Bayesian networks for DNA-based kinship analysis: Functionality and validation of the GENis missing person identification module

Ariel Chernomoretz[*1,2], Franco Marsico[3], Javier Iserte[2], Mariana Herrera Piñero[4], Maria Soledad Escobar[5], Manuel Balparda[5], Gustavo Sibilla[5]

[*] corresponding author: ariel@df.uba.ar

[1]Phys. Department, School of Sciences, University of Buenos Aires/IFIBA CONICET Argentina, [2]Leloir Institute Foundation, [3]UNPAZ, [4]Banco Nacional de Datos Geneticos, [5]Fundación Sadosky

GENis is a recently published open-source multi-tier information system developed to run forensic DNA databases. It relies on a Bayesian Networks framework and it is particularly well suited to efficiently perform large-size queries against databases of missing individuals. In this contribution we present a validation of the missing person identification capabilities of GENis. To that end we introduce *fbnet*, a free-software package written in the R statistical language that implements the complete GENis functionality to perform kinship analysis based on DNA profiles. With the aid of *fbnet*, we could validate likelihood ratios against estimations draw with *Familias* and *forrel* (two well-recognized R packages for kinship quantification) for complex pedigrees provided by the Argentinian reference databank (Banco Nacional de Datos Geneticos, BNDG). We found that our methodological approach presented an excellent performance in terms of accuracy and computation times.

INTRODUCTION

GENis is a highly customizable system composed of three different modules related to: (a) person identification and analysis of forensic evidence, (b) missing person identification, and (c) disaster victim identification [1]. The missing person identification (MPI) module was specifically developed to run automatic queries on family and missing person databases. For each family, the pedigree structure and available genotypes are integrated into a Bayesian network (BN). These modeling tools serve to represent joint probability distributions in a compact and efficient way, explicitly taking into account the statistical independence between random variables [2]. In our approach to the kinship analysis problem we used BN to infer genotype probability tables for the queried missing person (MP). The availability of these probability tables allows for subsequent rapid likelihood estimations for large MP databases.

METHODS

We implemented the complete GENis functionality to perform kinship analysis in *fbnet*, an open-source package written in the R statistical language, freely available from CRAN [3]. To speed-up calculations GENis and *fbnet* implemented a truncated version of the canonical stepwise mutational model where probabilities for rare mutations, i.e. more than a given number (L) of steps away from the original allelic value (i.e. the diagonal term), are neglected.

Throughout this contribution we used R packages *Familias*-v.2.4 [4] and *forrel*-v1.3 [5] in order to get reference values against which our LR estimations could be compared. Molecular markers considered for LR calculations are summarized in Sup Table 1.

RESULTS

Precision of the truncated step-wise mutational model

In order to assess for the accuracy of the truncated stepwise mutational model, we considered L=1, 2 and 4 and estimated with *fbnet* LR values for 1000 simulated profiles (marker setA in Sup Table 1) of unidentified persons (UPs). We considered ensembles for alternative scenarios, H1: UP is MP, and H0: UP is not MP. In Fig 1.A we showed the precision of these estimations, at a given tolerance level, compared against the software *Familia*s LR values for familyFF (Sup Fig 1.C). It can be appreciated that while L=1 is a rather rough approximation, LR estimations very rapidly level off at high precision levels for L=4. For instance, 97.7% and 99.6% agreement at a 10% tolerance level between *fbnet* and *Familias* LR estimations were observed for H0 and H1 simulated UP's respectively.

Speed-up factor of the bayesian network approach

LR estimations for large databases can be done very efficiently within the GENis/*fbnet* bayesian network approach, as the MP genotype probabilities, conditioned by the available evidence and pedigree relationships, are estimated only once for each family. In this way, *fbnet* presented near constant LR computational times for ensembles of simulated UP profiles. On the contrary, running times for other kinship analysis software, like *Familias* or *forrel* linearly increased with the size of the simulated ensemble. This behavior could be verified in Figure 1.B, where we showed *fbnet* and *forrel* LR computational times for 1000 UP simulations for families FD, FF and FJ (marker setB in Sup Table 1, pedigrees in Supplementary Figure 1). Noticeably, large speed-up factors (146x, 37x and 227x for families FD, FF and FJ respectively) were obtained using *fbnet* for simulations involving 1000 UPs (cpu: ryzen 5 2600x 3.6Ghz, 32 gb ram).

