Leiden University

Bioinformatics

Recognition of hybrid sequences that are generated during PCR

Name:	Nathan Hoogendorp
Studentnr:	S1914499
Date:	20/08/2018
1st supervisor:	MSc J. Hoogenboom (NFI)
2nd supervisor:	ing. K.J. van der Gaag (NFI)
3th supervisor:	Dr. E.M. Bakker (LIACS)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In forensic DNA analysis, a selection of highly polymorphic loci with high discriminative power are used for personal identification of the donor(s) of forensic DNA samples. These loci are amplified with the polymerase chain reaction (PCR) in order to obtain a DNA concentration that is sufficient for detection. During PCR, not only exact copies of the targeted DNA, but also hybrid fragments are formed. The sequences of these hybrids are a combination of two almost identical DNA fragments or a duplication or insertion of a subfragment in a single sequence. These hybrids became apparent after massively parallel sequencing (MPS), in which the sequences are read and counted. Prior to this study, PCR hybrids remained the most predominant type of noise that was not handled by the Forensic DNA Sequencing Tools (FDSTools). FDSTools is a software package that is used for the analysis of massively parallel sequencing data in a forensic setting. With this software package it is possible to recognize and correct for Short Tandem Repeats (STR) stutter and other PCR or sequencing noise. In this study, two methods were developed to simulate *in silico* all possible hybrids given one or two parent fragment sequences. Subsequently, these hybrids were recognized and marked if observed in the data. Thereafter, a least-squares model in conjunction with a feature set that characterizes the formation of the hybrids was built to predict hybrid fragments in unobserved data. The developed hybrid prediction model is able to quantitatively correct the majority of hybrids that are present in the data. The hybrid marking and prediction modeling tools will be included in a future version of FDSTools.

Contents

1 Introduction

1.1 Forensic DNA analysis

In the field of forensic DNA analysis, polymerase chain reaction (PCR) is used to duplicate highly polymorphic loci to attain a DNA concentration that is sufficient for individual identification of a person(s) in forensic DNA samples. This identification is based on short tandem repeats (STRs), the combination of the number of times a specific STR appears on a locus is unique for each person. Currently, capillary electrophoresis (CE) is the industry standard for analysing DNA samples. However, a more sensitive and revealing method called massively parallel sequencing (MPS) is the state of the art method to use in forensic DNA analysis. It has a higher discriminative power and exclusion rate than the current industry standard CE method, hence why it is expected to become the method of choice in the following decade [1]. Stutter artefacts were a known problem in DNA analysis, these artefacts interfered with the identification of the true allele(s). The artefacts are formed when a repeat is inserted to or deleted from an STR in the fragment of the true alleles, this occurs during the multiplication of the DNA strand(s) in a PCR. A software package called the Forensic DNA Sequencing Tools (FDSTools) has been developed to recognize and subsequently correct the stutter [2]. Solving that complication has lead to the recognition of additional PCR artefacts which are referred to as hybrids. These artefacts can be observed in MPS data. However, this is a labour intensive job when done manually. In this process, mistakes are easily made due to the excessive amount of data sequencing generates.

1.2 How are hybrids created

The first description of PCR hybrids dates back to 1989, in that research A.R. Shuldiner et al. observed that during a PCR DNA fragments were formed that consisted of a combination of two highly similar fragments, which we will refer to as parent A and parent B [3]. Moreover, a hybrid is formed when a DNA polymerase enzyme starts extending the complementary DNA fragment of parent A but does not fully extend the fragment, this results in an incomplete extended fragment which can serve as DNA template in the next PCR cycle. This can occur in one of two ways, either the polymerase enzyme pauses on the DNA fragment it is copying or the enzyme prematurely terminates the extension process [4]. In the next PCR cycle the incomplete extended fragment can bind to parent B since it is highly similar to parent A. Subsequently, a DNA polymerase enzyme extends the incomplete fragment with nucleotides complementary to parent B. The result is a DNA fragment which is a combination of parent A and parent B i.e. a hybrid fragment. Both parent fragments are true alleles present in the DNA sample, i.e. not artefacts.

This process is schematically shown in figure 1 on page 5. In this figure an tube containing: DNA polymerase, double-stranded DNA fragments and primers is set into a PCR machine to multiply the DNA and increase the DNA concentration (Fig 1A). Figure 1B displays a close-up of the annealing step of the PCR. In the denaturation step all double-stranded fragments are melted down to single-stranded DNA fragments, one single-stranded fragment is shown (parent A) in this figure. Additionally, a primer (blue) attaches to the strand and thereafter a polymerase enzyme attaches itself to the primer site. In the extension step of the PCR cycle (Fig 1C), the DNA polymerase starts extending parent A. Generally, the enzyme fully extends the primer fragment, there is however a chance that the enzyme pauses or stops. This can occur since polymerase is only a moderately processive enzyme [4], the result is an incomplete extension. Subsequently, the denaturation step is shown in 1D. In this figure another fragment is present, (parent B, complementary side not shown) which only differs from parent A on position 13 and 25. These differences, displayed in red in figure 1D. The nucleotides, displayed in green in figure 1D are the nucleotides that appear in both fragments, and are referred to as a 'crossing-over' window. A hybrid fragment can be formed when the end of the incomplete fragment falls within the window. During the next PCR cycle all double stranded fragments are again melted to single stranded fragments. The incomplete fragment from figure 1C can operate as a primer, this is shown in figure 1E. In this figure the incomplete fragment binds to parent B, which is complementary identical except for the nucleotide at position 13. A polymerase enzyme will subsequently extend the incomplete fragment. The result is displayed in figure 1F. This figure shows three fragments: parent A, parent B and a fragment which is a combination of the complement of the parent fragments i.e. the hybrid fragment. In the following

PCR cycles this hybrid fragment can now be multiplied. The formed hybrid fragment contains one distinct nucleotide from parent A and one from parent B. Another hybrid fragment could have been formed if the incomplete fragment was formed with parent B and subsequently extended with parent A.

In theory, it is possible that more than two parents are involved in the hybrid creation process. However, this is a case we do not consider in this study because it is not likely to occur within the 29 PCR cycles that are used in this study. Furthermore, if these fragments were to be formed the number of times they would be observed in the data would be negligible.

There is an alternative process of creating hybrids which involves the polymerase enzyme to switch templates during one PCR cycle. However, this process requires a high DNA concentration which is not applicable to forensic casework [5]. Therefore, this process is not considered in this study. Based on the results of the study of Meyerhans et al. [4] we expect to detect 5% of hybrid fragments in each dataset.

An additional process of forming hybrids is based on K-mers in a single fragment. A K-mer is a subfragment with a length of K nucleotides. In order to form a hybrid based on this method at least two copies of the same K-mer needs to be present. One of those copies, and the portion of the fragment between (if any) can be deleted or duplicated during a PCR cycle which results in a hybrid. When the K-mers are immediately adjacent, the resulting fragment can also be identified as stutter. This hybrid creation process is shown in figure 2 on page 6. The first two steps are identical to step A, B from figure 1 (page 5), only the parent fragment differs. In this example the size of the K-mer is four nucleotides and the corresponding K-mer fragment is *GATA*. To create a hybrid with this method the primer needs to be extended up to and including the first appearance of the K-mer, this is shown in figure 2A. Thereafter the denaturation step creates single-stranded DNA fragments, this is displayed in figure 2B. In the next PCR cycle the incomplete fragment binds to the second K-mer of the same parent. This partially-bound fragment can function as a primer in the current PCR cycle, this process is shown in figure 2C. Subsequently, the incomplete fragment is extended and a fragment containing only one copy of the K-mer is produced, visualized in figure 2D. The result is a hybrid fragment that is shorter in length compared to the parent fragment. A longer fragment can also be obtained if the incomplete fragment was extended up to and including the second K-mer and subsequently bound to the first K-mer in the next PCR cycle.

1.3 Accomplishments of previous studies

An important field that conceivably could explain the underlying mechanism as to why and how these hybrids are formed is the field of thermodynamics. Software packages such as UNAFold have been developed to simulate hybridization of one or two single-stranded nucleic acid fragments, this simulation is based on free energy minimization and a full melting profile of the fragment(s) [6]. Unfortunately, as stated in the previous section, the fragments obtained in forensic DNA samples are highly similar. As a consequence the obtained thermodynamic values are also highly similar. Therefore, tools based on thermodynamics alone cannot be used to determine which hybrids are the most likely to be formed, starting on parent A and 'crossing-over' to parent B or vice versa.

PCR hybrids are characterized differently in the field of metagenomics, here hybrids are called 'chimeras' or 'recombinants'. A study of T. Kanagawa showed that the formation of chimeras can be avoided by limiting the number of PCR cycles [7]. This is not feasible in forensic casework since often only low concentration DNA samples are available, as a consequence 29 PCR cycles need to be performed in order to obtain sufficient sensitivity for forensic trace material. An alternative solution from the field of metagenomics is to remove the hybrids from the sequencing data by using a tool called Chimera slayer. This solution is not applicable since it is trained on a dataset containing 16S rRNA genes, and does not work properly when the parent fragments are highly similar [8].

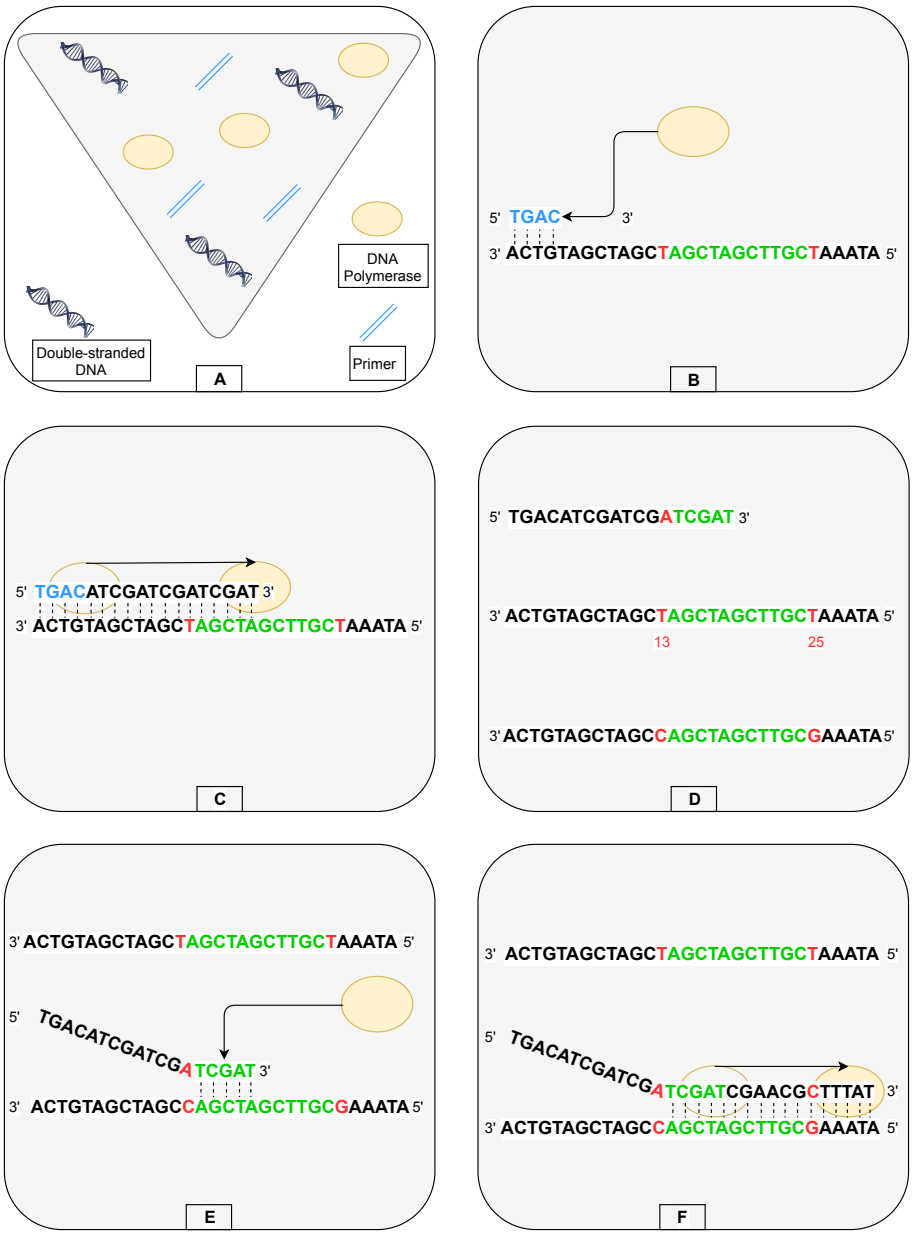


Figure 1: Overview of the creation of a hybrid fragment during PCR based on two single-stranded DNA fragments. **A)**: Bottom tip of an tube magnified. The fluid in the tip contains DNA fragments, primers and DNA polymerase. **B)**: Annealing step of PCR for a single-stranded DNA fragment. As result of the denaturation, two single-stranded DNA fragments are formed, the sequence of only one of those is shown. Subsequently, a primer binds to the fragment (blue sequence). Thereafter, a DNA polymerase enzyme attaches itself to the primer. **C)**: DNA polymerase starts extending the primer, creating the complementary DNA strand. The polymerase does not complete the extension, this results in an incomplete fragment. **D)**: Denaturation step of PCR, in this step all fragments are single-stranded. Another DNA fragment is present (complementary fragment not shown), this fragment differs on position 13 and 25 from the sequence above. These nucleotides are displayed in red, the section in between the distinct nucleotides shown in green is the region (window) were both sequences are identical. The shortest fragment is the incomplete extended fragment from C. **E)**: In the next annealing step, the incomplete fragment partially binds to another sequence. The incomplete sequence operates as primer for the sequence it is bound to. Subsequently, DNA polymerase attaches itself to the 'primer'. **F)**: The fragment is further extended, forming a combination of two fragments. In the following cycles this sequence can now also be multiplied.

1.4 Goal of this research

In this study we are going to design our own method to recognize, mark and predict hybrid artefacts in MPS data. This means that we have to simulate hybrids based on all possible combinations between different parent frag ble hybrids can subseq ly detected can then b ast-squares regression ı ; procedure could lead t Therefore, filtering th

Figure 2: Overview of the creation of a hybrid fragment during PCR based on one single-stranded DNA sequence. **A)**: The DNA fragment contains the K-mer *GATA*, which is present two times in the sequence (shown in red, the underscore marks the separate K-mers). The primer (blue fragment) is extended by polymerase up to and including the first appearance of the K-mer. This creates an incomplete complementary fragment. **B)**: Denaturation step of PCR, in this step all fragments are single-stranded. The shortest fragment is the incomplete extended fragment. **C)**: In the next annealing step, the incomplete fragment binds (shown in green) to the same fragment. However, in this cycle it binds to the second K-mer, resulting in a partial bind. The incomplete fragment operates as primer for the sequence it is bound to. Subsequently, DNA polymerase attaches itself to the 'primer'. **D)**: The fragment is extended and the result is a fragment that is four nucleotides shorter than the original fragment, this since it contains only one copy of the K-mer. In the next cycles this sequence can be multiplied.

2 Methods

This study is divided into three main sections:

1. Marking the possible hybrid sequences in MPS output data
2. Creating a feature dataset that characterizes the observed and not observed hybrids in the MPS data
3. Constructing and training a least-squares prediction model to correct for the hybrid artefacts

All scripts were written in the Python language and after validation will be included in the FDSTools software package [2]. The scripts were executed on a server with 8GB of memory and a total of 4 cores i.e. Intel(R) Xeon(R) Gold 6132 CPU with a processor base frequency of 2.60GHz. A virtual environment was set up as safety measure to prevent any alterations in the scripts from influencing the functionality of FDSTools because other research was conducted on the same server simultaneously.

2.1 Datasets

There were four DNA datasets available for which it was possible to mark hybrids and build prediction models, these datasets are described below:

1. Mitochondrial fragments (Mito): samples were analysed with the Mito-mini kit. This kit contains 10 (partially) overlapping fragments (markers) of the control region of the Mitochondrial genome. In total this dataset comprises of 37 mixed samples with an average sequence length of 126 nucleotides.
2. Mitochondrial fragments with low coverage (Mito-low): the samples in this dataset were also analysed with the Mito-mini kit, however the sequences are sequenced with lower coverage than Mito dataset. This results in less data (lower reads) compared to the Mito dataset. The Mito-low dataset consists of 15 mixed samples in which the average sequence length is 127 nucleotides.
3. STR fragments: samples were analysed with the PowerSeq Auto kit, this contains 22 autosomal markers, one Y-STR (for sex identification) and an Amelogenin marker (not an STR). A total of 465 samples were analysed, the sequences in these samples have an average sequence length of 160 nucleotides.
4. Microhaplotype/SNP fragments: the 1382 samples were analysed with a set of 19 unique microhaplotypes. These microhaplotypes are loci were single nucleotide polymorphisms (SNP's) are in close range of one another, some of the SNP's are tetra-allelic. This means that on the position of the SNP any of the four nitrogenous bases can be present [9]. In comparison to the other datasets this dataset contains the shortest sequences, the average length is 24 nucleotides, varying only a couple SNP's.

2.2 Marking possible hybrids in MPS data

The goal is to recognize which sequences present in an FDSTools output file could be explained as hybrid artefacts, and mark those and the corresponding parent sequences respectively. The FDSTools output file contains the name of the locus or fragment corresponding to each sequence, the DNA sequence, the total number of times that sequence has been observed (total reads), the total number of times the forward strand has been observed (total fw) and the total number of times the reverse strand has been observed (total rev). The algorithms that were used to simulate and mark the hybrids are displayed in algorithm 1, 2, page 9. Algorithm 1 and 2 are the two methods that are used to simulate all possible hybrid sequences. A condition is that only hybrids are simulated for parent sequences that have more than 100 reads, this since the creation of hybrids is unlikely to occur for sequences below this threshold with the current laboratory protocols. For the analysis of the hybrids the complementary fragment (i.e. the same orientation as the parents) is always used, this makes it more straightforward to compare it to the parent fragments. It is possible that hybrids are

simulated but not observed due to the depth at which the sequences are sequenced. A hybrid sequence can be simulated in one of two ways:

- 'Crossing-over' method (algorithm 1): This method makes all possible combinations of two single-stranded DNA sequences per marker/fragment and always sets one sequence as reference sequence and the other sequence to 'other' sequence. The difference in nucleotides between each combination is then calculated with the `call_variants` function that is built into FDSTools. This function projects all the differences in nucleotides onto the reference sequence. The output of the function contains the position, the nucleotide of the reference sequence and the nucleotide of the other sequence. An example of the output is 75C>G, this translates to: the 75^{th} position of the reference sequence contains a *C* and the other sequence a *G* nucleotide. Two hybrid sequences can be formed by 'crossing-over' from one parent sequence to the other between every adjacent differing nucleotide pair, one by 'crossing-over' from parent A to parent B and one for the reverse.

- K-mer method (algorithm 2): The K-mer method uses a sliding window of fixed length (K) to obtain subfragments (K-mers), if a subfragment is present more than once in the same sequence it is used for further analysis. In this study we have set the K-sizes accordingly:

  - Mito: K-size = 6
  - Mito-low: K-size = 6
  - STR: K-size = 10
  - Microhaplotype: K-size = 4

  The sizes of K are defined based on the most equal distribution of the number of observed and not observed hybrids that could be obtained for K-sizes in the range of 2 to 15. This process has been performed for two samples per dataset. Pairs of occurrences of K-mers are divided into three categories, the first category contains pairs of K-mers which have other nucleotides between them, this is defined as the 'apart' category. Furthermore there can be K-mers which have overlapping nucleotides this is defined as the 'overlap' category and the last category contains the K-mers which are 'adjacent'. The K-mers of the latter category can also be identified as STRs when the size of K is between two and five. All these categories are schematically displayed in figure 3. In that example, a total of six hybrids can be simulated from the three categories, one where the K-mer is inserted and one where the K-mer is deleted from the DNA fragment in each category. If the sequences are apart from each other, the fragment between the two copies of the K-mer is duplicated or deleted, along with one copy of the K-mer. For the overlapping category the nucleotides that overlap are duplicated or deleted and for the adjacent category one repeat is inserted or deleted.

Figure 3: Categories of K-mer pairs. DNA sequence (black line) containing different categories of K-mers which are shown as coloured blocks. The orange blocks (#1) illustrate the category 'apart', in this category the unique sequences have some nucleotides between them. The red blocks (#2) show the K-mers which have a 'overlapping' region and the green blocks (#3) are 'adjacent' i.e. STRs.

The possible hybrids are subsequently identified in the input file and are thereafter verified if the number of total reads of the hybrid is less than the total reads of the parent(s). Provided that these conditions are true the hybrid and corresponding parent(s) are marked with a unique number, e.g. `'Hybrid' [1], 'parent A' [1], 'parent B' [1]`. This shows that Hybrid 1 has two parents and was thus simulated by the *'crossing-over'* method, if it had only a parent A it would have been simulated with the *K-mer* method. The unique numbering of the hybrids is reset for each marker in the sample. In addition to the marking of the hybrids and parent(s) with a unique number, a percentage of total reads of the hybrid relative to the total reads of its parent(s) is calculated, which we call the hybrid ratio. The equation of the hybrid ratio is shown in equation 1. If a hybrid is simulated with the K-mer method then the number of reads of parent A is multiplied by two in order to calculate the hybrid ratio since there is no parent B. An example of the output is displayed as `['Hybrid', [1], 'Percentage reads parent A:', [['8.01']], 'Percentage reads parent B:', [['8.30']]]`. In this example the hybrid contains 8.01% of the total hybrids reads relative to parent A and 8.30% to parent B. When there is only a parent A the percentage of parent B is set to zero. The displayed percentages are the hybrid ratios multiplied by 100.

$$Hybrid\ ratio\ =\ \frac{Reads\ hybrid}{Reads\ parent\ A\ +\ Reads\ parent\ B} \tag{1}$$

Algorithm 1 Simulating possible hybrid sequences ('crossing-over' method)

Remove all sequences per marker that have less than 100 reads
for Every pair of two sequences (parent A, parent B) **do**
    Determine positions of all differences between the two parent sequences using one sequence as reference. ▷ Reference = parent A, Other = parent B
    for Every determined adjacent pair of nucleotide positions (pos1, pos2) **do**
        simulate hybrid: Start with parent A up to and including pos1, extend with parent B starting immediately after pos1
        simulate hybrid: Start with parent B up to and including pos1, extend with parent A starting immediately after pos1.
return Simulated hybrids

Algorithm 2 Simulating possible hybrid sequences (K-mer method)

for Every parent with more than 100 reads **do**
    Use sliding window to detect K-mers in sequence.
    for Every unique K-mer with count K-mer >= 2 **do**
        for Every pair of occurrences **do**
            if Occurrences of the K-mer is adjacent **then**
                Simulate hybrid: Find first appearance of K-mer and insert K-mer to parent sequence.
                Simulate hybrid: Find first appearance of K-mer and delete it from parent sequence.
            if Occurrences of the K-mer overlap **then**
                Simulate hybrid: Find first appearance of overlapping nucleotides and duplicate the overlapping section.
                Simulate hybrid: Find first appearance of overlapping nucleotides and delete it from parent sequence.
            if Occurrences of the K-mer is apart **then**
                Simulate hybrid: Find first appearance of K-mers apart from each other and insert section starting after first K-mer up to and including second K-mer and insert it after second K-mer.
                Simulate hybrid: Find first appearance of K-mers apart from each other and delete section from first K-mer up to second K-mer.
    return Simulated hybrids

In order to determine the percentage of hybrid sequences per dataset it is necessary to examine the total number of reads of observed hybrids per sample. Subsequently, a percentage of observed hybrids per sample can be obtained by dividing the total number of hybrid reads by the sum of the hybrids reads and the reads of the corresponding parent(s), this is displayed in equation 2. The total percentage of hybrid sequences per dataset is the average of all hybrid percentages per sample. If a sample does not contain any observed hybrids the percentage of this sample will be set to zero since a division by zero is undefined.

$$\%\ Observed\ hybrids = \frac{hybrid\ reads}{hybrid\ reads + parent(s)\ reads} \times 100 \tag{2}$$

2.3 Defining features

In order to build and train a hybrid prediction model it is necessary to define features that characterize the formation of the hybrids and the corresponding parents. These features need to contain the necessary information to predict the ratio of total reads of the hybrid sequence relative to the total reads of the parent(s), i.e. the hybrid ratio. All possible hybrids are simulated as described in the previous section 2.2 (page 7), except now only the two highest (most reads) sequences per marker instead of all sequences with more than 100 reads will be used. This because there are always more possible hybrids than will be observed (positive examples), the majority will therefore not be observed (negative examples). By including only the two highest sequences the amount of not observed sequences will be minimized. This is beneficial for the model since it would otherwise overfit on the negative examples and not be able to classify the positive examples. The hybrid ratios of the negative examples were set to zero.

