Forensic Genomics: Kin Privacy, Driftnets and Other Open Questions*

Frank Stajano
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
fms27@cam.ac.uk

Lucia Bianchi
Studio Legale Bianchi
Via G. di Vittorio
S. Casciano Val di Pesa
50026 Firenze
Italy
luciagbianchi@gmail.com

Pietro Liò
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
pietro.lio@cl.cam.ac.uk

Douwe Korff
London Metropolitan
University
Dept of Law, Governance and
International Relations
Ladbroke House
62-66 Highbury Grove
London N5 2AD
United Kingdom
d.korff@londonmet.ac.uk

ABSTRACT

DNA analysis is increasingly used in forensics, where it is being pushed as the holy grail of identification. But we are approaching a dramatic "phase change" as we move from genetics to genomics: when sequencing the entire genome of a person becomes sufficiently cheap as to become a routine operation, as is likely to happen in the coming decades, then each DNA examination will expose a wealth of very sensitive personal information about the examined individual, as well as her relatives. In this interdisciplinary discussion paper we highlight the complexity of DNA-related privacy issues as we move into the genomic (as opposed to genetic) era: the "driftnet" approach of comparing scene-of-crime samples against the DNA of the whole population rather than just against that of chosen suspects; the potential for errors in forensic DNA analysis and the consequences on security and privacy; the civil liberties implications of the interaction between medical and forensic applications of genomics. For example, your kin can provide valuable information in a database matching procedure against you even if you don't; and being able to read the whole of a sampled genome, rather than just 13 specific markers from it, provides information about the medical and physical characteristics of the individual.

Our aim is to offer a simple but thought-provoking and technically accurate summary of the many issues involved, hoping to stimulate an informed public debate on the statutes by which DNA collection, storage and processing should be regulated.

Categories and Subject Descriptors

K.4 [**Computing Milieux**]: Computers and Society.

General Terms

Legal Aspects, Security.

Keywords

Forensic genomics, DNA database, Kin privacy, Human Genome Project, Cold and hot hits, Data protection, Informational privacy.

1. INTRODUCTION AND MOTIVATION

Nowadays, genetic analysis of human samples is commonplace, both in forensics and in medicine. Specific positions in the chromosomes are examined: in forensics, to find matches with scene-of-crime samples; in medicine, to check for genes related to particular diseases.

The major qualitative change that is about to take place, enabled by advances in technology, is the switch from *genetics* to *genomics*: no longer just the analysis of a few specific points, amounting to a few thousand base pairs (on average one gene is about one thousand base pairs), but of the entire genome of a person (3 billion base pairs comprising about twenty thousand genes[1]). This will turn the tables entirely, and this paper explores the scenarios and the possible consequences.

An example of the growing importance of genomics in our society is given by the recent polemics following Nobel prize winner James Watson [2, 3] receiving a copy of his genome, recorded on two DVDs. The sequencing effort took only two months and just

*Revision 56 of 2008-09-02 16:25:51 +0100 (Tue, 02 Sep 2008).

WPES'08, October 27, 2008, Alexandria, Virginia, USA.

[1]The genes are only a small fraction of the total, compared to the vast amount of regulatory DNA that surrounds them.

under one million dollars, several orders of magnitude less than the effort for sequencing the first genome[2]. Watson chose to publish his genome for the benefit of researchers, with the exception of his apolipoprotein E gene sequence, the status of which he does not wish to know because it is linked to the likelihood of getting Alzheimer's disease. Watson was then quoted as saying that black Africans are not as intelligent as whites. But the analysis of his own genome by the company DeCode suggested that he probably had an African great-grandparent.

The motivations that drive this paper are as follows.

1. There are many privacy problems associated with forensic use of genetic information. There will be many more if we ever get to forensic use of *genomic* information. Some are obvious. This paper helps you acquire awareness and understanding of the less obvious ones.

2. The purpose of this paper is to foster an informed debate on what regulatory and legal constraints should be placed on the procedures involving genomic collection and manipulation in order to protect privacy and civil liberties.

3. This is not a paper of solutions but mostly a paper of unanswered questions about the effects in forensics of the transition from genetics to genomics. The point is that many are unaware of even the questions. If we don't discuss them now, and provide adequate regulatory answers, it will be impossible to put the genie back in the bottle. You can't "unpublish" your genome after it has been disclosed.

4. An interdisciplinary approach is required to make sense of this complex problem. Our professional backgrounds cover system security, law and bioinformatics.

In order to reach and engage as wide an audience as possible, we have arranged the first few sections of this paper as collections of concise, self-contained and easily digested information nuggets. These cover "important non-obvious facts" (section 2), "things that might one day happen" (section 3) and "open questions we should discuss" (section 4). For readability, we have attempted to stay away from domain-specific jargon in the short bullet points in these sections, although we stopped short of dumbing things down. In a subsequent section ("a closer look", section 5) we revisit some issues in greater depth.

2. IMPORTANT NON-OBVIOUS FACTS

1. The cost of sequencing a human genome, close to one million US$ in 2008, is expected to drop to 1,000 US$ in the near future[3]. meaning most first-world children could get it done at birth for medical reasons (predicting diseases and taking preventive measures) [37].

