EVOLUTION AND GENOMICS
Intensive and immersive training opportunities

2020 WORKSHOP ON POPULATION AND SPECIATION GENOMICS, CESKY KRUMLOV

The Multi-Species Coalescent (MSC) and its Application in Phylogenetics and Species Delimitation

L. Lacey Knowles

Dept. of Ecology and Evolutionary Biology
University of Michigan

Software: *Delineate*

Jeet Sukumaran
Dept. of Biology, Evolutionary Biology Program
San Diego State University

https://github.com/jeetsukumaran/delineate

Software: *Decrypt*

Arnaud Becheler
Dept. of Ecology and Evolutionary Biology
University of Michigan

https://becheler.github.io/pages/applications.html

Delimitation models that bring speciation to the multispecies coalescent

Inference of species boundaries (beyond the MSC)

Software: *Delineate* https://github.com/jeetsukumaran/delineate

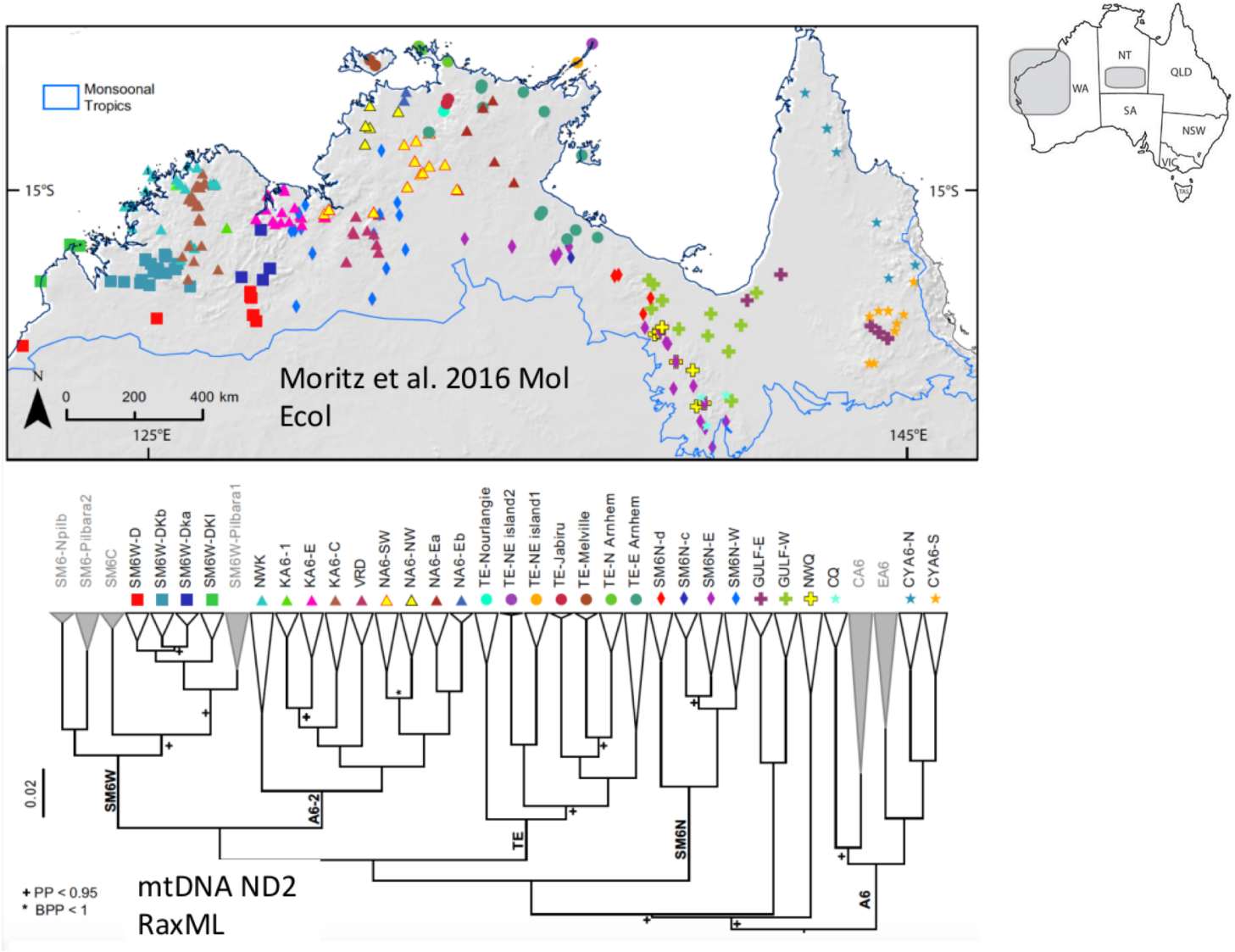
- Phylogenetic modeling approach that delineates species versus population lineages under a protracted speciation model

