European Commission

JRC TECHNICAL REPORT

Whole Genome Sequencing and forensics genomics

Angers, A., Drabek, J., Fabbri, M., Petrillo, M. and Querci, M.

2021

Joint Research Centre

EUR 30766 EN

Contact information
Name: Maddalena Querci
Address: Via E. Fermi, 2749 I-21027 Ispra (VA), Italy
Email: maddalena.querci@ec.europa.eu
Tel.: +39 033278-9308

EU Science Hub
https://ec.europa.eu/jrc

JRC125734

EUR 30766 EN

PDF ISBN 978-92-76-40265-7 ISSN 1831-9424 doi:10.2760/864087

Luxembourg: Publications Office of the European Union, 2021

How to cite this report: Angers, A., Drabek, J., Fabbri, M., Petrillo, M. and Querci, M., *Whole Genome Sequencing and forensics genomics,* EUR 30766 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40265-7, doi:10.2760/864087, JRC125734.

Contents

Authors

List of authors in alphabetical order:

Alexandre Angers, Publication Office of the European Union, Long Term Preservation Unit

Jiri Drabek, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Czech Republic

Marco Fabbri, European Commission Directorate General Joint Research Centre, Directorate F – Health, Consumers and Reference Materials, Knowledge for Health and Consumer Safety Unit

Mauro Petrillo, European Commission Directorate General Joint Research Centre, Directorate F – Health, Consumers and Reference Materials, Knowledge for Health and Consumer Safety Unit

Maddalena Querci, European Commission Directorate General Joint Research Centre, Directorate F – Health, Consumers and Reference Materials, Knowledge for Health and Consumer Safety Unit

Abstract

Advances in the massively parallel sequencing will increase the availability of human whole genome sequences. These are being produced by national and international initiatives, healthcare projects, research projects, and even direct-to-consumer genomics companies. This report aims to evaluate the potential impacts, in the field of forensics, of the information generated by whole genome sequencing and the large-scale availability of whole genome sequences.

Thus, this report provides a state of the art of use of large databases of human genome sequences to identify:

The donor of a given genome sequence, mostly in the context of wanted or missing persons.

The physical characteristics of the genome sequence donor, including medical conditions.

The report starts with the definition of the genotyping target, the human genome, and the description of the genotyping methods to analyse it, including massively parallel sequencing. It proceeds with a description of the forensic genetics work processes, quality issues, and logical evidence interpretation. Then, it tackles DNA as Big Data that enables genealogical, biogeographical, phenotypic, and health searches and that requires informed consent and safeguarding against privacy breaches. It then depicts the Prüm Convention as a current model of forensic use of genetic data in European countries. Finally, the report summarizes findings and concerns, and provides final recommendations.

1 Nucleic acids as the genotyping target

1.1 Nuclear and mitochondrial genome

The human genome is a complete set of deoxyribonucleic acid chains, coding for more than 20,000 human genes within the 23 chromosome pairs in the cell nuclei (nuclear genome, g) and in a small DNA molecule found in many copies within individual mitochondria (mitochondrial genome, mt) (Figure 1). Chromosomes 1 to 22 are autosomes, chromosome 23 is a sexual chromosome (gonosome, allosome) in two forms: X and Y.

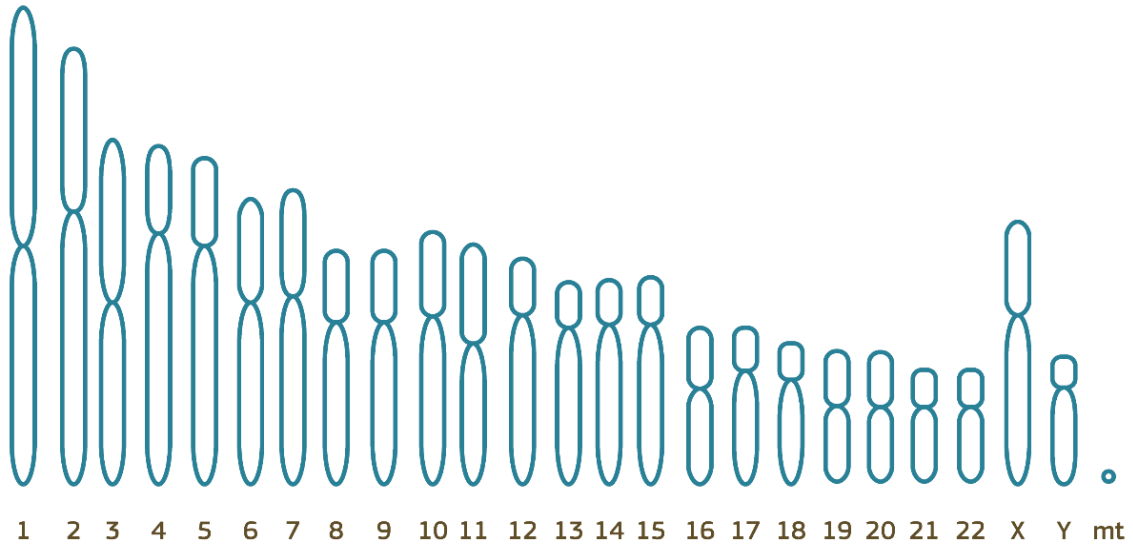

Figure 1: The set of chromosomes in the human genome

Autosomes are not inherited intact from each parent. Instead, each parent´s own pair of chromosomes is randomly combined into a new chromosome that is passed onto the child. Recombination occurs randomly, but nucleotides that are closer to one another on a chromosome are more likely to be inherited together. In contrast, nucleotides that are far apart are more likely to be separated by recombination. The probability of recombination between two nucleotides is quantified as their genetic distance, which is measured in centimorgans (cM). One cM equates to a 1% probability of recombination in a single generation.

Human genomes include both protein-coding and non-coding DNA. Non-coding DNA is known to have evolutionary and chromosome scaffolding functions, and can also serve for forensic purposes. Short tandem repeats, STRs, are an example of length polymorphism used in forensics that are located mostly in non-coding regions. STRs are microsatellites with the number of sequence repeats varying among individuals, effective for human identification purposes (see chapter 2.1 DNA profiling by STR genotyping). Haploid human genomes, which are contained in germ cells (the egg and sperm gamete cells created in the meiosis phase of sexual reproduction before fertilization creates a zygote) consist of on copy of each chromosome, representing $3x10^9$ DNA base pairs. In contrast, diploid genomes (found in somatic cells) have twice the DNA content. Thus, women possess 22 pairs of autosomes, mitochondrial DNA, and one pair of gonosomal chromosomes X, while men possess 22 pairs of autosomes, mitochondrial DNA, one chromosome X and one chromosome Y.

1.2 Ways of genetic information transfer

There are differences in transfer of genetic information from parents to offspring for different parts of the genome: while one copy of each autosome is transferred from the mother and one copy of each autosome is transferred from the father to their children (Figure 2), chromosome X of the father is transferred only to his daughters (Figure 3), mitochondrial DNA is transferred from mother only, but to both her sons and daughters (Figure 4), and chromosome Y is transferred only from father to sons (Figure 5).

Genetic lineage markers (chromosome Y and mitochondrial DNA) have additional specific characteristics. Apart from its pseudoautosomal regions in its very ends, chromosome Y does not recombine. Thus, it is inherited in an agnate way and follows the transfer from biological father to sons (with irregularities due only to mutations).

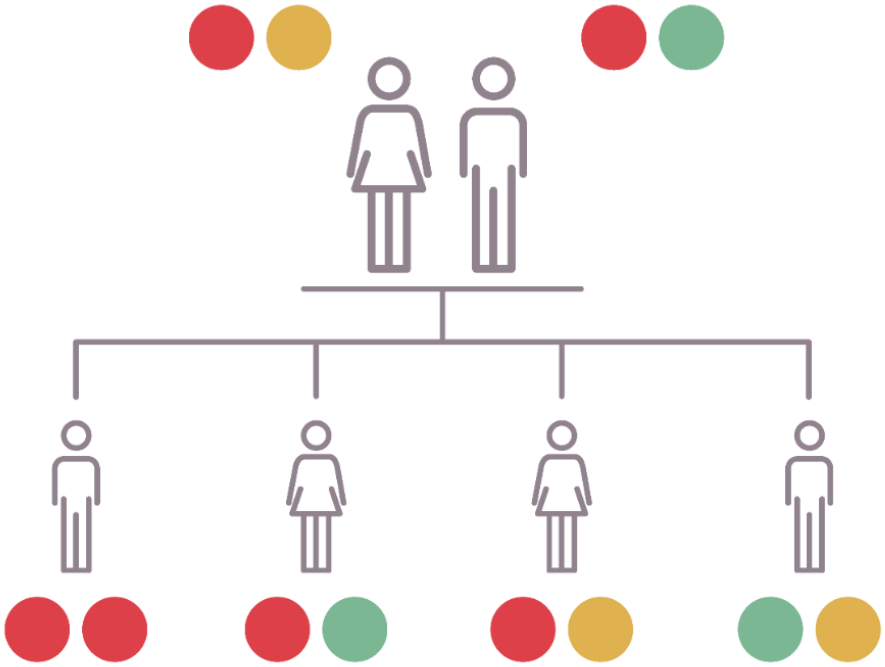

Figure 2: Transfer of autosomal markers between parents and their offspring (Adapted from original image of Halina Šimková with permission of author, https://www.slideserve.com/ronni/slide-pool-forenzn-genetiky-roz-en-soubor-obrazov-ch-sch-mat-k-publikaci)

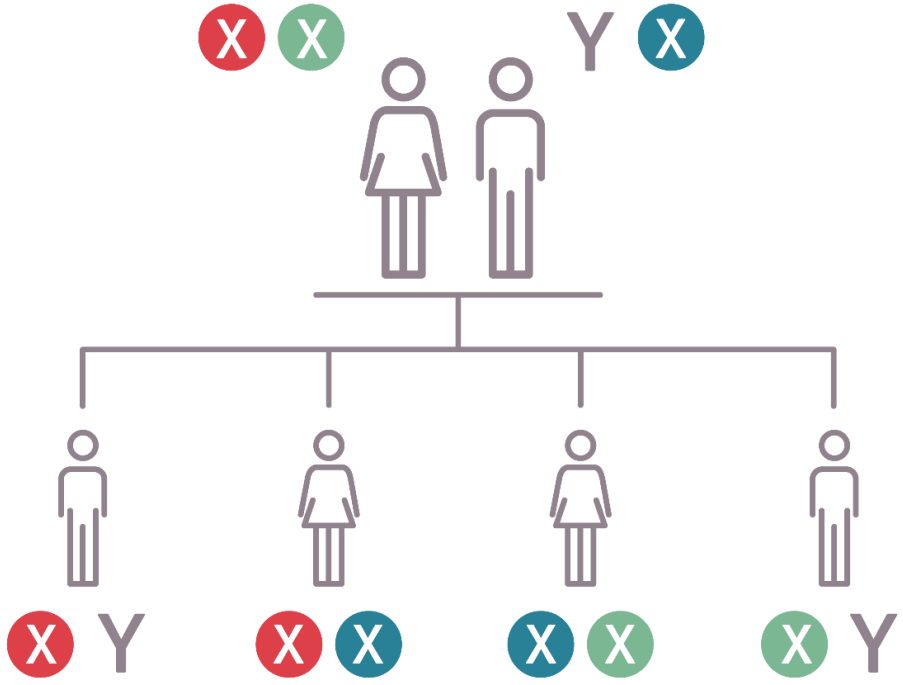

Figure 3: Transfer of the markers on chromosome X between parents and their offspring (Adapted from original image of Halina Šimková with permission of author, https://www.slideserve.com/ronni/slide-pool-forenzn-genetiky-roz-en-soubor-obrazov-ch-sch-mat-k-publikaci)

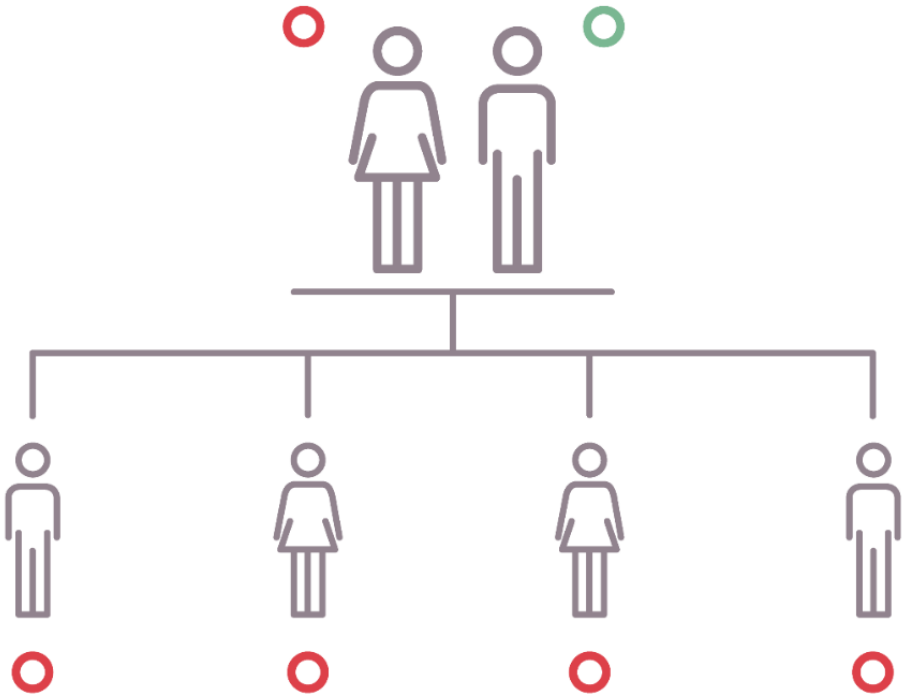

Figure 4: Transfer of markers on the mitochondria between parents and their offspring (Adapted from original image of Halina Šimková with permission of author, https://www.slideserve.com/ronni/slide-pool-forenzn-genetiky-roz-en-soubor-obrazov-ch-sch-mat-k-publikaci)

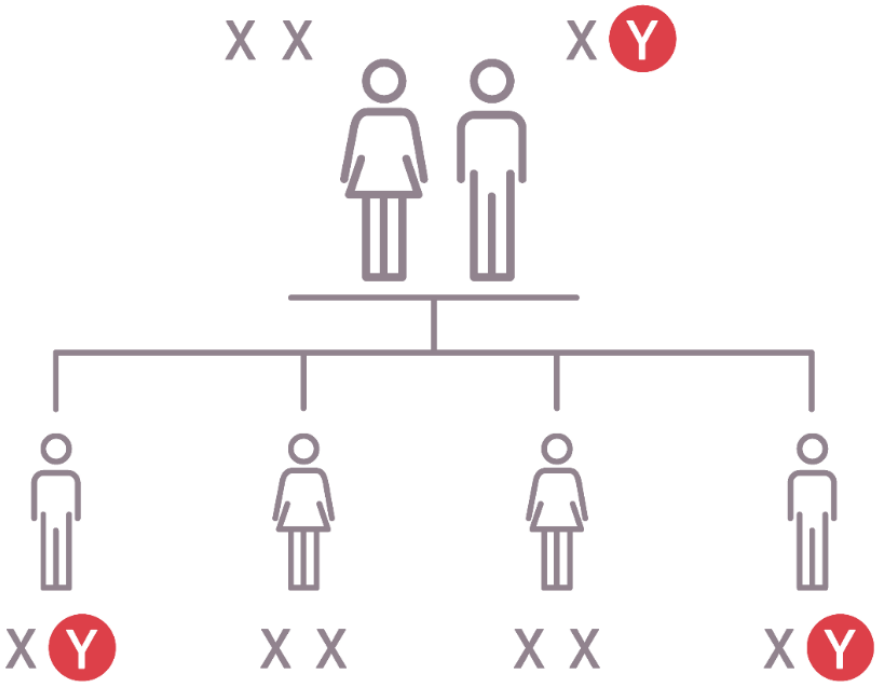

Figure 5: Transfer of markers on chromosome Y between parents and their offspring (Adapted from original image of Halina Šimková with permission of author, https://www.slideserve.com/ronni/slide-pool-forenzn-genetiky-roz-en-soubor-obrazov-ch-sch-mat-k-publikaci)

Mitochondrial DNA is present in multiple copies in every cell and has a physically stable circular form. Therefore, it can serve for forensic purposes in cases of decomposed human tissues and degraded DNA (Amorim et al.

2019). However, mitochondria do not have a DNA repair mechanism, so mutations can be spread clonally, thus generating mixtures of several genotypes in the tissue, a phenomenon known as heteroplasmy.

The different mode of transfer of different markers can be used in familial and genealogical searches over large pedigrees because each marker type brings a special type of information. For example, markers on chromosome X help distinguish hypotheses grandparent-grandchild *vs* uncle-niece *vs* half-siblings (Pinto et al. 2011) and markers on mitochondrial DNA can help distinguish hypotheses of non-maternal *vs* maternal lineage in missing person identification (Angers et al. 2019). Markers on chromosome Y can help not only in familial/genealogical searches (in cases when the male proband is unavailable but his male relatives are available for testing) but also in deconvolution of mixtures (mixture of male and female material after rape or even mixture of several males).

1.3 DNA as differentiation marker

DNA is a dense information medium (Bornholt et al. 2017): in computer terms, DNA has raw information storage limit of 1 exabyte/mm^3 (10^9 GB/mm^3). At the same time, DNA is a stable medium - in the living human body, it can be changed only by mutation. Somatic mutation in an individual cell can have an impact both on the personal life and on forensic DNA detection when it happens in cancer-driving genes and causes tumour growth. Then, the genotype of the tumour differs from the genotype of the rest of the body. A germinal mutation can affect the whole body of the offspring and can cause "mismatch" of the parental allele in a child despite the biological parentage.

Apart from mutation, DNA changes by degradation; it decomposes by the action of physicochemical factors (UV rays, gamma rays, and humidity) and biochemical factors (internal cell nucleases, microbes, and moulds). The degradation rate for DNA is slower than for proteins – the observed DNA half-life is over 500 years. This number must be interpreted with caution: though it is possible in ideal circumstances (e.g., in permafrost) to find analysable DNA in teeth or bones of Neanderthals (Gross 2020; Picin et al. 2020), it occurs quite frequently that DNA cannot be analysed on the crime scene deposit that is several hours old. This may be attributed not only to DNA degradation but also to non-deposition of trace, its secondary or tertiary transfer, and presence of mixture that complicates DNA analysis (Adamowicz et al. 2019; Bauer et al. 2020; Benschop et al. 2019; Coble and Bright 2019; Yang et al. 2019).

There are significant differences among the genomes of human individuals: 0.1% due to single-nucleotide variants (SNPs) and 0.6% when considering insertions and deletions (indels). For SNPs, there are parts of the human genome where every human being bears the same nucleotide (either A, C, T, or G) and parts of the human genome where, at a population level, either of the four nucleotides can be found. Most analyses estimate that SNPs occur on average 1 in 1000 base pairs, in the euchromatin human genome, although they do not occur at a uniform density. In indels, polymorphism is given not by change of nucleotides but by the insertion or deletion of DNA stretches. The range of what can be inserted starts with a single nucleotide, can increase to several nucleotides in tandem fashion in microsatellites or Short Tandem Repeats (STRs) and reaches megabase lengths in Copy Number Variants (CNVs). Special types of DNA insertion are the Long Interspersed Elements (LINEs) and Short Interspersed Elements (SINEs) as they are caused by retrotransposable - jumping - DNA: fragments of DNA that can transcribe themselves into RNA, reverse-transcribe into DNA, and then incorporate into new chromosome locations (Platt et al. 2018). The most common SINE in humans is the Alu repeat that is present in our genome in thousands of copies. Thus, it bears the advantage of multiplicity for identification and kinship testing of degraded DNA, in a similar way as mtDNA (Pineda et al. 2014; Ray et al. 2007).

When talking about tangible gene forms, alleles, polymorphism can reach dozens of variants, where every allele is composed of a stretch of polymorphic and non-polymorphic nucleotides (Figure 6). For forensic purposes, highly polymorphic loci are important because they bear the maximum differentiation power. Before DNA profiling of STR, the champion of polymorphism was the HLA complex – human histocompatibility complex of genes that evolved as part of the immune system to be polymorphic at a population level. This polymorphism matches the variability of microbes, allowing human populations to survive any plague until now. HLA was abandoned in forensic kinship testing in the 1990s but is still being tested in transplantation genetics because the mismatch between HLA haplotype of donor and recipient of transplant can cause graft rejection in case of e.g. organ transplantation and graft versus host reaction in case of bone marrow transplantation.

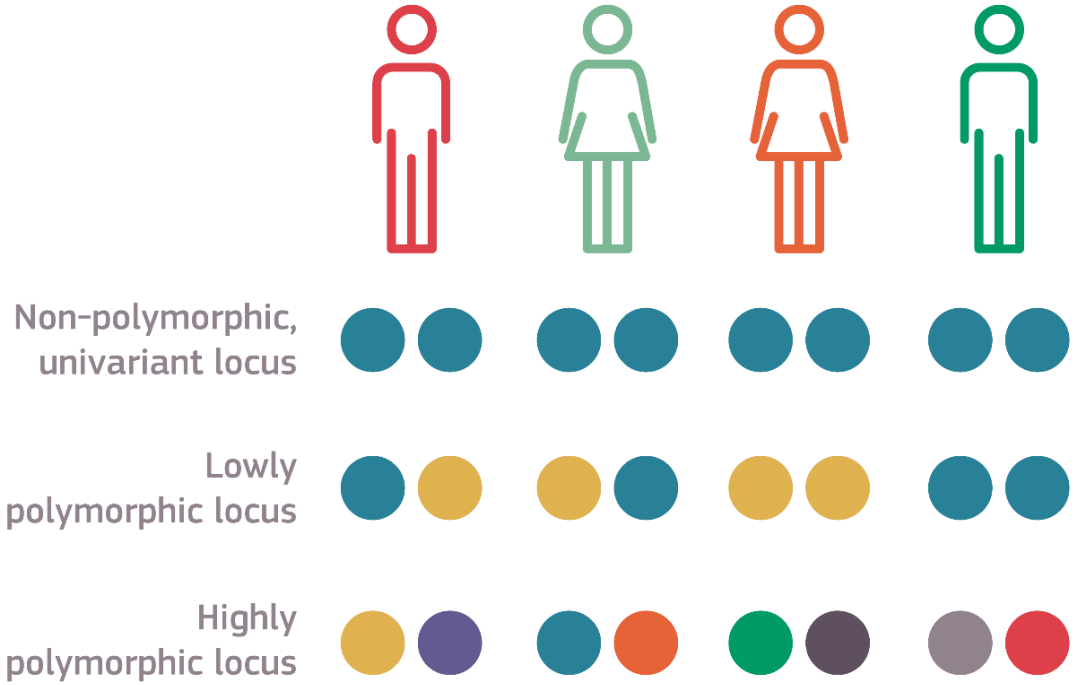

Figure 6: Loci with a different level of variability (Adapted from original image of Halina Šimková with permission of author, https://www.slideserve.com/ronni/slide-pool-forenzn-genetiky-roz-en-soubor-obrazov-ch-sch-mat-k-publikaci)

Although the human genome has been completely sequenced for practical purposes by the Human Genome Project in the turn of the 21st century, there are still hundreds of gaps in the sequence. These gaps are hard to sequence and map due to numerous repeats and other intractable sequence features (Demaerel et al. 2019). With cytogenetics staining, the unsequenced regions are visible as heterochromatin since they possess tightly packed or condensed form of DNA.

