

The Reactome Pathway Knowledgebase 2024

Marija Milacic¹, Deidre Beavers², Patrick Conley², Chuqiao Gong³, Marc Gillespie^{1,4}, Johannes Griss^{3,5}, Robin Haw¹, Bijay Jassal¹, Lisa Matthews⁶, Bruce May¹, Robert Petryszak², Eliot Ragueneau³, Karen Rothfels¹, Cristoffer Sevilla³, Veronica Shamovsky⁶, Ralf Stephan^{1,7}, Krishna Tiwari³, Thawfeek Varusai^{3,8}, Joel Weiser¹, Adam Wright¹, Guanming Wu^{2,*}, Lincoln Stein^{1,9,*}, Henning Hermjakob^{3,*} and Peter D'Eustachio^{6,*}

¹Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada

²Oregon Health and Science University, Portland, OR 97239, USA

³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

⁴College of Pharmacy and Health Sciences, St. John's University, Queens, NY 11439, USA

⁵Department of Dermatology, Medical University of Vienna, 1090 Vienna, Austria

⁶NYU Grossman School of Medicine, New York, NY 10016, USA

⁷Institute for Globally Distributed Open Research and Education (IGDORE)

⁸Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

⁹Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A1, Canada

*To whom correspondence should be addressed. Tel: +1 212 263 5779; Email: deustp01@med.nyu.edu

Correspondence may also be addressed to Henning Hermjakob. Tel: +44 1223 494 671; Email: hhe@ebi.ac.uk

Correspondence may also be addressed to Lincoln Stein. Tel: +1 416 673 8514; Email: Lincoln.Stein@oicr.on.ca

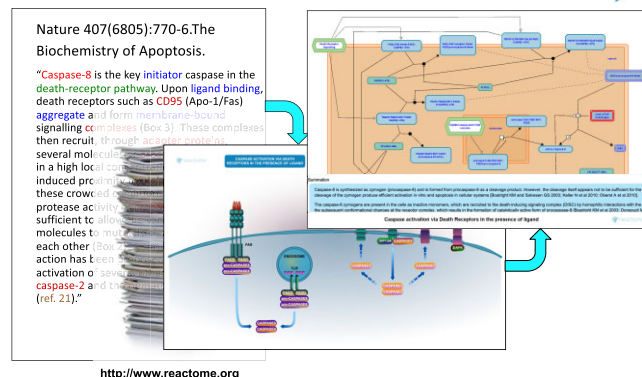
Correspondence may also be addressed to Guanming Wu. Tel: +1 503 494 4502; Fax: +1 503 346 6815 Email: wug@ohsu.edu

Abstract

The Reactome Knowledgebase (<https://reactome.org>), an Elixir and GCBR core biological data resource, provides manually curated molecular details of a broad range of normal and disease-related biological processes. Processes are annotated as an ordered network of molecular transformations in a single consistent data model. Reactome thus functions both as a digital archive of manually curated human biological processes and as a tool for discovering functional relationships in data such as gene expression profiles or somatic mutation catalogs from tumor cells. Here we review progress towards annotation of the entire human proteome, targeted annotation of disease-causing genetic variants of proteins and of small-molecule drugs in a pathway context, and towards supporting explicit annotation of cell- and tissue-specific pathways. Finally, we briefly discuss issues involved in making Reactome more fully interoperable with other related resources such as the Gene Ontology and maintaining the resulting community resource network.

Graphical abstract

Reactome: Computationally Accessible Pathway Reviews



Received: September 29, 2023. Revised: October 14, 2023. Editorial Decision: October 16, 2023. Accepted: October 20, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

At the cellular level, biological processes of cells and tissues can be represented by networks of molecular reactions that enable signal transduction, transport, DNA replication, protein synthesis, and intermediary metabolism. Various online resources capture this information at the level of individual reactions, such as Rhea (1), or at the level of reaction sequences covering many domains of biology, such as KEGG (2) or MetaCyc (3). The Reactome Knowledgebase is distinctive in focusing its manual annotation effort on a single species, *Homo sapiens*, and applying a single consistent data model across all domains of biology. Processes are systematically described in molecular detail to generate an ordered network of molecular transformations, resulting in an extended version of a classic metabolic map (4) generally compliant with the SBGN process description standard (5). The Reactome Knowledgebase systematically links human proteins to their molecular functions, providing a resource that is both a textbook of biological processes and a tool for discovering novel functional relationships in data such as tissue-, cell- or physiological state-specific gene expression, catalogs of somatic mutations in tumor cells, or likely effects of drugs based on their known interactions with proteins in annotated pathways.