Software: *Decrypt* https://becheler.github.io/pages/applications.html

- Model of the geography of genetic divergence under a spatially explicit coalescent to evaluate competing hypotheses about cryptic diversity (inferred under the MSC)

Seemingly fractal nature of species diversity

With more geographic sampling, get more and more species inferred under the MSC

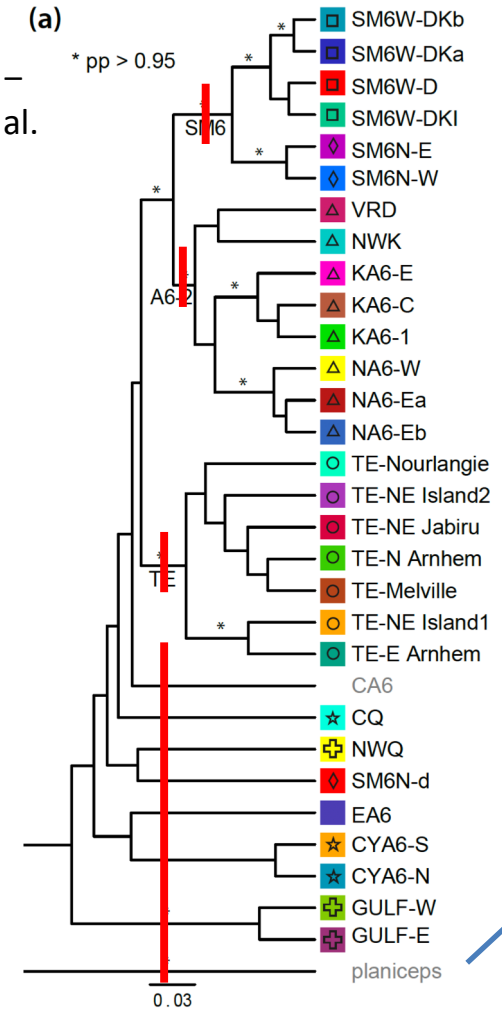


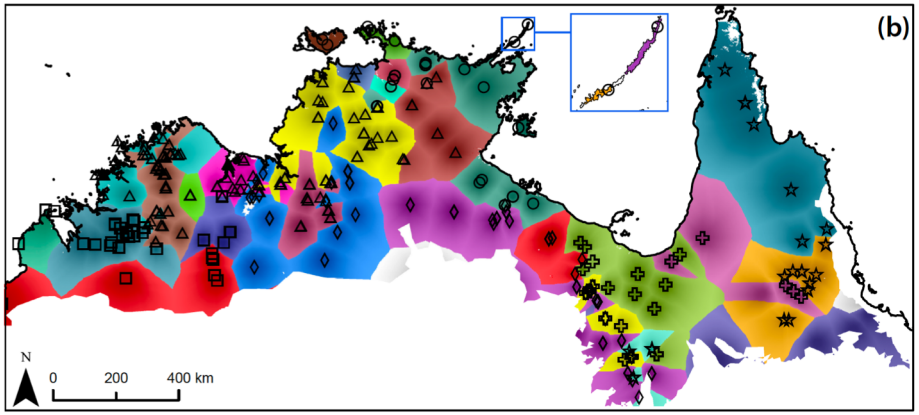
Seemingly fractal nature of species diversity

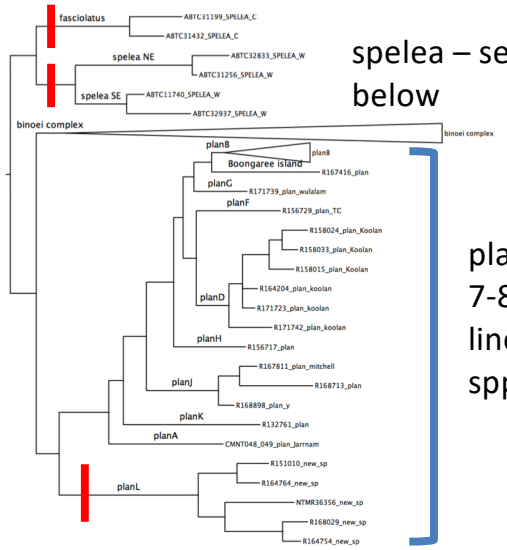
With more geographic sampling, get more and more species inferred under the MSC

Heteronotia binoei complex – species tree from Moritz et al. 2016 (7-8 loci) All lineages diagnosable using BPP.

