```r
# Supplementary material for:

# Combining contemporary and ancient DNA in population genetic
#   and phylogeographic studies

# by Miguel Navascués, Frantz Depaulis and Brent C. Emerson

# Molecular Ecology Resources, 2010



# SCRIPT FOR ABC    by Miguel Navascués
# ==============


# INTRODUCTION

# This script performs the three analyses presented in figure YY of the
# article. In these analyses the likelihood of the model given the data
# is approximated by a distance metric between the target data and some
# simulated data. Further details in main article and below. The objective
# of the analysis is to estimate the parameter theta=2*N*mu (where N is the
# effective population size and mu the mutation rate) of a constant
# size demographic model and a infinite site mutation model.


# HOW TO USE:

# 1) download external required programs and scripts and place them in the
#    working directory
# 2) edit the line of the script that sets R working directory [setwd()]
# 3) copy text and paste in R command prompt


# EXTERNAL REQUIREMENTS

# Coalescent simulator: COMPASS by Jakobsson_2009
#   http://www.egs.uu.se/evbiol/Research/JakobssonLab/compass.html

# RR Hudson's R script to read ms output (same format as COMPASS)
#   http://home.uchicago.edu/~rhudson1/source/mksamples.html

# MA Beaumont's R script to perform ABC rejection and regression steps
#   http://www.rubic.rdg.ac.uk/~mab/


# Sets working directory to the directory containg the executable
# for the coalescent simulator COMPASS, and the external R scripts
setwd("C:/directory/subdirectory/etcetera/REMEMBER TO EDIT THIS")

# Loads external R scripts
source("readms.output.R") # RR Hudson's R script
source("make_pd2.r")      # MA Beaumont's R script

# Creates graphic output file
pdf(file="FigureABC.pdf",width=6.28,height=8)
```

```r
par(mfrow=c(2,1),mar=c(4,4,0.3,0.05), oma=c(0.15,0.15,0.15,0.15))



# CHARACTERISTICS OF TARGET DATA

# Total number of samples
tot_sample_size <- 10

# Contemporary DNA number of samples
mDNA_sample_size <- 5

# Ancient DNA number of samples
aDNA_sample_size <- 5

# Age of ancient DNA samples (measured in N generations,
# where N is the effective population size)
aDNA_age <- 0.2

# Summary satistic (number of segregating sites)
target_seg_sites <- 20

##############################################################################

# ANALYSIS NUMBER 1
# ~~~~~~~~~~~~~~~~~

# Approximation of the likelihood as the proportion of simulations
# with exactly the same value for the summary statistic (number of
# segregating sites).

# Likelihood of 40 different values of theta will be evaluated
# the values are taken in a logarithm scale in the interval [1-100)
theta <- 10^seq( from=0, to=1.45, by=0.05 )

# A vector is created to store the likelihood values estimated for
# each value of theta considered
likelihood <- array( data=NA, length(theta) )

# Sets the number of simulation to perform for each value of theta evaluated
number_of_simulations <- 1000

