Draft of the Summary Report of 42-2017

Robust STR calling from High-throughput Sequencing Technologies

Federal Agency: National Institute of Justice
Federal Grant: 2014-DN-BX-K089

Applicant Contact Information
Name: Dr. Yaniv Erlich
Title: Core Member
Organization: New York Genome Center
Email: yaniv@nygenome.org
Phone: 617-913-1318
Address: 101 Avenue of the Americas, New York, NY 10013

Organization Point of Contact
Name: Desiree Douglas
Title: Manager, Sponsored Research
Email: ddouglas@nygenome.org
Phone: (646) 977-7051

Recipient organization:
Organization: New York Genome Center
Address: 101 Avenue of the Americas, New York, NY 10013

Submission date: Feb 20th, 2017
Project start date: January 1st, 2015
Project end date: December 31st, 2017

Funding opportunity number: "Research and Development in Forensic Science for Criminal Justice Purposes" CFDA No. 16.560

Signature of submitting official:

Desiree Douglas

Digitally signed by Desiree Douglas
Date: 2018.02.23 14:23:35 -05'00'

Purpose of the project

DNA forensic has become a pivotal tool in the criminal justice system. As of today, the US law enforcement agencies stores the DNA profiles of 17 million known suspects in a the federated CODIS database, which has assisted nearly 400,000 investigations. With this immense success, forensic labs have witnessed a constant annual increase in the number of samples received for DNA testing. However, technology has progressed at the same pace. In fact, despite breathtaking developments in genomics technologies, DNA profiling technology still mainly relies on cumbersome capillary electrophoresis from the 90s that suffers from a low throughput, limited resolution, and aging machines.

The overall arching goal of our project was to develop a range of cutting-edge algorithms, software, and datasets for forensic DNA technologies using high throughput sequencing. These sequencing platforms have become the ultimate backend for molecular biology experiments. They are already the standard in medical care for prenatal testing, identifying the genetic basis of severe pediatric conditions, and are utilized in diagnosing cancer. A growing trend is to use these machines to profile DNA from a single human cell, which is the holy grail for forensic cases that involve minute quantities. The power of these platforms stems from their ability to process a large number of DNA molecules in parallel. Due to constant improvements to these technologies, the costs of DNA sequencing have been falling exponentially in the last few years, from $10,000 to sequence 1 million DNA nucleotides in 2001 to less than $0.01 for the same amount of sequencing today. In addition, these sequencers improved their throughput and can read over a trillion DNA nucleotides per day. Taken together, these trends open the possibility to both lower the costs of forensic DNA fingerprinting, reduce laboratory backlog, and provide richer information about forensic samples.

Applying high throughput sequencing to DNA forensic necessitates solving several technical challenges. First, high throughput sequencers usually read the genome in short, random stretches. To explain that, think about the sentence "*the wheels on the bus go round and round.*" High throughput sequencers could report back the following stretches: "*the wheels on the bus g*", "*he bus go round and rou*", and "*s go round and round*". This short pieces are not very problematic to assemble is they are unique. However, DNA forensic databases solely relies on profiling genomic variations that are called Short Tandem Repeats (STRs). These repetitive elements consist of a short motif that reoccur multiple times at different length between individuals. For example, one person can exhibit "AC AC AC AC AC AC AC AC" in a specific region on one of his chromosomes, whereas another person can exhibit "AC AC AC AC". The challenge is to accurately determine the number of times that a motif occurs from these short stretches in order to implicate the suspect and not

confuse them with innocent people. To make things even harder, these repeats have the tendency to mutate while being processed for DNA sequencing, in what is known as "stutter noise". For example, preparing a sample harboring "AC AC AC AC" repeat for forensic profiling can induce also configurations such as: "AC AC AC" or "AC AC AC AC AC". Successful algorithms need to eliminate these configurations. Finally, DNA samples in crime scenes can be highly degraded or come in very low quantities such as a single drop of blood or from a fingerprint. In these cases, due to stochastic processes, the DNA sequencers might fail to read specific DNA regions and therefore miss STRs relevant for DNA profiling. The DNA forensic jargon dubs these cases as allelic dropouts.

