TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data

Luca De Sano*,[1], Giulio Caravagna*,[1], Daniele Ramazzotti*,[1], Alex Graudenzi[1], Giancarlo Mauri[1], Bud Mishra[2], and Marco Antoniotti[1,3]

[1]Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy.
[2]Courant Institute of Mathematical Sciences, New York University, New York, USA.
[3]Milan Center for Neuroscience, University of Milan-Bicocca, Italy.

Abstract

Motivation: We introduce TRONCO (TRanslational ONCOlogy), an open-source R package that implements the state-of-the-art algorithms for the inference of cancer progression models from (epi)genomic mutational profiles. TRONCO can be used to extract population-level models describing the trends of accumulation of alterations in a cohort of cross-sectional samples, e.g., retrieved from publicly available databases, and individual-level models that reveal the clonal evolutionary history in single cancer patients, when multiple samples, e.g., multiple biopsies or single-cell sequencing data, are available. The resulting models can provide key hints in uncovering the evolutionary trajectories of cancer, especially for precision medicine or personalized therapy.

Availability: TRONCO is released under the GPL license, and it is hosted at (Software section) http://bimib.disco.unimib.it/ and archived also at bioconductor.org.

Contact: tronco@disco.unimib.it

1 Introduction

Cancer develops through the successive expansions of clones in which certain (epi)genomic alterations, called *drivers*, confer a *fitness* advantage and progressively accumulate, in a context of overall scarcity of resources [4]. Therefore, it is possible to define *cancer progression models*, in the form of *probabilistic causal graphical models*, in which the conditional dependencies and the temporal ordering among these alterations is shown, revealing the likely evolutionary trajectories of cancer at the (epi)genome level.

We further distinguish [2]: (*i*) *ensemble-level* progression models, describing the statistical trends of accumulation of genomic alterations in a cohort of distinct cancer patients, and (*ii*) *individual-level* models, thus accounting for the specific evolutionary history of cancer clones in individual tumors.

*Equal contributors

Even if the inference of such models is further complicated by a series of theoretical and technical hurdles, such as, e.g., *intra-* and *inter-tumor heterogeneity* and the effective detection of drivers, it can benefit from the the increasing amount of *next-generation sequencing* (NGS) data, currently available through public projects such as The Cancer Genome Atlas (TCGA, `https://tcga-data.nci.nih.gov`). Usually, in such databases *cross-sectional* (epi)genomic profiles retrieved from single biopsies of cancer patients are provided, which can be used to extract ensemble-level models, yet higher resolution data such as *multiple-biopsies* or even *single-cell* sequencing data are bbecoming progressively more accessible and reliable, which can be used to infer individual-level models.

We here introduce `TRONCO` (TRanslational ONCOlogy), an R package aimed at the inference of cancer progression models from heterogeneous genomic data, in the form of alterations persistently present along tumor evolution. Currently, `TRONCO` provides the implementation of two distinct algorithms: (*i*) `CAPRESE` (CAncer PRogression Extraction with Single Edges [5]), and (*ii*) `CAPRI` (CAncer PRogression Inference [6]), both based on Suppes' theory of *probabilistic causation* [8], yet with distinct goals and properties (see Software Implementation).

`TRONCO`, in its current form and in perspective, should be thought of as a tool to provide the community with the implementation of up-to-date solutions to the progression inference problem. At the time of this writing, for instance, it is the ideal conclusive stage of a versatile and modular pipeline for the extraction of ensemble-level cancer progression models from cross-sectional data [2]. In such a pipeline input data are pre-processed to (*i*) stratify samples in tumor subtypes, (*ii*) select driver alterations and (*iii*) identify groups of fitness-equivalent (i.e., mutually exclusive) alterations, prior to the application of `CAPRI` algorithm. The resulting ensemble-level progression models can provide important indications on the evolutionary dynamics of cancer, with key repercussions on diagnostic and therapeutic spheres, especially in regard to precision medicine and personalized drug development.

From the complementary perspective, `TRONCO` can also exploit the `CAPRESE` algorithm to infer the clonal evolutionary history in single patients when multiple samples are available, as in the case of multiple biopsies and/or single-cell sequencing data, as long as the set of driver events is selected; see [2].

2 Software implementation

`TRONCO` implements a set of R functions to aid the user to extract a cancer progression model from genomic data. At a high-level, these function shall help to import, visualize and manipulate genomic profiles - regardless of their source of origin - eventually allowing to run the algorithms implemented in the package, and assess the confidence of a model.

Data loading and manipulation. Common formats to store data used to extract progression models can be natively imported in the tool. These include, for instance, the *Mutation Annotation Format* (MAF) for somatic mutations, as well as the *Genomic Identification of Significant Targets in Cancer* (GISTIC) format to store focal Copy Number Variations. The tool can exploit the Cbio portal for Cancer Genomics, which collects among others TCGA projects, to access freely available instances of such data [3].

`TRONCO` provides function for the preprocessing of the data in order to tidy them, i.e., select a certain subset of alterations, or samples or any abstraction which might be appropriate according to the cancer under study.

