**Supplementary Material 1: data analysis and sampling**

The original features from *Benschop and van der Linden et al.*[1] can be found in Supplementary Table 1, along with their description and the new names that we implemented for them. As the new feature names would be presented to users, it would be best for them to be as accurate and consistent as possible.

Supplementary Table 1: Overview of the 19 features and their descriptions. We have adopted new names for them to be more consistent with their definitions as these will be presented to users.

|  |  |  |
| --- | --- | --- |
| **Feature name [1]** | **New feature name** | **Description** |
| MAC | MAC | Maximum Allele Count |
| TAC | TAC | Total Allele Count |
| MinNOC\_CSF1PO | CSF1PO min. NOC | Minimal NOC based on locus CSF1PO (allele count at locus CSF1PO / 2, rounded up) |
| MinNOC\_D16S539 | D16S539 min. NOC | Minimal NOC based on locus D16S539 (allele count at locus D16S539 / 2, rounded up) |
| PercAF\_D1S1656 | D1S1656 perc. known alleles | Number of alleles at locus D1S1656 as a percentage of all known alleles at D1S1656 in the allele frequency file |
| AlleleCount\_D3S1358 | D3S1358 allele count | Allele count at locus D3S135 |
| AlleleCount\_D8S1179 | D8S1179 allele count | Allele count at locus D8S1179 |
| MinNOC\_Penta D | Penta D min. NOC | Minimal NOC based on locus Penta D (allele count at locus Penta D / 2, rounded up) |
| MinNOC\_Penta E | Penta E min. NOC | Minimal NOC based on locus Penta E (allele count at locus Penta E / 2, rounded up) |
| SumAF\_TH01 | TH01 sum of allele freq. | Sum of frequencies of the alleles at TH01 defined in the allele frequency file |
| AlleleCount\_TPOX | TPOX allele count | Allele count at locus TPOX |
| MinNOC\_TPOX | TPOX min. NOC | Minimal NOC based on locus TPOX (allele count at locus TPOX / 2, rounded up) |
| stdHeight\_vWA | vWA peak height variation | Standard deviation of peak heights at locus vWA(average variation from the mean peak height at locus vWA) |
| stdAllele | Allele count variation | Standard deviation of the number of alleles per locus (average variation from the mean number of alleles per locus) |
| MAC0 | Loci with 0 alleles | Number of loci with 0 alleles |
| MAC5-6 | Loci with 5-6 alleles | Number of loci with 5 or 6 alleles |
| peaksBelowRFU | Peaks below 800 RFU | Number of peaks below the stochastic threshold of 800 RFU |
| MatchProbability | Random match proba. | Probability of a random Dutch person matching to this DNA profile |
| MinNOC | Min. NOC | Minimal NOC (locus with lowest allele count / 2, rounded up) |

