
Article:
Polanski, Krzysztof, Gao, Bo, Mason, Sam A et al. (4 more authors) (2018) Bringing numerous methods for expression and promoter analysis to a public cloud computing service. Bioinformatics. 884–886. ISSN 1460-2059

https://doi.org/10.1093/bioinformatics/btx692

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Gene expression

Bringing numerous methods for expression and promoter analysis to a public cloud computing service

Krzysztof Polański [1], Bo Gao [1], Sam A. Mason [1], Paul Brown [3,4], Sascha Ott [2,4], Katherine J. Denby [5] and David L. Wild [1,4,*]

[1]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK, [2]Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK, [3]Department of Mathematics, University of Warwick, Coventry, CV4 7AL, UK, [4]Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, UK and [5]Department of Biology, University of York, York, YO10 5DD, UK.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Every year, a large number of novel algorithms are introduced to the scientific community for a myriad of applications, but using these across different research groups is often troublesome, due to suboptimal implementations and specific dependency requirements. This does not have to be the case, as public cloud computing services can easily house tractable implementations within self-contained dependency environments, making the methods easily accessible to a wider public. We have taken 14 popular methods, the majority related to expression data or promoter analysis, developed these up to a good implementation standard and housed the tools in isolated Docker containers which we integrated into the CyVerse Discovery Environment, making these easily usable for a wide community as part of the CyVerse UK project.

Availability: The integrated apps can be found at http://www.cyverse.org/discovery-environment, while the raw code is available at https://github.com/cyversewarwick and the corresponding Docker images are housed at https://hub.docker.com/r/cyversewarwick/

Supplementary information: Supplementary data are available at Bioinformatics online.

Contact: info@cyverse.warwick.ac.uk

1 Introduction

Experimental techniques keep evolving at a great pace, constantly increasing the range of research questions that can be posed to the data. This creates a need for computational methods to keep up, be it with the decreasing cost of a laboratory procedure making larger scale experimental designs possible (Windram *et al.*, 2012) or due to an improved technique leading to a drastic change in the character of the resulting data (Anders and Huber, 2010). As such, a huge number of novel algorithms are being created, but their upkeep and usability vary greatly. Some methods take on the form of established, regularly updated packages with extensive documentation that are easy to set up and use locally (Gentleman *et al.*, 2004), whilst others are but a set of scripts attached to a research paper with no documentation or subsequent upkeep, quickly becoming very difficult to run. Another common issue among algorithms that do not take on the form of dedicated software packages is the quality of the implementation, with the scripts often created in programmer-friendly environments, leading to less efficient implementations (Penfold *et al.*, 2012; Polanski *et al.*, 2014). These factors make the application of a number of very useful methods much more challenging than it has to be. Over the years, a number of freely accessible cloud computing services have been made available to the scientific community for data analysis purposes. Examples include iPlant, now rebranded to CyVerse (Goff *et al.*, 2011) and Galaxy (Hillman-Jackson *et al.*, 2012). CyVerse is a National Science Foundation (NSF)-funded cyberinfrastructure that democratises access to data storage space, HPC and cloud computing facilities. CyVerse provides three key services to its users: the cloud-based Data Store that enables scientists to store and share very large datasets; the user-friendly Discovery Environment, in which they can work individually

or collaboratively to analyse data using 'apps' built by the individual researchers or the wider community; and Atmosphere, through which users can access on-demand high-performance cloud computing power. To help spread effort, expertise and resources, CyVerse operates a distributed model within the US between TACC, Cold Spring Harbor Laboratory, and the University of Arizona, and its platform has been designed with extension and replication in mind. CyVerse middelware (computational software interfaces between services) enables integration of multiple data sources and HPC facilities to provide one simple user interface. Since its launch in 2008, more than 1800 researchers now use CyVerse. Such platforms make it possible to outsource the large computational burden of analysing big data. A large number of algorithms present on a single server create the need for separate environments to avoid dependency clashes, leading to the use of technologies such as Docker (Merkel, 2014). CyVerse UK is a joint effort between the Universities of Warwick, Liverpool and Nottingham, and the Earlham Institute, to create a UK node of the CyVerse collaborative. Here we document the result of a large body of work that has been carried out at the University of Warwick as part of the CyVerse UK initiative and made 14 popular tools much more available and easy to use.

