STATE OF NEW YORK

COUNTY COURT COUNTY OF SCHENECTADY

THE PEOPLE OF THE STATE OF NEW YORK

- against -

AFFIRMATION IN
RESPONSE TO MOTION

John Wakefield ,

Indictment #A-812-29
DA File #23796

Defendant.

Peter H. Willis, ESQ., an attorney duly admitted to practice before the Courts of New York, affirms under penalties of perjury pursuant to CPLR Section 2106 that the following statements are true:

A. I am an Assistant District Attorney in the Office of ROBERT M. CARNEY, District Attorney of the County of Schenectady, New York.

B. I submit this Affirmation in opposition to defendant's Motion, dated March 28, 2014.

C. This Affirmation is based upon information and belief, the source of my information and the basis for my belief being an examination of the files maintained by the Office of the District Attorney.

1. This affirmation is submitted in response to the defendant's notice of motion requesting the preclusion of testimony related to the analysis of certain biological evidence samples performed using the TrueAllele ™ Forensic Casework System a product of the Cybergenetics Corporation. Dr. Mark Perlin is the founder of Cybergenetics and is chiefly responsible for the development of the TrueAllele system.

2. TrueAllele ™ is a computer program that analyzes the data produced when DNA is extracted from a biological sample. TrueAllele represents an advanced approach to forensic DNA analysis that is both more objective than traditional approaches and more informative. Analyses performed by TrueAllele have been subjected numerous internal validations, as well as published validation studies in collaboration with the New York State Police Forensic Identification Center and the Virginia Department of Forensic Science and an anticipated publication in collaboration with the Kern

Regional Crime Laboratory in Bakersfield, California. Cybergenetics has also participated in validation studies with the Suffolk County Crime Laboratory in Hauppauge, New York, the Allegheny County Medical Examiner's Office in Pittsburg, PA, and forensic laboratories in the United Kingdom and Australia.

3. TrueAllele has been subjected to evaluation, and admitted into evidence, by courts in California, Pennsylvania and Virginia as well as in the United Kingdom.

4. TrueAllele has been approved by the New York State Commission on Forensic Science and the Commission's DNA subcommittee for use in forensic casework.

5. It is currently being used in forensic DNA analysis by authorities in Virginia, California and Pennsylvania.

6. In this case numerous biological samples were collected by the Schenectady Police Department and the New York State Police Forensic Investigation Center (NYSPFIC)[1]. Many of these items, along with appropriate control samples, were then analyzed for the presence of DNA.

7. At the NYSPFIC each sample underwent a process called Polymerase Chain Reaction (PCR) which amplifies the amount of DNA, followed by Short Tandem Repeat (STR) analysis.

8. STR analysis is the fundamental principle upon which forensic DNA casework is based. In this process after the DNA has been amplified a significant number of times it is analyzed to see how often certain areas of its sequence repeat themselves at different locations. There are 15 standard locations; each called a locus, within the DNA chain that are observed.[2] An individual's DNA profile is represented by the number of repetitions observed at each location.

9. To determine the number of repeating sections of DNA at each location the amplified material is injected into a DNA sequencer and subjected to a process called capillary electrophoresis.

---

[1] Some biological samples were collected directly by members of the Police Department. Those samples were submitted to the New York State Police Forensic Investigation Center in the form of cotton swabs that had come into contact with the certain specific items of evidence. Others, such as the samples at issue, were collected by employees of the New York State Police Forensic Center from evidence items submitted directly to the laboratory.

[2] The analyst also observes a locus called amelogin which identifies the sex of the contributor.

This process separates the DNA molecules by length, while a laser detects fluorescence intensity. The more genetic material present at a specific area, the greater the observed fluorescence. The end result of this process is the creation of what is known as an electropherogram (EPG) for each locus. (Fig 1. below is an example of an EPG from a generic DNA profile unrelated to this case at the D8S1179 locus) An EPG is in essence a line chart. The *y*-axis is a value of Relative Fluorescent Units (RFUs) which measures the intensity with which the sample was detected and derivatively the amount of genetic material present. The *x*-axis indicates the number of times the DNA repeated itself at the specific locus. An allele is represented by a peak in the graph.

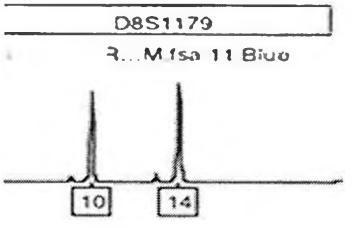


Fig. 1

10. As can be seen in figure 1 the electropherogram is showing two peaks, one at 10 and one at 14.

11. In the majority of instances a person has two numbers at each location; one inherited from their mother and one from their father. However in cases in which both parents contributed the same number, the profile will show only one number. The observed location (identified as D8S1179 in figure 1) is called a locus, and the different peaks are called alleles. The combination of an individual's allele pair at a specific locus is called a genotype. Therefore the genotype for this locus is a 10, 14. In a common DNA report each of the standard 15 locations will be listed in one column, and the individual-specific alleles in another. This is what is commonly referred to as an individual's DNA profile.

