Title: An artificial Neural Network system to identify alleles in reference electropherograms

Authors: Duncan Taylor, Ashleigh Harrison, David Powers

PII: S1872-4973(17)30144-8
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2017.07.002
Reference: FSIGEN 1743

To appear in: *Forensic Science International: Genetics*

Received date: 3-4-2017
Revised date: 20-5-2017
Accepted date: 6-7-2017

Please cite this article as: Duncan Taylor, Ashleigh Harrison, David Powers, An artificial Neural Network system to identify alleles in reference electropherograms, Forensic Science International: Geneticshttp://dx.doi.org/10.1016/j.fsigen.2017.07.002

An artificial Neural Network system to identify alleles in reference electropherograms

Author:

Duncan Taylor[1,2], Ashleigh Harrison[2] and David Powers[2]

1. Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
2. Flinders University, GPO Box 2100 Adelaide SA, Australia 5001

Corresponding author:

Duncan A. Taylor, PhD

Forensic Science South Australia

21 Divett Place

Adelaide SA

Australia 5000

Phone: +61-8 8226 7700

Fax: +61-8 8226 7777

Email: Duncan.Taylor@sa.gov.au

Highlights:

- We apply Artificial Neural Networks to electrophotographic data to identify features of reference STR DNA profiles
- The ANN system, coupled with a peak detection system, had a high degree of accuracy when assigning allelic peak centres
- There was a substantial improvement in performance in complex (overloaded) profiles compared to existing solutions
- The ANN system removes the need for an analytical threshold

Abstract

Electropherograms are produced in great numbers in forensic DNA laboratories as part of everyday criminal casework. Before the results of these electropherograms can be used they must be scrutinised by analysts to determine what the identified data tells them about the underlying DNA sequences and what is purely an artefact of the DNA profiling process. This process of interpreting the electropherograms can be time consuming and is prone to subjective differences between analysts. Recently it was demonstrated that artificial neural networks could be used to classify information within an electropherogram as allelic (i.e. representative of a DNA fragment present in the DNA extract) or as one of several different categories of artefactual fluorescence that arise as a result of generating an electropherogram. We extend that work here to demonstrate a series of algorithms and artificial neural networks that can be used to identify peaks on an electropherogram and classify them. We demonstrate the functioning of the system on several profiles and compare the results to a leading commercial DNA profile reading system.

Keywords: electropherogram; gel reading; artificial neural network; allele detection.

Introduction

A common task for any forensic DNA laboratory is the generation of short tandem repeat (STR) DNA profiles. Before these profiles can be used in interpretations they must be scrutinised by analysts to determine whether the information in the profile is representative of some component of DNA in the extract used to generate it, or if it is an artefactual product of the DNA profiling process. This task of 'reading' the electropherogram (EPG) can be time consuming and often leads to subjective differences between analysts. A recent work by Taylor *et al* [1] demonstrated an artificial neural network (ANN) that was trained on two good quality reference EPGs to classify data in the 6-FAM dye lane and then applied to a third (also good quality) EPG with reasonable success. Taylor *et al* [1] provided a proof of concept that ANN could be used to interpret EPGs, which we extend here by:

1) Increasing the amount of training data
2) Increasing the range of training EPG quality from completely blank to highly overloaded
3) Improving on the architecture of the ANN used
4) Training a series of ANN that are used on different areas of the EPG
5) Coupling the predictions of the ANNs with a peak detection algorithm originally designed for LCMS data [2, 3] and recently extended to DNA EPG data [4] to produce a peak detection and classification system

Having created the peak detection and classification system we trial it on several profiles and demonstrate the results, which we compare to the peaks flagged by Genemapper® ID-X (Life Technologies).

Method

ANN input data

An EPG consists of a measure of fluorescence (called relative florescence units, RFU) for a number of dye lanes at various timepoints (called scans). To classify each 'scan' it was deemed that the input data for ANN training would be the scan in question and 100 scans in either direction, in all dye lanes, which corresponds to approximately 8 base pairs (bp) in both directions. This information is presented diagrammatically in Figure 1 for a GlobalFiler™ DNA profile (which possesses six dyes). The result is 201 scans in each of six dyes, leading to 1206 inputs for each training set.

All profiles used in this work have been produced using the GlobalFiler™ DNA profiling system and run on a 3130xl genetic analyser (Life Technologies).

The number of outputs in the ANN depends on the number of categories of feature that we wish to classify. There is a trade-off between distinguishing large numbers of categories so as to provide ANN with many distinct (potentially diagnostic) data patterns, and the increase in

the required training data to generalise well using a large number of classifications. In this work we consider the following features:

- Baseline
- Allele
- Back Stutter (one repeat unit shorter than the allele)
- Pull-up
- Forward stutter (one repeat unit longer than the allele)
- Half Stutter (half a repeat unit, typically 2 base-pairs, shorter than the allele)

We break up the category of 'stutter' into three categories for two reasons. Firstly, they are each distinct in their relative position to allelic peaks and secondly, different loci in the EPG have different combinations of these stutter features. In our training we also classify double back stutter in the 'Back Stutter' category and stutter that is one and a half repeat units shorter than the allele as Half Stutter. There were 10 ANN trained for reading all GlobalFiler™ EPG data (Table 1).

There are a couple of points to note from the information provided in Table 1. Scans in each dye require their own training data as the position of the scan being classified within the 1206 input into the ANN varies (i.e. for the LIZ ANN the scan being classified would sit at position 905 in Figure 1, lower panel) and the pattern of pull-up from the surrounding dye lanes is different. In the cases of VIC, TAZ and SID, multiple ANN are required. In the case of VIC and SID this is due to the different stutters exhibited by different loci within the dyes, specifically while all STR loci exhibit back stutter and forward stutter, the amelogenin and Y-indel loci within VIC are not STR and so exhibit no stutters. Within SID, D1 is the only locus which exhibits half stutter. For TAZ all loci exhibit back stutter and forward stutter, but only SE33 also exhibits Half-stutter and so has its own ANN. Finally, D22 within TAZ is trained separately from all other TAZ loci as it has a different repeat motif (three base pairs rather than the four in others). The consequence of this is that the back stutter and forward stutter peaks will be a different number of scans apart from the allele in this locus than in any other.

Also, becoming apparent from Table 1 is that the training data used for loci includes those loci within the same dye that possess complete coverage of output types. For example, within the SID dye lane, locus D1 possesses 6 outputs and so can use all the data within the SID dye lane as training data. For loci D10, D12 and D2S1338 this will provide training data for the five outputs; baseline, allele, back stutter, pull-up and forward stutter. All half stutter learning for D1_ANN will be for data within the D1 region. However, the SID_ANN (encompassing D10, D12 and D2S1338) can only use data from D10, D12 and D2S1338 regions for training.

ANN training

The ANN architecture for all ANNs was the same, except for the number of outputs (general structure shown in Figure 2). There were 1206 input neurons, three hidden layers of 200, 50 and 20 neurons and 3-6 output neurons depending on the ANN being trained.

Activation functions for the three hidden layers were logistic sigmoid and the output layer was a softmax layer, allowing the outputs to be read as probabilities for the different categories. The cost function used was cross-entropy. Training was conducted using MATLAB R2016a (9.0.0.341360) [5] on an Intel® Xenon® with E3-1505M v5 CPU @ 2.80GHz and 64GB RAM with 64-bit Windows 10 Professional.

Training occurred in a stepped manner. Prior to training, all scan data was normalised in the manner described in [1]. Features were initially learned using stacked autoencoders, trained on 20 profiles. The weights and biases from these autoencoders were transferred to the ANN (with general structure shown in Fig 2) for fine tuning. The ANNs were used to classify scans in a further 30 profiles, and these were then assessed manually to correct any scans that were misclassified. The original 20 profiles and the additional 30 profiles (making a total of 50) were then used for further training of the ANN. Again, the resulting ANNs were used to classify scans in an additional 50 profiles, which were then manually corrected and used in further training of now 100 profiles. Training occurred for data-points between 3000 and 9000 scans, meaning that for the ANNs that utilise data from the full dye range, 600 000 training sets were available. In each instance data was divided into training, test and cross-validation sets (each containing 1206 inputs) in ratios of 70:15:15. The training was run for an indefinite number of epochs, stopping only when the performance of the system on the cross-validation data sub-set had not improved in 100 epochs (and the final weights and biases for the ANN being from the point where the cross-validation performance stopped improving).

Due to the method of training, exact training time cannot be provided, although in total each ANN would have undergone approximately 20 hours of training.

Peak classification using ANN

In order to compare results to other reading software programs a system of peak centre identification and classification was required. We employ the method of [4] whereby a likelihood ratio is assigned to each scan point:

$$LR_{PEAK} = \frac{\Pr(D \mid H_1)}{\Pr(D \mid H_2)}$$

where H_1 is that there is a peak centre at the scan point, H_2 is that there is not a peak centre at the scan point (either because there is no peak nearby, or a peak nearby but with a centre at a different scan point) and D is a window of data that acts as context (in the case of [4] and our work here the window was 21). We made the following adjustment to the algorithm:

- A peak centre was designated as any scan where the arbitrary value $LR_{PEAK} \geq 10$ was obtained
- We make a modification to the data prior to peak centre detection whereby the scan data has the mode of the data subtracted from itself, thereby centering the baseline of the scan data around zero rfu without losing morphological information.
- One base pair is approximately 15-20 scans in the profiles that we examined (this value does vary slightly from run to run depending on the speed at which the PCR fragments migrated through the instrument, which in turn is partially dependant on the surrounding environmental temperature). When two peaks were identified within 10 scans of each other, only the peak that corresponded to the highest fluorescent intensity was labelled and the other potential peak was removed.
- The amplitude of the peaks ('A' from [4]) was designated as the observed height (after baselining) of the scan being examined i.e the central scan for H_1 or the scan being considered a peak centre when not central in H_2

Once a vector of peak centres had been obtained they were compared to the results of the ANNs. For this task, we employed a simple algorithm whereby if the average of the probability of scans within $\pm x$ scans being allelic was greater than y then the peak was labelled, otherwise it was not labelled. We discuss more elegant (but more complex) systems of allele assignment that are possible in the discussions section of this paper. For this work, we use values of $x = 4$ and $y = 0.5$.