Our first set of studies focused on developing a robust algorithm to profile short tandem repeats from high throughput sequencing data. We sought for a method that can handle the short reading lengths of high throughput sequencing and can mitigate the stutter noise issues. In addition, we looked for extended capabilities that will make the algorithm more immune to allelic dropouts, such as completing missing alleles in a process known as genetic imputation, assigning the remaining alleles to a paternal of maternal chromosomes, or extracting more information for each present allele to compensate for the dropouts. As part of this process, as we developed the algorithm using a large collection of samples, we also created a high-quality reference set of forensic STRs that could be used to train the algorithm and serve as a community-wide resource to develop more algorithms.

Our second set of studies aimed discover more STRs that could help DNA forensic. High throughput sequencers have much higher capacity than legacy technologies. Therefore, future forensic work can leverage a much larger set of STRs to increase the reliability of matches to a suspect, boost their immunity to dropouts, and differentiate between samples when the DNA is found in a mixture. However, not all STRs have been created equally. Some STRs mutate very slow while others mutate quickly. The latter are in general more important for forensic analysis because they show higher variability between people, boosting their ability to differentiate between the true perpetrator and other suspects. These fast mutating STRs are specifically important for solving sex-crimes. In these cases, a small amount of the DNA of a male perpetrator is mixed with a large amount of a female victim, forcing the investigation to focus on the Y chromosome, which is a male specific DNA region and can be reliability differentiated from the victim's DNA. However, the Y chromosome does not recombine and therefore all of the paternal relatives of a perpetrator will display a highly similar – and even identical -- Y chromosome STR profiles (Y-STRs). Thus, limiting the resolution and the utility of the DNA evidence. Fast mutating Y-STRs increase the chance of finding differences between paternal relatives and implicating

only a single person. However, finding fast mutating STRs has been a slow process when conducted using legacy DNA profiling methods. To address this challenge, we instructed the algorithm developed in the previous aim to virtually scan every STR in the genome (and the Y chromosome). We then devised a series of genetic algorithms in order to infer the mutation rate of each STRs.

In our third set of studies, we developed a rapid and portable methodology that will allow to identify DNA using handheld high throughput DNA sequencers directly in crime scenes or an area affected by a mass disaster. Current DNA fingerprinting techniques usually require transporting samples to labs equipped with bulky equipment. As of today, the state of the art forensic genotyping platforms (e.g. DNAscan or RapidHIT 200) take about 90 minutes to process a DNA sample, weigh over 50 kilograms, have a capital cost of more than a quarter of a million dollars, and require about $300 to process a samples. These technological barriers limit DNA fingerprinting to high-latency applications and narrow its accessibility to specific societal entities such as law enforcement or security forces. We developed a portable and inexpensive strategy for re-identification of anonymous human DNA using a MinION sequencer (Oxford Nanopore Technologies). This device weighs less than 100 grams, costs $1000, and only requires a standard laptop to operate. The data is available in near real-time, with the first sequenced DNA fragments ready for analysis only a few seconds after the start of a sequencing run. Due to its small footprint and high robustness, it is possible to employ

