ETHICAL CONSIDERATIONS
FOR FORENSIC GENETIC FREQUENCY DATABASES

FIRST REPORT BY THE FORENSIC DATABASES ADVISORY BOARD (FDAB)*

(February 2023)

* In alphabetical order:

M.E. D'Amato[1], Y. Joly[2], V. Lynch[3], H. Machado[4], N. Scudder[5], M. Zieger[6]

[1] Forensic DNA Laboratory, Department of Biotechnology, Faculty of Natural Sciences, University of the Western Cape, Bellville 7535, South Africa.

[2]Centre of Genomics and Policy, McGill University, 740 avenue Dr. Penfield, Suite 5200, Montreal, Quebec, H3A 0G1, Canada.

[3]DNAforAfrica, Gordon Thomas Honeywell - Governmental Affairs, Cape Town, South Africa.

[4]Institute for Social Sciences, University of Minho, Campus de Gualtar, 4180-057 Braga, Portugal.

[5]Centre for Forensic Science, School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney, Sydney, Australia.

[6]Institute of Forensic Medicine, University of Bern, Murtenstrasse 26, CH-3008 Bern, Switzerland.

SUMMARY

This first report of the Forensic Databases Advisory Board (FDAB) aims to provide the International Society for Forensic Genetics (ISFG) with a framework to assess the ethical implications of hosting data from a variety of population groups on the forensic genetic frequency databases (FGFD) known as *Y-chromosome Haplotype Reference Database* (YHRD), the *EDNAP Mitochondrial DNA Population Database* (EMPOP) and the *STRs for Identity ENFSI Reference Database* (STRidER).These free FGFD were developed to serve as statistical support to the evaluation of forensic evidence (match, kinship) which benefits the advancement of science and society in general. Ethical considerations in this report include questions of data acquisition, data sensitivity and identifiability, and ethical principles to be implemented.

Submissions to the FGFD were classified with consideration to evolving ethical landmarks including those applicable to biomedical research, namely the *Declaration of Helsinki* (1975), the *UNESCO Universal Declaration on the Human Genome and Human Rights* (1997), as well as more specific *Forensic Science International: Genetics* (FSIG) guidelines published in 2010 and 2020. Further classification of data was defined by its source (academic, law enforcement, private, etc.). However, the complexity of the composition, variety of submissions and evolving ethics regulations pertinent to these FGFD rendered the composition of an appropriate ethical assessment framework challenging. It was therefore found that notwithstanding considerations proposed to the ISFG for an appropriate assessment of data held on the FGFD, a deeper debate is required on the role and acceptable alternatives to informed consent, the impact of including datasets on minorities and vulnerable groups, and other relevant ethical aspects for a complete risk-benefit assessment of data hosted on the FGFD.

Contents

1. Mandate of the FDAB

The Forensic Databases Advisory Board (FDAB) is an independent advisory board which was convened in January 2022 to provide evidence-based ethical advice to the International Society for Forensic Genetics (ISFG) and hence to the forensic community. None of the members of the FDAB have financial or other arrangements with the ISFG that would constitute a conflict of interest to report.

This first report of the FDAB seeks to define the ethical criteria as well as outline the processes for forensic genetic frequency databases to be curated, maintained, accessed and utilized by the forensic community under the banner of the ISFG. These non-commercial forensic genetic frequency databases, hosted and maintained by European academic institutions (collectively referred to as the 'forensic genetic frequency databases' (FGFD) for the purposes of this report, presently comprise the *Y-chromosome Haplotype Reference Database* (YHRD), the *EDNAP Mitochondrial DNA Population Database* (EMPOP) and the *STRs for Identity ENFSI Reference Database* (STRidER).

2. Objectives of the FDAB

The primary objective of the FDAB is to outline a methodology for the ISFG to evaluate the compliance of sampled population data contained in the FGFD, with defined ethical guidelines for current and future submissions, as well as legacy data[1]. To this end, this first report of the FDAB will provide an ethical framework which will help identify the relevant ethical questions to assess the ethical conformity of data residing on the FGFD, including, but not limited to, questions of data privacy, informed consent, and sensitivity of data.

Further, this report will provide an overview of the composition, contributions, access, control, and utilization of the various FGFD; discuss the ethical challenges raised in

[1] Legacy data refers to sampled population data entered onto the existing on the forensic genetic frequency databases prior to the publication of the ISFG 2020 ethical guidelines (D'Amato et al. 2020), when entry criteria were more permissive, non-existent or where the ethical process followed may not have been documented.

respect of the FGFD; outline an ethical framework for the evaluation of legacy, contemporary and future contributions to the FGFD and considerations regarding the retention and acceptance of submitted data to these databases.

While some of the considerations provided in the present report may also be relevant to other types of genetic databases, for the purposes of this report, only the three FGFD, namely the YHRD, EMPOP and STRidER, were specifically reviewed.

3.Forensic Genetic Frequency Databases (FGFD)

In contrast to forensic DNA databases,[2] the FGFD under consideration in this report are online, public, and subscription-based databases which serve the purpose of providing a basis for statistical considerations when a match between DNA profiles of crime scene samples and individuals (e.g., suspects) has already been established or when the probability of a certain kinship scenario must be assessed. While the FGFD provide frequencies of genetic characteristics in various populations, they do not allow the identification of specific individuals. In other words, the DNA profiles held on the FGFD constitute a representation of the background population and are used to calculate inferential statistics, which describe observable, relevant properties of the larger population. This provides the basis for forensic practitioners to analyze and interpret the DNA profile in question to determine the rarity of that profile in a particular population, metapopulation, region and/or globally. Without a statistical analysis that reflects how rare the DNA profile in question is, the assumption that the DNA profile of the person of interest, (e.g., a suspect) can be included as the same as that of the crime scene profile, has no probative evidentiary value (Balding and Steele, 2015; Bright et al.et al.2019).

The statistical methodology applied with reference to the FGFD assists forensic practitioners to establish a more accurate representation of each population by providing reliable statistical information for:

[2] Forensic DNA databases are State held computer databases containing DNA profiles obtained from crime scene samples and/or from individuals for the purpose of identifying who was present at a crime scene, linking several crime scenes together, or matching human remains to missing persons. While every country which administers a forensic DNA database has its own laws and regulations governing the DNA profile entry, expungement and retention criteria for its forensic DNA database, there is no single international overriding law regulating these practices.

(1) the estimation of the weight of evidence for matching DNA profiles in a crime; and
(2) the calculation of the probability of different hypotheses concerning filiation, e.g., for the identification of human remains or in paternity disputes.

While the main purpose of the FGFD is to assist forensic practitioners in the evaluation of DNA evidence, users accessing the FGFD encompass a wide community of stakeholders, which include researchers, educators, publishers and the general public.

In addition to the supply of frequency data for the evaluation of genetic findings, the FGFD presently operate under the mandate of the ISFG to assure a high quality of publicly available genetic frequency data. To this end, all autosomal STR, Y-STR and mtDNA data published in the ISFG-associated journal, FSIG, must undergo quality control at one of those databases (Gusmão et al. 2017). The same applies for the journal *Forensic Science International: Reports.*[3] Further, the *International Journal of Legal Medicine,* the official publication of the International Academy of Legal Medicine (Poetsch et al. 2012) also requests quality control from YHRD and EMPOP.

4.Nature of genetic data held on the FGFD: autosomal *vs* lineage markers

To lay out the criteria and provide considerations for the retention of legacy and contemporary data and incorporation of future data on the FGFD, it is important to define the nature and purpose of the genetic data residing on the FGFD as well as distinguish between lineage and autosomal markers. This distinction also provides the necessary background to discussing data sensitivity.

A haplotype (as a haploid genotype) is a combination of alleles, inherited together, from a parent to their child. Classic haplotypes in humans are the Y chromosome and the mitochondrial genome (also known as the 'mitogenome' or 'mtDNA'). Y chromosome haplotypes provide lineage markers to trace inheritance among paternally related men (Jobling and Tyler-Smith, 1995, 1997; Kayser and Sajantila, 2001). Mitochondrial haplotypes serve to trace the maternal lineage because mtDNA is always passed from a mother to her child (Amorim et al. 2019).

[3] FSI:R online Guide for Authors: https://www.sciencedirect.com/journal/forensic-science-international-reports.

In contrast to genetic markers located on the Y chromosome or the mitogenome (lineage markers), markers located on the diploid autosomal chromosomes (autosomal markers) are newly combined in every germ cell. This makes it extremely unlikely that children from the same parents will have the identical autosomal marker combination (except for monozygotic twins) using standard fragment analysis technology. Massive parallel sequencing (MPS) would even allow for the identification of monozygotic twins (Weber-Lehmann et al. 2014).

The different modes of inheritance described above have implications for the application of haplotype (lineage) markers and autosomal markers in forensics. Due to recombination, autosomal markers are newly combined in every individual and are therefore powerful for individual identification (Butler, 2010). In contrast, the lineage markers lack the individualization power of autosomal markers but have distinct advantages in certain cases, such as the potential to infer ancestry, both geographical and familial (Chaitanya et al. 2014, Syndercombe Court, 2021). The value of Y chromosome analysis has been demonstrated in cases of sexual assault, particularly where the amount of material left by a male assailant is limited in comparison with female DNA (Anderson and Balding, 2017; Kayser, 2017; Prinz and Sansone, 2021). Because the circular mitogenome molecule is very robust and exists in several hundred copies per cell, it is particularly useful when DNA is heavily degraded. This is the case with some biological materials such as hair shafts or with burnt human remains (Amorim et al. 2019). Other applications of mtDNA are in the triage of disaster victims and in the identification of bodies in mass graves (Syndercombe Court, 2021).

The different modes of inheritance between autosomal and lineage markers also impact the evaluation of corresponding DNA profiles or calculations in kinship cases. To calculate accurate family pedigree relationships, the frequency of a particular profile in the relevant population, be it an autosomal profile or a haplotype, must be estimated. This is straightforward for forensic autosomal STR markers, since they are all inherited independently from each other, i.e., the frequencies of the individual alleles in a population can be multiplied to obtain the frequency of a person's DNA profile (combination of alleles at several markers) in the relevant population. In contrast, the alleles of a certain haplotype are all inherited together. The frequency of the haplotype is therefore not just a multiplication of the frequencies of its different alleles. The frequency of the haplotype must be assessed as a whole. Due to the high variability of haplotypes, especially for larger modern Y-STR profiling kits including

rapidly mutating (RM) YSTRs, it was recommended that the reference data used to provide reliable frequencies be increased (Gusmão et al. 2017).

5.FGFD under the mandate of the ISFG

Below is a detailed description of the three forensic genetic frequency databases under the mandate of the ISFG, as well as the criteria for data submission for each.

