Clustering Single-Cell Electropherograms by Genotype Through Unsupervised Machine Learning

by

Leah O'Donnell

A thesis submitted in fulfillment of the requirements
for the MSc. degree in Applied Mathematics

at the

Hamilton Institute
Maynooth University
Maynooth, Co. Kildare, Ireland

May 2021

Under the supervision of
Prof. Ken Duffy

Contents

Abstract

Cells can be linked to the person who produced them by examining the information contained within their DNA. The challenge that a forensic analyst faces is to question whether a collection of cells obtained from a crime scene supports the hypothesis that a person of interest was present. The primary challenge is that cell samples collected at crime scenes typically contain material from an unknown number of genetic sources in an unknown mixture ratio. The standard genetic measurement protocol used in crime labs produces a single, combined signal for the entire collection of cells. If there are a small number of contributors, cells are in good condition, and the mixture ratio is not overly imbalanced, armed with this measurement, informative inference is possible for a trier of fact. If, however, the sample is complex, containing more than three genetic sources, or if the mixture ratio is highly imbalanced, or if genetic information within cells is degraded, the ability to confidently extract meaning from the measured signal is impaired. In high profile work published in the late 1990s it was demonstrated that genotype information could be extracted from individual cells. When used in a forensics context, single-cell methods offer a potential solution to the complex mixture problem by providing genetic information per-cell rather than solely for the whole collection. Advances in those measurement methods mean that single cell technologies may soon be practicable in crime labs. Significant challenges on the interpretation of the signals that result, however, remain. Instead of having a single high dimensional signal to assess, the trier of fact now has one for each cell. In the present thesis we take one step towards enabling the resolution of the complex mixture problem by proposing and assessing two methodologies that would facilitate the downstream analysis of genetic signal from a collection of single cells. Our goal is to query whether it is possible to use unsupervised machine learning to accurately and efficiently gather single cell signals into groups by genotype. If possible, it would greatly reduce the computational complexity of the evaluation of evidence and improve its accuracy. The results in this thesis suggest that this approach is viable and advances the potential of this societally important technology.

Acknowledgements

First and foremost I would like to thank my supervisor, Prof. Ken Duffy for his constant attention, encouragement and expert tutelage that has made this thesis what it is. He consistently provided his time regardless of how much work was on his own plate, somehow maintained a high level of patience when re-explaining something for the hundredth time, and frequently cultivated a refreshed excitement for the work through his own enthusiasm. I will be forever grateful for his supervision and in constant admiration of his persistent work effort. I would like to thank Prof. Catherine Grgicak who offered such positive encouragement in our monthly meetings, her warm voice always soothing my nervous presentations. I would also like to thank her student, Nidhi Sheth who helped explain the chemistry of our work and the rest of the team for all their contributions. This work was partially supported by NIJ2018-DU-BX-K0185 and NIJ2014-DN-BX-K026 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice.

I would like to thank my family, during this global pandemic my parents, Sean and Sue-Anne kindly welcomed me home and supplied me with copious amounts of tea, emotional support, sound boarding and proof reading. They spoiled their grateful baby and tolerated my many moods. My sister Rebecca, who although living in Germany, was always at the end of the phone to listen to me rant about the “stress” of completing this work while living back home with loving and caring parents (as any spoiled child would). My cousin Ciara, who regularly remembers to drop in and tell me I’m great.

Hamilton Institute as a whole deserves a recognition, an office space with the widest and most welcoming open door policy. Fellow students, post Docs, Doctors, Professors, and administrators all willing to offer help and advice along with emotional support and friendship. Thank you to Rosemary Hunt and Kate Moriarty for all the essential behind the scenes administration work that made everything run so smoothly. I would like to especially thank Bruna and Estevão for their constant help with writing code, no matter the problem (ashamedly even setting a working directory at one point) both were always patient with me and helped however they could. Eleni and HoChan who kept me company for “fresh air” breaks whenever I was feeling stress settle in. Maeve and Hannah who made sure to pop by my door and take me out for a cuppa, and Dáire who shared his (surprisingly tasty) vegan biscuits with me, offering his calming presence and advice whenever I should call.

My friends from outside work life, Darielle and Siobhan who live near me here in Tipp and during this pandemic I have been blessed to have such wonderful humans within my 5K radius. Tosin, Bassie, and Clair three purely amazing souls whom I was privileged to meet during my undergraduate degree here at MU, they made sure to ring, text or video call on the regular with a listening ear, advice, encouragement or just a good laugh. So many others have helped keep me sane through out this thesis, taking me on exciting adventures around our beautiful Éire, games nights, socially distanced walks, and endless phone calls. To all I have not mentioned by name, I thank you greatly.

Lastly I would like to thank myself, for certain there were many moments where I was tempted to pack it in or doubted my ability to contribute. I am proud of myself for completing, what was for me at times, an arduous journey.

1 An Introduction to Forensic DNA Analysis

To understand the importance of single-cell research within the field of forensics, we must first become accustomed with the current practices in forensic DNA analysis. This includes familiarising ourselves with the creation of a DNA profile: the type of DNA measured, the loci examined, the process in which it is measured and the complications that may arise. We will come to understand that when identifying a person by their DNA, we must focus our attention to short tandem repeats and that the allelic variation within a population is not unique, thus, to strengthen the distinction between individuals we now study a greater number of loci. Once a profile has been established, we will then gain insight into its interpretaion, identifying the true alleles from the stochastic artefacts that arise as a result of the measurement process.

Once we have a thorough understanding of DNA profiling, we must then consider the use of such a technique in the context of forensics DNA analysis. We will focus on the interpretation of a profile produced from crime scene DNA samples, in particular complex and low-template stains. The likelihood ratio is a powerful tool for evaluating the weight of crime scene evidence and so a brief understanding of its calculation is required. Although the interpretation of complex mixtures has advanced from traditional methods, often referred to as binary methods, many modern probabilistic genotyping software still employ various binary techniques and so we will review a selection of these.

1.1 Human Identification

1.1.1 Human DNA

Cells are the fundamental building blocks of humans, within the nucleus of each human cell is 46 chromosomes with one half inherited from the biological mother and one half from the biological father to give us 23 pairs [9]. These chromosomes are made up of tightly packed Deoxyribonucleic Acid (DNA) which holds the “instruction manual” for our cells.

Human DNA is stored in a coded fashion using a sequence of four bases- the nucleotides Thymine (T), Adenine (A), Cytosine (C) and Guanine (G). Ts can only pair with As, Cs only with Gs (and vice versa). This pairing is referred to as base pairing. Due to this base pairing, the unit of length of a strand of DNA is described as a number of base pairs (bp) [9].

The particular sequence of these bases determines the information available for building and maintaining an organism, they can be thought of as the words written in our "instruction manual". A gene is a DNA sequence that contains the instructions to make a protein, it encodes proteins. This coding DNA is not unique to an individual and accounts for about 2% of our DNA, giving way to seen traits such as eye colour or hair type [35].

1.1.2 Short Tandem Repeats

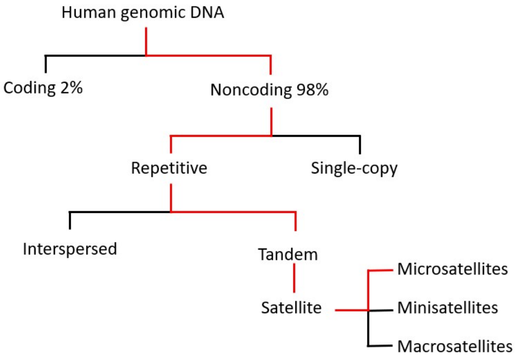


Figure 1.1: Composition of the human genome in terms of DNA classes. (Image adapted from [35].)

The remaining 98% is our non-coding DNA. Non-coding DNA, which was once dismissed as "useless junk DNA", is now broken up into sub-types with growing research into their uses [12]. Non-coding satellite DNA is used as a genetic marker for identifying one human from another. Satellite DNA consists of arrays of tandemly repeating non-coding DNA, or Short Tandem Repeats (STRs), that differs sufficiently in their base composition.

An STR occurs when a short segment of a DNA sequence, 2-6 base pairs, gets repeated back-to-back along a portion of a chromosome [23]. Fig. 1.2 gives an example of 3 different DNA sequences all containing an STR of four base pairs long, CTAA. The number of times this repeat is seen back-to-back is the allele/allelic variant [9]. In this way an allele can be thought of as a count; the allelic value represents the number of times an STR appears back-to-back in a DNA sequence at a given point or loci along a chromosome. For

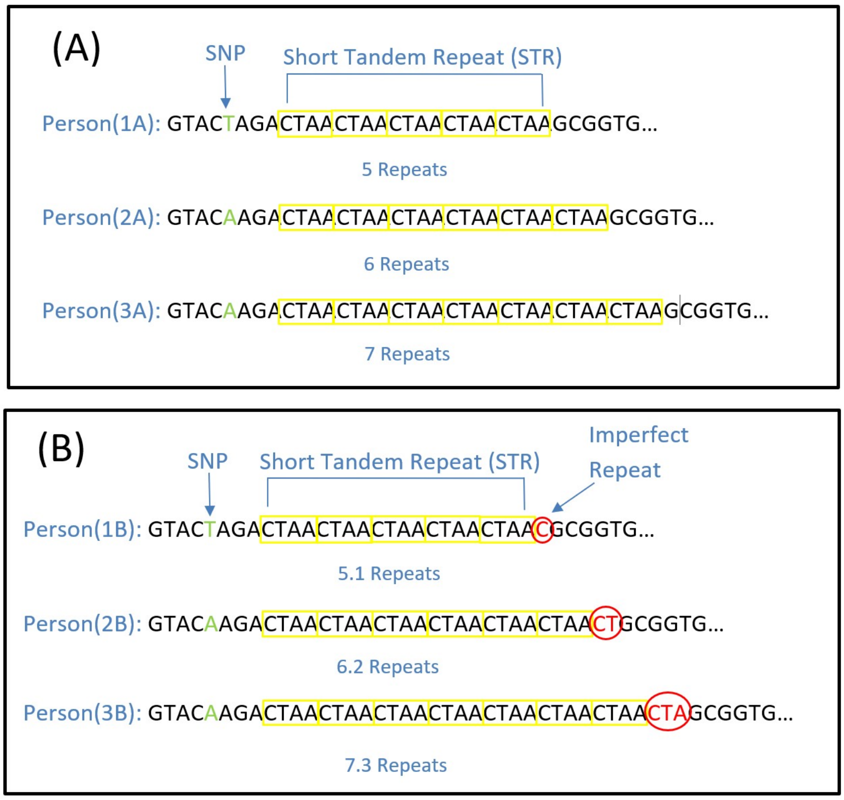

(A)
SNP
Short Tandem Repeat (STR)
Person(1A): GTACTAGACTAACTAACTAACTAACTAAGCGGTG...
5 Repeats
Person(2A): GTACAAGACTAACTAACTAACTAACTAACTAAGCGGTG...
6 Repeats
Person(3A): GTACAAGACTAACTAACTAACTAACTAACTAACTAAGCGGTG...
7 Repeats
(B)
SNP
Short Tandem Repeat (STR)
Imperfect Repeat
Person(1B): GTACTAGACTAACTAACTAACTAACTAACGCGGTG...
5.1 Repeats
Person(2B): GTACAAGACTAACTAACTAACTAACTAACTAACTGCGGTG...
6.2 Repeats
Person(3B): GTACAAGACTAACTAACTAACTAACTAACTAACTAACTAGCGGTG...
7.3 Repeats

Figure 1.2: Three different DNA sequences with a detectable STR. Each STR is 4 base pairs long (CTAA), and yet they differ in number of repeats. Figure (A) shows integer repeats only, while figure (B) shows imperfect repeats.

example, taking the sequences from Fig. 1.2(A) Person(1A) has an allele = 5, Person(2A) has an allele = 6 and Person(3A) has an allele = 7. Note that imperfect repeats can also occur and are recorded as modulo alleles, expressed as non-integer valued alleles seen in Fig. 1.2(B). Put more simply, Person(1B) has allele = 5.1 or 5 repeats + 1 bp, Person(2B) allele = 6.2 or 6 repeats + 2 bps and Person(3B) has allele = 7.3 or 7 repeats + 3 bps.

Humans are *diploid organisms* as we have two chromosomes, hence two alleles at each genetic locus with one allele inherited from the biological mother and one from the biological father. As a result, if an STR allelic measurement is made, two unordered numbers, one per chromosome, is recorded. Genotypes are described as *homozygous* if there are two identical alleles at a particular locus and as *heterozygous* if the two alleles differ [28].

1.1.3 Variation of STRs in the Human Population

Allelic variants found in humans are not entirely distinct, that is to say, we not only share genetic variants with our biological families but also with unrelated individuals. For example, both you and an unrelated stranger may have an allele = 8 for the locus CSF1PO. When looking at the larger picture, an array of loci across different chromosomes, the distinction between individuals can become clearer. Thanks to research such as The 1000 Genome Project, a wider view of allelic variation across 26 human populations has been studied, and in 2015 they found the typical difference between the genomes of two unrelated individuals was estimated at 20 million base pairs [1]. This high variability means that although individuals may share alleles, once examining a sufficient number of loci, the likelihood that two unrelated individuals have the same genome becomes negligible.

This large scale study by Auton et. al. [1] also highlights that alleles occur at different frequencies in different human populations, more specifically those of differing geographical origin. For example, if we look at an allele frequency table for 1036 unrelated samples in the U.S. population [57], it is far more likely to see the allelic variant 8 present at the locus CSF1PO in African-Americans than in Caucasians with frequencies 0.0556 and 0.0055, respectively.

Allele frequency tables have been established to estimate the likelihood of shared allelic variants among a population. These are compiled by taking a subset of a population, determining their genotypes and subsequently their allele frequency, where allele frequency is found on a locus basis as follows:

$$\text{Allele Frequency} = \frac{\text{Total Num. of Allele X Present}}{\text{Total Num. of All Alleles Present}}. \tag{1.1}$$

For clarity, if we take a subset of 20 people from a population, this means our total number of alleles present at any locus is 40 (2 alleles per individual). We would like to note that these 40 alleles are not expected to be distinct. If we find the allelic variant 8 appears at the locus CSF1PO 6 times within this sample, then the frequency of the allelic variant 8 at locus CSF1PO is said to be 6/40 = 0.15 for this subset of the population. This value is then assumed to be representative of the full population as an estimate for the expected frequency of said allele at said locus.

1.2 DNA Profiling and Measuring STRs

DNA profiling is the process of correlating an individual's identity to characteristics of their DNA. It can be used to aid forensic investigation and track down blood relatives. Customarily, this can be done as follows: i) take a sample of human cells such as a cheek swab or blood sample; ii) extract DNA from the cell by lysis; iii) quantise the DNA; iv) amplify the DNA via Polymerase Chain Reaction (PCR) to generate hundreds of millions of copies of that particular DNA segment [98]; and finally v) separate the DNA through gel electrophoresis, a traditional separation technique that is based on the movement of ions in an electric field - shorter STRs will move faster and further through the gel than longer ones. This method is often referred to as *bulk-processing* due to the fact it is applied to a collection of cells. An electropherogram (EPG) is a record produced as a result of electrophoresis. They are used to distinguish allelic variants by their lengths. An electropherogram is a plot of measured fluorescence intensity versus potential allelic variant.

1.2.1 Forensic DNA Profiling and the Core STR Loci

When entering the genome of an individual into a national or international database, generally used to link criminals, a common set of STR markers or core loci are required. Multiple loci are examined when testing human identity with the intention of reducing the possibility of a random match between unrelated individuals. As satellite DNA tolerates a high degree of variability, the core loci have been chosen as markers found between genes, allowing equivalent genetic information to be compared and shared [22]. As of January 1st 2017, the FBI's Combined DNA Index System (CODIS) has expanded their core loci to now include 21 STR markers: CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, D22S1045, and Amelogenin [55]. When creating DNA profiles we will examine 20 of these 21 core loci, excluding Amelogenin the sex determining marker. Fig. 1.3 is an example of an EPG, the 20 CODIS STR loci for a single source sample chosen at random from the PROVEDIt database [4].

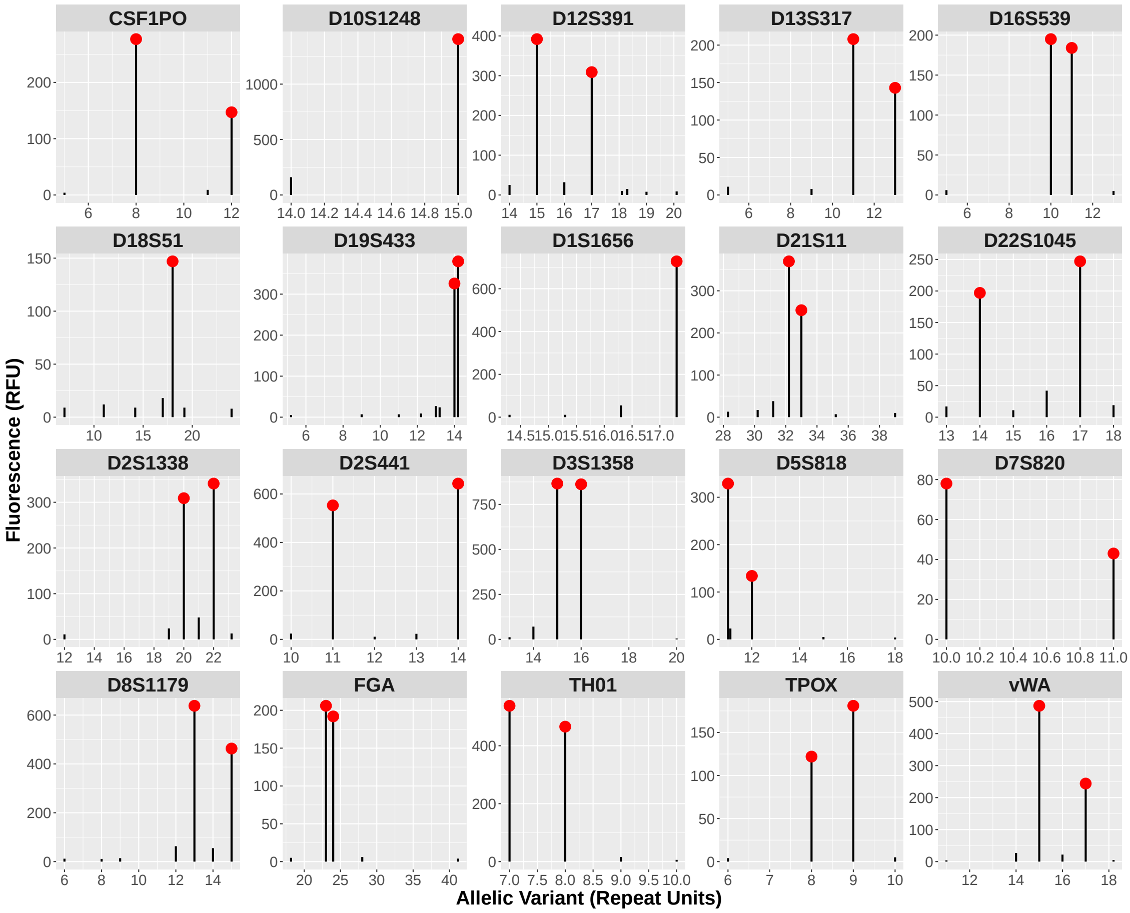


Figure 1.3: The 20 CODIS STR loci of Genotype 15, generated from the PROVEDIt database [4] (Sample.File: A02_RD14-0003-15d2U60-0.25GF-Q4.5_01.5sec.hid). Vertical bars indicated detected fluorescence at potential allelic variants and red dots indicate the true alleles. 0.25ng of DNA was extracted using the PicoPure kit and amplified via the GlobalFiler amplification kit (29 cycles).

1.2.2 Interpreting EPG Artefacts: Stochastic Effects

Fig. 1.4 shows six different EPGs where each one is an example of a common artefact. Fig. 1.4(A) is an EPG of the heterozygous locus D7S820, generated from a single source sample from the PROVEDIt database [4]. We have prior knowledge of the ground truth for this genotype and we know that for this locus the allelic variant $= (10, 11)$ as indicated by red dots. Fig. 1.4(A) is an ideal post-processing resultant EPG, as the only detected fluorescence is that of the true allelic variants, however this is not regularly the case.

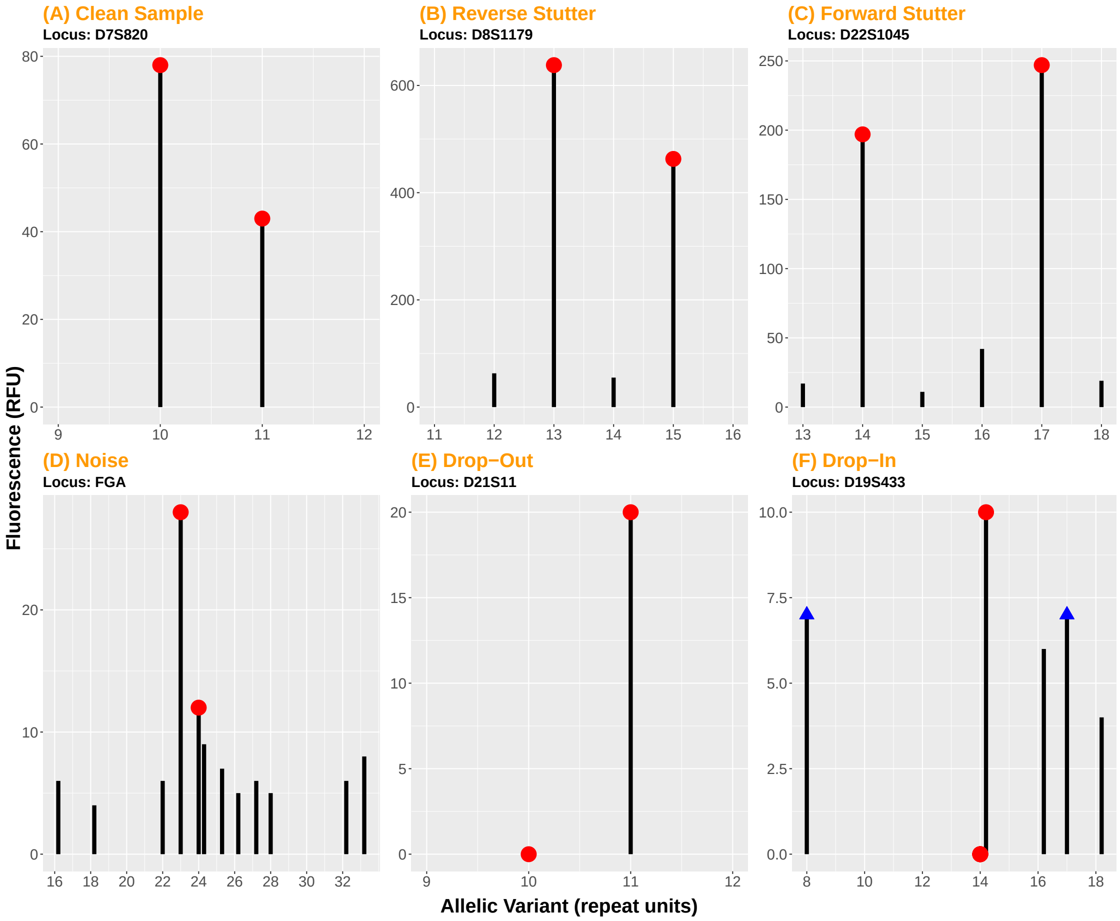


Figure 1.4: Examples of the various artefacts found in EPGs. All plots have been generated using three samples of Genotype 15 from the PROVEDIt database [4]. In all plots, (A through to F) true allelic variants for each loci are indicated by red dots while recorded fluorescence at an allelic variant are indicated by the black vertical bars. (A) is an example of the ideal case; a perfectly clean sample where the only fluorescence detected is that of the true alleles. (B) and (C) document the by-products of the PCR amplification process - reverse stutter and forward stutter, respectively. (D) is an example of a noisy EPG. (E) shows the common case of allelic drop-out while (F) shows that of drop-in, where the blue triangles indicate the presences of drop-in allelic variants.

Stutter

Reverse stutter (or back stutter) peaks are common and well documented by-products of the PCR amplification of STR regions typically occurring as strands which are one repeat unit shorter than the parental allele, (true allele − 1) [116]. Reverse stutter can be seen in Fig. 1.4(B) at 12 and 14 repeat units. Gill et. al. [45] experimentally observed that reverse stutter signal tends to be less than 15% of the detected parent allele peak. As stutter is such a prevalent feature of PCR, The SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories 2.6 recommends the inclusion of a laboratory established stutter-threshold [107]. In contrast, *forward stutter*, one repeat unit longer, and *double-back stutter*, two repeat units shorter, have been reported to occur less frequently than reverse stutter, with the latter occurring least frequently. [42, 121]

Forward stutter occurs when the polymerase enzyme slips forward and skips a template strand, (true allele + 1) [56, 117]. Forward stutter can be seen in Fig. 1.4(C) at 18 repeat units. Gibb et. al. [42] observed that parent peak height has an influence on the relative magnitude of forward stutter. Their results confirm that an all-loci encompassing guideline for the forward stutter in DNA mixture analyses does not reflect the nature of this artefact and proposed that it would be more appropriate for a general guideline per locus to be applied.

Double back stutter, (true allele − 2), can occur when the polymerase enzyme slips backward on an already back stuttered strand. Signal at 15 repeat units is observed in Fig. 1.4(C) and this can be classified as forward stutter from the 14 parent allele or double back stutter from the 16 reverse stutter allele. Weusten et. al. [121] found that the peak area of double stutters is generally below 1% of the allelic peak. The results of Gibb et. al. [42] shows that double-back stutter has the potential to interfere with interpretation of minor profiles within mixtures but to a lesser degree.

Noise

Noise peaks can be seen in Fig. 1.4(D). These are considered to be noise where, in the context of the experimental sciences, noise can refer to any unknown or random fluctuation of data that hinders inference of signal. Noise is an independent artefact that although does not occur at every potential allele position, it can occur at any potential allele position. It has also been experimentally observed that noise occurs at approximately 15% of potential allelic positions, and when it does occur it is best described by a log-normal distribution as found by Monich et. al. [82].

Amplification Failure

Total Failure of Amplification (TFA) may occur during the PCR amplification process of both single-cells and larger DNA samples, although it is seen much less frequently for the latter. It is noted by Piyamongkol et. al. [89] that generally TFA affects 5-10% of single-cells subjected to such an amplification process. They consider several likely causes for amplification failure: the isolated cell might have been lost during transfer to the PCR tube, the cell could have been anucleate, the cell may have been in the process of degeneration, or the DNA might not have been accessible to the PCR reagents due to a failure of cell lysis, however, they also acknowledged that it is immensely difficult to determine the reason for total amplification failure in all cases.

Allele Drop-out

Allele Drop-Out (ADO) is the failure to amplify one of the two alleles in a heterozygous cell due to one allele having insufficient amplification, a problem that is unique to PCR of minute quantities of DNA such as single-cells. ADO can affect either of the alleles of a given locus and occurs at random giving a heterozygous cell the appearance of homozygosity [89]. The frequency at which we see ADO is in part related to the choice of lysis used pre-amplification and Kim et. al. [68] found that DNA extraction using an alkaline lysis buffer results in less allele drop-out when compared to other methods of DNA extraction tested. An extended PCR protocol has also proven to reduce ADO frequencies as seen in [120]. An example of ADO can be seen in Fig. 1.4(E) where the ground truth for the locus D7S820 is $(10, 11)$ but only the allelic variant 11 has been amplified.

Allele Drop-in

Allele drop-in is the presence of an allele not associated with the sample and remains unexplained by the sample contributor. Drop-in is typically restricted to 1 or 2 alleles per profile. If multiple alleles are observed at more than two loci then these are considered as alleles from an extra contributor and analysis can proceed as a mixture of two or more contributors [46]. In Fig. 1.4(F) we observe potential drop-in indicated by the blue triangles. These “extra” alleles are only found at two loci and so we can attributed this to drop-in and not consider a potential DNA mixture. The authors of [46] and [8] suggest probabilistic approaches and likelihood ratio principles that can be applied to DNA profiles containing drop-in, both recommending that these probabilities be assessed and their relevance to a DNA profile be considered, with Gill et. al. [46] highlighting awareness that their approaches described are based on simplified assumptions.

1.3 Crime Scene DNA Evidence

When a crime has been committed it is often the case that DNA from an assailant has been left behind which, when profiled appropriately, which can aid in criminal investigation, be it linking a suspect to the crime or eliminating them. It has become a standard forensic technique to gather such DNA evidence for the wide spectrum of crime types, from volume crime such as burglary or autocrime to major crime. Properly collected and processed DNA evidence can be compared with known samples from specific suspect(s) or databases such as CODIS in the hopes of linking suspect(s) to a crime through the utilisation of a Likelihood Ratio (LR). The likelihood ratio test is an accepted approach to quantitatively assessing match information between a suspect and a piece of evidence found at the crime scene. It is the ratio of probabilities of observing the evidence under two distinct hypotheses,

$$\mathrm{LR} = \frac{\mathrm{P}(E|H_p)}{\mathrm{P}(E|H_d)}, \tag{1.2}$$

where H_p represents the prosecution's hypothesis, H_d the hypothesis of the defense and E represents the the evidence. The evidence in our case, will be a crime scene DNA profile or simply Crime Scene Profile (CSP) .

When establishing E, the DNA profile of a crime-stain, one is often faced with complex, Low-Template DNA (LTDNA) samples that exhibit allele sharing. The three main factors determining complexity are: i) the Number of Contributors (NoC) in the sample, the more contributors the more complex a sample is; ii) the mixture ratio of the sample, how much DNA is present from each contributor; and lastly iii) the quality of the DNA, how degraded the sample is [90].

Traditionally analysts have constructed the LR using a binary approach to interpret the TrueNoC of these complex LTDNA profiles, often applying a stochastic and/or analytical threshold along with other biological parameters such as Maximum Allele Count (MAC), heterozygote balance and stutter ratios [27]. As sensitivity and instrumentation continue to improve, there has been a movement from the binary methods of interpretation to probabilistic methods of interpretation, employing the development of Probabilistic Genotyping Software (PGS) such as STRmix, LRmix and CEESIt [91, 16, 104]. PSG's can determine the LR by factoring in the probabilities of allelic drop-in, drop-out, and some of the aforementioned biological parameters using mathematics to approximate what happens in a real mixture, but majoritively, they still require an assumption regarding the TrueNoC.

1.3.1 The Likelihood Ratio

For crime scene DNA samples, the weight of this evidence can be supplied by calculating the likelihood ratio. The numerator is the probability of observing the evidence given the prosecution's hypothesis and the denominator is the probability of observing the evidence given the defense's hypothesis. Likelihood ratios often make use of peak height information, assume a number of contributors to the mixture and always make use of the Person Of Interest's (POI's) genotype (the suspect) [19]. If the LR >1, the prosecution's hypotheses is supported by the evidence, conversely if the LR <1, the defense's hypotheses is supported by the evidence. For a LR = 1 neither hypotheses is supported by the evidence. LRs are complex, can be hard to phrase, and the development of the hypotheses may be difficult, particularly that of H_d. In the context of crime scene DNA evidence, equation (1.2) can be interpreted as such:

$$\frac{\mathrm{P}(E|H_p)}{\mathrm{P}(E|H_d)} = \frac{\mathrm{P}(EPG|POI, G_2, ...G_{n_p})}{\mathrm{P}(EPG|G_1, ..., G_{n_d})}, \tag{1.3}$$

where n_p and n_d are the prosecution and defenses assumed number of profiles contributing to crime scene EPG, and G is an unknown genotype.

The numerator, $\mathrm{P}(EPG|POI, G_2, ...G_{n_p})$ is the probability of seeing the EPG, given the genotype of a POI and n_p unknown genotypes. The denominator $\mathrm{P}(EPG|G_1, ..., G_{n_d})$ is the probability of seeing such an EPG given n_d unknown genotypes, typically assumed to be independent. G_1 is drawn with probabilities determined by allele frequency tables with the same ethnicity of the assumed POI, as is the case under the defenses hypothesis. We note that in practice n_p and n_d can be chosen to maximise the prosecution and defenses probabilities respectively and therefore are not necessarily equal [46, 104].

To understand the challenges faced when determining a LR, we must first consider the simplest case, $n_p = n_d = 1$ and build up our understanding from there. The LR will vary depending on the probabilistic model used for its calculation but at minimum any probabilistic model that aims to model an EPG requires true allele signal, stutter signal and noise. Many models have been developed to include a greater proportion of the information seen in an EPG incorporating forward stutter, drop-in, or drop-out [93, 67, 15, 109]. Let G be the set of genotypes such that $G = \{G_1, ..., G_n\}$. For clarity in notation, G_i is the full genotype on an individual, G_i^l is the genotype of an individual at a given locus. Moreover $G_i^l = \{A_{i_1}^l, A_{i_2}^l\}$ where $A_{i_1}^l$ and $A_{i_2}^l$ are the alleles at said locus. For any hypothesis H,

$$\mathrm{P}(EPG|H) = \sum_g \mathrm{P}(EPG|G = g)\mathrm{P}(G = g|H). \tag{1.4}$$

We will begin with a general description of $\mathrm{P}(EPG\ |G_i)$, the probability of seeing an EPG given a genotype G_i, employing a toy model for an EPG and then elaborating how the prosecution and defense compute their respective $\mathrm{P}(EPG\ |G)$ and $\mathrm{P}(G|H)$

Simple Model for an EPG

If we treat each locus as being probabilistically independent, to describe a model for a full EPG it is sufficient to restrict our attention to describing a model for a single EPG^l (single locus l). We will construct a toy model that only incorporates the key features: true allele signal, noise and reverse stutter.

True allele signal is the amount of fluorescence in RFU that comes as a result of detecting a true allelic variant during the process of electrophoresis. There exists insufficient characterisation of the true distribution of the signal detected at a true allelic variant (Z), with Peters et. al. [88] declaring it cannot be easily described by a simple distribution class. The gamma distribution has been adopted by Puch-Solis et. al. [95] as it gives a simple yet flexible class of unimodal and asymmetric densities that best fit their simulated data. However, it has been suggested by Bright et. al. [15] that one could determine the distribution directly when one has sufficient data to do so as it can vary with the quantity of DNA present.

Different loci have a different range of potential alleles and we will define the set of potential alleles for a given locus, l, as B^l. We will establish a toy model EPG^l_j that describes the signal recorded at allele $j \in \mathrm{B}^l$ for locus l as follows:

$$EPG^l_j = N_j + Z_1\mathbb{1}_{A^l_1=j} + Z_2\mathbb{1}_{A^l_2=j} + \lambda Z_1\mathbb{1}_{A^l_1=j-1} + \lambda Z_2\mathbb{1}_{A^l_2=j-1}, \tag{1.5}$$

where N_j is the noise recorded in RFU at allele j. In this model the occurrence of noise can be determined by a binomial distribution. Z_1 and Z_2 are the fluorescences recorded in RFU at true allelic variants. Here we assume Z follows a log-normal distribution as it appears to reasonably describe our data, seen in section 4.2. A^l_1 and A^l_2 are the true alleles for a given locus, λ is the stutter ratio and $\mathbb{1}$ is the indicator function.

This simple model can now be used to determine the probability of an EPG given a genotype, $\mathrm{P}(EPG^l\ |A^l_{i_1} = a^l_{i_1}, A^l_{i_1} = a^l_{i_2}) = \mathrm{P}(EPG^l\ |G^l_i)$. However, we are interested in the probability of the EPG given a genotype, $\mathrm{P}(EPG\ |G_i)$:

$$\mathrm{P}(EPG|G_i) = \prod_{l\in L} \mathrm{P}(EPG^l|G^l_i)\ , \tag{1.6}$$

where L is the set of all loci studied in a forensic DNA profile. L can be determined from CODIS or similar.

The Prosecution's Calculation

The prosecution is concerned with the probability of seeing the crime scene EPG given the genotype is that of the POI. Henceforth, the genotype of the POI shall be referred to as s. This yields:

$$\mathrm{P}(E|H_p) = \sum_g \mathrm{P}(EPG|G = s)\mathrm{P}(G = s|H_p). \tag{1.7}$$

If the genotype corresponds to a POI then A_1^l and A_2^l become fixed and there exists a genotype s such that $\mathrm{P}(G = s|H_p) = 1$ thus $\mathrm{P}(E|H_p)$ collapses to:

$$\mathrm{P}(E|H_p) = \mathrm{P}(EPG|G = s) = \prod_{l \in L} \mathrm{P}(EPG^l|A_1^l = s_1^l, A_2^l = s_2^l). \tag{1.8}$$

The Defence's Calculation

The defence is concerned with the probability that any other individual of the same gender and ethnicity as the POI could be responsible for the crime scene EPG. Under their hypothesis, there exists a distribution of G as determined by allele frequency tables which we call $\mathrm{P}(G = g|H_d)$. Applying this to equation (1.4) we get:

$$\mathrm{P}(E|H_d) = \sum_g \mathrm{P}(EPG\ |G = g)\mathrm{P}(G = g|H_d). \tag{1.9}$$

In principle, the defense's calculation must consider all possible combinations of allele frequencies for g, which is both the feature and the flaw. By design, the large combinatorial diversity is what makes for a strong candidate in identifying an individual but, conversely, choosing two alleles per loci with replacement yields a near infinite amount of possibilities for g;

$$\prod_{l \in L} \binom{|B^l| + 1}{2}. \tag{1.10}$$

To comprehend the sheer number of combinations, we can take a numerical example considering the cardinality of potential alleles per loci from our data and determine that there are 1.9×10^{64} possible combinations for g (see Appendix A). To try and put this number into perspective, there are more possible combinations than atoms on Earth (it is estimated that the Earth is comprised of 1.33×10^{50} atoms). In practice, this leads to a computationally infeasible sum and so instead of calculating equation (1.9) directly, $\mathrm{P}(E|H_d)$ is typically approximated using a Monte Carlo sampling algorithm [104, 16]. In its base form:

$$\sum_{g} \mathrm{P}(EPG|G=g)\mathrm{P}(G=g|H_d) \cong \frac{1}{M}\sum_{i=1}^{M} \mathrm{P}(EPG|G_i)\mathrm{P}(G_i|H_d). \tag{1.11}$$

We draw $G_1, ..., G_M$ independently from the distribution of $\mathrm{P}(G = g|H_d)$, to ensure they occur with the correct frequency. The strong law of large numbers [36] tell us as $M \to \infty$ the sample average will almost surely converge to the expected genotype.

Increasing the Number of Contributors

Satisfied with the LR calculation for $n_p = n_d = 1$, we now increase the complexity, considering multiple contributors in the sample. We will focus on $n_p = n_d = 2$ as it is easily generalised to higher n. The prosecutions hypothesis now becomes: there are two contributors, one of whom is the POI, the other is some unknown genotype ($H_p : G_1, G_2$ such that $G_1 =$ POI and $G_2 =$ Random Genotype). While the defenses hypothesis is now extended to the case of two independent unknown genotypes (H_d : G_1, G_2 are independent Random Genotypes). The LR can now be defined as:

$$\frac{\mathrm{P}(E|H_p)}{\mathrm{P}(E|H_d)} = \frac{\mathrm{P}(EPG|POI, G_2)}{\mathrm{P}(EPG|G_1, G_2)}. \tag{1.12}$$

Once the number of contributors has been assigned, the mixture ratio of the two genotypes must be accounted for in the calculation. When $n = 1$ it was reasonable to assume the DNA mass was present in the sample at every location however, for $n > 1$, we must also make a model decision about the proportion of the sample that comes from each genotype. Let Θ be a vector with components Θ_i, the mixture proportion of each contributor such that $0 < \Theta_i \leqslant 1$ for $i \in \{1, ..., n\}$ and $\sum_{i=1}^{n} \Theta_i = 1$. More precisely, for two contributors, $\Theta = \langle \Theta_1, \Theta_2 \rangle$ such that Θ_1 is the proportion of the sample contributed by person 1 and Θ_2 is the proportion of the sample contributed by person 2. The probability of the prosecution and defense hypotheses can be expressed as:

$$\frac{\mathrm{P}(E|H_p)}{\mathrm{P}(E|H_d)} = \frac{\mathrm{P}(EPG|G_1 = s, G_2 = g_2, \Theta = \theta)}{\mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta)}, \tag{1.13}$$

which, for either hypothesis H, can be expanded to:

$$\mathrm{P}(E|H) = \sum_{g_1}\sum_{g_2}\int_{\theta} \mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta)\mathrm{P}(G_1 = g_1, G_2 = g_2, \Theta = \theta|H). \tag{1.14}$$

The joint probability $\mathrm{P}(G_1 = g_1, G_2 = g_2, \Theta = \theta|H)$ simplifies to the product of probabilities due to independence of the random variables G_1, G_2, and Θ, yielding:

$$\sum_{g_1}\sum_{g_2}\int_{\theta} \mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta)\mathrm{P}(G_1 = g_1|H)\mathrm{P}(G_2 = g_2|H)\mathrm{P}(\Theta = \theta|H), \tag{1.15}$$

for the prosecution $G_1 = s$ and so $\mathrm{P}(G_1 = s|H) = 1$ as before. For the defense G_1, is determined by allele frequency tables. For both the prosecution and defense G_2 is determined by allele frequency tables and assumed to be independent of s and G_1. Again, to account for the vast number of combinations for G_1, and now G_2, a Monte Carlo approximation or similar can be made.

Focusing next on $\mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta)$ we can make one of two assumptions:

(1) $\Theta^l = \theta$ for all l

(2) Θ^l is drawn independently from distinct distributions.

