PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Recombulator-X: A fast and user-friendly tool for estimating X chromosome recombination rates in forensic genetics

Serena Aneli[1]*, Piero Fariselli[2], Elena Chierto[1], Carla Bini[3], Carlo Robino[1,4], Giovanni Birolo[2]*

1 Department of Public Health Sciences and Pediatrics, University of Turin, Turin, Italy, **2** Department of Medical Sciences, University of Turin, Turin, Italy, **3** Department of Medical and Surgical Sciences, Section of Legal Medicine, University of Bologna, Bologna, Italy, **4** S.C. Medicina Legale, AOU Città della Salute e della Scienza, Turin, Italy

* serena.aneli@unito.it (SA); giovanni.birolo@unito.it (GB)

Citation: Aneli S, Fariselli P, Chierto E, Bini C, Robino C, Birolo G (2023) Recombulator-X: A fast and user-friendly tool for estimating X chromosome recombination rates in forensic genetics. PLoS Comput Biol 19(9): e1011474. https://doi.org/10.1371/journal.pcbi.1011474

Editor: Jie Liu, University of Michigan, UNITED STATES

Received: April 28, 2023

Accepted: August 28, 2023

Published: September 18, 2023

Data Availability Statement: All underlying data can be found at Recombulator-X and it is freely available on GitHub and PyPI. Full documentation can be found on a dedicated website at https://serena-aneli.github.io/recombulator-x/.

Funding: This work was supported by Programma Nazionale della Ricerca PNR 2021-2027 e PON "Ricerca e Innovazione" 2014-2020 – progetti di ricerca su tematiche "Innovazione" e "Green" (to SA and CR) and by Fondi di Ricerca Locale (ex 60%) Department of Public Health Sciences and

Abstract

Genetic markers (especially short tandem repeats or STRs) located on the X chromosome are a valuable resource to solve complex kinship cases in forensic genetics in addition or alternatively to autosomal STRs. Groups of tightly linked markers are combined into haplotypes, thus increasing the discriminating power of tests. However, this approach requires precise knowledge of the recombination rates between adjacent markers. The International Society of Forensic Genetics recommends that recombination rate estimation on the X chromosome is performed from pedigree genetic data while taking into account the confounding effect of mutations. However, implementations that satisfy these requirements have several drawbacks: they were never publicly released, they are very slow and/or need cluster-level hardware and strong computational expertise to use. In order to address these key concerns we developed Recombulator-X, a new open-source Python tool. The most challenging issue, namely the running time, was addressed with dynamic programming techniques to greatly reduce the computational complexity of the algorithm. Compared to the previous methods, Recombulator-X reduces the estimation times from weeks or months to less than one hour for typical datasets. Moreover, the estimation process, including preprocessing, has been streamlined and packaged into a simple command-line tool that can be run on a normal PC. Where previous approaches were limited to small panels of STR markers (up to 15), our tool can handle greater numbers (up to 100) of mixed STR and non-STR markers. In conclusion, Recombulator-X makes the estimation process much simpler, faster and accessible to researchers without a computational background, hopefully spurring increased adoption of best practices.

Author summary

The X-chromosome is unique in the human genome. In males, the single copy of the X-chromosome is transmitted as a single unbroken DNA chunk to the females of the next

Pediatrics 2023 (to SA and CR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

generation, while in females the two chromosomal copies recombine and one of them is passed to both male and female descendants. Given this peculiar inheritance mode, X-chromosomal genetic markers are crucial for kinship analyses involving, for instance, half-sisters or deficiency paternity cases. In this situation, the recombination rates between genetic markers along the X chromosome need to be known to perform unbiased kinship analysis in forensics, which would be otherwise flawed due to the independence assumption. However, available implementations of computational methods for the estimation of recombination rates are lacking: they are slow, cumbersome and not open source. Thanks to algorithmic improvements and other optimization techniques, we were able to drastically reduce running time, also allowing us to handle more markers than previously feasible. Moreover, we extended previous methods, that were limited to Short Tandem Repeats (STR) markers, to handle any type of polymorphisms. We released our complete implementation as a Python module named Recombulator-X, which is the first open-source software for the estimation of recombination rates between markers along the X-chromosome.

Introduction

The analyses of DNA profiles for personal identification and kinship in forensic casework strongly rely on the biostatistical evaluation of the evidential weight under alternative mutually exclusive scenarios, whose specific probabilities are then combined into likelihood ratios (LRs). Short tandem repeats (STR) have been the markers of choice for such analyses due to their high discriminating capacity and genotyping ease using standard capillary electrophoresis typing techniques [1]. While autosomal DNA polymorphisms are most widely used in forensic practice thanks to their higher informativeness, particular caseworks require complementary information from other genomic regions including haploid markers. For instance, mitochondrial DNA is crucial when ancient or degraded genetic material is involved, while Y-chromosomal STRs are fundamental for the interpretation of mixtures involving a high ratio of female:male contribution [2]. Thanks to its unique features, halfway between autosomes and uniparental markers, the STR markers on the X chromosome (X-STRs) play a relevant role in challenging kinship testing, such as when the DNA from one of the parents is unavailable (kinship deficiency cases), half-sister or incest cases (S1 Fig and S1 Appendix). Moreover, in paternity analyses with inconclusive or statistically weak results, for instance in case of genetic inconsistencies or poor amplification from exhumed remains, adding X-STR markers can help in reaching an informative solution [3–7]. Over the last decade, use of sequencing-based techniques has become increasingly widespread in forensic genetics [8–11]. This brought back interest in other types of markers, especially SNPs, which can be analysed either in combination with STRs or alone [12–17]. Whilst more SNPs are necessary to reach the discrimination capacity of STRs, the possibility of genotyping large number of markers simultaneously from low quantities of input DNA or from degraded material has opened the floodgates to new forensic applications also based on SNPs, such as ancestry inference, DNA phenotyping and investigative genetic genealogy, the latter being specifically designed on dense SNP data [18–24]. Nevertheless, due to the almost exclusive attention traditionally given to STRs in forensics, most available tools do not support non-STR markers.

Forensic markers are located in the non-pseudoautosomal region of the X chromosome. This means that while females have two haplotypes, one inherited from the mother (the maternal haplotype) and one from the father (the paternal haplotype), males have just a single

haplotype inherited from the mother. This haplotype is a mixture of the mother's two haplotypes as a result of the recombination process. Since the genetic size of the X chromosome is about 155Mb [25] and assuming a 50Mb physical distance between markers to ensure independence, a maximum of 3–4 markers can be simultaneously analysed as independently segregating. For this reason, traditional analyses of highly polymorphic haplotypes, consisting of X-STR markers organised into "linkage groups" or "clusters", were devised in order to increase the evidential weight, which would be otherwise statistically inconclusive [6]. Nowadays four different X-STR linkage groups are routinely used for forensic applications [3, 26–44]. However, it has been shown that, while well spaced along the X chromosome, some of these linkage groups cannot be considered truly independent from each other [45–47]. The consequent violation of the independence assumption requires proper considerations in the biostatistical evaluation of kinship. Moreover, although it was originally assumed that recombination did not happen within linkage groups, later studies have demonstrated that, albeit rare, recombination may occur, thus motivating the evaluation of recombination rates for markers both between and within the same cluster [6, 48, 49]. Indeed, the latest recommendations of the International Society for Forensic Genetics (ISFG) about the use of X-STRs in kinship analyses clearly indicate the precise knowledge of recombination rates between markers included in in-house and commercial X-chromosomal multiplex PCR assays as a prerequisite to unbiased estimates of kinship likelihood ratios (LRs) [6]. The available software for kinship LR calculations, with FamLinkX being the most widely used, infers neither recombination nor mutation rates, which are instead expected to be known a priori [45, 46]. However, the evaluation of such measures is not straightforward and may appear computationally intensive. As also highlighted by recent works on the use of the X chromosome in forensics [3, 50], the analytical and statistical issues deriving from genetic linkage and the lack of software addressing such issues are actually hindering the proper applications by leading to significant biases in the quantification of the genetic evidence. For this reason, technical advancements in this field are highly encouraged [3, 6, 50].