146 samples
10 candidate species
30 lineages







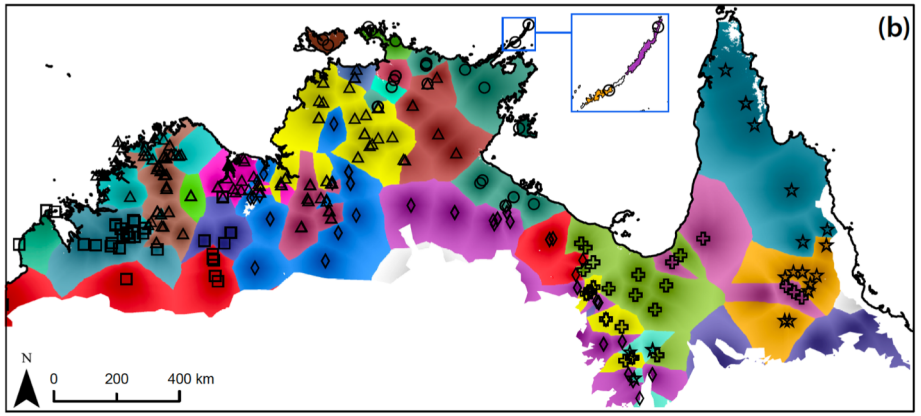
spelea – see below

planiceps – 7-8 lineages, 2 spp

Software: DECRYPT

Can the genetic structure detected under the MSC be explained by the intra-specific spatial structure?

That is, there is no justification for interpreting the detected structure as cryptic species.



DECRYPT analysis can be seen as a robustness analysis of the MSC for different realistic demographic histories and spatial sampling schemes.

Software: DECRYPT

Software components:

• demogenetic simulation core (i.e., demographic history and coalescent genealogies); C++ using the Quetzal library (see Becheler and Knowles (2020) Bioinformatics).

It generates a spatially explicit demographic history incorporating environmental heterogeneity (i.e., local carrying capacities, *k*, and migration rates, *m*, vary as a function of the underlying suitability of the environment), which includes random population persistence in unsuitable areas.

Data Requirements:

- Map of environmental heterogeneity File: australia.tiff
- Configuration file to set parameters of the spatial coalescent process File: spatial_process.ctl
- A configuration file for bpp File: bpp.ctl

Installation:

see instructions at https://becheler.github.io/pages/applications.html

(after installation)

From the terminal:

- Run spatial_process.ctl

```
cd path/to/sandbox chmod +x decrypt/spatial_process
mkdir output
./decrypt/spatial_process --config decrypt/example/spatial_process.ctl
```

If the program runs correctly, you should see in the terminal if the demographic history has been simulated, and then a progress bar that indicates how advanced the pseudo-observed data generation is:

```
--- Expanding demography
 --- Simulating coalescents
0% 10 20 30 40 50 60 70 80 90 100% |----|----|----|----|----|----|----|----|----|----|
**************************************************
```

Data Requirements:

- Map of environmental heterogeneity File: australia.tiff
- Configuration file to set parameters of the spatial coalescent process File: spatial_process.ctl
- A configuration file for bpp File: bpp.ctl

- Configuration file to set parameters of the spatial coalescent process File: spatial_process.ctl

```
# Geospatial file in tiff format
landscape=decrypt/example/australia_precipitation_6032.tif

# Number of sampling scheme simulations under 1 demographic history"
n_sim_gen=5

# Number of loci
n_loci=5

# Introduction point latitude
lat_0=-20.0

# Introduction point longitude
lon_0=125.0

# Number of gene copies at introduction point
N_0=1000

# Number of generations to simulate
duration=500

# Fixed sampling point latitude
lat_1=-20.0

# Fixed sampling point longitude
lon_1=125.0
```

