PURGd

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

Version 2.0

User's guide

Eugenio López-Cortegano

Diego Bersabé

Jinliang Wang

Aurora García-Dorado

Table of Contents

1. PURGd 2.0

PURGd is a software developed to detect purging and to estimate inbreeding-purging (IP) genetic parameters in pedigreed populations. The models and methods used in this software are described in García-Dorado *et al*. (2016) [4].

The main objective of this program is to estimate the effective purging coefficient (d_e, hereafter referred to as `d` for simplicity) described by García-Dorado [3], which is 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. Furthermore, the program estimates the regression coefficients on the purged inbreeding coefficient (g), here denoted `g(b)`, and on additional regressor variables, such as environmental factors or maternal inbreeding. This software also includes options to estimate parameters for purging models based on ancestral inbreeding, developed by Ballou [1] and Boakes & Wang [2].

Two alternative approaches are currently implemented:

1. **Linear regression method (LR)**: A Least Square (LS) `d` estimate is numerically searched. In this process, for each `d` value considered, a linear regression model is fitted for log-transformed fitness. When using this method, `-g(b)` gives an overestimate of the inbreeding load [4]. This approach cannot use data with fitness values less than or equal to zero.
2. **Numerical non-linear regression method (NNLR)**: The non-linear model for untransformed fitness is explored by numerically searching for the joint LS estimates of `d` and of the non-linear regression coefficients. In this method, `-g(b)` estimates the inbreeding load under the IP model [4].

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.

2. 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.

- **GNU/Linux**: An executable binary file of the program (`PURGd`) can be found in the `bin` folder. No installation is needed.
- **Microsoft Windows**: An executable binary file of the program (`PURGd.exe`) can be found in the `bin` folder. No installation is needed. See BOX 1 in Section 5 for a step-by-step explanation on how to run PURGd on Windows.

3. Program folders

Both the GNU/Linux and the Microsoft Windows versions of this software are distributed in a package that includes two folders. This is a quick overview on their content. More detailed information about each folder will be found in the following sections.

- `bin`: executable binary and settings files
- `input`: pedigree files and setfiles

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.

4. Input files

This program works with two kinds of input files: single pedigree files or setfiles. These files can be placed in the existing `input` folder for convenience, although this is not mandatory.

4.1. Single pedigree files

Files containing pedigree information must be in comma-separated values format (`.csv`) and have at least four columns, with the following precise order: identity of the individual (ID), mother ID, father ID, and the evaluation of fitness or of a fitness component trait. Fitness must be recorder using natural untransformed fitness numeric values (i.e., non-logarithmic), although it may be convenient scaling it to avoid large values (see --max-w0=NUM in section 5.2.1). However, 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 order 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 include a record for each individual (with its identity in the first column), including unrelated founders. We advise coding the identity of the parents of 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 may distort the estimates of inbreeding purging parameters.

Qilin.csv

```
ID,Dam,Sire,Longevity,YOB
1,0,0,1942,273
2,0,0,2106,273
3,0,0,2781,273
4,2,1,2051,275
5,3,1,2593,275
6,0,0,2399,273
7,2,1,4717,276
8,2,1,757,276
9,3,1,919,276
10,4,1,2655,277
11,4,1,2,277
12,5,1,518,277
13,5,1,422,277
14,5,1,1700,277
```

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

4.2. Setfiles

A setfile contains a list with the names of the pedigree files to be analysed. They are used to automatically process several pedigree files under the same running conditions (i.e., using the same parameters for the analysis). The present version of the program developed to run in Microsoft Windows has not been enabled to run using setfiles.

These setfiles must include a list of pedigree filenames (without the `.csv` extension), with one name per line. If the setfile and the respective single pedigree files are located in different directories, each pedigree filename must also be preceded by its path. It is encouraged to keep the setfile and the pedigree files in the same directory. See Figure 4.2 as an example.

The PURGd package includes several examples of single pedigree files and setfiles that can be found in the input folder. They can be checked as a reference in case of doubt.

Figure 4.2: Left: A setfile listing a series of pedigree files (`Qilin.csv`, `Phoenix.csv`, etc.) that are placed in the same folder as the setfile and will be analysed using the same running parameters; Right: The same setfile is shown using a spreadsheet program.

5. Running PURGd

PURGd must be run from the terminal (GNU/Linux) or the command prompt (Microsoft Windows; see BOX 1 for a step-by-step guide to run PURGd on Windows). It must be run from the `bin` folder described above and uses the following syntax:

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

`Options` are not required to run the program but, if specified, they will take preference over the default settings (see Sections 5.1 and 5.2 for more details).

