

# PURGd

A software to detect purging and to estimate  
inbreeding-purging genetic parameters  
in pedigreed populations

Version 2.3.2

User's guide updated in October 2022  
by Aurora García-Dorado



Eugenio López-Cortegano

Diego Bersabé

Jinliang Wang

Aurora García-Dorado

# Table of Contents

|      |   |    |
|------|---|----|
| 1.   | PURGd 2.3.2.....  | 3  |
| 1.   | Installation .....  | 5  |
| 2.   | Program folders .....   | 6  |
| 3.   | Input files .....   | 7  |
| 4.   | Running PURGd.....  | 8  |
| 4.1. | Program options and default choices.....                            | 10 |
| 4.2. | Modifying PURGd settings from the command line .....                | 12 |
| 5.   | Output files .....  | 15 |
| 5.1. | Output files with the estimates of the parameters of the model..... | 15 |
| 5.2. | Databases .....   | 17 |
| 6.   | About PURGd .....   | 18 |
| 7.   | References .....  | 19 |

## 1. PURGd 2.3.2

PURGd is a software developed to detect purging and to estimate the genetic parameters of the inbreeding-purging (IP) model introduced by García-Dorado [3]. The main objective of this program is to estimate both the rate of inbreeding depression and the effective purging coefficient ( $d_e$ , hereafter referred to as  $d$  for simplicity) which represents an overall genomic measure of the component of the deleterious effects that is only expressed in homozygosis, and is therefore responsible for purging under inbreeding.

PURGd analyzes pedigreed data for fitness traits using the models and methods described in García-Dorado *et al.* (2016) [4]. Traits that are not considered main fitness components can also be analyzed, although: i) the model assumes that recessive alleles have usually negative effects on the trait and low initial frequencies; ii) the rate of inbreeding depression will no longer be interpreted as an estimate of the inbreeding load (see below). This manual is written from the perspective of a fitness or fitness-trait analysis.

The program estimates the regression coefficients ( $b$ ) on the purged inbreeding coefficient ( $g$ ), here denoted  $b(g)$  and, optionally, on maternal purged inbreeding  $b(g_{dam})$ . Note that “ $-b(g)$ ” is the rate of inbreeding depression  $\delta$ , which estimates the inbreeding load of the fitness trait analyzed ( $B$ , classically measured in lethal equivalent units). Analogously, “ $-b(g_{dam})$ ”, is the rate of inbreeding depression ascribed to maternal genetic effects of the trait (i.e., the maternal component of the inbreeding load, when analyzing fitness traits). Furthermore, the program gives estimates for the regression coefficients on additional regressor variables, such as environmental factors. This software also includes options to estimate parameters for purging models based on ancestral inbreeding, developed by Ballou [1] and Boakes & Wang [2].

Previous versions of this software used two alternative estimation approaches: Linear regression (LR) and numerical non-linear regression (NNLR). In the first case, a linear regression model was fitted to log-transformed fitness data for each  $d$  value considered. It was warned that LR overestimates  $\delta$ , particularly when  $d$  is large [4]. Furthermore, this LR approach cannot use data with fitness values less than or equal to zero. The present version (PURGd 2.3) uses only the NNLR approach, where the non-linear (exponential) model for untransformed fitness is explored by numerically searching for the joint LS

estimates of  $d$  and of the non-linear regression coefficients ( $b(g)$  and coefficients on additional optional factors). Previous versions allowing LR estimation are still available.

PURGd also computes the standard Wright's inbreeding coefficient  $F$ , Ballou's ancestral inbreeding coefficient  $F_a$  [1] and García-Dorado's purged inbreeding coefficient  $g$  [3] for the individuals in the pedigree, as well as the effect of other genetic and environmental factors of interest introduced in the model.

## 1. Installation

The present software has been written in the C++ language and compiled with GNU g++ 7.2.1 under a GNU/Linux environment (Arch Linux) with kernel version 4.14.5. It has also been built for the Microsoft Windows platform under a MSYS2 environment using the MinGW-w64 toolchain. The package includes two executable files:

- **GNU/Linux:** An executable binary file of the program (PURGd) that can be found in the bin folder. No installation is needed.
- **Microsoft Windows:** An executable binary file of the program (PURGd.exe) that can be found in the bin folder. No installation is needed. See BOX 1 in Section 4 for a step-by-step explanation on how to run PURGd on Windows.