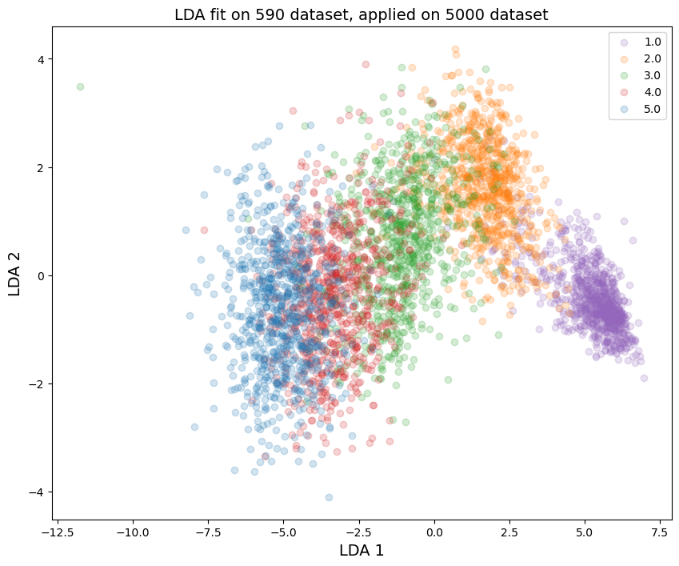
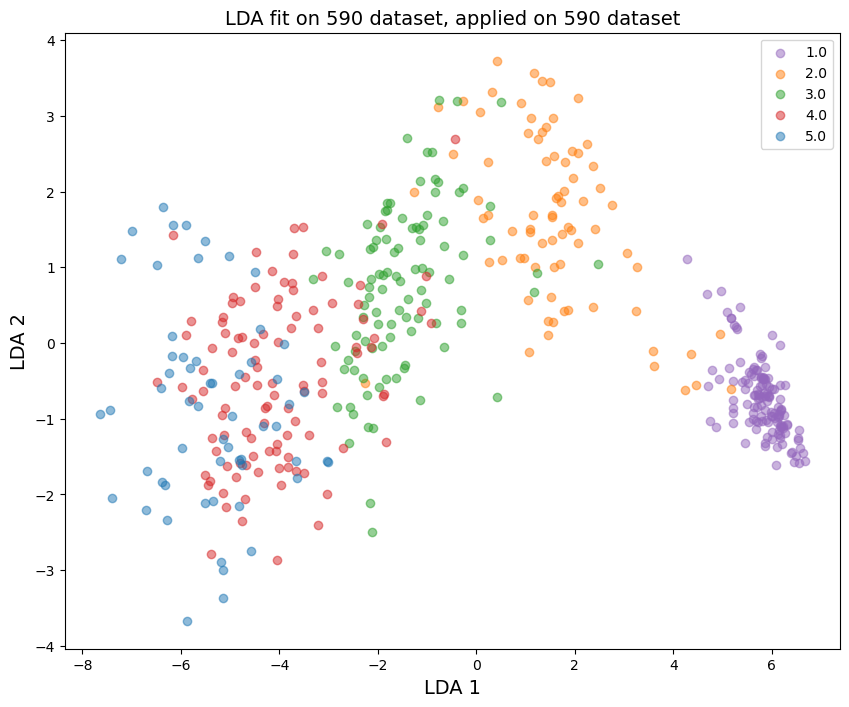
Additional data was sampled to handle the sparsity of the original dataset. Supplementary Table 2 details the parameters used to generate 5000 mixture profiles using a development version of DNAStatistX. For each number of contributors (1-5), 1000 profiles were generated. The used STR kit is PowerPlex Fusion 6CTM (PPF6C, Promega) with dye-specific detection thresholds as used by default at the NFI. Dutch population frequency data was used as described in Westen et al.. After generating a profile, LRs were calculated using each donor in a mixture as the person of interest under H1. Only mixtures for which all donors reached a minimum LR of 1000 were included in the dataset. LR calculations were performed using the true NOC under the propositions, using theta correction of 0.01 and using the kit settings for PPF6C as implemented in DNAStatistX.

Supplementary Table 2: Sampling parameters as used in the development version of DNAStatistX for simulation of the 5000 mixture profiles.

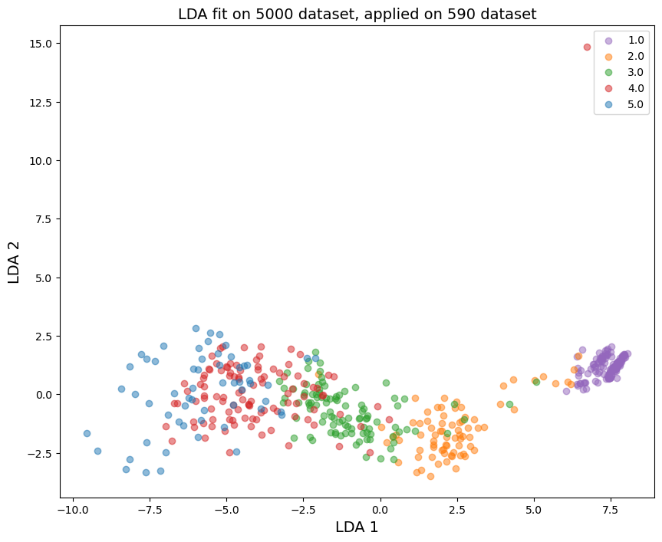
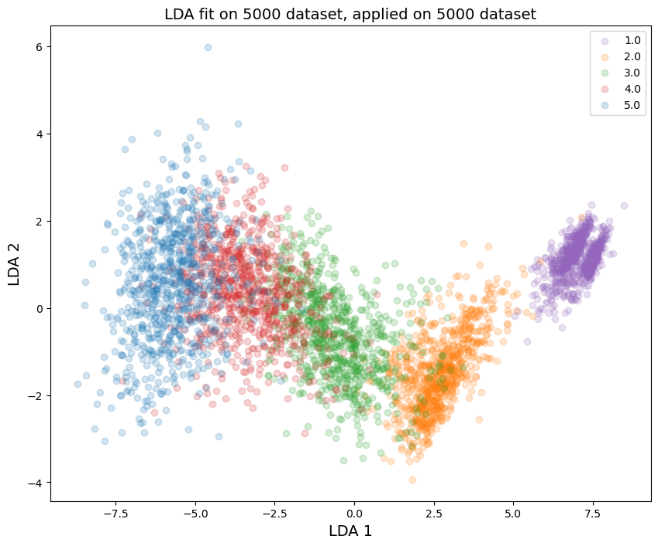
|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Drop-in prC | 0.05 |
| Drop-in lambda | 0.01 |
| Average peak heights | (100, 20000) |
| Variation coefficient peak heights | (0.1, 1.0) |
| Degradation | (0.4, 1.1) |

