Published online: February 1, 2018

EMBO reports

Science & Society

Uncharted waters

Next-generation sequencing and machine learning software allow forensic science to expand into phenotype prediction from DNA samples

Philip Hunter

Forensic molecular biology is entering uncharted waters with ethical and legal shallows to navigate. Next-generation sequencing and other technologies now enable determination of phenotypic characteristics and thereby ethnic and racial profiling from recovered DNA samples from crime scenes, which represents a major new frontier for forensic science. At the same time, two traditional areas of forensic biology—DNA profiling and entomology to establish the time of death—have also made major advances based on these new sequencing technologies. It also has allowed forensic biology to expand into new terrains such as the illegal trafficking of rhinoceros horns and elephant tusks.

> *"MPS has now further improved the accuracy and power of DNA typing"*

Until around a decade ago, the two principle areas of forensic biology were DNA profiling and entomology. In some countries, palynology was also used to analyse pollen and spores so as to track where a suspect or objects from a crime scene had been before the crime. Now, the availability of massively parallel sequencing (MPS) or next-generation sequencing (NGS) is driving genuinely new developments in forensic molecular biology. Another emerging area is the analysis of epigenetic markers on DNA samples, which can potentially yield valuable information about the age and lifestyle of suspects. "The most notable recent development is the widespread adoption of MPS or NGS for forensic purposes", said Dennis McNevin from the Forensic Genetics department at the University of Canberra, Australia. "This has been driven by the availability of what I would call 'forensically mature' benchtop MPS platforms".

The impact of next-generation sequencing

As a single targeted assay can cover the whole genome from field DNA samples, NGS has also enhanced traditional DNA profiling to identify suspects against a database. DNA profiling was first used in 1985 in an immigration test case by its inventor Alec Jeffreys at the University of Leicester, UK [1]. The technique, which exploited the then recently developed polymerase chain reaction (PCR), quickly gained traction for criminal investigations by exploiting microsatellites or short tandem repeats (STRs) of short DNA sequences of 1–7 bases that are repeated 5–100 times at each locus. STRs are scattered across the genome and have individual variations regarding the number of repeats. Each STR is polymorphic but typically has only a few alleles, each of which is shared by 5–20% of individuals in the human population. The power of DNA profiling comes from combining multiple STRs given that the variations are independent of each other: the probability that two individuals other than identical twins would share the same combination of STRs is the smaller the more segments are chosen for comparison. MPS has now further improved the accuracy and power of DNA typing. "The use of MPS has uncovered many more polymorphisms in STR alleles than was previously recognised", McNevin commented.

This has reduced the risk of a false identification to astronomically low levels, noted Peter Gunn from the Centre for Forensic Science at the University of Technology in Sydney, Australia. "Where 'traditional' STR typing has changed is in the number of loci being examined, which can now be as high as 21 in the current Australian standard level or even over 30 in some Chinese systems for example", he said. "The identification power of a full 21-locus profile is ridiculously huge—the chance of a random match is something like 1 in 10^{29} or 10^{30} which is a stupidly small number, when there are only about 8×10^9 people in the world. Where the advantage comes is that even when you get only a partial profile from a crime scene, you will usually get enough information to make a 'match' with a database sample". The other big gain is in sensitivity. "Template amounts of under 100 picograms will be sufficient to generate a full profile", Gunn explained. "That is only about 15 cells' worth of DNA".

> *"Height is another valuable variable for forensic investigations to construct a profile of a suspect, but it is much harder to predict than pigmentation..."*

McNevin also highlighted the equally important role of "probabilistic software" to compute probabilities or "likelihoods" that two individuals are the same based on their DNA profile. However, the growing use of probability in predictions is a two-edged sword, for while it is valuable in criminal investigations, it can be confusing for judges and juries during trials, as Gunn explained. "These techniques are seriously mathematical, so that even most forensic biologists are not equipped to understand them, making

Freelance journalist in London, UK. E-mail: ph@philiphunter.com
DOI 10.15252/embr.201845810

them a 'black box'", he explained. "This is challenged in court often as forensic biologists are presenting results when they don't actually understand how they got them. Then being probabilistic methods, if you run the program multiple times on the one set of results, you don't get the same answer each time. Close, usually, but not the same. Courts don't like that—they want certainty and we can't give them that".