- Configuration file continued:

```
# Number of gene copies to sample around center 1
n_sample_1=30

# Sampling radius around center 1
radius_sample_1=30.0

# Number of gene copies to sample in population 2
n_sample_2=30

# Sampling radius around center 2
radius_sample_2=30.0

# Population size under which random sampling is not considered
sampling_threshold=30

# Environmental threshold delimiting suitable (above threshold) and unsuitable (under threshol
suitability_threshold=26.4

# Carrying capacity in suitable areas
K_max=50

# Population persistance parameter in unsuitable areas
p=0.175

# Carrying capacity in unsuitable areas with probability p
K_min_a=1

# Carrying capacity in unsuitable areas with probability 1 - p
K_min_b=20
```

- Configuration file continued:

```
# Number of gene copies to sample around center 1
n_sample_1=30

# Sampling radius around center 1
radius_sample_1=30.0

# Number of gene copies to sample in population 2
n_sample_2=30

# Sampling radius around center 2
radius_sample_2=30.0

# Population size under which random sampling is not considered
sampling_threshold=30

# Environmental threshold delimiting suitable (above threshold) and unsuitable (under threshol
suitability_threshold=26.4

# Carrying capacity in suitable areas
K_max=50

# Population persistance parameter in unsuitable areas
p=0.175

# Carrying capacity in unsuitable areas with probability p
K_min_a=1

# Carrying capacity in unsuitable areas with probability 1 - p
K_min_b=20
```

(after installation)

From the terminal:

- Run spatial_process.ctl in decrypt

```
cd path/to/sandbox chmod +x decrypt/spatial_process
mkdir output
./decrypt/spatial_process --config decrypt/example/spatial_process.ctl
```

If the program runs correctly, you should see in the terminal if the demographic history has been simulated, and then a progress bar that indicates how advanced the pseudo-observed data generation is:

```
--- Expanding demography
 --- Simulating coalescents
0% 10 20 30 40 50 60 70 80 90 100% |----|----|----|----|----|----|----|----|----|----|
**************************************************
```

This program creates a bunch of files in the output directory that give access to various aspects of the demographic process. We will be able to visualize them later using the small R library decrypt/decrypt.R. Now we will focus on analyzing the simulated coalescents that have been stored in the output/test.db database.

(after installation of bpp; see https://github.com/bpp/bpp)

From the terminal:

- Preforming species delimitation under the MSC using BPP

If you copy the BPP executable into the sandbox directory, you can run the following command line:
python3 decrypt/decrypt.py -d output/test.db -l 100 -s 0.000001 -b bpp -c decrypt/example/bpp.ctl

The program will iterate through each gene genealogy simulated with spatial_process, evolve sequences along branches, and perform species delimitation on this pseudo-observed data.

When the BPP analysis is done, a dataframe data.txt is generated giving for each sampling scheme the probability to detect more than one species.

We can then use the R library to visualize the results.

Visualizing results in R studio:

