Preservation Of DNA Privacy During The Large Scale Detection Of Covid-19

Marcel Hollenstein, David Naccache, Peter B Rønne, Peter y A Ryan, Robert Weil, Ofer Yifrach-Stav

PRESERVATION OF DNA PRIVACY DURING THE LARGE SCALE DETECTION OF COVID-19

Marcel Hollenstein[4], David Naccache[1,3], Peter B. Rønne[2], Peter Y A Ryan[2], Robert Weil[5], and Ofer Yifrach-Stav[1]

[1] ÉNS (DI), Information Security Group, CNRS, PSL Research University, Paris, France
david.naccache@ens.fr , ofer.friedman@ens.fr
[2] SnT & University of Luxembourg, Luxembourg.
peter.roenne@uni.lu , peter.ryan@uni.lu
[3] School of Cyber Engineering, Xidian University, Xi'an, 710071, PR China
david@xidian.edu.cn
[4] Institut Pasteur, Laboratory for Bio-organic Chemistry of Nucleic Acids
Department of Chemistry and Structural Biology, Paris, France.
marcel.hollenstein@pasteur.fr
[5] Sorbonne University, Institut National de la Santé et de la Recherche Médicale (UMR1135), CNRS (ERL8255), Centre d'Immunologie et de Maladies Ifectueuses CIMI, Paris, France
robert.weil@upmc.fr

Abstract As humanity struggles to contain the global COVID-19 pandemic, privacy concerns are emerging regarding confinement, tracing and testing. The scientific debate concerning privacy of the COVID-19 tracing efforts has been intense, especially focusing on the choice between centralised and decentralised tracing apps. The privacy concerns regarding COVID-19 testing, however, have not received as much attention even though the privacy at stake is arguably even higher. COVID-19 tests require the collection of samples. Those samples possibly contain viral material but inevitably also human DNA. Patient DNA is not necessary for the test but it is technically impossible to avoid collecting it. The unlawful preservation, or misuse, of such samples at a massive scale may hence disclose patient DNA information with far-reaching privacy consequences.
Inspired by the cryptographic concept of "Indistinguishability under Chosen Plaintext Attack", this paper poses the blueprint of novel types of tests allowing to detect viral presence without leaving persisting traces of the patient's DNA.
Authors are listed in alphabetical order.

Acknowledgements
This work was supported by the Luxembourg National Research Fund (FNR) project SmartExit (14729565).
The authors thank Thibaut Heckmann, Head of Data Extraction Unit, and Audrey Gouello, DNA Expert, of the IRCGN (Institut de recherche criminelle de la gendarmerie nationale) for their enriching remarks.

1 Introduction and motivation

1.1 Privacy issues related to COVID-19

The current COVID-19 (Coronavirus Disease 2019) pandemic is rapidly spreading, significantly impacting healthcare systems. The disease is caused by the novel corona virus, also called SARS-COV-2 [Gor20]. SARS-COV-2 is a positive-sense single-stranded RNA virus. Stay-at-home and social distancing orders enforced in many countries are supporting the control of the disease's spread, while causing turmoil in the economic balance and in social structures [BBD+20]. Recently, we witness trends of resistance to abiding by lock-down policies, and particularly to obligatory mask-wearing in public [Pra20]. While in the United States some citizens claim that mandatory mask wearing is a violation of civil rights [And20], more than 70 countries have declared a "State of Emergency", increasing the governments' power. In some countries, such as Thailand [TNT20], Spain [KMG20] and Canada [Wil20] the state of emergency has been extended. In the attempt to control the virus' spread, mobile software applications (tracing apps) have been developed. These apps use digital tracking that monitor contact between individuals, so as to easily identify possible exposure to the virus. Some of these apps are based on tracking the geographical location of app users, thus raising privacy concerns.

The scientific debate concerning privacy of the COVID-19 tracing efforts has been intense, especially regarding the choice between centralised and decentralised tracing apps [Vau20], and it has had political implications, such as Germany changing to a decentralised approach [Sch20b]. Oddly, the privacy concerns regarding COVID-19 testing, however, .have not received as much attention even though the privacy at stake is arguably even higher, potentially compromising the privacy of one's DNA.

1.2 Testing for COVID-19 Infection

Rapid detection of cases and contacts is an essential component in controlling the pandemic's spread. In the US, the current estimation is that at least 500,000 COVID-19 tests will need to be performed daily to successfully reopen the economy [LVLM20].

There are currently two types of COVID-19 tests:

- *Molecular diagnostic tests* that detect the presence of SARS-COV-2 nucleic acids in human samples: A sample is taken using a narrow swab that is placed in the patient's nose or mouth. It is generally recommended to collect upper respiratory samples (nasopharyngeal swabs, oropharyngeal swabs, nasopharyngeal washes, and nasal aspirates), but when the patient exhibits productive cough, lower respiratory samples (sputum, BAL fluid, and tracheal aspirates) are sometimes used [UKK+20]. Polymerase chain reaction (PCR) is a process that causes a very small well-defined segment of DNA to be amplified, or multiplied many hundreds of thousands of times, so that there is enough of it to be detected and analyzed. Since the SARS-COV-2 virus does not contain DNA but only RNA, reverse transcription is used to convert the extracted RNA into DNA. The resulting DNA product is then subjected to a real-time PCR to determine whether sequences corresponding to the SARS-COV-2 RNA are present in the sample. The amplification of DNA is monitored in real time as the PCR reaction progresses using a fluorescent dye or a sequence-specific DNA probe labeled with a fluorescent molecule and a quencher molecule. The process is repeated for about 40 cycles until the viral cDNA can be detected [CGS+20].
- *Serological diagnostic tests* that identify antibodies to SARS-COV-2 in clinical specimens [WKLT20]. The current standard test for COVID-19 detection, qPCR is quick, sensitive and reliable, but can only tell if a person is currently infected. Detecting antibodies to SARS-COV-2 can tell a clinician if a patient has been infected with COVID-19 either currently, or in the past, depending on the type of immunoglobulins detected (IgM, IgG or both), and if the patient is immunized and thus protected against a second infection. In addition, identifying populations who have antibodies can facilitate research on the use of convalescent plasma in the development of a cure for COVID-19 [FA20].

1.3 The DNA Privacy Problem

In both types of tests, the collected specimen contains the tested person's DNA. DNA is the molecule that carries the genetic instructions of all living organisms. Screening of vast populations for the presence of the virus, inevitably means providing the testing agencies (clinics, governments, airlines, etc.) with sensitive genetic information on a considerable number of individuals. Even if the sample contains a very small amount of DNA, PCR can be used to amplify the DNA and reveal the genome [KYK+11]. A USB portable device, the MinION, developed by Zaaijer et al. [ZGS+17] can accurately identify human cells ("DNA re-identification") by comparing an unknown DNA sample to a collection of known DNA profiles, with 99.9% confidence, within three minutes of DNA sequencing.

DNA samples collected as part of COVID-19 tests are not supposed to be analyzed, and in any case, sensitive medical information is expected to be kept confidential. However, it is common practice to preserve medical samples to be used in further research or for prognosis monitoring.

Often times, tested individuals do not know how these samples will be used in the future. For example, in 2009, it was discovered that Texas had been collecting and storing blood and DNA samples taken from millions of newborns without the parents' knowledge or consent. These samples were used by the state for genetic experiments and for the set up of a database [Wal10].

DNA databases, or *biobanks*, are being maintained in many countries in the world [Wil18] and they are being used for forensic [KDK11] and research purposes. An analysis from 2012 indicates that there is no consensus on the the need for consent to use information in biobanks [MNMC12]. Despite the matter's sensitivity, the information can be accessed. In a research conducted in 2016, 95.7% of 46 biobanks surveyed by [CAD+16] gave other researchers permission to access their samples. Despite laws and regulations intended to prohibit the re-identification of anonymized data, such as Privacy Rule from HIPAA (Health Insurance Portability and Accountability Act of 1996), the Common Rule from the DHHS (Department of Health and Human Services), and the Human Subject Protection Regulations and 21st Century Cures Act from the FDA [Lee00], there have been numerous incidents of data breaches (including database hacks and ransomware attacks) in healthcare systems [Pec17, Mon17, CBC16].

DNA samples stored in databases are usually coded so as to reduce their identifiability. However, it is not impossible to track down the individual "behind" the DNA. For example, Malin and Sweeney (2000, 2001) [MS00, MS01] demonstrated that even if kept confidential, and without being linked to identifying personal or demographic details, DNA information can be traced to the tested patients using inferences drawn from the DNA information.

The fact that stored DNA can be linked to the person is especially problematic given the possible use there could be for this information. Knowing a person's DNA provides information about racial features, potential diseases or life span expectancy. This information can attribute to genetic discrimination. For example, an employer may refuse to hire someone based on the likelihood that they will become ill. Similarly, exposing one's genetic information may impact their eligibility to life or health insurance, or to increase their premia [BC+98]. In this sense, the potential transfer of sensitive genetic information to a third party raises significant ethical and legal issues [CT04]. Moreover, the collection of DNA into databases can *"raise human rights concerns, including potential misuse of government surveillance (for example, identification of relatives and non-paternity) and the risk of miscarriages of justice"* [Can16].

Beyond the discrimination against individuals based on their genetic profiles, human rights activists have been protesting against the mass collection of DNA samples from citizens by governments. In the past years, China has been collecting DNA samples from citizens as part of mandatory medical examinations [Fen17]. Human Rights Watch activists are worried about the use of this information for *"surveillance of persons because of ethnicity, religion, opinion or other protected exercise of rights like free speech"* [Haa17].

The gathering of DNA information by countries has raised concern regarding the way this information can be used to impact populations based on genetic characteristics. Moreover, this information may be deployed not only domestically, but internationally [Mos19a]. DNA information can be used to attack strategically identified persons, such as diplomats, politicians, high-ranking federal officials, or military leadership, or even to bio-engineer a disease that would be fatal to some races but not to others [Mos19b].