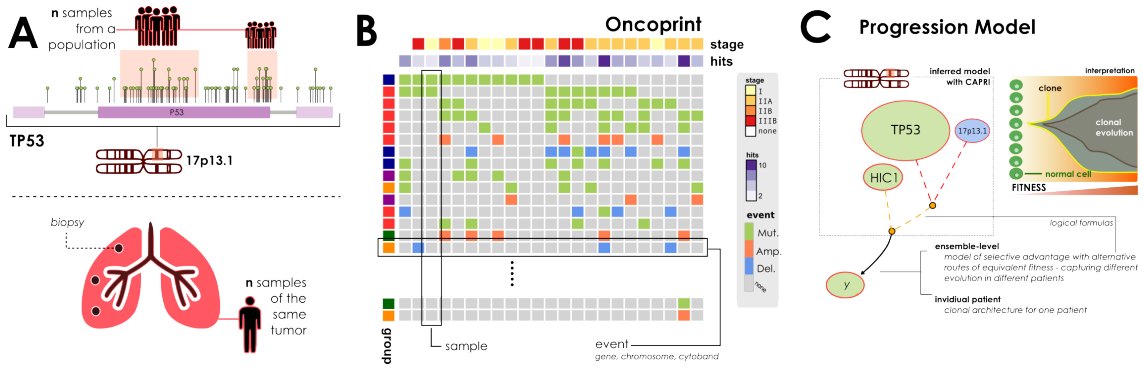

Figure 1: **(A)** TRONCO can process either alterations (e.g., somatic mutations or wider chromosomal lesions) in a cohort of independent samples (top lolliplot diagram), or a set of multiple snapshots from a unique patient (e.g., multi-region or single-cell, bottom panel). **(B)** *Oncoprints* allow the user to visualize the data that the tool is going to process. Regardless of the source, each row will represent a certain alteration - at a custom resolution which depend on the cancer under study - and each column a sample. **(C)** A model inferred with the tool might outline as cancer evolution might happen in the input ensemble or in a individual patient. Here, we show a hypothetical ensemble-level model predicting a selection pressure on two genes mapped to 17p13, TP53 and HIC1, as it may be inferred by analyzing samples harbouring either TP53/HIC1 mutations or homozygous deletions in the cytoband where any of these two genes map, i.e., here for purely explanatory cases we suppose just TP53, which maps to 17p13.1. The model suggests a trend of selection towards mutations in gene Y, which shall be interpreted as a set of preferential clonal expansions, which should be characteristic of the population of analyzed samples, involving alterations of the functions mapped to 17p13 and Y.

Visualization and interaction with other tools. TRONCO implements an *oncoprint* system to visualize and export the data that it is processing, see Figure 1. Datasets can be exported in formats for processing by other tools commonly used to, e.g., stratify input samples and detect groups of mutually exclusive alterations, which include the *Network Based Stratification* [9] and *MUTEX* [1] tools. By using other graph-plotting packages, TRONCO provides also function for visualization of the inferred models.

Model inference and confidence. TRONCO implements two distinct algorithms: (*i*) CAPRESE, which uses a *shrinkage*-like estimator to infer *tree*-models of progression, and (*ii*) CAPRI, which extracts more general *direct acyclic graphs (DAG)* - thus allowing for confluent evolution and complex hypothesis testing - by combining *bootstrap* and *maximum likelihood estimation* with regularization. A posteriori, TRONCO also allows to assess the confidence of the inferred models via (*i*) non-parametric, (*ii*) parametric and (*iii*) statistical bootstrap.

3 Discussion

TRONCO aims to become the bioinformatics reference tool of choice providing the community with the implementation of up-to-date statistical routines to understand the evolutionary trajectories of any cancer. Indeed, in terms of computational speed, scalability with respect to sample size, predictive accuracy and robustness against noise in the input data, the algorithms implemented in the tool is demonstrably to be the current state-of-the-art for the inference problem. The implementation in R makes straightforward the interaction of TRONCO with other common bioinformatics tools, possibly allowing the creation of a common suite of tools for cancer progression inference. The porting of the tool to other widely used bioinformatics platforms such as, e.g., Cytoscape [7] is under development.

Financial support. MA, GM, GC, AG, DR acknowledge Regione Lombardia (Italy) for the research projects RetroNet through the ASTIL Program [12-4-5148000-40]; U.A 053 and Network Enabled Drug Design project [ID14546A Rif SAL-7], Fondo Accordi Istituzionali 2009. BM acknowledges founding by the NSF grants CCF-0836649, CCF-0926166 and a NCI-PSOC grant.

References

[1] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, 16(1), 2015.

[2] G. Caravagna, A. Gradenzi, D. Ramazzotti, R. Sanz-Pamplona, G. Mauri, V. Moreno, M. Antoniotti, and B. Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Submitted, http://dx.doi.org/10.1101/027359*, 2015.

[3] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.

[4] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[5] Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PLOS ONE*, 9(12):e115570, 2014.

[6] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

[7] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[8] Patrick Suppes. *A Probabilistic Theory of Causality*. North-Holland Publishing Company Amsterdam, 1970.

[9] Xue Zhong, Hushan Yang, Shuyang Zhao, Yu Shyr, and Bingshan Li. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC Genomics*, 16(Suppl 7):S7, 2015.