Genetic Management Toolkit

Michael Raboin

August 14, 2015

Overview

The Genetic Management Toolkit is a web application developed at the Oregon National Primate Research Center (ONPRC) to facilitate some of the analyses that we perform regularly.

At present, the application is designed to support 3 functions:

1. Quality control of uploaded studbooks
2. Generation of Genetic Value Analysis Reports
3. Creation of potential breeding groups

Quality Control

Studbooks maintained by breeding colonies generally contain information of varying quality. The quality control functions of the toolkit check to ensure all animals listed as parents have their own line entries, check to ensure all parents have the appropriate sex listed, and check to ensure no animals are listed as both a sire and a dam.

Further quality control measures may be added later, such as checking to ensure parents' birthdates precede their children's.

Genetic Value Analysis Reports

The Genetic Value Analysis is a ranking scheme developed at ONPRC to indicate the relative breeding value of animals in the colony. The scheme uses the mean kinship for each animal to indicate how inter-related it is with the rest of the current breeding colony members. Genome uniqueness is used to provide an indication of whether or not an animal is likely to possess alleles at risk of being lost from the colony. Under the scheme, animals with low mean kinship or high genome uniqueness are ranked more highly.

Breeding Group Formation

One of the goals in breeding group formation is to avoid the potential for mating of closely related animals. Since behavioral concerns and housing constraints will also be taken into account in the group formation process, it is our goal to provide the largest number of animals possible from a list of candidates that can be housed together without risk of consanguineous mating. To that end, this function uses information from the Genetic Value Analysis to search for the largest combinations of animals that can be produced from a list of candidates.

For more information see:
Vinson A, Raboin MJ. 2015. Designing breeding groups to maximize genetic diversity in a large captive rhesus macaque colony: a practical approach. *J Am Assoc Lab Anim* **In Press**.

Input File Format

The Input File Format tab is the starting point for all analyses. The file should be a delimited, regular text file with a header row specifying the columns. The tab provides information on the allowable columns in input files, and how the columns will be used in quality control of the data. Quality control of studbook data occurs automatically upon file upload.

Presently, the only columns required are those specifying the Ego ID, Sire ID, Dam ID, and Sex. The remaining columns listed are optional, but will be used if they are present in the uploaded file. The table of the tab describes how these optional columns will be used. Additionally, the panel on the left of this tab provides options that can be used during the upload and QC process, such as specifying the field separator used in the uploaded file.

During quality control, a flag is added for the current, living population. This flag is generated based on the information columns provided and is fairly specific to how the breeding population is defined at ONPRC. Two of the options for specifying the population of interest can be toggled through this panel, however. Normally, the breeding colony is restricted to Indian-origin, SPF 4 animals. These two restrictions can be turned off by setting the options on this panel.

Additionally, the population of interest can be specified directly in either the input file, or entered on the Pedigree Browser tab.

Pedigree Browser

The Pedigree Browser tab allows the user to view the input data, specify a population to examine, and output the cleaned studbook or trimmed pedigree.

Upon uploading a studbook file, the data goes through the quality control process described on the Input File Format tab. The cleaned version is displayed on this tab. By default, the entire uploaded studbook will be available for viewing on this tab. The first 10 rows will automatically be shown, but this range can be adjusted using the input boxes at the top. The default setting of showing only the first 10 rows is due to the size of the full ONPRC studbook: loading all 32,000 animals can cause the application to be slow.

The tab also contains functionality for trimming the pedigree. By checking the provided box, the

studbook that was uploaded will be trimmed down to just the ancestors of the currently specified population. This will remove any lineages that haven't contributed to the focal group.

As stated above, this tab will also allow the user to directly specify a population to examine. In the top half of the tab, there is an input box to specify the focal set of animals. The population flag can be reset by adding the desired animal IDs to the box. Once the population flag has been set to the desired group of animals, all further analyses will be relative to this group.

Genetic Value Analysis

The Genetic Value Analysis tab provides all of the options needed for producing a genetic value analysis report. The specifics of generating the genetic value analysis report are described on the README tab.