Refusal to undergo medical tests or procedures because of the fear of exposing genetic information to hostile entities may hinder medical and scientific attempts to treat diseases and learn about them for the sake of mankind. Recently, for example, it has been reported that Israel revoked a deal with a company selling COVID-19 testing equipment out of concern about granting the company access to Israelis' genetic information [Sch20a]. Especially now, when COVID-19 tests are prevalent and may become mandatory in different settings (e.g. at airports) [Cri20], the need to find a way to conduct these tests without compromising our genetic information's safety arises. The fact that test samples are often sent to be analyzed in another country, e.g. [BBC20], reinforces the need to form a testing method that ensures that the samples sent do not contain identifiable DNA traces.

The rest of the paper is organized as follows: in the next section, we describe a theoretical safety model inspired by the cryptographic notion of Indistinguishability under Chosen Plaintext Attack. In Section 3, we describe a testing scheme applying the theoretical safety model to create privacy preserving medical tests, and elaborate on different approaches that could be applied. Finally, Section 4 concludes the paper.

2 Suggested Solution - Theoretical Model

Before describing the solution, let us provide the intuition behind our idea.

We aim to develop a DNA privacy-preserving test, or in other words, a test method that would yield the same results, but that the DNA in the specimen tested, or in any residue of the specimen collected, will be undetectable. We hence wish to develop a test procedure $\mathscr{T}$ such that:

$$\mathscr{T}(\text{DNA} \sharp V) = \{\mathsf{positive}, \mathsf{res}_{\mathsf{positive}}(\text{DNA}, V)\} \text{ and } \mathscr{T}(\text{DNA}) = \{\mathsf{negative}, \mathsf{res}_{\mathsf{negative}}(\text{DNA})\}$$

The positive and negative denote the virus' presence or absence, respectively (the test's result). $\sharp$ denotes mixing substances[6]. res denotes the test's residue, i.e. whatever is left after the test procedure has been completed. While some protocols may require this residue to be treated as bio-hazardous waste and be destroyed, residues are often preserved for future research or other purposes. Therefore, res is where unwanted DNA may be present, and is our main concern.

In COVID-19 tests, the sequence of operations is independent of the virus' presence. However, in other types of tests, the virus' presence may influence the sequence of operations done during the test, hence in all generality we distinguish two types of residues in our model. In any case, the residue will differ by the remains of the virus (or lack thereof).

The attacker's definition ($\mathscr{A}$) is inspired by the cryptographic notion of Indistinguishability under chosen-plaintext attack (IND-CPA):

$\mathscr{A}$ selects two DNA samples DNA_0 and DNA_1 and submits them to a challenger $\mathscr{C}$. $\mathscr{C}$ picks a random $b \in \{0,1\}$ and manufactures the samples:

$$\sigma_{b,\text{positive}} = \text{DNA}_b \sharp V \text{ and } \sigma_{b,\text{negative}} = \text{DNA}_b$$

$\mathscr{C}$ runs $\mathscr{T}$ on $\sigma_{b,\text{positive}}, \sigma_{b,\text{negative}}$ and gets:

$$\mathscr{T}(\sigma_{b,\text{positive}}) = \{\mathsf{positive}, \mathsf{res}_{\mathsf{positive}}\} \text{ and } \mathscr{T}(\sigma_{b,\text{negative}}) = \{\mathsf{negative}, \mathsf{res}_{\mathsf{negative}}\}$$

$\mathscr{A}$ gets $\{\mathsf{positive}, \mathsf{res}_{\mathsf{positive}}\}, \{\mathsf{negative}, \mathsf{res}_{\mathsf{negative}}\}$, performs any state of the art analyses and outputs a guess b'.

$\mathscr{A}$'s advantage is defined as:

$$Adv = 2|\Pr[b = b'] - \frac{1}{2}|$$

When $\mathscr{A}$ has no advantage in learning DNA information, his only strategy is to guess b' at random. In that case, $\Pr[b' = b] = \frac{1}{2}$ and hence $Adv = 0$.

When $\mathscr{A}$ always correctly determines b', we have $\Pr[b' = b] = 1$, i.e. $Adv = 1$.

Note that when $\mathscr{A}$ is always wrong, we have $\Pr[b' = b] = 0$ and hence $Adv = 1$. Such an $\mathscr{A}$ is effectively as powerful as an $\mathscr{A}$ who always finds the correct answer as it suffices to negate his response to get a perfect adversary.

In other words, $0 \leq Adv \leq 1$. The higher the advantage, the more powerful $\mathscr{A}$ is.

We define $\mathscr{T}$ as IND-C.DNA.A[7] secure[8] if $Adv < 10^{-3}$.

3 How to Build IND-C.DNA.A Tests?

We assume that the test procedure needs to be built in such a way that the tested person can trust that once the specimen is collected, the DNA will not be identifiable. Therefore, we suggest to add a testing procedure, step $\mathscr{T}_0$, which will be performed in the patient's presence. We suggest two approaches in which this step can be done: Mixing and Destroying.

[6] e.g. water $\sharp$ CO_2 = soda

[7] Indistinguishable under Chosen DNA Attack.

[8] The limit can be changed according to the test's acceptable level.

3.1 DNA Mixtures

We suggest to use a testing kit containing a mixture m of DNA samples, thus making it more difficult to analyze, or *profile*.