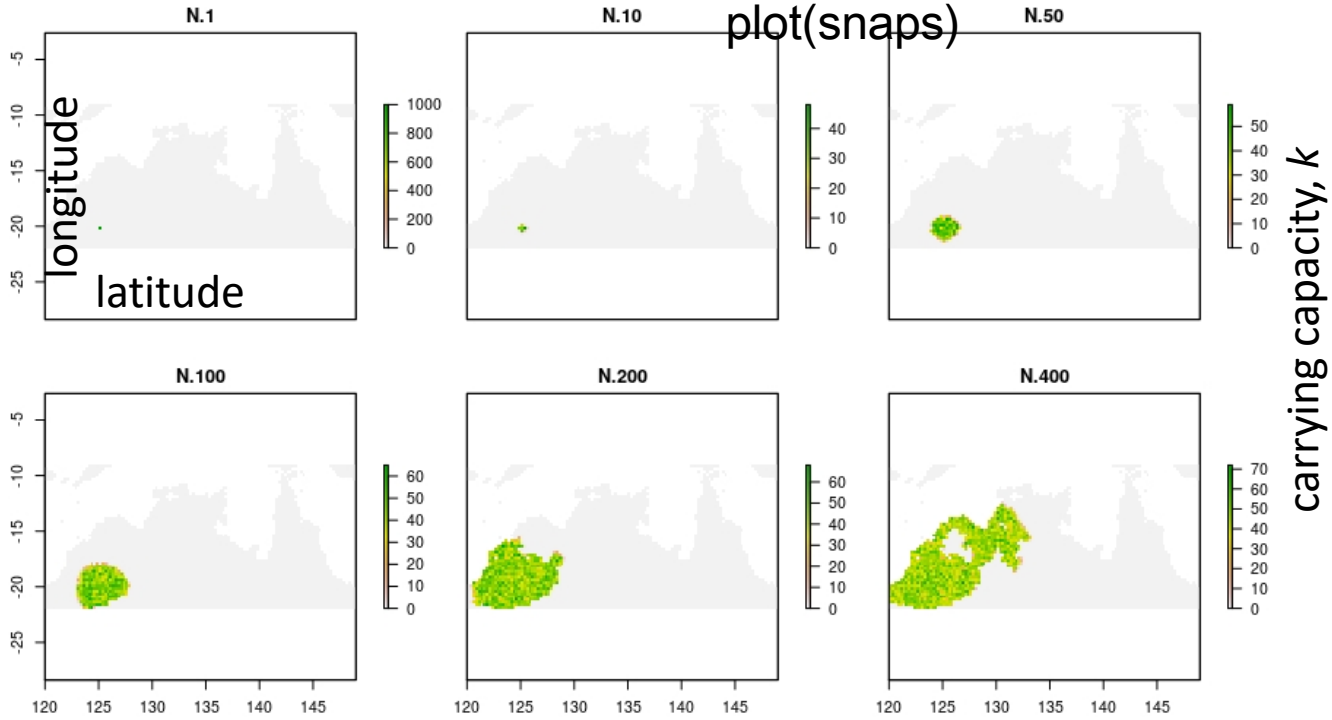
- Inspect the demographic history

  Snaps (snap shots of population carrying capacities at different time points of throughout the demographic simulation)

The parameter demography_out=output/N.tif in spatial_process.ctl creates a tif file to record the demographic history. It allows to inspect the effect of different parametrizations of the spatial process.

```
# Read the geotiff file created by spatial_process
history <- stack("output/N.tif")
# Subset history for different times and plot
snaps <- subset(history, list(1,10, 50, 100, 200,400))
plot(snaps)
```

N = # generations into simulation



Visualizing results in Movie:

- Inspect the demographic history

Sometimes it is easier to understand the process through an animation. We use here the command convert that is part of the ImageMagick package, which comes with many Linux distributions.

```
# Create a directory to store intermediary files
dir.create("movie") working_folder
<- paste0(getwd(),"/movie")
ordered_times <- 1:400
# Standardize the plots legends with an expected maximal N value in the dataset
# like the maximal carrying capacity
max_N_value <- 100
make_movie(history, ordered_times, max_N_value, working_folder)
```

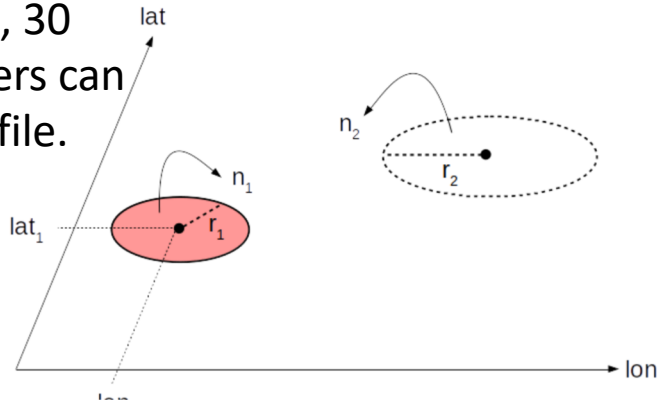
Visualizing results in R studio:

- Inspecting the Sampling Scheme:

In the spatial process configuration file, we limited the number of simulations to 5 sampling schemes, each one composed of:

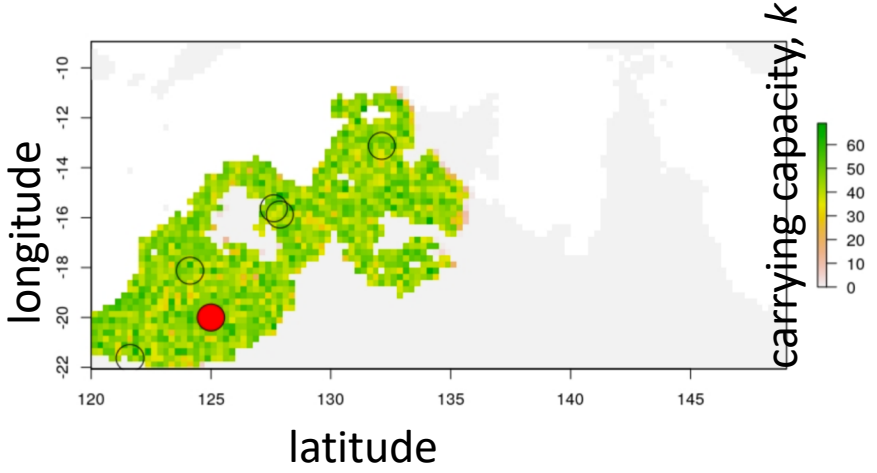
- 1 sampling cluster fixed on a given coordinate
- 1 sampling cluster that varies uniformly across the distribution area