In total, 29 features have been designed, these features were also squared, cubed, raised to the negative power 1, 2 and 3 in order to acquire different polynomial fits on the dataset. If a feature is raised to a negative power while zero, it was set to zero. A column of ones was added to the features in order to prevent the fit of the model being forced through the origin. All features were standardized prior to being used in the prediction model except the column of ones, these would otherwise be transformed to zeros (non-informative). The rest of the features are centered around 0 and have variance in the same order. In total, the number of features that were used is 175, an overview of all 29 features including an explanation is shown in appendix A, table 1 on page I.

2.3.1 Feature reduction

The challenge in implementing any machine learning method is to design features that explain the to-be-predicted variable. It is difficult to determine whether a single feature or a (sub)set of features contribute to a correct prediction. Therefore, we implemented a genetic algorithm [10] to test the contribution of the features to the accuracy of the prediction. The output of the algorithm is a subset of features that resulted in the lowest prediction error. The algorithm calculates a prediction error using all features (feature vector) and uses that value as baseline. Subsequently, the feature vector (subset) is randomly mutated (in- or excluding features), the error of this subset is then determined and compared to the error of the baseline. This process is performed 50 times (generations) using a population size of 100 individuals, meaning 100 mutually independent mutations are simulated every generation. In each generation the error for every individual is compared to the baseline error, only the individuals which have a lower error than the baseline error are saved and used in the next generation (fittest population). In the next generations the errors are compared to the lowest errors that were obtained in the previous generations. This feature selection method is performed for every dataset.

The genetic algorithm has slightly been modified so it could be applied to our data. The algorithm tries to predict class labels and therefore cannot be applied to a continuous dataset i.e. hybrid ratio. The logistic regression component of the algorithm has been replaced with a least-squares regression method, instead of optimizing the classification accuracy the model's performance is assessed by calculating the mean-squared error, displayed in equation 3. In this equation $\hat{y}_i$ is the predicted value of the i^{th} sample and y the actual hybrid ratio.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

2.4 Prediction model

In order to predict hybrid sequences in MPS data, least-squares regression models were trained for each dataset, separately in each individual marker (hereafter referred to as the overlay model) and for all markers combined (combined model). These models require the feature matrix, weights and the to be predicted variable which is the hybrid ratio. These matrices are schematically shown in table 1. The leftmost table is an $M \times N$ feature matrix containing one feature per column and a hybrid per row, the middle table (an $N \times 1$ column vector) contains the weights which will be calculated by the model and the rightmost table (an $M \times 1$ column vector) contains the to be predicted variable. This is a common representation of systems of linear equations. Such a system often does not have a solution, therefore the calculated weights can only approximate the hybrid ratio. The approximation is based on the minimization of the mean squared error, which is calculated as shown in equation 4. In this equation $\hat{\beta}$ is the predicted weight vector that minimizes the difference between the actual hybrid ratio (y) and the predicted hybrid ratio. In order to build the prediction models it is necessary to split the data in training and test sets, this is performed with the `train_test_split` function of the Python module *scikit-learn*. The latter function divides the data in a 80% training and 20% test set. Python's pseudorandom number generator was seeded with a fixed number.

Table 1: Least-squares model example illustrating a feature matrix (leftmost table), weights matrix (middle table) and hybrid ratio matrix (rightmost table). The hybrid ratios are the to be predicted variable.

M x N	**Feature1**	**Feature2**	**FeatureN**
Hybrid1	X_{11}	X_{12}	X_{1n}
Hybrid2	X_{21}	X_{22}	X_{2n}
HybridM	X_{m1}	X_{m2}	X_{mn}

$\cdot$

Weights
β_1
β_2
β_n

$=$

Hybrid ratio
y_1
y_2
y_n

$$\hat{\beta} = arg\ \min_{\beta} \left[\|y - X\beta\|^2\right] \tag{4}$$

To asses the quality of the obtained regression fits a coefficient of determination (R^2) and explained variance score are calculated per model. The R^2 equation is shown in equation 5, where $\hat{y}_i$ is the predicted value of the i^{th} sample and $\bar{y}$ is the mean. The explained variance equation is shown in equation 6 (page 12), here $\hat{y}$ is the predicted target output and y the correct output. For both measures, a score of one is for both measures the best possible result.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$Explained\ variance = 1 - \frac{Var(y-\hat{y})}{Var(y)} \tag{6}$$

2.4.1 Averaging hybrid ratio

The hybrids that are simulated with the K-mer method can have a relatively high hybrid ratio since these hybrids can also be classified as stutter, which are known to reach ratios of up to 20% [2]. We have therefore chosen to divide the ratios and features of these hybrids by the total number of times this hybrid can be simulated. This is illustrated in example in figure 4. This figure shows a parent sequence with three repeats and a hybrid sequence with two repeats, either the first, second or third repeat has been deleted from the parent sequence in order to simulate the hybrid. Therefore, it is possible to simulate the same hybrid three times. As a consequence, the hybrid ratio needs to be divided by three because the observed amount of the hybrid is triple the amount obtained by deleting any one repeat.

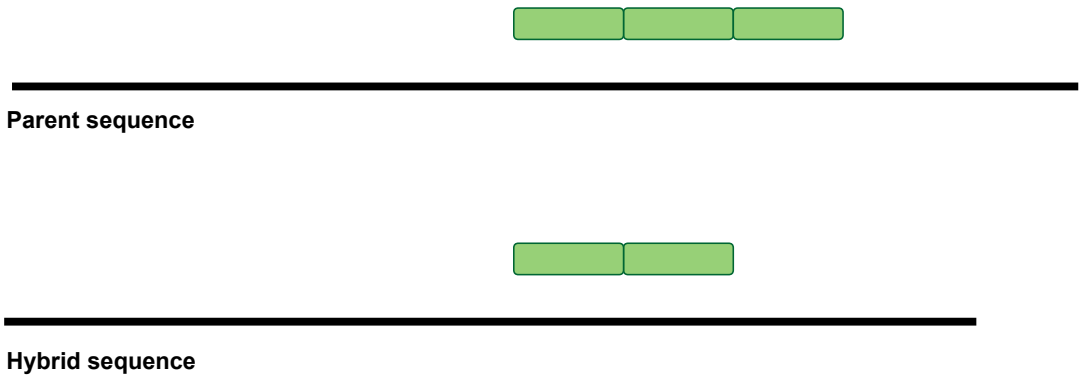


Figure 4: parent and hybrid sequence with K-mers. The K-mer is repeated three times in the parent sequence and two times in the hybrid, the hybrid can be simulated in three different ways.

3 Results & Discussion

In this chapter the obtained results are displayed and discussed. The first subsection describes the results of marking hybrid sequences in MPS data. Subsequently, the second subsection discusses the results of creating and evaluating the features that are used as input for the hybrid prediction model. To conclude, in the third subsection the hybrid prediction model and results are examined.

3.1 Marking hybrid sequences

3.1.1 Total number of observed and not observed hybrids

In order to mark hybrid sequences in MPS data all possible sequence combinations needed to be simulated. This has been accomplished by using two methods, the K-mer method and the 'crossing-over' method (section 2.2, page 7). If a certain hybrid was observed in the data it was marked as 'Hybrid' and the corresponding parent(s) as 'parent A' or 'parent B'. The hybrid ratio (percentage of hybrid reads relative to the parent(s)) was calculated and included in the marking of the hybrids. An overview of the total number of marked hybrids and the corresponding method is plotted per dataset and visualised as bar charts. It can be observed in figure 5 on page 15 that most of the hybrids for the Mito dataset were simulated with the 'crossing-over' method; a total of 28229 compared to a total of 6463 hybrids which were simulated with the K-mer method. However, the majority of the hybrids were not observed, 85.56% for the 'crossing-over' method compared to 94.40% of the K-mer method. This resulted in a total of 4076 marked hybrids for the 'crossing-over' method and 362 for the K-mer method respectively. In addition, the 'crossing-over' method contributed 91.84% to the total number of observed hybrids and the K-mer method thus 8.16%. It is apparent that more possible hybrids were simulated than there are found in the samples and that the 'crossing-over' method contributed to the majority of the observed hybrids. Therefore, we can conclude that the sequences in the Mito dataset contain more variation between pairs of parent sequences than there are K-mers present per parent. This is as expected since the sequences in the dataset do not contain a substantial amount of identical repetitive subfragments. These results confirms that both methods perform accordingly.

The Mito-low dataset contained similar sequences to the Mito dataset. However, in this dataset sequencing coverage is lower, this translates to the sequences having less reads compared to the Mito dataset. In addition, there is less variation because fewer sequences are read. It can be observed in figure 6 on page 15 that a total of 2887 hybrids were simulated of which 25.60% by the 'crossing-over' method and 74.40% by the K-mer method. Striking is that 51.42% of the total number of hybrids that were simulated with the 'crossing-over' method were observed while this was only 14.44% for the Mito dataset. The number of not observed hybrids simulated with the K-mer method i.e. 84.63% is still higher compared to the number of observed hybrids i.e. 31.16%. Consequently, the 'crossing-over' method contributed for 68.84% to the total number of observed hybrids and the K-mer method thus 31.16%. In addition, per method the percentage of observed hybrids for the K-mer method is 8.01%.

It can be observed in figure 7 on page 16 that in the STR dataset more observed and unobserved hybrids were simulated with the K-mer method than with the 'crossing-over' method. In total 147704 hybrids were simulated of which 15.73% with the 'crossing-over' method and 84.27% with the K-mer method. This difference can be explained due to the fact that the sequences in the STR dataset contain a large number of repetitive substring sequences. The K-mer method searches for these substrings (K-mers) and subsequently simulates hybrid sequences based on in which category the K-mer falls (section 2.2, page 8). In this dataset a total of 26.52% of all hybrids hybrids were simulated with the K-mer method that are observed and 57.75% that are not observed, compared to 2.86% and 12.87% hybrids that were simulated with the 'crossing-over' method respectively. The K-mer method contributed for 90.27% to the total number of observed hybrids and the 'crossing-over' method therefore 9.73%. It can be stated that for this dataset in particular the K-mer method really contributed to the number of detected hybrids.

An overview of the observed and not observed hybrids for the Microhaplotype dataset is shown in figure 8, page 16. In total 37591 were simulated of which 94.88% with the K-mer method and 5.12% with the 'crossing-over' method. It is evident that for this dataset a lot of possible hybrids were simulated with the K-mer method compared to the 'crossing-over' method. However, the majority of the hybrids of the K-mer method were not observed i.e. 98.36% as opposed to 44.92% of the hybrids that were simulated with the 'crossing-over' method. A possible explanation as to why the number of possible K-mer hybrids is high, is the window size of the K-mer method. The sequences in this dataset are approximately 24 nucleotides long, this limited the maximum size of **K** to 4, otherwise no repetitive substring sequences could be detected. As a consequence, it can be stated that the K-mer method is unsuitable for this dataset.

In the Microhaplotype dataset the number of sequences per marker is limited which restrained the number of hybrids that could be identified in the data. Nevertheless, the K-mer method contributed for 44.92%, which is comparable to the contribution of the 'crossing-over' method. Furthermore, 70.11% of the hybrids simulated by the 'crossing-over' method were observed, whereas only 3.08% of the hybrids simulated by the K-mer method were observed.

An overview of all aforementioned percentages is illustrated in table 2 on page 17. This table shows that for the Mito, Mito-low and Microhaplotype dataset the majority of the hybrids were simulated with the 'crossing-over' method (91.84%, 68.84% and 55.08% respectively) and that for the STR dataset the K-mer method simulated the most observed hybrids (90.27%). Additionally, only for the Microhaplotype and Mito-low dataset, the majority of the hybrids that were simulated with the 'crossing-method' were actually observed i.e. 70.11% and 51.42% respectively.

In order to visualize the number of observed and not observed hybrids per dataset a scatter plot has been created, this plot is shown in figure 9 on page 17. On the vertical axis of the plot the number of observed hybrids are displayed and on the horizontal axis the number of not observed hybrids. The circles in the plot represent all hybrid sequences that have been simulated per dataset, the squares represent the hybrids that have been simulated with the K-mer method and the triangles the hybrids that have been simulated with the 'crossing-over' method. It is apparent that the STR dataset contains the most observed and not observed hybrids, the majority of these hybrids were simulated with the K-mer method. The 'crossing-over' method of the STR dataset simulated approximately an equal amount of observed hybrids relative to the 'crossing-over' method of the Mito dataset. For the Microhaplotype dataset each method contributed approximately equally the number of observed hybrids, this in contrast to the number of not observed hybrids which were predominantly simulated with the K-mer method. The only dataset where the number of observed and not observed are to some extend within the same range is the Mito-low dataset. This because the data points is close to the reference line, this line shows were the number of observed and not observed hybrids are equal ($y = x$). The Mito and Microhaplotype dataset contain 12.79% and 6.52% observed hybrids relative to the percentage of not of observed hybrids.

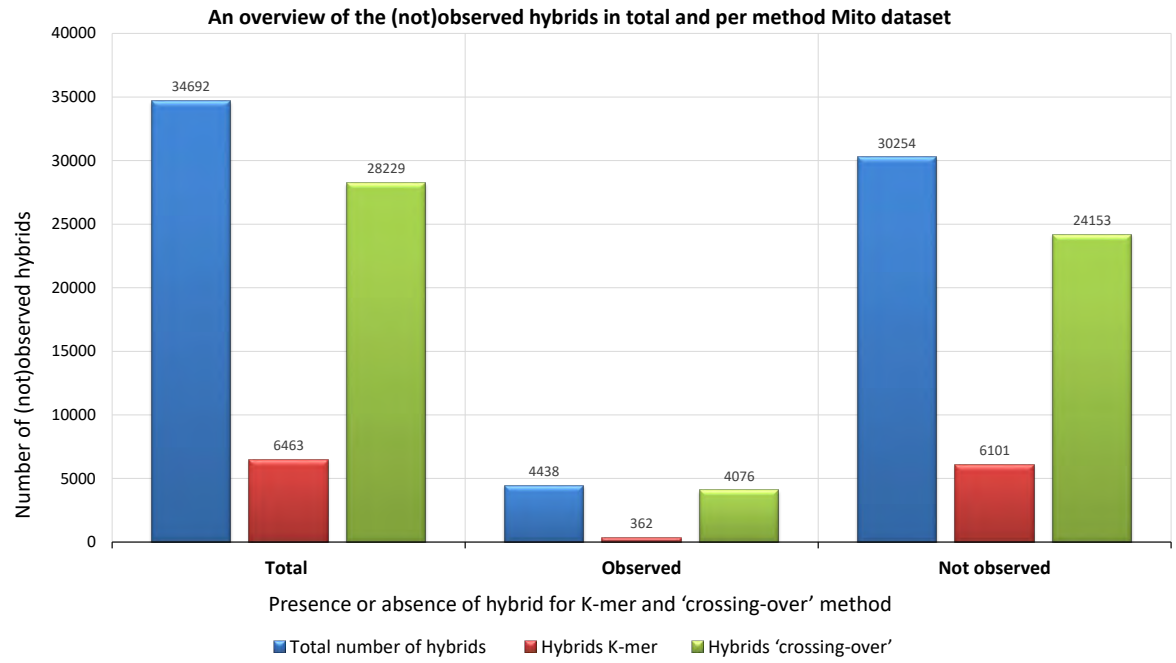


Figure 5: Number of total, observed and not observed hybrids in the Mito dataset.

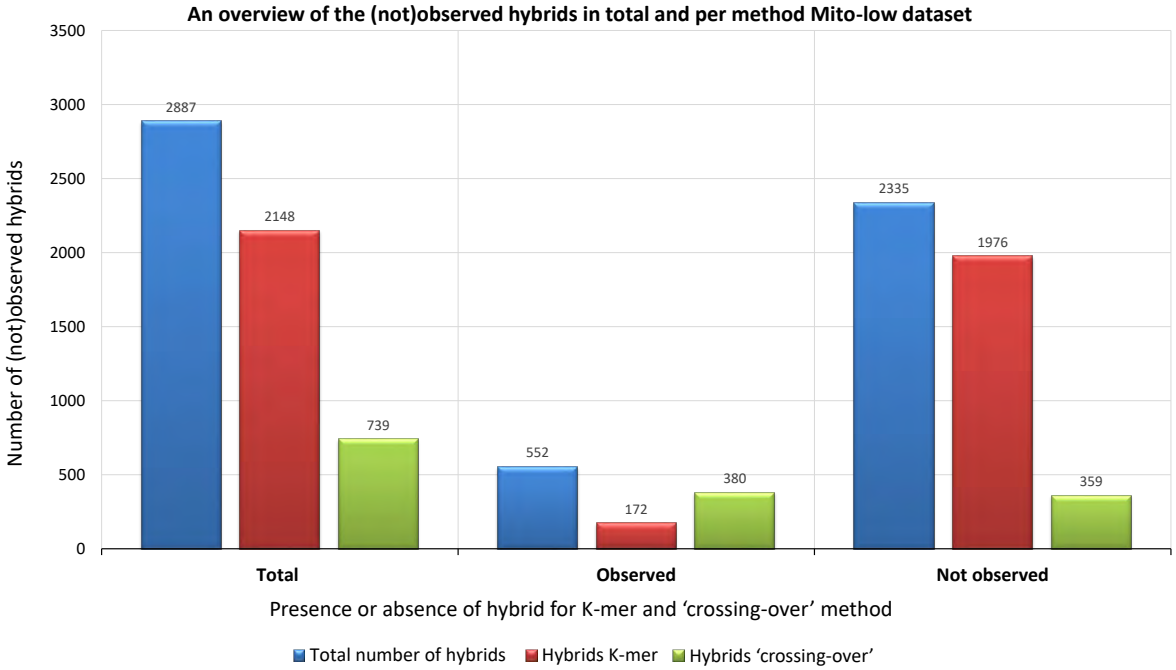


Figure 6: Number of total, observed and not observed hybrids in the Mito-low dataset.

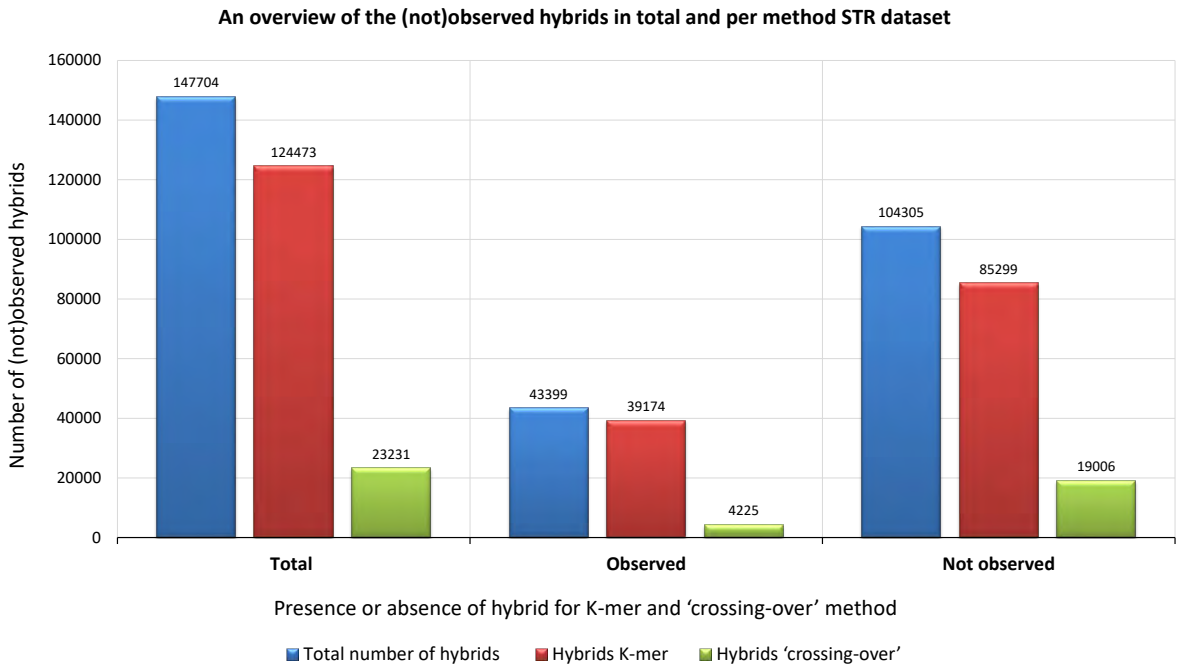


Figure 7: Number of total, observed and not observed hybrids in the STR dataset.

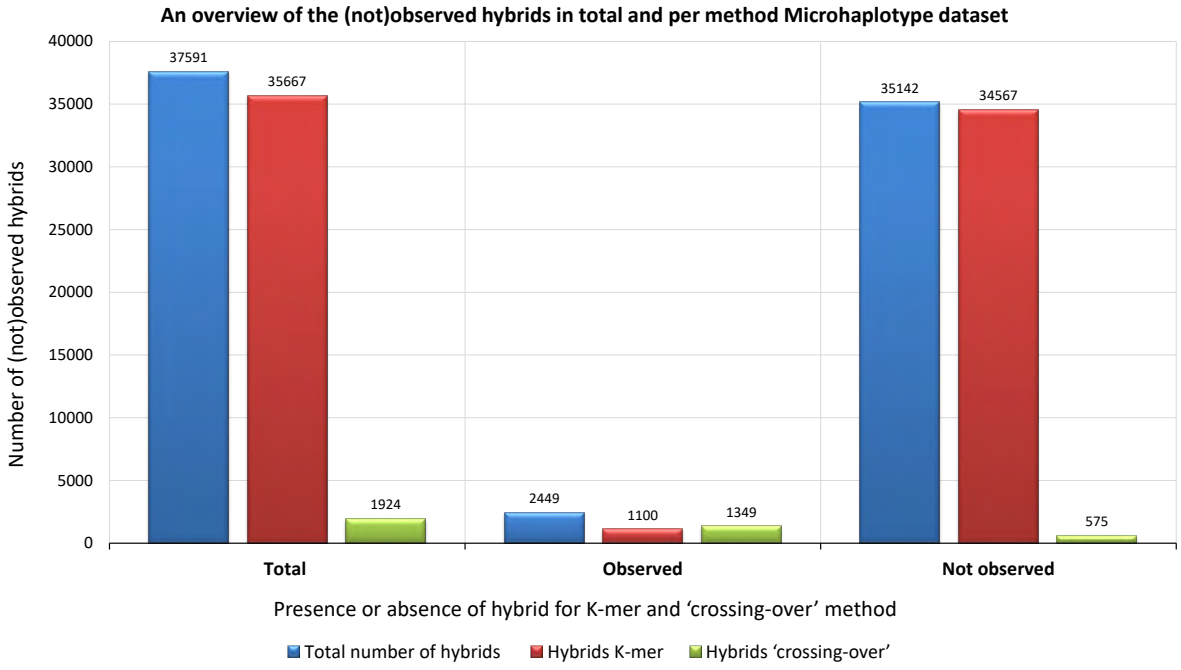


Figure 8: Number of total, observed and not observed hybrids in the Microhaplotype dataset.

Table 2: Overview of percentages of observed hybrids relative to the total number of observed hybrids and to the total number of hybrids per method.