2. Currently, forensic analysis is performed only on a small number of STR loci[4]; in other words, the matching is performed not on the whole genome but on an extremely small subset of specified locations, meaning that the genetic information currently acquired by forensic examiners is essentially useless for purposes other than identification [38, 39]. But we are heading towards a dramatic phase change: when decoding the entire genome of an individual becomes a routine operation, then genetic analysis of a biological sample will yield vast amounts of very intimate physiological and medical information—not just an identification fingerprint. Improper use of this information might have devastating effects on society.

3. Even if the low-level comparison of biological samples is scientifically accurate, the procedure for forensic analysis of DNA is prone to error at several stages: acquisition, labelling, statistical interpretation of the (probabilistic) result, legal requirements (variable by jurisdiction) on number of matches needed to validate a result. (See section 5.1.)

4. Finding genetic material at the scene of a crime (hair, blood, residue on toothbrushes and razors), like finding a fingerprint, may help identify a culprit. But DNA tells much more about the individual than a fingerprint does. It's not just a match on identity, it's the full genetic blueprint of the organism i.e. all the instructions for building an organism.

5. In the near future, the capabilities of forensic DNA science will be boosted by the refinement of three recent key innovations: thin film transistors (TFT) for *field (out of lab)* data acquisition, cheap and quick microarray-based genome sequencing, and artificial-intelligence methods for data analysis and interpretation. Together, they will provide genomics with effective cost, sequencing and analysis time reduction with respect to the past few decades.

6. Contrary to what happens with other biometric technologies used in forensic analysis, such as fingerprints and iris codes, with genetic material you can run much more complex queries than just "do these samples (the reference one and the one at the scene of crime) come from the same individual?". These include such diverse applications as "Is he the father? / Are they related?" and "what diseases is he likely to get?". Since your genes are not independently inherited, the presence of a non obvious character may be inferred from other more obvious characters.

7. KIN PRIVACY. You share much of your genome with your relatives. Therefore, even if you keep yours very private and never disclose it, much of what it contains can be inferred (probabilistically) by examining the genome of your relatives. (See section 5.2.)

8. Digital copying/reading of genomic information may not be error free and a potential mistake may be difficult to identify.

9. Internationally, several jurisdictions have laws or proposals dealing with at least some aspects of forensic and medical uses of genomic material. However, coverage of the relevant issues in national legislation is still sketchy and non-uniform, and many important data protection provisions are not fully enforced, even where they have been stated in principle. (See section 5.4.)

[2]The Human Genome Project produced the first complete genome sequence in 2003, based on samples from many anonymous donors. The first two genomes belonging to individuals were published within a few days of each other in 2007: Watson's was the second, the first being that of biologist and entrepreneur J. Craig Venter.

[3]Grants totalling over US$ 20 million have been awarded by NHGRI/NIH towards that goal [1].

[4]An STR (Short Tandem Repeat) is a class of DNA patterns. Seeking a match in a specified set of STR markers is a common forensic technique for the "genetic fingerprinting" of individuals.

3. THINGS THAT MIGHT ONE DAY HAPPEN

1. A private investigator collects some of the biological samples you leave daily anywhere you go (hairs, skin flakes etc) and obtains a full copy of your genome for his client—who then learns, among many other things, your susceptibility to Alzheimer's disease, cancer, food allergies, intolerance to chemicals and so forth, and whether you really are the father of your daughter.

2. Governments hold complete genomic databases of all their citizens, for combined reasons of national health and national security. This is considered sensitive personal information. However, tens of thousands of clerks have access to the databases. Every now and then, one of them loses a laptop with all the data, or a set of disks with an unencrypted copy of the database goes missing in the post. (They've already shown they're good at that [20, 21, 22, 42].)

3. Full genomic data of every individual is publicly available for medical research, but is claimed to have been anonymized. However, advances in computing one day allow anyone to run a simulation showing what any given genome would turn into, when growing as a full human body. There's even a slider for the age: see the person at 5 years old, 10, 20, 50, 80... Stalkers de-anonymize the genome of celebrities by recognizing them from this reconstruction. The secret police does the same with dissidents.

4. Insurers demand genetic prescreening before offering medical insurance. (In the US, GINA aims to prevent that; see section 5.4.) Even a national health service might require prescreening because emergency (and then chronic) treatment for a disease that could have been prevented is an unnecessary cost to society.

5. Your genome, like everyone else's, ends up being public knowledge. You are not able to get a date with the girl of your dreams because your DNA indicates you are not a good enough prospect.

6. Everyone, at birth, to ensure better medical assistance throughout their lifetime, has their genome sequenced and stored in a central database. A central refrigerated facility also stores some of their stem cells, to be used for regeneration of damaged organs. Given the history of security violations that have plagued any such large centralized facilities, it is unclear how unauthorized access to the database and stem cell bank will be prevented.

7. China's one-child policy and India's dowry traditions have already caused countless instances of female infanticide due to the preference of the parents for a baby boy heir. Once genome sequencing at birth is commonplace, the same mindset naturally leads to eugenetic infanticide: the killing at birth of any babies rated "not good enough"—for instance, those genetically more likely to develop certain diseases in later life.

