Mareile Kaufmann

24 DNA / big genome data

Abstract: DNA has gone big. Not only is our genome increasingly subject to surveillance, but DNA data are also analyzed with ever-changing digital instruments. Genomic analyses and predictions have become inseparable from sophisticated technologies that produce knowledge in criminologically relevant arenas including anything from prevention to forensic lab work and justice. Taking its vantage point in early criminological practices of measuring and counting, this chapter traces how DNA has risen to become a central type of biodata in today's surveillance societies.

Keywords: surveillance, police, forensics, predictions, biometrics

Our genomic material is increasingly registered and digitized. As Erin Murphy puts it: "Big genome data has arrived" (2018: e8). DNA is sequenced with ever-new techniques, stored in numerous databases, and processed with algorithms (see Algorithm by Leese). DNA data and the results of their analyses circulate in public and private domains, from medicine to ancestry research, but not least in the field of law enforcement. Studying the integration of digital technologies with genomics is relevant to criminologists because it influences forensic, police, and courtroom practices as well as raising issues of surveilling bodies and bio-data. What is more, the criminological engagement with these themes can be considered essential, because measuring bodies, collecting biometric data, and processing them for investigative purposes is intricately linked to criminology as a discipline.

Indeed, studying the use of DNA technologies in law enforcement is nothing peculiar in light of the early criminologists' project to scientize the discipline by measuring bodies and translating them into statistical systems. Throughout the 19th century a plethora of practices emerged that sought to metricize human bodies, many of them in the field of criminalistics. It was the rise of photography, but also instruments such as thermometers, dynanometers for force and weight, and stereographs for profiling skulls that sparked the development of metric systems, that is diverse measuring practices, for the analysis and categorization of humans. Cesare Lombroso's (1835–1909) infamous project of classifying criminal bodies by type comes to mind. However, contem-

Disclaimers and acknowledgments: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 947681).

 https://doi.org/10.1515/9783111062037-025

poraries such as Paul Broca (1824–1880), Francis Galton (1822–1911), Alphonse Bertillon (1853–1914) and Edmond Locard (1877–1966) were contributors to the anthropometric and taxonomic project at the time. Controversies over the 'correct' metric system, method, and classification kept emerging (see Gibson and Rafter, 2006). These controversies illustrate a key point about the use of biometrics in law enforcement at large: technology is not neutral—it is always tied to the design, underlying assumptions, and use by humans.

Today, the use of biometry for forensics, such as DNA profiling, follows much stricter protocol and forensic scientists are aware of the probabilistic nature of the knowledge they produce. And still, the ways in which genomic data travel across databases and are analyzed by software is highly dependent on the humans interacting with such technologies (see Databases by Bellanova).

The rise of DNA

In order to understand the ways in which DNA is relevant to law enforcement, it is necessary to understand some basics. The roots of DNA analysis lie in bio-chemistry and blood typing, but also in hereditary studies that emerged in the mid-19th century. The 'discovery' of DNA and its integration into scientific research was not the effort of a single scientist. Johann Friedrich Miescher discovered nucleic acids in 1869, which would later be termed DNA; in the mid 1900s Phoebus Levene identified several components that led to an understanding of DNA's functioning and Erwin Chargaff uncovered details of the structure of DNA (Pray, 2008). Building on the work of many others, James Watson and Francis Crick finally identified the double-stranded helix of DNA (1953). Another momentous finding was that repeating sequences on the genome could be 'informative markers' and unique to each individual (Jeffreys et al., 1985). Some of these markers would be placed in the non-coding regions of the genome. While the coding regions of the genome encode proteins for different bodily functions and traits, non-coding regions vary much more between individuals which is why they are suited for identification. Such variations show in short sequences, so-called Short Tandem Repeats (STRs). While early methods revealed four STRs, later sequencing procedures revealed 18, then 24/25 STRs. Early sequencing techniques visualized STRs as barcodes on membranes, which then had to be compared in an analog fashion. With the digitization of such data, DNA profiles could be more efficiently stored and exchanged across laboratories, police databases, and national borders. This means that DNA profiles could be more easily compared to identify matches between a DNA profile in a database and that of an unknown owner.