The complexity of a DNA mixture is determined by the number of people who contributed DNA to the mixture, the amount of DNA that each of them contributed, and the level of DNA degradation. More contributors make a mixture more complex, and therefore, more difficult to interpret [Pre19]. DNA profiling requires the comparison of short segments of DNA, called *alleles*, which vary from person to person. As part of the DNA profiling process, the DNA is amplified and the alleles are represented on a graph showing *peaks*. The positions of those peaks indicate which alleles are present, and thus the graph is a visual representation of the DNA in question. The DNA profiling task is based on the comparison of the pattern of those peaks. Small amounts of DNA derived from various contributors add "noise", called *drop-in*, which makes the comparison process more complicated. The greater the number of contributors is, the more complicated the task of identifying which peaks go with which contributor. In addition, the PCR process during copying reaction by the DNA polymerase creates small peaks, called *stutter products*, which are sometimes the same lengths as PCR products. This can make the determination of whether a small peak is a real peak from a minor contributor or a stutter products.

We propose, therefore, to increase this complexity by adding a DNA mixture to the specimen collected. Any analysis performed will be done on the mixture, and not on the individual DNA sample, thus making it more difficult to profile the DNA.

There are four possible scenarios in attempting to identify the DNA in the mixture:

- (A) The DNA of the victim x is known, and the composition of the mixture m is also known. The challenge is to determine if x is in the mixture $m \sharp x$ or not.
- (B) The DNA of the victim x is known, but the composition of the mixture m is unknown. The challenge is to determine if x is in the mixture $m \sharp x$ or not.
- (C) The DNA of the victim x is unknown, but the composition of the mixture m is known. The challenge is to isolate the DNA of the victim, x.
- (D) The DNA of the victim x is unknown, and the composition of the mixture m is also unknown. The challenge is to profile (separate) all the $\|m\|+1$ DNAs in the mixture $m \sharp x$ so as to learn the DNA of the victim x with probability $\frac{1}{\|m\|+1}$.

As is the case in cryptology, the above scenarii could be generalized and refined. For instance, one may consider a scenario where $\mathscr{A}$ is allowed to perform v (potentially adaptive) experiments with different mixtures $m_0, \ldots, m_{v-1}$ and an identical target DNA x etc. Whilst interesting in theory, we did not consider such extensions very relevant to "real world" settings.

scenario	**victim DNA** x	**hiding mixture** m	**attacker's goal**
(A)	known	known	confirm x
(B)	known	unknown	confirm x
(C)	unknown	known	learn x
(D)	unknown	unknown	learn[9] x

Table 1. Attack and protection scenarii.

Scenarios (A) and (B) represent a situation where the DNA of the individual is already known, and the challenge is to authenticate its presence in the mixture. Authentication methods are commonly used in the forensics field, where DNA found in a crime scene is compared to that of a suspect. For example, Homer et. al [HSR+08] have demonstrated that it is possible to identify the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA.

[9] with probability $\frac{1}{\|m\|+1}$.

In this paper, we address the option of attempting to identify the DNA of the individuals in the mixture when they are unknown to the attacker (scenarii (C) and (D)). Identification methods are intended to reveal the identity of one contributor in a mixture. Currently, these methods are achieved by comparing DNA samples to known profiles in a database. We propose solutions to prevent the possibility of identifying the genetic profile of an individual by an attacker.

Before we proceed, we would like to introduce a subtle distinction between the equality relationship ($=$) in mathematics and the chemical relationship $\simeq$ consisting in comparing two molecular mixtures.

We denote by $a = b$ an exact equality between the chemical components a and b. However, $a \simeq b$ will denote the fact that a cannot be distinguished from b using current laboratory equipment with very high probability (e.g. 99%).

Dilution (scenario (C)): The idea behind this technique is to add the sample into a pre-prepared mixture containing other samples or other DNAs, thus making DNA less identifiable: mislead $\mathscr{A}$ by reducing his advantage, exploiting the difference between $=$ and $\simeq$. The most plausible way to do so consists in adding to $\mathscr{T}$, a fixed mixture of k (e.g. $k = 20$) human DNAs taken from existing DNA samples, or animal DNA. Adding the DNA sample to a fixed mixture of DNA would make the process of identification significantly more complicated. However, using a fixed mixture of DNA grants a few possible advantages to $\mathscr{A}$. First, if the composition of the mixture is known to $\mathscr{A}$, identification of the added sample would be a relatively simple task. Second, even without being familiar with the mixture's composition, the characteristics of the contributors need to be taken into consideration to avoid easy identification. For example, race factors can influence the ease with which a sample can be identified. Thus, using a fixed DNA mixture will make the distinguishing of the DNA of the victim x more complicated, but not impossible.

Randomizing (scenario (D)): Another workaround, frequently used in cryptology, consists in adding randomness to $\mathscr{T}$. The idea behind randomizing is using a random mixture of DNA into which the sample is added. By doing so, any fixed DNA defines the distributions of residues $D_{\text{positive,DNA}} = \{\mathsf{res}_{\text{positive}}(\text{DNA}, V)\}$ and $D_{\text{negative,DNA}} = \{\mathsf{res}_{\text{negative}}(\text{DNA})\}$ obtained by testing this specific fixed DNA over and over again using $\mathscr{T}$.