12. The statistical rarity of an individual's DNA profile is a combination of how rare (or common) each of their allele combinations are at each of the fifteen loci relative to the general

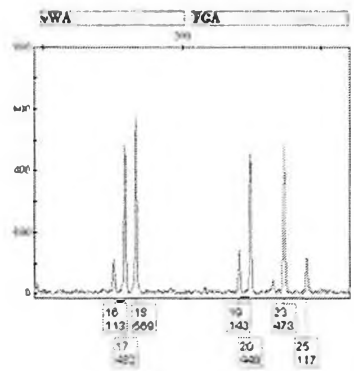
population.[3] The frequency with which each allele pair appears within the general population has been the subject of a considerable amount of scientific study over the last thirty years and is not in dispute in this case.

13. The above approach is a basic description of how to analyze a DNA profile of a biological sample containing a relatively large amount of DNA from a single contributor.

14. Figure 1 represents an allele pair that is likely from a single contributor. However the presence of two alleles could indicate the presence of two different contributors to the biological sample; one with a single 10 allele and one with a single 14 allele. In this example such a result is unlikely because the alleles both appear at relatively the same height, indicating they were detected with the same level of fluorescence and indicative of relatively equal amounts of genetic material. If they were from two different contributors the peaks would most likely be of observably different heights.

15. A biological sample with more than two contributors could show up to four different alleles.

Figure 2.



16. The issue of resolving a biological sample with more than one contributor is where TrueAllele provides a significant advancement over traditional approaches.

[3] This is more commonly referred to as a 'match statistic' in the context of forensic DNA casework and is meant to convey the strength of the probability that a profile comes from a given person relative to a coincidence.

17. As an example in figure 2 we can see two different loci, vWA and FGA. At vWA we can see three different peaks (a 16, 17 and 18) and at FGA we see four (19, 20, 23 and 25). This means that there are at least two contributors to both samples.

18. In our case the samples at issue are all DNA mixtures from two or more people.[4]

19. After the analyst at the NYSPFIC performed PCR and STR analysis of the samples in this case she issued conclusions for those mixture samples.

20. In each sample, except the swab from the outside rear portion of the victim's shirt collar, the analyst concluded that the victim was a major contributor to the sample. With regard to the shirt collar the analyst could not identify a major contributor.

21. With regard to the presence of the defendant's DNA in those mixtures, some of the conclusions included a statistical calculation called a Combined Probability of Inclusion (CPI). In this instance the analyst concluded that the defendant could not be excluded as a contributor to the sample and then expressed the odds that a randomly selected person could be included in the profile.[5]

22. Based on research performed by the People it was discovered that the TrueAllele™ program represented the most advance scientific approach to analyzing biological samples of mixed DNA.

23. The computer data produced by the DNA sequencing machine was then sent to Cybergenetics for additional analysis.[6]

24. Before discussing the approach taken by Cybergenetics through TrueAlle™ for assigning a statistical weight to the contributors to a sample of mixed DNA, it is important to understand

---

[4] The number of contributors is nominally determined by the maximum number of alleles at a specific locus within a sample. In a sample with four alleles present at a locus the number contributors would be expressed as two or more because that is the minimum number of contributors needed to reach four alleles. The sample could contain more contributors if one or more contributed only a single allele, or several of the contributors shared alleles. If that occurred it is possible that a three person mixture could result in the presence of only four alleles.

[5] The odds were 1 in 422 with respect to a sample taken from the victim's forearme and 1 in 1088 from a sample taken from the rear portion of the victim's shirt collar.

[6] In the form of a computer program called GeneMapperID™

how the CPI statistics and the other conclusions were reached by the NYSPFIC analyst, and what the drawbacks are to this approach when analyzing a DNA mixture sample.

25. The analyst for the NYSPFIC reported two CPI statistics, one for a sample of biological material taken from the victim's right forearm (identified as item 52F 1-2) and one taken from a cutting of the rear portion of the collar of the shirt the victim was wearing (identified as item 45A).

26. CPI is a simplistic formula that is used to assess the probability that an unrelated person's DNA profile who is not an actual contributor to a biological sample could be included (or not excluded) in the genetic profile culled from the sample.[7] In order to calculate a CPI statistic the analyst looks at the rate at which each allele appears within the general population, combines them in a standard equation, and arrives at the probability that a random and unrelated person could have contributed to the mixture.