The following section shows a comparative data analysis of the original dataset and the sampled dataset.

By fitting LDA on the original dataset of 590 samples (from here on referred to as “590-dataset”) and applying it to the dataset of the 5000 samples instances (from here on referred to as “5000-dataset”) as shown in Supplementary Figure 1, it appears that the 590-dataset captures a lot of the variance that is also present in the 5000-dataset. The spread of the 5000 samples is quite broad over the two LDA dimensions. However, looking at Supplementary Figure 2, it seems that by fitting LDA on the 5000-dataset, only LDA 1 captures a good spread of the 590-dataset. LDA 2 does not contain much differential information for the 590-dataset. This is to be expected as the 5000-dataset is artificially created and therefore might contain less unexpected variation which is present in the 590-dataset.

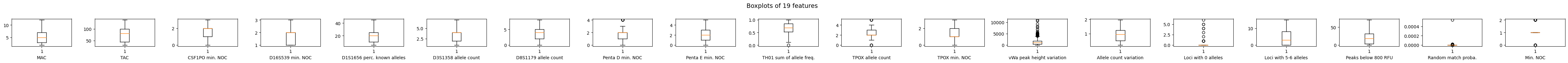


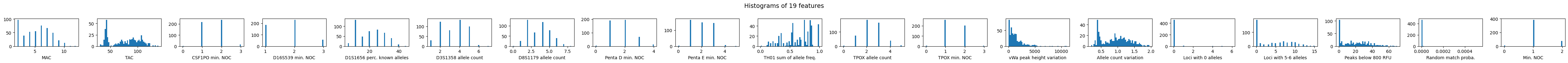
Supplementary Figure 1: LDA fit on the original dataset of 590 samples, applied to both that same dataset and to the dataset consisting of the 5000 sampled instances. DNA profiles consisting of one, two, three, four, or five donors are presented as purple, orange, green, red and blue circles, respectively.



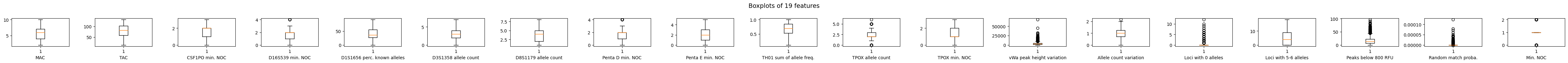
Supplementary Figure 2: LDA fit on the sampled dataset of 5000 samples, applied to both that same dataset and to the dataset consisting of the 590 original instances.

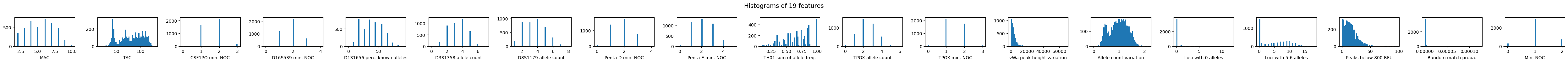
Plotting the 19 features in boxplots and histograms show that they are mainly not normally distributed and contain many outliers. This is true for both datasets as can be seen in Supplementary Figure 3 and Supplementary Figure 4.