Under (1) the mixture ratio is assumed to be equal across all loci, as assumed by [95, 29], yielding:

$$\mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta) = \prod_{l\in L} \mathrm{P}(EPG^l|G_1^l = g_1^l, G_2^l = g_2^l, \Theta^l = \theta), \tag{1.16}$$

Θ^l is fixed for each locus l. If there is a single mixture ratio across all loci one must first condition on θ, calculate the likelihood of the EPG given random g_1s and g_2s, then change θ, again calculate the likelihood of the EPG given random g_1s and g_2s and repeat until a sufficient number of θs have been calculated and then evaluate the product.

Alternatively, under (2) the mixture ratio can be assumed independent at each locus as was the assumption made in Model B by [106], yielding:

$$\mathrm{P}(EPG|G_1 = g_1, G_2 = g_2, \Theta = \theta) = \prod_{l\in L} \mathrm{P}(EPG^l|G_1^l = g_1^l, G_2^l = g_2^l, \Theta^l = \theta^l), \tag{1.17}$$

such an assumption allows one to independently compute the proportion at each locus and then evaluate the product directly. For computational simplicity many authors assume the mixture ratio is independent at each locus [87, 109].

We have already noted the number of genotypes is practically infinite but the combination of proportions is actually infinite and so we must approximate once again. Therefore, for multiple unknown contributors, with an unknown mixture ratio, if we take multiple selections of ratios and genotypes then, again by the strong law of large numbers, we may converge to the probabilities of the prosecution and the defense. It is clear to see that as the number of contributors increase the computational cost in determining the LR will become substantially greater.

1.3.2 Low-Template DNA

Low-Template DNA (LTDNA) refers to any situation where a small amount of DNA is present in a sample, such as a touch sample. Samples containing less than 100pg to 200pg of total DNA available for amplification fall into the range generally considered to be Low Copy Number (LCN) DNA by most practitioners [48]. Establishing a DNA profile from LTDNA may require additional PCR cycles in an attempt to compensate for the low starting template. LTDNA profiles can be characterised by frequent drop-out, low EPG peak heights, exaggerated peak imbalance and large stutter artefacts (due to the increased PCR cycles) [8], making it challenging to establish a clear and complete DNA profile for comparison with suspects.

1.3.3 Complex Mixtures

Currently there is no consensus on how to decide if a mixture (often referred to as an admixture) is “too” complex for interpretation, with different labs having different protocol on which samples they attempt to interpret [90]. Labs are increasingly developing new methods and tools to deal with these complex DNA mixture samples. These developments not only update the existing tool set but also introduce fundamental new approaches to forensic DNA analysis. To consider the challenges of interpreting complex DNA mixtures we will begin by taking a simple admixture, introduce varying levels of complexity and consider the challenges faced in identifying all genotypes present.

Number Of Contributors (NoC): Fig. 1.5 considers admixtures where contribution is in equal ratio and so to add complexity only the NoC has been increased. When interpreting these EPGs of the locus vWA, the black vertical bars indicate the fluorescence recorded at potential allelic variants while the pairwise coloured dots indicate the true allelic variant of each genotype present in the admixture. Interpretation of DNA mixtures with three or more contributors is challenging due to the inevitability of allele sharing. For example, the three distinct genotypes share the allelic variant 16 in Fig. 1.5(B) and so one could mistakenly determine this as a two source sample. Looking to (D), a five person admixture, one could determine this as a single source sample with true alleles $= (16, 18)$ repeat units accompanied by reverse stutter at 15 and 17 repeat units. The occurrence of such EPG artefacts along with common alleles means as the NoC increases one cannot by eye accurately determine the true NoC from a DNA admixture.

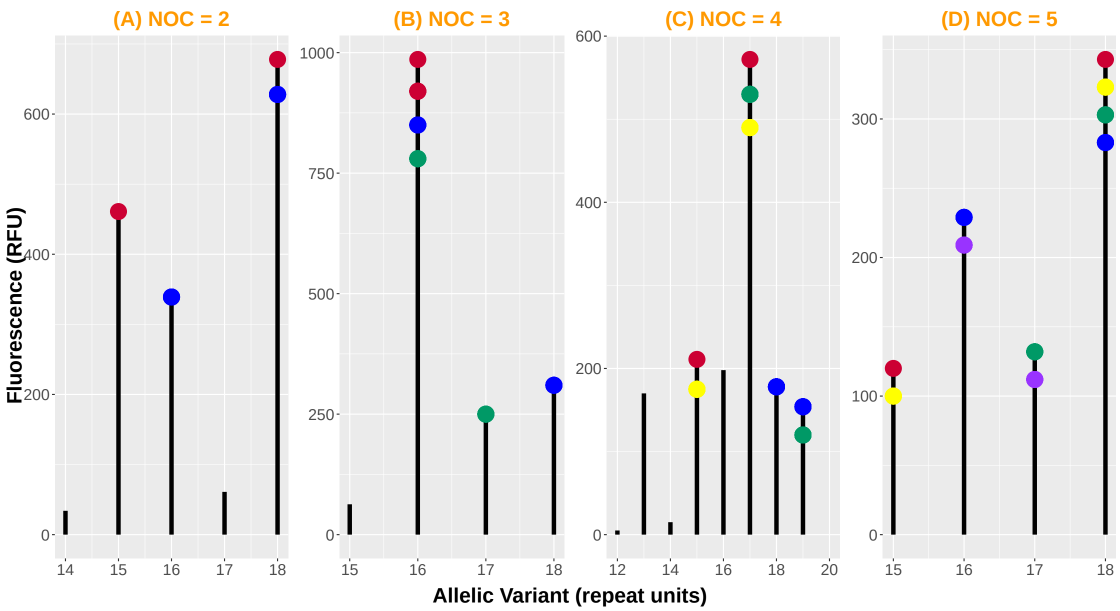


Figure 1.5: A study of complexity for an increase in the number of contributors at the locus vWA. These plots have been generated using true admixture samples from the PROVEDIt database [4]. For all admixtures selected, DNA contribution was in equal ratio. The various genotypes are indicated by dots, coloured pairwise and the vertical black bars represent recorded fluorescence's at potential allele variants. Selected samples are drawn from PROVEDIt_1 – 5-Person CSVs UnFiltered / PROVEDIt_1 – 5-Person CSVs UnFiltered_3500_F6C29cycles_hlfrxn / 2 – 5-Persons. Samples: (A): B01_RD14-0003-31_32-1;1-M1a-0.25GF-Q1.2_02.5sec.hid (B): B06_RD14-0003-46_47_48-1; 1; 1-M2a-0.375GF-Q0.4_02.5sec.hid (C): B01_RD14-0003-40_41_42_43-1; 1; 1; 1-M3a-0.5GF-Q0.9_02.5sec.hid (D): A05_RD14-0003-30_31_32_33_34-1; 1; 1; 1; 1-M3I22-0.315GF-Q1.3_01.5sec.hid

Mixture Ratio: Fig. 1.6 considers admixtures where the ratio of contribution has been varied, looking at cases where there may be a single major or single minor contributor. Interpretation of these EPGs is similar to above. One could determine that Fig. 1.6(A) is the result of a single source sample, with true alleles $(18, 19)$, stutter at 17 repeat units, and noise at 16 and 15 repeat units however, we know this not to be the case. When dealing with an admixture that has a minor contributor it can be the case that after quantisation, only the DNA of the major contributor remains. For this reason it is of great importance to produce more than one EPG in the hopes of capturing the minor contributors DNA, which may only be detectable in a handful of the EPGs. When the ratio of minor to major contributor's DNA is less than 1:10, Isaacson et. al. [61] finds that EPG measurements are not sensitive enough to detect the minor contributor DNA signatures. Considering Fig. 1.6(B) we see that in the presence of a major contributor, although detectable minor contributors could be disregarded as noise. Similarly, in the case of Fig. 1.6(C), multiple major contributors and a single minor, the minor contributor is completely indistinguishable from the major contributors. The unshared allele could be disregarded as noise while the allelic variant at 17 repeat units is shared with two major contributors.

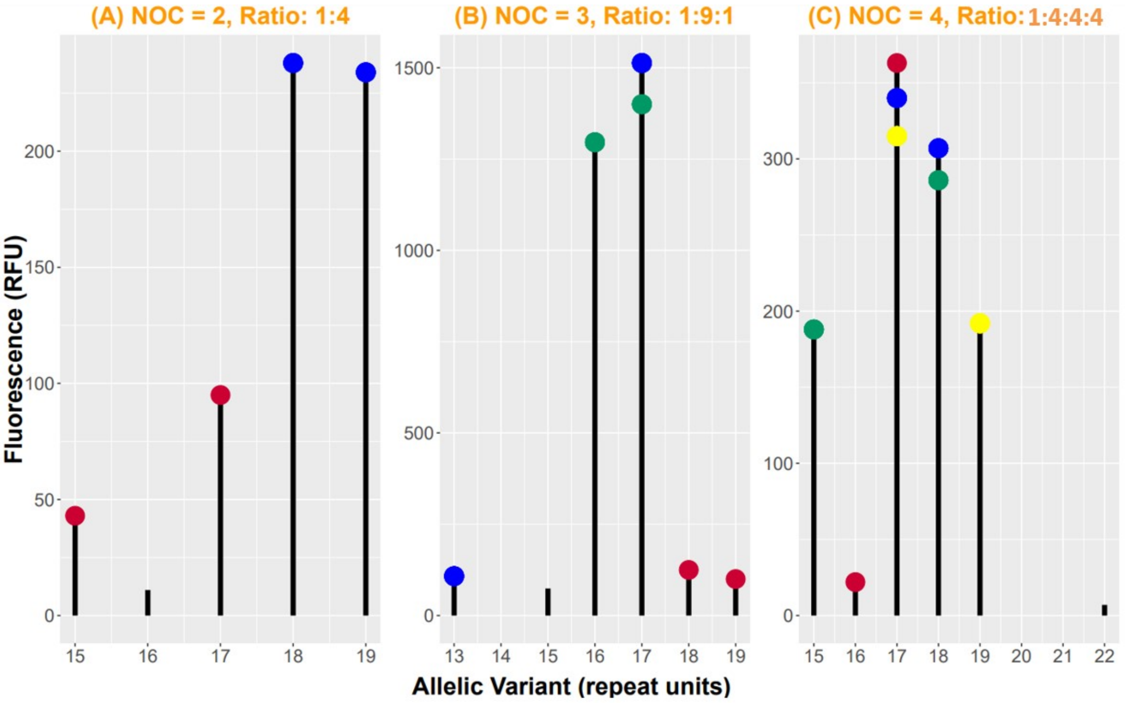


Figure 1.6: A study of complexity for various ratios of contribution at the locus vWA. (A) major-minor two source admixture in ratio 1 : 4, (B) two minor, one major three source admixture in ratio 1 : 9 : 1 and (C) one minor, three major four source admixture in ratio 1 : 4 : 4 : 4. These plots have been generated using true admixture samples from the PROVEDIt database [4]. The various genotypes are indicated by dots, coloured pairwise and the vertical black bars represent recorded fluorescence's at potential allele variants. Samples: (A): F01_RD14-0003-40_41-1;4-M1S30-0.625GF-Q3.2_06.5sec.hid (B): C03_RD14-0003-41_42_43-1; 9; 1-M2a-0.693GF-Q0.6_03.5sec.hid (C): C10_RD14-0003-48_49_50_29-1; 4; 4; 4-M3a-0.195GF-Q0.4_03.5sec.hid

1.3.4 Binary Methods of Interpretation

***Analytical and Stochastic Thresholds*:** When assessing the presence of an allele both an Analytical Threshold (AT) and a Stochastic Threshold (ST) are often used. The aim of introducing an AT is to exclude most of the baseline noise from analysis, while the aim of the ST is primarily to alert the DNA analyst that all of the DNA typing information may not have been detected for a given sample. The ST is defined as the peak height value below which it is reasonable to assume that allelic dropout may have occurred within a single-source sample [107]. The SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories 1.1 states that an AT should be established on signal-to-noise analysis of internally derived empirical data and should not be used for purposes of avoiding artefact labeling, which may result in loss of data should an exceedingly high AT be used [107]. Additionally, SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories 1.7.1 states that a ST must be based on empirical data derived within the laboratory and specific to the quantitation and amplification system (e.g kits) and the detection instrument used [107].

***Maximum Allele Count*:** The Maximum Allele Count (MAC) is a commonly used approach for estimating the minimum number of contributors to a mixture sample. This is done by first determining the locus that has the greatest number of allelic peaks and then counting these peaks. A single individual should only contribute a maximum of two peaks per locus and so should we find five peaks, one can infer the presence of at least two other contributors because for a two-person mixture, the expected number of alleles is four. Dembinski et. al. [32] found that the MAC method is not reliable beyond three person mixtures, highlighting the limited use of this binary approach.

***Heterozygote Balance*:** Heterozygote balance refers to the ratio of peak heights between the two alleles of a heterozygote. Balding et. al. [8] highlights LTDNA profiles experience exaggerated peak imbalance. This amplified imbalance can make it especially challenging to reliably pair alleles into major and minor genotypes when dealing with a mixture as reflected in [21].

***Stutter Ratio*:** As stuttering is a common artefact, it follows that predicting the rate of stuttering is important for interpretation of DNA profiles. A simple measure of how repetitive a strand is would be to consider its length [17]. It has been shown that when the structure of the STR region is simple there exists a linear relationship between the stutter ratio and the allele length [56, 117]. Klintschar et. al. [56] found that the number of uniform repeats is relevant for the degree of stutter and that polymerase slippage correlates to the length of the longest homogeneous repeat stretch. This assumption was also made by Walsh et. al. [117] after the determination of different sequence structures of vWA alleles. Specifically when considering LTDNA, we recognise the increased stutter peak heights as a result of the increased rounds of PCR, similarly observed in [8].

1.4 Goal of This Research

The capability of the LR is dependent on a combination of the NoC assumption and the ratio of contribution. As the complexity increases, these difficulties necessitate the use of statistical methods to make deductions. The goal of this research is to reduce the computational strain of the LR by restricting its calculation to the case of a single contributor, thus removing the approximation of proportionate contribution. We seek to develop methods that allow for an appropriate grouping of single-cells into their genotype by first separating all cells into distinct single sources and then creating an EPG from the signal produced by each individual source. Traditionally, EPGs have been created using bulk samples, that is to say EPG signal may stem from multiple sources and post-processing it is not possible to disaggregate into per-cell signal. By considering the complex mixture with a single-cell approach, we can overcome this signal separation problem, allowing for a much simpler assessment of an EPG from a single genetic source. We expect by doing so one will be able to determine the true NoC for mixtures of greater than three contributors, distinguish minor contributors more frequently, or when dealing with touch samples, produce a more complete DNA profile. If we can successfully group single-cells by genotype we can restrict the LR calculation to the case of a single contributor and although the defense's probability will still require some approximation, that of the prosecution's will be a straightforward computation. One of the challenges faced when dealing with single-cell DNA profiles is the sheer volume of data one has to work with. They must now simultaneously assess many EPGs from possibly distinct sources, creating a combinatorial explosion.

1.4.1 Resolution Achieved in This thesis

When treated in a forensic agnostic way, irrespective of their size, single-cell EPGs have a notably smaller similarity measure if they are from the same genetic source than those from distinct sources. By setting aside the clear distinction between loci recordings we can express an EPG as a high dimensional vector and subsequently visualise a collection of EPG-vectors in a low dimensional space. Such a visualisation allows an analyst determine the number of genetic sources comprised in the collection of EPGs. We establish that clustering complex mixtures is possible using a combination of dimensionality reduction, low dimensional visualisation and an appropriate similarity measure. We then demonstrate the capabilities of model-based clustering using two transformations of the data; first the recorded fluorescence of each EPG is normalised, second the logarithm base ten of each normalised fluorescence is taken.

1.5 Thesis outline

In this thesis, we have first introduced the reader to the the field of Forensic DNA Analysis, discussing terminology and common practices. Chapter 2, section 2.1 continues to summarise the current practices with a focus on the analysis techniques employed when faced with low-template and complex DNA profiles. We then shift our focus to the more recent studies of single-cells in a forensic context, providing a brief history of their role and the current analysis pipeline of singe-cell EPGs, see section 2.2. An overview of our data has been presented in Chapter 3, including a detailed description of the data, section 3.1, the application of a quality control, section 3.2, and a novel data transformation that allows us to manipulate the data into high dimensional vectors as opposed to a collection of per locus signal, section 3.3. In Chapter 4, we establish that distinct genotypes are distinguishable in this high dimensional format, section 4.1, and we determine that the logarithmic transformation of our data is reasonably Gaussian, section 4.2. We consider two methods that enable a 2D visualisation of our high dimensional EPG-vectors in section 4.3. In Chapter 5, we layout our experimental design, section 5.1, and propose two solutions for non-parametric unsupervised learning, sections 5.2 and 5.3, concluding with our results, section 5.4. And finally an in-depth discussion is carried out in Chapter 6.

2

History Of Single Cell Analysis

To fully appreciate the need for a single-cell analysis pipeline, we will first familiarise ourselves with the current bulk processing pipeline and the subsequent challenges faced, particularly when dealing with low-template and/or complex DNA mixtures. We will then give a brief historical account of the introduction of single-cell measurements into the field of forensic analysis, followed by a discussion as to why it has yet to be fully adopted. We then focus on the development of methods for the collection and processing of single-cells and subsequently the inference techniques that have been developed for the interpretation of the resulting collection of EPGs.

2.1 Current Practice for Determining the NoC in Complex, Low-Template DNA Samples

The interpretation of Low-Template DNA (LTDNA) crime scene samples can rely on the assumption that such a sample is in-fact composed of multiple genotypes originating from n unknown contributors. Determining an appropriate Number of Contributors (NoC) assumption is crucial for an effective use of a likelihood ratio and so, an analyst must first establish the LTDNA profile, followed by an analysis of such a profile to estimate the NoC.

2.1.1 Establishing Low-Template DNA Profiles

With the constant advances of instrumentation sensitivity, DNA profiles can be obtained from low level DNA amounts, corresponding to just a few cells. Gill et. al. [48] highlights that the starting quantity of DNA material is an important factor for successfully creating a full DNA profile, noting that if the amount of DNA is less than 100 pg then obtaining a full DNA profile by utilising normal PCR conditions becomes less likely. With the intention of enhancing detection sensitivity, they recommend increasing the number of PCR cycles from 28 to 34, a practice still seen today with certain kits recommending it as their standard. However, thanks to the increased number of cycles, these LTDNA profiles are subject to stochastic effects, such as allele dropout and highly variable stutter peak heights, making the interpretaion of these profiles substantially more challenging than those produced from high-template samples.

Even with this detection advancement, one must still consider at what point a detection technique can no longer deliver reliable results. Butler and Hill [21] examine the "Stop Testing" approach which reflects this concern. They consider the increased number of cycles to be an "enhanced interrogation technique", suggesting a follow up of further testing measures to avoid reporting incorrect results due to the observed stochastic effects. One such test is to include replicate PCR amplifications to produce a consensus profile. Replicate analysis involves dividing the DNA into several tubes prior to amplification to produce multiple EPGs, one per tube, for the same sample. The motivation for doing replicates is that the analyst can now compare multiple EPGs from the same sample, potentially providing more information about the genetic source as the stochastic effects will vary from one replicate to another. Gittelson et. al. [49] found that two replicates, each having an average allelic peak height as low as 43RFU, generally have a greater expected net gain than a single DNA analysis. They comment that it is a worthwhile technique to employ even when the expected average allelic peak is greater than 43RFU.

There has also been meaningful research into the use of direct PCR which allows for the generation of LTDNA profiles without the need to increase the PCR cycle number beyond the manufacture's recommendation. Sim et. al. [103] validated the feasibility of a direct PCR system for establishing a DNA database as a substitute for conventional PCR. They found a noticeable increase in detectable peak heights when using direct PCR. Templeton et. al. [111] found that omitting the DNA extraction step is the key success if there is limited sample DNA. They found that sufficient DNA can be isolated from enhanced touch samples, such as fingerprint deposits, using direct PCR but with its merits comes limitations. For instance, the quantification of the DNA sample cannot take place and there is no opportunity to remove potential PCR inhibitors. There is much research advocating widespread implementation of direct PCR into forensic laboratory practice [24, 6, 25] due to the faster sample turnaround times, improved results, and fiscal savings this method offers. Alternatively, Wong et. al. [123] offers a modified direct PCR amplification method, recommending the combined use of direct PCR and replicate PCR. Although this method adds 30 minutes to the processing time, with the inclusion of an additional pre-amplification step, they find this method offers a significant advantage in consensus-based interpretation of DNA mixture profiles.

2.1.2 Analysis of Low-Template DNA Profiles

Bio-statistical interpretation of LTDNA profiles was originally achieved by a binary interpretive approach and named as such since the probability of a crime scene profile given a proposed genotype was assigned zero (genotype exclusion) or one (genotype inclusion). Binary approaches include the application of an analytical and/or stochastic thresholds, stutter ratios, and peak height ratios, and therefore necessarily involves the assumption that all alleles are either unmistakable or may be masked by an artefact such as stutter [26, 20, 44]. The binary approach does not consider the stochastic effects, drop-in and drop-out, nor does it make full use of the available information regarding peak heights [19]. These shortcomings restrict the ability of binary methods and subsequently the Random Man Not Excluded (RMNE) calculation when dealing with complex LTDNA and consequently led to the development of probabilistic software that factors the probabilities of such stochastic effects. The Random Man Not Excluded (RMNE) or the Probability of Exclusion (PE) is a statistic that can be calculated and presented in a court of law that represents the weight of the evidence. The method establishes a nonexclusion by fragmenting the population into those that can not be excluded as a contributor [19]. It is a two step process where first the suspect is determined excluded or not, followed by a statistical calculation. Although conceptually equivalent to the likelihood ratio discussed in section 1.3.1, Gill et. al. [44] states that the RMNE simply does not make use of as much information contained in the signal as the LR approach, and later Gill et. al. [46] observes that though RMNE is still employed in practice, this method of evaluation is being replaced with the LR approach.

Moving away from binary methods, instead trying make use of more of the information contained in the signal, led to the development of what are called semi-continuous and fully-continuous models that can calculate rigours LR values, such as LikeLTd, Lab Retriever, LRmix Studio, EuroForMix, STRmix and CEESIt [94, 60, 39, 13, 16, 104] to name but a few. These are probabilistic models for the generation of EPGs given a genotype that try interpret the data by including both continuous and discrete features such as drop-out and drop-in [46, 104, 54]. Semi-continuous methods threshold the data and therefore only try probabilistically explain large effects while the fully-continuous methods intend to have a probabilistic model that interprets the raw EPG [5]. Semi-continuous models were initially in wider use primarily due to the accessibility of software being open source, secondly the fact that their algorithms and computations are more straightforward and tertiarily, their workings and results can be easily presented and discussed in courtrooms [15, 27]. However, fully-continuous approaches are seen as more powerful since they exploit more of the available LTDNA profile information and in recent years an increasing number of open source, fully-continuous software have been published [77, 5].

Alladio et. al. [5] found that fully-continuous models appear to be the most appropriate bio-statistical methodology to perform analysis of LTDNA mixtures. Although models such as these can account for both the qualitative and quantitative data provided by the EPGs, they rely on model specific parameterisation and various distributional assumptions of the signal intensities [47], for example STRmix performs Markov chain Monte Carlo simulations to estimate the distribution of peak heights [16], while EuroForMix uses the gamma distribution to model peak heights [13]. It is expected that different models will produce different LRs for a given sample and a given set of hypothesis.

2.2 Advances in Single-Cell Analysis

2.2.1 A Brief History of Single-Cell Analysis within the Field of Forensics

Looking to sidestep the issues presented when processing LTDNA, there has been a movement toward directly profiling single-cells from evidence material. The application of cellular separation techniques before the DNA extraction step offers to potentially reduce the complexity of downstream interpretation. The study of single-cells is not a new phenomenon, its analysis allows the study of cell-to-cell variation both within a cell population, e.g. organ or tissue of a distinct individual, and outside a cell population, e.g. comparing individuals for such ends as parental lineage. The study of single-cells has had significant implications for genetic disease diagnosis, drug development, in-depth analysis of stem cell differentiation, cancer and much more [40, 119, 72]. In its infancy, it had been suggested that an analysis of single-cells, particularly whole genome amplification of single-cells, could potentially aid forensic investigation [125], but it was Findlay et. al. [38] who first reported a system for determining STR profiles from single-cells using modern forensic techniques in 1997. Since this recommendation there has been extensive research to progress the pipeline for single-cell DNA typing [74, 73, 18, 41]. Johnson and Ferris [63] assess the application of single-cell gel electrophoresis to evaluate the postmortem cell death process, however this does not involve the profiling nor following analysis of crime scene stains. Dean et. al. [31] utilised an analytical technique in an attempt to group single-cells by genotype prior to PCR yet they do not establish single-cell EPGs, instead choosing to bulk process cells they believe to belong to a distinct source. Similarly Huffman et. al. [59] have demonstrated that it is possible to extract single source DNA autosomal STR profiles from the case of a two person admixture in equal ratio by introducing a mixture de-convolution technique, Direct Single Cell Subsampling (DSCS). Although interest in single-cell technologies for the forensic sciences has been rekindled more recently, the literature would suggest limited post-profiling analysis of single-cells, particularly in the case of sorting large numbers of profiles.

2.2.2 Collection of Single-Cells and Establishing EPGs

In the single-cell process, each cell is separated prior to DNA extraction. As a result, the data consists of n EPGs from n distinct cells, rather than one EPG from m not necessarily distinct cells. There exists many methods for cell separation, to list but a few: pico-pipetting [66], microfluidic cell sorting [70], Laser Capture Microdissection (LCM) [114], Fluorescence-Activated Cell Sorting (FACS) [31], and notably DEPArray technology [7], a microchip-based digital sorter, which combines microfluidic and microelectronic methods enabling precise image-based isolation of single-cells. This image-based isolation allows the analyst an opportunity to manually infer the quality of a cell. Cell features such as the nucleus or cell wall are tagged by various coloured dyes illuminating their presence (or lack there of). The analyst can then record cells where a nucleus has not been detected and choose to discard these cells from further study.

Post cell sorting, the processing pipeline of separated cells is akin to direct PCR wherein the cells are added directly to the amplification mixture, omitting the need for an extraction step. This requires the choice of extraction reagents to be compatible with forensically accepted PCR components, a choice that is cautiously determined, as DNA extraction has been shown to have the greatest contribution to loss of DNA yield [43]. Sheth et. al. [102] have very recently demonstrated that the Arcturus$^{®}$PicoPure$^{\text{TM}}$ extraction method resulted in the most promising DNA yield for the GlobalFiler$^{\text{TM}}$ STR loci amplified at half volume when assessing the compatibility of four extraction treatments on pico-pipetting single buccal cells. Their work focuses on optimising the chemistry so that the extraction and amplification process works in concert.

This pipeline is not frequently availed of within forensic practise largely due to two main factors: i) exuberant costs, a combination of the initial expense to separate cells coupled with the vastly increased number of cells undergoing individual rounds of PCR, resulting in a larger need of chemicals, and ii) a lack in development of suitable inference tools for analysing the vast quantity of resultant EPGs. Although we cannot directly affect the financial demand of a single-cell pipeline, we aim to introduce an appropriate analysis technique to handle the large volume of resultant EPGs.

2.2.3 Analysis of Single-Cell DNA Profiles

The Interpretation of single-cell EPGs (SC-EPG) is one of the same to that of LTDNA EPGs, true allele signal, increased stutter peaks, noise, high imbalanced stutter ratios and allelic drop-out/drop-in [59]. Anslinger et. al. [7] detected stutter peaks as high as 34% of the parent allele in single-cell profiles. When interpreting SC-EPGs while aware of occasional drop-in, we need not concern ourselves with the presence of more than a single contributor instead we must be acutely aware of the increased chance of drop-out

[89, 7]. Drop-out rates occur more frequently when amplifying low amounts of DNA and single-cells are no exception to the rule. Using LCM, Sanders et. al. [99] report drop-out rates of 26.9% for 75 sperm cells. Kim et. al. [68] reported an average drop-out rate of 18.3% for the five methods they tested, with a notable high of 43.9% for one of their methods. Anslinger et. al. [7] reported a drop-out rate of 18% per cell and detected stutter peaks as high as 34% of the parent allele in single-cell profiles.

2.2.4 Enabling the Assessment of Single-Cell EPGs

Both semi-continuous and fully-continuous interpretive approaches have been wholly established in the case bulk processed DNA samples but the literature would suggest this has yet to be considered for the interpretation of single-cell EPGs. We propose unsupervised machine learning techniques which will enable the assessment of groups of SC-EPGs by removing the computational hurdle of evaluating high volumes of EPGs from not necessarily distinct sources, and thus reducing the computational strain of the LR. The LR tends toward one as the number of contributors increases or the DNA template level decreases resulting in reduced inference power [88, 110]. Our method would allow for each genotype present in a mixture sample to be described by their own distinct collection of EPGs hence the LR computation would concede to its simplest case of a single contributor.

We suggest two methods: i) a two step process wherein the data must first be visualised and a number of contributors inferred followed by similarity based clustering to group EPGs by genotype. ii) An automated method that determines the number of contributors while simultaneously grouping EPGs by genotype. Both methods offer promising results with the latter marginally outperforming the former.

3

An Analytical Study of the Single-Cell Data

We intend to cluster single-cell EPGs by genotype and thus reduce the computational strain when computing the LR for a collection of single-cell EPGs. If we can successfully cluster single-cell EPGs into their distinct genotypes, we can restrict the LR calculation to the case of only one contributor. We are in the fortunate position where our wet lab. collaborator has created more than 500 single-cell EPGs with buccal epithelial cells from known genotypes. This allowed us create simulated admixtures where the true number of contributors and their ratio of contribution is known prior to clustering, of course this knowledge would not be available in real world situations. A full description of the the data including data specific drop-out rates and its processing pipeline have been included.

Not all EPGs are of a high quality, some contain little to no information and since the aim is to develop a method to effectively group single-cell EPGs by contributor, we shall use high-quality EPG data to validate the clustering performance against ground truth. We develop a method to discard scant EPGs. To determine the quality of an EPG, we aim to use information that is measurable even in the case of an unknown genotype, therefore we focus on the total fluorescence of a given EPG. For an EPG to be included in our study, we apply a high-pass filter to the total fluorescence of said EPG, one that corresponds to a minimum requirement of 20 true alleles being detected.

Once satisfied with the remaining collection of EPGs for our study, we consider alternative forms which we may choose to express the information of an EPG. In doing so, we have established a format in which one can store the recorded fluorescence in a location corresponding to an allelic variant of a loci within a high dimensional vector space. Although this representation permits the consideration of a wider range of methods for statistical comparison between single-cell EPG information, in doing so there is a trade off, one must sacrifice an awareness of both loci boundaries and allelic variant representation, that is to say, looking at a single EPG-vector one could not alone establish the original EPG.

3.1 Description of the Data

3.1.1 Collection and Processing of the Data

The data has been generated from single-cell experiments performed by our wet lab. collaborators, headed by Professor Catherine Grgicak, in the Department of Chemistry at Rutgers, Camden, New Jersey, USA. DNA was extracted from epithelial cells, namely saliva using the PicoPure® extraction kit [115] and processed as follows: i) single-cells were separated via micropipette, each cell was placed in a 96-well microtiter plate already containing 5μL of extraction mix; ii) the plate was then placed in a thermal cycler for 3 hours at 65°C, followed by a 10 minute 95°C incubation to inactivate Proteinase K used for cell lysis; iii) the genetic material of each single-cell was then amplified for 30 cycles using the GlobalFiler$^{\text{TM}}$ Amplification Kit [118]. The PCR reaction consisted of 5μL of extraction mix, 5μL of the master mix, and 2.5μL of buffer to bring the total reaction volume to 12.5μL; iv) the amplified fragments were separated and detected using capillary electrophoresis and laser induced fluorescence, respectively. All amplified products were prepped and run in the same way on the Thermofisher 3500 Genetic Analyzer® capillary electrophoresis machine; and finally v) all data was analysed in OSIRIS [51] at the lowest analytical threshold allowed (5RFU). The Genotype Tables, which included the File Name, Marker, Dye, Allele, Size and Height, were exported for further analysis.

3.1.2 Reading Tabulated Raw Data

The data is expected to be publicly available in the PROVEDIt database [4] after publications are complete and therefore we are providing a description of the current Grgicak lab. naming convention so that the interested reader can identify original data from which all figures are made. The *Sample.File* column contains a long character string which is the naming convention for each single-cell. We expect a unique character string to appear at most 24 times in this column as 24 markers have been recorded for each cell. The *Marker* column specifies the loci being examined and the *Dye* column indicates which colour DNA-binding dye has been added for the measurement of fluorescence. Columns *Allele X*, *Size X* and *Height X* are the measurement's recorded at each loci. Although we see Allele 1, Allele 2, and so on for each loci, these are in fact potential alleles not true ones. This numbering is such that Allele 1 is the first allelic variant where a peak has been detected for a locus, it may be noise, stutter or a true allele. Off-ladder readings are recorded as *OL*. *Size* refers to the length of an STR in bps and *Height* is the height of a peak in relative fluorescence units (RFU).

3.1.3 Grgicak Lab. Naming Convention of Single-Cell EPGs

Understanding the full naming convention is not paramount for our analysis of the data but we do require at minimum, an understanding of how one determines the genetic source identifier of an EPG by reading this character string. This is best done through example, see Fig. 3.1. We are mindful that this naming convention is subject to change as more data is made readily available, such as the inclusion of a cell type indicator once various other cell types (namely bloods and sperm) begin to be processed. At its core the naming convention will include the well plate number, the project number, the genetic source, the extraction kit, the amplification kit and the capillary number and so we will identify these in the current convention.

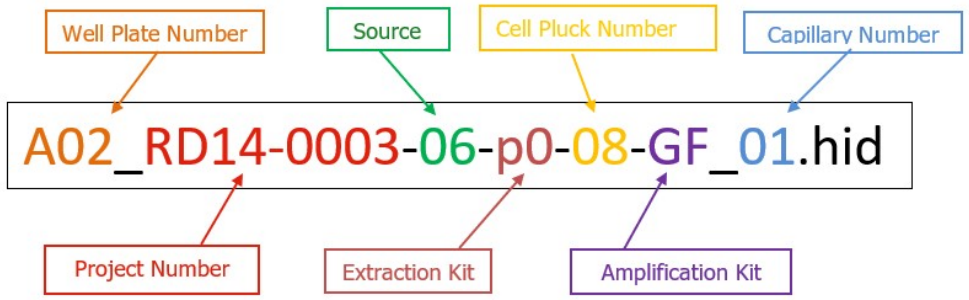


Figure 3.1: Example of how single-cell EPGs are currently named within the *Sample.File* column of the raw data.

Well Plate Number	The location in the 96 well plate that this single-cell was run.
Project Number	It is to be noted that each sample is designated by the combination of project number and source identifier.
Source	In this example, 06 is the source identifier, that is to say this cell has come from Genotype 06.
Extraction Kit	The type of kit used to extract the DNA, p0 indicates that the PicoPure kit has been used.
Cell Pluck Number	Which cell it is that has been plucked, for this example it was the 8th cell plucked by the micropipette.
Amplification Kit	The type of kit used to amplify the genetic material of the cell, GF indicates that the GlobalFiler^{TM} kit has been used.
Capillary Number:	The capillary this cell was run in, for this example the cell has been run in Capillary #01.

3.1.4 Description of All Sources

The GlobalFiler$^{\mathrm{TM}}$ kit allows the amplification of 24 loci: the 20 core CODIS loci, SE33, AMEL, Y-indel and DYS391 [108]. When plotting a complete EPG we have only included 21 loci. The loci AMEL, Y-indel and DYS391 have been removed for analytical interpretaion of the data as these are not autosomal STRs. DYS391 is an STR found on the Y-chromosome, Y-indel is a Single Nucleotide Polymorphism (SNP) on the Y-chromosome and AMEL is the amelogenin sex-determining locus. Although not a core CODIS locus, we have included the SE33 in our study. There has been recent studies into the reliability of the SE33 locus as a genetic marker for forensic DNA analysis systems [14, 11] with Karantzali et. al. [65] demonstrating further distinction between true and false matches in the case of paternity and full-sibling matches.

To visualise the raw data, we have plotted all single-cell EPGs overlaid for each genotype. This allows us a broad overview of potential complications in the downstream analysis. For example, Fig. 3.3 indicates potential allelic drop-in at the locus D2S441 for Genotype 02. We can trace back to EPGs associated with noisy loci for future reference, these EPGs have been recorded in Table 3.2. Ground truth for each genotype has been recorded in Table A.2, see Appendix A. For all genotypes at least 1 complete profile could be found and at least 4 profiles recorded no true allelic variants after applying a detection threshold (DT) of 30RFU to the data. On average the loci D10S1248 and D2S441 experience the lowest drop-out rates at 25% while the loci D7S820 and SE33 experience the highest drop-out rates at 56% and 57% respectively. Genotype 05 experienced both the highest average drop-out rates for heterozygote loci (57%) and homozygote loci (38%). A full description of drop-out rates can be found in Appendix A, section A.3.1.

Genotype	Total Num. EPGs	Num. Complete Profiles	Num. Profiles No Alleles Detected
01	111	11	4
02	110	5	6
05	100	1	11
06	102	5	20
07	103	3	5

Table 3.1: The number of EPGs per genotype, the number of profiles where all true alleles were detected above 30RFU AT and the number of profiles where no true alleles were recorded for each genotype.

Genotype	EPG	Locus
01	B01_RD16-0003-01-p0-81-GF_02.hid	D2S441
	E10_RD16-0003-01-p0-67-GF_05.hid	D2S441
02	H10_RD16-0003-02-p0-70-GF_08.hid	D2S441 D3S1358 D22S1045
	H11_RD16-0003-02-p0-77-GF_08.hid	D2S441 D3S1358 D10S1248
06	A05_RD16-0003-06-p0-31-GF_01.hid	D8S1179 D10S1248 D19S433

Table 3.2: EPGs believed to be associated with drop-in and the loci which they can be observed in.

Genotype 01

We observed both the greatest number of complete profiles and the least amount of empty profiles for this genotype. On average approximately 71% of the alleles could be obtained per single-cell for Genotype 01. Looking at all cells for this genotype, there is an average drop-out rate of 28% across all heterozygote loci. A notable high of 51% of expected alleles are absent from recordings at D7S820, while the lowest ADO for a heterozygote locus was 15% at D2S441. When studying drop-out of homozygote loci we can only determine if total ADO has occurred or not and we find that for the four homozygote loci of Genotype 01, D1S1656, D3S1358, D22S1045 and TH01, total ADO can be observed in 14%, 12%, 10% and 17% of profiles respectively. Looking to the locus D2S441 in Fig. 3.2, we suspect drop-in is responsible for the high recordings at 7.3 repeat units.

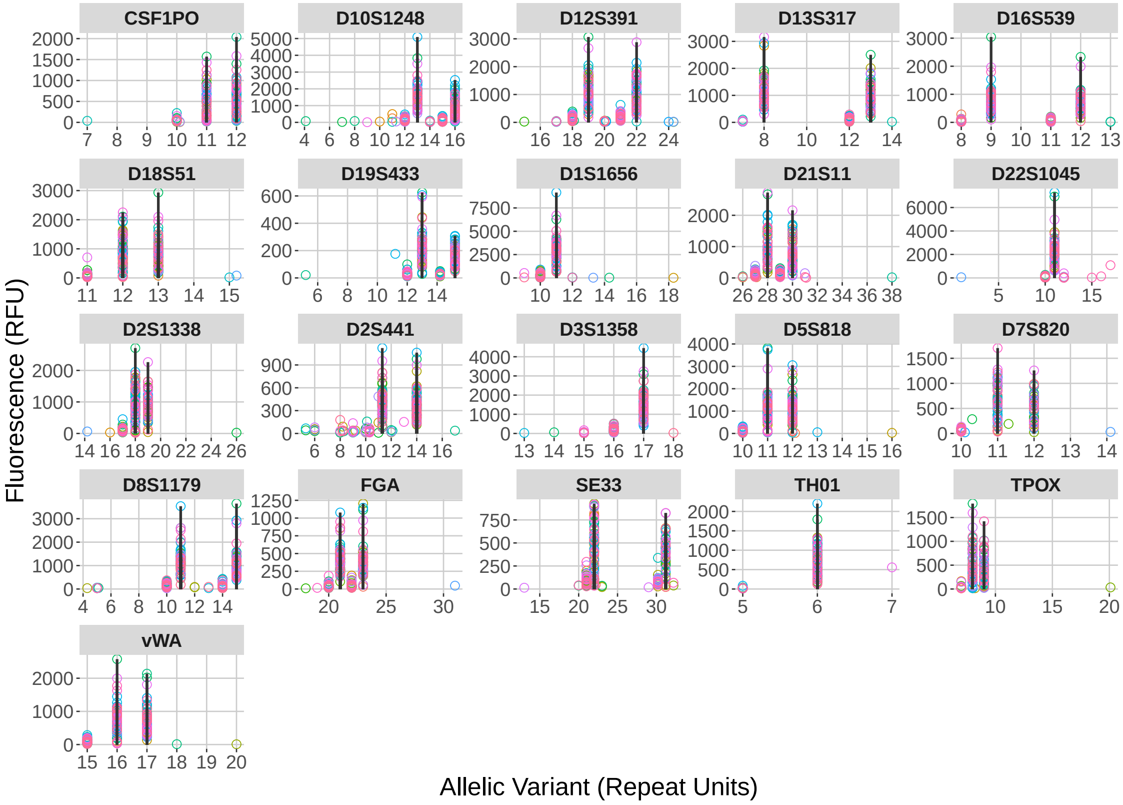


Figure 3.2: All wet lab. data for Genotype 01 plotted as points, coloured by cell. Ground truth is indicated by bold black vertical lines.