The software, its source files and documentation are also accessible from a long-term repository server (<https://gitlab.com/elcortegano/PURGd>), making it easier to interact with developers and users, who can submit their issues in a dedicated tracker.

## 2. Program folders

- Both the GNU/Linux and the Microsoft Windows packages include two folders:  
bin: including the executable binary file.
- input: including sample pedigree files corresponding to simulations analyzed by García-Dorado et al. [4]. Data in “D#” files were simulated assuming just drift, while data o “MSD#” files were simulated assuming mutation, selection (including purging) and drift. Two additional data files are included: one with two additional factors and one where the trait (survival) is measures as a percent.

Additionally, the package contains this user’s guide in pdf format and a copy of the License of the software PURGd as a text file.

### 3. Input files

Input files can be placed in the existing input folder for convenience, although this is not mandatory. They must be in comma-separated values format (.csv) and have at least four columns, with the following precise order: identity (ID) of the individual, mother ID, father ID, and the evaluation of the fitness-trait to be analyzed. Fitness must be recorded using the natural measure scale (i.e., not log-transformed). IDs can be numerical or strings of characters (excluding comma). The file must contain the names of these variables in the first line.

Individuals should be ordered in the file from older to younger. Individuals whose parents are not present in the first column (*i.e.*, in the individual identity column) are assumed to be unrelated. This is so even if several of these parents have been coded with the same identity. Therefore, the pedigree file should contain a record for each individual (with its identity in the first column), including unrelated founders. We advise coding the identity of the parents of assumedly unrelated founders or of individuals from the non-inbred base population as 0. Missing values for fitness or for additional factors must be coded as NA. (Figure 4.1).

Extra columns can be added containing additional causal factors to be fitted in the model. Note that including additional factors that are correlated to inbreeding (year of birth, for example) may distort the estimates of inbreeding purging parameters.

| Qilin.csv                 |    |    |   |   |      |     |
|---------------------------|----|----|---|---|------|-----|
| ID,Dam,Sire,Longevity,YOB | 1  | A  | B | C | D    | E   |
| 1,0,0,1942,273            | 2  | 1  | 0 | 0 | 1942 | 273 |
| 2,0,0,2106,273            | 3  | 2  | 0 | 0 | 2106 | 273 |
| 3,0,0,2781,273            | 4  | 3  | 0 | 0 | 2781 | 273 |
| 4,2,1,2051,275            | 5  | 4  | 2 | 1 | 2051 | 275 |
| 5,3,1,2593,275            | 6  | 5  | 3 | 1 | 2593 | 275 |
| 6,0,0,2399,273            | 7  | 6  | 0 | 0 | 2399 | 273 |
| 7,2,1,4717,276            | 8  | 7  | 2 | 1 | 4717 | 276 |
| 8,2,1,757,276             | 9  | 8  | 2 | 1 | 757  | 276 |
| 9,3,1,919,276             | 10 | 9  | 3 | 1 | 919  | 276 |
| 10,4,1,2655,277           | 11 | 10 | 4 | 1 | 2655 | 277 |
| 11,4,1,2,277              | 12 | 11 | 4 | 1 | 2    | 277 |
| 12,5,1,518,277            | 13 | 12 | 5 | 1 | 518  | 277 |
| 13,5,1,422,277            | 14 | 13 | 5 | 1 | 422  | 277 |

**Figure 4.1:** Left: A pedigree file (Qilin.csv) with data for a fitness component trait (longevity) is shown using a text editor, with no blanks, and comma (,) separated values; Right: The same file is shown using a spreadsheet program, such as LibreOffice Calc or Microsoft Excel.

## 4. Running PURGd

PURGd runs from the terminal (GNU/Linux) or the command prompt (Windows).

**See BOX 1 for a step-by-step guide to run PURGd on Microsoft Windows.**

PURGd must be run from the `bin` folder described above and uses the following syntax:

```
./PURGd [--options] PATH/FILE
```

Only the additional argument `FILE`, which is the name of the input file preceded by its absolute or relative path (`PATH`), is mandatory to run the program.

Options are not required to run the program but, if specified, they will take preference over the default settings (see Sections 4.1 and 4.2 for more details).