Options can also be specified in a settings file, which requires using the option `--config=settings.txt,` as explained in Section 5.3. In this case, 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. This input file can be a single pedigree file or a setfile (Section 4).

By entering:

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

the program will print a short manual to use it as a quick reference guide.

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

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

In this example, the program will use the NNLR method assuming the IP model, which are the default options, to analyse 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.

To use Ballou's model instead of the default IP model, the example above has to be slightly modified by specifying the `--ffa` argument in the command line:

```
./PURGd --ffa ../input/MSD1.csv
```

Now, PURGd will use the NNLR method for the analysis of the file `MSD1.csv`, but assuming Ballou's model instead of the IP one.

In the same way, in order to run the LR method assuming Ballou's model on the same input file, PURGd should be invoked as:

```
./PURGd --lr --ffa ../input/MSD1.csv
```

Check the following subsections to learn more about all the available settings and how to modify them to customise the analysis.

Once the analysis has finished, the software will print a short message in the terminal. The corresponding output files (see Section 6) 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, in the `input` folder. Let's assume that the name of your input file is `MSD1.csv`.

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 default IP model and the NNLR method (default options), you should type:

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

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

If, for example, you want to use Ballou's model instead of the IP one you will need to set that option by entering the `--ffa` argument when you run the program:

```
./PURGd --ffa ../input/MSD1.csv
```

Note: When typing the path to the input file, you may need try 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 "`..`").

5.1. Program options

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. In following subsections, it is explained how to change these options either from the command line (Section 5.2) or from the settings file (Section 5.3).

- **Method**: The statistical method used to estimate the parameters of the model. It can be the linear regression (LR) method, where the dependent variable is the natural logarithm of the fitness trait, or the numerical non-linear regression (NNLR) method. Default method is NNLR, as it provides an unbiased estimate of the inbreeding load.
- **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**: Compute ancestral inbreeding coefficients by using the gene dropping simulation process described in Suwanlee *et al*. (2007) [5]. By default, gene dropping simulations are disabled and expected ancestral inbreeding values are computed using Ballou's equation [1].
- **Initial average fitness**: A value for this parameter can be set by the user if there is information available. By default, it is computed as the average fitness of non-inbred individuals with non-inbred ancestors ($F = F_a = 0$). Alternatively, if the estimation method is set to NNLR, the initial average fitness can be numerically estimated as the remaining parameter of the model. Note that this last option can lead to some overfitting of the model and to some downward bias in the estimates of the inbreeding load and d, and is only recommended when the default setting produces average fitness estimates with too high sampling error.
- **Inbreeding load**: PURGd estimates `g(b)`, which under the NNLR method is an estimate of the inbreeding load (here `delta`) multiplied by (-1) (*i.e*., `delta` = - `g(b)`). Under the LR method, `-g(b)` is an overestimate of `delta` [4]. By default, `-g(b)` is estimated at the same time as the remaining parameters in the corresponding model.

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

Under the NNLR method, a `delta` value (for example, an estimate previously obtained using the option `--delta=NUM`) can settled by the user to be assigned to –`g(b)`. This option is also available for the LR method but in this case the present version of the program does not run or gives unreliable results.

- **Use of maternal effects**: Maternal effects can be used in the analysis by incorporating the inbreeding coefficient of the dams as an additional independent variable. 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. Their effect will be estimated as additional regression coefficients. By default, none is used.
- **Output**: Change 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 log file**: Save a log file that allows the user to keep track of the settings used for a given analysis. By default, no logs are saved.

Save a database: Save a database containing both the fitness and the inbreeding coefficients of the analysed individuals. By default, no databases are saved.

- **Accuracy for the search of the purging coefficient**: This parameter settles the width of each increase of the purging coefficient `d` during the search of its least square estimate. It can be modified if the user. By default, 0.01 is used.
- **Seed**: A seed is required to generate pseudorandom numbers during the analysis. It can be convenient to record the seed value in order to replicate the results obtained, especially for the NNLR method. Default seed is the current time.
- **Verbose mode**: Print a short summary in the terminal during program execution for each pedigree that is being analysed. It is disabled by default.