Suspect profiling from DNA

The same applies to other applications of MPS, which fall into three broad categories: phenotyping, estimation of age and determination of ancestry. "There are many custom panels developed by the forensic genetics community for identity, ancestry, phenotype, age estimation and tissue of origin", McNevin said. "All this has ushered in a new era of 'forensic intelligence' gathering where DNA is now no longer simply used for matching a crime scene sample with a suspect or database record. We can now determine the ancestry, pigmentation and age of an unknown DNA donor, with many more externally visible characteristics (EVCs) perhaps able to be predicted in the future. For example, I offer a service in my own lab where police agencies can send DNA for ancestry and phenotype analysis".

"DNA methylation is forensically very exciting as it may potentially provide a lot of additional hints for intelligence purposes…"

Prediction of phenotype covers a wide range of potential traits from skin pigmentation and eye colour to more complex behavioural features. "Right now, simple pigmentation categories, that is say blue eyes, brown hair or pale skin, can be predicted with a high degree of accuracy", said Susan Walsh from the Forensic & Investigative Sciences Program at Indiana University-Purdue University in Indianapolis, USA. "Of course, we consistently try to improve this all the time and my lab specifically works on improving pigmentation to yield a real colour prediction as opposed to a simple category".

Earlier attempts at predicting skin colour were fairly crude and failed to take account of either differences between major population groups, such as Europeans and Africans, or the variation within them. As Walsh noted, previous studies treated Europeans as one single group in their prediction analysis and ignored the many shades of colour among them, as well as different levels of tanning. Walsh and colleagues identified a set of 36 single nucleotide polymorphisms (SNPs) in 16 genes known to be associated with human pigmentation that would be predictive of skin colour in any human irrespective of origin [2]. Using the so-called Fitzpatrick scale to separate skin colour into three broad categories—light, intermediate dark and dark black—tests on 194 individuals yielded prediction accuracies measured on a curve plotting true against false positives of 0.92 for light skin, 0.74 for intermediate and 0.94 for dark, which the authors assert are higher than any obtained previously. The paper noted though that greater accuracy could almost certainly be achieved with additional but as yet unknown SNPs identified via future genome-wide association studies, especially for light skin colour.

Height is another valuable variable for forensic investigations to construct a profile of a suspect, but it is much harder to predict than pigmentation given the large number of genes involved. "Almost 10,000 DNA variants with a very small additive effect have been identified to be involved in the determination of human height and they still explain merely about 30% of the entire variance of this trait", commented Wojciech Branicki from the Central Forensic Laboratory of the Police in Warsaw, Poland.

His group has also been working on hair morphology, which is still complex genetically but much less so than height. Study of this particular trait highlighted two issues: that baldness in males depends on age and that the underlying genes vary between populations. "We showed that hair loss can be predicted in males but more accurate prediction is possible in a group of older men of European origin", Branicki said. "It means that the genetics of baldness is different in Asia and Europe and we have good predictors/markers for Europeans. However, baldness is a progressive trait and thus genotypes associated with a risk of male pattern baldness need time to develop the phenotype. Young individuals may still have a dense head of hair, but DNA will predict baldness because they will develop this phenotype in a couple of years. This demonstrates that in the case of some externally visible traits, in order to predict them accurately, we need to predict both ancestry and age first".

Epigenetic markers yield information

Accurate genetic markers of age have proved elusive, but recent analysis using MPS to identify relevant epigenetic DNA methylation patterns has yielded some promising results [3]. The underlying theory is that methylation patterns change with age in response to environmental cues and other factors. The study looked at 45 age-related CpG sites that are susceptible to DNA methylation and found that 23 of these could contribute to age prediction modelling using machine learning techniques to refine the link between actual age and methylation patterns. "Our group has now shown further that examining just 5 CpG sites in the human genome is sufficient to predict age with the accuracy of about 4 years", Branicki stated. "Accuracy is better in the case of younger individuals". Yet, since DNA methylation is tissue-specific, it may be necessary to identify different markers and algorithms to predict age from forensically relevant DNA sources such as saliva or semen.

"The major challenge for DNA 'forensic intelligence' is safeguarding genetic privacy."

