University of Zagreb   *University of Ljubljana*

MaGelLAn 1.0: a software to facilitate quantitative and population genetic analysis of maternal inheritance by combination of molecular and pedigree information

User's manual

Brajković Vladimir

**Thursday
10th July**

Literature

MaGelLAn Demo

Short communication

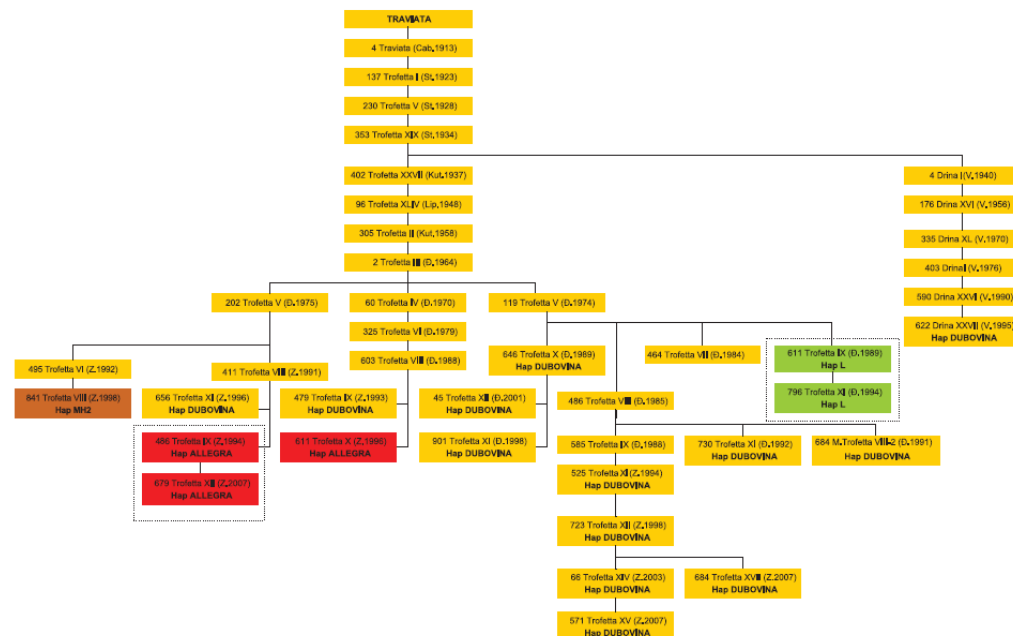Computational approach to utilisation of mitochondrial DNA in the verification of complex pedigree errors

Mato Čačić [a], Vlatka Cubric-Curik [b], Strahil Ristov [c,*], Ino Curik [b,*]

Fig. 2. Graphical illustration of the maternal lineage pedigree for the mare line *Traviata* with sequenced and a priori expected haplotype DUBOVINA (yellow), and non-concordant haplotypes ALLEGRA (red), L (green), and MH2 (brown). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Lipizzan horse pedigree

MaGelLAn 1.0

GSE Genetics Selection Evolution

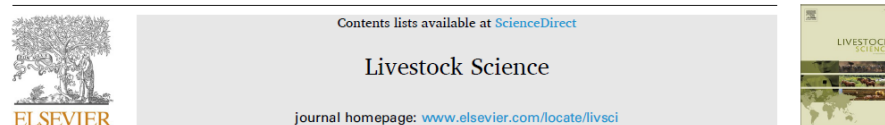SOFTWARE   Open Access

MaGelLAn 1.0: a software to facilitate quantitative and population genetic analysis of maternal inheritance by combination of molecular and pedigree information

Strahil Ristov[1*], Vladimir Brajkovic[2], Vlatka Cubric-Curik[2], Ivan Michieli[1] and Ino Curik[2]
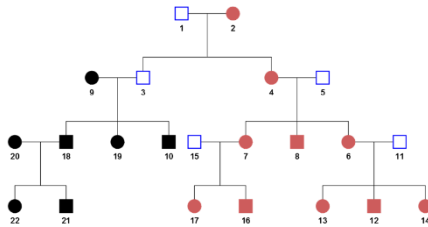
MaGelLAn 2.0/2.1

Mitogenome information in cattle breeding and conservation genetics: Developments and possibilities of the SNP chip

V. Brajkovic [a,1,*], D. Hršak [b,1,*], L. Bradić [a], K. Turkalj [a], D. Novosel [c], S. Ristov [b], P. Ajmone-Marsan [d], L. Colli [e], V. Cubric-Curik [a], J. Sölkner [f], I. Curik [a,1,*]

2

1. Introduction

MaGelLan 1.0

↓

maternal lineage analysis

↓

pedigree + **sequences (mitogenome)**

↓
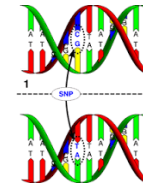
phenotypes + **polymorphism**

↓
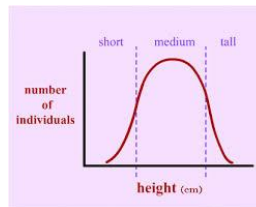
quanitative trait variation

↓

power of analaysis

2. Installation

- MaGelLAn 1.0 software

  - Project home page: http://lissp.irb.hr/software/magellan-1-0/

  - Article: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0242-9

  - Operating system(s): Platform independent

  - Programming language: **Python3**

  - Other requirements: None

  - License: The software is **free**

  - Any restrictions to use by non-academics: Written permission required

2. Installation MaGelLAn 1.0

Textpad 8.0

http://www.textpad.com/index.html

↓

Notepad++

https://notepad-plus-plus.org/

↓

R: The R Project for Statistical Computing

↓

Statistical Analysis Software, SAS

↓

Python 3.6.0

https://www.python.org/downloads/

↓

Command Prompt (cmd.exe)

↓

MaGelLAn 1.0

Preparing pedigree input file:
example.csv or example.txt

Install Python 3.6.0
☑ Add Python 3.6 to PATH

↓

```
C:\Users\Vladimir\Desktop\magellan.1.0\1. mag_verif>python mag_verif.py pedigree.txt
```
Run!