Supplementary Figure 3: Boxplots and histograms for each of the 19 features in the training set of the 590-dataset.





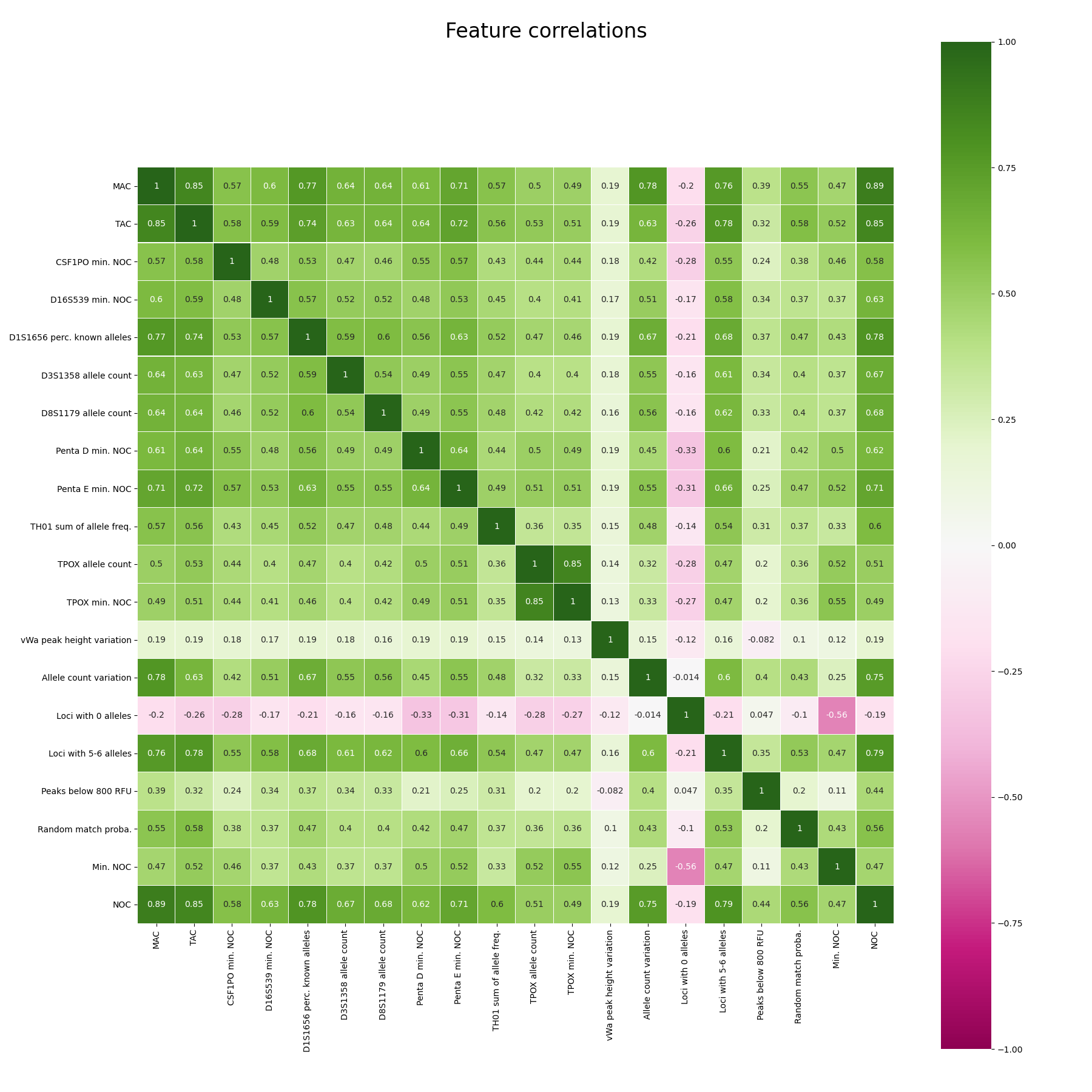