Recombination rates are known to vary across the human genome and cannot be automatically derived from combined linkage physical maps [51]. In the case of forensic X-STRs, recombination rates have been either inferred from population samples through high-density multi-point single nucleotide polymorphism (SNP) data [52] or directly estimated in large pedigree-based studies [44, 48, 49, 53]. However, while population-based approaches may suffer from long-term population size changes and selection effects, pedigree studies infer recombination across a few generations by directly observing the inheritance of alleles from parents to offspring [54]. Indeed, the ISFG's guidelines recommend that recombination rates should be primarily estimated from family-based studies [6]. For these reasons, further pedigree-based studies are expected in the future to comply with the steady increase in the number of X-chromosomal markers described for forensic applications [3] and to investigate possible population-specific variability in recombination rates [55].

Recombination between X-chromosomal markers only happens in female meiosis. This entails that only females can provide information on recombination events, while haploid males can be used to phase their mother/offspring. Such events are more easily observed between mother and sons, since genotyping sons immediately yields the recombined maternal haplotypes. Ideal linkage-informative families in pedigree studies are therefore three-generation families, including maternal grandfather, mother and one or more sons. In such families, labelled as *type I*, the mother can be phased using the grandfather and thus recombination events between the maternal haplotypes can be directly observed in the sons. Also informative are two-generation families consisting of one mother and two or more sons, labelled as *type II* (Fig 1A). Here the maternal haplotypes cannot be determined given the lack of the

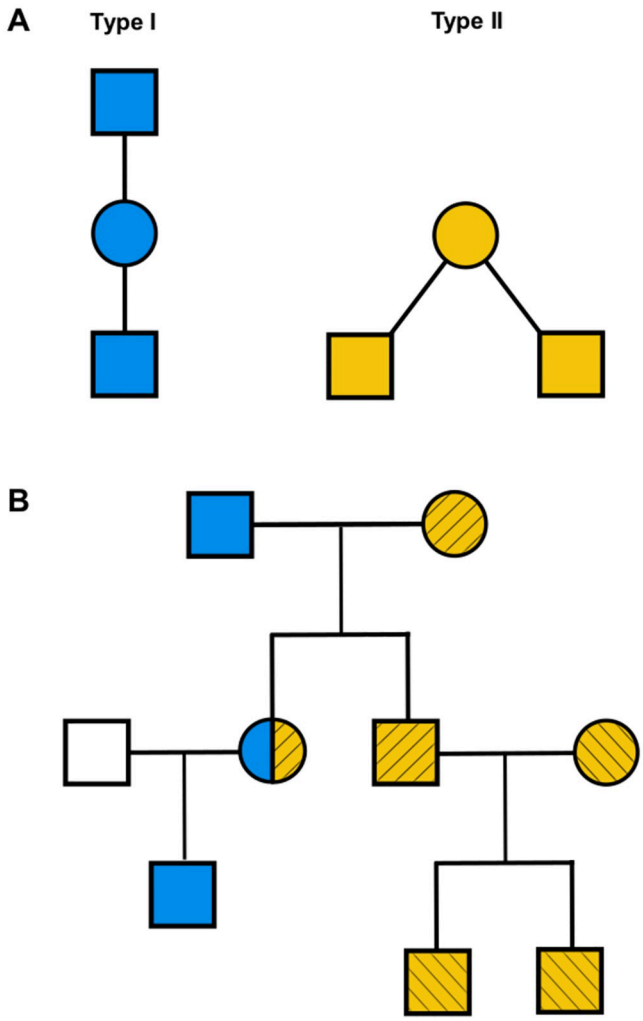

Fig 1. A) Three and two-generation family configurations useful for inferring recombination events (type I in blue on the left and type II in yellow on the right). B) An example pedigree in which one type I (in blue) and two type II families (in yellow with stripe patterns) can be extracted.

https://doi.org/10.1371/journal.pcbi.1011474.g001

grandfather. Hence the need for multiple sons to be evaluated together in order to discern, among all possible maternal phasings, those that can better explain multiple recombination events and thus also give information about recombination rates. Moreover, not only sons can be used: when the father genotype/haplotype is available, the maternal haplotype can also be retrieved from a daughter after phasing (Fig 1B).

The standard statistical approach for the estimation of recombination rates from pedigrees computes the likelihood of kinship by taking into account all possible recombinations within the maternal haplotype, resulting in the exponential complexity of the original algorithm [48]. Despite a new implementation in C++ that allows multi-core parallelization, this approach remains too slow to handle panels of more than 15 X-STRs [49]. Such limitation clashes with the increasing capability of forensic laboratories to simultaneously investigate larger panels of DNA markers favoured by advances in standard capillary electrophoresis typing techniques and the growing use of massively parallel sequencing (MPS) technology [8–17].

Moreover, the implementations of the estimation algorithm were never released to the public, even though they are available upon request from the authors (who kindly provided us with the original R script). They also do not include necessary steps such as data parsing and preprocessing, requiring some R programming knowledge from the user. All of these issues make performing the estimation on new datasets quite onerous, to the point that some recent studies resorted to less accurate but simpler approaches that can be solved manually [56–59].

In order to make the estimation of recombination rates between X chromosomal markers faster and more accessible, we developed the first open-source software with optimised algorithms that allows the user to perform the estimation from a standard pedigree file in just one command. The new algorithms implement the same statistical framework of the previous work [48], without approximations or limiting assumptions, but extending its applicability also to other types of polymorphisms (e.g., SNPs and INDELs). Taking advantage of dynamic programming and other optimization techniques, we were able to drastically reduce computational time, also allowing us to handle an increased number of markers than previously possible. Notably, this improvement was obtained without sacrificing accuracy and without increasing memory usage. We released this work as a Python module named "Recombulator-X", which is the first open-source software for the estimation of the recombination and mutation rates for all types of genetic markers Fig 2. Beyond the optimised implementations of the estimation method, it includes a command-line tool (requiring no programming knowledge), extensive documentation and usage examples, all available in a GitHub repository (https://github.com/serena-aneli/recombulator-x) and a dedicated website (https://serena-aneli.github.io/recombulator-x/).

Materials and methods

Recombulator-X follows the general statistical framework and estimation strategy introduced in [48]. There, the authors define a likelihood function that computes the exact probability of observing a pedigree, given recombination and mutation rates as parameters. Then, they use standard optimisation techniques (the L-BFGS-B method implemented in the optim function in R) to find those rates that maximise the likelihood of the dataset. Our main contribution is introducing a much faster implementation of the likelihood function (based on an optimised algorithm) which yields the exact same probability as the original, that is, without resorting to approximation.

Statistical framework and algorithmic optimization

We present here the statistical framework first introduced in [48] and show the parts that resulted amenable to optimization. A haplotype of STR markers can be described as a vector $x = (x_1, \ldots, x_n)$ of positive rational numbers (repeats can be fractionary). Let $m = (m^1, m^2)$ be the mother's haplotypes and c be the child's maternal haplotype (the one inherited from the mother through recombination of her haplotypes). Possible recombinations are represented by the inheritance vector $v = (v_1, \ldots, v_{n-1})$ where $v_i \in \{1, 2\}$ for all i (S2 Fig and S1 Appendix). Other parameters are the recombination and mutation rate vectors θ and μ of length $n - 1$ and n, respectively. Rates are all in the $\left(0, \frac{1}{2}\right]$ intervals. Then we can define the likelihood of observing a child c from a mother m by a specific inheritance vector v as follows:

$$\mathcal{L}(c \mid m, \theta, \mu, v) = \prod_{i=1}^{n-1}(\theta_i + (1 - 2\theta_i)\delta(v_i, v_{i+1})) \cdot \prod_{i=1}^{n}(\mu_i + (1 - 2\mu_i)\delta(m_i^{v_i}, c_i)) \quad (1)$$

where $\delta(a, b)$ is 1 when $a = b$ and 0 otherwise.