	Observed w.r.t. total observed	**Observed w.r.t total simulated per method**
K-mer method Mito	8.16	5.60
K-mer method Mito low	31.16	8.01
K-mer method STR	90.27	31.47
K-mer method Microhaplotype	44.92	3.08
'Crossing-over' method Mito	91.84	14.44
'Crossing-over' method Mito low	68.84	51.42
'Crossing-over' method STR	9.73	18.18
'Crossing-over' method Microhaplotype	55.08	70.11

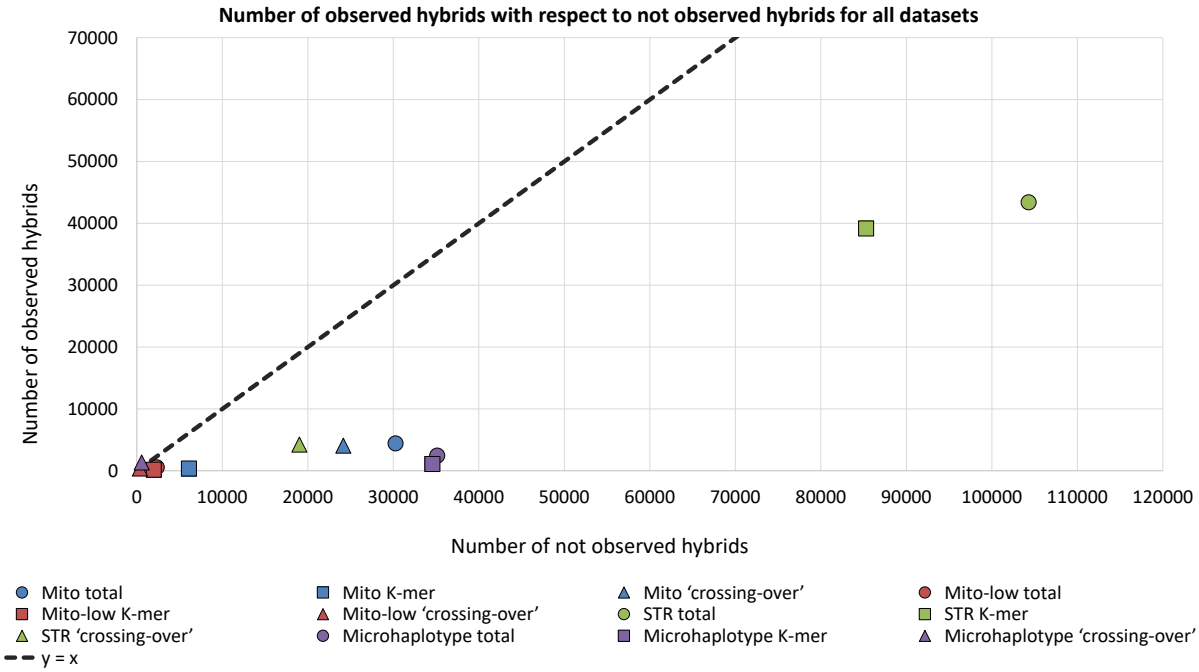


Figure 9: Overview of total number of observed hybrids w.r.t. not observed hybrids for each dataset. The total number of observed and not observed hybrids per dataset are displayed as circles, the number of hybrids simulated with the K-mer method as squares and the number of hybrids simulated with the 'crossing-over' method as triangles.

3.1.2 Percentage of hybrids per dataset

To determine whether the datasets contain 5% hybrid sequences as approximated by Meyerhans et al. [4], the percentage of observed hybrid sequences were first calculated per sample per dataset (equation 2, page 10). These percentages per sample were subsequently averaged to obtain the percentage of hybrids per dataset. An overview of the percentages per dataset is shown in table 3. This table shows that all datasets contain less than 5% hybrids, this can be explained by the fact that the 5% of observed hybrids as described by Meyerhans et al. represents an upper limit because they used high input DNA concentrations. The datasets with the lowest and highest percentage of hybrids are the Mito-low and STR dataset containing a percentage of 0.82 and 2.56 respectively. The Mito-low dataset is the dataset with the lowest concentration and coverage, it was therefore expected that this dataset would contain the smallest percentage of observed hybrids. In contrast, as described in section 3.1.1 (page 13), the STR dataset contains the most number of observed hybrids relative to the total number of simulated hybrids. Therefore it was expected that this dataset would contain the highest percentage of observed hybrids. Concerning the Mito and Microhaplotype dataset, they contain approximately the same percentage of hybrids. It was not expected that the Microhaplotype dataset would contain more observed hybrids compared to the Mito dataset since there is less variation between sequences of that dataset. Apparently, the difference in reads of the observed hybrids and corresponding parent(s) of the Microhaplotype dataset is less in comparison to the Mito dataset.

Table 3: Percentages of observed hybrids per dataset.

Dataset	**Percentage of observed hybrids**
Mito	1.43
Mito-low	0.82
STR	2.96
Microhaplotype	1.56

3.1.3 Total hybrids per marker Mito dataset

In the previous section we analysed how many hybrids were simulated in total with the K-mer and 'crossing-over' method and how many of these hybrids were detected in the datasets. To gain more insight in the data we have plotted the number of (not)observed hybrids with respect to the method the hybrid was simulated with, for each marker. This shows for which marker the most or least hybrids were simulated and the method it was simulated with. Figure 10 on page 19 displays the results of the Mito dataset. It can be seen that for the 'crossing-over' method most observed hybrids were simulated within *fragment 5*, this fragment contributed for 43.60% to the total number of observed hybrids for this method. The markers that simulated the second and third most observed hybrids were *fragments 7* and *8* which contributed 19.43% and 18.03% respectively. In this dataset there was only one fragment which did not produce any observed hybrids with the 'crossing-over' method, this was *fragment 9*. For the K-mer method most of the hybrids were simulated within *fragment 8*, this fragment accounted for 69.16% of the total number of observed hybrids for this method. There were three other markers *fragments 2, 7* and *10* for which the K-mer method simulated actually-observed hybrids, the remaining markers contributed therefore zero percent.

In order to determine the value of each hybrid simulation method, the number of unique observed hybrids that were simulated with the 'crossing-over' and K-mer method were examined. In the latter analysis, a hybrid is considered unique if it is exclusively simulated with one method. The results of this analysis can be seen in table 4, page 19. It is evident that both methods simulate unique hybrids, only for *fragment 8* and *10* there are hybrids i.e. 32.75% and 23.07% respectively that can be simulated with both methods. However, if hybrids were solely simulated with the 'crossing-over' method, a total of 222 hybrids would not have been recognised and marked.

The majority of the not observed hybrids for the 'crossing-over' method and K-mer method were simulated within *fragment 5* and *fragment 8* respectively. These fragments contributed for 33.13% and 28.31% to the total number of not observed hybrids for the latter specified methods. Nonetheless, both methods are still

suitable for these markers since 18% of *fragment 5* is observed for the 'crossing-over' method and 12.02% of *fragment 8* for the K-mer method. An overview of the number of observed hybrids versus the number of not observed hybrids per marker is displayed in figure 11 on page 20. The horizontal axis range (observed hybrids) is smaller than the vertical axis range (not observed hybrids). This demonstrates that the number of not observed hybrids is higher per marker than the number of observed hybrids. The only fragment that contained a relatively high amount of observed hybrids was *fragment 8* with 33.23% observed.

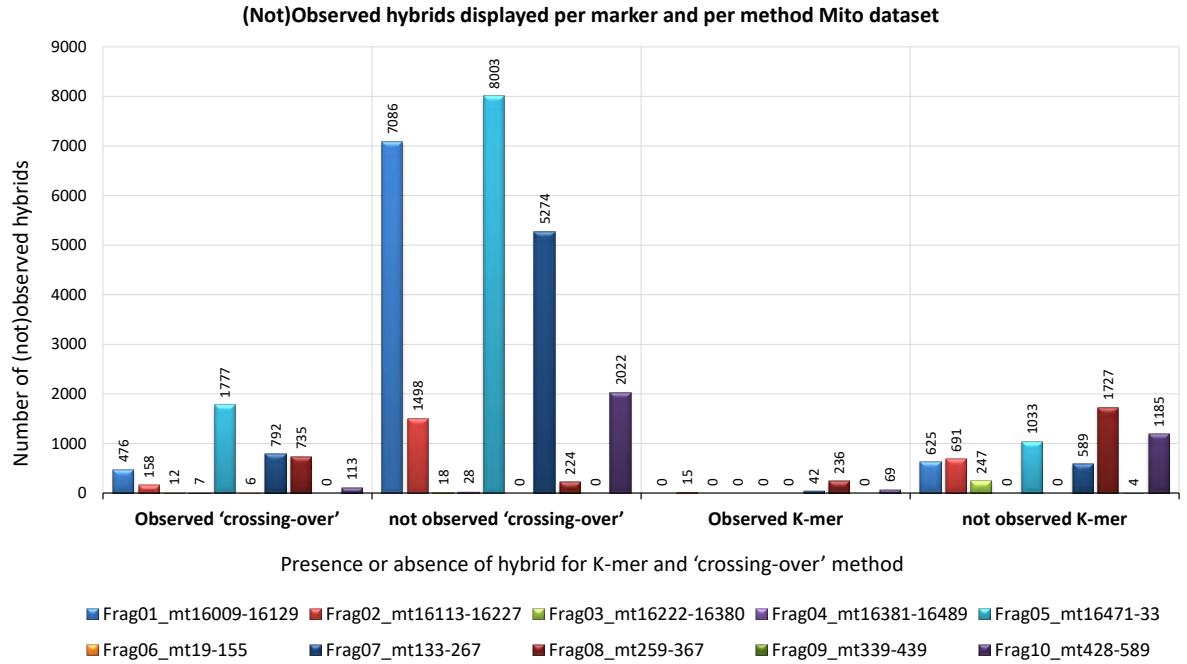


Figure 10: Number of observed and not observed hybrids in the Mito dataset, per marker for the K-mer and 'crossing-over' method.

Table 4: Number of observed hybrids simulated with 'crossing-over' and K-mer method per fragment in the Mito dataset.

	Unique 'crossing-over' method	**Unique K-mer method**	**Non-unique**
Frag01_mt16009-16129	476	0	0
Frag02_mt16113-16227	158	15	0
Frag03_mt16222-16380	12	0	0
Frag04_mt16381-16489	7	0	0
Frag05_mt16471-33	1777	0	0
Frag06_mt19-155	6	0	0
Frag07_mt133-267	792	42	0
Frag08_mt259-367	521	132	318
Frag10_mt428-589	92	48	42

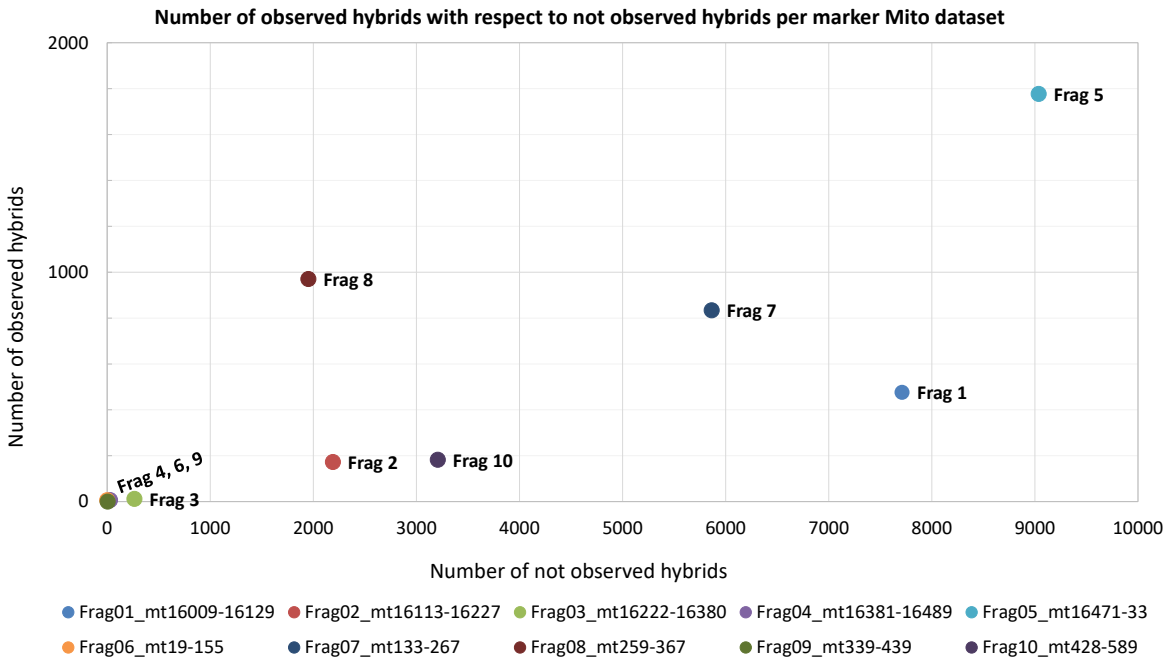


Figure 11: Number of observed w.r.t. not observed hybrids in the Mito dataset, per marker.

3.1.4 Total hybrids per marker Mito-low dataset

The K-mer results of the Mito-low dataset are comparable to the Mito dataset, this can be observed by comparing the right half of figure 12 (page, 21) with that of figure 10 on page 19. In this dataset *fragment 8* contributed 57.56% to the total number of observed K-mers compared to 65.19% in the Mito dataset. For the 'crossing-over' method most of the hybrids were also simulated within *fragment 8*, for this method the fragment contributed for 55.53% to the number of observed hybrids. An additional marker for which a relatively high amount of actually-seen hybrids was simulated is *fragment 2*, i.e. 24.21%. In order to asses the value of the methods, the numbers of unique observed hybrids were calculated. This is displayed in table 5 on page 21. The majority of the unique hybrid sequences were simulated with the 'crossing-over' method i.e. 53.44%, the K-mer method simulated 21.92% unique observed hybrids. This is still a substantial percentage relative to the total number of observed hybrids. Conclusively, the K-mer method can effectively be applied to the Mito-low dataset.

For the not observed hybrids, the larger part was simulated within *fragment 10* i.e 33.13% for the 'crossing-over' method and *fragment 8* i.e 28.31% for the K-mer method. However, within *fragment 10* a total of 7.39% hybrids were observed, this was 11.57% for *fragment 8*. It is therefore necessary to apply both methods for marking the hybrids in the Mito-low dataset. The number of observed hybrids versus not observed hybrids have been plotted in figure 13 on page 22. It can be seen in this figure that *fragment 7* simulates the most observed hybrids relative to the not observed hybrids i.e. 28.57% to 73.43% hybrids respectively. In addition, the percentages of observed hybrids is similar for *fragment 2* and *8* i.e. 19.19% and 28.11%. It is apparent that for all markers in this dataset more not observed that observed hybrids were simulated.

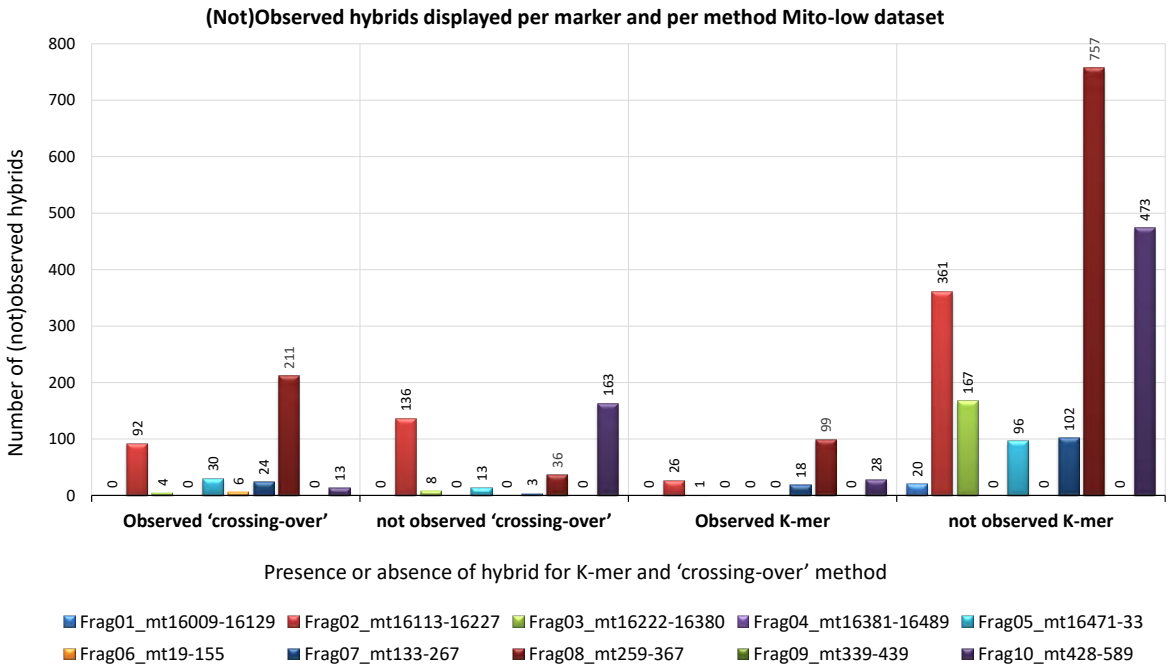


Figure 12: Number of observed and not observed hybrids in the Mito-low dataset, per marker for the K-mer and 'crossing-over' method.

Table 5: Number of observed hybrids simulated with 'crossing-over' and K-mer method per fragment for Mito-low dataset.

	Unique 'crossing-over' method	**Unique K-mer method**	**Non-unique**
Frag02_mt16113-16227	82	21	15
Frag03_mt16222-16380	4	1	0
Frag05_mt16471-33	30	0	0
Frag06_mt19-155	6	0	0
Frag07_mt133-267	24	18	0
Frag08_mt259-367	136	53	121
Frag10_mt428-589	13	28	0

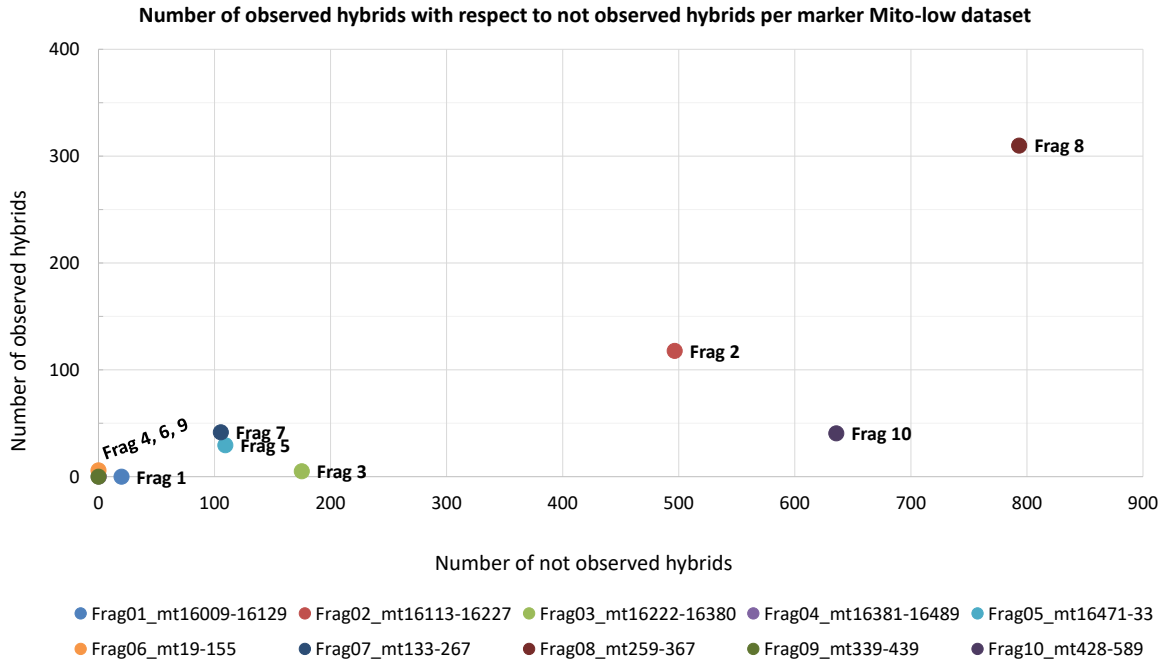


Figure 13: Number of observed w.r.t. not observed hybrids Mito -low dataset, per marker.

3.1.5 Total hybrids per marker STR dataset

For the STR dataset, the observed and not observed hybrids per marker are displayed in figure 14 on page 23. The marker *D12S391* simulated the most observed hybrids for both the K-mer method and the 'crossing-over' method i.e. 10.75% and 24.07% respectively. It is striking that the K-mer method simulates 2.17% to 9.50% actually-observed hybrids for each STR marker; a relatively equal performance for each marker. The only exclusion is the *Amel* marker in which no K-mer method hybrids are observed. This is probably due to *Amel* being the only marker that does not contain any repetitive subsections, it is more comparable to a random sequence. Additional markers for the 'crossing-over' method that simulated a relatively high amount of hybrids are *D21S11, PentaD* and *D5S818* i.e. 22.82%, 9.63% and 8.73%.

The number of unique observed hybrids per method, per fragment are displayed in table 6 on page 24. It is apparent that for this dataset the majority of the unique observed hybrids were simulated with the K-mer method i.e. 93.21% compared to 6.79% that were simulated with the 'crossing-over' method. The contribution of the 'crossing-over' method seems insignificant compared to the results of the K-mer method. However, if the 'crossing-over' method would not have been applied a total of 2781 hybrids would have not been marked. As a consequence, the K-mer method simulates the majority of the seen hybrids but, the 'crossing-over' method still contributes an adequate amount of hybrids that would not have been detected with the K-mer method.

Although, for the K-kmer method, the number of observed hybrids is very similar between markers, the number of hybrids that were not observed varies significantly. In this dataset, the markers that simulated the most not observed hybrids for the 'crossing-over' method are *Amel, D1S1656* and *D12S391*. These markers contributed for 29.54%, 15.59% and 14.14% respectively to the number of not observed hybrids for this method. The number of not observed hybrids for the K-mer method are more evenly distributed compared to the not observed 'crossing-over' hybrids. It is apparent that *D21S11* is the outlier, this marker contributed for 20.78% to the total number of not observed hybrids for the K-mer method. Indeed, marker *D21S11* turns out to contain the most repetitive subsections, which as a consequence resulted in the highest number of not observed hybrids. In addition to marker *D21S11*, the 'crossing-over' and K-mer method for markers *D13S317*, *D18S51*, *FGA*, *PentaD* and *vWA* simulated relatively more not observed hybrids compared to the

other markers i.e. 6.21% to 8.30%. The markers in which the most not observed hybrids for the 'crossing-over' method (*Amel*) and K-mer method (*D21S11*) were created contributed for 13.73% and 0.57% respectively to the number of observed hybrids. It is apparent that although hybrids were marked for the marker *Amel* using the 'crossing-over' method, that the method is not very efficient for this marker due to the high number of not observed simulated hybrids.

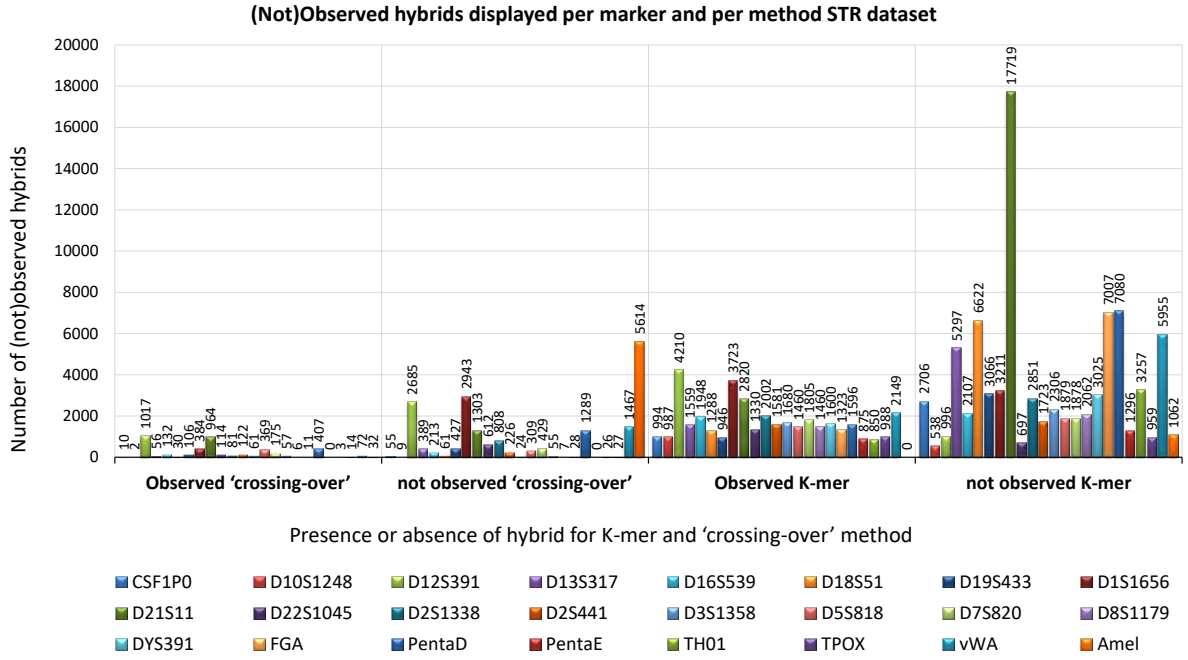


Figure 14: Number of observed and not observed hybrids STR dataset, per marker for the K-mer and 'crossing-over' method.