2 Tool Selection

Readily available tools on large cloud computing services predominantly center around high-throughput sequencing data analyses. Tools for complex analyses of expression data and regulatory sequences are underrepresented, while often requiring a large computational overhead in the case of more complex methodology. As such, a number of previously published, locally created methods were selected to produce a well-rounded time course expression data analysis package for inclusion into CyVerse. Differential expression can be handled with GP2S (Stegle *et al.*, 2010) in a typical control-treated scenario, or the gradient tool (Breeze *et al.*, 2011) to identify timing of first change in single condition datasets. Clustering can be performed with BHC (Cooke *et al.*, 2011) for a Bayesian hierarchical approach, TCAP (Kiddle *et al.*, 2010) to obtain complex regulatory modules with a rich information metric capable of capturing inversions and time shifts, or Wigwams (Polanski *et al.*, 2014) for co-regulation across subsets of multiple datasets. Transcription factor binding site overrepresentation of the resulting gene groups can be done with known sequences using the hypergeometric motif test (Breeze *et al.*, 2011) or *de novo* with MEME-LaB (Brown *et al.*, 2013), while a BiNGO-friendly (Maere *et al.*, 2005) output is created for GO term overrepresentation analyses to be carried out locally in Cytoscape. Causal network inference can be performed using CSI (Penfold and Wild, 2011), as well as its extensions to handle multiple data sets in a hierarchical structure (Penfold *et al.*, 2012) or across multiple species (Penfold *et al.*, 2015). The flow of the analysis of a time course dataset with the provided tools is shown in panel A of Figure S1. Other algorithms related to gene expression analysis have also been integrated: the reverse best hit orthologue detection and conserved promoter functionality of the APPLES suite (Baxter *et al.*, 2012), as well as footprint identification in DNase-seq data by Wellington (Piper *et al.*, 2013) along with its differential analysis extension Wellington-bootstrap (Piper *et al.*, 2015).

3 Deployment Standards

The first step in the preparation of each tool was to create efficient, stable implementations in freely available programming languages where this was not already the case. As such, the algorithms previously only available in Matlab (gradient tool, TCAP, Wigwams, CSI, hCSI, oCSI) were re-coded into Python, greatly decreasing their run time. Other algorithms which were previously available outside of Matlab (GP2S, hypergeometric motif test, APPLES) had their code bases refined to increase stability and user friendliness. Once implementations of adequate quality were available, the algorithms were housed in standalone Docker containers (Merkel, 2014). This makes the methods future-proof, with dependency problems solved proactively by encapsulating functional versions of the programs. This has already proven advantageous, with GP2S requiring a very particular Python 2.7 setup to function properly. The resulting Docker containers were imported into the CyVerse Discovery Environment and graphical, user-friendly apps were created. These can all be found by searching for 'uk cyverse' in the app window, and the algorithms all run remotely on CyVerse hardware. Each app links to exhaustive documentation and a set of test data, detailing in great depth how to format the input and what each available parameter is responsible for. The output is created with user friendliness in mind, and usually features some form of visualisation, often as interactive webapps. Excerpts of visual output produced by the programs can be seen in panels B and C of Figure S1.The complete results of each analysis get compressed into a single archive for ease of downloading to a local machine and further investigation. A tutorial on using the tools in the time course expression data pipeline and chaining these together using helper apps has also been created, and can be accessed at https://github.com/cyversewarwick/expression_tutorial. Every stage of the development process detailed above is publicly available - the raw code is housed on GitHub, which is chained to a DockerHub account, automatically rebuilding individual repositories when they get updated, with the newest versions of the images being used by CyVerse apps.

Funding

This work has been supported by the Biotechnology and Biological Sciences Research Council grant BB/M018431/1

Conflict of Interest: none declared.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, **11**(10), 1.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N. P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., *et al.* (2012). Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell*, **24**(10), 3949–3965.

Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y.-s., Penfold, C. A., Jenkins, D., *et al.* (2011). High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation. *The Plant Cell*, **23**(3), 873–894.

Brown, P., Baxter, L., Hickman, R., Beynon, J., Moore, J. D., and Ott, S. (2013). MEME-LaB: motif analysis in clusters. *Bioinformatics*, **29**(13), 1696–1697.

Cooke, E. J., Savage, R. S., Kirk, P. D., Darkins, R., and Wild, D. L. (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC bioinformatics*, **12**(1), 399.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10), R80.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., *et al.* (2011). The iPlant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science*, **2**, 34.

Hillman-Jackson, J., Clements, D., Blankenberg, D., Taylor, J., Nekrutenko, A., and Team, G. (2012). Using galaxy to perform large-scale interactive data analyses. *Current protocols in bioinformatics*, pages 10–5.

Kiddle, S. J., Windram, O. P., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., Denby, K. J., and Mukherjee, S. (2010). Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana. *Bioinformatics*, **26**(3), 355–362.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**(16), 3448–3449.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, **2014**(239), 2.

Penfold, C. A. and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface focus*, **1**(6), 857–870.

Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J., and Wild, D. L. (2012). Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, **28**(12), i233–i241.

Penfold, C. A., Millar, J. B., and Wild, D. L. (2015). Inferring orthologous gene regulatory networks using interspecies data fusion. *Bioinformatics*, **31**(12), i97–i105.

Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, page gkt850.

Piper, J., Assi, S. A., Cauchy, P., Ladroue, C., Cockerill, P. N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC genomics*, **16**(1), 1.

Polanski, K., Rhodes, J., Hill, C., Zhang, P., Jenkins, D. J., Kiddle, S. J., Jironkin, A., Beynon, J., Buchanan-Wollaston, V., Ott, S., *et al.* (2014). Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics*, **30**(7), 962–970.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, **17**(3), 355–367.

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., *et al.* (2012). Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *The Plant Cell*, **24**(9), 3530–3557.