27. This approach, while generally accepted, has severe limitations. The first limitation is that it treats each allele as equally likely to have come from either contributor. In looking at figure 2 at the FGA locus we can see four different peaks representing four different alleles. In this case it is readily apparent that two of the peaks, 19 and 25 share the same height and two others 20 and 23 share the same height. This indicates the presence of two donors to the sample, one with a 19, 25 genotype and one with a 20, 23.[8] However, when calculating a CPI statistic for this locus the formula assumes that it is equally likely that the contributors are 19, 20 and 23, 25; and 19, 23 and 20, 25.

28. This has the effect of greatly expanding the number of possible contributors to the sample by statistically including many more allele combinations than is supported by the evidence sample.

29. This effect is potentially beneficial and detrimental to a possible suspect. It is beneficial because it has the potential for rendering a match statistic that, because it includes the possibility of a greater number of allele combinations, is artificially lower. Additionally, by expanding the number of potential allele combinations it has also has the potential to include a non-contributor if his or her alleles

---

[7] This statistic can also be referred to as CPE or Combined Probability of Exclusion. CPE is the inverse of CPI and is expressed by the formula CPE = (1-CPI). As an example is the CPI was 1 out of 3 (1/3) CPE would be 2 out of 3 or 2/3.

[8] The amount of DNA a person contributes to a sample is relative to the intensity (in RFU's) present in the electropherogram.

happen to be present in the sample. Using the FGA locus from the electropherogram in figure 2 again as an example, a suspect with an allele combination of 19, 20 could not be excluded from sample even though it is unlikely that someone with that allele pair actually contributed to the sample.

Stochastic Thresholds and Drop-in/Drop-out

30. The other drawback of using a CPI/CPE method of analyzing a mixed sample of DNA is that it does not make use of all of the available data from the electropherogram.

31. The peaks in the electropherogram, as discussed above, are representations of the intensity with which the specific area of DNA is detected. The more DNA present, the greater the intensity and the higher the number of RFU's.

32. When a CPI statistic is reported the only peaks that are used in the calculation are those that rise above what is called a stochastic threshold of 150 RFU's. This means that peaks below that threshold are not included in the calculations.

33. Peaks representing alleles below that threshold are said to have "dropped-out" of the mixture. Drop out is a commonly confronted issue in cases in with low level samples of DNA.[9] Applying a stochastic threshold of 150 to the FGA locus in figure two would actually result in the exclusion from a CPI calculation of the alleles present at 19 and 25 because they are both under that threshold.

34. The effect of the use of the stochastic threshold is to effectively draw a line across the electropherogram at 150 RFU's; all of the peaks above the line, regardless of their height relative to each other, are equally weighted in the CPI calculation; all of the peaks below 150 are not included.

35. This approach means that large amounts of informative data are lost from the statistical analysis, especially from lower level samples of DNA.[10]

---

[9] Drop-in is the phenomena by which an allele, unassociated with any contributor to the sample, appears in the profile.

[10] By lower level amounts of DNA the people do not specifically refer to what is known as Low-Template DNA or LTDNA. LTDNA refers the analysis of extremely very low amounts of DNA through an increase number of PCR cycles. LTDNA analysis was not performed in this case.

36. This also leads to situations in which the lab will report the presence of alleles at a locus but will not include them in a CPI calculation. In addition to the stochastic threshold there is also what is called an analytic threshold that is set at 50 RFU's. When an analyst observes an allele above the analytic threshold, but below the stochastic threshold, it will be listed in their report, but the locus will be excluded from use in the CPI statistic. In this case that led to the exclusion of 10 out of the 15 loci from the sample taken from the outside portion of the rear part of the victim's shirt collar and 11 out of 15 from the sample taken from the victim's forearm.

37. This means that the majority of the information gained during the DNA analysis for these samples was not included in the statistical calculation. This is the essence of why a CPI match statistic does not do a good job of reporting the strength of the link between a potential contributor and a biological sample with a DNA mixture profile.

38. It is within this context that the People sought out a more advanced approach to analyzing the DNA evidence in this case and found Dr. Perlin's TrueAllele system.

TrueAllele

39. TrueAllele is a computer program that employs advanced statistical modeling that is specifically designed to deconvolute complex mixture samples. The program is a probabilistic model based on Bayes theorem and Markov Chain Monte Carlo ("MCMC") algorithms to infer a genotype at a specific locus from the reported data. The system is also accurately referred as a continuous model for forensic analysis.

