ms2gs:
A combined coalescence gene dropping tool for evaluating genomic selection in complex scenarios

Miguel Pérez-Enciso and Andrés Legarra

ICREA and Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain

INRA, UMR 1388 GENPHYSE, Castanet-Tolosan 31326, France

Contact
miguel.perez@uab.es

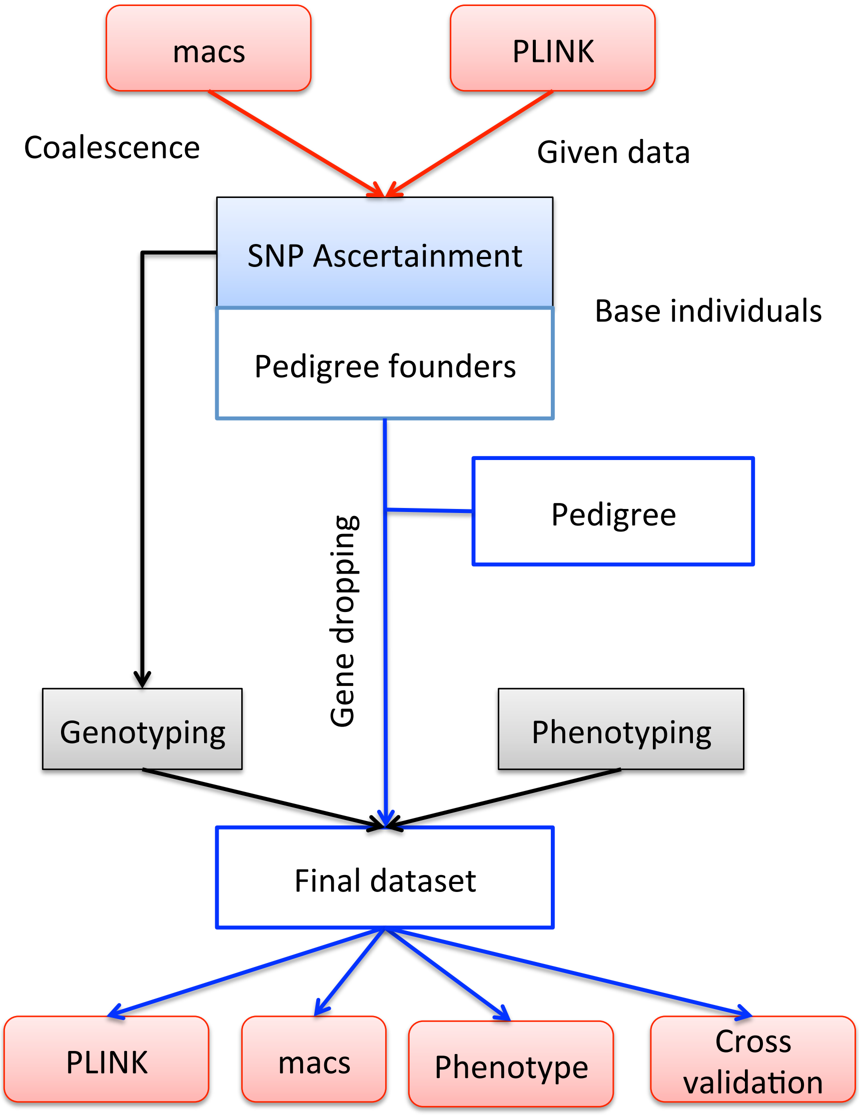
Overview

This program is a combined coalescence – gene dropping (i.e., backward - forward) simulator for complex traits. It therefore aims at combining the advantages of both approaches. It is primarily conceived for very short term recent scenarios (where mutation can be ignored) such as those that are of interest in animal and plant breeding. No new mutations are assumed to occur in the gene dropping part. The way ms2gs combines both strategies is to start with coalescence simulated genotypes, and follow a pedigree provided by the user. If no pedigree is provided, all samples are treated as unrelated (from a gene dropping perspective). The coalescence data are read from a file, in either ms or plink format, or a ms command can be included in the parameter file. This command is then executed automatically every new replicate. If the genotypes are provided in a file, either ms or plink format, the same starting genotypes are used in every replicate.