5
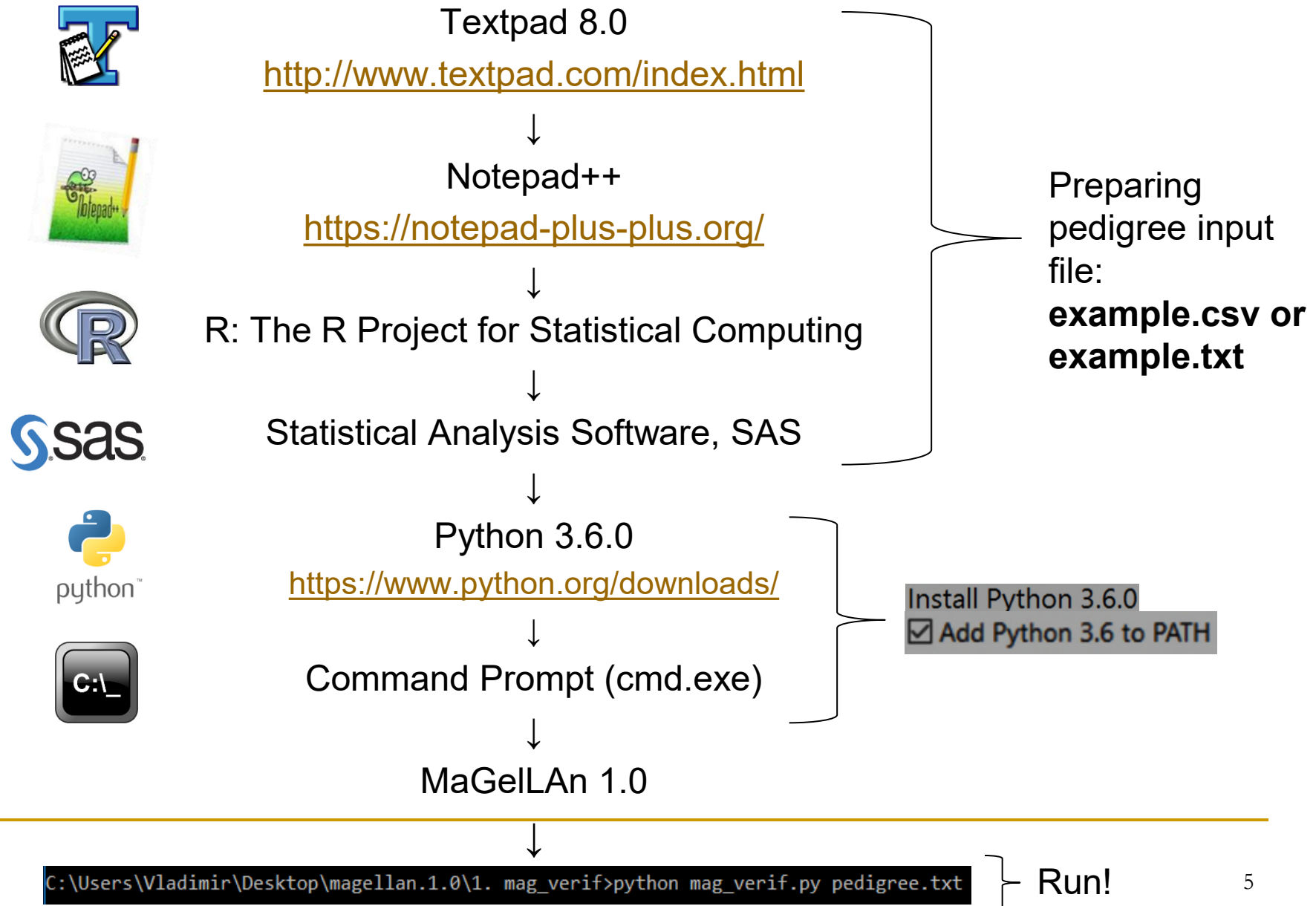
2. Installation MaGeILAn 2.0 – 2.1

7-Zip is a file archiver with a high compression ratio
https://www.7-zip.org/

TextPad - text editor
https://www.textpad.com/download#textpad9

Python 3.10.9 or some of the 3.10.x version — newer versions (3.11.x and above) are not compatible with the TensorFlow version listed below.

https://www.python.org/downloads/windows/
- when installing Python, mark "Add python.exe to Path"

Open command prompt in Windows or bash terminal in Linux/Mac/WSL:
pip install pandas
pip install dask
pip install dask[dataframe]
pip install PyQt6
pip install xgboost
pip install tensorflow #maybe
pip install matplotlib

2. Installation MaGelLAn 2.0 – 2.1

1.After installing packages, open CMD:
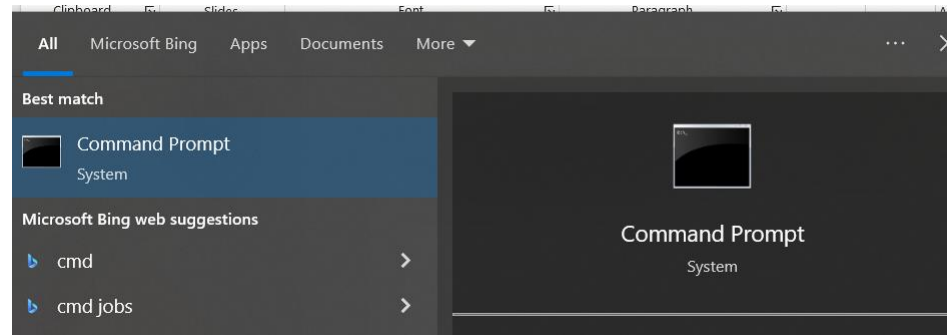
2. Copy address of Magellan folder as text:
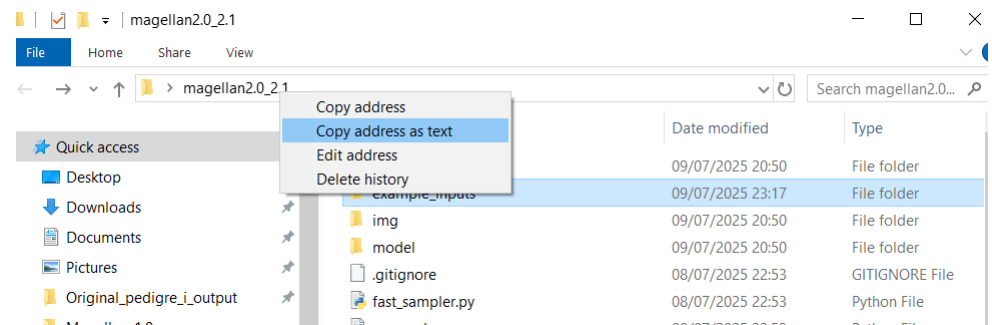
3. Type: cd C:\Users\Bella\Desktop\magellan2.0_2.1
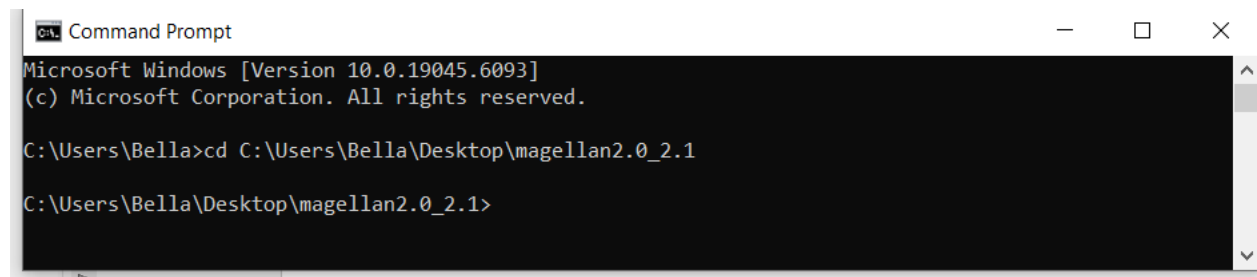
2. Installation MaGelLAn 2.0 – 2.1