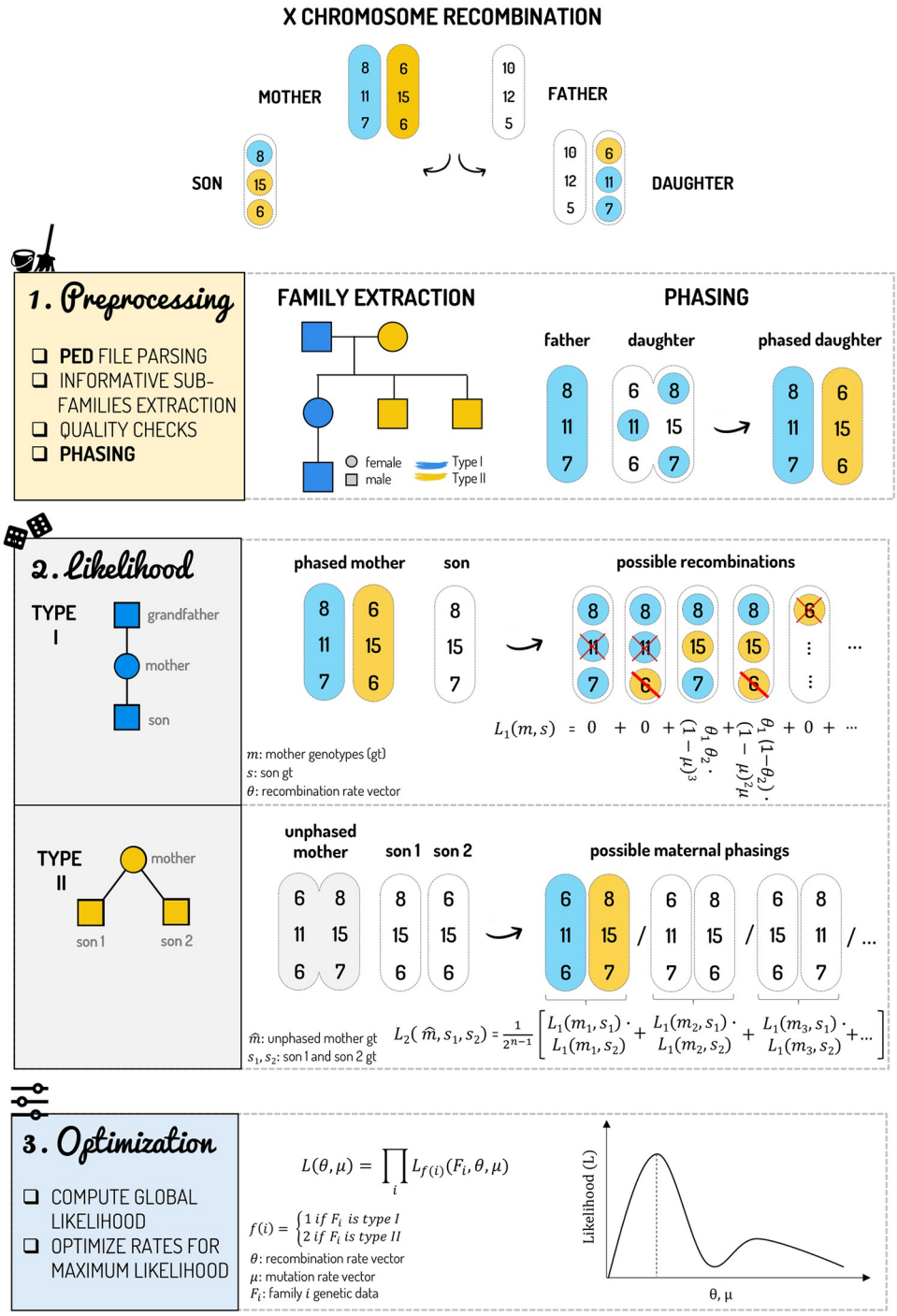

Fig 2. The steps of Recombulator-X are represented on the left column of the figure, while we reported a simplified example on the right using just three X-STRs. 1) Preprocessing: Recombulator-X reads the PED file, performs preliminary quality checks, extracts the informative type I and type II families and phases all the females, whenever their father is available. 2) Likelihood computation, depending on the family type: in the case of a phased mother (type I family), the likelihood (L_1) of each possible recombination is computed and summed up. Here, the red crosses indicate genetic incompatibilities (mutations greater than one repeat), while the single red lines correspond to compatible single-step mutations. When the grandfather is not available (and thus the mother cannot be phased, type II family), this process is repeated for each possible maternal phase (L_2). 3) In the last step—optimization—the likelihood of the entire dataset is computed by multiplying together the likelihood of each family and Recombulator-X searches the parameters (recombination and mutation rates) that maximize the global likelihood.

https://doi.org/10.1371/journal.pcbi.1011474.g002

Then, by summing over all possible 2^n inheritance vectors in $V = \{1, 2\}^n$, we obtain the likelihood of a child's haplotype:

$$\mathcal{L}(c \mid m, \theta, \mu) = \sum_{v \in V} \mathcal{L}(c \mid m, \theta, \mu, v) \tag{2}$$

A son's maternal haplotype can be observed directly from his genotype. For daughters, it can be inferred by subtraction of the father haplotype, when it is available.

This definition of likelihood is enough for type I families (Fig 1A), where the mother's haplotypes can be determined using the grandfather's haplotype and multiple children are handled as independent recombination events, thus multiplying their likelihood together:

$$\mathcal{L}(C \mid m, \theta, \mu) = \prod_{c \in C} \mathcal{L}(c \mid m, \theta, \mu) \tag{3}$$

where C is the set of the children's haplotypes.

For type II families, where only the mother's genotype is known but not her haplotypes, Eq 3 must be extended by further conditioning on the set M of all possible mother's haplotypes given her known genotype as follows:

$$\mathcal{L}(C \mid M, \theta, \mu) = \sum_{m \in M} \mathcal{L}(C \mid m, \theta, \mu) \tag{4}$$

The original exponential-time algorithm, which we will call the direct algorithm, iterates over all possible 2^n recombinations of n markers, following directly Eq 2. However, for any two inheritance vectors that are equal up to marker i, it can be seen that, in Eq 1, all products up to marker $i - 1$ are the same. These repeated sub-computations can be avoided with dynamic programming techniques and indeed by factoring them out we obtained an optimised linear time-complexity algorithm for computing Eqs 2 and 3 with no loss of precision, which we will call the *dynamic* algorithm. This is the case for type I families, where the mother's haplotypes are known. For type II families, where the grandfather is unavailable and only the mother's genotype is known, an additional iteration on all the possible $2^{(n-1)}$ maternal haplotypes is still needed as in Eq 4. Thus, type II families require exponential time also with the dynamic algorithm, albeit going from $O(2^{(2n-1)})$ of the original direct algorithm to $O(n2^{(n-1)})$. Our solution still provides an exponential speed-up with respect to the standard implementation (speed up is equal to $O\left(\frac{2^n}{n}\right)$).

The new likelihood function is then used to estimate the recombination and mutation rates in the same way as in the original paper, by finding a minimum of the negative log-likelihood of the dataset with the rates as parameters, employing the L-BFGS-B method for bound constraints (the rates must be positive and smaller than 0.5) as implemented in the scipy.optimize.minimize function in the SciPy Python package [60], version 1.10.1.

Likelihood implementations

Beyond the algorithmic improvements, other optimization techniques were explored to further reduce computation times. As a result, Recombulator-X includes multiple implementations of the likelihood computation, both of the original direct algorithm and the improved dynamic one: the *direct-loop*, *direct-numpy*, *dynamic* and *dynamic-numba* implementations.

The *direct-loop* is a straightforward implementation of the direct algorithm using loops, similar to the original R implementation. This version of the likelihood is arguably the simplest to understand and was thus used as a reference for testing the correctness of the more complex optimised versions. The same computation was also implemented using the fast vectorized

operations offered by the NumPy package version 1.23.5 [61], with the label *direct-numpy*. However, this implementation is still exponential in time and also in space, since it requires intermediate results to be stored as large multidimensional arrays.