1.4 Human transcriptome

In every moment, a portion of the DNA information from chromosomes is transcribed into RNA and a portion of RNA is translated into a protein. Thus, DNA testing provides static information while RNA and protein testing provide dynamic information. The human transcriptome is the set of coding and non-coding RNA transcripts in an individual or a population of cells (Pertea 2012). The use of RNA in forensics was forecast in 2007 (Bauer 2007), but sufficient data were gathered by the Genotype-Tissue Expression (GTEx) Consortium in 2015 (Mele et al. 2015). Based on their data and information from further biomedical studies, forensic applications of RNA testing started to appear, focusing on cell mixture separation, body fluid identification, and time of death estimation (Daca-Roszak and Zietkiewicz 2019; Vennemann and Huth 2014). Massively parallel RNA sequencing is in principle close to massively parallel DNA sequencing, so one would expect the same developments. However, RNA (except for small RNAs, like miRNA) is more fragile than DNA (Courts and Madea 2010). Thus, the use of RNA in forensics will be probably less common. When it is used, the privacy concern will follow the same principles as for DNA.

1.5 Human epigenome

Human epigenetics describes features of the human genome which are important in regulating gene expression, genome replication and other cellular processes but are not encoded directly in the A, C, T, and G genetic alphabet. They may be considered an interface between the genome and environmental factors. Epigenetic markers strengthen and weaken the transcription of specific genes but do not affect the sequence of DNA nucleotides. Instead, they transcend its primary DNA sequence, such as chromatin packaging, histone modifications, and DNA methylation. The latter (methylation of Cs in CpG regions) is a major form of epigenetic control over gene expression and one of the most studied topics in epigenetics.

In early germline cells, the genome has very low methylation levels. As development progresses, parental imprinting tags lead to increased methylation activity. Epigenetic patterns can be identified between tissues within an individual as well as between individuals themselves. They are not completely stable. The epigenome

is influenced by genotype and environmental factors like diet, toxins, lifestyle, and hormones, e.g., methyl-deficient diets can lead to hypomethylation of the epigenome (Koop et al. 2020).

The availability of methylation patterns in human genomes through the Human Epigenome Project (https://www.epigenome.org/) has a potential impact on our ability to understand and diagnose human diseases.

In forensics, it is essential for different analyses: epigenetic identification of body fluids (Lee et al. 2015), estimation of age from human cells (blood, sperm, or other cells) (Parson 2018), analysis of the parent-of-origin specific DNA methylation markers at imprinted loci for parentage testing and personal identification, differentiation between monozygotic twins, artificial DNA detection, and analyses of DNA methylation patterns in the promoter regions of circadian clock gene (Gršković et al. 2013).

1.6 Human microbiome

Human beings can also be identified by their ecological communities of commensal, symbiotic, and pathogenic microorganisms – the microbiota or microbiome - while microbial strain composition is more individualizing than that of a phylogeny (Woerner et al. 2019). Skin microbial communities, although personalized, vary systematically across body sites and time, with intrapersonal differences over time smaller than interpersonal ones (Tozzo et al. 2020). Microbiome sequencing can be also used for biodefense, real-time outbreak surveillance, bio-crime investigations, and metagenomics (cause of death and time since death).

The human transcriptome, epigenome, and microbiome are not further covered by this report despite their potential relevance to the topic of the availability of the whole genomes to forensics.

2 Technology of genotyping

2.1 DNA profiling by STR genotyping

Forensic Short Tandem Repeats (STR) markers were selected solely for identification purposes. Thus, they are present in non-coding regions of the human genome and their effect on phenotype is negligible.

DNA profiling by STR genotyping in the laboratory starts with DNA extractions. It is followed by quantification of the DNA, amplification of up to 27 human-specific Short Tandem Repeat loci (Wang et al. 2020), and separation of amplicons on a capillary column. The final step is the interpretation of the data, leading to reporting the DNA evidence to a court. Each of these steps was improved to increase the sensitivity of testing from minute trace samples (Linacre and Templeton 2014). In addition to autosomal STRs, genetic markers on the X and Y chromosomes are now used for both criminal and civil investigations, along with mitochondrial DNA. DNA profiling has set a standard of performance and quality in the forensic sciences.

Currently, it is used for:

- Identification of suspects,
- Identification of missing persons (Angers et al. 2019),
- Evidence of presence at crime scene(s), which may be the evidence of guilt,
- Exoneration of innocent parties,
- Testing for kinship, including incest and bloodline relations that serve in establishing the historical truth,
- Disaster victim identification (for terrorist attacks or natural catastrophes) (Prinz et al. 2007),
- Identification of person of origin in samples for medical testing (tissue for immunohistological testing, blood alcohol content testing, or other toxicological tests).

Any human biological sample with nucleated cells can be used for DNA profiling:

- Liquid or dry blood, saliva, sweat, and semen deposited on substrate,
- Epithelial cells shed or deposited by touch,
- Hard tissues (bones and teeth),
- Hairs with follicles,
- Pathological slides, formalin fixed paraffin embedded blocks, and cytological smears.

While traditional DNA profiling provides information only about STRs, microarray chips provide information about SNPs. Sequencing provides both SNPs and STRs information, while for STRs more information than just the length of DNA stretch is obtained because it reads the DNA sequence nucleotide by nucleotide.

Table 1: Schematic comparison of laboratory steps in STR profiling, microarray chip and massively parallel sequencing

DNA profiling	Microarray chip	Massively Parallel Sequencing
DNA extraction	DNA extraction	DNA extraction
DNA quantification	DNA quantification	DNA quantification
		Library preparation (and targeted sequence capture)
DNA amplification	DNA amplification	DNA amplification
Electrophoretic detection of amplicons	DNA fragmentation	Signal reading
	Hybridization and signal reading	Data cleaning
Allele and genotype assignment	Allele and genotype assignment	Allele and genotype assignment
Hypotheses definition	Hypotheses definition	Hypotheses definition
Likelihood ratio calculation	Likelihood ratio calculation	Likelihood ratio calculation

2.2 DNA genotyping by microarray chips

DNA is extracted, amplified, and cut into smaller pieces, which are then applied to a DNA chip (also known as microarray), a small glass slide with millions of microscopic beads on its surface. Each bead is attached to a probe, a short sequence of DNA that matches one of the genetic variants that is tested. The cut pieces of DNA stick to the matching DNA probes. A fluorescent label on each probe identifies which version of that genetic variant DNA bears.

Current microarrays can genotype up to a million SNPs (Illumina´s CytoSNP-850K array) and allow incorporation of Linkage disequilibrium into calculation; however, as they require a comparatively large input amount of DNA, they are used primarily by Direct to Customer (DTC) companies that can obtain DNA from saliva spit or buccal swab.

The power of microarrays for resolving kinship cases exceeds what is achievable by forensic STR kits and almost equals what can be achieved by whole genome sequencing. From the view of familial search and genealogical search capabilities, as few as 10,000s of SNPs spread over autosomes, gonosomes, and mtDNA are equivalent to a whole genome sequence (Kling 2019).

2.3 DNA genotyping by sequencing

Sequencing developed from humble beginnings sixty years ago. In the 1960s, 76 nucleotides of alanine tRNA were sequenced from 140 kg of yeast. In 1973, the Maxam and Gilbert DNA sequencing method had the pace of reading one base per month of hard work. Then, still in the 1970s, two duos of researchers, Maxam/Gilbert and Sanger/Coulson ingeniously converted 4-dimensional space of A-C-T-G nucleotides into 2-dimensional space of DNA strand length, detectable by electrophoresis (Shendure et al. 2017). This approach transformed the field. Many following inventions increased the sequencing speed, decreased the cost and decreased researchers´ hands-on time (e.g., switch from radioactivity to fluorescence, from native to genetically engineered polymerases, from flat gel to capillary, or from original to new fluorescent dyes).

However, the stepwise changes were overwhelmed by switching to the technology of massively parallel sequencing (MPS) that was for some time termed Next Generation Sequencing (NGS) before it was generally realized that breakthrough innovations in sequencing will not stop and “next” cannot be replaced by “next next” (Table 2).

Table 2: A brief history of nucleic acid sequencing

Generation	Launch	Method
I	1977	Sanger sequencing
	1986	PCR
I	1987	ABI370
I	1995	ABI Prism 310
I	1997	MegaBASE 1000, ABI Prism 3700
	2000	human genome sequenced
	2001	ABI 3730XL
II	2005	454
II	2006	MiSeq/HiSeq/Solexa
II	2007	SOLiD
II	2010	Ion Torrent
III	2010	PacBio
III	2014	Oxford Nanopore

As for other high-end genetics technologies, early adopters (Rogers 2010) were healthcare diagnostics laboratories while forensic geneticists belong to early majority category. Massively parallel sequencing increases the reach of forensic genetics (Budowle et al. 2017) by:

- Increasing the number and types of genetic markers that can be analysed, including molecular autopsy or pharmacogenomics (e.g., an ultrametabolizing mother with *CYP2D6* gene variant using codein-containing Tylenol can overdose her child with morphine during breastfeeding),
- Possibility to differentiate monozygotic twins (Wang et al. 2015; Yuan et al. 2020),
- Higher throughput of samples,
- Handling mixtures of DNA from several originators because not only the length of STR allele but also its sequence with microvariants is revealed,
- Targeting different organisms, including microbes with over 1,000,000 genes (used for investigation of bioterrorist attacks, bio-crimes, hoaxes, and inadvertent releases of biological agents),
- One unifying technology.

Thus, forensic genetics is on the way to forensic genomics. While genetics refers to the study of a restricted set of genes or genetic loci and the way that certain traits are passed down from one generation to another, genomics refers to a higher quantitative scale study of all of a person's genes (the genome); including interactions of those genes with each other and with the person's environment. Switching from genetics to genomics requires massively parallel sequencing. The immediate adoption of massively parallel sequencing in the forensic genetics field was hampered by issues of speed, databases, DNA input, and initial investment.

- **Speed**

While analysis of STR is done mostly by software in real-time and forensic scientists just check the presence of artefacts like stutter, bleedthrough, microvariant, allele or locus dropout, interpretation of the human sequenced genome requires days of analyses by pipelines of different (and not yet fully standardized) software on air-conditioned servers and a forensic scientist with a bioinformatics knowledge.

- **Databases**

It took some time in Europe and internationally to reach consensus regarding the choice of STR markers that can be shared through CODIS and Interpol. Once established, databases must be curated, and their IT continuously upgraded while the switch to a different set of markers will bring backward incompatibility. Given that most of the perpetrators are repeat offenders (https://www.cbsnews.com/news/once-a-criminal-always-a-criminal/), their absence of new markers in old DNA databases would negate the effect of a database as an investigative tool.

- **DNA input**

The input of DNA for STR typing is now equivalent to 10 cells (60 pg) while DNA input for massively parallel sequencing in its beginnings was three orders of magnitude higher.

- **Initial investment**

Though the price of whole genome sequencing decreases faster than what is described by Moore´s law (now US$ 1,000 for a whole genome with the expectation of US$ 100 per genome over the next 5 years with fixed costs in informed consent and DNA sample acquisition), the initial investment in a sequencer, processing computer, and data archiving computer (or cloud service) assuring the privacy of data is substantial[1].

There are commercial providers for whole genome sequencing[2] but their services do not fit the forensic genetics needs regarding data safety and DNA input.

Due to the reasons above, the forensic genetics community first opted for a slight increase in information content gained in comparison with STR typing with the same DNA input amount and without an increase of burden by lengthy lab work and elaborated bioinformatics interpretation pipelines. Verogen´s MiSeqFGx for SNPs and STRs have been commercially offered since 2015, with different kits having 58-63 STRs and 95-174 SNPs, Verogen´s ForenSeq DNA Signature Prep offers 27 A-STRs + 24 Y-STRs, and ThermoFisher´s Precision ID GlobalFiler NGS STR Panel offers 29 A-STRs + 1 Y-STR.

Sequencing of STRs reveals not only the polymorphism in the number of repeats, but also the tangible sequence within a repeat. For example, ACTGAC**T**G and ACTGAC**C**G are not distinguishable by amplification and capillary electrophoresis while by sequencing they are identified as two different alleles. Thus, sequencing provides more discrimination power even using the same number of markers. Many bioinformatics solutions for virtual STR extraction from the whole genome or its part are available (Liu and Harbison 2018):

- lobSTR (Gymrek et al. 2012)
- RepeatSeq (Highnam et al. 2013)
- STRviper (Cao et al. 2014)
- CoalescentSTR (Kojima et al. 2016)
- STRait Razor (Woerner et al. 2017)
- MyFlq (Van et al. 2014)
- STRinNGS (Friis et al. 2016)
- SEQ Mapper (Lee et al. 2017)
- FDStools (Hoogenboom et al. 2017)
- STRNaming (Hoogenboom et al. 2019)
- TSSV (Anvar et al. 2014)

Also, databases of allelic frequencies are available to allow calculation of likelihood ratio:

- genomAD http://gnomad.broadinstitute.org

[1] https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
[2] https://www.scienceexchange.com/marketplace/whole-genome-seq?page=1

- ExAC http://exac.broadinstitute.org
- 1000 genomes project (Abecasis et al. 2012; Siva 2008)
- Human Genome Diversity Project (Cavalli-Sforza 2005)
- OpenSNP project (Greshake et al. 2014)
- EMPOP for mtDNA (Prieto et al. 2011)
- ALFRED (Rajeevan et al. 2012), FROG-kb for SNP allelic frequencies (Kidd et al. 2018)
- STRidER for STR allelic frequencies (Bodner et al. 2016)
- STRseq for sequence variants (Gettings et al. 2017)
- STRs PopAffiliator for assigning to population (Pereira et al. 2011).

To standardize the sequencing output, the International Society for Forensic Genetics set minimal nomenclature requirements for MPS of STRs (Parson et al. 2016).

The use of MPS in forensics grows as reflected in an increasing number of articles published (Bruijns et al. 2018) and in the generally high acceptance of the technology across forensic laboratories (Alonso et al. 2017).

There are currently several sequencing platforms readily available (

Table 3) (Goldfeder et al. 2017).

Table 3: Parameters of sequencing platforms

Platform	Technique	Read length (nt)	Error rate (%)	Note
Sanger	PCR/chain termination	25-1,200	0.1	96 capillaries
454 (Roche)	Emulsion PCR/Sequencing by synthesis/pyrosequencing	100-1,000	0.1-1	unsupported technique
MiSeq/HiSeq/Solexa (Illumina/Verogen)	Bridge PCR/sequencing by synthesis/reverse termination	36-300	0.1-1	for SNP and indels (within WGS and WES); difficulty aligning short sequence reads
SOLiD (LifeTechnologies)	Sequencing by Oligonucleotide Ligation and Detection; Emulsion PCR/ligation/probing	35-75	0.1	derived from polony sequencing; run lasts one week
Ion Torrent PGM (ThermoFisher)	Emulsion PCR/ion-sensitive sequencing by synthesis/pH change	200-400	1-2	low start-up costs
Sequel II (PacBio)	Single Molecule, Real-Time (SMRT)/ zero-mode waveguide (ZMW) wells (without amplification)	8,000-20,000	10-15	large investment
MinION (Oxford Nanopore)	Ion current shift (without amplification)	9,545-200,000	5-40	portable

The current method of choice for MPS in forensic laboratories is Illumina´s MiSeq and Ion Torrent which may be changed with improving the analytical parameters and ratio price to yield.

3 Process of forensic genetics identification

During forensic genetics identification, the DNA profile from the crime scene or human remains is compared with a reference profile of suspect or pool of profiles in a criminalistic database (https://senseaboutscience.org/activities/-making-sense-of-forensic-genetics/). When alleles between two samples are the same in the currently tested number of STR loci (more than 13) - i.e., they match - then there is a piece of strong evidence that they derived from the same source. The person whose profile matches a crime scene profile may be involved in some way with the crime (e.g., by committing or aiding in the crime, or being present during the crime), but this would need to be established with further evidence. A profile match alone is not sufficient proof because their DNA profile may have occurred at the crime scene because of their innocent presence before the crime, secondary transfer, or laboratory contamination.

Many forensic assays have been developed using STRs, SNPs, indels, and microhaplotypes (de la Puente et al. 2016; Hao et al. 2019; Hou et al. 2014; Cho et al. 2014; Pang et al. 2020).

3.1 Comparison of DNA

Comparison of two sets of information about DNA sequence does not require two identical genotyping methods to be performed on both samples (Figure 7). Even when the same technique is used on both samples to be compared, the overlap of loci does not need to be absolute. It is sufficient to have 13 STRs or 50 SNPs available for comparison on both sides. When the number of markers is lower (e.g., only 6 STRs), a long list of database matches is produced that can be arrayed from the highest to lowest likelihood ratio. With the advancement of DNA extraction and MPS techniques, even degraded samples are sequenceable (Tillmar et al. 2019; Tillmar et al. 2020).

Generally, we can compare:

1. STR loci obtained by DNA profiling on sample 1 and the same STR loci obtained by DNA profiling on sample 2,
2. STR loci obtained by massively parallel sequencing on sample 1 and the same STR loci obtained by massively parallel sequencing on sample 2,
3. STR loci obtained by DNA profiling on sample 1 (in computer security terms, a beacon) and the same STR loci extracted from whole genome on sample 2,
4. SNP loci obtained by massively parallel sequencing (or microarray chip) on sample 1 and the same SNP loci extracted from whole genome (or set of SNPs on microarray chip) on sample 2,
5. STR loci obtained by DNA profiling (or massively parallel sequencing) on sample 1 and the SNP loci obtained by massively parallel sequencing (or microarray chip) on sample 2.

Though Copy Number Variants (repeats on the larger chromosomal scale) or whole genomes can be compared as well, their comparison does not have a practical significance.

A crime scene sample is rarely in an amount and quality sufficient to allow whole genome sequencing. In fact, by performing a whole genome sequencing of traces, plenty of information would be generated at the expense of quality of sequencing. This information would be later discarded during processing because it covers regions without variability or regions associated with heritable diseases. Thus, forensic MPS usually uses a sequence capture technique to filter genetic regions of interest before the real sequencing starts.

Regarding forensic use of whole genome databases, this will always involve a comparison of a subset from the investigator´s side to the whole set of nucleotide sequences from the database side. As a result, identification or hit will be provided (when achieved) the same way it is currently practised for forensic DNA databases. Information about other phenotypic characteristics will not be sought and will not be provided, as it would bring no new information for the purpose of the individual identification.

The 5th mentioned comparison method for disjoint loci (STRs vs SNPs) is the least intuitive. It relies on linkage disequilibrium, the non-random association of alleles at different loci in a given population. Thus, every marker in one person is not matched to the same marker of the other person but to the marker that is associated, linked with the same marker on the same chromosome.

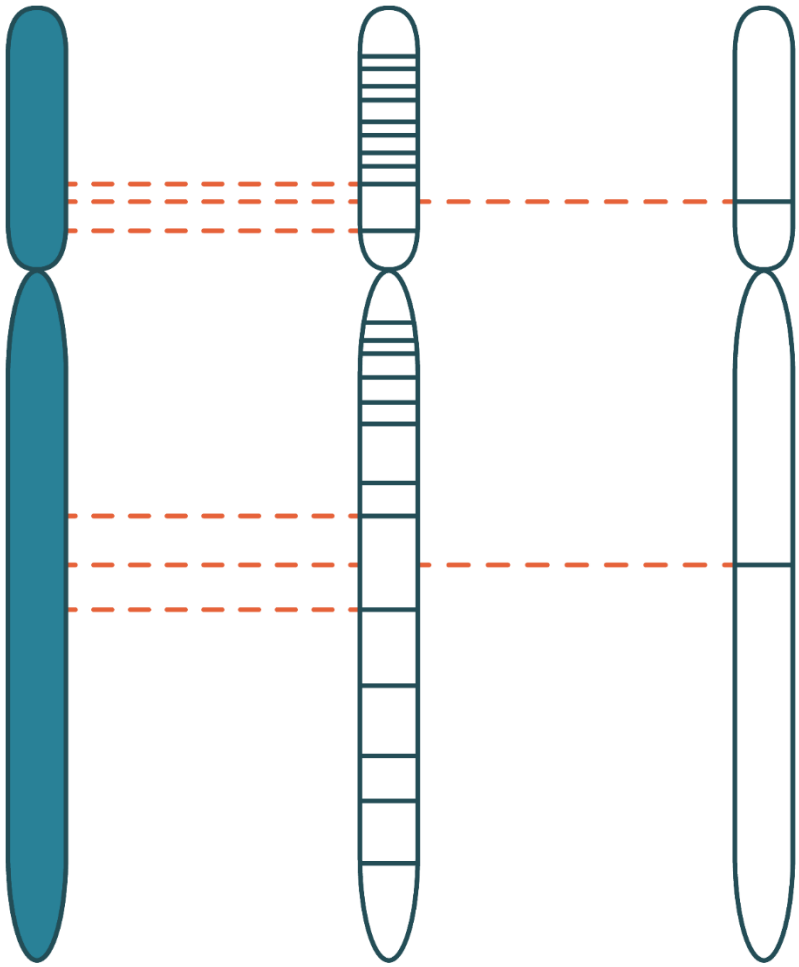

Figure 7: Schematic illustration of the ways to compare generic information between samples

Identification is called de-anonymization (or re-identification) when information is obtained about the identity of a person even though anagraphic details were removed on purpose (Hara et al. 2003). Vulnerability to de-anonymization is an inherent feature of high-dimensional data: not only genetic data (Bradley 2016; Humbert et al. 2015), but also social networks, location data, credit card data, browsing histories, writing style, source code (Caliskan et al. 2018), and compiled binaries. A small number of data points about an individual, none of which are uniquely identifying, are collectively equivalent to an identifier. This agrees with a theory that 33 bits of entropy are sufficient to identify an individual uniquely among the world's population (Narayanan and Shmatikov 2008).

Due to inheritance, a person can also be identified by knowledge about their relatives. Due to availability of other Big Data, a person can be identified by many other means, see the chapter Genomic Data is Big Data. Re-identification will likely become easier with time as the amount of publicly available data increases and technology improves (Cech 2019).

3.2 Quality of DNA data and profiling process

Such is the power of DNA to help identify, convict, and exonerate, that many non-specialists (including lawyers) perceive it to be infallible. A myth of the infallibility of DNA profiling may lead to overlooking possible errors and result in marginalization or even elimination of other types of evidence in court. Yet DNA evidence has a number of limitations and the cost of not being aware of them can result in the miscarriage of justice (Gill 2019). These limitations are inherent also to forensic genetics techniques. The availability of whole genome sequences of the large part of the population does not bring additional danger of judicial error in comparison to DNA profiling by STR genotyping.

Thus, the same highest quality standards must be striven in the whole process of forensic genetics (pre-laboratory, pre-sequencing, sequencing, and post- laboratory protocols) using whole genome sequencing. This can be achieved by top-down standardization by regulatory authorities and bottom-up standardization by the laboratories themselves.