8. KIN PRIVACY. The availability of personal genomes may provide a huge number of forensic markers. In theory each nucleotide in a genome can be used as a molecular marker. Close and distant relationships can be detected (see for instance the Romanov [4] and the Ashkenazi Jews [49] cases). Two aspects should be mentioned. First, compared to today's methods, using many more markers would extend the kinship that can provide valuable information in a database matching procedure. Secondly, reading the different parts of the available genomes would provide valuable hypotheses on the medical and physical characteristics of the individual ("it's probably a diabetic male with red hair"). These two aspects are not independent and may act synergistically, strengthening each other.

4. OPEN QUESTIONS WE SHOULD DISCUSS

1. If police routinely acquire DNA samples and plan to use them later for identification of suspects (because it's one of the most accurate methods available), should they be allowed to do anything with them other than checking whether they match against other samples? Should they be allowed to infer the gender, race, diseases of that person? If not, what technical framework should be in place to support a legal prohibition on such misuse? Is cryptography [7] the appropriate tool? (Discuss issues such as: store only a representative string, not the whole genome; certainly don't store the tissue sample (except for umbilical cord blood or just stem cells); anonymization; how hard it is to de-anonymize (particularly for studying recidive cancer); little tamper-proof analysing machine that only gives specific answers to allowed queries instead of a sequenced genome; procedure for the correct use of such a machine; etc.)

2. COLD AND HOT HITS. In the above situation, should police be allowed to test the sample found on the scene of crime only against the DNA of a small subset of suspects that were singled out through legally authorized investigation, or against that of everyone in the country? (See section 5.3.) To paraphrase Phil Zimmermann's 1996 senate testimony [14] during the great crypto wars of the last decade, it's the difference between fishing with a hook and line and fishing with a driftnet, "making a quantitative and qualitative Orwellian difference to the health of democracy".

3. In Communist East Germany, the Stasi secret police collected sweat samples of suspected dissidents in a giant "smell bank" of glass jars[5], in order to identify them and track them down with specially trained dogs if they went into hiding. In 2007, the same spine-chilling techniques were used by modern-day Germany for *preventive* tracking of G8 demonstrators [6]. Once the police builds a giant "DNA bank", what's to stop them using it to track down and crush political opposition to the regime of the day? To cite Zimmermann's passionate and inspirational testimony again [14]:

> This is unsettling because in a democracy, it is possible for bad people to occasionally get elected—sometimes very bad people. Normally, a well-functioning democracy has ways to remove these people from power. But the wrong technology infrastructure could allow such a future government to watch every move anyone makes to oppose it. It could very well be the last government we ever elect. When making public policy decisions about new technologies for the government, I think one should ask oneself which technologies would best strengthen the hand of a police state. Then, do

[5]Now on display at the Stasi museum in Berlin.

not allow the government to deploy those technologies. This is simply a matter of good civic hygiene.

For the sake of space we do not discuss other aspects that nevertheless we envisage may gain importance, such as access of genomic information as a source of power, effective research on aging based on genomics etc.

4. If police, or insurers, or your employer, want to know about your genome and you don't consent to it, they can get it from the hair / skin particles / sweat etc that you leave behind in daily life. Should they be allowed to use that? What regulations should govern the genomic analysis of biological samples that were "found", as opposed to being purposefully and willfully supplied by their owner?

5. KIN PRIVACY. If one sister wants to test for breast cancer likelihood (in order to undergo preventive surgery if necessary) and the other doesn't want to know (fearing depression if the answer is positive though not certain), can we protect the rights of both?

6. KIN PRIVACY. If you are under investigation, some of your distant relatives that you never even met might give the police some key information about you, or at least some working hypotheses. Therefore the police may get information about your genome even if you don't consent to disclosing it. And they don't have to get it from those distant relatives *today*: it's good enough if those relatives disclosed it in the past (perhaps in a country like the UK where your DNA is sampled even if you are stopped for a traffic offense) and it's now on file somewhere. Should all this be allowed?

7. KIN PRIVACY and monetization of desirable genetic traits. If your genome is worth money to Big Pharma because you were born with a valuable immunity, should you share the profits with your kin if you let Big Pharma analyze it for a fee? Should you be allowed to disclose your genome (and, implicitly, theirs) to Big Pharma if other relatives don't want to? Should you be allowed to sell it for profit at all? Is it moral? Who owns the rights—are they a "family heirloom"? [43]

8. What about transplanted organs? Should the recipient of a donated organ be allowed to profit by "selling" genetic features of the received organ? Shouldn't there be a wholesale ban on IPR-style monetization and monopolization of genomic features?

5. A CLOSER LOOK

5.1 The fallibility of the forensic DNA analysis procedure

Perhaps the first thing to say is that there are in fact *several* procedures, corresponding to the different types of "questions" one might wish to ask. The analysis procedure that must be followed to answer the question "does this blood stain found on the scene of crime come from this suspect?" (how likely it is that these two samples come from the same organism?) is not quite the same as the one for "to whom does this blood belong?" (who, of the people whose DNA is in the database, best matches this sample, and is the match good enough to say that it's really the same person and not just the closest among the ones we have on file?), and is certainly quite different from the one for "is he the father of this child?" (we know the two samples are definitely not from the same organism, but how likely it is that one organism generated the other?) and from "is this person at risk of developing this disease?" (not a comparison of samples but a search for specific genetic indicators).

The quantitative results from the matching operation are heavily dependent on the chosen statistical framework and particularly on the composition of the reference database. We shall examine this aspect later in section 5.3.