3. Type *dir* to see the content of folder:

4. Run Magellan by typing: python mag_gui.py

3. Software functionalities

- **MaGelLan** – suite of four modules:

  - *mag_verif*

  - *mag_stat*

  - *mag_calc*

  - *mag_sampl*
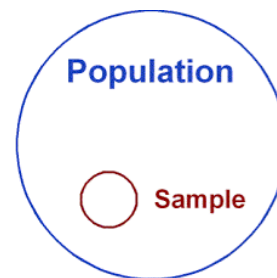
3. Software functionalities

Maternal lineages

↓

foun**ders** = 100

individu**als** = 5000

indivi**du**als** = 100

ML1   ML2   ML3
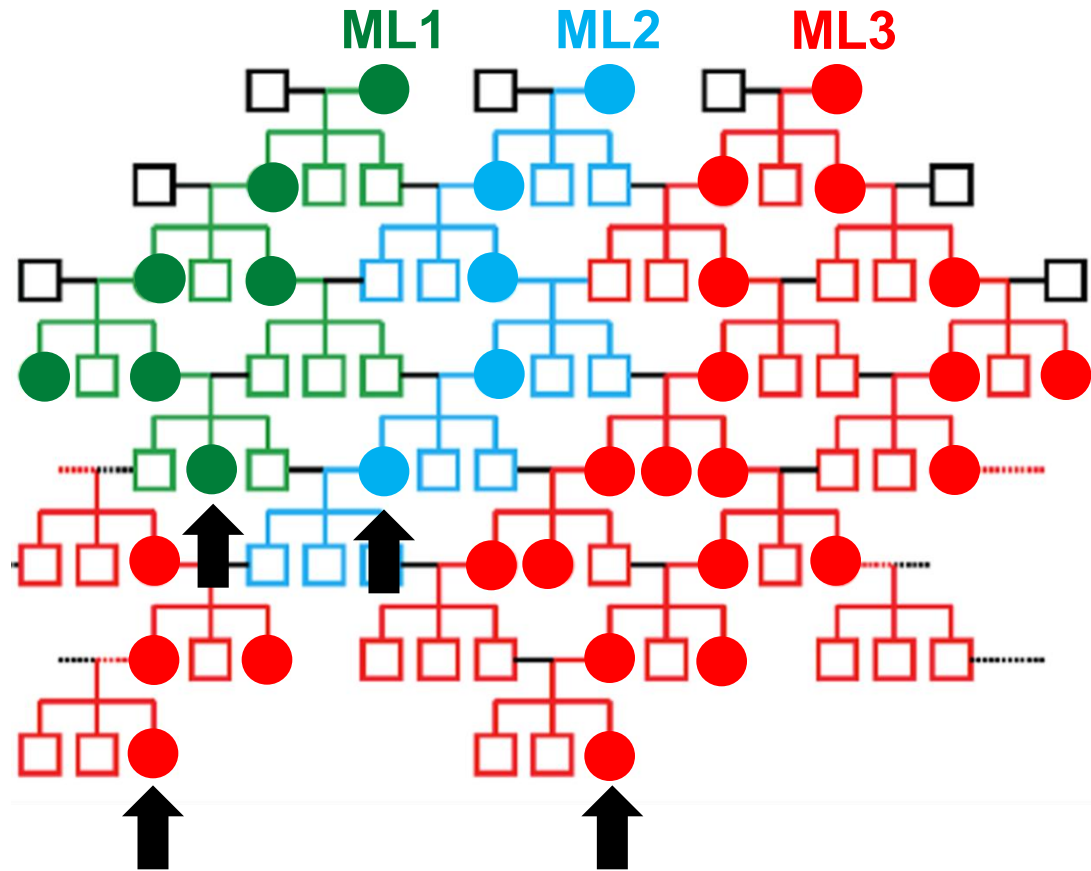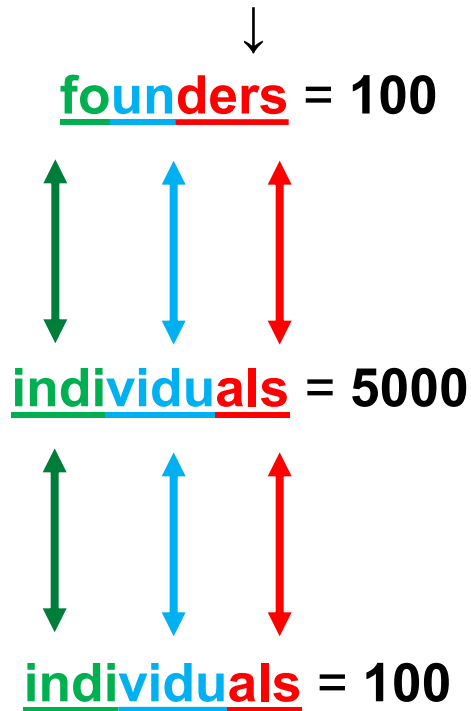
"Maternal inheritance without recombination"

MaGelLAn (Maternal Genealogy Lineage Analyser)

3.1. *mag_verif*

- **Verification** of the **pedigree correctness**

  - finds inconsistencies in female haplotype line

  - calculates haplotype error rates

  - search for standard pedigree errors:

  → 1. cycles in a pedigree – an individual repeated as its own parent

  → 2. gender inconsistencies

  → 3. when an individual is listed as a parent, but does not have its own record in the pedigree → non fatal error → „autocorrectional_log.txt"

- **Output files** are:

  - OutputVerif_Summary.txt

  - OutputVerif_ConflictingIndividuals.txt

  - OutputVerif_MisplacedBranches.txt

  - ERROR_ALERT.TXT

OutputVerif_Summary.txt

```
No. of individuals in pedigree                          4413
No. of distinct haplotypes                              0
No. of individuals with sequenced haplotype             0
No. of pairwise mismatches                              0
Maximal no. of conflicts / individual                   0 / NA
No. of informative individuals                          0
No. of individuals participating in conflicts           0
No. of conflicting or misplaced individuals             0
Pruned no. of conflicting individuals                   0
----------------------------------------------------------
There are no conflicts in the pedigree.
```

3.2. *mag_stat*

- Outputs various useful distributions of individuals over maternal pedigree lines.