3.3 Standards, certification of practitioners, and accreditation of laboratories (Wilson-Wilde 2018)

Standardisation allows high throughput, controlled quality, reliability of results, and compatibility across laboratories. The most important standards related to DNA-profiling are: ISO/IEC 19794-14:2013- Information Technology–Biometric Data Interchange Formats -- Part 14: DNA Data; ISO 18385:2016 -Minimizing the Risk of Human DNA Contamination in Products Used to Collect, Store and Analyse Biological Material for Forensic Purposes–Requirements; ISO/IEC 17025:2017 - General Requirements for the Competence of Testing and Calibration Laboratories; ISO 21043:2018 - Forensic Sciences.

The ISO/IEC 17025:2017 norm covers all aspects of the forensic testing laboratory and its accreditation requires documented compliance across all laboratory workflow steps (Validation of analytical methods and procedures, Equipment calibration testing and maintenance, Qualification of material, Traceability, Control of nonconforming testing, Qualification of personnel, Controlled environmental conditions, and Standard Operating Procedures). Regarding the sequencing method itself, a laboratory achieving accreditation must have successfully passed proficiency tests and collaborative exercises.

Stimulation (yet not a requirement) of laboratory accreditation is part of the European Forensic Science Area (EFSA) 2020 action, covering also the creation of best practice manuals for forensic disciplines, stimulating the exchange of forensic information from databases in the areas of weapons and ammunition, explosives and drugs, forensic awareness and training for law enforcement and justice communities, promoting and improving the exchange of forensic data via the 2005 Prüm Treaty.

3.4 Codes of conduct, the best laboratory practice

The International Society for Forensic Genetics (ISFG), European Network of Forensic Science Institutes (ENFSI) Expert DNA Working Group, INTERPOL (group 2015), and scientific bodies (i.e., Scientific Working Group on DNA Analysis Methods, SWGDAM in USA), provide recommendations regarding best laboratory practices (Adegoke et al. 2017; Bodner, Bastisch, Butler, Fimmers, Gill, Gusmao, Morling, Phillips, Prinz, Schneider and Parson 2016; Coble et al. 2016; Gjertson et al. 2007; Parson et al. 2014; Roewer et al. 2020; Samuel et al. 2018; Tillmar et al. 2017).

The results of forensic DNA analysis must be properly understood by the judicial system (to avoid CSI effect or worse, miscarriage of justice). Therefore, training of forensic practitioners, law enforcement, and justice in logical evidence interpretation is a must.

3.5 Logical, Bayesian way of evidence interpretation

To avoid confusion, misunderstanding, and misjudgement, DNA cannot be used as sole evidence and the database trawl match cannot be considered to imply identity (Evett 1995). Instead, logical – Bayesian – approach should be applied. Bayesian reasoning is defined by three rules for the reasonable and efficient handling of information:

1) Evidence interpretation is not performed in a vacuum, but always within the circumstances of the case. This is because DNA evidence without context may be presented in a strongly prosecution biased way. The expert must express the understanding of all relevant aspects of the case on which the interpretation is based, including a placement on hierarchy of propositions, a possibility of contamination, background DNA, secondary transfer, and mixture.
2) The expert considers the observations in the light of two hypotheses (scenarios, versions, or propositions): the prosecution hypothesis and the defence hypothesis. These hypotheses must be clearly stated.
3) The expert calculates the probability of evidence assuming that the prosecution hypothesis holds, and the probability of the same evidence assuming that the defence hypothesis holds, and puts these two probabilities into proportion. In other words, he/she calculates the likelihood ratio, LR, as the whole expert testimony. When hypotheses are changed by the claimant, the lawyer or the judge during the proceedings, the likelihood ratio must be recalculated.

The ability of an experienced forensic scientist to evaluate the results given the circumstances and propositions in a particular case and to present this to the court in a clear and concise way is very important for the legal process (Nordgaard and Rasmusson 2012; Robertson et al. 2016). Court officials can neither be expected to be able to interpret scientific data nor is it their task to do so. The duty of the court is rather to perform the ultimate evidence evaluation of all the information in the case combined, including judge experience and knowledge, police reports, statements from suspects and victims, witness reports, and forensic expert statements.

In due process, it is important to avoid the trap of transposing the conditional. Two sentences exemplify the issue of transposing the conditional: (1) the probability that an animal has four legs if it is a dog is approaching 100% does not mean the same thing as: (2) the probability that an animal is a dog if it has four legs is approaching 100%. Questions from attorneys and judges to experts are often wrongly framed as transposed conditionals (the probability of a suspect being at crime scene given the DNA match instead of the correct: the probability of the DNA match given the suspect being at crime scene).

4 Genomic Data is Big Data

Big Data are characterised by three Vs: volume, velocity, and variety. They have a high level of completeness (e.g., covering whole populations) that contains contextual information that can identify concrete and specific situations and individualise persons. Data are relational (possible to be compared to other resources) and flexible (able to incorporate new data). They rouse a mythological belief that large datasets offer a higher form of intelligence and knowledge that can generate insights previously impossible, with the aura of truth, objectivity, and accuracy, in striking resemblance to the CSI effect for DNA profiling.

Big Data can determine the risk that a specific individual will commit a crime or terrorist act; this may reinforce the surveillance of social groups and individuals who are vulnerable to police suspicion, thereby consolidating the social stigmatisation and reproduction of social inequalities. At the same time, it can produce evidence for the justice system and contribute to crime prevention and deterrence.

The phenomenon of Big Data emerges in the context of technological development and the growing importance of the digital world, which is associated with large-scale collection of citizens' data. It is a technique that aggregates and analyses a massive amount of data, converting it into algorithms, numerically categorised and identified by employing a calculated index, from which information can be extracted. It is being applied to several spheres of life, including commerce, consumption, health, social security, marketing, and immigration.

4.1 Big Data bears two risks: data silos and data misuse

An information silo is an insular management system in which one information subsystem is incapable of reciprocal operation with others that should be related. Thus, information is not adequately shared but rather remains sequestered within each subsystem, trapped within a container, like grain is trapped within a silo. Such data silos are an obstacle for data mining to make productive use of the data. Some biomedical data are not released because of substantiated or unsubstantiated fear from regulatory authorities while a problem of overly strict regulation may slow down and even stall innovation (Molnar-Gabor and Korbel 2020). To realise the potential of data, databases should be designed with FAIR principles in mind: data should be Findable, Accessible, Interoperable, and Reusable (Holub et al. 2018).

Data misuse can theoretically happen if the government or other agency gain absolute control over the private life of a person. Indeed, it is argued by some scholars that Big Data era is a post-privacy age (Cech 2019).

Patterns of data misuse may take the form of improper data profiling (e.g., it is possible to compile lists with the names and contact information of rape victims, the addresses and locations of domestic violence shelters, people with specific illnesses and more; data brokers can then obtain these lists and sell them to interested businesses), discrimination (e.g., in work offers), errors (e.g., in misinformation visualisation), political and social manipulation (e.g., by targeted spread of false news), and data breaches and cyber-attacks.

In recent years, there has been a dramatic increase in the amount of genetic information generated, analysed, shared, and stored by diverse individuals and entities (Clayton et al. 2019). DNA databases have speed, efficiency, automation, and accuracy that are unmatched in the history of policing. Now DNA data is Big Data that is part of a larger data ecosystem with greater possibilities for integration (Machado and Granja 2020). Though there were few reported cases of malicious DNA re-identifying to date, there is a danger that it will become profitable to decode DNA or re-identify genomic data for marketing, commercial, or other purposes. The easiest way to re-identify someone is when the metadata contained within the genomic data files (e.g., FASTQ, BAM, VCF) contain non-genomic personal information (zip code, birth date, and sex).

Surnames can be recovered from personal genomes by profiling STRs on the Y-chromosome and by querying recreational genetic genealogy databases (Gymrek et al. 2013). In 2013, there were 39,000 unique surnames in 135,000 records available to genealogical researchers. In trigonometry and geometry, triangulation is the process of determining the location of a point by forming triangles out of known points. Genetic triangulation is a term coined by genetic genealogist Bill Hurst to describe a method of determining the Y-STR or mtDNA ancestral haplotype using two or more data points. Combination of surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. This technique entirely relies on free, publicly accessible Internet resources.

A 12% success rate was achieved in recovering surnames of US Caucasian males (5% wrong surname and 83% referred as unknown); these results are relevant to socio-economic groups with high participation in these services – upper and middle-class US Caucasians.

The primary source of surname inference used to be Ysearch[3] and Sorenson Molecular Genealogy Foundation (SMGF). Ysearch, the free, public genetic-genealogy database, is no longer accessible as a result of the EU General Data Protection Regulation (GDPR) that went into effect on May 25th, 2018.

The assets of the SMGF were acquired by AncestryDNA in March 2012. AncestryDNA initially promised that the Sorenson database would continue to be available to researchers for the foreseeable future, but the website was taken down by AncestryDNA on 14 May 2015 after it had been used for "purposes other than that for which it was intended" and there are no plans to reinstate it.

In July 2020, FamilyTreeDNA, another commercial genetic testing company based in Houston, Texas, had more than 700,000 unique surnames.

4.2 Genealogical, familial, and biogeographical searches using consumer genetics databases

Consumer genetics (online genealogy) databases hold dense genotypes of millions of people, and the number is growing quickly (Kennett 2019). Collectively, they cover a larger part of human population than the police database CODIS[4] (Table 4).

Table 4: DTC and forensic databases amenable to genealogical testing

Database	Database size	Accessibility to forensic scientists
23andMe	10 million	The only way to access to data is by submitting a sample *via* a cheek swab or spit kit.
AncestryDNA	15 million	The only way to access to data is by submitting a sample *via* a cheek swab or spit kit.
CODIS	16 million	Per se.
FamilyTreeDNA	2 million	Law enforcement usage requires written permission from the company, as well as the required legal documentation.
GEDmatch	1 million	Created by Curtis Rogers and John Olson in 2010 as a public database where individuals from different testing companies could compare their DNA by downloading their raw data from a DTC company´s site and uploading it to a common database. After the Golden State Killer suspect was identified through the secret use of GEDmatch, the site´s administrators decided to explicitly allow law enforcement usage for homicides and sexual assaults. Even prior to implementing the new terms of use, GEDmatch explicitly stated that any data set to "public" would be searchable by anyone.
MyHeritage	2.5 million	Law enforcement usage requires written permission from the company, as well as a court order.
Parabon Nanolabs	unknown	Genetic data is kept on encrypted server accessible only to authorized employees, and the company´s GEDmatch accounts can only be accessed by the bioinformatics team and the lead genealogist, Dr. CeCe Moore.
Prüm Decision related national resources	>6.5 million	Note: Prüm Decision requires the establishment of national databases and automated access to data procedures to share data among them. Each country established a "Prüm national database". If the UK, Norway, Switzerland, and Liechtenstein are considered, the number must be doubled.

[3] www.ysearch.org
[4] https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart

UK NDNAD	5.9 million	Per se.
Others, e.g., CRI Genetics, Everlywell, FindMyPast, GenoPalate, LetsGetChecked, Living DNA	unknown	Unknown; note: apart from genealogy, they are offering DNA testing for food allergies, emotional health, and genetic disease risks.

DTC testing companies regulate themselves by adhering to 8 principles: transparency, consent, use and onward transfer, access/integrity/retention/deletion, accountability, security, privacy by design, and consumer education ((FPF) 2018).

In 2018, law enforcement agencies began using online genealogical databases to identify anonymous DNA via long-range familial searches. Genealogical analysis helped in the 11-M Madrid commuter train bomb attack investigation (Phillips et al. 2009) but the most media coverage was given to a successful identification of twelve-time murderer and fifty-time rapist, the Golden State Killer, a former police officer named Joseph James DeAngelo, Jr., ten years later (Cech 2019; Wickenheiser 2019b).

In May 2019, GEDmatch tightened its rules on privacy which were forecast to make it much more difficult for law enforcement agencies to find suspects using GEDmatch. Nevertheless, as of July 2020, GEDmatch has been used in at least 119 cold case arrests, most of which were the work of Parabon Nanolabs[5]. Until 2019, Parabon analysed more than 250 forensic samples (Greytak et al. 2018; Greytak et al. 2019).

The outcome is not guaranteed once a sample is uploaded to a database. However, the database query results are just clues from which in-depth genealogy and descendancy research must proceed to determine the possible identities of an unknown individual.

The complications can be: a lost identity through adoption, abandonment, anonymous gamete donation, and misattributed parentage. Also, populations founded by a small number of individuals can have low genetic diversity and high background relatedness (endogamy). In such subpopulations, individuals with a given relationship will share more DNA than in other populations. Endogamy manifests as a large number of matches, each sharing many small segments.

There are several algorithms and software available for inferring whole genome histories in large population datasets (Dodds et al. 2019; Dou et al. 2017; Hanghoj et al. 2019; Huff et al. 2011; Kelleher et al. 2019; Korneliussen and Moltke 2015; Manichaikul et al. 2010; Speidel et al. 2019). As is the case for identification by one to one comparison, linkage disequilibrium can be used for disjoint forensic and biomedical loci (Kim et al. 2018; Edge et al. 2017). Apart from the likelihood ratio approach, genetic genealogy relies on Identical By State (IBS) and Identical By Descend (IBD) measures.

The simulation results indicate that the traditional LR approach as a single source of classification is at least as good as the alternative approaches. However, it is both computer-intensive and sensitive to population frequencies as well as positions of the markers. The lowest false classification rate with a high true classification rate can be achieved when combining different classification approaches (Kling and Tillmar 2019).

IBD/IBS approach does not simply count the number of shared SNPs but relies on the fact that the recombination breaks up long stretches of DNA shared remain intact over the generations. Thus, more closely related people will share longer stretches of DNA that are IBD. The more recombination events that have occurred, the shorter the shared IBD segments will be. The number and length of IBD segments in cM can be used to approximate the degree of relationship.

Algorithms search for regions of the genome where two individuals share at least one allele at every SNP, segments must contain a minimal number of SNPs (e.g., 500) and be over a certain length (e.g., 6 cM). Thus, segments that are IBS are distinguished from ones identical by descent (IBD). When summed over all autosomes, the amount of shared IBD segments correlates with the degree of relatedness between two individuals.

While the average shared centiMorgans are shown in the Figure 8 below, in reality a range of values can be observed. For relationship of the first degree, it is 2,000-3,600 cM, for 2nd degree: 1,060-2,500 cM, 3rd: 425-1,500 cM, 4th: 160-950 cM, 5th: 65-650 cM, 6th: 0-375 cM, 7th: 0-245 cM, 8th: 0-185 cM (Greytak, Moore and Armentrout 2019). Each degree contains many relationship types that must be considered: a 5th-degree relative

[5] https://en.wikipedia.org/wiki/List_of_suspected_perpetrators_of_crimes_identified_with_GEDmatch

can be second cousin, first cousin twice removed, or half-first cousin once removed. Matches sharing 100 cM could be anywhere from 5th degree to 9th degree, with 6th degree being most likely. Ten per cent of 3rd cousins and 50% of 4th cousins share no detectable IBD segments.

Relationship degree	1	2	3	4	5	6	7	8	9
Percentage of shared DNA	50	25	12.5	6.25	3.125	1.563	0.781	0.391	0.195
Average shared centiMorgans	3400	1700	850	425	212.5	106.25	53.13	26.56	13.28

Relationships defined with respect to proband (white VI.2 box).
I.1 great-great-great-grand-father, **I.2** great-great-great-grand-mother, **II.1** great-great-grand-father, **II.2** great-great-grand-mother, **II.3** great-great-grand-uncle/aunt, **III.1** great-grand-father, **III.2** great-grand-mother, **III.3** great-grand-uncle/aunt, **III.4** first cousin thrice removed, **IV.1** grandfather, **IV.2** grandmother, **IV.3** grand-uncle/aunt, **IV.4** first cousin twice removed, **IV.5** second cousin twice removed, **V.1** father, **V.2** mother, **V.3** uncle/aunt, **V.4** first cousin once removed, **V.5** second cousin once removed, **V.6** third cousin once removed, **VI.1** half-brother/half-sister, **VI.2** proband or their monozygotic twin, **VI.3** brother/sister, **VI.4** first cousin, **VI.5** second cousin, **VI.6** third cousin, **VI.7** fourth cousin, **VII.1** half-niece/half-nephew, **VII.2** son/daughter, **VII.3** niece/nephew, **VII.4** first cousin once removed, **VII.5** second cousin once removed, **VII.6** third cousin once removed, **VIII.1** half-grand-niece/half-grand-nephew, **VIII.2** grand-son/grand-daughter, **VIII.3** grand-niece/grand-nephew, **VIII.4** first cousin twice removed, **VIII.5** second cousin twice removed, **VIII.6** third cousin twice removed

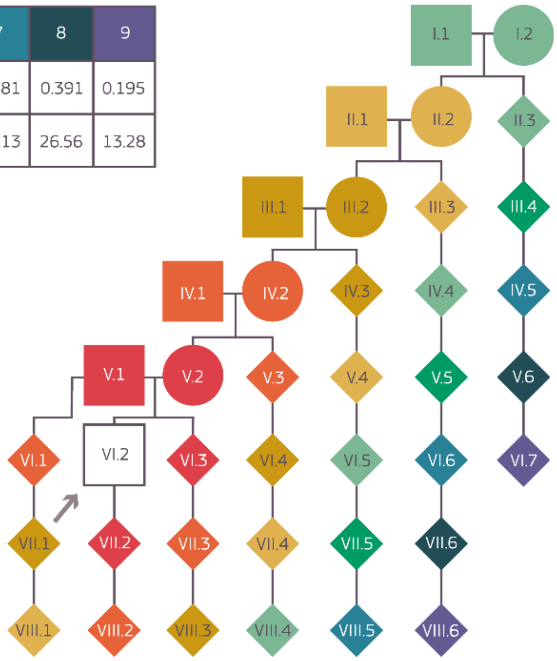

Figure 8: Relationships defined with respect to the proband

When uploading raw genotyping data, 23andMe will generate 10,000s of 2nd to 9th-degree cousin pairs within a heterogenous set of 5,000 Europeans (Henn et al. 2012). For the whole 23andMe database, the number of provided cousin matches is even higher.

At the same time, for over 50% of targets, their anonymous DNA can be identified (matched to the correct individual or same-sex sibling) when the genetic database includes just 1% of the population (Ellenbogen and Narayanan 2019). Each complete pedigree re-identification, using free online non-genetic or auxiliary resources were shown to take 3 to 7 hours by a single person (Gymrek, McGuire, Golan, Halperin and Erlich 2013; Lippert et al. 2017; Long et al. 2019). This time may be shortened with future software and Internet developments. The shared non-autosomal DNA, DNA on X-chromosome, Y-chromosome, and mtDNA can narrow the possible paths between matches and two-family pedigrees can intersect, triangulate. In a simulation, MyHeritage can individualize a person from single third cousin level relative match, given knowledge of the sex, location within 100 miles, and age within 5 years (Erlich et al. 2018).

Auxiliary information sources can be (Bonomi et al. 2020):

- Demographics, surnames (https://www.census.gov/data.html, www.PeopleFinders.com, https://www.ussearch.com/),
- Pedigree, family tree (https://pgp.med.harvard.edu),
- The Human Polymorphism Study Center (CEPH, http://www.cephb.fr),
- Gene expression (https://gtexportal.org/home/),
- Genotype data (https://www.opensnp.org),
- 1000 Genomes Project (https://www.internationalgenome.org),
- The database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/),
- Social relationships - Population registries and social networks (www.facebook.com, www.tiktok.com),
- Observable phenotypes - Social networks,

— Clinical data - Clinical data research networks, and

— Summary statistics (https://www.ukbiobank.ac.uk).

It should be noted that such databases grow not only in Europe and North America but also in China, India, and South Korea, so the development is global. By now, privacy concerns linked to the use of genetic data from DTC databases in police investigations is reduced by the fact that raw genetic data (that contains highly personal and health-related information) are not disclosed to law enforcement. Search results display only the length and chromosomal location of shared DNA blocks, which are used to determine approximate kinship relationships between individuals. Customer relations create an incentive for testing companies and GEDmatch to maintain current policies of not releasing raw data without consent. Finally, genetic genealogy is intended for investigative lead generation, not conviction. Genetic genealogy leads must be tested by direct DNA matching to samples from persons of interest using standard forensic identification loci; only matches obtained with these well-established methods will result in a continued investigation.

For illustration, Table 5 shows the tangible GEDmatch results of one of the authors of this report for the first 10 matches out of the 3,000 reported.

Table 5: Personal GEDmatch results

Largest Segment	Total cM	Overlap
32.8	47.2	178,541
18.9	33.4	299,605
19.7	27.1	139,411
18.1	26.2	70,643
14.4	23.2	71,338
14.3	22.5	70,367
13.9	22.4	68,765
13.1	22.1	70,671
13.9	22.1	306,992
13.7	21.7	304,133

Overlap is the number of positions that are in common between both kits, without regard to whether they match or not. The best match with 47.2 cM overlap can be with a 6th-degree relative. Thus, in this case, the GEDmatch report would not have been helpful for investigators.

On December 9, 2019, the GEDmatch was purchased by Verogen, Inc., a forensic genomics company whose focus is on human identification. There have been concerns that law enforcement will have greater access to GEDmatch user information. However, Verogen stated that it would fight all unauthorized law enforcement use. There has been a temporary drop in the database size because privacy policies in place in the various countries where GEDmatch users reside require citizens to specifically approve the transfer of their data to Verogen. As users grant permission, that data will again be visible on the site. Verogen has pledged to continue the GEDmatch philosophy of providing free services.

When a genealogical study extrapolates into the history of mankind to our roots in Africa, it can define the continental or even subcontinental population to which a person belongs, their biogeographical ancestry, BGA (Gannett 2014). While the concept of human race is an abandoned anthropological concept due to its World War II connotations and vague definition (see: “one definite and obvious consequence of the complexity of human demographic history is that races in any meaningful sense of term do not exist in the human species” (Goldstein and Chikhi 2002), BGA is a practical measure of the biological component of ethnicity, without cultural and religious confounders (Halder et al. 2008; Shriver et al. 2003).

It is useful not only for hobbyist ancestry-seekers but also for forensic geneticists. It can help them to focus on a population of an unknown suspect (unidentified individual) based on DNA in instances where a DNA profile is not found in DNA databases. By better targeting investigations, fewer innocent people are implicated. On top of the analysis that is available from DTC genomics companies as a part of the services purchased, there are currently several programs allowing to infer BGA based on Ancestry Informative Markers, AIM (for example Snipper (Phillips et al. 2007), STRUCTURE (Kaeuffer et al. 2007), GenoGeographer (Tvedebrink et al. 2017), and GRAF-pop (Jin et al. 2019)).

Non-commercial assays offered 34 SNPs, 46 Indels, and 127 SNPs with the probability that, in cross-validation, an unknown DNA sample of unambiguous continental origin will be assigned to the correct continent of origin reaching 99.9% (Schneider et al. 2019).

The commercial ForenSeq™ DNA Signature Prep Kit (Verogen, USA) distinguishes 4 populations (Europe, Africa, Americas, and East Asia) using 56 SNPs while Precision ID Ancestry Panel (ThermoFisher Scientific) distinguishes 7 populations (Europe, Africa, Americas, East Asia, Oceania, South Asia, and South-West Asia) using 165 SNPs. Precision parameters are not provided.

