Bioinformatics, 36(12), 2020, 3916–3917
doi: 10.1093/bioinformatics/btaa223
Advance Access Publication Date: 31 March 2020
Applications Note

Systems biology

corto: a lightweight R package for gene network inference and master regulator analysis

Daniele Mercatelli[1], Gonzalo Lopez-Garcia[2] and Federico M. Giorgi [1,*]

[1]Department of Pharmacy and Biotechnology, University of Bologna, Bologna 40126, Italy and [2]Genetics and Genomics Science Department, Ichan School of Medicine at Mount Sinai, New York City, NY 10029-5674, USA

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received and revised on January 30, 2020; editorial decision on March 24, 2020; accepted on March 26, 2020

Abstract

Motivation: Gene network inference and master regulator analysis (MRA) have been widely adopted to define specific transcriptional perturbations from gene expression signatures. Several tools exist to perform such analyses but most require a computer cluster or large amounts of RAM to be executed.

Results: We developed *corto*, a fast and lightweight R package to infer gene networks and perform MRA from gene expression data, with optional corrections for copy-number variations and able to run on signatures generated from RNA-Seq or ATAC-Seq data. We extensively benchmarked it to infer context-specific gene networks in 39 human tumor and 27 normal tissue datasets.

Availability and implementation: Cross-platform and multi-threaded R package on CRAN (stable version) https://cran.r-project.org/package=corto and Github (development release) https://github.com/federicogiorgi/corto.

Contact: federico.giorgi@unibo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of high-throughput methods to quantify transcript abundances has offered the possibility to measure gene co-expression across hundreds of samples and thousands of genes. The principle of co-expression has fueled the generation of several gene regulatory network representations over the past decade, and still constitutes the main source of genome-wide network inference pipelines (Mercatelli *et al.*, 2019a). Recently, scientists have developed tools to further leverage gene network models and interrogate them via MRA to identify regulatory subnetworks active in specific experimental conditions, tumor subtypes or even individual patients (Ding *et al.*, 2018), or to improve the readout in noisy or single-cell datasets by aggregating genes (Mercatelli *et al.*, 2019b).

The majority of current tools for gene network inference require either a computing cluster to be executed and/or a high amount of RAM. Therefore, we developed a lightweight R package-dubbed *corto* ('correlation tool') to infer significant, direct edges between a user-provided list of source genes ('centroids'), such as transcription factors (TFs). The algorithm underneath *corto* infers direct TF–target relationships by applying DPI on correlation triplets, as proposed in Reverter and Chan (2008). Our tool provides networks as Bioconductor *regulon* objects that can be immediately used in MRA by *corto* itself or by other tools, e.g. the VIPER pipeline (Alvarez *et al.*, 2016). We benchmarked *corto* across dozens of transcriptomics datasets, with respect to other analogous tools.

2 Functionalities

The *corto* algorithm expands the well-established pipeline of the public Java tool ARACNe-AP (Lachmann *et al.*, 2016), and allows the user to perform downstream MRA and visualization of master regulators. The R implementation of its algorithms is fully multi-threaded, with a user-friendly progress bar indicating the estimated time to completion. The functionalities of the package can be categorized as follows:

1. Gene network inference. This is based on optimized pairwise correlation, DPI and bootstrapping to evaluate the significant edges. Further details on the inference algorithm are available on the package vignettes and on the Github page.
2. Copy-number variation (CNV) correction. As the presence of CNVs can influence and bias the generation of gene networks (Schubert *et al.*, 2019), *corto* gives the optional possibility to use CNV data to correct target expression profiles via linear regression. An example of *corto* CNV-corrected network analysis in the The Cancer Genome Atlas(TCGA) Glioblastoma dataset is provided in the package vignette and in Supplementary Figure S1.
3. MRA. The *mra* function within our package calculates the enrichment of each TF-centered network on a user-selected signature, provided as two gene expression matrices (e.g. treatment

versus control). This kind of analysis is exemplified in detail in Supplementary Vignette 1.

4. Visualization of network enrichment. *corto* can generate MRA plots as visualized in Figure 1. For each most significant or user-specified centroid, the function visualizes the distribution of its targets across a signature, as red (positively correlated targets) or blue (negatively correlated) vertical bars. The normalized enrichment score and the corresponding *P*-value (based on permutation tests and signature sample shuffling) are shown. Also, the most correlated targets are shown as a mini network, connected to the centroid with pointed arrows (positively correlated) or blunted arrows (negatively correlated), and shown as red or blue if activated or repressed in the provided contrast.

Classic MRA is based on a provided gene network and on a gene expression differential signature, based on microarrays or RNA-Seq. To run *corto* on these data types, we suggest to use RMA and VST normalization, which have been shown to be optimal for co-expression analyses (Giorgi *et al.*, 2013). On top of these data types, *corto* can operate also on ATAC-Seq-derived signatures, which measure the differential chromatin status between two conditions, as shown in Supplementary Vignette 2.

