genotype$_{v}ariants$Documentation

Release 0.3.5

Ronak Shah

Sep 13, 2021

Contents:

CHAPTER 1

genotype_variants

Project to genotype SNV, INDELS and SV.

- Free software: Apache Software License 2.0
- Documentation: https://genotype-variants.readthedocs.io.

1.1 Features

Currently this module only supports genotyping and merging small variants (SNV and INDELS).

For this we have the following command line submodule called **small_variants**.

Which have the following sub-commands:

- **generate**: To run GetBaseCountMultiSample version 1.2.5 on given BAM files
- **merge**: To merge MAF format files w.r.t counts generated from the *generate* command.
- **all**: This will run both of the sub-commands above *generate* and *merge* togather.
- **multiple-samples**: This will run sub-commands *all* for multiple samples in the provided metadata file

Please read the USAGE (https://genotype-variants.readthedocs.io/en/latest/usage.html) section of the documentation for more information

Requires GetBaseCountMultiSample v1.2.4 and above

1.2 To Do

- Tagging genotyped files for thresholds
- Genotyping normal buffy coats
- Genotype structural variants calls

1.3 Credits

This package was created with Cookiecutter and the audreyr/cookiecutter-pypackage project template.

CHAPTER 2

Installation

2.1 Stable release

2.1.1 Requirements

- **Python 3**
- **click** (https://palletsprojects.com/p/click/)
- **click-log** (https://github.com/click-contrib/click-log)
- **pandas** (https://pandas.pydata.org/)

To install genotype_variants, run this command in your terminal:

```
$ pip install genotype_variants
```

This is the preferred method to install genotype_variants, as it will always install the most recent stable release.

If you don't have pip installed, this Python installation guide can guide you through the process.

2.2 From sources

The sources for genotype_variants can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/rhshah/genotype_variants
```

Or download the tarball:

```
$ curl -OJL https://github.com/rhshah/genotype_variants/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```

CHAPTER 3

Usage

Currently this module only supports genotyping and merging small variants (SNV and INDELS).

For this we have the following command line submodule called **small_variants**.

Which have the following sub-commands:

- *generate*: To run GetBaseCountMultiSample on given BAM files
- *merge*: To merge MAF format files w.r.t counts generated from the *generate* command.
- *all*: This will run both of the sub-commands above *generate* and *merge* togather.
- *multiple-samples*: This will run sub-commands *all* for multiple patients in the provided metadata file

3.1 generate

To use *small_variants generate* via command line here are the options:

```
> genotype_variants small_variants generate --help
Usage: genotype_variants small_variants generate [OPTIONS]

Command that helps to generate genotyped MAF, the output file will be
labelled with  patient identifier as prefix

Options:
-i, --input-maf PATH            Full path to small variants input file in
                                MAF format  [required]
-r, --reference-fasta PATH      Full path to reference file in FASTA format
                                [required]
-p, --patient-id TEXT           Alphanumeric string indicating patient
                                identifier  [required]
-b, --standard-bam PATH         Full path to standard bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
```

(continues on next page)

(continued from previous page)

```
-d, --duplex-bam PATH           Full path to duplex bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
-s, --simplex-bam PATH          Full path to simplex bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
-g, --gbcms-path PATH           Full path to GetBaseCountMultiSample
                                executable with fragment support  [required]
-fd, --filter-duplicate INTEGER
                                Filter duplicate parameter for
                                GetBaseCountMultiSample
-fc, --fragment-count INTEGER   Fragment Count parameter for
                                GetBaseCountMultiSample
-mapq, --mapping-quality INTEGER
                                Mapping quality for GetBaseCountMultiSample
-t, --threads INTEGER           Number of threads to use for
                                GetBaseCountMultiSample
-v, --verbosity LVL             Either CRITICAL, ERROR, WARNING, INFO or
                                DEBUG
--help                          Show this message and exit.
```

```
genotype_variants small_variants generate \
-i /path/to/input_maf \
-r /path/to/reference_fasta \
-g /path/to/GetBaseCountsMultiSample \
-p patient_id \
-b standard_bam \
-d duplex_bam \
-s simplex_bam
```

3.1.1 Expected Output

In the current worrking directory if the above command is executed you will find the following files:

- patient_id-STANDARD_genotyped.maf
- patient_id-DUPLEX_genotyped.maf
- patient_id-SIMPLEX_genotyped.maf

3.2 merge

To use *small_variants merge* via command line here are the options:

```
> genotype_variants small_variants merge --help
Usage: genotype_variants small_variants merge [OPTIONS]

Given original input MAF used as an input for GBCMS along with  GBCMS
generated output MAF for standard_bam, duplex_bam or simplex bam,  Merge
them into a single output MAF format.  If both duplex_bam and simplex_bam
based MAF are provided the program will generate merged genotypes as well.
The output file will be based on the give alphanumeric patient identifier
as suffix.
```

