MoPedS: Mongoose Pedigree Scripts

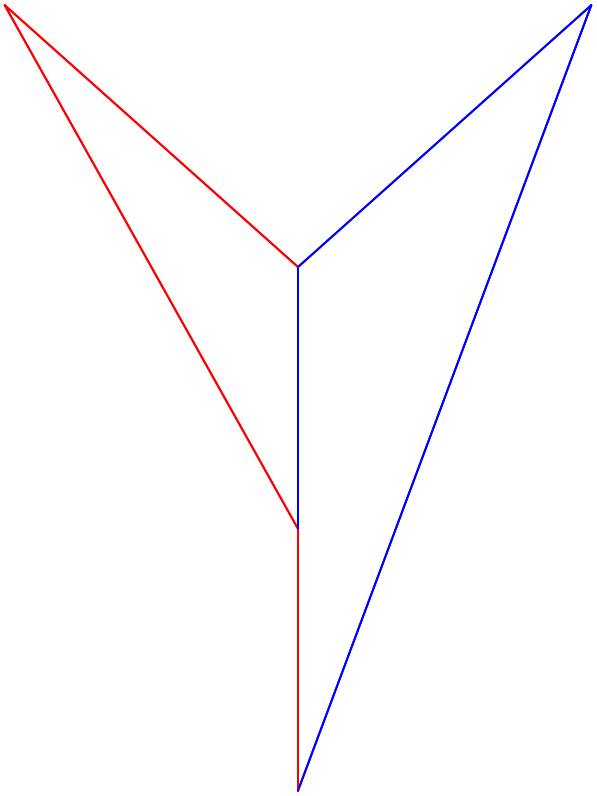
David Wells

20/07/2018

Abstract

The banded mongoose research project pedigree is constructed using the R package `MasterBayes` and the stand alone software COLONY2. MasterBayes takes advantage of phenotypic data to improve assignment acuracy while COLONY2 can identify imigrant siblings to prevent a downward bias in relatedness and inbreeding estimates. The mongoose pedigree scripts are designed to make updating the pedigree simple, including the use of priors so that small batches of newly genotyped individuals can be added to the pedigree.

Contents



Aim

Make an easy to use set of scripts to fit the banded mongoose pedigree. This should include checks of the data for errors and make it easy for the user to hunt down those errors so they can be corrected in the raw data. Also it needs to be easy to see how many mongooses have not been genotyped so effort can be made to genotype any missing individuals who are potential parents. The scripts also include a method to simulate a pedigree for power testing.

The aim is that it will more or less automate the process of checking for errors, correctly formatting the data, and fitting the pedigree in MasterBayes and Colony, checking that the MasterBayes model has converged, merging the two pedigrees, outputting summary statistics of the pedigree and a report of the whole process for posterity. It will also be possible to easily confirm the power of our parentage analysis through simulation.

Currently the scripts do not producea summary report of the pedigree. Although the process of updating the pedigree has (hopefully) been greatly simplified it still requires the user to have a reasonable understanding of R to confidently run, especially when assigning parentage through full siblings.

Set up

Download MoPedS from github.com/DAWells/MoPedS. The recomended folder structure is:

```
>MoPedS_id
  >Analysis
    MoPedS_1.R
    MoPedS_2.R
    MoPedS_3.R
```

```
    MoPedS_4.R
    MoPedS_5.R
    MoPedS_simulate_pedigree.R
    >colony
      colony_inputs_6-14.txt
      colony_inputs_17-23.txt
      colony_inputs_33-39.txt
  >data
    >rawdata
      BMRP_Genotype_data_month_year.csv
      error_rates.csv
      new_lhdata_month_year.xlsx
      previous_pedigree.csv
    >COLONY
```

All outputs will be in the folder "data"

MoPedS_1.R From MS access to R

Takes the excel output of the microsoft access file and prepares it for R. Including renaming the columns.

MoPedS_2.R Generate phenotypic data

Creates a phenotypic data file indicating the state of each individual at regular intervals. This interval is usually 30 days.