Meanwhile, since forensic analysis is an input to an inherently adversarial process that ultimately results in a winner and a loser, there are clear incentives for tampering with the procedure before and after the actual biotech phase.

At present, according to the procedure recommended by CODIS[6], DNA identification is obtained by matching 13 nuclear STR[7] markers of a victim's profile (personal items, like toothbrushes, and used razors) to a direct antemortem sample of the victim or to family references: either or both biological parents of the victim. Related cases include corpse identification, semen detection on underwear for suspected infidelity and autopsies for human identification following accident investigations. When a match is found, DNA typing[8] is performed again by a scientist or technician who does not know which sample he is processing: the sample is only identified by a bar code and no information is provided about the previous typing result. If a new PCR[9] analysis of the stored biological material confirms the match, fresh material is taken from the alleged suspect and analyzed in another laboratory. Only after a third confirmation of PCR results in matching 13 STR markers is the match reported to the relevant authority. But things vary by jurisdiction: German courts, for instance, generally consider five or six STRs to be sufficiently strong evidence of identity.

There is however no absolute assurance that the names attached to the barcodes that identify the samples are correct, nor that the names in the reference database are: when the enrollment procedure includes saliva swab testing of essentially any person arrested or stopped by the police, as happens in the UK, a number of labelling errors (both unintentional and intentional) are to be expected, and indeed have happened [8] on a large scale[10]. Felons may give the wrong name in the first place[11], or may resort to subterfuge or bribery to have the labels or the evidence bags swapped before the test tubes reach the lab. Insider fraud is always possible: given the right incentives, corrupt technicians or officers might edit database entries and swap or remove or add samples. The investigator may incorrectly assume that the hair fragments on X's razor belong to X whereas they don't (or because the razor found in X's bathroom wasn't even X's razor but someone else's). Someone asked to provide a DNA sample on several occasions may give a different false name each time, resulting in the database storing several identities (all wrong) for very similar sequences (extremely similar, in fact, as they come from the same individual). There may

[6]The Combined DNA Index System is the FBI-funded computer-based system that allows investigators to search DNA profiles.

[7]See footnote 4.

[8]Also called DNA testing or DNA profiling: a technique used to distinguish between individuals by comparing samples of their DNA.

[9]Polymerase Chain Reaction is a commonly used laboratory technique for isolating and exponentially amplifying a DNA sequence.

[10]About 550,000 files with wrong or misspelt names in the UK's DNA database of 4 million entries, which is the largest in the world at the time of writing.

[11]The wrong name may be a corrupted version of the correct one (e.g. misspelt, or incomplete, or scribbled in messy handwriting and then re-read as something different) or the name of another person, or a totally made-up one.

also be accidental or deliberate cross-contamination of samples of two different individuals during the acquisition phase, resulting in a given DNA sequence being recorded under different identities. See for instance police and statisticians errors reported for the O.J. Simpson case [44, 45, 46]. Finally, DNA can be newly synthesised and spread all around the crime scene (currently the cost is less than $0.55 per base pair, quickly decreasing).

5.2 Kin privacy

There is at the same time enormous similarity and enormous diversity between the genomes of two unrelated human beings. As for similarity, 99.9% of one individual's DNA sequences will be identical to that of another person. Of the 0.1% difference, over 80% will be single nucleotide polymorphisms (SNPs). An SNP is a single base substitution of one nucleotide with another, and both versions are observed in the general population at a frequency greater than 1%. Current estimates are that SNPs occur as frequently as every 100-300 bases. As for diversity, therefore, any two unrelated human beings differ by about 3 million distinct SNPs.

(Note on this subject that, to keep forensic identification applications separate from medical diagnostic ones, the standard for identification should be based exclusively on high variability markers that occur in non-coding DNA, i.e. portions of the genome not associated with identified functions.)

This diversity between individuals, however, gradually reduces for people with common ancestry [39], down to the case of monozygotic twins where it almost disappears. The similarities between the genomes of related individuals [47, 48], and the fact that knowing the genomes of an individual's relatives yields much of the information that one might wish to read from the genome of that individual, is what we define as **kin privacy**. Individuals may share different inherited blocks with relatives and distant ancestors. While it is sometimes difficult to pick out relatedness using only a few STR markers, the problem becomes much easier to solve if you know the entire genome sequences.

Monozygotic twins, which occur about once in every 250 births, start from the same zygote which then divides into two embryos; their genomes are initially copies of each other but they don't remain perfectly identical throughout the life of the twins owing to mutations and epigenetic effects [11], i.e. differing environmental factors, starting even in the womb, that affect the expression of the genes. Still, for practical purposes, despite the existence of such differences, current forensic DNA tests have a hard time distinguishing between monozygotic twins, as dramatically highlighted in the 2007 court case summarized below.

Holly Marie Adams testified having sexual intercourse, in the same month, with both Raymon Miller and his identical twin brother Richard Miller. When a baby was born, Adams filed suit against the twins to determine the identity of the natural father and to obtain a declaration of paternity. Court-ordered tests stated that "The probability that [Appellant] is the biological father of [K.A.A.] is identical to the probability that [Richard] is the biological father." [12]

Lei [13] observed that "there is currently no commercially available test that can determine which of the twin brothers passed his DNA to the child even though there are ways in which the genomes of identical twins differ" but suggested that DNA samples be stored from both twins, to be used when such a test is available in the future.