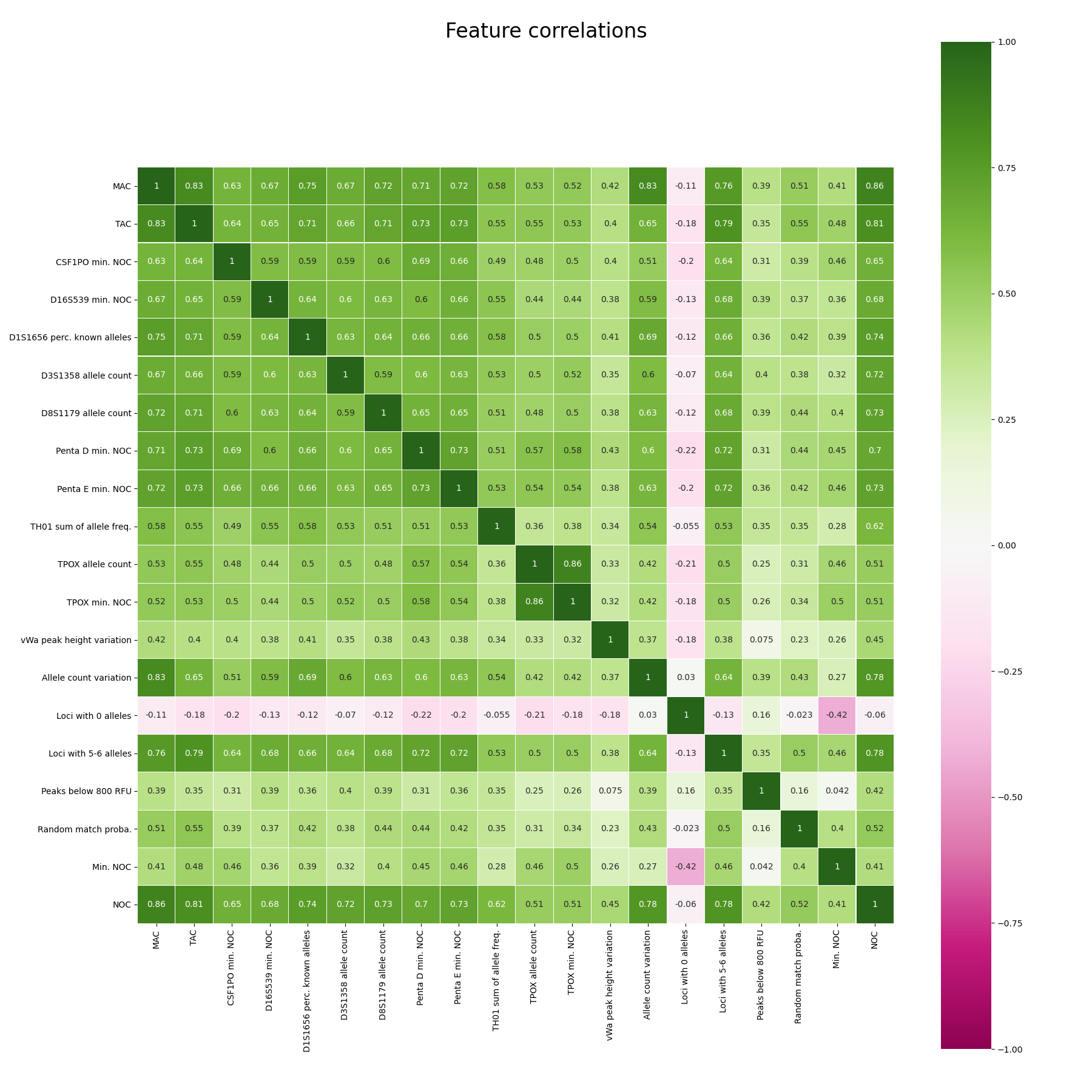
Supplementary Figure 4: Boxplots and histograms for each of the 19 features in the training set of the 5000-dataset.

