Generate a spatial genetic layer for species distribution modelling

This directory contains code and input data for generating a spatial genetic layer from kriged (interpolated) ancestry coefficient values across the seascape. This analysis uses methods of Chambers et al. 2023 and the R package `algatr`.

The example below uses DArTseq data for *Aipysurus laevis*. Input files were downloaded from the GitHub repository: sea-snake-dart. Environmental layers were downloaded from MARSPEC (via R). The northwest shelf shapefile was obtained from Vinay Udyawer.

The northwest shelf is the extent of the spatial genetic layer we will generate.

Prepare sample list

First, we will prepare a sample list which includes individuals of *A. laevis* from the northwest shelf.

```r
library(tidyverse)

laevis_popmap <-
  read.csv("./genetic_layer/laevis/sample_list/ALA-popmap-coords.csv",
           sep = ",", header = TRUE)

laevis_nw <- laevis_popmap %>%
  # Remove AL401 and AL404 - dropped samples
  subset(id!="AL401") %>%
  subset(id!="AL404") %>%
  # Keep only samples west of GoC
  subset(pop!="North_QLD") %>%
  subset(pop!="Gulf_of_Carpentaria") %>%
  subset(pop!="New_Caledonia")

laevis_nw <- laevis_nw %>%
  # unite first 3 column values separated with "-"
  unite("id_clean", species:targetid, remove = FALSE, sep = "-") %>%
  # arrange columns as desired
  select("id_clean","pop","locality","longitude","latitude")
laevis_nw

# create keep list for vcf
laevis_vcf_keep <- laevis_nw %>%
  select("id_clean")
# remove header from keep list
colnames(laevis_vcf_keep) <- NULL

### Not run ###
# write.table(laevis_vcf_keep,
#             file = "./genetic_layer/laevis/sample_list/ALA-nw.txt",
#             sep = ",",
#             row.names = FALSE,
#             quote = FALSE)
```

The written output will be used to subset the VCF file.

Subset the VCF file

Next, we will subset the VCF file by individual using `vcftools`. In bash, we run:

```
vcftools --vcf ./genetic_layer/laevis/vcf_file/ALA-stringent.highQ.filtered.vcf --keep
./genetic_layer/laevis/sample-list/ALA-nw.txt --recode --stdout > ./genetic_layer/laevis/vcf_file/ALA-
stringent.highQ.filtered.nw.keep.vcf
```

Process subsetted DArTseq data

After subsetting the VCF file to only include individuals of Olive sea snakes from the northwest shelf, we will prepare the data for downstream `algatr` analyses.

```
# load package
library(vcfR)
library(wingen)
library(algatr)

# load vcf
laevis_vcf <-
  vcfR::read.vcfR("./genetic_layer/laevis/vcf_file/ALA-stringent.highQ.filtered.nw.keep.vcf",
                  verbose = TRUE)

# convert to dosage
laevis_dosage <-
  wingen::vcf_to_dosage(laevis_vcf)

# impute any missing data with median
laevis_dos_imp <-
  algatr::simple_impute(laevis_dosage, FUN = median)
```

Prepare sample list for `algatr` analyses

Analyses in `algatr` require a sample list that has samples arranged in the same way as in the VCF file. To make sure we have that, we run:

```
# obtain list of samples from VCF
laevis_samples <- as.data.frame(colnames(laevis_vcf@gt)) # get sample names from ala_vcf
laevis_samples <- as.data.frame(laevis_samples[-1,]) # remove "FORMAT" colname
colnames(laevis_samples) <- "sample" # rename "FORMAT" as "sample"
laevis_samples

laevis_nw_coords <- laevis_nw %>%
  # arrange `laevis_nw_coords` as in samples in vcf file
  arrange(id_clean, laevis_samples$sample) %>%
  # select only long and lat cols
  select("longitude","latitude")
colnames(laevis_nw_coords) <- c("x","y") # rename cols
```

```
# project
laevis_nw_proj <- sf::st_as_sf(laevis_nw_coords, coords = c("x", "y"), crs = "epsg:4326")
laevis_nw_proj <- sf::st_transform(laevis_nw_proj, crs = 4326)
```

With the genetic input data prepared, we will proceed with preparing spatial input data.

Spatial: Marine environmental datasets and analysis extent

These environmental datasets can be accessed through the MARSPEC website. We will download it programmatically using the `sdmpredictors` R package.

```
# load packages
library(sdmpredictors)
library(raster)
library(sp)
library(dismo)

# MARSPEC data sets
# list names of MARSPEC layers (annual data)
mspec_names <- list_layers("MARSPEC", monthly = FALSE)$name

# all annual layers
mspec_annual <- list_layers("MARSPEC", monthly = FALSE)$layer_code

# download the layers -- already saved in marspec dir
mspec_layers <- load_layers(mspec_annual, datadir = "./genetic_layer/marspec/")
```