- In particular, lists all individuals belonging to a given **founder dam line** with the **corresponding haplotype**.

- Imputation of a haplotype to each individual in pedigree (**connection between phenotype and mitochondrial haplotype**)

- **Output files** are:
  - OutputStat_DamLineMembership_1.txt
  - OutputStat_DamLineMembership_2.txt
  - OutputStat_DamLineMembershipAllInRefPop.txt
  - OutputStat_DamLineMembershipFemaleOnlyInRefPop.txt
  - OutputStat_DamLinesWithFemalesInRefPop.txt
  - OutputStat_DamLinesWithOnlyMalesInRefPop.txt

OutputStat_DamLineMembership_1.txt

```
founder dam: 040000093111973        number of individuals in dam line = 8        haplotype = N/A
040000606937973
040000340162842
040000497533142
040000468688573
040000254837173
040000093111973
040000398973273
040000529104773
founder dam: 040000087046873        number of individuals in dam line = 7        haplotype = N/A
040000087046873
040000040417773
040000328567273
040000560171473
040000413949473
040000450971973
040000352568373
```

OutputStat_DamLineMembership_2.txt

```
founder dam:individual in dam line
840000000003841:840000000055163
840000000003841:840000000005502
840000000003841:840000000275936
840000000003841:840000000090827
040000571672743:040000694704845
040000571672743:040000758088634
040000571672743:040000345550211
040000571672743:040000620785742
040000571672743:040000305001244
840000000053381:840000000053381
840000000053381:040000566339973
840000000053381:840000000755396
840000000053381:040000331228362
840000000053381:040000569348773
```

OutputStat_DamLinesWithFemalesInRefPop.txt

```
276000970206864: no. of desc. in ref. pop. = 8
276000979313425: no. of desc. in ref. pop. = 3
756111900504092: no. of desc. in ref. pop. = 3
840000000011105: no. of desc. in ref. pop. = 4
756111900897729: no. of desc. in ref. pop. = 3
040000003577323: no. of desc. in ref. pop. = 3
040000053037464: no. of desc. in ref. pop. = 6
040000047670964: no. of desc. in ref. pop. = 6
756110154021003: no. of desc. in ref. pop. = 5
040000077586786: no. of desc. in ref. pop. = 2
040000025707749: no. of desc. in ref. pop. = 13
040000026691249: no. of desc. in ref. pop. = 4
840000000426072: no. of desc. in ref. pop. = 3
756110121091459: no. of desc. in ref. pop. = 4
```

3.3. *mag_calc*

- Calculates the **effective population** size separately for **founder dam lines**, **founder haplotype lines**, and **founder sire lines**.

- A meaningfull haplotype line N_e can be computed only if enough of the individuals are sampled.

- **Output file** is:
    - OutputCalc_InputAndResults.txt

OutputCalc_InputAndResults_1925_2002.txt

```
Records in the studbook = 4413
No. of individuals in the reference population (1925 - 2002) = 3764
No. of female individuals in the reference population = 2404
No. of founder dams = 881
No. of founder dam lines in reference population = 748
No. of founder dam lines including lines with only males in reference population = 870
No. of founder dam lines in reference population with samples = 0
No. of founder dam lines in reference population with only one sample = 0
Total number of samples (haplotyped individuals) = 0
No. of samples in reference population = 0:   dams = 0,  sires = 0
-------------------------------------------------------------|

Dam lines:

Probability of identity in founder dams = 0.0011350737797956867
Probability of identity of a dam line in reference population = 0.0019538982450214796
Increase in identity = 0.0008197549475726404
Effective dam line size = 1219.876748485727

Haplotype lines:

Probability of identity of a haplotype line in founder population = 0.0011350737797956858
Probability of identity of a haplotype line in reference population = 0.0019538982450214818
Increase in identity = 0.0008197549475726434
Effective haplotype line size = 1219.8767484857224
```

3.4. *mag_sampl*

- Accepts the optional **"available"** column with ('1' / anything) values.

- **Calculates** the **target number** of the **individuals** per dam line **for sequencing** within given **planned number of sequencing**.

- Restricts the target numbers to **available individuals** if available data is present.

- **Selects** the **candidates** that provide the **highest potential** for **haplotype diversity** within dam line.

- **Output files** are:
    - OutputSampl_IndividualsForSampling.txt
    - OutputSampl_DetailedInfo.txt
    - OutputSampl_AvailabilityRestrictions.txt

OutputSampl_DetailedInfo.txt

```
reference population size = 3764
dam lines in ref.pop. = 870
IndividualsMissingFounderDamInRefPopCount = 241
PreviouslySequencedInRefPop = 0
HowManyToSequence = 50
Number of lines without previous samples = 870
TargetSum (that remains to be proportionally divided among dam lines) = 0

FOUNDER DAM:
276000970206864 -> in ref.pop.: 8    targ: 0    prev: 0    todo:0
276000979313425 -> in ref.pop.: 3    targ: 0    prev: 0    todo:0
756111900504092 -> in ref.pop.: 3    targ: 0    prev: 0    todo:0
840000000011105 -> in ref.pop.: 4    targ: 0    prev: 0    todo:0
756111900897729 -> in ref.pop.: 3    targ: 0    prev: 0    todo:0
040000003577323 -> in ref.pop.: 3    targ: 0    prev: 0    todo:0
040000053037464 -> in ref.pop.: 6    targ: 0    prev: 0    todo:0
040000047670964 -> in ref.pop.: 6    targ: 0    prev: 0    todo:0
756110154021003 -> in ref.pop.: 5    targ: 0    prev: 0    todo:0
040000077586786 -> in ref.pop.: 2    targ: 0    prev: 0    todo:0
040000025707749 -> in ref.pop.: 13    targ: 0    prev: 0    todo:1
040000026691249 -> in ref.pop.: 4    targ: 0    prev: 0    todo:0
840000000426072 -> in ref.pop.: 3    targ: 0    prev: 0    todo:0
756110121091459 -> in ref.pop.: 4    targ: 0    prev: 0    todo:0
```