Ideally all 19 features should follow comparable distributions for both datasets. From visual inspection, it is clear that this is not the case for all these features. By using the two-sided Kolomogorov\_Smirnov (KS) statistic[[1]](#footnote-1), it was determined that 8 features do not appear to be drawn from the same distribution. The results of this statistic are listed in Supplementary Table 3, where a large KS statistic or small p-value (less than 1.0e-2) corresponds to rejecting the null hypothesis assuming the samples were drawn from the same distribution. For discrete variables, a different KS statistic implementation was used[[2]](#footnote-2).

Supplementary Table 3: KS statistic results for the 19 features comparing the 590- and 5000-datasets. Bold features have a large statistic value of small p-value, and therefore the null hypothesis is rejected.

|  |  |  |
| --- | --- | --- |
|  | **KS statistic** | **p-value** |
| **MAC** | **0.12** | **3.0e-7** |
| **TAC** | **0.09** | **1.2e-4** |
| CSF1PO min. NOC | 0.04 | 3.3e-1 |
| **D16S539 min. NOC** | **0.09** | **3.3e-4** |
| **D1S1656 perc. known alleles** | **0.63** | **4.7e-200** |
| D3S1358 allele count | 0.06 | 4.8e-2 |
| D8S1179 allele count | 0.06 | 2.5e-2 |
| Penta D min. NOC | 0.04 | 3.4e-1 |
| Penta E min. NOC | 0.06 | 3.3e-2 |
| TH01 sum of allele freq. | 0.07 | 1.9e-2 |
| TPOX allele count | 0.05 | 9.6e-2 |
| TPOX min. NOC | 0.02 | 9.9e-1 |
| **vWA peak height variation** | **0.42** | **5.7e-83** |
| **Allele count variation** | **0.13** | **7.2e-8** |
| Loci with 0 alleles | 0.05 | 1.6e-1 |
| Loci with 5-6 alleles | 0.05 | 9.0e-2 |
| **Peaks below 800 RFU** | **0.19** | **5.7e-18** |
| **Random match proba.** | **0.10** | **9.4e-5** |
| Min. NOC | 0.05 | 1.6e-1 |

For analyzing feature correlations, we applied Kendall rank correlation coefficient. It is suitable for features that are not normally distributed (as is assumed for Pearson), and is more robust to outliers [2, 3]. Most features are highly correlated as can be seen in Supplementary Figure 5. A slight decrease in correlation can be observed between the 590- and 5000-dataset, though most values still lie above 0.4.

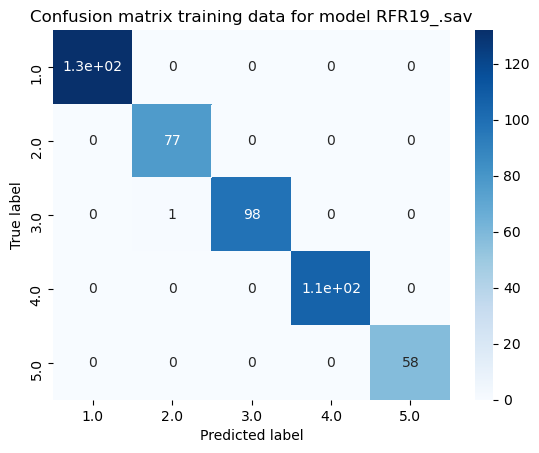
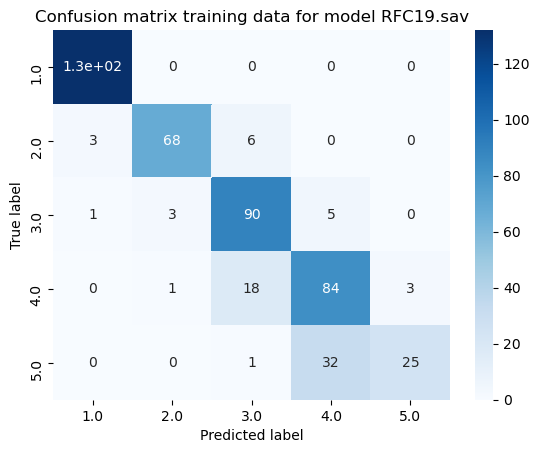


Supplementary Figure 5: Kendall feature correlations of the 19 features in the training set of the 590-dataset (left) and 5000-dataset (right).