The use of DNA in law enforcement

DNA matching was long contested as a reliable source of identification. Comparisons between DNA profiles, for example, never result in a 1:1 match. Instead, they work with approximations and DNA matches are expressed as probabilities. Michael Lynch et al. (2008) have traced the 'DNA wars' of the 1980s and 1990s, which revolved around controversies on sample collection and sampling methods, data availability and probabilistic modelling. It was thus necessary to develop protocol. The US National Research Council, for example, notes that such differences in technique can be documented in a laboratory protocol that "should not only specify steps and reagents, but also provide precise instructions for interpreting results, which is crucial for evaluating the reliability of a method" (NRC, 1992: 53).

Protocols or standard operating procedures, however, are enacted differently in practice (Lynch et al., 2008). They are appended, interpreted, and refined and many of such instances remain tacit—a known phenomenon amongst forensic staff (Lynch et al., 2008: 104). That is to say, even at the level of DNA matching, uncertainties exist, which is also expressed by the probabilistic communication of DNA 'matches.' That is, capturing and communicating these uncertainties is also a socio-technological practice.

By today, Next Generation Sequencing (NGS) established itself as an alternative to classic STR sequencing. NGS is a "high throughput method" utilizing "DNA sequencing technologies that are capable of processing multiple DNA sequences in parallel" (National Cancer Institute, n.d.). The result is a more comprehensive sequence that also allows for the discovery of so-called Single Nucleotide Polymorphisms (SNPs), which are genetic markers for specific traits. This can be useful, for example, in medicine for identifying a genetic disposition for a specific disease. Since the 2010s the method also gained traction in the forensic sciences and sparks growing interest. For example, NGS gives better results for degraded DNA samples (Lackey, 2017) or makes it easier to "distinguish the different DNA profiles in a complex mixture" (Bruijns et al., 2015: 2647). The latter is important, because most traces contain DNA from several owners and such mixtures are still a major challenge for forensic science. The more sequenced material available, the higher the likelihood of being able to distinguish DNA owners from each other. What is more, SNPs can be used as predictors for analyzing observable traits of an individual when a DNA match cannot be established. This process is called phenotyping. DNA profiles based on NGS can also be used for familial searches, which is to look for genetic relatedness to identify relatives of a potential suspect. Both will be discussed below.

While NGS is already an accepted sequencing technique in the global north, opinions about using DNA profiles for more than matching is divided among the police and forensic communities (Samuel and Prainsack, 2018). While some practitioners and product developers consider the predictive powers of such techniques a major progress in the field of forensics, others question the usability of their results and point to eth-

ical issues. Some countries, like the US, France, the UK, and the Netherlands have already integrated phenotyping and familial searches into forensic practice at the time of writing. Differences exist in terms of what type of phenotypic information can be used; especially phenotyping that can reveal medical conditions is more strictly regulated. Other countries have not yet adopted any of these techniques which could be for fiscal, practical, legal, or ethical reasons. Judging by the rising number of publications, products, and conference tracks about phenotyping and familial searches, however, an increase in their deployment is expectable.

The relationship between DNA and digital technologies: key phenomena

This entry fronts three phenomena as relevant research areas for criminologists. With the increasing integration of DNA and digital technologies these areas will keep on gaining traction. The most remarkable shift is the move from DNA as a tool for identification to a tool for prediction and investigation.

Databases

DNA profiles need storage in order to function as a source of identification. DNA data can now be stored in electronic databases, they can be analyzed with digital tools and exchanged between laboratories or police institutions. This leads to a rapid growth of DNA databases with many of them being, however, decentral. In the US, for example, such databases are run by different states, different public institutions, or commercial providers, because they grew out of different local forensic traditions. Here, we find many variations in information practices that influence the quality of the database: "It is flawed human beings that collect their information, write their operating codes, input their entries, maintain their systems, and search and retrieve their data. Databases may represent turbo-charged knowledge—but it is still human knowledge, just more powerful" (Murphy, 2010b: 825).