The dynamic programming algorithm is much faster with its linear complexity (for type I families), even when written using Python loops as in the *dynamic* implementation. However, it still benefits from being compiled with the Numba package (version 0.56.4) as the *dynamic-numba* implementation [62]. For type II families, to ameliorate the still exponential complexity, we introduced a further optimization, branching through the possible maternal phasings and computing partial likelihoods up to a certain marker, sharing part of the computations and discarding the branches with zero likelihood early.

Testing and benchmarking were performed by simulating random pedigrees from given recombination and mutation rates using generative functions (included in Recombulator-X). The procedure for generating the simulated pedigrees is the following: at first, the mother's haplotypes are generated with random markers and they are then recombined and mutated according to the given rates to generate the children's maternal haplotype. Then fathers and paternal haplotypes are randomly generated for female children. For type I families, a grandfather is added with one of the mother's haplotypes. Then haplotypes of females are sorted, effectively removing the phase information which will be recovered during preprocessing. This process is repeated for the desired number of families and the generated individuals are then written as a PED file. Datasets yielded by this procedure allow testing of the whole estimation process, from data loading to the recombination and mutation rate estimation.

All benchmarks were averaged across ten runs on a workstation with an Intel i9–12900F processor and 128Gb of RAM. The direct-numpy implementation does require a considerable amount of memory, especially for type II families, and thus an adequate machine was required for benchmarking; however, the actual requirements of the dynamic programming implementation are much lower, allowing Recombulator-X to run on non-workstation hardware for many typical use cases.

Extension to non-STR markers

We extend the statistical framework from [48] to handle panels of arbitrarily mixed STR and non-STR polymorphisms. Single base substitutions are expected to be represented as single-letter codes, but generic strings are accepted to accommodate for more complex non-STR polymorphisms like INDELs. Internally we extend the numeric representation of alleles by encoding unique non-STR alleles with decreasing negative integers. In the likelihood definition, we keep the recombination part unchanged since it is not affected by the type of marker, but we need to extend the mutation part. So we replace Eq 1 with the following:

$$\mathcal{L}(c \mid m, \theta, \mu, v) = \prod_{i=1}^{n-1}(\theta_i + (1-2\theta_i)\delta(v_i, v_{i+1})) \cdot \prod_{i=1}^{n} p_i^*(m_i^{v_i}, c_i))$$

where $p_i^*(a, b)$ is the probability of mutation from an allele a to an allele b, which is defined differently depending on the type of marker. For STR markers, we define:

$$p_i^{\mathrm{STR}}(a, b) \equiv \begin{cases} 1-\mu_i & a = b, \\ \mu_i & |a-b| = 1, \\ 0 & |a-b| \notin \{0, 1\}. \end{cases}$$

so that non-unit mutations (insertion or deletions of multiple or partial repeats) have zero probability since they are much less frequent than unit mutations. Instead for non-STR

polymorphic markers, we define:

$$p_i^{\mathrm{POLY}}(a,b) \equiv \begin{cases} 1-\mu_i & a=b, \\ \frac{9}{10}\mu_i & a \to b \text{ is a transition} \\ \frac{1}{10}\mu_i & \text{otherwise} \end{cases}$$

where transitions are single base mutations between purines (A and G) or pyrimidines (C and T), which are much more frequent than other single base substitutions or more complex polymorphisms [63].

Additional features

While the dynamic programming algorithm for the likelihood function is arguably the core of the package, Recombulator-X original contributions also include non-trivial dataset parsing and preprocessing functions. Thanks to those, datasets can be read from the standard PED format as used by the PLINK software [64], a simple textual format that encodes both arbitrary relatedness and genetic information. Preprocessing functions then 1) build an arbitrarily complex graph for each group of related individuals (S3 Fig and S1 Appendix); 2) extract all subgraphs that can be used as type I or II families; 3) phase mothers and daughters whenever possible and finally yield the processed dataset ready for the estimation. During this preprocessing some consistency checks take place, alerting the user of eventual problems with the data. All these steps are wrapped into a single command line tool, that takes a pedigree file as input and outputs the recombination and optionally the mutation rates. This tool allows the user to run the entire estimation without programming knowledge.

Results

The main contribution of this work is arguably the dynamic programming optimization in the likelihood computation for type I families, where the mother's phasing is known. The likelihood is defined as a sum over all the possible recombinations of the mother's haplotypes, which are 2^n for n markers. The exponential number of recombinations is the source of the exponential complexity of the original implementation. The key observation for the optimization is that the likelihood formula for n markers includes the likelihood formula for the first $n - 1$ markers *two times*. By avoiding the double computation of the likelihood of the first $n - 1$ markers, we can halve the computation time. The same can be done for the first $n - 1$ markers by avoiding the double computation of the likelihood of the first $n - 2$ markers. By repeating this process recursively, we avoid all repeated computations and obtain a linear time algorithm. More details on the dynamic optimization are available in the Methods section.

Similarly, the likelihood computation for type II families, where the mother's phasing is unknown, requires computing the type I likelihood for all possible phasings. Again these are exponential in the number of markers. The optimization of the type I likelihood computation strongly reduces the complexity, but performing an exponential number of linear-time subcomputations still yields an exponential-time algorithm. Unfortunately, the type II likelihood computation does not appear to be amenable to dynamic programming optimization and thus the presence of type II families is still a limiting factor for the number of markers that can be analysed.

In order to assess the reduction in the likelihood computation time and how it improves the whole recombination and mutation rate estimation process, we performed a series of benchmarks using simulated datasets with incremental number of markers.