Systematic comparisons of LR values

We considered 24 familial pedigrees from BNDG and randomly generated the genotype of available family contributors (marker setB in Sup Table 1, pedigrees in Supplementary Figure 2). For each pedigree we conditionally simulated 1000 MP genotypes and compared LR values against *forrel* estimations considering a stepwise mutational model (we used the truncated L=4 model for *fbnet*). Results were included in Sup. Table 1, where we reported the correlation of log(LR) *fbnet* and *forrel* estimations, the mean and maximum observed percentage differences, and the number of discrepancies at a given tolerance level. An overall high accordance level between reference and *fbnet* estimated values could be observed. In particular, almost perfect correlation values were reported for every analyzed pedigree and mean errors were well below 0.2% (they remained lesser than 0.1% for the majority of the analyzed families). In addition, through the 28000 sampled genotypes only 16 profiles (i.e. less than 0.06% of cases) presented a discrepancy greater than 5%.

DISCUSSION AND CONCLUSIONS

Overall we found an excellent agreement between GENis/*fbnet* estimations and the corresponding reference LR values for complex pedigrees. In particular we showed that the truncated stepwise mutational model with L=4

produced highly accurate LR estimations.

Particularly interesting for DNA database applications, our Bayesian network approach to the kinship analysis problem provided an extremely efficient methodology to evaluate LRs for large MP databases. This is so because MP genotype probabilities are estimated only once, at pedigree creation time. Of course, it could be eventually the case that MP genotypes presented rare alleles to be considered for LR estimations. For these uncommon cases a re-calculation step of the probabilities should be executed to accommodate the allelic probability value of the new rare allele. Still, an overall net speed-up gain is expected to occur as these uncommon circumstances would become rarer as the size of the database increases.

Last but not least, in this contribution we introduced the R open-source package *fbnet* in order to share with the community the functionality and main design principles behind GENis MPI module.

REFERENCES

[1] “GENis, an open-source multi-tier forensic DNA information system”, Chernomoretz et al, Forensic Science International:Reports (2020) http://doi.org/10.1016/j.fsir.2020.100132

[2] “Modeling and reasoning with bayesian networks”, Darwiche A., Cambridge University Press (2009)

[3] https://CRAN.R-project.org/package=fbnet

[4] “Relationship Inference with Familias and R”, Thore Egeland, Daniel Kling, Petter Mostad, Elsevier 2015.