We design $\mathscr{T}$ in such a way that $\forall$ DNA, the following distributions are indistinguishable:

$$\mathsf{res}_{\text{positive}}(\text{DNA}, V) \sim \mathsf{res}_{\text{positive}} \quad \text{and} \quad \mathsf{res}_{\text{negative}}(\text{DNA}) \sim \mathsf{res}_{\text{negative}}$$

This mixture is not known to $\mathscr{A}$, and even the number of contributors comprising the mixtures varies. This prevents $\mathscr{A}$ from learning through repeated experimentation. Randomizing makes the profiling task more complicated, because $\mathscr{A}$ will have no prior knowledge about the DNA characteristics. In DNA profiling, this is referred to as lack of *Framework of Circumstances*, [FSR18], which is one of the factors that hinder DNA profiling.

The model here assumes that when m is manufactured, each individual test kit is randomized (e.g. by the addition of a different random assortment of DNA material) so that different tests of the same patient will yield residues that do not leak information about the patient's DNA.

Current profiling methods in use in forensics allow analyzing a mixture of DNA and determining the number of DNA samples mixed into the mixture. However, separating an individual DNA from a mixture of an unknown number of contributors of unknown DNA profiles is a much more complicated task. In order to profile complex DNA mixtures, a software is used for computing the probability distribution for the number of contributors [TBB14]. If the number of contributors is unknown, the computational load will be considerably higher. Therefore, the Randomizing method could be applied to mask DNA and achieve our goal.

Another way of masking the DNA in question by adding randomization to the solution is to add an *allelic ladder* directly to the sample solution. An allelic ladder is an artificial mixture of the common alleles present in the human population, and it is commonly used to identify alleles in genetic profiles by comparison with peaks.

3.2 Destruction (all scenarii):

Another way to ensure that an attacker cannot access the DNA is to destroy it, thus making it unidentifiable. We start by observing that IND-C.DNA.A cannot exist if $\exists \mathrm{DNA}_0, \mathrm{DNA}_1$ such that:

$$\mathsf{res}_{\mathsf{positive}}(\mathrm{DNA}_0, V) \not\approx \mathsf{res}_{\mathsf{positive}}(\mathrm{DNA}_1, V) \text{ or } \mathsf{res}_{\mathsf{negative}}(\mathrm{DNA}_0) \not\approx \mathsf{res}_{\mathsf{negative}}(\mathrm{DNA}_1)$$

Simply because $\mathscr{A}$ can run the test by himself and compare the resulting residue to the challenger's residue, it follows that $\mathscr{T}_0$ must destroy human DNA while allowing subsequent testing of V.

To ensure that no DNA traces remain in the sample we suggest to destroy the DNA, leaving the RNA intact for the test. We propose, therefore, a model in which all human DNA is destroyed before the rt-PCR amplification. This can be done by treating the samples with DNase, an enzyme that selectively degrades DNA [ST14]. DNase eliminates DNA from RNA preparations prior to sensitive applications, such as rt-PCR. Within 10 to 20 minutes [HFK96, Bio20] there will be no identifiable DNA. To ensure that viral RNA remains intact, the DNase then needs to be inactivated by inclusion of removal reagents [Amb] or by using heat inactivation [WRSS00].

Following this process, we suggest a verification of the DNase's effectiveness. This verification process can be done by inducing reaction causing a colour change which could be visible by the patient. One way of doing so could be by applying gold nano-particles (AuNPs) interconnected by DNA duplexes. Without DNase activity, the AuNPs tend to cluster, displaying a blue colour. When DNase is present, it cleaves these duplexes, causing them to spread, which leads to a colour change from blue to red [HCPT17, BPE+08, XHM07]. Another method that could be applied is using a fluorophore-quencher system [SZZ+13]. In the absence of DNase, the fluorophore is in close proximity of the quencher and hence no fluorescence is visible. In the presence of DNase, DNA is hydrolyzed and the fluorophore is free to circulate in the solution, creating a fluroscentic reaction which can be easily monitored and detected as an output signal. The patient can witness the change in colour and be convinced that all DNA has been removed. Once this step is complete, the patient is free to go, and the sample is ready for testing.

Before we conclude, we have to consider the case of mutually distrusting parties. In this scenario, the tester wants to ensure that the test provides accurate information (i.e. that the virus will be detected, if present), while the patient may not trust the tester or the test kit. Theoretically speaking, we could have the patients provide their own DNase. Putting aside the practical logistic difficulties and unlikelihood of this solution, this solution poses a risk of a patient intentionally using a chemical killing both DNA and virus (for example, in a scenario of being virus-free as a condition to board a flight or enter a country). On the other hand, if the DNase is provided by the testing agency, the patient may suspect that another chemical is used that simply emulates the colour change. To solve this dilemma, we can have the patient be sampled twice using a classical "cut-&-choose" approach. To both samples the DNase and supplementary chemicals are added. The tested person then chooses one of the samples randomly, and adds his own chemical to verify the presence of DNase. The other sample is then used for testing. This allows the tested person to detect a maliciously generated test kit with probability $\frac{1}{2}$, and of course these odds can be improved, but at the cost of multiplying the number of samples used.

To ensure the integrity of the process, a positive process control method could be integrated. This could be done, for example, by identifying, at the end of the process, a specific reagent or a a known human target which will be deposited at the beginning of the analytical process. The presence of the target at the end of the process will allow concluding that a negative result obtained is stemming from absence of the virus, and not from a malfunction in the the reaction or process.

