**Urfus T., Pekařová M., Rejlová L., Záveská E., Weiser M., Josefiová J. & Chrtek J. (2024) *Urtica kioviensis*, a rare species of stinging nettle threatened by hybridization. – Preslia 96: 329–349.**

**Supplementary Data S2.** R code describing the classification of the non-reference individuals into taxa.

# Classification of the sampled individuals into *U. dioica* 2n, *U. kioviensis* and “hybrids” according to the relative fluorescence of nuclei

The classifier is built around an idea that the processes that form the distribution of relative fluorescence of nuclei values within population are complex and not necessarily lead to the normally-distributed values within populations. On the other hand, the differences between the populations, are (also thanks to Central Limit Theorem) random draws from the normal distribution. Therefore, we estimate the effect of population identity from a parametric model, i.e. as a random factor of a linear mixed-effect model, but we keep the structure of the data within the populations as they are represented in the reference populations. Essentially, in order to mimic the possible overall distribution of the relative fluorescence values for each of the taxa, we just shift the real reference populations by some constant, so that the constant comes from Normal distribution of zero mean and std. dev. as estimated by the linear mixed-effects model. Then, we use the simulated data to estimate the kernel density for a taxon at given relative nuclear fluorescence values.

```
# packages
library("lme4")

# check the names in the "pure" taxa dataset
# in this example, it should be:
# id_sample, population, taxon, rel_fluor
names(pure)


# make the model
model <- lmer(rel_fluor ~ taxon + (1|population), data= pure) #check the fit!
# extract the st. dev. at the population level
sdpop <- as.data.frame(VarCorr(model,comp="Std.Dev"))[1,"sdcor"]
# actual values of the random effect levels (populations)
re <- unlist(ranef(mod1)) #check this for the order of levels!

# simulate 99 replicates for each of the populations

# generate the random effects
popeff <- rnorm(x=4 * 99, mean=0, sd=sdpop) #4 populations, 99 replicates
# sort the data first
pure <- pure[order(pure$population), ]
# replicate the actual data
taxon_sim <- rep(pure$taxon, times=99)
population_sim <- rep(pure$population, times=99)
```

```
sample_id_sim <- rep(pure$sample_id_sim, times=99)
rel_fluor_rep <- rep(pure$rel_fluor, times=99)
is_sim <- rep(c(FALSE, TRUE), c(nrow(pure), 99 * nrow(pure)))
# simulate new data
rel_fluor_sim <- rep(NA, 99 * length(pure$rel_fluor))
for (quadruple in 0:98){
    current_popeffs <- popeff[(1:4) + quadruple * 4] - re) #check this for the order of leve
    addthis <- rep(current_popeffs, times=table(pure$population)) #check this for the order
    index <- (1:length(addthis)) + (length(addthis) * quadruple)
    rel_fluor_sim[index] <- rel_fluor_rep[index] + addthis
    #print(pg_sim)
}
# reconstruct the data frame
data_sim <- data.frame(taxon=taxon_sim,
    population=population_sim,
    sample_id=sample_id_sim,
    rel_fluor=rel_fluor_sim)
all_pure <- rbind(pure,data_sim)
all_pure$is_sim <- is_sim

# get the data for U. dioica
dio_sim <- all_pure[all_pure$taxon == "dioica", ]
# get the data for U. kioviensis
kio_sim <- all_pure[all_pure$taxon == "kioviensis", ]

# estimate the prob. densities
# try with different kernels and bandwidths, but it seems to be quite stable
d_e <- density(dio_sim$rel_fluor, kernel="e")
k_e <- density(kio_sim$rel_fluor, kernel="e")

## optionally, check the fit
#par(mfrow=c(1,2))
#hist(dio_sim$rel_fluor, freq=FALSE, main="U. dioica")
#lines(d_e)
#hist(kio_sim$rel_fluor, freq=FALSE, main="U. kioviensis")
#lines(k_e)
```

We linearly interpolate the estimated densities for the fluorescence values of the individuals to be classified. We rescale the densities into probabilities of an individual belonging to a certain taxon. The rescaling is based upon these formal priors:

1. An individual may belong either to one of the two reference taxa, or is a “hybrid”. Therefore, the probabilities of belonging into the groups for a given individual sum to one.

2. We have no prior information about relative fluorescence of nuclei of individ-

uals that do not belong to any of the reference taxa ("hybrids"). Therefore, the prior probability for an individual to be non-refrence ("hybrid") should be flat acrros the whole range of relative fluorescence of nuclei.

3. We have no prior information about true frequency of the taxa among the individuals to be classified. Therefore, the prior probabilities for an individual to belong to one of the three qroups should be equal.

```r
# check the names in the mixed populations dataset
names(mixpop)
```

```
## [1] "sample_id"   "population" "rel_fluor"
```

```r
# estimate the prob. of species identities
# set species prob. to 0 outside the classifier range

# kioviensis
is.kio <- approx(k_e, xout=mixpop$rel_fluor)
is.kio <- data.frame(is.kio)
is.kio$y[is.na(is.kio$y)] <- 0

# dioica
is.dio <- approx(d_e, xout=mixpop$rel_fluor)
is.dio <- data.frame(is.dio)
is.dio$y[is.na(is.dio$y)] <- 0

# rescale the probabilities so each of them sums to 1
hybrid <- rep(1 / nrow(mixpop), nrow(mixpop))
p_kio <- is.kio$y / sum(is.kio$y)
p_dio <- is.dio$y / sum(is.dio$y)

# probabilities for each individual sum to 1
p_tot <- hybrid + p_kio + p_dio
prob_kio <- p_kio / p_tot
prob_dio <- p_dio / p_tot
prob_hyb <- hybrid / p_tot

# assemble the data with the classification
classif <- data.frame(mixpop, prob_dio, prob_kio, prob_hyb)
```