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

ABSTRACT

Summary: Thousands of cancer exomes are currently being sequenced, yielding millions of non-synonymous single nucleotide variants (SNVs) of possible relevance to disease etiology. Here, we provide a software toolkit to prioritize SNVs based on their predicted contribution to tumorigenesis. It includes a database of precomputed, predictive features covering all positions in the annotated human exome and can be used either stand-alone or as part of a larger variant discovery pipeline.

Availability and Implementation: MySQL database, source code and binaries freely available for academic/government use at http://wiki.chasmsoftware.org. Source in Python and C++, Requires 32 or 64-bit Linux system (tested on Fedora Core 8,10,11 and Ubuntu 10), 2.5 ≤ Python < 3.0, MySQL server > 5.0, 60 GB available hard disk space (50 MB for software and data files, 40 GB for MySQL database dump when uncompressed), 2 GB of RAM.

Contact: karchin@jhu.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on April 13, 2011; revised on June 1, 2011; accepted on June 8, 2011

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) (Kaminker et al., 2007; Mooney et al., 2010; Ng and Henikoff, 2003; Sunyaev et al., 2001), to our knowledge there are no existing tools capable of fast classification of very large SNV datasets in cancer exomes.

We have previously developed a computational method Cancer-Specific High-throughput Annotation of Somatic Mutations (CHASM) (Carter et al., 2009, 2010) that predicts whether tumor-derived somatic missense mutations are important contributors to cancer cell fitness. 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.

PyInstaller 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 precomputed for each codon in all protein-coding exons of annotated human mRNA transcripts in the NCBI RefSeq, CCDS and EBI Ensembl databases (Bateman 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 Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
(2) Prepare an input file of missense SNVs to be classified. Each row contains a protein accession identifier, codon number, and reference and variant amino acid residues.

(3) Run the BuildClassifier program.
   • Generates a negative class of in silico passenger mutations by random nucleotide substitution in a library of expressed human mRNA transcripts from NCBI RefSeq, according to the distributions specified in the passenger mutation rate table (Supplementary Material).
   • Retrieves a list of predictive features for each passenger (and driver) in the training set from SNVBox.
   • Builds a Random Forest classifier, using waftles_learn.

(4) Run the RunChasm program.
   • Retrieves a feature list for all mutations supplied by the user.
   • Applies the trained classifier to generate a CHASM score for each variant.
   • Generates a second set of in silico passenger mutations, which (unlike the first set) is carefully filtered to eliminate mutations in any genes previously associated with cancer in either the Cancer Gene Census (Futreal et al., 2004), the COSMIC cancer gene list and all cancer (CA collection) genesets in MSigDB (Subramanian et al., 2005).
   • Filtered passengers are scored by the classifier to produce an empirical null distribution of variant scores.
   • This null score distribution is used to compute a P-value for each variant supplied by the user (fraction of filtered passengers having CHASM scores less than or equal to the score of the variant).
   • Benjamini-Hochberg multiple testing correction (Benjamini and Hochberg, 1995) is applied to the P-values.
   • Outputs a list of the user-supplied mutations, with CHASM scores, P-values and Benjamini-Hochberg estimated false discovery rate (FDR).
   • Outputs an ARRF formatted file of features for the submitted mutations.

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 type of interest. CHASM results are sensitive to this definition of missense mutations that match the mutation spectrum in a cancer discriminates between known drivers and a set of random passengers.

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 11.39 h 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