Genotype 02

Although we only obtain five complete profiles for Genotype 02, on average this genotype has the strongest profiles with a mean of approximately 78% of alleles obtained per single-cell, the highest for any genotype tested. The lowest average ADO across heterozygote loci is observed for this genotype at 22% and ranges from 13% at the locus D2S441 to 36% at the locus TPOX. Genotype 02 experiences the most homogeneity with total ADO ranging between 13% – 20% for the five homozygote loci, the second lowest average drop-out rate for a genotype. Two particularly noisy loci, D2S441 and D3S1358 can be observed in Fig. 3.3. Although the colouring looks similar, these recordings stem from two cells, again we suspect drop-in is accountable for these chaotic values.

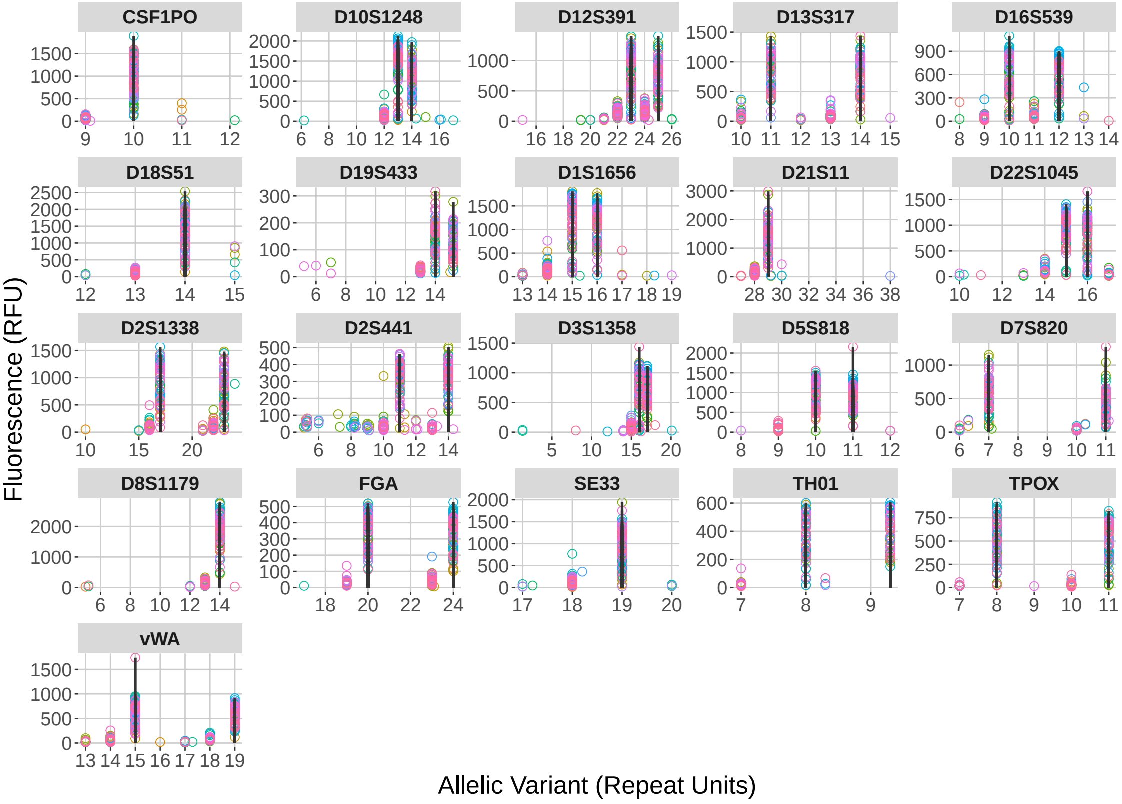


Figure 3.3: All wet lab. data for Genotype 02 plotted as points, coloured by cell. Ground truth is indicated by bold black vertical lines.

Genotype 05

The profiles established for Genotype 05 were of the poorest quality, on average approximately 45% of the alleles could be obtained per single-cell, only one complete profile was obtained and this genotype experiences an average drop-out rate of 57% across heterozygote loci and 38% across homozygote loci. In fact, of all the cells profiled for this genotype, less than half of the expected alleles were detected for 12 of the 17 heterozygote loci, with substantial ADO of 76% observed at D7S820. For the homozygote locus D21S11, true allelic variants are only observed in 50% of profiles.

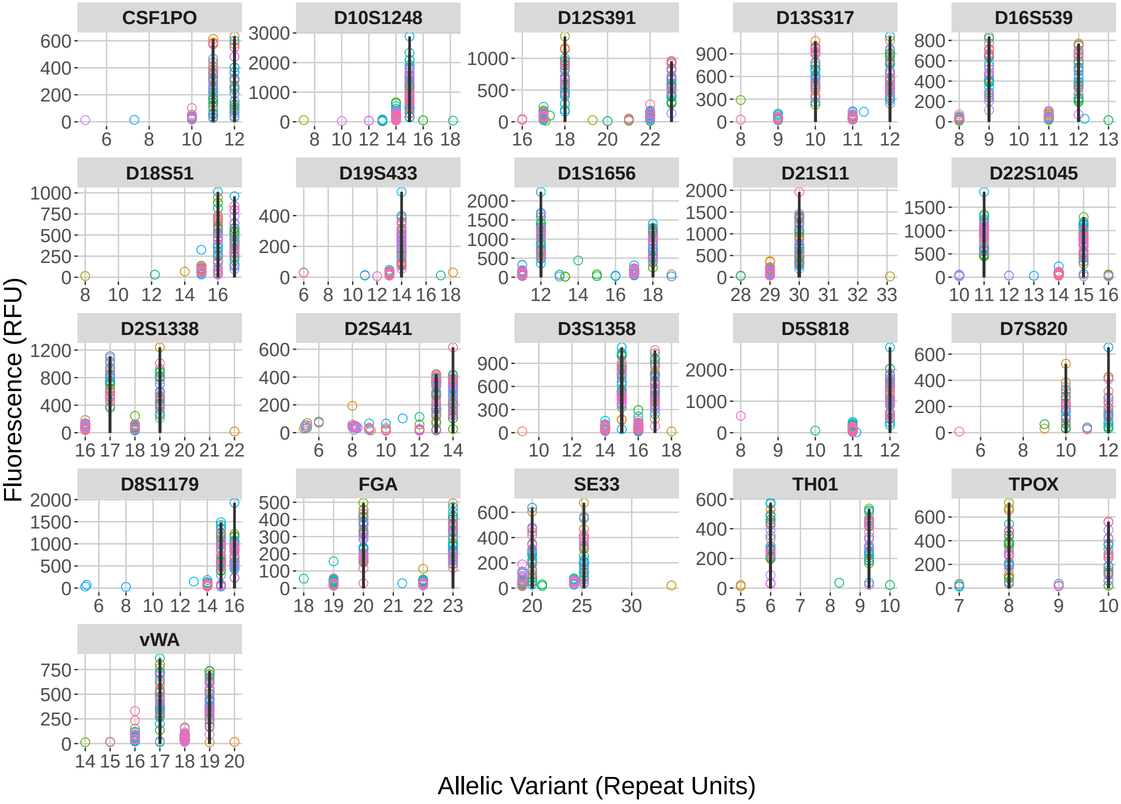


Figure 3.4: All wet lab. data for Genotype 05 plotted as points, coloured by cell. Ground truth is indicated by bold black vertical lines.

Genotype 06

For Genotype 06, on average approximately 50% of the alleles could be obtained per single-cell. 20 of the profiles established for Genotype 06 had no true alleic variants detected, the most observed for any genotype. Although not quite as frequent as Genotype 05, we still observe relatively high ADO across all loci for Genotype 06 with an average of 51% across heterozygote loci (a notable high of 61% at D2S1338) and 38% across homozygote loci.

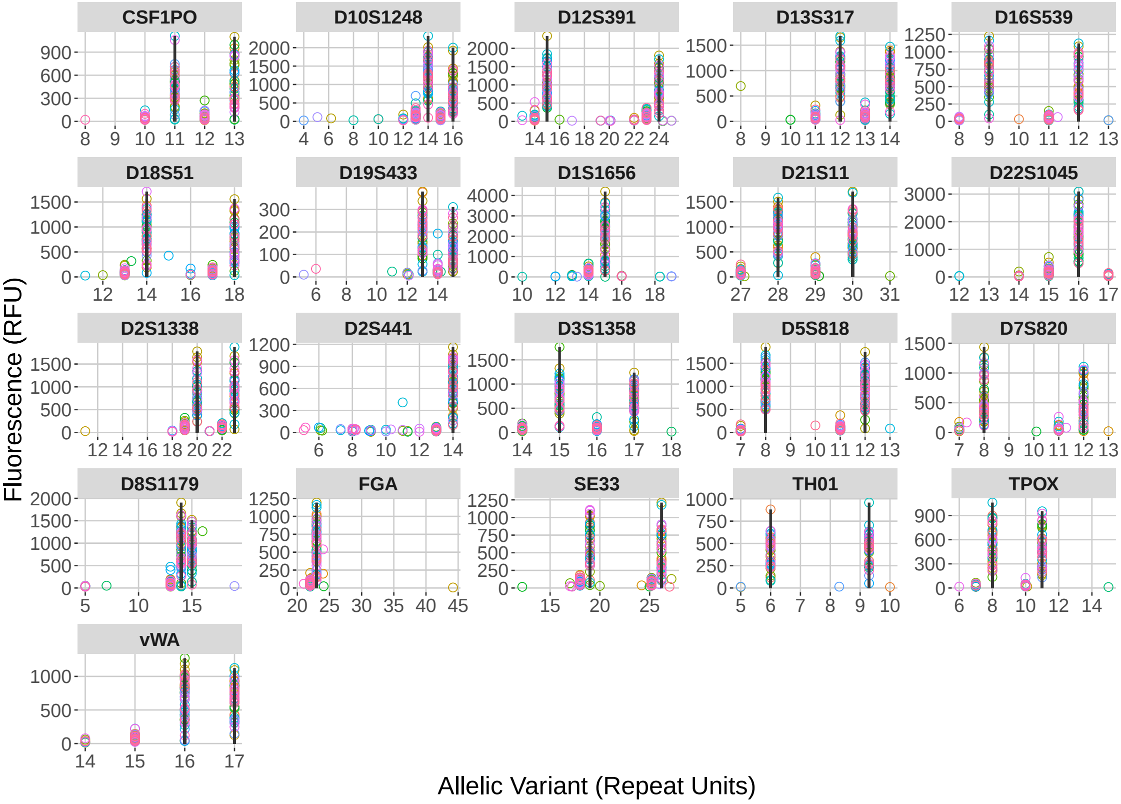


Figure 3.5: All wet lab. data for Genotype 06 plotted as points, coloured by cell. Ground truth is indicated by bold black vertical lines.

Genotype 07

For Genotype 07, on average approximately 59% of the alleles could be obtained per single-cell. D3S441 is the only homozygote locus and the true allelic variant could be determined in 82% of profiles. An average drop-out rate of 41% was observed across heterozygote loci, and the second highest drop-out seen for all genotypes was 67% at D7S820.

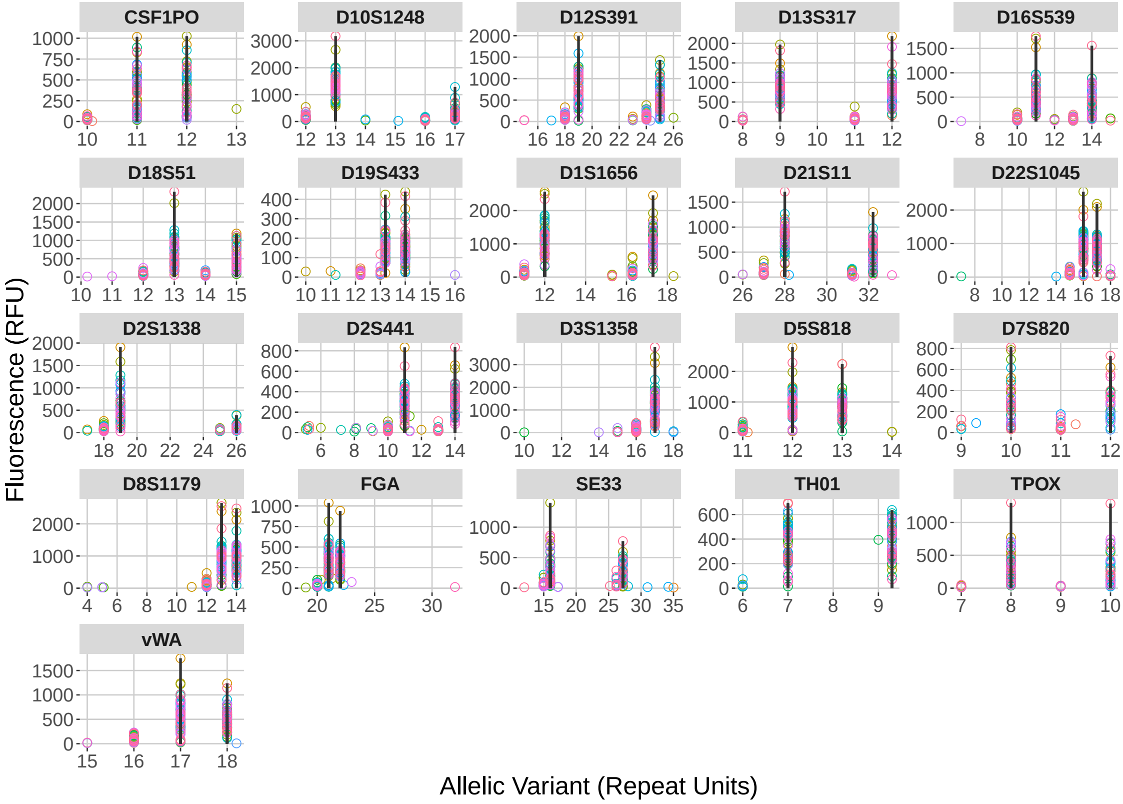


Figure 3.6: All wet lab. data for Genotype 07 plotted as points, coloured by cell. Ground truth is indicated by bold black vertical lines.

3.2 High-Pass Filter

We introduce a high-pass filter restricting the use of EPGs with very little fluorescence recorded which corresponds to restricting the use of EPG's where few true alleles have been recorded. We recognize that the decision to apply a high-pass EPG threshold for single-cell purposes may have implications to casework, though additional research would be required to characterize to what degree. However, we begin with high-quality data to develop and validate a method for forensic purposes, such that it is tested against real data that is known to exhibit signal from a "ground truth" contributor unambiguously. We have defined an EPGs intensity, I_k, as the sum of all peak heights, f_i, recorded for EPG_k, $i, k \in \mathbb{N}^+$ and we use this intensity as a proxy for the information contained in an EPG.

$$I_k = \sum_{i=1} f_i^{EPG_k}. \tag{3.1}$$

As this is known data, we can determine how many alleles were recovered above a DT of 30RFU, plotted against the EPG intensity I_k, seen in Fig. 3.7. We observe a correlation between how many alleles were recovered and an EPG's intensity, however, this is not linear as we might expect. In fact, as the number of true alleles recovered increases, the relationship becomes almost superlinear.

Analysis of the marginal distributions of true alleles suggests that they can be approximately modeled via a log-normal distribution. Thus, taking the logarithm of the data is a natural transformation to get the normality of these marginal distributions. Where the intensity J_k is determined considering the sum of log transformed peak heights,

$$J_k = \sum_{i=1} \log_{10}(f_i^{EPG_k}), \tag{3.2}$$

and we visualise in the same way as in Fig. 3.7, we find that there is a linear relationship between the EPG intensity, J_k, and the number of alleles recovered. This suggests to threshold the EPG quality based on the sum of the log transformed fluorescences. Here we chose to do so such that linear regression would suggest we remove EPGs for which 20 or less alleles have been recovered, as indicated by the red dashed line seen in Fig. 3.8.

Post filtering, a total of 351 EPGs remain, most of which have at least 20 true alleles detected. Three EPGs from Genotype 06 that have less than 20 true alleles detected escaped the high-pass filter, summarised in Table 3.3. Three of the EPGs associated with noisy loci are removed from the study by applying this filter to the data. The average drop-out rate post filtering decreased to less than 20% for all genotypes, with the locus D10S1248 still experiencing the lowest drop-out rate, reduced to an average of

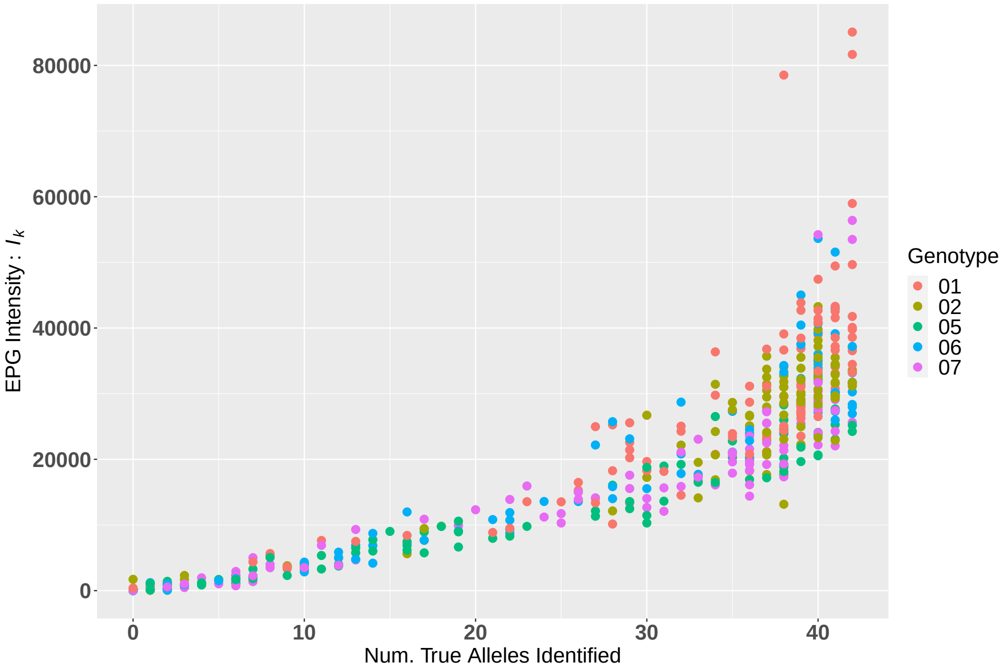


Figure 3.7: EPG intensity, I_k, versus the number of true alleles recorded for that EPG. We expect a linear relationship but observe as the number of true alleles present increases, the relationship becoming almost superlinear.

4.8%, while the locus D7S820 now experiences on average the highest drop-out rate at 36%. Genotype 05 still sees the highest drop-out rate for heterozygote loci (18.9%), while Genotype 06 sees the highest (3%) for homozygote loci. A full description of drop-out rates post filtering can be found in Appendix A, section A.3.2.

Genotype	**Total Num. EPGs**	**Num. Complete Profiles**	**Min. Num. Alleles Detected**
01	91	11	20
02	93	5	23
05	43	1	20
06	55	5	18
07	69	3	20

Table 3.3: The number of EPGs per genotype post filtering. The number of profiles where all true alleles were detected above a 30RFU AT was unchanged by filtering. The minimum number of alleles detected above a 30RFU AT is expected to be 20, however three profiles of Genotype 06 escaped the filtering.

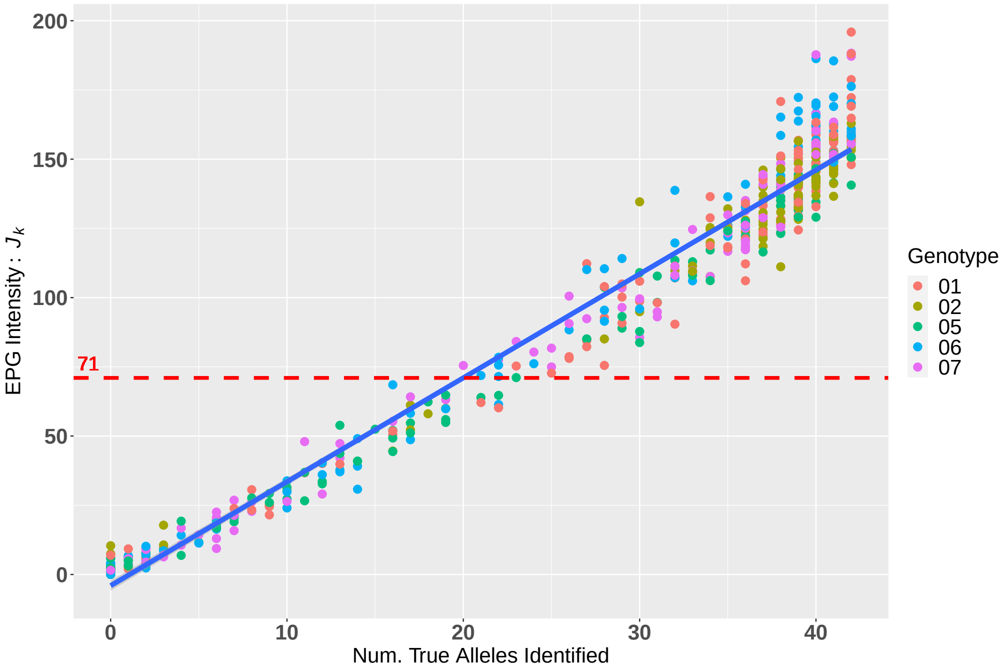


Figure 3.8: EPG intensity, J_k, versus the number of true alleles recorded for that EPG. We see a linear relationship as indicated by our linear model, the blue line. The red dashed line is a high-pass filter removing EPGs where 20 or less true alleles have been recovered for an EPG.

3.3 Creating EPG-vectors

An EPG can be described by a series of triples, $\langle l,\ a_i,\ f_i \rangle$, where l is the locus, $l \in \{1, ..., 21\}$, a_i is the allelic variant and f_i the corresponding fluorescence recorded at a_i, for $i \in \mathbb{N}^+$. In the standard paradigm, the fluorescence measurement at each locus of an EPG are treated differently and indeed typically they are measurements with different fluorophores, having many different ranges of intensity. In order to make these data comparable it makes sense to embed them in a single high dimensional space. The way in which we are going to do so is to take each potential allele location and assign it a unique vector index. Here in, we describe the process of doing so. It effectively amounts to concatenating the per locus EPGs in a consistent manner.

3.3.1 Vector Construction

We have chosen to construct forensic ignorant vectors such that one vector, V_k^G, will describe an EPG in full, dubbed EPG-vectors. G is the genotype id, $G \in \{1, 2, 5, 6, 7\}$ and $k \in \{1, ..., n_{EPG}\}$ where n_{EPG} is the total number of EPGs for genotype G. We say forensic ignorant as we have concatenated the fluorescences in such a way that one cannot readily determine at which loci a fluorescence was recorded thus treating an EPG as a single high dimensional signal. The method we use in creating EPG-vectors can be applied to any EPG data but the dimensions will be data specific. We construct V_k^G as follows:

Create a zero vector of length m, such that

$$m = \sum_{l=1}^{21} n_l. \tag{3.3}$$

where n_l is the data specific set of all potential allelic variants for the locus l such that

$$n_l = 4(\lceil a_{max}^l \rceil - \lfloor a_{min}^l \rfloor) + 1, \tag{3.4}$$

with a_{min}^l and a_{max}^l as the minimum and maximum allelic variants recorded for locus l across all genotypes in our data. As discussed in section 1.1.2, we can be faced with non-integer allelic variants and so to account for this we must take the floor and ceiling of our min and max respectively. It is also for this reason that we must multiply by a factor of 4. We have an offset of +1 to ensure we have the correct number of positions available. We choose $a_{min/max}^l$ across all genotypes present in the data to ensure $\mid V_k^G \mid$ is constant for all G and k.

To ensure each vector is comparable we must concatenate our loci consistently. The order we choose to concatenate is arbitrary but once selected it must remain constant. We have chosen the order: {CSF1PO, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, SE33, TH01, TPOX, vWA}

Now we can populate our vectors with their corresponding fluorescences. The element $v_{p_i^l} \in V_k^G$ is the fluorescence recorded at $\langle l,\ a_i \rangle$ for EPG k of genotype G. Moreover we can say $v_{p_i^l} = f_i^l$, where p_i^l is the position index such that

$$p_i^l = 4\lfloor a_i^l - a_{min}^l \rfloor + 10(a_i^l - a_{min}^l - \lfloor a_i^l - a_{min}^l \rfloor) + \chi, \tag{3.5}$$

where a_i^l is the i^{th} allelic variant, recorded at locus l. The first term in p_i^l accounts for the integer part of an allelic variant while the second term accounts for the non-integer part. If an allele is an integer the second term will be zero. Finally the third term, χ is an offset such that,

$$\chi = \begin{cases} 1 & if\ l = 1 \\ l - 1 + 4\sum\limits_{j=1}^{l-1}(\lceil a^j_{max} \rceil - \lfloor a^j_{min} \rfloor) & if\ l > 1 \end{cases} \tag{3.6}$$

3.3.2 EPG-vector Example

For this example we will only consider the positioning of the first four non-zero entries of V_1^7 as these vectors can be of lengths greater than 1000. This corresponds to taking the first EPG, A02_RD16-0003-07-p0-48-GF_01.hid, of Genotype 07 and determining the vector position for the fluorescence recorded at the loci CSF1PO and D1S1656.

l	a_1^l	f_1^l	a_2^l	f_2^1	a^l_{max}	a^l_{min}	n_l
1	11	23	12	351	13	7	25
2	16.3	41	17.3	643	20	8	49

Table 3.4: Data used to position and populate the first four non-zero entries in V_1^7. a_i^l and f_i^l have been taken from the EPG A02_RD16-0003-07-p0-48-GF_01.hid at the loci CSF1PO and D1S1656. $a^l_{max/min}$ have been determined across all our data (see Appendix A, Table A.1). n_l have been calculated using equation (3.4).

Using the data from Table 3.4 along with equation (3.5) we can assign f_i^l to its appropriate position in V_1^7. If we take $l = 1$ and $i = 1$ we find,

$$\begin{aligned} p_1^1 &= 4\lfloor a_1^1 - a^1_{min} \rfloor + 10(a_1^1 - a^1_{min} - \lfloor a_1^1 - a^1_{min} \rfloor) + 1 \\ &= 4\lfloor 11 - 7 \rfloor + 10(11 - 7 - \lfloor 11 - 7 \rfloor) + 1 \\ &= 17 \end{aligned}$$

$\Rightarrow v_{17} = 23 = f_1^1$, i.e the 17^{th} element of V_1^7 is the fluorescence recorded at the first allele detected at the locus CSF1PO. This process can be followed through to determine $v_{21} = 351 = f_1^2$, $v_{60} = 41 = f_2^1$, $v_{64} = 643 = f_2^2$. A visualisation of this can be seen in Fig. 3.9. These are high dimensional, sparsely non-zero vectors.

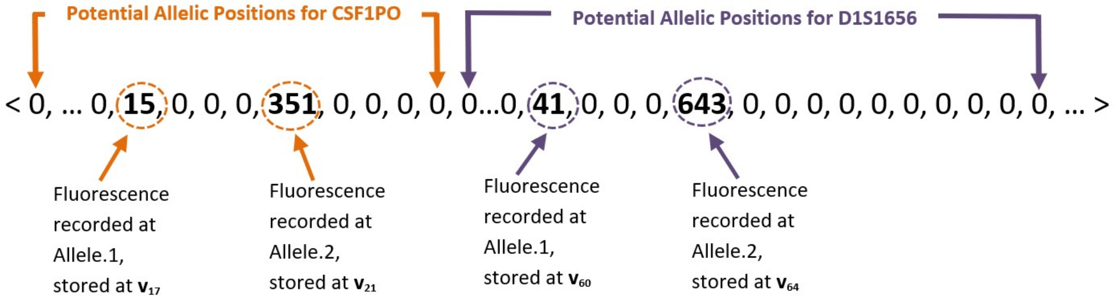

Potential Allelic Positions for CSF1PO
Potential Allelic Positions for D1S1656
< 0, … 0, 15, 0, 0, 0, 351, 0, 0, 0, 0, 0…0, 41, 0, 0, 0, 643, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, … >
Fluorescence recorded at Allele.1, stored at $\mathbf{v}_{17}$
Fluorescence recorded at Allele.2, stored at $\mathbf{v}_{21}$
Fluorescence recorded at Allele.1, stored at $\mathbf{v}_{60}$
Fluorescence recorded at Allele.2, stored at $\mathbf{v}_{64}$

Figure 3.9: Visual representation with commentary for the first 74 entries of V_1^7.

4

Distinguishing Single-Cell Electropherograms

We wish to determine if we can distinguish EPG-vectors by genotype using a similarity measure. We would have reason to believe they can be clustered by genotype if such a distinction can be made. By looking at the similarities across EPG-vectors using various metrics, we find the similarity of EPGs from the same genotypes are distinguishable from the similarity of those from distinct genotypes. Through an exploratory analysis via Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), we investigate if our single-cell EPG-vectors naturally fall into groups, that is trying to determine if an analyst can optically anticipate how many groups are present in an admixture. If this visualisation of our high dimensional vectors proves successful, one is led to believe that an analyst-in-the-loop solution is a reasonable end solution to our problem.

UMAP makes use of a distance metric to measure distances between our input data points, a metric we will select from our similarity analysis. PCA does not require a distance metric, however makes an assumption about the distribution of our input data. As PCA is a method that fits a Gaussian hyper-ellipsoid to the data, it assumes that the data is Normally distributed [112]. Although we will not be making use of such a fit, we must still consider the distribution of our data if we wish to make effective use of PCA as a visualisation tool. We will also consider non-similarity based methods of clustering, methods that make kindred assumptions about the Gaussianity of input data and so a study of the data distribution will prove useful for later solutions sought. We find evidence that the log of data is reasonably Gaussian and subsequently that PCA marginally outperforms UMAP as a visualisation tool for determining the potential number of groups in a complex DNA mixture.

4.1 Distinguishing Genotypes using Similarity Measures

A Euclidean distance is appropriate for data measured on the same scale, for which magnitudes are comparable [92]. When we consider our data, it is quite likely that two distinct vectors have high values yet originate from different contributors. If a Euclidean distance is chosen, observations with high values will be clustered together and those with low values will be clustered separately thus incorrectly grouping single-cells by their magnitude rather than their genotype.

We want to identify clusters of single-cells with the same overall profiles irrespective of their magnitudes as is often the case in gene expression data analysis [92] and so we need to consider metrics which forgo magnitude altogether. Cosine similarity relates observations by measuring the cosine of the angle between two non-zero vectors projected into a n-dimensional space, thus ignoring any reliance on magnitude. Observed values may be far apart in terms of a Euclidean distance but they may have a small angle between them implying high similarity [52]. Vectors with the same orientation have a cosine similarity of 1 while two vectors with a perpendicular orientation have a cosine similarity of 0. We will establish a cosine metric based on this logic that equates to saying EPG-vectors originating from the same genotype will lie close to 0 whereas EPG-vectors form different genotypes will lie close to 1. This is a cosine dissimilarity measure such that the cosine metric $= 1-$ the cosine similarity.

We can deduce from Fig. 4.1 that a cosine metric is capable of distinguishing EPG-vectors originating from one genotype to another. We have chosen three genotypes at random and compared the distributions of the “Self-Self” pairwise distances between the EPGs of a single genotype and “Self-Non-Self” pairwise distances between the EPGs of two distinct genotypes. We see extreme overlap in distinguishing Self-Self from Self-Non-Self when using a Euclidean distance, peaks lie relatively close to each other which is likely to cause problems. Comparing this to the use of a cosine dissimilarity measure we get peaks close to 0 in Self-Self and close to 1 for Self-Non-Self indicating that these are distinguishable.

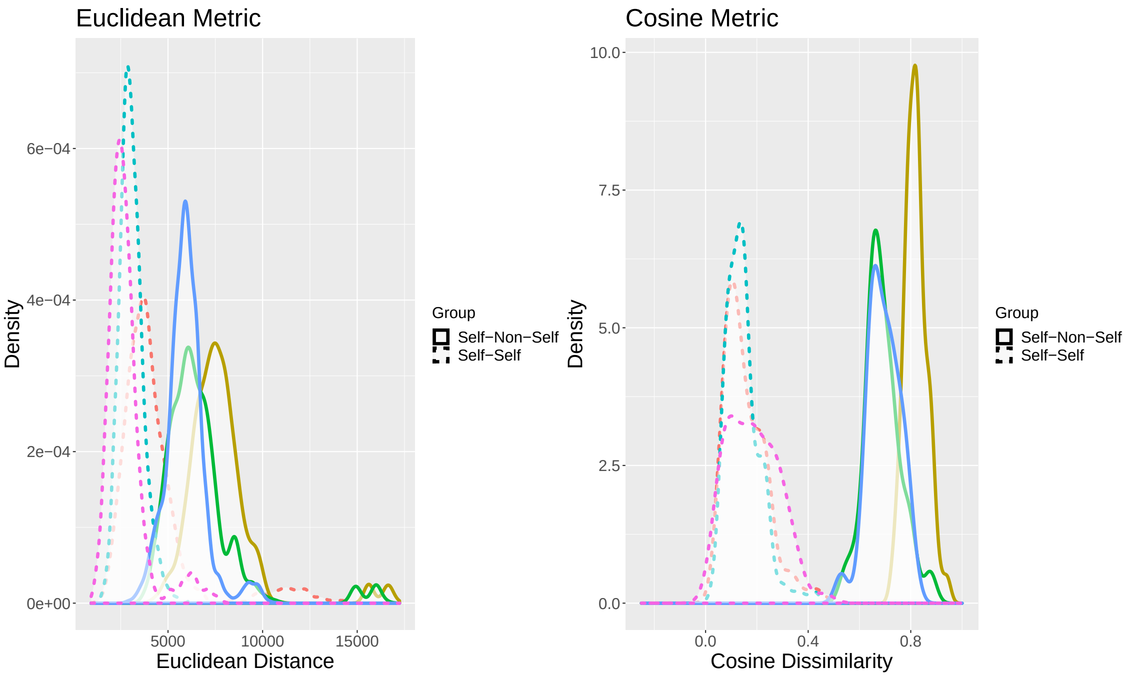


Figure 4.1: Distribution of the similarity measure between 60 pairs of single-cell EPG-vectors taken from the same genotype (Self-Self) or from distinct genotypes (Self-Non-Self). A high-pass filter as descibed in section 3.2 was applied before EPG-vector selection. We have used the Euclidean and the cosine metrics. We see much overlap in distinguishing Self-Self from Self-Non-Self when using a Euclidean distance, peaks lie relatively close to each other which is likely to cause problems. Comparing this to the use of a cosine dissimilarity measure, we get peaks close to 0 in Self-Self and close to 1 for Self-Non-Self, indicating that these are distinguishable.

4.2 Distribution of True Allele Peak Heights

Total EPG signal is dominated by true allele peak heights and so to determine which distribution best describes our data, we will focus on true allele signal. We consider both a Normal and a log-normal distribution. We compare these distributions on raw-signal recorded in RFU and on normalised-signal. We have normalised our fluorescences as follows:

$$\frac{f_i^{\mathrm{EPG}_k}}{I_k}, \tag{4.1}$$

where $f_i^{\mathrm{EPG}_k}$ are the fluorescences recorded for EPG_k, $i, k \in \mathbb{Z}^+$ and I_k is the intensity of EPG_k, see equation $\boxed{3.1}$.

We have shown log transformed true allele peak heights and log transformed normalised true allele peak heights are well described by the Normal distribution. The

Normal distribution provides good statistical fit where we transform our response variable (the signal) by taking the logarithm to the base 10 and find the best fit of the Normal distribution as indicated by the green dotted line seen in Fig. 4.2. We observe that this fit falls closely in line with our response variable, the red line, when compared to the best fit Normal of our raw-signal data or normalised-signal data, the blue dashed line.

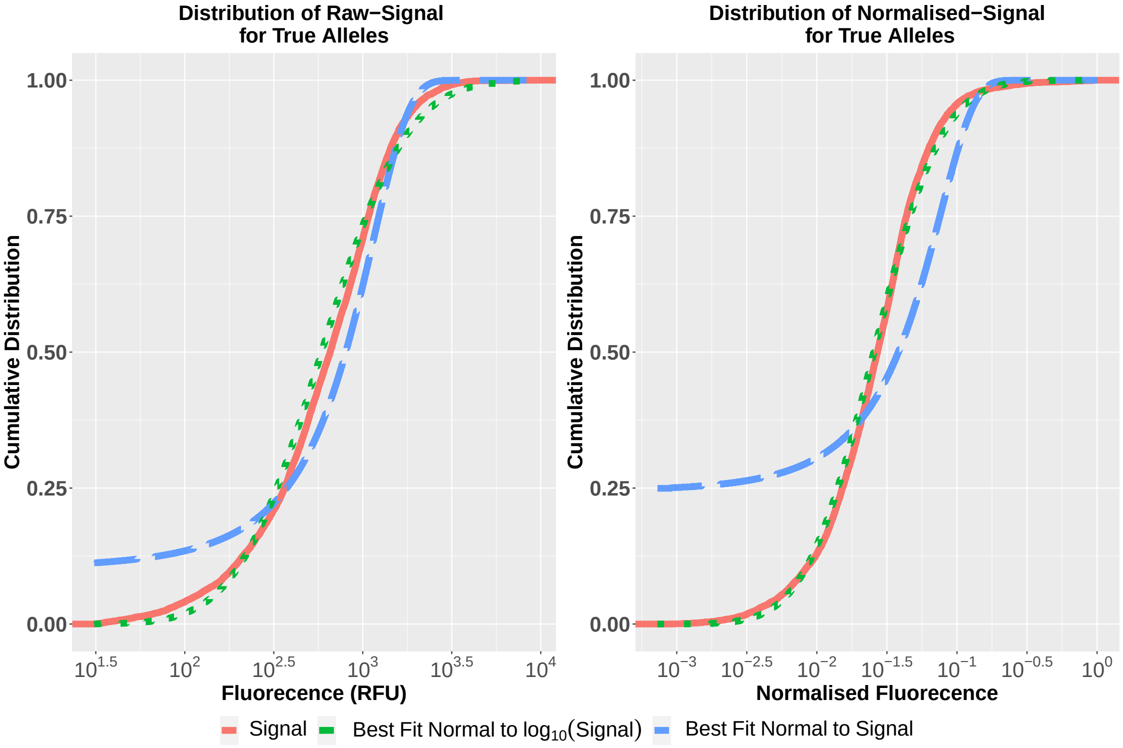


Figure 4.2: The Empirical cumulative distribution of true allele signal recorded (red) along with the best fit Normal to the signal (blue) and best fit Normal to the $\log_{10}$ of the signal (green). The parameters of both best fit models were determined using recorded true allele heights. The left plot is of the raw-signal while the right is of normalised-signal. The x-axis is taken logarithm base 10.

We have separated true alleles into one of three classes; i) Homozygous Alleles; ii) Heterozygous Non-Adjacent Alleles (alleles with a difference strictly greater than one); and iii) Heterozygous Adjacent Alleles (alleles with a difference of one). When considering alleles from class three, we have two sub-classes: Hetero-Adjacent Left and Hetero-Adjacent Right. We refer to the left as the smaller allelic variant of the two true alleles and the right as the larger allelic variant. In Fig. 4.3, we see the Normal distribution fitted to the $\log_{10}$ of the signal continues to show superiority over the Normal distribution fitted to the (untransformed) signal when we consider the distinct classes of true alleles present in our data, furthering the belief that log transformed signal is well described by the Normal distribution. As a result when using methods such as PCA or `mclust`, which assume that the data are normally distributed, we will take the logarithm base ten of our data (either raw or normalised) as the input.

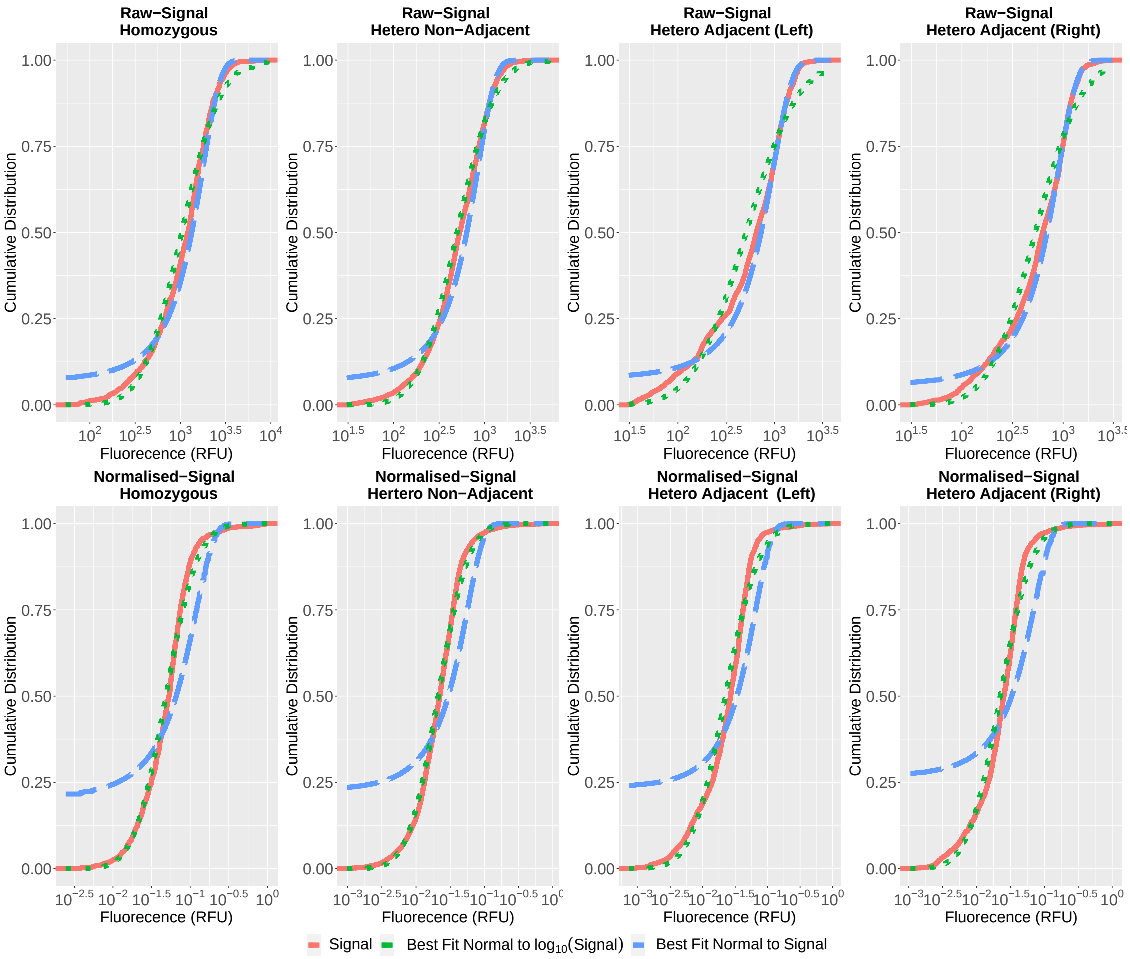


Figure 4.3: The Empirical cumulative distribution of true allele signal recorded, breaking into the four classes of true alleles for a closer examination of our claim that for normalised-signal a log-normal distribution best describes our data. Again, the x-axis is taken logarithm base 10. The top row of plots is of the raw-signal and the bottom row of plots is the corresponding normalised-signal.

