therapeutic hypothesis and explore options for alternative targets that interact with or share the same biological pathway with a known disease-causing protein or a protein for which there is already a known drug.

The next-generation platform

Since the last update, the Open Targets Platform underwent a major refactor of all its data and services in an attempt to unblock new target identification and prioritisation capabilities. Powered by a streamlined data model, the newly developed Spark extract-transform-load pipeline produces a set of datasets capturing all of the Platform knowledgebase. Notably, the target–disease evidence dataset now integrates potentially causal evidence from our 22 datasources in a harmonised format, simplifying the development of any downstream applications. Similarly, a newly designed GraphQL API serves the information to a brand new React web application that was released to the public in June 2021.

While the web application remains the Platform entry point for most users (Figure 1), a larger range of options exists for power users of the data. The GraphQL API might be a good option for users planning to query a few records in the data, with the downloadable datasets representing the most interesting option for most users with some programming experience. For cloud developers, the Platform

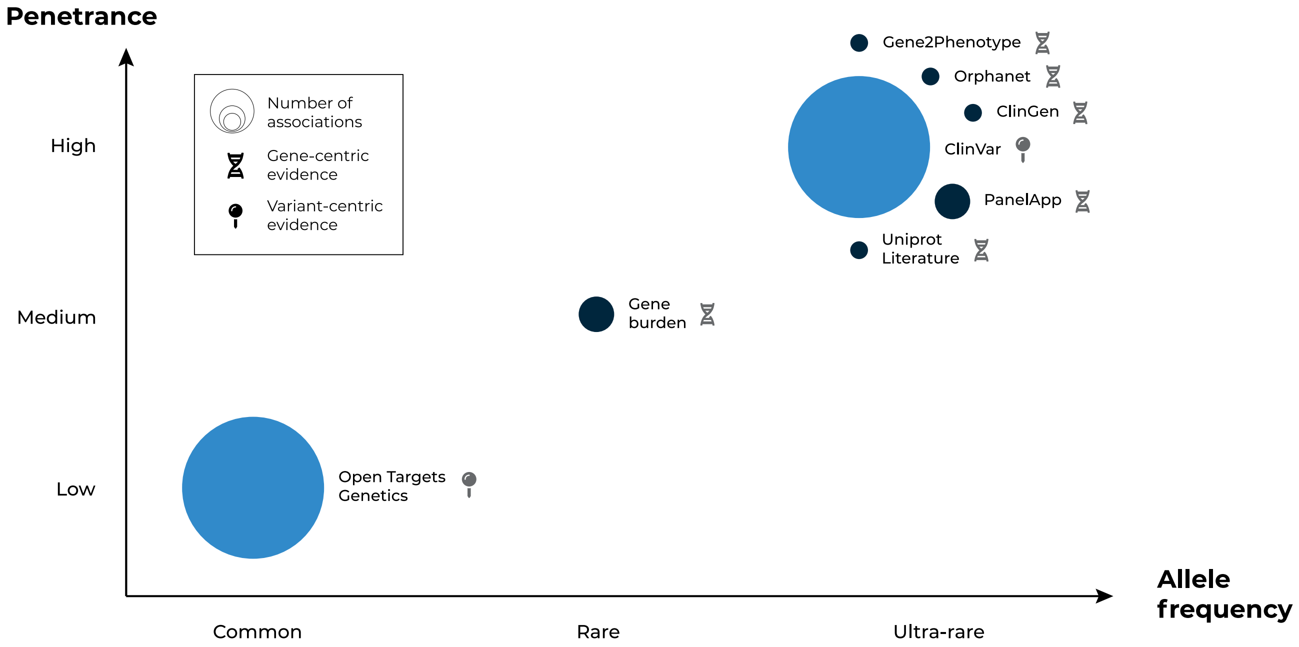

Figure 3. Schematic representation of germline genetic evidence in the Platform. Data sources—bubbles—are classified based on the predominant allelic frequency and penetrance of the reported genetic variation. The size of the bubbles represents the number of target–disease associations provided by each data source sorted into three bins: 1) 1000–10 000—from 1966 (ClinGen) to 8506 (Orphanet); 2) 10 000–100 000—with 27 162 target–disease associations from Gene Burden analyses to 40 446 from PanelApp; 3) and 100 000 + associations (Open Targets Genetics: 694 214; ClinVar: 1 541 903). Data sources are classified depending on whether they capture genetic variation at the variant (lighter blue) or gene level (dark blue).

