# Vcf2TrioGenotype - Utility to prepare the pedigree file suitable for FB-SCAT and scan\_trios programs

## Source code

Vcf2TrioGenotype.py

## Sample Vcf and PED files

Two 1000 Genomes Trios (CEU and YRI) were downloaded from <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/trio/>, merged and annotated with Anntools. Ten thousand lines are given as a sample VCF for you to practice. PED file for 2 trios is provided

## Purpose

1. Creates GT (genotype) file based on the VCF and PED files provided for each trio listed in PED file
2. Generates VARIANT file based on annotated VCF provided. By default, the snpEff/GATK annotation is expected, but can be configured to use other annotator.
3. See CONFIGURATION SECTION for details
4. Generated gene index file, i.e. start and end position of gene in the VCF file. Both positions are inclusive and start with one.

## Input

1. Formatted and sorted variant call format (VCF) file. (<http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41>). Use VCFTools (http://vcftools.sourceforge.net/) for format verification if in doubt. VCF file may be compressed. PED file. Must have the following columns IN THAT PARTICULAR ORDER:

* FamilyID (0=unknown)
* IndividualID (Must include all in the VCF file)
* PaternalID (0=unknown)
* MaternalID (0=unknown)
* Sex (1=male; 2=female; 0=unknown)
* Phenotype (1=control/unaffected, 2=case/affected, 0=unknown)

PED format is described at <http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped>

1. Base (no extension) of the output file
2. True for PASS only, False for All variants
3. True for minus 1 from Phenotype, as required by FB-SKAT. False will leave the Phenotype unchanged

## Output

1. For each complete trio (child, father and mother ID present in both PED and VCF files), program outputs one row, where columns are:
2. FamilyID, IndividualID, PaternalID, MaternalID, Sex, Phenotype, followed by Genotype1....GenotupeN
3. First six columns are drawn from the PED file, the genotypes are drawn from matching samples at VCF file.
4. Genotypes are coded as 0, 1 or 2 (no mutation, heterozygous, homozygous respectively)
5. Note! Families with more than one child are treated as 2 or more trios.

## Dependencies

Python 2.6 or higher, no special libraries are needed.

## Installation

No installation is needed. Copy the script to any directory at your computer and give it 755 permission.

To run

python Vcf2TrioGenotype.py vcf ped output True/False True/False