Supporting Information

Drug repurposing for the treatment of COVID-19: a knowledge graph approach

Vincent KC Yan[†], Xiaodong Li[†], Xuxiao Ye[†], Min Ou, Ruibang Luo, Qingpeng Zhang, Bo Tang, Benjamin J Cowling, Ivan Hung, Chung Wah Siu, Ian CK Wong, Reynold CK Cheng[*], Esther W Chan[*]

[†] Vincent KC Yan, Xiaodong Li and Xuxiao Ye contributed equally to this work. [*] Esther W Chan and Reynold CK Cheng share senior authorship.

Description of data sources included for creation of knowledge graph

- DrugBank database[1] is a comprehensive, freely accessible database containing information on drugs and drug targets. It currently contains 13,575 drug entries and is widely used by industry, medical practitioners and the general public. It has enabled the discovery and repurposing of a number of existing drugs to treat rare and newly identified illnesses and is a unique bioinformatics and cheminformatics resource.
- ClinicalTrials.gov[2] is a Web-based resource that provides patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions.
- Pharmacogenomics Knowledgebase (PharmGKB)[3] is a publicly available, online knowledgebase responsible for the aggregation, curation, integration, and dissemination of knowledge regarding the impact of human genetic variation on drug response.
- BindingDB[4] is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be candidate drug-targets with ligands that are small, drug-like molecules.
- Therapeutic Target Database (TTD)[5] involves information about (i) target-regulating microRNAs and transcription factors, (ii) target-interacting proteins, and (iii) patented agents and their targets (structures and experimental activity values if available), which can be conveniently retrieved and is further enriched with regulatory mechanisms or biochemical classes.
- BioGRID[6] is an interaction repository with data compiled through comprehensive curation efforts.
- Database of Interacting Proteins (DIP)[7] catalogues experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions.
- Human Protein Reference Database (HPRD)[8] represents a centralised platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks, and disease association for each protein in the human proteome.
- NCBI Entrez[9] is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centred and genomic mapping information, 3D structure data, PubMed MEDLINE, and more.
- Comparative Toxicogenomic Database (CTD)[10] provides manually curated information about chemical–gene/protein interactions, chemical–disease, and gene-disease relationships. These data are integrated with functional and pathway data to aid in the development of hypotheses about the mechanisms underlying environmentally influenced diseases.

- Human Phenotype Ontology (HPO)[11] is central in medical genetics and genomics It provides a standardized vocabulary of phenotypic abnormalities encountered in human disease and serves as a computational bridge between genome biology and clinical medicine.
- Disease Ontology (DO)[12] provides an open-source ontology for the integration of biomedical data that is associated with human disease.
- Medical Subject Headings (MeSH)[13] are the National Library of Medicine controlled vocabulary thesaurus used for indexing articles for PubMed.
- OpenKG[14] is maintained by several universities and companies from China, e.g., Tsinghua University and Huawei. The datasets are from different areas, and the datasets for research purposes include information about host, virus, drugs, gene and protein in JSON format.[15] However, the COVID-19 related data are limited and scattered.

The latest versions as of 10 August 2020 were used for all the data sources listed above.

Data integration

Detailed codes and algorithms used for the data integration process were documented and released for open access at https://github.com/Sheldon2016/covid19kg.

References

[1] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Res.* **2018**, *46*, D1074.

[2] National Library of Medicine. Clinical Trials Registry. https://clinicaltrials.gov/. accessed **2020**.

[3] M. Whirl-Carrillo, E. M. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. B. Altman, T. E. Klein, *Clin. Pharmacol. Ther.* **2012**, *92*, 414.

[4] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, *Nucleic Acids Res.* **2016**, *44*, D1045.

[5] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, *Nucleic Acids Res.* **2020**, *48*, D1031.

[6] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, *Nucleic Acids Res.* **2006**, *34*, D535.

[7] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, D. Eisenberg, *Nucleic Acids Res.* **2000**, *28*, 289.

[8] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, *Nucleic Acids Res.* **2009**, *37*, D767.

[9] NCBI Resource Coordinators, *Nucleic Acids Res.* **2017**, *45*, D12.

[10] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, C. J. Mattingly, *Nucleic Acids Res.* **2019**, *47*, D948.

[11] S. Köhler, L. Carmody, N. Vasilevsky, J. O B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, D. Osumi-Sutherland, V. Cipriani, J. P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A. C. Jansen, A. Muaz, W. H. Chang, J. Bergerson, S. J F. Laulederkind, Z. Yüksel, S. Beltran, A. F. Freeman, P. I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S. M. Bello, M. Sincan, K. Rageth, M. T. Wheeler, R. Oegema, H. Lourghi, M. G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R. C. Lovering, C.

Hajek, A. Olry, L. Notarangelo, M. Similuk, X. A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J. D. Milner, D. Leroux, C. F. Boerkoel, A. Klion, M. C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, P. N. Robinson, *Nucleic Acids Res.* **2019**, *47*, D1018.
[12] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, *Nucleic Acids Res.* **2019**, *47*, D955.
[13] National Library of Medicine. Medical Subject Headings. https://www.nlm.nih.gov/mesh/meshhome.html. accessed **2020**.
[14] OpenKG workgroup. OpenKG. http://www.openkg.cn/dataset/covid-19-research. accessed.
[15] W. Dai, M. Huang, Q. Wu, H. Cai, M. Sheng, X. Li, in *International Conference on Web Information Systems and Applications*, 2020, 314.

Table S1. Number of drug candidates ranked among top n% by our algorithm which are also under or completed clinical trial for the treatment of COVID-19. A total of 5,624 drug candidates were scored and ranked.

Percentage (%)	# of drugs under or completed clinical trial
1	14
2	28
3	33
4	39
5	44
6	58
7	61
8	67
9	71
10	74
11	79
12	85
13	90
14	102
15	108
16	112
17	120
18	123
19	126
20	130
21	132
22	140
23	141
24	141
25	146
26	151
27	158
28	161
29	163
30	164
31	168
32	171
33	173
34	174
35	178
36	179
37	182
38	184
39	190
40	193
41	194
42	195

43	196
44	196
45	199
46	202
47	202
48	206
49	208
50	209
51	209
52	210
53	212
54	212
55	216
56	216
57	218
58	219
59	222
60	226
61	231
62	234
63	234
64	238
65	239
66	244
67	244
68	244
69	247
70	248
71	249
72	250
73	250
74	253
75	254
76	266
77	267
78	268
79	269
80	269
81	269
82	270
83	271
84	272
85	273
86	273
87	277
88	277
89	277

90	278
91	278
92	280
93	280
94	280
95	283
96	283
97	283
98	284
99	284
100	289