SCIENTIFIC DATA

BREATHE: Biomedical Research Extensive Archive To Help Everyone

May 1, 2020

Abstract

Due to the outbreak of the novel coronavirus strain of disease SARS-CoV-2 causing COVID-19, there is a need to develop better language models to understand research literature from the biomedical domain. In response to this, we present BREATHE, a large-scale biomedical database containing entries from 11 major repositories of biomedical research. Our dataset contains both abstract and full body texts of biomedical papers since the 1600s up till the present time and is still being updated by our crawlers. This dataset can be used to train language models to better understand outcomes from biomedical research and uncover insights to combat the COVID-19 pandemic.

Background & Summary

(700 words maximum) An overview of the study design, the assay(s) performed, and the created data, including any background information needed to put this study in the context of previous work and the literature. The section should also briefly outline the broader goals that motivated the creation of this dataset and the potential reuse value. We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design. This section and the other main body sections of the manuscript should include citations to the literature as needed [1, 2]. References should be included within the manuscript file itself as our system cannot accept BibTeX bibliography files. Authors who wish to use BibTeX to prepare their references should therefore copy the reference list from the .bbl file that BibTeX generates and paste it into the main manuscript .tex file (and delete the associated \bibliography and \bibliographystyle commands).

Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental de-

sign, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the detailed section in our submission guidelines for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below). Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section.

Code availability

For all studies using custom code in the generation or processing of datasets, a statement must be included in the Methods section, under the subheading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

Data Records

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited numerically in the text of this section, for example [3, 4, 5, 6], and included in the main reference list as described below.. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record.

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each (please see Tables and Submitting Experimental Metadata, below). They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

Technical Validation

This section presents any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by up figures and tables, as needed. This is a required section; authors must present information justifying the reliability of their data.

SCIENTIFIC DATA

Usage Notes

The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our code availability policy for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

Acknowledgements

The Acknowledgements should contain text acknowledging non-author contributors. Acknowledgements should be brief, and should not include thanks to anonymous referees and editors or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions

Each author's contribution to the work should be described briefly, on a separate line, in the Author Contributions section.

Competing interests

A competing interests statement is required for all papers accepted by and published in *Scientific Data*. If there is no conflict of interest, a statement declaring this must still be included in the manuscript.

Figures and figures legends

Figure should be referred to using a consistent numbering scheme through the entire Data Descriptor. For initial submissions, authors may choose to supply this document as a single PDF with embedded figures, but separate figure image files must be provided for revisions and accepted manuscripts. In most cases, a Data Descriptor should not contain more than three figures, but more may be allowed when needed. We discourage the inclusion of figures in the Supplementary Information – all key figures should be included here in the main Figure section.

Figure legends begin with a brief title sentence for the whole figure and continue with a short description of what is shown in each panel, as well as explaining any symbols used. Legend must total no more than 350 words, and may contain literature references.

Tables

Authors are encouraged to provide one or more tables that provide basic information on the main 'inputs' to the study (e.g. samples, participants, or information sources) and the main data outputs of the study; also see the additional information on providing metadata on page 6. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Tables may be provided within the LaTeX document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included here. Generally, a Data Descriptor should have fewer than ten Tables, but more may be allowed when needed. Tables may be of any size, but only Tables which fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as 'Online-only' tables and numbered separately from the main table list e.g. 'Table 1, Table 2, Online-only Table 1' etc.

References

[1] Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).

[2] Wang, R. *et al.* PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat. Biotechnol* **30**, 135–137 (2012).

[3] Zhang, Q-L., Chen, J-Y., Lin, L-B., Wang, F., Guo, J., Deng, X-Y. Characterization of ladybird Henosepilachna vigintioctopunctata transcriptomes across various life stages. *figshare* https://doi.org/10.6084/m9.figshare.c.4064768.v3 (2018).

[4] *NCBI Sequence Read Archive* http://identifiers.org/ncbi/insdc.sra:SRP121625 (2017).

[5] Barbosa, P., Usie, A. and Ramos, A. M. Quercus suber isolate HL8, whole genome shotgun sequencing project. *GenBank* http://identifiers.org/ncbi/insdc:PKMF00000000 (2018).

[6] *DNA Data Bank of Japan* http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA004814 (2016).

Citing Data

In line with emerging industry-wide standards for data citation, references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the 'References' section in the same manner as a conventional literature reference. See the examples above.