covid19.analytics: An R Package to Obtain, Analyze and Visualize Data from the 2019 Coronavirus Disease Pandemic

21 January 2021

Summary

With the emergence of a new pandemic worldwide, novel strategies attempting to understand its evolution and dynamics, enpowered by computational techniques such as data analysis and computational modelling, have also emerged. Several initiatives under the umbrella of *open science* are contributing to tackle this unprecedented situation. In particular, the *R Language* (R Core Team 2016; Ihaka and Gentleman 1996) could be an excellent tool and ecosystem for approaches focusing on open science and reproducible research. Hence it is not surprising that with the onset of the pandemic, a large number of R packages and resources were made available for researches working in the pandemic.

In this paper, we present the `covid19.analytics` R package which allows users to access and analyze worldwide data from publicly available resources.

The `covid19.analytics` R package allows users to obtain live[1] worldwide data of reported cases from the novel *Coronavirus Disease* (COVID-19), as well as related datasets, such as, genomics data, testing and vaccinations records and historical pandemics records. It does this by accessing and retrieving data publicly available and published by several sources, such as:

- The "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" ("COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" 2020) for the worldwide and US data
- Health Canada ("Health Canada – COVID-19 cases in Canada" 2020), for Canada specific data
- The city of Toronto for Toronto data ("COVID-19: Status of Cases in Toronto" 2020)

[1] The data usually is accessible from the repositories with a 24 hours delay.

- Open Data Toronto for Toronto data ("Open Data Toronto – COVID-19 cases in the City of Toronto" 2020)
- Our World In Data for testing and vaccination data (Our World In Data 2020)
- NCBI servers for genomics datasets (NCBI Resource Coordinators 2017)
- Visual Capitalist infographics for historical pandemic records (Visual Capitalist 2020)

The package also provides basic analysis and visualization tools and functions to investigate these datasets and other ones structured in a similar fashion.

Statement of need

One important element to consider in the case of an ongoing and unfolding pandemic, is the rapid data accessibility and its prompt availability. One of the main goals of the `covid19.analytics` package is to make the latest data related to the COVID19 pandemic promptly available to researchers and the scientific community.

As today, there are a few dozen other packages also in the CRAN repository that allow users to gain access to different datasets of the COVID-19 pandemic. In some cases, some packages just provide access to data from specific geographical locations or the approach to the data structure in which the data is presented is different from the one presented here. Nevertheless, having a variety of packages from which users can try and probably combine, is an important and crucial element in data analysis. Moreover it has been reported different cases of data misuse/misinterpretation due to different issues, such as, erroneous metadata or data formats (Ledford and Van Noorden 2020) and in some cases ending in articles' retractions (Schriml et al. 2020). Therefore providing additional functionalities to check the integrity and consistency of the data, as our `covid19.analytics` package does is paramount. This is specially true in a situation where the unfolding of events and data availability is flowing so fast that sometimes is even hard to keep track of all the changes.

As today and to the best of our knowledge, there isn't any other R package that provides direct access to the COVID19 genomic data of the virus. The `covid19.analytics` package is indeed capable of retrieving genomics data, and it does that by incorporating a novel, more reliable and robust way of accessing and designing different pathways to the data sources.

Moreover, the `covid19.analytics` package offers a modular and versatile approach to the data, by allowing users to input their own data for which most of the functions in the package can be applied when the data is structured using a *time series* format as described in the package documentation.

Functionalities and Main Features

The `covid19.analytics` package is an open source tool, which its main goal is to offer users prompt and quick access to different data sources related to the unfolding pandemics. Its main implementation and API is the R package (Ponce 2020) which its stable version is part of CRAN (Hornik 2012).