TP401-6 cannot be linked to their genotype data because in the genetic data file they are given sexes. This must be dealt with by editing the raw data. Also note that this issue may make problems with linking the pup's to their heterozygosity in other studies.

General lhdata problem checks like this individual is over x years old, really? our last record form them is ymd. Any individual bourn twice.

MoPedS_3.R Fit `MasterBayes` pedigree

Uses the `pdata` file generated in MoDedS_2.R in combination with the genotype data to fit a pedigree using `MasterBayes`. This is the predominant way that parentage is assigned in the BMRP pedigree. Note, that this completely excludes immigrants or individuals with unknown dates of birth.

MoPeds_4.R Prepares the input for `COLONY`

COLONY is very particular about what inputs it will accept, hopefully this will make it easier. Note also that spaces in file names can sometimes make COLONY crash.

Colony 2

Colony2 can be run from the command line and this is probably the easiest way to ensure parameters are set the same across pedigree iterations. One of the advantages of COLONY is that it can identify shared parents

even when the parent is not present. COLONY is a very slow program to run, when the pedigree was first assembled it was done using Jinliang Wang's "super computer".

The purpose of using COLONY is to supplement MasterBayes, the main advantage is that it can identify full siblings when their parents are not sampled. This is most useful when immigrants join the population as it allows us to more accurately identify inbreeding instead of assuming that all immigrants are completely unrelated to the entire population including other immigrants.

When full siblings are identified despite sharing an unsampled parent, this inferred "dummy" parent is assigned a number preceded by either a # or * depending on whether they are a dam or sire. To prevent the same dummy parent number being used when updating the pedigree MoPedS_5 changes these so different runs of colony can be distinguished. Side note there seems to be a bug in COLONY that means you often can't name files with spaces in " ".

MoPedS_5.R combining the pedigrees

This script combines the `MasterBayes` and COLONY pedigrees and adds the consensus pedigree onto the existing pedigree. `MasterBayes` joint parents are accepted first, that is when the parent pair has probability above the threshold. Then if no parents have been assigned to an individual, marginal parents are accepted. This means that confident `MasterBayes` sires and dams are accepted separately. If an individual's parents are both still unassigned the COLONY identified parent pair is added if it exceeds the confidence threshold. Any parents still unassigned are then filled in from the confident COLONY maternity/paternity files. The final step is to incorporate full sibling identity.

This last step is tricky as COLONY doesn't use confidence thresholds in the way we have. This means that a full sibling assignment can be confident but COLONY has assigned them separate dummy parents. Therefore, I have left it to the user to identify and assign parents based on full sibships by hand. The ruling is described in the script file and in Jenny Sanderson's `banded mongoose pedigree.docx`. I would recommend simulating some pedigrees with full sibling immigrants to see how well COLONY identifies them as full siblings, see MoPedS_simulate_pedigree.R.

Checks still to do for MoPedS_5:

Combines the two pedigrees and produces summary statistics for them, and write a full report of the pedigree construction.

produce pedigree figures including the force directed graph.

No offspring should be confidently assigned to individuals with very little genetic data.

no body should be there own parent. Nobody should be the paretn of somebody older than themself.

Everybody in lhdata should be present in the pedigree under ID.

Check that the pedigree is "fixed".

MoPedS_simulate_pedigree.R

Simulate a pedigree based on the observed phenotypic data and use it to simulate genotypes. These can then be fed into the MoPedS pipeline and the resulting pedigree can be compared with the known true pedigree. Simple comparisons between the `MasterBayes` generated and True pedigree are implemented in this script file but note that this may be different to the final consensus pedigree generated in MoPedS_5.R.

It is simple to test the normal assignment of pedigrees but testing the identification of full siblings is more difficult. The simulated pedigree must be hand edited *before* the genotypes are simulated. I have only done this a very small amount but COLONY was not great at identifying immigrant siblings. Therefore I would recommend running some simulations to get a feel for best to assign parents based on their full sibling status.