**By entering:**

```
./PURGd --help
```

**the program will print a short manual to be used as a quick reference guide**, as well as to check the PURGd version being used.

A simple example on how to run PURGd is given below:

```
./PURGd ../input/MSD1.csv
```

In this example, the program will use the IP model, which is the default option, to analyze the data contained in the file `MSD1.csv`. Note that the relative path `../input/` needs to be typed because PURGd must be run from the `bin` folder and `MSD1.csv` is located in the `input` folder.

Once the analysis has finished, the software will print a short message in the terminal. The corresponding output files (see Section 5) will be saved in the `bin` folder by default.



### **BOX 1. A step-by-step guide to use PURGd on Microsoft Windows**

1. Download and unzip PURGd in any folder
2. Save your data file, with the correct .csv format (see section 3), in the input folder. Let's assume that the name of your input file is the MSD1.csv. sample file provided with the package.

#### **3. To run PURGd:**

Use the Windows File Explorer to move to the bin folder where you extracted PURGd. **To open the console, click on the secondary mouse button while holding the shift keyboard button at the same time and select “Open command window here” or “Open Power Shell”. Now, any command you type in the console will work from the PURGd bin directory.**

Now you are ready to run PURGd from the command prompt. For example, to analyse the data in your MSD1.csv file using the IP model (default option), you should type:

```
./PURGd ../input/MSD1.csv
```

where ../input calls the data from the input folder.

If, for example, you want to estimate the non-inbred average using NNLR as the remaining parameters of the model (recommended) and to run the program 50 times to evaluate how stable the output is, you need to set those options when you run the program by typing:

```
./PURGd --nruns=50 --w0 ../input/MSD1.csv
```

**Note:** When typing the path to the input file, you may need to check whether to use the forward slash (“/”) or the typical backslash (“\”) that is common in Windows environments. Also, note that some versions of the Windows command prompt do not admit the “./” characters before the keyword PURGd. Take care of typing blank spaces (in this example, between “./PURGd” and “..”).

## 4.1. Program options and default choices

PURGd ships with a complete set of predefined default values for the running parameters. Some of these values can be modified to change the way that input data are analysed. This section describes all the available options that can be modified by the user, as well as their corresponding default values. The next subsection (4.2) explains how to change these options from the command line.

- **Model:** The model assumed to predict the expected value of the fitness trait. It can be the inbreeding-purging (IP) model [3] or a model based on ancestral inbreeding: Ballou's [1], Boakes & Wang's or Ballou-Boakes & Wang's mixed model [2]. Default model is IP.
- **Genedrop:** Computes ancestral inbreeding coefficients by using the gene dropping simulation process described in Suwanlee *et al.* (2007) [6]. By default, gene dropping simulations are disabled and expected ancestral inbreeding values are computed using Ballou's equation [1].
- **Estimate of the initial average fitness ( $w_0$ ):** A value for the initial (non-inbred) average fitness parameter can be set by the user if there is information available. By default,  $w_0$  is computed as the average fitness of non-inbred individuals with non-inbred ancestors ( $F = F_a = 0$ ). This reduces the downwards bias in the estimate of the rate of inbreeding depression and the purging coefficient. However, under this default option, the remaining estimates and their error are conditional to this estimate of  $w_0$ , and they should be taken with caution if the standard error of  $w_0$  is large. In this case, we recommend to estimate  $w_0$  through NNLr together with the remaining parameters of the model. This last option can produce some overfitting of the model, leading to some downward bias in the estimates of  $\delta$  and  $d$ , but it is recommended whenever the default setting produces average fitness estimates with relatively large sampling error. See below how to set the maximum value explored to estimate  $w_0$  under this option.
- **rate of inbreeding depression :** PURGd estimates  $b(g)$  which, multiplied by -1, gives an estimate of the rate of inbreeding depression (*i.e.*,  $\delta = -g(b)$ ). By default,  $g(b)$  is estimated at the same time as the remaining parameters in the corresponding model during the NNLr procedure.

An option can also be run to obtain an estimate of  $b(g)$  using only non-purged individuals (those with non-inbred ancestors,  $F_a = 0$ ) and assuming  $d=0$ .