Reactome (version 86—September 2023) has entries for 11 148 protein-coding genes involved in 14 803 reactions annotated from 37 156 literature references (Table 1). These reactions are grouped into 2647 pathways (e.g. Interleukin-15 signaling) collected under 29 superpathways (e.g. Immune System) that describe normal cellular functions. A Disease superpathway includes pathways driven by germline and somatic mutations, and ones due to the actions of genes of infectious bacteria and viruses. Genetic disease annotations cover 4919 variant proteins and post-translationally modified forms of them, derived from 354 human gene products, and annotate 1544 disease-specific reactions tagged with 623 Disease Ontology terms (6). Infectious disease pathways include the effects of bacterial toxins, aspects of infection by *Leishmania*, *Listeria* and *Mycobacteria*, and viral infection mediated by influenza, HIV, human cytomegalovirus, and SARS-CoV-1 and -2. In addition, Reactome describes the modulating effects of 1119 drugs on both normal and disease processes.

Annotating the whole human proteome

The goal of the Reactome project is to describe the molecular function of every human protein in the context of a reaction network (7,8). The fraction of the human proteome annotated in Reactome is thus the core metric of progress; other measures are driven by work towards this goal. The 11148 protein gene products now annotated in Reactome are 56.2% of the 19 831 protein-coding genes predicted in the current (GRCh38.p14) human genome assembly (https://ensembl.org/Homo_sapiens/Info/Annotation). A recent survey (9) suggests that experimental evidence is available for ~68% of these predicted gene products, or 13 500, leaving ~2350 annotatable human proteins not yet in Reactome and ~6300 ‘dark’ proteins (10) whose functions are not yet directly accessible for annotation.

Placing proteins with unknown functions into the context of pathways using evidence from high-throughput surveys of protein/protein interactions and gene co-expression is a use-

ful strategy to predict the functions of these proteins. We first implemented this strategy in the FIViz tool (11). We have extended this work, in collaboration with the IDG/Illuminating the Druggable Genome project, to develop a robust, user-friendly web-based computational framework to associate human proteins not yet manually curated in Reactome with Reactome pathways. Our framework, trained with a random forest of 106 protein or gene pairwise relationship features, infers functional involvement of proteins in individual pathways based on pathway enrichment analysis and fuzzy logic based simulation. We validated these inferences by mining PubMed abstracts, analyzing independent single cell RNA-seq data and manually curating a sample of the inferred ‘dark’ protein pathway assignments. This framework is implemented as a web application, the Reactome IDG portal, <https://idg.reactome.org>. It can be applied to any protein not annotated in Reactome to identify candidate pathways in which the protein may function. It can also be applied to annotated proteins to identify crosstalk between pathways and fill other annotation gaps. Pilot work suggests that this strategy can generate ‘guilt by association’ relationships between about half the ~6300 ‘dark’ proteins and Reactome pathways (12,13).

Annotating germline and somatic genetic variation

The number of variants of human protein coding genes, whether discovered in targeted searches for causes of disease or inferred from genomic sequencing, is large and rapidly growing. Catalogs like ClinVar (14) and COSMIC (15) classify variants by their likely effects on protein structure and function, association with known disease phenotypes, and evidence quality.

