



Cold Spring Harbor Laboratory DNA LEARNING CENTER

MetaR: towards an R notebook with composable languages

Fabien Campagne^{1,2,3,*}, Alexander Pann¹, William ER Digan¹, Manuele Simi^{1,2}

¹The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, United States of America; ²Clinical Translational Science Center, Weill Cornell Medicine, New York, NY, United States of America;

³Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, United States of America. *To whom correspondence should be addressed:

fac2003@campagnelab.org

Keywords: Language Composition, Reproducibility, Language Workbench Technology

Data analysis tools have become essential to the study of biology. Here, we applied language workbench technology (LWT) to create data analysis languages tailored for biologists with a diverse range of experience: from beginners with no programming experience to expert bioinformaticians and statisticians.

A key novelty of our approach is its ability to blend user interface with scripting in a single platform with better language integration than possible with electronic notebooks. This new approach has several advantages over the state of the art: experts can design simplified data analysis languages that require no programming experience, and behave like graphical user interfaces, yet have the advantages of scripting. We report on such a simple language, called MetaR [1], which we have used to teach complete beginners how to call differentially expressed genes and build heatmaps. We found that beginners can complete this task in less than 2 hours with MetaR, when more traditional teaching with R and its packages would require several training sessions (6-24hrs). MetaR also seamlessly integrates with docker to enable reproducibility of analyses and simplified R package installations during training sessions.

We used the same approach to develop the first composable R language [1]. A composable language is a language that can be extended with micro-languages. We illustrate this capability with a Biomart micro-language designed to compose with R and help R programmers query Biomart interactively to assemble specific queries to retrieve data. (The same micro-language also composes with MetaR to help beginners query Biomart.)

Our teaching experience suggests that language design with LWT can be a compelling approach for developing intelligent data analysis tools and can accelerate training for common data analysis task. LWT offers an interactive environment with instant change detection and refreshing of plots and tables. Instant refresh was added to MetaR 2.0 and makes using popular R packages (e.g., tidy, dplyr, ggplots2) much more interactive. The talk will provide an introduction to LWT and MetaR and describe our latest interactive notebook results [1, 2, 3].

Software is distributed under the Apache 2.0 license and available at <https://github.com/CampagneLaboratory/MetaR> and <http://metaR.campagnelab.org>.

References: [1] <http://dx.doi.org/10.1101/030254> [2] Documentation <http://tinyurl.com/zx9wvjw> [3] BOSC 2016 Talk <https://youtu.be/bWYhuQCSagU>