3 Benchmarking

We tested *corto* extensively against the Java tool ARACNe-AP (Lachmann *et al.*, 2016) and other R-based tools for gene network inference, such as minet (Meyer *et al.*, 2008) and RTN (Castro *et al.*, 2016). Our analysis shows that *corto* is consistently at least 10× faster than its competitors even in single-thread mode, obtaining significantly similar networks (Supplementary Vignette 3).

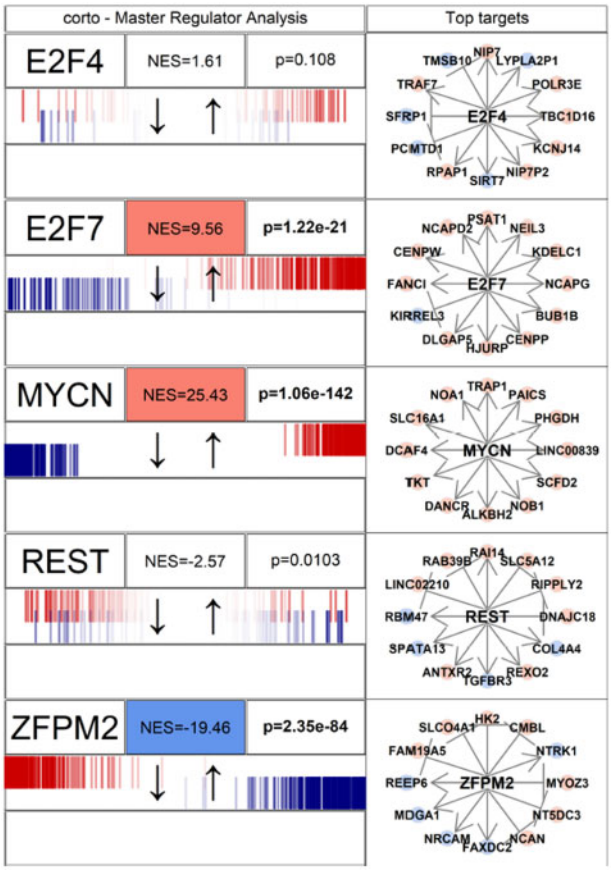

Fig. 1. Example of MRA performed by *corto*

Using the ENCODE ChIP-Seq dataset, we estimated the accuracy of *corto*, which is consistently above 75% in all tested cell lines (Supplementary Benchmark). We also detected a high similarity in MRA analyses executed with networks generated by *corto* and the other tool generating regulon objects, ARACNe-AP (Supplementary Vignette 4). More examples and details on *corto* and MRA are available in Supplementary File S1. In Figure 1, we tested a *corto*-generated neuroblastoma network (Kocak *et al.*, 2013) on a MYCN amplified versus not amplified signature, and visualize the results for selected TF networks. Unsurprisingly (more details in Supplementary Vignette 1), *corto* shows MYCN network amongst the most upregulated ones, together with the pro-proliferative E2F7, while E2F4 and REST are not affected within significance. On the other hand, the pro-differentiation TF ZFPM1 seems to be downregulated by the MYCN amplification. Top targets for each TF are shown.

4 Conclusions

We propose *corto* as a novel, lightweight and robust tool for rapid gene network inference and MRA in large-scale transcriptomics datasets. While benchmarked here in human RNA-Seq and ATAC-Seq data, there is no intrinsic limit to *corto* applications, which can be naturally extended to other organisms and quantitative biological data, such as proteomics datasets.

Funding

This research was supported by CINECA under the ISCRA initiative, grant numbers HP10CPQJBV and HP10CC5F89.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Alvarez,M.J. *et al.* (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, **48**, 838–847.

Castro,M.A. *et al.* (2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.*, **48**, 12–21.

Ding,H. *et al.* (2018) Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.*, **9**, 1471.

Giorgi,F.M. *et al.* (2013) Comparative study of RNA-seq-and microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics*, **29**, 717–724.

Kocak,H. *et al.* (2013) Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma. *Cell Death Dis.*, **4**, e586.

Lachmann,A. *et al.* (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, **32**, 2233–2235.

Mercatelli,D. *et al.* (2019a) Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1863**,194430–194444.

Mercatelli,D. *et al.* (2019b) Pan-cancer and single-cell modelling of genomic alterations through gene expression. *Front. Genet.*, **1863**, 671.

Meyer,P.E. *et al.* (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Reverter,A. and Chan,E.K. (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**, 2491–2497.

Schubert,M. *et al.* (2019) Gene networks in cancer are biased by aneuploidies and sample impurities. *Biochim. Biophys. Acta Gene Regul. Mech.*, 194444.