OutputSampl_IndividualsForSampling.txt

```
founder dam 040000025707749 : 040000567134473,

founder dam 040000024292449 : 040000647564843,

founder dam 040000062987186 : 040000199321586,

founder dam 040000002556523 : 040000251634126,

founder dam 840000000040077 : 840000000181926,

founder dam 840000000089500 : 840000000853876,

founder dam 040000002905723 : 040000340174334,

founder dam 040000052761964 : 040000143225961,

founder dam 040000432933532 : 040000243952345,
```

4. Example of the input pedigree file

- The input file must be in CSV (**comma-separated values**) → .csv or .txt
- The position of the information in columns must match the position of the corresponding keyword in the header!

x, extra information which I use to count number of individual
↑
header line → x,ID,father,mother,YOB,gender,haplotype,available,
1,Id100,0,0,2000,1,,0,
2,Id101,Id100,Id102,2005,2,hap1,1,
3,Id102,0,0,2001,2,hap1,0,

- **1 – male, 2 – female; 0 – missing parents;**

- **haplotype** - name tag of a haplotype; if present, haplotype data enables full functionality of all modules, however, modules are partially functional without it

- **available** - '1' for available, anything else for not available; exclusive use in module *mag_sampl*

5. Additional input files:

- **reference_years.txt**

    - stores the first and the last year of birth for the individuals included in the reference population

    - formatted as two numbers in two lines

    - used in modules *mag_ stat*, *mag_calc* and *mag_sampl*

- **planned_number_of_sequencings.txt**

    - stores the number of planned sequencings

    - formatted as one number in a single line

    - exclusive use in module *mag_sampl*

If any of the additional files is missing, the default values (coded in the scripts) are used.
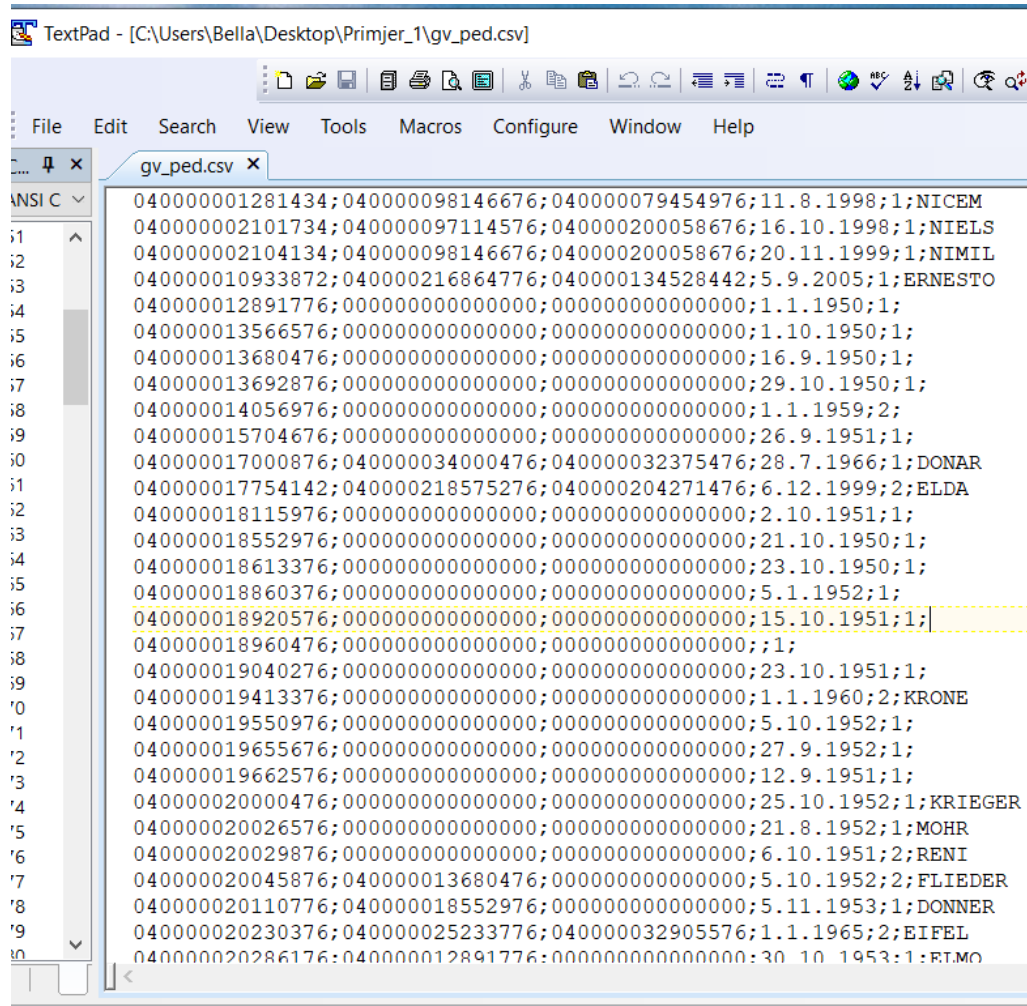
6. A typical workflow of the MaGelLan modules with a new pedigree
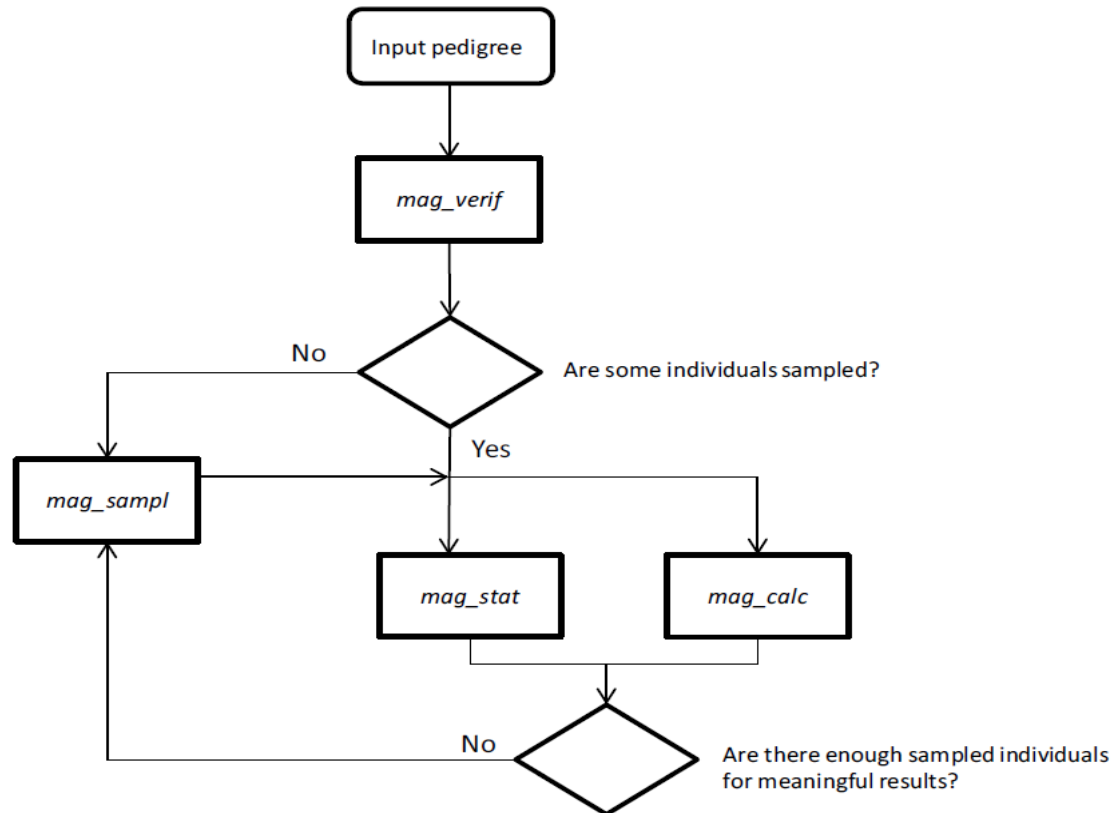
Each **module** performs the initial **verification** of pedigree correctness. In case of errors in the pedigree, module *mag_verif* outputs the description of errors, while other modules abort the execution and direct the user to module *mag_verif*.

