CHASMplus Documentation

Release 0.1.0

Collin Tokheim

May 19, 2020

Contents

Author Collin Tokheim

Contact ctokheim # jhu DOT edu

Source code GitHub

Q&A Biostars (tag: CHASMplus)

Large-scale cancer sequencing studies of patient cohorts have statistically implicated many cancer driver genes, with a long-tail of infrequently mutated genes. Here we present CHASMplus, a computational method to predict driver missense mutations, which is uniquely powered to identify rare driver mutations within the long-tail. We show that it substantially outperforms comparable methods across a wide variety of benchmark sets. Applied to 8,657 samples across 32 cancer types, CHASMplus identifies over 4,000 unique driver mutations in 240 genes, further distinguished by their specific cancer types. Our results support a prominent emerging role for rare driver mutations, with substantial variability in the frequency spectrum of drivers across cancer types. The trajectory of driver discovery may already be effectively saturated for certain cancer types, a finding with policy implications for future sequencing. As a resource to handle newly observed driver mutations, we systematically score every possible missense mutation across the genome and provide access to those scores through OpenCRAVAT.

Contents:

CHAPTER 1

Quick start (OpenCRAVAT & CHASMplus)

The easiest way to obtain CHASMplus scores is by using OpenCRAVAT to fetch precomputed scores. You will need python 3.5 or newer to use OpenCRAVAT.

1.1 Install OpenCRAVAT

You will first need to install the OpenCRAVAT python package, please follow the instructions on the OpenCRAVAT wiki:

Installation Instructions

1.2 Install CHASMplus annotators

OpenCRAVAT has a modular architecture to perform genomic variant interpretation including variant impact, annotation, and scoring. CHASMplus is one module available in the CRAVAT store. To install the CHASMplus module within OpenCRAVAT, please execute the following command:

```
$ cravat-admin install chasmplus
```

The above command may take a couple minutes and will install the pan-cancer model of CHASMplus scores. To install cancer type specific versions of CHASMplus, follow the following template:

```
$ cravat-admin install chasmplus_LUAD
```

where LUAD, the abbrevitation from the The Cancer Genome Atlas, designates lung adenocarcinoma. To see a full list of available annotators, issue the following commnad:

```
$ cravat-admin ls -a
```

1.3 Running CHASMplus

OpenCRAVAT takes as input either a VCF file or a simple tab-delimited text file. I will describe a simple example that uses the latter. The simple tab-delimited text file should contain a variant ID, chromosome (with "chr"), start position (1-based), strand, reference allele, alternate allele, and optional sample ID.:

```
var1        chr10   122050517       +       C       T
var2        chr11   124619643       +       G       A
var3        chr11   47358961        +       G       T
var4        chr11   90135669        +       C       T
var5        chr12   106978077       +       A       G
```

You can download an example input file here.

Note: By default, OpenCRAVAT processes variants on the hg38 reference genome. If you are using hg19 or hg18, please specify with the "-l" parameter your specific reference genome so that OpenCRAVAT will know to lift over your variants.

You can run CHASMplus by using the *cravat* command. For information about command line options, please see the command line help:

```
$ cravat -h
```

To obtain CHASMplus scores for pan-cancer (annotator "chasmplus") and lung adenocarcinoma (annotator "chasmplus_LUAD"), run the following command:

```
$ cravat -n MYRUN -t excel -a chasmplus chasmplus_LUAD -d output_directory input.txt
```

The above command will run all annotators (specified by the -a flag, multiple separated by a space) and save results to the directory named "output_directory". The "-t" option specifies the output to be saved as an excel file. The -n flag specifies the name of the run. Scores and p-values from CHASMplus are found in the "MYRUN.xlsx" file (or "MYRUN.tsv" if -t text is chosen). You should see the "Variant" excel sheet that contains columns like this:

```
CHASMplus                                   CHASMplus_LUAD
P-value Score   Transcript  All results P-value Score   Transcript  All results
0.399   0.048   ENST00000453444.6   ENST00000334433.7:(0.025:0.59),ENST00000358010.
↪5:(0.049:0.393),*ENST00000453444.6:(0.048:0.399),NM_001291876.1:(0.046:0.412),NM_
↪001291877.1:(0.045:0.418),NM_206861.2:(0.048:0.399),NM_206862.3:(0.025:0.59)    0.
↪644   0.013   ENST00000334433.7   *ENST00000334433.7:(0.013:0.644),ENST00000358010.
↪5:(0.023:0.478),ENST00000453444.6:(0.022:0.492),NM_001291876.1:(0.022:0.492),NM_
↪001291877.1:(0.022:0.492),NM_206861.2:(0.023:0.478),NM_206862.3:(0.013:0.644)
0.99    0.001   NM_052959.2 *NM_052959.2:(0.001:0.99)   0.945   0.002   NM_052959.2␣
↪*NM_052959.2:(0.002:0.945)
0.446   0.041   NM_001080547.1  ENST00000533968.1:(0.053:0.369),*NM_001080547.1:(0.
↪041:0.446),NM_003120.2:(0.049:0.393) 0.278   0.044   NM_001080547.1 ␣
↪ENST00000533968.1:(0.043:0.284),*NM_001080547.1:(0.044:0.278),NM_003120.2:(0.053:0.
↪224)
```