4.3 Visualisation of High Dimensional Data

When working with high dimensional data one may wish to project it to look at it in a low dimensional space and informative way. There are traditional methods, widely used in genealogical and genome-wide association studies [85, 80, 37] including Principal Component Analysis (PCA) [97], Independent Component Analysis (ICA) [84] and more modern methods, particularly driven by single-cell RNA sequencing data [10, 71, 75], which has led to a range of methodologies such as Uniform Manifold Approximation and Projection (UMAP) [78] and t-Distributed Stochastic Neighbour Embedding (t-SNE) [113]. We have shown PCA and UMAP when projecting our data in a low dimensional space, however a preliminary examination with ICA and t-SNE was also carried out. We found when transformed, the data seems reasonably Gaussian resulting in ICA plots that are very similar to the PCA and again t-SNE showed similar results to UMAP (data not shown). As one would expect, given the logarithm of the data look Gaussian, PCA does by far in a way, the best with the logarithm of both raw-signal and normalised-signal but, quite interestingly it is evident that there is more information to be gleaned from the PCA than the UMAP particularly for imbalanced mixtures. There is something to be learned by applying the PCA dimensional reduction techniques on the raw data too as it becomes apparent that the distance from the origin in a PCA plot is a good surrogate for EPG intensity.

4.3.1 Principal Component Analysis (PCA)

Principal Component Analysis is one of the oldest and most widely used techniques for interpretable dimensionality reduction of large datasets [86, 58, 50]. It is a method that identifies a new basis (one that is orthogonal) on which to represent the original data. The new coordinate system is determined sequentially such that the first dimension or Principal Component (PC) describes the greatest variance in the data, the second PC is computed with the constraints of being orthogonal to the first PC and describes the second greatest variance in the data and so on. These new variables are found as uncorrelated linear combinations of the original data set and so, to retain as much of the original variance as possible. This reduces to either solving an eigenvalue/eigenvector problem or, alternatively obtaining the Singular Value Decomposition (SVD) of the (centered) data matrix [64].

Although PCA as a descriptive tool needs no distributional assumptions, for inferential purposes PCA usually assumes the mean and variance are sufficient statistics to entirely describe the probability distribution of the data and the only zero-mean probability distribution that is fully described by the variance is the Gaussian distribution [112, 64]. PCA has a nice interpretation if the data is Gaussian; larger variance corresponds directly to more variability. Seen in section 4.2, the log transform of our data is reasonably Gaussian and so, any meaningful PCA analysis is going to take the data and log transform it before use.

When considering the number of PCs returned, Jolliffe et. at. [64] gives a detailed explanation of the underlying mathematical theory behind this choice, however, suffice to say the number of PCs returned equates to the rank r of the original data matrix where in general, the rank of an $m \times n$ matrix is $r \leq min\{m, n\}$ (or $r \leq min\{m-1, n\}$ for column-centered matrices). Genomic data frequently presents datasets where there are fewer individuals than variables hence, the number of individuals often dictates r in these types of data. This is in part due to technological advances enabling the ease of observing variables coupled with the high expense of repeating observations (i.e. in the context of our work, it is costly to process a single-cell but once processed we can observe a great many alleles).

By using a limited number of principal components each admixture can be represented by relatively fewer variables instead of thousands. Admixtures can then be explored graphically on a PCA plot of the individuals making it possible to visually assess similarities and differences between observations within an admixture and determine which, if any, individuals can be grouped [2]. Specifically, EPGs from the same genetic source will have little variability in their detected alleles thus, we expect they will lie close together when compared with EPGs from a distinct source.

Applying PCA

PCA was applied to our simulated admixtures using the R function `prcomp` [96], centering but not scaling the data. We will focus our attention on an example admixture of three sources (Genotypes $05, 06$ and 07) in equal ratio to discuss the functionality of PCA and `prcomp`. We found 60 principal components, called PC$1-60$. Note that we have obtained a number of dimensions which corresponds to the number of observations (and not the number of variables). This is due to the fact there is only 60 non-zero eigenvalues, hence 60 PCs accounts for the total variability of the admixture. Table 4.1 shows the importance of our first 6 principal components. PC1 explains 41.86% of the total variance, which means two-fifths of the information in these new variables can be described by just this one principal component. PC2 explains 31.57% of the total variance and so with just the first two principal components almost three-quarters of the overall variance has been encapsulated.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard Deviation	0.04061	0.03527	0.01114	0.009593	0.008567	0.00843
Proportion of Variance	0.41855	0.31566	0.03150	0.023360	0.018620	0.01803
Cumulative Proportion	0.41855	0.73421	0.76571	0.789070	0.807690	0.82572

Table 4.1: Summary statistics of the first 6 principal components.

We have plotted the individuals using the first two principal components in Fig. 4.4 of this simulated admixture with Genotype 05 represented by circles, Genotype 06 represented by triangles and Genotype 07 represented by squares. This plot has been coloured by EPG intensity. Although somewhat indicated on this plot (mostly for Genotype 05), if we plot the corresponding raw-signal PCA we see for certainty a correlation between an EPGs intensity and its distance from the origin. We notice three "spokes" almost stemming from the origin, with low intensity EPGs closest to the origin and as we move away from the origin EPG intensity increases. Due to the three-spoke nature of this plot, the analyst determines there are three genotypes contributing to this admixture.

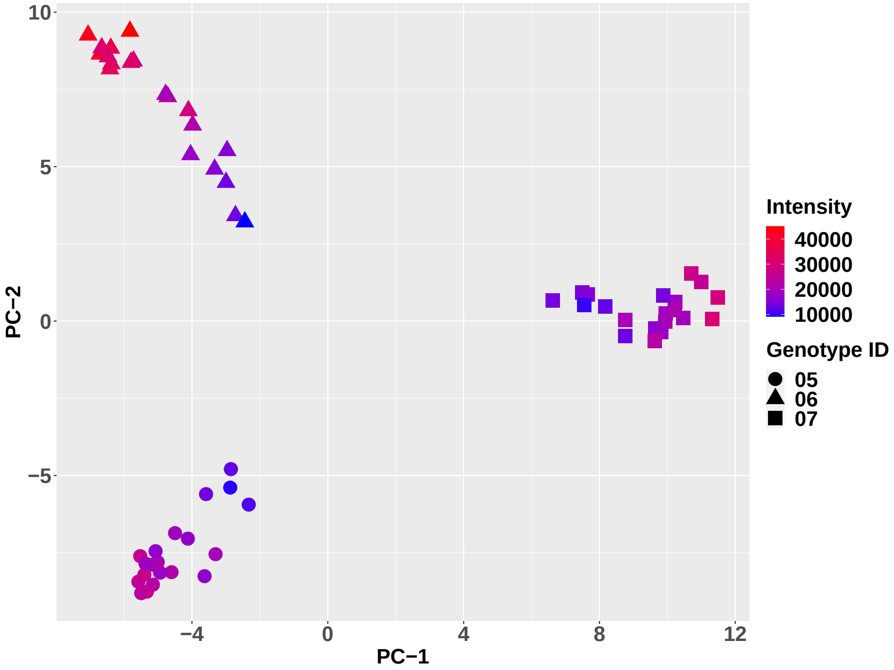


Figure 4.4: We have generated a PCA plot of the individuals for a three source admixture using the log transformed signal. Circles indicate Genotype 05, triangles Genotype 06, and squares Genotype 07. Individuals have been coloured by EPG intensity, I_k. We observe a three spoke nature indicating the presence of three genotypes in this admixture.

We observe the PCA plots of both the $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) present distinguishable clusters, particularly when multiple major contributors are involved in the admixture. As we can see in Fig. 4.5 when visualising an admixture with a single minor contributor, the $\log_{10}$(normalised-signal) presents a more prominent separation between clusters, conversely, when visualising an admixture with a single major contributor, all four genotypes are presented in clearly distinct clusters when plotting $\log_{10}$(raw-signal), where as Genotype 05 and Genotype 07 lie a little too close for distinction when plotting $\log_{10}$(normalised-signal). However, we note this observation is just as frequently seen in reverse. (See Appendix A, Fig. A.3 for examples of the converse). We note that for admixtures of four or five contributors, it is equally likely that $\log_{10}$(raw-

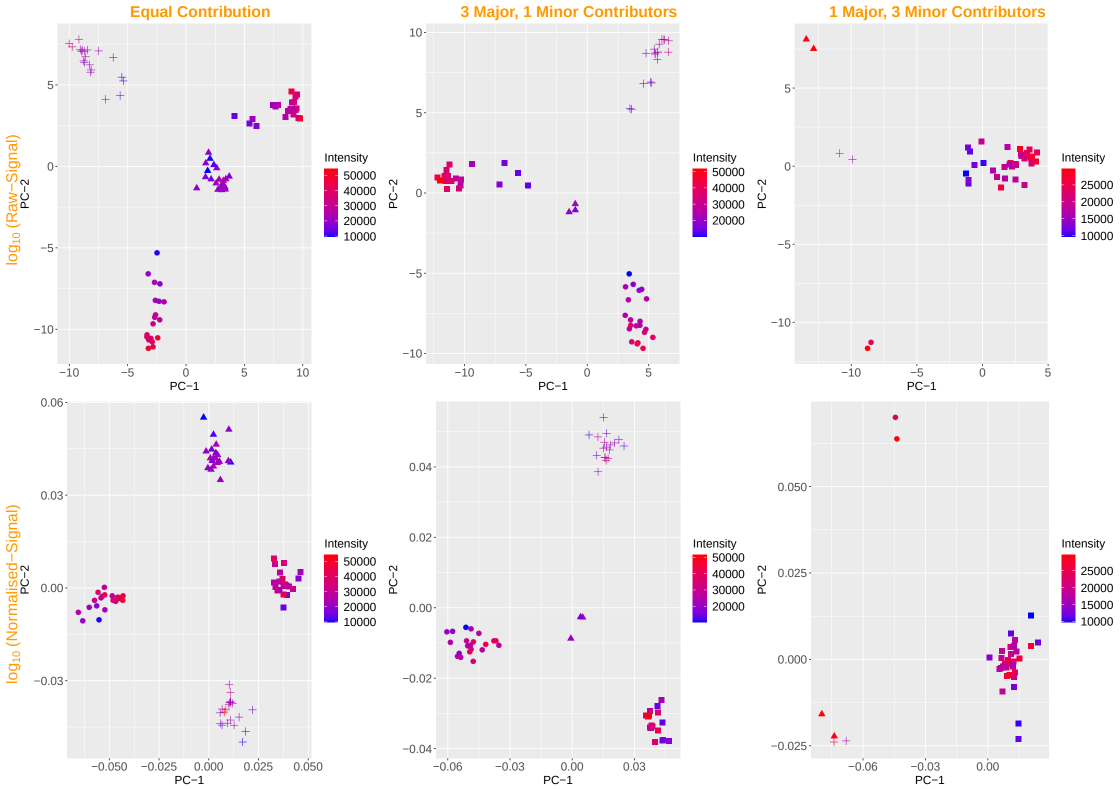


Figure 4.5: PCA Plots for simulated admixtures of 4 contributors in various mixture ratios. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06, and crosses are Genotype 07. Individuals have been coloured by EPG intensity, I_k. The first row of PCA plots have been generated using log transformed data. The second row of PCA plots, representing the same admixtures as above, are resultant of data that has undergone two transformation, first signal has been normalised, second the logarithm of normalised-signal has been taken. Each column of plots corresponds to a type of mixture ratio from equal contribution through to highly imbalanced. Four distinct groups can be determined for each mixture ratio when using the log transformed data. When plotting the PCA of the log transformed, normalised data for the highly imbalanced admixture one may incorrectly infer the number of contributors.

signal) or $\log_{10}$(normalised-signal) will present a more appropriate visualisation and so, it is our recommendation that the analyst should plot both and choose the larger number of potential contributors.

4.3.2 Uniform Manifold Approximation and Projection (UMAP)

One of the most recent techniques for both understanding and visualising large, high dimensional datasets is Uniform Manifold Approximation and Projection (UMAP), published in 2018 by McInnes et. al. [78]. In its simplest sense UMAP constructs a high dimensional graph representation of the data, then it optimises a low dimensional graph to be as structurally similar as possible [10]. UMAPs strong theoretical foundations allow the algorithm to strike a balance between emphasising local versus global structures. UMAP has rapidly been adopted by the population genetics community thanks to its non-linear neighbour graph-based dimensionality reduction. Among the many dissimilarities between PCA and UMAP there are two that we would like to highlight: i) UMAP does not make any assumption about the distribution of the data, so we need not transform our data when using UMAP; and ii) UMAP does not have a straight forward interpretation of distance once projected into a low-dimensional space, we gain no meaningful insight into the distance between observations in the high dimensional space when observing the distance between observations once projected. This second point is due to the fact that the UMAP algorithm focuses on preserving neighbourhood topology rather than absolute distance [33].

Applying UMAP

UMAP was applied to our simulated admixtures using the R package `umap` [69]. To first discuss the basic functionality of UMAP we will visualise the same example admixture seen in Fig. 4.4 using most of the default settings of UMAP. UMAP offers an array of metrics in R and for consistency with our earlier observation on its suitability as a measure of similarity of EPGs, we chose the cosine metric. UMAP has quite a few default configuration parameters. For instance, the parameter `n_components` determines how many dimensions `umap` returns with a default of two. As we intend to project our data into a two-dimensional space for visualisation ease, we will not alter the default option but we will consider altering the more commonly used parameters, `n_neighbors` and `min_dist`.

The number of approximate nearest neighbours used to construct the initial high dimensional graph corresponds to the `n_neighbor` parameter. In practice this parameter effectively controls how `umap` balances local and global structures. Low values will push more focus on the local structure while higher values will push the focus to the global structure. The default for `n_neighbors` is 15. The `min_dist` parameter controls how tightly `umap` “clumps” points together in the low dimensional graph with low values yielding tightly packed clusters and high values, looser clusters [10] with a default of 0.1. Fig. 4.6 is a plot of the observations, coloured by EPG intensity, in the new UMAP coordinate system which have been determined using a cosine metric and all other settings remain as the default. This time instead of a spoke shape, we see three distinct and tightly packed clusters but here the distance from the origin bears no significance to the intensity of an EPG, nor can we infer anything about the distance from one cluster to another.

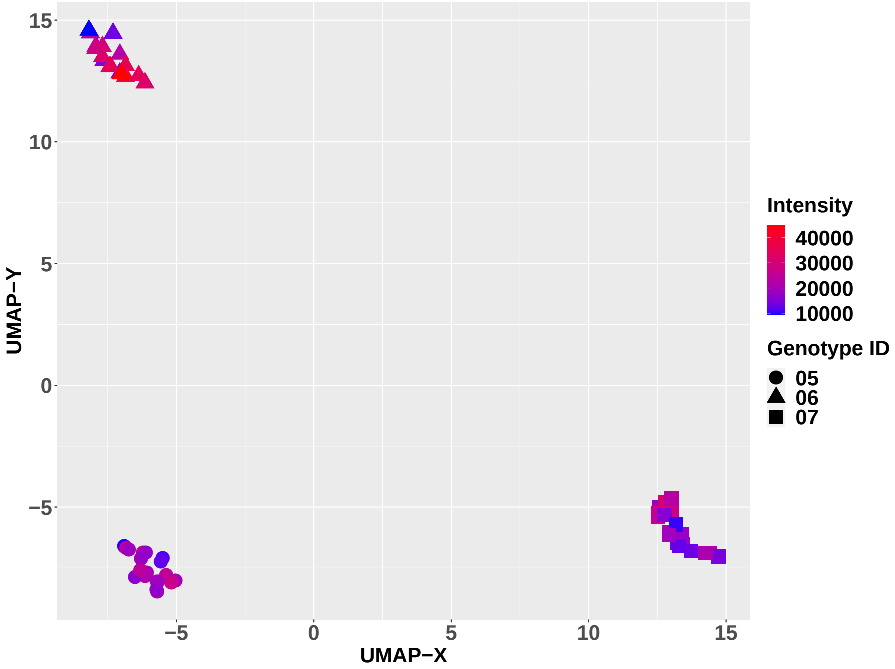


Figure 4.6: UMAP with default configuration settings and a cosine metric for a three source admixture, coloured by EPG intensity, I_k. UMAP has been run on the raw-signal.

Reporting the visualisation of the admixtures in Fig. 4.5, but now using UMAP (this time including the results when using raw-signal), we observe in Fig. 4.7 that UMAP presents similar visual results, with only small variations in the plot orientation for all data types (raw and transformed). This is largely due to the fact that UMAP makes no assumption regarding the distribution of the data. We also see that for admixtures where there is a single minor contributor, `umap` when used with default settings cannot separate this genotype from a major contributor. Similarly, when considering admixtures with a single major contributor, the analyst would not be able to identify any of the minor contributors due to the excessively loose cluster of the major contributor. To improve the use of UMAP as a visualisation tool we consider optimising the parameters, `n_neighbors` and `min_dist`.

First we alter the `n_neighbors` parameter, deciding to focus on smaller values as our aim is to improve the local structure. We evaluate the effects of choosing 2, 3 or 4 neighbours for admixtures in equal ratio, multiple major and one minor contributors

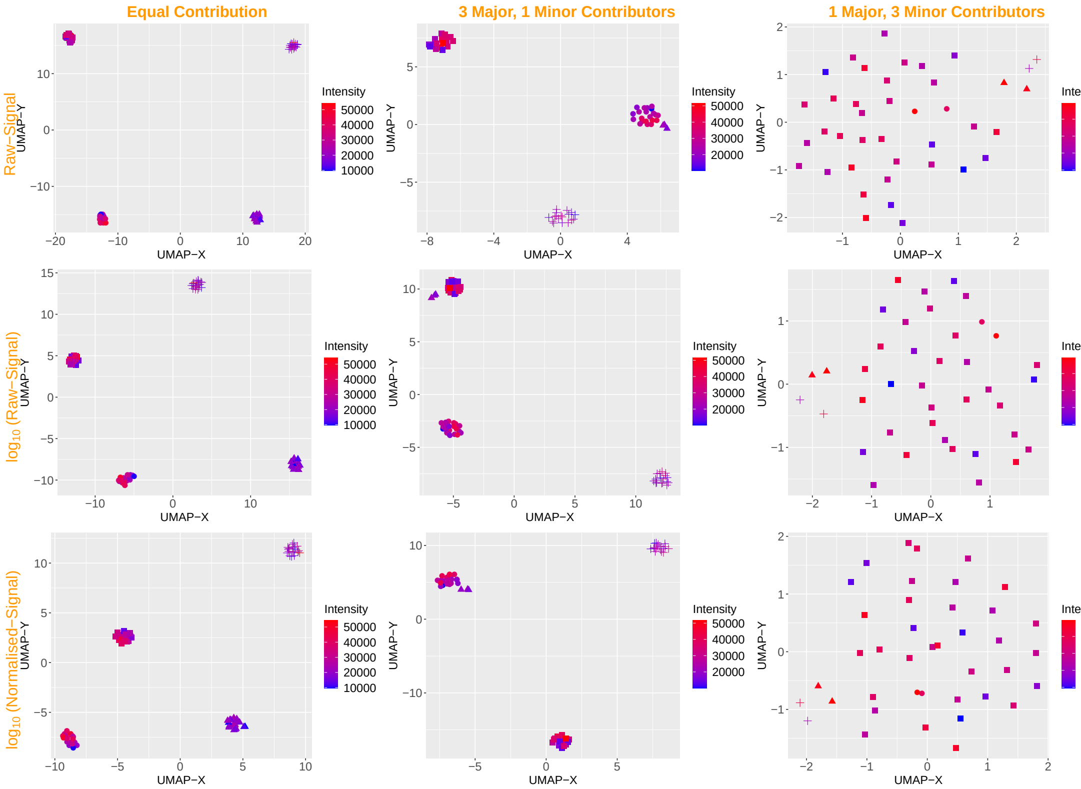


Figure 4.7: Corresponding UMAP plots for simulated admixtures in Fig. 4.5 of 4 contributors in various mixture ratios. These plots have been generated using a cosine metric and the default settings of `n_neighbors` = 15 and `min_dist` = 0.1. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06 and crosses are Genotype 07. Data types are grouped by each row of plots while mixture ratios are grouped by each column of plots. As we can see there is no distinguishable improvement on the quality of a UMAP plot when we plot raw-signal, the log of the signal or the log of the normalised-signal. Only the orientation of the clusters has varied. We notice with the default settings, UMAP performs sufficiently when in equal ratio but cannot identify a single or multiple number of minor contributors when in the presence of major contributors.

and one major multiple minor contributors. Fig. 4.8 shows the results of varying this parameter for a four person admixture. As we can see, two is too few and choosing three or four results in rather similar plots. Choosing `n_neighbors` = 3 (or 4) does not eliminate all problems but when handling an admixture with one minor contributor, we can now clearly identify all genotypes present. As a consequence however, when handling admixtures in equal ratio the clusters are no longer as tightly packed. We will next aim to empirically optimise the `min_dist` parameter which may remedy this problem. For admixtures with a single major and multiple minor contributors little improvement is observed. We also note that for two and three person admixtures no optimal number of neighbours could be found.

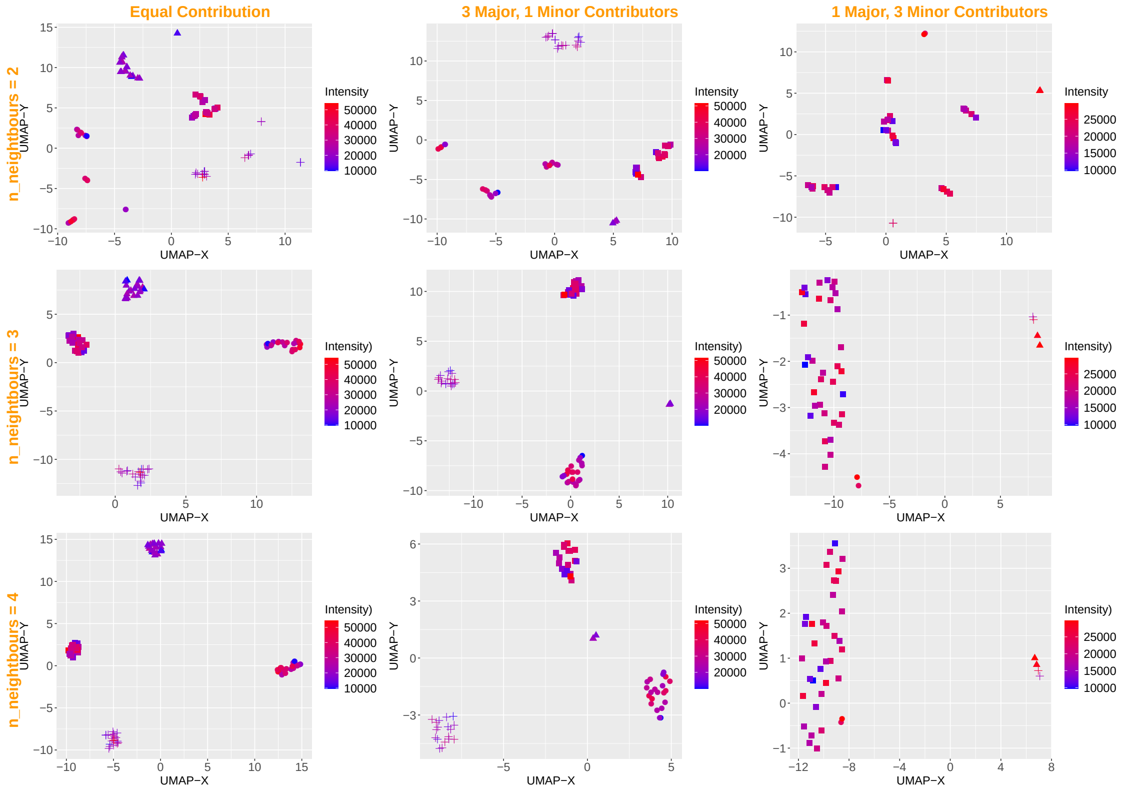


Figure 4.8: UMAP plots of 4 contributors in various mixture ratios with a cosine metric and the default setting `min_dist` $= 0.1$ but varying `n_neighbors` parameter from $2 - 4$. The plot rows indicate which `n_neighbors` parameter has been used. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06 and crosses are Genotype 07. `n_neighbors` $= 3$ and 4 present similar results. We see improved the visualisations of admixtures containing a single minor contributor. For a highly imbalanced mixture there is some improvement but one of the minor contributors is still completely masked by the major contributor.

Altering the `min_dist` parameter offers little remedy to the problem. To discuss the effects we have included a four person admixture example, Fig. 4.9. We considered decreasing the minimum distance close to zero (`min_dist` $= 0.00001$) with the intention of tightening the individual clusters when in equal ratio and highly imbalanced mixture ratio. However, this instead resulted in a slight separation of Genotype 02 as seen in the first plot of Fig. 4.9 and lessened the distinction between Genotype 05 and 07 in the highly imbalanced mixture (two of the minor contributors are now almost completely over-laid). We also considered increasing the minimum distance (`min_dist` $= 0.5$) with the intention of increasing the separation between the multiple minor contributors with no success as seen in the last plot of Fig. 4.9. In fact, regardless of the `min_dist` parameter, for highly imbalanced mixtures (column 3) minor contributors are indistinguishable from one another or the major contributor. We conclude, for the resulting UMAP visualisations to be as effective as our PCA plots, one would be required to modify potentially many more of UMAPs parameters without the promise of success. Therefore, we terminate our study of UMAP and regard PCA as the preferable of the two methods for visualising an admixture with the intention of determining the number of contributors present in the sample.

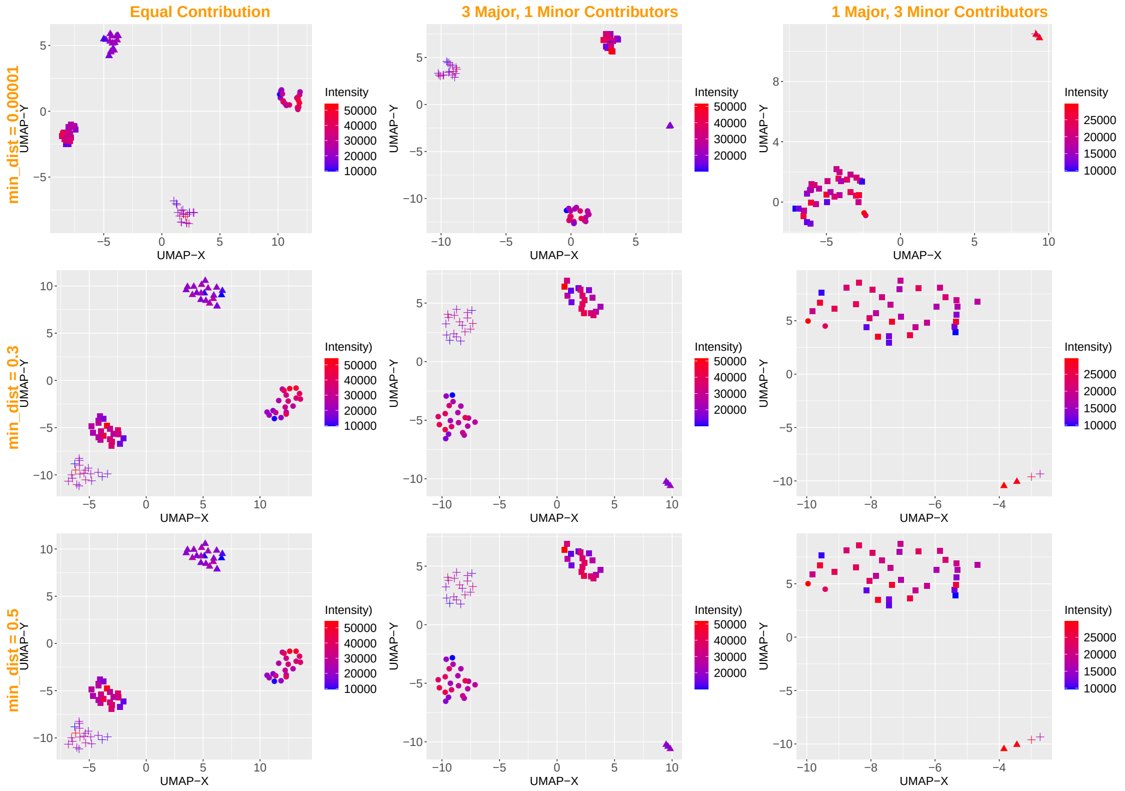


Figure 4.9: UMAP plots of 4 contributors in various mixture ratios with a cosine metric and `n_neighbors` = 3 but varying the `min_dist` parameter using 0.00001, 0.3, and 0.5. The plot rows indicate which `min_dist` parameter has been used. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06 and crosses are Genotype 07. An optimal `min_dist` that satisfies all mixture ratios cannot be recommended as minor contributors indistinguishable from one another when dealing with highly imbalanced mixtures.

5

Clustering Single-Cell Electropherograms

Satisfied that we can distinguish single-cell EPGs as seen in section 4.1, we now wish to cluster these single-cell electropherograms. We first considered traditional clustering methods that assume one has prior knowledge of how many groups/clusters they are looking for which is equivalent to knowing the true NoC in advance. This requires a two step process where first one must visualise the data with the aim of determining a potential number of clusters, followed by using a similarity based clustering algorithm to then do the clustering. Due to this first step we regard our traditional methods as the "Analyst-in-the-loop" solution. We then wished to automate this process, considering other methods that would simultaneously try to determine how many groups there are as well as group by membership which we regarded as a "computerised-end-to-end" solution.

5.1 Experimental Design

The likely use for single-cell methods in DNA forensics is when complex or LTDNA mixtures are anticipated. While not reported in this thesis, the Grgicak Lab. has made a range of low-template and complex true admixtures. We have mimicked these to evaluate the approaches considered in this thesis. We took real EPGs and simulated a range of low-template admixtures from low-template balanced with an increasing number of contributors through to highly imbalanced with an increasing number of contributors.

We simulate 11 types of admixtures (described in Table 5.1) as these are starting from simple through to highly complex. For each admixture, the complexity grows in two ways; one in the number of contributors and two in the imbalanced contribution ratio. For admixtures of three or more contributors we consider two types of imbalance, multiple major contributors and a single minor contributor (imbalanced) or a single major contributor and multiple minor contributors (highly imbalanced). We simulate 300 admixtures for each of the 11 types where ground truth is always known. We then run all admixtures through our respective clustering methods working strictly with our EPG-vectors.

NoC	Mixture Ratio	Num. EPGs	Shorthand
2	1:1	$\langle 20, 20\rangle$	N2R1
2	1:19	$\langle 2, 37\rangle$	N2R2
3	1:1:1	$\langle 20, 20, 20\rangle$	N3R1
3	1:9:10	$\langle 2, 18, 20\rangle$	N3R2
3	1:1:18	$\langle 2, 2, 36\rangle$	N3R3
4	1:1:1:1	$\langle 20, 20, 20, 20\rangle$	N4R1
4	1:6:6:7	$\langle 3, 18, 18, 21\rangle$	N4R2
4	1:1:1:17	$\langle 2, 2, 2, 34\rangle$	N4R3
5	1:1:1:1:1	$\langle 20, 20, 20, 20, 20\rangle$	N5R1
5	1:4:4:4:5	$\langle 4, 16, 20, 20, 20\rangle$	N5R2
5	1:1:1:1:16	$\langle 2, 2, 2, 2, 32\rangle$	N5R3

Table 5.1: Description of the 11 types of admixtures that have been simulated. When reading the shorthand NX for $X \in \{2, 3, 4, 5\}$ indicates the number of contributors found in the mixture. RY, for $Y \in \{1, 2, 3\}$ indicates the mixture ratio type. R1 = a balanced admixture, R2 = an imbalanced admixture (multiple major and a single minor contributor) and R3 = a highly imbalanced admixture (multiple minor, a single major contributor).

5.2 Analyst-in-the-loop Clustering of Single-Cell EPGs

Cluster analysis is the formal study of methods and algorithms for grouping or clustering, objects according to measured or perceived intrinsic characteristics or similarity [62]. Cluster analysis does not use categorical labels that tag objects with prior identifiers. The lack of categorical information is what distinguishes data clustering (unsupervised learning) from classification/discriminant analysis (supervised learning). Clustering has a long and rich history in a variety of scientific fields as it aims to find structure in data. We aim to cluster single-cell samples such that they are grouped according to their genotype.

K-means is one of the most popular and simple clustering algorithms, it was first published over 50 years ago [76], and despite countless clustering algorithms being published since, K-means is still widely used today. K-means clustering is a simple yet elegant approach of partitioning a data set into k distinct, non-overlapping clusters. To perform K-means clustering we must first specify the desired number of clusters k (Step 1) and then employing the `amap` R package, the K-means (`Kmeans`) algorithm will assign each observation, single-cell EPG-vectors in our case, to exactly one of the k clusters (Step 2).

K-means is an iterative algorithm which assigns data points to a cluster such that the sum of squares distance between the data points and a centroid is at a minimum [62]. The less variation within a cluster, the more similar the data points are within the same cluster. Traditionally this sum of square distance is established using a Euclidean distance but as we have seen in section 4.1 a Euclidean metric is inappropriate for distinguishing EPGs from distinct genotypes. As a result, when employing K-means we will choose to use a cosine similarity measure as we have previously determined this is an effective metric in distinguishing EPGs.

Step 1: Determine K

To determine an appropriate tool for visualising our data, an analysis of dimensionality reduction techniques was carried out in section 4.3, concluding in the use of PCA. Specifically, we suggest that the analyst should establish a PCA plot of both $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) and choose the larger number of potential contributors as their k value.

Step 2: K-means Clustering

The analyst runs K-means using the k chosen in step 1 and records the cluster assignment of each EPG. The result of this clustering can be visualised on the same PCA plot as before. Fig. 5.1 shows the clustering assignment of K-means when run on the raw-signal (and normalised-signal for comparison) of the three simulated admixtures from Fig. 4.5, in which the analyst observed four distinct clusters (i.e. $k = 4$). EPGs have now been coloured by their cluster assignment, and as before the genotype is indicated by shape. K-means has successfully grouped each EPG into their respective genotype for the balanced four person mixture. In this example, when faced with the imbalanced and highly imbalanced admixtures we see incorrect cluster assignment for both the raw and normalised signal. For corresponding two, three and five person admixtures see Appendix A, section A.5.

This particular example was selected to highlight the process of the analyst but it will not always be the case that the analyst will correctly assign k. Often times when dealing with admixtures where there are many contributors and/or major-minor contribution, the choice of k can become obscure as highlighted in Fig. 5.2, a four person imbalanced admixture. Distinct genotypes are indicated in Fig. 5.2 by the symbol shape as we have prior knowledge of ground truth. Without the genotype being thus indicated it is possible that an analyst could determine EPGs from Genotype 02 and Genotype 06 originate from the one genetic source as the minor contributor is completely masked by the major contributor in this visualisation. Similarly it can be the case that there exist a separation between EPGs originating from the same source, potentially leading to the case of the analyst inferring the NoC to be greater than ground truth. To quantitatively assess the impact of correct or incorrect NoC assignment, we perform an experiment where simulated admixtures are created and the analyst either infers TrueNoC/TrueNoC±1.

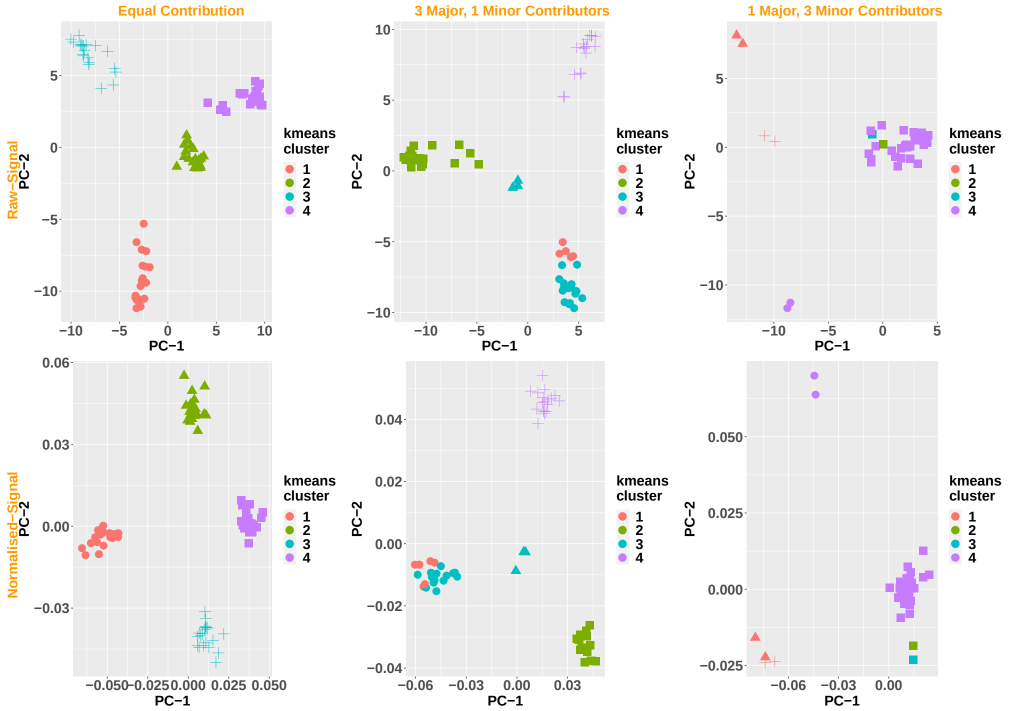


Figure 5.1: PCA plots coloured by K-means cluster assignment when choosing $k = 4$ for the simulated admixtures from Fig. 4.5. K-means was run on both raw-signal and normalised-signal however, we have plotted the $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) as we have previously observed log transformed data is preferential for low dimensional visualisation. K-means has correctly grouped EPGs by genotype for the balanced admixture.

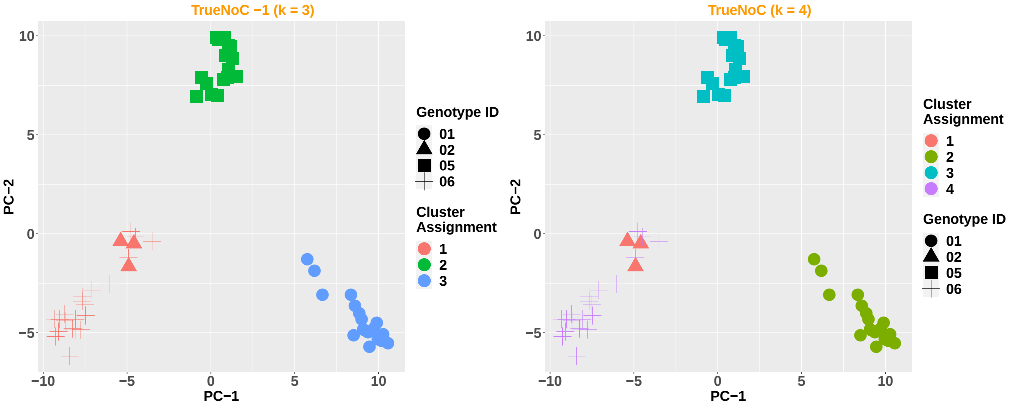


Figure 5.2: PCA plots coloured by the K-means cluster assignment when choosing $k = 3$ and 4 on a simulated admixture of four sources in an imbalanced mixture ratio. The plot has been generated using log10(raw-signal) while the K-means has been performed on the raw-signal. In the visualisation of this simulated admixture, the minor contributor (Genotype 02) is completely masked by one of the major contributors (Genotype 06) hence the analyst may infer $k = \text{TrueNoC}-1$.

5.3 Computerised-end-to-end Clustering of Single-Cell EPGs

We wish to consider the possibility of removing the analyst from the loop by investigating other methods of clustering that will simultaneously determine the number of clusters along with EPG cluster assignment. Finite mixture models have often been proposed and studied in the context of clustering [34, 30, 122] with more recent applications of such methodologies in the field of molecular biology, in particular, microarray and gene expression data [124, 79, 83]. Model-based clustering, also known as Mixture Models (MM), is a broad family of algorithms designed for modelling an unknown distribution as a mixture of distributions. The probability distribution of observed data is approximated by a statistical model and cluster analysis is performed by estimating the model parameters from the data where the parameters define clusters of similar observations [81]. Yeung et. al. [124] standardised the performance of model-based clustering on both synthetic and real gene expression data showing a key advantage of suggesting the number of clusters and an appropriate model when compared to leading heuristic clustering algorithms.