Note that the system of peak centre detection and allele assignment do not utilise an analytical threshold, all scan data is utilised in both the peak centre detection and the allele classification parts of the algorithm.

Comparison to Genemapper®

Five reference DNA profiles were chosen that represented a range of intensities, from very weak to very overloaded (example shown in Fig 3). None of these profiles were in the 100 profile set used to train the ANNs. Genemapper® ID-X was used to read these profiles using an analytical threshold of 50rfu (which is the standard threshold for reference profiles in the laboratory that they were produced). Back stutter, forward stutter, half stutter and double back stutter filters were applied in Genemapper® ID-X according to Forensic Science SA validation levels (the average of the detected instances plus three standard deviations on a per locus basis) and peak falling below these threshold will therefore not have been labelled by Genemapper® ID-X.

Results

ANN performance on test profiles

There are a number of ways in which the performance of an ANN can be shown. One common method is through the use of confusion matrices and summary statistics such as recall, precision, F score, markedness and informedness. Informedness is a measure of the power of the model to predict outcomes better than chance [6, 7]. Both a confusion matrix and measures of performance can be produced for all 10 ANNs and become somewhat overwhelming. We show (in Tables 2 and 3) the confusion matrix and summary statistics for all predictions using all ANNs and for all five profiles examined (profile by profile and ANN by ANN confusion matrices are given as supplementary material).

The informedness value of 0.868 is an approximate 8% improvement over the same value from the initial study [1], which demonstrates the benefit provided by the additional training data. Due to the increased complexity of assignments as the EPGs become more intense there is a trend of increasing both the precision/recall and the informedness values going from profile 1 to 5 (Fig 4).

In Fig 5a we show the results of the peak and allele assignments on profile one (the most intense, profile). The black lines show the positions of potential peaks as determined by the algorithm of [4] and that have been assigned as allelic by the ANNs. In Fig 5b we show the predictions for all scans for all of the possible outputs.

In Fig 6a we show the results of the peak and allele assignments on profile five (the last, lest intense, profile). Again, the black lines show the positions of potential peaks as determined by the algorithm of [4] and assigned as allelic by the ANNs. In Fig 6b we show the predictions for all scans for all of the possible outputs. The intensity (rfu) scale in Fig 6b has been reduced to show the baseline more clearly than in Fig 5a. The similar results for profiles 2, 3 and 4 are given as supplementary material.

Of interest in Figure 5 is that the ANN have identified some areas as allelic within the high level of noise present in the baseline. Take, for example, the peak that has been labelled as allelic at point A and F in Fig 5 (these are shown in close-up in Fig 7). Whilst very weak in comparison to the main DNA donor it does not sit in any stutter position (ruling out stutter) and no large peaks sit in the same scan range within different dye lanes (ruling out pull-up). An allelic peak as small as this within a very crowded baseline has a high probability of being overlooked, or simply removed along with all other small peaks, in a human read of the EPG.

In Fig 5, 49 peaks were flagged as allelic, and 42 of these were accepted (meaning 7 were not accepted). We show several of the flagged areas that were rejected in Figure 7.

In Fig 7:

- Interest point A appears to be an instance of a high pull-up peak being labelled as allelic
- Interest point B shows two low level peaks that have been flagged as allelic and do appear to be allelic (a low level second contributor to the profile)

- Interest point C shows a peak that has been flagged and rejected and appears to be due to an incomplete adenylation of the PCR products. This type of artefact was not one of the classification categories used in training the ANNs and it is therefore not surprising that it has been flagged as it is otherwise allelic in appearance
- Interest point D (not shown) was another instance of an incomplete adenylation being flagged
- Interest point E shows a pull-up affecting the morphology of the baseline and being flagged as a peak by the peak detection algorithm as it is within an area that has been assigned a high probability of being allelic by the ANN.
- Interest point F is an instance where a half stutter has been classified partly as half stutter and partly as allelic by the ANN and so has been flagged using the thresholds set for our work. Also shown are a low level peak that has been correctly flagged as allelic and an area of baseline that has been flagged as allelic, which appears to be simply a small 'mound' of baseline noise.
- Interest point G has flagged part of the leading edge of a peak as allelic. There is no apparent reason for this.

For profile five (results seen in Fig 6) there are three allelic peaks that were identifiable in the DNA profile that were not flagged as allelic as the ANN prediction probability was not high enough. One of these was at the low molecular weight area of the NED dye (point A in the upper pane of Fig 6) at 15rfu. One of these was in the high molecular weight area of the TAZ dye lane (point B in the upper pane of Fig 6) at 22rfu and the other in the high molecular weight area of the SID dye lane (point C in the upper pane of Fig 6) at 21rfu. In all instances, some probability was given to the allelic category, but more was given to the baseline category. Also, in all instances the peaks were flagged by the peak detection algorithm, it was the lack of probability assigned to the allelic category by the ANN that was the cause of the failure to be flagged. It is likely that more low-level training data would be required for the ANNs to identify these areas as allelic. Interestingly, other alleles at levels as low as 16rfu were classified as allelic.

In the profile seen in Fig 6, 37 peaks were flagged and each was accepted as allelic (i.e. there were no instances of areas of baseline noise that appeared to be flagged). Reading the same profile in Genemapper® ID-X identified 26 peaks. Analysing the profile in Fig 5 in Genemapper® ID-X flagged 109 non-artefact peaks, significantly more than the 49 flagged by the ANN system (especially when considering that the ANN had no analytical threshold, whereas Genemapper® ID-X had an analytical threshold of 50rfu, screening out many of the lower peaks). Fig 8 shows the results from the Genemapper® ID-X analysis of profile 1.

1

Table 4 gives the results of the ANN system and Genemapper® ID-X for the five test profiles (peak counts in Table 4 do not include the LIZ size standard lane).

ANN learned features

We can investigate, in a rough sense, the features of DNA profiles that the ANNs have identified as being partially diagnostic for classifying scan data into categories. Using the 6-FAM_ANN as an example we considered all 600 000 sets of training data (each consisting of 1206 inputs, that are made up of 201 inputs from each of the 6 dyes as shown in Figure 1) that yield an assignment of a chosen category above a set probability (for our example we arbitrarily choose 0.95). We then overlay the 1206 scan inputs (with mode subtracted so that the baseline sits at 0), broken into their separate dye lanes and display the results as a heat map. Figure 9 shows the results of the 494 796 training sets that were assigned as baseline.

Figure 9 shows that baseline is classified when there is a lack of fluorescent data at the position being classified and also when there is a general lack of data at the VIC, TAZ and SID dye lanes. Interestingly there is also a noticeable void in the 6-FAM to either side of the scan point being classified. This corresponds to the approximate size of a repeat motif, so that when baseline is assigned it is very common that baseline will also be present one repeat in either direction i.e had it been strongly fluorescent one repeat away then we would expect the central scan to be a stutter.

Figures 10, 11 and 12 show a similar layout of results as in Figure 9, but for the 25 892 training sets classified as allele, 10 333 as back stutter and 17 444 as pull-up.

As expected scans are classified as allelic (Fig 10) when there is fluorescence at the scan being classified and lack of fluorescence in all other areas (in adjacent areas in 6-FAM and other dyes). The main feature in the back stutter heatmap (Fig 11) is the fluorescence to the right of the scan being classified and again a lack of fluorescence in all other areas. The main feature in the pull-up heatmap (Figure 12) is the strong areas of fluorescence in the scan points corresponding to the scan being classified, but in the VIC, TAZ and SID dye lanes.

Discussion

Performance on profiles

We have demonstrated the performance of an ANN system, coupled with the Bayesian peak detection algorithm of [4] on EPGs that range from very weak to overloaded. Profiles 1 and 5 were deliberately chosen to be outside the intensity range that would normally be accepted for references in a Forensic setting. This choice was deliberate as we wished to explore the limits of performance of the ANNs to classify scan data in EPGs. Even at the extremes of intensity, and remembering that we have employed a single ANN for each area of the EPG (i.e. we do not utilise separate ANNs for weak and intense profiles) we feel that the performance is

admirable. Table 4 shows that in profiles 1 and 2 there was a substantial improvement (compared to Genemapper® ID-X) in the ability of the ANNs to distinguish alleles from artefacts. For the most intense profile, the practical result of this in a laboratory would be that an analyst is required to scrutinise 60 fewer peaks. At the other extreme (low intensities) the absence of an analytical threshold in the ANN system meant that additional allelic information could be identified. For profiles 3 and 4, which would typically be considered good quality profiles (i.e. not too intense so that the data become overloaded and not too weak so that data was at risk of not being detected) both the ANN system and Genemapper® ID-X performed roughly equally.

The performance of ANNs is highly dependent on the training data used. In this study, we provided 100 profiles as training data within which was a range of profile intensities. Figure 13 shows the spread of profile intensities (as a histogram of the most intense scan point, after baselining, between 3000 and 9000 scans).

In the five test profiles, two broad areas requiring improvement were immediately identified:

1) An improvement in the ability to distinguish low intensity peaks, <20rfu, from baseline
2) An ability to distinguish incomplete adenylation. Note that this second improvement would only be required if the ANNs were going to be used to interpret data in substandard (due to either overloading or run problems) EPG data.

In both cases additional training data targeting this type of profile could be acquired. An ideal workflow would be one where the ANN system was used to read EPGs for casework and any misclassifications would be automatically identified and corrected by the analyst and then fed back into growing training data sets. This system would, however, be susceptible to training an ANN that makes the same decisions as humans, even if this decision making was flawed in some way.