If a germline or somatic mutation leads to expression of a protein gene product that has lost its normal function or that has gained a novel one, annotation of the variant protein and of the resulting variant reaction in Reactome is straightforward in principle. Just as a co- or post-translational modification of a protein is annotated as replacement of a specified amino acid residue with a covalently modified one, germline and somatic genetic variants are annotated as the replacement of the amino acid residue normally found at a position by a different one. Our annotation strategy enables us to distinguish two types of reactions involving these variant proteins. If the variant protein has lost its normal function, reactions that require the wild-type protein as an input, catalyst, or regulator fail, having inputs but no outputs. If a variant gene product has gained a novel function or lost sensitivity to normal regulatory processes, reactions result with abnormal outputs or normal outputs under abnormal conditions (Figure 1).

Reactome thereby allows visualization of effects of mutations in a pathway context: what causally downstream processes might be expected to fail as the result of loss of function in a protein in the pathway? Do bypass processes exist that might relieve these effects? What abnormalities might be expected if a protein in a signaling process that is normally tightly regulated becomes constitutively active as the result of a gain of function mutation? Cross-referencing each variant reaction to its normal counterpart and thus to the pathway or pathways in which the normal counterpart functions allows the generation and side-by-side comparison of normal and variant/mutant reaction networks, enabling users to visualize possible disruptive effects of the mutated protein.

Table 1. Reactome content, version 78 (9/2021) versus 86 (9/2023)

Data type	Release 78	Release 86	Change
Human proteins	10 726	11 148	422
Proteoforms	29 466	30 338	872
Chemicals	1940	2025	85
Reactions	13 890	14 803	913
Human disease proteins	352	354	2
Disease variants	4603	4919	316
Chemical drugs	468	1033	565
Protein drugs	39	86	47
Literature references	34 025	37 156	3131

'Human proteins' is the number of human UniProt entries (not counting isoforms) annotated in Reactome. Each may be represented as multiple proteoforms to account for covalent modifications and subcellular locations. 'Disease variants' are ones whose germline or somatic variation gives rise to proteins with altered, pathogenic functions. 'Disease variants' is the total number of such variant alleles annotated in Reactome.