Practical – preparation of pedigree for Magellan format

- **1. Open pedigree "gv_ped.csv" using Textpad or Notepad**

- **2. Copy (CTRL + A) whole pedigree and paste it into Excel**

Practical

- **3. Import pedigree into Excel (mark column A → Data → Text to clumns**

Practical

- **4. Set Delimiter ":" to separate columns**

Practical

- **5. Set first four columns to the Text, and last two to General**

Practical

- **6. Now we have to extract year of birth**

	A	B	C	D	E	F
1	040000001281434	040000098146676	040000079454976	11.8.1998	1	NICEM
2	040000002101734	040000097114576	040000200058676	16.10.1998	1	NIELS
3	040000002104134	040000098146676	040000200058676	20.11.1999	1	NIMIL
4	040000010933872	040000216864776	040000134528442	5.9.2005	1	ERNESTO
5	040000012891776	000000000000000	000000000000000	1.1.1950	1	
6	040000013566576	000000000000000	000000000000000	1.10.1950	1	
7	040000013680476	000000000000000	000000000000000	16.9.1950	1	
8	040000013692876	000000000000000	000000000000000	29.10.1950	1	
9	040000014056976	000000000000000	000000000000000	1.1.1959	2	
10	040000015704676	000000000000000	000000000000000	26.9.1951	1	
11	040000017000876	04000034000476	04000032375476	28.7.1966	1	DONAR
12	040000017754142	040000218575276	040000204271476	6.12.1999	2	ELDA
13	040000018115976	000000000000000	000000000000000	2.10.1951	1	
14	040000018552976	000000000000000	000000000000000	21.10.1950	1	
15	040000018613376	000000000000000	000000000000000	23.10.1950	1	
16	040000018860376	000000000000000	000000000000000	5.1.1952	1	
17	040000018920576	000000000000000	000000000000000	15.10.1951	1	
18	040000018960476	000000000000000	000000000000000		1	
19	040000019040276	000000000000000	000000000000000	23.10.1951	1	
20	040000019413376	000000000000000	000000000000000	1.1.1960	2	KRONE
21	040000019550976	000000000000000	000000000000000	5.10.1952	1	
22	040000019655676	000000000000000	000000000000000	27.9.1952	1	
23	040000019662576	000000000000000	000000000000000	12.9.1951	1	
24	040000020000476	000000000000000	000000000000000	25.10.1952	1	KRIEGER
25	040000020026576	000000000000000	000000000000000	21.8.1952	1	MOHR
26	040000020029876	000000000000000	000000000000000	6.10.1951	2	RENI

Sheet1 +

29

Practical

- **7. Add two/three empty columns on the right side of column D**

Practical

■ **8. Separate column D (mark column D → Data → Text to clumns → separator "." → next**

■

Practical

■ **9. Leave "General" for all three columns → Finish**

Practical

■ **10. Subset only these columns:**

	A	B	C	D	E
1	040000001281434	040000098146676	040000079454976	1998	1
2	040000002101734	040000097114576	040000200058676	1998	1
3	040000002104134	040000098146676	040000200058676	1999	1
4	040000010933872	040000216864776	040000134528442	2005	1
5	040000012891776	000000000000000	000000000000000	1950	1
6	040000013566576	000000000000000	000000000000000	1950	1
7	040000013680476	000000000000000	000000000000000	1950	1
8	040000013692876	000000000000000	000000000000000	1950	1
9	040000014056976	000000000000000	000000000000000	1959	2
10	040000015704676	000000000000000	000000000000000	1951	1
11	040000017000876	040000034000476	040000032375476	1966	1
12	040000017754142	040000218575276	040000204271476	1999	2
13	040000018115976	000000000000000	000000000000000	1951	1
14	040000018552976	000000000000000	000000000000000	1950	1
15	040000018613376	000000000000000	000000000000000	1950	1
16	040000018860376	000000000000000	000000000000000	1952	1
17	040000018920576	000000000000000	000000000000000	1951	1
18	040000018960476	000000000000000	000000000000000		1
19	040000019040276	000000000000000	000000000000000	1951	1
20	040000019413376	000000000000000	000000000000000	1960	2
21	040000019550976	000000000000000	000000000000000	1952	1
22	040000019655676	000000000000000	000000000000000	1952	1
23	040000019662576	000000000000000	000000000000000	1951	1
24	040000020000476	000000000000000	000000000000000	1952	1
25	040000020026576	000000000000000	000000000000000	1952	1
26	040000020029876	000000000000000	000000000000000	1951	2