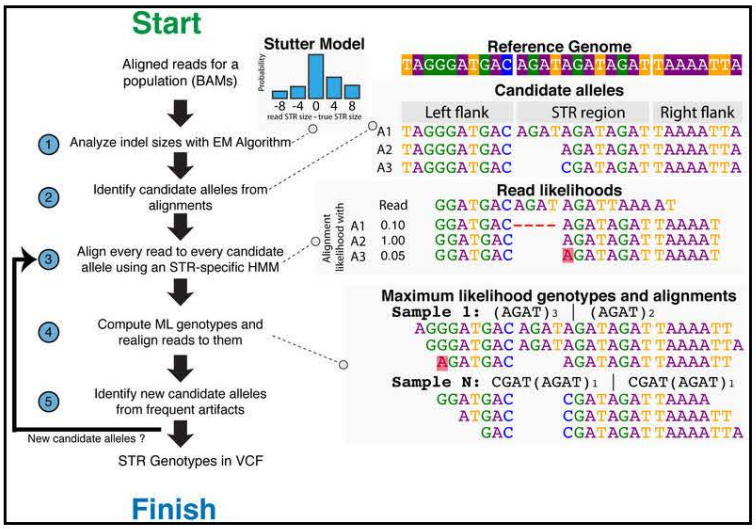

Figure 1: The basic steps of the HipSTR algorithm

MinION sequencers in the field with minimal investments.

In the following section, we will go through the design and methods, data analysis, findings, and implications for each one of our set of studies.

Robust algorithm for STR profiling from high throughput sequencing data

Methods

We developed a novel algorithm called haplotype inference and phasing for STRs (HipSTR) (**Figure 1**). Briefly, HipSTR begins by learning a parametric model that captures each STR's stutter noise profile. Using the genomic location of the repeat, it then harnesses this profile and a machine-learning approach called hidden Markov model (HMM) to realign the STR-containing reads to candidate haplotypes. This process further mitigates the effects of PCR stutter. The realignment framework is highly flexible and can integrate population-scale data from other individuals and phased scaffolds of other genetic variations to determine the most likely alleles, conferring even further robustness to the reported profile.

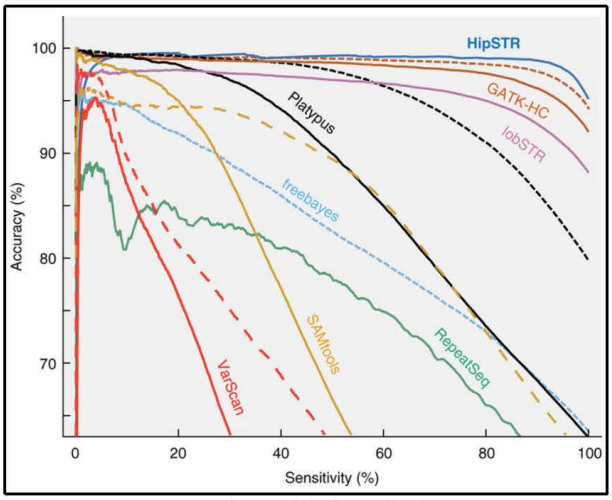

Figure 2: accuracy and sensitivity of HipSTR compared to various existing algorithms to profile STRs from sequencing data.

Data analysis and findings

We benchmarked HipSTR's accuracy by comparing its STR calls from high throughput sequencing data versus the current gold standard for STR profiling in legacy forensic platforms. To this end, we obtained 263 high throughput sequencing data sets from the Simons Genome Diversity Project (SGDP) that were sequenced with an Illumina, the most common and cost effective high throughput sequencer. A subset of these samples also had gold standard profiling for 600 STRs using the legacy forensic technology. It achieved an overall accuracy of 95.2% with the gold dataset. After filtering the 10% least confident genotypes, HipSTR again exhibited superior performance, and its accuracy improved to 98.9%, which is the same as the gold standard accuracy. For comparison, we also genotyped the same STRs with various existing algorithms to profile STRs. Under all settings, HipSTR outperformed all existing tools (**Figure 2**).

To further explore the performance of HipSTR with longer Illumina reads, we performed targeted high throughput sequencing of a panel of long forensic STRs in a single individual from our lab collection. The resulting HipSTR calls

perfectly matched the capillary results, demonstrating the power of the algorithm to profile forensic STR loci from high throughput sequencing data.