Bayes Theorem

40. Bayes theorem is based on the premise of using observed data to update our beliefs that an event has occurred. When Bayes theorem is applied it produces what is known as a likelihood ratio (LR). A LR is a ratio of the probability of a hypothesis being true before observing a specific set of data, with the probability of the hypothesis being true after observing the data. When it is applied in the context of DNA analysis the likelihood ratio expresses the odds with which a specific pattern of alleles are present after observing the data, compared to the normal distribution of the alleles within the

population. In the case of forensic DNA analyses the hypotheses are that a specific person (usually victim or suspect) contributed to a biological sample versus the hypotheses that a random person contributed to the sample.

41. Before looking at the electropherogram for a locus the likelihood that any one specific allele will be present is equal to the normal distribution of the allele within the population. As an example suppose that at the FGA locus a 7 allele occurs 5% (.05) of the time within the general population. This is the probability of the 7 occurring without looking at the data, or rather the prior odds. In calculating our likelihood ration this becomes our denominator. Once the electropherogram is observed suppose we find that there is an 85% (.85) chance that a 7 allele was actually present. This is referred to as the posterior odds. This number now becomes our numerator and the likelihood ration becomes equal to .85/.05. This means that it is 17 times more likely that a 7 is present at the FGA locus than a random allele.

42. The basic formula is; LR = posterior odds/prior odds

43. A likelihood ratio above 1 shows support for the hypothesis that an allele is present, a ratio of 1 is inconclusive, and anything under 1 does not support the hypothesis.

44. In the modeling process itself TrueAllele uses MCMC based algorithms in order to infer the most likely genotype at a locus. In order to infer a genotype at a locus the system proposes thousands of different possible allele combinations. These potential combinations are created and evaluated using an MCMC algorithm. For each proposed combination the system assigns a probability based on how well it fits the actual data. The combinations that better explain the data are given a higher probability. Most proposals do not fit data well and are be given a probability of at or near 0. The end result is that the system infers a genotype by selecting the explanation with the highest probability.

45. Using the FGA locus from figure 2 again as an example; the system would likely assign a very high probability to the allele pairs 19, 25 and 20, 25. Assume that the system assigned a probability of 50% (.5) that a contributor with a 19,25 genotype is present at that locus. If the probability that a 19, 25 genotype would occur randomly within the population is 5% (.05) then using our LR formula

(LR=posterior odds/prior odds) we find that the LR =.5/.05 or 10. This means that it is ten times more likely that a person with a 19, 25 genotype contributed to the sample than a random person.

46. In addition to inferring the genotype with the highest probability the system also records the probabilities assigned to every genotype combination. To produce a LR in a case such as this the system compares the probability of the inferred genotype to the probability of that genotype occurring in the general population at the suspect's genotype.

47. If a suspect had a 19, 25 genotype the system would compute the likelihood ratio as stated above. However if the suspect had a different genotype the system would report the probability of that genotype occurring from within the thousands of proposed solutions. This is what is meant by comparing the ratio at the suspect's genotype. If the suspect had only a 19 (most likely a lower probability that the originally proposed 19, 25 combination, but higher than most other potential combinations) the system would find the probability with which a single 19 was contributed to the mixture and compare that to the odds of a single 19 in the general population. Because a single 19 somewhat fits the data, but is not the best explanation, assume it gets assigned a probability of 10% (.10) and the rate that a 19 occurs in the general population is 5% (.05); therefore the LR = .10/.05 or 2. This would indicate some, but not much, support for the hypothesis that the suspect contributed to the sample.

48. Suppose the suspect had a genotype of 10, 15 at the FGA locus, and that like the other hypothetical genotypes a 10, 15 occurs 5% of the time in the general population. Because a contributor with a 10, 15 genotype would be an extremely poor fit for figure 2 it would be assigned a probability at or near zero. If we assume the assigned probability is not zero (if it is the LR is always 0) but rather .05% our LR would be .0005/.05 or .01. This would not indicate support for the hypothesis that the suspect contributed to the sample.[11]

49. The fundamental difference between using a CPI based stochastic threshold statistic and TrueAllele is that CPI is conditioned on the idea that an allele is either definitively present or absent from

[11] All numbers and probabilities used here are hypothetical and are used for illustrative purposes only; they do not represent actual probabilities.

a mixture, whereas TrueAllele determines the probability with which an allele is present. To make this inference TrueAllele looks at all of the reported data from the electropherogram. This includes every potential allele peak down to a level of approximately 10 RFU's. TrueAllele also takes into account the differences in allele peak heights in order to properly distribute alleles between potential contributors.

50. The defendant raises several issues with the methods used by TrueAllele that the People will address further on in this response. However in his description of the system he inaccurately portrays several aspects of the system that need to be brought to the Court's attention.

