Software for GBS-based relationship calculations using the KGD method

Author: Ken Dodds
Address: Invermay Agricultural Centre, Puddle Alley, Private Bag 50034, Mosgiel 9053, New Zealand
Email: ken.dodds at agresearch.co.nz
Date: 26 August 2015

Contents

Background

R code is available for the analysis of genotyping-by-sequencing (GBS) data, primarily to construct a genomic relationship matrix for the genotyped individuals. The code can be used on its own, or incorporated into other R programs. There are QC tools (primarily graphical output), relationship estimation tools, pedigree verification tools and pedigree 'mix and match' tools. The latter two operations require additional input information about the samples genotyped.

In this document, 'Individual' or 'sample' generally refers to the genotyping unit (possibly combined, if the same individual or sample is genotyped multiple times). Familial relationships are given the labels 'Father', 'Mother' and 'Offspring' (as appropriate).

The methods used are as described in Dodds *et al*. (in prep). Unless specified, relatedness estimates in this documentation refer to those using the 'G5' method of that paper, also referred to as the **K**inship using **G**BS with **D**epth adjustment (KGD) method.

The code is still undergoing development.

Program structure

There are two separate analysis program files, the first (GBS-Chip-Gmatrix.R) for genotype QC and relationship matrix construction and the second (GBSPedAssign.R) for pedigree verification and/or assignment, based on the related estimates. These programs can be invoked from another program file (using the *source* command), or users can insert all or parts of these programs into their own code. For the purposes of this documentation, it is assumed the first method is used, with calling program named GBSRun.R.

Calling program (GBSRun.R)

Variable / command	Type[1]	Description
genofile	V	Name (including path) of the genotype file. Default value is "HapMap.hmc.txt".
gform	V	Type of genotype file. Default is "uneak".
sampdepth.thresh	V	Minimum mean sample depth for retaining sample results. Default is 0.01.
source	C	Invoke GBS-Chip-Gmatrix.R code, to run QC procedures and define the genomic relationship matrix function (*calcG*)
calcG	C	Calculate genomic relationship matrices. May be invoked several times with different options.
pedfile	V	Name of file containing pedigree and/or parent group information
groupsfile	V	Name of file containing which individuals are in which parent groups
GCheck	V	The name (as a string) of the G matrix to use for parent verification or assignment This must be set before calling GBSPedAssign.R.
rel.thresh	V	The relatedness threshold to use for parent verification or assignment. This has a default value of 0.4.
mindepth.mm	V	Minimum depth to be used for calculating mismatch proportions in parent matching.
source	C	Invoke GBS-PedAssign.R code to verify parents (if given) or assign parents (if parent groups are given)

[1] Type is V for a variable to be set, or C for a command to be invoked or function to be run.

Relatedness estimation program (GBS-Chip-Gmatrix.R)

This program performs some QC diagnostics, rudimentary data cleaning and defining a function for relatedness estimation and reporting. Routines to check and report on positive and negative controls, based on a specified sample naming system, are yet to be included. Any procedures or output relating to depth are not implemented for chip data.

Samples with very low depth are dropped from the analyses. The threshold is a mean depth of *sampdepth.thresh* (default of 0.01, but can be set in the calling program) or with a maximum depth of one (including those with no genotype calls). Samples that are dropped are reported in the program output, as is the remaining number of samples.

SNPs with no data or with a MAF (minor allele frequency) of zero are dropped. The remaining number of SNPs is reported.

Some basic statistics are reported: Proportion of missing genotypes is the number of SNP x individual combinations with no allele calls; Mean sample depth is the average depth (number of reads of either allele) for a sample.

SampleStats.csv contains the identifier for each sample (as given in the genotype file), call rates for each sample and the mean sample depths (for GBS data).

AlleleFreq.png is a plot of allele frequencies calculated using different methods (and as given, if the uneak format is used).

CallRate.png shows a histogram of sample call rates (proportion of SNPs with a result for a sample).

SampDepth.png plots mean sample depth against median sample depth.

SampDepth-scored.png plots mean sample depth, over SNPs that are scored for the individual, against mean sample depth over all SNPs for the individual.