5.1.1. Options exclusive to the NNLR method

- **Number of runs**: Number of times to run the ABC algorithm of the NNLR method for each pedigree file. Results are averaged. 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**: Values of the initial average fitness can be explored from 0 to a maximum value when this parameter is being estimated. By default, 1.0 is the maximum allowed. Large maximum values for initial fitness can slow the program. Therefore, it is convenient to scale fitness by a constant factor to reduce maximum values.
- **Maximum value of the inbreeding depression rate**: Values of the inbreeding depression rate can be explored from 0 to a maximum value. By default, 10 is the maximum rate allowed.
- **Minimum and maximum values of the slope for other regression terms**: If additional factors are used, the values of their regression coefficients can be explored from a minimum to a maximum value. By default, this range is [−10,10].

5.2. Modifying PURGd settings from the command line

As mentioned in Section 5, PURGd default settings can be modified from the command line when the corresponding option is preceded by a double hyphen (`--`). Note that any command-line option that changes the value of a settings parameter will take preference over its default value or over the one given in the settings file (Section 5.3). These options are:

`--nnlr` / `--lr` / `--d=NUM`: Set the method to NNLR, LR, or compute g for a given value `NUM` of d, which must be between 0 and 0.5

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

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

`--w0=NUM`: Set the initial average fitness to `NUM`. Here, `NUM` can be a number, the character `a`, to compute the average fitness from individuals with $F = F_a = 0$, or the

character `b`, to estimate it as the remaining parameter of the model (only available for the NNLR method).

`--maternal`: Activate the use of maternal effects

`--factor.cols=`: Activate the use of additional factors. The equal sign must be followed by one or more numbers, separated by commas. Each of these numbers must match the column number of an additional factor in the input file.

`--factor.names=`: Set the name of the additional 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.

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

`--save-log=FILE`: Save a log file with custom filename `FILE`. If no name is specified (i.e., `--save-log`), a default filename is used instead.

`--save-db`: Save a database for each pedigree file

`--accuracy=NUM`: Set the accuracy of the purging coefficient estimate to `NUM`, which must be a positive number up to 0.5

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

`--verbose`: Enable the verbose mode

5.2.1. Modifying settings exclusive to the NNLR method from the command line

`--nruns=INT`: Set 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`: Set the maximum value to search for the estimate of the initial average fitness to `NUM`. Its default value is 1. A new max-w0 needs to be established when the scale of measure for fitness allows for values larger than 1. Note than when the max-w0 value established is large, the searching process may became too slow. In this case it is better to scale fitness values in the input file to reduce max-w0 (for example by dividing fitness by its maximum value so that max-w0 = 1).

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

`--delta=NUM`: Here, `NUM` can be the `delta` value to be assigned to –`g(b)`. This option is also available under the LR method, where –`g(b)` gives an overestimate of `delta`, but it is not reliable in the present version of the program. `NUM` can also be the character `n`, in order to estimate `b(g)` along other purging parameters. Alternatively, the option `--delta=s` allows to use PURGd to obtain an estimate of `b(g)` using `d`=0 and individuals with $F_a = 0$.

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

Finally, remember that the option `--help` will print a short summary on how to use the program, including a list of the most common options.

5.3. Modifying PURGd settings from the settings file

Program settings can also be modified using a settings file. The distributed package contains a settings text file (`settings.txt`) in the `bin` folder that can be altered to set the options described in Section 5.1. It is strongly advised to save a "read-only" copy of this template file as a backup before doing any modifications to it.

In order to run PURGd using the settings included in the settings file, the `--config` option must be used. For example, by entering:

```
./PURGd ––config=settings.txt ../input/MSD1.csv
```

the program will analyse the pedigree file `MSD1.csv` using the settings provided by `settings.txt`, which is located in the `bin` folder. Note that the `--config` option can be used in combination with any of the available command-line options described in Section 5.2. In that case, the additional command-line argument will take priority over the corresponding setting defined in the settings file. For example, if we run the program by entering:

```
./PURGd ––config=settings.txt ––lr ../input/set_MSD1.csv
```

PURGd will use the LR method in the analysis, no matter what method is specified in `settings.txt`. All of the other options will be set as specified in the `settings.txt` file.

The template settings file `settings.txt` includes a brief description and a list of valid values for each available option. Options are named in capitals and followed by an equal sign. If a valid value is written after the equal sign, it will override the corresponding default value for that option. Otherwise, the default value will be used.

A comprehensive list of the current available settings and their corresponding valid values (given in brackets) is given below.

