Manual for AllOr

Version 7.3

June 2022

Introduction

AllOr (**All**ele **Or**igin) is a program for assigning breed of origin to alleles of crossbred animals. AllOr was developed for crossbred dairy cows in a three way rotational crossbreeding with one parent purebred and assuming big proportion of purebred recent ancestors to be genotyped, e.g. most sires and maternal grandsires. It has been extended to be able to handle situation with both parents being crossbred but the accuracy of assignment has not been assessed for that situation. It can handle a maximum of 8 defined pure breeds. The program is written in Fortran.

AllOr requires the following input to work:

Genotypes of the crossbred animals, phased to two haplotypes. No missing genotypes allowed. The program runs on 1 chromosome at a time so the input haplotypes must be split into individual chromosomes before the analysis.

Genotypes of purebred animals from all breeds contributing to the crossbred, phased to two haplotypes. No missing genotypes are allowed. All genotypes, both for crossbred and purebred animals, must be imputed to the same set of SNPs. Genotypes of recent ancestors, for example sires and maternal grandsires, should be included as much as possible.

Pedigree file, connecting the crossbred animals to genotyped ancestors. (not strictly necessary, see details below)

Code indicating which pure breeds can contribute to each crossbred animal, see details in description of input files. If the contributing breeds can be read from the first digits of IDs in the pedigree, that can also be used for determining breed composition.

For further information write to jonh@qgg.au.dk

Input files

AllOr requires six input files. One file with parameters to control the program, paths to the other files and information about the analysis, one pedigree file, one file with haplotypes of crossbred animals and one with haplotypes purebred animals, one file with IDs and breed information of the purebred genotyped animals and one with codes indicating breed composition of the crossbred animals. Following are details on the necessary files.

control.txt

Fixed file name .

Should be placed in the folder were the program is run.

Text file with a minimum of 19 lines, giving AllOr information about the analyses, paths to other input files and some options. The lines should have format and information according to the following description. File names given in brackets refer to explanations of the files below.

<int>=positive integer, <stringX>= character string with a maximum of X characters, <real>=real number. All values should be separated by space. File names in brackets refer to description of the file below and the names in the provided example file.

<int> number of SNPs

<int> number of pure breeds

<string5> <int> code for pure breed 1; number of reference genotypes for breed 1

<string5> <int> code for pure breed 2; number of reference genotypes for breed 2

…. Eventually more lines if more breeds. Number of lines with pure breeds has to match number of pure breeds in line 2.

<int> Number of defined breed combinations of crossbred which will be defined in the lines below. This information is used to avoid assigning alleles to breeds that are not contributing to the crossbred animal.

<string5> <int> <int> ... One code and 2 or more integers. Breed code for first type of cross; number of possible sire breeds; sire breed(s), numbered according to order in lines 3-… For example "ABC 2 1 2" would mean that crossbreds with code ABC could have two possible paternal breeds, and those are the first two breeds given in the purebred code definition above.

<string5> <int> <int> ... One code and 2 or more integers. Breed code for first type of cross; number of possible dam breeds; dam breed(s) numbered according to order in lines 3-…

<string5> >int> <int> ... One code and 2 or more integers. Breed code for second type of cross; number of possible sire breeds; sire breed(s), numbered according to order in lines 3-…

<string5> <int> <int> ... One code and 2 or more integers. Breed code for second type of cross; number of possible dam breeds; dam breed(s) numbered according to order in lines 3-…

… Eventually more lines describing more types of crosses. In total, the lines describing the breed codes have to be exactly two times the number of defined breed combinations provided above. If the user does not want to restrict the breeds that can contribute to some or all animals, a breed dode can be defined where all breeds are defined as possible paternal and maternal breeds.

