Complex Portal 2022: new curation frontiers

Birgit H M Meldal, Livia Perfetto, Colin Combe, Tiago Lubiana, João Vitor Ferreira cavalcante, Hema Bye-A-Jee, Andra Waagmeester, Noemi Del-Toro, Anjali Shrivastava, Elisabeth Barrera, et al.

Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab991

Complex Portal 2022: new curation frontiers

Birgit H. M. Meldal [1], Livia Perfetto[1,2], Colin Combe[3], Tiago Lubiana[4], João Vitor Ferreira Cavalcante[5], Hema Bye-A-Jee[1], Andra Waagmeester [6], Noemi del-Toro [1], Anjali Shrivastava[1], Elisabeth Barrera[1], Edith Wong[7], Bernhard Mlecnik [8,9,10,11], Gabriela Bindea [8,9,10], Kalpana Panneerselvam[1], Egon Willighagen [12], Juri Rappsilber[3,13], Pablo Porras [1], Henning Hermjakob [1,*] and Sandra Orchard [1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Fondazione Human Technopole, 20157 Milan, Italy, [3]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK, [4]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, Av. Professor Lineu Prestes 580, CEP 05508-000 São Paulo SP, Brasil, [5]Bioinformatics Multidisciplinary Environment (BioME), Digital Metropolis Institute, Federal University of Rio Grande do Norte, Av. Odilon Gomes de Lima 1722, Capim Macio, 59078-400 Natal/RN, Brasil, [6]Micelio, Veltwijcklaan 305, 2180 Ekeren, Belgium, [7]Department of Genetics, School of Medicine, Stanford University, Palo Alto, CA, USA, [8]Laboratory of Integrative Cancer Immunology, INSERM, 75006 Paris, France, [9]Equipe Labellisée Ligue Contre le Cancer, 75006 Paris, France, [10]Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris, 75006 Paris, France, [11]Inovarion, 75005 Paris, France, [12]Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands and [13]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

Received September 15, 2021; Revised October 07, 2021; Editorial Decision October 08, 2021; Accepted October 10, 2021

ABSTRACT

The Complex Portal (www.ebi.ac.uk/complexportal) is a manually curated, encyclopaedic database of macromolecular complexes with known function from a range of model organisms. It summarizes complex composition, topology and function along with links to a large range of domain-specific resources (i.e. wwPDB, EMDB and Reactome). Since the last update in 2019, we have produced a first draft complexome for *Escherichia coli*, maintained and updated that of *Saccharomyces cerevisiae*, added over 40 coronavirus complexes and increased the human complexome to over 1100 complexes that include approximately 200 complexes that act as targets for viral proteins or are part of the immune system. The display of protein features in ComplexViewer has been improved and the participant table is now colour-coordinated with the nodes in ComplexViewer. Community collaboration has expanded, for example by contributing to an analysis of putative transcription cofactors and providing data accessible to semantic web tools through Wikidata which is now populated with manually curated Complex Portal content through a new bot. Our data license is now CC0 to encourage data reuse. Users are encouraged to get in touch, provide us with feedback and send curation requests through the 'Support' link.

INTRODUCTION

Protein complexes, stable functional assemblies consisting of two or more associated polypeptide chains, are responsible for driving and regulating many cellular processes. Multi-chain assemblies perform many functions, including (a) positioning molecules involved in the same process in close proximity (b) bringing structure to disordered regions of proteins and (c) creating novel substrate binding sites at subunit interfaces. These assemblies can contain additional molecules, such as nucleic acids and small molecules. In budding yeast (1), around one in three proteins have a function in stable heteromeric complexes and in bacteria around one in five (see below).

Although the existence of many well-studied protein complexes has been recognized for decades, existing manually curated, species-specific catalogues were either not regularly maintained, e.g. CYGD (2) for yeast, or entries were curated based on individual papers rather than amalgamated knowledge, e.g. CORUM (3) for mammalian species. The Complex Portal (www.ebi.ac.uk/complexportal) was created to meet this unmet need: It is a manually curated,

*To whom correspondence should be addressed. Tel: +44 1223494671; Email: hhe@ebi.ac.uk
Correspondence may also be addressed to Sandra Orchard. Email: orchard@ebi.ac.uk

encyclopaedic resource that provides stable identifiers and summarizes compositional, topological and functional aspects of stable macromolecular complexes from a selection of model organisms and organisms of special interest. It enables protein complex identification in large-scale data analyses, contributing to the study of complex evolution and increasing our understanding of cell biology (4–8).

Since the last update (9), the coverage of model organism complexomes, the compendium of known complexes for a given species, has increased significantly, with that of *Saccharomyces cerevisiae* being maintained and added to as more experimental data leads to the identification of new assemblies (1) and the completion of a first draft of the *Escherichia coli* complexome. Work is now focused on annotating human complexes, wherever possible in collaboration with other data resources or scientific groups. We have also responded to the ongoing SARS-CoV-2 pandemic by creating the complexome of this organism and also of related viruses, in order to contribute to global efforts to respond to this threat.

We have improved ComplexViewer so that it can display multiple features simultaneously and show links between participants of sub-complexes and other complex participants. We updated the participant table so that all participants are now colour-coordinated between ComplexViewer and the table (Figure 1). We have made updates to the ComplexTab format, adding a 'UniProt ID-only' column that lists the UniProt accession numbers (and their stoichiometry) for the protein participants of complexes, including those that are part of subcomplexes and molecule sets.

We have expanded our community collaboration by contributing to a large-scale transcription cofactor analysis, by working with Wikidata contributors who have written a bot that populates Wikidata with Complex Portal content for semantic web reuse and by collaborating with the Cytoscape ClueGO App developers to incorporate Complex Portal entries as a new ontology for complex enrichment analyses. Finally, we changed our content license to CC0 to improve data accessibility and re-use.