Practical

- **11. Prepare input for Magellan format:**

x	ID	father	mother	YOB	gender	haplotype	available
1	040000001281434	040000098146676	040000079454976	1998	1		
2	040000002101734	040000097114576	040000200058676	1998	1		
3	040000002104134	040000098146676	040000200058676	1999	1		
4	040000010933872	040000216864776	040000134528442	2005	1		
5	040000012891776	000000000000000	000000000000000	1950	1		
6	040000013566576	000000000000000	000000000000000	1950	1		
7	040000013680476	000000000000000	000000000000000	1950	1		
8	040000013692876	000000000000000	000000000000000	1950	1		
9	040000014056976	000000000000000	000000000000000	1959	2		
10	040000015704676	000000000000000	000000000000000	1951	1		
11	040000017000876	040000034000476	040000032375476	1966	1		
12	040000017754142	040000218575276	040000204271476	1999	2		
13	040000018115976	000000000000000	000000000000000	1951	1		
14	040000018552976	000000000000000	000000000000000	1950	1		
15	04000018613376	000000000000000	000000000000000	1950	1		
16	040000018860376	000000000000000	000000000000000	1952	1		
17	040000018920576	000000000000000	000000000000000	1951	1		
18	04000018960476	000000000000000	000000000000000		1		
19	040000019040276	000000000000000	000000000000000	1951	1		
20	040000019413376	000000000000000	000000000000000	1960	2		
21	040000019550976	000000000000000	000000000000000	1952	1		
22	040000019655676	000000000000000	000000000000000	1952	1		
23	040000019662576	000000000000000	000000000000000	1951	1		
24	040000020000476	000000000000000	000000000000000	1952	1		
25	000020026576	000000000000000	000000000000000	1952	1		

Practical

- **12. Replace all individuals 000000000000000 with 0**

Practical

- **13. Filter all "Blank" YEAR OF BIRTH and replace by 0; sort YOB from smallest to largest, notice interval 1950 - 2006**

Practical

- **14. Mark field A1, then CTRL+A to mark all columns and copy pedigree in Texpad**

Practical

- **15. Mark and copy "invisible" delimiter and replace with ",", and save as "ped_1_magellan.txt**

Practical

- **16. Open with magellan 2.0/2.1 GUI**
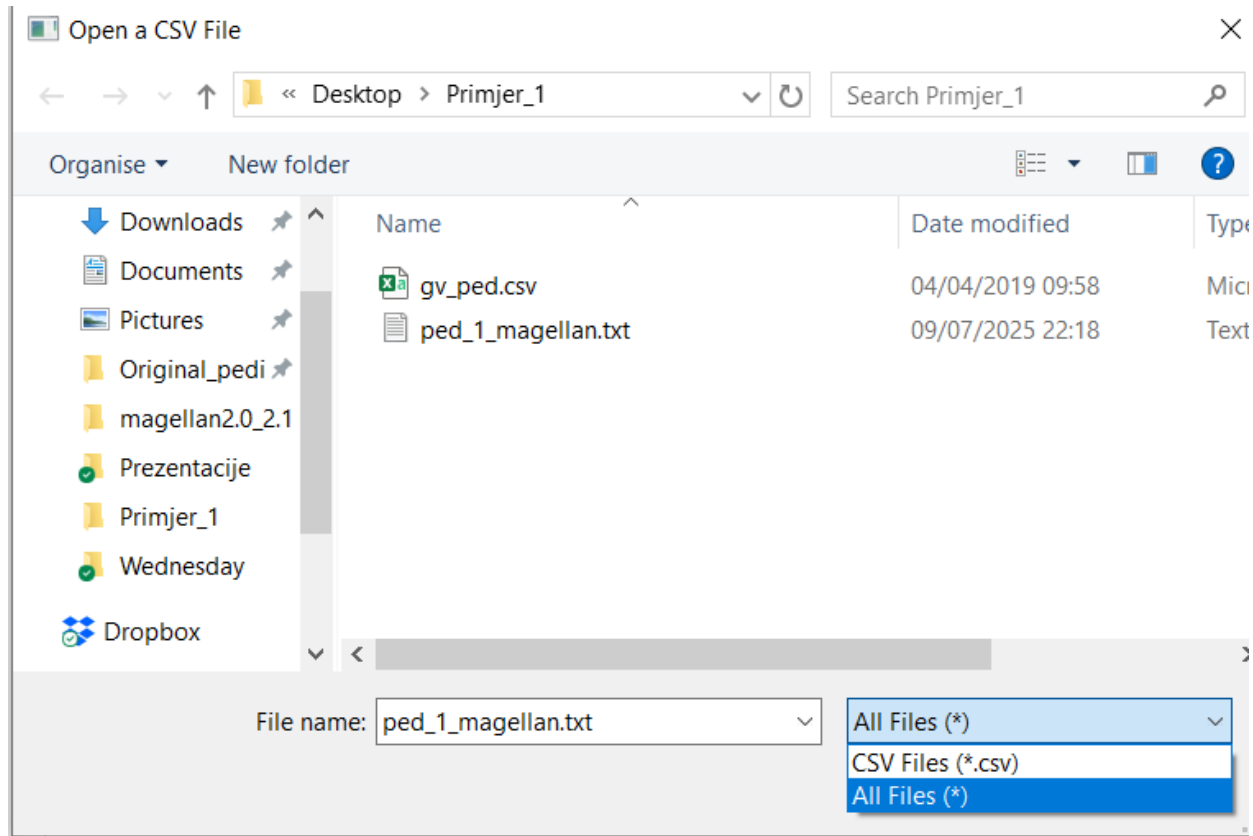
Practical