A  $\delta$  value can be settled by the user to be assigned to  $-g(b)$ , for example, an estimate previously obtained using only non-purged individuals.

The program can also be run for a  $d$  value specified by the user.

- **Use of maternal effects on inbreeding depression:** Maternal effects on inbreeding depression can be incorporated in the analysis by using the purging inbreeding coefficient of the dams as an additional independent variable (always assuming the same  $d$  value for maternal and non-maternal components). By default, they are not used.
- **Use of additional factors:** Several other numerical factors can be incorporated into the model to reduce the estimation noise from non-genetic sources. They can be regressor variables or categorical factors. Their effect will be estimated as additional regression coefficients or as a set of fixed effects, respectively. By default, none is used.
- **Output:** Sets the path where output files are saved. This custom directory must exist before running the program. Using absolute paths is recommended over relative paths. By default, output is stored in `bin`.
- **Save a database:** Saves a file containing both the fitness and the inbreeding coefficients of the analyzed individuals. By default, no databases are saved.
- **Accuracy for the search of the purging coefficient:** Sets the width of each increase of the purging coefficient  $d$  during the numerical search of its least square estimate. It can be modified if the user. By default, 0.01 is used.
- **Seed:** Sets a seed value, required to generate pseudorandom numbers during the analysis. It can be convenient set a seed in order to replicate the results obtained. Default seed is the current time.
- **Verbose mode:** Prints a short summary in the terminal during program execution for each pedigree that is being analysed. It is disabled by default.
- **Number of runs:** Sets the number of times to run the NNLR method. If more than one, results are given averaged over runs, together with the corresponding standard

deviation (in the next row) that measures the stability of the NNLR results. **Note that these standard deviations cannot be used to estimate the standard errors** of the estimates, as they do not include the component due to sampling error that accounts for different estimates being obtained from different data sets sampled from the same population. By default, only one run is used. It is convenient to run each analysis several times to check the stability of the results.

- **Maximum value of the initial average fitness:** When fitness is estimated during the NNLR procedure, the values of the initial average fitness are explored from 0 to a maximum value. By default, 1.0 is the maximum allowed, which is appropriate for relative fitness or for standard viability measures. When the fitness trait can take values larger than 1, the user needs to introduce the maximum value to be explored when searching for the  $w_0$  estimate. Then, the program will automatically scale the accuracy during the NNLR search by multiplying the default width of the  $w_0$  search steps (0.01) by the maximum  $w_0$  value provided by the user, so that the program is not slowed.
- **Maximum value of the rate of inbreeding depression:** Values for the estimate of  $\delta$  are explored from 0 to a maximum value (the default is 10).
- **Minimum and maximum values of the slope for other regression terms:** If additional factors are used, the values of their regression coefficients or effects can be explored from a minimum to a maximum value. By default, this range is  $[-10,10]$ .

Options modified from the command line prevail over the default options.

## 4.2. Modifying PURGd settings from the command line

PURGd default options, can be modified from the command line, where each option is preceded by a double hyphen (--). Note that any command-line option will take preference over its default value. These options are:

`--d=NUM`: Performs the Inbreeding-Purging analysis assuming a given `d` value (NUM), which must be between 0 and 0.5.

`--ip / --ffa / --fa / --faffa`: Sets the purging model to inbreeding-purging, Ballou's, Boakes & Wang's, or their mixed model, respectively [2, 5]

`--genedrop=NUM`: Activates the gene dropping simulations using NUM iterations. If no value is specified (i.e., `--genedrop`), a default number of  $10^6$  iterations is used.

`--w0`: When typing `--w0`, PURGd estimates the original non-inbred average fitness during NNLR, as the remaining parameters. If `--w0` is not typed, `w0` is estimated as the average value for individuals with  $F=Fa=0$ , and is given with its standard error.

`--w0=NUM`: Sets the initial average fitness to the value NUM.

`--maternal`: Activates the use of maternal purged inbreeding effects

`--factor.cols=`: Activates the use of additional regressor variables. The equal sign must be followed by one or more numbers separated by commas, matching the column number of an additional factor in the input file. This option needs to be used together with the `--factor.names` option.

`--factor.names=NAME`: Sets the names of the additional regressor factors. The equal sign must be followed by one or more names, separated by commas. Each name is assigned to each factor following the order specified with the `--factor.cols` option.

