R-WALKTHROUGH

Quantifying the individual impact of artificial barriers in freshwaters:

A standardized and absolute genetic index of fragmentation

Jérôme G. Prunier [1] *, Camille Poesy [1], Vincent Dubut [2], Charlotte Veyssière [3], Géraldine Loot [3],

Nicolas Poulet [4] & Simon Blanchet [1,3]*

[1] Centre National de la Recherche Scientifique (CNRS), Université Paul Sabatier (UPS); Station d'Ecologie Théorique et Expérimentale, UMR 5321, F-09200 Moulis, France

[2] Aix Marseille Univ, CNRS, IRD, Avignon Univ, IMBE, Marseille, France

[3] CNRS, UPS, École Nationale de Formation Agronomique (ENFA) ; UMR 5174 EDB (Laboratoire Évolution & Diversité Biologique), 118 route de Narbonne, F-31062 Toulouse cedex 4, France

[4] Pôle écohydraulique AFB-IMT, allée du Pr Camille Soula, 31400 Toulouse, France

*Corresponding authors:
Jérôme G. Prunier and Simon Blanchet
Station d'Ecologie Théorique et Expérimentale, UMR 5321, F-09200 Moulis, France
2 route du CNRS, 09200 Moulis
Phone: (+33)561040361
E-mail: jerome.prunier@gmail.com, simon.blanchet@sete.cnrs.fr

The F_{INDEX} computation was automated within a user-friendly R-function. The FINDEX() R-function is embedded within the "FINDEXpackage.rda" file (http://www.jeromeprunier.fr/Tools.html). This .rda file also contains all RandomForest predictions required to compute the F_{INDEX}. Users are invited to download this .rda file within their working directory, and to install required R-libraries ("adegenet", "randomForest", "mmod", "lme4" and "reshape2").

```
# Installing required libraries (if needed)
    install.packages(c("adegenet ", " randomForest ","mmod","lme4","reshape2"),dependencies=TRUE)

# Going to working directory (where " FINDEXpackage.rda" is located)
    setwd('path_to_working_directory')

# Load .rda file to import the FINDEX() R-function and all other required files in the R-environment
    load("FINDEXpackage.rda")
```

Users are then simply expected to provide one or several empirical genotypic datasets (in the *genepop* format with a ".gen" extension) and a parameter file (a dataframe in ".txt" format) with a line for each combination of both an obstacle and a species and at least **eight mandatory columns**:

- "**Species**": the name of the considered species (factor);
- "**Barrier**": the name of the considered obstacle (factor);
- "**Upstream**": the name of the upstream population as found in the corresponding *genepop* file (factor);
- "**Downstream**": the name of the downstream population as found in the corresponding *genepop* file (factor);
- "**Creation**": the date of barrier creation (numeric);
- "**Generations**": the number of generations elapsed since barrier creation (numeric; this number of generations is to be estimated from the life-history traits of the considered species);
- "**Genepop**": the name of the *genepop* file from which genotypes are to be extracted (with or without the ".gen" extension); see https://genepop.curtin.edu.au/help_input.html#Input for help with the formatting of *genepop* files;
- "**Digits**": the number of characters used to code an allele in the *genepop* file.

Note that the parameter file may contain any additional column, provided that these columns add information about **barriers** (for instance, height, spatial localization, etc.).

In an illustrative purpose, we provide a parameter file "*Prunier2018_illustration.txt*" as well as two *genepop* files ("*Phoxinus.gen*" and "*Gobio.gen*") to test the FINDEX R-function. These files are provided as a compressed zip-file (http://www.jeromeprunier.fr/Tools.html).
The "*Prunier2018_illustration.txt*" looks like the following:

River	**Species**	**Barrier**	**Upstream**	**Downstream**	**Creation**	**Generations**	**Genepop**	**Digits**
Cele	Gobio	SCC	SCCup	SCCdown	1960	20	Gobio.gen	3
Viaur	Gobio	CAP	CAPup	CAPdown	1700	124	Gobio	3
Cele	Phoxinus	SCC	SCCup29	SCCdown30	1960	25	Phoxinus.gen	3
Viaur	Phoxinus	CAP	CAPup30	CAPdown30	1700	155	Phoxinus	3

It contains the eight mandatory columns (in bold) and an additional column indicating the river in which each barrier is located.
In the parameter file, the names of *genepop* files (Genepop column) can be provided with or without the .gen extension.

In the "*Phoxinus.gen*" file, genotypes are provided with the **individual numbering** directly attached to the population name (e.g., SCCup**01**, CAPdown**02**, etc.). In that case, the name of each population (Upstream and Downstream columns) is the name of the last genotype in that population (e.g., SCCup29, CAPdown30, etc.).

In the "*Gobio.gen*" file, genotypes are either provided with the **individual numbering** separated from the population name by an underscore (e.g., SCCup**_01**, SCCdown**_02**, etc.), in which case the name of each population (Upstream and Downstream columns) is the part of the genotype name located before the underscore (e.g., SCCup, SCCdown), or without any individual numbering, in which case the name of each population (Upstream and Downstream columns) is directly the name of genotypes.