Sequencing the entire genome of the two brothers may identify a mutation that occurred at the time the embryo split in two. So genomics has more chances than simple genetic testing of providing a solution to the case.

Our opinion is that the availability of ever more discriminating technical tests may draw attention away from the ethical issues. In a case like the one above, for example, assuming that none of the three people involved had any way of knowing, short of a genomic test, whether fertilization occurred during intercourse with one or the other twin, why should one of the twins be any less responsible for the fatherhood than the other?

Another fundamental kin privacy issue in forensic genomics, already anticipated in section 2, is the technical possibility of finding culprits, who are not in the DNA database, through partial matching against the DNA of a relative who is in the database, as lucidly explained by Bieber et al. [15]. They report of two murder cases from the 1980s, one from the USA and one from the UK, that were solved in the 2000s by matching scene-of-crime samples against the respective countries' DNA databases; in both cases a close but not exact match was found—with the nephew of the British killer and the brother of the American one. The relatives of the matching individuals were investigated until the culprits confessed. Here we stress that the transition from forensic DNA (as in [15]) to forensic genomics will extend enormously the power of identifying distant relatives.

There is a clear tension here between bringing criminals to justice and willingly deploying a system that relies on systematic investigation of innocents in order to catch the guilty. It is also a concern that you can be genetically investigated even if you have done nothing to end up in the DNA database—all that is needed is for one of your relatives to have been "sampled" once, perhaps just for a speeding offence. Of particular worry is the circumstance that, even if you live in a jurisdiction with strong protection for genetic privacy, you might be framed by investigations on distant family members who years ago emigrated to a country such as the UK where the national DNA database is aggressively filled up with samples from almost anyone who gets stopped by the police.

There are important kinship and population-related aspects of the personalised genome medicine. Human populations have different propensities/susceptibilities for many genetic diseases. If budget has to be allocated for research on a genetic disease, certain populations may get more benefit than others [18, 19]. It may also be embarrassing to belong to the group more susceptible to neurological diseases or genetically-inherited behavioural disorders. Therefore, although the genomic test is at the level of an individual, the consequences may be at ethnic level. Note that if you are susceptible for a genetic disease you may be required to prove you have not got the disease even if you do not show any symptoms.

5.3 Cold and hot hits

The statistics behind forensic genomics are sufficiently subtle and complex that even the experts have long held conflicting opinions on how to interpret them correctly. (This perhaps says something about the difficulty of the task faced by members of the public called upon to serve in a jury.)

An important qualitative distinction is between cold and hot hits. A hot hit is what you get when some non-genomic evidence already points at a particular suspect and then a sample from that specific suspect is checked against a crime scene sample to determine whether the two match. A cold hit, instead, is when you match the crime scene sample against all the people in your database, regardless of whether they are suspected of anything or not. Using Zimmermann's cited metaphor (section 4), it's the difference between fishing with hook and line and fishing with a driftnet.

Statistics has often been presented as a science that can be bent in different directions depending on the circumstances and opportunities. Two quick examples: one, Bayesian reasoning has been

banned from courts (as in the Regina vs Adams case [50]); two, the different conclusions reached by Bayesians and frequentists in the O.J. Simpson case [46] have been considered as a failure of statistics.

Here we should distinguish the statistical treatment for hot and cold hits. In case of hot hit we may carry out standard likelihood ratio (LR) testing of the hypothesis that the suspect is the source of a stain found on the scene of crime [5]. We give here a brief explanation of LR, suggesting that the reader consult the references [5, 9, 16, 17] for the detailed mathematical theory. LR testing is a powerful test in which competing hypotheses H_1 (the suspect is the source of the stain) and H_0 (its complement) are compared using a statistic based on the ratio of the maximum likelihoods (l_0, l_1) under each hypothesis; for example, $2\delta = 2\ln(l_1/l_0)$ which follows a chi-square distribution. Results can be expressed in terms of p-values, the probability of the statistic being at least as extreme as observed when H_0 is true: low p-values (e.g. < 0.05 in standard statistical practice) suggest rejection of H_0 in favour of H_1.

In case of cold hit search, initially the American National Research Council (NRC) hypothesized that using a larger database made you more likely to mistakenly identify an innocent as a culprit ("the np rule", where n is the size of the database and p the probability of a match) [10]. But one striking aspect is that the database search not only points to a suspect but also eliminates as possible culprits all the other persons in the database because their DNA profile differs from that of the crime sample. Two British statisticians, Balding and Donnelly, have provided statistical support for this effect [5, 9, 16], showing that the DNA evidence is somewhat stronger when the suspect is identified by DNA database matching first (cold hits) than when identified by non DNA evidence (hot hits) and subsequently found to match the profile of the crime sample. It is the effect of ruling out others which makes the DNA evidence stronger after a database search. If we consider a database comprising the entire population of the world, a unique match would indicate that we found the criminal.

We stress here that, if personal genomic information were available, the accuracy of the identification of the suspect and of ruling out others would be much greater than in the current situation. The resolving power of genomics would allow the search for the culprit to be restricted quickly to certain ethnic groups and populations. This power would defeat any argument in favour of the "np rule" of the NRC.