The weight of evidence for tangible BGA can be several orders of magnitude lower than the one for human identification. In contrast to likelihood ratios (LRs) provided by a full profile of 13 STRs in identity testing reaching at least 10^{14}, LRs in biogeographical testing can be as low as 1.5 (e.g., the ancestry of a tangible sample from a man, originating from Afghanistan, with two competing hypotheses of continental origin, Europe *vs* South Asia, with original probabilities 50%:50% is, using BGA testing, changed to 39.2%:60.8%).

As an example, Figure 9 shows the MyHeritage results for the same individual analysed in Table 5.

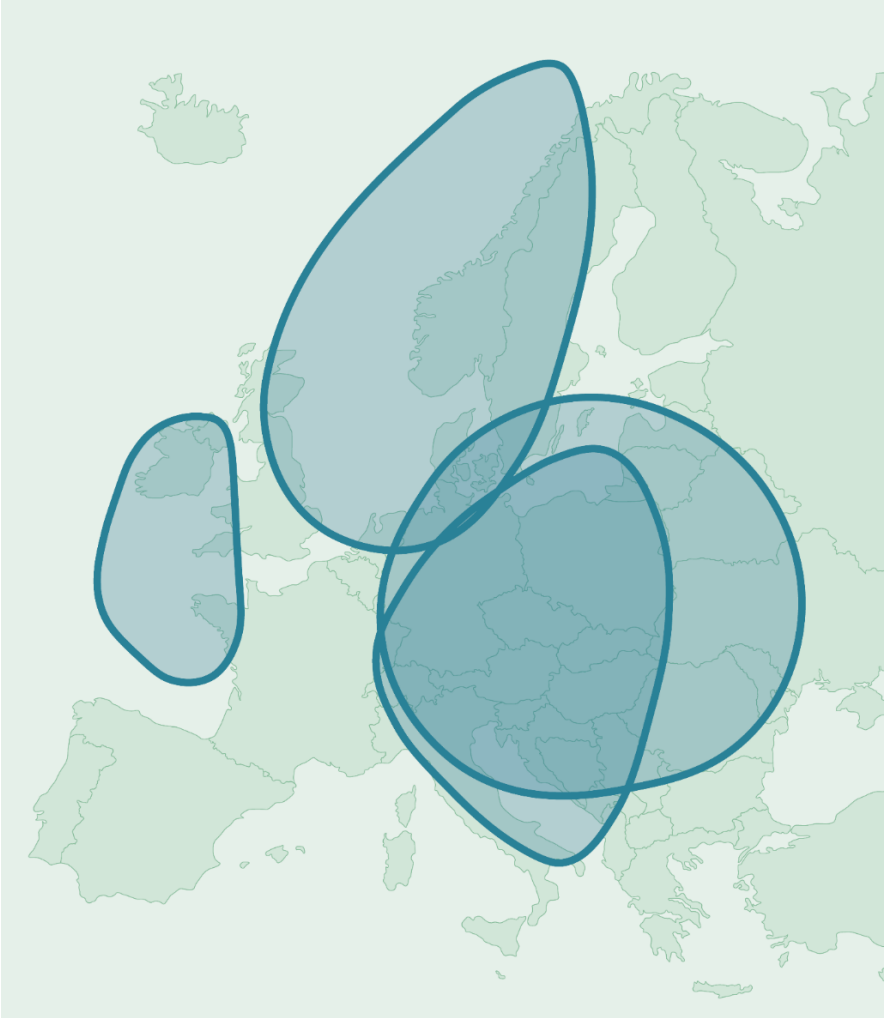

Figure 9: Example of results obtained from MyHeritage for an individual

With assembling of the 100% European ancestry by Eastern Europe 83.3% (Non-Balkan 65. 6%, Balkan 17.7%), Northern and Western Europe 16.7% (Scandinavia: 10.8%; Ireland + Scotland + Wales: 5.9%), the analysis provided information that is useful not only at inter-continental level but also provides some clues about the intra-continental position. However, the exact country of origin cannot be identified.

4.3 DNA as phenotypic or biometric data

Most of the European countries currently have functional national forensic DNA databases which store the DNA profiles of suspected or convicted criminals and DNA traces from crime scenes in the form of short tandem repeats (STRs) while some databases also include DNA profiles of victims and volunteers. When there are no known suspects of a crime, traditional forensic DNA testing uses these databases to compare an STR profile obtained from a crime scene with all STR profiles in the database to see whether there is a match.

Originally, STR profiles used in forensics were derived from markers located in non-coding DNA regions. It was argued that the police does not need more information than what is needed to identify the individual. As such, it is perceived that no information regarding disease or personal characteristics can be inferred. This was not completely true, as exemplified by the allele of the STR locus *TH01* where association was found with type I diabetes. When this allele is revealed to forensic scientists during DNA profiling, they can infer (if they wish so) that the person has increased risk of diabetes by 0.4% to 0.52% (Benecke 2002). Thus, the practical predictive value is absent and the possibility of data misuse remains in the theoretical plane (Bennett and Todd 1996).

With the advent of MPS, the efficient generation of data at the nucleotide level beyond that of STR profiles only has allowed laboratories to produce wide-ranging additional DNA information, including related to coding regions of the genome. Thus, it is increasingly hard to insist that only non-coding DNA can be used for criminal investigation. Indeed, in recent years, DNA biometrics have started to be used.

Biometrics means body measurements and calculations related to human individualising characteristics. Primary biometrics are DNA profiling and fingerprinting. However, there are currently other possibilities: palm veins, palm print, hand geometry, iris recognition, retina, body odour, computer typing rhythm, gait, voice, and face recognition. Historically, it is a continuation of the Bertillonage. The Bertillonage used to be the method of forensic identification founded by Alphonse Bertillon, based on anthropometric measurements of height, reach, trunk, length of head, width of head, right ear, left foot, left middle finger, and left forearm[6].

The ideal biometrical marker should be with a high differentiation power, easily and reproducibly testable, hard for someone to change on purpose, and stable during human life. Some of the biometrics (Externally visible characteristics, EVC) have a high level of heritability and can be considered phenotype, in the meaning of genotype expression. However, there is a fundamental distinction between DNA profiling by STR and Forensic DNA Phenotyping (FDP). In STR, we are comparing identical things, DNA profile with DNA profile. On the other hand, in FDP, we are comparing a command with its execution: the process of expression of information encoded in the language of nucleic acids, in DNA, starts by transcribing its parts to RNA and follows with translating them to the language of proteins.

The process of expression of information encoded in DNA is mostly interactively multigenic and is affected by the environment (for example, the embryo is affected by uterus environment, the child by vaccination and illnesses, and the adult by lifestyle and nutrition) and chance (unpredictable, non-influenceable, or stochastic changes at the molecular level). Only naïve genetic determinism neglects the role of environment and chance in the phenotypic markers. Also, phenotypic characteristics are not stable and can be changed on purpose.

Hair colour can change from childhood to adulthood, and can be affected by UV exposition, colouring, highlighting, straightening, and wearing a wig. Eyes colour can be masked by coloured contact lenses or reflective glasses. Faces change with ageing, injury, plastic surgery, cosmetics, and by the effects of drug abuse or smoking. It can be camouflaged by gluing a moustache or a beard, by wearing carnival masks, niqabs, burkas, balaclavas, bandannas, scarves, or anti-SARS-CoV-2 masks.

Nevertheless, most of the perpetrators act spontaneously, without preparation by disguising themselves and the mode of using FDP in forensics focuses on visible characteristics searchable in the population of suspects during the investigation and not during a criminal act. FDP is used in cases where DNA of the perpetrator is present on the crime scene, but the profile (neither of the individual nor of their kindred) is not found in DNA databases and investigators need to narrow the pool of suspects. The same applies for identification of cadavers that are decomposed or devastated by explosive blast, aeroplane crash, passing of human body through dam turbine, etc.

[6] https://journals.openedition.org/criminocorpus/2970?lang=en

FDP may be considered a quantified and better-qualified equivalent of eye witnessing. For comparison, the Innocence Project in USA revealed that 70% of the erroneous verdicts retrospectively identified by DNA profiling had been reached because of false eyewitness reports[7].

For evaluation of evidence strength of FDP in the investigative phase of forensic work, it is accepted to use other statistics than likelihood ratio (Table 6). It is because they are more practical and because the likelihood ratio for FDP is sometimes as low as 2 (compare with LR for STRs full profile from a non-degraded, non-mixed sample surpassing 10^{13}).

Table 6: Contingency table or confusion matrix

		True phenotype	
	Total population	Phenotype present	Phenotype absent
Predicted phenotype	Predicted present	True positive	False positive
	Predicted absent	False negative	True negative

A Receiver Operating Curve (ROC) is a graph that plots True Positive Rate of detection against False Positive Rate of detection. True positive rate is also known as sensitivity, recall, or probability of detection and is calculated as follows: True positive rate = (True positive) / (True positive + False negative). False positive rate is also known as the probability of false alarm and can be calculated as False positive rate = (False positive) / (False positive + True negative).

AUC is one number that combines parameters of test sensitivity (True positive rate) and specificity (True negative rate, measures the proportion of actual negatives that are correctly identified as such). AUC means area under ROC, with values between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier.

Precision or Positive predictive value (PPV) is (True positive) / (True positive + False positive) while Negative predictive value (NPV) is (True negative) / (True negative + False negative). PPV can be considered as the posterior probability that is obtained by combining prior probability and likelihood ratio using Bayes' theorem.

The pioneering FDP work was started by the Human Longevity Company led by Craig Venter. They tried to identify individuals by trait (facial structure, voice, eye colour, skin colour, height, weight, and BMI) using WGS data from 100 ng of DNA (Lippert et al., 2017). For a large fraction of the traits, their predictive accuracy beyond ancestry and demographic information was limited.

Most media attention was received by a company named Parabon Labs with their Snapshot DNA Phenotyping Service claiming to be able to predict genetic ancestry, eye colour, hair colour, skin colour, freckles, and face shape in individuals from any ethnic background, even individuals with mixed ancestry. However, there is not any peer-reviewed data available to substantiate their claims.

The European synonym for FDP studies is a project with acronym VISAGE, Visible Attributes through Genomics (http://www.visage-h2020.eu/). While most of the FDP results presented in the next tables were achieved by VISAGE members (Palencia-Madrid et al. 2020), other scientific groups provide valuable data as well (Montesanto et al. 2020) and the list has no intention to be comprehensive.

The obvious externally visible characteristics are hair colour, skin colour, eye colour, and face. The AUC values lie in the range 0.64–0.94 for hair colour (Table 7), 0.72–0.99 for skin colour (Table 8) and 0.74–0.99 for eye colour (Table 9), depending on the predictive model and colour category used. For face, AUC reaches 0.8 (Claes et al. 2014; Sero et al. 2019; Xiong et al. 2019). These numbers correspond to the likelihood ratio between 2 and 200.

[7] https://www.innocenceproject.org/all-cases/

Table 7: Hair colour FDP parameters. AUC, Area Under receiver operating Curve, combines parameters of test sensitivity and specificity. It lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier. PPV, Positive predictive value, is (True positive)/(True positive + False positive) while NPV, Negative predictive value, is (True negative)/(True negative + False negative). For example, when HIrisPlex kit result indicates red hair colour then you may be 73% sure that the hair colour is indeed red (adapted from Schneider et al. 2019).

Kit	Number of SNPs	Red			Black			Blond			Brown		
		AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV
HIrisPlex/ HIrisPlex-S	24	0.92	0.73	0.97	0.83	0.7	0.91	0.8	0.63	0.79	0.72	0.58	0.72
SHEP 1. 2	12	0.94	n.a.	n.a.	0.86	n.a.	n.a.	0.84	n.a.	n.a.	0.64	n.a.	n.a.
ForenSeq™ DNA Signature Prep Kit (Verogen)	24	0.9	0.77	0.91	0.78	0.45	0.9	0.75	0.67	0.72	0.72	0.21	0.91
Identify (Identitas)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Parabon Snapshot (Parabon Nanolabs)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table 8: Skin colour FDP parameters. AUC, Area Under receiver operating Curve, combines parameters of test sensitivity and specificity. It lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier. PPV, Positive predictive value, is (True positive)/(True positive + False positive) while NPV, Negative predictive value, is (True negative)/(True negative + False negative). For example, when HIrisPlex-S kit result indicates that colour of the skin is not dark to black then you may be 99% sure about that (adapted from Schneider et al. 2019).

Kit	Number of SNPs	Very light			Light			Intermediate			Dark			Dark to black		
		AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV
HIrisPlex-S	41	0.74	0.4	0.94	0.72	0.6	0.72	0.73	0.6	0.73	0.88	0.34	0.98	0.96	0.81	0.99
SHEP 1. 2. 4	10	0.99	n.a.	n.a.	n.a.	n.a.	n.a.	0.8	n.a.	n.a.	n.a.	n.a.	n.a.	0.97	n.a.	n.a.
Identify (Identitas)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Parabon Snapshot (Parabon Nanolabs)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table 9: Eye colour FDP parameters. AUC, Area Under receiver operating Curve, combines parameters of test sensitivity and specificity. It lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier. PPV, Positive predictive value, is (True positive)/(True positive + False positive) while NPV, Negative predictive value, is (True negative)/(True negative + False negative). For example, when ForenSeq DNA Signature Prep Kit result indicates that the eye colour is intermediate then you may be 8.5% sure about that (adapted from Schneider et al. 2019).

Kit	Number of SNPs	Blue			Brown			Green-hazel			Intermediate		
		AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV
IrisPlex/ HIrisPlex/ HIrisPlex-S	6/ 24/ 41	0.94	0.9	0.9	0.95	0.77	0.96	n.a.	n.a.	n.a.	0.74	0.085	0.96
SHEP 1. 2	23	0.999	0.8	0.99	0.99	0.51	0.96	0.82	0.55	0.93	n.a.	n.a.	n.a.
ForenSeq™ DNA Signature Prep Kit (Verogen)	24	0.94	0.9	0.9	0.95	0.77	0.96	n.a.	n.a.	n.a.	0.74	0.085	0.96
Identify (Identitas)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Parabon Snapshot (Parabon Nanolabs)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Face

The human face represents a combined set of highly heritable phenotypes, but knowledge of its genetic architecture remains limited, despite the relevance for forensics and despite claims of commercial companies that can do reliable genetic face predictions. An international study defined 78 facial shape phenotypes from 3-dimensional facial images and identified 24 associated genetic loci (Xiong et al. 2019). A global map of derived polygenic face scores assembled facial features in major continental groups consistent with anthropological knowledge. These results substantially advance our understanding of the genetic basis underlying human facial variation but are not yet ready for forensic use.

Additional FDP markers are eyebrow colour, height, and the presence of freckles.

Eyebrow colour

Eyebrow colour may share a large genetic component with scalp hair colour, which has an estimated heritability of up to 90%. Phenotypic relationship between eyebrow and scalp hair colour is not perfect, suggesting the existence of overlapping and unique genetic components for both traits.

Eyebrow colour was graded into four broad ordinal categories (red, blond, brown, and black) by using photonumeric scales. A model including 25 SNPs achieved prediction accuracies expressed as AUC of 0.701 for blond, 0.620 for brown, and 0.674 for black eyebrow (Peng et al. 2019).

Height or stature

689 SNPs provided an AUC of 0.79, while a subset of 412 SNPs achieved 0.76 (Liu et al. 2019).

Freckles

Freckles or ephelides are hyperpigmented spots on the skin observed in Europeans and Asians. Pigmentation genes explain a significant portion of freckles heritability. Non-freckled, medium-freckled, and heavy-freckled people can be predicted with AUC = 0.75, 0.66, and 0.79 (Kukla-Bartoszek et al. 2019).

Other phenotypic markers

Estimating a person's age from DNA found at the crime scene is also part of forensic DNA phenotyping but will not be considered in this report, as it differs from the forensic DNA phenotyping in its molecular basis, analytical methods, and sample requirements. The same holds for human odour. The human odour is not an externally

visible but an Externally Sniffable Characteristics (ESC), detectable by nose of specially trained dogs or by techniques like mass spectrometry. Its heritable part is not currently sufficiently studied: it would definitively include human major histocompatibility complex (HLA) and maybe also human skin microbiome.

Legislative frameworks

Since 2019, forensic DNA phenotyping is explicitly permitted by law in the Netherlands and Slovakia and practised in compliance with existing laws in Austria, the Czech Republic, Germany (with the exception of biogeographical ancestry), Hungary, Poland, Spain, Sweden, and the United Kingdom. In Switzerland, it is forbidden under current law, but the legalization of FDP is currently being considered (http://www.visage-h2020.eu/Report_regulatory_landscape_FDP_in_Europe2.pdf).

In countries where FDP is used in criminal investigations, genetic information about SNPs is stored de-centralised in laboratories performing the analysis. This requires consideration of how such findings should be stored, who should have access to them and how findings (which are highly probabilistic and predictive and raise issues of discrimination) should be communicated with operative police work. Even members of the police are affected by the CSI effect, as exemplified by the refusal of the criminalistic technicians to provide their DNA profiles into the elimination database. Recommendations and training are provided by the VISAGE consortium, regarding FDP interpretation including software use, the circumstances under which the use of FDP by law enforcement officers can be justified, and the need for a transparent evaluation of any costs and benefits of FDP use in the criminal justice system http://www.visage-h2020.eu/PDF/Delliverable_5.2_for_online_publication_vo1.pdf.

4.4 DNA as health data

The advent of massively parallel DNA sequencing has propelled clinical genetics into a new era, with unprecedented scope and comprehensiveness. It ended the stepwise or reflexive partial genotyping, “diagnostic odyssey”, for many patients and families, expanded the known phenotypes of countless disorders, and led to new disease gene discoveries. It is estimated that MPS provides 5 times more diagnosis at 1/3 cost of the previous standard of care. Genomics is being integrated into personalized medicine with opportunities in pharmacogenomics, undiagnosed diseases in critically ill newborns, tumour sequencing for tailor-made treatment, germline risk prediction, implementation research (local adaptation, sustainability of evidence-based interventions, scaling up healthcare interventions across health plans, and stopping of suboptimal care), and capturing evidence (expert panels and the ClinVar database). Clinical utility of computational phenotyping for genetic diseases is increasingly appreciated and global clinical collaborations on identifiable patient data are on the way, e.g., The Minerva Initiative (Nellaker et al. 2019).

Nevertheless, clinical genomics still fails to identify the molecular cause in many patients who clearly exhibit genetic/syndromic conditions, while at the same time fails to assign a clinical significance to the found sequence variants (Grody 2019).

Large whole genome sequencing projects

Thus, larger projects with the aim to do whole genome sequencing for representative populations emerged around the world (Table 10).

Table 10: Large genome sequencing projects

Country/Continent	Genomics initiative
Australia	Australian Genomics Health Futures Mission
Dubai, United Arab Emirates	Dubai Genomics
Estonia	Personalized Medicine Programme
Europe	European 1+ Million Genomes Initiative
France	French Plan for Genomic Medicine 2025
China	100,000 Genomes Project
Japan	Initiative on Rare and Undiagnosed Diseases (IRUD)
Saudi Arabia	Saudi Human Genome Program
Turkey	Turkish Genome Project
United Kingdom	100,000 Genomes Project
United States	All of Us Research Program

The European 1+ Million Genomes Initiative "Towards access to at least 1 million sequenced genomes in the EU by 2022" with 21 signatory Member States (including UK) and Norway had the ambition to sequence the same number of persons as all the other projects together. It wants to improve disease prevention, allow for more personalized treatments (therapies, medicines, and interventions), enable better diagnostics, provide a sufficient scale for new clinically impactful research, and make more efficient use of resources in healthcare that would never suffice. Among the objectives, there is the building of technical infrastructure allowing for secure, federated access to genomic data.

Health information is considered the most sensitive private information. However, it is shared by a surprisingly large number of individuals and institutions[8] (Figure 10).

[8] https://thedatamap.org/map2013/index.php

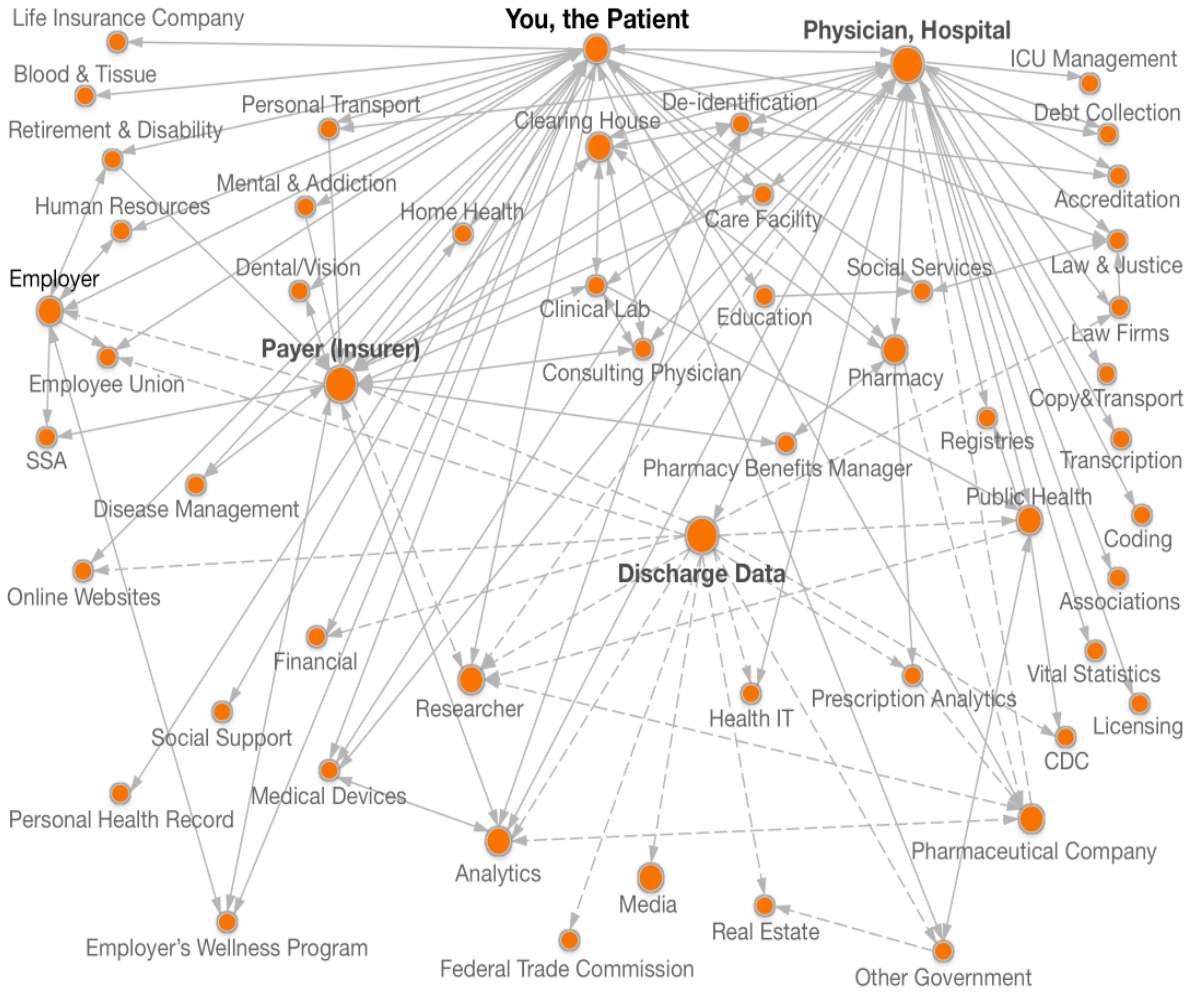

Figure 10: Legal sharing of health information (Figure adapted with the permission of the author Latanya Sweeney, https://thedatamap.org/map2013/index.php)

Moreover, health information is marketed, with the top buyers of Personal Health Information being Truven Health Analytics, Optuminsight (Ingenix), Milliman, WebMD Health, IMS Health (SDI Health and Verispan), Intellimed International, Service Employees International Union (SEIU), DataBay Resources, iVantage Health Analytics (Health Info Technics), and Health Market Science https://thedatamap.org/map2013/statebuyers.php.