In Europe, the PRÜM treaty was created in 2005 to establish a more streamlined system for DNA databases across Europe, even requiring member states to set up DNA databases for information exchange (European Data Protection Supervisor, 2007, no page number). However, while centralized working groups and competence teams for support were established by the EU, the different stages of database development, problems with the availability and exchange of information, but also high costs accompany the implementation of the system (Jones, 2012).

Since DNA databases also form the fundament for algorithmic analysis, the composition of stored data moves to the fore. Here, scholars have voiced concern, for example, over the categorization and overrepresentation of minorities in such databases

(Murphy and Tong, 2020; Skinner 2020). The ways in which DNA databanks are organized, how and whether they speak to each other, need continued scholarly engagement in order to understand the role of digital DNA systems in law enforcement.

Phenotyping

A phenotype relates to an individual's observable features. In genetic phenotyping, the abovementioned SNPs work as predictors for analyzing traits of an individual. This procedure can amongst other things be used for medical, but also for forensic purposes. For the latter traits such as hair-, eye- and skin-color are of interest. Phenotyping efforts can also be directed at predicting age or geographical heritage by continental or sub-regional type (Kayser et al., 2023). When used for law enforcement, phenotyping is supposed to assist in investigations, especially when a DNA match cannot be established.

For phenotyping one needs large data pools that allow for a comparison of genes known to impact, for example, pigmentation with data of individuals whose geno- and phenotype is already known. An algorithm infers from these datasets which genetic marker is likely to influence what type of pigmentation. In order to obtain a graded (i. e., not continuous) visualization within the software, pigmentation is categorized. This procedure, as Roos Hopman criticizes, reflects everyday racial categories, such as 'white,' 'intermediate,' and 'black' and inscribes them into software (2023). Some phenotyping software solutions, such as Parabon Snapshot®, also seek to predict the shape of the face. For this, facial morphologies are captured and standardized, cleaned of irregularities, symmetrized, and rendered comparable (Hopman, 2023). Similar to the process above, an algorithm infers which genomic markers likely express typical morphological features. All these physiognomic results are of predictive nature: they do not describe an actual person, but predict a face (Wienroth, 2018). While practitioners are mainly concerned about the practicability of such predictions (a person can dye their hair, wear contact lenses or makeup), more fundamental critique is directed at the making and the effects of 'typing.' Indeed, not an individual's face, but a pheno*type* is produced: "The focus shifts from the genetically unique to the genetically common" (Hopman, 2020). The ambition of phenotyping may be accuracy, but such predictions never articulate accuracy, only an appearance of truth that results in categorical predictions (Hopman, 2020). Though the logic of the category was already operative in earlier biometric practices (Cole, 2001), acts of categorizing bodies based on genetic make-up express new generalizations that are now part of identification procedures. Or, as Hopman and M'charek write: "the promise of individuality, namely the face of the suspect, is quintessentially about producing the collective" (2020: 460). Not only is the net for investigation broadened, drawing entire collectives into the realm of suspicion, but the concept of the collective opens up for categorizations based on race (Hopman and M'charek, 2020). The future of phenotyping will see efforts to type

more and different characteristics, the implications of which are important to understand and discuss in detail.

Familial searches

First implemented in the UK in 2002, more and more countries are adopting familial searches, that is biological relatedness of DNA samples. In most countries, however, its use is restricted to serious cases (Granja, 2021). The technique gained traction with the advent of NGS, because the higher number of identified STRs and SNPs enables a search for genetic relatedness. Originally, a matching exercise of two DNA samples would aim to identify an as exact as possible match of 26 alleles. The more alleles match, the higher is the likelihood that two DNA samples stem from the same person. With the rise of familial searches, however, a conscious decision is taken to compare more comprehensive genomic material to focus specifically on genetic mutations or part-matches that can indicate biological relatedness. The technique, then, creates a larger pool of people from which new leads for investigation can be inferred. However, this also means that more individuals are drawn into such investigative processes. As Erin Murphy summarizes it: familial searches do not provide matches but only 'gestures'; they "point only to a list of possible suspects, all but one of whom definitely did not leave the evidence" (2010a: 313). Consequently, familial searches "affect more people, because they cast the net indiscriminately and widely" (Murphy, 2010a).