When the analysis is begun, it will generate a genetic value analysis for the currently-specified population in the pedigree. If no population has been specified, the entire pedigree will go into the analysis. This can be problematic, as the function for calculating the pairwise kinship matrix cannot handle large pedigrees. The kinship calculation is known to handle pedigree files containing up to 6000 individuals. It will not, however, handle the whole ONPRC rhesus studbook (~24,000 animals). The exact maximum pedigree size is not currently known and will need to be tested. Due to these problems, the input studbook will automatically be trimmed to the ancestors of the currently-specified population before the genetic value analysis is begun.

The genome uniqueness threshold input box allows the user to specify what constitutes a 'unique' allele in the gene-drop simulation. The algorithm description later in this document provides a more in-depth explanation of how the genome uniqueness calculation uses this information. By default, the gene-drop simulation underlying the genome uniqueness calculation considers an individual as unique if no other members of the current population have inherited the same allele during an iteration of the gene-drop. This can be adjusted using the drop-down box to allow up to four other animals to have inherited the allele and still consider it unique.

After the report has been generated, it can be subset to view a specific group of the animals using the text input box. Both the currently-viewed subset and the full report can be exported to a file from here.

Summary Statistics

The Summary Statistics tab may, at some point, be merged with the ORIP reporting tab. Currently this tab provides some descriptions of the population being examined. This tab may be expanded on in the future to contain additional information relevant for management.

After the genetic value analysis has been run, this tab will be populated with a number of statistics, tables and histograms. The tab will report the number of known founders, the estimate of founder equivalents, and the estimate of founder genome equivalents. The tab will generate a table providing the mean and quartiles for the individual mean kinship and genome uniqueness values. Lastly, the tab will also display histograms of the individual mean kinship values, and their corresponding z-scores.

Breeding Group Formation

The last major function of the R-package is to aid in generating breeding groups that avoid inter-animal relatedness. This tab allows you to build a number of breeding groups from a specified list of candidate animals. It also has an option to build a group by adding animals from a list of candidates to a currently-existing group.

In the top half of the tab, there are entry boxes and menus to adjust the options of the analysis. By default, the analysis will ignore relatedness between animals that is more distant than the second cousin level, pairwise relatedness involving an animal under 1 year of age, and relatedness between all females. All of these options can be adjusted before the analysis is run, however.

If the desire is to add animals to an existing group, the IDs of the candidate animals can be entered into the first text box (just as they would be if a new group were being generated). The IDs of the current group members can be added into the second text box. It should be noted that it will cause an error if the provided candidate animals or current group member IDs are not part of the population for which a genetic value analysis was run. The kinship matrix produced for this analysis provides the pairwise kinship values used by the group formation functions.

After the simulation is done, the first group will be displayed automatically. The group being displayed can be changed with the drop-down menu. Whichever group is currently being displayed can then be downloaded with the export button.

ORIP Reporting

The ORIP Reporting tab will eventually contain information for reporting to the Office of Research Infrastructure Programs (ORIP). This tab may end up being merged with the Summary Statistics tab and contain a number of statistics, tables and histograms. Alternatively, this may contain a subset of information from the Summary Statistics tab presented as a formatted report that can be exported and submitted to ORIP. The exact information that needs to be submitted for ORIP recordkeeping is still under discussion.

README

This tab contains more in-depth descriptions of how the Genetic Value Analysis is created, and how breeding groups are formed by the program.

Algorithm: Breeding Group Formation

The group formation process is accomplished by using an algorithm for determining the maximal independent set (MIS). In graph theory, a maximal independent set is the largest set of vertices in a graph where no two share an edge. In breeding group formation, the vertices are animals, and the edges are the kinships that need to be considered. For a given group of animals and pairwise kinships, there are potentially many maximal independent sets, depending on which animals are included or excluded from the final group. In order to effectively sample the set of MISs, we use random selection of

animals and repeat the MIS generation numerous times. This allows us to sample a number of MISs and then choose the one that best fits our selection criteria. For our purposes, we want the largest group that can be formed from this set of animals, where none have concerning relatedness to each other.

The algorithm requires several pieces of information:

1. The candidate animals
2. A matrix of pairwise kinships between candidate animals
3. The number of groups desired from the list of candidate animals
4. The number of simulations to run.
    - This is equivalent to the number of random MISs to generate and compare.
5. Information on which inter-animal relationships (if any) should be ignored.

Data Pre-processing

Before the group formation algorithm begins generating MISs, the data is pre-processed to remove any animals and pairwise kinships that should not be considered.