CONTENT

Curation update

A protein complex is a functional, biological entity that contains two or more macromolecules (proteins, nucleic acids or small molecules) for which there is experimental or inferred evidence that these molecules stably interact with each other. As of release 241 (18 October 2021), 3572 complexes from 26 species have been curated and released. Each entry is species-specific, describing the complex composition, topology and function and linking out to external databases that provide further domain-specific information, such as structural details from wwPDB (10) or EMDB (11) or the role of the complex in metabolic reactions or signalling pathways in Reactome (12). Complex components are linked to primary reference resources; UniProt (13,14) for proteins, ChEBI (14) for small molecules and RNAcentral (15) for noncoding RNAs. Complexes that are also participants of larger complexes are linked to their own Complex Portal entries. In principle, we create separate entries for each compositional variant of a complex. However, some complexes, mainly the ribosomes, contain many participants that are potentially coded by two alternative, paralogous genes. In these cases, we create molecule sets, identified by identifiers of type EBI-{1–9}, containing the UniProt IDs of each of the paralogous proteins. On the website they appear as unlinked concatenations of gene symbols and species names, e.g. 'rps4a_rps4b_yeast'.

Since our last update three years ago, we have focused on a number of curation priorities:

Escherichia coli complexome. In December 2019, we released the first version of the *E. coli* K12 (NCBI reference taxon: 83333) complexome currently consisting of 321 complexes. This work was based on extensive literature mining and subsequent comparison with existing resources, in this case primarily UniProt KB and EcoCyc (16). This systematically annotated set of *E. coli* complexes includes, at the time of writing, 786 unique proteins (18% of the proteome). 95% (305) of complexes contain 5 or fewer proteins (median = 3) (Figure 2). 87% unique proteins (681/786) are found in only one complex and 9% protein (74/786) in two complexes with the remaining 31 proteins found in more than two complexes. This distribution is similar to what we saw for yeast (1) except that fewer proteins in general were found in heteromeric complexes (18% versus 32%).

As with the recently completed *S. cerevisiae* complexome, a watchlist of additional potential candidate complexes exists and these are being added to the dataset on an ongoing basis, if and when they are experimentally verified. Also, it must be recognised that *E. coli* K12 is a non-pathogenic laboratory strain and many proteins are cryptic or have been engineered out of the strain. Complexes containing such proteins are therefore absent from this model organism. For example, *E. coli* K12 does not express the PhnE permease due to the presence of an 8 bp insertion in *phnE* (17), therefore the phosphonate ABC transporter complex is not formed. To capture complexes that contain proteins not present in the K12 strain but that are essential for fully understanding the life-cycle of wild-type strains of these Gram-negative bacteria, the complex and its protein components are mapped to the species level for *E. coli*, NCBI taxon ID:562. Examples include the phosphonate ABC transporter complex (CPX-4382) and heat-labile enterotoxin IIB complex (CPX-2304).

Coronavirus complexes. In response to the COVID-19 pandemic we have curated coronavirus complexes as well as human targets of viral proteins (18). To date, we have released 21 SARS-CoV-2 complexes (taxon ID: 2697049), 16 SARS-CoV complexes (taxon ID: 694009) and 17 MERs-CoV complexes (taxon IDs: 1263720 and 1235996 [no reference proteome available]). They include mixed-species complexes of the viral Spike proteins with their host receptors ACE2 (Figure 1) and DPP4, respectively. As new experimental evidence emerges frequently for these complexes we have made use of our versioning protocol and updated complex components, stoichiometry and function. For example, new evidence from crystal and SAXS analyses led to an update of the stoichiometry of the nsp7-nsp8 primase (CPX-5690) from 8:8 to 2:2 while the function of the nsp10-nsp14 complex (CPX-5692) was updated from