In addition to this, the package has a few more adds-on:

- A central GitHub repository, https://github.com/mponce0/covid19.analytics where the latest development version and source code of the package are available. Users can also submit tickets for bugs, suggestions or comments using the "issues" tab from this repository.
- A rendered version with live examples and documentation also hosted at GitHub pages, https://mponce0.github.io/covid19.analytics/
- A dashboard for interactive usage of the package with extended capabilities for users without any coding expertise, https://covid19analytics.scinet.utoronto.ca
- A backup data repository hosted at GitHub, https://github.com/mponce0/covid19analytics.datasets – where replicas of the live datasets are stored for redundancy and robust accessibility sake (see Figure 1).

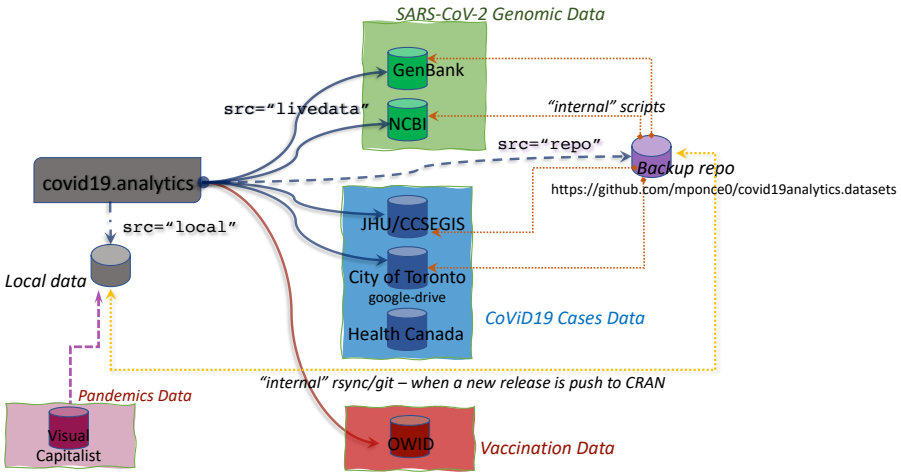

Figure 1: Schematic of the data acquisition flows between the `covid19.analytics` package and the different sources of data. Dark and solid/dashed lines represent API functions provided by the package accessible to the users. Dotted lines are "internal" mechanisms employed by the package to synchronize and update replicas of the data.

The package also incorporates a dashboard to facilitate the access to its function-

alities to less experienced users (see Figure 2). This can be accessed interactively in a local session or through the public deployed dashboard at the URL described above.

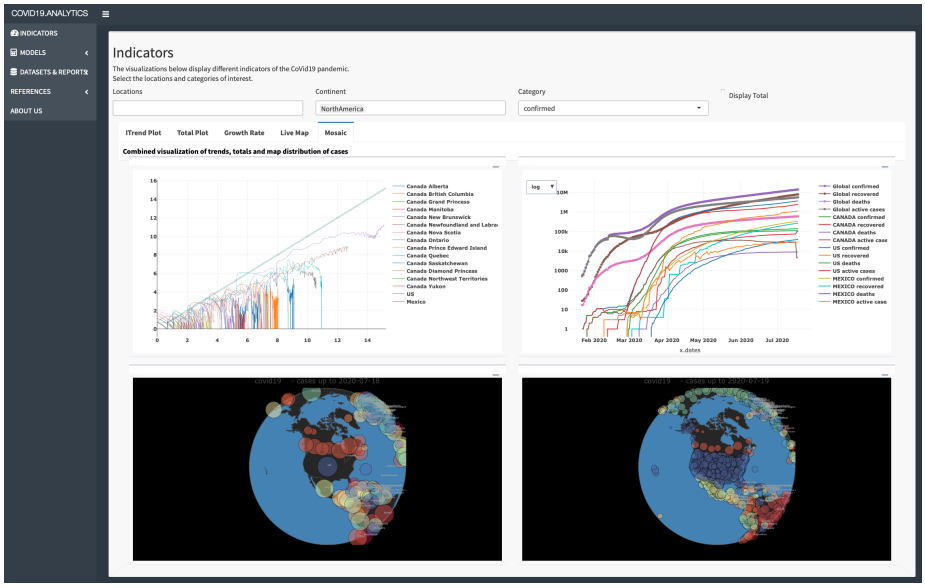

Figure 2: Screenshot of a `covid19.analytics` dashboard implementation – the dashboard can be used through a web-interface or deployed locally in the users' machine.