51. The first is contained in paragraph 30 of defense counsel's affirmation in which it is stated that "The final result is a number in the form of a likelihood ratio which represents a ration between the probability of a match between the DNA profile inferred by TrueAllele and a known reference DNA profile and the probability of a coincidental match between the DNA profile and a randomly selected individual within a particular population. *To arrive at the number TrueAllele calculates probabilities using undisclosed computerized mathematical formulas not previously used in forensic analysis."*

52. Because of how the paragraph is written the People are unable to discern exactly what the defense claims are the "undisclosed computerized formulas not previously used in forensic analysis." The basic formula for calculating a likelihood ratio has been present for hundreds of years and has been the recommended statistic for mixtures since 2006.[12]

53. The use of Markov Chain Monte Carlo algorithms also cannot be considered undisclosed or not previously used in forensic analysis. MCMC methods were developed by physicists working in Los Alamos, New Mexico during World War II and have been used continuously since then in a wide variety of disciplines.[13]

54. In paragraph 31and 32 the defendant discusses and contrasts the approach taken by TrueAllele verses conventional forensic analysis with respect to stochastic thresholds and drop-out. The

---

[12] Gill et al. (2006) DNA Commission of the International Society of Forensic Genetics: Recommendations of the interpretation of mixtures. *Forensic Sci. Int.* 160: 90-101

[13] Robert, C., Casella, G., A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science1* (2011) vol. 0, No. 00-14.

defense is accurate that TrueAllele does not use a stochastic threshold for "scoring" alleles. The reason that conventional analysis employees a stochastic threshold is because "it may be difficult to distinguish, true low-level peaks from technical artifacts." That statement, which the People agree with, necessarily implies that true low level peaks, representing real information, do exist below the stochastic threshold. It also highlights the problem in conventional analysis that much data is lost because of a level of uncertainty. Conventional analysis disregards data while knowing that it contains accurate information.[14]

55. The defendant criticizes TrueAllele for "ignoring" the stochastic threshold. TrueAllele interprets all of the accumulated data on an electropherogram in order to reach an answer based on all of the information. The manner in which the system accounts for low level peaks is to assign correspondingly lower probabilities to lower peaks. In this way TrueAllele models the uncertainty by giving those peaks less influence on the final outcome.

56. The defendant also claims that there is no "evidence in the literature that TrueAllele directly takes into account allelic drop-in or drop-out." This statement is both a misinterpretation of how TrueAllele models data from an electropherogram, and is inaccurate. First, "drop-out" is a term associated with alleles whose peaks do not rise above the stochastic threshold. Unlike conventional analysis which treats an allele as either definitively present or absent (the result when it has dropped out) TrueAllele, as discussed above, models those low-level peaks by assigning lower probabilities. TrueAllele also accounts for alleles that do not present any peak about by again assigning a correspondingly lower probability.

57. Dr. Perlin has published several papers that mathematically describe the manner in which the system takes this into account. The People would direct the defendant to the article "An information Gap in DNA Evidence Interpretation" written by Dr. Perlin and Dr. Alexander Sinelnikov and published

---

[14] Rakay et. al., conclude that lowering an analytic threshold to 10 RFU's reduced drop-out rates by a factor of 100 without significantly increasing rates of erroneous noise detection. Rakay et. al., Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out. *Forensic Sci. Int. Gen.* 6 (2012) 723-728.

in the *PLOS One* Journal on December 16, 2009.[15] In that article Dr. Perlin describes how the mathematical model employed by TrueAllele accounts for low-level peaks and other events that occur during the stochastic process. The article the system is modeled in such a way that "quantitative STR data can convey their uncertainty via the data variance into a genotype. Greater genotype uncertainty is represented by a more diffuse probability distribution. And... a less certain genotype pmf [which] generally reduces match LR information." Perlin MW, Sinelnikov A (2009) p. 5.

58. The *PlosOne* paper is far from the only time that Dr. Perlin has published mathematical descriptions of how the system works. In addition to foundational articles such as the one cited above, the system is mathematically described in the validation studies that are attached to this affidavit and which will be discussed further.

Validation Studies

59. A validation study is the process of taking known DNA profiles and subjecting them to TrueAllele analysis to see if the system returns accurate results. The TrueAllele system for forensic casework has been the subject of three different validation studies performed in collaboration with the NYSPFIC.[16] The first of these studies entitled "Validating TrueAllele DNA mixture interpretation" was published in the *Journal of Forensic Sciences* in November of 2011. Dr. Perlin and other employees of Cybergenetics co-authored this study with Dr. Barry Duceman, the head of the NYSPFIC's DNA section and Jamie Bellrose from the Northeast Regional Forensic Institute at the State University of New York at Albany. It evaluated two person mixture samples from adjudicated cases worked on by analysts at the FIC. The study showed that TrueAllele match statistics results not only concurred with the previous results, but offered much more informative match statistics in every case.[17]

---

[15] All publications cited within this affidavit have been compiled into an appendix for the Court's reference.