The increase in publicly owned databases, but also recreational databases, such as *FamilyTreeDNA*, *23andMe*, and *Ancestry* that offer the analysis of an individual's genealogy as a service, add a new dimension to familial searches. In the latter, individuals chose to let their DNA be analyzed to learn more about their genetic 'roots' based on a wider net for genetic overlaps. These are long-range familial searches that are also called 'forensic genealogy' when adopted by law enforcement. The use of recreational databases for forensic purposes enlarges datasets and data subjects that can be used for investigation. At the same time, this means that the "involvement with the criminal justice system is no longer a pre-requisite to participate in law enforcement searches" (Granja, 2021: 339). This and the probabilistic nature of the technique are ethical concerns and produce a clear shift in the use of DNA in law enforcement. A different dimension of this trend is that genetic knowledge is increasingly commodified, creating issues of transparency, ownership, and protection of genomic data.

Conclusion

The integration of genetics with digital technologies produces new phenomena, practices, and politics in the field of law enforcement. While the project of metrifying and typing human bodies is as old as the discipline of criminology, it takes new shape with DNA data 'going big.' For a long time, the main use of DNA was identification,

whereby the likely match between several samples is assessed. Today, the practice of DNA matching is joined by several big data practices (see Big Data by Završnik). Not only do databases grow in size and number, but today's sequencing methods can reveal ever-more comprehensive information as well as genetic markers for specific traits. While such advances generate new avenues for investigation and identification, for example from sample mixtures, they also introduce a double shift: from ascertaining identity towards prediction, and from a focus on individuals to a focus on collectives. New investigative methods such as familial searches or the prediction of body 'types' are symptomatic of this shift. These developments are accompanied by a range of issues relevant to criminologists:

- How are databases created and maintained? Can and should databases communicate with each other?
- How do new software models generate phenotypes and what exactly are they 'typing'? What assumptions and decisions are folded into these models? What are the promises and pitfalls of predicting types altogether?
- Which relations are traced in familial searches and forensic genealogy? What does it mean that genetic relation, history, and geography are now a part of police investigation?
- How are digital tools for DNA analysis used in practice? How do genomic prediction and uncertainty influence a context as sensitive as law enforcement? What are the effects of the drift from the individual to the collective?
- How are forensic technologies spreading? There is literature on the 'professional gaze' in forensic genetics (Granja and Machado, 2019; Samuel and Prainsack, 2018). Yet, how do attitudes change over time?
- Much has been written about the ethical responsibilities and controversies in forensic genetics (e.g., Wienroth et al., 2021, Roux et al., 2022), but what *ethos* do genomic technologies perform?
- What is the role of commercial providers in genetic knowledge-making?
- What are legal and regulatory challenges (cf. Scudder et al., 2018)?

Suggested reading

Granja, R. (2021). Long-range familial searches in recreational DNA databases: Expansion of affected populations, the participatory turn, and the co-production of biovalue. *New Genetics and Society*, 40(3), 331 – 352.

Hopman, R. (2023). The face as folded object: Race and the problems with 'progress' in forensic DNA phenotyping. *Social Studies of Science*, 53(6), 869 – 890. https://doi.org/10.1177/03063127211035562

Murphy, E. (2010). Relative doubt: Familial searches of DNA databases. *Michigan Law Review*, 109(3), 291 – 348.

References

Bruijns, B., Tiggelaar, R., & Gardeniers, H. (2018). Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, 39, 2642 – 2654.

Cole, S. (2001). *Suspect Identities. A History of Fingerprinting and Criminal Identification.* Cambridge, MA: Harvard University Press.

European Data Protection Supervisor. (2007). *Cross-border cooperation (Prüm Treaty).* https://edps.europa.eu/data-protection/our-work/publications/opinions/cross-border-cooperation-pr%C3%BCm-treaty_en (Accessed 5 September 2022)

Gibson, M., & Rafter, N. (2006). Editors' introduction. In D. Lombroso (ed.), *Criminal Man* (pp. 1 – 35). Durham, NC: Duke University Press.

Granja, R. (2021). Long-range familial searches in recreational DNA databases: Expansion of affected populations, the participatory turn, and the co-production of biovalue. *New Genetics and Society*, 40(3), 331 – 352.