5.1 STRs for Identity ENFSI Reference Database (STRidER)

The STRidER database is the expanded and enhanced version of the ENFSI (European Network of Forensic Science Institutes) STRbASE (2004-2016). It is currently hosted and curated by the Institute of Legal Medicine (GMI), Medical University of Innsbruck, Austria. STRidER has been active since 2016 and is the frequency database and platform for autosomal STR data quality control (QC) (Bodner et al. 2016). There is an apparent need for QC since it has been reported that between July 2017 and July 2019 only 48 of 147 submitted datasets passed QC (Bodner and Parson, 2020). Currently STRidER contains aggregated frequency datasets from 19 European countries, Saudi Arabia and Thailand, for 40 autosomal STR loci, compiled from approximately 10,000 individuals.[4] The database does not contain information about contributors. STRidER data can be used for frequency estimates of query profiles (Probability of Random Match) overall, at continental level or in national populations of choice. As no individual genotypes are stored, no number of matches will be returned (this contrasts with the lineage databases). STRidER only returns a frequency estimate in the chosen population. STRidER only holds continental or national datasets and data is not organized according to metapopulations or ethnic groups.

STRidER users do not need to register with the host to conduct queries and the allele frequency data can be downloaded for use with external calculation software as aggregated frequency files. Since its installation in 2016, STRidER further requires

[4] https://strider.online/frequencies (accessed March 25, 2022). Some countries have their own statistical databases. For example, in South Africa the allele frequency database is known as the National DNA Statistics Database (NDSD).

data submitters to complete an informed consent checkbox and provide the applicable ethical clearance approval or code to proceed (Bodner et al. 2016).[5]

5.2 Y-chromosome Haplotype Reference Database (YHRD)

YHRD is an online Y chromosome haplotype reference database historically hosted by Charité – Universitätsmedizin Berlin (for updates see https://yhrd.org/pages/disclaimer), currently curated by forensic geneticists Sascha Willuweit (affiliated with Landeskriminalamt Berlin, Germany) and Lutz Roewer (affiliated with Charité University Hospital, Berlin, Germany). The YHRD has been active since 1st August 1999 and was made available online in 2000 (Roewer et al. 2001; Willuweit and Roewer, 2015). The YHRD was initially established from a collaboration of 25 forensic genetics, medical genetics and anthropological genetics institutions across Europe, Latin and North America and Asia, which collectively contributed 3825 male Y-STR haplotypes typed in 48 population samples on occasion of the 1st Y-User Workshop in Berlin, Germany on April 19-20th, 1996 with the aim of creating a forensic Y chromosome frequency database (Kayser et al. 1997). Since that time, the YHRD has grown exponentially, with contributions from across the world (Syndercombe Court, 2021) and is currently in Release 68 (November 2022). The YHRD is used worldwide as the current standard reference database to calculate the weight of evidence of matching Y-STR profiles (Schiermeier, 2021; Roewer and Willuweit, 2023).

The collection of populations on the YHRD are classified using both geopolitical (by country) and anthropological/biological (by metapopulation) criteria. Currently, 141 countries are represented in the YHRD. Metapopulations are defined based on both linguistic (preferred) and geographical determinants, and the YHRD currently holds 37 different metapopulations. These metapopulations are organized hierarchically. As an example, the Western European metapopulation is part of the European metapopulation that is again part of the Eurasian metapopulation. Approximately 40 percent of the database entries are from the East Asian metapopulation (of which approximately two-thirds are of Han Chinese descent); about 30 percent from the

[5] https://strider.online/contribute (accessed February 8, 2023).

Eurasian metapopulation (the largest group being Western Europeans with about one-third) and about 20 percent from admixed populations (more than half of which are U.S. American). People from African, Afro-Asiatic, Native American, Australian Aborigine and Inuit, Yupik, and Unangan populations make up the remaining 10 percent of the database.[6]

YHRD curators have recently introduced changes to the YHRD (YHRD Release 67, February 2022) removing the "ancestry" function from the database which could previously be used to infer, to some degree, the paternal biogeographical ancestry of a query sample. The resolution of the query results has also been reduced, so that queries now provide haplotype frequency results by country and metapopulation only. These changes were affected to enable users to see all studies that contribute to a country's or metapopulation's dataset and at the same time minimize the potential for the identification of a specific population group from one particular study.[7]

Today the YHRD holds over 340 000 individual Y-chromosomal haplotypes consisting of at least 9 markers (minimal dataset). About 100 000 of these haplotypes consist of 27 Y-STR markers (YFiler Plus dataset). Direct submissions per country are listed with the respective submitter's ethics practice or disclaimer and affiliation details, while submissions that have been subjected to peer review are listed as a bibliography and linked to countries of origin. From 2022, contributors are asked to provide a blank informed consent form and information about ethics approval upon data submission.[8]

The YHRDs website (www.yhrd.org) is publicly available:

(1) to generate Y-STR haplotype frequency estimates for Y-STR haplotypes for use in the quantitative assessment of matches in forensic and kinship casework (not restricted to registered users).

(2) for the assessment of male population stratification among world-wide populations as far as reflected by Y-STR and Y-SNP[9] frequency distributions (restricted to registered users); and

[6] https://yhrd.org/pages/resources/metapopulations (accessed March 25, 2022).

[7] Roewer L, Willuweit S. Y-chromosomale Analyse in der Praxis - Interpretation und Biostatistik mit Hilfe der YHRD Datenbank. Workshop presentation. 42nd joined Spurenworkshop of the German Society for Legal Medicine and the German Stain Commission (February 09, 2022).

[8] https://yhrd.org/pages/help/contribute (accessed February 8, 2023).

[9] Single nucleotide Polymorphism. SNPs are used to determine Y chromosomal haplogroups. Haplogroup distributions differ between populations and geographic regions. They can be used to describe the history of human evolution or gain probabilistic hints towards the most probable biogeographical, paternal origin of a biologically male individual.

(3) for the provision of advanced tools and further resources concerning Y-STRs and Y-SNPs (e.g., a tool for Y-STR mixture analysis; not restricted to registered users).[10]

Notwithstanding the above objectives, the primary purpose of the YHRD is for the estimation of the frequency of a certain Y-STR haplotype in a male population relevant to a crime scene sample. Y-STRs are also frequently used in the examination of a paternal lineage in both criminal (e.g., Kayser et al. 2017) and civil investigations. For example, in cases of inheritance where individuals may want to claim against a deceased estate where family pedigree has been disputed, Y-STRs can be helpful when the putative father is deceased or missing, and a reliable estimation of the profile frequency is crucial.

YHRD users can freely query the database by entering a specific Y-STR-profile. The database will return how often it contains the query profile in the entire database or within a user-selected metapopulation or national dataset. In the absence of additional information about the origin of the person of interest (e.g., a suspect), the relevant reference population for such a frequency query is usually the one at the place where the alleged crime occurred (Roewer et al. 2020). The YHRD will also provide an estimate based on different calculation methods for the frequency of the haplotype within the population selected by the user. It is important to note that no genetic data can be downloaded from the YHRD.

The Scientific Working Group on DNA Analysis Methods (SWGDAM) published Y chromosome interpretation guidelines in 2014 (SWGDAM, 2014) and recently updated them (SWGDAM, 2022), outlining the principles for using Y-STR analysis as an additional tool to autosomal DNA, or instead of autosomal DNA, when needing to detect male DNA in a mixture with an abundance of female DNA. The DNA Commission of the ISFG has also provided recommendations on the interpretation of forensic Y-STRs setting down certain basic requirements for reference Y-STR databases, namely full specific kit-haplotypes from randomly selected individuals and meta- and sub-population data and population datasets of sufficient size to ensure representation (Roewer et al. 2020). Both SWGDAM and the ISFG, as well as various national guidelines, refer to the YHRD as the relevant data repository for haplotype frequency estimation (SWGDAM 2014 and 2022, Roewer et al. 2020).

[10] https://yhrd.org/ (Accessed February 8, 2023).

5.3. EDNAP Mitochondrial DNA Population Database (EMPOP)

The EMPOP database is currently hosted and curated by the Institute of Legal Medicine (GMI), Innsbruck Medical University, Austria. EMPOP holds partial (control region, non-coding ~1200 bp) or full mitochondrial genome sequences (encompassing coding and non-coding regions) of more than 48,000 individuals.[11] The study participants are grouped in large metapopulations. Approximately 46 percent of the database entries are from the West Eurasian metapopulation, 19 percent from the Asian metapopulation (subdivided into South Asian, Southeast Asian and East Asian), 17 percent from the Native American metapopulation and 11 percent from Sub-Saharan African metapopulation. The remaining samples in the database originate from the Oceania metapopulation and from admixed individuals.[12]

Since 2020, in line with the FSIG and FSIR guidelines (D'Amato et al. 2020), for submission to EMPOP, data contributors must confirm by a 'check box' that the submitted data was generated according to applicable national laws and that ethics approval was obtained. The name of the relevant ethics committee and application number must also be provided.[13] Submitters are further requested to provide a blank version of the study information provided to the participants and the informed consent form (personal communication, requirement not stated on website). Contributors and the accession numbers of their submitted datasets are also accessible to registered users.

Users of EMPOP must first register with the host before they can access the information in the database, and are required to provide an e-mail address, name and affiliation. Registration is confirmed via a link sent to the provided e-mail address. No genetic data can be downloaded from EMPOP, but the database can be queried for matches and close matches.

The database can be interrogated by entering a mitochondrial query sequence. The database will show how many matches to the queried sequence it holds, in which population sample and metapopulation, and at which frequency. EMPOP also returns

[11] For the difference in terms of information content, see infra at 6.2.2.
[12] https://empop.online/empop_stats (accessed February 8, 2023).
[13] https://empop.online/contribute (website accessible for registered users only, accessed February 8, 2023).

a map, depicting from which population the matching database sequences originate.[14] For every dataset, the submitter and publication of the original data is shown.

6. Ethical challenges affecting the FGFD

The primary ethical concerns affecting the FGFD can be summarized as follows:

(1) Validity of practices in the data acquisition process (data acquisition).

(2) Potential for re-identification of contributing human subjects, including the potential for linking the data in the FGFD with data held in another database or in publications (data identifiability for individuals, groups of individuals or biogeography).

(3) The potential inference of a medical condition, lifestyle and other personal or physical characteristics derived from individual data (data sensitivity).

(4) Special ethical requirements pertaining to vulnerable groups.

Each of these concerns will be addressed separately below.

6.1. Data acquisition

6.1.1. A variety of sources

The processes involved in the acquisition of samples and data, as well as governance of those samples and data are often subject to regulations imposed by different national and international bodies. It is thus important to understand both the nature of the contribution as well as the samples' origin.

A substantial portion of the datasets on the FGFDs stem from academic research studies.[15] The collection of genetic data in the context of research projects is subject to international bioethical principles and national regulations. Prior to the collection of biological material from humans for the purposes of research, the approval of the research project by the relevant ethics review board/s is generally required. This approval effectively endorses the research proposal that ensures voluntary

[14] Personal communication from the curator: This map will be removed in the next version of EMPOP.