Once these illustration files have been extracted within the working directory, users simply have to run the FINDEX() R-function as follows:

```
# Running the FINDEX() function on illustration files
    FINDEX_test=FINDEX(input_file="Prunier2018_illustration.txt")
```

A "verbose" option can be turned off ("verbose=FALSE"), in which case no information about the computation progress will be provided.

This command will launch the computation of the F_{INDEX} and return a list with two elements: "output1" and "output2".

"output1" is a dataframe providing results for each combination of both an obstacle and a species, with at least the following columns (as well as any additional (non-mandatory) column from the input file):
- "Species": the name of the considered species (factor)
- "Barrier": the name of the considered obstacle (factor)
- "Upstream": the name of the upstream population (factor)
- "Dowstream": the name of the downstream population (factor)
- "Creation": the date of barrier creation (numeric).
- "Generations": the number of generations elapsed since barrier creation (numeric)
- "He_Up": the expected heterozygosity in the upstream population (numeric)
- "He_Down": the expected heterozygosity in the downstream population (numeric)
- "He": the averaged expected heterozygosity across populations (numeric)
- "Obs_PhiST": the observed measure of φ'st (numeric)
- "Obs_GST2": the observed measure of G''st (numeric)
- "Findex": the computed F_{INDEX} (numeric)
- "ciFindex" : the 95% confidence interval about the F_{INDEX} (numeric)

By default, "output2" is a dataframe providing results **for each obstacle** with F_{INDEX} values averaged across species, with at least the following columns:
- "Barrier": the name of the considered obstacle (factor)
- "Creation": the date of barrier creation (numeric)
- "Findex": the computed F_{INDEX} (numeric).
- "ciFindex" : the 95% confidence interval about F_{INDEX} (numeric).

The FINDEX() function can also be used to compile F_{INDEX} values across other factors than obstacles (the default). In that case, users will have to use the "CompileBy" option and indicate which additional column from the input parameter file will be used for averaging. This column must be coded as factors. From the illustration files, F_{INDEX} values could for instance be compiled across rivers, as follows:

```
# Running the FINDEX() function on illustration files with Findex values compiled across rivers
    FINDEX_test=FINDEX(input_file="Prunier2018_illustration.txt", CompileBy="River")
```

Here, the "output2" dataframe will provide results for each river, with F_{INDEX} values averaged across both obstacles and species.

Eventually, the results can be plotted using the following commands:

```
# Loading required libraries (to be installed if needed)
    require("ggplot2")
    require("viridis")


# Plotting results for each combination of both an obstacle and a species
    ggplot(FINDEX_test$output1,  aes(x=reorder(Barrier,  as.numeric(FINDEX_test$output1$Creation)),  y=
            Findex,group=Species, fill= Findex))+
        geom_bar(width = 0.9,stat="identity", position=position_dodge(.9))+
        geom_text(aes(y = 105, label = Species, angle=0),position = position_dodge(width = .9))+
        coord_cartesian(ylim=c(0,105)) +
        geom_errorbar(aes(ymin=        Findex-ciFindex,        ymax=        Findex+ciFindex),        width=0,
            position=position_dodge(.9))+
        scale_fill_viridis(discrete = FALSE,begin = 0.1,end=0.9,limits=c(0,100))+
        geom_hline(yintercept=c(0,20,90,100), linetype="dashed", color = "black")+
        scale_x_discrete(name ="Obstacles")


# Plotting results for each obstacle with FINDEX averaged across species
    ggplot(FINDEX_test$output2,  aes(x=reorder(Barrier,  as.numeric(FINDEX_test$output2$Creation)),  y=
            Findex,group=Barrier, fill= Findex)) +
        geom_bar(width = 0.9,stat="identity", position=position_dodge(preserve = 'single'))+
        coord_cartesian(ylim=c(0,105)) +
        geom_errorbar(aes(ymin=        Findex-ciFindex,        ymax=        Findex+ciFindex),        width=0.1,
            position=position_dodge(width = 0,preserve = 'single'))+
        scale_fill_viridis(discrete = FALSE,begin = 0.1,end=0.9,limits=c(0,100))+
        geom_hline(yintercept=c(0,20,90,100), linetype="dashed", color = "black")+
        scale_x_discrete(name ="Obstacles")
```

Note that users also have the possibility to provide their own GD_{min} and GD_{max} values, in the form of two dataframes, each with two columns ('GST2' and 'PhiST') and the same number of rows as the input file (NA values are allowed, in which case the predicted GD_{min} and/or GD_{max} values would be used). These options are yet to be used with caution (see main manuscript for details).

```
# Creating GDmin and GDmax dataframes and running the FINDEX() function
    GDMIN=data.frame(GST2= rep(0.01,4),PhiST= rep(0.02,4))
    GDMAX=data.frame(GST2=rep(0.2,4),PhiST= rep(0.3,4))
    FINDEX_test=FINDEX(input_file="Prunier2018_illustration.txt",GDmin=GDMIN,GDmax=GDMAX)
```