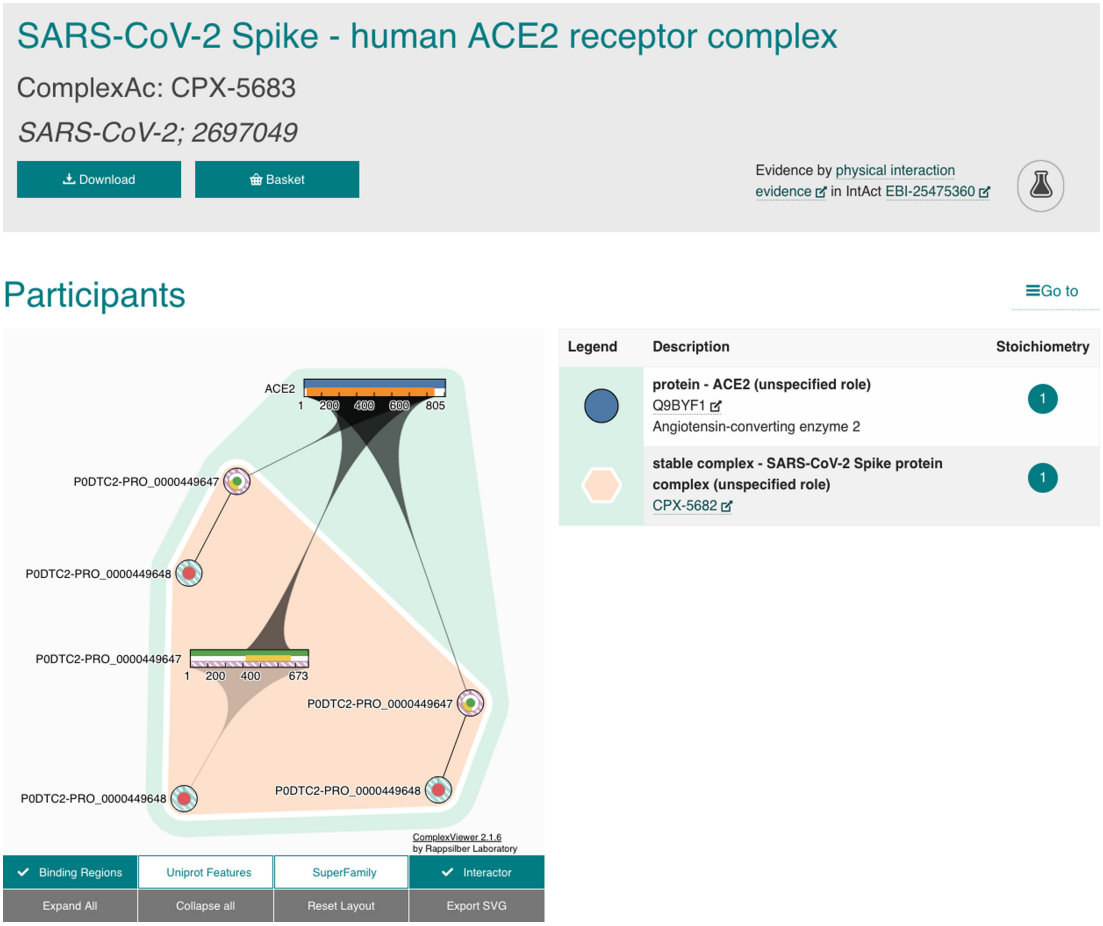

Figure 1. Top of the details page for SARS-CoV-2 Spike–human ACE2 complex displaying some of the new features available in the ComplexViewer (left side) and Participant table (right side).

a guanine-N7 methyltransferase to a 3′-5′ exoribonuclease (GO:0000175). We continue to add new interaction evidence from IntAct (19) and structural evidence from wwPDB (20) and EMDB (21) at each release. These complexes are already used as cross-references by several external databases such as the IMEx resources (22), Gene Ontology (23,24), MatrixDB (25), Reactome, SIGNOR (26), UniProtKB and WikiPathways (27) thus allowing for a more integratable set of coronavirus-related information to be freely available to the research community.

Human complexes. Efforts are now focused on expanding the collection of manually curated human complexes. Exactly how many assemblies comprise the human complexome is a question still very much open to debate. We anticipate that most, but not all, of the intracellular complexes we have identified in *S. cerevisiae* are conserved in multicellular organisms, but human complexes often contain additional protein components or the existence of paralogous proteins lead to increased numbers of complex variants. One simple example of the latter case is the replication protein A complex, a single heterotrimer in yeast (CPX-21), but in humans one protein (P15927/Q13156) has been duplicated resulting in two heterotrimeric complex variants (CPX-1878/CPX-1879). Additionally, there will be an appreciable number of transmembrane and extra-cellular complexes with a role in the immune response, inter- and intracellular communication and signalling. A data mining exercise, undertaken in collaboration with the UniProt group, of information embedded in UniProtKB records has suggested that there may be at least 4000 different assemblies in the human complexome. To date, we have released 1255 human complexes including almost 200 that are either targets of viral proteins or play a role in the immune system, including the B-cell and T-cell receptor complexes, the interferon-receptor family of complexes and the complete complement cascade of the innate immune system. We are also focusing on dimeric transcription factor complexes, complementing a recent revision of the human transcription factor proteins undertaken by members of the Gene Ontology and Gene Regulation Consortia (28).

Linking chemistry to biology: enhancing enzyme annotation. Many enzymes are found in multi-chain complexes, which may serve to bring together multiple enzymes associated with a specific metabolic pathway, bring regulatory subunits in close proximity to the catalytic chain, enable the coordination of ligand binding, or create new binding sites in subunit interfaces. In order to improve our annotation of these assemblies, we are now adding cross-references to the Rhea