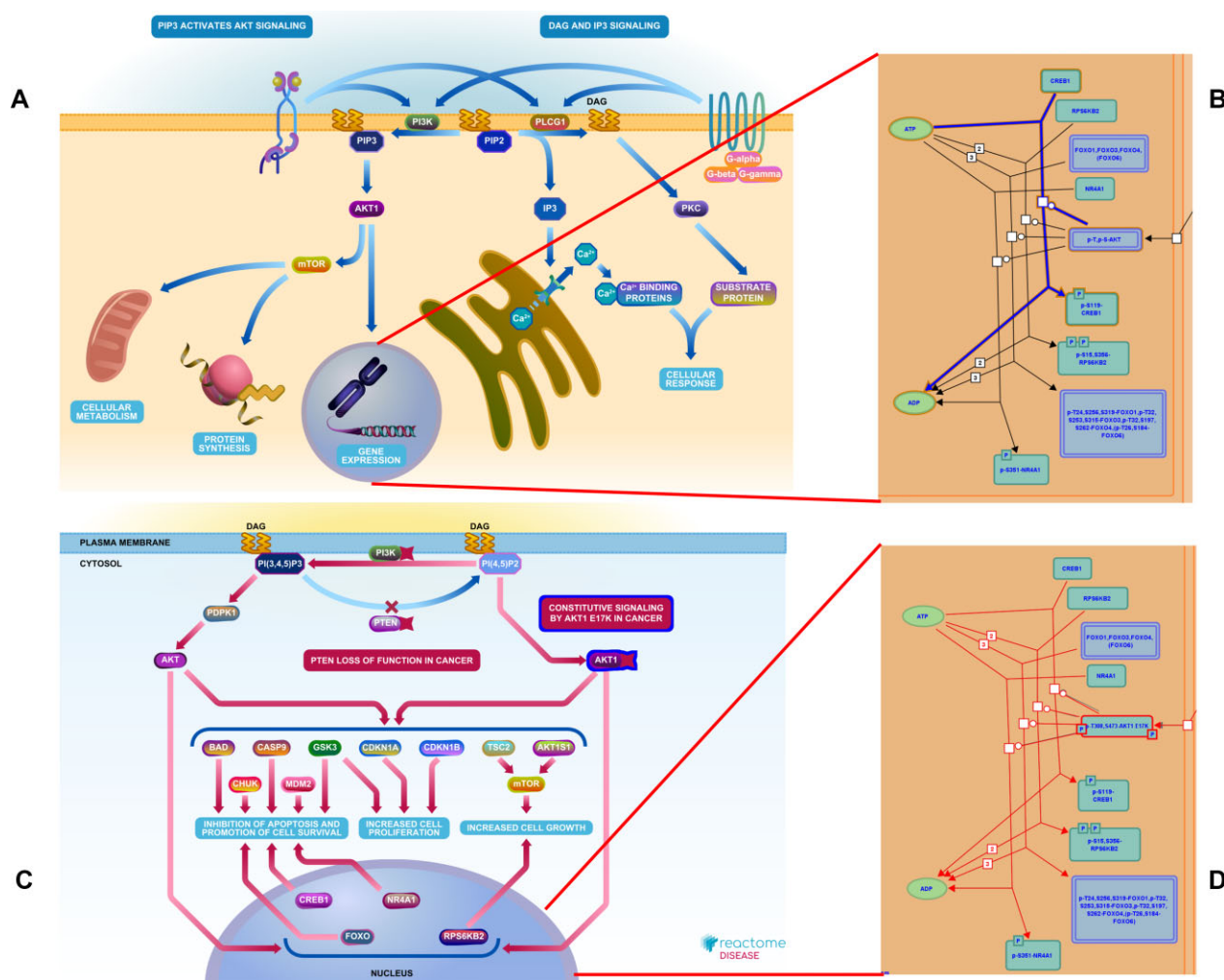


Figure 1. Disease variant-associated disease reactions in Reactome pathway diagrams. The pathway, 'Intracellular signaling by second messengers (R-HSA-9006925)' (A) includes a subpathway, 'AKT phosphorylates targets in the nucleus (R-HSA-198693)' (B) that shows normally regulated signaling mediated by AKT. The disease counterpart of the signaling pathway, 'PI3K/AKT Signaling in Cancer (R-HSA-2219528)' (C) includes a subpathway, 'Constitutive Signaling by AKT1 E17K in Cancer (R-HSA-5674400)' (D) that shows constitutive AKT signaling due to a missense mutation that enables AKT to bypass normal negative regulation and become constitutively active.

The Reactome curation process, however, cannot scale to encompass existing catalogs of pathogenic mutations and their continued rapid growth. Rather, our curation efforts have focussed on mining these catalogs to identify well-characterized 'type' variants, each associated with a qualitatively distinct molecular phenotype such as changed catalytic

activity, insensitivity to a normal negative regulator, sensitivity to a drug that has no effect on its normal counterpart, or gain of resistance to a drug that inhibits other variants. Our annotation of the pathway 'Signaling by ERBB2 in Cancer' (R-HSA-1227990), for example, illustrates the range of mechanisms by which normal ERBB2 signaling is disrupted and the