[15] Accessed 2 November 2022: YHRD shows a total of 363 accession numbers for direct submissions and 719 listed publications with at least 1 accession number each, while EMPOP shows a total of 35903 submitted profiles, of which 48.22% are linked to scientific publications, 46.22% are direct submissions and for 6.75 % no information is provided. Contributions and contributors are not listed in STRidER database.

participation, confidentiality, informed consent, and often includes other applicable references to risks and benefits of participation in the study. The possibility of withdrawing from the study, agreement to sharing the samples/results with other researchers (including internationally), and the agreement to the results being made available in a public database are clauses often included in the informed consents.

Regulatory frameworks in a few countries offer guidance on minimal consent requirements, e.g., the revised 2018 Common Rule 45-CFR-46.116 (US Dept. of Health and Human Services, 2017).

Law enforcement agencies, state laboratories and other organizations (some of which have been listed below) are also regular contributors to the FGFD. For this purpose, they have contributed a variety of data from samples.[16] Below is a summary of the type of samples collected for data stored in the three FGFDs, as well as the universal principles governing scientific research with human subjects and their adoption by the forensic genetics' academic community. A separate section summarizes the regulation of non-academic sources of samples.

Data in the FGFD are submitted by a variety of contributors:

- Academics
- Law enforcement agencies
- Corporate/manufacturers
- Service/private labs that may possibly include some direct-to-consumer genetic testing (DTC-GT) companies
- Non-governmental organizations
- Not indicated

The origin of the biological material is also indicated:

- Research study participants
- Convicted criminals and suspects
- Purchased / Biobanks
- Not declared

[16] YHRD accompanies direct submissions with the statements "The submitter confirmed the legal permission to submit this data to YHRD and to comply with good scientific practice (Deutsche Forschungsgemeinschaft 2019). When this data was submitted, an exemplarily informed consent form and a confirmation of an approval by an ethics board, committee or organization was provided to us.", and in some cases "The submitter confirmed the legal permission to submit this data to YHRD and to comply with good scientific practice (Deutsche Forschungsgemeinschaft 2019). Since this data was submitted more than 10 years ago, it is treated as legacy data." EMPOP and STRidER do not provide additional information.

6.1.2. Regulations and practices on data acquisition

The *Declaration of Helsinki* 1964 version[17] stated the principles of free informed consent and the right to withdraw from the study. The subsequent 1975 amendments of the Declaration incorporated the need for an independent review board. These basic principles were further consolidated and promoted by the CIOMS (Council for International Organizations of Medical Sciences), *Proposed International Ethical Guidelines for Biomedical Research Involving Human Subjects* (1982), and by UNESCO's *Universal Declaration on the Human Genome and Human Rights* (1997). In summary, it became clear from the date these universal principles were adopted, they needed to be guaranteed to participants in the research context. These basic principles are:

(1) Voluntary participation with informed consent unless an ethics waiver of consent was obtained from an ethical committee before collection (1964).
(2) Personal data to be held confidential as per applicable legal requirements (1964).
(3) The collection of samples must be approved by an ethics board, or an ethics waiver must be obtained by an ethics committee (1975).
(4) Individuals whose samples are collected shall not be subjected to discrimination based on genetic characteristics that are intended to infringe or have the effect of infringing human rights, fundamental freedoms, and/or human dignity (1997).

Note: The applicability of such rights to unidentifiable data subjects is unlikely. However, if the data was originally derived from biological samples and not through secondary use of existing data, then those rights would likely apply if the relevant international principles mentioned above were in force at the time when the samples were initially collected. The applicability of the above stated rights and the categorization of human biological data or samples may vary across jurisdictions. The relevant technical vocabulary related to privacy, security, research ethics and human subjects may also vary across different regulations and are summarized and regularly updated on an international level by the Global Alliance for Health and Genomics

[17] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

organization (GA4HG, 2019). In the present report, terminology from the current version of the GA4HG document has been adopted.

The collection of biological samples can also be mandated by legislative authority, such as from suspects and convicted offenders. These legislative mandates generally prescribe how biological samples for forensic purposes are collected, managed, retained and destroyed.[18] Inclusion of pseudonymized or anonymous samples from a national law enforcement DNA database in an FGFD would raise significant ethical issues, given the high likelihood that the contributors may not have freely consented to inclusion.

The journals that publish scientific work with data accessible in the FGFD[19] have established publication guidelines with varying levels of ethical requirements. For instance, the FSIG 2010 guidelines (Carracedo et al. 2010) state under "Ethical Requirements" that: "Informed consent and/or specific approval of a recognized ethical committee are required and must be stated in the text", and that authors must only provide anonymized data. The suggested practice is not well aligned with standard globally adopted principles of voluntary participation, as the obtention of both informed consent and approval by an ethics board should generally be documented (Declaration of Helsinki 1975 and subsequent versions, CIOMS 1982 and subsequent versions, UNESCO 1997, etc.). Exceptions to this practice would be determined by a waiver to consent issued by an ethics board, or an explanation letter indicating the absence of an ethics regulatory board at the time.

A recent review highlighted the poor ethics reporting (or compliance) gathered from forensic journals (Bonsu et al.et al. 2022) in the period 2010-2019, but this review was not limited to genetics papers. The journal *Legal Medicine* requires an explicit statement about informed consent, but not about review boards[20]. The *Journal of*

[18] Legislation includes the UK Police and Criminal Evidence Act 1984 ("PACE") that regulates the powers of police in relation to the taking and retention of DNA, Protection of Freedoms Act 2012 ("PoFA"), which amended PACE to introduce a restricted retention framework within which the forensic DNA profile records of individuals who have not been convicted must be expunged from the National DNA Database; the Criminal Law (Forensic Procedures) Amendment Act 37 of 2013 (South Africa); the US DNA Identification Act of 1994; Canada's DNA Identification Act (SC 1998, c 37) and Australia's Crimes Act 1914 (Cth), Part 1D.

[19] The journals with the highest number of contributions are: *Legal Medicine, International Journal of Legal Medicine, Journal of Forensic Sciences, Forensic Science International* and *Forensic Science International: Genetics.*

[20] https://www.elsevier.com/journals/legal-medicine/1344-6223/guide-for-authors, see ethics in publishing

Forensic Sciences shows an identical statement as that of *Legal Medicine*[21] but adds a requirement of compliance with the Declaration of Helsinki. The *International Journal of Legal Medicine* 2012 guidelines request "the description of the ethical considerations" and provides a fictional example with approval by an ethics board in accordance with the Declaration of Helsinki (Poetsch et al 2012). Even though the Declaration of Helsinki implies obtaining informed consent and review or approval by a relevant board, the latter has been mostly absent from published statements. This may be due to a lack of knowledge on the ethics review process, or the research consent requirement formulated in the Declaration of Helsinki, lack of enforcement of ethics practice by the journals, or lack of legislative or ethical frameworks in some countries at the time. In the latter case, a publicly available author statement indicating the ethical or legislative status at the time of sample/data collection is needed.

Current comprehensive guidelines *in Forensic Science International: Genetics (FSIG)* and *Forensic Science International: Reports* (D'Amato et al.2020) are linked to the general Elsevier guidelines, which requires confirmation of informed consent and ethics approval from the authors, similar to the *International Journal of Legal Medicine*.[22]

6.2. Data sensitivity and identifiability

Further concerns regarding the data held on FGFD relate to genetic privacy and the disclosure of sensitive information. To determine the level of protection required for genetic data held in the FGFD, it is important to understand the nature of the genetic data that is being processed, namely:

(1) What kind of information is contained in the data? (Data sensitivity).[23]
(2) Can the data be easily related to a specific person? (Data identifiability).

[21] The author should ensure that the work described has been carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans.

[22] https://www.springer.com/journal/414/submissionguidelines#Instructions%20for%20Authors_Informed%20consent

[23] Sensitive information could be medical or phenotypic information, information about family or even about a person's lifestyle.

The answer to these questions is not the same for the distinct types of data found on the YHRD, EMPOP and STRidER and will therefore be discussed separately below.

6.2.1. Data on STRidER

6.2.1.1. Data sensitivity

STRidER hosts aggregate data for forensic autosomal STR loci with decoupled locus-to-locus information (Bodner et al. 2016). Since no individual genotypes are registered in STRidER, the assessment of the sensitivity of the data is less relevant. In addition, autosomal STRs contain very little to no functionally relevant information. The intronic location of some autosomal markers make them potential candidates subjected to selection pressure. Associations of phenotypic traits detected for a few intronic forensic STRs (Wyner et al 2020) were not identified as causative or predictive, but rather, related to possible regulatory roles or contributed to polygenic traits. The role of STRs as modulators of gene expression has been proposed (Gymrek et al. 2016, Fotsing et al. 2019). However, this would not imply any risk of trait inference from the database, given the aggregated data format used by STRidER.

6.2.1.2. Data identifiability

STRidER only retains aggregated allele frequency data. No individual genotypes are stored, making individual matches or inferences about family members impossible.

6.2.2. Data on EMPOP and YHRD

6.2.2.1. Data sensitivity

The STR markers on the Y chromosome used in forensics carry no medical information *per se*. However, a signal drop-out in the analysis might indicate a chromosomal deletion potentially causative of infertility and which may be unknown to the sample donor (Cooke, 1999; Carvalho et al. 2011, Krausz and Casamonti 2017). The YHRD does not, however, permit explicit searches for such deletions.

With respect to medical information, mtDNA sequences can be considered as more sensitive than Y-STR profiles. However, a clear distinction must be made between the commonly used control region sequences that are highly polymorphic and carry only

sparse medically relevant information and whole mitogenome sequences, encoding more than 30 functionally relevant genes. Mutations in those genes are responsible for a broad variety of diseases and may have high diagnostic value (Habbane et al. 2021). A search via the well-known NIH tools BLAST (Basic Local Alignment Search Tool) and GenBank (an annotated collection of all publicly available DNA sequences) would, however, be more efficient for such a medically motivated candidate search. The curators of EMPOP are aware of the existence of pathogenic variants in their collection (Marshall et al. 2020), albeit at a frequency of less than 0.5%. To avoid using sensitive information in forensic practice, it has been suggested that a filtering of such information in the search algorithms be implemented (Marshall et al 2020).

6.2.2.2. Data identifiability

Y-STR haplotypes are inherited, in most cases, unchanged from father to son. However, mutation rates for modern profiling kits such as YFilerPlus can be more than 1 mutation per 10 generations or in other words: more than 1 out of 10 paternities are expected to show a mutation on the Y chromosome. Due to these high mutation rates, all those larger, modern Y-STR-profiles are rare within a population and simulations suggest that two matching modern Y-STR profiles originate most likely from two men related over less than 20 generations (Andersen and Balding 2017). Y-STR profiles can therefore provide information on kinship, which makes them a potential tool for familial searching. The routine use of mtDNA in forensic casework is not as extensive as the use of STRs and has for a long time been limited to the non-coding control region (CR). However, the progress in sequencing technologies (Canale et al. 2022) and the persistent decline in experimental costs, make the routine use of whole mtDNA information a possibility in the near future. The risk of kinship inferences that could be drawn from mtDNA sequences is smaller compared to Y-STRs, due to lower mutation rates (Andersen and Balding, 2018). As hundreds of copies of every mitogenome are expected to exist in a population, most of them only very distantly related, inferences of family relations would therefore be less reliable. This renders such an approach virtually useless for law enforcement.