On the other hand we should be aware of the privacy risks before handing over to the police the enormous resolving power of genomics.

5.4 Regulatory issues in international law

Several international instruments already prohibit any discrimination based on genetic data [23, 24, 25]; the Convention on Human Rights and Biomedicine (the Oviedo Convention) furthermore allows the carrying out of predictive genetic tests for medical purposes only.

The US Genetic Information Nondiscrimination Act (GINA) will similarly prohibit price differentiation and discrimination based on genetic data by insurers and employers, when it comes into force (probably later in 2008) [26].

Data protection law in Europe also requires strong protection of genetic data, as emphasised by the EU's "Article 29 Working Party" in its 2004 Working Document on the issue [27], and in other related documents [28, 29]. These stress inter alia that, in the WP's view:

- genetic data are "particularly sensitive" personal data; even as yet unmatched DNA samples (and the collection of such samples) should be regulated in accordance with data protection law and principles;
- genetic (and biometric) data pose special risks in respect of possibilities for unwarranted linking or "matching" of data in different databases[12], the more so since they can be surreptitiously obtained;
- to counter this, the purpose-specification and limitation principle should be very strictly applied, and the use of genetic data must be subject to strict tests of necessity and proportionality;
- special, clear and precise rules and effective procedures are needed to regulate the use of genetic data[13]; this should include special "prior checks" by the data protection authorities before the establishment of any genetic database is permitted;
- fully free and informed explicit consent is required for any diagnostic or predictive genetic tests for medical purposes;
- the processing of genetic data in the field of employment should be prohibited in principle, with only extremely rare exceptions (and no exceptions for predictive use of genetic information); the processing of genetic data in the field of insurance should be prohibited in principle and only authorised under really exceptional circumstances, clearly provided for by law; and
- the blanket implementation of mass genetic screening is unlawful.

However, in many respects the Art. 29 WP still only identified issues and questions, without providing conclusive answers. This happened, for instance, with:

- the question of whether a person may be forced to disclose his/her genetic data to blood relatives, where such data are relevant in view of safeguarding their health;
- the exercise of the right, inside a group, not to know one's genetic data; and
- bio-banks. In this respect, the WP mentioned, on the one hand, that "the issue of prescribing practices applying anonymisation could be a possibility to address issues from the data protection perspective." However, it then also noted that "there has been evidence that stored DNA is capable of being linked to a particular person—provided certain additional knowledge is available, even though it may not be stored in a directly person-specific way."

It would appear that, in some respects, the WP and the national data protection authorities are still not fully informed of the scientific facts, or not fully equipped to understand them and their

[12] They insightfully remark [29] that "The centralised storage of biometric data also increases the risk of the use of biometric data as a key to interconnecting different databases that could lead to detailed profiles of an individual's habits both in the public and in the private sector. Moreover, the question of compatible purpose raises the issue of interoperability of different systems using biometrics. The necessary *standardisation for interoperability could lead to greater interlinking* between databases." (emphasis added).

[13] This flows from wider human rights law [30], and in particular the European Convention on Human Rights.

implications in this field. Cf., e.g., the WP's comment that "templates [generated by biometrics] and their digital representations", if "processed with mathematical manipulations (encryption, algorithms or hashfunctions), using different parameters for every biometric product in use", will "avoid the combination of personal data from several databases through the comparison of templates or digital representations." [29, p. 10]. This overestimates the capacity of such techniques (manipulations) to prevent re-identification: secure anonymization is notoriously an unsolved research problem in information security.

Moreover, the EU's data protection directive and national laws contain many vague and open-ended provisions and are often not fully enforced [31]. There are loopholes and dubious practices—including practices (e.g., on data sharing) that are actually encouraged by (some) governments [32]. There have for instance been calls, in particular in the UK, for a comprehensive DNA register on everyone in the country (or at least known to be in the country)[14]. These calls are directly contrary to the principle laid down by the WP, set out in the final point in the first of the two bulleted lists above.

Going in the opposite direction, on 27 May 2005 the Prüm Treaty was signed by Germany, Spain, France, Luxembourg, Netherlands, Austria and Belgium[15]. It covers a series of justice and home affairs issues including the "exchange of information". For example, Articles 2–12 allow direct access by the law enforcement agencies in the participating states to each other's databases on DNA, fingerprints and vehicle registration, on a "hit/no-hit" basis. If there is a "hit" the file is provided. Indeed, the Treaty requires the establishment, in the participating States, of certain databases, including a DNA database, and imposes a duty on participating States to obtain DNA from "particular individual[s]" (note: not necessarily suspects) if no DNA is available in the national database. Terrorism is explicitly included in the remit. In June 2007, the Council agreed to integrate the main provisions of the Prüm Convention into the EU's legal framework, to enable wider exchanges between all EU Member States of biometric data (DNA and fingerprints) in the fight against terrorism and cross border crime. All EU Member States will therefore be required to set up DNA databases. Note that the process under which arrangements between small groups of countries are subsequently extended to all, without proper debate, has been strongly criticised.

In summary, although the broad parameters for regulation are known and turn around established principles of human rights, data protection and ethics, there is clearly still a dire need for clear, precise, yet workable regulation in this area. Without it, fundamental rights and basic ethical standards are seriously at risk.