We adopt a model-based clustering method which considers the data as coming from a distribution that is mixture of two or more Gaussian distributions, employing the R package `mclust` [100]. This is a popular R package for model-based clustering, classification, and density estimation based on finite Gaussian mixture modelling. Each component k is modeled by a Gaussian distribution, characterised by the mean vector, μ_k, the covariance matrix, Σ_k, and an associated probability in the mixture (each observation has a probability of belonging to each cluster). These parameters are estimated using the Expectation and Maximisation (EM) algorithm and each cluster k is centered at μ_k, with increased density for points near the mean. The geometric features of each cluster, the shape, volume, and orientation, are determined by Σ_k [3]. In addition, the `mclust` package provides functions for performing the EM algorithm to different Gaussian mixture models, for simulating data as well as for visualising fitted models along with clustering, classification and density estimation results [101].

5.3.1 Running Mclust

The `Mclust` function assumes that in each cluster k the data follows a Gaussian distribution and so we compare the performance of `Mclust` on $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) as the log transformation of our data is well described by the Normal distribution (see in section 4.2). Fig. 5.3 is a visual representation of the EPG cluster assignment for the four person admixtures considered in Fig. 4.5. For corresponding two, three, and five person admixtures see Appendix A, section A.6. In this example, we observe that `Mclust` has correctly clustered EPGs by distinct genotype for both $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) however, we will come to see that `Mclust` run on the

$\log_{10}$(normalised-signal) consistently outperforms the alternatives.

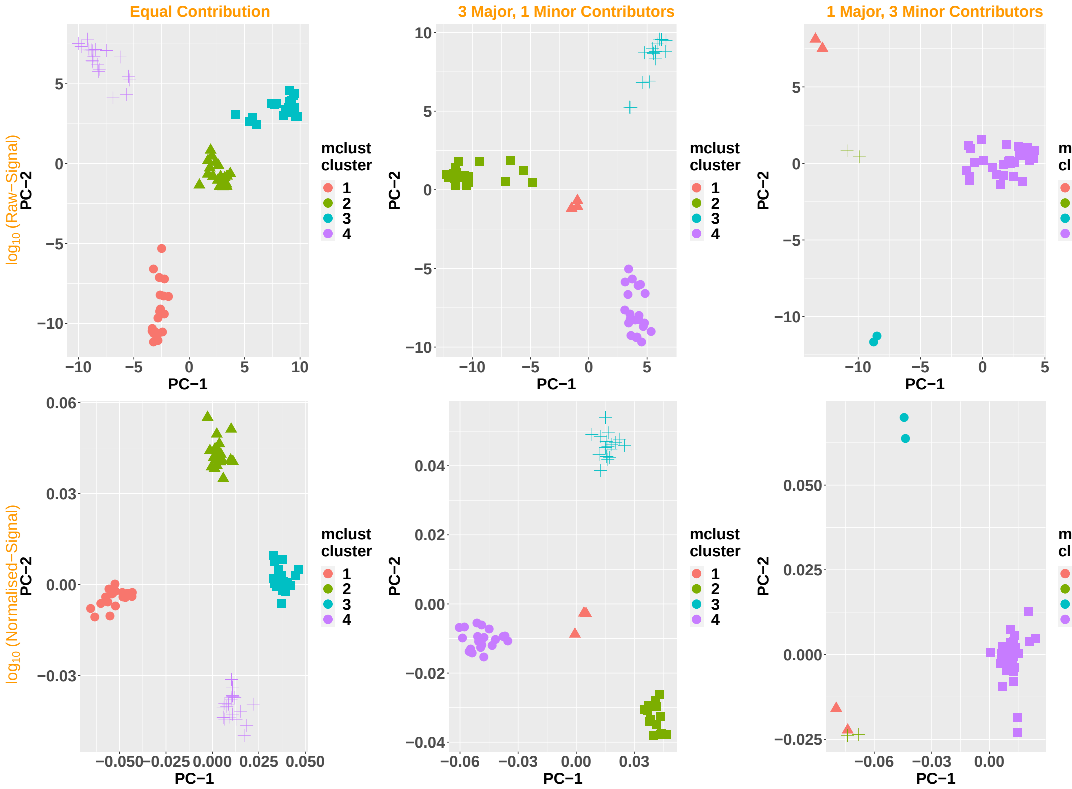


Figure 5.3: PCA plots coloured by the `Mclust` classification of EPGs for the simulated admixtures from Fig. 4.5. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06, and crosses are Genotype 07. `Mclust` has correctly clustered EPGs by distinct genotype for both $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) for all types of admixture.

To generate the results presented above in Fig. 5.3, we only provide the data to `Mclust`. Then, the optimal model is selected based in the Bayesian Information Criterion (BIC). Initially, many finite Gaussian mixture models are fitted with different numbers of clusters and covariance structures, from the simplest compound symmetric structure of the EII model, through to the complex unstructerd VVV model (A full description of models fitted can be found in Table 3 of [101]) and then, all models fitted are compared via BIC [101]. Hence, the final model is which has the optimal BIC given a specific number of clusters and covariance structure. Further, the idea of using BIC to select the best model is to balance the number of clusters and the increase in the log-likelihood function due to the addition of more components in the model. A plot of the BIC traces (see Fig. 5.4 for the twice transformed signal model selection plot) for all models considered is then obtained. There is a clear indication of a four-component mixture with covariances having spherical distributions and unequal shape and volume (VII). The VII and EII (spherical distribution with equal shape and volume) are most prevalent selections across all admixtures seen in Table 5.1

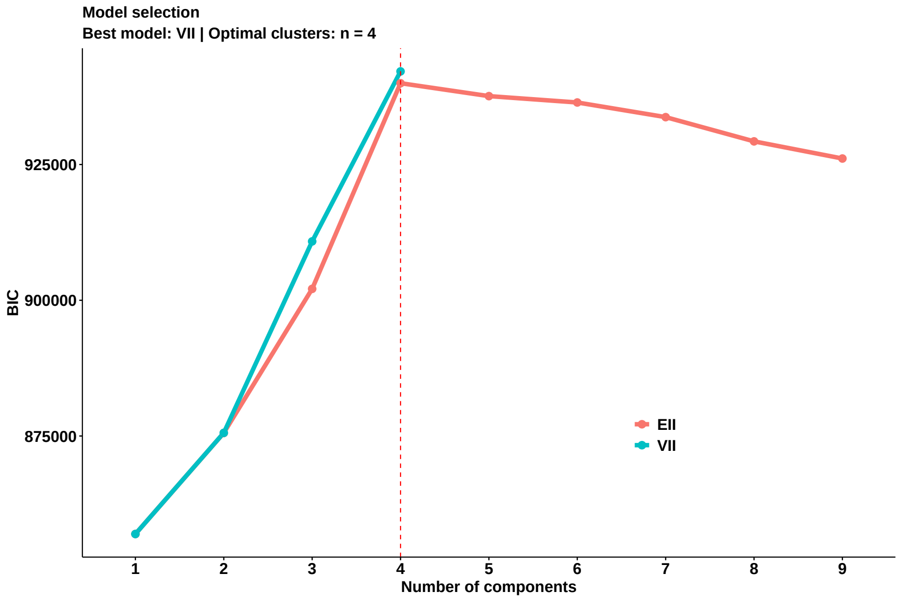


Figure 5.4: BIC plot for the model fitted to the log10(normalised-signal) for the four person admixture in equal contribution observed in Fig. 5.3. This plot shows both, the optimal number of clusters and which model has been selected. The VII model was selected. This is a Gaussian mixture model, which assumes a multivariate Normal distribution for each cluster, that has clusters with spherical shape and varying volume (see Table 3 and Fig. 2 of [101] for more details).

5.4 Results

When clustering, we encounter incorrect groupings and we first consider the source of error. We determine two causes, mis-clustering and over-clustering which are responsible for incorrect EPG classification. We then analyse the results of the analyst-in-the-loop clustering when we have conditionally assigned the correct number of groups and compare this with the computerised-end-to-end clustering of EPGs.

5.4.1 Source of Incorrect Grouping

When analysing our results for either K-means or `Mclust`, we must first define the types of errors we encounter. We observe two types of error:

(1) mis-clustering

(2) over-clustering

These are not orthogonal results both can, and do, occur in a clustering event.

Mis-clustering

We define mis-clustering as an incident were two or more distinct genotypes are found in one cluster (Fig. 5.5). We consider an incident of mis-clustering with greater concern as this can lead to an incorrect description of a genotype. If EPGs from two (or more) distinct genotypes are clustered together, this may lead to mistaken genetic identification in a downstream pipeline.

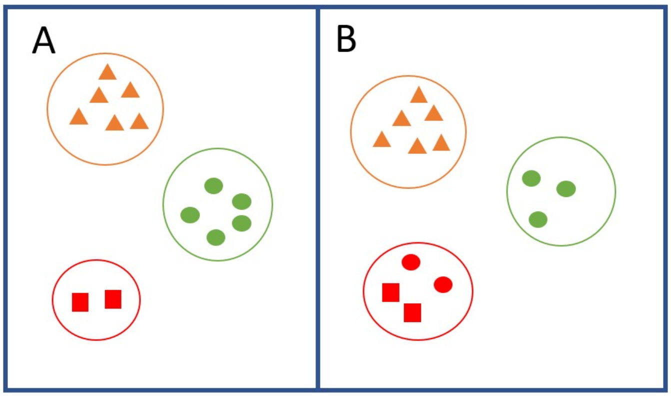


Figure 5.5: An example of mis-clustering. Shape corresponds to the genotype and the colour corresponds to the cluster assignment. (A) is the correct cluster assignment for each genotype. For the same collection of EPGs, (B) shows an example of mis-clustering. Squares and some circles have been grouped in one cluster, the red cluster.

Over-clustering

We define over-clustering as an incident where a single genotype has been grouped into two or more distinct clusters (Fig. 5.6). As an error we believe it will not have as significant an impact as mis-clustering in downstream interpretaion. At most, when there has been no mis-clustering, over-clustering will imply the presence of too many contributors.

We would like to note when interpreting incidents of over-clustering there is some overlap with incidents of mis-clustering. More precisely, should the example in Fig. 5.5 occur, this will also be recorded as an event of over-clustering since the genotype indicated by circles now appears in more than one distinct cluster. Both mis-clustering and true over-clustering (such as the case in Fig. 5.6) can occur in a single event and so we cannot separate simply the recordings of such incidents. We ask the reader to be mindful of this and be aware when interpreting results, if we see a high volume of mis-clustering, we will see similar or higher results for over-clustering. However, should we see a low volume of

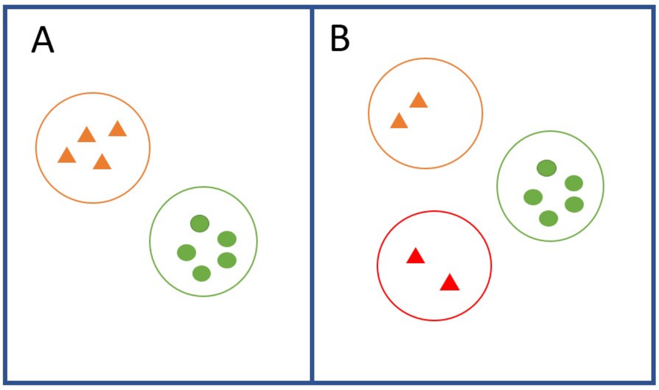


Figure 5.6: An example of over-clustering. Shape corresponds to the genotype and the colour corresponds to the cluster assignment. (A) is the correct cluster assignment for each genotype. For the same collection of EPGs, (B) shows an example of over-clustering. Triangles have been split into two distinct clusters, now the red and orange clusters.

mis-clustering but a high volume of over-clustering, then we can be assured the majority of this over-clustering is due to incidents of true over-clustering.

5.4.2 K-means Results

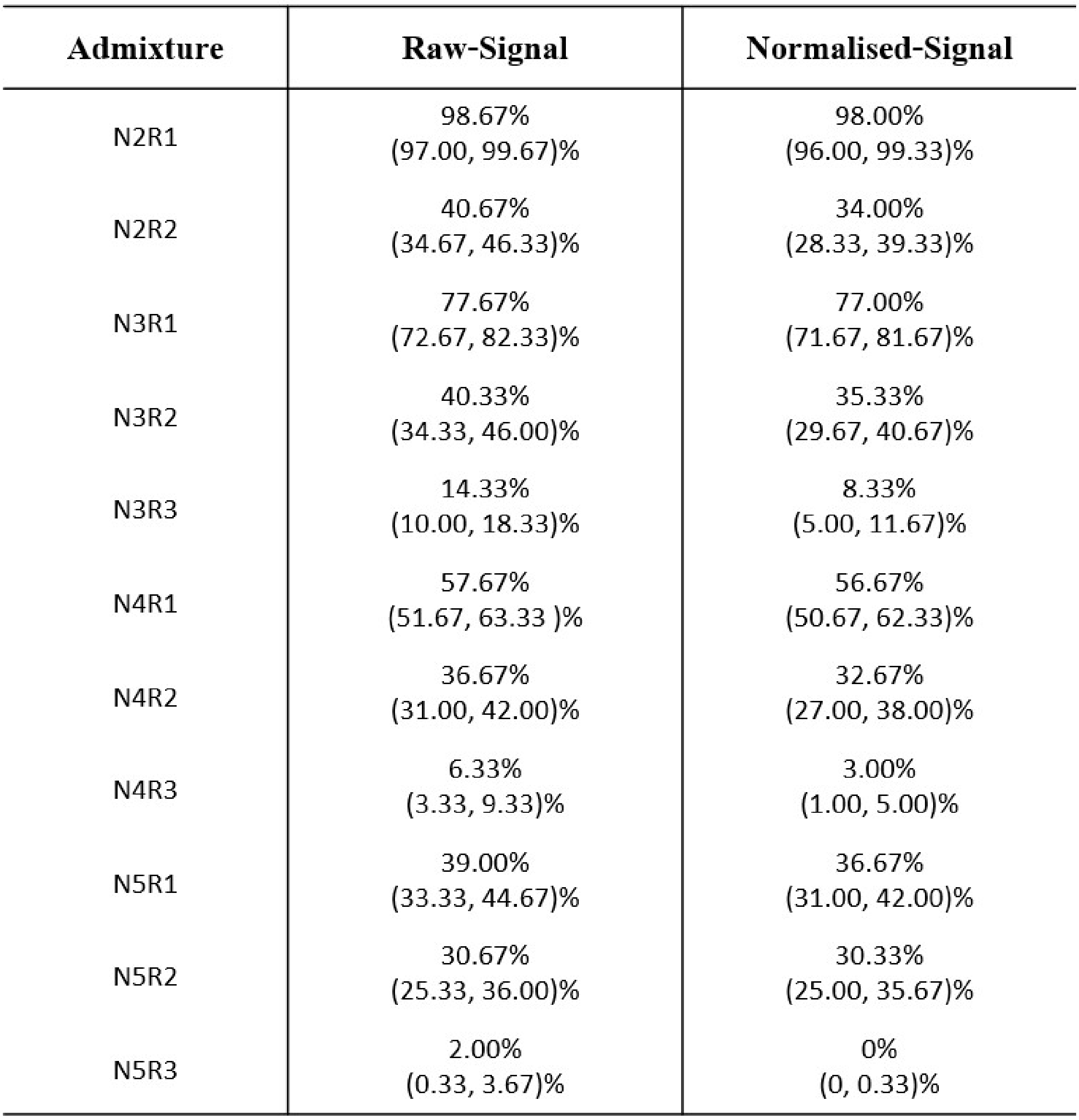
When performing K-means clustering we considered the situation where an analyst assigns the TrueNoC, TrueNoC−1 or TrueNoC+1. If the analyst has underestimated how many groups there are (TrueNoC−1), then it necessarily results in mis-clustering. If the analyst over estimates how many groups there are (TrueNoC+1), it does reduce the amount of mis-clustering that occurs but unsurprisingly the frequency of over-clustering undoubtedly increases. For this reason we have focused our attention on results where the analyst has correctly determined how many groups there are.

We have compared the results of running `Kmeans` on raw-signal and normalised-signal (see Table 5.2). Complexity can increase in one of two ways, the number of contributors increases or the mixture becomes more imbalance. K-means performs exceptionally well when faced with two person admixtures of equal contribution with a 98% success rate. However, the performance decrease acutely with every increase of complexity but most notably when faced with admixtures comprised of one major and multiple minor contributors, K-means correctly groups EPGs by genotype less than 15% of the time.

We consider which, if either error is most prevalent, see Table 5.4 and Table 5.3. We observe frequent mis-clustering, increasingly so with increasing complexity. When faced with admixtures of four or less contributors in equal mixture ratio, more than half of the time kmeans can successfully group EPGs by genotype, yet when faced with admixtures of five contributors in any mixture ratio, kmeans miss-clusters more than 60% of the time.

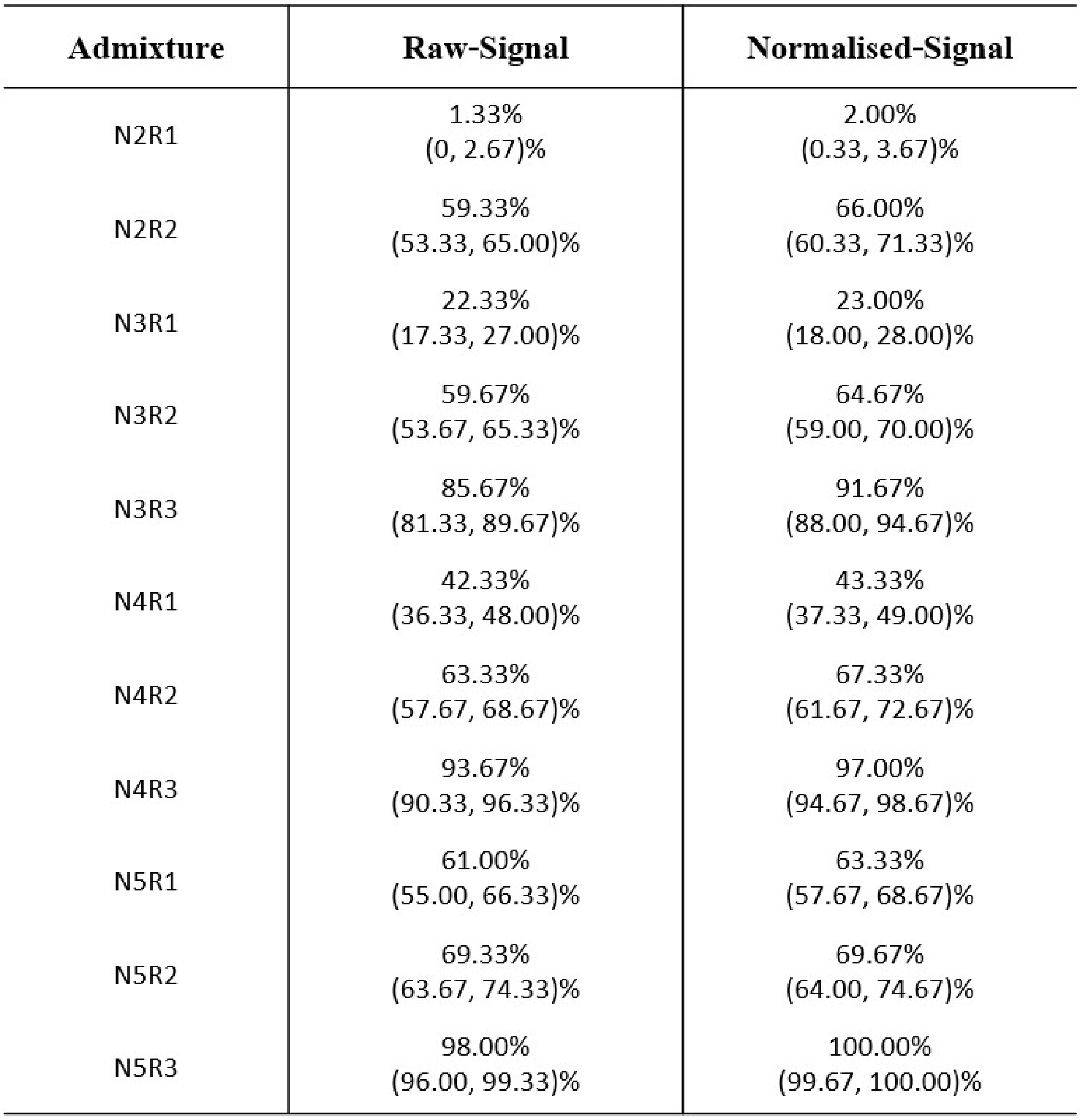
We have included two examples of the post analyst-in-the-loop processing. The first one is a balanced two person admixture that was successfully grouped by genotype and the second, a highly imbalanced five person admixture that includes an incident of mis-clustering. Fig. 5.7 is a plot of all EPGs involved in the first example over-laid. We have coloured EPGs by their cluster assignment which manifests the presence of two genotypes but without any colouring the interpretaion of such an EPG plot would be highly similar to that of a bulk processed sample. For a closer examination we can plot the over-laid EPGs of each individual cluster and have done so in Fig. 5.8 and Fig. 5.9. By experimental design, ground truth is always known so we have included this in our plots, indicated by the black vertical bars yet, even without these the analyst could almost read of the genotype of each cluster directly.

Fig. 5.10 is a plot of all EPGs involved in the the second example over-laid and again coloured by cluster assignment. One could almost certainly infer something about the genotype of the major contributor, however, interpretation regarding the minor contributors from this plot alone is challenging. By plotting the over-laid EPGs of the individual clusters, it becomes suspect that EPGs have been incorrectly grouped in cluster 1, Fig. 5.11 where we have identified genotype by shape and EPG by colour. We can see that cluster 1 is in fact an incident of mis-clustering as the EPGs of two minor contributors have been grouped together.



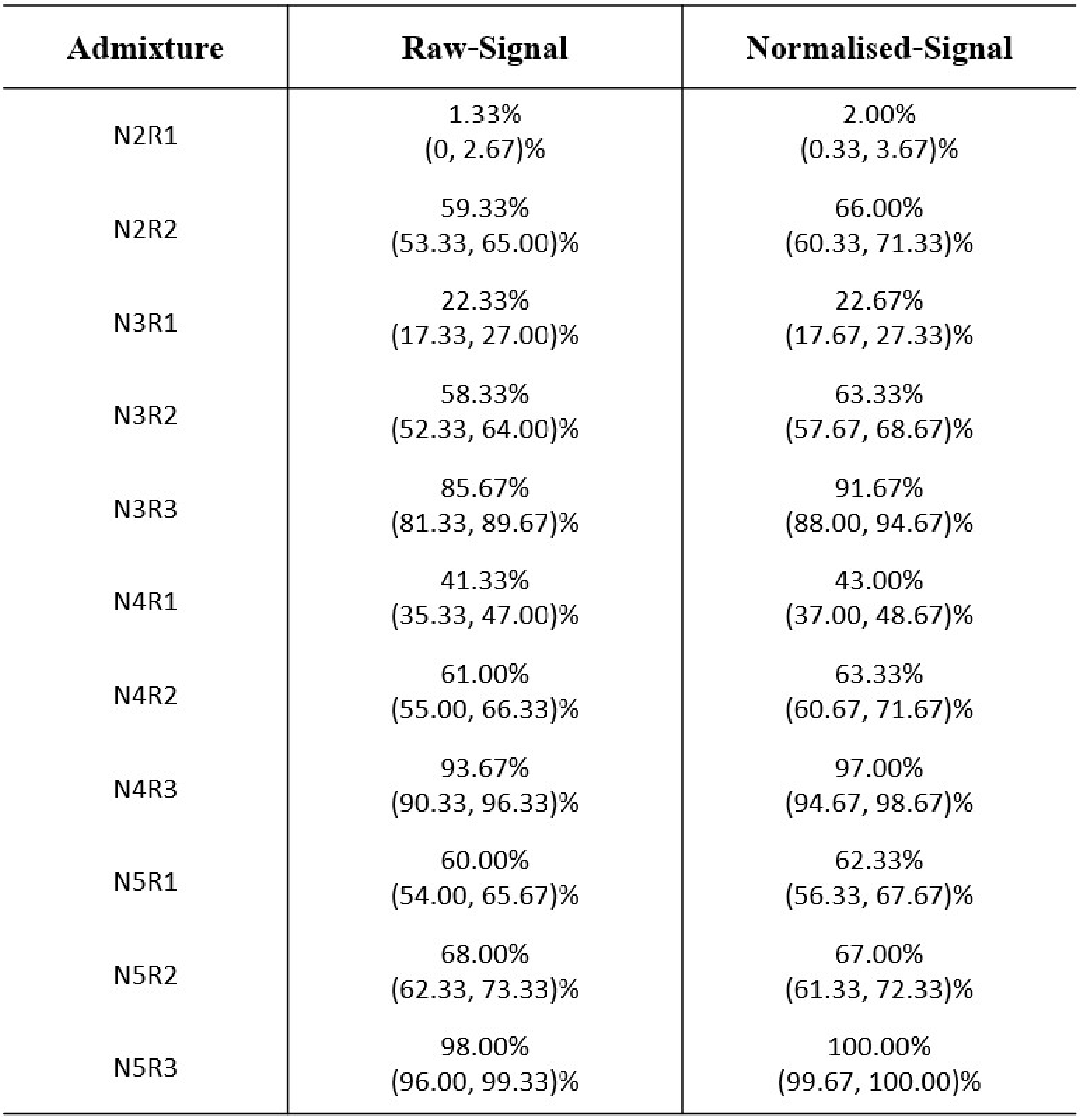
Admixture	Raw-Signal	Normalised-Signal
N2R1	98.67% (97.00, 99.67)%	98.00% (96.00, 99.33)%
N2R2	40.67% (34.67, 46.33)%	34.00% (28.33, 39.33)%
N3R1	77.67% (72.67, 82.33)%	77.00% (71.67, 81.67)%
N3R2	40.33% (34.33, 46.00)%	35.33% (29.67, 40.67)%
N3R3	14.33% (10.00, 18.33)%	8.33% (5.00, 11.67)%
N4R1	57.67% (51.67, 63.33)%	56.67% (50.67, 62.33)%
N4R2	36.67% (31.00, 42.00)%	32.67% (27.00, 38.00)%
N4R3	6.33% (3.33, 9.33)%	3.00% (1.00, 5.00)%
N5R1	39.00% (33.33, 44.67)%	36.67% (31.00, 42.00)%
N5R2	30.67% (25.33, 36.00)%	30.33% (25.00, 35.67)%
N5R3	2.00% (0.33, 3.67)%	0% (0, 0.33)%

Table 5.2: The percentage of Kmeans runs where no mis-clustering or over-clustering has occurred for raw and normalised signal with a 95% binomial confidence interval. In general, raw-signal outperforms normalised-signal. A full description of the admixtures can be found in Table 5.1.



Admixture	Raw-Signal	Normalised-Signal
N2R1	1.33% (0, 2.67)%	2.00% (0.33, 3.67)%
N2R2	59.33% (53.33, 65.00)%	66.00% (60.33, 71.33)%
N3R1	22.33% (17.33, 27.00)%	23.00% (18.00, 28.00)%
N3R2	59.67% (53.67, 65.33)%	64.67% (59.00, 70.00)%
N3R3	85.67% (81.33, 89.67)%	91.67% (88.00, 94.67)%
N4R1	42.33% (36.33, 48.00)%	43.33% (37.33, 49.00)%
N4R2	63.33% (57.67, 68.67)%	67.33% (61.67, 72.67)%
N4R3	93.67% (90.33, 96.33)%	97.00% (94.67, 98.67)%
N5R1	61.00% (55.00, 66.33)%	63.33% (57.67, 68.67)%
N5R2	69.33% (63.67, 74.33)%	69.67% (64.00, 74.67)%
N5R3	98.00% (96.00, 99.33)%	100.00% (99.67, 100.00)%

Table 5.3: The percentage of `Kmeans` runs where mis-clustering has occurred for raw and normalised signal with a 95% binomial confidence interval. We note that mis-clustering and over-clustering are not orthogonal events, both can occur in an individual run. We observe mis-clustering is the preponderant issue when clustering via K-means. A full description of the admixtures can be found in Table 5.1.



Admixture	Raw-Signal	Normalised-Signal
N2R1	1.33% (0, 2.67)%	2.00% (0.33, 3.67)%
N2R2	59.33% (53.33, 65.00)%	66.00% (60.33, 71.33)%
N3R1	22.33% (17.33, 27.00)%	22.67% (17.67, 27.33)%
N3R2	58.33% (52.33, 64.00)%	63.33% (57.67, 68.67)%
N3R3	85.67% (81.33, 89.67)%	91.67% (88.00, 94.67)%
N4R1	41.33% (35.33, 47.00)%	43.00% (37.00, 48.67)%
N4R2	61.00% (55.00, 66.33)%	63.33% (60.67, 71.67)%
N4R3	93.67% (90.33, 96.33)%	97.00% (94.67, 98.67)%
N5R1	60.00% (54.00, 65.67)%	62.33% (56.33, 67.67)%
N5R2	68.00% (62.33, 73.33)%	67.00% (61.33, 72.33)%
N5R3	98.00% (96.00, 99.33)%	100.00% (99.67, 100.00)%

Table 5.4: The percentage of `Kmeans` runs where over-clustering has occurred for raw and normalised signal with a 95% binomial confidence interval. We note that mis-clustering and over-clustering are not orthogonal events, both can occur in an individual run. A full description of the admixtures can be found in Table 5.1.

Example 1

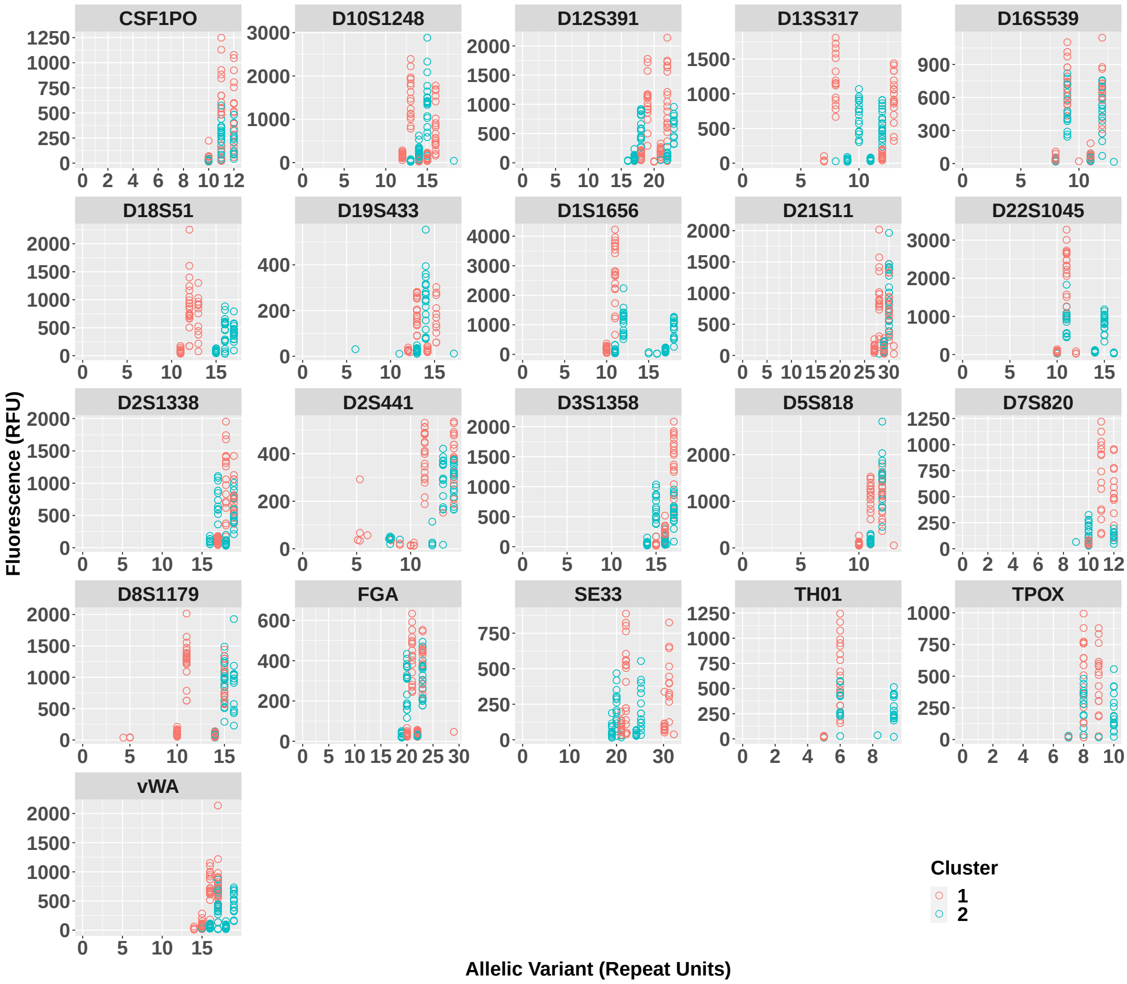


Figure 5.7: Plot of all EPGs present in a balanced admixture of two contributors (Genotype 01 and Genotype 05). EPGs have been coloured by their K-means cluster assignment. This is an example of a correct clustering result. Without colouring EPGs, interpretaion of such a plot would be similar to that of bulk processed sample but with colouring we can see an indication of two genotypes present.

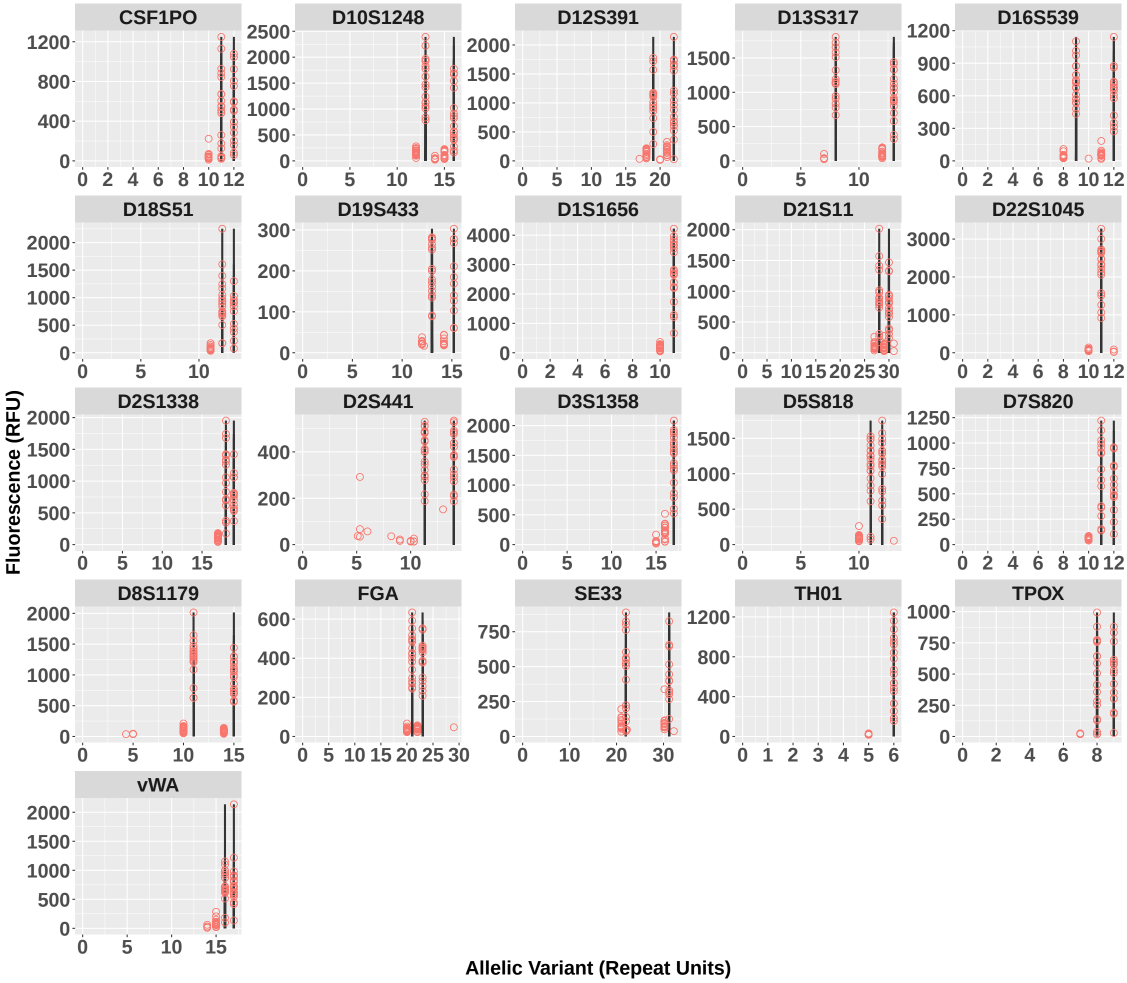


Figure 5.8: Plot of all EPGs assigned to cluster 1. As ground truth is known we have included the known genotype, indicated by the black vertical bars. By plotting cluster 1 on its own we can visually confirm that K-means has correctly assigned all EPGs from this genotype (Genotype 01) to a distinct cluster. One could almost read the genotype of this cluster directly from the plot without including ground truth.

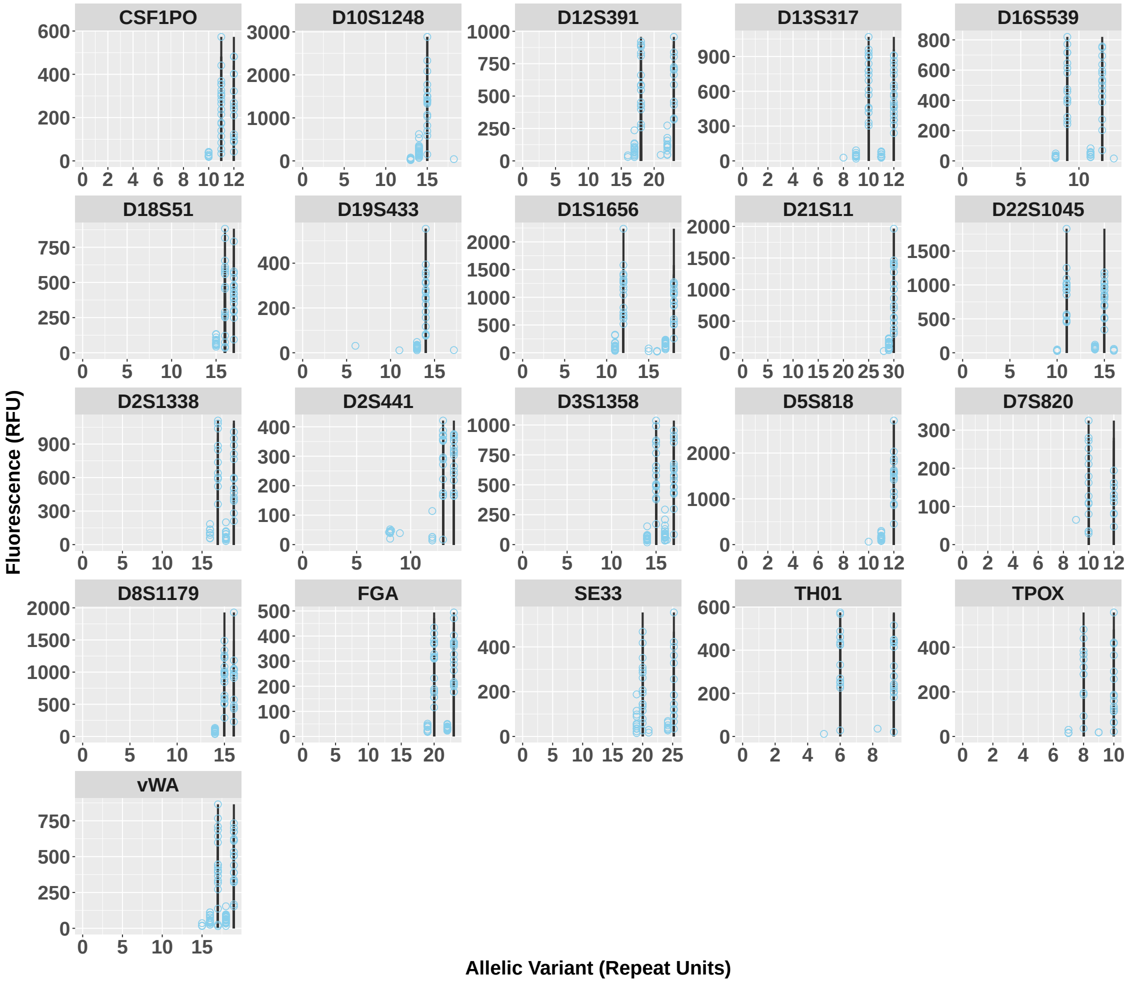


Figure 5.9: Plot of all EPGs assigned to cluster 2. As ground truth is known we have included the known genotype, indicated by the black vertical bars. Similarly, by plotting cluster 2 on its own we can visually confirm that K-means has correctly assigned all EPGs from this genotype (Genotype 05) to a distinct cluster. One could almost read the genotype of this cluster directly from the plot without including ground truth.