The program generates complex phenotypes using a variety of options. There is extreme flexibility as to the number of causal loci and their effects: specific effects and qtl positions can be provided, or the program chooses locations from the whole genome or specified regions and effects can be sampled according to specified distributions. Effects are adjusted such that the desired heritability is obtained. Binary traits are implemented via a threshold model. Dominance and additive effects are implemented, but no epistasis yet. Any number of chromosomes and varying recombination rates along each chromosome can be specified.

By default, the program estimates breeding values using BLUP and genomic selection (G-BLUP) using three SNP sets: (1) sequence (using all SNPs); (2) array (ascertained SNPs), (3) only the causal SNPs. The SNPs to be included in an array can be specified either from a file or chosen randomly according to several filtering criteria, such as minimum allele frequency or a percentage of SNPs. SNPs can be ascertained from all individuals in the base population (i.e., those whose genotypes are simulated with the coalescence) or from subset. Note that, if no pedigree file is provided, BLUP cannot be run.

The Figure shows a scheme of ms2gs workflow. First, genotype data of base individuals are generated with the coalescence, or read from a plink or ms format file. Part of the base individuals are used for SNP ascertainment for the chip and part as founder individuals in the pedigree. Next, gene dropping is performed following a user provided pedigree. The whole pedigree individuals plus pedigree founders are 'genotyped' with the chip, the causal mutations and all SNPs (sequence). In addition, phenotypes are generated given the causal genotypes, the specified distribution of QTL effects and desired heritability. The program finally provides genotype and phenotype output in either Plink or ms format, and statistics regarding cross-validation or goodness of fit.



Installation

To compile, simply type

```
make
```

An executable ms2gs is produced. To install in /opt/ type

```
sudo make install
```

Afterwards, you may want to remove intermediate files with
make clean

In addition, ms2gs requires either an input file in ms or plink format. For that, we recommend macs program (https://github.com/gchen98/macs). For use of macs see below.Alternatively, you can specify the call to any other coalescence program that outputs ms format within the parameter file.

Running ms2gs

The minimum command to run ms2gs is:

```
./ms2gs –i PARFILE
```

where PARFILE is the parameter file. Only -i flag is mandatory, optional flags are:

```
./ms2gs -iP ARFILE-quiet -h -niter NITER-seed SEED -ms MSFILE -snp SNPFILE -rsnp
SNPREGFILE
```

where '-quiet' suppresses STDOUT output, -h prints help and exit, NITER is the number of replicates, SEED is random generator seed, MSFILE is the ms file containing the genotypes, SNPFILE is the file containing the chip SNPs, and SNPREGFILE is the file containing the regions where SNPs are sampled from. All options can be also specified in the parameter file, except '-quiet' and '-h'.

WARNING: in case of conflict, command line options override those in the parameter file.

VERY IMPORTANT: ms2gs does NOT run in windows and will never do even if it may compile because the code uses system calls specific to linux. So far tested, it does run in mac if compiled with gfortran.

The ms format and ms commands

The ms format has been the standard approach of coalescence programs, starting with the widely popular ms program by coalescence inventor Rick Hudson (http://home.uchicago.edu/rhudson1/). The format is:

```
ms-command
seed
[blank line]
//
segsites: n_snps
positions: pos1 ... pos_n_snp
haplotype1
haplotype2
...
haplotype_n
```

where positions are given in relative positions from 0 to 1; ms2gs adjusts these positions to actual base pair positions by multiplying those by the chromosome length specified in par file. Haplotype is a list of SNP alleles as 0 and 1 without spaces, 0 for the ancestral allele and 1 for the derived one. The length in characters is therefore the number of SNPs. The program assumes that two consecutive haplotypes make up a diploid individual genome. To specify several chromosomes, a single file with several runs of the coalescence (one pertaining to each chr) must be specified. The program automatically recognizes whether one or more chromosomes are run.

You can specify ms data to feed ms2gs in two ways. Either in section MSFILE you specify a file with the output of a previous ms run or with a MSCOMMAND option. In this latter case, a system call is specified in each iterate of ms2gs and the command is run. A typical MSCOMMAND line using macs software (https://github.com/gchen98/macs) can be

```
macs200 100000 -i 3 -t .002 -r .001 -h 0.01 2>/dev/null | msformatter
2>/dev/null
```

where 200 is the number of sequences (ie, 100 diploid individuals), 100000 is the length of the chromosomes simulated, three chromosomes (ie, three replicates of the program are run, -i 3), the global variability is 0.002 per base pair per 4N generations, and recombination rate is 0.002; flag -h specifies how many nucleotides history is tracked for optimizing running time (see the macs publication). The rest of the command simply allows for a ms format file to be written. Note that macs should be in the main pathway. If in the working directory, you can use

```
./macs200 100000 -i 3 -t .002 -r .001 -h 0.01 2>/dev/null | ./msformatter
2>/dev/null
```

command instead. For a comprehensive list of ms commands and examples, read the excellent documentation of Hudson's ms software (http://home.uchicago.edu/rhudson1/source/mksamples.html). Note that original ms program requires parameters along the whole DNA length, whereas macs reads parameter values per base pair.

