Population Structure and Phylogenetics

Graham Etherington[1]

[1]The Earlham Institute

Oct 25, 2022

1 Works for me Share

dx.doi.org/10.17504/protocols.io.bretm3en

Graham Etherington
The Earlham Institute

ABSTRACT

The European polecat (*Mustela putorius*) is a mammalian predator which breeds across much of Europe east to central Asia. In Great Britain, following years of persecution the European polecat has recently undergone a population increase due to legal protection and its range now overlaps that of feral domestic ferrets (*Mustela putorius furo*). During this range expansion, European polecats hybridised with feral domestic ferrets producing viable offspring. Here we carry out population-level whole genome sequencing on domestic ferrets, British European polecats, and European polecats from the European mainland and find high degrees of genome introgression in British polecats outside their previous stronghold, even in those individuals phenotyped as 'pure' polecats.

DOI

dx.doi.org/10.17504/protocols.io.bretm3en

EXTERNAL LINK

https://doi.org/10.1093/jhered/esac038

PROTOCOL CITATION

Graham Etherington 2022. Population Structure and Phylogenetics. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bretm3en

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Etherington GJ, Ciezarek A, Shaw R, Michaux J, Croose E, Haerty W, Palma FD, Extensive genome introgression between domestic ferret and European polecat during population recovery in Great Britain. Journal of Heredity 113(5). doi: 10.1093/jhered/esac038

CREATED

Jan 13, 2021

LAST MODIFIED

Oct 25, 2022

PROTOCOL INTEGER ID

46259

Structure analysis

1 RAxML phylogeny

1.1 Run RAxML in two stages. The first is the 'check' and 'parse' options that check the alignment and creates an *raxml.rba binary alignment file. It also estimates the resources needed to run RAxML efficiently (cores and memory)

```
ALIGNMENT=$1 #An alignment in fasta format

source RAxML-NG-0.9.0
raxml-ng-mpi --check --msa $ALIGNMENT --model GTR+G --
prefix T1
raxml-ng-mpi --parse --msa $ALIGNMENT --model GTR+G --
prefix T2
```

1.2 Using the resources advised by the previous step, run final step to produce the phylogeny.

```
source RAxML-NG-0.9.0
raxml-ng-mpi --all --msa T2.raxml.rba --outgroup weasel
--prefix raxml_all --threads 1 --seed 2 --bs-metric
fbp,tbe
```

Citation: Graham Etherington Population Structure and Phylogenetics https://dx.doi.org/10.17504/protocols.io.bretm3en

2 Treemix phylogeny

2.1 The first step is to create a 'clust' file. The clust file contains three columns, where the first and second column both indicate the name of the individual and the third column indicates the taxon name. This can be achieved by using a combination of bcftools and awk.

```
source bcftools-1.9 bcftools query -l$FILE.vcf.gz |
awk'{split($1,pop,"."); print $1"\t"$1"\t"pop[2]}'>
mustelids.clust
```

2.2 Next, we need as input a VCF file that has been pruned for linkage-disequilibrium (LD).
The input for this is the 'all_samples_genotyped_snps.vcf' file produced here:
https://www.protocols.io/view/gatk-nuclear-variant-discovery-and-consensus-assem-bqzgmx3w?step=2.2

We filtered for an LD of 0.8 over windows of 5Kb

```
source bcftools-1.9
bcftools +prune -l 0.8 -w 5000
all_samples_genotyped_snps.vcf -Oz -o
all_samples_genotyped_snps_ld_0.8.vcf.gz
```

2.3 Remove any missing data

```
source vcftools-0.1.13
 vcftools --gzvcf
all_samples_genotyped_snps_ld_0.8.vcf.gz --max-missing 1
--recode --stdout | gzip > all_samples_no_missing.vcf.gz
```

2.4 Convert the VCF file into TreeMix format
The files required can be found here:
https://bitbucket.org/nygcresearch/treemix/downloads/plink2treemix.py
https://github.com/speciationgenomics/scripts/blob/master/vcf2treemix.sh

```
source plink-1.90
source vcftools-0.1.13
./vcf2treemix.sh all_samples_no_missing.vcf.gz
mustelids.clust
```

2.5 Then run TreeMix with 1 to 5 migration edges (-m parameter)

```
source treemix-1.13
for i in {1..5}
 do
  treemix -i all_samples_no_missing.vcf.gz -m $i -o
all_samples_$i -root weasel -bootstrap -k 500 >
treemix_${i}_log &
 done
```

3

To create ADMIXTURE plots for our data, we used the following pipeline:
BCFTOOLS prune for linkage, reformat the into Plink format, run Structure over a sensible range of K, chooseK and plot with Structure

3.1 Take the linkage pruned dataset from 'Phylogenetics, Step 3.2, and handle missing genotypes

```
source plink-1.90
plink --bfile dom_ep_snps_plink_stucture --geno 0.999 --
make-bed --out dom_ep_snps_plink_stucture_pruned
```

3.2 Then run on k=2-5

```
K=$1
source admixture-1.3.0
admixture --cv -j6 dom_ep_snps_plink_stucture_pruned.bed
$K
```

To identify the best value of K, I used the --cv for bootstrap and CV estimates

3.3 grep out the CV errors to find the best (highest) value of K

Citation: Graham Etherington Population Structure and Phylogenetics https://dx.doi.org/10.17504/protocols.io.bretm3en

```
grep "CV" *.out | awk '{print $3,$4}' | cut -c 2-4,6,7-
20 > ADMIXTURE.cv.error
```

3.4 Plot all of the Admixture plots together.

```
Rscript plotADMIXTURE.R -p
dom_ep_snps_plink_stucture_pruned -i k5_sample.map -k 5
-l euro,welsh,english,hybrid,domestic
```

Citation: Graham Etherington Population Structure and Phylogenetics https://dx.doi.org/10.17504/protocols.io.bretm3en