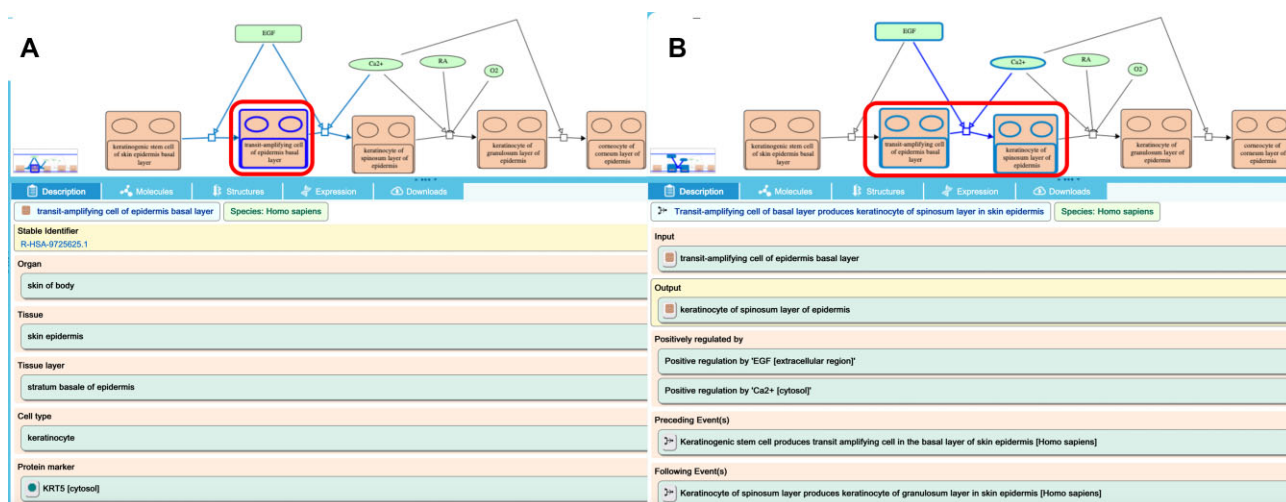


Figure 2. Annotation of cell- and tissue-specific reactions. A cell lineage path, ‘Development of keratinocytes in interfollicular skin epidermis’ (R-HSA-972554) is represented as four cell development steps, equivalent to reactions. Each step transforms an input cell, equivalent to a physical entity, to an output cell, regulated by physical entities such as epidermal growth factor (EGF) and calcium ions (Ca²⁺). **(A)** A cell icon, ‘transit-amplifying cell of epidermis basal layer’ (R-HSA-9725625), highlighted with a red box in the pathway diagram panel (above) is annotated with cell and tissue ontology terms to identify and distinguish it from other kinds of cells and with protein and RNA markers that are key to its function in the process annotated here, listed in the details panel below the diagram. **(B)** A cell development step, ‘Transit-amplifying cell of basal layer produces keratinocyte of spinosum layer in skin epidermis’ (R-HSA-9727354) highlighted with a red box in the pathway diagram panel (above) is annotated with regulators of the transformation and identities of preceding and following development steps in the details panel below the diagram.

types of small-molecule drugs that target these ERBB2 variants. In this way, Reactome enables a clinical investigator with a novel variant that can be classified based on the appropriate catalog, or perhaps by a computational tool such as AlphaMissense (16), to use Reactome to form specific, testable hypotheses for its likely effects at a pathway level and its drug susceptibility.

Drugs

In the same way that the Reactome data model allows the easy extension of annotation of covalent modifications of proteins to capture changes due to genetic variation, the model allows the effects of drugs, both small molecules and macromolecules such as RNAs and monoclonal antibodies, to be annotated in two-step reaction sequences. In the first, the drug binds its target human protein. In the second, the drug:protein complex regulates the reaction that the protein would otherwise mediate. Using this model, we have annotated, so far, the effects of 1033 chemical drugs and 86 protein drugs. As in the case of genetic variant proteins, the network organization of Reactome allows easy visualization of downstream effects of drugs. And also as in the case of genetic variants, scaling even this basic annotation process to encompass the vast and growing catalogs of drugs and their targets is impractical. To date, we have focused manual annotation on drugs and processes of immediate clinical interest, such as coronavirus infection (R-HSA-9679191) and fibrin clot formation in blood coagulation (R-HSA-140877). Also, data visualization features, such as automatic overlay of drug interactions on the top of pathway diagrams, have been implemented in the Reactome main web site and the Reactome IDG portal, assisting researchers to infer the potential impacts of drugs on pathways without manual curation.

Building on our annotations of drug interactions with their human protein targets, we have begun to annotate the entire

ADME (absorption, distribution, metabolism, excretion) life cycle of selected drugs, e.g. ribavirin (R-HSA-9755088), in collaboration with PharmGKB (17), potentially facilitating the visualization of interactions and off-target effects of drugs of interest.

Extensions of Reactome

Molecular annotations above the single-cell level

The Reactome project, at its inception, was envisioned as building a comprehensive parts list of all reactions enabled by human proteins, onto which a user could overlay a list of proteins expressed in a tissue of interest under physiological conditions of interest to infer a tissue- and condition-specific reaction network (7,8). As discussed above, experimental data needed to construct a full list is lacking, and computational tools are insufficient to fill gaps and infer tissue or cell type specific reaction networks.

