BOLD.R

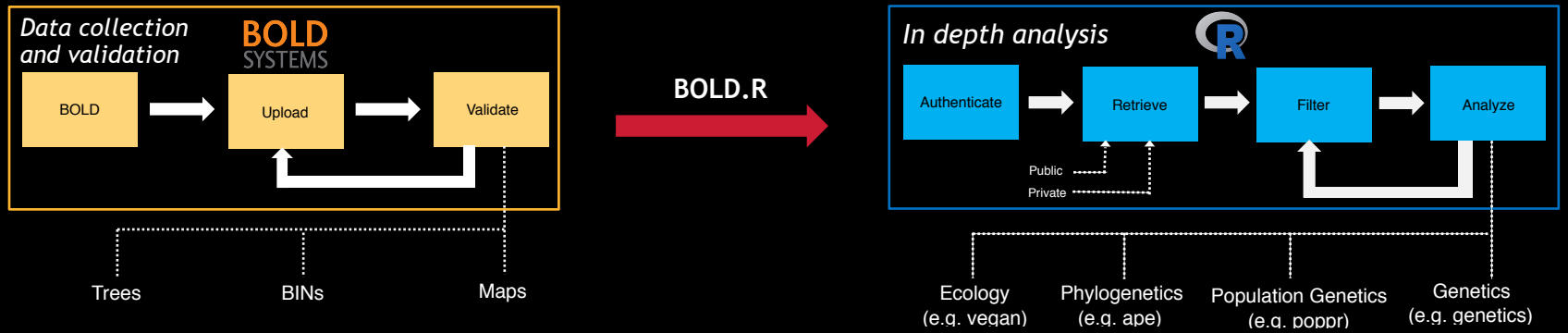
Nishan Mudalige [1, 2, 3], Megan Milton [1] and Sujeevan Ratnasingham [1]

International HPC Summer School 2018, Ostrava, Czech Republic

[1] Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, N1G 2W1, Canada.
[2] Department of Mathematics and Statistics, University of Guelph, Guelph, ON, N1G 2W1, Canada.
[3] Ontario Graduate Scholar

BOLD.R:

A software package to interface with BOLD through R



Background

The International Barcode of Life project (iBOL) continuously records and catalogs species information and stores data on the Barcode of Life Database system (BOLD). Advances in DNA analysis have lead to a rapid increase in the volume of data available for researchers involved with the iBOL project to study. As a consequence, modern statistical techniques are playing an increasingly important role in the analysis of such large volumes of data. A popular software package for statistical analysis is R, however existing methods to retrieve data from BOLD into R are inconvenient, time-consuming or return limited information. Therefore we have introduced a more accessible system to provide convenient and direct access to the data stored on BOLD into R.