The haplotypes registered in EMPOP and YHRD are not linked to personal identifiers, such as name or birthday. Re-identification is therefore highly unlikely and even a warrant from a law enforcement agency, forcing the database curators to surrender all

registered metadata for a certain sample, would be extremely unlikely to lead to the identification of an individual. However, a match on EMPOP or YHRD could point to a particular study and the law enforcement agency might try to obtain a warrant to order the disclosure of the identity of the respective study participants directly from the research institution that conducted the original study. In a parallel situation, a commercial genetic genealogy database in the United States has already been obligated by warrant to disclose the identity of users to a law enforcement agency (Guerrini et. al. 2021). It has also been demonstrated in the past, that people can be identified from Y-STR profiles by database cross-matching; but in those instances, the researchers used larger marker sets and had additional metadata available, such as age and surnames that could be linked with the data to a particular population group with well documented genealogical records (Gymrek et al. 2013).

In other words, the identifiability of data on EMPOP and YHRD depends on the design of the original study by the submitter.[24] However, to assess whether data on the EMPOP and YHRD is considered personal data, it is insufficient that re-identification be a remote or theoretical possibility. If such re-identification is reasonably unlikely, or should not be reasonably expected, then the information will not be considered personal data by most data privacy laws. Recital 26 of General Data Protection Regulation (GDPR) of the EU provides an example of this requirement.[25] Another more universal example is the definition of re-identification in the Guidance Note on Big Data from the United Nations Development Group[26].

Two ways of re-identification were indicated in previous paragraphs, namely by either (1) highly specialized experts in genetics investing a considerable amount of time, such as demonstrated by Gymrek et al. (2013) or by (2) law enforcement agencies through search warrants directed towards the authors of the original studies. Neither approach appears to meet the reasonableness standard: By way of example, re-

[24] Explanatory example: If a match on the YHRD can be connected to a particular study, the respective haplotype was observed only once in this study and the identifying metadata is still available from the submitting lab, the identity of the study participant could potentially be revealed e.g., through a warrant.

[25] The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes." (Recital 26, GDPR).

[26] https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf (accessed May 10, 2022).

identification as demonstrated with Y-STRs by Gymrek et al. (2013) through databases for genetic genealogy is highly unlikely, because YHRD does not hold any meta-data, other than the information from which population study the samples originated. Y-STR profiles are to some degree correlated with patrilinear inherited surnames. Gymrek et al. (2013) used this correlation to infer surnames from public databases holding both surnames and Y-STR profiles. This step was crucial for their successful re-identification attempt but would not be possible with the limited data available on YHRD.

Executing a search warrant in a different country is unusual and generally complex. This, in practice, limits the ability of law enforcement agencies to access original research data to instances where research is conducted, or research data is stored, in that same jurisdiction. Even though the seizure of research data is possible in some countries (e.g., Poland, Surmiak, 2020), it is highly controversial and is only rarely practiced (Drake and Maundrell, 2016). Such a case is pending before the German Constitutional Court (Bögelein et al. 2021) and received significant media attention. In the US, research data can and should be protected from disclosure upon court order by certificates of confidentiality (Wolf et al. 2015). Notwithstanding these examples, identification through the seizure of research data could only be successful if researchers retained personal identifiers of study participants at the research institute, which will often not be the case.

In summary, taking the “reasonable effort” criterium into account, re-identification of individuals from FGFD entries is highly unlikely. The de-identified entries on the FGFDs can therefore be considered anonymized.

However, given that genetic data can always be considered “intrinsically self-identifying" to some degree (Wjst ,2010) and given the increasing capacities of data analytics and the increase in the sharing of anonymized data through open-access models in research (Shabani and Marelli, 2019), database administrators should remain careful and look prospectively at potential future re-identification risks and consider what they can do to provide reasonable protection to their data.

Staying alert is necessary in view of the increasing possibilities of data linkage. Data linkage refers to associating an individual’s data with similar data on other platforms where it is possible to infer additional information about that individual. For instance, numerous submissions of mtDNA sequence data to EMPOP are also available in GenBank. In addition, haplotype data, autosomal allele frequency data, and additional

information about haplogroup or biogeographical origin is directly available through the original publications in scientific journals.

It is anticipated that a growing amount of data will become accessible, with appropriate levels of restriction to protect data privacy, in the near future. The universal trend to ensure the transparency and reproducibility of scientific information has reached mainstream forensic genetics journals, which now require statements relating to data accessibility.[27]

6.3. Inference of biogeographical ancestry

Biogeographical ancestry inference depends on the ancestral groups within the relevant populations. With relation to haplotype databases such as EMPOP or YHRD, the potential biogeographical origin of paternal or maternal ancestors is indicated by the respective haplogroup.[28] With release R67 (Feb 2022), the YHRD curators removed the ancestry function based on SNP-based haplogroup assignment. Previously, this function could be used to infer biogeographical origin to some degree.[29] Autosomal forensic STR markers also show some potential for inference of biogeographical ancestry (Pereira et al. 2011, Phillips et al. 2014, Algee-Hewitt 2016). Ethical issues of biogeographical ancestry inference may therefore arise: a key issue is that so-called “mixed populations” in reference data lead to erroneous test results (Pfaffelhuber et al. 2020) and overly optimistic reliability estimates, and yet they are more likely to be trusted than other evidence (Wienroth et al. 2021). Other ethical concerns relate to the potential risk of racial profiling (Toom et al. 2016).

6.4. Familial genetic information

Haplotype testing may raise ethical concerns relating to the identification of (untested) relatives of the tested individual. In this context, the debate over genetic privacy

[27] IJLM and FSIG (and other Springer and Elsevier journals) require that the data is available in either a public repository or database or data repository with doi identifiier. Most often the access to data in repositories such as Dryad or Zenodo are granted after compliance statements such as non-commercial use of the data, no sharing with 3rd parties and agreement of no disclosure of the data.

[28] For example, Y haplogroup I1 would suggest a paternal biogeographical ancestry in Northern Europe. The haplogroup to which a certain haplotype belongs can e.g., be predicted from Y-STRs by dedicated freely available bioinformatics tools (e.g. https://www.nevgen.org/ or http://www.hprg.com/hapest5/, both accessed on February 8, 2023) or it can be determined in a more targeted way from Y-SNPs or mtDNA sequences. Taking the example of Y haplogroup I1 again, a SNP characterizing this haplogroup branch would be M253 (http://www.phylotree.org/Y/tree/I.htm, accessed on February 8, 2023).

[29] https://yhrd.org/pages/resources/release_history (accessed February 8, 2023).

revolves around the question of familial searching which constitutes an intrusion on the right to privacy of a certain group of individuals — relatives of persons of interest in an investigation — primarily because of their genetic association with someone (Granja and Machado, 2019; Greely et al. 2006; Murphy, 2010; Samuel and Kennett, 2021). The disclosure of sensitive information poses the problem of the latent risk of revealing information that can be implied in the results of familial searching, for example, the potential revelation of genetic information that challenges relationships in ways which might produce disruptive implications in established kinship structures (Haimes, 2006; Kim et al. 2011; Suter, 2010). Therefore, the possibility of using familial genetic information for law enforcement in a manner that is proportionate with the objective of such endeavors must be carefully assessed, including in relation to effective informed consent.

6.5. Consent and representation: controversies

The YHRD and EMPOP contain data sets from vulnerable populations[30] (Lipphardt et al. 2021) such as Roma people or Native Americans. Recently, commentaries suggested that some haplotypes submitted to the online, publicly available YHRD database from minority groups, such as Uyghurs and Roma people, were collected without proper, informed consent (Lipphardt et al. 2021; Normile, 2021; Moreau, 2019; Schiermeier, 2021; Wee, 2019), or that the procedures used to obtain these haplotypes were inadequate (Forzano et al. 2021). The use of unsound ethical practices in collecting information from minorities / vulnerable groups led to the retraction of several papers from peer-review journals (e.g. Zhang et al. 2021 Nothnagel et al. 2022). While commendable and ethically warranted, it should be noted that such an endeavor will contribute to a lack of appropriate population data for these minority groups when estimating the frequency of a haplotype match, leading to

[30] We follow the guidelines provided in the Report of the International Bioethics Commitee of UNESCO (IBC) (2013) that suggest that attempts to define vulnerability in general risk drawing the concept too widely or too narrowly, thereby triggering disputes rather than resolving them. In most cases, however, it is relatively easy to recognize vulnerability when it arises: something fundamental is indeed at stake - human dignity, human rights and fundamental freedoms. The issue of social, political and environmental determinants is particularly complex and involves the fundamental matter of justice in the relations between individuals, groups and States. Many individuals, groups and populations nowadays become especially vulnerable because of factors created and implemented by other human beings, in many cases in blatant violation of fundamental human rights. Social vulnerability is a phenomenon determined by the structure of people's and communities' daily lives. Situations of social vulnerability usually interfere with the self-determination of individuals and lead to significantly increased exposure to risks caused by social exclusion. (https://unesdoc.unesco.org/ark:/48223/pf0000219494)

a lower and therefore adverse frequency for the suspect's haplotype. In other words: the strength of evidence of this match will be overestimated, which is not in the interest of the alleged suspect from a given equity seeking population group (Syndercombe-Court, 2021).

Further, it should be noted that concerns about potential collection of samples without consent do not exclusively pertain to FGFD. Regrettably. the unethical collection of biological samples has been reported regarding data held in distinct types of genetic databases which involved collecting samples from Indigenous communities and minority populations around the world (Claw et al. 2018; Claw et al. 2021; D'Amato et al. 2020; Ortiz-Prado et al. 2020; Garrison, 2013).

7. Ethical principles to be implemented

The *Universal Declaration on the Human Genome and Human Right*s (United Nations, 1997) and the *International Declaration on Human Genetic Data* (UNESCO, 2003) apply to the collection, processing, use and storage of human genetic data and biological samples. The *International Declaration on Human Genetic Data* applies "except in the investigation, detection and prosecution of criminal offences and in parentage testing that are subject to domestic law that is consistent with the international law of human rights".[31]

Prior to these declarations, the ethical practices in biomedical research have been guided by the *Nuremberg Code* 1947 (USGPO 1946-49) and the *Declaration of Helsinki* (WMA 1964 and subsequent versions). Today, it is widely recognized that human genetic data constitutes sensitive personal information (Article 9, GDPR; Wan et al. 2022). A small amount of genetic data can uniquely identify individuals and be predictive of future health conditions, which among other concerns:

(1) may have a significant impact not only on individuals but also on families, and in some instances on whole groups.