As genomics can reveal a lot of health data, genetic data sharing policies are already well established (Table 11).

Table 11: Genetic data sharing policies

OECD Recommendation on Human Biobanks and Genetic Research Databases (HBGRD)	Recommendations on HBGRD were adopted in 2009 by the OECD council as a non-legally binding contract: research must respect the participants and be conducted in a manner that upholds human dignity, fundamental freedoms, and human rights and be carried out by responsible researchers.
Public Health Genomics European Network (PHGEN) II	PHGENII produced the European Best Practice Guidelines for Quality Assurance, Provision and Use of Genome-based Information and Technologies. It introduced the notion of primary data subject and family members due to hereditary feature of certain information which is of shared value for a family.
Global Alliance for Genetics and Health (GA4GH)	GA4GH is an international non-profit alliance, stressing confidentiality, integrity, availability of data, privacy of individuals, families, and communities whose genetic data are shared. If the data are coded or anonymized, it should take place at the earliest opportunity consistent with use of the authorized purposes. Data Stewards should provide a clear summary or description of the coding of the anonymization process that should ensure that further robust data linkage would not be possible.
NIH Genomic Data Sharing (GDS) Policy and Policy for Sharing Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)	NIH GDS Policy applies to NIH-funded research. NIH expects that informed consent for future research use and broad data sharing will have been obtained even if the cell lines or clinical specimens are re-identified. A risk of re-identification must be conveyed to prospective subjects in consent process because identification of specific individuals from raw genotype-phenotype data is feasible and increasingly straightforward.
Welcome Trust Sanger Institute Data Sharing Policy	The Welcome Trust Sanger Institute is a non-profit British trust. It lists factors that can contribute to lowering the risk of re-identification; acknowledges crosswalk from medical bioethics to forensic bioethics (Wickenheiser 2019a) and derivation of medical conditions in a forensics context.
American Society for Human Genetics (ASHG) core principles (Directors 2019)	The American Society for Human Genetics is a professional genetics society. It requires individuals to have a right to maintain the confidentiality of their own genetic information and not be compelled to disclose it. Entities holding human genomic data must take robust measures to protect the confidentiality of individual´s medical and genetic information. The users of genomic research participants´ information should assess the risks and benefits for both the participants and the society. The nature of genetic analyses in context should determine which privacy protections and data-sharing practices are appropriate – when it is desirable and appropriate for genetic information to be treated the same way as other biological, health, or personal information and when there are factors that require genetic information to be treated differently from other forms of health data. Research policies should both facilitate data sharing and protect the confidentiality of research participants' medical and genetic data in a way that both advances research and respects participants´ preferences.

In healthcare, genetic data are subjected to informed consent and (pseudo)anonymization.

Informed consent

The Nuremberg Code of 1947, the ten commandments of human subject research, was drafted in the aftermath of World War II during the Nuremberg War Crime Trials, in which Nazi doctors were charged with conducting murderous and torturous human experiments in the concentration camps. It states that explicit, voluntary, and informed consent from patients is required for human experimentation. A human research subject authorizes the anticipated medical procedure or research study prior to its performance and requires that the permission be given voluntarily and with knowledge of the facts, risks, and benefits to the individual human research subject.

A typical informed consent for genetic testing consists of:

- A general description of the test, including the purpose of the test and the condition for which the testing is being performed.
- How the test will be carried out (from which biological sample).
- What the test results mean, including positive and negative results, and the potential for uninformative results or incorrect results such as false positives or false negatives.
- Any physical or emotional risks associated with the test.
- Whether the results can be used for research purposes.
- Whether the results might provide information about other family members´ health, including the risk of developing a particular condition or the possibility of having affected children.
- How and to whom test results will be reported and under what circumstances results can be disclosed (for example, to health insurance providers).
- What will happen to the test specimen after the test is complete.
- What will happen to the results when reanalysed using new knowledge.
- Acknowledgement that the person requesting the testing has had the opportunity to discuss the test with a healthcare professional.
- The individual´s and witnesses´ signatures.

When we acknowledge the blurring boundaries between DNA-based information inside and outside of forensic databases[9] (Bradbury et al. 2019), we should think about updating the informed consent for forensic use. It is good to avoid a potential future function creep when information is used for other purposes than those originally specified.

For example, in the GEDmatch disclaimer, it is stated that some of the possible uses of Genealogy Data by any registered user of GEDmatch include but are not limited to:

- Discovery of identity, even if there is an alias, unidentifiable e-mail address, and other obscuring information.
- Finding genetic matches (individuals that share DNA).
- Paternity and maternity testing.
- Discovery of unknown or unidentified children, parents, or siblings.
- Discovery of other genetic and genealogical relatives, including both known and unknown or unexpected genetic and genealogical relatives.
- Discovery of ethnic background.
- Discovery of a genetic relationship between parents.
- Discovery of biological sex.
- Discovery of medical information or physical traits.
- Obtaining an email address.

[9] http://www.visage-h2020.eu/PDF/Delliverable_5.2_for_online_publication_vo1.pdf

— Familial searching by third parties such as law enforcement agencies to identify the perpetrator of a crime, or to identify remains.

Anonymisation

Anonymised data must be entirely stripped of any identifiable information, making it impossible to derive insights on a tangible individual, even by the person or entity who performed the anonymisation. In other words, anonymisation cannot be reversed.

The term pseudonymisation is more properly applied to the situation where someone masks or replaces the name or date of birth of the tester with a unique identifier that will accompany the sample through the whole process of sequencing. Pseudonymisation under GDPR (Article 4(5)) is defined as "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".

Though there are different anonymisation strategies (Kushida et al. 2012), in fact, anonymisation of the whole genome is not possible (Ohm 2010), given that the purpose of clinical genetics and genetic genealogy conflicts with the very concept of anonymisation. Therefore, there are attempts to preserve privacy of genomic data by more complicated means, using cryptographic techniques.

Techniques for preserving the privacy of genomic data

To preserve privacy of genomic data, one must ensure data security (blocking unauthorised user´s access to original data) and data anonymisation (protecting the identity/presence of the individual in shared data).

Access control grants access only to authorised users (trust but verify). This approach has been used in data repositories, such as dbGaP[10].

Cryptographic primitives are low-level cryptographic algorithms used to build cryptographic systems to provide information security. Different cryptographic privacy-preserving primitives developed for genomic data are: k-anonymity, Differential privacy, Homomorphic encryption, Secure cryptographic hardware, Secure multi-party computation (Al Aziz et al. 2019), Probabilistic modelling, and the Internet of Things approach.

k-anonymity transforms data such that, for each record in the output, there are $k - 1$ other records with the same set of quasi-identifiers (achieved by generalisation and suppression of SNPs) (Malin 2005).

Differential privacy adds a controlled amount of noise on the disclosure of data or any query result (Raisaro et al. 2018).

Homomorphic encryption preserves certain structures that allow arithmetic operations to be performed directly on its ciphertext (Kim et al. 2017).

Secure cryptographic hardware uses hardware to complement software for data encryption and protection. It is most often implemented as part of the processor's instruction set and comes in the form of cryptographic coprocessors, accelerators, chip cards, and smart cards (for example as Software Guard Extensions (Chen et al. 2017a; Chen et al. 2017b), ARM TrustZone, IBM cryptographic coprocessor, and Field-programmable gate array).

Secure multi-party computation is an arrangement for multiple parties collaborating in computation a function on their data while keeping them private except for the computation result. The most popular way of implementing are garbled circuit, homomorphic encryption, and secret sharing (Bogdanov et al. 2018).

Probabilistic modelling combines minimal amounts of perturbation with Bayesian statistics and Markov Chain Monte Carlo techniques (Simmons et al. 2019).

Internet of Things approach separates data and users and focuses on the transmission and sharing security of data while providing users with mining methods (Wu et al. 2020).

Despite the availability of privacy-enhancing technologies, there is a gap between current approaches and their applicability because of their impracticability and trade-off between security and efficiency. Most existing frameworks for secure computation are not scalable. They incur significant computational overhead for large-scale and complex data analysis tasks, which are common in biomedical data analysis.

Though some solutions exist, e.g., genomic global positioning system GPS (Kim et al. 2019) that applies the multilateration technique commonly used in the GPS to genomic data or some solutions may arise as result of Secure Genome Analysis competition organized by integrating Data for Analysis, Anonymization and Sharing (iDASH), it is desirable that they are under the umbrella of international standard-setting organisations for

[10] https://www.ncbi.nlm.nih.gov/gap/

genomics research pipelines, such as the Global Alliance for Genomics and Health (GA4GH) and the MPEG-G Consortium (Berger and Cho 2019).

Online data are not safe

We may distinguish between the risk for disclosure of the identity of an individual, information about the individual's health and behaviour, including previously unknown phenotypes, and information about individual's blood relatives. Unfortunately, all these risks were already realised.

Privacy attacks can be performed by techniques of identity tracing, attribute disclosure, and completion attack (Bonomi, Huang and Ohno-Machado 2020; Yakubu and Chen 2020) (Table 12):

Table 12: Attacks on privacy of genetic data

	Adversary background knowledge	IT inference technique	Adversary goal, prediction
Identity tracing attacks	Victim's demographic data Victim's surname, Y-chromosome haplotypes and demographic data VCF file of the victim's genome, # of individuals in the beacon, SFS of the population in the beacon VCF files of people from the victim's population, corresponding MAF and LD and high-order correlation Victim's genotype, familial relations, demographic data (location, age, and sex)	DNA matching Data linkage, e.g., demographic data or social interaction matching Search space pruning Statistical hypothesis testing, Likelihood-ratio test, and High-order Markov chain model DNA phenotyping Pedigree/genealogy	Re-identify participants of the personal genome project Triangulate identity from surnames and Y chromosomes Re-identifying individuals and their relatives within a beacon Re-identify an individual within a data set in a beacon Re-identify an individual by long-range familial search
Attribute disclosure (phenotype inference) attack	Victim's SNP profile, GWAS statistics: set of SNPs Warfarin dose predicting model, victim's demographic data and dosage Victim's anonymised genotype and phenotype, SNP-trait association	Statistical testing, distance measure Model inversion, genotype imputation, kinship, linkage disequilibrium Statistical methods based on SNP correlation Data mining and matching techniques, deterministic proofs of study inclusion Summary statistics, machine learning methods	Determine the presence of an individual in a GWAS Infer sensitive genetic markers of an individual from a warfarin dosage pharmacogenetic model Predict an individual's predisposition to Alzheimer's disease
Completion attacks	Family members pedigree structure, genotypes (linkage disequilibrium between haplotypes) Family members familial relationships, genotypes (linkage disequilibrium between SNPs, Minor allele frequencies) Family members familial relationships, genotypes (linkage disequilibrium between SNPs, Minor allele frequencies) and phenotypes Pedigree structure, victim's partial genotype (high order correlation between SNPs) and phenotypes Genotypes and phenotypes of the victim and relatives, and SNP-trait association from GWAS	Genotype imputation, haplotype sharing graph Belief propagation, factor graph	Infer haplotypes of ungenotyped individuals from their relatives' genetic information Infer an individual's genotype from their relatives' genomes Infer an individual's genotype from their relatives' genomes and phenotypes Reconstruct missing parts of an individual's genotype Predict the genotypes and traits of individuals based on publicly available genome data and traits released by individuals or their relatives

To illustrate the scale of the problem, Figure 11 reproduces the statistics regarding general GDPR breaches, health/genomics data included in Statista[11] at the time of this report.

[11] https://www.statista.com/chart/20566/personal-data-breaches-notified-per-eea-jurisdiction/

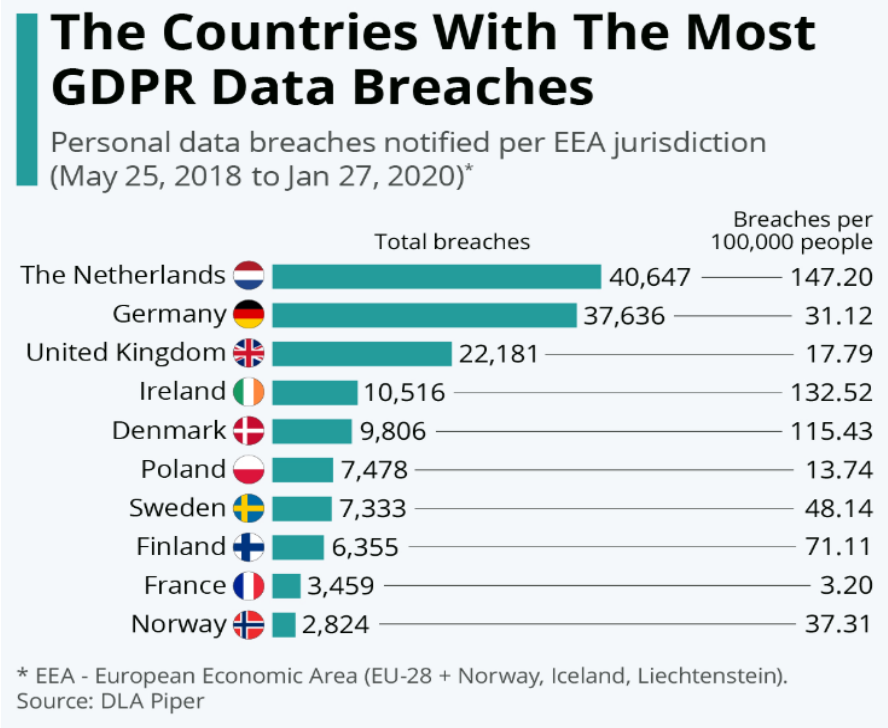

Figure 11: GDPR data breaches (Source: https://www.statista.com/chart/20566/personal-data-breaches-notified-per-eea-jurisdiction/)

U.S. Department of Health and Human Services Office for Civil Rights runs the webpage Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information[12] (Figure 12, Figure 13).

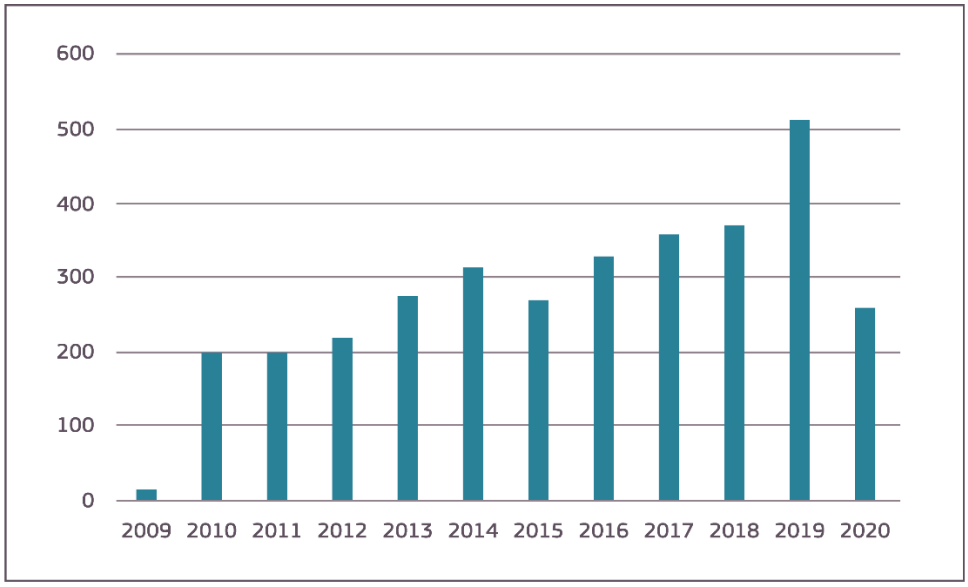

Figure 12: Number of data breach incidents according to the Breach Portal website

[12] https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (accessed 17th July 2020)

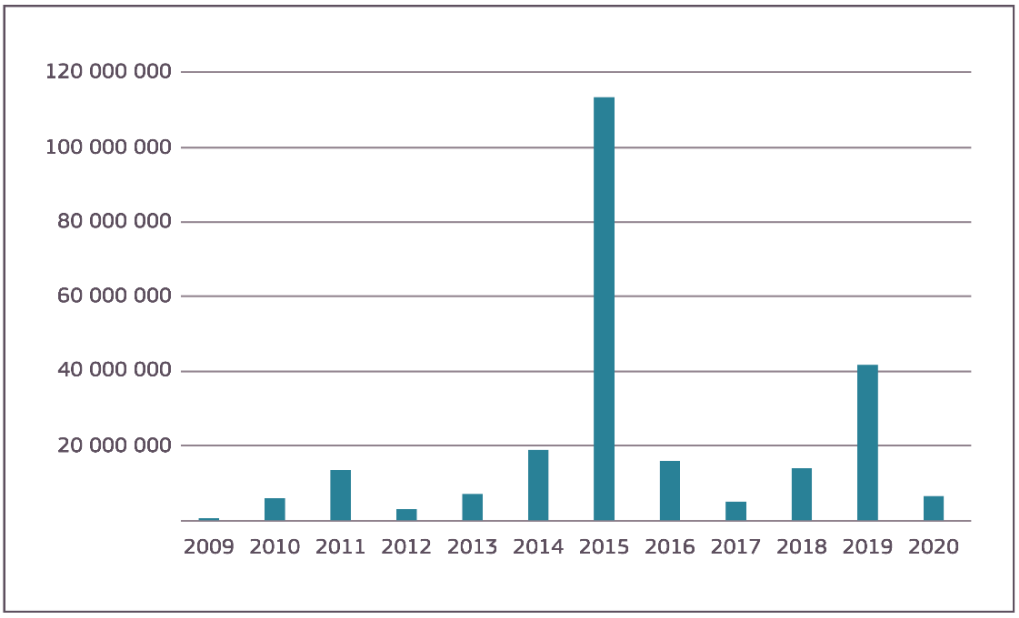

Figure 13: Number of persons affected by data breach incidents according to the Breach Portal website

The outliers of years 2015 and 2019 are caused by data breaches in companies with databases affecting over 10,000,000 people (in 2015: Anthem Inc., Premera Blue Cross, and Excellus Health Plan, Inc.; in 2019: Optum360, LLC and Laboratory Corporation of America Holdings, LabCorp).

In Singapore, a successful cyber-attack was carried out on SingHealth. In this breach, data of 1.6 million Singaporean consisting of 1/3 of the population have been uploaded to the darknet.

These were cases for general biomedical data. The most recent genealogical example happened on July 19 2020, when GEDmatch experienced a security breach orchestrated through a sophisticated attack on a server via an existing user account. The website was taken down. As a result of this breach, all user permissions were reset, making all profiles visible to all users. This was the case for approximately 3 hours. During this time, users who did not opt-in for law enforcement matching were available for law enforcement matching and, conversely, all law enforcement profiles were made visible to GEDmatch users. Anybody with the proper skill could have downloaded all the data available.

Then, GEDmatch was the target of a second breach in which all user permissions were set to opt-out of law enforcement matching. DNA information was not compromised, as GEDmatch does not store raw DNA files on the site (when data are uploaded, the information is encoded, and the raw file deleted).

However, MyHeritage customers who are also GEDmatch users were the target of a phishing scam.

The examples illustrated quite clearly that genetic data on the Internet are not safe. Also, failure of protective mechanisms of such extent refreshes the call for universal forensic genetics database (Hazel et al. 2018) that was implemented so far only in Kuwait.

European collaboration in forensic genetics

Police investigators across Europe cooperate in the field of forensic genetics in the framework of the Prüm Convention, and soon in the framework of the Central Schengen Information System[13]. Many of the established communication channels, safety measures, quality checks, and logistic procedures can be used in the current or amended form to allow investigators to use the Whole Genome Sequencing (WGS) data. Also, it is expected that the so-far encountered challenges in the implementation (IT problems, privacy and data protection issues,

[13] The Central Schengen Information System was established to contribute to police and law enforcement cooperation between Member States and to support external border control. Since 2018, EU Regulation 2018/162 added the possibility to introduce DNA profiles for missing person's alerts. It is foreseen in the near future that persons can be sought for their own protection or for the protection of society against threat to public order or public security, and that DNA profiles are allowed in the cases when other biometrics means (photographs, facial images, and dactyloscopic data) fail (Angers et al. 2019).

legal issues, national structures, lack of information, lack of human resources and funding) will be met also in the case of WGS implementation and previous experience would be helpful (Angers et al. 2019).

The Prüm Convention and DNA information exchange

The Prüm Convention is an expansion of the Hague Principle of Availability. It is a law enforcement treaty signed on 27 May 2005 by Austria, Belgium, France, Germany, Luxembourg, the Netherlands, and Spain in the town of Prüm in Germany, open to all members of the European Union, 24 of which are currently members. Core elements of the Convention were picked up by Prüm Decision, EU Council Decision 2008/615/JHA on 23 June 2008 on the stepping up of cross-border cooperation, particularly in combatting terrorism and cross-border crime.

The Prüm Convention is the result of scientific, ethical, and political decisions: scientific - because they compare DNA profiles using scientific technique; ethical - because the infrastructure is not error-free (i.e., false positives, potential of abuse of power or data misuse); and political - because it is intended as a voluntary cooperation mechanism (Matos 2019).

The Convention was adopted so as to enable the signatories to exchange data regarding DNA, dactyloscopic data and vehicle registration of concerned persons and to cooperate against terrorism (by the deployment of armed sky marshals on flights between signatory states, joint police patrols, entry of armed police forces into the territory of another state during a hot pursuit, and cooperation in case of mass events or disasters). On the contrary, DNA exchange is considered the most sensitive issue because it contains far more information than a simple fingerprint.

The Prüm regime allows the exchange of forensic DNA information across the national databases of 24 EU Member States. Prior to Prüm, countries sent entire files, including personal data, from one country to another when they collaborated on an investigation. Now with Prüm, the identity of the originator of a DNA profile is not disclosed to authorities in other countries unless the information held in both countries' data repositories (here called "Prüm databases") indicates a match. Matching is checked by running informatic software (here called "Prüm software"). Only then such identifying personal information will cross borders. This underscores the importance of well-designed (in terms of privacy by default and privacy by design) systems for digital data exchange which can reduce privacy risks if implemented well.

The Prüm software was developed jointly by DNA and IT experts from the Bundeskriminalamt (BKA) in Germany, the Ministry of the Interior of Austria and the Netherlands Forensic Institute in the Netherlands. From the Prüm database of a country, DNA profiles[14] can be sent to other countries for comparison to their Prüm databases. A country can decide to send a DNA profile to one or more selected countries or to all operational countries to which it is connected.