IMPORTANT: For a best usage of ms2gs, you should be comfortable with the coalescence simulation. An excellent practical introduction is the ms manual by Rick Hudson.

WARNING: For speed and number of digits in positions, we strongly recommend macs over original Hudson's ms program.

The PLINK format

If section PLINKFILE is specified with value plinkfile, files plinkfile.ped and plinkfile.map must exist (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml). The format for map file is

```
chr   snp_id      pos_m       pos_bp
```

with as many rows as SNPs, and where chr is chromosome id (integer number from 1 onwards), snp_id is SNP id, pos_m is position in Morgans and pos_bp is position in base pair units. The ped file is

```
Family_id      id      paternal_id      maternal_id  sex  phenotype allele1_snp1  allele2_snp1
allele1_snp2 ...
```

NOTE: snp_id, pos_m, family_id, paternal_id, maternal_id, sex and phenotype can be 0 and are irrelevant to ms2gs.

WARNING: chr must be an integer 1,2, ... nchr. Alleles must be coded 1 and 2, and no missing values are allowed.

How individual ids are managed

The program assumes individual ids are numbered consecutively from 1 onwards, and that the first individuals are those in the ms file, which are the founder or base population individuals. By default, ms2gs assumes diploid individuals and that two sequences in the ms file correspond to one individual, although this can be changed if section INBRED_FOUNDERS is present. The program recognizes how many pairs of sequences are in the file and automatically assigns them sire and dam as unknown. Suppose the ms file contains 10 individuals (20 sequences), the following pedigree file

```
1 0 0
2 0 0
5 0 0
11 1 2
12 5 2
```

means that the first three individuals in the pedigree correspond to haplotypes in ms files 1,2 (ind 1) 3,4 (ind 2) and 9,10 (ind 5), which produce offspring 11 and 12 based on their parents genotypes, and so on. Internally, the rest of ms data is disregarded except for two purposes:

1- To estimate QTL effects: There are several options to specify QTL effects (as explained above) and some need ascertaining causal SNP frequencies. The individuals to calculate these frequencies from the ms file are determined with section QTL_INDS in parameter file, where the first and last individual are specified.

2- To ascertain SNPs: One of the interesting features of ms2gs is that it allows a 'realistic' way (or at least no too simplistic) to mimic SNP ascertainment bias. The individual range where SNPs are ascertained is determined in section ASCERTAINMENT_INDS in par file, filtered by allele frequency if desired (ASCERTAINMENT_P). This allows to ascertain SNPs in a sample of individuals but use other set for gene dropping. For instance, id 1 to id n can be used for ASCERTAINMENT_INDS, and pedigree start with id n+1. This permits to simulate the realistic approach of no individual used for SNP ascertainment to participate as well in the actual pedigree.

IMPORTANT: It is possible to specify homozygous base individuals with section INBRED_FOUNDERS in the par file. If this section is present, then one line in ms corresponds to one individual (a dihaploid). This option is useful to study typical plant designs.

Specifying QTL architecture

There are four ways to determine QTL effects:

1. Additive and dominant QTL effects are specified for each QTL in sections QTL_EFFECT_A and QTL_EFFECT_D.
2. Additive and dominant QTL variances are specified for each QTL in sections QTL_VA and QTL_VD. Then, conditional on QTL frequencies and variances, the additive and dominant values are solved for.
3. The distribution of additive and dominant effects be specified in sections QTL_DISTRIBUTION_A and QTL_DISTRIBUTION_D. In this case, ms2gs samples a value for the additive or dominant effects for each QTL, using one of three distributions (uniform, normal or gamma).
4. The distribution of additive and dominant variances can be specified in sections QTL_DISTRIBUTION_VA and QTL_DISTRIBUTION_VD. In this case, ms2gs samples a value for the additive or dominant variance for each QTL, using one of three distributions (uniform, normal or gamma). Then, conditional on QTL frequencies and variances, the additive and dominant values are solved for.