[16] TrueAllele has been used by the NYS Police for the last ten years in order to automate their system of uploading single source genetic profiles to CODIS and NDIS. It has also been the subject of validation studies for that use however those are not included with this response.

[17] Perlin, MW, Duceman B, Validating TrueAllele DNA mixture interpretation, *J. Forensic Sci,* Nov 2011, Vol. 56. No. 6. p 1442-1443.

60. The second validation study performed in collaboration with the NYSPFIC was published two years later in November of 2013, again in the *Journal of Forensic Sciences*. This study looked at more mixture samples than the first study. Again the study found that TrueAllele was in agreement with the previous results arrived at by FIC analysts. The study also examined mixture samples in which an analyst was unable to develop a statistic. In each of those case TrueAllele was able to produce a match statistic.

61. The third New York study was conducted by Jay Caponera, MS from the NYSPFIC independently of Dr. Perlin and Cybergenetics. The results of this study were initially presented at the 2014 Conference of the American Academy of Forensic Science (AAFS) in February.[18] In evaluating two person mixtures the study found that the match statistics produced by TrueAllele were both reproducible and specific to all known donor profiles. The study also analyzed known non-donor reference profiles. It found that the mean difference in likelihood ratio between a known donor and a known non-contributor was 36.7 log units.[19] The study also found that in every instance in which a non-donor profile was evaluated using TrueAllele a likelihood ratio with a negative association was produced. The average being on the order of -6.76 log units for two-person mixtures and -3.5 log units for three person mixtures.[20] This is important because it shows that TrueAllele is extremely effective in determining when a person is not a contributor to a mixture.

62. The validation studies conducted with the crime laboratories in Virginia and Kern County California produced similar results to the New York Studies. The results of the Virginia study were published this spring in *PlosOne*.[21] The results from Kern County were presented at the same 2014

---

[18] Caponera, J. "Evaluating the Specificity of Genotypic Inference with TrueAllele Casework Software" *AAFS Annual Conference*. (2014) Seattle, WA

[19] A log unit is basically how many 0's a number has, so in this case the mean difference was on the order of a 1 followed by 36.7 zero's which is also referred to an undecillion. As a reference a trillion (1,000,000,000) is a 1 followed by nine zero's. This is a complicated way of saying that the statistical support for the inclusion of known donor's to a sample dwarfed the support for the inclusion of known non-donors.

[20] A negative log unit refers to how many zeros occur in front of the decimal point.

[21] Perlin, MW., et al. "TrueAllele Casework on Virginia DNA Mixture Evidence: Computer and Manual Interpretation in 72 Reported Criminal Cases" (2014) *PloSONE* 9(3): e92837. Doi: 10.137/journal.pone.0092837.

AAFS Conference as the latest New York study.[22] It is the People's understanding that in addition to the validation study performed in collaboration with Cybergenetics, the Virginia Department of Forensic Science has performed its own internal validation of the TrueAllele system with similar results.

63. The defendant briefly makes reference to the guidelines set forth by the Scientific Working on DNA Analysis Methods (SWGDAM). SWGDAM is a group of approximately 50 scientists representing federal, state and local forensic DNA laboratories in the United States and Canada that develops guidance and publications for the forensic DNA community. A SWGDAM committee was formed in 2007 to offer guidance on how to interpret mixed DNA profiles. In 2010 the committee published the guidance document referenced by the defendant. The defendant claims that the use of TrueAllele does not comport with the SWGDAM recommendations. While it is accurate that SWGDAM recommends the use of a stochastic threshold in traditional analysis, it makes an explicit exemption to that recommendation for the use of a probabilistic system. Guidline 3.2.2 states;

> "If a stochastic threshold based on peak height is not used in the evaluation of DNA typing results, the laboratory must establish alternative criteria (e.g., quantitation values or use of a probabilistic genotype approach) for addressing potential stochastic amplification. The criteria must be supported by empirical data and internal validation and must be documented in the standard operating procedures."

64. TrueAllele uses a probabilistic genotype approach, and the numerous validation studies that TrueAllele has undergone exactly follow the approach recommended by SWGDAM.

Disclosure of Source Codes

65. The defendant claims that TrueAllele's accuracy cannot be properly evaluated unless the actual computer code that runs the system is disclosed to the defense. The code is written in MATLAB ™ programming language and approximately 170,000 lines long.

66. The defense request for disclosure of the source code must be denied for several reasons

---

[22] Perlin MW., et al. "Assessing TrueAllele Genotype Identification on DNA/Mixtures Containing up to five Unknown Contributors." *AAFS Annual Conference.* (2014) Seattle, WA

67. First and foremost the defendant has provided no authority for this request. He has cited no case law or any provision of the Criminal Procedure Law that authorizes this disclosure. In the absence of any such authority the request must be denied as beyond the scope of CPL § 240.20.

