VCF2Linkdatagen

Catherine Bromhead, Melanie Bahlo, 8/7/11. Updated by Katherine Smith 15 May 2012. Updated by Tom Scerri/Rick Tankard. 3rd February 2016.

Email bug reports to Melanie Bahlo bahlo@wehi.edu.au.

Vcf2linkdatagen.pl is a Perl script to create a BRLMM genotype file from a VCF file. This script is tailored towards VCF files created by samtools mpileup (http://samtools.sourceforge.net/mpileup.shtml) or GATK's unifiedGenotyper. Detailed instructions for running the correct samtools pipeline, vcf2linkdatagen.pl and linkdatagen.pl can be found in the file "MPS linkage quick start.pdf" available on the linkdatagen web site. This document includes example commands.

Vcf2linkdatagen.pl discards SNPs where one or more detected alleles do not match the alleles specified in the annotation file. It also discards SNPs that do not meet optional depth or quality thresholds, SNPs not present in the annotation file of interest or SNPs whose population frequency is missing for the HapMap population of interest.

Vcf2linkdatagen.pl takes user-defined quality thresholds or reverts to defaults (see Optional Parameters table below).

```
Usage for single VCF file (command written on a single line):

vcf2linkdatagen.pl -annotfile <filename> -missingness <float> \
-variantCaller <string> -pop <string> -mindepth 10 -min_MQ 10 \
-min_FQ 10 -minP_strandbias <float> -minP_baseQbias <float> \
-minP_mapQbias <float> -minP_enddistbias <float> file_in.vcf \
> out.brlmm


Usage for multiple VCF files:

vcf2linkdatagen.pl -annotfile <filename> -missingness <float> \
-variantCaller <string> -pop <string> -idlist <filename> \
-mindepth 10 -min_MQ 10 -min_FQ 10 -minP_strandbias <float> \
-minP_baseQbias <float> -minP_mapQbias <float> \
-minP_enddistbias <float> \
> out.brlmm


For single VCF file usage, file_in.vcf is the vcf file you wish to
convert to BRLMM genotype calls.  If you put a - in this space instead
of a filename the program will take STDIN as input.

-annotfile: annotation file listing SNP details. This must be either
annotHapMap2U.txt or annotHapMap3U.txt. These files can be downloaded
from the linkdatagen website.
```

OPTIONAL PARAMETERS:

Option	Default	Description
-variantCaller or just -vc	N/A	Program used to create the BRLMM file; currently two options [mpileup | unifiedGenotyper]. Alternative short forms can be used, [mp | ug]
-idlist	N/A	a file containing a list of paths to input VCF files. If there is only one, do not specify -idlist. Instead, provide the name of this VCF file before >
-missingness	1	The maximum proportion of missing genotype calls for a SNP to be output to the brlmm file. This parameter should only be used when reading in multiple VCF files. If missingness is set to 1, all SNPs will be output to the brlmm file.
-pop	CEU	A three-letter code specifying a HapMap population to use as a source of population allele frequencies. Choices are ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI and YRI. Specify the same -pop option that you intend to specify when running linkdatagen.pl.
-min_MQ	10	minimum root mean square mapping quality.
-min_FQ	10	minimum absolute value of consensus quality.
-mindepth	10	minimum read depth. Here read depth is taken as a sum of the DP4 values and not as the DP field, as the DP4 field counts only high quality base calls.
-minP_strandbias	0.0001	minimum p value for strand bias (exact test)
-minP_baseQbias	1e-100	minimum p value for baseQ bias (t-test)
-minP_mapQbias	0	minimum p value for mapQ bias (t-test)
-minP_enddistbias	0.0001	minimum p value for tail distance bias (t-test)

Vcf2linkdatagen.pl input: VCF file produced using samtools.
Example:

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER INFO    FORMAT
chr1    123456 .        C       .       9.02    .       DP=1;AF1=8.363e-
05;CI95=0.5,0.5;DP4=0,1,0,0;MQ=60;FQ=-6.98      PL      0
chr1    234567 .        G       .       9.02    .       DP=1;AF1=8.393e-
05;CI95=0.5,0.5;DP4=0,1,0,0;MQ=30;FQ=-6.99      PL      0
```

See http://samtools.sourceforge.net/mpileup.shtml for descriptions of fields in the VCF file.

Vcf2linkdatagen.pl output: BRLMM genotype file containing one column of SNP identifiers followed by one column of genotype calls for each sample. BRLMM genotype calls are:

0 AA call
1 AB call

2 BB call
-1 No Call.

VCF entries whose quality scores and depth fall below specified thresholds will have genotypes set to "No Call" (-1).

LIMITATIONS

Vcf2linkdatagen.pl cannot process multi-sample VCF files. This limitation may be removed in future.

REFERENCES

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009, 25:2078-9.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D and Daly M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 2011, 43:491-498.