Assigning peaks centres as alleles

Earlier we mentioned that there are more elegant ways of assessing allelic peak centres than the $\pm x$ scans having an allelic probability $>y$. One possibility is to use a post-processing ANN. Such an ANN would take into account the predictions (outputs) of the ANN shown in Fig 2 for some number of surrounding scan points (we use x again here to be analogous to our current method) as inputs, and pass the information through the ANN to the output layer (which would be two nodes that specified the probability of the scan being a peak centre, or, not a peak centre). The post-processing ANN would then be able to learn to identify features based on information such as:

1) If scans approximately one STR repeat unit upstream are being classified as stutter then it is likely that the scan of interest is allelic. Similar relationships would exist for forward and half stutter peak and alleles
2) If the surrounding scans have been classified with high probability as allelic then it is more likely that the scan of interest is allelic also (as features in EPGs tend to span 10 to 15 scans)
3) The peak detection probability could be used as an input into the ANN

which were not available to the ANN in Fig 2. We show an example of a post-processing ANN in Fig 14 using outputs from a five output ANN as inputs.

Using the post-processing ANN seen in Fig 14 would then provide a posterior probability for each scan point as being an allelic peak centre or not being a peak centre. Then a user would simply provide a probabilistic cut-off for when they wished data to be labelled. This would be one step closer to the world without an analytical threshold written about in [1].

Potential ANN improvements

The ANN used in this work are relatively simple, fully connected, feed forward ANNs. The idea of utilising surrounding scan data predictions (which gave rise to the idea of the post-processing ANN seen in Fig 14) could also be incorporated into a ANN in the form of a recurrent neural network, RNN (see [8, 9] for explanations of training and use). A RNN utilises the outputs of previous training sets as inputs in future training sets. It is similar to the ANN in Fig 14 except that it can only use data from preceding scan points (i.e. in Fig 14 this would be scan points from -x to 0). The advantage of RNNs is that the weights and biases for both the scan input data and the previous predictions can be adjusted all together in the one analysis, hence giving the whole system a greater capacity for learning features.

Another ANN feature that could be utilised in this type of analysis is convolutional neural networks (see [10] for an example of the use of convolution NN on visual data). Convolutional NNs have the ability to learn localised features within the input data that may occur at differing points. This would potentially be useful in DNA profile data: for example, a 'peak' feature could be learned from all data (in all dye lanes) and hence obtain the benefit of additional training material for each of the 10 ANNs.

Conclusion

Artificial neural networks are a tool that is becoming more and more popular to find patterns in, and make predictions from, large amounts of data. Electrophoretic signals that make up EPGs are a perfect candidate for applying ANN to reduce the subjective and laborious task of manually classifying data as allelic or artefactual. The current system is still basic in terms of ANN architecture complexity, but already demonstrates power to identify and classify peaks.

For the more difficult profile (those that are very weak or overloaded) the system is already performing better than currently available alternatives. The greatest benefits for an ANN system such as the one we have demonstrated will be in the interpretation of complex mixed DNA profiles. In these situations there is a genuine difficulty in distinguishing weaker contributors in a mixed DNA profile from artefacts that appear at the same intensity (such as pull-up peaks present because of the intense peaks of major DNA contributors). This is an area we intend to explore next and are hopeful to demonstrate a comparison of ANN assignment and human interpretation (rather than just comparison to other software performance), which we feel is the ultimate test.

Supplementary material

- Profile by profile and ANN by ANN confusion matrices.
- Results of peak detection and assignment for profiles 2, 3 and 4 (as in Fig 5 and 6)

Acknowledgements

Points of view in this document are those of the author and do not necessarily represent the official position or policies of their organisations.

References:

[1] Taylor D, Powers D. Teaching artificial intelligence to read electropherograms. Forensic Science International: Genetics. 2016;25:10-8.

[2] Woldegebriel M, Gonsalves J, Asten Av, Vivo-Truyols G. Robust Bayesian Algorithm for Targeted Compound Screening in Forensic Toxicology. Analytical Chemistry. 2016;88:2421-30.

[3] Woldegebriel M, Vivo-Truyols G. Probabilistic Model for Untargeted Peak Detection in LC-MS Using Bayesian Statistics. Analytical Chemistry. 2015;87:7345-55.

[4] Woldegebriel M, Asten Av, Kloosterman A, Vivó-Truyols G. Probabilistic Peak Detection in CE-LIF for STR DNA Typing. ELECTROPHORESIS. 2017.

[5] MATLAB Neural Network Toolbox Release 2016a. Massachusetts, United States: The MathWorks Inc.

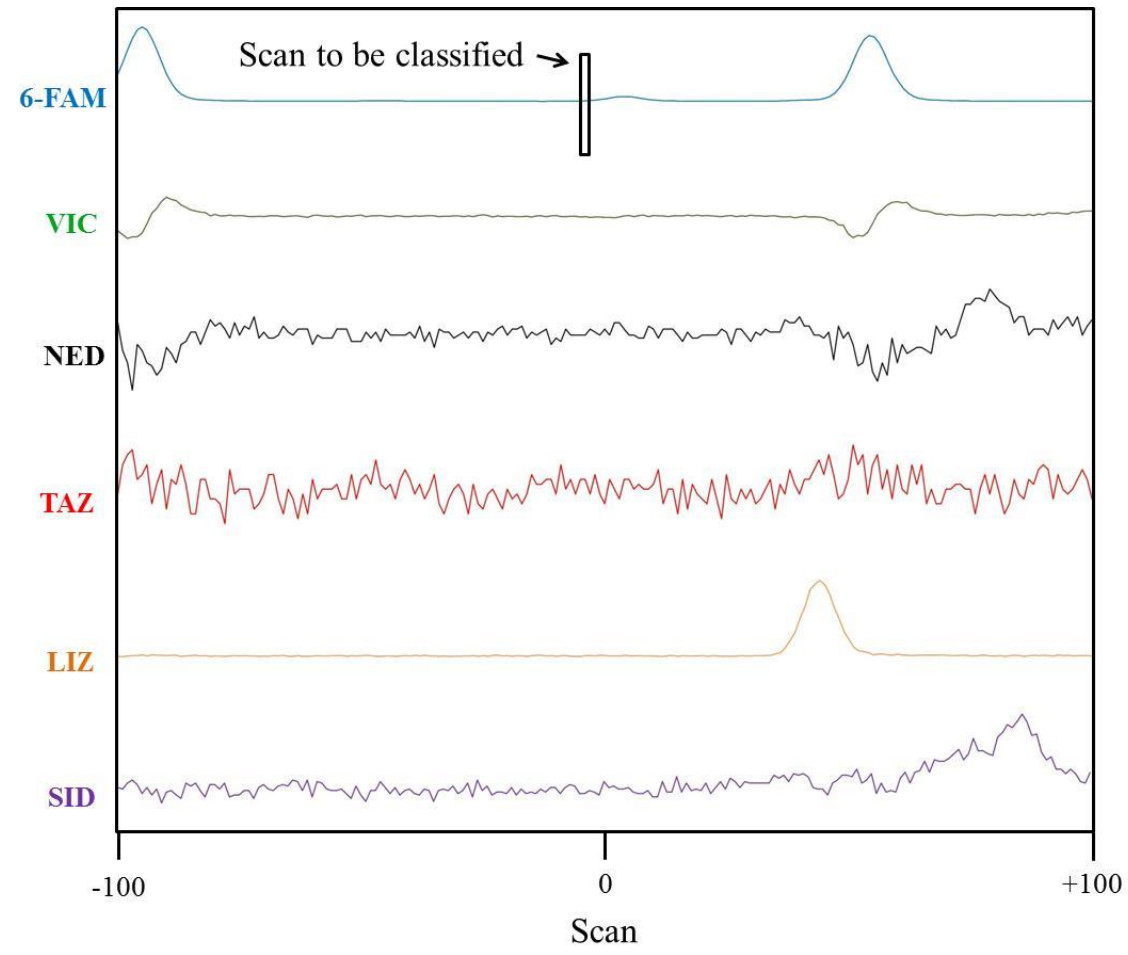
[6] Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011;2:37-63.

[7] Powers D. Evaluation Evaluation a Monte carlo study. European Conference on Artifical Intelligence, ArXiv: 150400854 [csAl]. 2008.

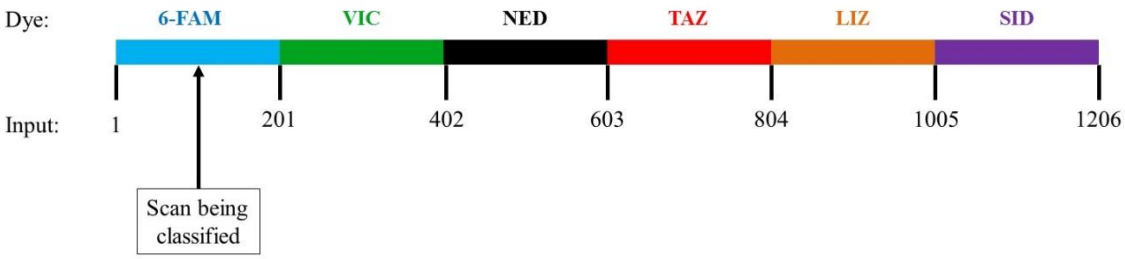
[8] Sutskever I. Training Recurrent Neural Networks. Toronto: University of Toronto; 2013.

[9] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9:1735-80.

[10] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv: 14114389v4. 2016.