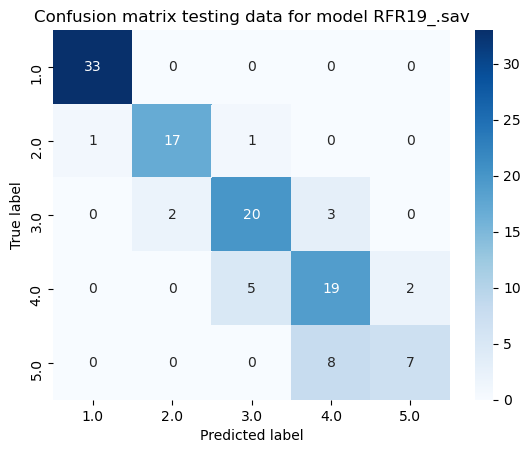
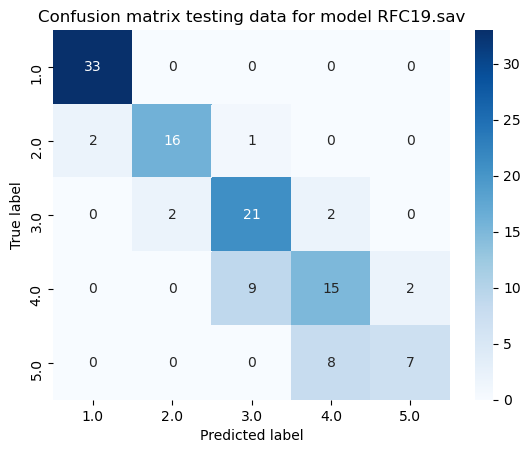
In a short benchmarking study, we demonstrated the ability of regression to outperform classification for NOC estimation. Supplementary Table 4 shows which models were compared, while Supplementary Figure 6 and Supplementary Figure 7 list the results on the training- and test data respectively.

Supplementary Table 4: Models and model parameters used for a short benchmarking study to A) explore if the new merged 5590-dataset has benefits for performance on the original 590-dataset, and B) compare the original model with a default regressor.

|  |  |  |
| --- | --- | --- |
|  | **RFC19 model** | **RFR19 model** |
| Model type | Random forest classifier[[3]](#footnote-3) | Random forest regressor[[4]](#footnote-4) |
| Model parameters | As described in [1] | Defaults |
| Dataset used for A) | 590-dataset | 590-dataset |
| Dataset used for B) | 5590-dataset (merged 590- and 5000-datasets) | 5590-dataset (merged 590- and 5000-datasets) |



Supplementary Figure 6: Confusion matrices of the RFC19 (left) and RFR19 (right) models on the training data from the 590-dataset. The regression model performs almost perfectly (99% accuracy), where the classifier performs a bit worse (85%), and predicts some instances more than one class off.

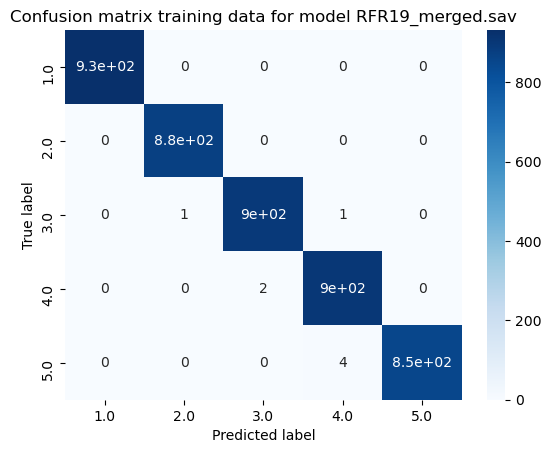
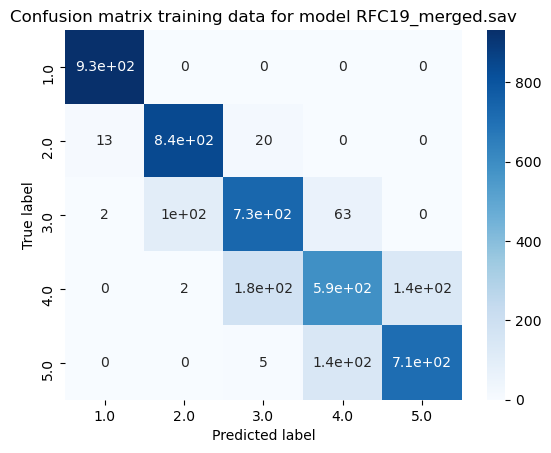


Supplementary Figure 7: Confusion matrices of the RFC19 (left) and RFR19 (right) models on the test data from the 590-dataset. The regressor obtains similar or better results than the classifier (81% accuracy versus 78%). Especially for profiles of 4 donors, there is consistent better performance from the regression model. For profiles of 2, 3, and 5 donors, the number of correct predictions is on average the same for the classifier and the regressor. Multiple fits show that this varies by one or two both in favor of the classifier and the regressor.

