taboo Documentation

Release 0.0.1

Robin Andeer

November 30, 2015

Contents

Taboo is a simple genotype comparison tool. It can handle multiple VCF files with multiple samples. Taboo is extendible to allow for plugins that customize the output of the comparison.

Taboo is exclusively a command line utility.

CHAPTER 1

Motivation

Comparing variants between samples and VCF files is a common task. However, I haven't found *the* VCF comparison tool yet.

Programs that are often recommended include *vcf-compare*, *vcfgtcompare*, *BEDTools*, and *GATK*. VCFTools had wierd output and wrote most useful data to a log file and therefore data couldn't be piped to a subsequent filter process. Vcflib was difficult to set up and required VCF files to be gzipped and indexed. GATK seemed too heavy-handed for such a simple task.

Therefore I decided to develop my own **simple** genotype comparison tools in Python. It will focus on transparency and easy of use.

CHAPTER 2

Installation

Taboo is not distributed on *pip*, so to install it run:

```
$ pip install https://github.com/Clinical-Genomics/taboo/zipball/master
```

2.1 Dependencies

- VCFTools. I know, hypocrisy right? In my defence, I only use it to easily sort VCF files.
- PyVCF. I generally think the module is a little over-designed with custom classes galore. However, there are enough benefits and conveniences like "walk_together" included to not use it.

CHAPTER 3

Usage

The main objective of the package is comparison of genotypes between samples. The package handles multi-sample VCFs as well as multiple single-sample VCFs. The important thing is that they are sorted using the same key. The simplest way to do so is to use "vcf-sort" from the VCFTools library:

```
$ vcf-sort /path/to/sample.vcf > /path/to/sample.sorted.vcf
```

To compare each genotype across all samples in all files, issue the command:

```
$ ls
sample1.sorted.vcf sample2.sorted.vcf
$ taboo compare sample*.sorted.vcf > results.txt
```

You can then continue filtering the output as you wish. It might be interesting to:

```
$ grep discordant results.txt
```

3.1 What does it compare?

Each comparison module is built as a plugin that can be turned on/off and additional plugins can be installed using *pip*. The builtin comparators include:

- quality: the quality of the genotype call (GQ)

3.2 Clinical Genomics

Initially, some parts of the package will deal with tasks more or less specific to Clinical Genomics.

1. A MAF Excel report can be converted to a VCF file. This enables standardized comparison of 2 VCFs.
2. Trimming of large VCF files down to the variants of interest. This will use RS numbers as identifiers but could be expanded to chromosome, start, ref(, and alt).
3. Splitting of multi-sample VCFs into multiple single-sample VCFs. This feature might not be needed in the future.

CHAPTER 4

Contributing

There's no point in contributing at the moment. I need to first make sure I have a grasp on the scope of the project.