The last line above will write the layers in the `marspec` directory for ease of access since the layers need to be downloaded only once. Next, we introduce the extent of our analyses - the northwest shelf.

```
# load shapefile
nw_shelf <- sf::st_read("./genetic_layer/nw_shapefile/NWShelf.shp", quiet = TRUE) %>% sf::st_transform(4326)

# limit environmental layers to NW shelf boundary
mspec_spatrast <- raster::crop(mspec_layers, nw_shelf)
mspec_spatrast <- terra::rast(mspec_spatrast) # global; convert RasterBrick to SpatRaster
mspec_shelf <- terra::mask(mspec_spatrast, mask = terra::vect(nw_shelf))

# preview with laevis points
plot(mspec_shelf[[14]]) # 14 = sea surface temp (annual mean)
points(laevis_nw_coords, pch = 19)

# environmental PCA
mspec_sh_pcs <- RStoolbox::rasterPCA(mspec_shelf, spca = TRUE)
mspec_sh_stack <- raster::stack(mspec_sh_pcs$map)
mspec_sh_stack
```

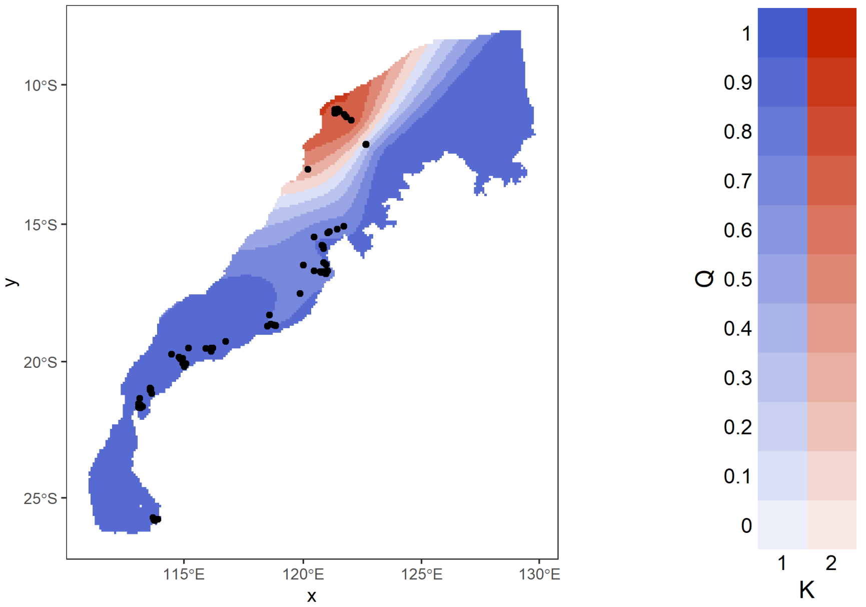
The last few lines of code above under `# environmetal PCA` summarises the variation across the environmental variables. We will use environmental PC1 in downstream analyses.

Interpolate ancestry coefficients across the seascape

With our genetic and spatial data prepared, we can now perform K-estimation and kriging to generate our spatial genetic layer.

```
# run TESS to estimate best K, use manual K selection
laevis_tess <-
  algatr::tess_ktest(gen = laevis_dos_imp,
                     coords = laevis_nw_proj,
                     Kvals = 1:7, # 7 pops
                     ploidy = 2,
                     K_selection = "manual") # choose K = 2

# get TESS object and best K from results (i.e., laevis_tess)
laevis_tessobj <- laevis_tess$tess3_obj
laevis_bestK <- laevis_tess[["K"]]

# get matrix of ancestry coefficients
laevis_qmat <-
  tess3r::qmatrix(tess3 = laevis_tessobj,
                  K = laevis_bestK)

# prepare grid for kriging

# grid
mspec_sh_krig <- (mspec_sh_stack[[1]]) # environmental PC1, nw shape
y_krig_sh_raster <- raster::projectRaster(mspec_sh_krig, crs = "epsg:3112")

# sample coordinates
x <- sf::st_as_sf(laevis_nw_coords, coords = c("x", "y"), crs = 4326)
x_proj <- sf::st_transform(x, crs = 3112) # EPSG:3112; GDA94 Geoscience Australia projection

# krig
z_krig_sh_admix <-
  algatr::tess_krig(qmat = laevis_qmat,
                    coords = x_proj,
                    grid = y_krig_sh_raster)

# map the krig
z_map_sh_admix <-
  algatr::tess_ggplot(z_krig_sh_admix,
                      plot_method = "maxQ",
                      plot_axes = TRUE,
                      coords = x_proj)

# plot
plot(z_map_sh_admix)
```

Here is a plot of `z_map_sh_admix`:



Interpolated ancestry coefficient values across the northwest shelf at K = 2.

We now have our spatial genetic layer for *A. laevis*. We also want to save our output as a `.csv` file so we can use it as input for species distribution modelling.

```
### write: save as csv ###
anc <- raster::raster(z_krig_sh_admix$K2) # take either K; K1 is inverse of K2
anc <- terra::rast(anc)
anc <- terra::project(anc, "+proj=longlat +datum=WGS84")
df_anc <- as.data.frame(anc, xy = TRUE)
write.csv(df_anc, file = paste0("./genetic_layer/output/laevis_K",laevis_bestK,".csv"))
```