[5] https://CRAN.R-project.org/package=forrel

FIGURES

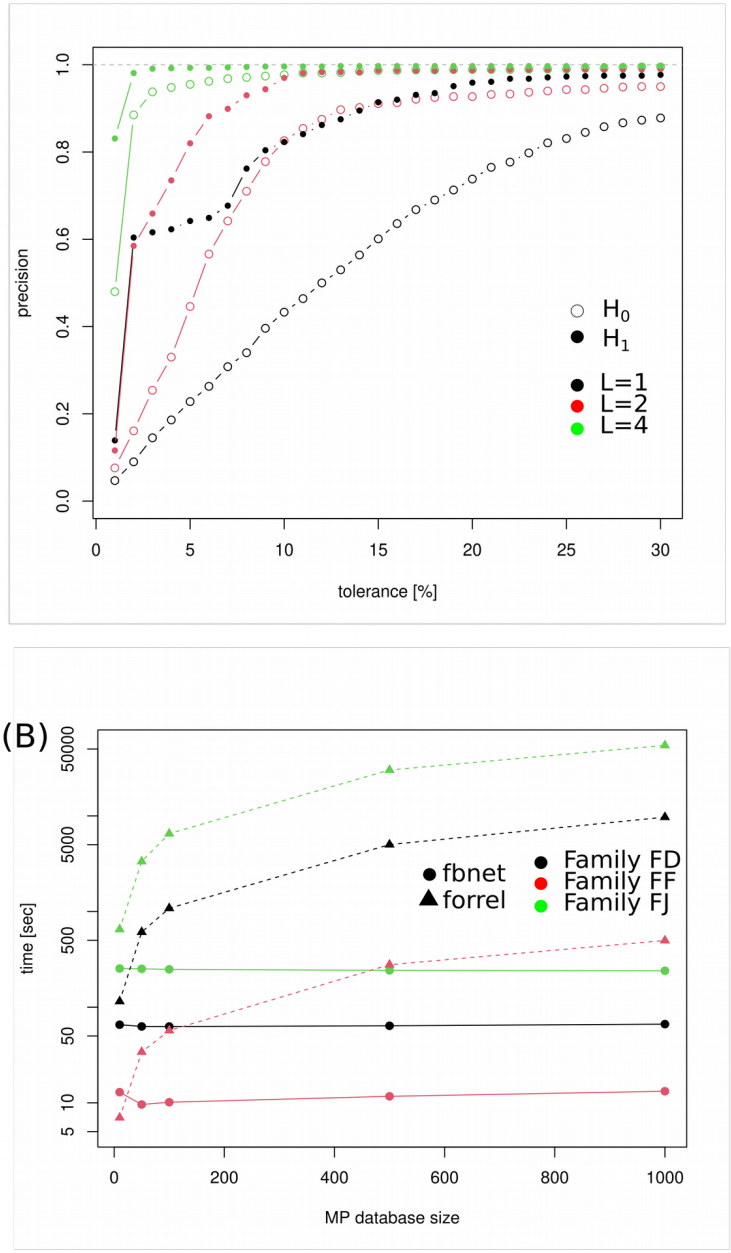


Figure 1. Panel A. Precision of *fbnet* LR estimation (i.e. fraction of acceptable estimations) for Family FJ (see SupFig 1) as a function of the considered tolerance percentage. Values for L=1, 2 and 4 truncated stepwise mutational model are displayed with black, red and green symbols respectively. Solid and empty symbols were used for ensembles of simulated genotypes generated under H1 and H0 hypothesis respectively. Panel B: Running times for *fbnet* and *forrel* LR estimation as a function of the considered ensemble size.

Supplementary Material

Bayesian networks for DNA-based kinship analysis: Functionality and validation of the GENis missing person identification module

Ariel Chernomoretz[*1,2], Franco Marsico[3], Javier Iserte[2], Mariana Herrera Piñero[4], Maria Soledad Escobar[5], Manuel Balparda[5], Gustavo Sibilla[5]

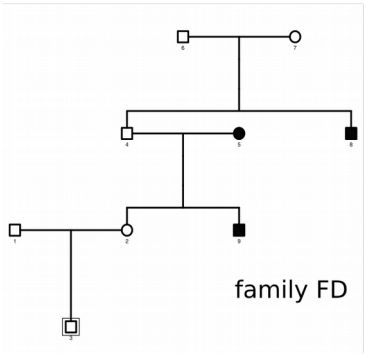
Supplementary Table 1: Table of molecular markers considered in this contribution. Marker set A was used for assessing the precision of the truncated stepwise mutational model comparing *fbnet* and *Familias* results for Family FF (Figure 1). Marker set B was used for the computation of running times and LRs for the 24 familial pedigrees of the BNDG (Table 1) comparing *forrel* with *fbnet*.

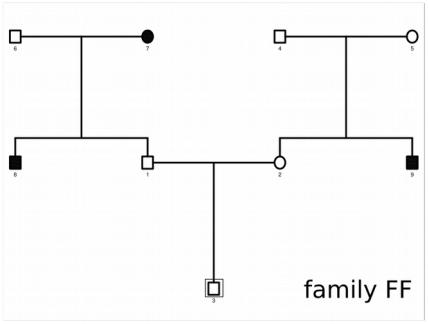
Marker	**num alleles**	**Marker set A**	**Marker set B**
D2S1338	13		x
D2S1360	14		x
D2S441	11		x
D3S1358	12	x	x
D3S1744	11		x
D4S2366	12		x
D5S2500	10		x
D5S818	10	x	x
D6S474	8		x
D7S1517	15		x
D7S820	14	x	x
D8S1132	15		x
D8S1179	12	x	x
D10S1248	8		x
D10S2325	13		x
D13S317	11	x	x
D16S539	10	x	x
D18S51	25	x	
D19S433	15		x
D21S11	32	x	
D22S1045	8		x
CSF1PO	12	x	x
FGA	27	x	
Penta D	18	x	
Penta E	23	x	
TH01	10	x	x
TPOX	13	x	x
VWA	13	x	x