data was included in 2022 as a public dataset in Google BigQuery and AWS Open Data. A comprehensive description of the Platform data and how to access it can be found in the renovated documentation site (https://platform-docs.opentargets.org).

The Open Targets community

The expansion of the Open Targets Platform was paralleled by an expansion of our outreach activities. In particular, we created the Open Targets Community forum (community.opentargets.org) in April 2021, to host discussions, foster collaborations, and leverage feedback from our users. We also developed our training materials to support users in a variety of mediums and access levels. For example, working with EMBL-EBI's training team, we hosted a series of webinars, which were repurposed intro training courses on the Platform and its API (https://www.ebi.ac.uk/training/online/courses/open-targets-quick-tour/). We also featured posts on the Open Targets blog detailing how to access and use the data in the Platform, the architecture of our re-written pipeline, and our plans for the future of the Platform.

DISCUSSION

Target identification and prioritisation remain fundamental challenges in the goal of developing safe and effective medicines. As a core resource of the Open Targets consortium, the Platform aims to capitalise upon the expertise and data produced in the public–private partnership, as well as from the public realm, to provide a systematic resource in the pre-competitive target discovery space. The continuous generation of potential causal evidence from increasingly complicated data models, granularity of disease characterisation, and large population-scale datasets, forced us to redesign the core functionalities of the Platform leading to the more modern resource described here. The adoption of these new technologies and frameworks will speed up future development, enhancing the open source contributions to the codebase (https://github.com/opentargets), and facilitate the creation of separate instances of the Platform. It will allow us to bring in more complex datasets (such as temporal single cell transcriptomics) and introduce sophisticated ways for users to build therapeutic hypotheses.

DATA AVAILABILITY

All code is available in GitHub: https://github.com/opentargets.

ACKNOWLEDGEMENTS

The authors would like to thank the Platform users, data providers and open source contributors to the codebase. We also thank all the partners (Wellcome Sanger Institute, EMBL-EBI, GSK, Bristol Myers Squibb, Sanofi and Pfizer) and our Scientific Advisory Board for the crucial strategy discussions. We thank our previous partners Takeda for their contributions and the Hyve for their contributions to the codebase. This research was funded in part by a Wellcome Trust grant [Grant number 206194]. For the purpose of Open Access, the authors have applied a CC-BY

public copyright licence to any author-accepted manuscript version arising from this submission.

FUNDING

Wellcome Trust [206194]; Open Targets. Funding for open access charge: Open Targets.

Conflict of interest statement. None declared.

