Pipeline to make a consensus tree from bootstrapped datasets

Author: Marilijn van Rumpt - marilijn@live.nl (2024)

Project description:

To get an overview of relations between different dog breeds, they can be placed in a phylogenetic tree. This pipeline makes a consensus tree of a SNP dataset.

Steps performed by this tool

(x = number of chosen iterations by user)

- Produce x SNP datasets by bootstrapping with resampling the SNPs of the samples (in .bim file)
- Kinship matrices are made from these bootstrapped SNP datasets.
- Of each of these matrices, a phylogenetic tree is made.
- The produced phylogenetic trees are combined to make a consensus tree in newick format.

Command line utility quality_control.sh

Input files should be in same folder as consensus.sh script. In this folder should also be the scripts folder.

Usage: bash consensus.sh [-f|o|t|i|g|h]

Command line tool to create a consensus tree from bootstrapped datasets

- Syntax options:
  - -f <prefix_filename> Specify prefix for .bed + .fam +. bim file. Obligatory.
  - -o <prefix_filename> Specify prefix for output files. Obligatory.
  - -t <method_tree_construction> Specify method of tree construction: phylip or biopython"
  - -i <number_of_iterations> Specify number of iterations
  - -g <outgroup_sample_ID> Specify Sample ID of outgroup sample"
  - -h Print the help overview
- Examples:
  - bash consensus.sh -f inputfile -t phylip -i 100 -g Coyote_1 -o newfilename"
  - bash consensus.sh -f inputfile -t biopython -i 50 -g 93754 -o newfilename"
- Dependencies needed:
  - python3
    - with packages: numpy and biopython (only if chosen tree construction method is biopython)
  - plink 1.9 (included in this tool)
  - Phylip's programs neighbor and consense (included in this tool)

Useful information and tips:

- Tree construction method phylip is generally faster than biopython
  - tip: to get an approximation of how long the script will take, you can do a test run with a low iteration number.
- More iterations means more trustworthy tree. For example 100 iterations can make a reliable tree.
- If the tree is made by using tree construction method phylip, the numbers shown in this file are bootstrap values, not branch lengths

- If the tree is made by using tree construction method biopython, both information about bootstrap values as branch lengths are in the newick file.
- bootstrap values = proportion of times (proportion of the number of chosen iterations) that each group appeared in the input trees
- Think about which samples and SNPs you want to use for the tree.
    - You might want to use only the SNPs in common between different datasets, if you want to merge these and make a consensus tree.
    - You generally only want to use autosomal SNPs, not chromosome X (non-pseudo autosomal part) and Y
    - Samples kan have different relatedness to each other, which can be checked by e.g. --make-king-table by plink

Operating system and used program versions

- consensus.sh
    - This command line utility operates in Linux and is tested in Ubuntu, so it is advised to use Ubuntu. If you have a Windows system, a Windows Subsystem for Linux (WSL) should be used. Installation information can be found on the ubuntu website.
- consensus_GB.sh
    - This utility operates in git bash.
- Used versions to test this tool:
    - Ubuntu 22.04.2 LTS
        - GLIBC 2.35-0
    - Git bash 2.40.1.windows.1
    - Python 3.10.12
        - numpy 1.25.2
        - biopython 1.81
    - Programs neighbor and consense from phylip 3.698

File descriptions:

- BootstrapSamples.py
    - Makes x (amount of iterations) new SNP lists by bootstrapping with resampling over the SNPs in the original .bim file. This resampled list with SNPs is put in a new file.
- MakeTree.py
    - Makes a phylogenetic tree newick file from a distance matrix, using biopython
- MakeConsensusTree.py
    - Makes a consensus phylogenetic tree newick file from multiple trees, using biopython
- ReformatDist.py
    - Reformats the distance matrix and makes temporary sample IDs, so the matrix can be used by the PHYLIP package
- UpdateSampleIDs.py
    - Changes the temporary sample IDs in the newick file to the original sample IDs
- Directory temp_files
    - In this directory the temporary files made by the tool are placed
    - There should be no files in this folder after running the tool. However, if an error occurred or if the run was stopped prematurely, there can be files in this folder. These should be removed to prevent build-up of files. This folder gets automatically emptied at the start of a new run of the tool.