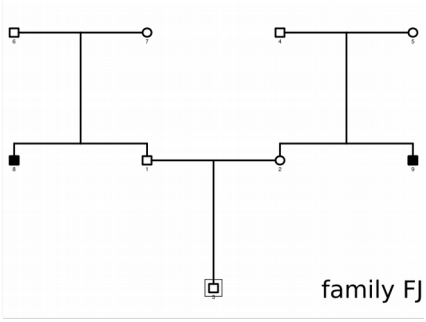
Supplementary Table 2: Comparison between *forrel* and *fbnet* LR estimations for 1000 MP conditioned simulation for 24 families of the BNDG (see Sup Fig 2 for the complete set of pedigrees). Asterisks denote pedigrees with paternal branch excluded. The number of family members, genotyped individuals and the correlation between LogLR values estimated by *forrel* and *fbnet* (L=4 truncated stepwise mutational model) are displayed in the first three columns of the table. Mean and maximum error percentages are displayed in the fourth and fifth columns. The number of discrepancies found at 1%, 2% and 5% tolerance thresholds are reported in the last three columns of the table.

	#members	#genotyped	Corr Log[LR]	mean error [%]	max error [%]	# dev>1%	#dev>2%	#dev>5%
F1	12	5	0.99999269	0.17	74.50	10	2	1
F1*	9	3	0.99999794	0.08	14.42	5	3	2
F2	16	9	0.99999994	0.06	2.61	5	1	0
F2*	12	6	0.99999991	0.06	1.86	5	0	0
F3	9	5	0.99999988	0.09	6.42	6	4	1
F3*	9	5	0.99999932	0.07	13.18	6	2	1
F4*	6	3	0.99999981	0.04	5.43	2	2	1
F5	9	5	0.99999998	0.04	1.79	3	0	0
F5*	9	5	0.9999999	0.05	2.33	4	1	0
F6	9	3	0.99999963	0.11	9.80	14	5	2
F6*	6	2	0.99999983	0.07	2.38	6	2	0
F7	14	8	0.99999983	0.10	4.68	7	4	0
F7*	11	6	0.99999974	0.15	2.03	4	1	0
F8	6	2	0.99999991	0.06	1.51	4	0	0
F9	7	5	1	0.02	0.04	0	0	0
F10	11	5	0.99999959	0.11	10.38	5	2	2
F10*	8	3	0.99999859	0.09	20.73	7	3	2
F11	13	6	0.99999989	0.12	5.95	2	2	1
F11*	8	3	0.99999996	0.07	1.42	3	0	0
F12	5	1	0.99999951	0.08	5.82	8	5	2
F13	11	5	0.99999983	0.09	4.69	5	4	0
F13*	6	2	0.99999983	0.05	4.49	2	2	0
F14	16	8	0.99999997	0.07	1.80	2	0	0
F15*	17	8	0.99999968	0.11	9.04	4	1	1