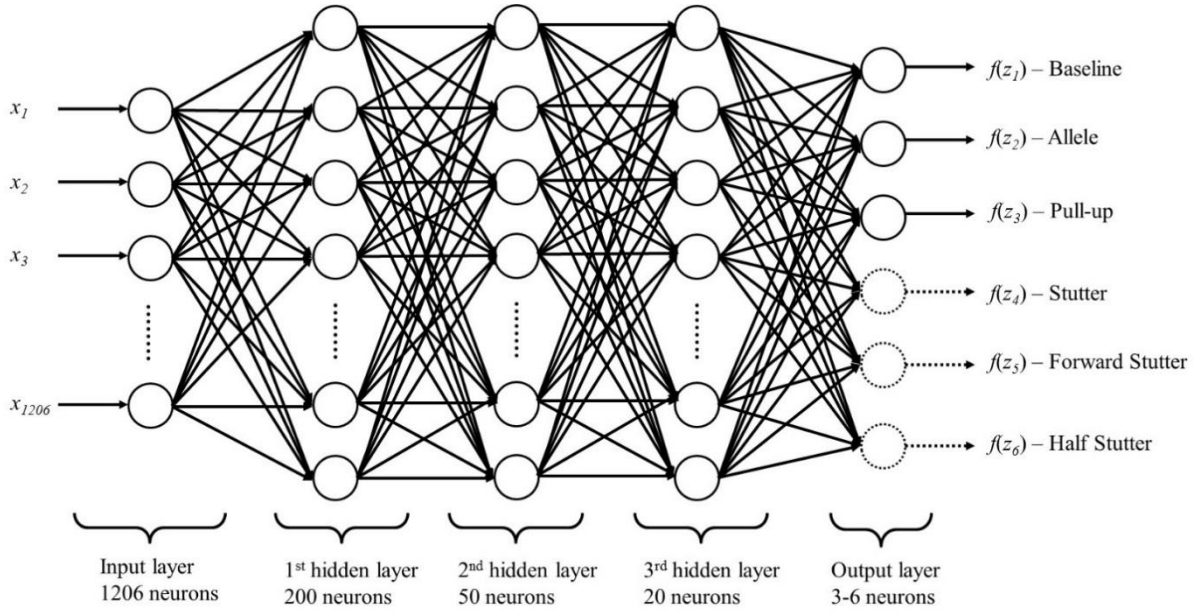


<InlineImage1>



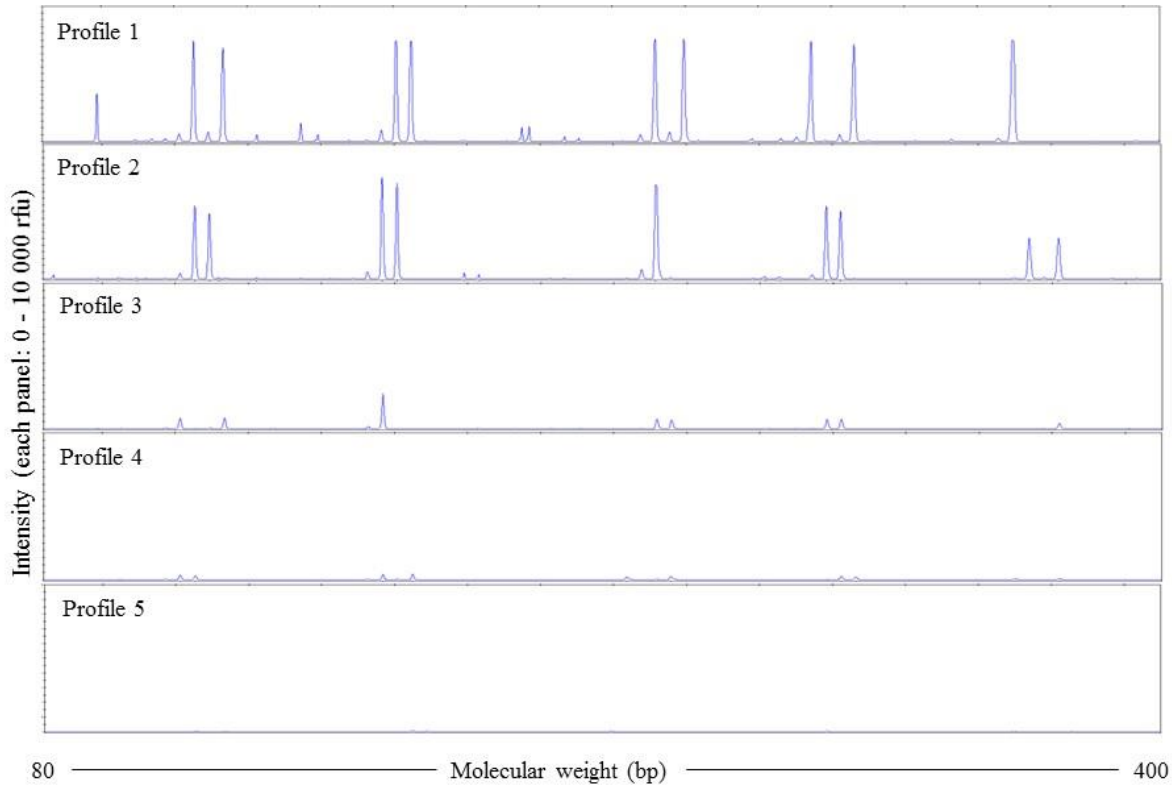
<InlineImage2>

Figure 1: Data used as input to classify central scan point in the 6-FAM dye lane in EPG context (upper) and in the context of an input for an ANN (lower)

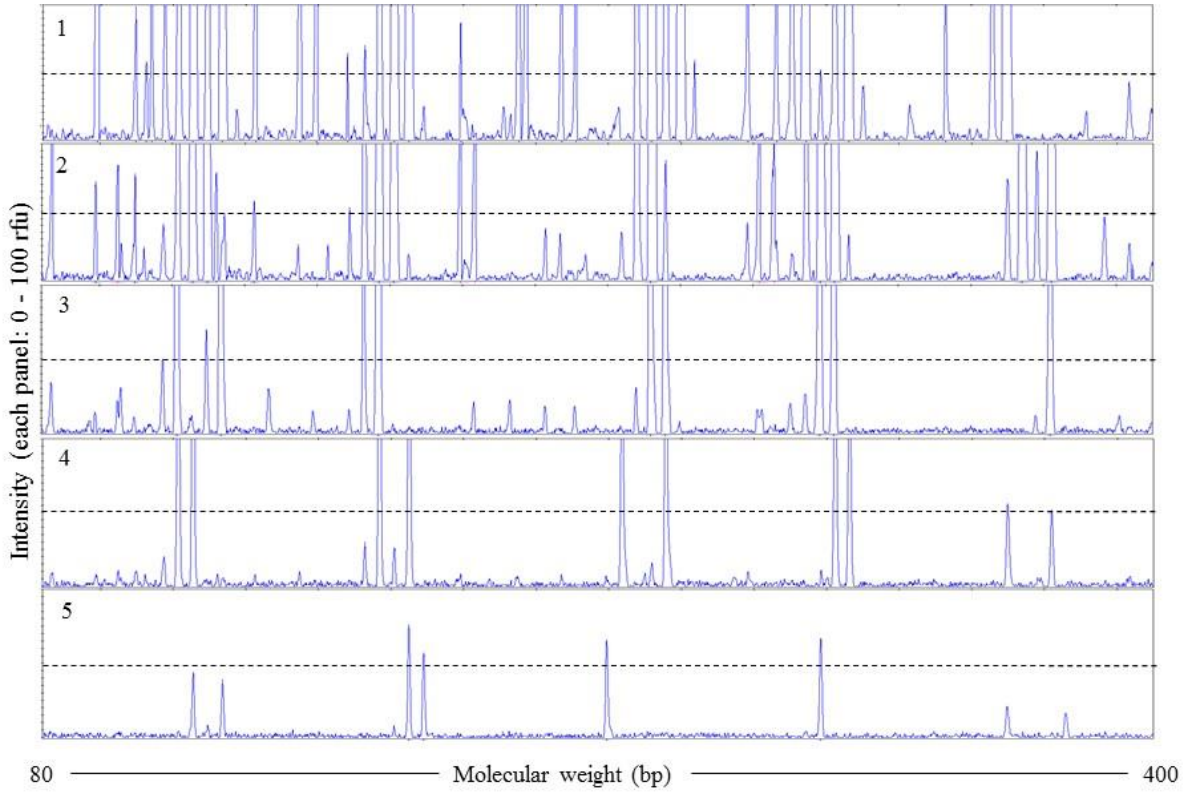


<InlineImage3>

Figure 2: ANN structure used within this study. Dashed outputs are present in only some of the ANNs