6. CONCLUSIONS

As the adoption of DNA in forensic contexts grows rapidly, some countries (notably the UK) are building up very large DNA databases. But this is happening without an informed and widespread debate.

Many members of the public are simply misinformed and do not understand what is at stake: the DNA database is considered as "just another intrusive government database"[16], and it is. But there is a fundamental, qualitative difference between genomics and the other biometrics used for identification: ignoring the fact that the genome of an individual contains enormous amounts of private medical and ancestry-related information about its owner and his or her family leads to a grossly inaccurate and dangerous underestimation of the privacy problems involved.

We stress that, while genomic testing may bring great benefits to medicine, we would face great civil liberties threats if police were to upgrade from their current CODIS (or similar sets of markers) to genomics, which are much more privacy-invasive. Given that there is an increasing awareness of the genomic differences between humans, it is important that ethnic-specific susceptibilities not be included in any police-linked based database, which should contain a similar number of people for each ethnic group.

The technical issue we highlighted about hot and cold hits is also at the core of the civil liberties problem: while on one hand the driftnet technique can be shown to yield more accurate results, on the other hand we feel it is inappropriate and unfair for an honest citizen to have their genome forensically inspected even when there is no evidence whatsoever of them having committed a crime.

Is GINA enough? GINA particularly focuses on inappropriate use of genetic information in health insurance and job recruitment. But it is also necessary to restrict the circulation of genomic information of individuals only to authorised personnel, and to ensure that what is collected for one purpose (e.g. medicine) can't be used for another (e.g. forensic investigation). Therefore we see GINA as a first step towards the recognition of the fact that genome information is a fundamental inalienable aspect of the dignity and privacy of each of us and of our roots and genetic history. An informed public debate is necessary and this paper is our contribution towards raising awareness and understanding of the core issues to be discussed.

[14]The press reported on a senior judge calling for a national DNA database [33] and on a senior police officer making a similar call [34]. These calls have so far been rejected by the Government [35]. For a general, critical discussion and extensive information, see the Genewatch website [36].

[15]Other countries joined later: Italy, for example, did on 30 October 2007.

[16]As would be the ones of fingerprints or iris codes that are also being built under the excuse of spotting terrorists at border control.

Acknowledgements

Pietro Liò acknowledges funding from the EC's IST SOCIALNETS project (http://www.social-nets.eu).
We are grateful to Alastair Beresford, Mike Roe and Markus Kuhn for their comments.

7. REFERENCES

[1] "New Grants Drive Development Of Rapid, Cost-Effective Sequencing Technologies", Medical News Today, 2008-08-23, http://www.medicalnewstoday.com/articles/118963.php

[2] http://www.iht.com/articles/2007/06/01/america/dna.php

[3] http://www.nytimes.com/2007/12/12/science/12watson.html

[4] Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K. "Identification of the remains of the Romanov family by DNA analysis". *Nat Genet.* 1994 Feb;6(2):130-5.

[5] Balding, D. J. and Donnelly, P. (1995). "Inference in forensic identification". *Journal of the Royal Statistical Society*, Series A 158, 21–53.

[6] BBC News. "Germany adopts Stasi scent tactic". 2007-05-23. http://news.bbc.co.uk/1/hi/world/europe/6683803.stm

[7] Philip Bohannon, Markus Jakobsson, Sukamol Srikwan. "Cryptographic Approaches to Privacy in Forensic DNA Databases". In *Public Key Cryptography 2000*, pp. 373–390. Jan, 2000.

[8] Helm, T.: "Outrage at 500,000 DNA database mistakes". *Telegraph*, 2007-08-27

http://www.telegraph.co.uk/news/main.jhtml?xml=/news/2007/08/27/ndna127.xml

[9] Balding D., Donnelly P. (1996) "Evaluating DNA Profile Evidence When the Suspect Is Identified Through a Database Search". *J Forensic Sci* 41, 603–607.

[10] National Research Council (NRC). "The evaluation of forensic DNA evidence". Technical report, National Academy Press, Washington D. C., 1996.

[11] Singh SM, Murphy B, O'Reilly R. "Epigenetic contributors to the discordance of monozygotic twins". *Clin Genet.* 2002 Aug;62(2):97-103.

[12] Phillip R. Garrison (Judge), Adams vs Miller and Miller, Case number 27188, 2007-03-14, Missouri Court of Appeals Southern District, http://www.courts.mo.gov/Courts/PubOpinions.nsf/8e937ac7ce0301288625661f004bc963/67393c1c232272598625729e0075ff81?OpenDocument.

[13] Hsien-Hsien Lei, "Genetic Differences Between Identical Twins", http://www.eyeondna.com/2008/02/page/2/, 2008-02-20.

[14] Testimony of Philip R. Zimmermann to the Subcommittee on Science, Technology, and Space of the US Senate Committee on Commerce, Science, and Transportation. 1996-06-26. http://www.philzimmermann.com/EN/testimony/index.html

[15] Frederick R Bieber, Charles H Brenner and David Lazer, "Finding Criminals Through DNA of Their Relatives", *Sciencexpress*, 2006-05-11.

[16] Donnelly, P. and Friedman, D. "DNA database searches and the legal consumption of scientific evidence", *Michigan Law Review* 97, 931–984, 1999.