Table 6: Number of observed hybrids simulated with 'crossing-over' and K-mer method per fragment for STR dataset.

	Unique 'crossing-over' method	Unique K-mer method	Non-unique
CSF1P0	5	989	10
D10S1248	1	986	2
D12S391	807	4070	350
D13S317	37	1545	33
D16S539	66	1898	116
D18S51	18	1282	18
D19S433	104	944	4
D1S1656	335	3691	81
D21S11	520	2541	723
D22S1045	49	1285	110
D2S1338	53	1984	46
D2S441	54	1523	126
D3S1358	24	1651	66
D5S818	156	1282	391
D7S820	75	1733	172
D8S1179	35	1443	39
DYS391	0	1595	11
FGA	10	1322	2
PentaD	323	1542	138
PentaE	0	875	0
TH01	1	848	34
TPOX	71	2148	18
vWA	5	979	2
Amel	32	0	0

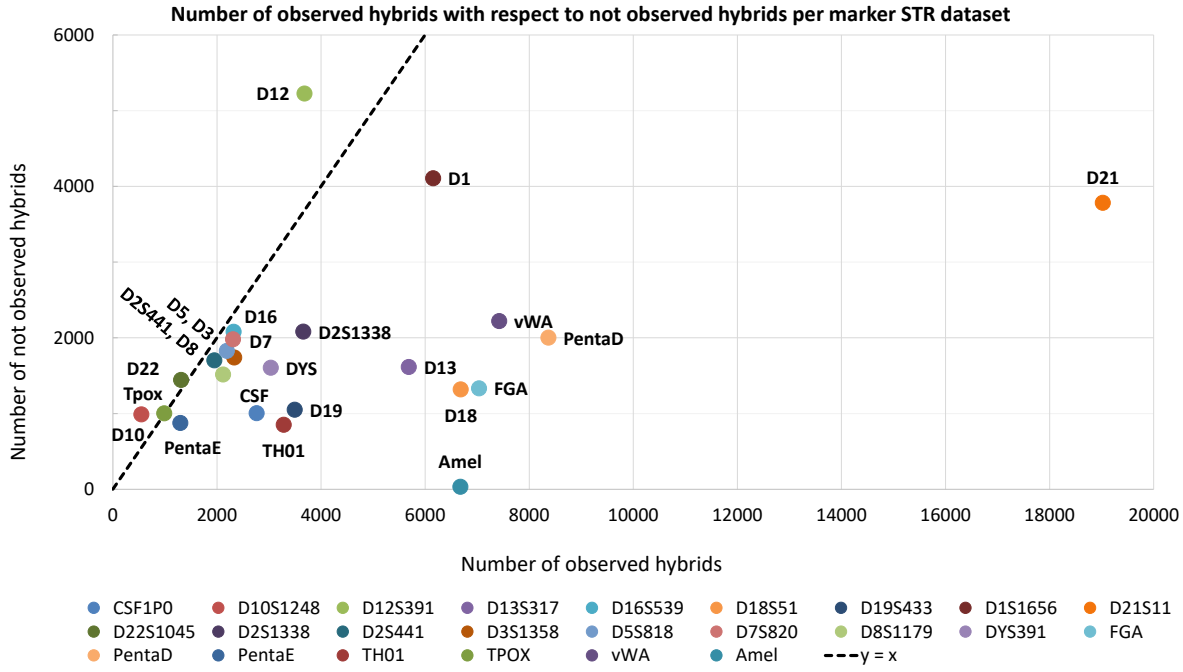


Figure 15: Number of observed w.r.t. not observed hybrids STR dataset, per marker.

3.1.6 Total hybrids per marker Microhaplotype dataset

Regarding the Microhaplotype dataset, it is clear from figure 8 on page 16 that both the 'crossing-over' method and the K-mer method did not simulate a considerable amount of observed hybrids. The contribution per marker is shown in figure 16, page 26. This figure shows that most of the observed hybrids for the 'crossing-over' method were simulated within the *rs4652604* marker, this marker accounts for 44.03% of the total number of observed hybrids for this method. Additionally, marker *rs35474228* contributed 50.55% to the total number of observed hybrids for the K-mer method. There were only four other markers for the K-mer method that simulated observed hybrids namely *rs1721827*, *rs6595279*, *rs650433* and *rs7632479*. However, each of these individual markers did not contribute a significant amount of observed hybrids i.e. 0.18%, 5.55%, 0.82% and 0.18% of hybrids simulated by the method. For the 'crossing-over' method eight additional markers contributed to the number of total observed hybrids. Regarding these markers, three markers contributed more than 10% each to the number of observed hybrids i.e. *rs13145525* 12.08%, *rs6595279* 12.68% and *rs6504633* for 24.83%.

The number of unique hybrids per marker for the 'crossing-over' method and K-mer method are displayed in table 7, page 26. Overall, the percentage of unique observed hybrids that were simulated with the 'crossing-over' method is 55.08%, as a result the percentage that is simulated with the K-mer method is 44.92%. This demonstrates that each method contributed relatively equally to the number of unique hybrids. Therefore, this dataset benefits optimally by using both hybrid simulation methods.

For the number of not observed hybrids, marker *rs13145525* accounts for 87.13% of the total number of not observed hybrids for the 'crossing-over' method. The 'crossing-over' method did not produce many observed nor not observed hybrids. This is an indication that there is not enough variety between a pair of sequences in the dataset to simulate more hybrids. This is supported by the fact that the sequences differ only a couple SNPs.

The K-mer method does not perform better for the observed hybrids. In this method the marker *rs35474228* contributes 49.35% to the total number of observed hybrids of this method. This is a higher contribution for the number of observed hybrids compared to the 'crossing-over' method. Nonetheless, the total number of observed hybrids for the K-mer method is lower. The hybrids that are not observed and produced with the K-mer method were for the majority simulated within marker *rs6504633*. This marker contributed for 30.03% to the total number of not observed hybrids for this method. The markers which produced the most not observed hybrids, contributed for 3.66% and 24.55% to the number of observed hybrids for that marker respectively. The contribution of *rs6504633* appears insignificant but relative to the total number of observed hybrids for the K-mer method this marker contributes 35.82%. An overview of the total number of observed hybrids relative to the total number of not observed hybrids per marker is shown in figure 17 on page 27. This figure shows that marker *rs6595279* is the only marker for which the number of not observed versus observed hybrids is balanced i.e. 51.42% and 48.58%. The outlier, as we have shown in this section is *rs6504633*, for this marker the ratio observed to not observed hybrids is 6.56% to 93.44% respectively.

It is interesting to examine the parent sequences of the markers that simulated a significant amount of unique hybrids i.e. *rs35474228* for the K-mer method and *rs6504633* for both the 'crossing-over' and K-mer method. These sequences are displayed in item 1, 2 of list 1 on page 26 respectively. It is apparent that the sequence of item 1 contains many repetitive subsections when a K-size of four is used. It is therefore that many hybrids can be simulated for the K-mer method for marker *rs35474228*. Concerning the sequence of item 2, it contains some repetitive sections but apparently there is also an adequate amount of variation between the parent sequences of this marker. As a consequence an equal amount of hybrids were simulated in marker *rs6504633* with the K-mer and 'crossing-over' method. Additionally, it is interesting to analyse the sequences simulated with the K-mer method which are observed and not observed and determine what can be a possible explanation for this. Examples of sequences that are observed and not observed are displayed in item 3 and 4 of list 1 respectively. It is apparent that the not observed hybrid (item 4) contains more nucleotides than the observed hybrid (item 3). In fact, the unobserved hybrid is 2.21 times longer compared to the average nucleotide length of the hybrids in this dataset (section 2.2, page 8). This while the observed hybrid is only 1.46 times longer. It is less likely that the longer hybrid will be created and is therefore not observed in the Microhaplotype dataset.

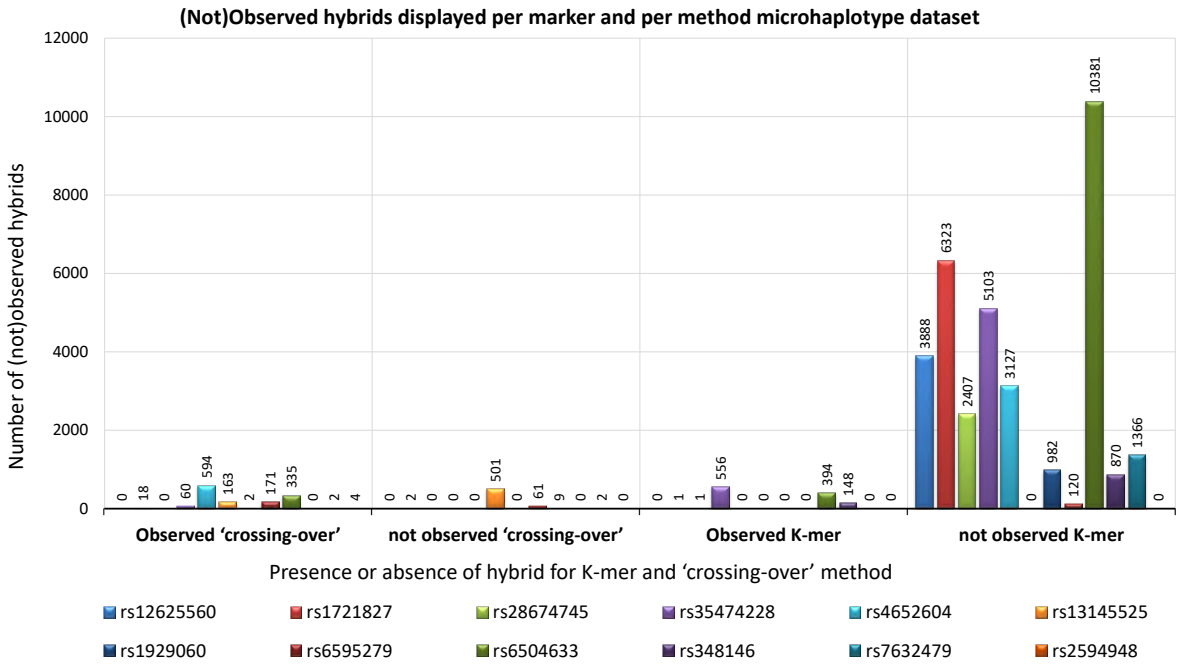


Figure 16: Number of observed and not observed hybrids Microhaplotype dataset, per marker for the K-mer and 'crossing-over' method.

Table 7: Number of observed hybrids simulated with 'crossing-over' and K-mer method per fragment for Microhaplotype dataset.

	Unique 'crossing-over' method	**Unique K-mer method**	**Non-unique**
rs1721827	18	1	0
rs28674745	0	1	0
rs35474228	60	556	0
rs4652604	594	0	0
rs13145525	163	0	0
rs1929060	2	0	0
rs6595279	171	0	0
rs6504633	335	394	0
rs348146	0	148	0
rs7632479	2	0	0
rs2594948	4	0	0

List of parent, observed and not observed hybrid sequences

1. **rs35474228, parent:**
   GGGGCGTCTGTTGGGGGGACCTGGCGTCATTACC

2. **rs6504633, parent:**
   GAAGGCCAGGGAGGTGAAGGGGGGGAAGGAGGTTT

3. **rs6504633, observed hybrid:**
   GAAGGCCAGGGAGGTGAAGGGGGGGGAAGGAGGTTT

4. **rs6504633, unobserved hybrid:**
   GAAGGCCAGGGAGGTGAGTGGGGAGAAGGAGGTGAGTGGGGAGAAGGAGGTTT

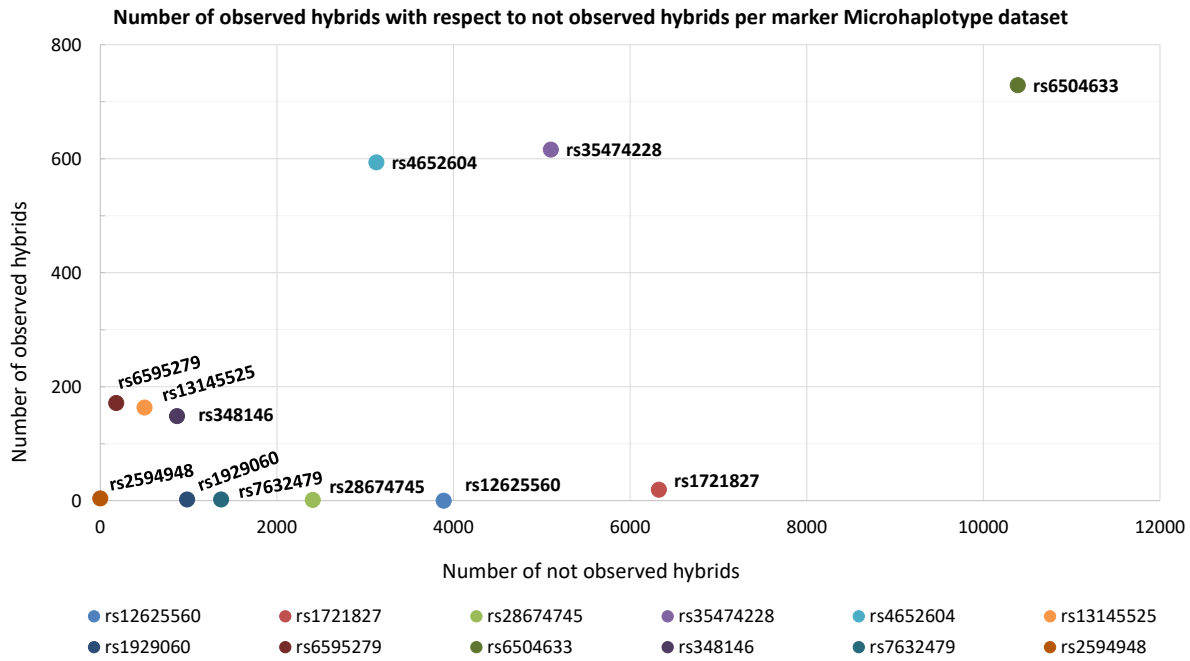


Figure 17: Number of observed w.r.t. not observed hybrids Microhaplotype dataset, per marker.

3.2 Genetic feature selection algorithm

In this study, a total of 175 features have been defined that characterize the formation of hybrid sequences. These features were implemented in a least-squares regression model (chapter 2.3, page 10) to quantitatively predict hybrid ratios. A genetic feature selection algorithm [10] was used to select features that contributed to the lowest classification error and exclude the rest. This feature reduction step can increase the computation speed and accuracy of the hybrid prediction model, a subset of features can contain more descriptive information to predict the hybrid ratio than all features combined. In this section the results of the feature selection algorithm are discussed.

3.2.1 Optimal feature set

The features have been designed in order to predict the hybrid ratio (number of reads of a hybrid sequence relative to the reads of the parent(s)). The feature selection algorithm has been executed three times, this because the feature selection algorithm and the following steps initialize randomly. Therefore, the selected feature subset and accuracy can differ per instance. Subsequently, the feature subset that resulted in the lowest prediction error was used in the prediction model. If the error was lowest when all features were included then all features were included in the model.

The algorithm starts with calculating the validation error including all features i.e. -2.33×10^{-5} for the Mito-dataset. Thereafter, subsets of features were simulated and the corresponding errors evaluated. The subset with the lowest prediction error consisted of 88 features and a validation error of -2.91×10^{-5}. Consequently, the feature selection algorithm has reduced the number of features by 87 however was not able to reduce the validation error on the dataset. Therefore, all features were included for the prediction model of the Mito dataset. In total, 26 features were selected by the genetic algorithm in all three instances, an overview of these features is shown in table B.1 on page II. These features presumably contain the most information for predicting the hybrid ratio for this dataset.

For the Mito-low dataset the validation error including all features was -7.88×10^{-6}, the genetic algorithm selected a total of 85 features and reduced the error to -5.22×10^{-6}. There were 14 features that were selected in all three experiments, these features are shown in B.2 on page II. We have tested the genetic algorithm including only these features but this did not further reduce the validation error.

In total 81 features were selected by the genetic algorithm to obtain the lowest validation error in the STR dataset. However, this validation error was higher than the baseline error which included all features. By selecting the 81 features the error increased from -2.84×10^{-6} to -2.89×10^{-6}, this is an error incrementation of 4.57×10^{-8}. A total of 21 features were detected in all three experiments, these are shown in B.3 on page II.

In the dataset containing Microhaplotype sequences a total of 96 features were selected which reduced the validation error from -3.95×10^{-6} to -3.74×10^{-6}, this corresponds to an error reduction of 2.10×10^{-7}. In total 30 features were selected in all three experiments, which are are displayed in table B.4 on page IV.

Nearly all hybrids in the STR dataset were simulated with the K-mer method, which makes this dataset different compared to the other datasets. Nonetheless, this dataset has the lowest validation error, this is probably due to the large amount of data it contains. Consequently, the hybrid prediction model for the STR dataset should be the most accurate model. This in contrast to the Mito dataset, which has the highest validation error. As a consequence, the Mito dataset is the most difficult dataset to predict the hybrid ratio.

3.3 Hybrid prediction model

In this section the results of the hybrid prediction model are displayed and discussed per dataset. For each dataset two hybrid prediction models were build, one where the model is trained and tested on all markers combined i.e. *Combined model* and one where the model is trained and tested per individual marker i.e. *individual marker model* or *Overlay model*. The overlay model is the result of the individual predictions per marker superimposed on each other.

3.3.1 Mito prediction

In order to build a hybrid prediction model that is able to recognise hybrid sequences in MPS data it is necessary that the model is trained on hybrids that have been observed (positive examples). An overview of the number of positive examples per marker is shown in table 8. A minimal of 10 positive examples per marker was set as cut-off value. Below this cut-off, the model cannot be accurately trained and tested. As a result, *fragments 1, 4, 6* and *fragment 9* were excluded in the hybrid prediction models.

Table 8: Number of hybrids that have been observed per marker for the Mito dataset.

Markers	Number of positive examples
Frag01_mt16009-16129	0
Frag02_mt16113-16227	15
Frag03_mt16222-16380	12
Frag04_mt16381-16489	0
Frag05_mt16471-33	30
Frag06_mt19-155	6
Frag07_mt133-267	106
Frag08_mt259-367	110
Frag09_mt339-439	0
Frag10_mt428-589	66

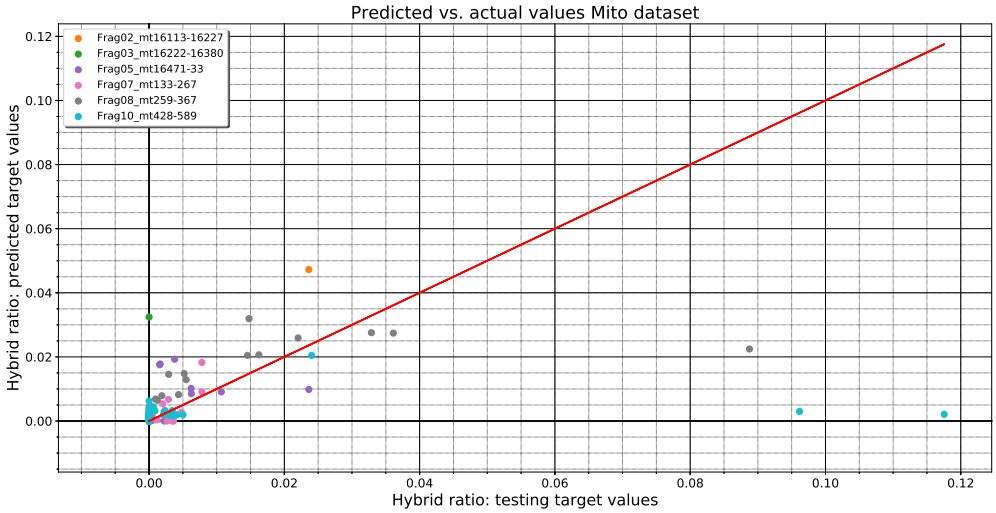
A division of the data in a training and test set was performed to create a hybrid ratio prediction model. The effect of splitting the data on the number of positive and negative data points for both the combined and individual marker models is displayed in table 9. There is a class imbalance between the positive and negative examples, the number of negative examples is always higher than the number of positive examples. As a result, the model is predominantly trained on the not observed hybrids. The fragment with the highest imbalance is *fragment 3* and lowest *fragment 7*, the imbalance range is between 3.62% and 29.41% positive examples per fragment respectively. A lower class imbalance is more likely to result in a more accurate hybrid prediction.

Table 9: Number of positive and negative hybrids included in the training and test set for the Mito dataset.

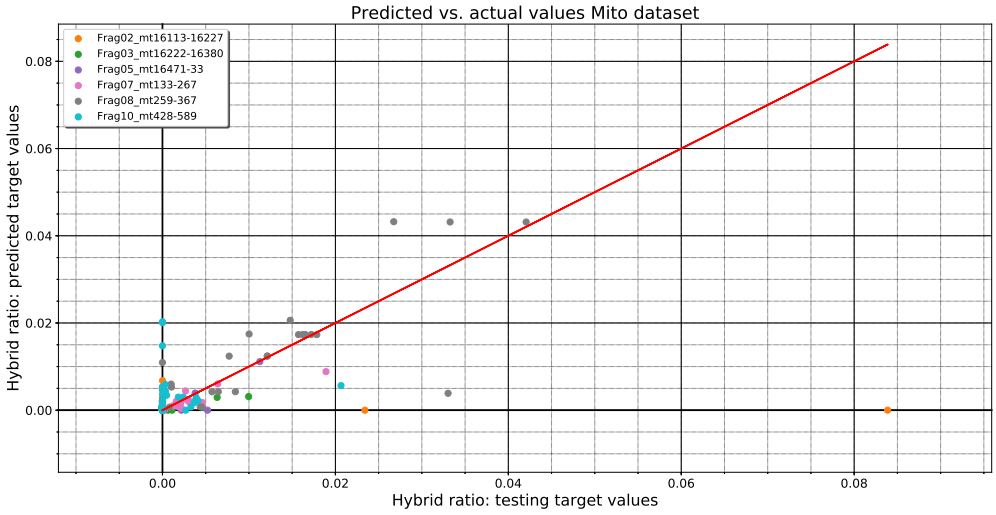
Markers	**Training set pos. examples**	**Training set neg. examples**	**Test set pos. examples**	**Test set neg. examples**
Frag02_mt16113-16227	13	216	2	56
Frag03_mt16222-16380	8	213	4	52
Frag05_mt16471-33	26	124	4	34
Frag07_mt133-267	85	204	21	52
Frag08_mt259-367	81	448	29	104
Frag10_mt428-589	52	485	14	121
Combined model	270	1688	69	421

The results of the hybrid ratio predictions of the combined and overlay models for the test set are shown in figure 18 on page 30, for the individual plots per marker see appendix C on page V. The vertical axis of these plots displays the hybrid ratio that is predicted by the least-squares model and the horizontal axis the true hybrid ratios. A reference line (red line) is drawn where $\hat{y} = y$. The closer in range the data points are to the line the more accurate the prediction of the model is.