<int> Number of lines in pedigree file (not used, but a number has to be provided)

<int> Depth for looking for ancestor genotypes, max 5

<string60> Name and path of pedigree file (ped.txt)

<string25> Code for missing parent in the pedigree file

<string60> Name and path of file with IDs and breed codes of purebred (boacode_pb.txt)

<string60> Name and path of file with purebred haplotypes (haplotypes_pb.txt)

<int> Number of crossbred animals to analyze

<string60> Name and path of file with IDs and breed codes of crossbred (boacode_cb.txt)

<string60> Name of file with haplotypes of crossbred (haplotype_cb.txt)

<int> <int> Window length and number shifted. For example 100 5. See more on recommended values in "Notes on parameters and input" below.

<real> Proportion of positions that need to match to say that haplotypes within window (e.g. 0.99). See more on recommended values in “Notes on parameters and input” below.

<int> Gap-filling. Either 0,1 or 2, where 1 is recommended. Controls if gaps are filled based on neighboring loci. 0=no fill 1=only if <2*WL and assigned on both sides, 2: also filled in in ends. Other numbers are also accepted here but result in different parts of the program to be skipped and is only meant for debugging.

<string1> y/n Output a file, breedoforigin2.txt, with code for possibilities for unassigned loci

ped.txt

Name and path can be changed in control.txt

Pedigree file. No header. One line per individual with three values: ID, sireID, damID, space separated. IDs can be numerical or alphanumerical, maximum 25 characters. Code for missing parents can be set in control.txt. Animals with genotype should be included with their own line, if they are just as parents, the pedigree relationship is not used to facilitate assignment.

boacode_pb.txt

Name can be changed in control.txt

File with number of lines equal to number of purebred animals with genotype in the reference. One line per purebred animal with animal ID and breed code as described in control.txt. No header, the two values space separated. If crossbred code in boacode_cb.txt for some animals is not defined and breed composition is to be determined from IDs in pedigree, the codes for the purebreds need to match the first three characters in IDs of purebred from the breed.

haplotype_pb.txt

Name can be changed in control.txt

Two lines per purebred animal in the same order as boacode_pb.txt. Each line has one haplotype coded as 1/0, space separated. No header.

boacode_cb.txt

Name can be changed in control.txt

File with number of lines equal to number of crossbred animals with genotype. One line per crossbred animal with animal ID and breed composition code as described in control.txt, space separated. No header. The reason for including codes for crossbred combination is to be able to exclude breeds that are not expected to contribute to the individual crossbred. If breed composition is completely unknown, the animal can still be included with breed code defined to have all breeds included. If the provided breed code in this file is not defined in the control file, the program will try to find the breed information from first three characters of ancestors in the pedigree, that is the first three characters are assumed to match the purebred breed codes. If that does not work the program will not exclude any breeds, which is not likely to result in accurate estimates of breed of origin.

haplotype_cb.txt

Name can be changed in control.txt

Two lines per crossbred animal in the same order as boacode_cb.txt. Each line has one haplotype coded as 1/0, space separated. No header.

Notes on parameters and input

Genotypes of purebred

How many genotypes are needed from the pure breeds has not been investigated thoroughly. To give some numbers, in testing the program on real data 2500 genotypes from each breed gave low (<1%) percentage of alleles not assigned to breeds for analysis of relatively simple crosses of three dairy cow breeds. Reducing the number of genotypes from one of the pure breeds at a time, from 2500 to 200 had limited impact when the breed with reduced number was had been separated from the other breed for a long time. However, 1-2% of alleles changed assignment or were unassigned when the breed with reduced number was shared fairly recent genetics with another breed in the evaluation. All genotypes of purebred are useful for assignment but genotypes of ancestors of the crossbred animals are more useful, the more related the better.

Pedigree

The pedigree is only used to connect crossbred animals to genotyped direct ancestors and traces the pedigree from crossbred animal for a maximum of 5 generations. Including pedigree information on animals that are not direct ancestors of the CB animals is therefore not useful. If genotyped (crossbred or purebred) animal is only in the pedigree as parent, not with a line for itself, then the program will not be able to make use of the relationship. Crossbred animals can be included in the analysis even if they have no genotyped ancestors or no pedigree information, but the assignment will be less accurate. If someone wants to try without using pedigree information then that is possible, but some file needs to be included for pedigree file, for example just list of IDs of crossbreed with 0 0 for parents. If the three first characters in the IDs are breed codes given for purebred codes in control, that information can be used instead of providing breed codes of purebred.