# In this loop, for each value of theta the specified number of simulations
# is run with the specified characteristics of the target data using COMPASS.
# The proportion of simulations with the same number of segregating sites
# than the target data is stred in the vector 'likelihood'
for ( i in 1:length(theta) ){
  # Message to report progress of simulations
  cat(paste("ANALYSIS NUMBER 1: simulation",i,"\n"))

  # Generation of the string containing the comand line to be passed to the
  # operative system for the execution of COMPASS
  compass_command <- paste("compass ",  tot_sample_size,
                                        number_of_simulations,
                                        " -t ", theta[i],
                                        " -h 0.0 ", mDNA_sample_size,
```

```
                                        " -h ", aDNA_age, aDNA_sample_size,
                                        " > sim.txt")

  # Runs COMPASS (automatically detects operative system and runs COMPASS
  # with the appropriate function)
  if(.Platform$OS.type == "windows") shell( compass_command )
  if(.Platform$OS.type == "unix") {
    compass_command <- paste( "./", compass_command, sep="" )
    system( compass_command )
  }

  # Reads COMPASS output file using Hudson function 'read.ms.output'
  cat(paste("   reading COMPASS output file: progress... "))
  sim_results <- read.ms.output(file="sim.txt")

  # Counts the number of simulations with the same number of segegating
  # sites than the target data and calculated the proportion
  likelihood[i] <- length( which(sim_results$segsites==target_seg_sites) )
  likelihood[i] <- likelihood[i]/number_of_simulations
}

# Plots the estimated likelihoods in function of the theta
# at logarithmic scale
log_theta <- expression(log*" "*theta1)
plot( log10(theta), likelihood,
                    type="h",
                    ylab="likelikelihood",
                    lwd=3,
                    cex.lab=1.4,
                    xlim=c(0,1.5),
                    xlab=log_theta)
text(1.5, max(likelihood)[1], "a", cex=1.5)

# Obtains maximum likelihood point estimate of theta
round(theta[which(likelihood==max(likelihood))])

# Reports the proportion of simulations rejected
cat( paste( sum(likelihood)*number_of_simulations,
            " simulations kept of ",number_of_simulations*length(theta),"\n"))

##############################################################################

# ANALYSIS NUMBER 2
# ~~~~~~~~~~~~~~~~~

# ABC analysis using the rejection method as in Pritchard_1999. The number
# of segregating sites is the only summary statistic used as example but
# more satistics should be used in a real analysis.

# Sets the number of simulations to perform in total
number_of_simulations <- 30000

# The value of the parameter theta for each simulation is taken from
# a prior distribution. In this example, a uniform prior on the logarithm
# of theta is chosen.
theta <- 10^runif(number_of_simulations,min=-0.1,max=1.6)
# Note: additional parameters would be also sampled from priors
```

```
# for more complex models

# Stores all parameters into a matrix
parameters <- cbind(theta) # for additional parameters: cbind(theta,p1,p2)

# Writes the values of theta into a file that will be used by COMPASS
# to make simulations from those values (function 'tbs', see manual of
# COMPASS for further details)
write( t(parameters), file="parameters.txt", ncol=dim(parameters)[2])


# Generation of the string containing the comand line to be passed to the
# operative system for the execution of COMPASS
compass_command <- paste("compass ",  tot_sample_size,
                                      number_of_simulations,
                                      " -t tbs ",
                                      " -h 0.0 ", mDNA_sample_size,
                                      " -h ", aDNA_age, aDNA_sample_size,
                                      " < parameters.txt > sim.txt")

# Runs COMPASS
if(.Platform$OS.type == "windows") shell( compass_command )
if(.Platform$OS.type == "unix") {
  compass_command <- paste( "./", compass_command, sep="" )
  system( compass_command )
}

# Reads COMPASS output file
sim_results <- read.ms.output(file="sim.txt")

# IMPORTANT NOTE: The current example, which uses only the number of
# segregating sites, does not perform any calculation at this point.
# However, for a real analysis additional summary statistics would be
# calculated for each simulated data set. Some R packages for the
# analysis of population genetics data (such as pegas, Paradis_2009)
# might be useful for this.

# Stores the number of segregating sites of the simulations in a matrix
sim_seg_sites <- data.matrix(sim_results$segsites)

# Performs the rejection step using M Beaumont function.
# Note that the proportion of simulations accepted (50%) is anormally
# high compared to the values normally used. This has been chosen for a
# better illustration of the improvement obtained by using the regression
# algorithm.
ABC_rejec_results <- makepd4( target_seg_sites,  # target summary statistics
                              log10(parameters), # parameter values (sims)
                              sim_seg_sites,     # summary statistics (sims)
                              tol=0.5,  # proportion of sims kept
                              rej=T)    # perform only rejection algorithm

# Obtains point estimate of theta as median of posterior distribution
10^median(ABC_rejec_results$x)

# Reports the proportion of simulations rejected
cat( paste( number_of_simulations*0.5,
            " simulations kept of ", number_of_simulations, "\n"))
```

```r
##############################################################################

# ANALYSIS NUMBER 3
# ~~~~~~~~~~~~~~~~~