In all cases, environmental variance is adjusted such that the broad heritability is as specified (H2).This is done by assuming complete equilibrium between QTL. If section ADJUST_VE is present, then exact genetic variance is computed with actual haplotypes and environmental variance is adjusted such that h2 value is exactly as defined in the base population. In practice, we did not find large differences with ADJUST_VE but this may vary in each case.

WARNING: Only one of the options above should be present in the par file, but different options can be used for additive and variance effects, e.g., QTL_DISTRIBUTION_D and QTL_EFFECT_A.

Specifying QTL positions and frequencies

The user must specify the number of QTL in NQTL section. Then, by default, ms2gs samples QTL positions randomly among all available SNPs, irrespective of frequency. However, the precise causal SNPs can be determined in section QTL_POS. Alternatively, QTL can be sampled only among SNPs in predetermined regions with section QTLREGFILE. This can be done eg to mimic that only genes can contain QTL, and specify genes in given genome regions.

By default, no restriction on QTL frequency is set up, but can this be restricted with section QTL_FREQ_RANGE.

WARNING: ms2gs samples QTL positions given conditions setup by the user, but if these cannot be fulfilled (e.g. a large number of QTL in a very narrow region) the program crashes.

Determining SNPs to be used in genomic selection

By default, ms2gs predicts breeding values with BLUP, all SNPs, all causal SNPs and a set of ascertained SNPs to mimic a chip. The SNPs used in the array are sampled randomly using either of these two options:

1. The chip SNPs can be determined in a file using section SNPFILE. This file contains the precise set of SNPs to be used in GBLUP.
2. The chip SNPs can be sampled from all SNPs subject to conditions:
   - ASCERTAINMENT_INDS: SNPs are chosen from those segregating in the set of individuals specified in this section.
   - ASCERTAINMENT_MAF: SNPs must have at least minimum MAF specified.
   - ASCERTAINMENT_P: a percentage p of the possible SNPs is sampled.
   - SNPREGFILE: potential SNPs must be located within regions specified in this file.

WARNING: ms2gs samples SNPs given conditions setup by the user, but if these cannot be fulfilled the program crashes.

Simulating causal SNPs when only sparse genotyping is available

It is often the case where the user has access to high or low density genotyping, but no sequence, and the interest is to simulate QTL conditional on these data. ms2gs provides a relatively simple approach specified with SIMULATE_QTL_SNP option. In section `SIMULATE_QTL_SNP_R2` the r2 (the average disequilibrium between the causal locus and the nearest left SNP) is specified. Genotypes of a hypothetical causal QTL are simulated conditional on allele frequencies of the extant SNP and following a distribution for the QTL frequency, which can be uniform SIMULATE_QTL_SNP_FUNIFORM)or following a neutral distribution (SIMULATE_QTL_SNP_FNEUTRAL). In the former case, the lower and upper bounds of the ancestral allele must be specified. In the latter case, the program samples an allele frequency from the expected spectrum under a neutral scenario, that is, q=1/f. The rest of options for QTL work as usual, that is, number or QTL regions. In this setting, the NOSEQ option is recommended, as only chip data is available. Note nevertheless that this option can be combined with the availability of complete sequence as well, but we do not think is worthwhile in this case.

Evaluating BLUP and GBLUP

ms2gs perform a goodness of fit and crossvalidation analysis. In the first case, the correlation between true and predicted breeding values, obtained with the complete phenotype dataset, is printed for each of the methods. This analysis can be suppressed with the NOFIT option.

In addition, or instead, ms2gs also performs crossvalidation, computing correlation between true and predicted breeding values obtained when phenotypes are removed from these individuals. The range of individuals to be removed as specified in section IDXVAL, a percentage PXVAL of these is removed, and NXVAL replicates are run. For instance, if one is interested in predicting only the youngest animals breeding values, say id 100 to id 150 one could enter

```
IDXVAL
100   150
PXVAL
1
NXVAL
1
```

which means that all ids 100-150 are removed and one round of cross validation is run.

By default, all individuals in the pedigree are used for both BLUP and GBLUP. However, with section KILLINDS, the range of individuals removed from GBLUP analyses is specified.

WARNING: KILLINDS affect only GBLUP, BLUP is computed using all pedigree and phenotypes present in the pedigree.