1. DNA profiles that meet the Prüm inclusion rules are copied from the National Criminalistic DNA database to the Prüm database of a country at a predetermined frequency. The Prüm database of a country can either be a physical copy or a view of the National Criminalistic DNA database.
2. From the Prüm database, DNA profiles can be sent to other countries using the Communication Tool, which converts the DNA profile into an encrypted e-mail attachment that is sent to one or more other countries via the secure European TESTA network (https://ec.europa.eu/isa2/solutions/testa_en). Encryption is done through Secure/ Multipurpose Internet Mail Extensions (sMIME).
3. The e-mail arrives at the e-mail server of the requested country.
4. The Communication Tool of the requested country picks up the e-mail attachment and decrypts it.
5. The Communication Tool of the requested country puts the DNA profile in the Request and Response Database of that country.

[14] In forensics genetics, STR DNA profiles can be expressed as alphanumeric text, formatted using an XML mark-up language in accordance with norm ISO 19794-14. The mitochondrial DNA sequence can be expressed in the FASTQ format that covers not only the sequence composition but also quality metrics. Accordingly, the communication between two whole genome sequencing datasets can run in FASTQ format or metadata-enhanced BAM format. From BAM format, FASTQ information is extractable and the updated form of BAM, uBAM or unmapped BAM, allows a data manager to attach metadata to the reads as early in the analysis process as possible. Communication between two types of sequence data (e.g., STR profile on one side and WGS on the other side) would need translational interface or both sides agreeing to the same standards (Gettings et al. 2019).

6. The Matching Tool of the requested country picks up the DNA profile from the Request and Response Database and compares the DNA profile with the Prüm database of that country.
7. The Matching Tool puts the result of the comparison (HIT or NO-HIT) back in the Request and Response Database of the requested country, where it can be viewed via the Graphical Use Interface. Any match of more than 6 STR loci is reported. However, even in the case of a HIT notification, the response does not contain any personal information. Further investigations will be carried out by an identification number that allows the exchange of more detailed information on the results by specially authorized officers (Angers et al. 2019).
8. The Communication Tool of the requested country picks up the result of the comparison and converts the result of the comparison into an encrypted e-mail attachment.
9. The e-mail is sent to the requesting country via the secure European TESTA network.
10. The e-mail arrives at the e-mail server of the requesting country.
11. The Communication Tool of the requesting country picks up the e-mail attachment and decrypts it.
12. The Communication Tool of the requesting country puts the result of the comparison in the Request and Response Database of that country, where the results can be viewed via the Graphical User Interface by both the requesting and requested country.

If we follow 2010 data for the Netherlands, we can see what happens after the match is found (Figure 14) (Toom 2018; Toom et al. 2019).

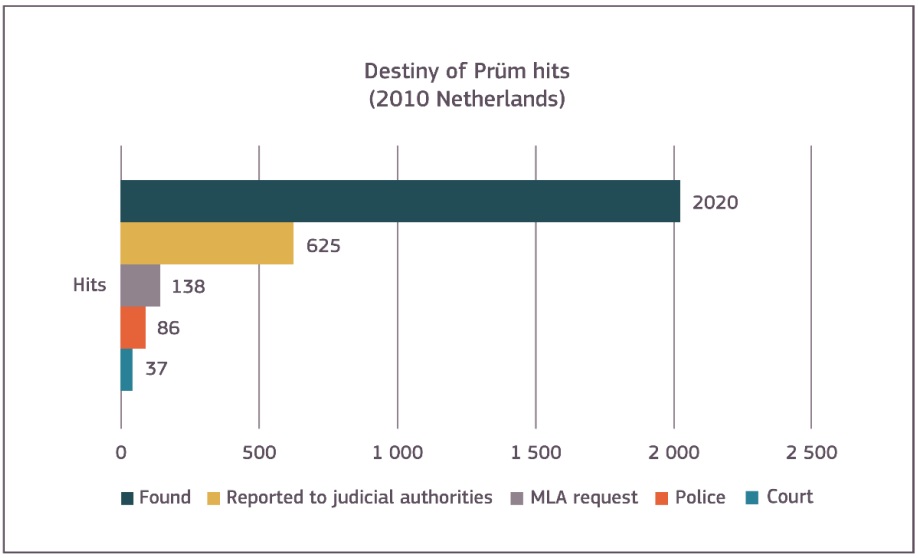

Figure 14: What happens following a DNA hit in requests under the Prüm regime

Only 37/2020 = 1.8% of the hits reach the Court stage. This may seem very high drop-out with low utility of cross-border DNA data exchange. However, it can be partially explained by reported matches with low likelihood ratio (e.g., 6 STR loci match), further hit selection, evaluation, and prioritization with regards to investigator´s tactic, case development, legal aspects, and time distance from crime.

Data suggest a trend for West and Central European countries to concentrate the majority of Prüm matches domestically, while DNA databases of the Eastern European countries tend to contribute with profiles of people that match crime scene traces in other countries (Santos and Machado 2017).

However, the official reports are not comprehensible enough to deeply analyse the type and amount of information that is exchanged between the Prüm system members and to provide an evidence-based assessment of its functioning and societal benefits with detail (Toom, Granja and Ludwig 2019).

To make cross-border exchange of DNA data including WGS even more transparent and accountable, the Council of Europe´s 12 principles of Good Governance should be applied[15].

[15] Council of Europe´s 12 principles of Good Governance: 1. Participation, Representation, Fair Conduct of Elections 2. Responsiveness 3. Efficiency and Effectiveness 4. Openness and Transparency 5. Rule of Law 6. Ethical Conduct 7. Competence and Capacity 8. Innovation and Openness to Change 9. Sustainability and Long-Term Orientation 10. Sound Financial Management 11. Human Rights, Cultural Diversity and Social Cohesion 12. Accountability. See https://www.coe.int/en/web/good-governance/12-principles.

5 Concerns

There are two sets of societal goods that must be kept in balance in assessing forensic genomics: a safe society without threat of serious crimes against human dignity, established by timely and efficient identification of offenders by police investigators and their subject to justice versus fundamental human rights: freedom, autonomy, privacy, presumption of innocence, and equality on the side of every individual. This is a right versus right dilemma when technological hubris, in which scientific certainty and technological robustness are overstated and social consent is assumed without discussion, should be avoided (Williams and Wienroth 2017).

Balance is sought not only for results but also for the process. The Council of Europe proposes that "the principle of proportionality requires that there be a reasonable relationship between a particular objective to be achieved and the means used to achieve that objective." While investigators look for a murderer, they are dealing with a suspect whose rights also constitute the building blocks of our democratic societies.

Many of the concerns are not specific to whole genome sequencing but apply to other Big Data technologies (non-genetic biometric technologies, facial recognition, etc.).

Criminal justice focuses on uncovering the truth and protecting the rights of innocent people together with identifying and punishing offenders and deterring crime. Upholding people's safety and combatting crime are common goods that justify constraints on individual rights. At the same time, defendants must be presumed to be innocent until proven guilty and special attention must be given to procedures that protect defendants against the possibility of error and ensure equal access to evidence both for defence and for prosecution.

Members of our societies whose friends and family were affected by severe crimes that may previously not have been solved will benefit from progress in the investigation of these crimes. They may also benefit from cases being solved faster. Finally, the use of massively parallel sequencing could help to strengthen trust in the criminal justice system and could have a positive effect on community life and shared values.

In EU countries, there were 3,993 police-recorded intentional homicides and 583,000 assaults in 2018[16]. The percentage of solved homicide cases ranges from 77% in the Netherlands, through 90% in the Czech Republic[17] to 98% in Finland (Liem et al. 2019). These numbers compare favourably with 64% of cleared homicides for USA (Smith and Cooper 2013) but still, there are every year dozens of murderers at large, ready to commit further crimes.

Until now, the expansion of the scientific toolbox available to users of forensic genetic technologies increased the effectiveness of investigations and also exonerated the wrongly convicted individuals.

This led to a significant increase of biological samples being taken, stored, and analysed during an investigation, as well as to a widening of the categories of individuals who are subject to forensic DNA profiling. Still, greater effectiveness in identifying guilty persons is needed and whole genome sequencing can be one of the means to achieve this.

There are many concerns, both justified and unjustified, regarding forensic genetics and forensic genomics particularly[18]:

Accountability and transparency

If whole genome data outside of centralized DNA databases are used for forensic purposes, then control over data scale and quality may not lie in the hands of public authorities but private companies or even single individuals. This can create accountability and transparency deficits.

Autonomy

Collection, retention, and use of DNA without the consent of those from whom they were taken or retrieved, along with the information routinely derived from them, is a breach of individual human autonomy, even if it is performed by a state.

Commercialisation

Forensic genetics technology development for law enforcement is a public function and should not rely exclusively on commercial providers. Accountable public bodies need to play an important role in setting standards for technology validation and deployment. It is desirable that they also provide services themselves to follow commitment to public interest without commercial profits.

[16] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Crime_statistics
[17] https://www.policie.cz/clanek/statistiky-vrazd.aspx
[18] http://www.visage-h2020.eu/PDF/Delliverable_5.2_for_online_publication_vo1.pdf

Data protection

The General Data Protection Regulation (GDPR) is not applicable to genetic data processing by competent authorities for law enforcement purposes. The EU Police Directive from May 2018 provides exceptions in cases where notification would impede public interest in an ongoing investigation.

However, GDPR´s principle of data minimisation, purpose limitation, and storage limitation is present also in Council Framework Decision 2008/977/JHA. Protection against unauthorised or unlawful processing and accidental loss, destruction or damage, and transparency must be applied.

Distributed data

Forensic DNA Phenotyping data used to prevent, solve, and punish crimes exist decentrally in local laboratories, people's personal electronic devices, and servers of commercial companies. The supervision that is applied to centralised forensic is missing.

DTC and medical databases

When police search medical, research, and DTC genetic databases to allow comparisons with results obtained from the crime scene samples, then this may reduce the willingness of individuals to donate samples to medical databases, to do business for DTC companies, or to trust key social institutions.

Genetic inquiries can blur boundaries between application fields. There may be medical issues relevant to a criminal investigation, medical information and genetic information may be used in combination to identify the human body, or ancestral lineage may be relevant both for biomedical research and for forensic identification. After court approval, forensic science practitioners will possess information about a sample donor that is unknown to the donor themselves. Sample donors may not want to know that information and may not want to share this information with others.

Clinical, research, DTC, and forensic genomics seem to converge into translational genomics (Wolf et al. 2020).

Effectiveness

It is difficult to establish the utility of various DNA profiling and databasing techniques and their impact on crime detection. Although proponents of forensic genomics stress notions of efficiency, relevant data have proven to be difficult to capture and hard to interpret (Williams and Wienroth 2017).

Massively parallel sequencing contributions to investigating crimes has yet to be evidenced properly (Phillips, et al. 2009).

Ethnic discrimination

Biogeographical ancestry can lead to racial profiling based on cultural bias at individual and institutional levels in law enforcement and security agencies who deploy forensic genetics technologies. Any technology is always technology in practice, deployed in social and logistic contexts.

Ethnicity is frequently problematized as causing unequal treatment of those sampled by police forces or security agencies. It may reinforce the stigmatization or even aggravate vulnerabilities of minorities. If improperly used, Forensic DNA Phenotyping can lead to vindication of the mistaken belief of a biological basis of race.

If MPS were used to suspect a group of people around ethnic, religious or cultural lines (Queiros 2019), then this is likely to have a destructive effect on social cohesion, especially if these minorities were associated with a higher prevalence of crime and thus, higher prior probability. These risks relate to the context of unintentional but structural racism, which is rather embodied by our societal and political institutions and shared practices.

External visible characteristics

Genetic diseases and their predispositions are excluded from forensic DNA phenotyping, as it is generally accepted that their forensic use would disproportionately violate privacy. The view that Externally visible characteristics (EVC) are non–privacy-sensitive markers, similar to eyewitness but more reliable and, as such, there is no reason for excessive privacy concerns, is not universally shared (Toom et al. 2016).

The use of FDP technologies represents a clear departure from earlier assurances to critics that forensic genetics only uses non-coding parts of the genome that do not hold any significant information about the individual and its relatives.

Freedom, dignity, and bodily integrity

The principle of inherent worth of a human being in a community of equal beings and ethical freedom for self-fulfilment is a right of the human person. The bodily integrity concerns the inviolability of the physical body and emphasizes the importance of personal autonomy, self-ownership, and self-determination of human beings over their own bodies.

Forensic (scientific, economic and social) interests have to be subordinated to the dignity and bodily integrity of the human person.

Genealogical witnessing

In some jurisdictions, the right against self-incrimination in the court is expanded to the right to refuse to testify as a witness in case one is a first degree relative of the accused. However, this right is usually excepted by severe offences like murder or human trafficking. From this point of view, genealogical witnessing helping investigators to focus on the suspect is not against the law.

Genetic surprises

A familial search may reveal non-paternity and BGA inference may reveal unknown and potentially unwanted information to the suspected perpetrator.

In an illustrative example, probable Adolf Hitler´s Y chromosome haplotype E1b1b is rare in Western Europeans but common among the Berber tribes of Morocco, Algeria, Libya and Tunisia. It is also one of the major founding lineages of the Ashkenazi and Sephardic Jews. In other words, Hitler's paternal lineage may have included Jewish or African ancestors (https://www.discovermagazine.com/health/hitlers-jewish-genes).

Justice

Justice may be retributive, procedural, and distributive.

Retributive justice or corrective justice is concerned with the righting of wrongs through the operations and outcomes of the criminal justice system. The development of DNA profiling and databasing has contributed hugely to crime control and to retributive justice outcomes by providing actionable intelligence to investigators and evidence to the court.

Procedural justice is closely connected to the due-process in guaranteeing that all those subject to legal proceedings are given their full rights. Procedural justice benefited from the introduction of DNA profiling into modern criminal justice systems by a) the scientific basis and standardized laboratory application of DNA technology in contrast to other less well evidence-based forms of forensic technology, b) the use of post-conviction forensic DNA analysis, which has made possible the exoneration of individuals whose convictions were not secured originally (see Innocence Project).

Distributive or allocative justice concerns the extent to which benefits and burdens are shared equally among all members of a community. Equality is expressed in the EU Charter of Fundamental Rights, Article 21 - Non-discrimination as "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited". However, it can also be interpreted as equal access to the benefits of technology uses and equal protection against any potentially damaging effects of their application.

Misinterpretation of evidence

Misinterpretation, following unnecessary false leads, unwarranted privacy breaches and miscarriage of justice can occur when evidence is given inappropriate weight.

Privacy, genetic surveillance, and genetic policing

The right to privacy is reaffirmed by the Human Right Council in Resolution 28/16 Article 12 of the Universal Declaration of Human Rights (UDHR) and Article 17 of the International Covenant on Civil and Political Rights (ICCPR). Privacy can be seen as a social license to carry out a limited number of acts free from communal, public, and governmental scrutiny. It encompasses notions of confidentiality, secrecy, anonymity, data protection, data security, fair information processes, decisional autonomy, and freedom from unwanted intrusion.

Massively parallel sequencing allows new forms of biological surveillance of citizens, residents, visitors, and migrants. When used in actual casework, it could interfere with people's privacy by including innocent people in the investigation as suspects or witnesses. The inclusion of innocent people into criminal investigations is unavoidable. However, if MPS were to increase the scope of people who are included in investigations, then this may have a negative impact not only on privacy but also on family lives. Similarly, if it increased the proportion of members of stigmatised or marginalised groups and minorities among those innocently implicated in investigations, this would have a negative effect on security as a whole.

Quality of evidence

Quality of evidence is influenced by crime scene deposition of sample, evidence handling, genetic analysis, interpretation of findings, evaluation of evidence, and presentation of findings at the court.

Self-identifiability of DNA

Large-scale genetic data sets or whole genomes are frequently shared with other research groups and even released on the Internet to allow for secondary analyses (Wjst 2010). Study participants are usually not informed about such data sharing because data sets are assumed to be anonymous after stripping off personal identifiers. However, the assumption of anonymity of genetic data sets is not fulfilled because genetic data are intrinsically self-identifying. The types of forensic comparison described in chapter "Process of forensic genetics identification" (Section 3) can be performed by any person in the know without police authority.

Social solidarity

Public interest and public goods such as security, safety, and justice can justify the application of MPS because of the ethical principle of social solidarity, obligating persons to assist the criminal justice system in their work. Nevertheless, there is a need to consider potential conflict with individual rights.

6 Summary with recommendations

The human genome consists of autosomal chromosomes 1 to 22, gonosomal chromosomes X and Y, and mitochondrial DNA, mtDNA. Each part of the human genome has a special way of transfer from parental to offspring generations that can be used in genealogical investigations during human identification.

DNA is a dense and stable information medium with significant differences among the genomes of human individuals: 0.1% differs due to single-nucleotide variants, SNPs, and 0.6% due to insertions and deletions, indels. These differences have differentiation power that can be used in forensics.

Forensic Short Tandem Repeat (STR) markers are without any coding function, chosen solely for identification purposes. DNA profiling by STR genotyping was optimised to handle minute biological trace samples and became a standard of performance and quality in the forensic sciences. It can be used for identification of suspects, evidence of presence at crime scenes, exoneration of innocents, testing for kinship, and Disaster Victim Identification.

Microarray chip, an alternative genotyping technology, provides information about Single Nucleotide Polymorphism, SNPs. It requires a comparatively large amount of template DNA but can provide genealogical and phenotypic information.

Massively Parallel Sequencing provides information both for SNPs and STRs while for STRs, information is richer than the one obtained by DNA profiling. It can be applied not only to human genome but also to human transcriptome, epigenome, and microbiome. All mentioned –omes also have individualizing capability and similar guidance as for the human genome should be applied to them. Forensic scientists are technically prepared to use MPS technology.

In forensic DNA comparison, we can compare STR profile or SNP profile on one person to STR profile or SNP profile of another person or persons. Regarding forensic use of whole genome databases, it will always be a comparison of a subset to the whole set of nucleotide sequences. Profiles can be obtained by different genotyping methods, mentioned above. Also, a comparison of two sets of information about DNA sequence does not require the same loci testing to be performed on both samples.

A myth of the infallibility of DNA profiling may lead to overlooking possible errors and result in the marginalization of other types of evidence in court or even miscarriage of justice. Availability of whole genome sequences from the large part of the population does not bring a new danger of judicial error. Thus, the same high quality standards must be striven in the whole process of forensic genetics using whole genome sequencing. This can be achieved by: codes of conduct, certification of practitioners and accreditation of laboratories according to international standards and norms (i.e., ISO17025), the best laboratory practice, and logical, Bayesian way of evidence interpretation. DNA cannot be used as sole evidence and the database trawl match cannot be considered to imply identity. Instead, three rules must be followed: 1) Evidence interpretation is always performed within the circumstances of the case, 2) The expert considers the observations in light of two scenarios: the prosecution hypothesis and the defence hypothesis. These hypotheses must be clearly stated, 3) The expert calculates the likelihood ratio by calculating the probability of evidence assuming that the prosecution hypothesis holds, and the probability of the same evidence assuming that the defence hypothesis holds, and putting these two probabilities into proportion. When hypotheses are changed, the likelihood ratio must be recalculated.

Big Data, to which whole genome sequences belong, are characterised by relational and flexible information in high volume, velocity, and variety. They can determine the risk that a specific individual will commit a crime or terrorist act, produce evidence in the justice system, and may contribute to crime prevention and deterrence. Big Data bears the risk of data siloing and data misuse.

To avoid data siloing, databases should be designed, following FAIR data principles in mind (to make data Findable, Accessible, Interoperable and Reusable). To avoid data misuse (e.g.,, by re-identifying data source for marketing, commercial, or other purposes), organisational and technical measures must be applied.

Direct to Customer (DTC) genetics databases hold dense genotypes of millions of people, and the number is growing quickly, now exceeding at least twice the number that is available in forensic genetics databases. Since 2018, law enforcement agencies succeeded to use online genealogical databases to identify anonymous DNA via long-range familial searches. However, even long-range familial searches can bring no result despite using auxiliary information sources on the Internet. Also, the availability of DTC genomic data for future forensic investigation is questioned because of entrepreneurial, legal, and data hacking concerns.

With the advent of MPS, the efficient generation of data at the nucleotide level beyond that of STR profiles alone has allowed laboratories to produce much more wide-ranging DNA information, including regions coding for Forensic DNA Phenotypes (FDPs). This type of genetic testing can provide valuable information to target perpetrators whose DNA profile is absent in forensic databases.

The most researched externally visible characteristics are hair colour, skin colour, eye colour, and shape of the face. The Area Under Receiver Operating Curve values lie in the range 0.74–0.99 for eye colour, 0.64–0.94 for hair colour, and 0.72–0.99 for skin colour. For face, AUC reaches 0.8. These numbers correspond to the likelihood ratio between 2 and 200, several orders of magnitude lower than for STR profiling. For eyebrow colour, height, and freckles, AUC numbers are even lower.

Since 2019, forensic DNA phenotyping is explicitly permitted or practised for serious crimes in compliance with existing laws in many European countries. There, genetic information about SNPs is de-centrally stored in laboratories performing the analysis, and therefore does not contribute to the establishment of new kinds of criminalistic DNA databases. This requires consideration about data storage, access, and results communication to avoid issues of discrimination. Education is needed while recommendations and training on how to avoid petrification of racial stereotyping are provided by the VISAGE consortium (http://www.visage-h2020.eu/).

Special care must be taken in the analysis of crime scene samples where the members of a minority ethnic community become persons of interest as a result of the application of biogeographical ancestry (BGA) testing. The CSI effect still allocates more investigative weight as well as more social freight accompanying any type of DNA evidence, despite FDP and BGA having entirely different information value than DNA profiling. For some critical observers, promises about the potential utility and informativity of FDP may worsen existing forms of unequal treatment based on ethnicity and perceived "race" and produce genetic "facts" based on culturally conceived and unreflective assumptions.

The advent of MPS in clinical genetics ended the stepwise partial genotyping for many patients and families, expanded the known phenotypes of countless disorders, and led to new disease gene discoveries. Nevertheless, it still fails to identify the molecular cause of many patients who clearly exhibit genetic/syndromic conditions, while at the same time fails to assign a clinical significance to the found sequence variants. Thus, many large projects with the aim to perform whole genome sequencing for representative populations emerged around the world. The European 1+ Million Genomes Initiative with 21 signatory Member States and Norway has the ambition to sequence the same number of persons as all the other sequencing projects together. Even so, the number of individuals is smaller than what is currently available in Direct to Customer genetics companies in the form of chip-derived SNPs.

Health information is shared by a surprisingly large number of individuals and institutions, despite its generally held sensitivity. Nevertheless, genetic data sharing policies are already well established and genetic data are subjected to informed consent and (pseudo)anonymization.

A human research subject authorises the anticipated genetic testing prior to its performance and requires that the permission be given voluntarily and with knowledge of the facts, risks, and benefits to the individual human research subject. Due to crossing boundaries between databases, informed consent should be updated to include articulation of person´s will regarding their data use for investigative purposes.

Even when genetic data are stripped of all identifiable metadata, anonymisation can be reversed by the nature of DNA as the carrier of the data representing the individual.

The more appropriate term pseudonymisation refers to 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.

Despite the attempts to preserve the privacy of genomic data by cryptographic techniques, their deployment is not sufficient due to technical reasons and many privacy breaches of healthcare data so far occurred and can occur for genomics data by identity tracing, attribute disclosure, and completion attack.