Note that we could also audit test kits in general, but this relies on trusted third parties, or a public procedure.

4 Discussion

This paper described a methodology in which biological specimens containing DNA taken from patients can be processed while securing the safety and confidentiality of the DNA information contained in the specimen. Our model relieves the patient of the expectation to *trust* the testing entity.

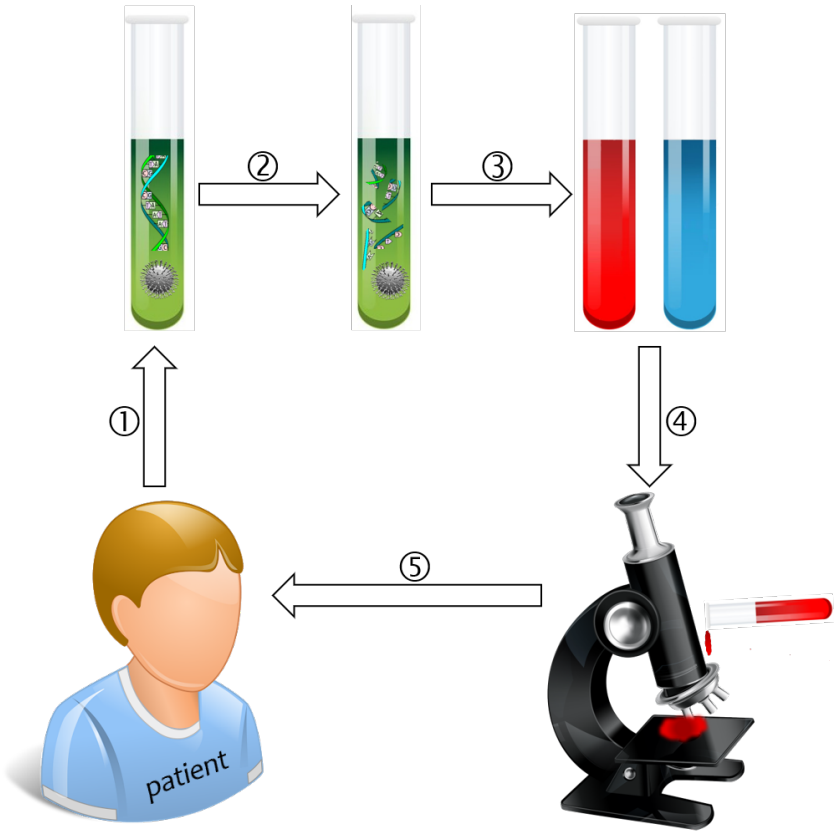

Figure 1. Testing following DNA destruction: ① Sample is taken from the patient using swab ⇒ ② DNase is applied to the sample, destroying DNA ⇒ ③ A colourimetric method is applied ⇒ ④ This demonstrates that no DNA traces are present (Blue indicates DNA traces, red means that DNA was properly destroyed) ⇒ ⑤ The patient is now convinced that the sample can be analyzed without risk of exposing his DNA.

While the medical system is still based on the patients' confidence and trust in the clinician, in the past years there has been a shift towards more informed patients expecting to have more involvement and control over processes and decision-making [RC06].

Mass COVID-19 testing performed all over the world nowadays highlights the needs for more secure types of tests. The model proposed in this paper can be applied not only for COVID-19 , but for other types of tests where DNA is extracted but not necessary to obtain test results. As DNA is present in any specimen collected from the human body, every lab test has the potential of exposing one's sensitive genetic information. The idea described in this paper will need to be adapted so as to provide protection to other types of tests.

While public awareness regarding the need to protect genetic information grows, the ability to perform successful profiling using smaller amounts of DNA increases. The recent developments in the field of DNA profiling now allows to analyze even minute amounts of DNA, called *trace DNA* or *touch DNA*. Small amounts of DNA can be found on any surface; people shed DNA on any object or surface they touch. In this sense, one may claim that protecting DNA information is impossible. However, it is important to note the difference between analyzing *trace amounts* of DNA, and analyzing the content of a test tube containing body fluids or mucosa. Let us consider a hypothetical case in which an attacker is interested in the DNA of a specific person. An attacker could try to retrieve trace DNA from objects touched by that person, a cup the person drank from etc., but the amount of DNA retrieved would be much smaller, and the process of analyzing this DNA would be significantly more difficult. In addition, it is important to note that since, in the case of medical tests, the specimen is sent for analysis in a lab, the likelihood of the DNA to be profiled increases, thus increasing the risk.

At this stage, we only offer a theoretical blueprint. Future work will include laboratory experiments demonstrating that the validity of the test is not negatively impacted by the added security phase $\mathscr{T}_0$ (i.e. DNA mixture or DNA destruction).

References

Amb. ThermoFisher Scientific Ambion. *Now it's easy to make your RNA free of genomic DNA contamination and ready for RT-PCR*. https://tinyurl.com/COVID19-Amb (accessed June 10, 2020).

And20. Scottie Andrew. *The psychology behind why some people won't wear masks*, 2020. https://tinyurl.com/COVID19-And20 (accessed June 18, 2020).