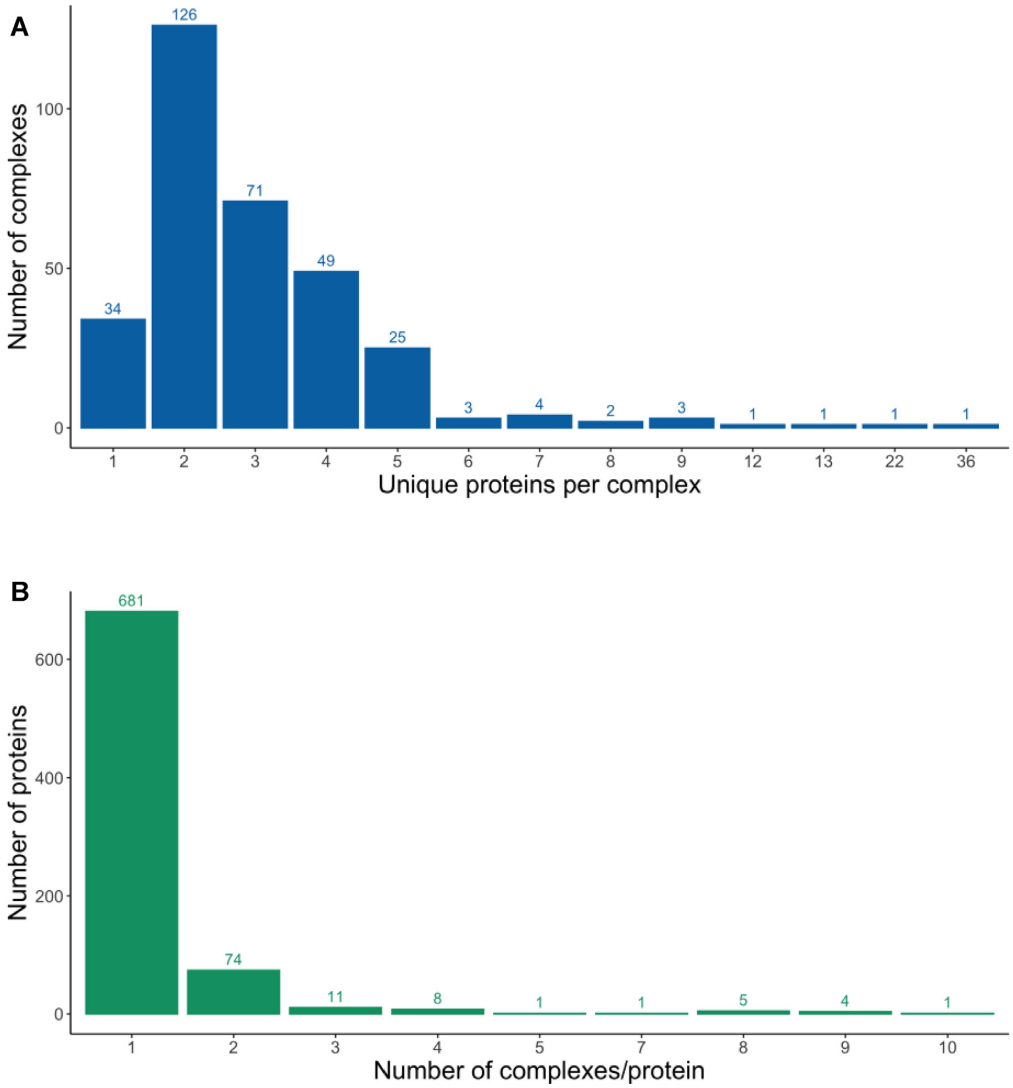

Figure 2. Most *E. coli* proteins are found in only one or two different complexes (**A**) and most *E. coli* complexes contain five or fewer unique proteins (**B**).

knowledgebase of biochemical reactions (29) where it has been demonstrated that the enzyme has this activity in the context of a given complex. Rhea uses the chemical ontology ChEBI to describe reaction participants, their chemical structures and chemical transformations and enables a more granular description of the enzyme/substrate interaction. For example, human serine palmitoyltransferase exists as four variant complexes (CPX-6663, CPX-6664, CPX-6665, CPX-6681) which share a common E.C. number (2.3.1.50). However, the substrate specificities of these complexes vary, due to the positioning of side chain residues of differing amino acids in the alternate complex components, and this can be fully described by the addition of the appropriate Rhea cross-references. Rhea cross-references are currently available in download files, and will be visualised on the website in the near future.

Enzymes are not the only complex components which bind small molecules. Rhea encompasses transporter activity within its concept of a reaction, and cross-references are increasingly being added to complexes involved in transportation. The ligands that are bound by receptors are also systematically captured as a 'ligand' annotation topic and are described by both a recommended, human-readable protein, peptide or small molecule name and a UniProtKB or ChEBI accession number, as appropriate.

Defining curation practice for obsoleting complexes and versioning. When the research community's understanding of a complex's existence changes to the extent that we need to delete an existing entry, or we decide to merge a sub-complex into a larger assembly, the original entries remain available in previous release files accessible via our ftp repository (http://ftp.ebi.ac.uk/pub/databases/intact/complex/). If a complex has been merged into an existing entry, the accession number of the obsoleted complex is added to the complex it has been merged into as a secondary identifier, allowing an external user to still retrieve an entry. Secondary identifiers are available in all download files, while the website currently only displays the primary accession number (e.g. CPX-3042 is now part of

CPX-2161). This enables us to adhere to the FAIR principle of data Findability. More minor updates, for example the identification of an additional participant, are indicated by entry versioning, as previously described (1,9).

WEBSITE SEARCH AND DATA VISUALIZATION

ComplexViewer and participant table

We have made a number of improvements to the ComplexViewer (3) (Figure 1):

- *Zooming into a protein sequence*: clicking on a protein icon expands it into a short sequence bar. Clicking into the bar brings up a pop-up menu from which the sequence can be expanded further to four different zoom levels or collapsed again into the original circular icon. This feature is now available on touchscreens.
- *Subcomplexes*: if a complex has another complex as a participant, all proteins are now displayed and grouped by subcomplex membership using differently coloured backgrounds for each subcomplex (e.g. CPX-1556).
- *Binding features in subcomplexes:* binding features are displayed between any participants in a complex, which now includes participants which are part of a subcomplex and bind another molecule outside the subcomplex, e.g. the binding of human ACE2 receptor to the SARS-CoV-2 Spike S1 protein chain (CPX-5683) (Figure 1).
- *Undefined binding regions:* are now represented as full length features of the participant and the range is hatched, while defined ranges are filled solidly (e.g. CPX-2158).
- *Multiple feature types and ranges:* are displayed as separate tracks in the bar representation of a protein and separate wedges or circles within the circular icon (e.g. CPX-1919 or CPX-1003). Different feature types can be turned on and off using a new set of buttons below the viewer window.

Additionally, the participant table is now colour-coordinated, matching the node colours of proteins or background colours of complexes as participants.

Website updates

We have refreshed and updated the Home, About and Documentation pages, including adding more information about our curation practices, handling of edge cases and documentation about our file formats. These new pages are also linked to our GitHub repository which allows us to make any updates easier and quicker.

DATA ACCESS OPTIONS

ComplexTab

Complexes may contain molecules other than proteins, such as nucleic acids or small molecules, but some users, for example mass spectrometry proteomics scientists, are only interested in the protein components. Protein complexes can be subcomponents of larger assemblies, and users previously needed to parse the files separately to retrieve the participant of these subcomplexes. In response to requests from this user community, we have added a new column to the tab-delimited format, ComplexTab, that contains a list of UniProt accession numbers (including isoform or post-processes chain extensions) in pipe-separated style with stoichiometry in parentheses.

Data licencing

To ensure our data is available for reuse by all interested parties, and in line with EMBL-EBI policy, our licence has been updated to Creative Commons Public Domain (CC0) License (https://creativecommons.org/publicdomain/zero/1.0/). This applies to all Complex Portal data, i.e. PSI-MI XML3.0 (30), MI-JSON and ComplexTab files (9), as well as data directly accessed via web pages/services.