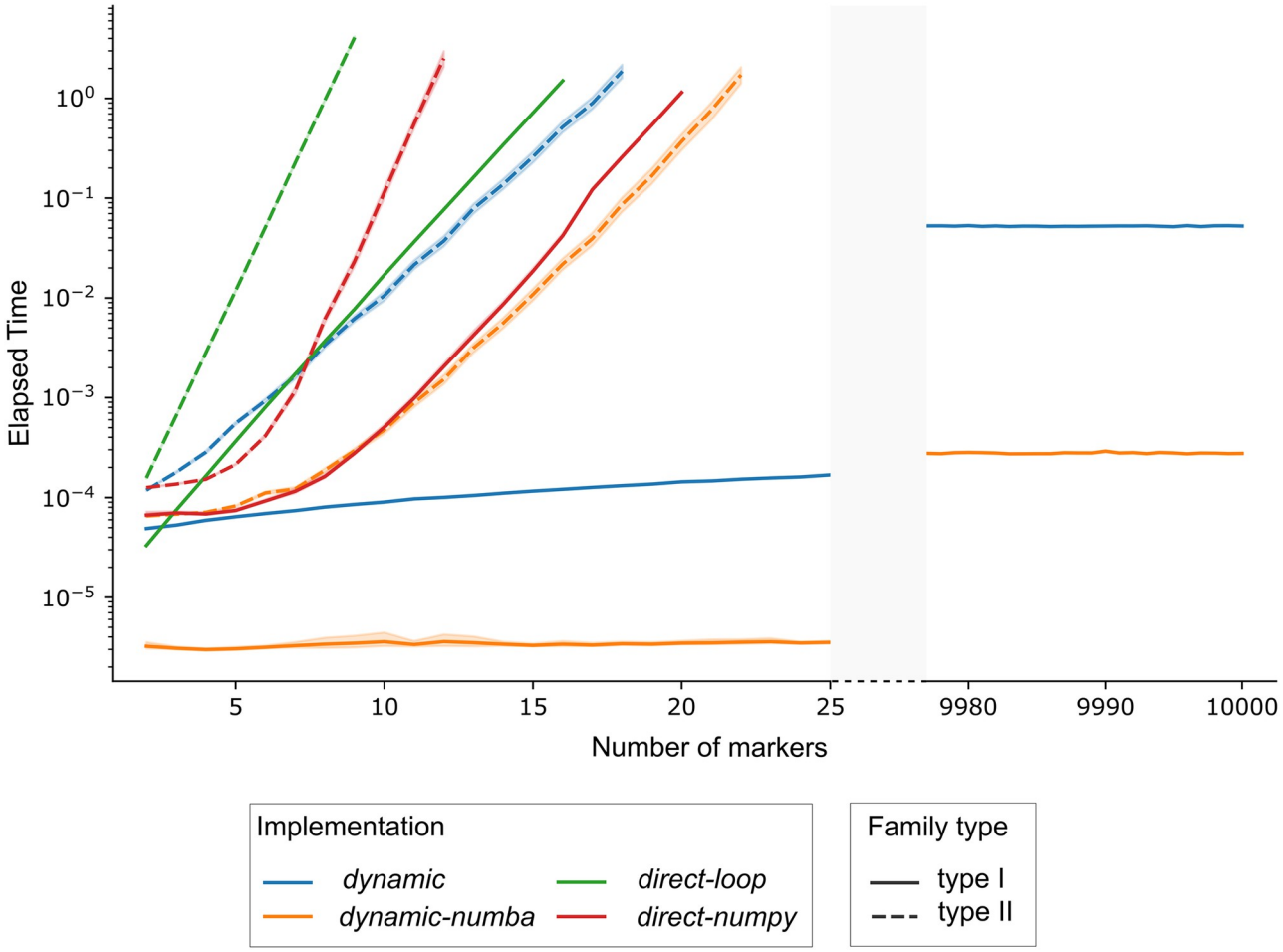

Fig 3. Mean time needed to compute the likelihood for one family typed over up to 10,000 markers. Each implementation is represented with a different colour, while the linestyle refers to family types. The y axis is in log scale. For each implementation, the number of markers was progressively increased until the computation time went above one second per family.

https://doi.org/10.1371/journal.pcbi.1011474.g003

The first benchmark compares the different algorithms and implementations of the likelihood function that are included in Recombulator-X. To see how the computation time is affected by the number of markers, we measured the average time to compute the likelihood of type I or type II simulated families in Fig 3 (see also S1 File and S1 Table). The exponential complexity of the direct algorithm is clearly visible, both for type I and type II families. The dynamic programming algorithm shows instead its linear complexity for type I families, allowing a virtually unlimited number of markers. Unfortunately, type II families remain problematic, even though the dynamic programming numba-optimised version is able to handle ten more markers than the numpy-vectorized direct implementation in the same time (from 11 to 21 markers).

After testing the likelihood function implementations, we benchmarked the entire optimization procedure using the fastest implementation, dynamic-numba, in order to see how the improvements to the likelihood computations impact the whole process of recombination and mutation rate estimation. The results are reported in Fig 4.

Unsurprisingly, when considering only type I families, the estimation is very fast. For 100 families, the estimation of up to 100 markers is almost instantaneous, taking less than two

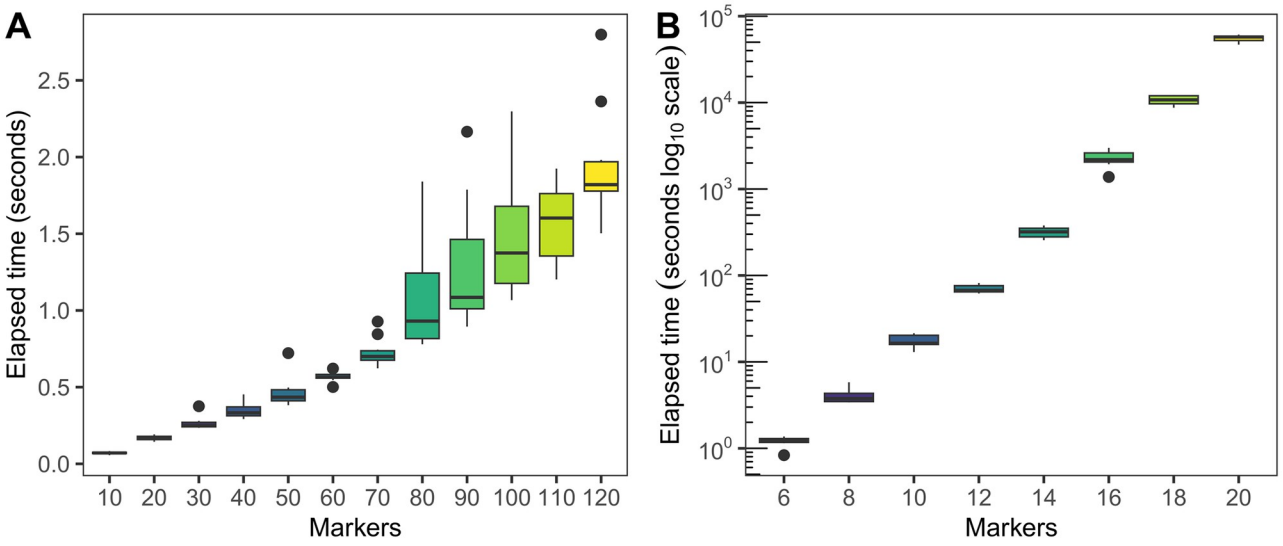

Fig 4. Recombination and mutation estimation times using the fastest (dynamic-numba) likelihood implementation depending on the number of markers. A simulated dataset of 100 type I families (A) and one of 100 type I and 100 type II families (B) were tested. Times are in seconds, with a logarithmic axis for the right panel.

https://doi.org/10.1371/journal.pcbi.1011474.g004

seconds. However, at around 130 markers, we start having issues where the optimization fails to converge. Even raising the iteration limit for convergence and trying the other optimization methods available in SciPy did not allow the estimation to converge.

When also type II families are involved, the exponential complexity of the likelihood function poses a hard limit to the number of markers. For a dataset of 100 type I and 100 type II families, the computation times are much longer, roughly doubling for each additional marker (as expected). We stopped testing at 20 markers, where the average computational time was 16 hours.

While a direct comparison of Recombulator-X with previous software is difficult (the genetic data from the two previous studies are not publicly available), we also simulated datasets with the same size as in previous studies to have a comparison in more realistic scenarios. Nothnagel and colleagues analysed 216 type I and 185 type II families genotyped with a panel of 12 markers [48]. While they did not report the time required, they felt the need to say that a faster implementation was needed. Our method only takes 3.5 minutes on a simulated dataset with the same number of families. In a later analysis, involving 54 type I and 104 type II families with a panel of 15 markers, the authors developed a much faster parallel version in C++. However, due to the increased number of markers and the exponential complexity, the estimation process took a few months on a highly parallel computing systems [49]. Our method takes 20 minutes to carry on the same task. While these times are not completely comparable given that they ran on different datasets, on different hardware and in different languages, it is clear that such a decrease in time cannot be attributed to those factors alone and that Recombulator-X, even with its current limitations, brings a substantial improvement over previous methods.

Note that this speed-up in computation does not entail any trade-off in spatial complexity. On the contrary, Recombulator-X has a minimal memory usage: in the previous two simulated examples, the maximum memory occupation were 197Mb and 202Mb for the 12 and 15 marker datasets, respectively.

Discussion

The proper biostatistical evaluation of the evidential weight in personal identification and kinship tests when dealing with X chromosome markers is a nagging problem in forensics, due to physical linkages [3, 6, 45, 50, 65]. Despite being crucial for unbiased formulations of the evidential weight, as also highlighted by the International Society of Forensic Genetics [6], few biostatistical tools for the evaluation of recombination rates between adjacent forensic markers along the X chromosome are available today.

Routine kinship analyses rely almost exclusively on commercial kits, such as the commonly used Argus X-12 QS which consists of 12 X-STRs [3, 26–44]. However, current implementations of state-of-the-art statistical framework for estimation from pedigrees, besides being quite onerous to use, are already very slow for 12 markers and so unsuitable for larger panels without the availability of large computational resources [48, 49]. Consequently, many recent works have been limited to a "manual" evaluation of the recombination rate which does not consider the mutation probability [56–59].