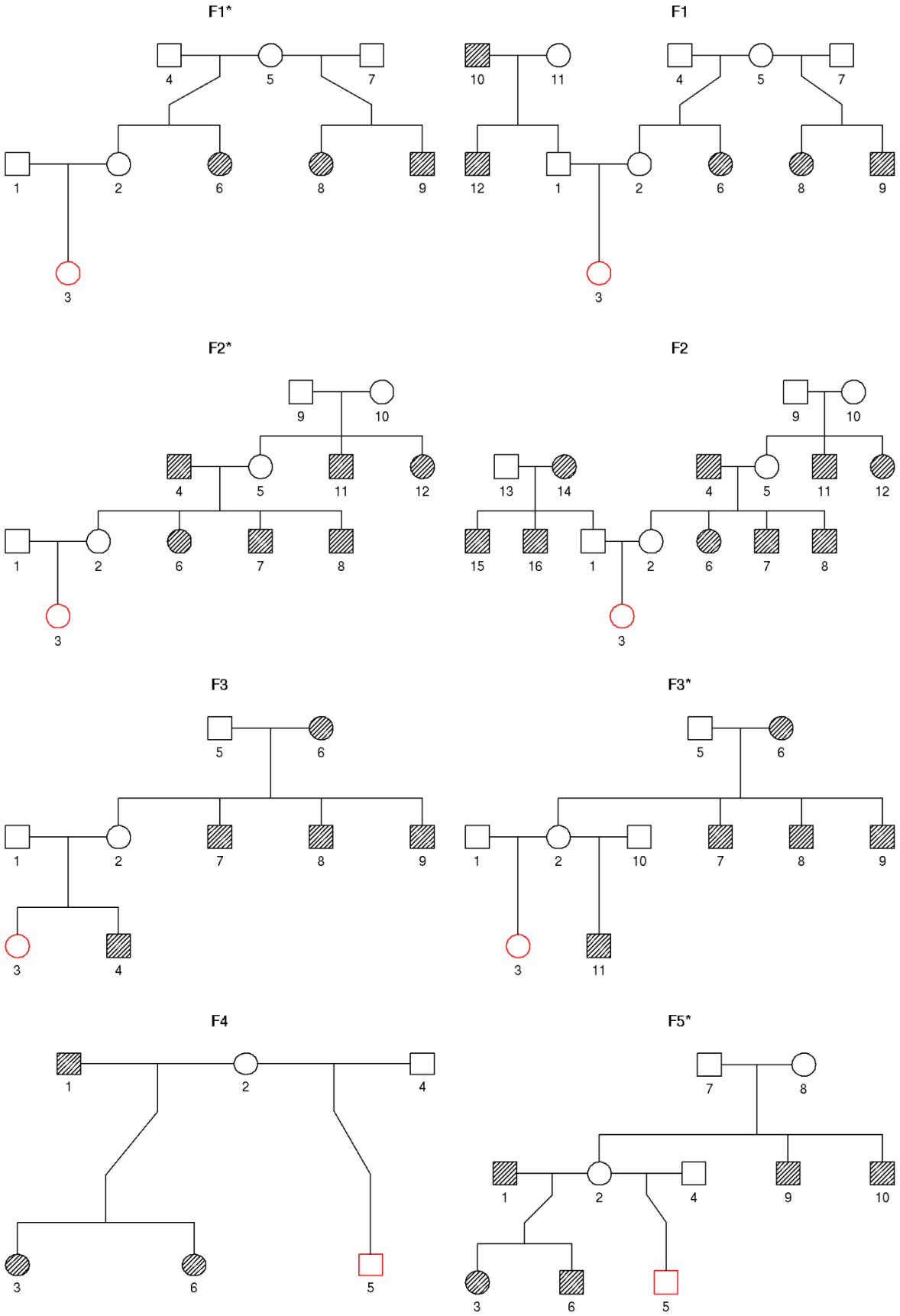
Supplementary Figure 1 - Pedigrees considered for precision and execution time evaluation of *fbnet*. Genotyped individuals and the family missing person are depicted with shadowed and double stroked symbols respectively.