Example 2

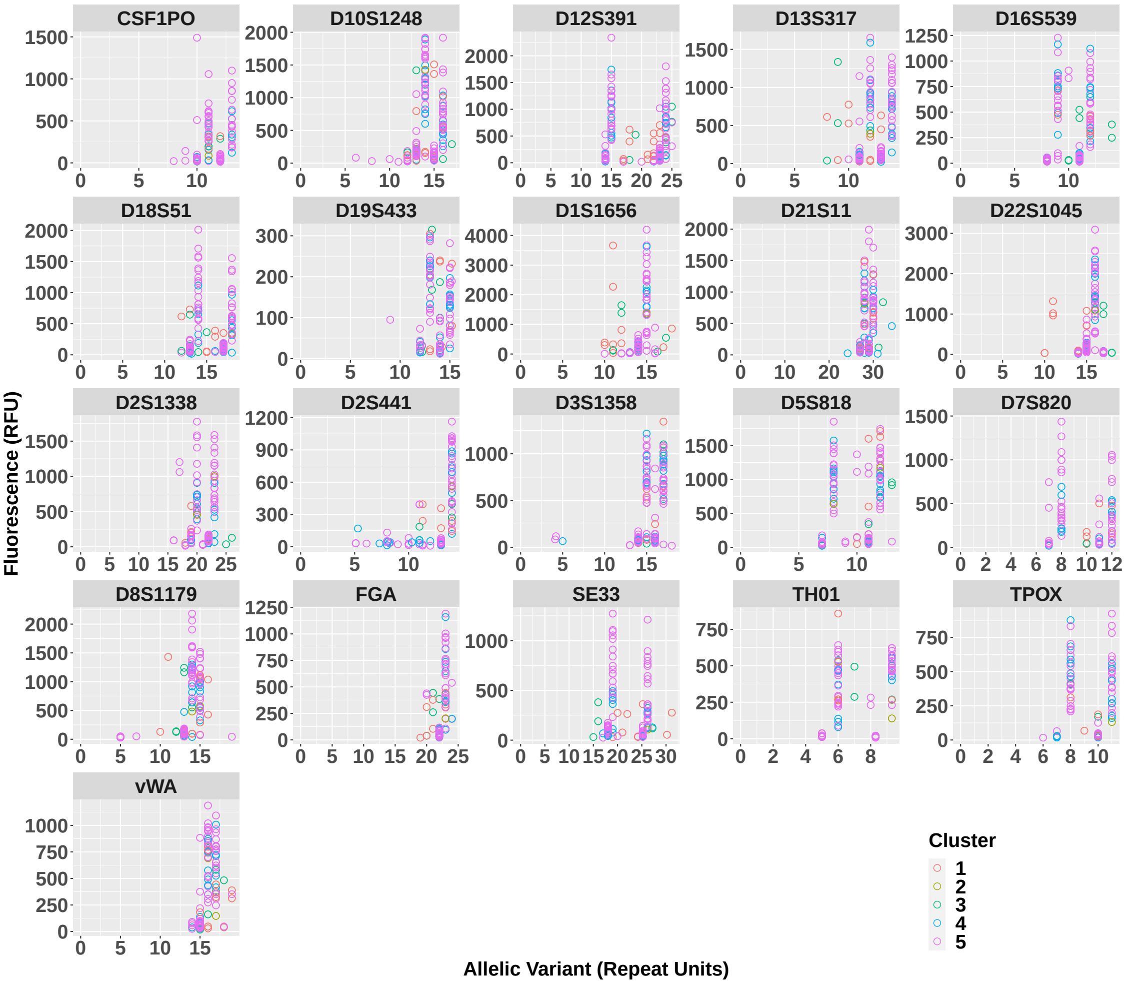


Figure 5.10: Plot of all EPGs present in a highly imbalanced admixture of five contributors. EPGs have been coloured by their K-means cluster assignment. This is an example where frequent mis-clustering occurred. With colouring, we can infer something about the major contributor, but little about the minor contributors.

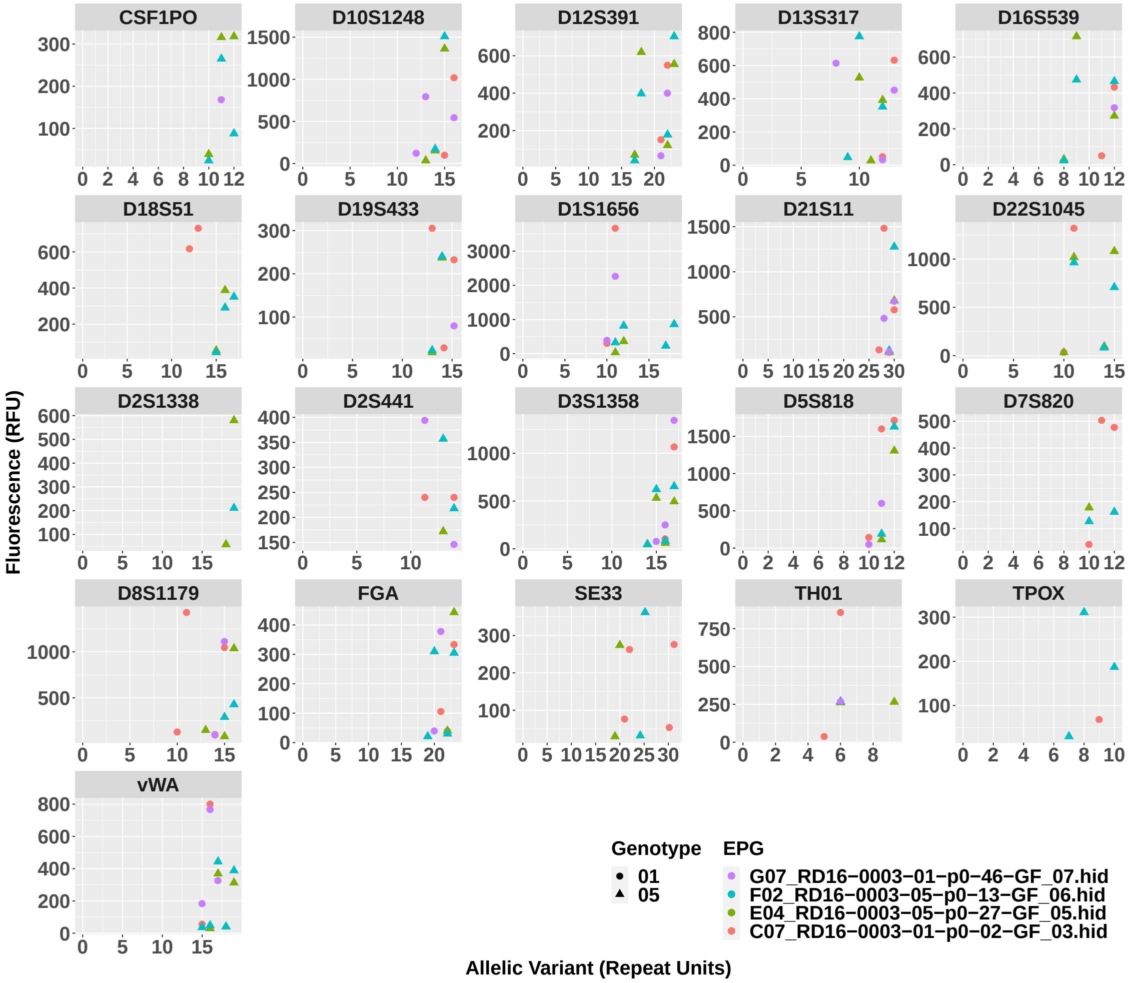


Figure 5.11: Plot of all EPGs assigned to cluster 1. As ground truth is known we can identify EPGs from different genetic sources as indicated by the shape. In this example, the EPGs from two minor contributors have been grouped together. One could not attempt to correctly read off the genotype if ground truth was unknown.

5.4.3 Mclust Results

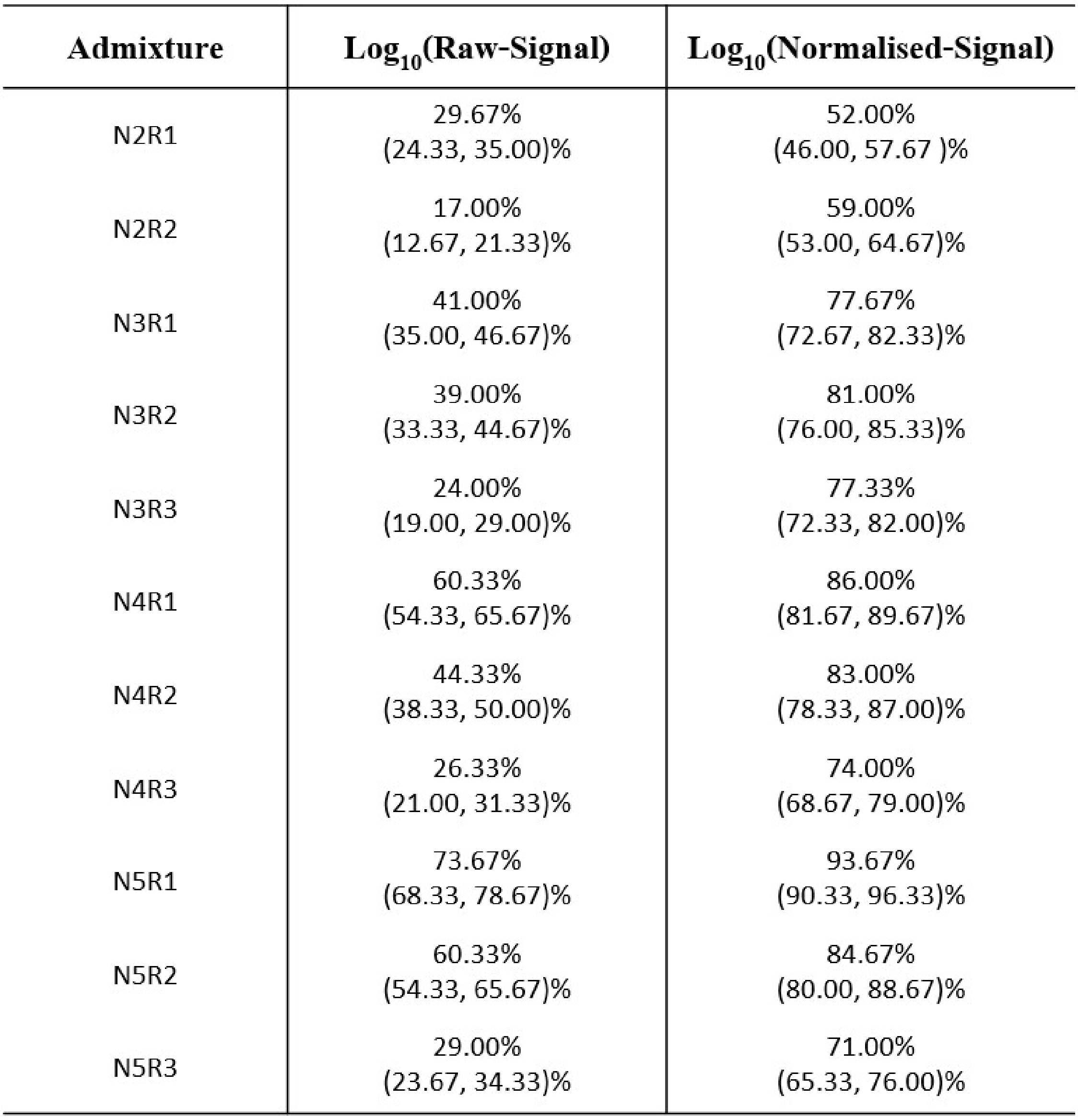
We have compared the performance of Mclust when using $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal), see Table 5.5. Using Mclust on the twice transformed data consistently outperforms the use of log transformed data. When we consider "perfect" results (i.e no mis-clustering or over-clustering has occurred) Mclust out performs K-means when clustering all but a two person balanced admixture. In fact, it is this admixture that performs the poorest when using Mclust, with just over a 50% perfect-clustering rate. Similar to K-means, we see a decrease in perfect clustering as we increase complexity however, this decrease in performance is minimal in comparison, with the exception of a three person imbalanced mixture out performing a three person balanced mixture by 3.33%. We note the exceptional improvement of our computerised end-to-end solution over our analyst-in-the-loop solution when grouping the most complex mixture ratio, a five person highly imbalance mixture, where Mclust successfully assigns EPGs to their distinct genotype almost five times more frequently than kmeans.

As we have stated mis-clustering is the more severe error and so we consider the frequency of mis-clustering compared with the frequency of over-clustering when applying mclust (see Tables 5.6, 5.7 respectively). We discover that over-clustering occurs far more frequently than mis-clustering, noting that although Mclust appears to under-perform when faced with balanced two person mixtures, there is infrequent mis-clustering when using the log transformed data ($< 5\%$) and none when using the twice transformed data. Moreover, when compared to kmeans, we are seeing a vast improvement on the rate of mis-clustering, only when faced with highly imbalanced mixtures do we see the mis-clustering rate exceed 5%.

To consider the severity of over-clustering, we examine how frequent Mclust finds X number of clusters for each mixture type, see Fig. 5.12. For highly imbalanced mixtures, when fewer clusters have been determined than there are contributors in the admixture this is necessarily due to mis-clustering. When we observe over-clustering, Mclust most frequently separates at most one genotype into two distinct clusters returning TrueNoC+1 groupings. For mixtures of three or more genotypes, Mclust determines the correct NoC more than 75% of the time, with the exception of a five person highly imbalanced mixtures, which falls just short of this. Balanced mixtures of two contributors experience the most regular over-clustering and about an 1/8 of the time, both genotypes are being separated into two distinct clusters resulting in TrueNoC+2 groupings.

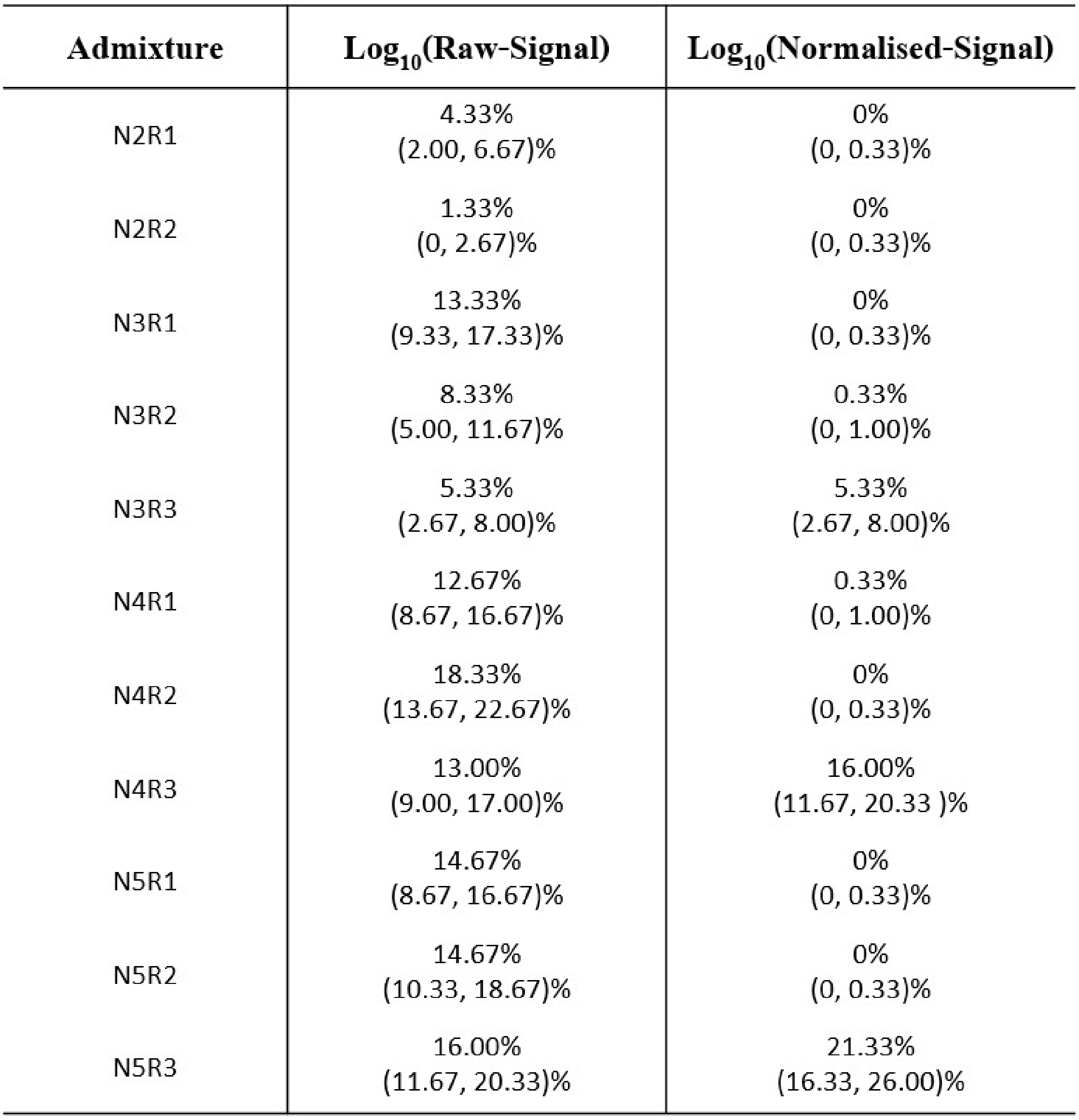
We have included two examples of the computerised end-to-end clustering results, first a highly imbalanced five person admixture where EPGs have been correctly grouped by genotype followed by an balanced two person admixture where an incident of over-clustering has occurred. By plotting all EPGs over laid and colouring by their cluster assignment, the genotype of the major contributors is apparent in Fig. 5.13. However, to make observations regarding the minor contributors, all clusters must be plotted individually, so we have plotted one such cluster, Fig. 5.14. By plotting the superposition of these two EPGs we are gaining coverage of the dropped alleles resulting in a more complete image of the true genotype as indicated by the black vertical bars.

Our second example considers the case of over-clustering, two sources in equal contribution have been assigned to three distinct clusters in Fig. 5.15. This time instead of plotting each cluster individually, we have plotted the superposition of clusters one and two as we can see much overlap in signal of EPGs assigned to these clusters. By over laying all EPGs from cluster one and two, as shown in Fig. 5.16, we see an almost perfect alignment of all EPGs and the true genotype with the exception of the noisy locus D2S442. Such a plot indicates that over-clustering has occurred and that these EPGs do in fact originate from the one genetic source.



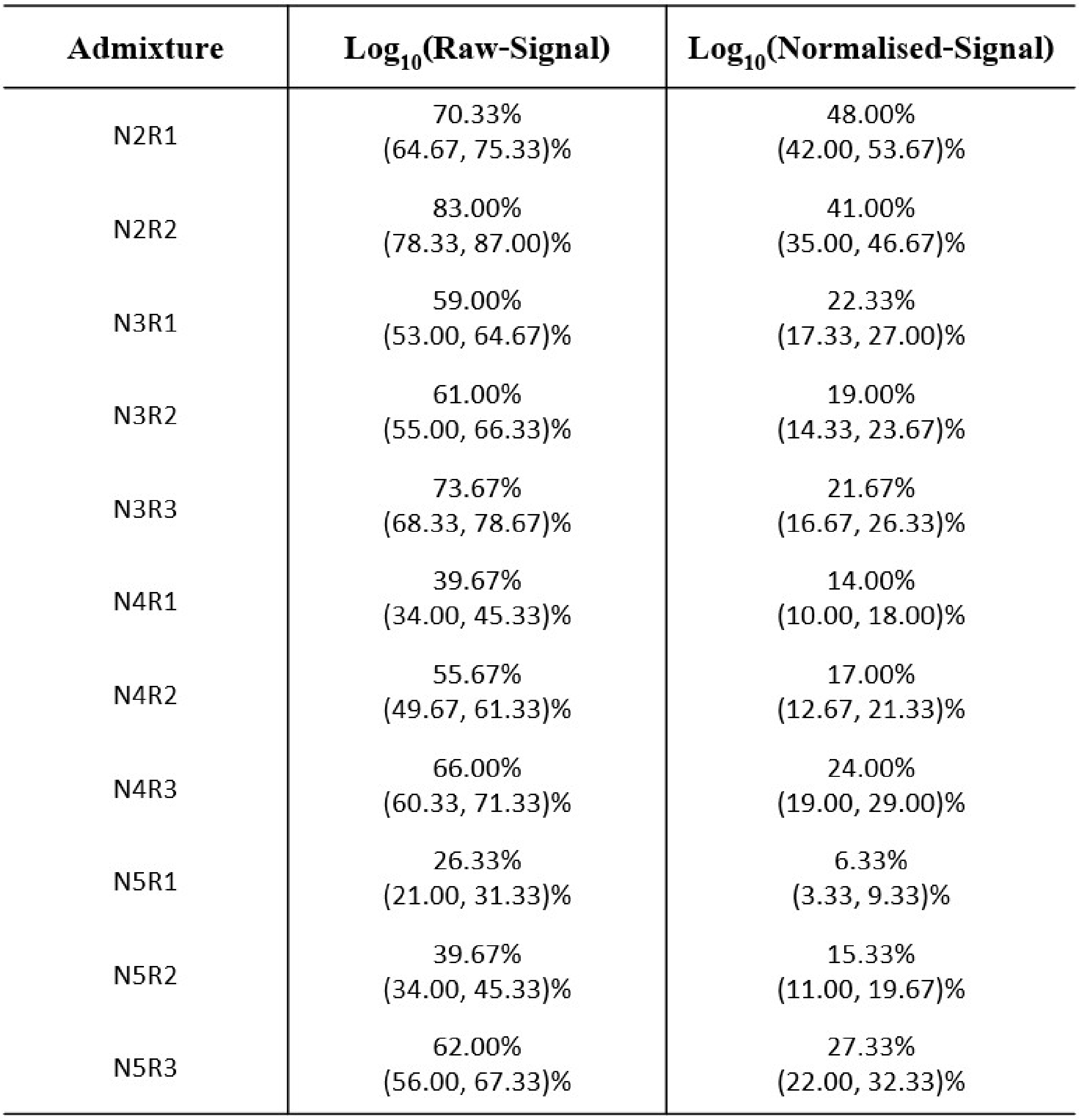
Admixture	Log_{10}(Raw-Signal)	Log_{10}(Normalised-Signal)
N2R1	29.67% (24.33, 35.00)%	52.00% (46.00, 57.67)%
N2R2	17.00% (12.67, 21.33)%	59.00% (53.00, 64.67)%
N3R1	41.00% (35.00, 46.67)%	77.67% (72.67, 82.33)%
N3R2	39.00% (33.33, 44.67)%	81.00% (76.00, 85.33)%
N3R3	24.00% (19.00, 29.00)%	77.33% (72.33, 82.00)%
N4R1	60.33% (54.33, 65.67)%	86.00% (81.67, 89.67)%
N4R2	44.33% (38.33, 50.00)%	83.00% (78.33, 87.00)%
N4R3	26.33% (21.00, 31.33)%	74.00% (68.67, 79.00)%
N5R1	73.67% (68.33, 78.67)%	93.67% (90.33, 96.33)%
N5R2	60.33% (54.33, 65.67)%	84.67% (80.00, 88.67)%
N5R3	29.00% (23.67, 34.33)%	71.00% (65.33, 76.00)%

Table 5.5: The percentage of `Mclust` runs where no mis-clustering or over-clustering has occurred with a 95% binomial confidence interval using the log transformed signal and the twice transformed signal where in, first signal is normalised, second the logarithm is taken. A full description of the admixtures can be found in Table 5.1.



Admixture	Log_{10}(Raw-Signal)	Log_{10}(Normalised-Signal)
N2R1	4.33% (2.00, 6.67)%	0% (0, 0.33)%
N2R2	1.33% (0, 2.67)%	0% (0, 0.33)%
N3R1	13.33% (9.33, 17.33)%	0% (0, 0.33)%
N3R2	8.33% (5.00, 11.67)%	0.33% (0, 1.00)%
N3R3	5.33% (2.67, 8.00)%	5.33% (2.67, 8.00)%
N4R1	12.67% (8.67, 16.67)%	0.33% (0, 1.00)%
N4R2	18.33% (13.67, 22.67)%	0% (0, 0.33)%
N4R3	13.00% (9.00, 17.00)%	16.00% (11.67, 20.33)%
N5R1	14.67% (8.67, 16.67)%	0% (0, 0.33)%
N5R2	14.67% (10.33, 18.67)%	0% (0, 0.33)%
N5R3	16.00% (11.67, 20.33)%	21.33% (16.33, 26.00)%

Table 5.6: The percentage of `Mclust` runs where mis-clustering has occurred with a 95% binomial confidence interval using the log transformed signal and the twice transformed signal where in, first signal is normalised, second the logarithm is taken. We note over-clustering and mis-clustering are not orthogonal events, both can occur in an individual run. A full description of the admixtures can be found in Table 5.1.



Admixture	Log_{10}(Raw-Signal)	Log_{10}(Normalised-Signal)
N2R1	70.33% (64.67, 75.33)%	48.00% (42.00, 53.67)%
N2R2	83.00% (78.33, 87.00)%	41.00% (35.00, 46.67)%
N3R1	59.00% (53.00, 64.67)%	22.33% (17.33, 27.00)%
N3R2	61.00% (55.00, 66.33)%	19.00% (14.33, 23.67)%
N3R3	73.67% (68.33, 78.67)%	21.67% (16.67, 26.33)%
N4R1	39.67% (34.00, 45.33)%	14.00% (10.00, 18.00)%
N4R2	55.67% (49.67, 61.33)%	17.00% (12.67, 21.33)%
N4R3	66.00% (60.33, 71.33)%	24.00% (19.00, 29.00)%
N5R1	26.33% (21.00, 31.33)%	6.33% (3.33, 9.33)%
N5R2	39.67% (34.00, 45.33)%	15.33% (11.00, 19.67)%
N5R3	62.00% (56.00, 67.33)%	27.33% (22.00, 32.33)%

Table 5.7: The percentage of `Mclust` runs where over clustering has occurred with a 95% binomial confidence interval using the log transformed signal and the twice transformed signal where in, first signal is normalised, second the logarithm is taken. We note over-clustering and mis-clustering are not orthogonal events, both can occur in an individual run. We observe over-clustering as the preponderant issue when clustering via `Mclust`. A full description of the admixtures can be found in Table 5.1.

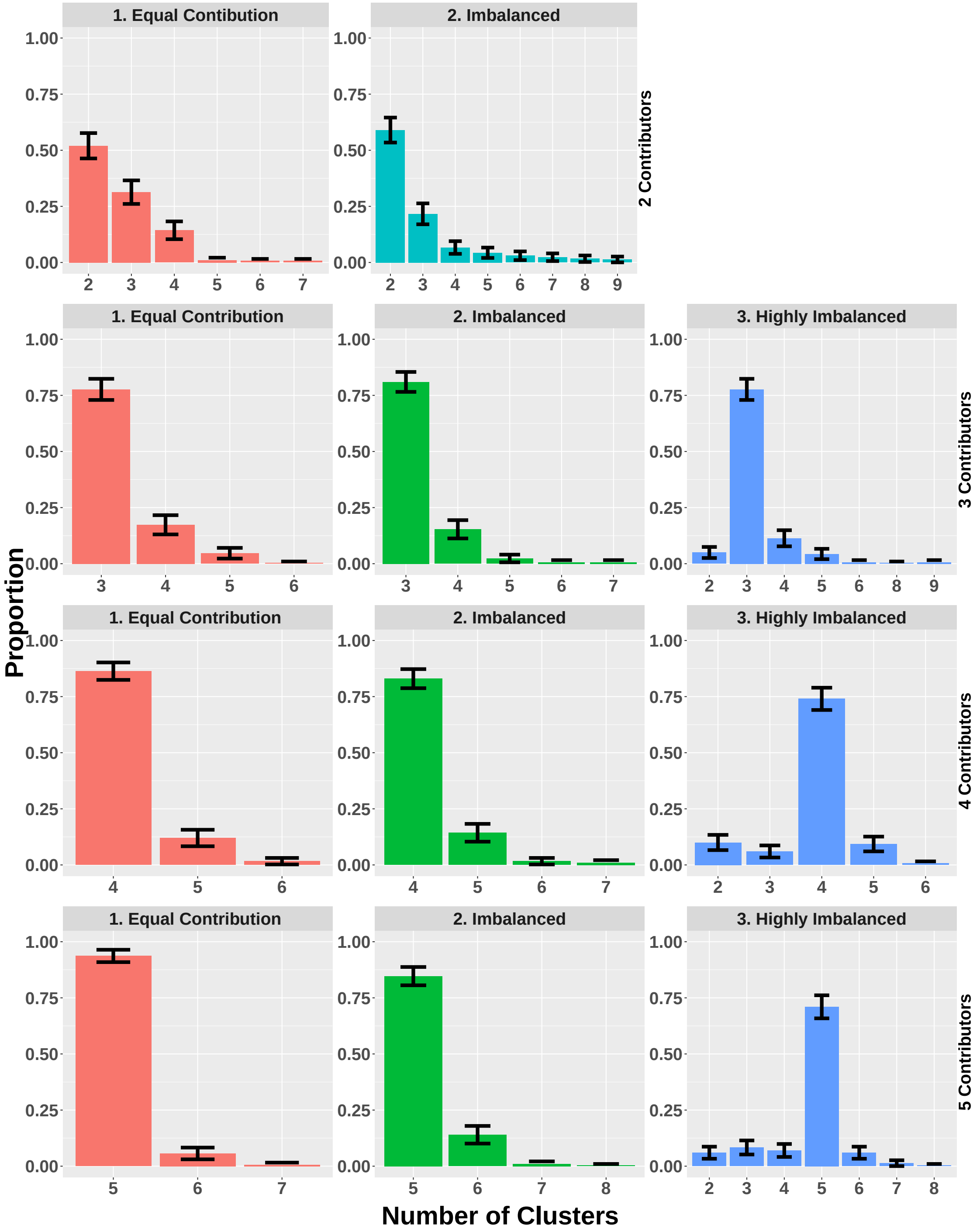


Figure 5.12: A study of the degree of over-clustering for twice transformed data. Here each row corresponds to the number of contributors while the columns correspond to the mixture ratio. When considering the highly imbalanced results we would like to highlight when fewer cluster than contributors have been observed this is necessarily due to mis-clustering. When we are seeing over-clustering, `Mclust` most frequently separates at most one genotype resulting in one extra cluster. Error-bars were determined using a binomial confidence interval.

Example 1

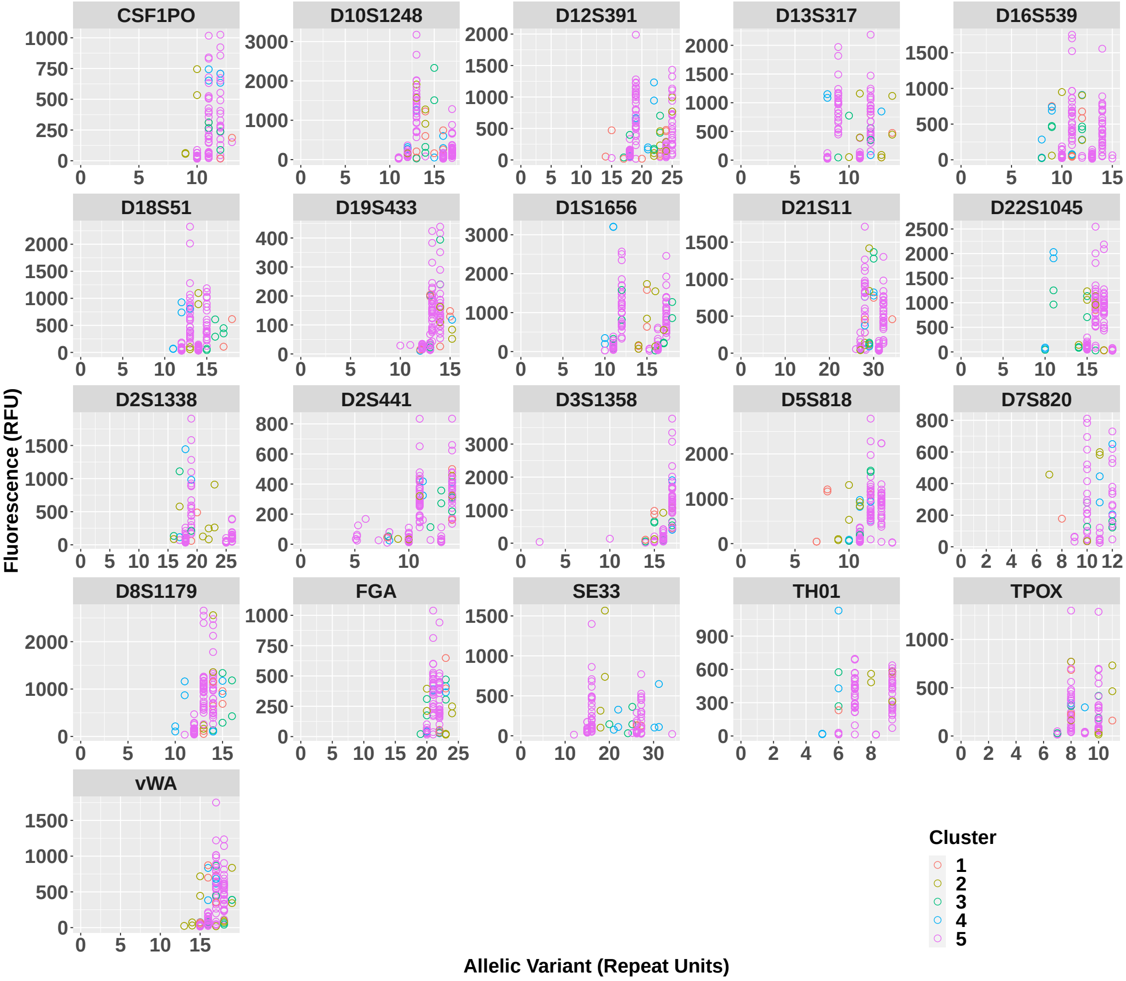


Figure 5.13: Plot of all EPGs present in a highly imbalanced admixture of five contributors. EPGs have been coloured by their `Mclust` cluster assignment. This is an example of a correct clustering result. Without colouring EPGs, interpretaion of such a plot would be similar to that of bulk processed sample but with colouring we can see a clear indication of major contributors genotype. To see any indication of the minor contributors we must plot the individual clusters.

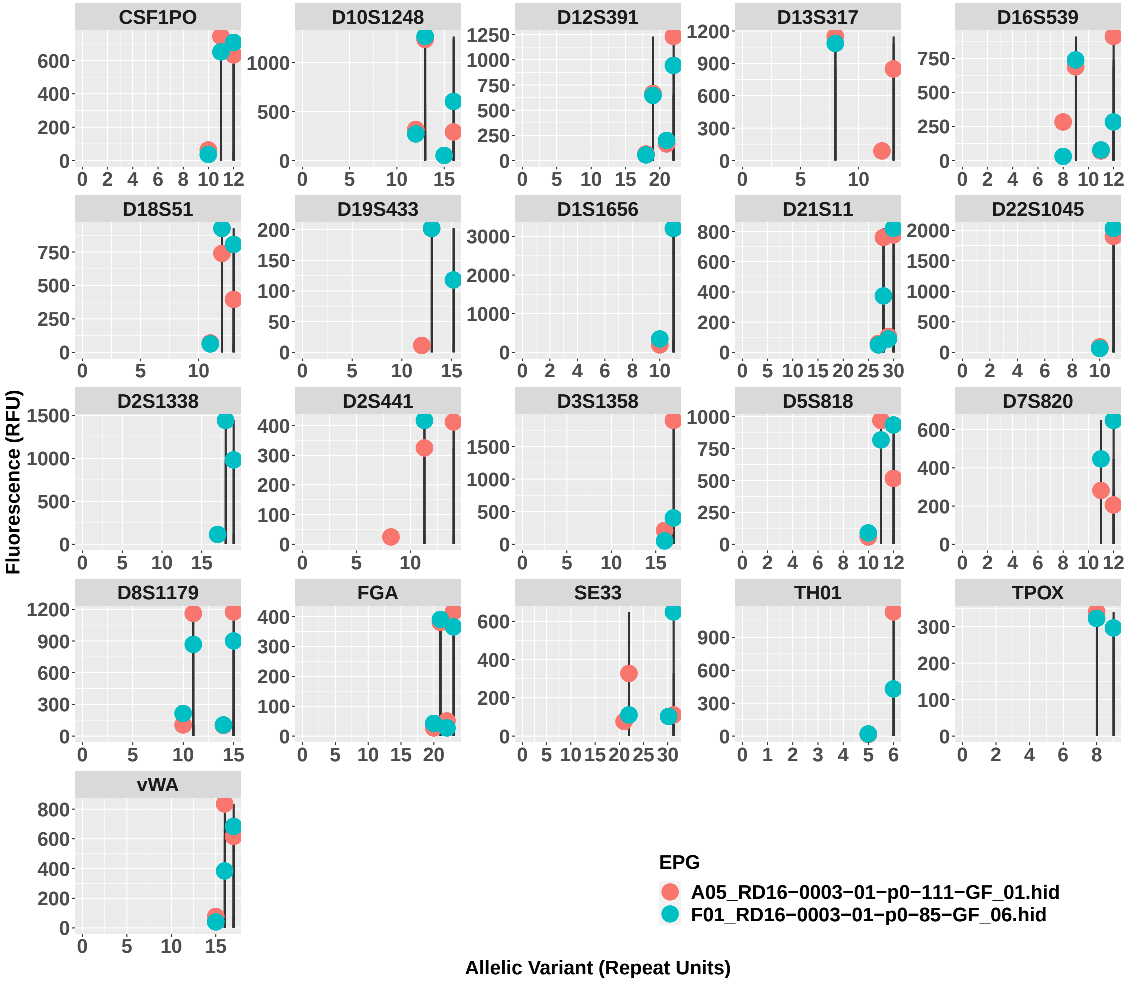


Figure 5.14: Plot of all EPGs assigned to cluster 1. As ground truth is known we have included the known genotype, indicated by the black vertical bars. By plotting cluster 1 on its own we can visually confirm that `Mclust` has correctly assigned all EPGs from this genotype (Genotype 01) to a distinct cluster. For this example we have used colour to indicate which results come from which EPG. By doing so we can emphasise that regardless of drop-out (see loci D2S1338 and D19S433) a complete profile can be obtained from just these two correctly grouped EPGs.

Example 2

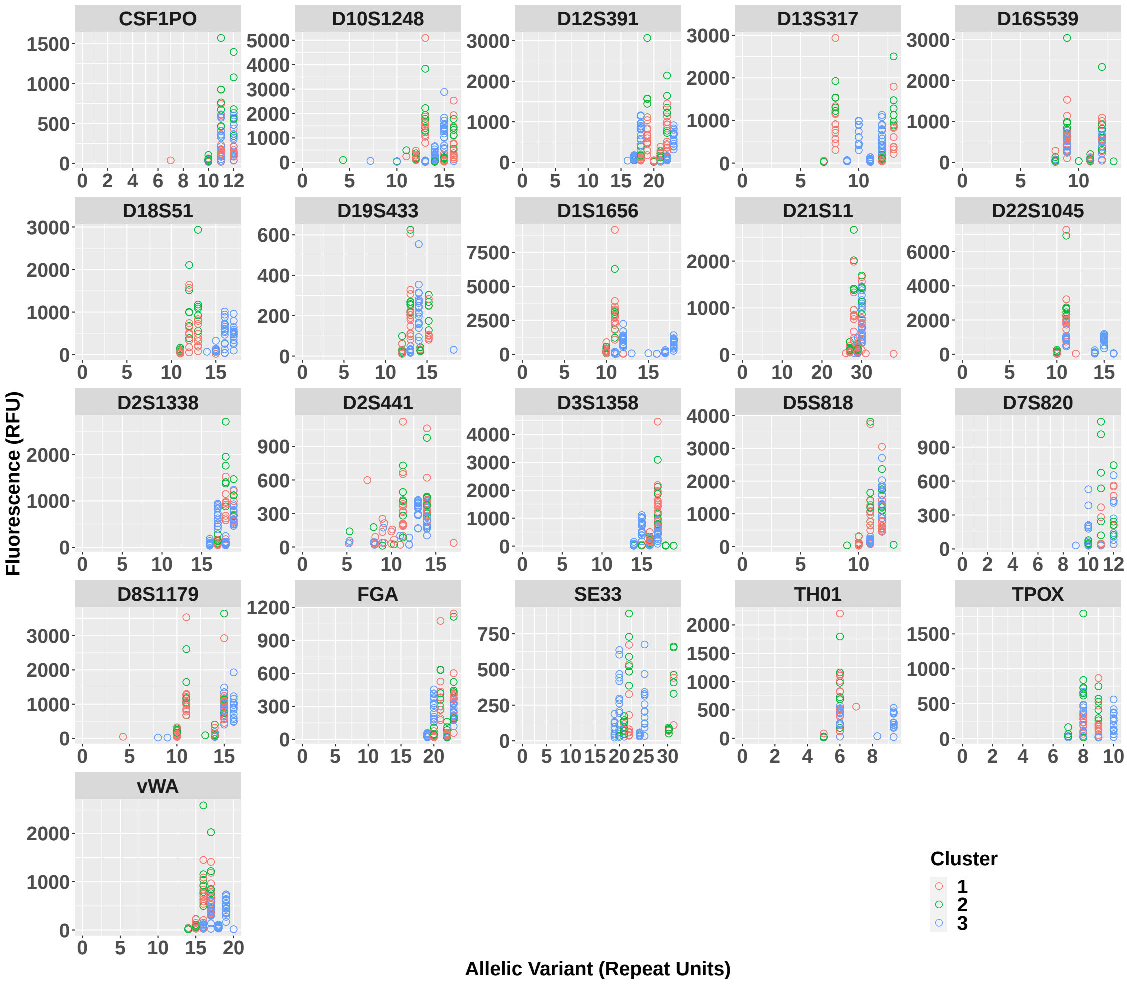


Figure 5.15: Plot of all EPGs present in a balanced admixture of two contributors. EPGs have been coloured by their `Mclust` cluster assignment. This is an example where one genotype has been over-clustered, Genotype 01 has been split into two distinct clusters, group 1 and group 2.

