GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework

Saket Kumar Choudhary and Santosh B Noronha

bioRxiv first posted online October 19, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/010538

GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework

Saket K Choudhary and Santosh B Noronha

{saket.kumar, noronha}@iitb.ac.in
Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai, India

Abstract

Identification of driver mutations can lead to a better understanding of the molecular mechanisms associated with cancer. This can be a first step towards developing diagnostic and prognostic markers. Various driver mutation prediction tools rely on different algorithm for prediction and hence there is little consensus in the predictions. The input and output formats vary across the tools. It has been suggested that an ensemble approach that takes into account various prediction scores might perform better. There is a need for a tool that can run multiple such tools on a dataset in a more accessible and modular manner, whose output can then be combined to select consensus drivers.

We developed wrappers for various driver mutation predictions tools using Galaxy based framework. In order to perform predictions using multiple tools on the same dataset, we also developed Galaxy based workflows to convert VCF format to tool specific formats. The tools are publicly available at: `https://github.com/saketkc/galaxy_tools` *The workflows are available at:* `https://github.com/saketkc/galaxy_tools/tree/master/workflows`

I. Introduction

Cancer is known to be arise from genomic aberrations[1]. With the advent of Next Generation Sequencing, it has been possible to profile the genomes from cancer patients and perform an in-depth analysis to yield insights that could be used for therapeutic, diagnostic and prognostic applications. Many of these profiles have also been released inn public domain, such as TCGA [2]. Given huge datasets such as these, the challenge has been to make sense out of them.

It has been suggested that Cancer is the outcome of Darwinian evolution at the cell level involving genetic variation. Natural selection acts on the phenotype level, thus positively selecting the cells which have acquired those mutations, that allow such cells to proliferate. Cells with such mutations lead to abnormal growth in the tumors. This growth may turn out to be controlled such as in the case of skin moles or it may end up leading to cancer tissues

These set of mutations are the 'drivers' of cancer that confer selective advantages to the cell to grow autonomously or alternatively to prevent cell death by affecting the apoptosis pathways ultimately leading to the positive selection of the cell. [3].

The 'passengers' do not confer any growth advantage to the cells.'Drivers' thus, by definition are found in 'cancer' genes. the 'passengers' are known to be distributed randomly [3]. The 'cancer genes' are known to contribute

Identification of these set of driver mutations across various cancer types, would possibly act as a set of prognosis markers be-

sides acting like therapeutic targets. Different approaches have been used to differentiate drivers from passengers. The challenge lies in correctly classifying thousands of mutations generated out of whole genome or exome sequencing as drivers or passengers.

Multiple approaches have been used for predicting driver mutations. These use information ranging from difference in mutational frequency among drivers and passengers to calculating functional impact scores , besides using a curated dataset of driver and passenger mutation for training classifiers [4].

Different tools use different methods and input formats. This often makes the task of biologists difficult. These tools, owing to different underlying approach used for prediction give non-consensus predictions for the same set of mutations. Galaxy [5, 6, 7] is an open source framework that allows running Bioinformatics tools in a reproducible manner. Galaxy thus provides a user friendly graphical interface accessible through a web browser for running such tools in a reproducible manner. Galaxy is used extensively for anlayzing next-generation sequencing datasets[8]. However few driver mutation predictor tools have been developed for Galaxy.

II. Methods

We developed wrappers around tools described in Table 1 leveraging Galaxy's Toolshed framework[9]. These tools are freely available from the toolshed `https://toolshed.g2.bx.psu.edu/` and are open source.

One of the standard formats used to store mutation data is the VCF(Variant Call Format) [10]. In order to streamline the process of generating tool specific inputs, we developed workflows in Galaxy that take a VCF as an input and convert it to the tool specific input format.

There have been previous attempts at comparing such softwares [11] [4]. It has also been suggested that an aggregate approach such as the Condel[12] program which uses a weighted average score of SIFT, Polyphen and Mutation Assessor to make predictions and is shown to outperform each individual method. GalDrive aims at streamlining the process of comparing the methods by providing tools in Galaxy that can be used to run such comparison pipelines in a reproducible manner

III. Results & Discussion

The GalDrive toolbox provides set of tools and workflows for running comparative pipelines for driver mutation prediction in a reproducible and flexible manner. Utilizing Galaxy's set of internal tools with certain custom developed tools we created workflows that allows direct utilization of VCF file, which is a more standard format as inputs to these workflows. This can serve as a powerful tool to the biologists, who might not be acquainted with command line tools to run the methods, rest aside the need to perform various format interconversions.

These workflows can further be utilized to generate ensemble scores by combining output scores of two or more tools' output.