The growing use of next-generation sequencing technologies in the forensic fields, with the possibility of combining thousands of markers together, requires the development of new biostatistical frameworks scalable to a higher number of genetic markers [9]. Moreover, many commercially available NGS-based kits allow to combine STRs and other non-traditional markers, such as SNPs or INDELs [12–17]. In particular, SNPs have been increasingly appealing thanks to their technical features and informational power: their smaller amplicon size is crucial with samples of low quantity and poor quality (this is relevant since the majority of forensic analyses involves degraded DNA) [66] and they provide insight for predicting human appearance and the biogeographical origin of unknown sample donors or deceased/missing persons [67, 68], thus ultimately resulting in new investigative leads. Additionally, given their lower mutation rate when compared to STRs, they were shown to be helpful in solving kinship cases [69, 70]. Notably, the latest application of SNPs is investigative genetic genealogy where dense SNP data are jointly analysed to infer distant relationships (which in forensics indicate relatedness exceeding that of first cousins) [71]. For these reasons, an increasing number of commercial NGS-based kits have included X chromosomal SNPs and/or STRs to address complex kinship scenarios [14, 18, 72–78]. Nevertheless, complex kinship cases relying on many and mixed types of X chromosomal genetic markers cannot be addressed using the previous implementations for the inference of recombination rates, which are used, albeit with limitations, for STR markers.

In order to overcome these issues, we developed Recombulator-X, the first open-source tool for rapidly inferring X chromosome recombination rates. Our optimised algorithm is substantially faster than existing gold-standard methods, with no loss of accuracy since it is based on the same statistical framework. Performing the estimation on standard panels of 12 markers on a new dataset can now be done in minutes instead of days or weeks on a single PC. This will also enable new studies to experiment with larger panels than previously possible, going from a practical limit of around 15 markers to more than 25 for general datasets and one hundred when considering only type I families. Moreover, the extension to mixed STR and non-STR markers is especially relevant to enable sequencing-based panels.

No less important from a practical point of view is that the full implementation and source code (including dataset parsing and preprocessing) are available as a Python package. The repository also includes documentation, usage examples and a command line tool, greatly simplifying the estimation process for a non-technical user. This, together with the lower computational requirements, will encourage the use of the gold standard estimation technique that can account for mutations instead of the simpler but biased frequency calculation that are still commonly used in the research community.

For all these reasons, we hope that Recombulator-X might transform the estimation of recombination rates from an arduous process requiring specialised expertise and hardware to a routine computational analysis that anyone can perform.

Supporting information

S1 Appendix. Supplementary Materials.
(PDF)

S1 Fig. X chromosomal informative pedigrees. Examples of kinship cases where X chromosomal markers may be informative: half-sisters (A), deficiency paternity test (B) and incest cases (C). The individuals whose genotype needs to be assessed to resolve the kinship case are in yellow.
(PDF)

S2 Fig. Inheritance vector. One of the possible inheritance vectors for a given mother-son pair.
(PDF)

S3 Fig. Pedigree example. Graph of a simulated family.
(PDF)

S1 Table. A short version of the implementations benchmark. Means and standard deviations of the running time needed to compute the likelihood for one family typed over an increasing number of markers. The complete version of this table is reported in S1 File. For each implementation, the number of markers was progressively increased until the computation time went above one second per family.
(PDF)

S1 File. Implementations benchmark. Means and standard deviations of the running time needed to compute the likelihood for one family typed over up to 10,000 markers. For each implementation, the number of markers was progressively increased until the computation time went above one second per family.
(XLSX)

Acknowledgments

We would like to thank Prof. Michael Nothnagel for kindly providing us with his original code and answering our questions.

Author Contributions

Conceptualization: Serena Aneli, Piero Fariselli, Carlo Robino, Giovanni Birolo.

Formal analysis: Serena Aneli, Giovanni Birolo.

Funding acquisition: Carlo Robino.

Investigation: Serena Aneli, Giovanni Birolo.

Methodology: Piero Fariselli, Giovanni Birolo.

Resources: Elena Chierto, Carla Bini.

Supervision: Piero Fariselli, Carlo Robino, Giovanni Birolo.

Visualization: Serena Aneli.

Writing – original draft: Serena Aneli, Giovanni Birolo.

Writing – review & editing: Serena Aneli, Piero Fariselli, Elena Chierto, Carla Bini, Carlo Robino, Giovanni Birolo.

References

1. Butler JM. Short tandem repeat typing technologies used in human identity testing. Biotechniques. 2007; 43(4):ii–v. https://doi.org/10.2144/000112582 PMID: 18019344
2. Butler MJ. Advanced Topics in Forensic DNA Typing: Interpretation. Elsevier; 2015.
3. Gomes I, Pinto N, Antão-Sousa S, Gomes V, Gusmão L, Amorim A. Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. Frontiers in Genetics. 2020; 11. https://doi.org/10.3389/fgene.2020.00926
4. Szibor R. X-chromosomal markers: past, present and future. Forensic Sci Int Genet. 2007; 1(2):93–99. https://doi.org/10.1016/j.fsigen.2007.03.003 PMID: 19083736
5. Pinto N, Gusmão L, Amorim A. X-chromosome markers in kinship testing: A generalisation of the IBD approach identifying situations where their contribution is crucial. Forensic Science International: Genetics. 2011; 5(1):27–32. https://doi.org/10.1016/j.fsigen.2010.01.011 PMID: 20457080
6. Tillmar AO, Kling D, Butler JM, Parson W, Prinz M, Schneider PM, et al. DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. Forensic Sci Int Genet. 2017; 29:269–275. https://doi.org/10.1016/j.fsigen.2017.05.005 PMID: 28544956
7. Pinto N, Silva PV, Amorim A. A general method to assess the utility of the X-chromosomal markers in kinship testing. Forensic Sci Int Genet. 2012; 6(2):198–207. https://doi.org/10.1016/j.fsigen.2011.04.014 PMID: 21592877
8. Bruijns B, Tiggelaar R, Gardeniers H. Massively parallel sequencing techniques for forensics: A review. ELECTROPHORESIS. 2018; 39(21):2642–2654. https://doi.org/10.1002/elps.201800082 PMID: 30101986
9. Ballard D, Winkler-Galicki J, Wesoły J. Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. Int J Legal Med. 2020; 134(4):1291–1303. https://doi.org/10.1007/s00414-020-02294-0 PMID: 32451905
10. Alonso A, Barrio PA, Müller P, Köcher S, Berger B, Martin P, et al. Current state-of-art of STR sequencing in forensic genetics. Electrophoresis. 2018; 39(21):2655–2668. https://doi.org/10.1002/elps.201800030 PMID: 29750373
11. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, et al. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet. 2016; 22:54–63. https://doi.org/10.1016/j.fsigen.2016.01.009 PMID: 26844919
12. Churchill JD, Schmedes SE, King JL, Budowle B. Evaluation of the Illumina Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling. Forensic Sci Int Genet. 2016; 20:20–29. https://doi.org/10.1016/j.fsigen.2015.09.009 PMID: 26433485
13. Novroski NMM, Cihlar JC. Evolution of single-nucleotide polymorphism use in forensic genetics. WIREs Forensic Science. 2022; 4(6). https://doi.org/10.1002/wfs2.1459
14. Stephens KM, Barta R, Fleming K, Perez JC, Wu SF, Snedecor J, et al. Developmental validation of the ForenSeq MainstAY kit, MiSeq FGx sequencing system and ForenSeq Universal Analysis Software. Forensic Sci Int Genet. 2023; 64:102851. https://doi.org/10.1016/j.fsigen.2023.102851 PMID: 36907074
15. van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JFJ, et al. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system. Forensic Sci Int Genet. 2016; 24. https://doi.org/10.1016/j.fsigen.2016.05.016 PMID: 27347657
16. Turchi C, Previderè C, Bini C, Eugenia C, Grignani P, Manfredi A, et al. Assessment of the Precision ID Identity Panel kit on challenging forensic samples. Forensic Sci Int Genet. 2020; 49:102400. https://doi.org/10.1016/j.fsigen.2020.102400 PMID: 33075733
17. Frégeau CJ. Validation of the Verogen ForenSeq DNA Signature Prep kit/Primer Mix B for phenotypic and biogeographical ancestry predictions using the Micro MiSeq Flow Cells. Forensic Sci Int Genet. 2021; 53. PMID: 34058534
18. Tillmar A, Sturk-Andreaggi K, Daniels-Higginbotham J, Thomas JT, Marshall C. The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications. Genes. 2021; 12(12). https://doi.org/10.3390/genes12121968 PMID: 34946917