`METHOD=`: Set the method used to estimate the purging parameters: `[NNLR]` or `[LR]`. A third option, `[d <d-value>]`, computes *g* for a given value of *d*, which must be between 0 and 0.5. For example, `METHOD=d 0.15` computes *g* assuming `d` = 0.15.

`MODEL=`: Set the model used to predict the expected value of the fitness trait: inbreeding-purging `[IP]`, Ballou's `[BA]`, Boakes & Wang's `[BW]`, or Ballou-Boakes & Wang's mixed model `[MX]`

`GENEDROP=`: Set the number of iterations to run the gene dropping simulation. If left blank, gene dropping is not used and expected ancestral inbreeding values are computed using Ballou's equation [1].

`W0=`: Set the initial average fitness to a known value. Alternatively, using the character `[a]` allows to compute the initial average fitness from individuals with $F = F_a = 0$. A third option, using the character `[b]`, allows to estimate `W0` as the remaining parameter of the model (only available for the NNLR method).

`BG=`: Using the character `[n]` allows to estimate it `b(g)` at the same time as the remaining parameters of the model. Using the character `[s]`, allows to use PURGd to estimate `b(g)` by using `d`=0 and individuals with $F_a = 0$. Under the NNLR option, the inbreeding load (`delta`) can be provided by the user. Then, PURGd settles `b(g)` = `-delta`. This option is also available for the LR method but is not reliable at present.

`MATERNAL=`: Introduce maternal effects as a factor `[1]` or not `[0]`

`ADDITIONAL_FACTORS=`: Number(s) of the column(s) containing the additional factor(s) in the input file, separated by commas

`NAME_OF_ADDITIONAL_FACTORS=`: Name(s) of the additional factor(s) in the input file, separated by commas

`OUTPUT=`: Set the path to store the output files. It must be entered in double quotes (e.g., "`../my_output_folder`"). If left blank, output will be saved in `bin`.

`SAVE_LOG=`: Save a log file with a default `[1]` or custom `[FILE]` filename, or not `[0]`. Custom filenames cannot include the path character ("`/`") and must be entered in double quotes (e.g., "`my_log_file`").

`SAVE_DATABASES=`: Save databases `[1]` or not `[0]`

`ACCURACY=`: Set the accuracy of the purging coefficient estimate, which must be a positive number up to 0.5

`SEED=`: Set the seed value, which must be a positive integer. If left blank, the current time is used.

`VERBOSE=`: Enable the verbose mode `[1]` or not `[0]`

5.3.1. Modifying settings exclusive to the NNLR method from the settings file

`NRUNS=`: Set the number of runs of the ABC algorithm

`MAX_W0=`: Set the maximum value to search for the estimate of the initial average fitness

`MAX_BG=`: Set the maximum value to search for the estimate of the inbreeding depression rate

`RANGE_FACTORS=`: Set the minimum and maximum values to explore for the effects of additional factors. These two values must be separated by a comma.

6. 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: Each output file is named after its corresponding input file and the chosen model and method for the analysis. This means that any new analysis with the same input, 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 four kind of output files, which are described below.

6.1. Output files for estimates of the parameters of the model

Analysis performed using the LR or NNLR method will always save a file with the `_[model]_[method].csv` extension in the output folder. Two output sets are included in this file. The first set shows the pertinent results for the analysis performed considering purging, while the second one refers to an analogous analysis assuming no purging. Comparing these two analyses shows how far fitting improves by considering purging. When the option to estimate the inbreeding load using exclusively individuals with non-inbred ancestors is enabled, only results assuming no purging are displayed.

The standard output consists of the following columns:

- **Pedigree file**: The name of the input pedigree file
- **Analysis**: The name of the corresponding analysis performed for each input file. It can be "Inbreeding-purging model", "Ballou's model", "Boakes & Wang's model" or "Mixed model" for the analysis considering purging. 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 $d = 0$ in this column. In these cases, this value should be disregarded.
- **RSS**: The residual sum of squares
- **p-value (bootstrap)**: The *p*-value of a bootstrap analysis to test for the statistical significance of the estimate of d (or of the regression coefficients accounting for purging in F_a-models) against the value under the non-purging null hypothesis (i.e., zero). This *p*-value only appears in the row for the purging analysis.
- **p-value (F)**: The *p*-value for the Snedecor's *F*-test of significance for the linear regression model being fitted. Only displayed for the LR method.
- **aR2**: The adjusted coefficient of determination for the linear regression model being fitted. Only displayed for the LR method.
- **AICc**: The corrected Akaike's Information Criterion, assuming normality for residual errors
- **ln(W0)** or **W0**: The initial non-inbred mean for log-fitness or for untransformed fitness
- **b[factor]**: The value of the regression coefficient for each factor included in the analysis, including that for the purged inbreeding coefficient term *g* or those for the regressor variables defined in terms of F_a
- **SD[parameter]**: The standard deviation for the estimated parameters when available.
- **p-value (t)**: The *p*-value for the Student's *t*-test of significance for each regression coefficient in the linear regression model. Only displayed for the LR method.