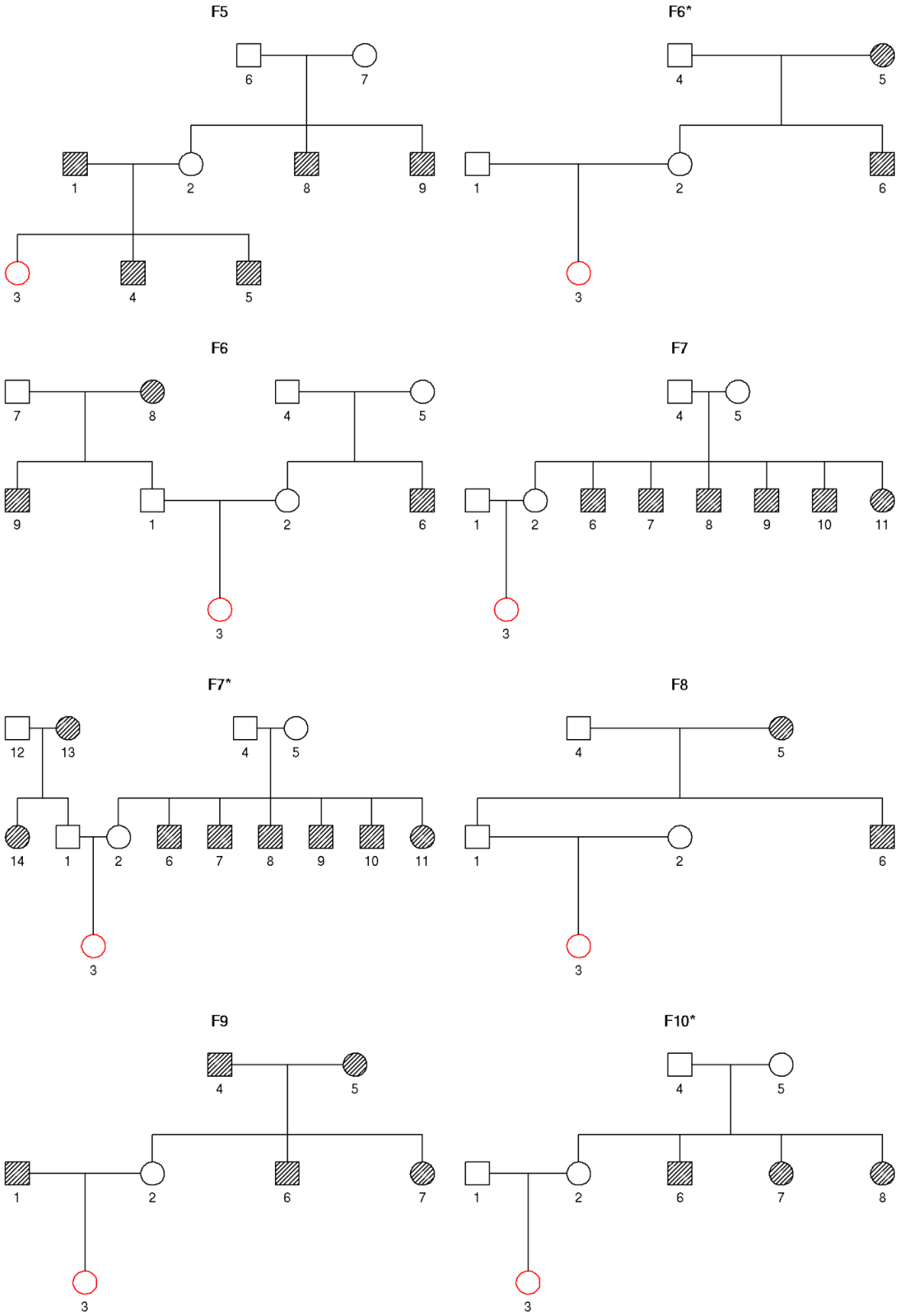




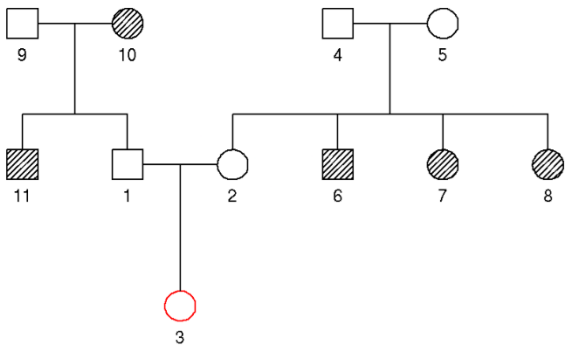
Supplementary Figure 2 – BNDG pedigrees



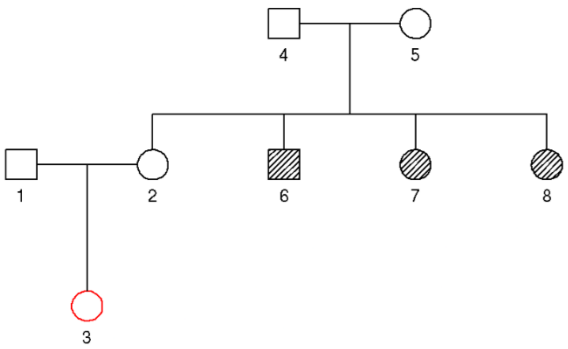
Supplementary Figure 2 - contd



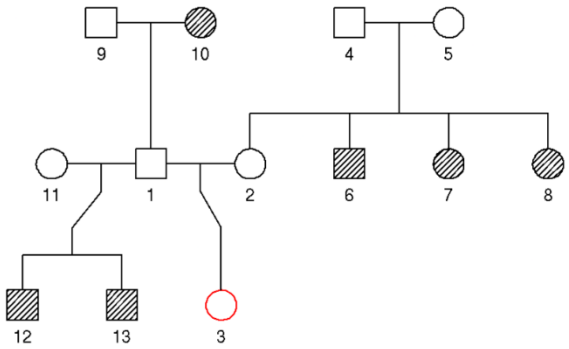