To work around this limitation we have extended the event class of the Reactome data model and added new visualization features in the Reactome web site to allow explicit annotation of cell- and tissue-specific reactions. Annotation has focused initially on differentiation processes. In the same way that molecular events capture the transformation of input physical entities into output ones by means of molecular functions such as catalysis, transport, and binding, cell development steps are grouped into cell lineage paths. A new entity type, cell, allows annotation of a cell type with terms from cell and tissue ontologies (Cell Ontology (18), and Uberon (19), respectively) and with associated proteins and RNAs as markers that identify and individuate the cell (Figure 2A). Development steps have cell types as inputs and outputs and are regulated by small molecules such as calcium ions and proteins such as growth factors (Figure 2B). These molecular attributes and their functions link cell- and tissue-level annotations of differentiation processes to the subcellu-

lar molecular processes already systematically annotated in Reactome.

Releasing material with limited review

To accommodate difficulties in getting external reviews of new and revised material, we are allowing release of limited numbers of events without such review, tagged to indicate their potentially lower reliability. Material that has received full internal and external expert reviews (all released content before June 2023) has a 5-star review status rating. New material that has been curated and fully reviewed by a second curator but for which we have been unable to obtain an expert external review after six months is released with a 3-star rating. Previously fully-reviewed (5-star review status) events whose key attributes (participating entities for a reaction; participating reactions for a pathway) have been changed are re-released with a 4-star rating after independent internal review.

Three- and four-star events (<https://reactome.org/userguide/review-status>) were first released in June 2023 (V85). So far, this policy change has allowed the release of 122 finished events without external review (three-star), e.g. 'NFE2L2 regulating MDR associated enzymes', and of 12 previously released events that have been revised but not re-reviewed (four-star), e.g. 'SUMOylation of immune response proteins'.

Interoperability and resilience of data resources

Information resources like Reactome are most useful not in isolation but as parts of an integrated resource community. This kind of interoperability is a central feature of Reactome: reliable and widely-used reference resources provide the core information on which Reactome event annotations are assembled. For example, canonical forms of proteins are taken from UniProt (20), then annotated locally to describe changes from that canonical form, e.g. covalent or genetic modifications, and to add reaction-specific attributes such as subcellular localization. The controlled vocabulary of molecular functions, biological processes, and cellular components, and the logical relationships among these terms are taken from the Gene Ontology (9). Small molecules are from ChEBI (21) and the chemical equations for charge- and mass-balanced reactions occurring at physiological pH are from RHEA (1). As part of our own literature-based curation process, if we discover a discrepancy between information in one of these resources and new information we are annotating, we consult with the other resources to resolve the discrepancy. We also regularly check our annotations against these other resources to uncover and resolve discrepancies due to changes in the latter.

This approach to designing data resources has yielded a high level of interoperability among them. A user, viewing the UniProt entry for a protein or the ChEBI entry for a chemical, can navigate to the Reactome representations of reactions involving those entities. The first part of this dynamic, resource integration to promote interoperability, is codified in the FAIR (Findable, Accessible, Interoperable, Reusable - 22) and TRUST (Transparency, Responsibility, User community, Sustainability, Technology - https://datascience.nih.gov/sites/default/files/NIH_Workshop_on_Trustworthy_Data_Repositories_Report_7-8-2019%20FINAL.pdf) principles, and entities

like ELIXIR Core Data Resources (23) and the Global Core Biodata Coalition - <https://globalbiodata.org/> promote it. Reactome is a part of the ELIXIR infrastructure and is a Global Core Data Resource.