`--cfactor.cols=`: Activate the use of additional categorical factors. The equal sign must be followed by one or more numbers, separated by commas matching the column number of an additional categorical factor in the input file. This option needs to be used together with the `--cfactor.names` option below.

`--cfactor.names=NAME`: Sets the name of the additional categorical factors. The equal sign must be followed by one or more names, separated by commas. Each name is assigned to each factor following the order specified with the `--cfactor.cols` option.

`--output=PATH`: Changes the path where output files will be saved to PATH

`--save-db`: Saves a database that contains the values for fitness and for the different coefficients ( $g$  computed using the  $d$  value estimated or the one provided by the user).

--accuracy=NUM: Sets the accuracy (steps width) in the search of the estimate of  $d$  to NUM, which must be a positive number up to 0.5

--seed=INT: Sets the seed value to INT, which must be a positive integer

--verbose: Enables the verbose mode

--nruns=INT: Sets the number of runs of the ABC algorithm to INT, which must be a positive integer. Due to the stochastic nature of this approach, it is very convenient running each analysis several times to check the stability of the results.

--max-w0=NUM: Sets the maximum value to search for the estimate of the initial average fitness to NUM. Its default value is 1. A new max-w0 value needs to be established when the scale of measure for fitness gives values larger than 1.

--max-delta=NUM: Sets the maximum value to search for the estimate of the rate of inbreeding depression .

--delta=NUM: Assigns to  $g(b)$  the value “-NUM”. NUM can also be the character n, in order to estimate  $b(g)$  with the other purging parameters during NNLR (default option). Alternatively, the option --delta=s allows to use PURGd to obtain an estimate of  $b(g)$  assuming  $d=0$  and using only individuals with  $F_a = 0$ .

--factor.range=NUM1,NUM2: Sets the minimum and maximum values to explore for the effects of additional factors to NUM1 and NUM2, respectively

--help: Prints a short summary on how to use the program, including a list of the most common options.

## 5. Output files

Output files are stored in the bin folder or in the custom output directory once the program finishes the analysis. As with the input pedigree files, output files are also in csv format, so that they can be easily converted for a friendly view when opened with a spreadsheet program. For example, with Microsoft Excel 2007 or later, select the first column of the file, go to the DATA tab and choose, in this order: “text in columns” - “delimited” - “comma values”.

**Important note:** Any new analysis with the same model and method will overwrite the existing output file. Note that, in these cases, the new output files may not be saved if the old output is still open while the program is running the new analysis.

PURGd generates three kind of output files, which are described below.

### 5.1. Output files with the estimates of the parameters of the model

Each analysis performed will save the estimates obtained in a file with the name the [input file name]\_[model].csv in the bin or output folder.

When NNLR is run just once, only two output sets are included in this file, each in a different row. The first row shows the results for the analysis performed considering purging, while the second one refers to an analogous analysis assuming no purging ( $d=0$ ). Comparing these two analyses shows how far fitting improves by considering purging. When the option to estimate  $\delta$  using exclusively individuals with non-inbred ancestors ( $F_a=0$ ) is enabled, only results assuming no purging are displayed.

When NNLR is run several times (see --nruns option), one additional row is included below each analysis showing **the standard deviation of estimates from different runs**, which **should be used just to evaluate the stability of NNLR estimates. Not to compute the standard error of the estimates.**

The standard output consists of the following columns:

- **Analysis:** The name of the model used in the analysis. For the analysis considering purging, this name indicates the “Inbreeding-purging model”, “Ballou’s model”,

“Boakes & Wang’s model” or “Mixed model” [2,3,5]. The analysis assuming no purging is always labelled as “No purging model”.

