Decision making in large scale comparisons for missing person identification

The last military civic dictatorship in Argentina left tens of thousands of kidnapped and disappeared people. Many victims were pregnant women. Rapes in captivity were also common, frequently resulting in pregnancies. Children kidnapped with their parents or born in captivity were killed or delivered to families related to or from the military forces and their identities were forged. In the large number of cases both biological parents of the kidnapped children were murdered and their bodies still remain missing. As the regime eventually ended by 1983, grandmothers of the abducted children (organized in the association Abuelas de Plaza de Mayo) started to enquire the scientific community and the new democratic government to aid them in the search for their missing grandchildren [1]. In 1987 the Banco Nacional de Datos Genéticos (BNDG) was created through the passing of specific laws. Since then, the BNDG has collected DNA from relatives (mainly grandparents, uncles/aunts and siblings) of appropriated children, and performed genetic analyses on thousands of children suspected to be one of the missing grandchildren.

At the moment, the BNDG database has more than ten thousand samples from people who were born in the civil-dictatorship period and doubt about their true identity (Persons of Interest, POI) and more than one thousand of individuals of reference (IRs) that make up around three hundred family groups. For all POI both mitochondrial (mtDNA) and autosomal short tandem repeats (autosomal STRs) profiles were analyzed. For all IRs autosomal STRs and mtDNA of the maternal side of the pedigrees were analyzed. As counterpart, sexual chromosomes genetic profiles (X-STRs and Y-STRs) are used in particular cases when it is possible and necessary.

The use of DNA databases for missing person identification (MPI) and disaster victim identification (DVI) has gained relevance in the last decades [2,3]. From the very beginning, the identification of the missing grandchildren of Argentina played a key role in the development of statistical analysis that could guarantee the identification of the children when there was a lack of genetic information from both disappeared parents [1]. The evaluation of the DNA-based capability of identification was previously studied in the BNDG [4]. To assess the power in an identification case, authors advocated the combined use of two statistics: the Power of Exclusion (PE) and the Power of Inclusion (PI). The PE is the probability that the genotypes of a random person are incompatible with the family genetic data and the PI is the probability that the likelihood ratio will exceed a LR threshold of 10000, assuming that POI is indeed the missing person (MP). Those family groups with low EP and IP have poor statistical power. It is a non-trivial problem to evaluate when enough genetic data has been collected to enable a robust conclusion when testing for a match. To this end, conditional simulations were performed using all the available information from each pedigree. It was the first time that such type of analysis was systematically implemented in a large dataset. One of the main conclusions of this work was the identification of family groups that require additional genetic data to enable a positive identification. Such type of analysis are eligible in order to make decisions, particularly when we are dealing with large scale comparison and complex family groups in MPI cases. The advantages and limitations of this types of approaches are very well discussed by Slooten et al. [Slooten, 2019].

In large scale comparisons there are still several complications that the forensic scientists are faced with, such as complex family structures, large-scale comparisons and inconsistencies due to mutations. Having low statistical power implies a risk in the DNA-based identification allowing the occurrence of false positives (when PE=0) and/or having high probability of losing an identification. The establishment of an LR decision threshold (LR_{dt}) for each family group that minimize the probability of losing an identification whilst obtaining a manageable number of matches that exceeds this threshold (doubted positives, DP) is somewhat of black art. It has to be considered that massive comparisons carried on in MPI the DP are re-analyzed by geneticists and more genetic and non-genetic data is incorporated to the analysis in order to avoid false positives. Once having a manageable number of positives for a specific family group, decision could be made about incorporating more autosomal STRs markers, Y-STRs or X-STRs [5] into the analysis.

Here we worked in a general method in order to achieve a better decision making criteria when dealing with large scale comparison and family groups with low statistical power in MPI cases. We made a general description of the detection capability of the family groups present in the BNDG and then, we proposed a strategy for working with those groups with low PI and/or low PE. This criteria was based in the assumption that the probability of missing an identification caused by a low LR value obtained when the POI is the MP should be minimal. Moreover, we discussed the utility of the application of lineage genetic markers (mtDNA) and non-genetic information in databases searching.

[1] Penchaszadeh, V. Use of dna identification in human rights, work to reunite families in latin america, in els, john wiley & sons, ltd (2001).

[2] Baeta, M. et al. Digging up the recent spanish memory: genetic identification of human remains from mass graves of the spanish civil war and posterior dictatorship. Forensic Sci. Int. Genet. 19, 272–279 (2015).

[3] Vullo, C. M. & et al. Ghep-isfg collaborative simulated exercise for dvi/mpi: Lessons learned about large-scale profile database comparisons. Forensic Sci. Int. Genet. 21, 45–53, DOI: https://10.1186/2041-2223-2-15 (2016).

[4] Kling et al. Evaluating the statistical power of DNA-based identification, exemplified by 'The missing grandchildren of Argentina'. Forensic Sci. Int. Genet. DOI: 10.1016/j.fsigen.2017.08.006 (2017)

[5] Slooten K. Likelihood ratio distributions and the (ir)relevance of error rates. Forensic Sci. Int. Genet. doi: 10.1016/j.fsigen.2019.102173. (2019)