Interoperability, however, creates a high level of interdependence: if one resource cannot maintain its current data and add new information as it accrues, or if in response to availability of new data types and user needs a resource changes the range (scope) or degree of specificity (granularity) of its annotations, the resulting gaps and discrepancies propagate to all of the resources. In the past two years, for example, 1990 (4.5%) of the 43300 terms in the Gene Ontology have been changed by additions, obsoletions, or merges (9). And UniProt, enforcing the standard that there should be one canonical protein entry for each gene product, has collapsed their previous collection of hundreds of HLA-A, B, C and D genes into four, each with an extraordinary number of polymorphic alleles on the basis of extensive resequencing of this region of the human genome (24). Management of such changes in all of the interdependent resources is essential to maintain both the quality of individual resources and their interoperability. But this requires continuing, skilled (expensive) manual work and can have unpredictable downstream effects on usage of annotated data for studies such as gene overexpression analyses. These are hard problems but collaborative work aimed at aligning Reactome process-description pathways content systematically with GO activity-flow models (25,26) and at coordinating annotation of Reactome human pathways with corresponding mouse ones to develop a resource to analyze disease phenotypes between the two species (27) promises to provide good solutions to them.

Conclusions

The Reactome Knowledgebase of the molecular details of human biological processes continues to grow in size and scope. Since the last NAR update, Reactome has extended its annotations of genetic variation associated with disease, and of small-molecule drugs, including ones that target the protein products of genetic variants. Software developments include a protocol to expedite release of selected new and updated information without full external review, and a framework for cell type- and tissue type-specific annotations. Finally, we consider how to best maintain and improve Reactome's interoperability with other resources such as GO and UniProt in accord with FAIR and TRUST principles.

Data availability

Reactome is open-source and open-access. All Reactome data are available in various formats from our downloads page (<https://reactome.org/download-data>). A history tool is under development to enable users to track changes in our data over time. All software is available from our GitHub repositories (<https://github.com/reactome> and <https://github.com/reactome-pwp>), under terms that allow for free reuse and redistribution.

We have created Zenodo packages for the versions of our software and data discussed in the article:

- Pathway Browser: 10.5281/zenodo.10022792 (Frontend for the PathwayBrowser)

- Data-content: 10.5281/zenodo.10022911 (Frontend for the search, detail, schema and icon library pages)
- CuratorTool: 10.5281/zenodo.10022856 (Local software to create reactome data)
- Content-Service: 10.5281/zenodo.10022866 (Backend API to access content of Reactome)
- Analysis-Service: 10.5281/zenodo.10022873 (Backend API to perform analysis on Reactome)
- Reactome Data 86: 10.5281/zenodo.10018440 (Data dump of version 86, both Neo4j and MySQL database)

Acknowledgements

We are grateful to the 820 expert scientists who have collaborated with us as external authors and reviewers of Reactome content since 2002. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

The development of Reactome is supported by grants from the National Institutes of Health [U41HG003751, U24HG012198, U24HG011851, U54GM114833, U01CA239069]; European Bioinformatics Institute (EMBL-EBI), Open Targets (The Target Validation Platform) and Medicine by Design (University of Toronto). Funding for open access charge: National Institutes of Health [U24HG012198].

Conflict of interest statement

None declared.