- **d coefficient:** The estimated (or assumed) value of the effective purging coefficient. Note that output files for analysis based on  $F_a$ -models will display a value of  $d = 0$  that should be disregarded.
- **RSS:** The residual sum of squares

\*\*\*\*\*

- **WARNING: Errors have been detected in the calculation of AICc and Chi2 under some options, as well as in the P-values for very large CHI2 values. Although the consequences are often irrelevant, I have included in the package an output-patch excel file where RSS can be pasted to obtain correct results.**

- **AICc:** The corrected Akaike’s Information Criterion (which assumes normality for residual errors).
- **RL:** The ratio of the likelihood of the purging (IP) model over that of the non-purging ( $d=0$ ) model,
- **Chi2:** a statistic to test the null hypothesis  $d=0$  against  $d>0$ , to be computed as

$$\text{Chi2} = 2[\text{Loglikelihood}(\text{analysis estimating } d) - \text{Loglikelihood}(\text{analysis with } d=0)]$$

This statistic distributes approximately  $\chi^2$  with parameter  $v = K(\text{IP}) - K(d=0) = 1$  under the null hypothesis that  $d=0$ .

**p-value** for the test that  $d > 0$  in the case of the Inbreeding-Purging model, or for non-null regression values of the purging regression terms in the remaining models.

\*\*\*\*\*

- **w0:** The initial non-inbred mean for fitness
- **SE(w0):** The empirical standard error for  $W_0$  when estimated using individuals with  $F=F_a=0$ .
- **b(factor):** The value of the regression coefficient for each factor included in the analysis. In the IP analysis,  $-b(g)$  estimates the  $\delta$  due to direct genetic effects and  $-b(g_{\text{dam}})$  estimates the rate of inbreeding depression ascribed to maternal effects ( $\delta_{\text{dam}}$ ).



## 5.2. Databases

A databases with the [input file name]\_[model].csv extension in their filename will be saved in the output folder if the corresponding option is specified. Database includes the following columns:

- **ID**: Identity of the individual
- **W**: Fitness values, as used in the analysis
- **F**: Standard inbreeding coefficient
- **g(d)**: Purged inbreeding coefficient, computed with the estimate obtained for  $d$
- **Fa**: Ancestral inbreeding coefficient
- **Fa(genedrop)**: Ancestral inbreeding coefficient estimated by gene dropping. This column is only shown if the gene dropping option was enabled.

Furthermore, if maternal and/or other factors are included in the model, additional columns will contain their values.

Note that individuals with unknown fitness will not appear in this output file.

## 6. About PURGd

The first version of this software, PURGd 1.0, was developed by Eugenio López-Cortegano, Jinliang Wang, and Aurora García-Dorado.

The current version of PURGd is 2.3, dated 1/6/2020. It has been developed by Eugenio López-Cortegano, Diego Bersabé, Jinliang Wang, and Aurora García-Dorado. It is available from <https://www.ucm.es/genetical/mecanismos>

PURGd is a free software oriented to research, with non-commercial use, and it is distributed under the terms described in the PURGd License.txt file.

### **If you use PURGd in your research, cite:**

García-Dorado A, Wang J, López-Cortegano E (2016). Predictive model and software for inbreeding-purging analysis of pedigreed populations. *G3 (Bethesda)* **6**: 3593–3601.

Users are encouraged to request additional features on the software and to report bugs. In that case, please contact Eugenio López-Cortegano ([e.lopez@uvigo.es](mailto:e.lopez@uvigo.es)) or Aurora García-Dorado ([augardo@ucm.es](mailto:augardo@ucm.es)).

This work was funded by grant CGL2015-53274-P and by an FPI research fellowship (BES-2012-055006) from MINECO (Spanish Government).

## 7. References

- [1] Ballou JD (1997). Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. *J Hered* **88**: 169–178.
- [2] Boakes E, Wang J (2005). A simulation study on detecting purging of inbreeding depression in captive populations. *Genet Res* **86**: 139–148.
- [3] García-Dorado A (2012). Understanding and predicting the fitness decline of shrunk populations: inbreeding, purging, mutation, and standard selection. *Genetics* **190**: 1461–1476.
- [4] García-Dorado A, Wang J, López-Cortegano E (2016). Predictive model and software for inbreeding-purging analysis of pedigreed populations. *G3 (Bethesda)* **6**: 3593–3601.
- [5] López-Cortegano, E., Bersabé, D., Wang, J., & García-Dorado, A. (2018). Detection of genetic purging and predictive value of purging parameters estimated in pedigreed populations. *Heredity*, 121(1), 38-51.
- [6] Suwanlee S, Baumung R, Sölkner J, Curik I (2007). Evaluation of ancestral inbreeding coefficients: Ballou's formula versus gene dropping. *Conserv Genet* **8**: 489–495.