Next, we evaluated HipSTR's ability to report not only length polymorphisms but also full STR haplotypes. About half of the STRs in the genome display a repeat structure that includes short interruptions to the recurrent motif. For example, one individual can exhibit the following allele: "AC AC AC TT AC" and other individual might exhibit the following allele: "AC TT AC AC AC". Current STR profiling methods would report the same STR length for both individuals, omitting data that could help the investigation. As HipSTR reports accurate STR sequences, we sought to test its accuracy using a trio that was sequenced by Illumina. For ~70,700 STRs that passed our filters, at least two alleles had identical lengths but different sequences. Only 304 (0.4%) of these STRs were inconsistent with the laws of Mendelian inheritance, highlighting the accuracy of the reported sequence variations.

Implications for forensic investigations

Our analysis shows that HipSTR is the best software so far to profile STRs form high throughput sequencing data. It deals well with stutter noise, provides more accurate genotypes, and reports the actual sequence of the STR alleles, which enables more information from low quality samples regarding the identity of the person. In overall, the software has the potential to facilitate the integration of high throughput sequencing platforms as part of DNA forensics.

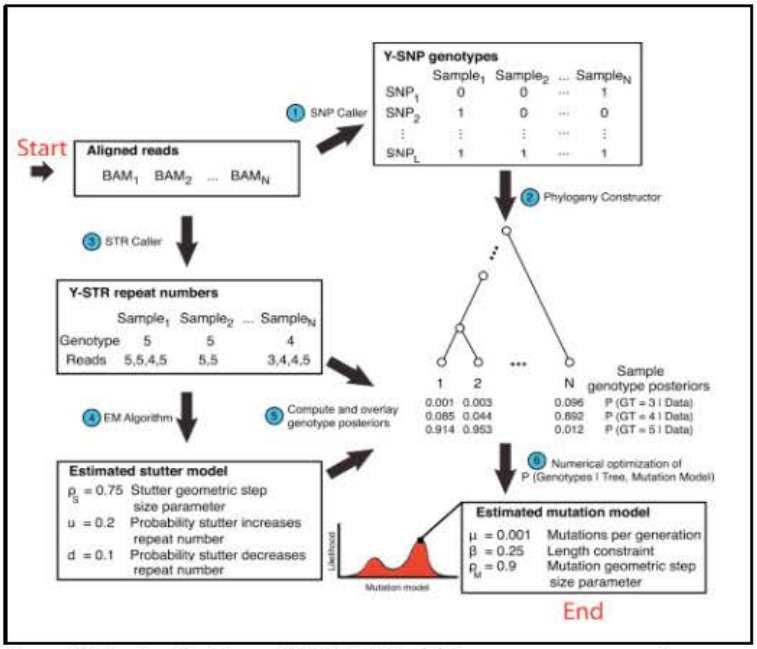

Figure 3: the basic steps of MUTEA for Y chromosome samples

As subsidiary impact, we created an STR call set for the Simons Diversity Genomes Project, which includes most of the CODIS and Y-STR markers for a diverse set of populations. The set can be used as a benchmark for additional STR

genotyping algorithms, facilitating additional tools in the area. The software is available as an open source to facilitate adoption by the forensic community.