SampDepthHist.png is a histogram of mean sample depths

SampDepthCR.png plots mean sample depth against call rate.

SNPDepthHist.png is a histogram of SNP depths (number of reads of either allele averaged over samples)

SNPDepth.png plots SNP depth against mean SNP depth over samples that are called for that SNP. This may reveal SNPs that are called infrequently, but when they are called have good depth (these SNPs may be near the boundary of a size selection step in the laboratory).

finplot.png plots Hardy-Weinberg disequilibrium (HWD) against MAF, shaded by the SNP depth. HWD is the proportion of (reference allele) homozygotes minus the expected proportion (under Hardy-Weinberg equilibrium). HWD is the same whichever allele is used in the calculation. The ‘fin plot’ may reveal sets of SNPs that do not follow Mendelian inheritance, for example apparent SNPs in duplicated regions.

HWdisMAFsig.png is similar to the fin pot, but with shading by the likelihood ratio test statistic for HWD.

LRT-QQ.png is a QQ plot for the likelihood ratio test statistic for HWD.

LRT-hist.png is a histogram of the likelihood ratio test statistic for HWD.

MAF.png is a histogram of the MAFs for each SNP (based on observed genotypes).

A function, *calcG* is defined.
Usage: calcG(snpsubset,sfx="",puse,indsubset,depth.min=0,depth.max=Inf,npc=0)
Arguments:

snpsubset a vector of integers (between 1 and *nsnps,* inclusive) of the SNPs to use in the calculation. The default is to use all SNPs.

sfx A suffix to use in output file names to identify which function call has produced that output.

puse a vector of (reference) allele frequencies to use in the calculations. The default is to use allele frequencies calculated on the basis of allele counts.

indsubset a vector of integers (between 1 and *nind,* inclusive) of the individuals for which relatedness matrices will be calculated. The default is to calculate for all individuals.
depth.min The minimum depth for a SNP result for an individual to be used.
depth.max The maximum depth for a SNP result for an individual to be used.
npc The number of principal components of the ‘G5’ relatedness matrix to display. If *npc*=0 (the default) the principal component analysis is omitted.

Value: a list of relatedness structures: G1, G4d (diagonal elements of G4), G5 and PC, the output of the principal components analysis (if *npc*>0). The G*n* relatedness matrices are described in Dodds *et al*. (in prep).
Details: The function also produces a set of output (files), as follows.

Basic information: Number of SNPs, number of individuals, mean depth, mean self-relatedness (using G5).

MAF<sfx>.png is a histogram of the MAFs for each SNP used (based on observed genotypes). Only produced if a subset of SNPs is used.

HighRelatedness.csv contains pairs of samples and their G5 relatedness, where this relatedness is > *hirel.thresh* (default value of *hirel.thresh* is 0.9).

Heatmap-G5<sfx>.png is a heatmap plot using G5 relatedness

G<sfx>-diag.png is a plot of diagonal elements (self-relatedness estimates) of G4 against those of G5 (illustrating the effect of correcting for depth).

G<sfx>diagdepth.png is a plot of diagonal elements of G5 against the logged sample depth. We do not expect there to be a relationship between these variables (unless planned) so this serves as a diagnostic for e.g. non-Mendelian SNPs and/or the assumption of random sampling of alleles during sequencing.

PC1v2G5<sfx>.png (if *npc*>0) is a plot of 2nd versus the 1st principal components.

PCG5<sfx>.pdf (if *npc*>2) is a scatterplot matrix of the first *npc* principal components.

There is a vector *fcolo* (length *nind*) of colours to be used for the individuals in these plots. It defaults to all black, but can be reset after sourcing the program and before calling *calcG*.

Pedigree program (GBSPedAssign.R)

This program uses a relatedness matrix to verify given pedigrees and/or to find the best matching parents from groups of potential parents. Both these tasks require a pedigree file (with name given in *pedfile*). For parent matching a groups file (with name given in *groupsfile*) is also required. See below for the formats for these files. Father (Mother) verification is undertaken if the pedigree file contains a FatherID (MotherID) variable. Father (Mother) matching is undertaken if a groups file is given and the pedigree file contains a FatherGroup (MotherGroup) variable.