References

1. Bansal,P, Morgat,A., Axelsen,K.B., Muthukrishnan,V., Coudert,E., Aimo,L., Hyka-Nouspikel,N., Gasteiger,E., Kerhornou,A., Neto,T.B., *et al.* (2022) Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.*, 50, D693–D700.
2. Kanehisa,M., Furumichi,M., Sato,Y., Kawashima,M. and Ishiguro-Watanabe,M. (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*, 51, D587–D592.
3. Caspi,R., Billington,R., Keseler,I.M., Kothari,A., Krummenacker,M., Midford,P.E., Ong,W.K., Paley,S., Subhraveti,P. and Karp,P.D. (2020) The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.*, 48, D445–D453.
4. Gillespie,M., Jassal,B., Stephan,R., Milacic,M., Rothfels,K., Senff-Ribeiro,A., Griss,J., Sevilla,C., Matthews,L., Gong,C., *et al.* (2022) The Reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, 50, D687–D692.
5. Le Novère,N., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A., Demir,E., Demir,K., Aladjem,M.I., Wimalaratne,S.M., *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, 27, 735–741.
6. Kibbe,W.A., Arze,C., Felix,V., Mitraka,E., Bolton,E., Fu,G., Mungall,C.J., Binder,J.X., Malone,J., Vasant,D., *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, 43, D1071–D1078.
7. Joshi-Tope,G., Vastrik,I., Gopinath,G.R., Matthews,L., Schmidt,E., Gillespie,M., D'Eustachio,P., Jassal,B., Lewis,S., Wu,G., *et al.* (2003) The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harb. Symp. Quant. Biol.*, 68, 237–243.
8. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G., Matthews,L., *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33, D428–D432.
9. Gene Ontology Consortium (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, 224, iyad031.
10. Oprea,T.I., Bologa,C.G., Brunak,S., Campbell,A., Gan,G.N., Gaulton,A., Gomez,S.M., Guha,R., Hersey,A., Holmes,J., *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.*, 17, 317–332.
11. Wu,G. and Haw,R. (2017) Functional interaction network construction and analysis for disease discovery. *Methods Mol. Biol.*, 1558, 235–253.
12. Beavers,D., Brunson,T., Sanati,N., Matthews,L., Haw,R., Shorser,S., Sevilla,C., Viteri,G., Conley,P., Rothfels,K., *et al.* (2023) Illuminate the Functions of Dark Proteins Using the Reactome-IDG Web Portal. *Curr Protoc*, 3, e845.
13. Brunson,T., Sanati,N., Matthews,L., Haw,R., Beavers,D., Shorser,S., Sevilla,C., Viteri,G., Conley,P., Rothfels,K., *et al.* (2023) Illuminating dark proteins using Reactome pathways. bioRxiv doi: <https://doi.org/10.1101/2023.06.05.543335>, 5 June 2023, preprint: not peer-reviewed.
14. 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.
15. 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.
16. Cheng,J., Novati,G., Pan,J., Bycroft,C., Žemgulytė,A., Applebaum,T., Pritzel,A., Wong,L.H., Zielinski,M., Sargeant,T., *et al.* (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381, eadg7492.
17. Gong,L., Whirl-Carrillo,M. and Klein,T.E. (2021) PharmGKB, an Integrated Resource of Pharmacogenomic Knowledge. *Curr Protoc*, 1, e226.
18. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntinivajai,S., *et al.* (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*, 7, 44.
19. Haendel,M.A., Balhoff,J.P., Bastian,F.B., Blackburn,D.C., Blake,J.A., Bradford,Y., Comte,A., Dahdul,W.M., Dececchi,T.A., Druzinsky,R.E., *et al.* (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semantics*, 5, 21.
20. UniProt Consortium. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, 51, D523–D531.
21. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, 44, D1214–D1219.
22. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data.*, 3, 160018.
23. Durinx,C., McEntyre,J., Appel,R., Apweiler,R., Barlow,M., Blomberg,N., Cook,C., Gasteiger,E., Kim,J.H., Lopez,R., *et al.* (2016) Identifying ELIXIR Core Data Resources. *F1000Res*, 5, ELIXIR-2422.
24. Robinson,J., Guethlein,L.A., Cereb,N., Yang,S.Y., Norman,P.J., Marsh,S.G.E. and Parham,P. (2017) Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.*, 13, e1006862.
25. Thomas,P.D., Hill,D.P., Mi,H., Osumi-Sutherland,D., Van Auken,K., Carbon,S., Balhoff,J.P., Albou,L.P., Good,B., Gaudet,P., *et al.* (2019) Gene Ontology Causal Activity Modeling (GO-CAM)

- moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.*, **51**, 1429–1433.
26. Good,B.M., Van Auken,K., Hill,D.P., Mi,H., Carbon,S., Balhoff,J.P., Albou,L.P., Thomas,P.D., Mungall,C.J., Blake,J.A., *et al.* (2021) Reactome and the Gene Ontology: digital convergence of data resources. *Bioinformatics*, **37**, 3343–3348.
27. Hill,D.P., Drabkin,H.J., Smith,C.L., Van Auken,K.M. and D'Eustachio,P. (2023) Biochemical pathways represented by Gene Ontology-Causal Activity Models identify distinct phenotypes resulting from mutations in pathways. *Genetics*, **225**, iyad152.