Implementing sequencing and genotyping errors

NGS data are quite noisy, yet their errors are complex and difficult to simulate; ms2gs provides a very simple way to simulate errors, which is based on the observation that there is usually a bias towards the reference allele (0) and against heterozygous SNPs. This is governed by parameter lambda, which is specified in section P_ERROR_NGS, and ranges from 0 (no error) to 1. Imputation errors in sequence are specified in P_ERROR_IMP. Chip genotyping errors in P_ERROR_CHIP.

Parameter file

The parameter file has different sections, which names **MUSTBE IN UPPERCASE**. The section names are below. See also the examples in the documentation. Comments can be included with '!' or '#' in the start of a line, in any line of the par file.

```
# this is a comment
! this is also a comment
NOBLUP    !--> does not perform blup
NOCAUSAL  !--> does not perform gblup with causal mutations
NOSEQ     !--> does not perform gblup with sequence data
NOCHIP    !--> does not perform gblup with snp data
NOFIT     !--> does not perform fit analysis

NXVAL     !-->no of xvalidation iterates
# this is a comment but not recommended for legibility
nxval

IDXVAL    !--> sample ids range to perform xvalidation
idx1 idx2

PXVAL     !--> % of samples in idx1-idx2 range deleted for xvalidation
pxval
```

```
NITER    !--> no. of iterates, overriden by command line -iter
niter

MSCOMMAND !--> ms command that can be called from system, overrides MSFILE
mscommand
# a simplest ms command for ms2gs usage can be
# macs 10 30000 -t 0.001 2>/dev/null | msformatter 2>/dev/null
# which simulates 10 sequences 30kb long with diversity 0.001

MSFILE   !--> file with ms format data
msfile

PLINKFILE !-->plink file prefix to read genotypes(instead of MSFILE)
plinkfile! filespfile.ped and pfile.map must exist

OUTFILE   !-->outfile prefix
outfile

PEDFILE   !-->pedfile to carry out gene dropping w/o selection
pedfile

SNPFILE   !-->snpfile with snps to be used in chip
snpfile   !    overrides ASCERTAINMENT section

SNPREGFILE !--> SNPREG FILE,: chr, pos1, pos
snpregfile

KILLINDS   !-->ind range removed for GBLUP analyses
idk1 idk2

INBRED_FOUNDERS !--> homozygous founders (no. sequences = no. baseinds)

CM2MB !--> general cM2Mb ratio [1 by default]
cm2mb

MAPFILE    !-->mapfile with local xover rates, default otherwise
mapfile

PRINT_SOL !--> prints snp solutions for every replicate and method

PRINT_PLINK !-->prints plink files every replicate

PRINT_YG    !--> prints phenotype, true breeding value and ebv file

NCHR  !--> no. ofchrs
nchr

CHR_LENGHTS !-->chr lengths in bp, all in one line
chr1_lengthchr2_length ...

H2     !-->broad heritability of the trait
h2

BINARY_TRAIT  !-->incidence if a threshold trait simulated
Incidence

ADJUST_VE   !-->var e is adjusted to match exactly expected h2

KILLINDS   !--> deletes marker info for those ids range in GBLUP
idk1  idk2

NQTL    !--> no. ofqtls
```

```
nqtl

QTL_INDS    !--> QTL parameters computed using ind idq1 to idq2
idq1  idq2

QTL_SIGN    !--> probability of derived allele being deleterious [0.5]
pderived

SIMULATE_QTL_SNP_R2 !-->simul qtl snps with r2 nearest marker
r2

SIMULATE_QTL_SNP_FNEUTRAL !-->simul qtl freq neutral distribution

SIMULATE_QTL_SNP_FUNIFORM !-->simul qtl uniform freq frq1 frq2
frq1  frq2

QTL_POS !--> fixed qtl positions, ipos is snp order within chr
qtl1_chr qtl1_ipos
qtl2_chr qtl2_ipos
...

QTLREGFILE    !-->qtl positions are sampled within regions in this file
qtlregfile

QTL_FREQ_RANGE !-->QTL freq range on which causal snps are chosen from
frq1 frq2

QTL_VA !-->specified additive variance per qtl
qtl1_va qtl2_va ...

QTL_VD  !-->specified dominant variance per qtl
qtl1_vd qtl2_vd ...

QTL_EFFECT_A !-->specified additive effects per qtl
qtl1_a qtl2_a ...

QTL_EFFECT_D  !-->specified dominant effects per qtl
qtl1_d qtl2_d ...

!--> in sections below distribution can be one of
! u b1 b2   (uniform, lower bound, upper bound)
! n mu var  (normal, mean and variance)
! g s b     (gamma, shape and rate parameters)

QTL_DISTRIBUTION_A  !--> QTL add effects are sampled from a distribution
[u, l_bound, u_bound], [n, mu, var], [g, s, b]

QTL_DISTRIBUTION_D !--> QTL dom effects are sampled from a distribution
[u, l_bound, u_bound], [n, mu, var], [g, s, b]

QTL_DISTRIBUTION_VA  !--> QTL add variances are sampled from a distribution
[u, l_bound, u_bound], [n, mu, var], [g, s, b]

QTL_DISTRIBUTION_VD  !--> QTL dom variances are sampled from a distribution
[u, l_bound, u_bound], [n, mu, var], [g, s, b]

ASCERTAINMENT_INDS  !-->snps are ascertained from ids ida1 to ida2
ida1 ida2

ASCERTAINMENT_MAF  !-->snps are ascertained with minimum maf
maf

ASCERTAINMENT_P  !--> a fraction p of snps is chosen
```

```
p

SEED  !--> random seed can be overriden in command line by -seed
seed!    WARNING: does not affect ms program, use –s seed flag in macs

KMIN !--> min allele count for a snp to be considered
kmin

P_ERROR_CHIP !--> genotyping error
perr_chip

P_ERROR_NGS  !--> base sequencing error
perr_ngs

P_ERROR_IMP  !-->imputation error
perr_imp
```

