CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer

Wing Chung Wong1,†, Dewey Kim1,†, Hannah Carter1, Mark Diekhans2, Michael C. Ryan3 and Rachel Karchin1,∗

1Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, 2Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA and 3In Silico Solutions, Fairfax, VA, USA

1 INTRODUCTION

A fundamental goal of modern cancer genomics studies is to understand how alterations in DNA sequence contribute to disease susceptibility and prognosis. Targeted whole-exome deep sequencing is now affordable for many academic labs and the multitude of studies underway is yielding datasets of unprecedented magnitude. While researchers have previously developed methods to computationally predict the impact of single nucleotide variants (SNVs) of possible relevance to disease etiology. Here, we describe a software package that implements the CHASM method. The package includes a database of pre-computed predictive features called SNVBox that facilitates rapid feature retrieval and classification of very large SNV datasets. Furthermore, the features in SNVBox can be generally used to aid in the development of new classification algorithms that predict the impact of either germline or somatic SNVs.

2 METHODS AND IMPLEMENTATION

CHASM is an open-source collection of Python and C++ programs that takes a list of somatic missense mutations as input and ranks them according to their likely tumorigenic impact. It includes a curated set of driver mutations culled from the COSMIC database (Forbes et al., 2008), which is used as a positive class for training a Random Forest classifier (Amit and Geman, 1997; Breiman, 2001). The negative class of mutations is generated in silico according to an estimated distribution of benign (passenger) variation, matched to the tumor type of interest. Users have the option to use their own estimates of passenger variant frequencies or to select from a library of pre-computed passenger frequency tables for several common cancers.

Pyninstaller 1.4 was used to package Python source into dynamically linked, executable binaries. The SnvGet, Build Classifier and RunChasm executables are run by the user on the command line, while the others are called internally. The statically compiled C++ executable waffles_learn from the WAFFLES machine learning library is also called internally.

SNVBox is an MySQL database of 86 predictive features relevant to the biological impact of an SNV. The features have been pre-computed for each codon in all protein-coding exons of annotated human mRNA transcripts in the NCBI RefSeq, CCDS and EBI Ensembl databases (Birney et al., 2004; Pruitt et al., 2007, 2009). The SnvGet program enables fast retrieval of selected features from the database for classifier training and scoring of mutations input by the user.

3 WORKFLOW

(1) Prepare an input file of estimated passenger mutation rates in the cancer of interest. Optionally, select from one of several pre-computed passenger rate tables.
The CHASM/SNVBox toolkit is the first distributable software developed to allow users to run on their local custom datasets (e.g. thousands to millions of SNVs discovered in sequencing projects). Some researchers have handling large size custom datasets (reviewed in Karchin (2009)), these are not currently capable of estimates of the spectrum (Supplementary Material).

Variant calls from their sequencing data to make the best possible type of interest. CHASM results are sensitive to this definition of cancer. The learning task of the Random Forest classifier is to 2148

4 DISCUSSION

The CHASM/SNVBox toolkit is the first distributable software package that specifically targets somatic missense mutations in cancer. The learning task of the Random Forest classifier is to discriminate between known drivers and a set of random passenger missense mutations that match the mutation spectrum in a cancer type of interest. CHASM results are sensitive to this definition of mutation spectrum and users are encouraged to use the somatic variant calls from their sequencing data to make the best possible estimates of the spectrum (Supplementary Material).

While many SNV classifiers are available through web interfaces reviewed in Karchin (2009), these are not currently capable of handling large size custom datasets (e.g. thousands to millions of SNVs discovered in sequencing projects). Some researchers have developed distributable packages that users can run on their local system to enable high-throughput SNV processing. These packages depend on third-party databases (sequences, alignments, protein structures, specialized protein annotations) and third-party software packages. The popular PolyPhen system, for example, requires installation of 10 third-party software packages, in addition to three Perl modules. To our knowledge, all available SNV classification tools base their inferences on predictive features computed when a custom dataset is input to the system (almost always using third-party databases and software). In contrast, the predictive features available in SNVBox (also calculated with many third-party tools) have been exhaustively pre-computed, allowing rapid retrieval for a custom dataset. In benchmark testing, retrieval of 86 features for one million SNVs took 11.39h on a Dell R900 server with two AMD Opteron dual-core 64 bit CPUs and 16 GBs of RAM. CHASM score computation for these one million mutations took an additional 10 min and 33 s.

Finally, the predictive features available in SNVBox were designed to be useful for classification of both germline and somatic SNVs. We hope that SNVBox will enable design of new, improved machine learning algorithms to predict the impact of SNVs.

Funding: National Cancer Institute, National Institutes of Health grant (CA135877 to R.K., in part); National Science Foundation CAREER award (DBI 0845275 to R.K., in part); DOD NDSEG Fellowship 32 CFR 168a (to H.C., in part).

Conflict of Interest: none declared.

REFERENCES