REFERENCES

1. Ochoa,D., Karim,M., Ghoussaini,M., Hulcoop,D.G., McDonagh,E.M. and Dunham,I. (2022) Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat. Rev. Drug Discov.*, **21**, 551.
2. Nelson,M.R., Tipney,H., Painter,J.L., Shen,J., Nicoletti,P., Shen,Y., Floratos,A., Sham,P.C., Li,M.J., Wang,J. *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.*, **47**, 856–860.
3. King,E.A., Davis,J.W. and Degner,J.F. (2019) Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.*, **15**, e1008489.
4. Ochoa,D., Hercules,A., Carmona,M., Suveges,D., Gonzalez-Uriarte,A., Malangone,C., Miranda,A., Fumis,L., Carvalho-Silva,D., Spitzer,M. *et al.* (2021) Open targets platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.
5. Attwood,M.M., Fabbro,D., Sokolov,A.V., Knapp,S. and Schiöth,H.B. (2021) Trends in kinase drug discovery: targets, indications and inhibitor design. *Nat. Rev. Drug Discov.*, **20**, 839–861.
6. Pietzner,M., Wheeler,E., Carrasco-Zanini,J., Cortes,A., Koprulu,M., Wörheide,M.A., Oerton,E., Cook,J., Stewart,I.D., Kerrison,N.D. *et al.* (2021) Mapping the proteo-genomic convergence of human diseases. *Science*, **374**, eabj1541.
7. Geleta,D., Nikolov,A., Edwards,G., Gogleva,A., Jackson,R., Jansson,E., Lamov,A., Nilsson,S., Pettersson,M., Poroshin,V. *et al.* (2021) Biological insights knowledge graph: an integrated knowledge graph to support drug development. bioRxiv doi: http://doi.org/10.1101/2021.10.28.466262, 01 November 2021, preprint: not peer reviewed.
8. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
9. Ferguson,C., Araújo,D., Faulk,L., Gou,Y., Hamelers,A., Huang,Z., Ide-Smith,M., Levchenko,M., Marinos,N., Nambiar,R. *et al.* (2021) Europe PMC in 2020. *Nucleic Acids Res.*, **49**, D1507–D1514.
10. Cezard,T., Cunningham,F., Hunt,S.E., Koylass,B., Kumar,N., Saunders,G., Shen,A., Silva,A.F., Tsukanov,K., Venkataraman,S. *et al.* (2022) The European variation archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.*, **50**, D1216–D1220.
11. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Félix,E., Magariños,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
12. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Füllgrabe,A., Fuentes,A.M.-P., George,N. *et al.* (2018) Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
13. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
14. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
15. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
16. Ghoussaini,M., Mountjoy,E., Carmona,M., Peat,G., Schmidt,E.M., Hercules,A., Fumis,L., Miranda,A., Carvalho-Silva,D., Buniello,A. *et al.* (2021) Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.*, **49**, D1311–D1320.
17. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
18. Kurki,M.I., Karjalainen,J., Palta,P., Sipilä,T.P., Kristiansson,K., Donner,K., Reeve,M.P., Laivuori,H., Aavikko,M., Kaunisto,M.A. *et al.* (2022) FinnGen: unique genetic insights from combining isolated population and national health register data. medRxiv doi: http://doi.org/10.1101/2022.03.03.22271360, 06 March 2022, preprint: not peer reviewed.
19. Mountjoy,E., Schmidt,E.M., Carmona,M., Schwartzentruber,J., Peat,G., Miranda,A., Fumis,L., Hayhurst,J., Buniello,A., Karim,M.A. *et al.* (2021) An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.*, **53**, 1527–1533.
20. GTEx Consortium. (2020) The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
21. Wang,Q., Dhindsa,R.S., Carss,K., Harper,A.R., Nag,A., Tachmazidou,I., Vitsios,D., Deevi,S.V.V., Mackay,A., Muthas,D. *et al.* (2021) Rare variant contribution to human disease in 281,104 UK biobank exomes. *Nature*, **597**, 527–532.
22. Backman,J.D., Li,A.H., Marcketta,A., Sun,D., Mbatchou,J., Kessler,M.D., Benner,C., Liu,D., Locke,A.E., Balasubramanian,S. *et al.* (2021) Exome sequencing and analysis of 454,787 UK biobank participants. *Nature*, **599**, 628–634.
23. Karczewski,K.J., Solomonson,M., Chao,K.R., Goodrich,J.K., Tiao,G., Lu,W., Riley-Gillis,B.M., Tsai,E.A., Kim,H.I., Zheng,X. *et al.* (2022) Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK biobank exomes. *Cell Genomics*, **2**, 100168.
24. Strande,N.T., Riggs,E.R., Buchanan,A.H., Ceyhan-Birsoy,O., DiStefano,M., Dwight,S.S., Goldstein,J., Ghosh,R., Seifert,B.A., Sneddon,T.P. *et al.* (2017) Evaluating the clinical validity of gene–disease associations: an evidence-based framework developed by the clinical genome resource. *Am. J. Hum. Genet.*, **100**, 895–906.
25. Rodwell,C. and Aymé,S. (2015) Rare disease policies to improve care for patients in europe. *Biochim. Biophys. Acta*, **1852**, 2329–2335.
26. DiStefano,M.T., Goehringer,S., Babb,L., Alkuraya,F.S., Amberger,J., Amin,M., Austin-Tse,C., Balzotti,M., Berg,J.S., Birney,E. *et al.* (2022) The gene curation coalition: a global effort to harmonize gene–disease evidence resources. *Genet. Med.*, **24**, 1732–1742.
27. Landrum,M.J., Chitipiralla,S., Brown,G.R., Chen,C., Gu,B., Hart,J., Hoffman,D., Jang,W., Kaur,K., Liu,C. *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.*, **48**, D835–D844.
28. Tamborero,D., Rubio-Perez,C., Deu-Pons,J., Schroeder,M.P., Vivancos,A., Rovira,A., Tusquets,I., Albanell,J., Rodon,J., Tabernero,J. *et al.* (2018) Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.*, **10**, 25.
29. Martínez-Jiménez,F., Muiños,F., Sentís,I., Deu-Pons,J., Reyes-Salazar,I., Arnedo-Pac,C., Mularoni,L., Pich,O., Bonet,J., Kranas,H. *et al.* (2020) A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*, **20**, 555–572.
30. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
31. Kafkas,Ş., Dunham,I. and McEntyre,J. (2017) Literature evidence in open targets - a target validation platform. *J. Biomed. Semantics*, **8**, 20.
32. Schneider,M., Radoux,C.J., Hercules,A., Ochoa,D., Dunham,I., Zalmas,L.-P., Hessler,G., Ruf,S., Shanmugasundaram,V., Hann,M.M. *et al.* (2021) The PROTACtable genome. *Nat. Rev. Drug Discov.*, **20**, 789–797.
33. Skuta,C., Popr,M., Muller,T., Jindrich,J., Kahle,M., Sedlak,D., Svozil,D. and Bartunek,P. (2017) Probes &Drugs portal: an interactive, open data resource for chemical biology. *Nat. Methods*, **14**, 759–760.

34. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
35. Le Mercier,P., Bolleman,J., de Castro,E., Gasteiger,E., Bansal,P., Auchincloss,A.H., Boutet,E., Breuza,L., Casals-Casas,C., Estreicher,A *et al.* (2022) SwissBioPics-an interactive library of cell images for the visualization of subcellular location data. *Database*, **2022**, baac026.
36. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
37. Hunter,F.M.I., Bento,A.P., Bosc,N., Gaulton,A., Hersey,A. and Leach,A.R. (2021) Drug safety data curation and modeling in chembl: boxed warnings and withdrawn drugs. *Chem. Res. Toxicol.*, **34**, 385–395.
38. Barrio-Hernandez,I., Schwartzentruber,J., Shrivastava,A., del-Toro,N., Zhang,Q., Bradley,G., Hermjakob,H., Orchard,S., Dunham,I., Anderson,C.A. *et al.* (2021) Network expansion of genetic associations defines a pleiotropy map of human cell biology. bioRxiv doi: http://doi.org/10.1101/2021.07.19.452924, 19 July 2021, preprint: not peer reviewed.
39. Göös,H., Kinnunen,M., Salokas,K., Tan,Z., Liu,X., Yadav,L., Zhang,Q., Wei,G.-H. and Varjosalo,M. (2022) Human transcription factor protein interaction networks. *Nat. Commun.*, **13**, 766.
40. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
41. Licata,L., Lo Surdo,P., Iannuccelli,M., Palma,A., Micarelli,E., Perfetto,L., Peluso,D., Calderone,A., Castagnoli,L. and Cesareni,G. (2020) SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update. *Nucleic Acids Res.*, **48**, D504–D510.