WARNING: SEED does not affect the coalescence program. In macs, you should use -s seed, to initialize the seed. In ms, use -seed flag.

ms file / Plink file (MSFILE / PLINKFILE, mandatory unless MSCOMMAND)

These files contain the starting genotypes. Only MSFILE or PLINKFILE can be specified. In the ms format, several chromosomes are identified by several runs of the program. The actual SNP positions are obtained by multiplying the 0-1 value in the ms file with the chromosome length specified in the parameter file. For plink files, only the prefix needs to be specified, and files prefix.ped and prefix.map are assumed to be present.Either MSFILE, PLINKFILE or MSCOMMAND are mandatory.

WARNING!original ms program by Hudson provides very few digits in the SNP positions and this may result indifferent SNPs to be localized in the same position causing unpredictable behavior. **We do recommend macs**(https://github.com/gchen98/macs) program instead.

Pedigree file (PEDFILE, mandatory)

Contains the pedigree used to carry out the gene dropping (forward) simulation. Founders should be in ms file (two haplotypes per individual). Ids should be numbered 1 to n, unknown parents are coded as 0. The format is:

id idsire iddam

SNP file (SNPFILE, optional)

Contains the exact SNPs to be used in the chip. The format is

chr isnp

where isnp is order of SNP in that chromosome, one row per SNP.

WARNING: This option is not tested with SIMULATE_QTL_SNP, as this implies inserting SNPs internally, and order may vary.

WARNING: It overrides the rest of ascertainment sections.

SNP region file (SNPREGFILE, optional)

Specifies regions from which chip SNPs can be sampled from. If not present, SNPs are sampled from the whole available set. The format is

```
chr bp1 bp2
```

where bp1 and bp2 are region limits in base pairs.

QTL region file (QTLREGFILE, optional)

Specifies regions from where QTLs are sampled from. If not present, causal SNPs are sampled from the whole available set. The format is

```
chr    bp1    bp2
```

where bp1 and bp2 are region limits in base pairs.

Map file (MAPFILE, optional)

The file contains local recombination maps. The format is

```
chr     bp    local_cM2Mb
```

which means that recombination rate between the preceding bound and bp (base pair limit) is local_cm2Mb. The first bound specifies between position 0 and first bp bound. Unspecified regions are assumed to have default cM2Mb ratio (1 cM/Mb).

WARNING: Map file is used only for gene dropping simulation; coalescence macs and ms have options to simulate recombination hotspots.

Output

The program prints on STDOUT the main results for genome, qtl positions and effects, and so on. If one is interested say in getting the cross-validation results only, simply execute

```
./ms2gs -i ms2gs.par | grep  XVAL
```

Most of output except warnings can be suppressed with the -quiet option in the command line:

```
./ms2gs -i ms2gs.par -quiet
```

If PRINT_SOL option, the program prints SNP solutions for every criterion (chip, sequence or qtl).
If PRINT_YG, the program prints SNP phenotypes, true and predicted breeding values per individual.
If PRINT_PLINK, the program prints genotype data for every criterion (chip, sequence or qtl) in plink format.