68. The defendant claims that he needs the source code because "No one, outside of Cybergenetics, knows exactly how TrueAllele interprets complex STR DNA mixtures and, thereby assigns them evidentiary weight. ***It is therefore not possible to verify the accuracy of the components of the TrueAllele system."*** This statement is completely incorrect and misleading. The accuracy of the has been repeatedly verified during each of the validation studies. In Commonwealth v. Foley, 38 A.3d 882, 2012 PA Super 31., the Court in affirmed the use of TrueAllele in a murder conviction. The Court rejected the 'source code argument stating, "[Defendant's] third reason for exclusion is misleading because scientists can validate the reliability of a computerized process even if the 'source code' underlying that process is not available to the public." Id., at 885. The Court then noted the validation studies reported in the "Information Gap in DNA Evidence Interpretation" article published in *PlosOne* in 2009 and the first of the studies done in collaboration with the NYSPFIC.

69. This same holding was endorsed by the Court in Commonwealth of Virginia v. Mathew Brady, case Nos. CR11-465-01,-02,-03 and 04 an CR11-494-01,-02,-03 and -04 (oral decision July 26, 2013; written decision December 7, 2013) when it held that "much is made of the inability to thoroughly test the TrueAllele protocol, because its source code is unknown, but the Court relies on the observation by the Pennsylvania Court, and I quote: validation studies are the best tests of the reliability of the source codes. In common parlance, the question is does it work." Brady oral transcript p. 7.[23]

70. The reality is that nearly every step in forensic DNA analysis depends to some degree on a computer code of one kind or another. The PCR amplification process is fully automated and completed using a machine called a thermocycler, the entirety of which is carried out based on a computer code. The DNA sequencer used in the STR portion of the process is also fully automated and controlled,

[23] This transcript and the Court's written decision denying the defendant's motion to preclude TrueAllele are part of the appendix.

at its most basic level, by a computer code. Finally, the electropherograms are created by another computer program, referred to earlier, called GeneMapper ID™ which interprets the information gained by the laser and transforms it into a graphical form. The People are unaware of, and the defense has not cited, any evaluation by a Court or a member of the scientific community in which an instrument's source code has been examined in order to determine reliability, while eschewing the results of a validation study. Each of these computer codes is arguably just as important to the accuracy of the process as is the one underlying TrueAllele.

71. Even if the defense were to examine the code for TrueAllele and found no issue, only a validation study could properly demonstrate that the code actually produced accurate results.

72. The request for the source code also completely ignores the fact that the mathematical basis for the system has been published in numerous scientific articles. Neither defense counsel's affidavit, nor the affidavit of the defendant's expert Dr. Chakraborty, acknowledges reading these articles. There is no claim that the description of the system, as set out in the available scientific literature, is inadequate to assess its accuracy.

73. In fact Dr. Chakraborty's affidavit states, in reference to his work on the DNA subcommittee, that "My position on the subcommittee allowed me to meet with and speak to Dr. Perlin on several occasions and to gain deep insight into the science and mathematics which underpin TrueAllele." Despite Dr. Chakraborty's admitted understanding of the system, he has failed to identify a single concrete deficiency in how TrueAllele operates.

74. The source codes are also definitively not in the People's possession. They are confidential trade secrets held solely by Cybergenetics which is located in Pittsburg, Pennsylvania. The People are not required to obtain documents from sources beyond their control. CPL § 240.20(2), People v. Flynn, 79 N.Y.2d 879 (1992), Matter of Phillips v. Ramsey, 839 NYS2d 223(2nd Dept., 2007). This applies to government agencies such as the New York State Department of Motor Vehicles and New York City Office of the Chief Medical Examiner as well as private entities. Flynn supra., People v. Wright, 639 N.Y.S.2d 361 (2nd Dept., 1996), People v. Bynes, 598 N.Y.S.2d 217 (1st Dept., 1993).

75. It would also be against public policy to require the disclosure of the source code. Cybergenetics and Dr. Perlin have been developing TrueAllele for commercial use for over the last 15 years. Disclosure of the source code would allow any number of competing entities to copy and market a similar product. There are a number of entities that market commercial products that purport to analyze the same type of data as TrueAllele and disclosure of TrueAllele's source code would eliminate its competitive advantage. This would be bad for public policy because it would eliminate the incentive for private enterprises to develop and market useful and informative forensic DNA analysis software.

76. The elimination of private incentive to invest in forensic DNA technology would be crushing for the forensic community. Every instrument and the majority of software used in forensic DNA analysis is manufactured by a private corporation. Without private investment forensic DNA research would be limited to government funded programs. Private investment into DNA analysis has benefitted victims and defendant's alike and should promoted and protected.