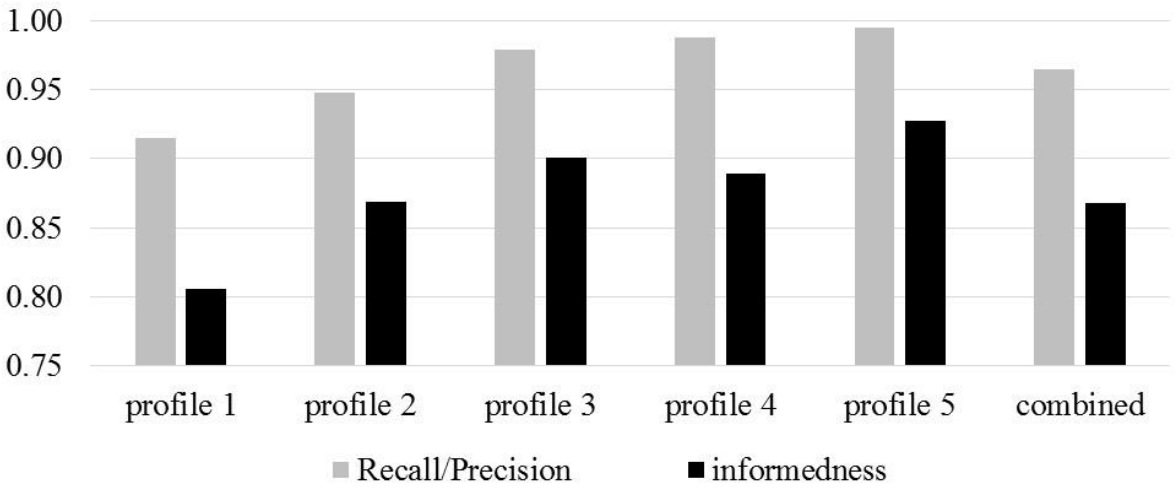


<InlineImage4>



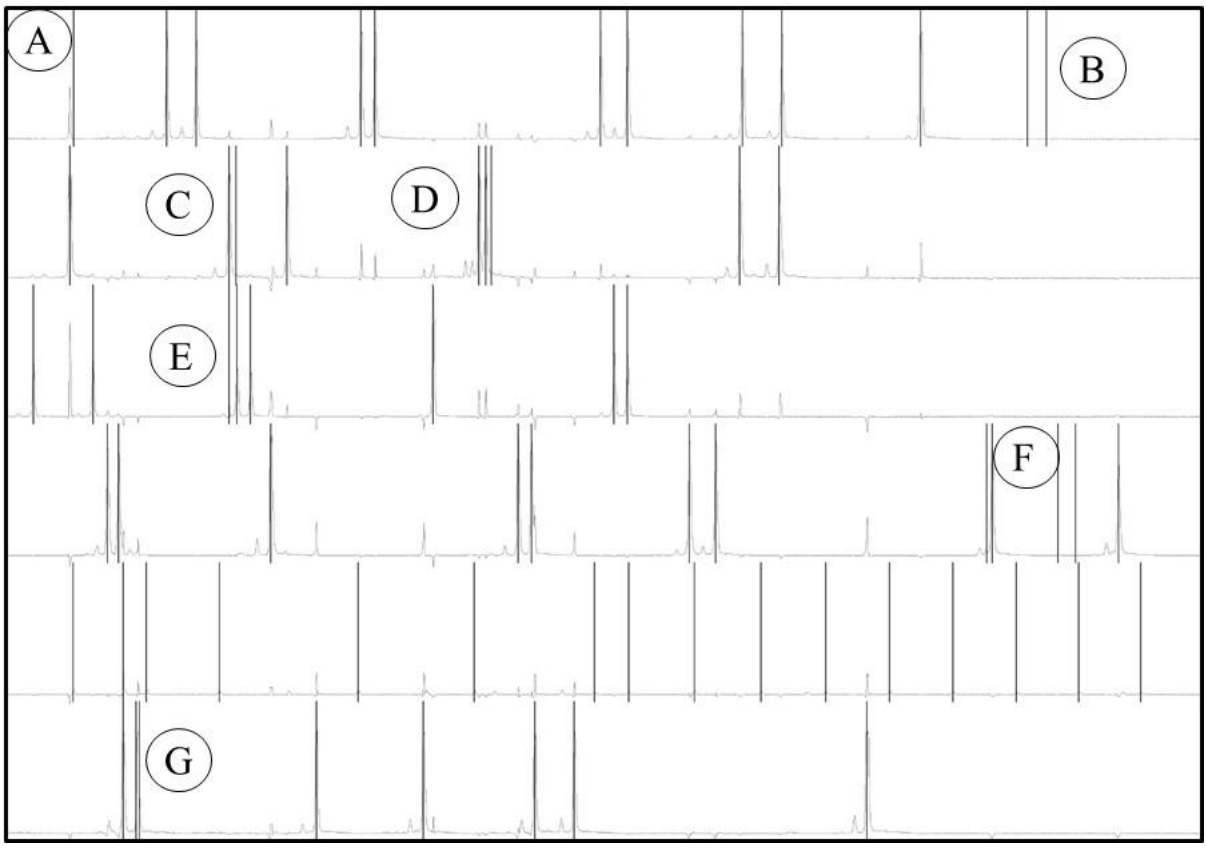
<InlineImage5>

Figure 3: 6-FAM dye of the five profiles chosen for comparison ranging from strong (top) to weak (bottom) in intensity. Because of the range of scales the profiles are shown twice, once at the full rfu scale (0 to 10 000rfu shown in the upper panel) and one at a reduced scale (0 to 100rfu in the lower panel). The analytical threshold of 50rfu is shown on the lower panel as a dashed line.

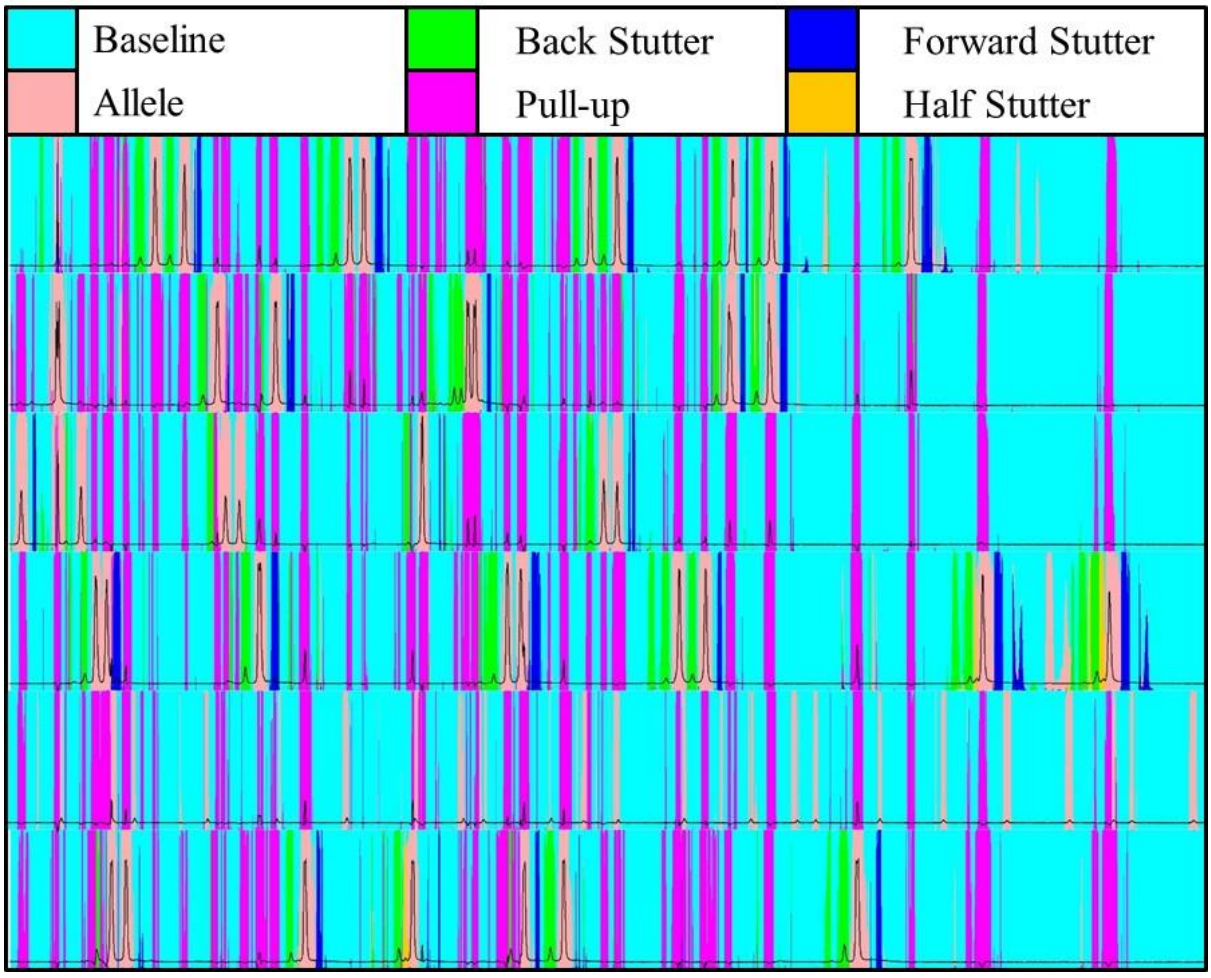


<InlineImage6>

Figure 4: Recall/precision (grey) and informedness (black) for each profile and the five profiles combined