19. Gorden EM, Greytak EM, Sturk-Andreaggi K, Cady J, McMahon TP, Armentrout S, et al. Extended kinship analysis of historical remains using SNP capture. Forensic Sci Int Genet. 2022; 57:102636. https://doi.org/10.1016/j.fsigen.2021.102636 PMID: 34896972

20. King JL, Churchill JD, Novroski NMM, Zeng X, Warshauer DH, Seah LH, et al. Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit. Forensic Sci Int Genet. 2018; 36:60–76. https://doi.org/10.1016/j.fsigen.2018.06.005 PMID: 29935396

21. Churchill JD, Novroski NMM, King JL, Seah LH, Budowle B. Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. Forensic Sci Int Genet. 2017; 30:81–92. https://doi.org/10.1016/j.fsigen.2017.06.004 PMID: 28651097

22. Dash HR, Avila E, Jena SR, Kaitholia K, Agarwal R, Alho CS, et al. Forensic characterization of 124 SNPs in the central Indian population using precision ID Identity Panel through next-generation sequencing. Int J Legal Med. 2022; 136(2):465–473. https://doi.org/10.1007/s00414-021-02742-5 PMID: 34748086

23. Pereira V, Mogensen HS, Børsting C, Morling N. Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. Forensic Sci Int Genet. 2017; 28:138–145. https://doi.org/10.1016/j.fsigen.2017.02.013 PMID: 28273506

24. de la Puente M, Phillips C, Xavier C, Amigo J, Carracedo A, Parson W, et al. Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. Forensic Sci Int Genet. 2020; 45:102213. https://doi.org/10.1016/j.fsigen.2019.102213 PMID: 31835179

25. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. The DNA sequence of the human X chromosome. Nature. 2005; 434(7031):325–337. https://doi.org/10.1038/nature03440 PMID: 15772651

26. Bergseth EF, Tillmar A, Haddeland PJT, Kling D. Extended population genetic analysis of 12 X-STRs—Exemplified using a Norwegian population sample. Forensic Sci Int Genet. 2022; 60. https://doi.org/10.1016/j.fsigen.2022.102745 PMID: 35870434

27. Robino C, Lacerenza D, Aneli S, Di Gaetano C, Matullo G, Robledo R, et al. Allele and haplotype diversity of 12 X-STRs in Sardinia. Forensic Sci Int Genet. 2018; 33:e1–e3. https://doi.org/10.1016/j.fsigen.2017.12.002 PMID: 29221994

28. Zidkova A, Capek P, Horinek A, Coufalova P. Investigator Argus X-12 study on the population of Czech Republic: comparison of linked and unlinked X-STRs for kinship analysis. Electrophoresis. 2014; 35(14):1989–1992. https://doi.org/10.1002/elps.201400046 PMID: 24789012

29. Elakkary S, Hoffmeister-Ullerich S, Schulze C, Seif E, Sheta A, Hering S, et al. Genetic polymorphisms of twelve X-STRs of the investigator Argus X-12 kit and additional six X-STR centromere region loci in an Egyptian population sample. Forensic Sci Int Genet. 2014; 11:26–30. https://doi.org/10.1016/j.fsigen.2014.02.007 PMID: 24632058

30. Salvador JM, Apaga DLT, Delfin FC, Calacal GC, Dennis SE, De Ungria MCA. Filipino DNA variation at 12 X-chromosome short tandem repeat markers. Forensic Sci Int Genet. 2018; 636:e8–e12. https://doi.org/10.1016/j.fsigen.2018.06.008 PMID: 29909139

31. Bini C, Riccardi LN, Ceccardi S, Carano F, Sarno S, Luiselli D, et al. Expanding X-chromosomal forensic haplotype frequencies database: Italian population data of four linkage groups. Forensic Sci Int Genet. 2015; 15:127–130. https://doi.org/10.1016/j.fsigen.2014.11.008 PMID: 25435156

32. Martinez G, Schaller C, Nazar P, Brondani A, del Rio NG, Bolea M, et al. X-chromosomal haplotype frequencies of four linkage groups in a population of Argentina. Forensic Science International: Genetics Supplement Series. 2015; 5:e524–e526.

33. Uchigasaki S, Tie J, Takahashi D. Genetic analysis of twelve X-chromosomal STRs in Japanese and Chinese populations. Mol Biol Rep. 2013; 40(4):3193–3196. https://doi.org/10.1007/s11033-012-2394-1 PMID: 23275196

34. Vongpaisarnsin K, Boonlert A, Rasmeepaisarn K, Dangkao P. Genetic variation study of 12 X chromosomal STR in central Thailand population. Int J Legal Med. 2016; 130(6):1497–1499. https://doi.org/10.1007/s00414-016-1363-y PMID: 27059997

35. Cainé L, Costa S, Pinheiro MF. Population data of 12 X-STR loci in a North of Portugal sample. International Journal of Legal Medicine. 2013; 127(1):63–64. https://doi.org/10.1007/s00414-012-0672-z PMID: 22297426

36. Veselinović I, Vapa D, Djan M, Veličković N, Veljović T, Petrić G. Genetic analysis of 12 X-STR loci in the Serbian population from Vojvodina Province. Int J Legal Med. 2018; 132(2):405–408. https://doi.org/10.1007/s00414-017-1677-4 PMID: 28868569

37. Sufian A, Hosen MI, Fatema K, Hossain T, Hasan MM, Mazumder AK, et al. Genetic diversity study on 12 X-STR loci of investigator Argus X STR kit in Bangladeshi population. International Journal of Legal Medicine. 2017; 131(4):963–965. https://doi.org/10.1007/s00414-016-1513-2 PMID: 27933412

38. Almarri MA, Lootah RA. Allelic and haplotype diversity of 12 X-STRs in the United Arab Emirates. Forensic Sci Int Genet. 2017; 33:e4–e6. https://doi.org/10.1016/j.fsigen.2017.12.013 PMID: 29305242

39. Cortés-Trujillo I, Zuñiga-Chiquette F, Ramos-González B, Chávez-Briones MdL, Islas-González KL, Betancourt-Guerra DA, et al. Allele and haplotype frequencies of 12 X-STRs in Mexican population. Forensic Sci Int Genet. 2019; 38:e11–e13. https://doi.org/10.1016/j.fsigen.2018.10.012 PMID: 30389253

40. García MG, Catanesi CI, Penacino GA, Gusmão L, Pinto N. X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. Forensic Sci Int Genet. 2019; 41:e8–e13. https://doi.org/10.1016/j.fsigen.2019.04.005 PMID: 31085140

41. Bottinelli M, Gouy A, Utz S, Zieger M. Population genetic analysis of 12 X-chromosomal STRs in a Swiss sample. Int J Legal Med. 2022; 136(2):561–563. https://doi.org/10.1007/s00414-021-02684-y PMID: 34420081

42. Pinto N, Pereira V, Tomas C, Loiola S, Carvalho EF, Modesti N, et al. Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study. Forensic Sci Int Genet. 2020; 46:102258. https://doi.org/10.1016/j.fsigen.2020.102258 PMID: 32066109

43. Hakim HM, Khan HO, Ismail SA, Lalung J, Kofi AE, Aziz MY, et al. Population data and genetic characteristics of 12 X-STR loci using the Investigator Argus X-12 Quality Sensor kit for the Kedayan population of Borneo in Malaysia. Int J Legal Med. 2021; 135(4):1433–1435. https://doi.org/10.1007/s00414-021-02577-0 PMID: 33782746

44. Bini C, Di Nunzio C, Aneli S, Sarno S, Alù M, Carnevali E, et al. Analysis of recombination and mutation events for 12 X-Chr STR loci: A collaborative family study of the Italian Speaking Working Group Ge.F.I. Forensic Science International: Genetics Supplement Series. 2019; 7(1):398–400.