# ABC analysis using the regression method proposed by Beaumont_2002.
# All the initial steps are the same as the ANALYSYS NUMBER 2 so the same
# simulations will be used here. The methods are different after the
# rejection step and this is starting point for this final analysis.

# Performs the rejection and regression steps using M Beaumont function
ABC_regres_results <- makepd4( target_seg_sites,  # target summary statistics
                               log10(parameters), # parameter values (sims)
                               sim_seg_sites,     # summary statistics (sims)
                               tol=0.5,  # proportion of sims kept
                               rej=F)    # performs regression algorithm

# Plots the estimated posterior (calculated from the rejection algorithm or
# with the regression algorithm) and prior probability functions for the
# paramater theta at logarithmic scale
prior                <- density(log10(parameters),from=0,to=1.5)
posterior_rejection  <- density(ABC_rejec_results$x,from=0,to=1.5)
posterior_regression <- density(ABC_regres_results$x,from=0,to=1.5)
# Adds to the plot the estimated posterior from the regression algorithm
plot(posterior_regression$x, posterior_regression$y,
                             ylab="probability density",
                             xlab=log_theta,
                             cex.lab=1.4,
                             type="l", xlim=c(0,1.5),lwd=1.5)
# Adds to the plot the estimated posterior from the rejection algorithm
lines(posterior_rejection$x, posterior_rejection$y,
                             type="l",
                             lwd=1.5,
                             lty="dashed")
# Adds to the plot the prior
lines(prior$x, prior$y, lty="dotted", lwd=1.5)
legend(-0.1,max(posterior_regression$y)[1],
       c("prior","posterior (rejection)","posterior (regression)"),
       lwd=1.5,bty="n",
       lty=c("dotted","dashed","solid"))
text(1.5, max(posterior_regression$y)[1], "b", cex=1.5)


# Obtains point estimate of theta as median of posterior distribution
10^median(ABC_regres_results$x)

# Reports the proportion of simulations rejected
cat( paste( number_of_simulations*0.5,
            " simulations kept of ", number_of_simulations, "\n"))

##############################################################################

# Closes graphical device (i.e. output file)
dev.off(which = dev.cur())
```

```
# REFERENCES

# @article{Beaumont_2002,
#  title = {Approximate Bayesian computation in population genetics},
#  volume = {162},
#  url = {http://www.genetics.org/cgi/content/abstract/162/4/2025},
#  number = {4},
#  journal = {Genetics},
#  author = {Mark A. Beaumont and Wenyang Zhang and David J. Balding},
#  month = dec,
#  year = {2002},
#  pages = {2025–2035}
# }

# @article{Jakobsson_2009,
#  title = {{COMPASS:} a program for generating serial samples under
#           an infinite sites model},
#  volume = {25},
#  url = {http://dx.doi.org/10.1093/bioinformatics/btp534},
#  doi = {10.1093/bioinformatics/btp534},
#  number = {21},
#  journal = {Bioinformatics},
#  author = {Mattias Jakobsson},
#  month = sep,
#  year = {2009},
#  pages = {2845–2847}
# }

# @misc{Paradis_2009,
#  title = {pegas: Population and Evolutionary Genetics Analysis System},
#  url = {http://ape.mpl.ird.fr/pegas/pegas.html},
#  author = {E. Paradis},
#  year = {2009}
# }

# @article{Pritchard_1999,
#  title = {Population growth of human Y chromosomes:
#           a study of Y chromosome microsatellites},
#  volume = {16},
#  url = {http://mbe.oxfordjournals.org/cgi/content/abstract/16/12/1791},
#  number = {12},
#  journal = {Molecular Biology and Evolution},
#  author = {J. K. Pritchard and M. T. Seielstad
#            and A. {Perez-Lezaun} and M. W. Feldman},
#  month = dec,
#  year = {1999},
#  pages = {1791–1798}
# }
```