Granja, R., & Machado, H. (2019). Ethical controversies of familial searching: The views of stakeholders in the United Kingdom and in Poland. *Science, Technology, & Human Values*, 44(6), 1068 – 1092.

Hopman, R. (2020). Opening up forensic DNA phenotyping: The logics of accuracy, commonality and valuing. *New Genetics and Society*, 39(4), 424 – 440.

Hopman, R. (2023). The face as folded object: Race and the problems with 'progress' in forensic DNA phenotyping. *Social Studies of Science*, 53(6), 869 – 890. https://doi.org/10.1177/03063127211035562

Hopman, R., & M'charek, A. (2020). Facing the unknown suspect: Forensic DNA phenotyping and the oscillation between the individual and the collective. *BioSocieties*, 15(3), 438 – 462. https://doi.org/10.1057/s41292-020-00190-9.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314, 67 – 73.

Jones, C. (2012). Complex, technologically fraught and expensive' – The problematic implementation of the Prüm Decision. *Statewatch Journal*, 22(1), 1 – 7 https://www.statewatch.org/media/documents/analyses/no-197-prum-implementation.pdf (Accessed 10 November 2023)

Kayser, M., Branicki, W., Parson, W., & Phillips, C. (2023). Recent advances in forensic DNA phenotyping of appearance, ancestry and age. *Forensic Science International: Genetics*, 65: 102870.

Lackey, A. (2017). *Capillary Electrophoresis (CE) or Next-Generation Sequencing (NGS) for STR Analysis?* ThermoFisher Scientific. https://www.thermofisher.com/blog/behindthebench/capillary-electrophoresis-or-next-generation-sequencing-for-str-analysis/ (accessed 9 November 2023)

Lynch, M., Cole, S. A., McNally, R., & Jordan, K. (2008). *Truth Machine. The Contentious History of DNA Fingerprinting.* Chicago: The University of Chicago Press.

Murphy, E. (2010a). Relative doubt: Familial searches of DNA databases. *Michigan Law Review*, 109(3), 291 – 348.

Murphy, E. (2010b). Databases, doctrine, and constitutional criminal procedure. *Fordham Urban Law Journal*, 37(3), 803 – 836.

Murphy, E. (2018). Law and policy oversight of familial searches in recreational genealogy databases. *Forensic Science International*, 292, e5–e9.

Murphy, E., & Tong, J. H. (2020). The racial composition of forensic DNA databases. *California Law Review*, 108, 1847 – 1911.

National Cancer Institute. (n.d.). *Massively Parallel Sequencing.* Genetics Dictionary. Available at: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/expand/M (3 September 2022)

National Research Council. (1992). *DNA Technology in Forensic Science.* Washington, DC: National Academy Press.

Pray, L. (2008). Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1): 100.

Roux, C., Bucht, R., Crispino, F., De Forest, P., Lennard, C., Margot, P., Miranda, M. D., et al. (2022). The Sydney Declaration – Revisiting the essence of forensic science through its fundamental principles. *Forensic Science International*, 332, 111182.

Samuel, G., & Prainsack, B. (2018). Forensic DNA phenotyping in Europe: Views "on the ground" from those who have a professional stake in the technology. *New Genetics and Society*, 38(2), 119 – 141.

Scudder, N., McNevin, D., Kelty, S. F., Walsh, S. J., & Robertson, J. (2018). Massively parallel sequencing and the emergence of forensic genomics: Defining the policy and legal issues for law enforcement. *Science & Justice: Journal of the Forensic Science Society*, 58(2), 153 – 158.

Skinner, D. (2020). Race, racism and identification in the era of technosecurity. *Science as Culture*, 29(1), 77 – 99.

Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171, 737 – 738.

Wienroth, M. (2018). Governing anticipatory technology practices. Forensic DNA phenotyping and the forensic genetics community in Europe. *New Genetics and Society*, 37(2), 137 – 152.

Wienroth, M., Granja, R., Lipphardt, V., Amoako, E. N., & McCartney, C. (2021). Ethics as lived practice. Anticipatory capacity and ethical decision-making in forensic genetics. *Genes*, 12(12), 1868.