Discovery of additional forensic STRs by considering each STR in the genome

Methods

To find fast-mutating STRs that could expand forensic panels, we developed a method to estimate the mutation rate of each STR in the genome. Our model, called MUTEA, took into account three aspects of the mutational dynamic of STR sequences, namely their tendency to grow by a specific number of steps, bias towards a certain allelic length, and the likely mutation rate per generation, which was the parameter of interest (**Figure 3**). Usually, mutation rates are inferred by enumerating the number of differences between a large group of parent and children genomes. However, large-scale projects usually focus on sequencing unrelated individuals rather than parent-child trios, reducing the availability of these datasets. In addition, STR-specific mutation rates from parent-child trios tend to be imprecise. For example, a highly mutating STR typically displays a rate of up to 1 mutation in 100 transmissions. If we observe only 100 trios, which is quite typical in high throughput studies, we can see sometimes zero mutations due to stochastic noise, which could lead to the wrong conclusion that this STR is not informative. To overcome this problem, we sought for a method that can use population-scale sequencing datasets of unrelated individuals that are readily available for various diverse populations. The challenge in this approach is to devise a molecular clock that can somehow calibrate the differences in the STR lengths between individuals that have various levels of distant relationships into the expect mutation rate per generation. We solved this problem by first devising a molecular clock that counts the number of point mutations around the STRs of a pair of individuals. The rate of point mutation per generation is known from previous studies (about 3 mutations every 100 million nucleotides per generation). Therefore, by counting the number of point mutations, we estimated the number of generations elapsed between a pair of individuals and their most

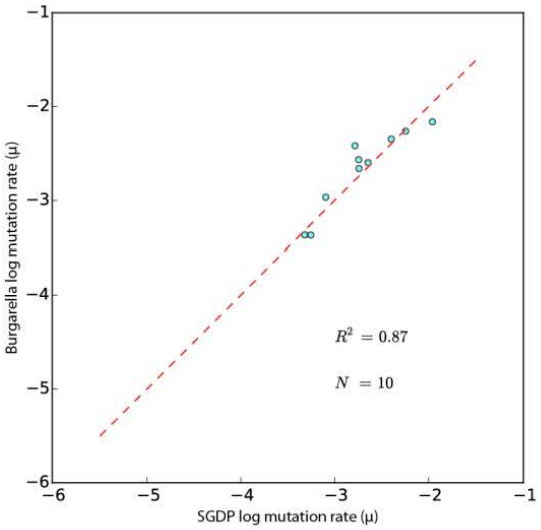

Figure 4: MUTEA estimates (x-axis) versus known mutation rates for 10 commonly used Y-STRs in forensic.

common recent ancestor. Then, we calibrated the differences in their STR alleles into the rate per generation. Repeating this process over and over with various pairs of individuals conferred a reliable estimate of the mutation rate.

Data analysis and findings

We ran MUTEA on whole genome sequencing (WGS) datasets from the 1000 Genomes and the SMGF collection using HipSTR. In order to validate MUTEA, we first processed the CODIS markers currently used in forensic, whose per-generation mutation rate has been extensively studied and established. We found a good correlation (R^2 = 81%) between MUTEA estimates and previous estimates. In addition, we also compared the MUTEA estimates for a panel of 10 Y-STRs commonly used in forensic analysis whose mutation rate was studied using 5000 father-son duos (**Figure 4**). Again, we found excellent concordance (R^2 = 0.87) between MUTEA and traditional trio-based methods.

Encouraged by the quality of our method, we scanned each STR region in the human genome that can analyzed with high throughput sequencing. We inferred the mutation rate for 251,510 STRs on the autosome and additional 4500 STRs on the Y-STRs.

Finally, we also assessed the ability to recover Y-STRs allelic dropouts using genetic imputation. This capability could be useful in forensic cases involving a highly degraded male sample, from which it would be difficult to obtain complete Y-STR profiles. We assessed the accuracy of our algorithm by imputing the 1000 Genomes individuals for the PowerPlex Y23 panel, a set of markers regularly used in forensic cases involving sex crimes. In over 100 iterations, we randomly constructed reference panels of 500 samples and used MUTEA to impute simulated allelic dropouts for a distinct set of 70 samples from the sequencing data. Despite the small size of the reference panel, we were able to correctly impute an average of 66% of the genotypes without any quality filtration. Discarding imputed genotypes with quality score below 70% resulted in an overall accuracy of 88% and retained about 40% of the calls.