If a data point is above the line it will be filtered. In contrast, if a data point is below the line it will not be completely filtered away, however it is an improvement compared to not being filtered at all. This because the hybrid might be corrected below the interpretation threshold. By comparing the models visually it is apparent that in the combined model, hybrids in *fragment 10* are under-predicted as is the data point from *fragment 8* of which the true hybrid ratio is highest. These data points, with the highest true hybrid ratio, are the most important to predict correctly since these can be interpreted as true alleles in DNA analysis. The under-predicted hybrids are more accurately predicted in the overlay model. This can also be observed in table 10 on page 31. In this table the coefficient of determination (R^2) and explained variance are displayed for the individual markers and the combined model. By examining these scores it is clear that the models of the individual markers *fragments 3, 6, 7* and *fragment 8* are more accurate compared to the combined model. For the other markers (fragment 2 and fragment 10) the results are suboptimal, these markers were more accurately predicted in the combined model than in the individual model, this can be due to stochastic variance in the PCR.



(a) Hybrid prediction model based on all markers combined of the Mito dataset.



(b) Hybrid prediction model based separate markers superimposed on each other for the Mito dataset.

Figure 18: Combined (18a) and Overlay (18b) hybrid prediction model for the Mito dataset. The horizontal axis represents the true hybrid ratios and the vertical axis the predicted hybrid ratio. The red line is a reference line where $\hat{y} = y$. The closer the data points are to the reference line, the more accurately these points are predicted by the model.

Table 10: Error measurement, coefficient of determination and explained variance measurements for the Mito dataset.

Markers	$\mathbf{R^2}$	**Explained variance**
Frag02_mt16113-16227	-3.56×10^{-2}	-1.73×10^{-2}
Frag03_mt16222-16380	5.47×10^{-1}	5.56×10^{-1}
Frag05_mt16471-33	7.99×10^{-1}	8.04×10^{-1}
Frag07_mt133-267	7.22×10^{-1}	7.30×10^{-1}
Frag08_mt259-367	7.37×10^{-1}	7.42×10^{-1}
Frag10_mt428-589	-1.97	-1.83
Combined model	1.55×10^{-1}	1.55×10^{-1}

The effect of the hybrid prediction model correction has been visualised in figure 19, the horizontal axis illustrates the hybrid ratio before the application of the correction model and the vertical axis the hybrid ratio after the correction. The triangle shapes in the figure represent the individual markers that resulted in the best prediction and the circles the markers that performed best when included in the combined model. It is apparent that for the Mito dataset the model does not correct the high hybrid (>0.08) ratios effectively. The majority of the hybrids that are corrected are within the range of 0.015 to 0.04. However the prediction model frequently over-corrects data points (hybrid ratio below zero). The hybrid ratio of these points will be set to zero since a negative hybrid ratio is not possible. If a hybrid corresponds to a true allele of a minor contributor in a forensic DNA sample, it is possible that a percentage of that allele is over-corrected. As a consequence, this allele will possibly not be detected in the DNA analysis. The marker that is most often over-corrected is *fragment 8* i.e. 49.62%. In total, 47.10% of all data points are over-corrected, 8.51% is corrected and 44.39% has not been affected by the prediction model. The latter category is the most difficult to predict and to correct for because the data points are not affected by the correction model. The marker that is most unaffected by the prediction model is *fragment 10*. The data points of this marker are closest to the reference line ($y = x$), the hybrid ratio after correction is (almost) identical to the hybrid ratio before the correction.

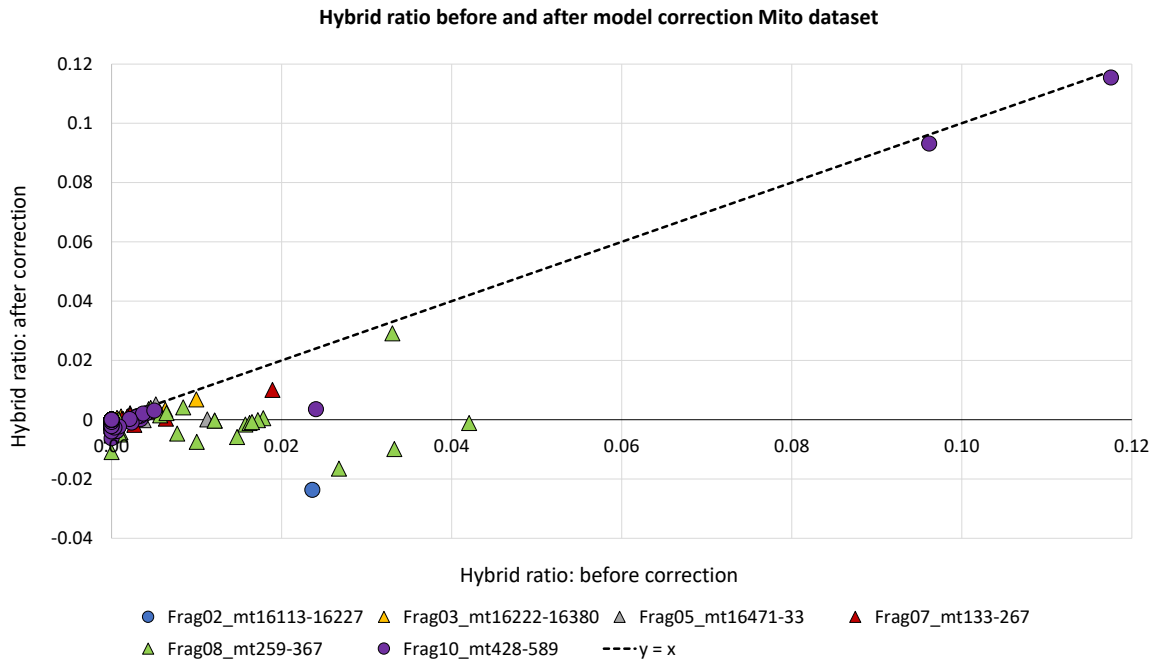


Figure 19: Hybrid ratio prediction correction of the test set of the Mito dataset. The horizontal axis represents the hybrid ratio before correction and the vertical axis the hybrid ratio after correction. The triangle shapes are corrected using the individual models (trained and tested per marker) and the circles are the markers that are included in the combined model.

3.3.2 Mito-low prediction

For the Mito-low dataset the number of positive examples per marker are shown in table 11. Half of the markers have more than ten positive examples i.e. *fragments 2, 5, 7, 8* and *fragment 10*, the other markers were excluded in the prediction models. Subsequently, the data was split in a training and a test set of which the result is displayed in table 12. The markers with the most and least imbalanced training set are *fragment 10* and *7*, these fragments have a range between 9.35% and 26.67% positive examples per fragment respectively.

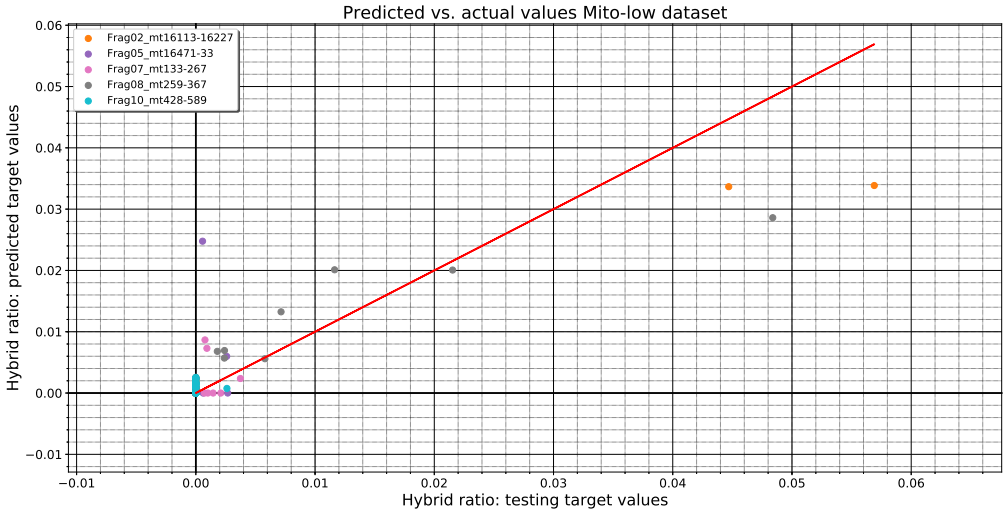
Table 11: Number of hybrids that have been observed per marker for the Mito-low dataset.

Markers	**Number of positive examples**
Frag01_mt16009-16129	0
Frag02_mt16113-16227	16
Frag03_mt16222-16380	5
Frag04_mt16381-16489	0
Frag05_mt16471-33	16
Frag06_mt19-155	6
Frag07_mt133-267	38
Frag08_mt259-367	50
Frag09_mt339-439	0
Frag10_mt428-589	33

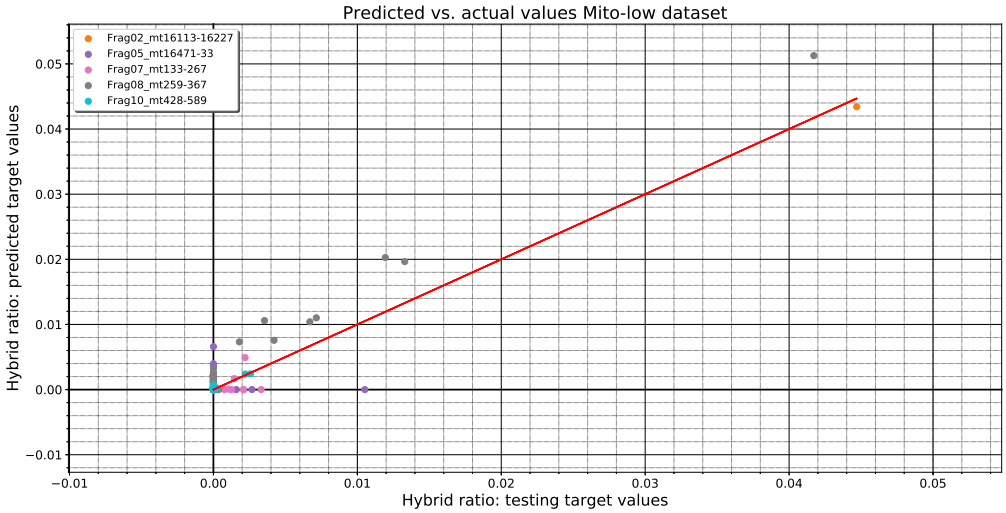
Table 12: Number of positive and negative hybrids included in the training and test set for the Mito-low dataset.

Markers	**Training set pos. examples**	**Training set neg. examples**	**Test set pos. examples**	**Test set neg. examples**
Frag02_mt16113-16227	14	111	2	30
Frag05_mt16471-33	12	66	4	16
Frag07_mt133-267	28	77	10	17
Frag08_mt259-367	41	205	9	53
Frag10_mt428-589	26	252	7	63
Combined model	128	706	25	184

The combined and overlay hybrid prediction models are displayed in figure 20 on page 34, the horizontal axis represents the actual hybrid value and the vertical axis the predicted hybrid ratio. A significant number of data points are close to the origin. The hybrid ratios of some of these points are predicted as zero while they have a positive hybrid ratio, this also occurs vice versa. The hybrid ratios that are more accurately predicted in the overlay model compared to the combined model correspond to *fragments 2, 8* and *fragment 10*. This is confirmed by the R^2 and explained variance scores of these markers, displayed in table 13, page 35. The model is accurate in predicting the hybrid ratio for *fragments 2, 8* and *10* per individual marker, *fragment 5* and *7* are more accurate predicted in the combined model. It is remarkable that for the Mito-low dataset *fragment 2* is accurately predicted while in the Mito dataset this fragment the predictions are better predicted in the combined model, this also applies to *fragment 10*. In addition, *fragment 5* and*7* are better predicted individually in the Mito prediction model and *fragment 8* is adequately predicted in both the Mito and Mito-low dataset.



(a) Hybrid prediction model based on all markers combined of the Mito-low dataset



(b) Hybrid prediction model based separate markers superimposed on each other for the Mito-low dataset

Figure 20: Combined (20a) and Overlay (20b) hybrid prediction model for the Mito-low dataset. The horizontal axis represents the true hybrid ratios and the vertical axis the predicted hybrid ratio. The red line is a reference line where $\hat{y} = y$. The closer the data points are to the reference line, the more accurately these points are predicted by the model.

Table 13: Error measurement, coefficient of determination and explained variance measurements for the Mito-low dataset

Markers	$\mathbf{R^2}$	**Explained variance**
Frag02_mt16113-16227	9.98×10^{-1}	9.99×10^{-1}
Frag05_mt16471-33	$-7.78E \times 10^{-1}$	-7.78×10^{-1}
Frag07_mt133-267	-3.77×10^{-1}	-3.08×10^{-1}
Frag08_mt259-367	7.88×10^{-1}	8.82×10^{-1}
Frag10_mt428-589	8.91×10^{-1}	9.13×10^{-1}
Combined model	7.44×10^{-1}	7.49×10^{-1}

The effects of the prediction corrections on the Mito-low dataset are visualised in figure 21. In this figure the horizontal axis represents the hybrid ratio before the corrections of the prediction model and the vertical axis the hybrid ratio after the correction. It is clear that the the two highest data points (hybrid ratio between 0.04 and 0.05) are adequately corrected. These data points are the most important to correct because these can be mistakenly identified as true alleles, which can result in the false determination of additional contributor(s) in DNA analysis of forensic casework samples. Furthermore, *fragments 7, 8* and *fragment 10* are occasionally over-corrected, of which the majority can be assigned to *fragment 8* i.e. 77.42% and the highest over correction is applied to a data point of *fragment 5* i.e. -0.024. In total, 63.59% of all data points are over-corrected, 2.30% is corrected and 34.11% of the points is not corrected. Conclusively, about a third of all hybrids is unaffected by the prediction model, but there is a large degree of overcorrection for some fragments.

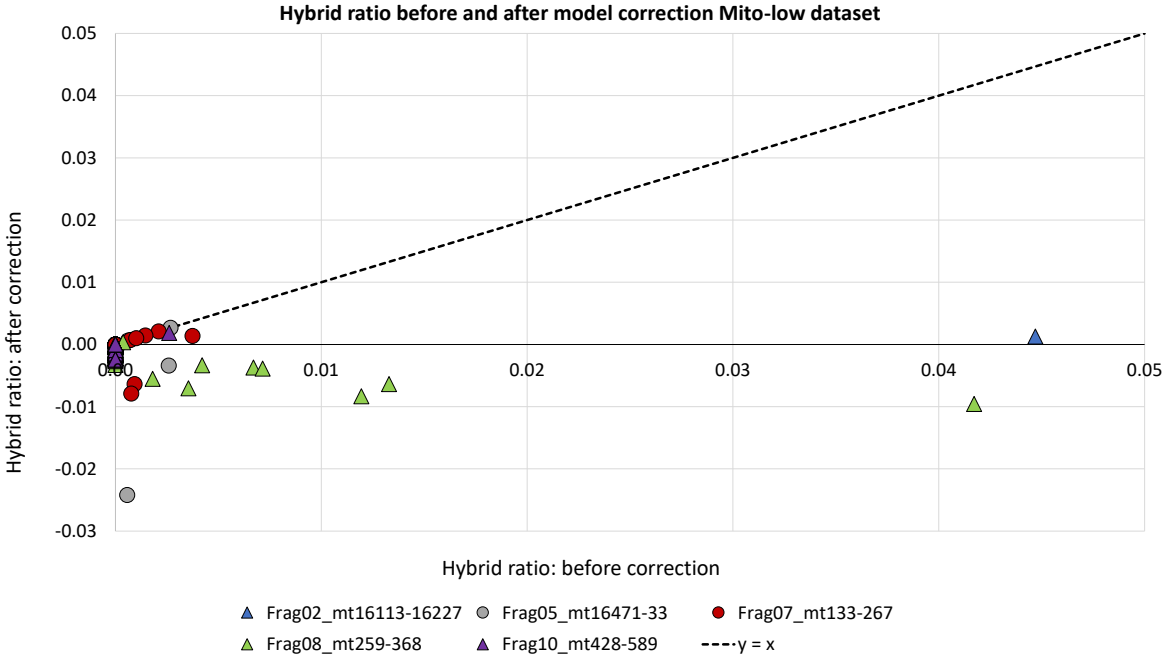


Figure 21: Hybrid ratio prediction correction of the test set of the Mito-low dataset. The horizontal axis represents the hybrid ratio before correction and the vertical axis the hybrid ratio after correction. The triangles shapes are the markers that obtained the best predictions when trained on the individual markers, the circles represent the markers that performed best in the combined model.

3.3.3 Mito-combined prediction

The applied corrections on the Mito and Mito-low dataset are adequate however a more accurate model can possibly be obtained by combining these two datasets. The advantage of a combined model is that instead of four models all hybrids can be predicted on two models. An additional advantage is that the training and test set for this model is larger compared to the training and test set for the individual datasets, all data can subsequently be analysed simultaneously. This can result in additional markers being included in the prediction model as a consequence of the number of positive examples exceeding the cut-off value i.e. ten positive examples. The number of positive examples per marker is displayed in table 14. This table shows that in the Mito-combined dataset *fragment 6* is included in the prediction model, while it was excluded in both the Mito and Mito-low datasets. The fragments that are not included in the prediction model are *fragments 1, 4* and *fragment 9*, these fragments did not at all simulate any hybrid examples that were observed in the Mito or Mito-low dataset. Based on the available data, it is unlikely that hybrid artefacts are formed within these markers. For the markers that were included in the prediction model, a train-test split was made of the data, this split is displayed in table 15. The fragments that have the most and least imbalanced training set are *fragment 3* and *7* with a range between 3.29% and 27.60% positive examples per fragment respectively. Striking is that *fragment 6* only contains positive examples, as a consequence the individual model may not be able to predict a hybrid ratio of zero for this marker.

Table 14: Number of hybrids that have been observed per marker for the Mito-combined dataset

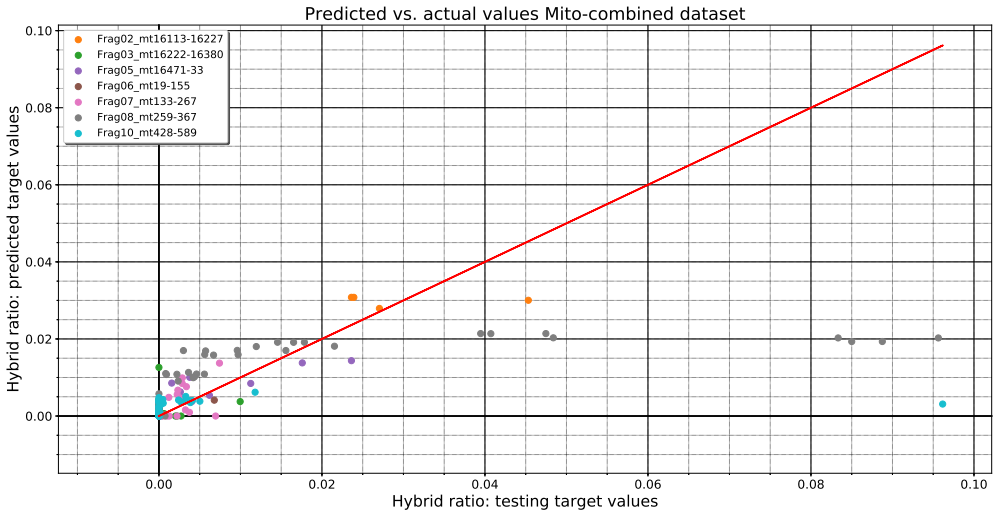
Markers	**Number of positive examples**
Frag01_mt16009-16129	0
Frag02_mt16113-16227	31
Frag03_mt16222-16380	17
Frag04_mt16381-16489	0
Frag05_mt16471-33	46
Frag06_mt19-155	12
Frag07_mt133-267	144
Frag08_mt259-367	160
Frag09_mt339-439	0
Frag10_mt428-589	99

Table 15: Number of positive and negative hybrids included in the training and test set for the Mito-combined dataset

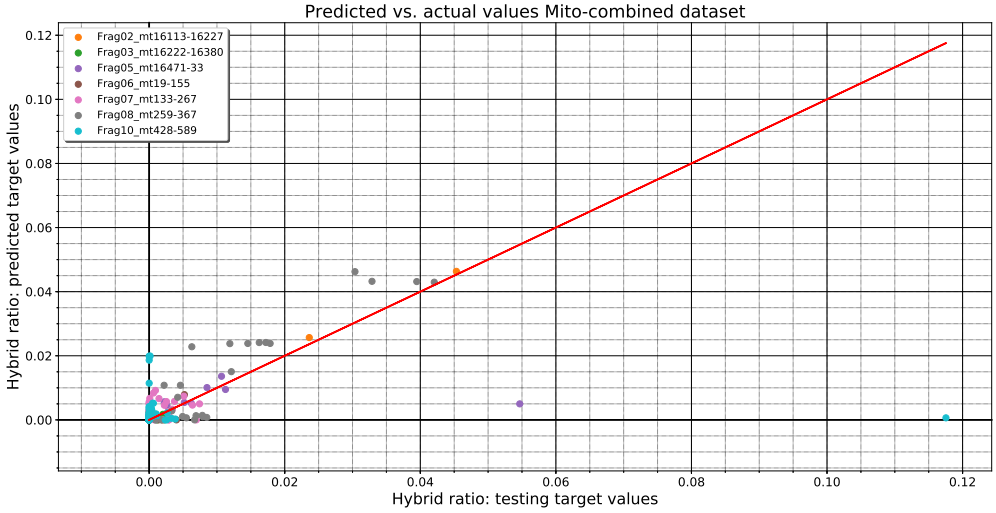
Markers	**Training set pos. examples**	**Training set neg. examples**	**Test set pos. examples**	**Test set neg. examples**
Frag02_mt16113-16227	27	328	4	85
Frag03_mt16222-16380	12	353	5	87
Frag05_mt16471-33	36	192	10	48
Frag06_mt19-155	9	0	3	0
Frag07_mt133-267	109	286	35	64
Frag08_mt259-367	125	651	35	159
Frag10_mt428-589	80	736	19	185
Combined model	395	2551	114	623

The hybrid prediction results for all markers combined and per individual marker are shown in figure 22 on page 37. The only markers that were more accurately predicted when trained on the individual markers were *fragment 2* and *fragment 8* (figure 22b), the other markers performed better when incorporated in a model including all markers (22a). The accuracy of the predictions of the models is displayed in table 16, page 38. It is noticeable that the R^2 and explained variance of *fragment 2* and *8* are accurate for the prediction model when trained on the individual corresponding markers and that the scores are almost identical to the scores of the Mito-low dataset (table 13, page 35). The hybrids of the other markers are more accurately predicted in the combined model.

For the current Mito and Mito-low dataset it is preferred to predict hybrid ratios per dataset as opposed to combining these and subsequently predict the hybrid ratios. The constructed models per dataset are more precise in predicting the hybrid ratios. This approach is at the expense of *fragment 6*. However, the R^2 and explained variance scores are low for this fragment, which as a consequence makes it insignificant for predicting the hybrid ratio in the overlay model.



(a) Hybrid prediction model based on all markers combined of the Mito-combined dataset



(b) Hybrid prediction model based separate markers superimposed on each other for the Mito-combined dataset

Figure 22: Combined (22a) and Overlay (22b) hybrid prediction model for the Mito-combined dataset. The horizontal axis represents the true hybrid ratios and the vertical axis the predicted hybrid ratio. The red line is a reference line where $\hat{y} = y$. The closer the data points are to the reference line, the more accurately these points are predicted by the model.