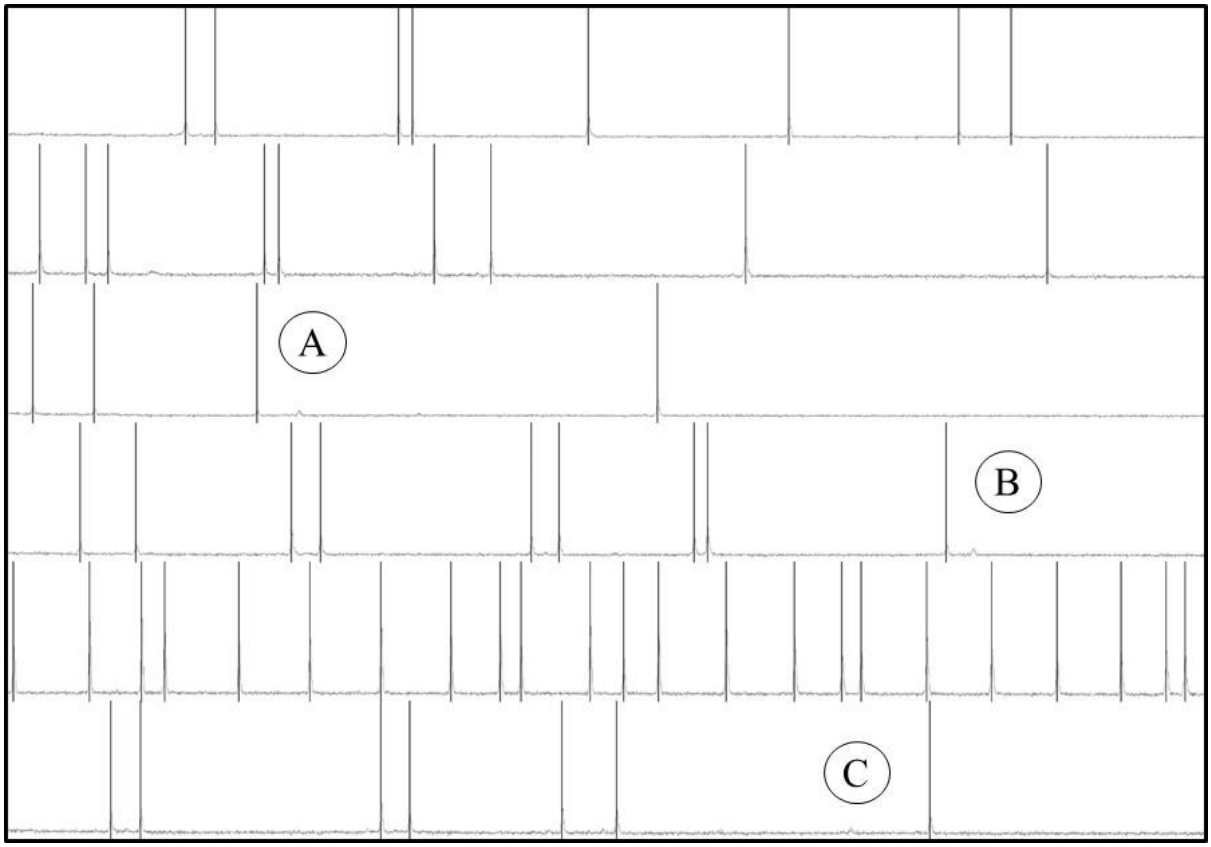


<InlineImage7>

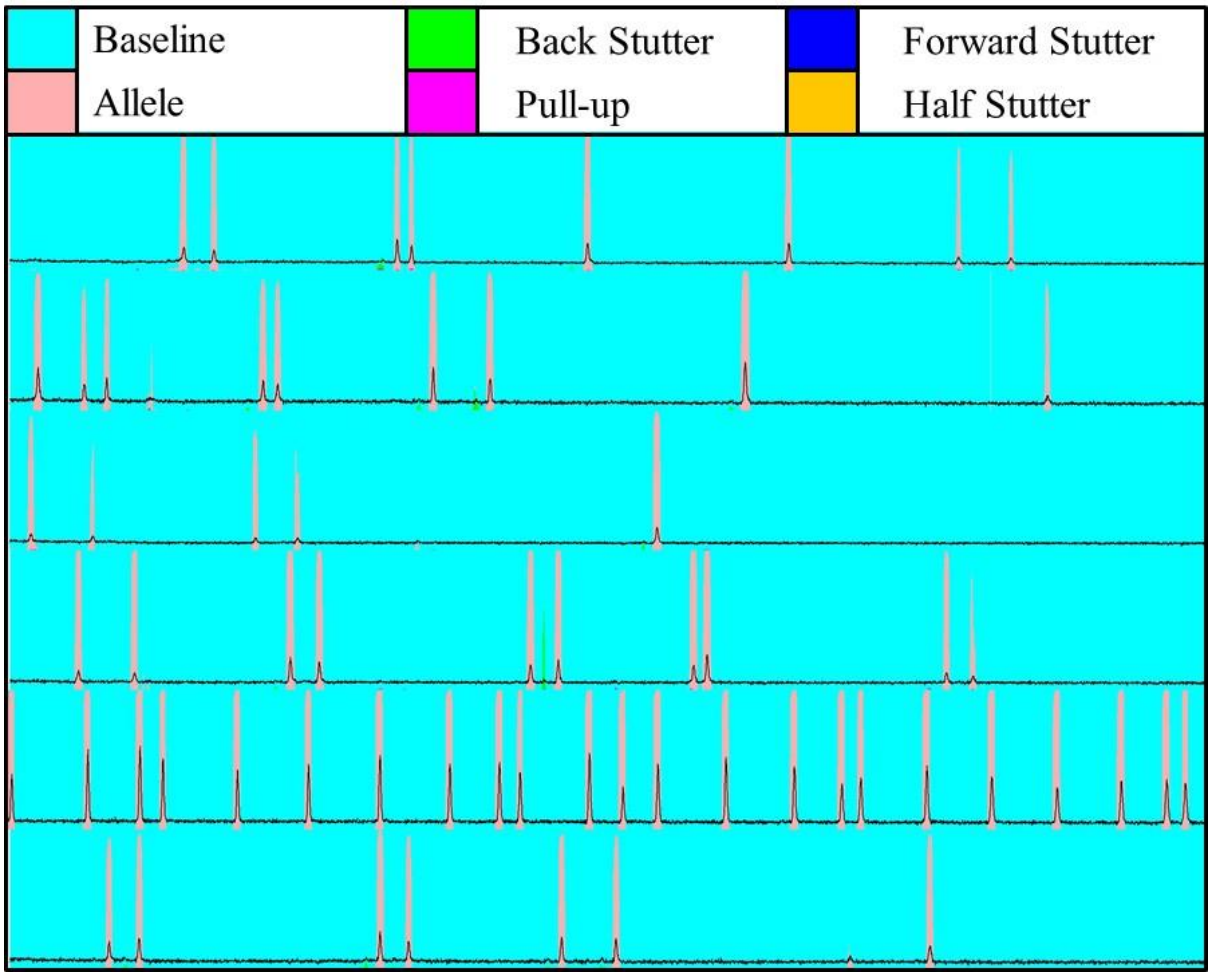


<InlineImage8>

Figure 5: Results of peak detection and classification (black lines) in upper pane and probabilistic results of ANN assignments in lower pane for profile 1 in test set. Intensity scale is from 0 to 10 000 rfu. Points of interest A to G have been identified that are discussed in the text.

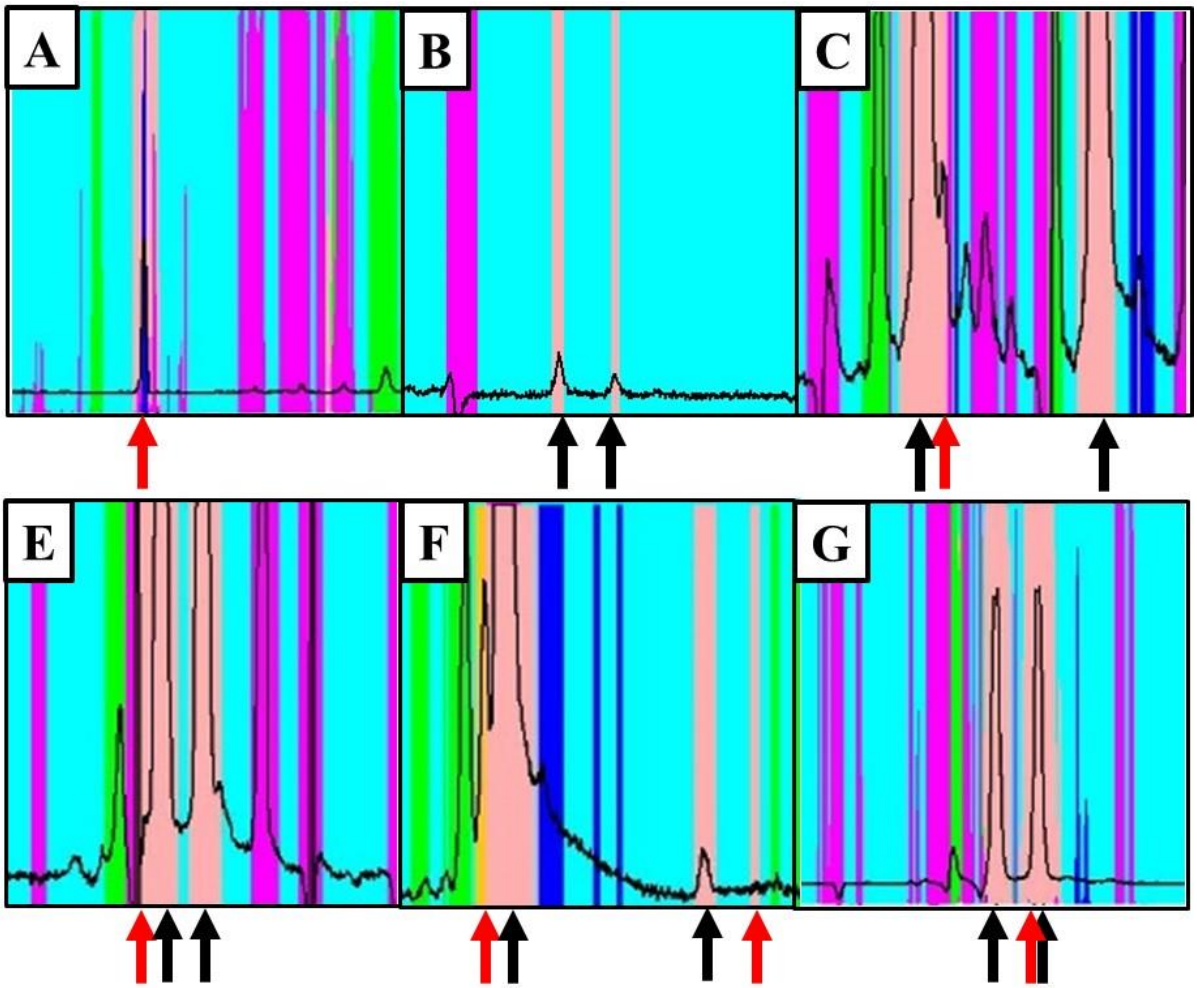


<InlineImage9>



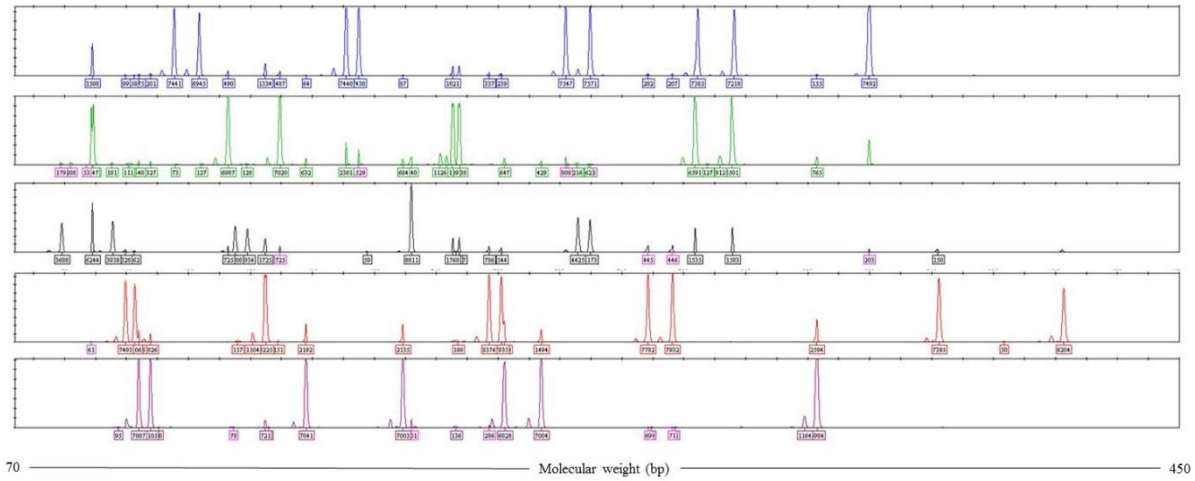
<InlineImage10>

Figure 6: Results of peak detection and classification (black lines) in upper pane and probabilistic results of ANN assignments in lower pane for profile 5 in test set. Intensity scale is from 0 to 500 rfu. Points of interest A to C have been identified that are discussed in the text.