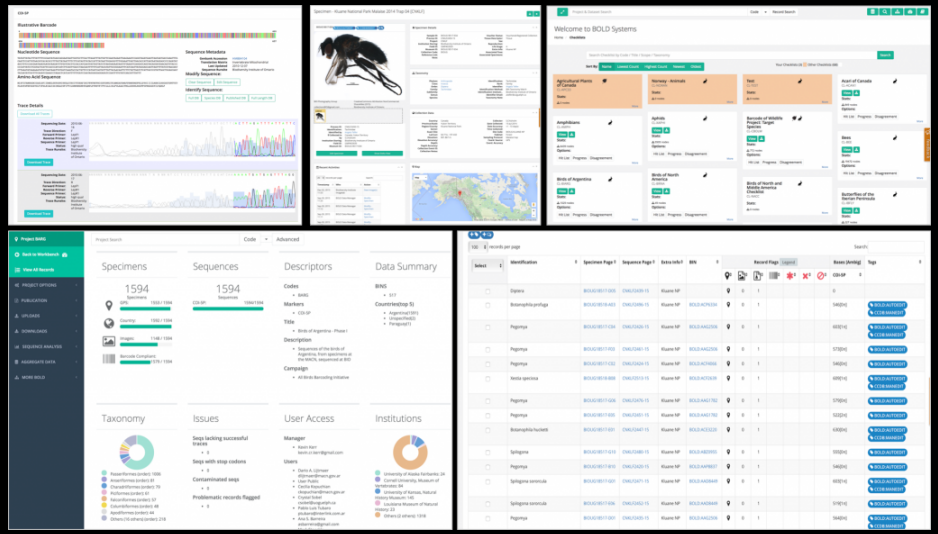
Significance

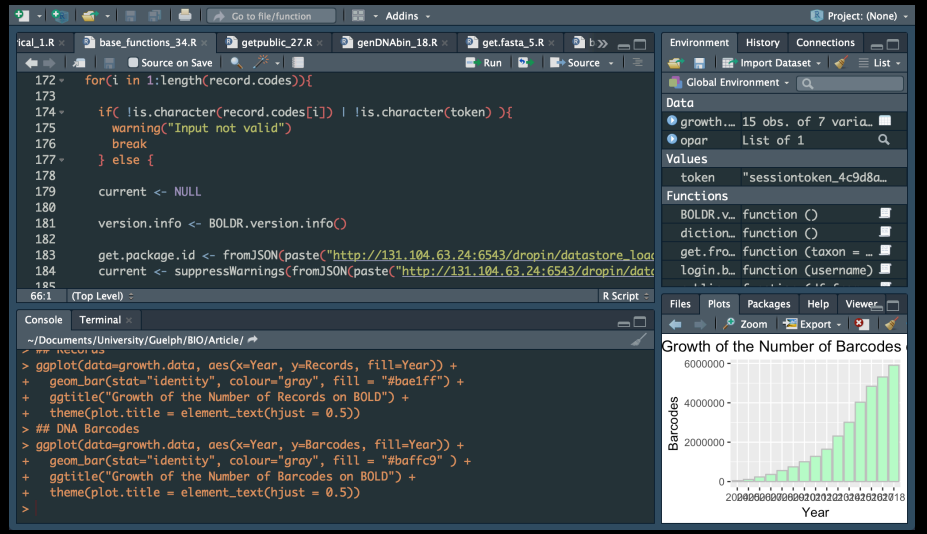
We developed an R library called BOLD.R which allows users to access data directly from BOLD into R via current APIs maintained by BOLD. Users can access their own private data by logging in to BOLD using BOLD.R, or they can access public data without the need to login. All data accessed using BOLD.R is stored in R with a consistent internal structure and this allows the user to employ the suite of functions provided by BOLD.R along with existing packages to perform statistical analysis on DNA sequences.

BOLDR is a powerful R library which bridges the gap between data storage and data analysis by providing the user with the ability to access and analyze large amounts of private and public data on BOLD directly through R.

BOLD.R:

A software package to interface with BOLD through R





Advantages of R

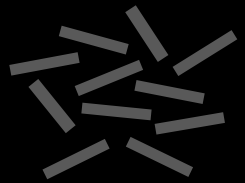
- *Cost*:
  R is free open-source software.
- *Development environment*:
  R is built specifically for statistical analysis by statisticians and scientific researchers.
- *Extensive*:
  R is the most complete statistical analysis package available.
- *Efficient and Optimized*:
  Base R Is very efficient on resources and many libraries are often continuously updated until optimized.
- *Comprehensive*:
  R has a large selection of comprehensive libraries that can be used to perform generic functions as well as be used for specific research objectives.
- *Visual*:
  R offers exceptional graphical capabilities.
- *Format support*:
  R offers users support for many file formats (input and output).
- *Cross-platform*:
  R is available for all major platforms (Windows, Mac, Linux).

BOLD.R:

A software package to interface with BOLD through R

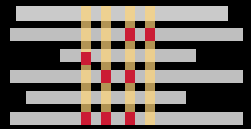
Chemical process | Computational processes | (Future)

Sequencing → Read Alignment → Substitution Matrix → Classification → Reporting → Statistical Analysis



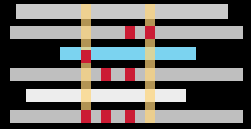
Copy the target gene millions of times and sequence using only that gene and produce a single DNA string.