(continues on next page)

(continued from previous page)

```
Options:
-i, --input-maf PATH            Full path to small variants input file in
                                MAF format used for input to GBCMS for
                                generating genotypes
-std, --input-standard-maf PATH
                                Full path to small variants input file in
                                MAF format generated by GBCMS for
                                standard_bam
-d, --input-duplex-maf PATH     Full path to small variants input file in
                                MAF format generated by GBCMS for duplex_bam
-s, --input-simplex-maf PATH    Full path to small variants input file in
                                MAF format generated by GBCMS for
                                simplex_bam
-p, --patient-id TEXT           Alphanumeric string indicating patient
                                identifier  [required]
-v, --verbosity LVL             Either CRITICAL, ERROR, WARNING, INFO or
                                DEBUG
--help                          Show this message and exit.
```

```
genotype_variants small_variants merge \
-i /path/to/input_maf \
-std /path/to/standard_bam_genotyped_maf \
-d /path/to/duplex_bam_genotyped_maf \
-s /path/to/simplex_bam_genotyped_maf \
-p patient_id \
```

3.2.1 Expected Output

In the current worrking directory if the above command is executed you will find the following files:

- patient_id-ORG-STD-SIMPLEX-DUPLEX_genotyped.maf

If only input_maf with duplex_bam_genotyped_maf and simplex_bam_genotyped_maf is given then the output file will be:

- patient_id-ORG-SIMPLEX-DUPLEX_genotyped.maf

If only standard_bam_genotyped_maf with duplex_bam_genotyped_maf and simplex_bam_genotyped_maf is given then the output file will be:

- patient_id-STD-SIMPLEX-DUPLEX_genotyped.maf

If only duplex_bam_genotyped_maf and simplex_bam_genotyped_maf is given then the output file will be:

- patient_id-SIMPLEX-DUPLEX_genotyped.maf

3.3 all

To use *small_variants all* via command line here are the options:

```
> genotype_variants small_variants all --help
Usage: genotype_variants small_variants all [OPTIONS]

Command that helps to generate genotyped MAF and merge the genotyped MAF.
```

(continues on next page)

(continued from previous page)

```
the output file will be labelled with patient identifier as prefix

Options:
-i, --input-maf PATH            Full path to small variants input file in
                                MAF format  [required]
-r, --reference-fasta PATH      Full path to reference file in FASTA format
                                [required]
-p, --patient-id TEXT           Alphanumeric string indicating patient
                                identifier  [required]
-b, --standard-bam PATH         Full path to standard bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
-d, --duplex-bam PATH           Full path to duplex bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
-s, --simplex-bam PATH          Full path to simplex bam file, Note: This
                                option assumes that the .bai file is present
                                at same location as the bam file
-g, --gbcms-path PATH           Full path to GetBaseCountMultiSample
                                executable with fragment support  [required]
-fd, --filter-duplicate INTEGER
                                Filter duplicate parameter for
                                GetBaseCountMultiSample
-fc, --fragment-count INTEGER   Fragment Count parameter for
                                GetBaseCountMultiSample
-mapq, --mapping-quality INTEGER
                                Mapping quality for GetBaseCountMultiSample
-t, --threads INTEGER           Number of threads to use for
                                GetBaseCountMultiSample
-v, --verbosity LVL             Either CRITICAL, ERROR, WARNING, INFO or
                                DEBUG
--help                          Show this message and exit.
```

```
genotype_variants small_variants all \
-i /path/to/input_maf \
-r /path/to/reference_fasta \
-g /path/to/GetBaseCountsMultiSample \
-p patient_id \
-b standard_bam \
-d duplex_bam \
-s simplex_bam
```

3.3.1 Expected Output

Please refer to the *generate* and *merge* usage for the expected output.

3.4 multiple-samples

To use *small_variants multiple-samples* via command line here are the options:

```
genotype_variants small_variants multiple-samples --help
Usage: genotype_variants small_variants multiple-samples [OPTIONS]
```

(continues on next page)

