Description of the dataset

1. Features590\_278\_original.txt (590 samples, 278 features)
   1. The original Features590 files
2. Features590\_278\_match\_prob.txt (590 samples, 278 features)
   1. For consistency changed: *MatchKans* to *MatchProbability*
3. Features590\_278\_edited.txt (590 samples, 278 features)
   1. Corrected feature *PercAF\_<locus>* (alleles at locus / alleles in AF file for locus)
4. Features5000\_278.txt (5000 samples, 278 features)
   1. Same as 3, but the samples are created in DNAStatistX, 1000 for each NOC
5. Features590\_19.txt (590 samples, 19 features)
   1. Same as 3, but selected the 19 features
6. Features5000\_19.txt
   1. Same as 4, but selected the 19 features

Description by Jerry for generating the 5000 samples:

*“Ik heb gebruikgemaakt van een optie om DNA-profielen te genereren, welke in een developmentversie van DNAStatistX zit. Het genereert realistische profielen omdat hiervoor gebruik wordt gemaakt van hetzelfde model waarmee ook de bewijskrachtberekeningen worden gedaan (EuroForMix model). Het programma genereert eerst tussen opgegeven grenswaarden random parameters voor het model (gemiddelde piekhoogte, piekhoogte variatie, mate van degradatie en mengverhoudingen van de donoren). Vervolgens genereert het één voor één random genotypes voor de donoren, gebruikmakend van het populatiefrequentiebestand. De piekhoogtes van de allelen worden daarbij random gegenereerd op basis van het EuroForMix model. Tot slot worden er (eventueel, at random) drop-in allelen toegevoegd, waarbij opnieuw gebruikgemaakt wordt van het populatiefrequentiebestand en het EuroForMix model voor de piekhoogte. En als extra checkt het nu ook de LR van elke donor en wordt een geheel nieuw profiel gegenereerd als iemand een LR van minder dan 1000 krijgt.”*

Small analysis of chosen features.

When running a randomforestregressor 50 times on the 590 dataset, and 10 times on the 5000 dataset, as well as the partial correlation for both. The top 25 features are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| RandomForestRegressor top features (590) | RandomForestRegressor top features (5000) | Partial Correlation top features (590) | Partial Correlation top features (5000) |
| **MAC5-6** | **MAC5-6** | **MAC** | **MAC** |
| **TAC** | **meanAllele** | **TAC** | **TAC** |
| **meanAllele** | **TAC** | **stdAllele** | **stdAllele** |
| **MAC1-2** | **MAC3-4** | **MAC5-6** | **MAC5-6** |
| **stdAllele** | **MAC** | LowAF\_Penta E | **peaksAboveRFU** |
| **MAC** | **stdAllele** | ***MatchProbability*** | **MAC7-8** |
| **MAC3-4** | **peaksBelowRFU** | **PercAF\_TH01** | stdHeight |
| ***MatchProbability*** | **MAC7-8** | stdHeight\_vWA | PercAF\_Penta E |
| maxHeight\_D2S441 | Below/AboveRFU | **peaksAboveRFU** | medianHeight |
| **minHeight\_D12S391** | Above/BelowRFU | **peaksBelowRFU** | **MAC0** |
| **stdHeight\_D2S441** | **MAC1-2** | **MAC0** | **minHeight** |
| SumAF\_D10S1248 | minHeight\_D1S1656 | **MAC1-2** | MAC9 |
| **SumAF\_D18S51** | minHeight\_D21S11 | MinNOC\_Penta D | **MinNOC\_D13S317** |
| SumAF\_SE33 | minHeight\_D2S441 | **stdHeight\_D2S441** | LowAF\_D2S1338 |
| minHeight\_Penta E | **minHeight** | maxHeight\_CSF1PO | AlleleCount\_SE33 |
| medianHeight\_D7S820 | minHeight\_FGA | maxHeight\_TH01 | **MAC1-2** |
| SumAF\_D12S391 | minHeight\_vWA | **SumAF\_vWA** | MinNOC |
| **minHeight\_D18S51** | minHeight\_D8S1179 | medianHeight\_D21S11 | medianHeight\_D8S1179 |
| SumAF\_D1S1656 | medianHeight\_vWA | stdHeight\_D16S539 | ***MatchProbability*** |
| SumAF\_D21S11 | minHeight\_D2S1338 | **medianAllele** | **PercAF\_TH01** |
| medianHeight\_Penta E | **minHeight\_D12S391** | AlleleCount\_D8S1179 | PercAF\_D3S1358 |
| SumAF\_D16S539 | **minHeight\_D18S51** | PercAF\_D8S1179 | AlleleCount\_D16S539 |
| maxHeight\_vWA | minHeight\_SE33 | medianHeight\_D5S818 | **SumAF\_vWA** |
| minHeight\_TH01 | **SumAF\_D18S51** | MinNOC\_D2S441 | **medianAllele** |
| SumAF\_TH01 | minHeight\_D7S820 | **MinNOC\_D13S317** | LowAF\_D18S51 |

5 features are always important: MAC, TAC, MAC5-6, MAC1-2, stdAllele

1 feature that was in ¾ methods: MatchProbability

Add features:

* Peaks below the detection threshold (can be sampled in EuroForMix, but not available for the original data)

It would improve readability if features were grouped logically. For example, put all profile-specific values first, then locus-specific ones.