COLLABORATION AND COMMUNITY INVOLVEMENT

Identification of putative transcription cofactor complexes through the Gene Regulation Ensemble Effort for the Knowledge Commons (GREEKC) collaboration

The GREEKc collaboration was an EC-funded COST action aimed at integrating data and knowledge pertaining to gene regulation, of which the Complex Portal was an active participant throughout. As part of this effort, Velthuijs *et al.* (31) have used the Complex Portal as a verification dataset in their analysis of potential transcription cofactors. By combining inferred complexes from hu.MAP 2.0 (32), curated complexes from CORUM (33) and curated physical interaction data from IntAct [this volume NAR paper] and BioGrid (34) with a selected set of transcription-related Gene Ontology terms we have identified more than 1500 putative transcription cofactors. 415 of these are already participants of complexes in Complex Portal, and the remaining proteins will be curated into Complex Portal if they are identified as components of complexes.

Integration of complexes into Wikidata

Wikidata (https://www.wikidata.org/) is part of the infrastructure provided by the Wikimedia Foundation and a sister project of Wikipedia; it initially provided a semantic web infrastructure for encyclopedic knowledge to be used in Wikipedia (35), but has since gained traction as a generic linked open resource; over the past years a number of life sciences resources have been aligned to Wikidata, including subsets of UniProt, the Gene Ontology and ChEBI.

After we released the first eleven SARS-CoV-2 complexes in April 2020 we joined the COVID19 Virtual Elixir BioHackathon 2020 (https://github.com/virtual-biohackathons/covid-19-bh20/wiki). During this hackathon we added Wikidata entities for these eleven complexes using a semi-automated curation pipeline using OpenRefine (https://openrefine.org/) and the Wikidata Integrator Python module (36). Through this integration into Wikidata, Complex Portal identifiers (https://www.wikidata.org/wiki/Property:P7718) for SARS-CoV-2 complexes as well as selected human complexes, which had already been manually added to Wikidata, were immediately used in the WikiPathways COVID-19 Pathways Collection (http://covid.wikipathways.org/)

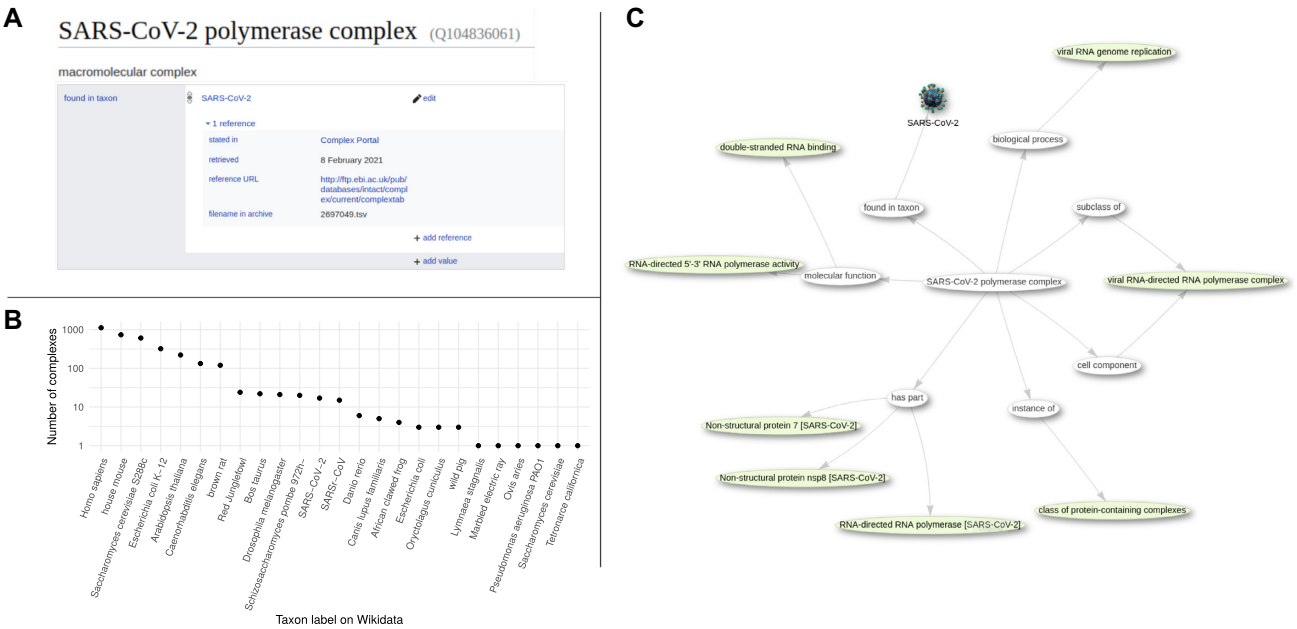

Figure 3. Complex Portal and Wikidata. (**A**) example of an entry assertion in Wikidata with provenance pointing to Complex Portal (Q104836061). (**B**) Number of protein complexes in Wikidata per taxon (https://w.wiki/3ggX). (**C**) Subset of Wikidata connected to the SARS-CoV-2 polymerase complex (https://w.wiki/3eta)

(36) which is part of the COVID-19 Disease Map project (https://covid19map.elixir-luxembourg.org/minerva/) (37) that also includes the Reactome COVID-19 project. As an extension to this initial SARS-CoV-2-focused collaboration we have subsequently developed a Wikidata Complex Bot, reconciling the entries in Complex Portal to Wikidata. The Complex Bot parses the Complex Portal releases and enriches the Wikidata environment, connecting proteins by their common presence in complexes, linking the new entries to existing entries and matching GO complexes with their Wikidata IDs. As a result, Complex Portal data is now available on the semantic web and updated regularly in line with our regular data releases and integrated with related UniProt, GO and ChEBI data in semantic web format (Figure 3). This enables rapid integration with other linked-data sources using Wikidata as a proxy. To be able to run bots on Wikidata a bot-flag was requested (https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/ComplexPortalBot) and approved after an online vetting process, where the rationale and the software was assessed. The open source code of the bot can be found at https://github.com/lubianat/complex_bot. Currently, the bot runs manually; after each Complex Portal update a request is sent to the bot developer to run the bot on the recent update. We are working towards an automatic integration of the two processes, similar to the continuous integration applied elsewhere (36). Another use of the Complex Portal data in Wikidata is the visualisation of complex information with Scholia (38). For example, https://scholia.toolforge.org/complex/Q104836061 shows information for the SARS-CoV-2 polymerase complex (CPX-5742). Similarly, protein pages show in which complexes they participate (e.g. https://scholia.toolforge.org/protein/Q90038963 for SARS-CoV-2 NSP7).