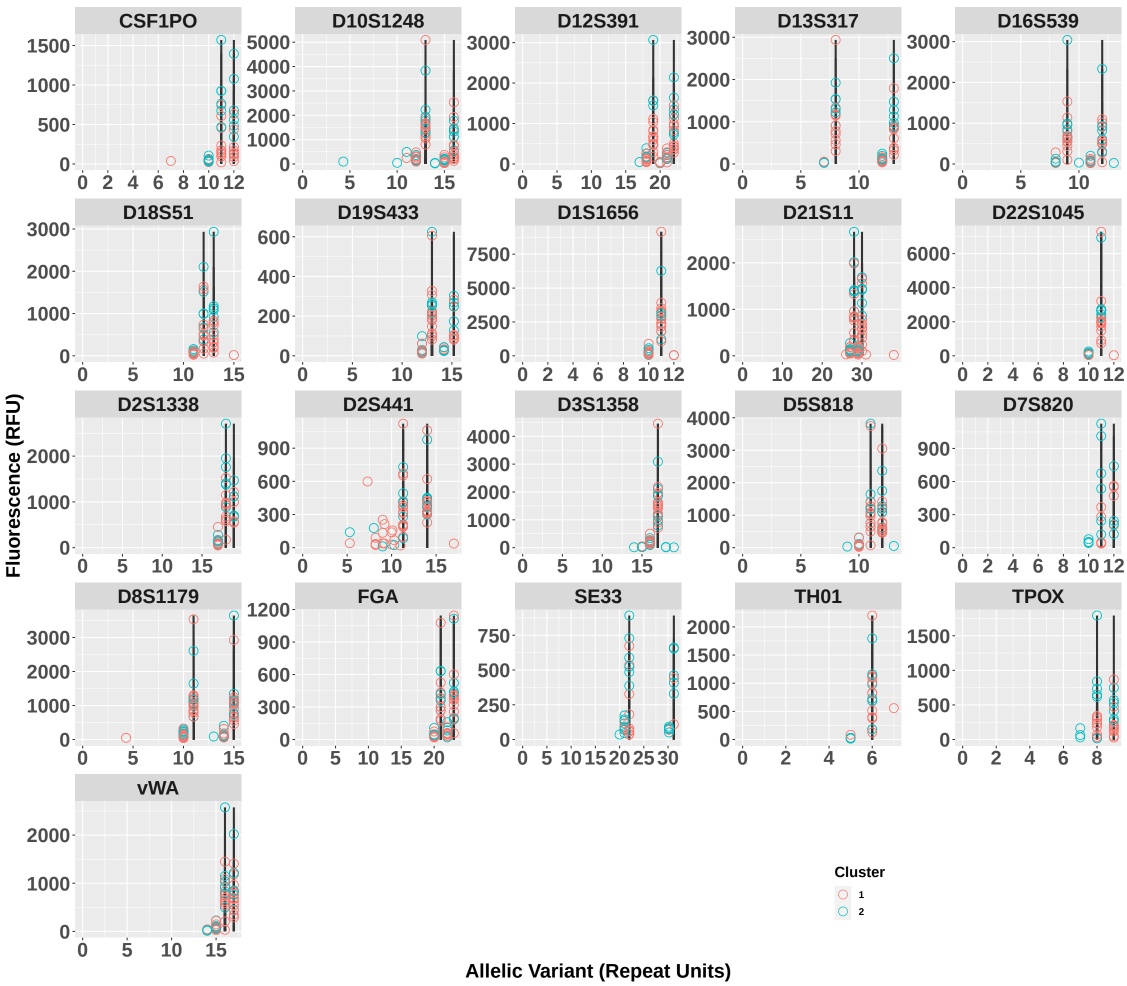


Figure 5.16: Plot of all EPGs assigned to cluster 1 and 2 over-laid. Aside from the particularly noisy locus D2S441, even without ground truth indicated, one could visually determine these two clusters originate from the same genetic source, as EPGs from both clusters stack almost seamlessly together.

6 Conclusion

6.1 Discussion and Future Work

The assessment of single-cell EPGs offers the potential to solve the complex mixture problem, but in order to make it practically reasonable, advances are needed in the analytical pipeline. Here we make a step in that direction by demonstrating that after single-cell EPGs have been created, unsupervised machine learning is capable of grouping them with reasonable accuracy by genotype for assessment, once in a vectorised format rather than the per locus EPG. If clustering is correct, we can reduce the computational strain of the likelihood ratio to the case of a single contributor, thereby removing the added complexity of multiple contributors in an unknown mixture ratio for both the prosecution and the defenses calculation thus allowing for a straightforward calculation by the prosecution. The probabilistic software CEESIt [104] is a fully continuous method that currently calculates the LR (among other statistics) of a bulk processed sample, which is undergoing developments that will further its usability (or establish a sister software) by enabling a similar calculation when instead faced with a collection of single-cell EPGs. We believe the proposed grouping of single-cell EPGs will be well suited for the downstream computations of such a software.

Even though we did not set out to call a genotype, and ultimately one would take the fluorescence measurements of all the EPGs in a given group followed by a detailed LR computation, one could consider using the cluster results to effectively do a binary interpretation, simple allele calling. If the analyst were to do this, they would do so based on the centroid of a K-means cluster or mean and variance of an `Mclust` grouping. We do not claim this as an proxy solution to the LR computation, instead accentuating the extensive information that can be gleamed from these cluster results. However, visual inspection of cluster contents, similar to those seen in Fig. 5.8 and Fig. 5.14, suggests the viability of this approach.

Based on this initial study, the primary failure mode of the analyst-in-the-loop solution appears to be mis-clustering (i.e. EPGs from more than one genotype have been assigned to one cluster) by the `Kmeans` function. When mis-clustering does occur spiking a cluster, particularly in the case of a minor contributor spiking the grouping of a major contributor, may still lead the cluster to give accurate inferences with possibly lower the LRs. By plotting the over laid EPGs of such a cluster it is sometimes evident that this grouping cannot be explained by a single genotype and so other possibilities suggest themselves. One could further investigate a possible mis-clustering by re-applying one of the solutions put forward in this work (example follows), or perhaps consider determining the NoC using the traditional Maximum Allele Count (MAC) method or a program such as a single-cell counterpart to NOCIt [105].

Under the assumption that the analyst has identified a grouping in which mis-clustering has occurred, they may wish to repeat either clustering process, that is plot the PCA of the suspect cluster, determine how many groupings should be assigned and run K-means or feed this new grouping through `Mclust`. Taking the example of one of the miss-clustered groupings from a five person highly imbalanced admixture (see Fig. 5.11, section 5.4.2) we have considered one such manual intervention. We have plotted the PCA of the log transformed signal, coloured by the second round K-means cluster assignment, which was run on the raw signal. We chose $k = 2$, which is the TrueNoC, and $k = 3$ due to the configuration of the PCA, Fig. 6.1. When $k = 2$, K-means correctly assigned both sets of EPGs to their genotype in this second run, however, it is unlikely the analyst would have determined $k = 2$ due to the vast separation between the EPGs of Genotype 01 but surprisingly, when using k= 3 the algorithm over-clustered Genotype 05 instead. For comparison we have taken the same incorrect grouping, twice transformed the signal and fed this new grouping into `Mclust`. Again we see over-clustering this time splitting Genotype 01 into two distinct clusters.

Alternatively, the analyst could consider forensic relevant techniques applied to the true signal to debate the number of contributors present such as the binary maximum allele count evaluation (see section 1.3.3 for more detail). We have taken the same miss-clustered example as above and plotted the over-laid signal of all four EPGs (indicated by colour) along with a DT set at 30RFU in Fig. 6.2. The locus D8S1179 has the greatest number of peaks above this DT, with all six peaks greater than 30RFU and so by MAC we can certainly confirm there is more than one contributor within this grouping. This has been presented as a suggestive example but with a further study one could determine a more appropriate DT and subsequently use MAC to determine if there is in fact two contributors within this grouping as MAC has been shown to be effective for mixtures of three or less contributors [32].

Another avenue one could pursue is the use additional computational methods that infer the number of contributors of a DNA sample such as NOCIt [105]. NOCIt is an

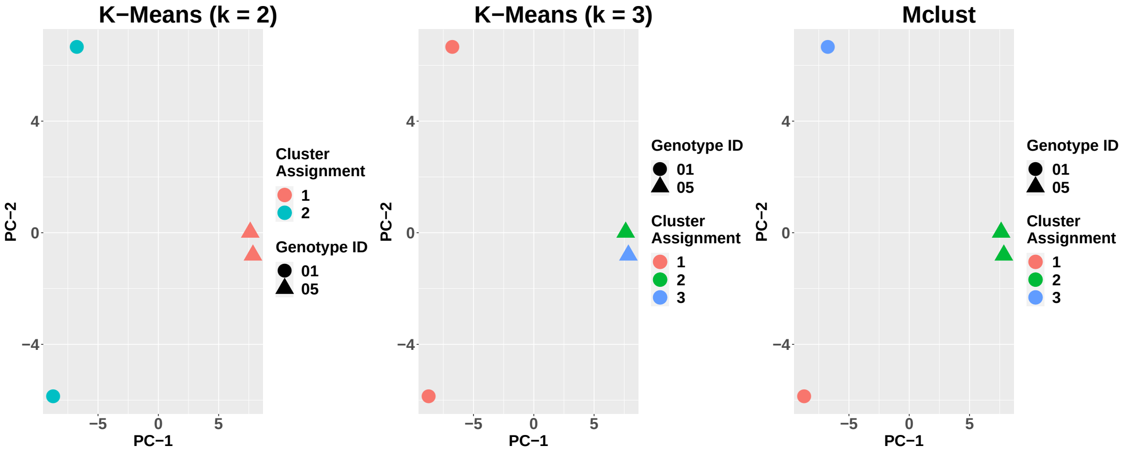


Figure 6.1: One may wish to repeat either clustering process on a grouping they believe to contain more than one distinct genotype. We have taken the miss-clustered example from Fig. 5.11, re-run K-means assuming $k = 2$ on the raw signal and plotted the PCA using the log transformed signal of these four EPGs coloured by cluster assignment. We note that the analyst may not have inferred k correctly from this PCA plot considering the large separation between the two EPGs from Genotype 01 and so we have included the result of re-running the analyst-in-the-loop solution with $k = 3$ which unexpectedly resulted in an over-clustering of Genotype 05. We have also re-run `Mclust` on the twice transformed signal for comparison and this has also resulted in an over-clustering, however here, as expected Genotype 01 has been split into two distinct groups.

algorithm that calculates the *a posteriori* probability (APP) on the number of contributors given an EPG, taking into account signal peak heights, population allele frequencies, allele drop-out and stutter [105, 53]. Although originally designed for bulk processed samples, it would not be challenging for future versions of NOCIt (or alternate software) to instead handle a grouping of single-cell EPGs in a similar manner. Notably, NOCIt can determine with high accuracy if a sample contains one or two contributors, becoming less confident as the number of contributors increases [53]. In this way, when faced with a grouping believed to contain more than one distinct genotypes, a program similar to NOCIt could offer confident clarity on the matter. By applying such a software to a suspect grouping, the analyst at minimum could confidently determine if a mis-clustering has occurred or not.

In contrast to the preponderant source of error when implementing our analyst-in-the-loop solution, the computerised-end-to-end solution seems to experience a far greater level of over-clustering (i.e. EPGs from one distinct genotype have been grouped into two or more clusters) resulting in an over estimation of the number of contributors. Once satisfied that each grouping contains at most one genotype, done so by applying one of the possible methods suggested above or otherwise, a resolution to over-clustering could

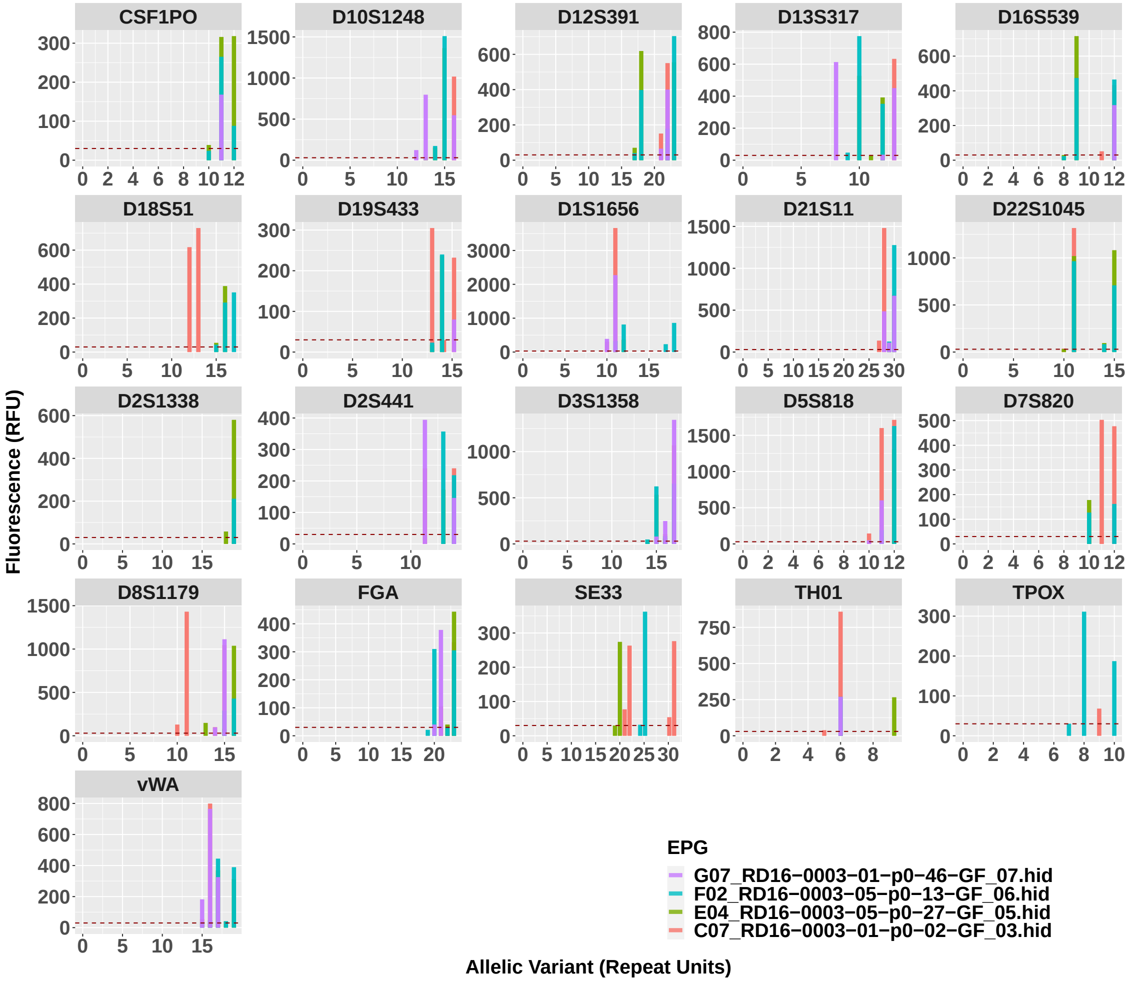


Figure 6.2: One may wish to confirm the presence of additional genotypes within a grouping they suspect to be the result of mis-clustering using maximum allele count. We have taken the miss-clustered example from Fig. 5.11, plotted the over laid peak heights of detected signal for the four EPGs found in this grouping. Signal has been coloured by EPG and an analytical threshold of 30RFU is indicated by the horizontal red dashed line. By employing the MAC method (see section 1.3.3) for each locus we see at most, four substantial peaks indicating the presence of more than one genotype.

be determined by employing pairwise choosing of the groupings coupled with a similar aforementioned technique to determine a cluster as a single contributor grouping. If we consider the example of over-clustering in Fig. 5.15, a balanced two person mixture consisting of 40 EPGs was fragmented into three distinct clusters, where cluster 1 and 2 contained EPGs from Genotype 01 and cluster 3 contained EPGs from Genotype 05. If the analyst were to join cluster 1 with cluster 2 and apply either the MAC technique or a single-cell NOCIt program to this new grouping, they could determine with high accuracy that this new grouping is in fact described by a distinct genotype. Without our proposed grouping methodology, one could consider the pairwise combinations of each EPG, resulting in ${}^{40}C_2 = 780$ possible combinations, but by considering the groupings as opposed to distinct EPGs, the number of pairwise combinations is reduced drastically to ${}^{3}C_2 = 3$. In reality the number of possible combinations can be reduced even further as this process can be done sequentially, that is each time a pair of clusters is amalgamated, the number of groups is instantly reduced.

To validate the clustering performance against ground truth it was necessary to remove low quality EPGs, whether they came about because the cells were degraded or because further optimisation of the extraction and amplification chemistry might lead to improved signal quality is unknown but work is ongoing. Areas for future work involve testing on EPGs that have been generated using the extraction methods suggested by Sheth et. al. [102] for epithelial cells, evaluating the cluster performance when faced with low quality and/or degraded EPGs. We would also like to extend the study to include the grouping of single-cell EPGs generated from blood, another common sample type found at crime scenes and sex cells, such as sperm which only contain half the genome. As a result, even if high quality EPGs can be made, there are issues of imputation in the case of sex related crimes.

6.2 Conclusion

We have shown that single-cell EPG-vectors are distinguishable from distinct genetic sources when using a cosine dissimilarity measure and subsequently can be clustered by employing one of two solutions, an analyst-in-the-loop solution or the highly effective computerised-end-to-end solution. We can visually infer the NoC of a balanced mixture with high accuracy, an imbalanced mixture with moderately high accuracy, and a highly imbalanced mixture with accuracy (less so for the case of five contributors) by applying PCA to the log transform of our high dimensional EPG-vectors and plotting the first two principal components. Alternatively, we could use a more modern technique such as UMAP which works well on balanced mixtures, but less impressively on imbalanced or highly imbalanced admixtures. Informally the author of umap is apparently aware the performance of his function is not well suited to small data sets, and mixtures of these types that we have simulated in general contain between 40 and 80 EPGs.

The computerised-end-to-end solution has shown to perform exceedingly well on log transformed signal that has first been normalised, particularly when compared to the capabilities of the analyst-in-the-loop solution for complex mixtures. Although occasional mis-clustering (where EPGs from two or more contributors have been grouped together), and frequent over-clustering (where a single genotype has been grouped into two distinct clusters) is observed, we believe this issue will be highly manageable in downstream interpretation, as the clustering can be refined by returning to the consideration of the forensic information contained in the single-cell EPGs. There is good reason to hope that single-cell EPG technologies can be made practical in future crime scene investigation.

A

Appendix A

A.1 Potential Allele Information Per Locus

Locus - l	**Set of Potential Alleles -** B^l	**Cardinality -** $\|B^l\|$
CSF1P0	$B^1 = \{7.0, 7.1, 7.2, 7.3, 8.0, ...12.3, 13.0\}$	24
D1S1656	$B^2 = \{8.0, 8.1, 8.2, 8.3, 9.0, ...19.3.20.0\}$	48
D2S1338	$B^3 = \{11.0, 11.1, 11.2, 11.3, 12.0, ...25.3, 26.0\}$	60
D2S441	$B^4 = \{5.0, 5.1, 5.2, 5.3, 6.0, ...16.3, 17.0\}$	48
D3S1358	$B^5 = \{1.0, 1.1, 1.2, 1.3, 2.0, ...18.3, 19.0\}$	72
D5S818	$B^6 = \{7.0, 7.1, 7.2, 7.3, 8.0, ...15.3, 16.0\}$	36
D7S820	$B^7 = \{6.0, 6.1, 6.2, 6.3, 7.0, ...12.3, 13.0\}$	28
D8S1179	$B^8 = \{4.0, 4.1, 4.2, 4.3, 5.0, ...18.3, 19.0\}$	60
D10S1248	$B^9 = \{3.0, 3.1, 3.2, 3.3, 4.0, ...18.3, 19.0\}$	64
D12S391	$B^{10} = \{14.0, 14.1, 14.2, 14.3, 15.0, ...26.3, 27.0\}$	52
D13S317	$B^{11} = \{7.0, 7.1, 7.2, 7.3, 8.0, ...14.3, 15.0\}$	32
D16S539	$B^{12} = \{8.0, 8.1, 8.2, 8.3, 9.0, ...14.3, 15.0\}$	28
D18S51	$B^{13} = \{9.0, 9.1, 9.2, 9.3, 10.0, ...17.3, 18.0\}$	36
D19S433	$B^{14} = \{6.0, 6.1, 6.2, 6.3, 7.0, ...18.3, 19.0\}$	52
D21S11	$B^{15} = \{24.0, 24.1, 24.2, 24.3, 25.0, ...37.3, 38.0\}$	56
D22S1045	$B^{16} = \{1.0, 1.1, 1.2, 1.3, 2.0, ...17.3, 18.0\}$	68
FGA	$B^{17} = \{16.0, 16.1, 16.2, 16.3, 17.0, ...44.3, 45.0\}$	116
SE33	$B^{18} = \{12.0, 12.1, 12.2, 12.3, 13.0, ...34.3, 35.0\}$	92
TH01	$B^{19} = \{5.0, 5.1, 5.2, 5.3, 6.0, ...9.3, 10.0\}$	20
TPOX	$B^{20} = \{6.0, 6.1, 6.2, 6.3, 7.0, ...20.3, 21.0\}$	60
vWA	$B^{21} = \{13.0, 13.1, 13.2, 13.3, 14.0, ...22.3, 23.0\}$	40

Table A.1: Set of potential alleles for each locus determined empirically from the (unfiltered) data along with the set cardinality.

A.2 Ground Truth for All Genotypes

Locus	Geno 01	Geno 02	Geno 05	Geno 06	Geno 07
CSF1P0	11, 12	10, 10	11, 12	11, 13	11, 12
D1S1656	11, 11	15, 16	12, 18	15, 15	12, 17.3
D2S1338	18, 19	17, 23	17, 19	20, 23	19, 26
D2S441	11.3, 14	11, 14	13, 14	14, 14	11, 14
D3S1358	17, 17	16, 17	15, 17	15, 17	17, 17
D5S818	11, 12	10, 11	12, 12	8, 12	12, 13
D7S820	11, 12	7, 11	10, 12	8, 12	10, 12
D8S1179	11, 15	14, 14	15, 16	14, 15	13, 14
D10S1248	13, 16	13, 14	15, 15	14, 16	13, 17
D12S391	19, 22	23, 25	18, 23	15, 24	19, 25
D13S317	8, 13	11, 14	10, 12	12, 14	9, 12
D16S539	9, 12	10, 12	9, 12	9, 12	11, 14
D18S51	12, 13	14, 14	16, 17	14, 18	13, 15
D19S433	13, 15.2	14, 15.2	14, 14	13, 15	13.2, 15
D21S11	28, 30	29, 29	30, 30	28, 30	28, 32.2
D22S1045	11, 11	15, 16	11, 15	16, 16	16, 17
FGA	21, 23	20, 24	20, 23	23, 23	21, 22
SE33	22, 31.2	19, 19	20, 25.2	19, 26.2	16, 27.2
TH01	6, 6	8, 9.3	6, 9.3	6, 9.3	7, 9.3
TPOX	8, 9	8, 11	8, 10	8, 11	8, 10
vWA	16, 17	15, 19	17, 19	16, 17	17, 18

Table A.2: Ground truth for each genotype present in our study.

A.3 Drop-Out Recordings

We have undertaken a preliminary empirical study of drop-out rates by means of a binary approach. An allele is counted if it exceeds a standard Analytical Threshold (AT) of 30RFU. When studying homozygote loci, we only consider the case of total allele drop-out and if signal is detected above the AT we count this as one true allele has been recovered. We examine these data specific drop-out rates for both the unfiltered and filtered data.

A.3.1 Drop-Out Pre-Filtering

Locus	Geno 01	Geno 02	Geno 05	Geno 06	Geno 07	Average per locus
CSF1P0	29%		62%	59%	49%	50%
D1S1656		17%	54%		37%	36%
D2S1338	34%	33%	69%	61%	62%	52%
D2S441	15%	13%	45%		28%	25%
D3S1358		15%	47%	46%		36%
D5S818	16%	16%		45%	31%	27%
D7S820	51%	28%	76%	59%	67%	56%
D8S1179	19%		48%	38%	31%	34%
D10S1248	18%	13%		39%	31%	25%
D12S391	28%	21%	63%	52%	45%	42%
D13S317	30%	24%	60%	52%	42%	42%
D16S539	30%	27%	63%	56%	41%	43%
D18S51	28%		54%	54%	47%	46%
D19S433	29%	26%		42%	28%	31%
D21S11	23%			50%	42%	38%
D22S1045		14%	42%		25%	27%
FGA	28%	22%	49%		36%	34%
SE33	46%		66%	59%	56%	57%
TH01		35%	61%	50%	35%	45%
TPOX	38%	36%	67%	58%	53%	49%
vWA	17%	19%	52%	45%	31%	33%
Average per Geno	28%	22%	57%	51%	41%	

Table A.3: Percentage of alleles that drop-out per heterozygote loci for each genotype. An analytical threshold of 30RFU has been applied to the data, any signal below this is considered to be noise. Red indicates the highest drop-out observed for each genotype.

Locus	**Geno 01**	**Geno 02**	**Geno 05**	**Geno 06**	**Geno 07**	**Average per locus**
CSF1P0		16%				16%
D1S1656	14%			41%		29%
D2S441				35%		35%
D3S1358	12%				18%	15%
D5S818			42%			42%
D8S1179		13%				13%
D10S1248			29%			29%
D18S51		14%				14%
D19S433			31%			31%
D21S11		14%	50%			32%
D22S1045	10%			32%		21%
FGA				37%		37%
SE33		20%				20%
TH01	17%					17%
Average per Geno	13%	15%	38%	36%	18%	

Table A.4: Percentage of total drop-out observed per homozygote loci for each genotype. An analytical threshold of 30RFU has been applied to the data, any signal below this is considered to be noise. Red indicates the highest drop-out observed for each genotype. For homozygote loci we only consider the case of total allele drop-out.

A.3.2 Drop-Out Post-Filtering

Locus	Geno 01	Geno 02	Geno 05	Geno 06	Geno 07	Average per locus
CSF1P0	17%		16%	29%	26%	22%
D1S1656		5%	10%		12%	9%
D2S1338	21%	22%	36%	28%	43%	30%
D2S441	4%	2%	10%		5%	5.3%
D3S1358		7%	6%	14%		9%
D5S818	5%	5%		9%	4%	5.8%
D7S820	40%	14%	47%	27%	55%	36.6%
D8S1179	7%		8%	5%	4%	6%
D10S1248	7%	3%		3%	6%	4.8%
D12S391	13%	9%	27%	15%	18%	16.4%
D13S317	16%	11%	16%	19%	19%	16.2%
D16S539	15%	15%	22%	22%	15%	17.8%
D18S51	13%		14%	21%	24%	18%
D19S433	17%	17%		8%	7%	12.3%
D21S11	10%			17%	17%	14.7%
D22S1045		5%	8%		2%	5%
FGA	14%	9%	8%		10%	10.3%
SE33	36%		28%	26%	36%	31.5%
TH01		25%	24%	17%	13%	19.8%
TPOX	25%	26%	28%	24%	31%	26.8%
vWA	6%	7%	13%	12%	7%	9%
Average per Geno	15.6%	11.4%	18.9%	17.4%	17.7%	

Table A.5: Percentage of alleles that drop-out per heterozygote loci for each genotype after applying a high-pass filter. EPGs who's sum of log transformed signal is less than 71 have been removed from the study. Red indicates the highest drop-out observed for each genotype.

Locus	**Geno 01**	**Geno 02**	**Geno 05**	**Geno 06**	**Geno 07**	**Average per locus**
CSF1P0		2%				2%
D1S1656	1%			4%		2.5%
D2S441				2%		2%
D3S1358	0%				1%	0.5%
D5S818			7%			7%
D8S1179		0%				0%
D10S1248			2%			2%
D18S51		1%				1%
D19S433			0%			0%
D21S11		2%	2%			2%
D22S1045	1%			2%		1.5%
FGA				4%		4%
SE33		5%				5%
TH01	3%					3%
Average per Geno	1.3%	2%	2.8%	3%	1%	

Table A.6: Percentage of total drop-out observed per homozygote loci for each genotype after applying a high-pass filter. EPGs who's sum of log transformed signal is less than 71 have been removed from the study. Red indicates the highest drop-out observed for each genotype. For homozygote loci we only consider the case of total allele drop-out.

A.4 Accompanying PCA Plots

Accompanying plots for section 4.3.1, comparing $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal) visualisations of our data on a PCA plot for two, three, four and five person admixtures in various mixture ratios. We determine the analyst should plot both and choose the larger number of clusters as it is equally likely a plot of either data transformation will out perform the other.

A.4.1 PCA Plots for Two Person Admixtures

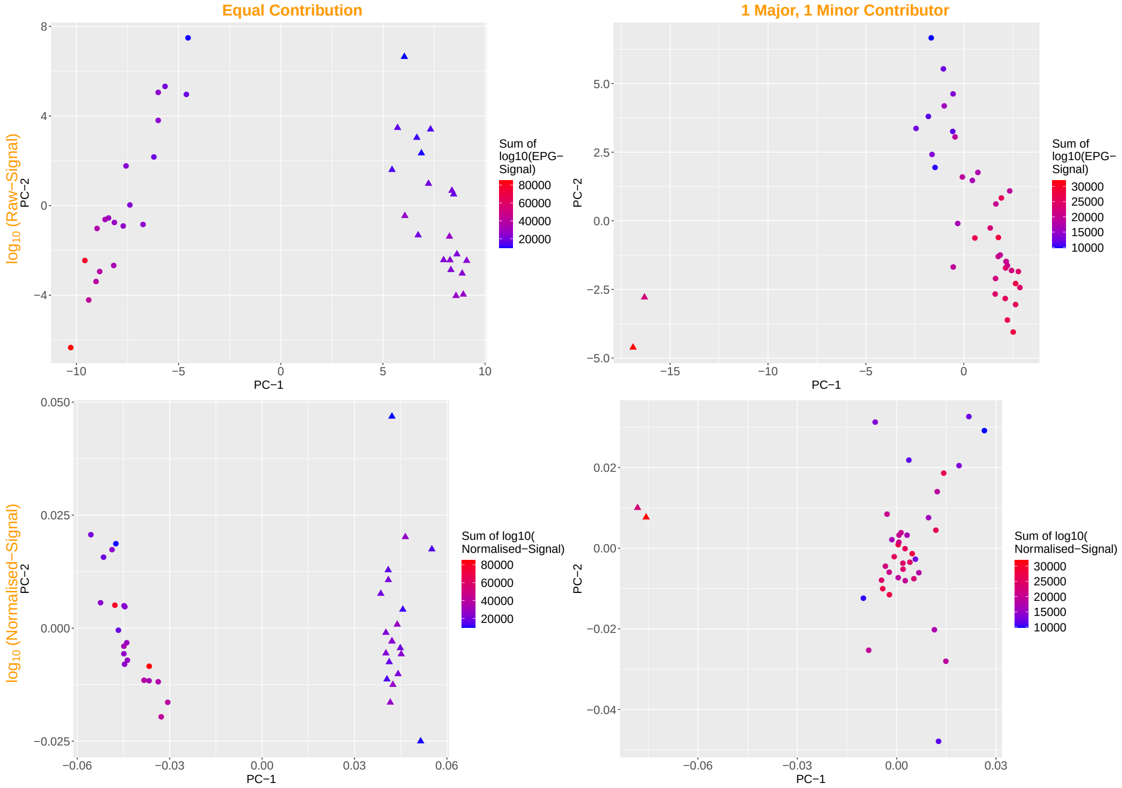


Figure A.1: PCA plots for simulated admixtures of 2 contributors in various mixture ratios. Circles are Genotype 02 and triangles are Genotype 07. Individuals have been coloured by EPG intensity. The first row of PCA plots have been generated using log transformed data. The second row of PCA plots, representing the same admixtures as seen in the above, are resultant of data that has undergone two transformations, first signal has been normalised, second the logarithm of normalised signal has be taken. Each column of plots corresponds to a type of mixture ratio from equal contribution through to highly imbalanced. Both data transformations perform equally sufficiently as two distinct groups are present in each plot.

A.4.2 PCA Plots for Three Person Admixtures

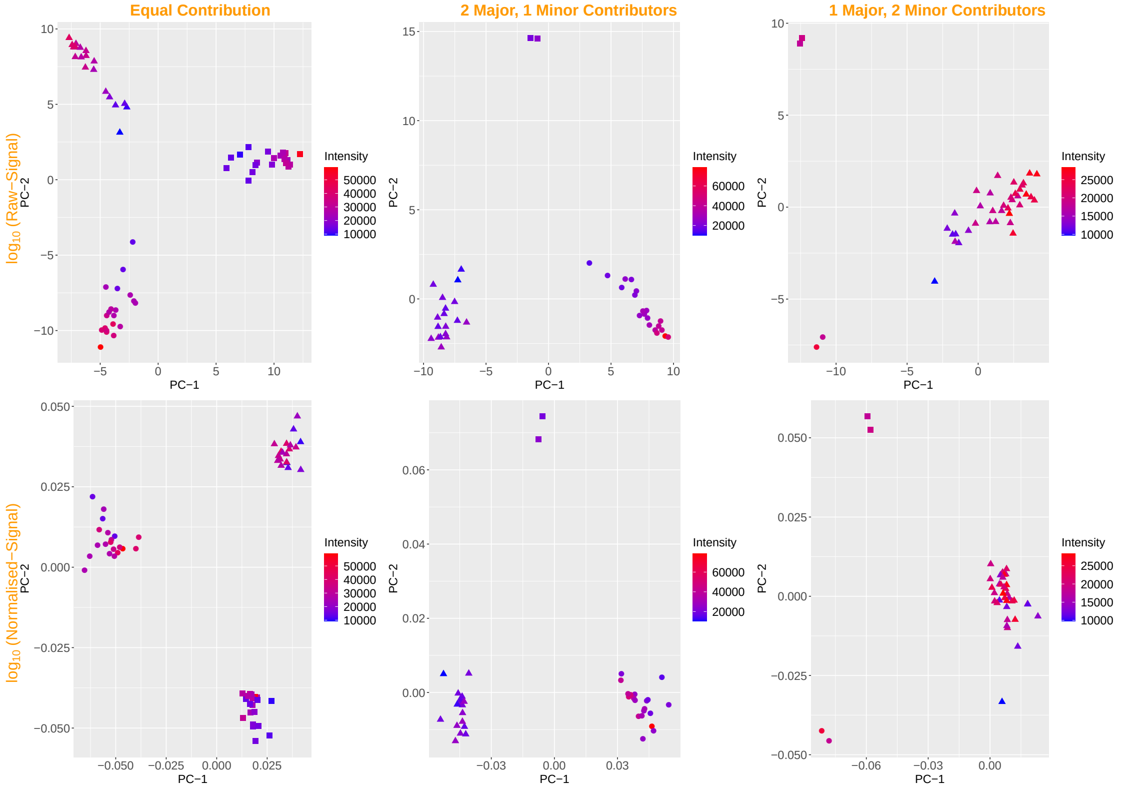


Figure A.2: PCA plots for simulated admixtures of 3 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 06, and squares are Genotype 07. Individuals have been coloured by EPG intensity. The first row of PCA plots have been generated using log transformed data. The second row of PCA plots, representing the same admixtures as seen in the row above, are resultant of data that has undergone two transformations, first signal has been normalised, second the logarithm of normalised signal has be taken. Each column of plots corresponds to a type of mixture ratio from equal contribution through to highly imbalanced. Both data transformations perform equally sufficiently with three distinct groups present in each plot.

A.4.3 PCA Plots for Four Person Admixtures

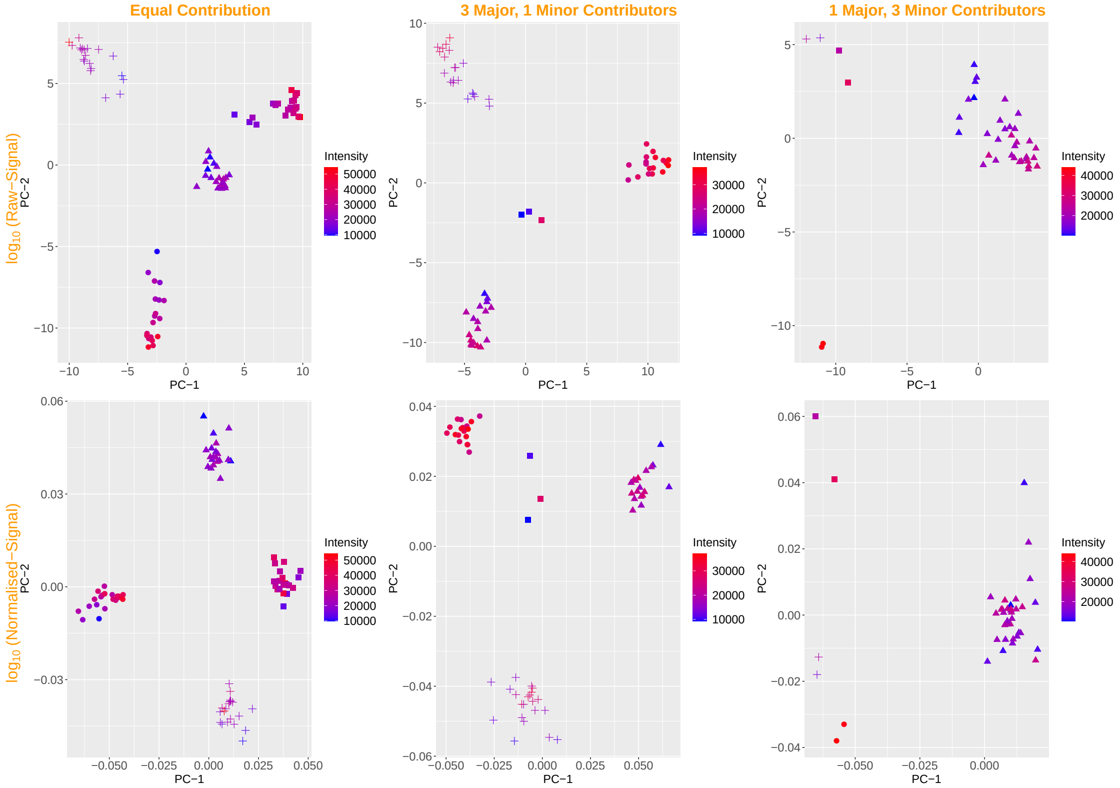


Figure A.3: PCA plots for simulated admixtures of 4 contributors in various mixture ratios. Circles are Genotype 02, triangles are Genotype 05, squares are Genotype 06, and crosses are Genotype 07. Individuals have been coloured by EPG intensity. The first row of PCA plots have been generated using log transformed data. The second row of PCA plots, representing the same admixtures as seen in the row above, are resultant of data that has undergone two transformations, first signal has been normalised, second the logarithm of normalised signal has be taken. Each column of plots corresponds to a type of mixture ratio from equal contribution through to highly imbalanced. Four distinct groups can be determined for each mixture ratio when using the log transformed normalised data. When plotting the PCA of the log transformed data for the highly imbalanced mixture ratio one may incorrectly infer the number of contributors due to the close proximity of Genotype 05 and Genotype 06.

A.4.4 PCA Plots for Five Person Admixtures

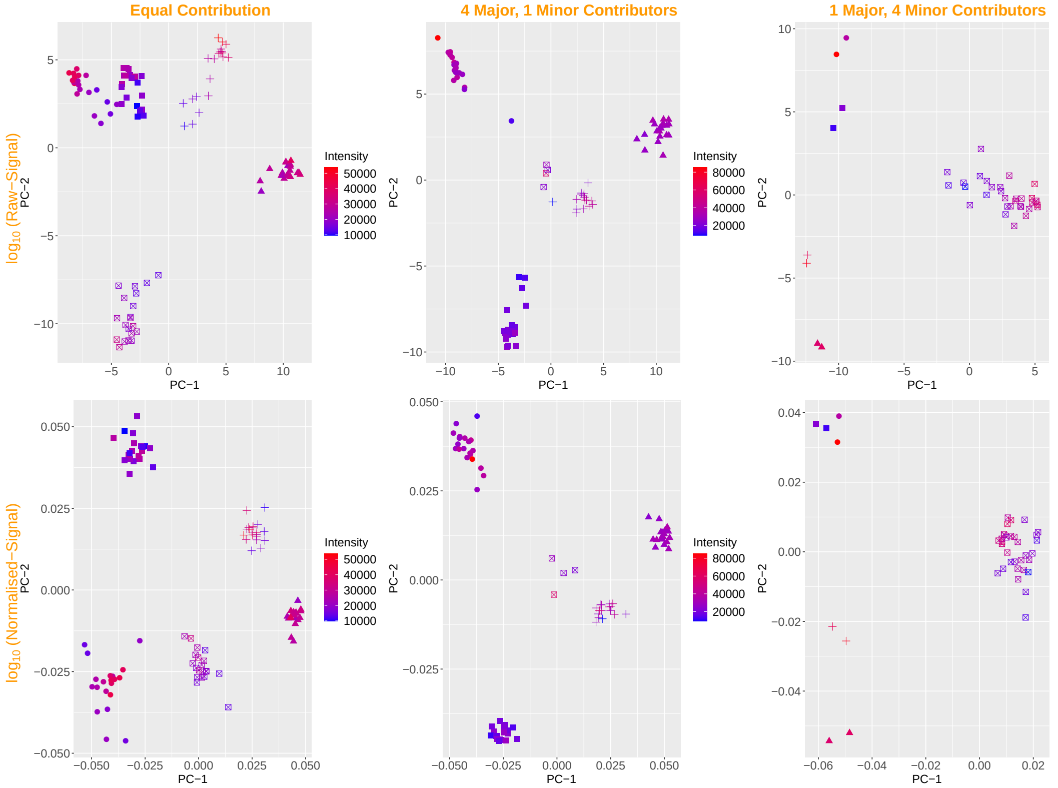


Figure A.4: PCA plots for simulated admixtures of 5 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 02, squares are Genotype 05, crosses are Genotype 06 and boxes with an x are Genotype 07. Individuals have been coloured by EPG intensity. The first row of PCA plots have been generated using log transformed data. The second row of PCA plots, representing the same admixtures as above, are resultant of data that has undergone two transformations, first signal has been normalised, second the logarithm of normalised signal has be taken. Each column of plots corresponds to a type of mixture ratio from equal contribution through to highly imbalanced.