Within a radius of 30km each of these coordinates, 30 individuals are sampled uniformly. These parameters can be change in the spatial_process.ctl configuration file.




```
data <- read.csv("data.txt")
mask <- history[[nlayers(history)]]
x0 <- data.frame("lon" = c(125), "lat" = c(-20))
plot_sampling_scheme(mask, x0=x0, r0=30000, x=data[,c('lon','lat')], r=30000,
proj4string=crs(mask))
```

The previous lines allow to plot the fixed sampling cluster, in red, and the 5 varying clusters with their respective radius, in black, on top of the spatial distribution of the population sizes at sampling time, in colors.

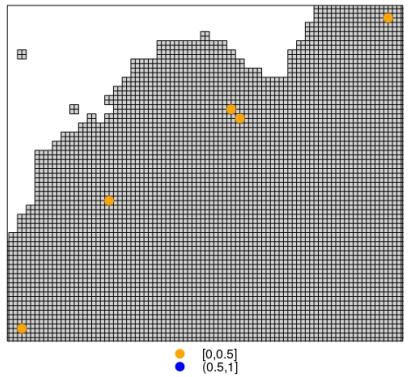


Visualizing results in R studio:

- Posterior Probability

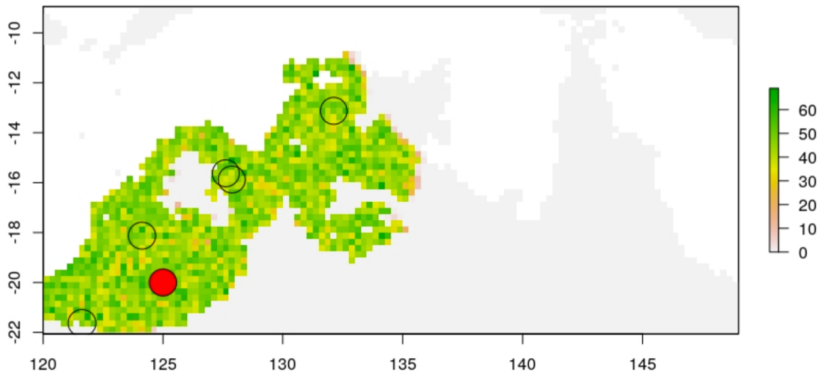
In short

To visualize the combined effects of departures from the MSC model hypothesis (i.e., rejection of a single species) and sampling scheme, you can either look at the raw posterior probabilities, or perform a spatial interpolation of this probability.




```
mask2 <- disaggregate(mask,fact=2)
raw_posterior_probability(data=data, mask=mask2, proj4string=crs(mask))
interpolate_posterior_probability(data=data, mask=mask2, x0=x0, proj4string=crs(mask))
```

Of course, the example that was developed here focus on a quite recent history, *i.e.* 400 generations, and then test only 5 alternative sampling points. This is computationally tractable for a demo, but it is not an ideal situation to perform a spatial interpolation, so we will show the related figure.

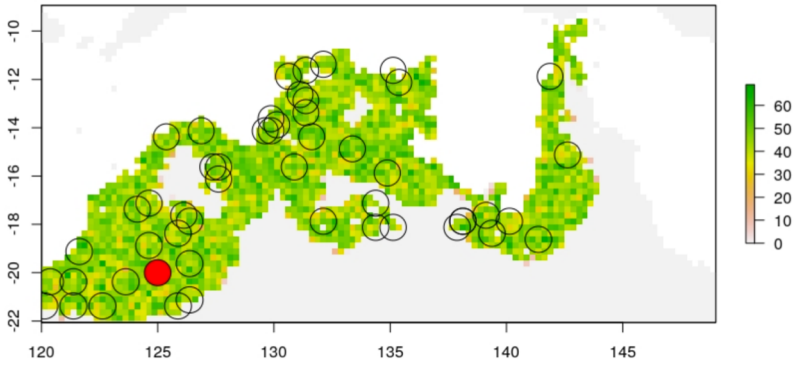


Visualizing results in R studio:

- Posterior Probability

Larger dataset

We provide in the decrypt/example directory two supplementary files giving the results of a more substantial analysis on longer times with more intensive sampling scheme:



You can run:

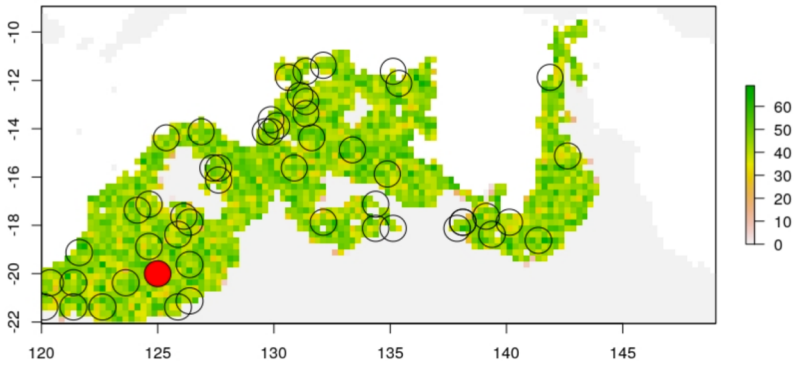
```
data <- read.csv("decrypt/example/data_extract.txt",header=TRUE)
mask <- raster("decrypt/example/last_N.tif")

x0 <- data.frame("lon" = c(125), "lat" = c(-20))
plot_sampling_scheme(mask, x0=x0, r0=30000, x=data[,c('lon','lat')], r=30000,
proj4string=crs(mask))
 mask2 <- disaggregate(mask,fact=2)
 raw_posterior_probability(data=data, mask=mask2, proj4string=crs(mask))
interpolate_posterior_probability(data=data, mask=mask2, x0=x0, proj4string=crs(mask))
```

Visualizing results in R studio:

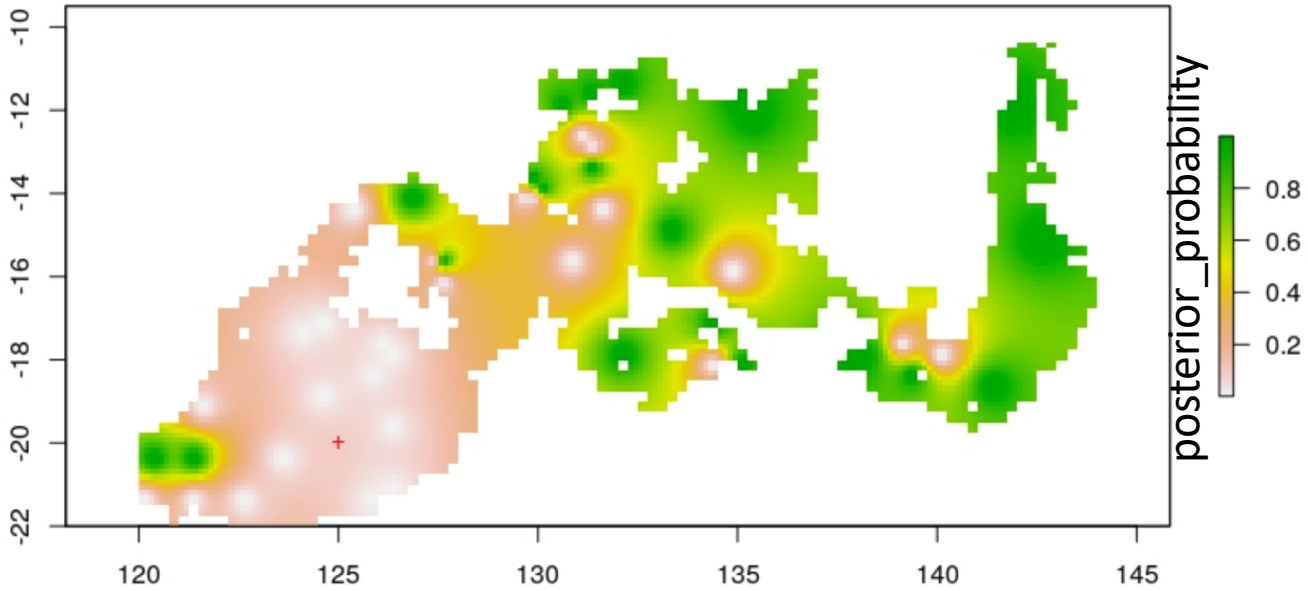
- Posterior Probability

Larger dataset
We provide in the decrypt/example directory two supplementary files giving the results of a more substantial analysis on longer times with more intensive sampling scheme:



