Guide to extracting HapMap SNP genotypes from MPS data sets and preparing them for analysis using vcf2linkdatagen.pl and linkdatagen.pl

Katherine Smith
Updated by Rick Tankard, with help from Thomas Scerri
Email bug reports and questions to Melanie Bahlo bahlo@wehi.edu.au.
Updated 3rd February 2016

Note: the commands listed here are examples. You will need to modify filenames and options according to your circumstances. Please consult the vcf2linkdatagen.pl and linkdatagen.pl documentation for a detailed description of options available.

Download the following files from
http://bioinf.wehi.edu.au/software/linkdatagen/:

- The Perl script vcf2linkdatagen.pl
- The Perl script linkdatagen.pl
- The HapMap Phase II and/or HapMap Phase III annotation files:
    - The gzipped file annotHapMap2U.txt.gz contains annotHapMap2U.txt
    - The gzipped file annotHapMap3U.txt.gz contains annotHapMap3U.txt
      You may also wish to download the documentation files for the two Perl scripts.
    - For GATK genotyping, gzipped VCF files for:
        - HapMap2: dbsnp_137.hg19_HapMap2L.vcf.gz
        - HapMap3: dbsnp_137.hg19_HapMap3L.vcf.gz

Genotyping can be performed with either SAMtools or GATK. GATK is our preferred caller but can be more difficult to set up than SAMtools. If performing genotyping with SAMtools, download and install version 0.1.19 of the SAMtools package from http://sourceforge.net/projects/samtools/files/samtools/0.1.19/. Please note that newer versions of SAMtools will not work with the commands presented in this guide. If performing genotyping with GATK, you will be required to download the VCFs for GATK from the LINKDATAGEN website.

Align your MPS reads to version hg19 of the human genome (hg19.fa) using your favourite aligner. The annotation files will not work correctly with other versions of the genome as the SNPs will have different coordinates and/or the references (chromosomes) will have different names. Vcf2linkdatagen.pl assumes that chromosomes are named chr1, chr2, … chrX rather than 1,2,…,X. Output your alignment in SAM format, convert to BAM and process as you would for variant calling. This should include duplicate removal; Picard's MarkDuplicates can perform this task (http://broadinstitute.github.io/picard/). Local realignment and quality

score recalibration (such as with GATK https://www.broadinstitute.org/gatk/) may give minor improvements to genotyping.

Before you begin, decide which HapMap population you want to use to obtain population allele frequencies for linkage analysis, then choose the appropriate annotation (HapMap II or HapMap III). The eleven populations available have the codes ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI and YRI. Most populations only have frequencies available for the HapMap III annotation. CHB, CEU, JPT and YRI frequencies are available for both annotations, but the HapMap II annotation is recommended for theses populations as this contains more markers than HapMap III. For the rest of this example, we will use the HapMap II annotation and specify CEU population allele frequencies.

For each sample sequenced, obtain genotypes at the location of HapMap SNPs for Phase II or Phase III, as desired. This can be performed with SAMtools (easier) or GATK UnifiedGenotyper (preferred, but can be difficult to set up if not using GATK already):

SAMtools 0.1.19 example command (SAMtools 1.0.0+ will not work):

```
samtools mpileup -d10000 -q13 -Q13 -gf hg19.fa -l
annotHapMap2U.txt sample1.bam | bcftools view -cg -t0.5 - >
sample1.HM.vcf
```

GATK UnifiedGenotyper command:

```
java -Xmx4g -jar ${GATK_jar_file} -R ucsc.hg19.fasta -T
UnifiedGenotyper -I sample1.bam -o sample1.ug.vcf -
stand_call_conf 0 -stand_emit_conf 0 -dcov 500 --dbsnp
dbsnp_137.hg19_HapMap2L.vcf -L dbsnp_137.hg19_HapMap2L.vcf --
genotyping_mode GENOTYPE_GIVEN_ALLELES --alleles
dbsnp_137.hg19_HapMap2L.vcf --output_mode
EMIT_ALL_CONFIDENT_SITES
```

This command genotypes HapMap Phase II SNPs. It assumes that hg19.fa, annotHapMap2U.txt, dbsnp_137.hg19_HapMap2L.vcf and sample1.bam are all stored in the current working directory, but may be set to other locations by their path. The bash variable `${GATK_jar_file}` should be set to the path of the GATK .jar file or replaced with the path directly.