CHASMplus scores are provided in a transcript specific manner, with the score for the default selected transcript shown in the "Score", "P-value", and "Transcript" columns. Scores for other transcripts are listed in the "All results" column.

1.4 Interpretation

CHASMplus scores range from 0 to 1, with higher scores meaning more likely to be a cancer driver mutation. If you are looking to identify a discrete set of putative driver mutations, then we suggest that you correct for multiple hypothesis testing. We recommend using the Benjamini-Hochberg (BH) procedure for controling the false discovery rate. You will need to use an external package to do this, e.g., the *p.adjust* function in R. False discovery rate adjustments will likely be added in the future.

1.5 Further documentation

For further advanced features of OpenCRAVAT, please see the OpenCRAVAT wiki.

CHAPTER 2

Available CHASMplus models

CHASMplus can perform predictions either using a cancer type-specific model or in a "pan-cancer" manner by consider multiple cancer types together. Pan-cancer is a useful default if a matching cancer type is not available from The Cancer Genome Atlas (TCGA). We have made the following results available through OpenCRAVAT:

Annotator name	Data source	Cancer type
chasmplus	TCGA	Pan-cancer (multiple cancer types)
chasmplus_LAML	TCGA	Acute Myeloid Leukemia
chasmplus_ACC	TCGA	Adrenocortical carcinoma
chasmplus_BLCA	TCGA	Bladder Urothelial Carcinoma
chasmplus_LGG	TCGA	Brain Lower Grade Glioma
chasmplus_BRCA	TCGA	Breast invasive carcinoma
chasmplus_CESC	TCGA	Cervical squamous cell carcinoma and endocervical adenocarcinoma
chasmplus_CHOL	TCGA	Cholangiocarcinoma
chasmplus_COAD	TCGA	Colon adenocarcinoma
chasmplus_ESCA	TCGA	Esophageal carcinoma
chasmplus_GBM	TCGA	Glioblastoma multiforme
chasmplus_HNSC	TCGA	Head and Neck squamous cell carcinoma
chasmplus_KICH	TCGA	Kidney Chromophobe
chasmplus_KIRC	TCGA	Kidney renal clear cell carcinoma
chasmplus_KIRP	TCGA	Kidney renal papillary cell carcinoma
chasmplus_LIHC	TCGA	Liver hepatocellular carcinoma
chasmplus_LUAD	TCGA	Lung adenocarcinoma
chasmplus_LUSC	TCGA	Lung squamous cell carcinoma
chasmplus_DLBC	TCGA	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
chasmplus_MESO	TCGA	Mesothelioma
chasmplus_OV	TCGA	Ovarian serous cystadenocarcinoma
chasmplus_PAAD	TCGA	Pancreatic adenocarcinoma
chasmplus_PCPG	TCGA	Pheochromocytoma and Paraganglioma
chasmplus_PRAD	TCGA	Prostate adenocarcinoma

Continued on next page

Table 1 – continued from previous page

Annotator name	Data source	Cancer type
chasmplus_READ	TCGA	Rectum adenocarcinoma
chasmplus_SARC	TCGA	Sarcoma
chasmplus_SKCM	TCGA	Skin Cutaneous Melanoma
chasmplus_STAD	TCGA	Stomach adenocarcinoma
chasmplus_TGCT	TCGA	Testicular Germ Cell Tumors
chasmplus_THYM	TCGA	Thymoma
chasmplus_THCA	TCGA	Thyroid carcinoma
chasmplus_UCS	TCGA	Uterine Carcinosarcoma
chasmplus_UCEC	TCGA	Uterine Corpus Endometrial Carcinoma
chasmplus_UVM	TCGA	Uveal Melanoma

CHAPTER 3

Advanced: download (source)

3.1 CHASMplus releases

- CHASMplus v1.0.0 - 8/17/2018 - Initial release

3.2 Necessary additional code

- 20/20+ code produces driver gene scores. Please follow installation instructions from the 20/20+ website.
- SNVBox code fetches the features used by CHASMplus from a MySQL database

3.3 Necessary data files

- SNVBox MySQL database
- Pre-computed scores data set
- Reference SNVBox transcripts in BED format

CHAPTER 4

Advanced: installation (source)

CHASMplus is only intended to be ran on linux operating systems and on a compute server.