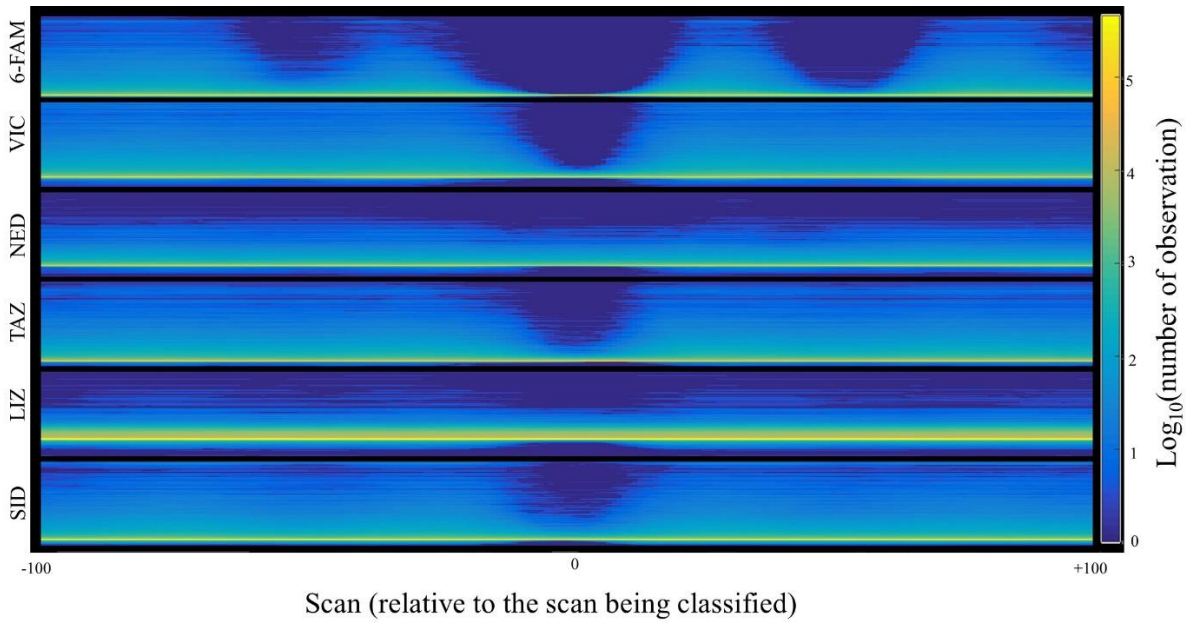
<InlineImage11>

Figure 7: Scan points identified as allele peak centres that were accepted (black arrow) or rejected (red arrow) by human interpretation



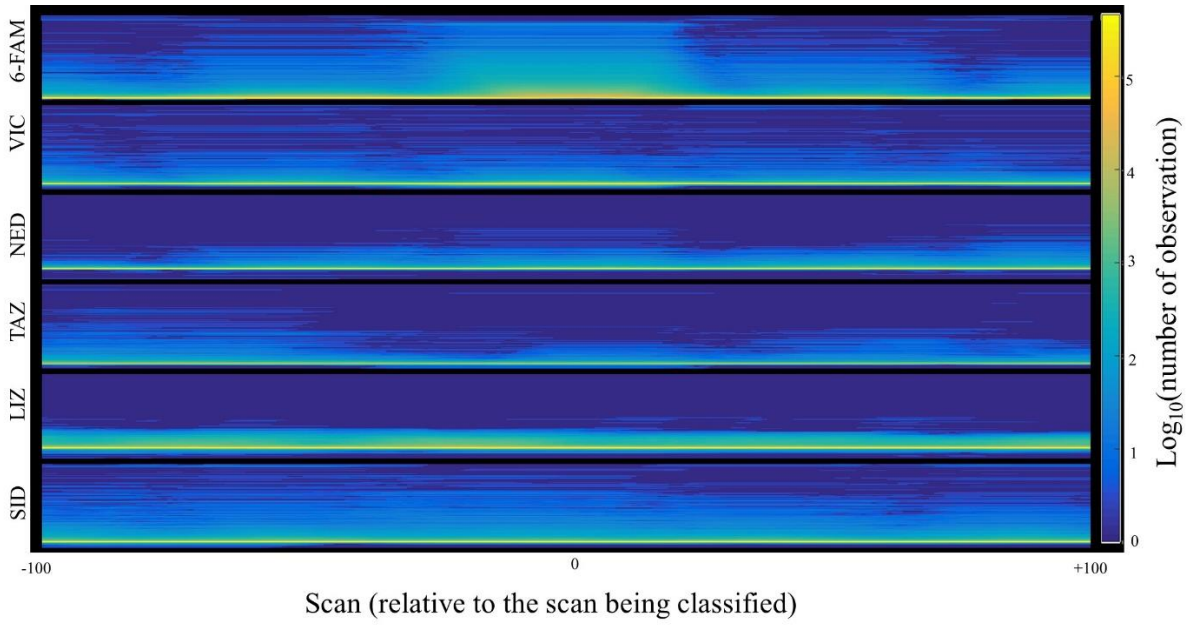
<InlineImage12>

Figure 8: Profile 1 in Fig 5 as read by Genemapper® ID-X using an analytical threshold of 50rfu and with stutter filters present. Value in boxes are peak heights (rfu). Pink boxes represent peaks identified by Genemapper® ID-X as artefacts.



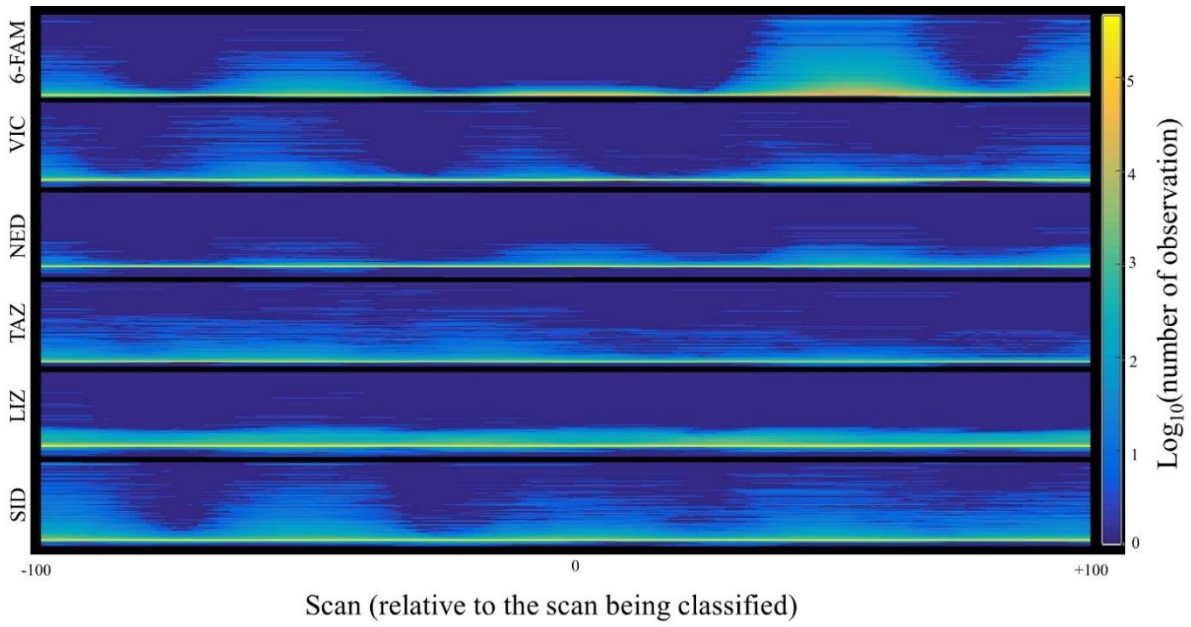
<InlineImage13>

Figure 9: Heat map showing the 1206 inputs (201 for each of the 6 dye lanes) for all training sets classified as baseline with greater than 0.95 probability. Each dye lane has a scale from 0 to 10 000 rfu.