45. Kling D, Tillmar A, Egeland T, Mostad P. A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. Int J Legal Med. 2015; 129(5):943–954. https://doi.org/10.1007/s00414-014-1117-7 PMID: 25425094

46. Kling D, Dell'Amico B, Tillmar AO. FamLinkX—implementation of a general model for likelihood computations for X-chromosomal marker data. Forensic Sci Int Genet. 2015; 17:1–7. https://doi.org/10.1016/j.fsigen.2015.02.007 PMID: 25771099

47. Tillmar AO, Egeland T, Lindblom B, Holmlund G, Mostad P. Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. Forensic Sci Int Genet. 2011; 5(5):506–511. https://doi.org/10.1016/j.fsigen.2010.11.004 PMID: 21167800

48. Nothnagel M, Szibor R, Vollrath O, Augustin C, Edelmann J, Geppert M, et al. Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. Forensic Sci Int Genet. 2012; 6(6):778–784. https://doi.org/10.1016/j.fsigen.2012.02.015 PMID: 22459949

49. Diegoli TM, Rohde H, Borowski S, Krawczak M, Coble MD, Nothnagel M. Genetic mapping of 15 human X chromosomal forensic short tandem repeat (STR) loci by means of multi-core parallelization. Forensic Sci Int Genet. 2016; 25:39–44. https://doi.org/10.1016/j.fsigen.2016.07.004 PMID: 27497644

50. Pereira V, Gusmão L. The X-Chromosomal STRs in Forensic Genetics: X Chromosome STRs. In: Taylor & Francis Group, editor. Forensic DNA Analysis. Apple Academic Press; 2021. p. 21.

51. Machado FB, Medina-Acosta E. Genetic map of human X-linked microsatellites used in forensic practice. Forensic Sci Int Genet. 2009; 3(3):202–204. https://doi.org/10.1016/j.fsigen.2008.10.006 PMID: 19414170

52. Phillips C, Ballard D, Gill P, Court DS, Carracedo A, Lareu MV. The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. Forensic Sci Int Genet. 2012; 6(3):354–365. https://doi.org/10.1016/j.fsigen.2011.07.012 PMID: 21871851

53. Inturri S, Menegon S, Amoroso A, Torre C, Robino C. Linkage and linkage disequilibrium analysis of X-STRs in Italian families. Forensic Sci Int Genet. 2011; 5(2):152–154. https://doi.org/10.1016/j.fsigen.2010.10.012 PMID: 21087904

54. Peñalba JV, Wolf JBW. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet. 2020; 21(8):476–492. https://doi.org/10.1038/s41576-020-0240-1 PMID: 32472059

55. Graffelman J, Balding DJ, Gonzalez-Neira A, Bertranpetit J. Variation in estimated recombination rates across human populations. Hum Genet. 2007; 122(3-4):301–310. https://doi.org/10.1007/s00439-007-0391-6 PMID: 17609980

56. Xiao C, Yang X, Liu H, Liu C, Yu Z, Chen L, et al. Validation and forensic application of a new 19 X-STR loci multiplex system. Leg Med. 2021; 53:101957. https://doi.org/10.1016/j.legalmed.2021.101957

57. Perera N, Wijithalal R, Galhena G, Ranawaka G. Linkage, recombination and mutation rate analyses of 16 X-chromosomal STR loci in Sri Lankan Sinhalese pedigrees. Int J Legal Med. 2022; 136(2):415–422. https://doi.org/10.1007/s00414-021-02762-1 PMID: 35022841

58. Yang Q, Qian J, Shao C, Yao Y, Zhou Z, Xu H, et al. Identification and Characterization of Nine Novel X-Chromosomal Short Tandem Repeats on Xp21.1, Xq21.31, and Xq23 Regions. Front Genet. 2021; 12:784605. https://doi.org/10.3389/fgene.2021.784605 PMID: 34868274

59. Song F, Wei X, Zhou C, Wang S, Deng C, Liao M, et al. Resolving the recombination pattern of 38 X-STRs from Chinese Han three-generation pedigrees. Leg Med. 2022; 59:102135. https://doi.org/10.1016/j.legalmed.2022.102135 PMID: 36029693

60. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17(3):261–272. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543

61. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020; 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2 PMID: 32939066

62. Lam SK, Pitrou A, Seibert S. Numba: a LLVM-based Python JIT compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. No. Article 7 in LLVM'15. New York, NY, USA: Association for Computing Machinery; 2015. p. 1–6.

63. Rosenberg MS, Subramanian S, Kumar S. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol. 2003; 20(6):988–993. https://doi.org/10.1093/molbev/msg113 PMID: 12716982

64. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4(1). https://doi.org/10.1186/s13742-015-0047-8

65. Garcia FM, Bessa BGO, Dos Santos EVW, Pereira JDP, Alves LNR, Vianna LA, et al. Forensic Applications of Markers Present on the X Chromosome. Genes. 2022; 13(9). https://doi.org/10.3390/genes13091597 PMID: 36140765

66. Gettings KB, Kiesler KM, Vallone PM. Performance of a next generation sequencing SNP assay on degraded DNA. Forensic Sci Int Genet. 2015; 19:1–9. https://doi.org/10.1016/j.fsigen.2015.04.010 PMID: 26036183

67. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015; 18. https://doi.org/10.1016/j.fsigen.2015.02.003 PMID: 25716572

68. Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet. 2015; 18:49–65. https://doi.org/10.1016/j.fsigen.2015.05.012 PMID: 26013312

69. Amorim A, Pereira L. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic Sci Int. 2005; 150(1). https://doi.org/10.1016/j.forsciint.2004.06.018 PMID: 15837005

70. Yagasaki K, Mabuchi A, Higashino T, Wong JH, Nishida N, Fujimoto A, et al. Practical forensic use of kinship determination using high-density SNP profiling based on a microarray platform, focusing on low-quantity DNA. Forensic Sci Int Genet. 2022; 61:102752. https://doi.org/10.1016/j.fsigen.2022.102752 PMID: 35987117

71. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: Current methods, knowledge and practice. Forensic Sci Int Genet. 2021; 52:102474. https://doi.org/10.1016/j.fsigen.2021.102474 PMID: 33592389

72. Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, et al. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Sci Int Genet. 2017; 28:52–70. https://doi.org/10.1016/j.fsigen.2017.01.011 PMID: 28171784

73. Köcher S, Müller P, Berger B, Bodner M, Parson W, Roewer L, et al. Inter-laboratory validation study of the ForenSeq DNA Signature Prep Kit. Forensic Sci Int Genet. 2018; 36:77–85. https://doi.org/10.1016/j.fsigen.2018.05.007 PMID: 29945120

74. Zhang S, Bian Y, Chen A, Zheng H, Gao Y, Hou Y, et al. Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM. Forensic Sci Int Genet. 2017; 27:50–57. https://doi.org/10.1016/j.fsigen.2016.12.003 PMID: 27951431

75. Fattorini P, Previderé C, Carboni I, Marrubini G, Sorçaburu-Cigliero S, Grignani P, et al. Performance of the ForenSeq DNA Signature Prep kit on highly degraded samples. Electrophoresis. 2017; 38(8):1163–1174. https://doi.org/10.1002/elps.201600290 PMID: 28078776

76. Hollard C, Ausset L, Chantrel Y, Jullien S, Clot M, Faivre M, et al. Automation and developmental validation of the ForenSeq DNA Signature Preparation kit for high-throughput analysis in forensic laboratories. Forensic Sci Int Genet. 2019; 40:37–45. https://doi.org/10.1016/j.fsigen.2019.01.010 PMID: 30739830

77. Li R, Shen X, Chen H, Peng D, Wu R, Sun H. Developmental validation of the MGIEasy Signature Identification Library Prep Kit, an all-in-one multiplex system for forensic applications. Int J Legal Med. 2021; 135(3):739–753. https://doi.org/10.1007/s00414-021-02507-0 PMID: 33523251

78. Peck MA, Koeppel AF, Gorden EM, Bouchet JL, Heaton MC, Russell DA, et al. Internal validation of the ForenSeq kintelligence kit for application to forensic genetic genealogy. Forensic Genom. 2022; 2 (4):103–114. https://doi.org/10.1089/forensic.2022.0014