4.1 Releases

CHASMplus can be downloaded on github.

4.2 Package requirements

4.2.1 CHASMplus Environment

We recommend using conda to install the CHASMplus dependencies.

```
$ conda env create -f environment.yml  # create environment for CHASMplus
$ source activate CHASMplus  # activate environment for CHASMplus
```

Make sure the CHASMplus environment is activated when you want to run CHASMplus.

4.2.2 20/20+

You will need to download the 2020plus github repository. Please follow the installation instructions from the 20/20+ website.

Set the directory of 20/20+ in the configuration file for CHASMplus. You can find this configuration file within the CHASMplus directory at chasm2/data/config.yaml.

```
twentyTwentyPlus: /path/to/2020plus  # set this directory
```

Check your PATH variable

Make sure that you have add the 20/20+ directory to your *PATH* variable. If you have done this correctly, the following command should print the location of the 2020plus.py script.

```
$ which 2020plus.py
```

4.2.3 SNVBox database (MySQL)

Features for mutations CHASMplus are obtained can also be prepared by directly using a MySQL database. A MySQL dump of the SNVBox database contains features used for our study. The SNVBox database has a fairly large file size, you may want to directly download and upload to MySQL.

```
$ wget http://karchinlab.org/data/CHASMplus/SNVBox_chasmplus.sql.gz
$ gunzip SNVBox_chasmplus.sql.gz
$ mysql [options] < SNVBox_chasm2.sql
```

This will create a database named mupit_modbase, where [options] is the necessary MySQL parameters to login. You will need sufficient privileges on your MySQL database to CREATE a new database. If everything worked properly, you should see a database named "SNVBox_20161028_sandbox".

4.2.4 SNVBox code

The next step is to download the code that fetches features from the SNVBox database. Please download the code from here, or use wget:

```
$ wget http://karchinlab.org/data/CHASMplus/SNVBox.tar.gz
```

The next step is to set the configuration file (snv_box.conf) to point towards the established database in the previous section. Specifically, change the db.user, db.password, and db.host to point towards your own mysql user name, mysql password, and mysql host.

The last step is to set the CHASMplus configuration file to point towards the path of the snvGetGenomic command within the SNVBox code. The yaml configuration file is found within the CHASMplus directory at chasm2/data/config.yaml.

```
snvGetGenomic: /path/to/SNVBox/snvGetGenomic  # set this path
```

CHAPTER 5

Advanced: Tutorial (source)

To come

CHAPTER 6

FAQ

Who should I contact if I encounter a problem?

If you believe your problem may be encountered by other users, please post the question on biostars. Check to make sure your question has not been already answered by looking at posts with the tag CHASMplus. Otherwise, create a new post with the CHASMplus tag. We will be checking biostars for questions. You may also contact me directly at ctokheim AT jhu dot edu.

Are the p-values by CHASMplus valid for targeted gene panels?

The p-values reported from CHASMplus are based on whole-exome sequencing studies. If your mutations comes from a targeted gene panel, CHASMplus cannot capture ahead of time what are the specific genes being assessed. To get an accurate estimate of statistical significance, you will need to use the source code version of CHASMplus to perform a customized analysis. Documentation on how to do this will be added in the future.

Where can I obtain the training data for CHASMplus?

You can obtain the set of mutations used for training from here.

I want to compare my method to CHASMplus. How should I do it?

I recommend using the precomputed scores available through OpenCRAVAT [see *Quick start (OpenCRAVAT & CHASMplus)*]. Scores in the precompute were generatured using gene-hold out cross-validation, so there is no issue when evaluating performance about training set overlap leading to overfitting. However, the scores do reflect training based on data from The Cancer Genome Atlas (TCGA). If a new method is trained using more data than is available from the TCGA, then it is recommended to create a new CHASMplus model based on the larger data set by using the CHASMplus source code.

CHAPTER 7

Releases

- CHASMplus v1.0.0 - 8/17/2018 - Initial release

CHAPTER 8

Citation

The manuscript is currently submitted. Please cite the biorXiv paper for now:

Tokheim, C., & Karchin, R. (2018). Enhanced context reveals the scope of somatic missense mutations driving human cancers. bioRxiv, 313296.