Additional usage and collaborations

Complex Portal identifiers are already being used as the preferred identifiers for complex entities in, among others, IMEx, Gene Ontology, *Saccharomyces* Genome Database (39) and SIGNOR curation efforts and more recently for causal interaction curation efforts based on MI2CAST curation guidelines (40), WikiPathways (27) annotations (via the BridgeDb (41) mapping service) and the COVID-19 Disease Map project (37). The use of Complex Portal identifiers in WikiPathways was enabled by manually creating complex entities in Wikidata with cross-references to Complex Portal, an effort that preceded and initiated our Wikidata bot development. An ongoing collaboration with the UniProt team will drive further work on the human complexome and an enhanced import of data from the Complex Portal into UniProt records is under active discussion.

Complex Portal data, together with molecular interaction data from IntAct, Reactome and SIGNOR, is integrated into the Open Targets (42) partnership that uses human genetics and genomics data for systematic drug target identification and prioritisation, via our bespoke graph database. Additionally, we provide a JSON file from the same graph database, containing protein-to-complex mappings.

Functional analysis through ClueGO

The ClueGO App (43) is a powerful platform for functional enrichment analysis within Cytoscape (44). Bespoke Complex Portal ontology files have been created for a selected number of species with the Complex Portal complexes being

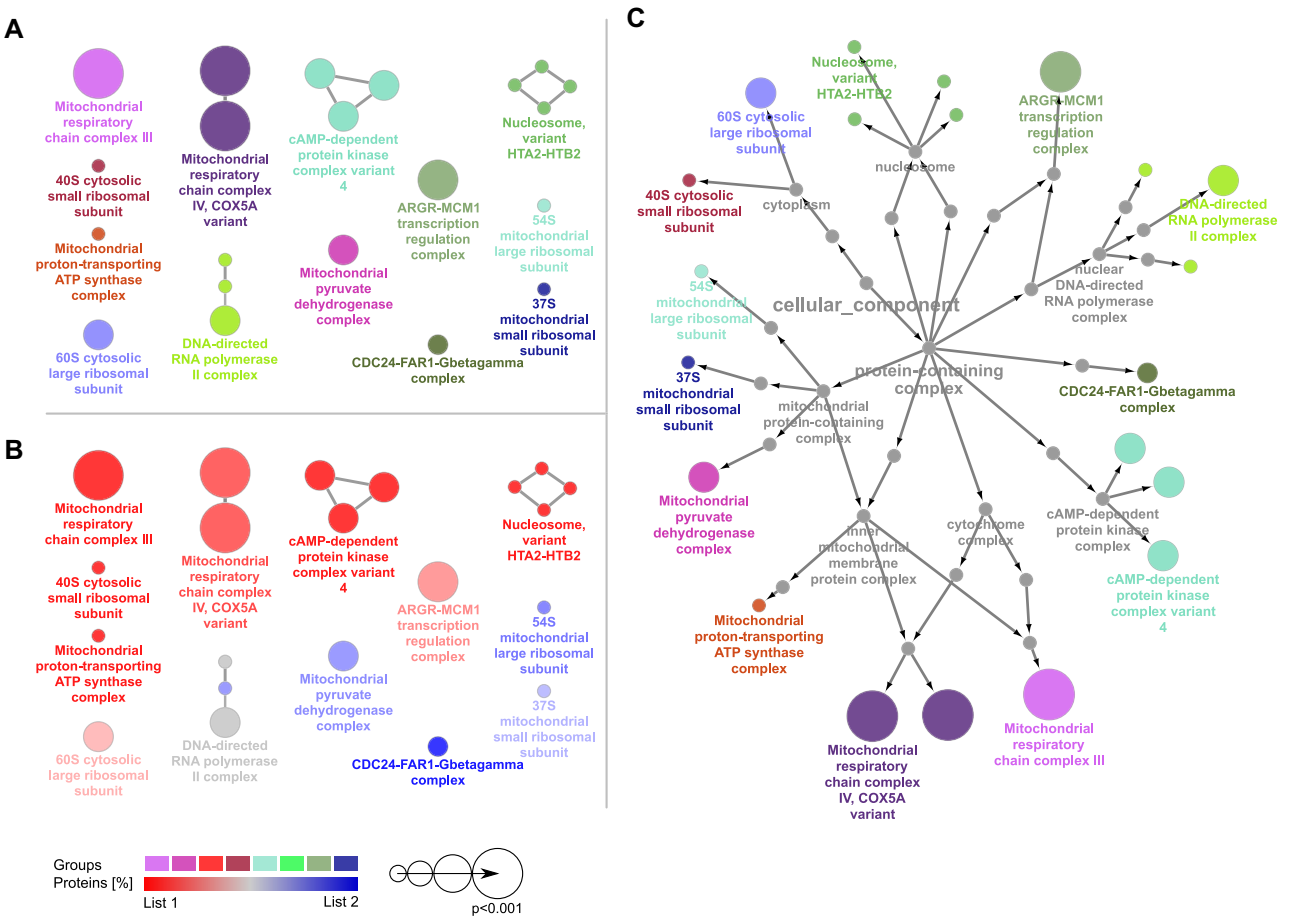

Figure 4. Output of an enrichment analysis for Complex Portal complexes of *S. cerevisiae* proteins using ClueGO. Two different lists of proteins were compared. (**A**) functionally grouped complexes, (**B**) distribution (%) of the proteins from list 1 (red) and 2 (blue) within each of the complexes from (A), (**C**) relationship of complexes based on the Cellular Component part of the Gene Ontology Tree.

leaf nodes of the Gene Ontology Cellular Component class. These new ontology files are available starting with ClueGO version 2.5.8 and allow users to conduct enrichment analyses for complex composition, see Figure 4. We will further collaborate with the ClueGO team to extend the Complex Portal Ontology, and to create new visualizations for Complex Portal data.