Example files:

In the consensus_files folder there is a folder with example files:

- SNP dataset with multiple breeds (289 breeds, and wolves and coyotes)
    - This dataset contains only autosomal snps and the common SNPs between platforms from which the data comes.
    - Per breed 5 dogs are present, except for coyotes, of which 4 animals are present
    - The samples are selected based on lowest kinship, to minimalize the number of first or second degree relationships between samples.

- The consensus tree made out of this dataset (made using phylip)
- A file with sample IDs in tree and their corresponding original sample IDs
    - This was made, so the breeds in the tree are easier to recognize based on the sample name
- Breeds_tree.txt is a file with breeds in the SNP dataset

Plink settings

- --chr-set 38 is used in command (not --dog)
    - By doing this, the chromosome coding will remain the same (all in numbers from 1 to 42).
    - Dog has 38 autosomes
- --allow-no-sex
    - prevents errors because of missing sex in .fam or .ped file
- --distance triangle 1-ibs
    - for making a distance matrix with lower-triangular format
    - 1-ibs is used to express distances as genomic proportions (1 minus the identity-by-state value)
- --distance square 1-ibs
    - for making a distance matrix with square symmetric format
    - 1-ibs is used to express distances as genomic proportions (1 minus the identity-by-state value)

Output (file) descriptions

- Log file: contains output of performed checks and the plink log's
- file_consensus_tree.newick: contains the consensus tree in newick format
    - this tree can be visualized by programs such as ITOL (online tool), dendroscope, figtree etc.
    - If the tree is made by using tree construction method phylip, the numbers shown in this file are bootstrap values, not branch lengths
    - If the tree is made by using tree construction method biopython, both information about bootstrap values as branch lengths are in the newick file.
    - bootstrap values = proportion of times (proportion of the number of chosen iterations) that each group appeared in the input trees

Steps performed by consensus tool:

x = number of chosen iterations

Steps performed by the consensus command line utility, if chosen tree construction method is biopython:

1. BootstrapSamples.py to make bootstrapped datasets
    - Makes x new list files with random chosen SNPs from .bim file, by bootstrapping with resampling
2. Plink using --distance triangle 1-ibs and --extract listfile
    - To make x distance matrices in lower triangle format
    - The SNP lists made in step 1 are used to make a distance matrix only based on the SNPs in the list file. By doing that, x unique distance matrices are made, by using a different SNP set every time.
    - Executing x times:
        - plink --bim inputfile.bim --fam inputfile.fam --bed inputfile.bed --extract snp_list --distance triangle 1-ibs --allow-no-sex --chr-set 38 --out new_file
3. MakeTree.py to make phylogenetic trees from the distance matrices, using biopython
    - To make x phylogenetic rooted trees by using the neighborjoin method
4. MakeConsensusTree.py to make a consensus tree from the x phylogenetic trees, using biopython
    - The consensus tree is made by using the majority rule

Steps performed by the consensus command line utility, if chosen tree construction method is phylip:

1. BootstrapSamples.py to make bootstrapped datasets
    - Makes x new list files with random chosen SNPs from .bim file, by bootstrapping with resampling
2. Plink using --distance square 1-ibs and --extract listfile
    - To make x distance matrices in square symmetrical format

    - The SNP lists made in step 1 are used to make a distance matrix only based on the SNPs in the list file. By doing that, x unique distance matrices are made, by using a different SNP set every time.
    - Executing x times:
        - plink --bim inputfile.bim --fam inputfile.fam --bed inputfile.bed --extract snp_list --distance square 1-ibs --allow-no-sex --chr-set 38 --out new_file
3. ReformatDist.py to reformat the distance matrix to phylip format
    - The distance matrices made in step 2, need to be adjusted, so they can be used as input in the phylip program
    - The phylip format does not allow for sample IDs longer than 10 characters and wants a specific format for the IDs, so the sample ids are recoded to a temporary id. This is written to a file, so they can be reversed later.
4. neighbor program from phylip to make phylogenetic trees from the x reformatted distance matrices
    - Used settings are:
        - O - Outgroup is used
        - J - Input order of species is randomized
        - M - Multiple distance matrices are analyzed
5. consense program from phylip to make a consensus tree from the x trees produced in step 4
    - Used settings are:
        - R - Trees are treated as rooted
        - 3 - Tree does not get printed
        - 2 - Progress of run is not printed
6. UpdateSampleIDs to revert the temporary sample IDs in the consensus newick file back to the original IDs

Credits

This project is part of the Expertise Centre Genetics of Companion Animals (Faculty veterinary medicine, Utrecht University).

License