gsalib Documentation

Release 1.0.0

Michael Yourshaw

Mar 21, 2018

Contents:

CHAPTER 1

gsalib: Python/pandas package of utility functions for GATK

`gsalib` makes it easy for Python users to analyze metrics reports created by the Broad Institute's Genome Analysis Toolkit (GATK). The Broad provides an R library called gsalib that allows you to load GATKReport files into R for further analysis (https://gatkforums.broadinstitute.org/gatk/discussion/1244/what-is-the-gatkreport-file-format). Python `gsalib` is an adaptation of the R libray that allows you to load GATKReport files into Python/pandas DataFrames.

Neither the R nor Python versions of `gsalib` support the samtools.metrics reports created by Picard Tools. To analyze Picard reports with Python, consider using the `picard.parse` function in the Crimson module.

1.1 Features

- Enables analysis of GATK reports with powerful pandas DataFrames and plotting
- Reads GATKReport versions 0.x and 1.x
- Compatible with Python >=2.7 and >=3.4

1.2 Installation

Install `gsalib` by running

```
pip install gsalib
```

1.3 Example

Read a report and get a table's DataFrame:

```
from gsalib import GatkReport

report = GatkReport('/path/to/gsalib/test/test_v1.0_gatkreport.table')
table = report.tables['ExampleTable']
```

1.4 Documentation

https://gsalib.readthedocs.io/en/latest/

1.5 Contribute

- Issue Tracker: https://github.com/myourshaw/gsalib/issues
- Source Code: https://github.com/myourshaw/gsalib

1.6 License

The project is licensed under the MIT license.

CHAPTER 2

gsalib tutorial

2.1 Introduction

This tutorial is adapded from What is the GATKReport file format? by Geraldine Van der Auwera.

A GATK Report is simply a text document that contains a well-formatted, easy to read representation of some tabular data. Many GATK tools output their results as GATK Reports. A report contains one or more individual GATK report tables.

Here's a simple example (note that the format varies depending on the report version):

```
#:GATKReport.v1.0:2
#:GATKTable:true:2:9:%.18E:%.15f:;
#:GATKTable:ErrorRatePerCycle:The error rate per sequenced position in the reads
cycle  errorrate.61PA8.7         qualavg.61PA8.7
0      7.451835696110506E-3      25.474613284804366
1      2.362777171937477E-3      29.844949954504095
2      9.087604507451836E-4      32.875909752547310
3      5.452562704471102E-4      34.498999090081895
4      9.087604507451836E-4      35.148316651501370
5      5.452562704471102E-4      36.072234352256190
6      5.452562704471102E-4      36.121724890829700
7      5.452562704471102E-4      36.191048034934500
8      5.452562704471102E-4      36.003457059679770

#:GATKTable:false:2:3:%s:%c:;
#:GATKTable:TableName:Description
key    column
1:1000  T
1:1001  A
1:1002  C
```

This report contains two individual GATK report tables. A report file begins with a report header that contains the report version and, in later versions, the number of tables. Every table begins with a header for its metadata and then a header for its name and description. The next row contains the column names followed by the data.

The Python module `gsalib` allows you to load GATK Report files into Python/pandas DataFrames for further analysis. Here are the simple steps to get `gsalib`, install it, and load a report.

2.2 1. Get the gsalib module from PyPI

Install `gsalib` by running on the command line:

```
pip install gsalib
```

2.3 2. Start Python (or open a Python notebook in Jupyter)

```
$ python3
Python 3.6.4 (default, Dec 19 2017, 11:33:49)
[GCC 6.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

2.4 3. Load the GatkReport object from gsalib

```
>>> from gsalib import GatkReport
```

2.5 4. Finally, load the GATKReport file and have fun

`gsalib` has one class, `GatkReport`, that is a dict-like container for all of the tables in a GATK Report file. The `GatkReport.tables` attribute is a key-value object where the key is the table name and the value is a pandas DataFrame that contains the table's data. Note that if a report contains more than one table with the same name the keys will be uniquified as `table`, `table.1`, etc.:

```
>>> d = GatkReport('/path/to/gsalib/test/test_v1.0_gatkreport.table')
>>> for table in d.tables:
...     d[table].describe()


          cycle  errorrate.61PA8.7  qualavg.61PA8.7
count  9.000000           9.000000         9.000000
mean   4.000000           0.001595        33.581250
std    2.738613           0.002273         3.687989
min    0.000000           0.000545        25.474613
25%    2.000000           0.000545        32.875910
50%    4.000000           0.000545        35.148317
75%    6.000000           0.000909        36.072234
max    8.000000           0.007452        36.191048
           key column
count        3      3
unique       3      3
top     1:1000      A
freq         1      1
```

For more examples, *see* gsalib/examples, which contains:

reshape_concordance_table Given a GATK Report generated by GATK GenotypeConcordance this function reshapes the concordance for a specified sample into a matrix with the EvalGenotypes in rows and the CompGenotypes in columns.

summarize_varianteval (Python3 only) Summarize several tables produced by GATK VariantEval into a VariantEvalMetricsSummary table as described in (howto) Evaluate a callset with VariantEval.

CHAPTER 3

GatkReport class