[31] As countries develop legislation to govern DNA databases, it is important that civil society is engaged in the debate about what safeguards are needed to protect human rights. The International Declaration on Human Genetic Data, which was adopted unanimously at UNESCO's 32nd General Conference on 16 October 2003, applies to the collection, processing, use and storage of human genetic data and biological samples, "except in the investigation, detection and prosecution of criminal offences and in parentage testing that are subject to domestic law that is consistent with the international law of human rights". Therefore, more clarity is needed about what national legislation and international standards are required to protect such rights. Important principles enshrined in the United Nations Declaration on Human Rights include the right to privacy and a family life (Article 12), equality before the law (Article 7) and the right to the presumption of innocence and a fair trial (Article 11). The United Nations Convention on the Rights of the Child is also relevant to children subject to forensic DNA analysis.

(2) may contain information, the significance of which is not necessarily known at the present time.
(3) may have cultural significance for persons and groups (D'Amato et al. 2020).

In summary, the nature of DNA, including Y chromosome and mitochondrial DNA, requires an understanding by the individual of the implications of providing genetic samples/data not only to themselves, but to their families and wider community. In the following sections, ethical considerations have been submitted to the ISFG, indicating a proposed framework against which the ethics of various data stored on the FGFD can be measured.

7.1. Legacy data

Under this category of datasets, the curators of the FGFD distinguished between:
(1) data from samples collected even decades before the onset of the FGFD; and
(2) data from samples collected with the purpose of contributing information to the FGFD.

YHRD refers to "legacy samples" as those datasets that were submitted "more than 10 years ago", following the German Research Foundation (2019) code of conduct.
For practical purposes, the FDAB suggests adopting temporal categories to data for samples collected before 2021 (as per latest FSIG and FSIR guidelines (D'Amato et al. 2020) and to classify the data in terms of the associated risk (low-medium-high), according to the

1. risk of infringement of ethical principles
2. risk of re-identification
3. source of data/samples
4. provider of data/samples.

For these categorizations, the landmarks used are

1. the *Declaration of Helsink*i (1964, 1975 and later updates),
2. UNESCO *Universal Declaration on the Human Genome and Human Right*s (1997)
3. the *Guidelines for the publication of genetic population data* adopted by the forensic genetics' community (Carracedo et al. 2010 and later updates).

Considerations as to the treatment of data and associated risk category follow below.

7.1.1. Datasets from samples collected pre-1964:

These samples were collected prior to the establishment of the generalized ethical principles in the *Declaration of Helsinki*: *ethical principles for medical research involving human subjects* (WMA, 1964). The identifiability risk for sample donors from this period could be considered negligible after six decades from collection, with most donors likely to be deceased by now. Therefore, a general retrospective risk assessment for datasets derived from those samples is not necessary. However, removal of the dataset could still be warranted if misconduct upon sample collection or data generation is reasonably suspected.

7.1.2. Datasets from samples collected between 1964 and 1997:

The *Declaration of Helsinki* (WMA 1964) first established the principles of free informed consent and it was first amended in 1975 adding a requirement of approval by an ethics review committee (WMA 1975). The principles outlined by the Declaration of Helsinki were already well recognized in the biomedical field before 1997. However, no comparable universal framework, specifically dedicated to non-medical human genetic research, existed. The *UNESCO Universal Declaration on the Human Genome and Human Rights* (1997) is the first official declaration on a global scale,[32] establishing the principle of informed consent for genetic research and encouraging states to require that all human genetics research projects be approved by ethics committees.

Even though it can be assumed that most samples collected in this period were collected with the consent of the donors, the practice of including statements of consent and approval by ethics boards in publications was not widely established. In addition, a substantial proportion of samples from this period appear to have been re-purposed from their original aim to genotyping the data with current forensic markers, given that the first FGFD was only established in 1999.

[32] The HUGO ELSI committee included these principles in the Statement on the Principled Conduct of Genetics published in 1995, available at: http://hrlibrary.umn.edu/instree/geneticsresearch.html. We adopted the UNESCO 1997 declaration for the more comprehensive framework provided.

Need for a broader community debate:

The FDAB acknowledges the difficulty in such a complex policy framing exercise that would provide an ideal framework to satisfy all theoretical scenarios and was not able to reach a consensus as to whether datasets derived from samples collected before 1997 should undergo a general risk assessment. Two different standpoints could be adopted depending on whether:

(1) the ethical standards laid for biomedical research (*Declaration of Hel*sinki) are applied analogously to research on human population genetics (datasets are categorized as high risk if not collected in accordance with the valid *Declaration of Helsinki* principles at the time of collection);

(2) the guidelines stated by the publishers of the data at the time the samples were collected are applied, since no explicit universal guidelines for non-medical population genetics existed, (low risk of having contravened the guidelines, due to lower standards at that time).

The FDAB recommends inviting the wider forensic community to provide feedback on this specific issue for further consideration.

7.1.3. Datasets from samples collected between 1997-2009:

These samples were mostly collected with the specific purpose of forensic statistics applications, after the *UNESCO Universal Declaration on the Human Genome and Human Rights* and before the first explicit ethical guidelines in a forensic genetics journal (Carracedo et al. 2010) were published, requesting informed consent and/or specific approval of a recognized ethical committee and anonymized data. This period can be considered as a transition phase for bioethics in forensic genetics.

Locally, the enforcement of ethical principles often depends on the respective legislation in place. While very influential, the UNESCO Declaration has no binding character. The only binding international treaty protecting human rights in biomedicine is the Oviedo Convention of the Council of Europe, adopted in 1999 (only binding for 29 states, having ratified it). Given the time needed for adoption into domestic legislation, ethics boards might not yet have been established in many countries in the late 1990s to early 2000s.

Notwithstanding the persisting heterogeneity in ethics regulations at the domestic level during this period, ethical requirements of consent and ethics review for human genetics research were sufficiently visible in international texts which diligent, ethical,

data collectors should have been aware of. An example of early international guidance principles are those issued by the International *Human Genome Organisation* (HUGO) ethics committee, which played a key role in the development of early ethical norms on genomic research. HUGO issued their first *Statement on the Principled Conduct of Genetics Research,* setting out the principles of informed consent, confidentiality, and review by an ethical committee by the end of 1995 (HUGO 1995).

In 1998 HUGO adopted a *Statement on DNA Sampling: Control and Access* (Knoppers et al.1998). The statement provides some indication of how to address the case of pre-UNESCO legacy collection, and recommends that:

- ‘‘Routine samples, obtained during medical care and stored, may be used for research if: there is general notification of such a policy, the patient has not objected, and the sample to be used by the researcher has been coded or anonymized. Routine samples obtained during medical care and stored before such notification of such a policy may be used for research if the sample has been anonymized prior to use.”
- “Research samples obtained with consent and stored may be used for other research if there is general notification of such a policy, the participant has not yet objected, and the sample to be used by the researcher has been coded or anonymized. For the use of research, samples obtained before notification of a policy may be used for other research purposes if they have been coded or anonymized prior to use.’’

This statement reflects the widespread belief at that point in time that legacy samples, once anonymized or coded, no longer needed to be accompanied by additional documentation. Ethics guidelines issued in 2003 by the European Society of Human Genetics (ESHG) (ESHG, 2003) also adopted this position. However, with the implementation of the GDPR in Europe in 2018, the notion of what samples/data could be considered sufficiently protected to waive documentation has become more restrictive.

The risk of having contravened ethical practices upon data collection in this period should be categorized as high, medium, or low. Upon randomly screening database entries, it was noted that a considerable proportion of samples collected in this period only refer to informed consent or compliance with the *Declaration of Helsink*i with no

explicit reference to an ethics board (WMA 1975 and subsequent versions[33] call for both informed consent and approval by ethics board/s). Vague and imprecise statements regarding informed consent related to datasets submitted in this period should be considered as signaling a potential high-risk that the sample collection or data acquisition did not follow ethics requirements. Missing statements regarding ethics board approval should only be classified as medium risk. However, the risk assessment should also give special consideration to the recommendations in place concerning the re-use of de-identified data from previous sample collections, such as the ones from the HUGO ethics committee or ESHG.

7.1.4. Datasets from samples collected between 2010-2020:

These samples were collected at a point in time when standardized national and international regulations, and/or explicit indications in journals on submission guidance (Carracedo et al. 2010) were well known. However, it has been noted that the enforcement/recording of ethical practices was poor during this period (Bonsu et al. 2022). Submissions with incomplete statements or records, such as consent alone, no ethics committee approval, or no explanation of why ethics committee approval is not necessary, and submissions with no tangible record of consent, such as an oral communication, should be considered high-risk.

Table 1. Scheme for assessment of risk according to ethics compliance

Samples temporal categories	**Ethics practices**		***Risk Assessment***
Period	***Informed Consent***	***Ethics Board/s***	
Prior to 1997	No consensus reached by the FDAB board.		
1997-2010	No	No	High
	Yes	No	Medium
	Yes	Yes	Low
2010-2020	No	No	High
	Yes	No	High
	Yes	Yes	Low

[33] All versions are available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (accessed February. 2023).

7.2. Contemporary and future population studies

Samples collected from 2021 onwards are regulated by the Guidelines and Recommendations published in 2020 in the journals *Forensic Science International: Genetics* and in *Forensic Science International: Reports* (D'Amato et al. 2020). These guidelines require that recorded informed consent and recorded approval by ethics board/s be disclosed, code / number of the approved revision provided, as well as an indication of the date of collection, or alternatively, disclosure of a waiver or justification for not adhering to the above. The latter becomes crucial in order to place the collections in a temporal framework for risk assessment as suggested in Table 1.
These ethical requirements should be stated on the FGFD websites, as academic submissions are expanding the peer review publications to non-forensic journals.

7.3. Special categories: datasets from non-academic contributors

These types of datasets are listed as direct submissions in the FGFD (see examples in section 6.1). However, a few direct submissions are also found among early legacy samples, subjected to peer-review publication before the onset of the FGFD. The collection of samples that are not regulated by ethics boards but subjected to national legislation should be carefully considered. Some of those datasets might have been generated based on domestic legislation, without a requirement to obtain informed consent. While applying a simple blanket principle to all samples in this category is not possible, article 9 of UNESCO 1997 provides that "limitations to the principles of consent and confidentiality may only be prescribed by law, for compelling reasons within the bounds of public international law and the international law of human rights".

7.4. Special categories: vulnerable groups and minorities

A generally applicable authoritative definition of a 'minority group' is proposed in the 1977 study prepared for the UN-Commission on Prevention of Discrimination and Protection of minorities by Francesco Capotorti (Capotorti, 1977) as:
"A group numerically inferior to the rest of the population of a State, in a non-dominant position, whose members - being nationals of the State - possess ethnic, religious or linguistic characteristics differing from those of the rest of the population and show, if

only implicitly, a sense of solidarity, directed towards preserving their culture, traditions, religion or language."

The protection of vulnerable groups[34] and minorities raise specific ethical concerns and therefore proper ethical practices in the collection of these samples must be guaranteed (Jackson-Preece, 2014). Full compliance with the 2020 guidelines (D'Amato et al. 2020), and in addition, special consideration to the UNESCO 1997 Art.19a (i) should be taken. In cases of organized minorities (i.e., that have their own councils and legal representatives), a group consent to participate in research prior to individual consent may be mandatory (e.g., SAN Code of Ethics). Any missing information regarding ethics compliance must be regarded as high-risk, regardless of the temporal categories described above.