When using a setfile as an input (only Linux version), the corresponding output will also store the output of every pedigree that was listed in the setfile. See Figure 6.1 below:

	A	B	C	D	E	F	G	H	I
1	Pedigree file	Analysis	d coefficient	RSS	AICc	p-value (bootstrap)	W0	SD(W0)	b(g)
2	Mirkwood_Spider	Inbreeding-purging model	0.0021	0.580193	-101623	0.5087	0.810258	0.0632063	-0.3526
3		No purging model	0	0.580044	-103857		0.810258	0.0632063	-0.340051
4	Cirith_Ungol_Spider	Inbreeding-purging model	0.11	1.37672	-49.7773	0.5273	0.846213	0.0275694	-0.75
5		No purging model	0	1.37663	-51999		0.846213	0.0275694	-0.74

Figure 6.1: An output file obtained by running PURGd to estimate the purging coefficient d using the IP model with the NNLR method on a setfile.

6.2. Output files with inbreeding coefficients computed using `d` values specified by the user

Files with the `_g(d).csv` extension in their filename will be stored in the output folder when the option to compute the purged inbreeding coefficient *g* for a given value of `d` is used (see Sections 5.2 and 5.3). A single output file is created for each pedigree and contains the following fixed number of columns for every individual:

- **ID**: Identity of the individual
- **F**: Standard inbreeding coefficient
- **g(d)**: Purged inbreeding coefficient, computed with the `d` value specified by the user
- **Accumulated purge**: Measured as the reduction of *g* for increasing values of *F*. It is computed as $1 - g/F$.
- **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.

6.3. Log files

A log file with the `_[model]_[method]_log.csv` extension in its filename will be saved in the output folder for both LR and NNLR estimation methods if the corresponding option is specified in the program settings. Alternatively, the log file can be named using a custom filename (see Sections 5.2 and 5.3). Log files follow the same format as settings files and include a list of the actual settings used in the analysis.

6.4. Databases

Databases with the `_data_[method].csv` extension in their filename will be saved in the output folder for both LR and NNLR estimation methods if the corresponding option is specified in the program settings (see Sections 5.2 and 5.3). A single separate database is created for each pedigree file analysed and includes the following columns:

- **ID**: Identity of the individual

- **W** or **ln(W)**: Fitness (NNLR method) or log-fitness (LR method) values, as used in the analysis
- **F**: Standard inbreeding coefficient
- **g(d)**: Purged inbreeding coefficient, computed with the estimate obtained for *d*
- **Accumulated purge**: Measured as the reduction of *g* for increasing values of *F*. It is computed as $1 - g/F$.
- **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 coefficients. These extra columns are:

- **gdam(d)**: Maternal purged inbreeding coefficient, computed with the estimate obtained for *d*
- **Effects of additional factors** in the input

Note that individuals excluded from the analysis (i.e., individuals with unknown fitness) will not appear in this output file.

7. Performance

PURGd is an efficient software that runs quickly and requires low memory usage, so it can be run in a desktop computer or a laptop. Details on a test of performance are shown below.

We measured the actual CPU execution time as the sum of the user and system time, which means that we do not show results for execution time in real (clock) time, as it can be affected by other concurrent processes, including input/output. This CPU time could be defined as the time used *within* the process. We also measured memory usage through the maximum resident set size memory consumed, which is the portion of main memory (RAM) employed by the process. These values were computed for simulated pedigrees with different number of individuals per generation (N) and different depth (t, in generations) using the linear regression (LR) and the numerical non-linear regression (NNLR) methods available in PURGd.

Figure 7 shows that CPU time increases linearly with the product $N\,t$, that is, with the total number of individuals in the input pedigree file, at a rate that is much higher for the NNLR method ($\sim 1.1 \times 10^{-3}$ seconds / individual) than for the LR method ($\sim 7.3 \times 10^{-4}$ seconds /individual). In practice, both methods are very fast: a large pedigree file of about 5000 individuals can be analysed in a few seconds using any of them.