BBC20. BBC. *Coronavirus: UK sent 50,000 Covid-19 samples to US for testing*, 2020. https://www.bbc.com/news/uk-52603566 (accessed June 10, 2020).

BBD+20. Scott Baker, Nicholas Bloom, Steven J Davis, Kyle Kost, Marco Sammon, and Tasaneeya Viratyosin. The unprecedented stock market reaction to COVID-19. COVID *Economics: Vetted and Real-Time Papers*, 1(3), 2020.

BC+98. Space Studies Board, National Research Council, et al. *Privacy Issues in Biomedical and Clinical Research. In: A strategy for research in space biology and medicine in the new century*. National Academies Press, 1998.

Bio20. New-England Biolabs. *A Typical DNase I Reaction Protocol (M0303)*, 2020. https://tinyurl.com/COVID19-BIO20 (accessed June 2, 2020).

BPE+08. Pedro Baptista, Eulália Pereira, Peter Eaton, Gonçalo Doria, Adelaide Miranda, Inês Gomes, Pedro Quaresma, and Ricardo Franco. Gold nanoparticles for the development of clinical diagnosis methods. *Analytical and bioanalytical chemistry*, 391(3):943–950, 2008.

CAD+16. Marco Capocasa, Paolo Anagnostou, Flavio D'Abramo, Giulia Matteucci, Valentina Dominici, Giovanni Destro Bisol, and Fabrizio Rufo. Samples and data accessibility in research biobanks: an explorative survey. *PeerJ*, 4:e1613, 2016.

Can16. Joseph A Cannataci. Report of the special rapporteur on the right to privacy. *Human Rights Council*, 2016.

CBC16. CBC. *[Alberta Health Services notifies almost 13,000 patients of privacy breach*, 2016. https://tinyurl.com/COVID19-CBC16 (accessed May 16, 2020),.

CGS+20. Linda J Carter, Linda V Garner, Jeffrey W Smoot, Yingzhu Li, Qiongqiong Zhou, Catherine J Saveson, Janet M Sasso, Anne C Gregg, Divya J Soares, Tiffany R Beskid, et al. Assay techniques and test development for covid-19 diagnosis, 2020.

Cri20. Karla Cripps. *[Airline passengers undergo Covid-19 blood tests before boarding*, 2020. https://tinyurl.com/COVID19-Cri20 (accessed April 24, 2020).

CT04. Anne Cambon-Thomsen. The social and ethical issues of post-genomic human biobanks. *Nature Reviews Genetics*, 5(11):866–873, 2004.

FA20. Food and Drug Administration. *Coronavirus (*COVID-19*) Update: serological test validation and education efforts*, 2020 (accessed April 23, 2020). https://tinyurl.com/COVID19-FA20.

Fen17. Emily Feng. *[China authorities mandated to collect DNA from Xinjiang residents*, 2017. https://tinyurl.com/COVID19-FEN17 (accessed May 16, 2020).

FSR18. UK Government Publications Forensic-Science-Regulator. *DNA mixture interpretation, FSR-G-222*, 2018. https://tinyurl.com/COVID19-FSR2018 (accessed June 23, 2020).

Gor20. Alexander E Gorbalenya. Severe acute respiratory syndrome-related coronavirus–the species and its viruses, a statement of the coronavirus study group. *BioRxiv*, 2020.

Haa17. Benjamin Haas. *[Chinese authorities collecting DNA from all residents of Xinjiang*, 2017. https://tinyurl.com/COVID19-HAA17 (accessed May 16, 2020).

HCPT17. Yue He, Fen Cheng, Dai-Wen Pang, and Hong-Wu Tang. Colorimetric and visual determination of dnase i activity using gold nanoparticles as an indicator. *Microchimica Acta*, 184(1):101–106, 2017.

HFK96. Zeqi Huang, Michael J Fasco, and Laurence S Kaminsky. Optimization of dnase i removal of contaminating dna from rna for use in quantitative rna-pcr. *Biotechniques*, 20(6):1012–1020, 1996.

HSR+08. Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8), 2008.

KDK11. Manfred Kayser and Peter De Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179, 2011.

KMG20. Elizabeth Melimopoulos Kate Mayberry and Mersiha Gadzo. *Spain extends coronavirus emergency until June 21: Live updates*, 2020. https://tinyurl.com/COVID19-KMG20 (accessed June 18, 2020).

KYK+11. Min-Jung Kang, Hannah Yu, Sook-Kyung Kim, Sang-Ryoul Park, and Inchul Yang. Quantification of trace-level dna by real-time whole genome amplification. *PloS one*, 6(12), 2011.

Lee00. Bonnie M. Lee. *[Comparison of FDA and HHS Human Subject Protection Regulations*, 2000. https://tinyurl.com/COVID19-LEE00 (accessed May 16, 2020).

LVLM20. By Ethan Lee and The Harvard Crimson Virginia L. Ma. *[At Least 500,000 Tests Needed Per Day to Reopen Economy, Harvard Researchers Say*, 2020. https://tinyurl.com/COVID19-LVLM20 (accessed April 27, 2020).