7.5. Factors to consider for risk assessment of data held on the FGFD

For every dataset that is in the medium to high-risk range according to the table above, the FDAB strongly recommends an individual, documented risk-benefit analysis based on the principles stated in Table 2. Results with a negative outcome in the risk-benefit analysis should be removed from the databases.

Table 2. Risk factors to assess

Scientific value	- Datasets of low scientific value should not be retained if in the high-risk range. This includes data with poor collection record, suspected sampling bias, modified political boundaries since collection for national datasets, etc. - For datasets with high scientific value, an individual risk-benefit analysis is necessary. High scientific value could e.g., be due to: i) Scarce data for the population group of interest. ii) Exceptional marker set studied
Data submitters	- For data from non-academic submitters a higher risk of contravening ethics principles can be assumed. - For direct submissions without prior control through a publication process a higher risk can be assumed.

[34] For a definition of vulnerable groups, see footnote 25.

Lack of ethics board approval	- Samples collected after 2010 with no statement on ethics board approval are at high risk (unless exempted by national legislation e.g., the United States when samples are de-identified). - For samples taken based on legal provisions that did not require ethics board approval and/or consent, a higher risk than for samples with ethics board approval can be assumed.
Source of the data collection	- Data with an undefined submitter should be discarded. - For samples purchased or otherwise provided from biobanks a higher risk can be assumed. First, the samples might have been collected without informed consent, but with blanket consent to research in general. Second, those samples might be distributed widely for the assessment of different research topics, facilitating re-identification through data linkage.
Study design and type of data	- Does the study design or type of data entail an elevated risk of re-identification? Possible contributing factors include: i) Many markers analyzed ii) Poor pseudonymization[35] iii) Submission of data from the same source also held in other data repositories iv) Poor privacy rights in the respective country
Vulnerability of the study population	- Does the study include samples from vulnerable individuals or groups (e.g., ethnic minorities, prisoners, patients)?

For an operational risk assessment of entries or datasets in the FGFD, an evaluation scheme has been proposed in Appendix 1. Possible alternative actions for samples with detected risk are developed below.

[35] According to the GDPR Art 4(5), pseudonymization is " the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

7.6. Strategies for medium to high-risk datasets

It is recommended that custodians of the FGFD consider discarding datasets, requiring additional documentation and/or retaining data based on a case-by-case assessment of risk, informed by the above framework. Alternatively, in lieu of discarding a medium- to high-risk dataset, a retrospective waiver for consent and ethics approval or clearance can be requested from the relevant ethics regulatory board(s) in respect of samples collected under higher risk. If the samples were collected in more than one country, the waiver should be requested in all relevant countries. The request may indicate that the process of collection of samples for which individual information is available in the FGFD does not contravene the basic ethical principles indicated in sections above.

If a retrospective waiver is requested, authors must also consider obtaining an explicit agreement from self-organized vulnerable population groups through their representatives, to retain these datasets in the databases. The latter should be considered at an equal (or even higher) level of importance as obtaining a retrospective waiver from an ethical review board. The decision of retention of information of vulnerable groups and minorities should be carefully assessed, with appropriate weight given to the benefit to these societies (a fairer estimation of the weight of evidence) and ideally, involve them in the process of assessment. Their representation must be carefully evaluated, as minorities distributed across several jurisdictions, may exhibit strong population differentiation across borders (e.g., Roma: Martinez-Cruz et al. 2015; Ena et al. 2022, Halilović et al. in press). A further consideration is that, discarding information from the FGFD may contribute to making minority populations more vulnerable in courts of law, as the weight of the evidence against them would be inflated in case of a match, should the minority population not be represented in the relevant FGFD.

7.7. Additional considerations for curators of the FGFD

(1) To maintain transparency, exercise due diligence and promote public trust, database users should be registered. Registrations should be linked to a valid e-mail address or mobile service, and access and use of the databases should be conditional upon accepting the terms of use of the FGFD. The purpose of

accessing the database (e.g., research, private practice, education and/or law enforcement) should be collected prior to the search and unlinked from the user's identity. Summary statistics on the purpose of access should be published on an ongoing or regular basis.

(2) Public disclosure of contributors and their affiliations for all entries.

(3) Assessment of vulnerability of the study population by the curators with availability of this information reflected on the FGFD websites. To assist with this type of assessment, a checkpoint process could be implemented which requires: 1. a declaration from submitters on the content of submission (i.e., that it contains data from minorities or vulnerable groups), 2. an additional check by curators for the presence of widely known minorities, and 3. permission granted to users to report to curators concerns regarding datasets of minorities or vulnerable groups if observed in the FGFD websites.

(4) Example/s of acceptable consents could be provided by the database curators as a guideline to submitters, including agreement to make sample contributor's own DNA results/ profiles/ etc. available in public databases and/or subjected to scientific research. Such examples should be available for download.

(5) Given that biogeographical ancestry (BGA) analysis for the purpose of law enforcement is often controversial and even prohibited in some countries, inference of biogeographical ancestry by maps or by listing matches in all available metapopulations and/or national populations, should be accompanied by a disclaimer in the terms and conditions, indicating that biogeographical ancestry assessment might be illegal in some countries.

(7) Be transparent about the processes of data collection and governance, including granting or denial of access to law enforcement agencies and about access criteria for database users.

(8) Conduct regular independent audits for compliance with ethical standards.

(9) Note the importance of FGFD influence in ensuring good research practices in forensic genetics and, through that, promoting fairer justice systems more broadly. Given the extensive amount of work involved in FGFD assessment such as this one, the ISFG should consider the feasibility of seeking support (financial, technical, or administrative) to FGFD data custodians to ensure that the necessary resources are available to comply with the FDAB's considerations

(10) The FGFD curators should provide periodic reports to the ISFG and FDAB detailing their experiences in adopting these considerations the adopted measures in place, any 'in progress' / non-adopted measures as well as any difficulties encountered.

7.8. Areas for further ethical consideration

The FDAB group urges the curators of the FGFD to undertake periodic ethics assessment which should cover a broad range of evolving areas of interest to the ISFG, including:

(1) Transparency in areas such as greater traceability of ethics approvals and requests for law enforcement access.
(2) Consistency in ethical processes globally, including drawing attention to irregularities in genetic sample collection which may threaten human dignity and damage the trust citizens have placed in genetics.
(3) Function creep related to the scope of consent for donors in areas such as broader commercial or law enforcement use.
(4) Ongoing monitoring of re-identifiability threshold and techniques applicable to the data types they are hosting on their database so that vulnerable datasets can rapidly be put in controlled access and be provided the protection warranted by privacy legislation.
(5) New accountability deficits, including limited professional guidance on international sharing of genetic data in the forensic field.

8. Conclusion

While it cannot be disputed that a core purpose of FGFD is to promote more effective justice and protection of human society, if the collection, retention, and disclosure of individual's data on the FGFD compromises individual human rights and research ethics norms, including for vulnerable and minority groups, it has failed to achieve that purpose. The approach outlined in this report seeks a way forward that balances these

complex interests, ensuring the ethical principles of data collection, retention and disclosure are followed while maintaining the purpose and application of the FGFD where possible. Notwithstanding the recommendations made in this report, a deeper debate is required on the role and acceptable alternatives to informed consent, the impact of including datasets of minorities and vulnerable groups, privacy and identifiability risks of various datasets and other relevant considerations for a risk-benefit assessment of data hosted on the FGFD.

Such important challenges require the consideration of a broader ethical and privacy assessment approach to FGFD, such that the data deposition and data access processes are developed in partnership with relevant professionals and stakeholders, or through external audits, rather than placing ethics responsibility solely on the FGFD users and/or curators. The current limitation of focus also narrows ethical discussion by shielding other substantive political and societal issues from critical scrutiny, such as public interest issues, societal good, and public trust in genetics. The aim of such interventions is to help render beneficial technologies socially acceptable rather than shutting down viable lines of development because of claims of their potential ethical or wider social hazards.

While novel and oncoming technologies will pose additional ethical challenges, the need to deal with past and current issues has motivated the drafting of this report. This report also lays a foundation to deal with future challenges, which need to be seen as a moving target requiring constant monitoring and adapting to emerging legal, ethical, and social issues. It is furthermore vital for the forensic genetics community to develop an understanding of the various social and ethical concerns associated with genetic technologies. These and other ethical challenges relevant to the FGFD, necessitate a continuous discussion, this first report representing a starting point in such a journey.

References:

Algee-Hewitt B.F.B., Edge M.D., Kim J., Li J.Z., Rosenberg N.A. 2016.Individual identifiability predicts population identifiability in forensic microsatellite markers. *Current Biology* 26, 935-942.

Amorim A., Fernandes T., Taveira N. 2019. Mitochondrial DNA in human identification: a review. PeerJ 7, e7314. http://doi.org/10.7717/peerj.7314.

Andersen M.M., Balding D.J. 2017. How convincing is a matching Y-chromosome profile?, PLOS Genetics 13(11), e1007028. https://doi.org/10.1371/journal.pgen.1007028.

Andersen M.M., Balding D.J. 2018. How many individuals share a mitochondrial genome?, PLOS Genetics 14(11), e1007774. https://doi.org/10.1371/journal.pgen.1007774.

Anderson M.M., Eriksen P.S. and Morling N. 2013. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. J. Theor. Biol. 329, 39–51 https://doi.org/10.1016/j.jtbi.2013.03.009.

Balding D.J, and Steel C.D. 2015. Weight-of-evidence for forensic DNA profiles. 2nd Edition. John Wiley & Sons Ltd, Chichester, UK.

Bodner M., Bastisch, I., Butler, J.M., Fimmers, R., Gill, P., Gusmão, L., Morling, N., Phillips, C., Prinz, M., Schneider, P.M., Parson, W. 2016. Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)., Forensic Sci. Int. Genet. 24, 97-102 DOI: 10.1016/j.fsigen.2016.06.008.

Bodner M. & Parson W. 2020. The STRidER Report on Two Years of Quality Control of Autosomal STR Population Datasets. *Genes,* 11**,** 901. https://doi:10.3390/genes11080901.

Bögelein N., Golla S., Lehmann L., & Leimbach K. 2021. When the Police Are at the Door and Want the Interview Data: Situating the Ethics and Law of Qualitative Radicalization Research. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 22(3). https://doi.org/10.17169/fqs-22.3.3681.

Bonsu D, Afoakwah, C, Abedi M, Higgins D, Austin J. 2022. Ethics Reporting in Forensic Science Research Publications – A Review, Forensic Science International, 2022,111290, https://doi.org/10.1016/j.forsciint.2022.111290.

Bright J.-A., et al. 2019. "The interpretation of forensic DNA profiles: an historical perspective." Journal of the Royal Society of New Zealand, 1-15.

Butler J. M. 2010. Fundamentals of DNA typing. Academic Press.