The Prüm Decision on cross-border cooperation in combating terrorism and crime set the working system for sharing fingerprints, vehicle registration, and DNA profiles. By now, it allows the exchange of forensic DNA information across the national databases of 24 EU Member States more efficiently while simultaneously more restricted than it was prior to Prüm. This privacy by default and privacy by design system for digital data exchange on secure European TESTA network reduced privacy risks while providing many hits to investigators. Together with experience from DNA profile exchange within the Central Schengen Information System, it may serve well as a template for establishing a functional way of data transfer between European genomic databases and forensic investigators. Nevertheless, it is felt by some researchers that in Prüm, there is a deficit of transparency and accountability that may be cancelled out by application of the Council of Europe´s 12 principles of Good Governance, allowing to evaluate legal oversight regimes (https://www.coe.int/en/web/good-governance/12-principles) (Kennett 2019).

In the regulation of forensic genomics, two public goods must be balanced: societal good regarding public safety and individual good regarding privacy (https://www.wordclouds.com/, Figure 15). Any recommendations regarding forensic use of genomics databases should maximize the potential of whole genome sequencing for mankind and minimize the negative impact for individual rights. The aim of any interventions should be to render technologies socially acceptable rather than shutting down viable lines of development because of assertions of their potential ethical hazards.

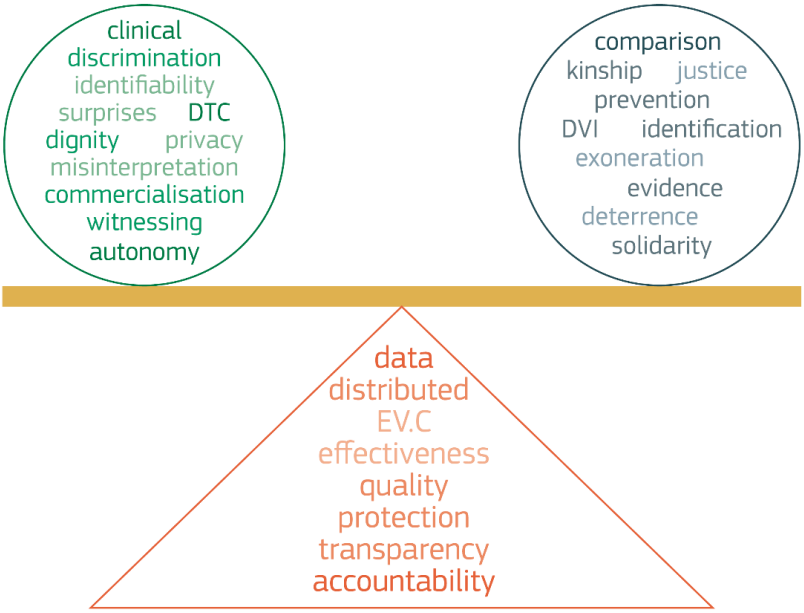

Figure 15: Forensic genomics balance word cloud

Clinical, research, DTC, and forensic genomics seem to converge into translational genomics. It would be advantageous to have integrating rules for genomic research, clinical care, public health screening, DTC testing, and forensic profiling with overarching fit-for-purpose recommendations over differing legal frameworks across domains. Simplification, reduction, and harmonisation of law across domains will aid understanding and compliance. In case of conflict between rules, the rules that are more protective of the individual´s rights and interests should apply.

To reduce the risk of misuse of genomic information (by privacy breaches), risk management means can be applied. The risk is calculated as the probability of the adverse event multiplied by consequence of this adverse event. Such calculated risk may be compared with other Big Data risks. However, human irrationality must come into computation as well. The typical person is willing to accept a much higher level of risk if they choose the risk themselves than if the risk is imposed on them by someone else (compare smoking death risk *vs* nuclear accident death risk 50,000:1) and each person uses his value function, which corrects the objective measures of risk.

Risk management can be performed by reducing the probability occurrence of an adverse event (prevention) or by reducing the severity of the consequences of an adverse event (protection). The risk must be reduced to a level where risk mitigation costs become disproportionate to the relevant risk mitigation (principle ALARA, As Low As Reasonably Achievable). When translated into money (insurance) terms, the costs of corrective actions to reduce the risks of adverse events should not outweigh the benefits of reducing their risk. On the societal level, there should be a proportional distribution of risk *vs* compensation and benefits.

References

(FPF), T. F. O. P. F. Privacy Best Practices for Consumer Genetic Testing Services. T.F.O.P.F. (FPF), 2018.

ABECASIS, G. R., A. AUTON, L. D. BROOKS, M. A. DEPRISTO, et al. An integrated map of genetic variation from 1,092 human genomes. Nature, Nov 1 2012, 491(7422), 56-65.

ADAMOWICZ, M., J. CLARKE, T. RAMBO, H. MAKAM, et al. Validation of MaSTR (TM) software: Extensive study of fully-continuous probabilistic mixture analysis using PowerPlex (R) Fusion 2 - 5 contributor mixtures. Forensic Science International Genetics Supplement Series, Dec 2019, 7(1), 641-643.

ADEGOKE, A., G. AL-NAJJAR, T. FALCHETTA, E. HICKOK, et al. Establishing best practice for forensic DNA databases. F.G.P. INITIATIVE, 2017.

AL AZIZ, M. M., M. N. SADAT, D. ALHADIDI, S. WANG, et al. Privacy-preserving techniques of genomic data-a survey. Briefings in Bioinformatics, May 2019, 20(3), 887-895.

ALONSO, A., P. MULLER, L. ROEWER, S. WILLUWEIT, et al. European survey on forensic applications of massively parallel sequencing. Forensic Science International-Genetics, Jul 2017, 29, E23-E25.

AMORIM, A., T. FERNANDES AND N. TAVEIRA Mitochondrial DNA in human identification: a review. Peerj, Aug 2019, 7, e7314.

ANGERS, A., D. M. KAGKLI, L. OLIVA, M. PETRILLO, et al. Study on DNA Profiling Technology for its Implementation in the Central Schengen Information System. Luxembourg: P.O.O.T.E. UNION, 2019. ISBN 978-92-76-07983-5.

ANVAR, S. Y., K. J. VAN DER GAAG, J. W. VAN DER HEIJDEN, M. H. VELTROP, et al. TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes. Bioinformatics (Oxford, England), 2014, 30(12), 1651-1659.

BAUER, D. W., N. BUTT, J. M. HORNYAK AND M. W. PERLIN Validating TrueAllele Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors. Journal of Forensic Sciences, Mar 2020, 65(2), 380-398.

BAUER, M. RNA in forensic science. Forensic Science International-Genetics, Mar 2007, 1(1), 69-74.

BENECKE, M. Coding or non-coding, that is the question: having solved the last technical hurdles to extract DNA information from virtually any biological material, forensic biologists now have to ponder the ethical and social questions of using information from exonic DNA. EMBO Rep, Jun 2002, 3(6), 498-502.

BENNETT, S. T. AND J. A. TODD Human type 1 diabetes and the insulin gene: principles of mapping polygenes. Annu Rev Genet, 1996, 30, 343-370.

BENSCHOP, C. C. G., A. NIJVELD, F. E. DUIJS AND T. SIJEN An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors. Forensic Science International-Genetics, Sep 2019, 42, 31-38.

BERGER, B. AND H. CHO Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol, Jul 2 2019, 20(1), 128.

BODNER, M., I. BASTISCH, J. M. BUTLER, R. FIMMERS, et al. 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 Science International: Genetics, 2016, 24, 97-102.

BOGDANOV, D., L. KAMM, S. LAUR AND V. SOKK Implementation and Evaluation of an Algorithm for Cryptographically Private Principal Component Analysis on Genomic Data. IEEE/ACM Trans Comput Biol Bioinform, Sep-Oct 2018, 15(5), 1427-1432.

BONOMI, L., Y. X. HUANG AND L. OHNO-MACHADO Privacy challenges and research opportunities for genomic data sharing. Nature Genetics, 2020, 52, 646–654.

BORNHOLT, J., R. LOPEZ, D. M. CARMEAN, L. CEZE, et al. Toward a DNA-based archival storage system. Ieee Micro, May-Jun 2017, 37(3), 98-104.

BRADBURY, C., A. KOTTGEN AND F. STAUBACH Off-target phenotypes in forensic DNA phenotyping and biogeographic ancestry inference: A resource. Forensic Science International-Genetics, Jan 2019, 38, 93-104.

BRADLEY, S. Realistic DNA De-anonymization using Phenotypic Prediction. ArXiv, 07/25 2016, 1607.07501.

BRUIJNS, B., R. TIGGELAAR AND H. GARDENIERS Massively parallel sequencing techniques for forensics: A review. Electrophoresis: an international journal, 2018, 39(21), 2642-2654.

BUDOWLE, B., S. E. SCHMEDES AND F. R. WENDT Increasing the reach of forensic genetics with massively parallel sequencing. Forensic Science Medicine and Pathology, Sep 2017, 13(3), 342-349.

CALISKAN, A., F. YAMAGUCHI, E. DAUBER, R. E. HARANG, et al. When coding style survives compilation: de-anonymizing programmers from executable binaries. ArXiv, 2018, abs/1512.08546, 1-15.

CAO, M. D., E. TASKER, K. WILLADSEN, M. IMELFORT, et al. Inferring short tandem repeat variation from paired-end short reads. Nucleic Acids Research, 2014, 42(3).

CAVALLI-SFORZA, L. L. The Human Genome Diversity Project: past, present and future. Nature Reviews Genetics, 2005, 6(4), 333-340.

CECH, M. Genetic Privacy in the "Big Biology" Era: The "Autonomous" Human Subject. Hastings Law Journal, Apr 2019, 70(3), 851-886.

CLAES, P., H. HILL AND M. D. SHRIVER Toward DNA-based facial composites: preliminary results and validation. Forensic Sci Int Genet, Nov 2014, 13, 208-216.

COBLE, M. D. AND J. A. BRIGHT Probabilistic genotyping software: An overview. Forensic Science International-Genetics, Jan 2019, 38, 219-224.

COBLE, M. D., J. BUCKLETON, J. M. BUTLER, T. EGELAND, et al. DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications. Forensic Science International: Genetics, 2016, 25, 191-197.

COURTS, C. AND B. MADEA Micro-RNA - A potential for forensic science? Forensic Science International, 2010, 203(1-3), 106-111.

DACA-ROSZAK, P. AND E. ZIETKIEWICZ Transcriptome variation in human populations and its potential application in forensics. Journal of Applied Genetics, Nov 2019, 60(3-4), 319-328.

DE LA PUENTE, M., C. SANTOS, M. FONDEVILA, L. MANZO, et al. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. Forensic Science International: Genetics, 2016, 22, 81-88.

DEMAEREL, W., Y. MOSTOVOY, F. YILMAZ, L. VERVOORT, et al. The 22q11 low copy repeats are characterized by unprecedented size and structural variability. Genome Research, Sep 2019, 29(9), 1389-1401.

DODDS, K. G., J. C. MCEWAN, R. BRAUNING, T. C. VAN STIJN, et al. Exclusion and Genomic Relatedness Methods for Assignment of Parentage Using Genotyping-by-Sequencing Data. G3 (Bethesda), Oct 7 2019, 9(10), 3239-3247.

DOU, J., B. SUN, X. SIM, J. D. HUGHES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet, Sep 2017, 13(9), e1007021.

EDGE, M. D., B. F. B. ALGEE-HEWITT, T. J. PEMBERTON, J. Z. LI, et al. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. Proc Natl Acad Sci U S A, May 30 2017, 114(22), 5671-5676.

ELLENBOGEN, P. AND A. R. V. NARAYANAN Identification of Anonymous DNA Using Genealogical Triangulation. bioRxiv, 2019.

EVETT, I. W. Avoiding the Transposed Conditional. Science & Justice, 1995, 35(2), 127-131.

FRIIS, S. L., A. BUCHARD, E. ROCKENBAUER, C. BORSTING, et al. Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs. Forensic Science International-Genetics, Mar 2016, 21, 68-75.

GANNETT, L. Biogeographical ancestry and race. Stud Hist Philos Biol Biomed Sci, Sep 2014, 47 Pt A, 173-184.

GETTINGS, K. B., D. BALLARD, M. BODNER, L. A. BORSUK, et al. Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Science International-Genetics, Nov 2019, 43.

GETTINGS, K. B., L. A. BORSUK, D. BALLARD, M. BODNER, et al. STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. Forensic Science International: Genetics, 2017, 31, 111-117.

GILL, P. DNA evidence and miscarriages of justice. Forensic Science International, Jan 2019, 294, E1-E3.

GJERTSON, D. W., C. H. BRENNER, M. P. BAUR, A. CARRACEDO, et al. ISFG: Recommendations on biostatistics in paternity testing. Forensic Science International: Genetics, 2007, 1(3-4), 223-231.

GOLDFEDER, R. L., D. P. WALL, M. J. KHOURY, J. P. A. IOANNIDIS, et al. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. American Journal of Epidemiology, Oct 2017, 186(8), 1000-1009.

GOLDSTEIN, D. B. AND L. CHIKHI Human migrations and population structure: What we know and why it matters. Annual Review of Genomics and Human Genetics, 2002, 3, 129-152.

GRESHAKE, B., P. E. BAYER, H. RAUSCH AND J. REDA openSNP - a crowdsourced web resource for personal genomics. PLoS ONE, 2014, 9(3), e89204.

GREYTAK, E. M., D. H. KAYE, B. BUDOWLE, C. MOORE, et al. Privacy and genetic genealogy data. Science, Aug 2018, 361(6405), 857-857.

GREYTAK, E. M., C. MOORE AND S. L. ARMENTROUT Genetic genealogy for cold case and active investigations. Forensic Science International, Jun 2019, 299, 103-113.

GRODY, W. W. The transformation of medical genetics by clinical genomics: hubris meets humility. Genetics in Medicine, Sep 2019, 21(9), 1916-1926.

GROSS, M. Digging deeper into human evolution. Current Biology, May 2020, 30(9), R371-R374.

GRŠKOVIĆ, B., D. ZRNEC, S. VICKOVIĆ, M. POPOVIĆ, et al. DNA methylation: the future of crime scene investigation? Mol Biol Rep, Jul 2013, 40(7), 4349-4360.

GYMREK, M., D. GOLAN, S. ROSSET AND Y. ERLICH lobSTR: A short tandem repeat profiler for personal genomes. Genome Research, 2012, 22(6), 1154-1162.

GYMREK, M., A. L. MCGUIRE, D. GOLAN, E. HALPERIN, et al. Identifying Personal Genomes by Surname Inference. Science, Jan 2013, 339(6117), 321-324.

HALDER, I., M. SHRIVER, M. THOMAS, J. R. FERNANDEZ, et al. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat, May 2008, 29(5), 648-658.

HANGHOJ, K., I. MOLTKE, P. A. ANDERSEN, A. MANICA, et al. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. Gigascience, May 1 2019, 8(5), giz034.

HAO, W. Q., J. LIU, L. JIANG, J. P. HAN, et al. Exploring the ancestry differentiation and inference capacity of the 28-plex AISNPs. International Journal of Legal Medicine, Jul 2019, 133(4), 975-982.

HARA, K., K. OHE, T. KADOWAKI, N. KATO, et al. Establishment of a method of anonymization of DNA samples in genetic research. Journal of Human Genetics, 2003, 48(6), 327-330.

HAZEL, J. W., E. W. CLAYTON, B. A. MALIN AND C. SLOBOGIN Is it time for a universal genetic forensic database? Science, 2018, 362(6417), 898-900.

HENN, B. M., L. HON, J. M. MACPHERSON, N. ERIKSSON, et al. Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLoS ONE, Apr 2012, 7(4), e34267.

HIGHNAM, G., C. FRANCK, A. MARTIN, C. STEPHENS, et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Research, Jan 2013, 41(1), e32.

HOLUB, P., F. KOHLMAYER, F. PRASSER, M. T. MAYRHOFER, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreserv Biobank, Apr 2018, 16(2), 97-105.

HOOGENBOOM, J., K. J. VAN DER GAAG, R. H. DE LEEUW, T. SIJEN, et al. FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. Forensic Science International-Genetics, 2017, 27, 27-40.

HOOGENBOOM, J., K. J. VAN DER GAAG AND T. SIJEN STRNaming: Standardised STR sequence allele naming to simplify MPS data analysis and interpretation. Forensic Science International Genetics Supplement Series, Dec 2019, 7(1), 436-437.

HOU, G. W., X. H. JIANG, Y. Y. YANG, F. JIA, et al. A 21-Locus Autosomal SNP Multiplex and its Application in Forensic Science. Journal of Forensic Sciences, Jan 2014, 59(1), 5-14.

HUFF, C. D., D. J. WITHERSPOON, T. S. SIMONSON, J. XING, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Research, 2011, 21(5), 768-774.

HUMBERT, M., K. HUGUENIN, J. HUGONOT, E. AYDAY, et al. De-anonymizing Genomic Databases Using Phenotypic Traits. In *15th Privacy Enhancing Technologies Symposium (PETS)*. Philadelphia, PA, United States, 2015, vol. 2015, p. 99-114.

CHEN, F., C. WANG, W. DAI, X. JIANG, et al. PRESAGE: PRivacy-preserving gEnetic testing via SoftwAre Guard Extension. BMC Med Genomics, Jul 26 2017a, 10(Suppl 2), 48.

CHEN, F., S. WANG, X. JIANG, S. DING, et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. Bioinformatics, Mar 15 2017b, 33(6), 871-878.

CHO, S., H. J. YU, J. HAN, Y. KIM, et al. Forensic application of SNP-based resequencing array for individual identification. Forensic Science International-Genetics, Nov 2014, 13, 45-52.

JIN, Y., A. A. SCHAFFER, M. FEOLO, J. B. HOLMES, et al. GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. G3 (Bethesda), Aug 8 2019, 9(8), 2447-2461.

KAEUFFER, R., D. RÉALE, D. W. COLTMAN AND D. PONTIER Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity (Edinb), Oct 2007, 99(4), 374-380.

KELLEHER, J., Y. WONG, A. W. WOHNS, C. FADIL, et al. Inferring whole-genome histories in large population datasets. Nature Genetics, Sep 2019, 51(9), 1330-1338.

KENNETT, D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. Forensic Science International, 2019, 301, 107-117.

KIDD, K. K., U. SOUNDARARAJAN, H. RAJEEVAN, A. J. PAKSTIS, et al. The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb. Forensic Science International-Genetics, 2018, 33, 33-37.

KIM, J., M. D. EDGE, B. F. B. ALGEE-HEWITT, J. Z. LI, et al. Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci. Cell, Oct 18 2018, 175(3), 848-858.e846.

KIM, K., H. BAIK, C. S. JANG, J. K. ROH, et al. Genomic GPS: using genetic distance from individuals to public data for genomic analysis without disclosing personal genomes. Genome Biology, Aug 2019, 20(1), 175.

KIM, M., Y. SONG AND J. H. CHEON Secure searching of biomarkers through hybrid homomorphic encryption scheme. BMC Med Genomics, Jul 26 2017, 10(Suppl 2), 42.

KLING, D. On the use of dense sets of SNP markers and their potential in relationship inference. Forensic Sci Int Genet, Mar 2019, 39, 19-31.

KLING, D. AND A. TILLMAR Forensic genealogy-A comparison of methods to infer distant relationships based on dense SNP data. Forensic Science International-Genetics, 2019, 42, 113-124.

KOJIMA, K., Y. KAWAI, N. NARIAI, T. MIMORI, et al. Short tandem repeat number estimation from paired-end reads for multiple individuals by considering coalescent tree. BMC Genomics, Aug 2016, 17.

KOOP, B. E., F. MAYER, T. GÜNDÜZ, J. BLUM, et al. Postmortem age estimation via DNA methylation analysis in buccal swabs from corpses in different stages of decomposition-a "proof of principle" study. Int J Legal Med, Jul 7 2020.

KORNELIUSSEN, T. S. AND I. MOLTKE NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics, Dec 15 2015, 31(24), 4009-4011.

KUKLA-BARTOSZEK, M., E. POSPIECH, A. WOZNIAK, M. BORON, et al. DNA-based predictive models for the presence of freckles. Forensic Science International-Genetics, Sep 2019, 42, 252-259.

KUSHIDA, C. A., D. A. NICHOLS, R. JADRNICEK, R. MILLER, et al. Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. Medical Care, Jul 2012, 50(7), S82-S101.

LEE, H. Y., J. H. AN, S. E. JUNG, Y. N. OH, et al. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. Forensic Sci Int Genet, Jul 2015, 17, 17-24.

LEE, J. C. I., B. TSENG, L. K. CHANG AND A. LINACRE SEQ Mapper: A DNA sequence searching tool for massively parallel sequencing data. Forensic Science International-Genetics, Jan 2017, 26, 66-69.

LIEM, M., K. SUONPAA, M. LEHTI, J. KIVIVUORI, et al. Homicide clearance in Western Europe. European Journal of Criminology, Jan 2019, 16(1), 81-101.

LINACRE, A. AND J. E. L. TEMPLETON Forensic DNA profiling: state of the art. Research and Reports in Forensic Medical Science, 2014, 4, 25-36.

LIPPERT, C., R. SABATINI, M. C. MAHER, E. Y. KANG, et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proc.Natl.Acad.Sci U.S.A, 2017, 114(38), 10166-10171.

LIU, F., K. ZHONG, X. JING, A. G. UITTERLINDEN, et al. Update on the predictability of tall stature from DNA markers in Europeans. Forensic Sci Int Genet, Sep 2019, 42, 8-13.

LIU, Y. Y. AND S. HARBISON A review of bioinformatic methods for forensic DNA analyses. Forensic Science International-Genetics, Mar 2018, 33, 117-128.

LONG, G. S., M. HUSSEN, J. DENCH AND S. ARIS-BROSOU Identifying genetic determinants of complex phenotypes from whole genome sequence data. BMC Genomics, Jun 2019, 20, 470.

MACHADO, H. AND R. GRANJA. DNA Databases and Big Data. In *Forensic Genetics in the Governance of Crime.* Singapore: Palgrave Pivot, 2020.

MALIN, B. A. Protecting genomic sequence anonymity with generalization lattices. Methods Inf Med, 2005, 44(5), 687-692.

MANICHAIKUL, A., J. C. MYCHALECKYJ, S. S. RICH, K. DALY, et al. Robust relationship inference in genome-wide association studies. Bioinformatics, Nov 2010, 26(22), 2867-2873.

MATOS, S. Privacy and data protection in the surveillance society: The case of the Prüm system. Journal of Forensic and Legal Medicine, Aug 2019, 66, 155-161.

MELE, M., P. G. FERREIRA, F. REVERTER, D. S. DELUCA, et al. The human transcriptome across tissues and individuals. Science, May 2015, 348(6235), 660-665.

MONTESANTO, A., P. D'AQUILA, V. LAGANI, E. PAPARAZZO, et al. A New Robust Epigenetic Model for Forensic Age Prediction. Journal of Forensic Sciences, Sep 2020, 65(5), 1424-1431.

NARAYANAN, A. AND V. SHMATIKOV *Robust De-anonymization of Large Sparse Datasets*. Edtion ed., 2008. 111-125 p. ISBN 978-0-7695-3168-7.

NELKIN, D. AND S. LINDEE *The DNA mystique: The gene as a cultural icon*. Edtion ed. New York: WH Freeman, 1995. 276 p. ISBN ISBN 0-7167-2907-9.