Align reads to reference reference library



Locate substitution sites and employ error model to discard unreliable substitutions

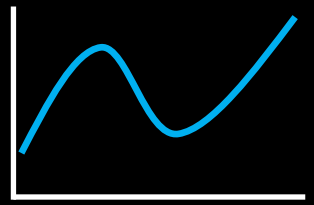
General similarity search pipeline



Apply classifier engine to generate species designation and associated confidence

One or more species matches with confidence and decision support information

Domain-specific assignment pipeline



Harness the power of HPC by running scripts via cloud computing on BOLD hardware or on the hardware of an HPC partner (such as Compute Canada).

Statistical Analysis

HPC

(BOLD hardware: 1024 cores, 64 TB RAM, 200 TB) + (The cloud: 6000+ cores, 384 TB RAM, 100 TB)

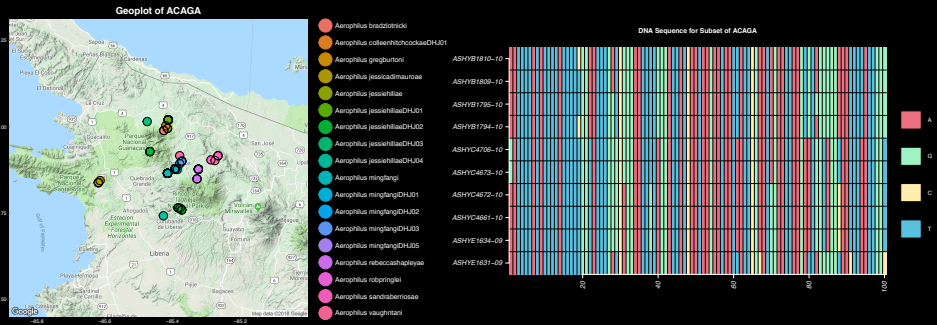
UNIVERSITY of GUELPH

BOLD.R:

A software package to interface with BOLD through R

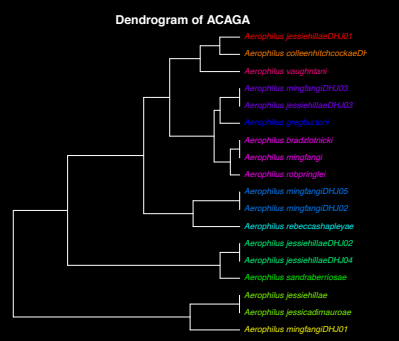
Centre for Biodiversity Genomics

An alpha version of BOLD.R is available for download at: http://www.boldsystems.org/BOLD.R