WARNING: As these files can be large, it is recommended to use them only with one or a few iterates.

Examples and practicals

A set of ms2gs files is in the examples directory. File ms2gs.par is a generic parameter file, and can be readily used for testing but is not realistic, accompanying files are ms2gs.ped, ms2gs.out(an output from an iterate), and ms2gs.ms (the output of a coalescent run). Files starting with 'hs' concern a half sib population, whereas mag* files concern MAGIC lines.

By default, chip SNPs and QTLs are uniformly distributed throughout the genome. To fine-tune the number of SNPs you desire in the array, modify ASCERTAINMENT_MAF (the minimum allele frequency) and ASCERTAINMENT_P (the fraction of SNPs fulfilling the MAF criterion that are selected). By default, MAF is 0 and P is 1, ie, no SNP selection. Even without macs installed, you can still run ms2gs using the files provided (*.ms) as input. Note though that this implies using the same input per replicate.

Unequal QTL distribution

This can be achieved with QTLREGFILE section. In the hs1pop.par example, QTLs can be distributed only in 30 regions of 10 kb each, as determined by hs.qtlreg file. To make QTL be distributed uniformly along the whole genome, simply comment the section title, as is done in file hs*.par files.

RAD sequencing

RAD genotyping can be mimicked by specifying a SNP region file in SNPREGFILE section, and suppressing all ASCERTAINMENT_* sections. The region file should be of segments of length equivalent to RADs, eg, a few kb. This makes that all SNPs within those regions are used in the chip option. Check hs1rad.par file.

SNP ascertainment in the same or different populations

Fileshs1pop.par and hs2pop.par present two examples with one or two populations, each of 250 diploid individuals. In file hs1pop.par SNPs are ascertained from the first 250 individuals, but the pedigree starts in individual 251 (hs.ped file), but they all belong to the same population, as specified by the ms command. You can simulate a complex scenario to mimic that SNPs are ascertained in a variety of populations. In file hs2pop.par two populations are simulated each of 250, individuals. In this case SNPs are ascertained in pop 1 and the pedigree comprises only individuals in pop 2.

Effect of SNP density

I am afraid that the only way so far is to run ms2gs several times and modify ASCERTAINMENT_Pand ASCERTAINMENT_MAF values. Probably you should use the NOSEQ and NOBLUP options to avoid re-computing these options every time.

Printing outputs

Use the PRINT_YG, PRINT_PLINK and PRINT_SOL options. These options print all relevant data and can be used eg to test other GS or GWAS methods. Since they print a concatenated file with all files per iterate, **it is recommended to use these options with one iterate only.**

MAGIC line pedigree

MAGIC (Multiparent Advanced Generation Inter-Cross) lines have become a popular tool in some plant species, and consistof crosses between multiple inbred lines followed by inbreeding. To simulate MAGIC lines with ms2gs, one option is to simulate with the coalescence several populations sampling one sequence per population and making it dihaploid with INBRED_FOUNDERS section, whereas the crossing and inbreeding is specified with the pedigree file. Further, we make use of option KILLINDS to analyze only the final 190 inbred individuals. We compare two designs starting with 20 inbred lines (magic.par file), a F2 followed by self fertilization (magic.ril.ped) or crosses between all pairs of lines followed by 5 generations of random mating (magic.ran.ped).

No pedigree

If no pedigree file is specified, ms2gs assumes unrelated individuals (from a pedigree point of view).Logically,BLUP evaluation makes no sense in this case. For that, comment out the PEDFILE section.

Acknowledgments

Work was funded by project AGL2013-41834-R to MPE. AL acknowledges financing from the project GenSSeq of the INRA metaprogramSelGen (France).

Citations

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Research**19**, 136–42.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics**18**, 337–338.

Pérez-Enciso M, Rincón JC, Legarra A. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. GenetSelEvol. 47:43. doi: 10.1186/s12711-015-0117-5.

Pérez-Enciso M, A. Legarra A.A combined coalescence gene-dropping tool for evaluating genomic selection in complex scenarios (ms2gs). J Anim Breed Genet (submitted)