Before calling the program the variable *GCheck* must be set to the name (as a string) of the G matrix to use. In addition, *rel.thresh* may be set to override the default relatedness value of 0.4 for declaring a parentage match, and *mindepth.mm* may be set to override the default minimum depth (5) for calculating mismatch rates for parentage matching. For parent matching, no parentage is declared as true or false, the program just reports the closest matching parents and their relatedness statistics.

This program outputs summary statistics and a number of files. The %s of verified fathers and mothers are given, as well as the mean relatedness estimates for matching and non-matching fathers and mothers. There is also the number of offspring and mean relatedness in full-sib families (for families with at least 2 offspring) and the mean relatedness between any of these offspring from families with different parents (i.e. not a full- or half-sib). For parent matching, a mismatch statistic is calculated. This is the proportion of apparent (i.e. using observed genotypes) opposing homozygotes, but using a minimum depth of *mindepth.mm* for both individuals being compared. The files, where relevant, are as follows:

PedVerify.csv returns the pedigree file with additional columns, as shown below:

Variable name	Description
FatherRel	Relatedness estimate between individual and it's specified father
FatherMatch	TRUE if *FatherRel* > *rel.thresh*
MotherRel	Relatedness estimate between individual and it's specified mother
MotherMatch	TRUE if *MotherRel* > *rel.thresh*

FatherVerify.png is a scatterplot matrix showing *FatherRel* (see above), the position of the individual in the pedigree file and the position of the recorded father in the pedigree file. This is useful for seeing the distribution of relatedness values, and possibly for detecting sample tracking issues (if the order in the pedigree file relates to the order samples are processed at a particular stage).

MotherVerify.png is a scatterplot matrix like FatherVerify.png but for mother verification.

FatherMatches.csv shows the results of the father matching. It returns the first two columns of the pedigree file with additional columns, as shown below:

Variable name	Description
BestFatherMatch	IndivID of the father from the *FatherGroup* having the highest estimated relatedness to the individual
FatherMatch2nd	IndivID of the father from the *FatherGroup* having the 2nd highest estimated relatedness to the individual
Fatherrel	The estimated relatedness for
Fatherrel2nd	The estimated relatedness for *FatherMatch2nd*
mmrateFather	The (raw) mismatch rate for *BestFatherMatch*

MotherMatches.csv shows the results of the mother matching (with columns as for FatherMatches.csv but for mothers instead of fathers).

GroupsParentCounts.csv returns the groups file with additional columns, as shown below:

Variable name	Description
FatherFreq	Number of offspring where this father is the *BestFatherMatch* in this group
MotherFreq	Number of offspring where this mother is the *BestMotherMatch* in this group

BestFatherMatches.png is a plot of the mismatch rate for *BestFatherMatch* against the estimated relatedness (*Fatherrel*).

BestMotherMatches.png is a plot of the mismatch rate for *BestMotherMatch* against the estimated relatedness (*Motherrel*).

Input formats

The genotype input format is set with *gform*, one of “uneak” (the default), “Tassel” or “Chip”.

GBS via UNEAK

The default input format (‘uneak’) is a ‘hapmap count’ formatted file as produced by the UNEAK pipeline (Lu *et al*. 2013). This is a tab-separated flat text file with the first column being the SNP identifier, then a column for each genotyped individual (or sample, or other genotyping unit), followed by 5 columns of summary information (HetCount_allele1, HetCount_allele2, Count_allele1, Count_allele2, Frequency). Only the last of these 5 is used. Each row is for a different SNP. The column for each individual contains the genotype information as the allele depth (number of reads of that allele) for the ‘reference’ and ‘alternate’ alleles, respectively. The designation of reference and alternate is arbitrary for this software. The numbers of reads are separated by a pipe symbol (“|”). There is a header line, which, for the genotype columns, is taken as the identifiers of the individuals.

GBS via Tassel

An alternative format(‘Tassel’) is available that may be easier to use for GBS data that has been manipulated in Tassel. It is similar to the uneak format, but allele depths in a genotype are separated by a comma (“,”), has two columns before genotype data, and no columns following the genotype data. The first two columns are the chromosome and position (which together serve as the SNP identifier), respectively. As with the “uneak” format, this is a tab- separated flat text file with a header row.