NELLAKER, C., F. S. ALKURAYA, G. BAYNAM, R. A. BERNIER, et al. Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative. Frontiers in Genetics, Jul 2019, 10, 611.

NORDGAARD, A. AND B. RASMUSSON The likelihood ratio as value of evidence-more than a question of numbers. Law Probability & Risk, Dec 2012, 11(4), 303-315.

OHM, P. Broken promises of privacy: responding to the surprising failure of anonymization. Ucla Law Review, Aug 2010, 57(6), 1701-1777.

PALENCIA-MADRID, L., C. XAVIER, M. DE LA PUENTE, C. HOHOFF, et al. Evaluation of the VISAGE Basic Tool for Appearance and Ancestry Prediction Using PowerSeq Chemistry on the MiSeq FGx System. Genes, Jun 2020, 11(6).

PANG, J. B., M. RAO, Q. F. CHEN, A. Q. JI, et al. A 124-plex Microhaplotype Panel Based on Next-generation Sequencing Developed for Forensic Applications. Sci Rep, Feb 6 2020, 10(1), 1945.

PARSON, W. Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi) Genomics: A Mini-Review. Gerontology, 2018, 64(4), 326-332.

PARSON, W., D. BALLARD, B. BUDOWLE, J. M. BUTLER, et al. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Science International: Genetics, 2016, 22, 54-63.

PARSON, W., L. GUSMAO, D. R. HARES, J. A. IRWIN, et al. DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing. Forensic Science International-Genetics, 2014, 13, 134-142.

PENG, F. D., G. ZHU, P. G. HYSI, R. J. ELLER, et al. Genome-Wide Association Studies Identify Multiple Genetic Loci Influencing Eyebrow Color Variation in Europeans. Journal of Investigative Dermatology, Jul 2019, 139(7), 1601-1605.

PEREIRA, L., F. ALSHAMALI, R. ANDREASSEN, R. BALLARD, et al. PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. International Journal of Legal Medicine, 2011, 125(5), 629-636.

PERTEA, M. The Human Transcriptome: An Unfinished Story. Genes, Sep 2012, 3(3), 344-360.

PHILLIPS, C., L. PRIETO, M. FONDEVILA, A. SALAS, et al. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation. PLoS ONE, 2009, 4(8).

PHILLIPS, C., A. SALAS, J. J. SÁNCHEZ, M. FONDEVILA, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet, Dec 2007, 1(3-4), 273-280.

PICIN, A., M. HAJDINJAK, W. NOWACZEWSKA, S. BENAZZI, et al. New perspectives on Neanderthal dispersal and turnover from Stajnia Cave (Poland). Scientific Reports, 2020, 10(1), 14788.

PINEDA, G. M., A. H. MONTGOMERY, R. THOMPSON, B. INDEST, et al. Development and validation of InnoQuant™, a sensitive human DNA quantitation and degradation assessment method for forensic samples using high copy number mobile elements Alu and SVA. Forensic Sci Int Genet, Nov 2014, 13, 224-235.

PINTO, N., L. GUSMAO AND A. AMORIM X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. Forensic Science International: Genetics, 2011, 5(1), 27-32.

PLATT, R. N., 2ND, M. W. VANDEWEGE AND D. A. RAY Mammalian transposable elements and their impacts on genome evolution. Chromosome Res, Mar 2018, 26(1-2), 25-43.

PRIETO, L., B. ZIMMERMANN, A. GOIOS, A. RODRIGUEZ-MONGE, et al. The GHEP-EMPOP collaboration on mtDNA population data-A new resource for forensic casework. Forensic Science International-Genetics, Mar 2011, 5(2), 146-151.

PRINZ, M., A. CARRACEDO, W. R. MAYR, N. MORLING, et al. DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). Forensic Sci Int Genet, Mar 2007, 1(1), 3-12.

QUEIROS, F. The visibilities and invisibilities of race entangled with forensic DNA phenotyping technology. Journal of Forensic and Legal Medicine, Nov 2019, 68.

RAISARO, J. L., C. GWANGBAE, S. PRADERVAND, R. COLSENET, et al. Protecting Privacy and Security of Genomic Data in i2b2 with Homomorphic Encryption and Differential Privacy. IEEE/ACM Trans Comput Biol Bioinform, Sep-Oct 2018, 15(5), 1413-1426.

RAJEEVAN, H., U. SOUNDARARAJAN, J. R. KIDD, A. J. PAKSTIS, et al. ALFRED: an allele frequency resource for research and teaching. Nucleic Acids Research, Jan 2012, 40(D1), D1010-D1015.

RAY, D. A., J. A. WALKER AND M. A. BATZER Mobile element-based forensic genomics. Mutat Res, Mar 1 2007, 616(1-2), 24-33.

ROBERTSON, B., G. A. VIGNAUX AND C. E. H. BERGER *Interpreting evidence: Evaluating Forensic Science in the Courtroom, 2nd edition*. Edtion ed. Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, Ltd., 2016. 1-216 p. ISBN ISBN: 978-1-118-49243-7.

ROEWER, L., M. M. ANDERSEN, J. BALLANTYNE, J. M. BUTLER, et al. 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, Jun 4 2020, 48, 102308.

ROGERS, E. M. *Diffusion of innovations*. Edtion ed.: Simon and Schuster, 2010. 518 p. ISBN 1451602472.

SAMUEL, G., H. C. HOWARD, M. CORNEL, C. VAN EL, et al. A response to the forensic genetics policy initiative's report "Establishing Best Practice for Forensic DNA Databases". Forensic Science International-Genetics, Sep 2018, 36, E19-E21.

SANTOS, F. AND H. MACHADO Patterns of exchange of forensic DNA data in the European Union through the Prüm system. Science & Justice, Jul 2017, 57(4), 307-313.

SERO, D., A. ZAIDI, J. LI, J. D. WHITE, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun, Jun 11 2019, 10(1), 2557.

SHENDURE, J., S. BALASUBRAMANIAN, G. M. CHURCH, W. GILBERT, et al. DNA sequencing at 40: past, present and future. Nature, Oct 2017, 550(7676), 345-353.

SHRIVER, M. D., E. J. PARRA, S. DIOS, C. BONILLA, et al. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet, Apr 2003, 112(4), 387-399.

SCHNEIDER, P. M., B. PRAINSACK AND M. KAYSER The Use of Forensic DNA Phenotyping in Predicting Appearance and Biogeographic Ancestry. Deutsches Arzteblatt International, Dec 2019, 116(51-52), 873-880.

SIMMONS, S., B. BERGER AND C. SAHINALP Protecting Genomic Data Privacy with Probabilistic Modeling. Pac Symp Biocomput, 2019, 24, 403-414.

SIVA, N. 1000 Genomes project. Nat Biotechnol, Mar 2008, 26(3), 256.

SMITH, E. L. AND A. COOPER. Homicide in the U.S. Known to Law Enforcement, 2011. In O.O.J.P. U.S. DEPARTMENT OF JUSTICE, BUREAU OF JUSTICE STATISTICS. U.S., 2013, p. 1-18.

SOLETO MUNOZ, H. AND A. FIODOROVA DNA and Law Enforcement in the European Union: Tools and Human Rights Protection. Utrecht Law Review, 01/31 2014, 10, 149-162.

SPEIDEL, L., M. FOREST, S. N. SHI AND S. R. MYERS A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics, Sep 2019, 51(9), 1321–1329.

TILLMAR, A., I. GRANDELL AND K. MONTELIUS DNA identification of compromised samples with massive parallel sequencing. Forensic Sciences Research, 2019, 4(4), 331-336.

TILLMAR, A., P. SJÖLUND, B. LUNDQVIST, T. KLIPPMARK, et al. Whole-genome sequencing of human remains to enable genealogy DNA database searches - A case report. Forensic Sci Int Genet, May 2020, 46, 102233.

TILLMAR, A. O., D. KLING, J. M. BUTLER, W. PARSON, et al. DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. Forensic Science International: Genetics, 2017, 29, 269-275.

TOOM, V. *Cross-border exchange and comparison of forensic DNA data in the context of the Prüm Decision*. Edition ed.: Directorate-General for Internal Policies of the Union (European Parliament), 2018. ISBN 978-92-846-3096-7.

TOOM, V., R. GRANJA AND A. LUDWIG The Prüm Decisions as an Aspirational regime: Reviewing a Decade of Cross-Border Exchange and Comparison of Forensic DNA Data. Forensic Science International-Genetics, Jul 2019, 41, 50-57.

TOOM, V., M. WIENROTH, A. M'CHAREK, B. PRAINSACK, et al. 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, May 2016, 22, E1-E4.

TOZZO, P., G. D'ANGIOLELLA, P. BRUN, I. CASTAGLIUOLO, et al. Skin Microbiome Analysis for Forensic Human Identification: What Do We Know So Far? Microorganisms, Jun 9 2020, 8(6), 873.

TVEDEBRINK, T., P. S. ERIKSEN, H. S. MOGENSEN AND N. MORLING GenoGeographer - A tool for genogeographic inference. Forensic Science International Genetics Supplement Series, 2017, 6, E463-E465.

VAN, N. C., M. VANDEWOESTYNE, C. W. VAN, D. DEFORCE, et al. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. Forensic Science International: Genetics, 2014, 9, 1-8.

VENNEMANN, M. AND A. HUTH *RNA Profiling A New Tool in Forensic Science*. edited by X. MALLETT, T. BLYTHE AND R. BERRY. Edtion ed., 2014. 289-304 p. ISBN 978-1-4398-2516-7; 978-1-4398-2514-3.

WANG, S., F. SONG, M. XIE, K. ZHANG, et al. Evaluation of a six-dye multiplex composed of 27 markers for forensic analysis and databasing. Mol Genet Genomic Med, Jul 16 2020, e1419.

WANG, Z., R. ZHU, S. ZHANG, Y. BIAN, et al. Differentiating between monozygotic twins through next-generation mitochondrial genome sequencing. Anal Biochem, Dec 1 2015, 490, 1-6.

WICKENHEISER, R. A. A crosswalk from medical bioethics to Forensic Bioethics. Forensic Science International: Synergy, 2019/01/01/ 2019a, 1, 35-44.

WICKENHEISER, R. A. Forensic genealogy, bioethics and the Golden State Killer case. Forensic Science International: Synergy, 2019/01/01/ 2019b, 1, 114-125.

WILLIAMS, R. AND M. WIENROTH Social and ethical aspects of forensic genetics: A critical review. Forensic Sci Rev, Jul 2017, 29(2), 145-169.

WILSON-WILDE, L. The international development of forensic science standards - A review. Forensic Science International, Jul 2018, 288, 1-9.

WJST, M. Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. BMC Medical Ethics, Dec 2010, 11.

WOERNER, A. E., J. L. KING AND B. BUDOWLE Fast STR allele identification with STRait Razor 3.0. Forensic Science International: Genetics, 2017, 30, 18-23.

WOERNER, A. E., N. M. M. NOVROSKI, F. R. WENDT, A. AMBERS, et al. Forensic human identification with targeted microbiome markers using nearest neighbour classification. Forensic Sci Int Genet, Jan 2019, 38, 130-139.

WOLF, S. M., P. N. OSSORIO, S. A. BERRY, H. T. GREELY, et al. Integrating Rules for Genomic Research, Clinical Care, Public Health Screening and DTC Testing: Creating Translational Law for Translational Genomics. Journal of Law Medicine & Ethics, Mar 2020, 48(1), 69-86.

WU, X., Y. T. ZHANG, A. M. WANG, M. Y. SHI, et al. MNSSp3: Medical big data privacy protection platform based on Internet of things. Neural Computing & Applications, 2020.

XIONG, Z., G. DANKOVA, L. J. HOWE, M. K. LEE, et al. Novel genetic loci affecting facial shape variation in humans. Elife, Nov 26 2019, 8.

YAKUBU, A. M. AND Y. P. P. CHEN Ensuring privacy and security of genomic data and functionalities. Briefings in Bioinformatics, Mar 2020, 21(2), 511-526.

YANG, J. W., D. H. LIN, C. W. DENG, Z. LI, et al. The advances in DNA mixture interpretation. Forensic Science International, Aug 2019, 301, 101-106.

YUAN, L., X. CHEN, Z. LIU, Q. LIU, et al. Identification of the perpetrator among identical twins using next-generation sequencing technology: A case report. Forensic Sci Int Genet, Jan 2020, 44, 102167.

Glossary of terms

Adenine	A purine base; one of the building blocks of DNA; abbreviated A.
Allele	One of two or more alternative forms of a gene. In DNA profiling, the definition is extended to any DNA region used for analysis.
Amplification	Increasing the number of copies of a DNA region, usually by PCR.
Autosome	Any chromosome other than the X or Y.
Base	Purine (A or G) or pyrimidine (T or C) part of nucleotide.
Base pair	Two complementary nucleotides in double-stranded DNA; these are AT or GC.
Blind proficiency test	A proficiency test in which the laboratory personnel do not know that a test is being conducted.
Bayes theorem	Alternatively, Bayes law or Bayes rule, describes the probability of an event, based on prior knowledge of conditions that are related to the event. Let P be probability, H is hypothesis/scenario, E is evidence, and "|" means on the condition that. Then, in mathematical notation, Bayes theorem is P(H|E) = P(H)* P(B|H)/P(E). In forensic context with prosecution and defense hypotheses, it is used in the form P(H1|E)/P(H2|E) = (P(H1)* P(E|H1)/P(E))/(P(H2)* P(E|H2)/P(E)) = P(H1)* P(E|H1))/(P(H2)* P(E|H2).
Chromatin	A substance within a chromosome consisting of DNA and protein.
Chromosome	A physical structure in the cell nucleus, made of DNA, RNA, and proteins. The genes are arranged in linear order along the chromosome.
Crossing over	The exchange of parts between homologous chromosomes during meiosis; recombination.
CSI effect	Unrealistic expectation, forensic imaginary about capabilities of forensic science and forensic scientists built by watching CSI: Crime Scene Investigation series that depict forensic science in an exaggerated way.
Cytosine	A pyrimidine base; one of the building blocks of DNA; abbreviated C.
Degradation	The breaking down of DNA by chemical or physical means.
Denaturation	A process in which proteins or nucleic acids lose the quaternary structure, tertiary structure, and secondary structure which is present in their native state; in case of DNA, it is a reversible separation of a double-stranded DNA into single strands.
Deoxyribonucleic acid (DNA)	The genetic material; in native conformation a double helix composed of two complementary chains of paired nucleotides.
Diploid	Organism or cell having two sets of chromosomes (in contrast to haploid).
DNA polymerase	The enzyme that catalyses the synthesis of double-stranded DNA based on template.
Electrophoresis	A technique in which different charged molecules are separated by their rate of movement in an electric field. DNA bears negative charge and thus, travels to anode.
Enzyme	A biocatalytical protein that is capable of speeding up, and therefore facilitating, a specific chemical reaction.
Euchromatin	A lightly packed form of chromatin that comprises the most active (transcribed) portion of the genome within the cell nucleus. 92% of the human genome is euchromatic.
F statistics	Wright's measures of inbreeding and population structure.
Gamete	A haploid reproductive cell; sperm or egg.
Gametic equilibrium	See Linkage equilibrium.

Gel	A semisolid medium used to separate charged molecules (including DNA) by electrophoresis.
Gene	The basic unit of heredity; a functional sequence of DNA in a chromosome, coding for protein.
Gene frequency	The relative frequency (proportion) of an allele in a population.
Genetic drift	Random fluctuation in allele frequencies.
Genome	The total (haploid) genetic makeup of an organism. In the human being this comprises 3*109 base pairs.
Genotype	The genetic makeup of an organism, as distinguished from its physical appearance (phenotype). The word may be used to designate any number of loci, from one to the total number.
Guanine	A purine base; one of the building blocks of DNA; abbreviated G.
Haploid	Organism or cell having one set of chromosomes, as a gamete (in contrast to diploid).
Haplotype	Haploid genotype, a group of alleles in an organism that are inherited together from a single parent because they are close to each other on a chromosome.
Hardy-Weinberg proportions	Population genetics model stating that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences (such as genetic drift, mate choice, assortative mating, natural selection, sexual selection, mutation, gene flow, meiotic drive, genetic hitchhiking, population bottleneck, founder effect, and inbreeding). Model is used in calculation of denominator of likelihood ratio.
Heterochromatin	A tightly packed form of DNA or condensed DNA in chromosome.
Heteroplasmy	The presence of more than one type of organellular genome (mitochondrial DNA or plastid DNA) within a cell or individual.
Heterozygosity	The proportion of a population that is heterozygous for a particular locus.
Heterozygote	A fertilized egg (zygote) with two different alleles at a designated locus; by extension, the individual that develops from such a zygote.
Heterozygous	Having different alleles at a particular locus (in contrast to homozygous).
Homozygote	A fertilized egg (zygote) with two identical alleles at a designated locus; by extension, the individual that develops from such a zygote.
Homozygous	Having the same allele at a particular locus (in contrast to heterozygous).
Inbreeding coefficient	The probability that two homologous genes in an individual are identical by descent, descended from the same gene in an ancestor; a measure of the proportion by which the heterozygosity is reduced by inbreeding; designated by F.
Indel	Insertion or deletion of nucleotide or nucleotide stretch in DNA sequence.
Kinship coefficient	The probability that two randomly chosen genes, one from each of two individuals in a population, are identical; equivalent to the inbreeding coefficient of an offspring.
Likelihood ratio	The number x that summarizes and weighs results of forensic test in statement like: “The probability of the evidence is x times more likely if the stain came from Mr Smith than if it came from an unknown unrelated individual”. It is wrong to say “likelihood ratio x is the probability that the stain came from Mr Smith” because we must always include the conditioning statement. It is recommended to use the condition “if” when using a likelihood ratio to avoid this trap.
Linkage	Inheritance of two or more genes on the same chromosome.

Linkage disequilibrium	The state in which two or more loci in a gamete are not in random proportions (i.e., the gamete frequency is not the product of the allele frequencies; abbreviated LD), the non-random association of alleles at different loci in a given population.
Linkage equilibrium	The state in which two or more loci in a gamete are in random proportions (i.e., the gamete frequency is the product of the allele frequencies; abbreviated LE).
Locus (plural Loci)	The physical location, address of a gene on a chromosome.
Massively parallel sequencing	Also called Next Generation Sequencing; the method of high-throughput DNA sequencing to determine the entire genomic sequence of a person or organism. This method processes millions of reads, or DNA sequences, in parallel instead of processing single amplicons that generate a consensus sequence.
Match	Two DNA profiles are declared to match when they are indistinguishable in genetic type. For loci with discrete alleles, two samples match when they display the same set of alleles.
Meiosis	The reductive divisions of one cell into two that occur in the development of a sperm or egg, during which the chromosome number is halved.
Microhaplotype	The combination of two or more closely linked SNPs within DNA segments of about 200 base pairs.
Negative predictive value	The probability that subjects with a negative test truly do not have the phenotypic feature.
Nucleic acid	DNA or RNA.
Nucleotide	A unit of nucleic acid composed of phosphate, a sugar, and a purine or pyrimidine base.
Phenotype	The observable physical properties of an organism (appearance, development, and behaviour). It may be externally visible, as eye colour, or observed by a special technique, as blood groups or enzymes. It is a manifestation of both the genotype and environmental factors.
Polymerase chain reaction	Molecular xerox, an in vitro process for making many copies of a fragment of DNA; abbreviated PCR.
Polymorphism	The presence of more than one allele at a locus in a population.
Positive predictive value	The probability that subjects with a positive test truly have the phenotypic feature.
Proband	A person being investigated in genealogical/familial study
Proficiency test	A test to evaluate the quality of performance of a laboratory.
Prosecutor´s fallacy	See Transposing the conditional.
Purine	The larger of the two kinds of bases found in DNA and RNA; A and G are purines.
Pyrimidine	The smaller of the two kinds of bases found in DNA and RNA; C, T, (and U in RNA) are pyrimidines.
Quality assurance	A program conducted by a laboratory to ensure accuracy and reliability of tests performed; abbreviated QA.
Quality audit	A systematic and independent examination and evaluation of a laboratory's operations.
Quality control	Activities used to monitor the quality of DNA typing to satisfy specified criteria; abbreviated QC.
Random-match	A match in the DNA profiles of two samples of DNA, where one is drawn at random from the population.
Random-match probability	The probability that the DNA in a random sample from the population has the same profile as the DNA in the evidence sample.

Random sample	A sample chosen so that each sample of the population has a known (preferably equal) chance of being represented.
Ribonucleic acid	A class of nucleic acid, synthesized based on DNA template and is part of the process of translating a DNA sequence into a phenotype; abbreviated RNA.
Sex chromosomes	The X and Y chromosomes in humans.
Short tandem repeat	Multiple copies of an identical DNA sequence arranged in direct succession in a particular region of the chromosome in which the repeat units are three, four, or five base pairs; abbreviated STR.
Somatic cells	Cells other than those in the cellular ancestry of egg and sperm.
Thymine	A pyrimidine base; one of the building blocks of DNA; abbreviated T.
Transcription	The copying of DNA into RNA, the first step in gene expression.
Transposing the conditional	A logical fallacy whereupon a conditional probability is equivocated with its inverse: given two events A and B, the probability of A happening given that B has happened is assumed to be about the same as the probability of B given A. (P(A|B) is assumed to be equal to P(B|A)).
Translation	Messenger RNA is decoded to produce a specific amino acid chain (polypeptide), the second step in gene expression.
Zygote	The diploid cell resulting from the fusion of egg and sperm.

List of abbreviations and definitions

A	Adenine
AIM	Ancestry Informative Markers
ALARA	As Low As Reasonably Achievable
ASHG	American Society for Human Genetics
AUC	Area Under receiver operating Curve
Bp	Basepair
C	Cytosine
cM	CentiMorgan
CODIS	Combined DNA Index System (USA)
DNA	Deoxyribonucleic acid
DTC	Direct to Customer genetic company
E	Evidence
EFSA	European Forensic Science Area (not to be confused with European Food Safety Authority)
ENFSI	European Network of Forensic Science Institutes
ESC	External Sniffable Characteristics
EVC	External Visual Characteristics
FDP	Forensic DNA Phenotyping
G	Guanine
GA4GH	Global Alliance for Genomics and Health
GDPR	General Data Protection Regulation
GDS	Genomic Data Sharing
GWAS	Genome Wide Association Study
H	Hypothesis, version, or scenario
IBD	Identical by Descent
IBS	Identical by State
iDASH	integrating Data for Analysis, Anonymization and SHaring
LD	Linkage disequilibrium
LINE	Long Interspersed Element
LR	Likelihood ratio
MAP	Mutual assistance procedures
MLA	Mutual legal agreement
MPS	Massively Parallel Sequencing
NCP	National contact point
NDNAD	National DNA Database of England and Wales
NIH	National Institutes of Health
NPV	Negative predictive value
Nt	Nucleotide
P	Probability

PPV	Positive predictive value
ROC	Receiver Operating Curve
SINE	Short Interspersed Element
sMIME	Secure/Multipurpose Internet Mail Extensions
SMRT	Single Molecule, Real-Time
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
T	Thymine
TESTA	Trans European Services for Telematics between Administrations
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
ZMW	Zero-Mode Waveguide

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30766-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/864087

ISBN 978-92-76-40265-7