[17] Storvik, G. and Egeland, TB. "The DNA database search controversy revisited: Bridging the Bayesian - Frequentistic gap". *Biometrics* 63 pp 922–925, 2007.

[18] Pan Q, Luo X, Chegini N. "Genomic and proteomic profiling I: leiomyomas in African Americans and Caucasians". *Reprod Biol Endocrinol.* 2007 Aug 23;5:34.

[19] Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF. "Ethnic-difference markers for use in mapping by admixture linkage disequilibrium". *Am J Hum Genet.* 2002 Mar;70(3):737-50. Epub 2002 Feb 11.

[20] Collins T, "Loss of 1.3 million sensitive medical files in the US", *ComputerWeekly*, July 2007, http://www.computerweekly.com/blogs/tony_collins/2007/07/loss-of-13-million-sensitive-m.html.

[21] BBC, "Thousands of driver details lost", 2007-12-11, http://news.bbc.co.uk/1/hi/northern_ireland/7138408.stm.

[22] B. Heffernan and E. Kennedy, "Alert as 170,000 blood donor files are stolen", 2008-02-20, http://www.independent.ie/national-news/alert-as-170000-blood-donor-files-are-stolen-1294079.html.

[23] Council of Europe, European Convention on Bio-medicine (ETS 164, the "Oviedo Convention"), Art. 11.

[24] EU Charter of Fundamental Rights, Art. 21.

[25] UNESCO, "Universal Declaration on Human Genome and Human Rights", Article 6.

[26] "US House Passes Genetic Discrimination Bill", *Genome Web News*, New York, 1 May 2008.

[27] Article 29 Working Party (WP), "Working Document on Genetic Data" (WP 91 of 17 March 2004).

[28] Article 29 Working Party (WP), "Opinion 6/2000 on the Genome Issue" (WP 34 of 13th July 2000)

[29] Article 29 Working Party (WP), "Working Document on Biometrics" (WP 80 of 1 August 2003).

[30] D. Korff, "The need to apply UK data protection law in accordance with European law", *Data Protection Law & Policy*, May 2008.

[31] D. Korff, "Study on Implementation of Data Protection Directive – Comparative Summary of National Laws", Study for the EC Commission, 2003.

[32] R Anderson, I Brown, R Clayton, T Dowty, D Korff, E Munro. "Children's Databases – Safety and Privacy" (2006), FIPR study for the UK Information Commissioner.

[33] BBC, "All UK must be on DNA database", BBC News, 5 September 2007 http://news.bbc.co.uk/1/hi/uk/6979138.stm

[34] "Britain needs DNA database, says officer who headed Sally Anne murder inquiry", *Times Online*, 22 February 2008.

[35] "Mandatory DNA database rejected", BBC News, 23 February 2008 http://news.bbc.co.uk/1/hi/uk/7260164.stm.

[36] Genewatch, "The UK Police National DNA Database" http://www.genewatch.org/sub-539478.

[37] "The 100 dollars Genome". *Technology Review*, published by MIT, April 17, 2008. http://www.technologyreview.com/Biotech/20640/page1/

[38] P. Kuzniar, E. Jastrzebska , R. Ploski. "Validation of nine non-CODIS STR loci for forensic use in a population from Central Poland". *Forensic Science International*, Volume 159 , Issue 2–3 , Pages 258–260.

[39] J S. Barnholtz-Sloan, R Chakraborty, T A. Sellers and A G.Schwartz. "Examining Population Stratification via Individual Ancestry Estimates versus Self-Reported Race". *Cancer Epidemiology Biomarkers and Prevention* Vol. 14, 1545–1551, June 2005

[40] Harris MA et al., Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". *Nucleic Acids Res.* 2004 Jan 1;32(Database issue):D258-61.

[41] Ashburner M, Lewis S. "On ontologies for biologists: the Gene Ontology–untangling the web". *Novartis Found Symp.* 2002;247:66-80; discussion 80-3, 84-90, 244-52.

[42] Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. "Percolation of annotation errors through hierarchically structured protein sequence databases." *Math Biosci.* 2005 Feb;193(2):223-34.

[43] Jacoby E. "Chemogenomics: drug discovery's panacea?" *Mol Biosyst.* 2006 May;2(5):218-20. Epub 2006 Mar 30.

[44] "Simpson lawyers switch emphasis to police errors". *USA Today* http://www.usatoday.com/news/index/nns133.htm

[45] "What is the chance of your being guilty?" *Financial Times* 19 June 2003. http://www.johnkay.com/society/287

[46] "Evaluating Legal Evidence". Department of Computer Science, Queen Mary University 9 May 2008. http://www.dcs.qmul.ac.uk/researchgp/spotlight/legal.html

[47] Levinson DF, Holmans P. "The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees." *BMC Genet.* 2005 Dec 30;6 Suppl 1:S6.

[48] Gasbarra D, Pirinen M, Sillanpää MJ, Arjas E. "Estimating genealogies from linked marker data: a Bayesian approach." *BMC Bioinformatics.* 2007 Oct 25;8:41.

[49] Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M, Oppenheim A. "The Y chromosome pool of Jews as part of the genetic landscape of the Middle East." *Am J Hum Genet.* 2001 Nov;69(5):1095-112.

[50] R v Doheny and Adams [1997] 1 Crim App R 369.