Canale L. C., Parson W., & Holland M. M. 2022. The time is now for ubiquitous forensic mtMPS analysis. Wiley Interdisciplinary Reviews: Forensic Science, 4(1), e1431. https://doi.org/10.1002/wfs2.1431.

Caportoti F. 1977. The International Protection of Persons Belonging to Ethnic, Religious and Linguistic Minorities since 1919. United Nations Economic and Social Council.

Carracedo A, Butler JM, Gusmão L, Parson W, Roewer L, Schneider PM. 2010. Publication of population data for forensic purposes. Forensic Sci Int Genet. 4(3),145-147. doi: 10.1016/j.fsigen.2010.02.001.

Carvalho C.M., Zhang F., Lupski J.R. 2011. Structural variation of the human genome: mechanisms, assays, and role in male infertility. Syst. Biol. Reprod. Med. 57, 3–16 https://doi.org/10.3109/19396368.2010.527427.

Chaitanya L., Van Oven M., Weiler N., Harteveld J., Wirken L., Sijen T., De Knijff P. Kayser M. 2014. Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level. Forensic Sci Int Genet, 11, 39-51. https://doi:10.1016/j.fsigen.2014.02.010.

Claw K.G, Dundas N., Parrish M.S., Begay R.L., Teller T.L., Garrison N.A., Sage F. 2021. Perspectives on Genetic Research: Results From a Survey of Navajo Community Members. *Front Genet.*, 12, 7345292021 https://doi:10.3389/fgene.2021.734529.

Claw K.G., Anderson M.Z., Begay R.L. *et al.* 2018. A framework for enhancing ethical genomic research with Indigenous communities. *Nat Commun* 9, 2957 https://doi.org/10.1038/s41467-018-05188-3.

Cooke H. 1999. Y chromosome and male infertility. Rev. Reprod. 4, 5-10 https://doi.org/10.1530/ror.0.0040005.

Coppola L, Cianflone A, Grimaldi AM, et al. 2019. Biobanking in health care: evolution and future directions. *J Transl Med*. 17(1),172. doi:10.1186/s12967-019-1922-3.

Council for International Organizations of Medical Sciences (CIOMS); 1982. Proposed International Guidelines for Biomedical Research Involving Human Subjects. Geneva, 1982.

D'Amato M. E., Bodner M, Butler, J. M, Gusmão L., Linacre A, Parson W. Schneider, P. M, Vallone, P, Carracedo, A. 2020. Ethical publication of research on genetics and genomics of biological material: guidelines and recommendations, Forensic Science International: Genetics, 20:102299, doi:10.1016/j.fsigen.2020.102299 , and Forensic Science International: Reports, 2, 100091. https://doi.org/10.1016/j.fsir.2020.100091.

Drake K. & Maundrell R., 2016. Researcher-Participant Privilege, Confidentiality, and the Jailhouse Blues. McGill JL & Health 10,1.

El-Haj N.A. 2007. The genetic reinscription of race. Annual Review of Anthropology 36(1), 283–300.

Ena G.F., Aizpurua-Iraola J, Font-Porterias N, Calafell F, Comas D. 2022. Population Genetics of the European Roma - A Review, Genes 13(11), 2068.

European Society of Human Genetics. 2003.Data storage and DNA banking for biomedical research: technical, social and ethical issues. Eur J Hum Genet 11, S8–S10. https://doi.org/10.1038/sj.ejhg.5201115.

Forzano F, Genuardi M, Moreau Y. 2021. ESHG warns against misuses of genetic tests and biobanks for discrimination purposes. Eur J Hum Genet 29, 894–896 https://doi.org/10.1038/s41431-020-00786-6.

Fotsing S. F, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. 2019. The impact of short tandem repeat variation on gene expression. Nature genetics, 51(11), 1652-1659. https://doi.org/10.1038/s41588-019-0521-9.

GA4HG 2019. Global Alliance for Genomics and Health: Data Privacy and Security Policy. Available at: https://www.ga4gh.org/wp-content/uploads/GA4GH-Data-Privacy-and-Security-Policy_FINAL-August-2019_wPolicyVersionsUPDATED.docx-2.pdf

Garrison N.A. 2013. Genomic Justice for Native Americans: Impact of the Havasupai Case on Genetic Research. Sci Technol Human Values 38(2), 201-223. doi: 10.1177/0162243912470009.

German Research Foundation. 2019. Guidelines for Safeguarding Good Research Practice: Code of Conduct. Doi: 10.5281/zenodo.3923602.

Granja R, Machado, H. 2019. Ethical controversies of familial searching: The views of stakeholders in the United Kingdom and in Poland, Science, Technology & HumanValues, 44(6), 1068-1092. https://doi.org/10.1177/0162243919828219

Greely H, Riordan D, Garrison N, Mountain, J. 2006. Family Ties: The Use of DNA Offender Databases to Catch Offenders' Kin, Journal of Law, Medicine & Ethics 34(2), 248-262. DOI: 10.1111/j.1748-720X.2006.00031.x

Gusmão L, Butler JM, Linacre A, Parson W, Roewer L, Schneider PM, Carracedo A. 2017. Revised guidelines for the publication of genetic population data, Forensic Sci Int Genet 30, 160-163. https://doi.org/10.1016/j.fsigen.2017.06.007

Guerrini C J, Wickenheiser R, Bettinger B, McGuire A. L, Fullerton S. 2021. Four misconceptions about investigative genetic genealogy, Journal of Law and the Biosciences,8(1), lsab001, https://doi.org/10.1093/jlb/lsab001

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. Science.339(6117),321-4. doi: 10.1126/science.1229566. PMID: 23329047.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly M. J, Price A. L, Pritchard J. K, Sharp A. J, Erlich Y. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*, *48*(1), 22–29. https://doi.org/10.1038/ng.3461.

Habbane M, Montoya J, Rhouda T, Sbaoui Y, Radallah D, Emperador S. 2021. Human Mitochondrial DNA: Particularities and Diseases. Biomedicines 9(10),1364. https://doi.org/10.3390/biomedicines9101364.

Haimes E. 2006. Social and Ethical Issues in the Use of Familial Searching in Forensic Investigations: Insights from Family and Kinship Studies. Journal of Law, Medicine & Ethics 34(2), 263-76. DOI: 10.1111/j.1748-720X.2006.00032.x.

Halilović E, Ahmić A, Kalajdžić A, Ismailović A, Čakar J, Lasić L, Pilav A, Džehverović M, Pojskić N. 2022. Paternal genetic structure of the Bosnian-Herzegovinian Roma: A Y-chromosomal STR study. American Journal of Human Biology 34:e23719.https://doi.org/10.1002/ajhb.23719

HUGO, Statement on the Principled Conduct of Genetics Research, December 1995, Eubios Journal of Asian and International Bioethics,:59-60. https://www.eubios.info/HUGO.htm

Jackson-Preece J. 2014. Beyond the (Non)definition of minority, ECMI Brief # 30, European Centre for Minority Issues (ECMI), available on https://www.files.ethz.ch/isn/177881/issue%20brief%20nr.30.pdf

Jobling M.A. & Tyler-Smith C. 1995. Fathers and sons: the Y chromosome and human evolution. Trends Genet. 11, 449–456. https://doi.org/10.1016/S0168-9525(00)89144-1

Jobling M.A, Pandya, A,Tyler-Smith C. 1997. The Y chromosome in forensic analysis and paternity testing. Int. J. Leg. Med. 110 118-124. https://doi.org/10.1007/s004140050050

Kayser M. 2017. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. ;136(5),621-635. doi: 10.1007/s00439-017-1776-9.

Kayser M., Caglià A., Corach D., Fretwell N., Gehrig C., Graziosi G., Heidorn F., Herrmann S., Herzog B., Hidding M., Honda K., Jobling M., Krawczak M., Leim K., Meuser S., Meyer E., Oesterreich W., Pandya A., Parson W., Penacino G., Perez-Lezaun A., Piccinini A., Prinz M., Schmitt C. and Roewer L. (1997), 'Evaluation of Y-chromosomal STRs: a multicenter study.', Int J Legal Med 110(3):125-133, 141-149. https://doi.org/10.1007/s004140050051

Kayser M & de Knijff P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics 12(3): 179-192. https://doi.org/10.1038/nrg2952.

Kayser M. & Sajantila A. 2001 Mutations at Y-STR loci: implications for paternity testing and forensic analysis. Forensic Sci. Int. 118, 116–121 https://doi.org/10.1016/S0379-0738(00)00480-1.

Kayser M. & Schneider P. 2009. DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. Forensic Science International: Genetics 3(3), 154–161.

Kim J, Mammo D, Siegel M, Katsanis S. 2011. Policy Implications for Familial Searching. Investig Genet. 2(1), 1-22. doi: 10.1186/2041-2223-2-22.

Knoppers BM, Hirtle M, Lormeau S, Laberge CM, Laflamme M. 1998. HUGO Ethics Committee. HUGO Ethics Committee Statement on DNA sampling: control and access. Genetic Resour. 11(2), 43-44. Available at: http://hrlibrary.umn.edu/instree/dnastatement.html.

Krausz C, Casamonti E. 2017. Spermatogenic failure and the Y chromosome. Hum Genet 136, 637–655 https://doi.org/10.1007/s00439-017-1793-8.

Lipphardt V, Surdu M, Ellebrecht N, Pfaffelhuber P, Wienroth M, Rappold G. A. 2021. Europe's Roma people are vulnerable to poor practice in genetics, Nature599, 368-371. https://www.nature.com/articles/d41586-021-03416-3.

Marshall C, Sturk-Andreaggi K., Ring J. D., Dür A., & Parson W. 2020. Pathogenic Variant Filtering for Mitochondrial Genome Haplotype Reporting. *Genes*, *11*(10), 1140. https://doi.org/10.3390/genes11101140.

Martínez-Cruz B, Mendizabal I, Harmant C, de Pablo R, Ioana M, Angelicheva D, Kouvatsi A, Makukh H, Netea MG, Pamjav H, Zalán A, Tournev I, Marushiakova E, Popov V, Bertranpetit J, Kalaydjieva L, Quintana-Murci L, Comas D, Genographic-Consortium. 2016. Origins, admixture and founder lineages in European Roma, Eur J Hum Genet 24(6):937-43 DOI: 10.1038/ejhg.2015.201.

Moreau Y. 2019. Crack down on genomic surveillance. Nature 576, 36-38. https://doi.org/10.1038/d41586-019-03687-x.

Murphy E. 2010. Relative Doubt: Familial Searches of DNA Databases. Michigan Law Review 109 (3), 291-348.

Nothnagel, M., Fan, G., Guo, F. *et al.* 2022. Retraction Note to: Revisiting the male genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes. *Hum Genet* **141**, 175–176 https://doi.org/10.1007/s00439-021-02413-w . Available at: https://link.springer.com/article/10.1007/s00439-021-02413-w.

Normile, D. 2021.Genetic papers containing data from China's ethnic minorities draw fire. Science Insider doi: 10.1126/science.abl8764.