Implications

We reported the mutation rate of each STR in the genome that is accessible to Illumina sequencing. This set may enable finding new fast mutating markers for forensic applications. Specially, we identified 100 new rapidly mutating Y-STRs and refined the mutation rate estimators for existing ones. This set has the potential to provide better discrimination between close paternal relatives in sexual assault cases. On the basis of the entire Y-STR set reported by our study, we expect roughly one de-novo mutation (new change that does not exist in the parental genome) to occur every four generations. In addition, from WGS data, one also expects to identify approximately one de novo point mutation every

2.85 generations, resulting in a 60% theoretical probability of differentiating between a father and son's Y chromosome by high-throughput sequencing. With increased interest in high-throughput sequencing among the forensics community, our results suggest that whole Y chromosome sequencing can achieve better patrilineal discrimination than common panel-based methods. Of course, the main caveat is that sequencing technology is at least an order of magnitude more expensive than a panel-based approach. However, if the current trajectory of dedining sequencing costs continues, random sequencing to discriminate between closely patrilineally related individuals might soon become economically viable.

Our imputation results show that the accuracy is still not feasible to rescue allelic dropouts in forensic investigation. However, we envision that a larger panel of tens of thousands of male genomes will substantially increase the ability to correctly impute Y-STRs. Such panels are currently under constructions by various biobanks and should be available for forensic research. Thus, our algorithm might be able in a few years to rescue highly degraded or low copy number samples sequenced with high throughput methods. The full datasets of our analysis was published in the scientific literature to facilitate dissemination knowledge.

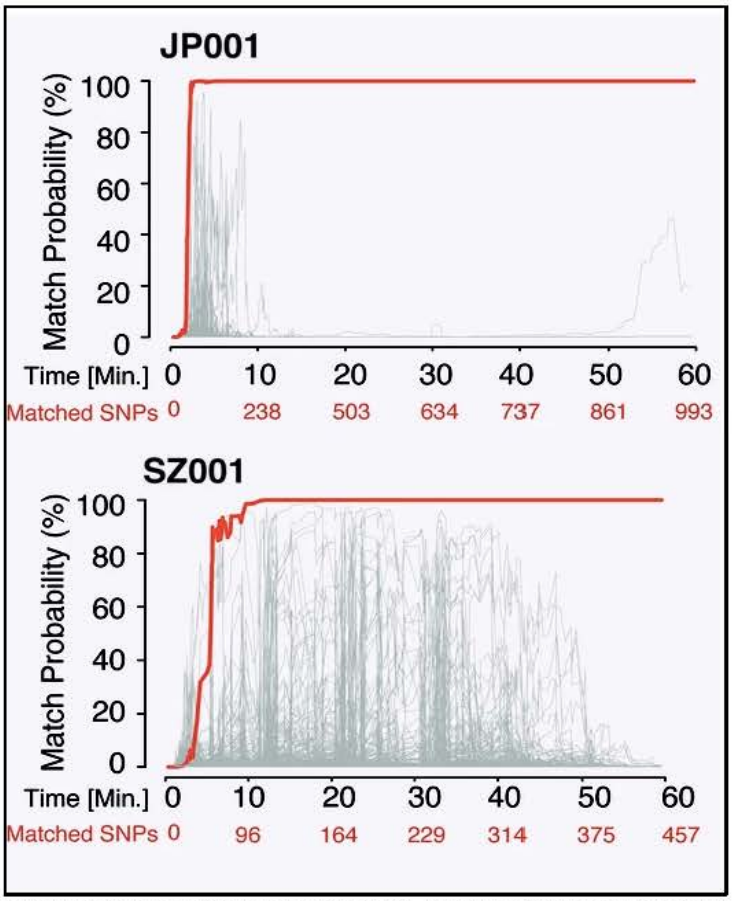

Figure 5: the match probability (red) of the male and the female (bottom) samples as a function of sequencing time using handheld sequencer. Three to ten minutes are sufficient to infer identify. In gray, results for the 30,000 by stander samples in our database.