Chip

Fully recorded genotypes can be entered via the “Chip” format. This comma-separated format has results for each individual in the rows and SNP results in a column. There is a header row (SNP identifiers) and the first column contains individual identifiers. Subsequent columns contain the SNP results. Genotype data is given in 0/1/2 format, representing first homozygote, heterozygote and second homozygote, respectively. Designation of which allele is the ‘first’ is arbitrary.

Pedigree file

An optional pedigree file can be given, and will be used to verify or find parent matches. This is a comma separated file (csv). The columns of this file are specified below. The names must be exactly as specified. Additional columns may be present in the file.

Variable name	Required?	Description
IndivID	Y	identifies individuals in the pedigree and groups files
seqID	Y	matches *IndivID* to the identifier in the genotype file
FatherID	N	Recorded *IndivID* of father
MotherID	N	Recorded *IndivID of mother*
FatherGroup	N	Group label for group of potential fathers
MotherGroup	N	Group label for group of potential mothers

Father and mother group labels should be distinct. FatherID and MotherID should be the same type as IndivID (i.e. all numeric or all character) when read with R’s read.csv command (with default settings).

Groups file

If parent matching is required, then a groups file describing the group labels in the pedigree file is required. This is a comma separated file (csv). The columns (both required) of this file are specified below. The names must be exactly as specified. Additional columns may be present in the file.

Variable name	Description
IndivID	identifier for potential parent, matching *IndivID* in the pedigree file
ParGroup	Group label for the group that *IndivID* belongs to

There should be one row for each group a potential parent belongs to.

Example

Files in directory :
GBSRun.R HapMap.hmc.txt.gz Ped-GBS.csv Ped-Groups.csv

GBSRun.R

```
genofile <- "HapMap.hmc.txt.gz"

source("/Code/GBS-Chip-Gmatrix.R")
Gfull <- calcG()
GHWdgm.05 <- calcG(which(HWdis > -0.05),"HWdgm.05", npc=4)  # recalculate using Hardy-Weinberg disequilibrium cut-off at -0.05

pedfile <- "Ped-GBS.csv"
groupsfile <- "Ped-Groups.csv"

rel.thresh <- 0.2
GCheck <- "GHWdgm.05$G5"
source("/Code/GBSPedAssign.R")
```

linux command:

```
R CMD BATCH --no-save GBSRun.R &
```

Files in directory after running code:

```
AlleleFreq.png           GHWdgm.05diagdepth.png   MotherVerify.png
BestFatherMatches.png    GHWdgm.05-diag.png       PC1v2G5HWdgm.05.png
BestMotherMatches.png    GroupsParentCounts.csv   PCG5HWdgm.05.pdf
CallRate.png             HapMap.hmc.txt.gz        Ped-GBS.csv
FatherMatches.csv        Heatmap-G5HWdgm.05.png   Ped-Groups.csv
FatherVerify.png         Heatmap-G5.png           PedVerify.csv
finplot.png              HighRelatedness.csv      SampDepthCR.png
GBSRun.R                 HWdisMAFsig.png          SampDepthHist.png
GBSRun.Rout              LRT-hist.png             SampDepth.png
GcompareHWdgm.05.png     LRT-QQ.png               SampDepth-scored.png
Gcompare.png             MAFHWdgm.05.png          SampleStats.csv
Gdiagdepth.png           MAF.png                  SNPDepthHist.png
G-diag.png               MotherMatches.csv        SNPDepth.png
```

Acknowledgements

This software development was supported by the Ministry of Business, Innovation and Employment via its funding of the “Genomics for Production & Security in a Biological Economy” programme (Contract ID C10X1306). AgResearch colleagues have provided testing and suggested improvements of this software.

References

Dodds, K G, McEwan, J C, Brauning, R, Anderson, R A, Van Stijn, T C, Kristjánsson, T and Clarke, S M (in prep) Construction of relatedness matrices using genotyping-by-sequencing data.

Lu, F, Lipka, A E, Glaubitz, J, Elshire, R, Cherney, J H, Casler, M D, Buckler, E S and Costich, D E (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics* **9**, e1003215.