- **17. Notice error alert**
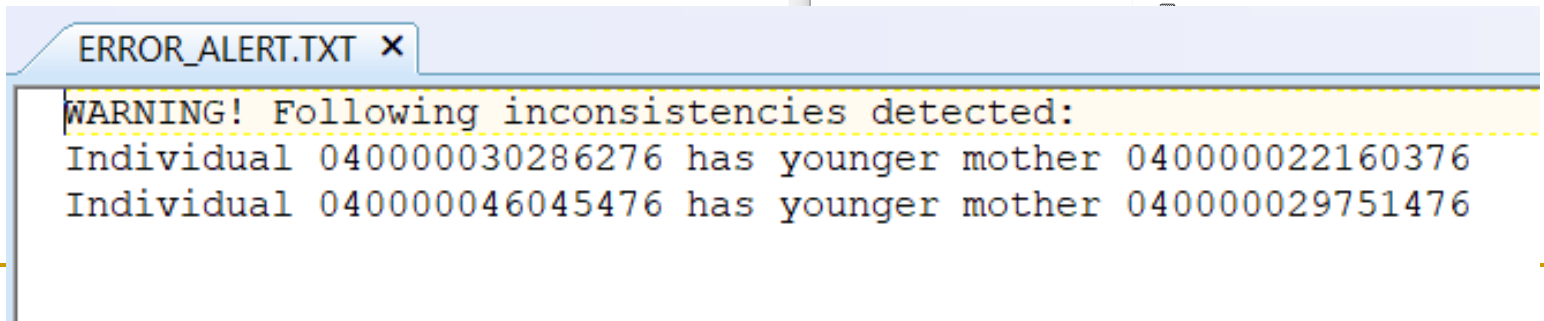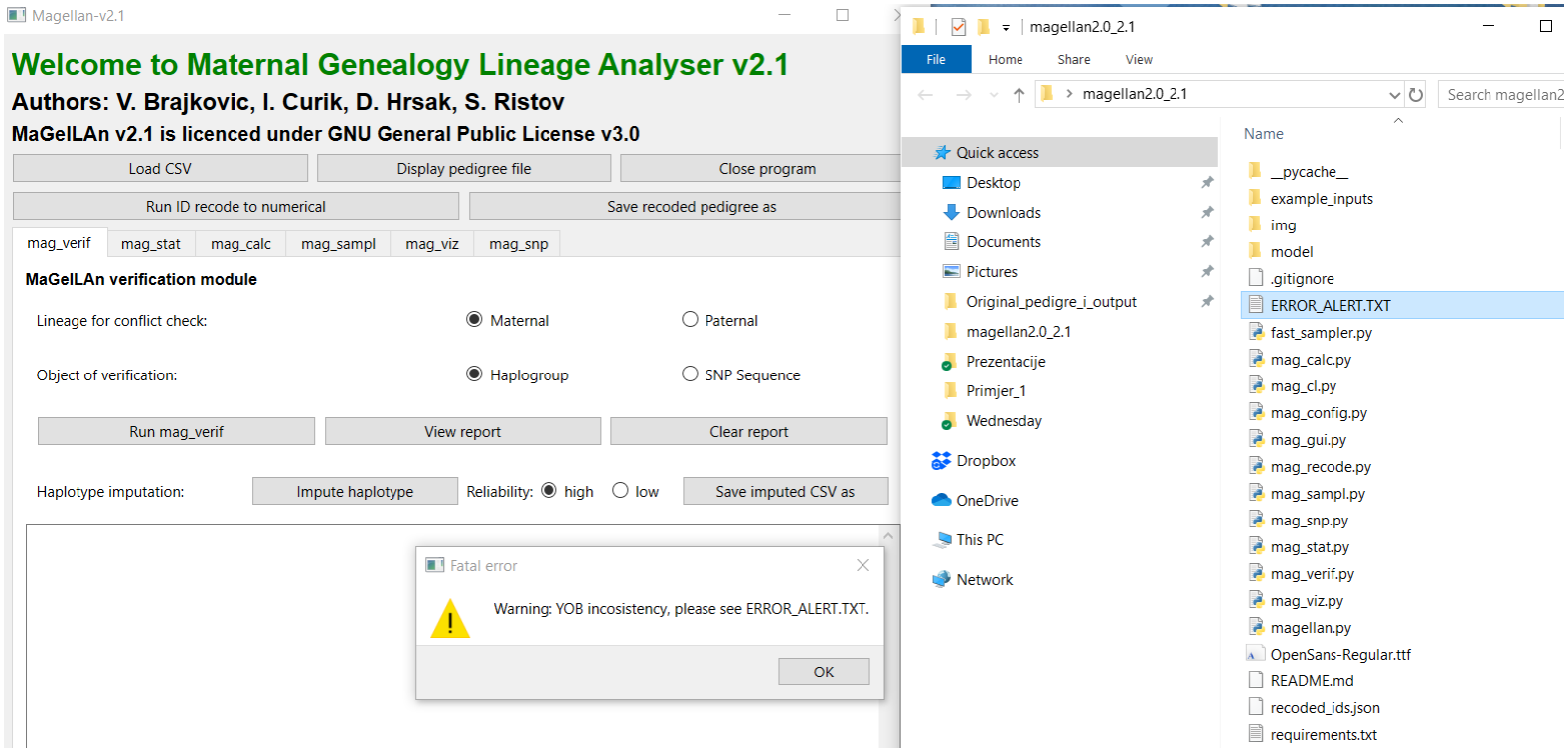
Practical

- **18. Notice error alert – replace year of birth of younger mothers with 0, search**

ERROR_ALERT.TXT ×

```
WARNING! Following inconsistencies detected:
Individual 040000030286276 has younger mother 040000022160376
Individual 040000046045476 has younger mother 040000029751476
```

x   ID   father   mother   YOB gender haplotype available

```
273,040000030286276,040000023500676,040000022160376,1959,2,,
275,040000030413576,040000024340576,0,1959,2,,
```

ID

mother

```
54,040000022030576,0,0,1960,2,,
56,040000022160376,0,0,1960,2,,
91,040000024519176,0,0,1960,2,,
```

mother

Practical

- **19. Load corrected pedigree, Run mag_verif, View report**

Practical

- **20. Display pedigree file**

Practical

- **21. Run ID to recode to numerical and Save recoded pedigree**

Practical

■ **22. Mag_stat module for the reference year 2000 - 2006**

Practical

- **23. Mag_calc module for the reference year 2000 - 2006**

Load CSV	Display pedigree file	Close program
Run ID recode to numerical		Save recoded pedigree as

mag_verif	mag_stat	mag_calc	mag_sampl	mag_viz	mag_snp

MaGelLAn calculation module

Start reference year: 2000   [Set]

End reference year: 2006   [Set]

Select lineage: ● Maternal   ○ Paternal

Run mag_calc	View report	Clear report

```
Records in the studbook = 1647
No. of individuals in the reference population (2000 - 2006) = 85
No. of female individuals in the reference population = 15
No. of founder dams = 270
No. of founder dam lines in reference population = 13
No. of founder dam lines including lines with only males in reference population = 40
No. of founder dam lines in reference population with samples = 0
No. of founder dam lines in reference population with only one sample = 0
Total number of samples (haplotyped individuals) = 0
No. of samples in reference population = 0:   dams = 0,  sires = 0
----------------------------------------------------------------

Dam lines:

Probability of identity in founder dams = 0.003703703703703704
Probability of identity of a dam line in reference population = 0.08444444444444445
Increase in identity = 0.08104089219330854
Effective dam line size = 12.339449541284406

Haplotype lines:
```

46

Practical

- **24. Mag_viz**