Table 16: Error measurement, coefficient of determination and explained variance measurements for the Mito-combined dataset

Markers	$\mathbf{R^2}$	**Explained variance**
Frag02_mt16113-16227	9.97×10^{-1}	9.97×10^{-1}
Frag03_mt16222-16380	-6.67×10^{-1}	-6.63×10^{-1}
Frag05_mt16471-33	2.18×10^{-1}	2.27×10^{-1}
Frag06_mt19-155	-2.28	-1.96
Frag07_mt133-267	-1.15	-8.75×10^{-1}
Frag08_mt259-367	7.83×10^{-1}	7.91×10^{-1}
Frag10_mt428-589	-9.98×10^{-2}	-9.95×10^{-2}
Combined model	3.98×10^{-1}	3.98×10^{-1}

The hybrid predictions of the most accurate model per marker have been applied as corrections to the test set and visualized in figure 23. The horizontal axis of the figure describes the hybrid ratio before applying the prediction model corrections and the vertical axis the hybrid ratio after the corrections. It is evident that 57.74% of *fragment 8* is over-corrected and that the data point with the highest hybrid ratio i.e. *fragment 10* is barely affected by the applied correction. Furthermore, the only markers of which the hybrid ratios are more accurately predicted when trained on the individual markers are *fragment 3* and *8*, the hybrid ratio of the other markers are more accurately predicted in the combined model. The hybrid ratios that are most difficult to predict correspond to *fragment 10*, this is illustrated in figure 19 (page 32) and 23. Fragment 10 contains the highest hybrid ratios (between 0.095 and 0.12), however these ratios are not corrected by the prediction models even though these are the most important hybrids to filter. In total, 34.51% of all hybrids in the Mito-combined dataset is over-corrected, 5.88% is corrected and 59.61% has not been affected by the predictions of the model. These results confirm that combining the Mito and Mito-low datasets for hybrid prediction is less optimal compared to developing models for the individual datasets.

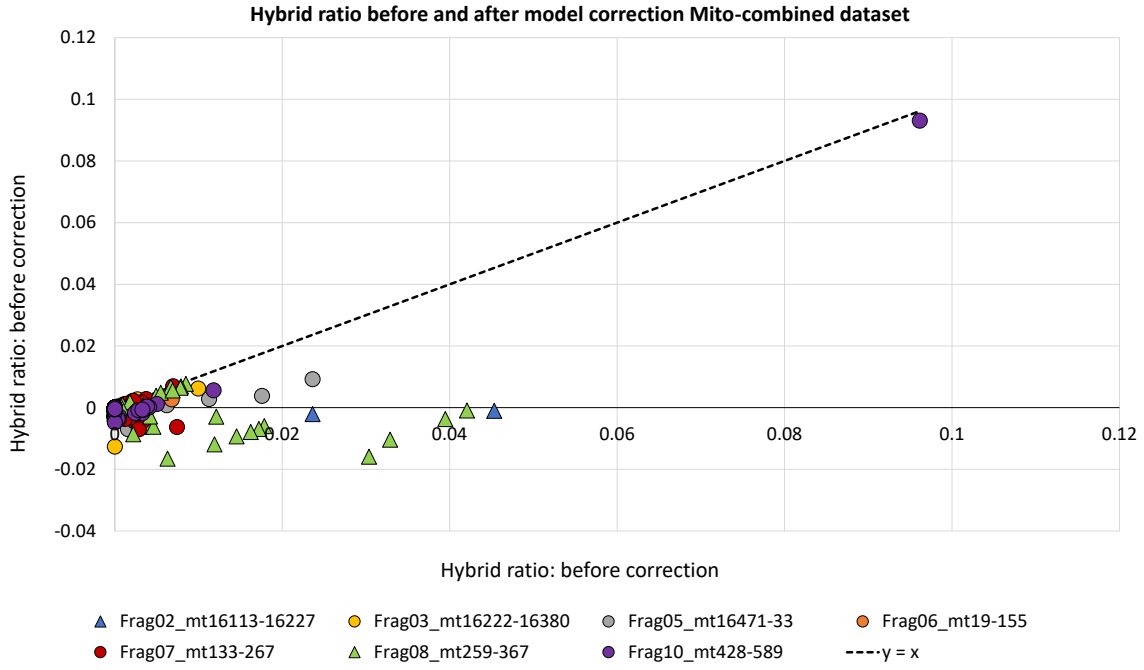


Figure 23: Hybrid ratio prediction correction of the test set of the Mito-combined dataset. The horizontal axis represents the hybrid ratio before correction and the vertical axis the hybrid ratio after correction. The triangle shapes are the individual models (trained and tested per marker) and the circles are the markers that are included in the combined model.

3.3.4 STR prediction

The STR dataset contains more samples and longer sequences relative to the Mito and Mito-low dataset. Therefore, there are significantly more positive examples per marker simulated, this is displayed in table 17. It is apparent that *Amel* contains the smallest number of positive examples i.e. 34, this is due to the fact that *Amel* does not contain any repetitive subsections and in general resembles more or less a random sequence. As a consequence, it was expected that *Amel* generated the smallest number of positive hybrids. The hybrid simulating algorithm simulated a substantial amount of positive examples i.e. 845 to 4014 for the remaining set of markers. All markers were therefore included in the prediction model for the STR dataset. Subsequently, a split of the data in training and test set was performed. The number of positive and negative examples per marker for the training and test set is depicted in table 18, page 40. Marker *Amel* has the most imbalanced training set containing 0.342272% positive examples, this in contrast to *D12S391* contains 69.29% positive examples. For the latter marker the number of positive examples in the training set are therefore overrepresented.

Table 17: Number of hybrids that have been observed per marker for the STR dataset.

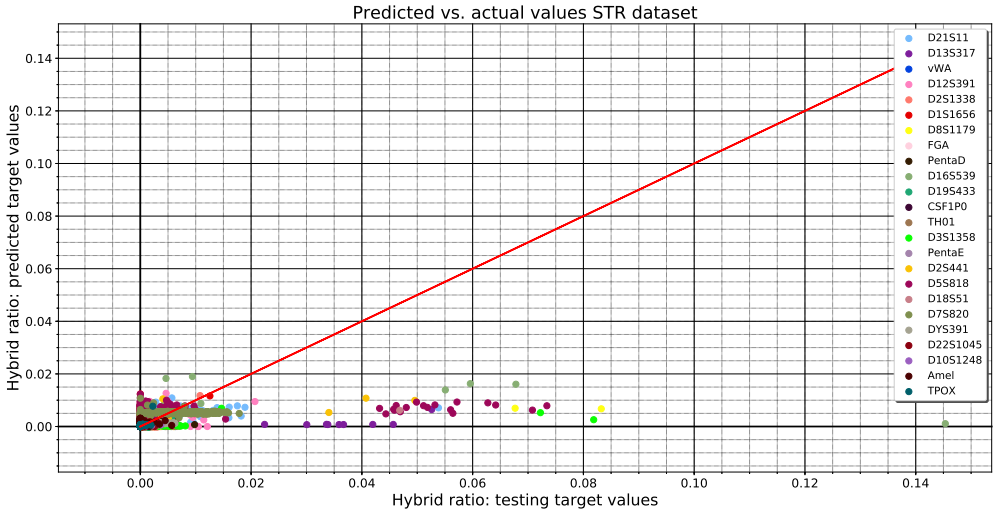
Markers	**Number of positive examples**
CSF1P0	1005
D10S1248	1130
D12S391	2829
D13S317	1819
D16S539	1990
D18S51	1465
D19S433	1057
D1S1656	4014
D21S11	2697
D22S1045	1359
D2S1338	1911
D2S441	1684
D3S1358	1905
D5S818	1683
D7S820	3551
D8S1179	1489
DYS391	1603
FGA	1466
PentaD	2750
PentaE	881
TH01	845
TPOX	1006
vWA	2068
Amel	34

The results of the hybrid prediction models for the STR dataset are shown in figure 24, page 41. On the horizontal axis of these figures the actual hybrid ratios are displayed, the vertical axis represents the predicted hybrid ratios. The prediction model that is trained on all markers combined (24a) under-predicts all hybrids that have a ratio higher than 0.02. For these data points the model is not able to make valid predictions. This is unfortunate since these data points are more important to recognize and filter, this because the higher the hybrid ratio the more likely it is to be interpreted as true allele in forensic DNA analysis. The markers that correspond to the hybrids with the highest ratio are *D3S1358, D8S1179, D5S818* and *D13S317*, the majority of the remaining data-points are centred around the origin. In the overlay model (24b) the hybrids with a ratio of more than 0.02 are more accurately predicted compared to the combined model (close in range of red line). However, the highest data point of this model e.g. hybrid ratio of 0.18 is under-predicted. In general, the overlay model predicts the higher hybrid ratios more accurately compared to the combined model. This as opposed to the lower ratios (≤ 0.02), some of these hybrids are significantly over and under-predicted in the overlay model i.e. *D13S317, D12S391, D1S1656, D8S1179, D5S818* and *Amel*.

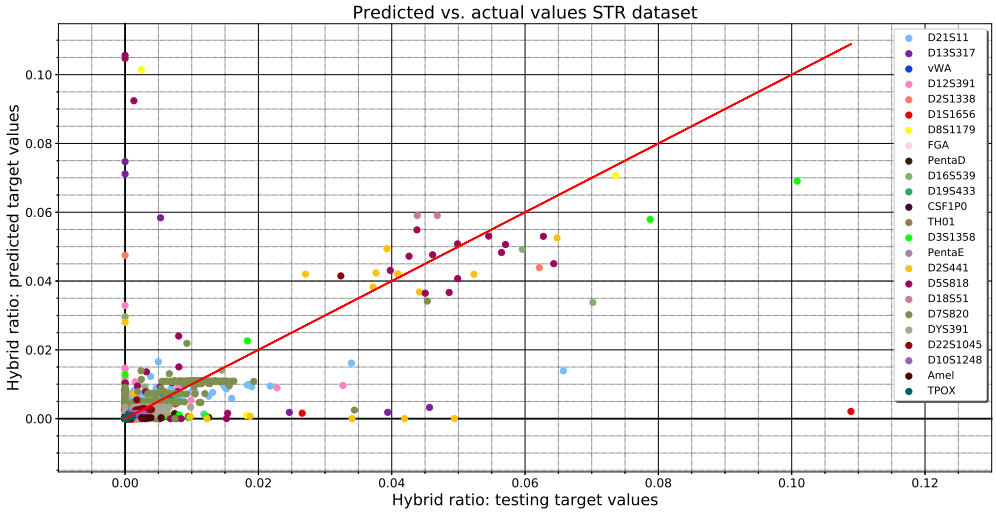
These results can be verified by the prediction accuracy of the combined and overlay model which are displayed in table 19 on page 42. The R^2 and explained variance score confirm that the above mentioned markers are more accurately predicted in the combined model. From the remaining markers there are eight markers that are predicted accurately i.e. *D19S433, CSF1P0, D3S1358, PentaE, D18S51, D22S1045, D10S1248* and *TPOX*. In addition, also the explained variance of these markers is above 0.85. Therefore, more than 85% of the variance of the data of the markers can be explained by the model. Striking is that for marker *D12S391* which had the most positive examples in the training set the hybrid prediction model trained on the individual marker performs poorly. A possible explanation is that due to an over-representation of the positive examples the model is not able to accurately predict negative examples. This can be seen in figure 24b (41). The hybrid ratio for negative examples for marker *D12S391* are indeed incorrect predicted, these ratios are over-predicted.

Table 18: Number of positive and negative hybrids included in the training and test set for the STR dataset.

Markers	**Training set pos. examples**	**Training set neg. examples**	**Test set pos. examples**	**Test set neg. examples**
CSF1P0	809	4040	196	1017
D10S1248	915	1473	215	383
D12S391	2272	1007	557	263
D13S317	1479	8318	340	2110
D16S539	1590	3398	400	847
D18S51	1174	11483	291	2874
D19S433	845	9226	212	2306
D1S1656	3214	3696	800	928
D21S11	2133	10243	564	2530
D22S1045	1096	1368	263	354
D2S1338	1538	4668	373	1179
D2S441	1339	4361	345	1081
D3S1358	1533	6134	372	1545
D5S818	1350	5493	333	1378
D7S820	2825	6308	726	1558
D8S1179	1196	5442	293	1367
DYS391	1267	5209	336	1284
FGA	1180	6412	286	1613
PentaD	2204	13908	546	3483
PentaE	712	3590	169	907
TH01	683	9905	162	2485
TPOX	814	3368	192	854
vWA	1647	11290	421	2814
Amel	27	7907	7	1977
Combined model	33776	148324	8465	37060



(a) Hybrid prediction model based on all markers combined of the STR dataset



(b) Hybrid prediction model based separate markers superimposed on each other for the STR dataset

Figure 24: Combined (24a) and Overlay (24b) hybrid prediction model for the STR dataset. The horizontal axis represents the true hybrid ratios and the vertical axis the predicted hybrid ratio. The red line is a reference line where $\hat{y} = y$. The closer the data points are to the reference line, the more accurately these points are predicted by the model.

The corrections of the most accurate prediction model (individual or combined) per marker have been visualized in figure 25 on page 42. It is evident that the majority of the hybrids have a low ratio (≤ 0.02) of which the greater part is (over)corrected. Of the hybrids that have a higher ratio almost all hybrids are corrected. However, this correction is generally minor, this applies to markers *D8S1179* and *D5S818*. The most effective corrections were applied to marker *D2S441*, all actual hybrid ratios in the range of 0.04 to 0.11 have been reduced to a ratio of approximately 0.01 to 0.03. In total, 54.93% of all data points are over-corrected, 9.96% is corrected and 35.11% has been unaffected by the prediction model.

Table 19: Error measurement, coefficient of determination and explained variance measurements for the STR dataset.

Markers	$\mathbf{R^2}$	**Explained variance**
CSF1P0	9.32×10^{-1}	9.32×10^{-1}
D10S1248	9.60×10^{-1}	9.60×10^{-1}
D12S391	-3.52×10^{-1}	-3.38×10^{-1}
D13S317	-2.98	-2.97
D16S539	7.19×10^{-1}	7.19×10^{-1}
D18S51	8.92×10^{-1}	8.92×10^{-1}
D19S433	8.72×10^{-1}	8.75×10^{-1}
D1S1656	1.03×10^{-1}	1.03×10^{-1}
D21S11	4.54×10^{-1}	4.56×10^{-1}
D22S1045	9.26×10^{-1}	9.26×10^{-1}
D2S1338	2.85×10^{-1}	2.94×10^{-1}
D2S441	6.73×10^{-1}	6.74×10^{-1}
D3S1358	8.70×10^{-1}	8.70×10^{-1}
D5S818	5.59×10^{-2}	6.02×10^{-2}
D7S820	8.32×10^{-1}	8.33×10^{-1}
D8S1179	-8.03×10^{-1}	-8.01×10^{-1}
DYS391	5.84×10^{-1}	5.88×10^{-1}
FGA	5.87×10^{-1}	5.95×10^{-1}
PentaD	2.49×10^{-1}	2.51×10^{-1}
PentaE	8.96×10^{-1}	8.97×10^{-1}
TH01	3.62×10^{-1}	3.88×10^{-1}
TPOX	9.47×10^{-1}	9.48×10^{-1}
vWA	4.29×10^{-1}	4.32×10^{-1}
Amel	4.26×10^{-2}	4.36×10^{-2}
Combined model	1.89×10^{-1}	1.91×10^{-1}

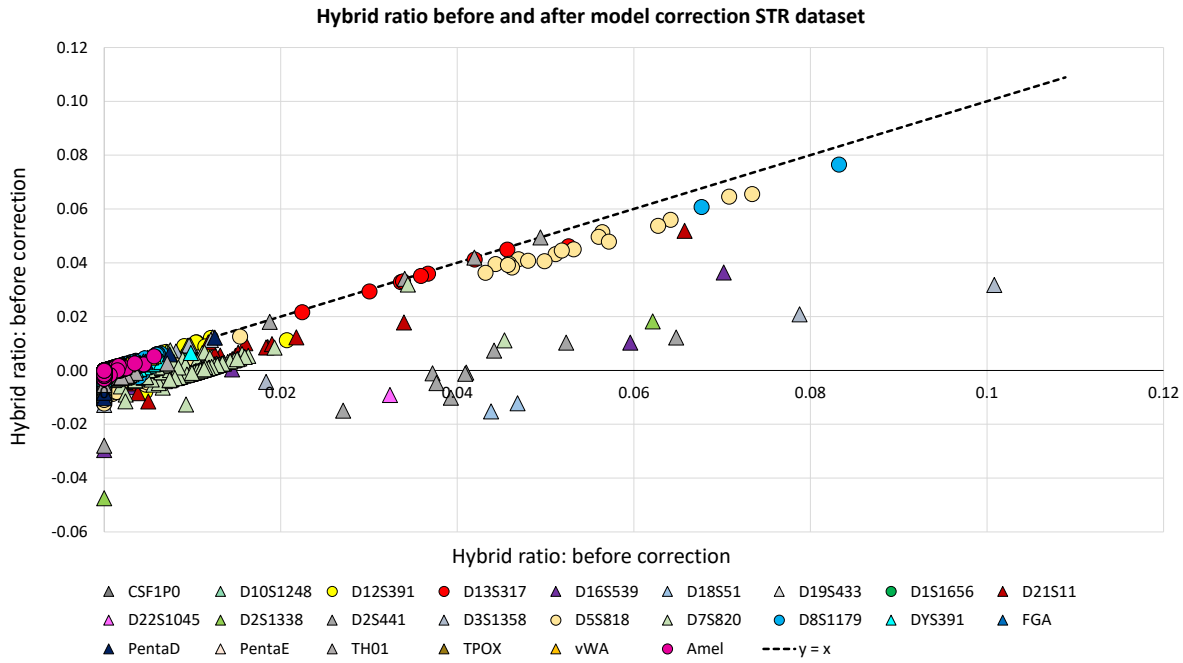


Figure 25: Hybrid ratio prediction correction of the test set of the STR dataset. The horizontal axis represents the hybrid ratio before correction and the vertical axis the hybrid ratio after correction. The triangle shapes are the individual models (trained and tested per marker) and the circles are the markers that are included in the combined model.

3.3.5 Microhaplotype prediction

The Microhaplotype dataset contains the most samples compared to the Mito, Mito-low and STR dataset. However, the Microhaplotype dataset also contains the shortest sequences. Therefore, there is less variation in one sequence or between sequence pairs and consequently less combinations are possible to form hybrids compared to the other datasets. This can be observed in table 20. In total, the Microhaplotype dataset comprises of 21 markers of which the majority did not simulate any positive example with the K-mer method or 'crossing-over' method. There were only 7 markers that resulted in more hybrids than the cut-off value, marker *RS6504633* contained the most positive examples i.e. 729. The train-test split of the markers that were included in the prediction model are illustrated in table 21. It is apparent that marker *rs1721827* has the most imbalanced training set i.e. 0.34% positive examples respectively. The most balanced training set corresponds to marker *rs6595279*, this marker contains 47.86% positive examples.

Table 20: Number of hybrids that have been observed per marker for the Microhaplotype dataset.

Marker	**Number of positive examples**
rs6504633	729
rs6595279	171
rs2594948	4
rs1721827	19
rs7610981	0
rs7610981V2	0
rs4332095	0
rs6870979	0
rs4652604	594
rs348146	148
rs1929060	2
rs35474228	590
rs58329356	0
rs12625560	0
rs13145525	163
rs1738442	0
rs7218712	0
rs28674745	1
rs28674745V2	0
rs1612734	0
rs7632479	2

Table 21: Number of positive and negative hybrids included in the training and test set for the Microhaplotype dataset.

Markers	**Training set pos. examples**	**Training set neg. examples**	**Test set pos. examples**	**Test set neg. examples**
rs6504633	570	8325	159	2065
rs6595279	134	146	37	33
rs1721827	17	5058	2	1267
rs4652604	474	1684	120	420
rs348146	118	696	30	174
rs35474228	472	3955	118	989
rs13145525	130	399	33	100
Combined model	1923	20257	491	5054

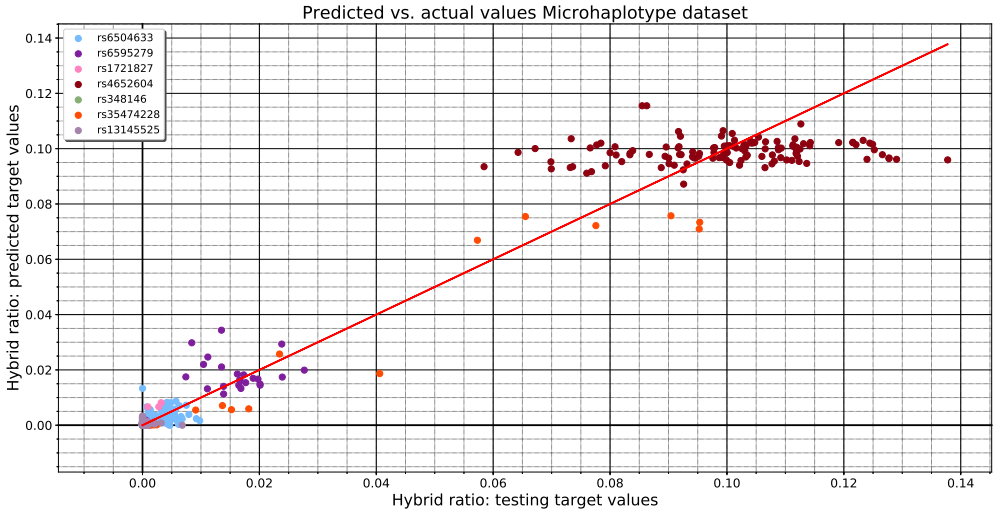
The hybrid ratio predictions for the combined marker model and per-marker model are presented in figure 26 on page 45. On the horizontal axis the actual hybrid ratios are displayed and on the vertical axis the predicted hybrid ratios. The accuracy of the hybrid predictions of both models is very similar, the main difference being the hybrid ratio prediction of marker *rs35474228*. This marker is adequately predicted in the combined model, however in the individual model some hybrids, with a ratio between 0.04 an 0.10 are significantly under predicted. Striking is marker *rs4652604*, this marker contains the most variation in both models in the true hybrid ratio direction, while the predicted hybrid ratios contains significantly less variation. The actual hybrid ratios for the latter mentioned marker vary between 0.05 and 0.14. All these simulated hybrids could be traced back to a single parent-hybrid combination. A possible explanation for the hybrid ratio variance within this marker could be stochastic variation that occurs during the PCR process. The same hybrid is formed earlier in one sample relative to another sample, as a consequence, this hybrid is multiplied more.

For each individual marker model and the combined model the R^2 and explained variance score were calculated in order to determine which model results in a more accurate hybrid ratio prediction, see table 22. It is clear that for the Microhaplotype dataset all markers are more accurately predicted in the combined model than the individual marker models. Additionally, markers *rs6504633, rs6595279* and *rs4652604* do perform exceptionally well in both models, nonetheless the combined model is more accurate.

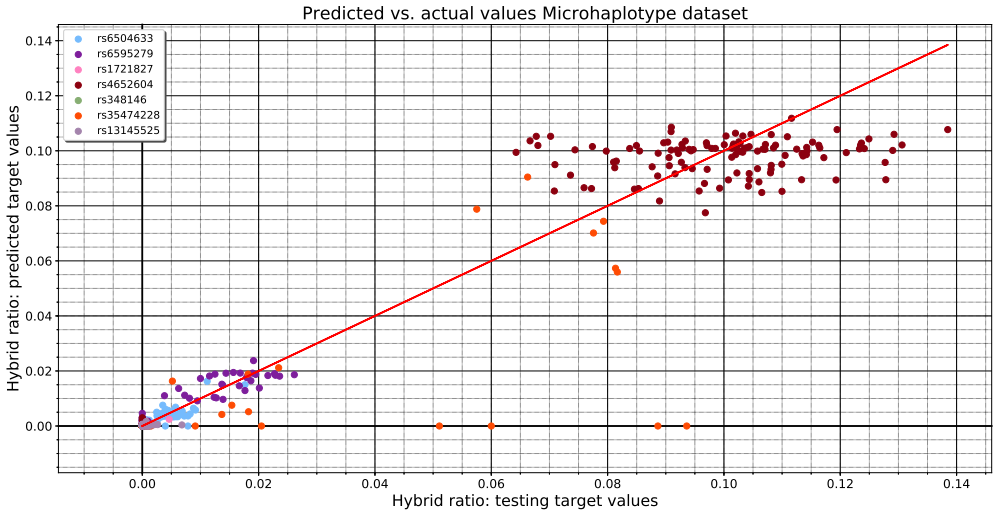
Table 22: Error measurement, coefficient of determination and explained variance measurements for the Microhaplotype dataset.