This in turn leads to another forensic application of DNA methylation: the accurate assessment of body fluids recovered from a crime scene to reconstruct the course of events. Until recently, this involved analysis of proteins or RNAs associated with specific cell types, but these are prone to degradation. DNA methylation patterns are more durable and thus offer an alternative method to look for differences between cell types [4]. The challenges include verifying that the methylation patterns themselves are robust against exogenous exposure and that they are sufficiently independent of other factors, such as health and age.

Indeed, methylation patterns can potentially yield more phenotypic information than just age or tissue origin. "DNA methylation is forensically very exciting as it may potentially provide a lot of additional hints for intelligence purposes, for example about life style of an unknown individual", Branicki said. Lifestyle factors might also have an impact on another forensic phenotype target, which is not purely a function of genetics: facial structure. Yet, there is still much work in progress and much has been subject to overhyped or premature claims of success. "We need to process a huge amount of genetic information, so it is not just a matter of collecting the data, but analysing it so massive computing power is necessary", Gunn explained. But such power is now available and the possibility of facial morphology prediction is emerging. Mark Barash, one of Gunn's colleagues, has identified associations between craniofacial traits and 12 SNPs located in 12 genes and intergenic regions. "The last two years have seen DNA phenotyping generally become an actual tool in forensic investigation, primarily in the United States", said Robert Oldt, a forensic researcher at Arizona State University in the United States. "Just a few months ago, a 2009 Louisiana cold case was solved after a facial composite of the murderer was generated via DNA phenotyping, and other cases in the last year have utilized genome-produced sketches as an attempt to identify suspects. This technology has also been used in reconstructing the appearance of unidentified victims".

In the Louisiana case, a murder investigation went cold for lack of any tangible clues until a facial sketch and characteristic profile were built by Parabon NanoLabs based on DNA from the crime scene. NanoLabs' sketch indicated the suspect was a white male, contradicting previous assumptions based on uncorroborated circumstantial evidence that the murder had been committed by a Hispanic male. This led to identification and an arrest (https://www.forensicmag.com/news/2017/07/tip-dna-phenotype-snapshot-leads-arrest-2009-murder-cold-case).

Legal and ethical challenges

However, facial reconstruction as well as the increasing power and scope of MPS in general raises wider ethical and legal issues. "This trait evokes particularly strong emotions in forensics as prediction of this trait could potentially lead to direct identification of an unknown individual just based on results of DNA analysis", commented Branicki about facial reconstruction. "The major challenge for DNA 'forensic intelligence' is safeguarding genetic privacy", McNevin explained. "As more of the human genome is sequenced in a forensic context, then more attributes of the donor are revealed. Up until now, it was thought that STRs used universally for identity were largely uninformative for these attributes but even this paradigm is being challenged with new research suggesting that STRs have a regulatory role in gene expression".

> *"... DNA-based techniques in general promise to transform investigation of many crimes involving theft or killing of endangered wildlife species ..."*

Now the addition of markers for various phenotypes is yielding further information about the donor, some of which may be associated with health-related phenotypes. "[O]ne question is whether such phenotypes should be revealed to the donor", McNevin said. "Another question is under what circumstances phenotypes should be collected, retained and/or destroyed. Should all DNA collected from crime scenes be subject to a full suite of phenotype analyses, for example, or should those analyses be reserved for instances where there are no suspects and no database matches?"

Covering all areas of life

At least such concerns do not apply to non-human DNA from crime scenes, which have also attracted growing interest for their potential to yield information about both crime and suspects. "We believe that Non-Human Forensic Genetics (NHFG) will overtake Human Forensic Genetics (HFG), just because in the world there is much more non-human than human genetic material and so NHFG can provide more evidence than HFG", commented Miguel Arenas, a specialist in forensic genetics from the Facultad de Biología at the University of Vigo in Spain. But Arenas added that NHFG was at a much earlier stage of development than HFG. "NHFG has to make more progress in some crucial aspects for forensics, with still very limited reference genetic sequences for many organisms".

NHFG can potentially cover the whole of life, according to Antonio Amorim from the Population Genetics & Evolution Research Group at the University of Porto in Portugal. "The first division would be between viruses and the rest of life, then among the latter pro- and eukaryotes and so on", he said. Some of the early successes of NHFG have come from the high end of eukaryotes through analysis of both plant and animal remains. Domestic pets, for instance, can provide valuable samples attached to clothing of victims or suspects; Oldt noted two cases where cat hairs played a role in murder investigations. One of these occurred more than 20 years ago in Canada when fairly straightforward analysis of cat hairs obtained from the body of the suspect helped secure his conviction [5].