Window length (WL)

What is the optimum WL might depend on the situation. Testing indicated that 100 was better than 150 and 200 on simulated simple crosses and WL of 50 has also given good results in tests by others. The program is generally faster with shorter WL and expected to assign higher proportion to definite breed. Risk of errors might however increase with shorter WL. For many generation of crossbreeding short WL (<100) is probably more appropriate but for simple (e.g. F1) crosses with many genotypes of recent purebred ancestors included in the analyses longer WL is probably saver strategy.

Number of SNPs shift between rounds (NS)

Low number increases accuracy but running time is proportional to 1/NS so NS=1 takes 5 times longer than NS=5. Testing showed considerably better results with NS=5 than NS=20.

Proportion to match

In order to allow some flexibility, for example genotype errors in the haplotype comparison step, the limit of proportion of alleles that need to match for haplotypes to be considered to match can be less than 1. A value of 0.99 was used when testing the program, which means 1 allele was allowed not to match across a window of 100 SNP. Note that keeping the proportion in 0.99 with window length <100 will not allow any mismatch. Lower proportion is expected to reduce number of unassigned alleles, but might increase risk of errors.

Output files

allor.log

Log file with information from the running of the program.

breedoforigin.txt

File with the estimated breed of origin of the alleles. The format is the same as haplotype_cb.txt, that is two lines per crossbred animal, each line with one number for each SNP on the chromosome. The numbers indicate the estimated origin of the respective allele in haplotype_cb.txt, 1 being the first breed defined in control.txt, 2 the next and so on. Alleles that could not be assigned to one breed are indicated with the number 9.

breedoforigin2.txt

Only written if the last line in control.txt starts with "y"

File with the estimated breed of origin of the alleles with alternative coding where information about probability of breeds for the unassigned alleles is included. The format is the same as breedoforigin.txt, that is two lines per crossbred animal, each line with one number for each SNP on the chromosome. The numbers indicate the estimated origin of the respective allele in haplotype_cb.txt. The coding is translated in bo2code.txt. Note that for example "3" in breedoforigin2.txt does not mean the allele is from breed 3.

bo2code.txt

Explanation of the breed code in breedoforigin2.txt. Each line starts with a integer code that matches the code in breedoforigin2.txt, followed with one number for each pure breed, giving the probability that the allele is from that breed. Alleles that are assigned in breedoforigin.txt have probability 1 for one of the breeds, 0 for others. If two breeds are possible, the two will generally have 0.5, and other breeds 0 for example. More than one code can result in the same probabilities.

Propassigned_animal.txt

One line for every crossbred animal with the following information:
Animal ID
%assigned haplotype1
%assigned haplotype2
Number of SNPs of haplotype 1 assigned to pure breed 1
Number of SNPs of haplotype 1 assigned to pure breed 2
…
Number of alleles from haplotype 1 not assigned to breeds
Number of SNPs of haplotype 2 assigned to pure breed 1
Number of SNPs of haplotype 2 assigned to pure breed 2
…
Number of alleles from haplotype 2 not assigned to breeds

snpsummary.txt

Percentage assigned for each marker: three numbers in every line:
SNP number, % assigned haplotype 1, % assigned haplotype 2.

Example of control.txt

```
2168
3
AAAA 1200
BBBB 1200
CCCC 1200
4
BBAA 1 2
BBAA 1 1
CCBA 1 3
CCBA 2 1 2
AACB 1 1
AACB 3 1 2 3
BBAC 1 2
BBAC 3 1 2 3
33600
4
ped.txt
0
boacode_pb.txt
haplotype_pb.txt
1000
boacode_cb.txt
haplotype_cb.txt
100 5
0.99
1
y
```