F10



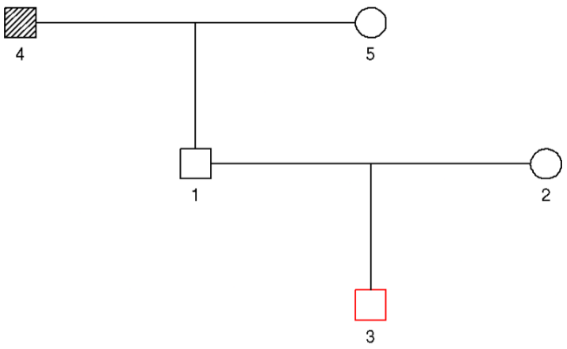
F11*



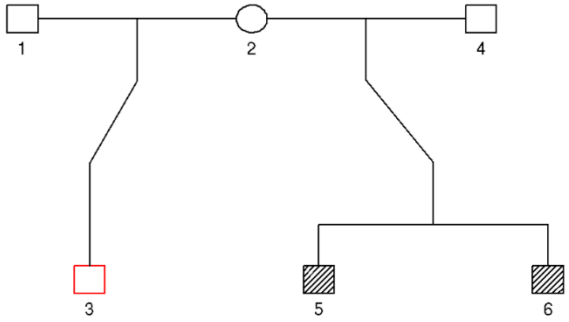
F11



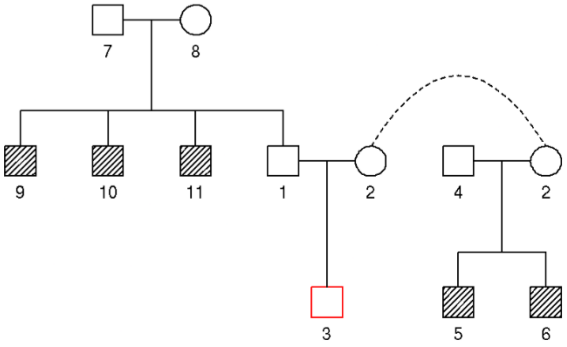
F12



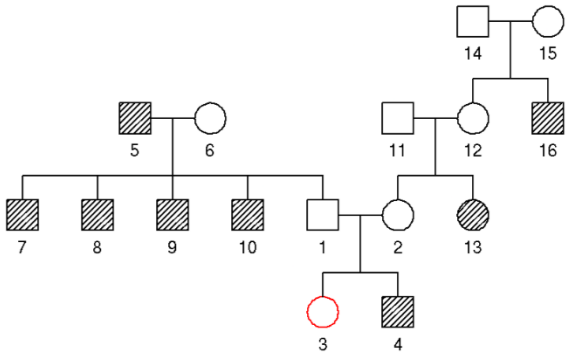
F13*



F13



F14



F15