MNMC12. Zubin Master, Erin Nelson, Blake Murdoch, and Timothy Caulfield. Biobanks, consent and claims of consensus. *Nature Methods*, 9(9):885, 2012.

Mon17. Kate Monica. *[79K Patients Affected by Emory Healthcare Data Breach.*, 2017. https://tinyurl.com/COVID19-MON17 (accessed May 16, 2020),.

Mos19a. Steven W Mosher. *[China's Ploy to Establish a Global DNA Database The first in a 3-part series on the regime's DNA collection project*, 2019. https://tinyurl.com/COVID19-MOS19a (accessed May 16, 2020).

Mos19b. Steven W Mosher. *[What Will China Do With Your DNA? China's Fourth Magic Weapon, Part III: Bioweapons*, 2019. https://tinyurl.com/COVID19-MOS19b (accessed May 16, 2020).

MS00. Bradley Malin and Latanya Sweeney. Determining the identifiability of dna database entries. In *Proceedings of the AMIA Symposium*, page 537. American Medical Informatics Association, 2000.

MS01. Bradley Malin and Latanya Sweeney. Re-identification of dna through an automated linkage process. In *Proceedings of the AMIA Symposium*, page 423. American Medical Informatics Association, 2001.

Pec17. Alexandra Wilson Pecci. *[Healthcare Data Breaches Up 40% Since 2015*, 2017. https://tinyurl.com/COVID19-Pec17/ (accessed May 16, 2020).

Pra20. Ritu Prasad. *Coronavirus: Why is there a US backlash to masks?*, 2020. https://tinyurl.com/COVID19-Pra20 (accessed June 18, 2020).

Pre19. Rich Press. *DNA Mixtures: A Forensic Science Explainer, NIST US Government Publications*, 2019. https://tinyurl.com/COVID19-Pre19 (accessed June 24, 2020).

RC06. Rosemary Rowe and Michael Calnan. Trust relations in health care—the new agenda. *The European Journal of Public Health*, 16(1):4–6, 2006.

Sch20a. Julia Schulman. *Israel Renegotiates COVID-19 Testing Lab Deal With China*, 2020. https://tinyurl.com/COVID19-Sch20 (accessed May 15, 2020).

Sch20b. Matthew S. Schwartz. *Germany Backs Away From Compiling Coronavirus Contacts In A Central Database*, 2020. https://tinyurl.com/COVID19-Sch20b (accessed June 18, 2020).

ST14. Shinobu Sato and Shigeori Takenaka. Highly sensitive nuclease assays based on chemically modified dna or rna. *Sensors*, 14(7):12437–12450, 2014.

SZZ+13. Xin Su, Chen Zhang, Xiaocui Zhu, Simin Fang, Rui Weng, Xianjin Xiao, and Meiping Zhao. Simultaneous fluorescence imaging of the activities of dnases and 3' exonucleases in living cells with chimeric oligonucleotide probes. *Analytical chemistry*, 85(20):9939–9946, 2013.

TBB14. Duncan Taylor, Jo-Anne Bright, and John Buckleton. Interpreting forensic dna profiling evidence without specifying the number of contributors. *Forensic Science International: Genetics*, 13:269–280, 2014.

TNT20. The-National-Thailand. *Country under state of emergency until June 30*, 2020. https://https://tinyurl.com/COVID19-TNT20 (accessed June 18, 2020).

UKK+20. Buddhisha Udugama, Pranav Kadhiresan, Hannah N Kozlowski, Ayden Malekjahani, Matthew Osborne, Vanessa YC Li, Hongmin Chen, Samira Mubareka, Jonathan B Gubbay, and Warren CW Chan. Diagnosing covid-19: the disease and tools for detection. *ACS nano*, 14(4):3822–3835, 2020.

Vau20. Serge Vaudenay. Centralized or decentralized. *The contact tracing dilemma*, 2020.

Wal10. Ann Waldo. The texas newborn bloodspot saga has reached a sad—and preventable—conclusion. *Genomics Law Report*, 16:1–45, 2010.

Wil18. Jane Williams. *[Exploring The World's Largest Biobanks*, 2018. https://tinyurl.com/COVID19-WIL18/ (accessed May 16, 2020).

Wil20. Codi Wilson. *Ontario plans to extend state of emergency until middle of July*, 2020. https://tinyurl.com/COVID19-Wil20 (accessed June 18, 2020).

WKLT20. Yishan Wang, Hanyujie Kang, Xuefeng Liu, and Zhaohui Tong. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-COV-2 outbreak. *Journal of medical virology*, 2020.

WRSS00. Ilse Wiame, Serge Remy, Rony Swennen, and László Sági. Irreversible heat inactivation of dnase i without rna degradation. *Biotechniques*, 29(2):252–256, 2000.

XHM07. Xiaoyang Xu, Min Su Han, and Chad A Mirkin. A gold-nanoparticle-based real-time colorimetric screening method for endonuclease activity and inhibition. *Angewandte Chemie International Edition*, 46(19):3468–3470, 2007.

ZGS+17. Sophie Zaaijer, Assaf Gordon, Daniel Speyer, Robert Piccone, Simon Cornelis Groen, and Yaniv Erlich. Rapid re-identification of human samples using portable dna sequencing. *Elife*, 6:e27798, 2017.