Next, run vcf2linkdatagen.pl to convert variant calls from VCF to BRLMM format. The BRLMM file will contain a column of SNP IDs followed by a column of genotypes for each individual. Genotypes will be coded 0 for AA, 1 for AB, 2 for BB or -1 for missing. A is the alphabetically lowest allele and B the alphabetically highest allele on the plus strand, i.e. for a C/T SNP 0=CC and 2=TT.

For a single VCF file (i.e. only one person was sequenced from the pedigree) with genotypes called by SAMtools mpileup:

```
vcf2linkdatagen.pl -variantCaller mpileup -annotfile
annotHapMap2U.txt -pop CEU -mindepth 10 -missingness 0
MySNPs.raw.vcf > MySNPs.brlmm
```

If GATK UnifiedGenotyper was used then set the -variantCaller option to unifiedGenotyper instead.

If you sequenced more than one individual from the pedigree, you will have multiple VCF files. Create a text file that lists the path to each VCF file on a separate line, and specify the name of this file using the -idlist argument:

```
vcf2linkdatagen.pl -variantCaller mpileup -annotfile
annotHapMap2U.txt -pop CEU -mindepth 10 -missingness 0 -idlist
MyVCFlist.txt > MySNPs.brlmm
```

`-mindepth 10` specifies that a SNP's genotypes should only be printed if the SNP location has a minimum read depth (coverage) of ten high-quality reads. `-pop CEU` specifies that CEU population frequencies should be used. These are the default settings.

The `-missingness` argument specifies the maximum proportion of missing genotype calls for a SNP to be output to the BRLMM file. `-missingness 0` specifies that only SNPs with non-missing genotypes for all samples should be printed to the BRLMM file. The default is `-missingness 1`, meaning that any amount of missingness is allowed.

Type `vcf2linkdatagen.pl -help` to learn more about these and other options.

Note: Other variant callers, such as GATK HaplotyperCaller, also produce VCF files. However, we have not yet extended vcf2linkdatagen.pl to work with these. The script currently requires the presence of the FQ tag in the INFO field created by SAMtools (see http://samtools.sourceforge.net/mpileup.shtml) or GATK specific tags when using the UnifiedGenotyper. Optional filters may also require SAMtools-specific tags, e.g. DP4 for depth filtering.

Create a pedigree file and a whichsamples file for linkdatagen.pl. The whichsamples file consists of a single line of space-separated numbers; the *k*th number indicates the column number of the BRLMM file that contains genotypes for the individual specified in the *k*th row of your pedigree file. The first column of genotypes in the BRLMM file is column 1. Refer to the linkdatagen.pl documentation for more detail.

You are now ready to run linkdatagen.pl to create input files for linkage analysis. Here is an example command that will generate MERLIN (see http://www.sph.umich.edu/csg/abecasis/merlin/download/) input files. It assumes that annotHapMap2U.txt and MySNPs.brlmm are stored in the current working directory, but may be set to other locations by their path.

```
linkdatagen.pl -data m -pedfile MyPed.ped -whichSamplesFile
MyWS.ws -callFile MySNPs.brlmm -annotFile annotHapMap2U.txt -pop
CEU -binsize 0.3 -MendelErrors removeSNPs -prog me -outputDir
MyPed_HapMap2 > MyPed_HapMap2_me.out
```

linkdatagen.pl can create input files for many other programs, e.g. ALLEGRO. Refer to the documentation for further details concerning the `-prog` option and other options available.

You may wish to specify the `-removeWFHBS` option, which requests that uninformative SNPs (those for which all samples within a family have identical homozygous genotypes) are discarded. This should not be done if you only have genotypes from affected samples. Refer to section '4.3. Within/whole-family homozygosity-by-state ("uninformative") markers' in the linkdatagen.pl documentation for further details.

If you are performing parametric linkage analysis, don't forget to create a file specifying the genetic model (see http://www.sph.umich.edu/csg/abecasis/merlin/tour/parametric.html). You are now ready to perform linkage analysis using MERLIN.