The maximum resident set size memory used by the larger pedigree file ($N\,t = 5000$) were 37.80 and 26.76 Megabytes for the LR and NNLR methods, respectively. It can be concluded that no RAM problems are expected to happen when handling real pedigree files, which are usually smaller than these test files. Moreover, no memory leaks have been detected, so running pedigree files continuously using setfiles will not require any additional memory.

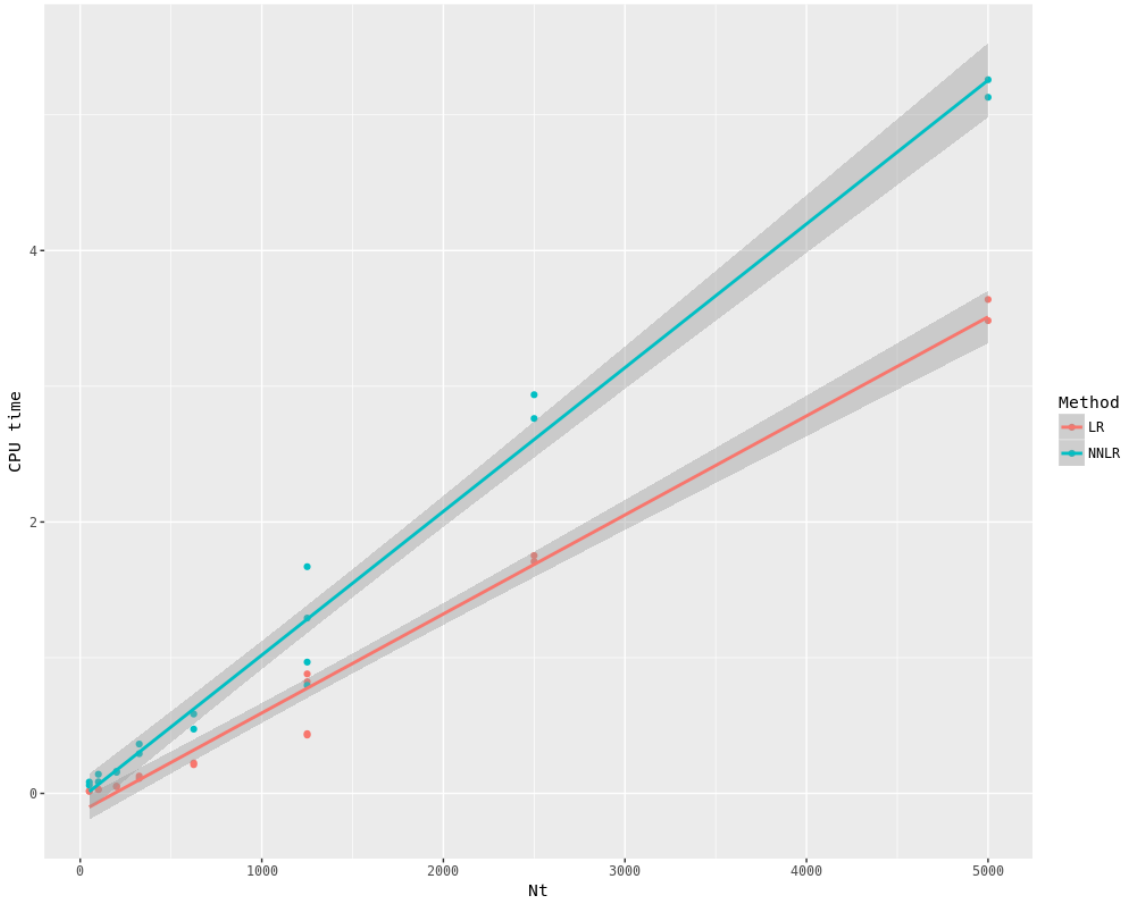

Figure 7: CPU time (in seconds) increase with the number of individuals in the pedigree ($N\,t$). The CPU time of the numerical non-linear regression method (NNLR, in blue) increases with $N\,t$ more steeply than that of the linear regression method (LR, in red).

8. 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 2.0, dated 12/01/2018. It has been developed by Eugenio López-Cortegano, Diego Bersabé, Jinliang Wang, and Aurora García-Dorado. It is available from https://www.ucm.es/genetica1/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), Aurora García-Dorado (augardo@ucm.es), or Diego Bersabé (diebersa@ucm.es).

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

9. 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] 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.