IV. Acknowledgments

We would like to thank the Galaxy Team for their valuable help on the mailing-list during the process of development.

References

[1] T. Sjöblom, S. Jones, *et al.*, "The consensus coding sequences of human breast and colorectal cancers.," *Science (New York, N.Y.)*, vol. 314, pp. 268–74, Oct. 2006.

[2] J. N. Weinstein, E. a. Collisson, *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project.," *Nature genetics*, vol. 45, pp. 1113–20, Oct. 2013.

[3] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome.," *Nature*, vol. 458, pp. 719–24, Apr. 2009.

[4] J. Zhang, J. Liu, *et al.*, "Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing.," *Briefings in bioinformatics*, vol. 15, pp. 244–55, Mar. 2014.

[5] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational

Table 1: *Tools Wrapped in Galaxy*

Tool Name	ToolShed URL	Reference
CHASM	`https://toolshed.g2.bx.psu.edu/view/saket-choudhary/chasm_webservice`	[13]
FATHMM	`https://toolshed.g2.bx.psu.edu/view/saket-choudhary/fathmm_web`	[14, 15, 16]
Mutation Assessor	`https://toolshed.g2.bx.psu.edu/view/saket-choudhary/mutationassessor_web`	[17, 18]
SIFT	`https://testtoolshed.g2.bx.psu.edu/view/saket-choudhary/sift_web`	[19, 20, 21, 22, 23]
Polyphen2	`https://testtoolshed.g2.bx.psu.edu/view/saket-choudhary/polyphen2`	[24]

research in the life sciences," *Genome Biol*, vol. 11, no. 8, p. R86, 2010.

[6] D. Blankenberg, G. V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," *Current protocols in molecular biology*, pp. 19–10, 2010.

[7] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. C. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[8] D. Blankenberg and J. Hillman-Jackson, "Analysis of next-generation sequencing data using galaxy," in *Stem Cell Transcriptional Networks*, pp. 21–43, Springer, 2014.

[9] D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor, and A. Nekrutenko, "Dissemination of scientific software with Galaxy ToolShed.," *Genome biology*, vol. 15, p. 403, Jan. 2014.

[10] F. F. T. Team, "Variant call format." `http://samtools.github.io/hts-specs/VCFv4.1.pdf`.

[11] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, and Z. Zhang, "Assessment of computational methods for predicting the effects of missense mutations in human cancers.," *BMC genomics*, vol. 14 Suppl 3, p. S7, Jan. 2013.

[12] A. González-Pérez and N. López-Bigas, "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.," *American journal of human genetics*, vol. 88, pp. 440–9, Apr. 2011.

[13] C. Douville, H. Carter, R. Kim, N. Niknafs, M. Diekhans, P. D. Stenson, D. N. Cooper, M. Ryan, and R. Karchin, "CRAVAT: cancer-related analysis of variants toolkit.," *Bioinformatics (Oxford, England)*, vol. 29, pp. 647–8, Mar. 2013.

[14] H. a. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. a. Barker, K. J. Edwards, I. N. M. Day, and T. R. Gaunt, "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.," *Human mutation*, vol. 34, pp. 57–65, Jan. 2013.

[15] H. a. Shihab, J. Gough, D. N. Cooper, I. N. M. Day, and T. R. Gaunt, "Predicting the functional consequences of cancer-associated amino acid substitutions.," *Bioinformatics (Oxford, England)*, vol. 29, pp. 1504–10, June 2013.

[16] H. a. Shihab, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, and T. R. Gaunt, "Ranking non-synonymous single nucleotide polymorphisms based on disease concepts.," *Human genomics*, vol. 8, p. 11, Jan. 2014.

[17] B. Reva, Y. Antipin, and C. Sander, "Determinants of protein function revealed by combinatorial entropy optimization.," *Genome biology*, vol. 8, p. R232, Jan. 2007.

[18] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics.," *Nucleic acids research*, vol. 39, p. e118, Sept. 2011.

[19] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.," *Nature protocols*, vol. 4, pp. 1073–81, Jan. 2009.

[20] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions.," *Genome research*, vol. 11, pp. 863–74, May 2001.

[21] P. C. Ng and S. Henikoff, "Accounting for human polymorphisms predicted to affect protein function.," *Genome research*, vol. 12, pp. 436–46, Mar. 2002.

[22] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function.," *Annual review of genomics and human genetics*, vol. 7, pp. 61–80, Jan. 2006.

[23] P. C. Ng, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, pp. 3812–3814, July 2003.

[24] I. a. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations.," *Nature methods*, vol. 7, pp. 248–9, Apr. 2010.