SUMMARY AND FUTURE PLANS

With the completion of the draft complexomes for *S. cerevisiae* and *E. coli* we are now fully focusing on completing a first draft of the human complexome. We are looking to extend our collaborations with other resources to increase our coverage of other model organisms, a paradigm successfully initiated with *Saccharomyces cerevisiae* (39). We are currently focusing on immune system complexes through collaborations with WikiPathways, the COVID-19 Disease Map project and the Cellxgene initiative (45).

We are developing an import pipeline for heteromeric structures from PDBe which will speed up the manual part of the curation process by populating all standardized, structured fields directly from the PDB files. We are also working with the group of Colin Logie (https://molbiology.science.ru.nl/about/molecular-biology/colin-logie/) who is studying the relation between chromatin structure and transcription and is providing extensive lists of important curation targets. The group is developing a pipeline to identify additional potential protein complexes with a role in cotranscription which will be further evaluated by manual curation in the Complex Portal.

We actively encourage curation requests and user feedback which will improve our databases and services. Please contact the Molecular Interaction Team via our support page at https://www.ebi.ac.uk/support/complexportal. Information about curation is provided at https://www.ebi.ac.uk/complexportal/documentation. Extensive training material on how to best use our resource is available at https://bit.ly/Complex-Portal-training.

DATA AVAILABILITY

The Complex Portal is a community project. Developers can contribute to the code at https://github.com/Complex-Portal/complex-portal-view.

Data can be accessed either via our ftp site (ftp.ebi.ac.uk/pub/databases/intact/complex/current/) or our REST API (https://www.ebi.ac.uk/intact/complex-ws/).

Lists of putative complexes on our 'watch lists' are available on request (https://www.ebi.ac.uk/support/complexportal).

Wikidata Integrator python module: https://github.com/SuLab/WikidataIntegrator.

Wikidata bot: https://github.com/lubianat/complex_bot.

ACKNOWLEDGEMENTS

We would like to acknowledge Andrew Su for providing virtual workspace and staff for collaboration, Sabah Ul-Hasan for proofreading and improving the manuscript and Eliot Ragueneau for advising on aspects of the improvements to ComplexViewer.

FUNDING

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI); Open Targets [OTAR-044, OTAR-048]; National Eye Institute; National Human Genome Research Institute; National Heart, Lung, and Blood Institute; National Institute of Allergy and Infectious Diseases; National Institute of Diabetes and Digestive and Kidney Diseases; National Institute of General Medical Sciences; by National Cancer Institute; National Institute On Aging; National Institute of Mental Health of the National Institutes of Health [U24HG007822 to S.O., H.B.]; National Human Genome Research Institute [U41HG002273 to S.O., H.B.]; National Institute of General Medical Sciences [R01GM080646 to S.O., H.B., P20GM103446 to S.O., H.B.] (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health); Wellcome Trust [218294 to C.C, J.R.]; Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust [203149]; INSERM [to G.B., B.M.]; São Paulo Research Foundation [2019/26284-1 to T.L.]; Alfred P. Sloan Foundation [G-2019-11458 to E.W.]; National Institute of General Medical Sciences [R01 GM089820 to A.W.]. Funding for open access charge: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI).

Conflict of interest statement. None declared.

REFERENCES

1. Meldal,B.H.M., Pons,C., Perfetto,L., Del-Toro,N., Wong,E., Aloy,P., Hermjakob,H., Orchard,S. and Porras,P. (2021) Analysing the yeast complexome-the complex portal rising to the challenge. *Nucleic. Acids. Res.*, **49**, 3156–3167.
2. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., García-Martínez,J., Pérez-Ortín,J.E. *et al.* (2005) CYGD: the comprehensive yeast genome database. *Nucleic. Acids. Res.*, **33**, D364–D368.
3. Combe,C.W., Sivade,M.D., Hermjakob,H., Heimbach,J., Meldal,B.H.M., Micklem,G., Orchard,S. and Rappsilber,J. (2017) ComplexViewer: visualization of curated macromolecular complexes. *Bioinformatics*, **33**, 3673–3675.
4. Sartori,P. and Leibler,S. (2020) Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 114–120.
5. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L.Y., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
6. Liebeskind,B.J., Aldrich,R.W. and Marcotte,E.M. (2019) Ancestral reconstruction of protein interaction networks. *PLoS Comput. Biol.*, **15**, e1007396.
7. Taggart,J.C. and Li,G.-W. (2018) Production of protein-complex components is stoichiometric and lacks general feedback regulation in eukaryotes. *Cell Syst.*, **7**, 580–589.
8. Michalak,W., Tsiamis,V., Schwämmle,V. and Rogowska-Wrzesińska,A. (2019) ComplexBrowser: a tool for identification and quantification of protein complexes in large-scale proteomics datasets. *Mol. Cell. Proteomics*, **18**, 2324–2334.
9. Meldal,B.H.M., Bye-A-Jee,H., Gajdoš,L., Hammerová,Z., Horácková,A., Melicher,F., Perfetto,L., Pokorný,D., Lopez,M.R., Türková,A. *et al.* (2019) Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic. Acids. Res.*, **47**, D550–D558.
10. Velankar,S., Burley,S.K., Kurisu,G., Hoch,J.C. and Markley,J.L. (2021) The Protein Data Bank Archive. *Methods Mol. Biol.*, **2305**, 3–21.
11. Abbott,S., Iudin,A., Korir,P.K., Somasundharam,S. and Patwardhan,A. (2018) EMDB Web Resources. *Curr. Protoc. Bioinformatics*, **61**, 5.10.1–5.10.12.
12. 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.
13. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
14. 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.
15. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.
16. Keseler,I.M., Gama-Castro,S., Mackie,A., Billington,R., Bonavides-Martínez,C., Caspi,R., Kothari,A., Krummenacker,M., Midford,P.E., Muñiz-Rascado,L. *et al.* (2021) The EcoCyc Database in 2021. *Front. Microbiol.*, **12**, 711077.
17. Stasi,R., Neves,H.I. and Spira,B. (2019) Phosphate uptake by the phosphonate transport system PhnCDE. *BMC Microbiol.*, **19**, 79.
18. Perfetto,L., Pastrello,C., Del-Toro,N., Duesbury,M., Iannuccelli,M., Kotlyar,M., Licata,L., Meldal,B., Panneerselvam,K., Panni,S. *et al.* (2020) The IMEx coronavirus interactome: an evolving map of Coronaviridae-host molecular interactions. *Database*, **2020**, baaa096.
19. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
20. Young,J.Y., Berrisford,J. and Chen,M. (2021) wwPDB biocuration: on the front line of structural biology. *Nat. Methods*, **18**, 431–432.
21. Chiu,W., Schmid,M.F., Pintilie,G.D. and Lawson,C.L. (2021) Evolution of standardization and dissemination of cryo-EM structures and data jointly by the community, PDB, and EMDB. *J. Biol. Chem.*, **296**, 100560.
22. Porras,P., Barrera,E., Bridge,A., Del-Toro,N., Cesareni,G., Duesbury,M., Hermjakob,H., Iannuccelli,M., Jurisica,I., Kotlyar,M. *et al.* (2020) Towards a unified open access dataset of molecular interactions. *Nat. Commun.*, **11**, 6144.
23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
24. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
25. Bertholier,C., Vallet,S.D., Deniaud,M., Clerc,O. and Ricard-Blum,S. (2021) Building protein-protein and protein-glycosaminoglycan interaction networks using MatrixDB, the extracellular matrix interaction database. *Curr. Protoc.*, **1**, e47.
26. 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.

27. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., Miller,R.A., Digles,D., Lopes,E.N., Ehrhart,F. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
28. Lovering,R.C., Gaudet,P., Acencio,M.L., Ignatchenko,A., Jolma,A., Fornes,O., Kuiper,M., Kulakovskiy,I.V., Lægreid,A., Martin,M.J. *et al.* (2020) A GO catalogue of human DNA-binding transcription factors. bioRxiv doi: https://doi.org/10.1101/2020.10.28.359232, 28 October 2020, preprint: not peer reviewed.
29. Lombardot,T., Morgat,A., Axelsen,K.B., Aimo,L., Hyka-Nouspikel,N., Niknejad,A., Ignatchenko,A., Xenarios,I., Coudert,E., Redaschi,N. *et al.* (2019) Updates in Rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res.*, **47**, D596–D600.
30. Sivade Dumousseau,M., Alonso-López,D., Ammari,M., Bradley,G., Campbell,N.H., Ceol,A., Cesareni,G., Combe,C., De Las Rivas,J., Del-Toro,N. *et al.* (2018) Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**, 134.
31. Velthuijs,N., Meldal,B., Geessinck,Q., Porras,P., Medvedeva,Y., Zubritskiy,A., Orchard,S. and Logie,C. (2021) Integration of transcription coregulator complexes with sequence-specific DNA-binding factor interactomes. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1864**, 194749.
32. Drew,K., Wallingford,J.B. and Marcotte,E.M. (2021) hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.*, **17**, e10016.
33. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.*, **47**, D559–D563.
34. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B.-J., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F. *et al.* (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.
35. Vrandečić,D. and Krötzsch,M. (2014) Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, **57**, 78–85.
36. Waagmeester,A., Stupp,G., Burgstaller-Muehlbacher,S., Good,B.M., Griffith,M., Griffith,O.L., Hanspers,K., Hermjakob,H., Hudson,T.S., Hybiske,K. *et al.* (2020) Wikidata as a knowledge graph for the life sciences. *Elife*, **9**, e52614.
37. Ostaszewski,M., Mazein,A., Gillespie,M.E., Kuperstein,I., Niarakis,A., Hermjakob,H., Pico,A.R., Willighagen,E.L., Evelo,C.T., Hasenauer,J. *et al.* (2020) COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci Data*, **7**, 136.
38. Nielsen,F.Å., Mietchen,D. and Willighagen,E. (2017) Scholia, Scientometrics and Wikidata. In: Blomqvist,E., Hose,K., Paulheim,H., Ławrynowicz,A., Ciravegna,F. and Hartig,O. (eds). *The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. Lecture Notes in Computer Science*. Vol. **10577**, Springer, Cham.
39. Wong,E.D., Skrzypek,M.S., Weng,S., Binkley,G., Meldal,B.H.M., Perfetto,L., Orchard,S.E., Engel,S.R., Cherry,J.M. and SGD Project (2019) Integration of macromolecular complex data into the Saccharomyces Genome Database. *Database*, **2019**, baz008.
40. Touré,V., Vercruysse,S., Acencio,M.L., Lovering,R.C., Orchard,S., Bradley,G., Casals-Casas,C., Chaouiya,C., Del-Toro,N., Flobak,Å. *et al.* (2021) The minimum information about a Molecular Interaction CAusal STatement (MI2CAST). *Bioinformatics*, **36**, 5712–5718.
41. van Iersel,M.P., Pico,A.R., Kelder,T., Gao,J., Ho,I., Hanspers,K., Conklin,B.R. and Evelo,C.T. (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**, 5.
42. 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.
43. Bindea,G., Mlecnik,B., Hackl,H., Charoentong,P., Tosolini,M., Kirilovsky,A., Fridman,W.-H., Pagès,F., Trajanoski,Z. and Galon,J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.
44. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
45. Megill,C., Martin,B., Weaver,C., Bell,S., Prins,L., Badajoz,S., McCandless,B., Pisco,A.O., Kinsella,M., Griffin,F. *et al.* (2021) Cellxgene: A performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv doi: https://doi.org/10.1101/2021.04.05.438318, 06 April 2021, preprint: not peer reviewed.