Specifically:

1. The candidate animals provided are checked, and any that were designated as low-value by the genetic value analysis will be removed from further consideration.
    - This behavior can be toggled off to allow low-value animals in the formation process
2. The pairwise kinship data is filtered down to only the kinship between candidate animals.
3. If an age threshold has been set, kinships involving animals below the threshold will be filtered out.
    - This allows the algorithm to ignore young animals, as young animals typically go to whatever social group their dam does.
    - By default, we ignore animals under 1 year of age
4. Pairwise kinships below the specified level will be filtered out.
    - By default, we ignore relatedness more distant than 2nd cousin
5. Pairwise kinships between females will be filtered out
    - This allows females of the same matriline to be part of the same group like they would be in the wild.
    - This behavior can be toggled off to prevent relatedness between females.

Random Maximum Independent Set Generation

After any animals and relationships that should be ignored are removed from the dataset, the algorithm begins using the remaining animals and kinship information to generate potential groups.

The algorithm proceeds by the following steps:

1. For I iterations:
    a. Generate N empty sets, where N is the desired number of groups to be created.
    b. While there are candidate animals remaining:
        i. Pick an animal A randomly from the set of candidate animals
        ii. Choose a group G randomly from one of the N groups, and assign A to it

iii. Remove animal A from consideration for all N groups
iv. Remove all animals related to A from consideration from for group G

c. Score the groups that were generated
   - For our purposes, we calculate the average group size

d. If the score of the new groups is higher than groups that were previously generated, save the new groups.

2. Return the currently saved groups
   - This should be the best groups encountered in I iterations.

Algorithm: Genome Uniqueness

Genome uniqueness is calculated through the use of a gene-drop simulation to estimate how frequently an animal will possess founder alleles not present in other members of the focal population, or present in a specified number or fewer.

The gene-drop simulation used by the web application is a vectorized version and is shown in the figure below. In an un-vectorized version, if 5000 gene-drop simulations are desired for the estimation process, the population had to be iterated over 5000 times. Since each iteration of the gene-drop is independent, the process can be vectorized so that each element of a vector represents 1 iteration of the gene-drop simulation. In the vectorized version, the population is iterated over once, regardless of the number of simulations desired. This drastically reduces the amount of time necessary for the program to run.

Overview

The basic steps of the gene-drop are:

1. Each founder is assigned two unique alleles
2. For each subsequent generation:
   a. Assign genotypes to each member of the generation
      - For each animal, find the genotypes of the parents, and select one allele from each parent randomly.

Once every animal has been assigned a genotype by mendelian inheritance tally the number of unique alleles possessed by each member of the focal population. In the case of this algorithm, we do allow the ‘uniqueness’ threshold to be adjusted so that an allele can be considered unique if it is possessed by N or fewer other members of the focal population.

Vectorized Gene-Drop Details

The vectorized gene-drop simulation follows the same basic process described above. The difference is that instead of dropping one allele at a time, and repeating the simulation N times, the vectorized version drops N independent alleles one time.

In the vectorized version, each animal has a vector of paternally inherited alleles and a vector of maternally inherited alleles. For each offspring, a random combination of these alleles is produced and dropped down to the offspring by the process below and shown in the following figure:

1. To start the simulation, each founder is assigned two unique founding alleles. N-element vectors

are created of these alleles, where N is the desired number of simulations. In the example below, this founder was assigned the unique founder alleles 1 & 2 and 5 simulations were desired.

2. Each time alleles need to be dropped from parent to offspring, a unique transmission vector is created representing whether or not an allele was passed to that offspring. The vector is generated to contain a random combination of 0's and 1's. The animal's paternally inherited alleles are then multiplied by the transmission vector, while the maternally inherited alleles are multiplied by the compliment of the transmission vector.
3. To generate the final set of alleles received by the offspring, the maternal and paternal allele vectors are added together.
4. The result is a vector of alleles that this offspring has received from this parent.

Once allele vectors have been generated for every animal in the pedigree, the focal population can be subset out. Within this population of allele vectors, unique alleles can be determined:

For each position on the allele vectors (1:N) - Gather each animal's two alleles - If the number of other animals possessing that allele is equal to, or below the threshold, score the allele as unique (1) - Otherwise, score the allele as non-unique (0)

Once every position on each animal's two allele vector's has been scored, sum all of the scores for an animal and divide by the total number of alleles being considered (2 * number of simulations).