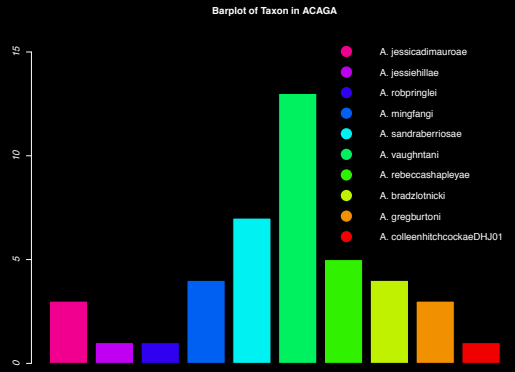


(a) Spatial Distribution

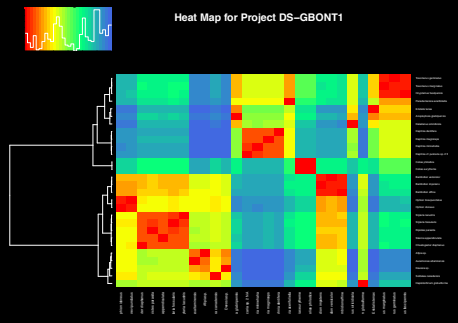
(b) Substitution



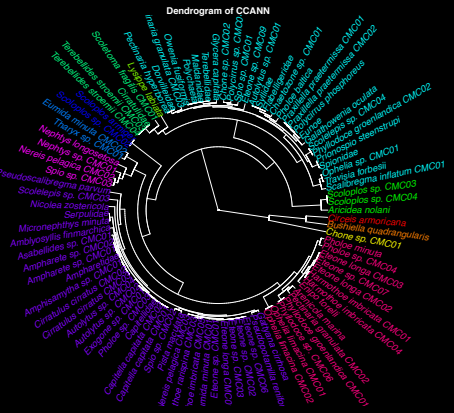
(c) Phylogeny



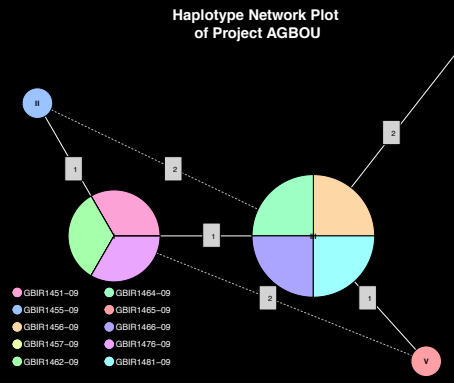
(d) Composition



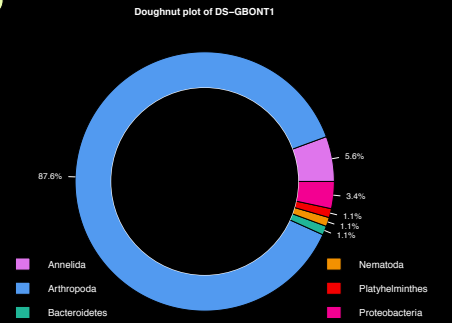
(e) Clustering



(f) Phylogeny



(g) Network



(g) Composition

BOLD.R makes it significantly easier to use the vast array of libraries available in R.
We used BOLD.R along with existing R libraries to obtain all of the plots and figures above.

BOLD.R:

A software package to interface with BOLD through R

Long Term Plan

- *Employ the analytical engine of BOLD:*
  Develop the BOLD API and BOLD.R to harness the power of the analytical engine of BOLD through cloud computing.
- *Establish a script repository:*
  Users can store their own scripts privately or make scripts public for others to use.
- Integrate popular scripts as Plugins on BOLD.

Summary

BOLD.R is a powerful, flexible and convenient package which provides the user with the ability to access and analyze large amounts of private and public data on BOLD directly through R. BOLD Systems and its partners plan to introduce the capacity to perform intense analytical tasks on data stored on BOLD by providing the user with access to BOLD Systems HPC hardware through BOLD.R. This integration will allow users to analyze extremely large datasets related to DNA barcode records. BOLD.R can therefore become an invaluable tool to assist researchers make informed decisions through the automation of data processes through the computational power of HPC.

References

[1.] De Vries, A. et al. (2016) Package 'ggdendro'.
[2.] Galili, T. (2015) Package 'dendextend'.
[3.] Kahle, D. Et al. (2016) Package 'ggmap'.
[4.] Lang, L. et al. (2018) Package 'RCurl'.
[5.] Ooms, J (2014) Package 'jsonlite'.
[6.] Paradis, E., et al. (2017) Package 'ape'.
[7.] Paradis, E., et al. (2017) Package 'pegas'.
[8.] Schmidt et al. (2017) Package 'getPass'.
[9.] Van der Loo, M. et. Al (2017) Package 'stringdist'.
[10.] Warnes, G. R. et. al. (2015) Package 'gtools'.
[11.] Warnes, G. R. et. al. (2016) Package 'gplots'.
[12.] Wickham, H (2016) Package 'plyr'.

This project is funded and supported by