Markers	$\mathbf{R^2}$	**Explained variance**
rs6504633	8.51×10^{-1}	8.51×10^{-1}
rs6595279	8.72×10^{-1}	8.76×10^{-1}
rs1721827	7.16×10^{-1}	7.16×10^{-1}
rs4652604	9.69×10^{-1}	9.70×10^{-1}
rs348146	1.90×10^{-1}	1.92×10^{-1}
rs35474228	5.45×10^{-1}	5.47×10^{-1}
rs13145525	2.77×10^{-2}	2.79×10^{-2}
Combined model	9.74×10^{-1}	9.74×10^{-1}

The results of the hybrid predictions of the combined model (26a, page 45) have been converted to a correction model, this model is displayed in figure 27 on page 46. The majority of the data points with a ratio smaller than 0.01 are within close range of the reference line, as a consequence these points will not be corrected by the prediction model. However, the hybrids with a ratio between 0.06 and 0.14 are all corrected adequately. In particular, marker *rs4652604* is significantly corrected, 54.20% of this marker is over-corrected, 11.95% is corrected and 89.05% is unaffected by the prediction model. For the complete correction model, a total of 53.33% of all data points are over-corrected, 6.47% is corrected and 40.20% is unaffected by the hybrid ratio prediction model.



(a) Hybrid prediction model based on all markers combined of the Microhaplotype dataset



(b) Hybrid prediction model based separate markers superimposed on each other for the Microhaplotype dataset

Figure 26: Combined (26a) and Overlay (26b) hybrid prediction model for the Microhaplotype dataset. The horizontal axis represents the true hybrid ratios and the vertical axis the predicted hybrid ratio. The red line is a reference line where $\hat{y} = y$. The closer the data points are to the reference line, the more accurately these points are predicted by the model.

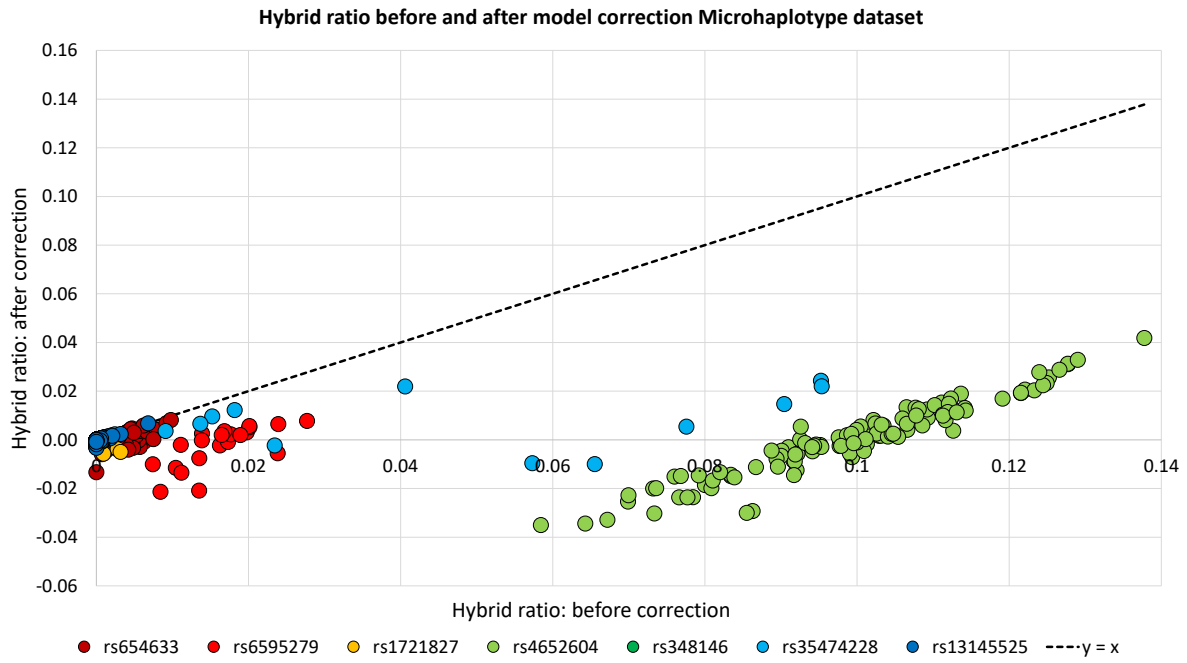


Figure 27: Hybrid ratio prediction correction of the test set of the Microhaplotype dataset. The horizontal axis represents the hybrid ratio before correction and the vertical axis. The prediction corrections are based on the combined model.

3.3.6 Hybrid prediction correction summary

In order to compare the hybrid ratio corrections of all datasets, an overview of all aforementioned percentages is displayed in table 23. The hybrid ratios of the Mito-low dataset are more accurately predicted compared to the other datasets. The larger part of the data points of this dataset are over-corrected, these corrections will all be set to zero in the finished version of the automated prediction and correction model. As a result these hybrids will be filtered from the dataset. The correction model for the Mito-combined dataset performs worse than the individual correction model for the Mito and Mito-low dataset. This because approximately three-fifths of all data points are not corrected at all. In addition, about a third of the points have been unaffected by the prediction model and only 2.30% is corrected. This confirms that hybrid ratios of the Mito and Mito-low datasets are more accurately predicted independently. The over-correction percentages for the Mito, STR and Microhaplotype are all very similar. Of these datasets the majority of the data points are corrected in the STR dataset, approximately 10%.

Table 23: Overview of the results of the prediction model applied corrections to each individual dataset displayed in percentages.

Dataset	Over-corrected	Corrected	Identical
Mito	47.10	8.51	44.39
Mito-low	63.59	2.30	34.11
Mito-combined	34.51	5.88	59.61
STR	54.93	9.96	35.11
Microhaplotype	53.33	6.47	40.2

4 Conclusion

Using MPS for forensic DNA analysis reveals artefactual sequences that originate from the PCR. These artefacts were already existing in traditional CE-based profiles, but with the exception of stutter artefacts, these were not visible due to the lower sensitivity of CE and the inability to detect sequence variation. The most predominant type of artefacts used to be the stutter artefacts. A software package called FDSTools was developed by J. Hoogenboom [2] to apply a correction to these artefacts. This has lead to the recognition of additional interferences by a type of artefacts called 'hybrid sequences'. Hybrid artefacts can wrongfully be interpreted as true alleles, which in forensic DNA analysis can result in overestimating the number of contributors, they can even overshadow true minor contributors which makes it difficult to characterize these. Therefore, in this study, a method has been developed to recognize and predict these hybrid artefacts. The creation of these artefacts could not be avoided by adjusting the PCR protocol, an *in silica* solution was developed in this study.

A hybrid sequence can be formed in two ways i.e. the 'crossing-over' method or the K-mer method. For the recognition of the artefacts all possible hybrid sequences were simulated per sample and subsequently marked if present. This procedure was performed on four datasets: Mito, Mito-low, STR and Microhaplotype. The number of not observed hybrids was always significantly higher compared to the number of observed hybrids, this illustrates that there are always more possible hybrids than will be present in the dataset. The contribution of both methods was checked by examining the number of unique (mutually exclusive) observed hybrids each method simulated. This analysis showed that the 'crossing-over' method contributed to the majority of observed hybrids in the Mito and Mito-low dataset i.e. 91.84% and 68.84% respectively. In addition, the K-mer method contributed to the majority of the observed hybrids of the STR dataset i.e. 90.27% and the 'crossing-over' and K-mer methods contributed equally to the number of observed hybrids of the Microhaplotype dataset i.e. 55.08% and 44.92%. Conclusively, both methods are necessary in order to obtain the most exhaustive hybrid marking tool.

It is not possible to calculate an error of the hybrid marking method for the datasets since the hybrids are not known prior to the marking procedure. In order to calculate an error a synthetic dataset where all the hybrids are known needs to be created. This possibility is further discussed in chapter 5. To quantify the marking and to allow for quantitative correction, a least-squares hybrid prediction model was constructed. This model predicts the hybrid ratio of artefacts based on a feature set that characterizes the formation of the hybrids; observed (positive examples) and not observed hybrids (negative examples) were included in this model. A total of 175 features were created for the prediction model. Subsequently, a genetic algorithm was used to simultaneously lower the prediction error of the model and reduced the number of features to obtain a more computationally efficient model. This worked successfully for the Mito-low and Microhaplotype dataset, the number of features were reduced to 85 and 96 respectively. For the other datasets all 175 features were used for the hybrid ratio prediction model.

The ratios of certain hybrids were not accurately predicted, this can be due to the combination of the least-squares model and features being sub-optimal for predicting these hybrids. In addition, stochastic variance in PCR can result in variance in the dataset for identical hybrid-parent combinations, complicating the prediction of the hybrid ratio for this combination. A hybrid ratio can either be predicted by a model that is specifically trained on that marker or by a model that is trained on all markers of a particular dataset combined. The hybrid ratio prediction model that is most accurate in predicting the ratios is selected as correction model for that marker. Accurate correction of hybrids with high ratios is most important since these interfere most with the interpretation of the true alleles, these hybrids are best corrected in the Mito-low, STR and Microhaplotype dataset. Overall, the majority of the hybrids in the test set of the Mito-low dataset are (over)corrected i.e. 65.59%. However, the Mito-low dataset contained the least amount of observed hybrids. This as opposed to the STR dataset of which 64.89% of the data points were (over)corrected. For the Mito and Microhaplotype datasets 55.91% and 59.8% of the data points of the test set were (over)corrected.

Although the developed recognition and prediction model is already able to correct the for the majority of the data, with further development it can be applied to reduce the number of hybrid sequences in MPS data even more. Conclusively, the hybrid prediction model will be automated and included in FDSTools so it can be a component of the forensic DNA analysis pipeline for casework samples.

5 Future work

The aim of this study was to develop an automated hybrid artefact recognition, prediction and correction tool. Currently, a model has been established that is able to display the predicted hybrid ratio for each hybrid sequence. However, the hybrid sequences are not automatically corrected in the data as of right now. The finished product of this study is a tool that is incorporated in the FDSTools pipeline that is able to filter all possible hybrids sequences that are present in the data.

In order to automate the correction process, a threshold range for the hybrid ratio needs to be determined for which a sequence can be marked as hybrid sequence. All hybrid ratios that fall within this range can be marked as hybrids and filtered from the data. A set of test and validation experiments needs to be performed to set this threshold. This can be accomplished with a synthetic created dataset of which all hybrids are known. Subsequently, the hybrid prediction method simulates the hybrids, calculates the corresponding features and predicts the hybrid ratios based on the pre-trained model. Thereafter, the predicted hybrids can be compared to the true hybrids and a threshold range wherein prediction is sufficiently accurate. As a consequence, the model can be included in the FDStools pipeline.

The least-squares regression model that is used to predict the hybrid ratios of DNA sequences in sample data has been selected because of the simplicity of the model. The functionality of the model is straightforward and can be explained in court if elaboration on the data processing section of MPS is necessary. However, a drawback of the model is that it is linear and is therefore limited to linear relationships and is sensitive to outliers. If there is a nonlinear relationship the model can give inaccurate predictions, all features were raised to the negative power of one, two and three and also squared and cubed in an attempt to compensate for this. An implication of a linear model is that it considers all features to be uncorrelated even though this is often not true. Additionally, a linear model cannot make combinations of features but rather includes all the provided features. An alternative nonlinear model that can cope with these drawbacks is the `DecisionTreeRegressor` of the scikit-learn package of python [11]. The predominant advantages of this model are that it is not affected by outliers and that nonlinear relationships between parameters do not affect the performance of the tree [12].

It requires minimal effort to change from one prediction model to another in code, however the model will still need to be optimized and validated. The hybrid ratio prediction model using least-squares regression will therefore be automated and included in FDSTools, while leaving the option open to switch to a different model like the decision tree regressor. If the latter model proves to be the more accurate hybrid ratio predictor it will be applied instead of the least-squares model.

6 Acknowledgements

First of all, I would like to thank my supervisor Jerry Hoogenboom. If it was not for you I would not have started the Bioinformatics Master in the first place. It was two years ago when we met at the NFI, when I was doing my bachelor thesis in bloodstain pattern analysis. During that project you learned me how to program and analyse my own data. By learning me how to program you gave my future career a new direction. The job offers in the forensic field are very minimal but with a new skill like programming I could develop and specialize myself further. I am very grateful that you gave me this opportunity and supported me during the study. In addition, I would like to thank Kristiaan van der Gaag for his valuable input and for the discussions we had during this project, the results section of the thesis would not have been as complete without your input.

I expand my thanks to the LUMC for providing the STR and Microhaplotype datasets, without the datasets I would not have been able to test my model on other data than the Mito datasets. This would have limited the research possibilities and results of this study.

I would also like to thank Team Research for listening and advising me when necessary. We knew how to keep each other motivated and when it was time for a break. At last, I would like to thank my family and friends, their support was indispensable and a valuable motivation during the study.

7 References

[1] Kristiaan J. van der Gaag et al. "Massively parallel sequencing of short tandem repeats—Population data and mixture analysis results for the PowerSeq^{TM} system". In: *Forensic Science International: Genetics* 24 (Sept. 2016), pp. 86–96. ISSN: 18724973. URL: http://linkinghub.elsevier.com/retrieve/pii/S1872497316300941.

[2] Jerry Hoogenboom et al. "FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise." In: *Forensic science international. Genetics* 27 (Mar. 2017), pp. 27–40. ISSN: 1878-0326. URL: http://www.ncbi.nlm.nih.gov/pubmed/27914278.

[3] A R Shuldiner, A Nirula, and J Roth. "Hybrid DNA artifact from PCR of closely related target sequences." In: *Nucleic acids research* 17.11 (June 1989), p. 4409. ISSN: 0305-1048. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC317979.

[4] A Meyerhans, J P Vartanian, and S Wain-Hobson. "DNA recombination during PCR." In: *Nucleic acids research* 18.7 (Apr. 1990), pp. 1687–91. ISSN: 0305-1048. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC330584.

[5] Shannon J. Odelberg et al. "Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I". In: *Nucleic Acids Research* 23.11 (June 1995), pp. 2049–2057. ISSN: 0305-1048. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/23.11.2049.

[6] Nicholas R. Markham and Michael Zuker. "UNAFold". In: Humana Press, 2008, pp. 3–31. URL: http://link.springer.com/10.1007/978-1-60327-429-6_1.

[7] Takahiro Kanagawa. "Bias and artifacts in multitemplate polymerase chain reactions (PCR)". In: *Journal of Bioscience and Bioengineering* 96.4 (Jan. 2003), pp. 317–323. ISSN: 1389-1723. URL: https://www.sciencedirect.com/science/article/pii/S1389172303901307.

[8] Brian J Haas et al. "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons." In: *Genome research* 21.3 (Mar. 2011), pp. 494–504. ISSN: 1549-5469. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3044863.

[9] Kristiaan J van der Gaag et al. "Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts." In: *Forensic science international. Genetics* 35 (July 2018), pp. 169–175. ISSN: 1878-0326. URL: http://www.ncbi.nlm.nih.gov/pubmed/29852469.

[10] S. Colianni. *Genetic Algorithm Feature Selection*. 2017. URL: https://github.com/scoliann/GeneticAlgorithmFeatureSelection.

[11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[12] Alexandru G. Floares, George A. Calin, and Florin B. Manolache. "Bigger Data Is Better for Molecular Diagnosis Tests Based on Decision Trees". In: Springer, Cham, 2016, pp. 288–295. URL: http://link.springer.com/10.1007/978-3-319-40973-3_29.

Appendices

Appendix

A Feature overview

Table A 1: Overview of designed features including explanation how these are calculated.

'Ratio' features are always calculated by dividing the smaller number by the larger number	
Feature	**Explanation**
Length_parentA	The sequence length of parentA
Length_parentB	The sequence length of parentB
Length_Hybrid	The sequence length of the hybrid
Ratio_parent	The ratio of the total reads of one parent divided by the total reads of the other parent
ReverseA_ForwardA	The ratio of the reverse reads and the forward reads of parentA
ReverseB_ForwardB	The ratio of the reverse reads and the forward reads of parentB
ForwardA_ForwardB	The ratio of the forward reads of parentA and the forward reads of parentB
ReverseA_ReverseB	The ratio of the reverse reads of parentA and the reverse reads of parentB
TotalReverse_TotalForward	The ratio of the total reverse reads and forward reads of parentA and parentB
Range_window	The size of the window in which the hybrid can be formed
Window_ratio_ref	The ratio of the size of the window relative to the size of the reference sequence
Window_ratio_other	The ratio of the size of the window relative to the size of the other sequence
GC_ratio	The ratio of GC nucleotides in the window
Start_pos_ref	The start position of the window on the reference sequence
End_pos_ref	The end position of the window on the reference sequence
Start_pos_other	The start position of the window on the other sequence
End_pos_other	The end position of the window on the other sequence
Start_pos_ref_ratio	The ratio of the start position of the reference sequence relative to the total length of the reference sequence
End_pos_ref_ratio	The ratio of the end position of the reference sequence relative to the total length of the reference sequence
Start_pos_other_ratio	The ratio of the start position of the other sequence relative to the total length of the other sequence
End_pos_other_ratio	The ratio of the end position of the other sequence relative to the total length of the other sequence
TM_window	The melting temperature of the nucleotides in the window
TM_ref	The melting temperature of the reference sequence
TM_other	The melting temperature of the other sequence
TM_Hybrid	The melting temperature of the hybrid sequence
TM_ratio_ref	The ratio of the melting temperature of the window relative to the melting temperature of the reference sequence
TM_ratio_other	The ratio of the melting temperature of the window relative to the melting temperature of the other sequence
Minor_parent_ratio	The ratio of the parent with the lower total reads relative to the total reads of both parents
Major_parent_ratio	The ratio of the parent with the higher total reads relative to the total reads of both parents

B Feature selection genetic algorithm

Table B.1: Most informative features Mito dataset

Mito dataset features	**Exponent** X = feature
End_pos_other_ratio	X
Length_parentA	X^2
Length_Hybrid	X^2
ReverseA_ForwardB	X^2
Minor_parent_ratio	X^2
Ratio_parent	X^3
ReverseA_ForwardB	X^3
GC_ratio	X^3
Start_pos_ref	X^3
Start_pos_ref_ratio	X^3
ReverseA_ForwardA	$1/X$
ForwardA_ForwardB	$1/X$
TotalReverse_TotalForward	X
End_pos_other	$1/X$
Start_pos_ref_ratio	$1/X$
ReverseB_ForwardB	$1/X^2$
Range_window_minus_two	$1/X^2$
Window_ratio_ref	X
Start_pos_other	X/X^2
TM_window	$1/X^2$
TM_ratio_other	$1/X^2$
TotalReverse_TotalForward	X
GC_ratio	$1/X^3$
Minor_parent_ratio	$1/X^3$
Ones	X

Table B.2: Most informative features Mito-low dataset

Mito-low dataset features	**Exponent** X = feature
Length_Hybrid	X
TM_other	X
TM_ratio_ref	X
Window_ratio_other	X^2
End_pos_other	X^2
Major_parent_ratio	X^2
Start_pos_other	X^3
Length_parentA	$X \mid 1/X$
Length_Hybrid	$1/X$
Start_pos_other	$1/X^2$
TM_Hybrid	$1/X^2$
Minor_parent_ratio	$1/X^2$
Window_ratio_ref	$1/X^3$
End_pos_ref	$1/X^3$

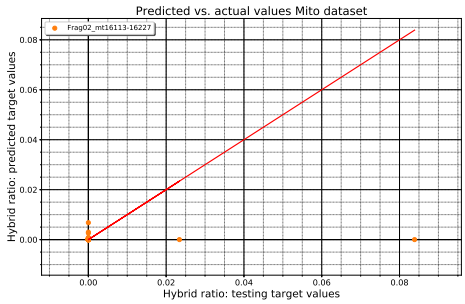
Table B.3: Most informative features STR dataset

STR dataset features	**Exponent** X = feature
Length_parentA	X
TotalReverse_TotalForward	X
GC_ratio	X
Major_parent_ratio	X
TM_Hybrid	X^2
TM_ratio_other	X^2
Window_ratio_ref	X^3
TM_Hybrid	X^2
Length_parentB	$1/X$
Window_ratio_ref	$1/X$
GC_ratio	$1/X$
TM_window	$1/X$
TM_other	$1/X$
ReverseA_ForwardA	$1/X^2$
TM_ratio_ref	$1/X^2$
ForwardA_ForwardB	$1/X^3$
Window_ratio_ref	$1/X^3$
Window_ratio_other	$1/X^3$
GC_ratio	$1/X^3$
TM_ratio_ref	$1/X^3$
Minor_parent_ratio	$1/X^3$

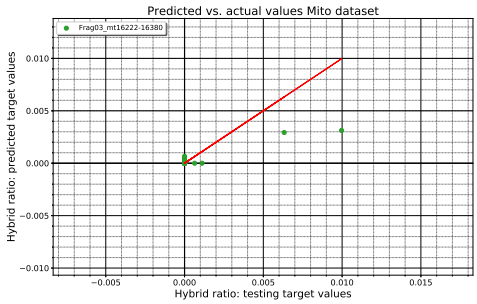
Table B.4: Most informative features Microhaplotype dataset

Microhaplotype dataset features	**Exponent** X = feature
Length_parentB	X
Ratio_parent	X
Range_window	X
GC_ratio	X
TM_other	X
TM_Hybrid	X
Window_ratio_other	X^2
GC_ratio	X^2
TM_ratio_other	X^2
ReverseA_ForwardA	X^3
Range_window	X^3
Start_pos_ref_ratio	X^3
TM_other	X^3
ReverseB_ForwardB	$1/X$
Window_ratio_other	$1/X$
TM_other_minus	$1/X$
TM_ratio_other	$1/X$
ReverseB_ForwardB	$1/X^2$
ForwardA_ForwardB	$1/X^2$
Window_ratio_other	$1/X^2$
GC_ratio	$1/X^2$
TM_window	$1/X^2$
TM_ref	$1/X^2$
TM_ratio_other	$1/X^2$
Length_Hybrid	$1/X^3$
Window_ratio_other	$1/X^3$
End_pos_ref	$1/X^3$
TM_ref	$1/X^3$
TM_ratio_ref	$1/X^3$
Minor_parent_ratio	$1/X^3$

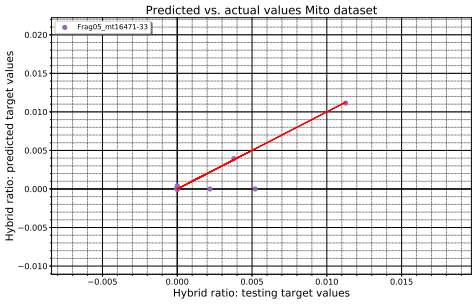
C Least-squares fit per marker Mito dataset



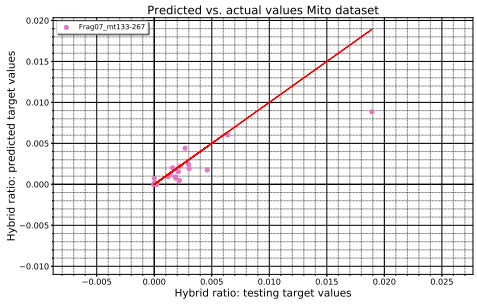
(a) Hybrid ratio prediction: Fragment 2



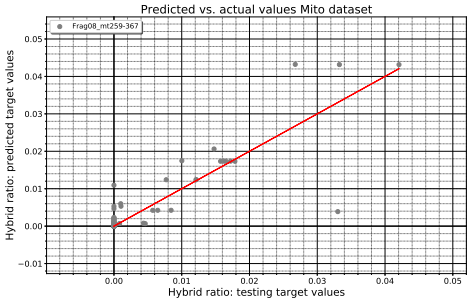
(b) Hybrid ratio prediction: Fragment 3



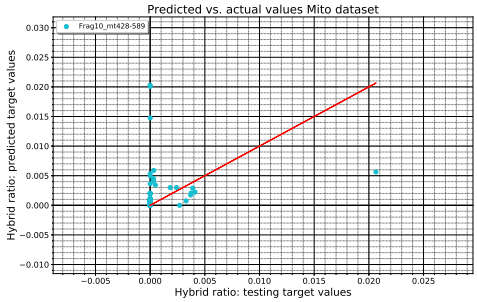
(c) Hybrid ratio prediction: Fragment 5



(d) Hybrid ratio prediction: Fragment 7



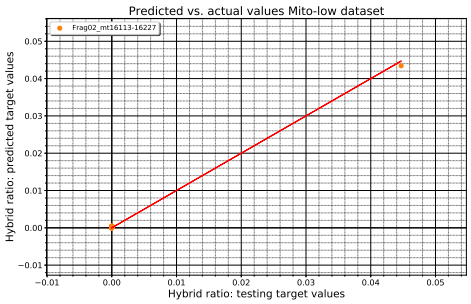
(e) Hybrid ratio prediction: Fragment 8



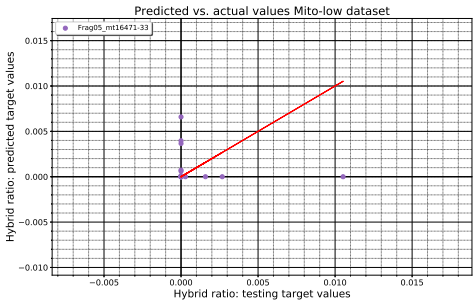
(f) Hybrid ratio prediction: Fragment 10

Figure C.1: Predicted hybrid ratio w.r.t. true hybrid ratio displayed per marker for Mito dataset. The red line is a reference line where $\hat{y} = y$, the closer the points are to the line the more accurate the prediction is.