More recently, microbial forensics, covering viruses and prokaryotes as well as single-celled eukaryotes, has developed sufficiently to play a significant role in real cases. "The amount of molecular microbial data collected from any sample can be huge and today it is crucial to collect and analyse such data quickly", explained Ricardo Araujo, a researcher from the Institute of Molecular Pathology and Immunology of the University of Porto (Ipatimup) in Portugal. "Illicit action can have consequences in the microbial populations and some effects can only be detected in a crime scene for a short period of time while contamination should be clearly controlled and detected". He cited some advances, such as tools based on machine learning, that can provide accurate assessments of how long a murder victim has been dead at the point of discovery [6].

Tracking poachers and traffickers

Furthermore, microbial forensics is being applied in some areas where traditional investigative techniques have failed to make much impact, such as illegal trafficking in exotic animals. "Once we understand the role of each microorganism in the environment and its distribution across crime scenes, it can help determine the origin of illicit products, revealing routes taken for delivery and even pursuing environmental crimes that are difficult to track today", Araujo said. In fact, DNA-based techniques

in general promise to transform investigation of many crimes involving theft or killing of endangered wildlife species, as reflected in formation of a dedicated body, the Society for Wildlife Forensic Science (www.wildlifeforensicscience.org).

"Global public interest in wildlife crime is definitely increasing, but this is not well reflected yet in resourcing for investigations" said Sherryn Ciavaglia, Wildlife DNA Forensics – Diagnostics, Wildlife & Molecular Biology Section at the Science and Advice for Scottish Agriculture (SASA). "The large amount of money involved in trafficking high profile endangered species, such as rhinos and elephants, is the reason why organised crime is now well recognised as being associated with these activities".

But the promise of molecular techniques encourages devotion of greater resources. "I am confident that both forensic methods applied to criminal prosecutions and supporting research that helps provide investigative leads can and do make an impact to help protect endangered species", Ciavaglia commented. "There is some really ground-breaking work being done by the Centre for Conservation Biology, University of Washington in Seattle, using DNA analysis to trace elephant tusks seized in illegal shipments back to point of origin in Africa and also to link multiple shipments to each other, helping to provide intelligence about global organised criminal trade networks". South African authorities have built a large database of DNA samples from African rhinoceroses, called RhoDIS, and taught rangers forensic crime scene practices to collect samples from carcasses killed by poachers. The aim of these efforts is not just to secure evidence to track and convict individual poachers, but to curb the increasing illegal trade in rhinoceros horns and elephant tusks [7,8].

Ciavaglia suggested that such forensic methods were already having a deterrent effect. This is also a factor for other forms of crime, given the growing number of cases that went cold only to be reopened and solved later with the application of newly developed molecular techniques to samples collected at the time. Given the current rate of progress and what is in the research pipeline, the role of molecular techniques is set to proliferate further in deterrence, investigation and conviction.

References

1. Jeffreys AJ, Brookfield JF, Semeonoff R (1985) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* 317: 818 – 819
2. Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, Koller J, Kovatsi L, Wollstein A, Branicki W *et al* (2017) Global skin colour prediction from DNA. *Hum Genet* 136: 847 – 863
3. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Court DS (2017) DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci Int Genet* 28: 225 – 236
4. Forat S, Huettel B, Reinhardt R, Fimmers R, Haidl G, Denschlag D, Olek K (2016) Methylation markers for the identification of body fluids and tissues from forensic trace evidence. *PLoS One* 11: e0147973
5. Kolata G (1997) Cat hair finds way into courtroom in canadian murder trial. *The New York Times*, April 24
6. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, Lenmts NH (2016) A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One* 11: e0167370
7. Harper C, Ludwig A, Clarke A, Makgopela K, Yurchenko A, Guthrie A, Dobrynin P, Tamazian G, Emslie R, van Heerden M *et al* (2017) Robust forensic matching of confiscated horns to individual poached African rhinoceros. *Curr Biol* 28: R13 – R14
8. Kolata G (2018) In Africa, geneticists are hunting poachers. *The New York Times*, Jan 8