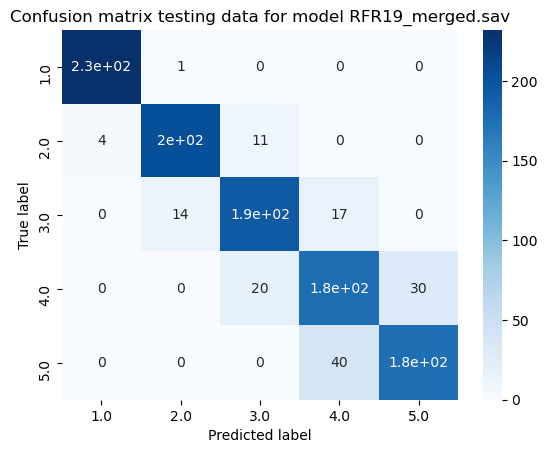
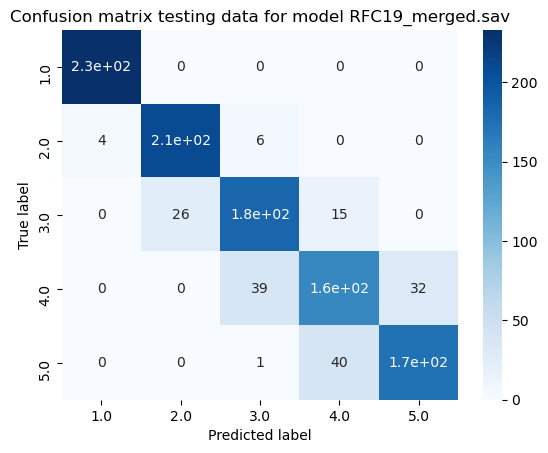
The larger discrepancy between the train and test performance for the regressor is due to overfitting. The default parameters of the regressor in comparison to the optimized classification parameters are more tuned towards larger datasets. These results show promise that a regression model will outperform a classification model once put through more rigorous training.

A similar set-up was used to verify that using the new 5000 samples improved performance, especially on the original dataset. Supplementary Table 4 shows which models and datasets were compared, while Supplementary Figure 8 and Supplementary Figure 9 list the results on the training- and test data respectively.





Supplementary Figure 8: Confusion matrices of the RFC19\_merged (left) and RFR19\_merged (right) models on the training data from the 5590-dataset.



Supplementary Figure 9: Confusion matrices of the RFC19\_merged (left) and RFR19\_merged (right) models on the test data from the 5590-dataset. Accuracy now lies at about 85% for classification, and 88% for regression. Improvement lies mainly for predicting 3 and 4 contributors.

Additional analysis summarized in 5 shows that the regression model trained on the 5590-dataset performs about equally well on samples originating from the 590- and 5000-dataset, while the classification model performs slightly worse on samples from the 590-dataset. The overall performance of the regression model is also slightly better, showing its potential for future applications.

Supplementary Table 5: Performance of the RFC19\_merged and RFR19\_merged models on the test data from the 5590-dataset. We compare the accuracy on samples that originate from the original 590-dataset and the 5000-dataset separately.

|  |  |  |
| --- | --- | --- |
|  | **RFC19\_merged** | **RFR19\_merged** |
| Total test accuracy | 85% | 88% |
| Test accuracy on samples from 590-dataset | 82% | 86% |
| Test accuracy on samples from 5000-dataset | 86% | 88% |

1. Benschop, C.C.G., van der Linden, J., et al., *Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach.* Forensic Science International: Genetics, 2019. **43**: p. 102150.

2. KENDALL, M.G., *A NEW MEASURE OF RANK CORRELATION.* Biometrika, 1938. **30**(1-2): p. 81-93.

3. Molnar, C., et al., *Pitfalls to Avoid when Interpreting Machine Learning Models.* ArXiv, 2020. **abs/2007.04131**.

1. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks\_2samp.html [↑](#footnote-ref-1)
2. https://rdrr.io/cran/dgof/man/ks.test.html [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [↑](#footnote-ref-3)
4. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.htmls [↑](#footnote-ref-4)