Ortiz-Prado E, Simbaña-Rivera K, Gómez-Barreno L. Tamariz L, Lister A, Baca J. C, Norris A, Adana-Diaz, L. 2020. Potential research ethics violations against an

indigenous tribe in Ecuador: a mixed methods approach. BMC Med Ethics 21**,** 100 https://doi.org/10.1186/s12910-020-00542-x.

Pereira L, Alshamali F, Andreassen R, Ballard R, Chantratita W, Cho NS, Coudray C, Dugoujon JM, Espinoza M, González-Andrade F, Hadi S, Immel UD, Marian C, Gonzalez-Martin A, Mertens G, Parson W, Perone C, Prieto L, Takeshita H, Rangel Villalobos H, Zeng Z, Zhivotovsky L, Camacho R, Fonseca NA. 2011. PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. Int J Legal Med. 125(5),629-636. doi: 10.1007/s00414-010-0472-2.

Phillips C, Gelabert-Besada M, Fernandez-Formoso L, García-Magariños M, Santos C, Fondevila M, Ballard D, Syndercombe Court D, Carracedo A, Lareu MV. 2014. "New turns from old STaRs": enhancing the capabilities of forensic short tandem repeat analysis. Electrophoresis. 35(21-22), 3173-3187. doi: 10.1002/elps.201400095.

Pfaffelhuber P, Grundner-Culemann F, Lipphardt V, Baumdicker F. 2020. How to choose sets of ancestry informative markers: A supervised feature selection approach. Forensic Sci Int Genet.; 46,102259. doi: 10.1016/j.fsigen.2020.102259.

Poetsch M, Bajanowski T, Pfeiffer H. 2012. The publication of population genetic data in the International Journal of Legal Medicine: guidelines, International Journal of Legal Medicine 126(4), 489-490. https://doi.org/10.1007/s00414-012-0700-z.

Prinz M, Sansone M. 2001. Y chromosome-specific short tandem repeats in forensic casework. Croat Med J, 42, 288-291. PMID: 11387641.

Regulation, Protection. 2016."Regulation (EU) 2016/679 of the European Parliament and of the Council." *Regulation (eu)* 679, 2016. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679#page=38

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A. et al. 2001. Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci. Int. 118, 106-113 https://doi.org/10.1016/S0379-0738(00)00478-3.

Roewer L, Andersen M., Ballantyne J, Butler J, Caliebe A, Corach D, D'Amato ME, Gusmão L, Hou Y, De K, Parson W, Prinz M, Schneider PM, Taylor D, Vennemann M, Willuweit S. 2020, DNA commission of the International Society of Forensic Genetics

(ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis, Forensic Sci. Int. Genet. 48, 102308 http://dx.doi.org/10.1016/j.fsigen.2020.102308

Roewer, L and Willuweit, S (2023) Y Chromosome Databases. In: Max M. Houck (ed.) Encyclopedia of Forensic Sciences, Third Edition, vol. 4, pp. 837–840. Oxford: Elsevier.

Samuel G.& Kennett D. 2021. Problematizing consent: searching genetic genealogy databases for law enforcement purposes. New Genetics and Society, *40*(3), 284-304. https://doi.org/10.1080/14636778.2020.1843149 .

Shabani M & Marelli L. 2019. Re-identifiability of genomic data and the GDPR, EMBO reports 20(6), e48316, https://doi.org/10.15252/embr.201948316.

South African San Institute. 2017. The San Code of Research Ethics. Kimberly, South Africa. Available at: https://www.globalcodeofconduct.org/wp-content/uploads/2018/04/San-Code-of-RESEARCH-Ethics-Booklet_English.pdf.

Schiermeier Q. 2021.Forensic database challenged over ethics of handling. Nature 594, 320-322. doi: https://doi.org/10.1038/d41586-021-01584-w.

SWGDAM 2014. Interpretation guidelines for Y chromosome STR typing. http://media.wix.com/ugd/4344b0_da25419ba2dd4363bc4e5e8fe7025882.pdf [accessed 4 March 2022].

SGDAM 2022. Interpretation Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories. https://www.swgdam.org/_files/ugd/4344b0_bc90bcfef52c43559aa28618ef87c424.pdf [accessed 28 July 2022].http://mathgene.usc.es/snipper/Surmiak A. 2020. Should we Maintain or Break Confidentiality? The Choices Made by Social Researchers in the Context of Law Violation and Harm. *J Acad Ethics* 18, 229-247 https://doi.org/10.1007/s10805-019-09336-2.

Suter S. 2010. All in the Family: Privacy and DNA Familial Searching.Harvard Journal of Law & Technology 23 (2), 309-99.

Syndercombe Court D. 2021. The Y chromosome and its use in forensic DNA analysis. Emerg Top Life Sci.5(3), 427-441. doi: 10.1042/ETLS20200339.

The Nuremberg Code. In: Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10. Nuremberg, October 1946–April 1949. Washington, D.C.: U.S. G.P.O, 1949–1953. Available at: https://www.fhi360.org/sites/all/libraries/webpages/fhi-retc2/Resources/nuremburg_code.pdf; full volumen available at https://collections.nlm.nih.gov/catalog/nlm:nlmuid-01130400RX2-mvpart.

Toom V, Wienroth M, M'charek A, et al. 2016. Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively: Reply to 'Forensic DNA phenotyping: Predicting human appearance from crime scene material for investigative purposes' by Manfred Kayser. *Forensic Science International: Genetics* 22: e1–e4.

UNESCO. International Bioethics Committee. 2013. The principle of respect for human vulnerability and personal integrity: report of the International Bioethics Committee of UNESCO (IBC). UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000219494.

UNESCO. 1997. Universal Declaration on the Human Genome and Human Rights. Available at:https://unesdoc.unesco.org/ark:/48223/pf0000110220.

UNESCO. 2003. International Declaration on Human Genetic Data. Available at: https:// unesdoc.unesco.org/ark:/48223/pf0000133171.

US Department of Health and Human Services. 2017. Federal Policy for the Protection of Human Subjects, section XIV. General Requirements for Informed Consent. Available at: https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects#p-818 and https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html

U.S. Government. Permissible Medical Experiments. In: Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No.10: Nuremberg U.S. Government Printing Office (n.d.), 2,181-182 October 1946–April 1949. Washington.

Wan, Z., Hazel, J.W., Clayton, E.W. *et al.* 2022. Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet* 23, 429–445 https://doi.org/10.1038/s41576-022-00455-y.

Weber-Lehmann J, Schilling E, Gradl G, et al. 2014. Finding the needle in the haystack: differentiating 'identical' twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet 9, 42–46.

Wee S.L. 2019. China uses DNA to track its people, with the help of American Expertise. *The New York Times*, 1–9. Available from: https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html .

Wienroth M, Granja R, Lipphardt V, Nsiah Amoako E, McCartney C. 2021. Ethics as Lived Practice. Anticipatory Capacity and Ethical Decision-Making in Forensic Genetics. Genes 12(12), 1868. https://doi.org/10.3390/genes12121868

Willuweit, S. & Roewer, L. 2015. The new Y chromosome haplotype reference database. Forensic Sci. Int.: Genet. 15, 43-48 https://doi.org/10.1016/j.fsigen.2014.11.024

Wjst M, 2010. Caught you: threats to confidentiality due to the public release of large-scale genetic data sets,BMC MED. ETHICS, 11(1), 21. https://doi.org/10.1186/1472-6939-11-21.

Wolf L. E., Patel M. J., Williams Tarver B. A., Austin J. L., Dame L. A. & Beskow L. M. 2015. Certificates of Confidentiality: Protecting Human Subject Research Data in Law and Practice. J Law Med Ethics, 43, 594-609.

World Medical Association.1964. Declaration of Helsinki: Ethical principles for medical research involving human subjects. Recommendations guiding doctors in clinical research. Available at https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

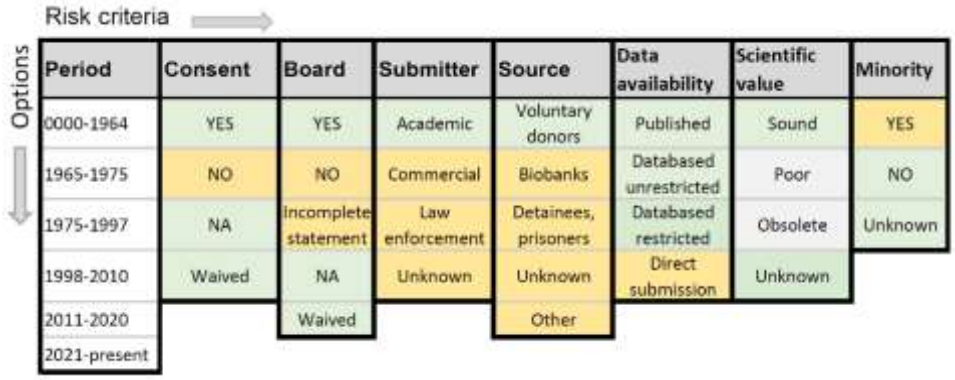
World Medical Association. 1975. Declaration of Helsinki: Recommendations guiding medical doctors in biomedical research involving human subjects. Available at https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Wyner N., Barash M, McNevin D. 2020. Forensic Autosomal Short Tandem Repeats and Their Potential Association with Phenotype. *Frontiers in Genetics,*11, 884. https://doi.org/10.3389/fgene.2020.00884.

Zhang, D., Cao, G., Xie, M. *et al.* 2021. RETRACTED ARTICLE: Y Chromosomal STR haplotypes in Chinese Uyghur, Kazakh and Hui ethnic groups and genetic features of DYS448 null allele and DYS19 duplicated allele. *Int J Legal Med* 135, 1119 https://doi.org/10.1007/s00414-019-02049-6. Available at: https://link.springer.com/article/10.1007/s00414-019-02049-6.

Appendix 1.

The following table shows a series of criteria (Period, Consent, Board, etc) and the possible alternative options for each, in the corresponding columns.



Risk criteria →

Options ↓

Period	Consent	Board	Submitter	Source	Data availability	Scientific value	Minority
0000-1964	YES	YES	Academic	Voluntary donors	Published	Sound	YES
1965-1975	NO	NO	Commercial	Biobanks	Databased unrestricted	Poor	NO
1975-1997	NA	Incomplete statement	Law enforcement	Detainees, prisoners	Databased restricted	Obsolete	Unknown
1998-2010	Waived	NA	Unknown	Unknown	Direct submission	Unknown	
2011-2020		Waived		Other			
2021-present							

An interactive version with further explanations is available at https://uwcacza-my.sharepoint.com/:x:/g/personal/medamato_uwc_ac_za/ETqd-rIiJ2lFotGJz02a_WEBfpmXhaXdrQpJQEQ6oiwm1g?e=TH4UW2

A case-by-case proportionate assessment of existing risk factors and justifications for a given dataset will need to be carried out in order to make this decision. The presence of a risk factor (yellow color) does not necessarily entail the destruction of samples. The color code may change according to specific conditions of the collected dataset.