A.5 K-means Classification for Two, Three and Five Contributors

Accompanying plots for section 5.4.2, comparing K-means cluster assignment on raw-signal and normalised-signal when the correct k has been chosen. (PCA plots are of $\log_{10}$(raw-signal) or $\log_{10}$(normalised-signal) as we have previously observed log transformed data is preferential for low dimensional visualisation)

A.5.1 K-means: Two Person Admixtures

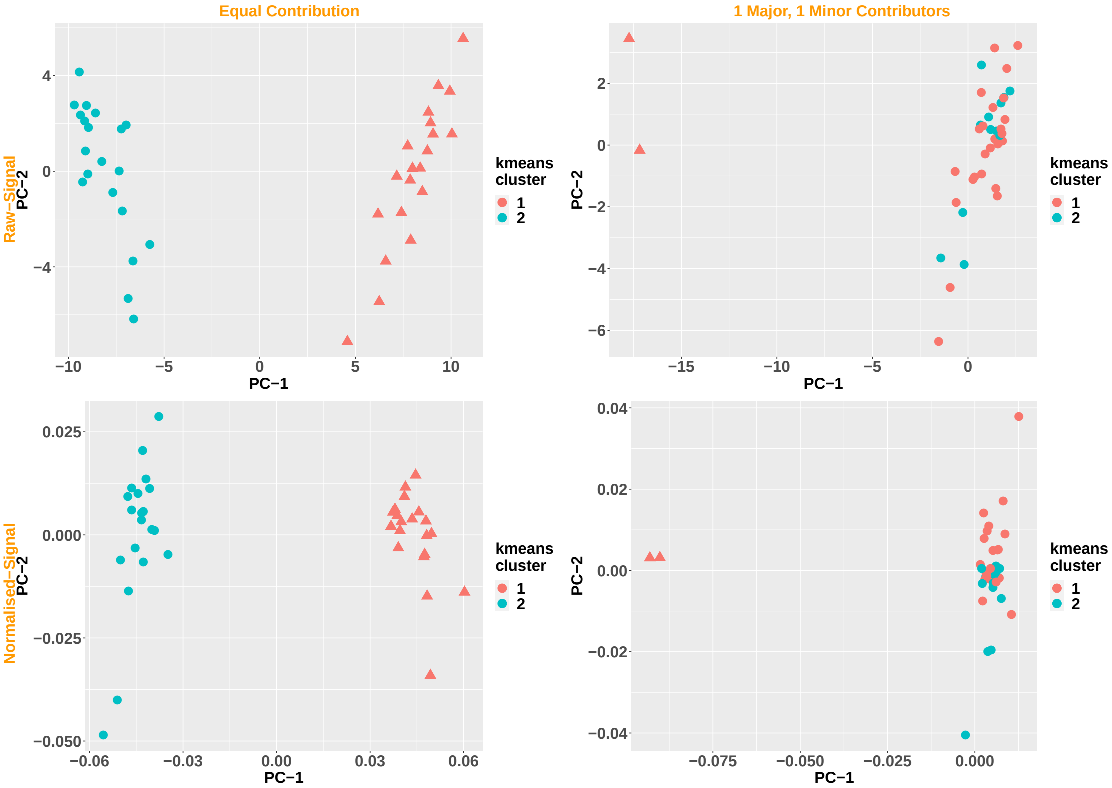


Figure A.5: PCA plots for simulated admixtures of 2 contributors in various ratios. Circles are Genotype 02 and triangles are Genotype 07. We have assumed the analyst determined k correctly, colour corresponds to the K-means cluster classification. The first column of plots is an admixture in equal ratio, K-means successfully assigns EPGs to their genotype. The second column is an admixture with one major and one minor contributors, K-means was unable to distinguish the minor contributor from the major and subsequently over clustered the major contributor.

A.5.2 K-means: Three Person Admixtures

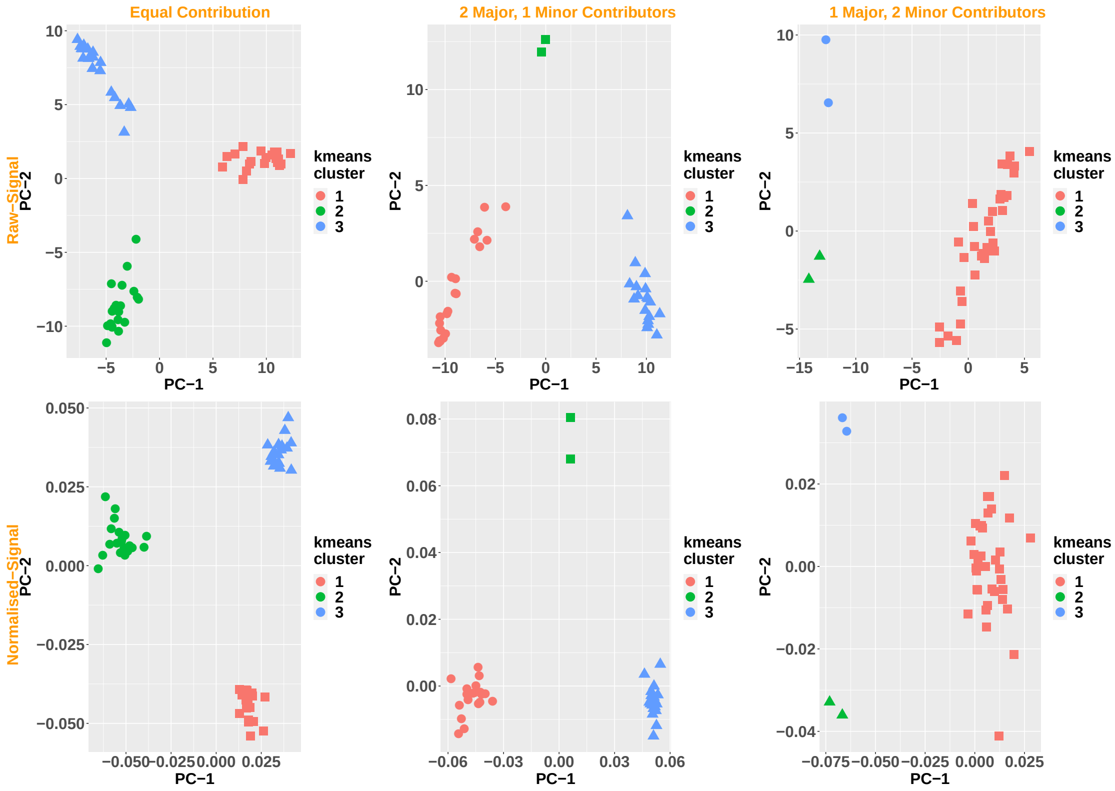


Figure A.6: PCA plots for simulated admixtures of 3 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 06, and squares are Genotype 07. We have assumed the analyst determined k correctly, colour corresponds to the K-means cluster classification. K-means has successfully assigned EPGs to their genotype for all three example admixtures.

A.5.3 K-means: Five Person Admixtures

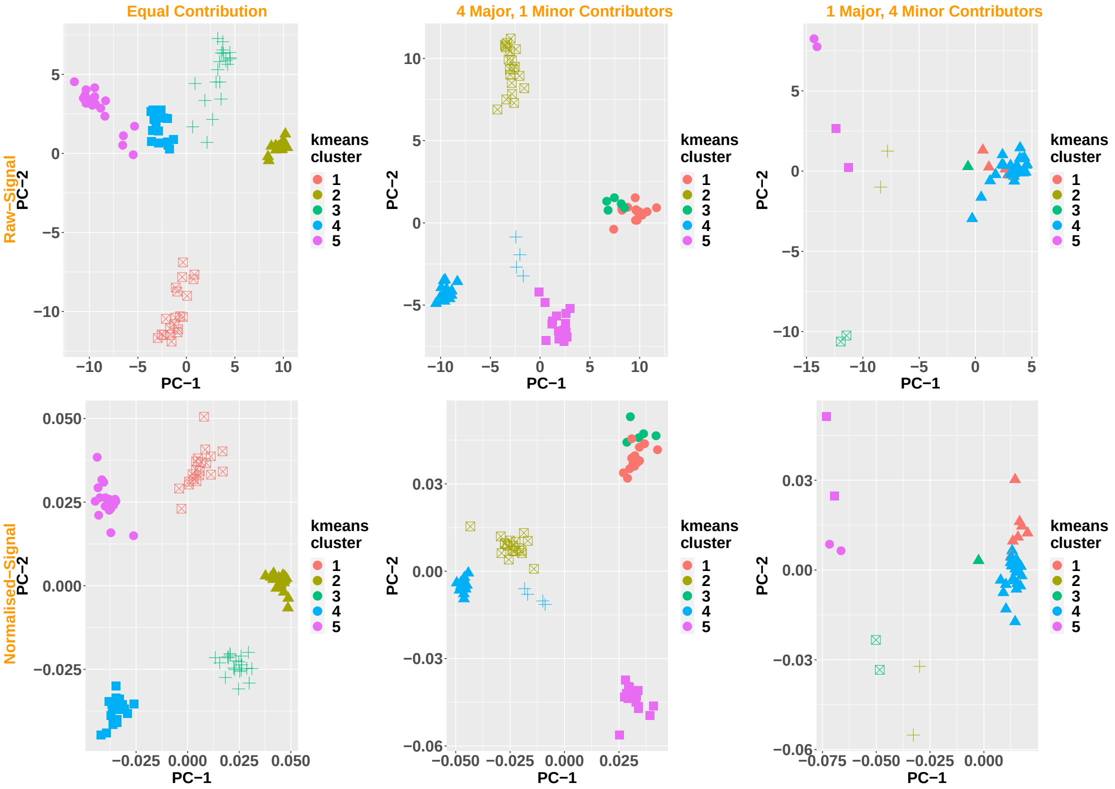


Figure A.7: PCA plots for simulated admixtures of 5 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 02, squares are Genotype 05, crosses are Genotype 06 and boxes with an x are Genotype 07. We have assumed the analyst determined k correctly, colour corresponds to the K-means cluster classification. The first column of plots is an admixture in equal ratio, K-means successfully assigns EPGs to their genotype. K-means incorrectly clusters EPGs for imbalanced and highly imbalanced admixtures of five contributors (columns two and three).

A.6 Mclust Classification for Two, Three and Five Contributors

Accompanying plots for section 5.4.3, comparing mclust cluster assignment on $\log_{10}$(raw-signal) and $\log_{10}$(normalised-signal).

A.6.1 Mclust: Two Person Admixtures

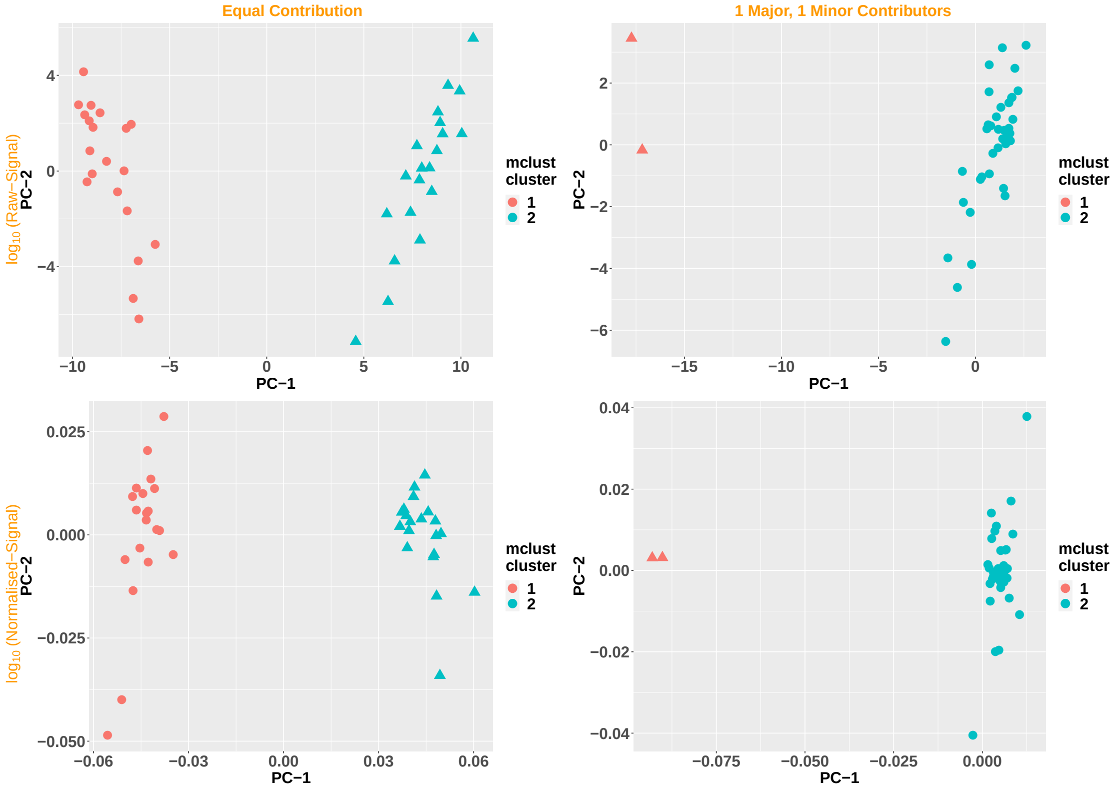


Figure A.8: PCA plots for simulated admixtures of 2 contributors in various ratios. Circles are Genotype 02 and triangles are Genotype 07. `Mclust` has correctly determined the number of clusters and EPG classification for both balance and imbalanced two person admixture examples, irrelevant of which data transformation was applied.

A.6.2 Mclust: Three Person Admixtures

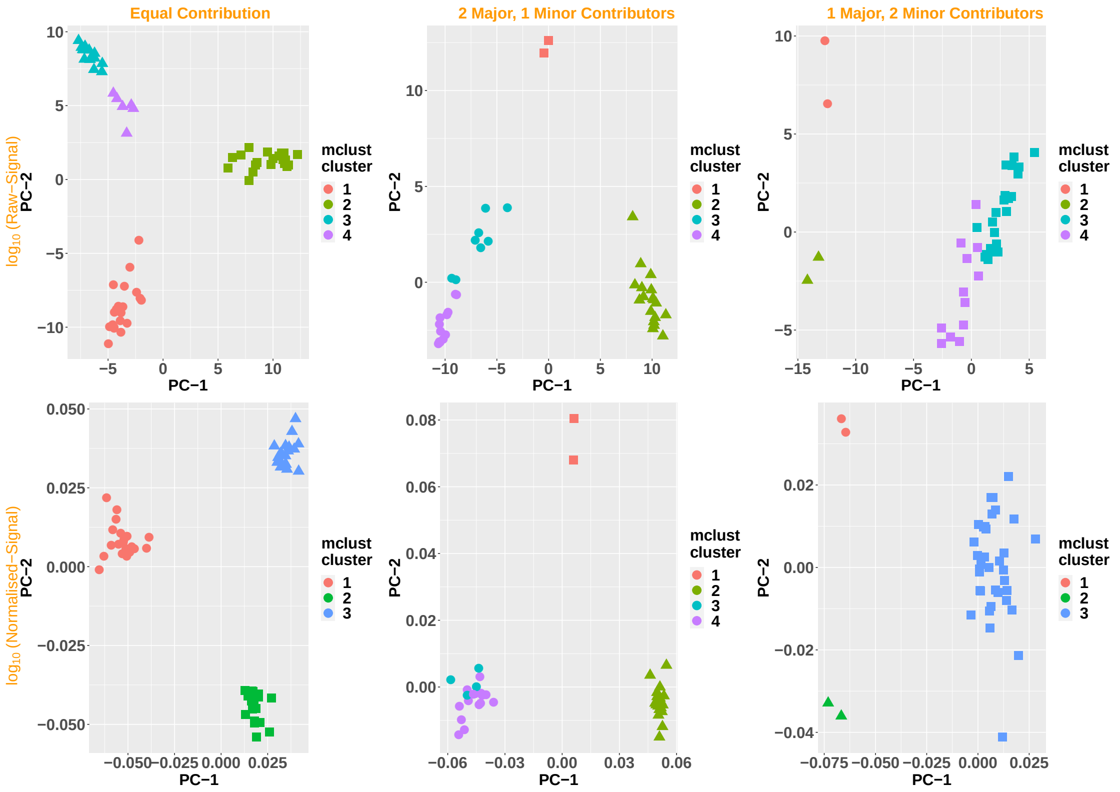


Figure A.9: PCA plots for simulated admixtures of 3 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 06, and squares are Genotype 07. We see consistent over-clustering when using log transformed signal. When using twice transformed signal `Mclust` correctly determined the number of clusters and EPG classification for admixtures in equal contribution and highly imbalanced (column one and three).

A.6.3 Mclust: Five Person Admixtures

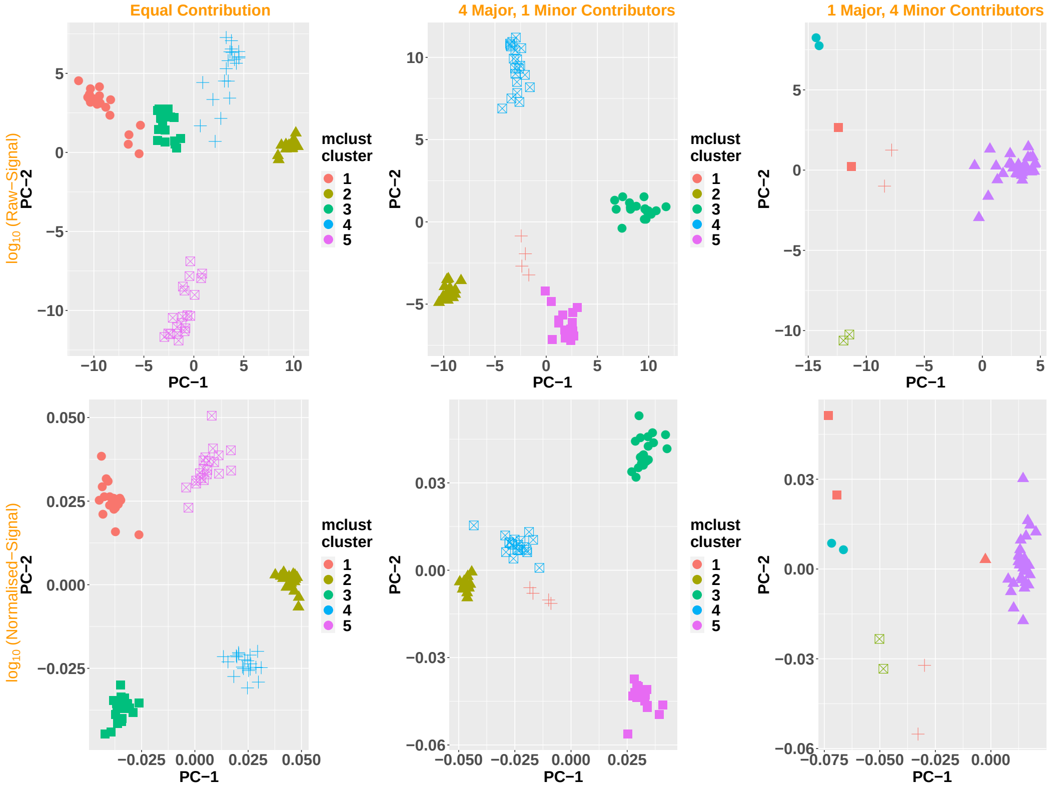


Figure A.10: PCA plots for simulated admixtures of 5 contributors in various ratios. Circles are Genotype 01, triangles are Genotype 02, squares are Genotype 05, crosses are Genotype 06 and boxes with an x are Genotype 07. `Mclust` has correctly determined the number of clusters and EPG classification for both balance and imbalanced five person admixture examples, irrelevant of which data transformation was applied (columns one and two). When handling the highly imbalanced five person example, `Mclust` has incorrectly assigned two minor contributors to one cluster when using log transformed data. When working with the twice transformed data `Mclust` incorrectly assigns three distinct genotypes to one cluster, two minor contributors and one EPG from the major contributor.

Bibliography

[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015.

[2] H. Abdi and L. J. Williams. Principal component analysis, 2010.

[3] B. Abu-Jamous, R. Fa, and A. K. Nandi. *Integrative Cluster Analysis in Bioinformatics.* John Wiley & Sons, Ltd, Chichester, UK, Mar 2015.

[4] L. E. Alfonse, A. D. Garrett, D. S. Lun, K. R. Duffy, and C. M. Grgicak. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Science International: Genetics*, 32:62–70, Jan 2018.

[5] E. Alladio, M. Omedei, S. Cisana, G. D'Amico, D. Caneparo, M. Vincenti, and P. Garofano. DNA mixtures interpretation – A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples. *Forensic Science International: Genetics*, 37:143–150, Nov 2018.

[6] A. Ambers, R. Wiley, N. Novroski, and B. Budowle. Direct PCR amplification of DNA from human bloodstains, saliva, and touch samples collected with microFLOQ ® swabs. *Forensic Science International: Genetics*, 32:80–87, Jan 2018.

[7] K. Anslinger, M. Graw, and B. Bayer. Deconvolution of blood-blood mixtures using DEPArrayTM separated single cell STR profiling. *Rechtsmedizin*, 29(1):30–40, Feb 2019.

[8] D. J. Balding and J. Buckleton. Interpreting low template DNA profiles. *Forensic Science International: Genetics*, 4(1):1–10, Dec 2009.

[9] B. Barnes and J. Dupré. *Genomes and What to Make of Them.* University of Chicago Press, 2008.

[10] E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, Jan 2019.

[11] M. A. Bhinder, M. Y. Zahoor, H. Sadia, M. Qasim, R. Perveen, G. M. Anjum, M. Iqbal, N. Ullah, W. Shehzad, M. Tariq, and A. M. Waryah. Evaluation of the facial nerve and internal auditory canal cross-sectional areas on three-dimensional fast imaging employing steady-state acquisition magnetic resonance imaging in Bell's palsy. *Turkish Journal of Medical Science*, 48(3), Jun 2018.

[12] C. Biémont and C. Vieira. Junk DNA as an evolutionary force. *Nature*, 443(7111):521–524, Oct 2006.

[13] Ø. Bleka, G. Storvik, and P. Gill. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Science International: Genetics*, 21:35–44, Mar 2016.

[14] L. A. Borsuk, K. B. Gettings, C. R. Steffen, K. M. Kiesler, and P. M. Vallone. Sequence-based US population data for the SE33 locus. *Electrophoresis*, 39(21):2694–2701, Nov 2018.

[15] J.-A. Bright, D. Taylor, J. M. Curran, and J. S. Buckleton. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2):296–304, Feb 2013.

[16] J.-A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, and J. Buckleton. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. *Forensic Science International: Genetics*, 23:226–239, Jul 2016.

[17] C. Brookes, J.-A. Bright, S. Harbison, and J. Buckleton. Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1):58–63, Jan 2012.

[18] S. Brück, H. Evers, F. Heidorn, U. Müller, R. Kilper, and M. A. Verhoff. Single Cells for Forensic DNA Analysis-From Evidence Material to Test Tube. *Journal of Forensic Sciences*, 56(1):176–180, Jan 2011.

[19] J. Buckleton and J. Curran. A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics*, 2(4):343–348, Sep 2008.

[20] B. Budowle, A. J. Onorato, T. F. Callaghan, A. D. Manna, A. M. Gross, R. A. Guerrieri, J. C. Luttman, and D. L. McClure. Mixture Interpretation: Defining the Relevant Features for Guidelines for the Assessment of Mixed DNA Profiles in Forensic Casework. *Journal of Forensic Sciences*, 54(4):810–821, Jul 2009.

[21] J. Butler and C. Hill. Scientific issues with analysis of low amounts of dna. *Profiles in DNA*, 13(1), 2010.

[22] J. M. Butler. Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. *Journal of Forensic Sciences*, 51(2):253–265, Mar 2006.

[23] J. M. J. Butler. *Forensic DNA typing: biology, technology, and genetics of STR markers*, volume 1. Elsevier, 2005.

[24] S. E. Cavanaugh and A. S. Bathrick. Direct PCR amplification of forensic touch and other challenging DNA samples: A review. *Forensic Science International: Genetics*, 32:40–49, Jan 2018.

[25] K. W. Y. Chong, Y. Wong, B. K. Ng, W. S. H. Lim, A. R. Rosli, and C. K. C. Syn. A practical study on direct PCR amplification using the GlobalFiler™ PCR Amplification Kit on human bloodstains collected with microFLOQ™ Direct swabs. *Forensic Science International*, 300:43–50, Jul 2019.

[26] T. Clayton, J. Whitaker, R. Sparkes, and P. Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1):55–70, Jan 1998.

[27] M. D. Coble and J.-A. Bright. Probabilistic genotyping software: An overview. *Forensic Science International: Genetics*, 38:219–224, Jan 2019.

[28] N. R. Council et al. *DNA Technology in Forensic Science.* National Academies Press, Washington, D.C., Jan 1992.

[29] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of forensic DNA mixtures with artefacts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):1–48, Jan 2015.

[30] N. E. Day. Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, 56(3):463, Dec 1969.

[31] L. Dean, Y. J. Kwon, M. K. Philpott, C. E. Stanciu, S. J. Seashols-Williams, T. Dawson Cruz, J. Sturgill, and C. J. Ehrhardt. Separation of uncompromised whole blood mixtures for single source STR profiling using fluorescently-labeled human leukocyte antigen (HLA) probes and fluorescence activated cell sorting (FACS). *Forensic Science International: Genetics*, 17:8–16, Jul 2015.

[32] G. M. Dembinski, C. Sobieralski, and C. J. Picard. Estimation of the number of contributors of theoretical mixture profiles based on allele counting: Does increasing the number of loci increase success rate of estimates? *Forensic Science International: Genetics*, 33:24–32, Mar 2018.

[33] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, Jan 2021.

[34] A. W. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, pages 362–375, 1965.

[35] G. Elgar. The Evolution of the Genome * Edited by T. Ryan Gregory * Elsevier Academic Press, Inc., London, UK; 2005; ISBN 0-12-301463-8; 768 pp.; 39.99; Hardback. *Briefings in Functional Genomics and Proteomics*, 4(4):377–378, Feb 2006.

[36] N. Etemadi. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 55(1):119–122, Feb 1981.

[37] C. Feng, S. Liu, H. Zhang, R. Guan, D. Li, F. Zhou, Y. Liang, and X. Feng. Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study. *International Journal of Molecular Sciences*, 21(6):2181, Mar 2020.

[38] I. Findlay, A. Taylor, P. Quirke, R. Frazier, and A. Urquhart. DNA fingerprinting from single cells. *Nature*, 389(6651):555–556, Oct 1997.

[39] P. Garofano, D. Caneparo, G. D'Amico, M. Vincenti, and E. Alladio. An alternative application of the consensus method to DNA typing interpretation for Low Template-DNA mixtures. *Forensic Science International: Genetics Supplement Series*, 5:e422–e424, Dec 2015.

[40] C. Gedik, S. Ewen, and A. Collins. Single-cell Gel Electrophoresis Applied to the Analysis of UV-C Damage and Its Repair in Human Cells. *International Journal of Radiation Biology*, 62(3):313–320, Jan 1992.

[41] T. Geng, R. Novak, and R. A. Mathies. Single-Cell Forensic Short Tandem Repeat Typing within Microfluidic Droplets. *Analytical Chemistry*, 86(1):703–712, Jan 2014.

[42] A. J. Gibb, A.-L. Huell, M. C. Simmons, and R. M. Brown. Characterisation of forward stutter in the AmpFlSTR® SGM Plus® PCR. *Science & Justice*, 49(1):24–31, Mar 2009.

[43] P. Gill. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, Jan 2005.

[44] P. Gill, C. Brenner, J. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. Schneider, and B. Weir. DNA commission of the International Society

of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2-3):90–101, Jul 2006.

[45] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, and J. Lambert. Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics*, 2(2):91–103, Mar 2008.

[46] P. Gill, L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, and B. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International: Genetics*, 6(6):679–688, Dec 2012.

[47] P. Gill and H. Haned. A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Science International: Genetics*, 7(2):251–263, Feb 2013.

[48] P. Gill, J. Whitaker, C. Flaxman, N. Brown, and J. Buckleton. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, 112(1):17–40, Jul 2000.

[49] S. Gittelson, C. R. Steffen, and M. D. Coble. Low-template dna: A single dna analysis or two replicates? *Forensic science international*, 264:139–145, Jul 2016.

[50] C. Goodall and I. T. Jolliffe. Principal Component Analysis. *Technometrics*, 30(3):351, Aug 1988.

[51] R. M. Goor, L. F. Neall, D. Hoffman, and S. T. Sherry. A mathematical approach to the analysis of multiplex dna profiles. *Bulletin of mathematical biology*, 73(8):1909–1931, 2011.

[52] P. Grabusts. Distance Metrics Selection Validity in Cluster Analysis. *Scientific Journal of Riga Technical University. Computer Sciences*, 45(1), Jan 2011.

[53] C. M. Grgicak, S. Karkar, X. Yearwood-Garcia, L. E. Alfonse, K. R. Duffy, and D. S. Lun. A large-scale validation of NOCIt's a posteriori probability of the number of contributors and its integration into forensic interpretation pipelines. *Forensic Science International: Genetics*, 47:102296, Jul 2020.

[54] H. Haned, P. Gill, K. Lohmueller, K. Inman, and N. Rudin. Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations. *Science & Justice*, 56(2):104–108, Mar 2016.

[55] D. R. Hares. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Science international. Genetics*, 17:33–34, Jul 2015.

[56] R. Hedell, C. Dufva, R. Ansell, P. Mostad, and J. Hedman. Enhanced low-template DNA analysis conditions and investigation of allele dropout patterns. *Forensic Science International: Genetics*, 14:61–75, Jan 2015.

[57] C. R. Hill, D. L. Duewer, M. C. Kline, M. D. Coble, and J. M. Butler. U.S. population data for 29 autosomal STR loci. *Forensic Science International: Genetics*, 7(3):e82–e83, May 2013.

[58] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[59] K. Huffman, E. Hanson, and J. Ballantyne. Recovery of single source DNA profiles from mixtures by direct single cell subsampling and simplified micromanipulation. *Science & Justice*, 61(1):13–25, Jan 2021.

[60] K. Inman, N. Rudin, K. Cheng, C. Robinson, A. Kirschner, L. Inman-Semerau, and K. E. Lohmueller. Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles. *BMC Bioinformatics*, 16(1):298, Dec 2015.

[61] J. Isaacson, E. Schwoebel, A. Shcherbina, D. Ricke, J. Harper, M. Petrovick, J. Bobrow, T. Boettcher, B. Helfer, C. Zook, and E. Wack. Robust detection of individual forensic profiles in DNA mixtures. *Forensic Science International: Genetics*, 14:31–37, Jan 2015.

[62] A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, Jun 2010.

[63] L. A. Johnson and J. A. Ferris. Analysis of postmortem DNA degradation by single-cell gel electrophoresis. *Forensic Science International*, 126(1):43–47, Mar 2002.

[64] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, Apr 2016.

[65] E. Karantzali, P. Rosmaraki, A. Kotsakis, M.-G. Le Roux-Le Pajolec, and G. Fitsialos. The effect of FBI CODIS Core STR Loci expansion on familial DNA database searching. *Forensic Science International: Genetics*, 43:102129, Nov 2019.

[66] Y. Kasai, S. Sakuma, and F. Arai. Isolation of single motile cells using a high-speed picoliter pipette. *Microfluidics and Nanofluidics*, 23(2):18, Feb 2019.

[67] H. Kelly, J.-A. Bright, J. Curran, and J. Buckleton. The interpretation of low level DNA mixtures. *Forensic Science International: Genetics*, 6(2):191–197, Mar 2012.

[68] S. A. Kim, J. A. Yoon, M. J. Kang, Y. M. Choi, S. J. Chae, and S. Y. Moon. An efficient and reliable DNA extraction method for preimplantation genetic diagnosis: a comparison of allele drop out and amplification rates using different single cell lysis methods. *Fertility and Sterility*, 92(2):814–818, Aug 2009.

[69] T. Konopka. *umap: Uniform Manifold Approximation and Projection*, 2020. R package version 0.2.4.1.

[70] S. Köster, F. E. Angilè, H. Duan, J. J. Agresti, A. Wintner, C. Schmitz, A. C. Rowat, C. A. Merten, D. Pisignano, A. D. Griffiths, and D. A. Weitz. Drop-based microfluidic devices for encapsulation of single cells. *Lab on a Chip*, 8(7):1110, 2008.

[71] B. B. Lake, S. Chen, M. Hoshi, N. Plongthongkum, D. Salamon, A. Knoten, A. Vijayan, R. Venkatesh, E. H. Kim, D. Gao, J. Gaut, K. Zhang, and S. Jain. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nature Communications*, 10(1):2832, Dec 2019.

[72] D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, P. Yaswen, A. Goga, and Z. Werb. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526(7571):131–135, Oct 2015.

[73] S. Leonov, E. Zemskova, and P. Ivanov. LMD-assisted single cell DNA typing of forensic biological evidence: Issues of the cell type and sample condition. *Forensic Science International: Genetics Supplement Series*, 3(1):e47–e48, Dec 2011.

[74] C. Li, B. Qi, A. Ji, X. Xu, and L. Hu. The combination of single cell micromanipulation with LV-PCR system and its application in forensic science. *Forensic Science International: Genetics Supplement Series*, 2(1):516–517, Dec 2009.

[75] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245, Mar 2019.

[76] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[77] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, and K. Tamaki. Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model. *PloS one*, 12(11):e0188183, Nov 2017.

[78] L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, Sep 2018.

[79] G. J. McLachlan, R. W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, Mar 2002.

[80] G. McVean. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, 5(10):e1000686, Oct 2009.

[81] M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, Sep 2002.

[82] U. J. Mönich, K. Duffy, M. Médard, V. Cadambe, L. E. Alfonse, and C. Grgicak. Probabilistic characterisation of baseline noise in STR profiles. *Forensic Science International: Genetics*, 19:107–122, Nov 2015.

[83] S. K. Ng, G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng. A Mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22(14):1745–1752, Jul 2006.

[84] K. Nordhausen and H. Oja. Independent component analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1440, Sep 2018.

[85] N. Patterson, A. L. Price, and D. Reich. Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[86] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, Nov 1901.

[87] M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, and B. W. Duceman. Validating TrueAllele ® DNA Mixture Interpretation* ,†. *Journal of Forensic Sciences*, 56(6):1430–1447, Nov 2011.

[88] K. C. Peters, H. Swaminathan, J. Sheehan, K. R. Duffy, D. S. Lun, and C. M. Grgicak. Production of high-fidelity electropherograms results in improved and consistent DNA interpretation: Standardizing the forensic validation process. *Forensic Science International: Genetics*, 31:160–170, Nov 2017.

[89] W. Piyamongkol. Detailed investigation of factors influencing amplification efficiency and allele drop-out in single cell PCR: implications for preimplantation genetic diagnosis. *Molecular Human Reproduction*, 9(7):411–420, Jul 2003.

[90] R. Press. DNA Mixtures: A Forensic Science Explainer, 2019.

[91] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, F. Álvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, M. Farfán, M. Fenger-Grøn, J. García-Ganivet, E. González-Moya, L. Hombreiro, M. Lareu, B. Martínez-Jarreta, S. Merigioli, P. Milans del Bosch, N. Morling, M. Muñoz-Nieto, E. Ortega-González, S. Pedrosa, R. Pérez, C. Solís, I. Yurrebaso, and P. Gill. Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles. *Forensic Science International: Genetics*, 9:47–54, Mar 2014.

[92] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8(1):111, 2007.

[93] R. Puch-Solis. A dropin peak height model. *Forensic Science International: Genetics*, 11:80–84, Jul 2014.

[94] R. Puch-Solis and T. Clayton. Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software. *Forensic Science International: Genetics*, 11:220–228, Jul 2014.

[95] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, and D. Balding. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics*, 7(5):555–563, Sep 2013.

[96] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.

[97] M. Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, Mar 2008.

[98] L. Roewer. DNA fingerprinting in forensics: past, present, future. *Investigative Genetics*, 4(1):22, 2013.

[99] C. T. Sanders, N. Sanchez, J. Ballantyne, and D. A. Peterson. Laser Microdissection Separation of Pure Spermatozoa from Epithelial Cells for Short Tandem Repeat Analysis*. *Journal of Forensic Sciences*, 51(4):748–757, Jul 2006.

[100] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.

[101] L. Scrucca, M. Fop, B. Murphy, T., and E. Raftery, Adrian. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289, 2016.

[102] N. Sheth, H. Swaminathan, A. J. Gonzalez, K. R. Duffy, and C. M. Grgicak. Towards developing forensically relevant single-cell pipelines by incorporating direct-to-PCR extraction: compatibility, signal quality, and allele detection. *International Journal of Legal Medicine*, Jan 2021.

[103] J. E. Sim, S. J. Park, H. C. Lee, S.-Y. Kim, J. Y. Kim, and S. H. Lee. High-Throughput STR Analysis for DNA Database Using Direct PCR ,. *Journal of Forensic Sciences*, 58(4):989–992, Jul 2013.

[104] H. Swaminathan, A. Garg, C. M. Grgicak, M. Medard, and D. S. Lun. CEESIt: A computational tool for the interpretation of STR mixtures. *Forensic Science International: Genetics*, 22:149–160, May 2016.

[105] H. Swaminathan, C. M. Grgicak, M. Medard, and D. S. Lun. NOC It : A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. *Forensic Science International: Genetics*, 16:172–180, May 2015.

[106] H. Swaminathan, M. O. Qureshi, C. M. Grgicak, K. Duffy, and D. S. Lun. Four model variants within a continuous forensic DNA mixture interpretation framework: Effects on evidential inference and reporting. *PloS one*, 13(11):e0207599, Nov 2018.

[107] SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, 2010.

[108] J. Y. Y. Tan, Y. P. Tan, S. Ng, A. S. Tay, Y. H. Phua, W. J. Tan, T. Y. R. Ong, L. M. Chua, and C. K. C. Syn. A preliminary evaluation study of new generation multiplex STR kits comprising of the CODIS core loci and the European Standard Set loci. *Journal of Forensic and Legal Medicine*, 52:16–23, Nov 2017.

[109] D. Taylor, J.-A. Bright, and J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5):516–528, Sep 2013.

[110] D. Taylor, J. Buckleton, and I. Evett. Testing likelihood ratios produced from complex DNA profiles. *Forensic Science International: Genetics*, 16:165–171, May 2015.

[111] J. E. Templeton, D. Taylor, O. Handt, and A. Linacre. Typing DNA profiles from previously enhanced fingerprints using direct PCR. *Forensic Science International: Genetics*, 29:276–282, Jul 2017.

[112] M. Üzümcü, A. F. Frangi, M. Sonka, J. H. C. Reiber, and B. P. F. Lelieveldt. ICA vs. PCA Active Appearance Models: Application to Cardiac MR Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 451–458. Springer, 2003.

[113] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[114] M. Vandewoestyne and D. Deforce. Laser capture microdissection in forensic research: a review. *International Journal of Legal Medicine*, 124(6):513–521, Nov 2010.

[115] M. Vandewoestyne, F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce. Evaluation of three DNA extraction protocols for forensic STR typing after laser capture microdissection. *Forensic Science International: Genetics*, 6(2):258–262, Mar 2012.

[116] S. B. Vilsen, T. Tvedebrink, P. S. Eriksen, C. Bøsting, C. Hussing, H. S. Mogensen, and N. Morling. Stutter analysis of complex STR MPS data. *Forensic Science International: Genetics*, 35:107–112, Jul 2018.

[117] P. S. Walsh, N. J. Fildes, and R. Reynolds. Sequence Analysis and Characterization of Stutter Products at the Tetranucleotide Repeat Locus VWA. *Nucleic Acids Research*, 24(14):2807–2812, Jul 1996.

[118] D. Y. Wang, S. Gopinath, R. E. Lagacé, W. Norona, L. K. Hennessy, M. L. Short, and J. J. Mulero. Developmental validation of the globalfiler® express pcr amplification kit: a 6-dye multiplex assay for the direct amplification of reference samples. *Forensic Science International: Genetics*, 19:148–155, 2015.

[119] Y. Wang, X. Tang, X. Feng, C. Liu, P. Chen, D. Chen, and B.-F. Liu. A microfluidic digital single-cell assay for the evaluation of anticancer drugs. *Analytical and Bioanalytical Chemistry*, 407(4):1139–1148, Feb 2015.

[120] N. E. Weiler, A. S. Matai, and T. Sijen. Extended PCR conditions to reduce dropout frequencies in low template STR typing including unequal mixtures. *Forensic Science International: Genetics*, 6(1):102–107, Jan 2012.

[121] J. Weusten and J. Herbergs. A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications. *Forensic Science International: Genetics*, 6(1):17–25, Jan 2012.

[122] J. H. Wolfe. Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research*, 5(3):329–350, Apr 1970.

[123] Y. Wong, B. K. Ng, K. W. Y. Chong, W. S. H. Lim, A. R. Rosli, J. J. Tay, W. W. X. Lim, A. Q. H. Ng, A. G. L. Tan, E. H. Q. Ng, and C. K. C. Syn. A modified direct PCR amplification method using the GlobalFiler™ PCR Amplification Kit on bloodstains collected using microFLOQ™ direct swabs. *Forensic Science International: Genetics Supplement Series*, 7(1):30–32, Dec 2019.

[124] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, Oct 2001.

[125] L. Zhang, X. Cui, K. Schmitt, R. Hubert, W. Navidi, and N. Arnheim. Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences*, 89(13):5847–5851, Jul 1992.