Rapid re-identification of human samples using portable DNA sequencing

Methods

To develop a rapid and portable method for DNA re-identification, we focused on MinION sequencer. While being extremely small (100g), cheap ($1000), and portable, MinION exhibits two challenges: first, it has a high error rate of 5–15%, which is two orders of magnitude beyond the expected differences between any two individuals. Second, MinION

have much lower throughput compared to Illumina sequencers. Therefore, MinION sequencing results with random bits and pieces of the human genome with many gaps in between. While this could be solved by polymerase chain reaction (PCR) that can direct the sequencer into specific region, we avoided this solution as it would come with cumbersome latency reduced portability.

To address these limitations, we developed a statistical method, called MinION sketching (**Figure 5**). Our method computes the probability that the sketch matches or no match to a record in forensic database. Our sketching method integrated the frequency of each potential allele was discovered by the sketch, the prior probability that a sample matches an entry in the reference database, and the probability of an error. The statistical approach sequentially updates the match/no-match probability with every new input from the sequencer until a reliable conclusion could be reached.

Data analysis and findings

To simulate a forensic scenario, we constructed a large-scale reference databases of genomic datasets that contained data on 31,000 individuals. For each individual, we collected genome-wide genotyping array that includes ~700,000 point mutations (SNPs). These array files came from real individuals tested by Direct-to-Consumer (DTC) companies such as 23andMe, AncestryDNA, and FamilyTreeDNAN.

Next, we sketched samples of a Northern European female and a Northern European-Italian-Ashkenazi male with the latest MinION technology. Importantly, we were able to re-identify these two samples after less than 5 min of MinION sketching. The presence of the other 31,000 individuals did not confuse our method. In fact, the level of evidence reached such a strong likelihood that we estimated that even if the entire earth population was in the database, we could still implicate these two profiles uniquely. We also repeated the analysis with another sample of a mixed Ashkenazi-Uzbeki male but a previous version of the MinION technology. Here, it took us 13 minutes to identify the individual. Finally, we repeated the analysis with a sample whose son and granddaughter are in the database. As expected, our method implicated the sample of interest but not the other family members.

Implications for forensics

Our strategy highlights the possibility of using the MinION in combination with other off-the-shelf equipment to build an inexpensive system for low latency DNA fingerprinting. This may open new possibilities for security and law enforcement applications. Near-real-time DNA surveillance can serve as a tool for identification of victims after a mass disaster or for border control to fight human trafficking. Portable fingerprinting by MinION sketching can also offer great

advantages for regular forensic casework, but integration might be more challenging. Existing forensic databases only hold STR profiles of individuals that are not compatible with our MinION sketches. However, with the continuous drop in costs for sequencing and genotyping arrays, developing combatable databases might be more economically feasible in the future. From a legal perspective, forensic use of MinION sketching seems to comply with US federal and state statutes. Previous scholarly work has postulated that genotyping DNA of abandoned material by law enforcement is lawful in all US states, permitting the generation of MinION sketches from such material by forensic teams. A more complicated aspect is the legality of constructing genome-wide genotyping databases by compulsory DNA collection from arrestees, similar to current forensic databases. Indeed, genome-wide genotyping would collect far more information than the small number of STR markers that are currently used. However, the US legal framework mostly places restrictions on the purpose of compulsory DNA collection rather than the extent of genotyping. In a major decision, the Supreme Court of the United States recently stated that DNA collection from arrestees indeed constitutes a search under the 4th Amendment; but if DNA is solely used for identification, this procedure is reasonable and viewed as a sophisticated version of standard arrestee booking procedures such as fingerprint collection. The Court further stated that genotyping markers that are associated with disease predisposition is not a significant invasion of privacy as long as they are not tested for that end. Therefore, it seems that a genome-wide forensic database would not violate the 4th Amendment. However, the ultimate decision needs to take privacy concerns of the public versus the benefit for forensic applications. Our algorithm was published as an open source to facilitate adoption by the community.