Another unique feature of this package is the ability of incorporating models to estimate the disease spread by using the actual data. The `covid19.analytics` package allows users to model the dispersion of the disease by implementing a simple *Susceptible-Infected-Recovered* (SIR) model (Kermack and McKendrick 1927; Smith, Moore, and others 2004). The package can also generate multiple SIR models by varying some of the parameters used to select the actual data to be used in the model and some additional visualizations tools.

There are several plotting and visualization functions within the `covid19.analytics` package, both generating static and interactive visualizations.

Further details, documentation, examples and tutorials of the `covid19.analytics` package can be found on the package repository and (Ponce and Sandhel 2020).

Users can submit bug reports, improvements requests or seek for support using the "issues tracker" tab from the GitHub repository. Contributions from users and the community are welcome and can be done using "pull requests" from the repository as well. We will continue working on adding and developing new features to the package, a glimpse and brief description about these can be found on the repository's Projects area.

Acknowledgments

MP wants to thank all his colleagues at SciNet, especially Daniel Gruner for his continuous and unconditional support, and Marco Saldarriaga who helped us setting up the VM for installing the shiny server hosting the covid19.analytics dashboard web interface.

References

"COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University." 2020. https://github.com/CSSEGISandData/COVID-19.

"COVID-19: Status of Cases in Toronto." 2020. https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/.

"Health Canada – COVID-19 cases in Canada." 2020. https://health-infobase.canada.ca/src/data/covidLive.

Hornik, Kurt. 2012. "The Comprehensive R Archive Network." *WIREs Computational Statistics* 4 (4): 394–98. https://doi.org/10.1002/wics.1212.

Ihaka, Ross, and Robert Gentleman. 1996. "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5 (3): 299–314. https://doi.org/10.1080/10618600.1996.10474713.

Kermack, William Ogilvy, and Anderson G McKendrick. 1927. "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115 (772): 700–721.

Ledford, Heidi, and Richard Van Noorden. 2020. "High-Profile Coronavirus Retractions Raise Concerns About Data Oversight." *Nature* 582 (7811): 160–60. https://doi.org/10.1038/d41586-020-01695-w.

NCBI Resource Coordinators. 2017. "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–D13. https://doi.org/10.1093/nar/gkx1095.

"Open Data Toronto – COVID-19 cases in the City of Toronto." 2020. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

Our World In Data. 2020. "Our World in Data – COVID-19 Data Repository." https://github.com/owid/covid-19-data/.

Ponce, Marcelo. 2020. *covid19.analytics: Load and Analyze Live Data from the CoViD-19 Pandemic*. https://CRAN.R-project.org/package=covid19.analytics.

Ponce, Marcelo, and Amit Sandhel. 2020. "covid19.analytics: An R Package to Obtain, Analyze and Visualize Data from the Coronavirus Disease Pandemic." http://arxiv.org/abs/2009.01091.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schriml, Lynn M., Maria Chuvochina, Neil Davies, Emiley A. Eloe-Fadrosh, Robert D. Finn, Philip Hugenholtz, Christopher I. Hunter, et al. 2020. "COVID-19 pandemic reveals the peril of ignoring metadata standards." *Scientific Data* 7 (1): 188. https://doi.org/10.1038/s41597-020-0524-5.

Smith, David, Lang Moore, and others. 2004. "The SIR model for spread of disease: the differential equation model." *Loci.(originally Convergence.)*. https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model.

Visual Capitalist. 2020. "Visualizing the History of Pandemics & The Race to Save Lives: Comparing Vaccine Development Timelines." https://www.visualcapitalist.com/history-of-pandemics-deadliest/.