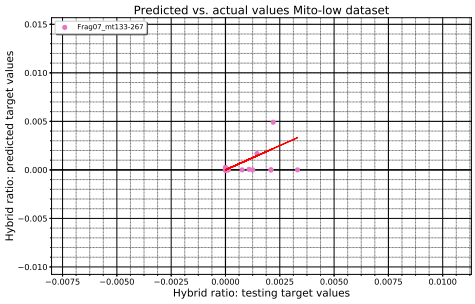
D Least-squares fit per marker Mito-low dataset



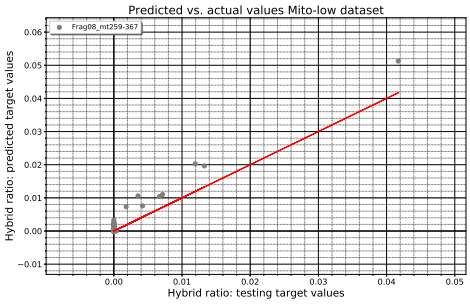
(a) Hybrid ratio prediction: Fragment 2



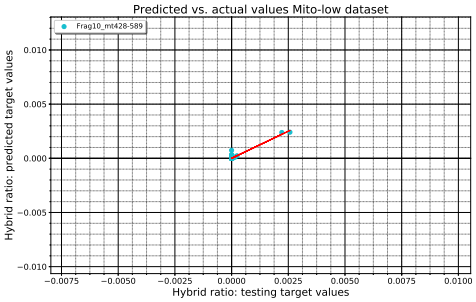
(b) Hybrid ratio prediction: Fragment 5



(c) Hybrid ratio prediction: Fragment 7



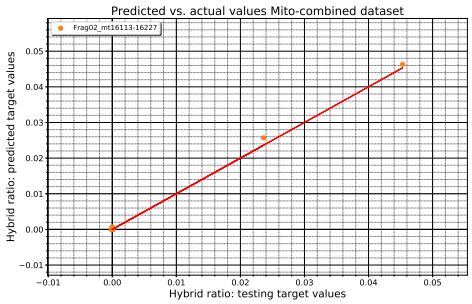
(d) Hybrid ratio prediction: Fragment 8



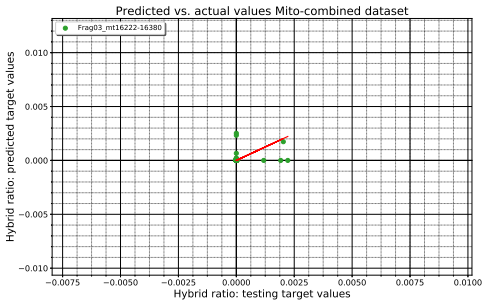
(e) Hybrid ratio prediction: Fragment 10

Figure D.1: Predicted hybrid ratio w.r.t. true hybrid ratio displayed per marker for Mito-low dataset. The red line is a reference line where $\hat{y} = y$, the closer the points are to the line the more accurate the prediction is.

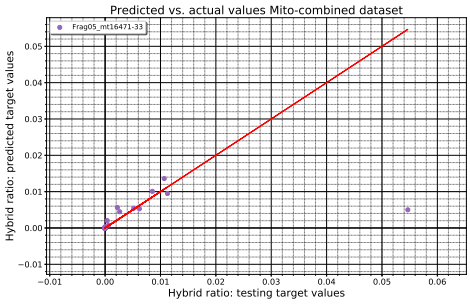
E Least-squares fit per marker Mito combined dataset



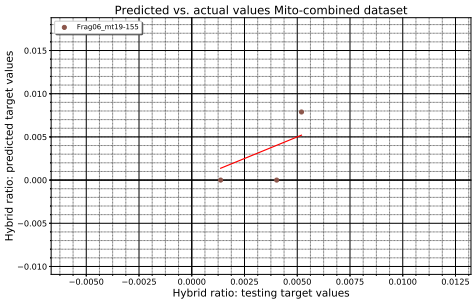
(a) Hybrid ratio prediction: Fragment 1



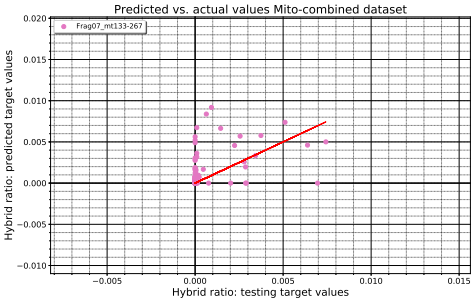
(b) Hybrid ratio prediction: Fragment 3



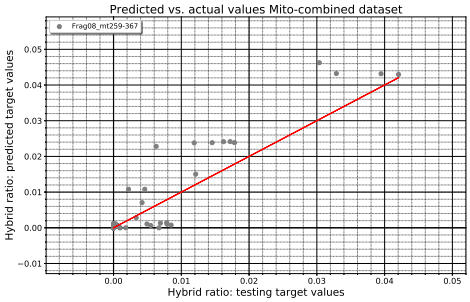
(c) Hybrid ratio prediction: Fragment 5



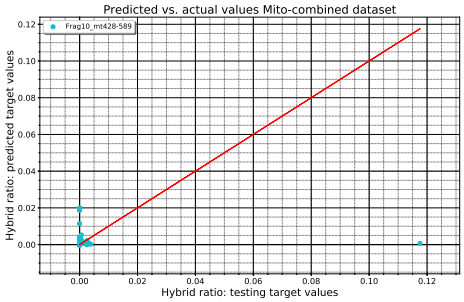
(d) Hybrid ratio prediction: Fragment 6



(e) Hybrid ratio prediction: Fragment 7



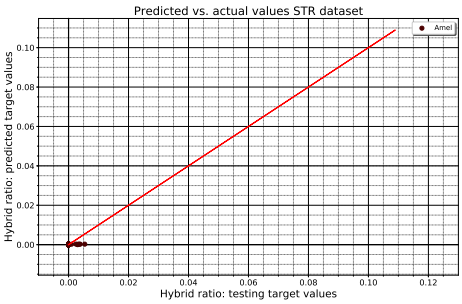
(f) Hybrid ratio prediction: Fragment 8



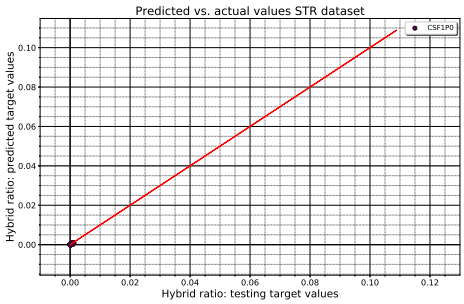
(g) Hybrid ratio prediction: Fragment 10

Figure E.1: Predicted hybrid ratio w.r.t. true hybrid ratio displayed per marker for Mito-combined dataset. The red line is a reference line where $\hat{y} = y$, the closer the points are to the line the more accurate the prediction is.

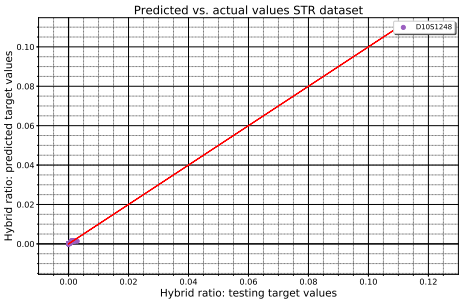
F Least square fit per marker STR dataset



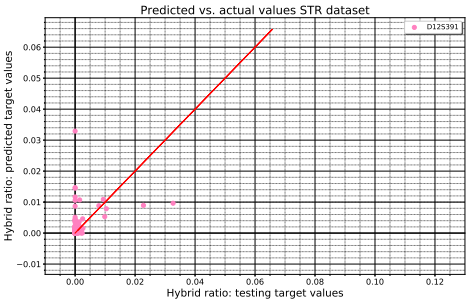
(a) Hybrid ratio prediction: Amel



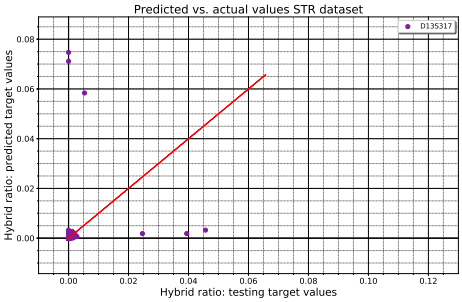
(b) Hybrid ratio prediction: CSF1P0



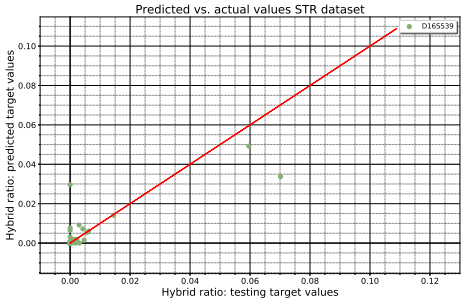
(c) Hybrid ratio prediction: D10S1248



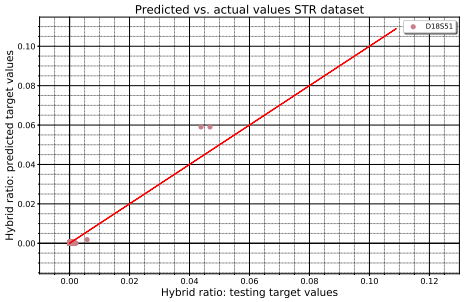
(d) Hybrid ratio prediction: D12S391



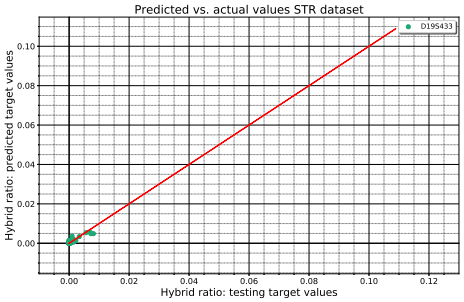
(e) Hybrid ratio prediction: D13S317



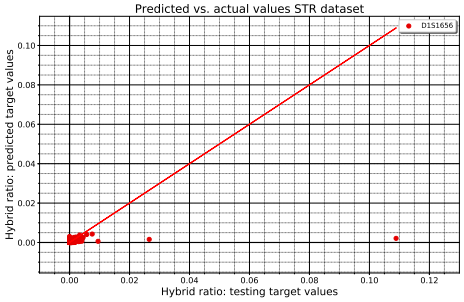
(f) Hybrid ratio prediction: D16S539



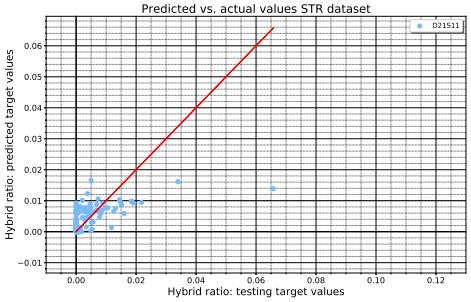
(g) Hybrid ratio prediction: D18S51



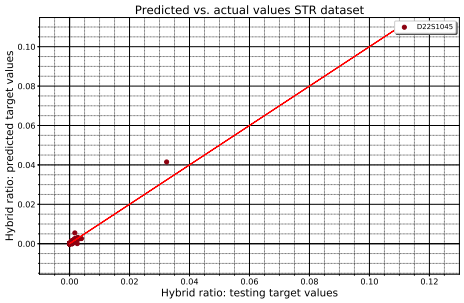
(h) Hybrid ratio prediction: D19S433



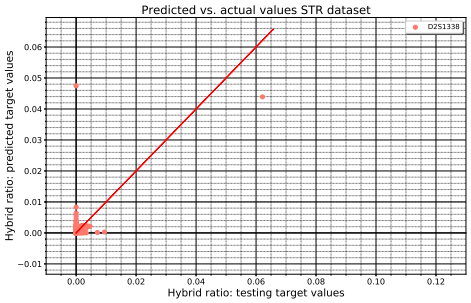
(i) Hybrid ratio prediction: D1S1656



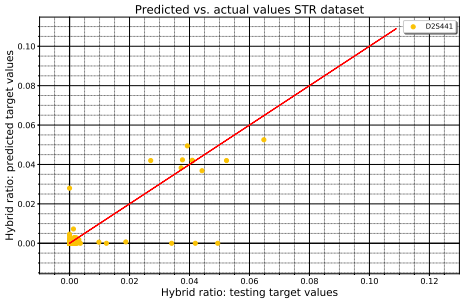
(j) Hybrid ratio prediction: D21S11



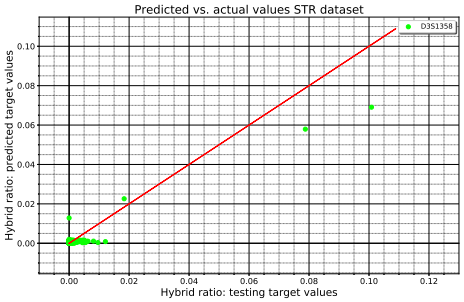
(k) Hybrid ratio prediction: D22S1045



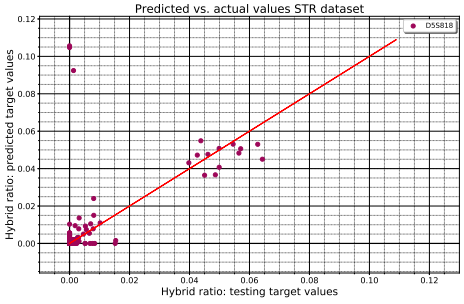
(l) Hybrid ratio prediction: D2S1338



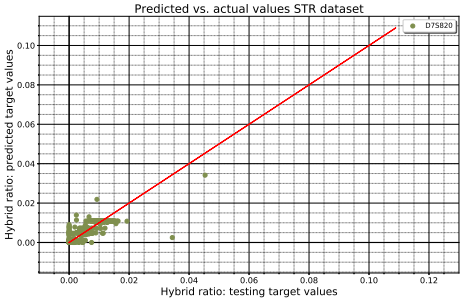
(m) Hybrid ratio prediction: D2S441



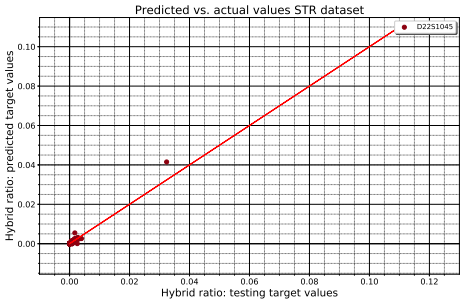
(n) Hybrid ratio prediction: D3S1358



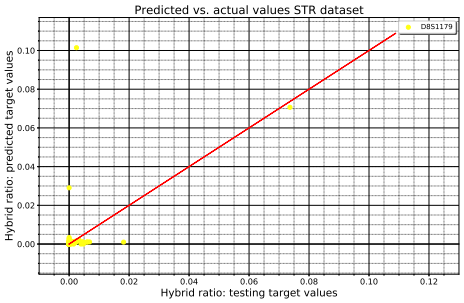
(o) Hybrid ratio prediction: D5S818



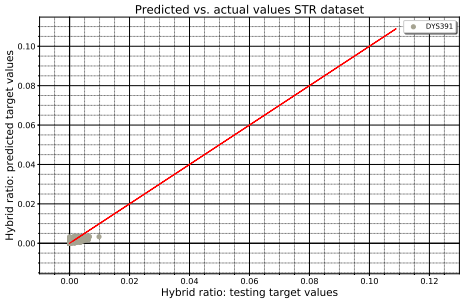
(p) Hybrid ratio prediction: D7S820



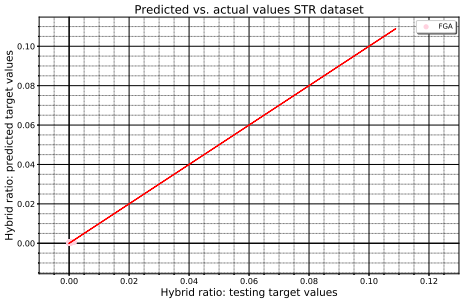
(q) Hybrid ratio prediction: D22S1045



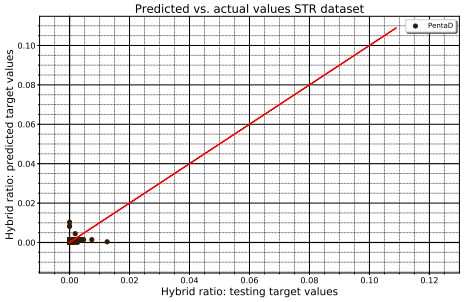
(r) Hybrid ratio prediction: D8S1179



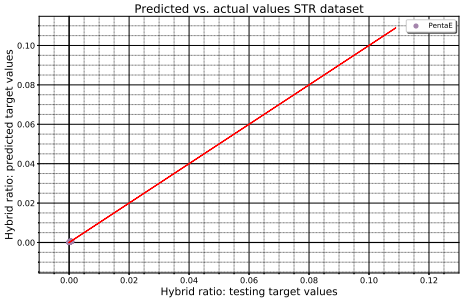
(s) Hybrid ratio prediction: DYS391



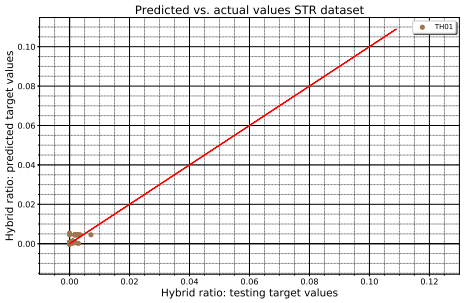
(t) Hybrid ratio prediction: FGA



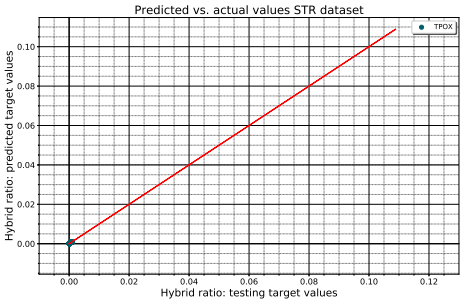
(u) Hybrid ratio prediction: PentaD



(v) Hybrid ratio prediction: PentaE



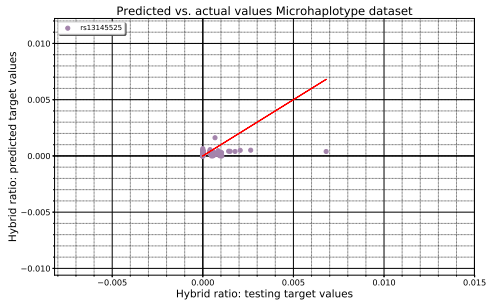
(w) Hybrid ratio prediction: TH01



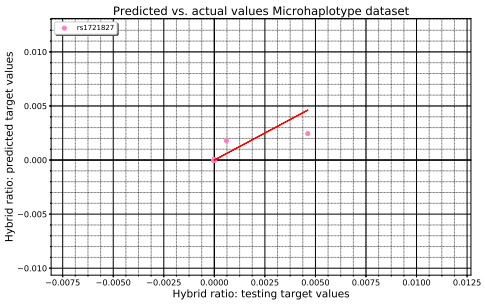
(x) Hybrid ratio prediction: TPOX

Figure F.1: Predicted hybrid ratio w.r.t. true hybrid ratio displayed per marker for STR dataset. The red line is a reference line where $\hat{y} = y$, the closer the points are to the line the more accurate the prediction is.

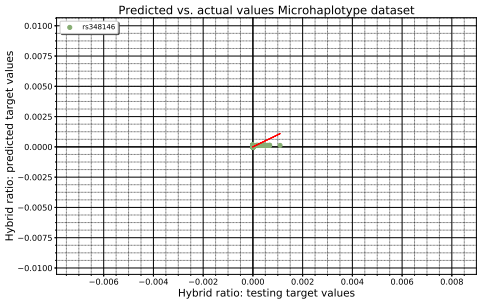
G Least-squares fit per marker Microhaplotype dataset



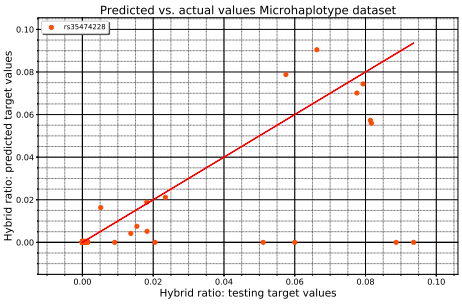
(a) Hybrid ratio prediction: rs13145525



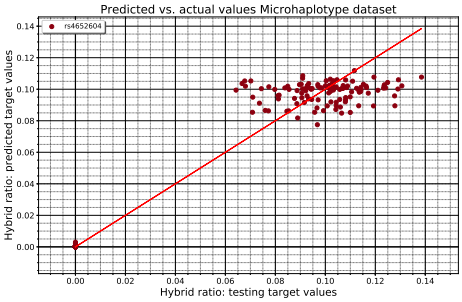
(b) Hybrid ratio prediction: rs1721827



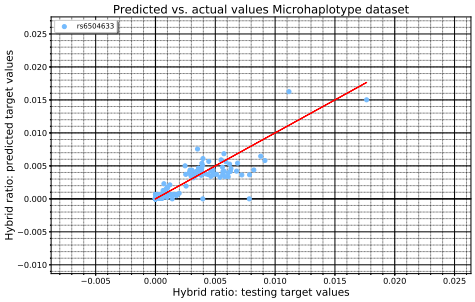
(c) Hybrid ratio prediction: rs348146



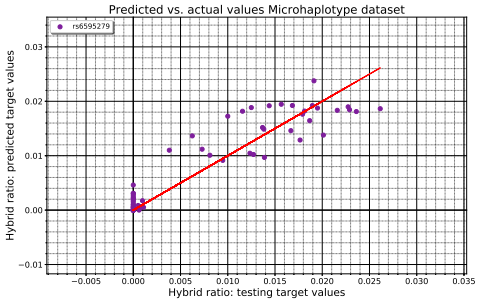
(d) Hybrid ratio prediction: rs35474228



(e) Hybrid ratio prediction: rs4652604



(f) Hybrid ratio prediction: rs6504633



(g) Hybrid ratio prediction: rs6595279

Figure G.1: Predicted hybrid ratio w.r.t. true hybrid ratio displayed per marker for Microhaplotype dataset. The red line is a reference line where $\hat{y} = y$, the closer the points are to the line the more accurate the prediction is.