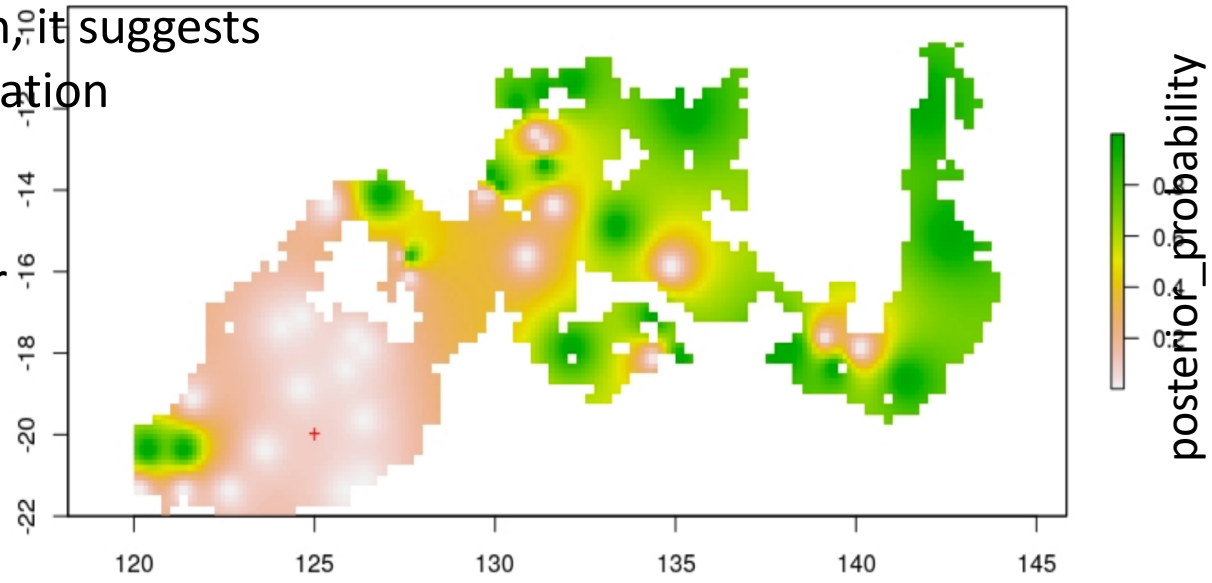
You can run:

interpolate_posterior_probability(data=data, mask=mask2, x0=x0, proj4string=crs(mask))



Interpreting results

• If you have two sampled places for empirical data (the red-x and some other location), and analysis of you get a posterior probability in bpp that it is two (not one species) in the empirical data– that is pp ≈ 1.0, but the sampling site corresponds to an area in green, it suggests the degree of genetic differentiation is consistent with *intraspecific IBD* without having to invoke a higher order process (i.e., evolution of reproductive isolation) to explain the observed differentiation.



• Conversely, If the sample of your empirical data corresponds to an area with low pp in the simulation (i.e., the light color), it suggests that something other than intraspecific IBD has contributed to the differentiation – i.e., there is something interesting that suggests the differentiation is not consistent with intraspecific expectations.

Adapting the simulation:

Besides changing the lanscape tiff file, you can also modify the spatial process parameters in the decrypt/example/spatial_process.ctl file: parameters description is given below.

> **Caution** changing the sampling parameters in the configuration file like the number of loci or the number of gene copies, requires to also modify the bpp.ctl file: BPP does not detect these parameters automatically.

Also, modification of the spatial model itself *or the sampling scheme*
requires modifying the C++ source code. It should be reasonably easy to do if you know C++. The model specification is in the decrypt/cpp/spatial_simulator.h file in the code project, check it out.

Software: *Decrypt* https://becheler.github.io/pages/applications.html

- Model of the geography of genetic divergence under a spatially explicit coalescent to evaluate competing hypotheses about cryptic diversity (inferred under the MSC)