(continued from previous page)

```
Command that helps to generate genotyped MAF and  merge the genotyped MAF
for multiple patients. the output file will be labelled with sample
identifier as prefix

Expected header of metadata_file in any order: sample_id maf standard_bam
duplex_bam simplex_bam

For maf, standard_bam, duplex_bam and simplex_bam please include full path
to the file.

Options:
-i, --input-metadata PATH       Full path to metadata file in TSV/EXCEL
                                format, with following headers: sample_id,
                                maf, standard_bam, duplex_bam, simplex_bam.
                                Make sure to use full paths inside the
                                metadata file  [required]
-r, --reference-fasta PATH      Full path to reference file in FASTA format
                                [required]
-g, --gbcms-path PATH           Full path to GetBaseCountMultiSample
                                executable with fragment support  [required]
-fd, --filter-duplicate INTEGER
                                Filter duplicate parameter for
                                GetBaseCountMultiSample
-fc, --fragment-count INTEGER   Fragment Count parameter for
                                GetBaseCountMultiSample
-mapq, --mapping-quality INTEGER
                                Mapping quality for GetBaseCountMultiSample
-t, --threads INTEGER           Number of threads to use for
                                GetBaseCountMultiSample
-v, --verbosity LVL             Either CRITICAL, ERROR, WARNING, INFO or
                                DEBUG
--help                          Show this message and exit.
```

```
genotype_variants small_variants multiple-samples \
-i /path/to/input_metadata \
-r /path/to/reference_fasta \
-g /path/to/GetBaseCountsMultiSample
```

3.4.1 Expected Output

Please refer to the *generate* and *merge* usage for the expected output.

To use genotype_variants in a project:

```
import genotype_variants
```

CHAPTER 4

Contributing

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given.

You can contribute in many ways:

4.1 Types of Contributions

4.1.1 Report Bugs

Report bugs at https://github.com/rhshah/genotype_variants/issues.

If you are reporting a bug, please include:

- Your operating system name and version.
- Any details about your local setup that might be helpful in troubleshooting.
- Detailed steps to reproduce the bug.

4.1.2 Fix Bugs

Look through the GitHub issues for bugs. Anything tagged with "bug" and "help wanted" is open to whoever wants to implement it.

4.1.3 Implement Features

Look through the GitHub issues for features. Anything tagged with "enhancement" and "help wanted" is open to whoever wants to implement it.

4.1.4 Write Documentation

genotype_variants could always use more documentation, whether as part of the official genotype_variants docs, in docstrings, or even on the web in blog posts, articles, and such.

4.1.5 Submit Feedback

The best way to send feedback is to file an issue at https://github.com/rhshah/genotype_variants/issues.

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.
- Remember that this is a volunteer-driven project, and that contributions are welcome :)

4.2 Get Started!

Ready to contribute? Here's how to set up *genotype_variants* for local development.

1. Fork the *genotype_variants* repo on GitHub.
2. Clone your fork locally:

```
$ git clone git@github.com:your_name_here/genotype_variants.git
```

3. Install your local copy into a virtualenv. Assuming you have virtualenvwrapper installed, this is how you set up your fork for local development:

```
$ mkvirtualenv genotype_variants
$ cd genotype_variants/
$ python setup.py develop
```

4. Create a branch for local development:

```
$ git checkout -b name-of-your-bugfix-or-feature
```

   Now you can make your changes locally.

5. When you're done making changes, check that your changes pass flake8 and the tests, including testing other Python versions with tox:

```
$ flake8 genotype_variants tests
$ python setup.py test or pytest
$ tox
```

   To get flake8 and tox, just pip install them into your virtualenv.

6. Commit your changes and push your branch to GitHub:

```
$ git add .
$ git commit -m "Your detailed description of your changes."
$ git push origin name-of-your-bugfix-or-feature
```

7. Submit a pull request through the GitHub website.

4.3 Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. The pull request should include tests.
2. If the pull request adds functionality, the docs should be updated. Put your new functionality into a function with a docstring, and add the feature to the list in README.rst.
3. The pull request should work for Python 3.5, 3.6, 3.7 and 3.8, and for PyPy. Check https://travis-ci.org/rhshah/genotype_variants/pull_requests and make sure that the tests pass for all supported Python versions.

4.4 Tips

To run a subset of tests:

```
$ python -m unittest tests.test_genotype_variants
```

4.5 Deploying

A reminder for the maintainers on how to deploy. Make sure all your changes are committed (including an entry in HISTORY.rst). Then run:

```
$ bump2version patch # possible: major / minor / patch
$ git push
$ git push --tags
```

Travis will then deploy to PyPI if tests pass.

CHAPTER 5

Credits

5.1 Development Lead

- Ronak Shah <rons.shah@gmail.com>

5.2 Contributors

None yet. Why not be the first?

CHAPTER 6

History

6.1 0.3.0 (2020-04-10)

- Release with merge for standard BAM maf and Input MAF. Converted multiple-patient to multiple-sample

6.2 0.2.1 (2020-04-09)

- Release bug fixes, where simplex numbers are listed as duplex and vice versa, during running *all* command.

6.3 0.2.0 (2020-04-08)

- Release with multiple-patient command.

6.4 0.1.0 (2020-01-30)

- First release on PyPI.

CHAPTER 7

Indices and tables

- genindex
- modindex
- search