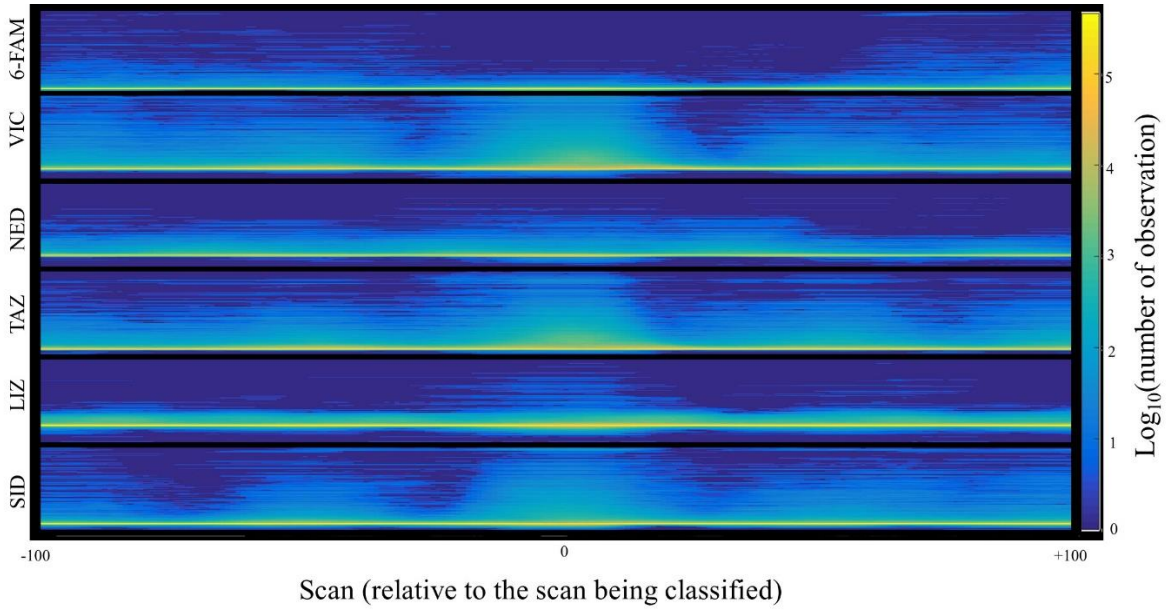


<InlineImage14>

Figure 10: Heat map showing inputs for all training sets classified as allele with greater than 0.95 probability. Each dye lane has a scale from 0 to 10 000 rfu.

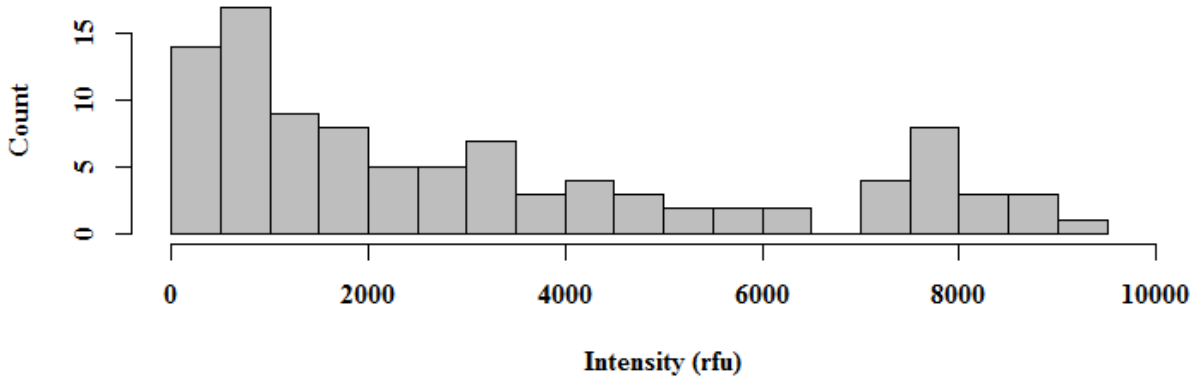


<InlineImage15>*Figure 11: Heat map showing inputs for all training sets classified as stutter with greater than 0.95 probability.*



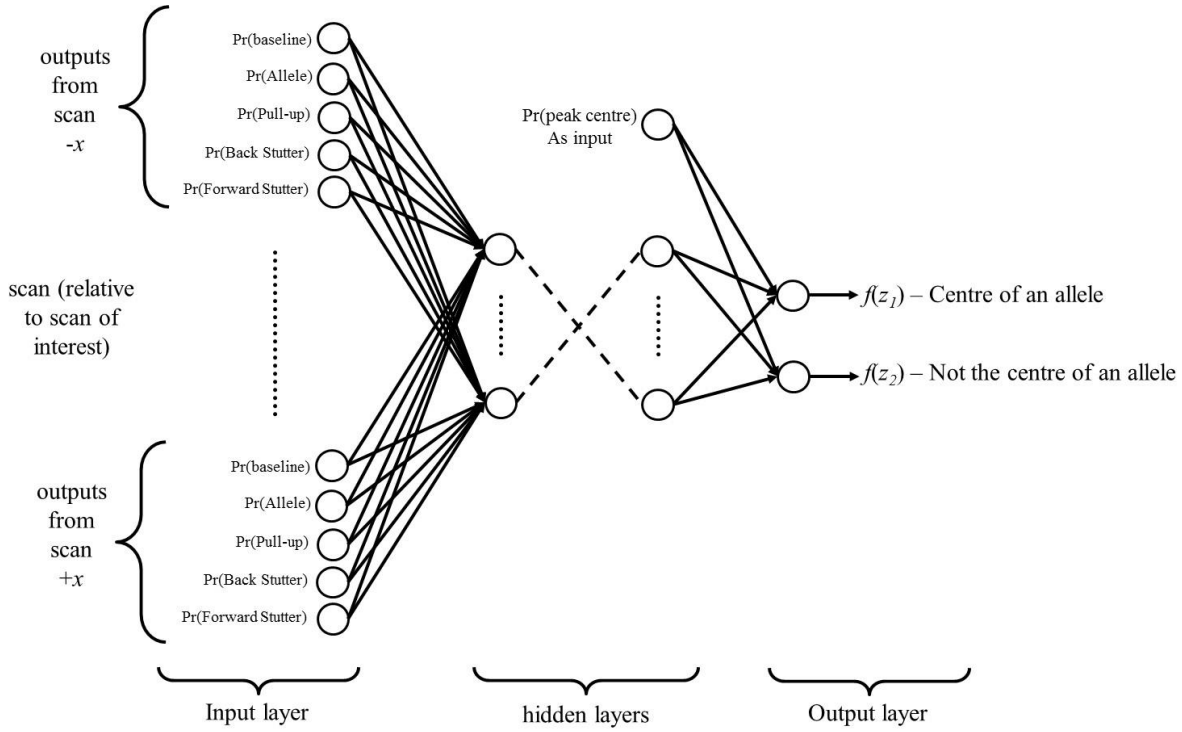
<InlineImage16>

Figure 12: Heat map showing inputs for all training sets classified as pull-up with greater than 0.95 probability.



<InlineImage17>

Figure 13: Spread of profile intensities used in this study.



<InlineImage18>

Figure 14: Example of a post-processing ANN that could be used to calculate the probability of allelic centres using output information from ANNs as seen in Fig 2 and peak centre probabilities as inputs

ANN Name (DYE)	**Applicable Loci**	**Outputs**	**Training data**
6-FAM_ANN (6-FAM)	D3S1358, vWA, D16, CSF1PO, TPOX	Baseline Allele Back Stutter	All of 6-FAM

		Pull-up Forward stutter	
VIC_ANN (VIC)	D8, D21, D18, DYS391	Baseline Allele Back Stutter Pull-up Forward stutter	All of VIC
GENDER_ANN (VIC)	Y-indel, Amelogenin	Baseline Allele Pull-up	Y-indel and Amelogenin
NED_ANN (NED)	D2S441, D19, TH01, FGA	Baseline Allele Back Stutter Pull-up Forward stutter	All of NED
SE33_ANN (TAZ)	SE33	Baseline Allele Back Stutter Pull-up Forward stutter Half Stutter	All of TAZ except D22
D22_ANN (TAZ)	D22	Baseline Allele Back Stutter Pull-up Forward stutter	D22
TAZ_ANN (TAZ)	D5, D13, D7	Baseline Allele Back Stutter Pull-up Forward stutter	D5, D13 and D7
LIZ_ANN (LIZ)	Size standard	Baseline Allele Pull-up	All of LIZ

D1_ANN (SID)	D1	Baseline Allele Back Stutter Pull-up Forward stutter Half Stutter	All of SID
SID_ANN (SID)	D10, D12, D2S1338	Baseline Allele Back Stutter Pull-up Forward stutter	All of SID except D1

Table 1: ANNs trained to classify entire GloablFiler™ EPG

	Baseline	Allele	Stutter	Pull-up	Forward Stutter	Half Stutter	Correctly Classified (%)
Baseline	182585	676	456	2285	372	75	97.9%
Allele	721	12157	45	185	33	5	92.5%
Stutter	569	18	4067	158	3	20	84.1%
Pull-up	1463	58	107	11928	40	7	87.7%
Forward Stutter	304	0	3	90	1018	0	71.9%
Half Stutter	15	9	1	0	0	90	78.3%
Totals	185657	12918	4679	14646	1466	197	96.5%

Table 2: Confusion matrix from trained ANNs applied to the five test profiles. Columns represents ground truth responses and rows are predicted classifications.

	Baseline	Allele	Stutter	Pull-up	Forward Stutter	Half Stutter	Total
Recall	0.983	0.941	0.869	0.814	0.694	0.457	0.965
Precision	0.979	0.925	0.841	0.877	0.719	0.783	0.965
F score	0.981	0.933	0.855	0.845	0.707	0.577	0.790
Markedness	0.887	0.921	0.838	0.864	0.717	0.782	0.885
Informedness	0.870	0.936	0.866	0.806	0.693	0.457	0.868

Table 3: diagnostics for performance of ANNs shown in Figure 2.

	ANN system		Genemapper® ID-X	
Profile	Peaks flagged as allelic	Peaks accepted as allelic	Peaks flagged as allelic	Peaks accepted as allelic
1	49	42	109	40

2	44	44	72	44
3	42	41	42	42
4	40	40	40	40
5	37	37	26	26

Table 4: Results of the ANN system and Genemapper® ID-X identifying potentially allelic peaks for the five test profiles