77. Cybergentics and TrueAllele are also not alone as entities that do make the source code for their software publicly available. This includes government agencies such as the New York City Office of Chief Medical Examiner which has developed The Forensic Statistical Tool (FST), STRmix™, developed by Environmental Science and Research in New Zealand, Forensic Science South Australia, and the Australian National Institute of Forensic Science and corporations such as Applied Biosystems (GeneMapper ID-X) and Forensics LLC (Armed Expert); all are closed source programs.[24]

78. TrueAllele has also been evaluated by many members of the scientific community who do not have access to its source code. These include Drs. David Balding and Christopher Steele of the UCL Genetics Institute, University College London in the UK; [25] Drs. John Buckleton, Hannah Kelly, Jo-Ann Bright and James Curran from Environmental Science and Research in New Zealand and the Department of Statistics at the University of Auckland and Dr. Duncan Taylor from Forensic Science

---

[24] Though developed in Australia and New Zealand STRmix is currently being marketed in the United States and England. http://strmix.esr.cri.nz/node/11

[25] Steele, Christopher D., Balding, David J., "Statistical Evaluation of Forensic DNA Profile Evidence", *Annu. Rev. Stat. Appl.* 2014. 1:361-84

South Austrailia;[26],[27] Dr. John Butler and Dr. Michael Coble from the National Institute of Standards (NIST);[28],[29] Dr. Jack Ballantyne from the University of Central Florida and the chair of the New York State Commission on Forensic Science DNA subcommittee[30]; and Dr. Susan Greenspoon from Virginia's Department of Forensic Science, and Drs. Ruth Dickover and Kevin Miller from the Kern Regional Laboratory in Bakersfield California.

79. TrueAllele was also evaluated by the members of the New York State Commission on Forensic Science's DNA subcommittee without disclosure of its source code. The Commission on Forensic Science is empowered by Executive Law 49-B to develop standards and a program of accreditation for all forensic laboratories in New York. Accreditation is granted through the DNA subcommittee which also advises the Commission on any matter related to the implementation of scientific controls and quality assurance procedures for the performance of forensic DNA analysis. When the subcommittee makes a recommendation to the full Commission it is binding and must be either adopted or sent back to the subcommittee for further study, but it cannot be rejected. This means that the vote in the subcommittee is in almost all instances controlling. Because of this the subcommittee is comprised of highly regarded members of the scientific community including Dr. Ballantyne and, at the time TrueAllele was presented, Dr. Chakraborty. At no time during the period from 2002, when Dr. Perlin began presenting the features of TrueAllele to the subcommittee, and 2011 when the subcommittee made a binding recommendation to the Commission to approve TrueAllele for forensic casework, did Dr. Chakraborty ever request to see the source code.[31] It is only now, when acting as a paid defense expert, has Dr. Chakraborty called for the production of the source code.

---

[26] Kelly, Hannah et. al "A comparison of statistical models for the analysis of complex forensic DNA profiles", *Science and Justice* 54 (2014) 66-70.

[27] Taylor, Duncan et. al., "The interpretation of single source and mixed DNA profiles", *Forensic Science International: Genetics*. Vol. 7, Issue 5, September 2013 p. 516-528.

[28] http://www.cstl.nist.gov/strbase/pub_pres/ISFG2011-Coble-TrueAllele.pdf

[29] http://www.cstl.nist.gov/strbase/training/ISHI2012-MixtureWorkshop-Statistics.pdf

[30] Ballantyne, J., Hanson, E.K., and Perlin, M.W. "DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. *Science and Justice,* 53(2): 103-114, 2013.

[31] TrueAllele was first presented to the committee in 2002 for use as an automated data review system. It was approved for that use in 2006.

80. For all of the reasons set forth above the People ask that the Court deny the defendant's motion to compel the production of any source codes related to the TrueAllele system.

Acceptance in the Scientific Community

81. TrueAllele is accepted within the scientific community. As set forth in the preceding section numerous members of the scientific community have evaluated TrueAllele and positively commented on its methods and reliability. Before addressing some of those comments it is necessary to place them in the proper context. As DNA profiling has advanced more small and more complicated samples of DNA are now able to be analyzed than ever before.

> "Alternatives to the binary model [CPI] were experimented with in the late 1990s. Two options were tabled; 1. The fully continuous model, and 2) A model that is partially continuous based on allowing a probability for dropout and drop-in.
>
> Neither of these models saw any large-scale deployment throughout the early 2000s. This has come in for justifiable criticism. [citation omitted] However there has been a new and highly welcome forward movement n the late 2000s driven by the creation of continuous software (TrueAllele™) by Perlin et. al., and efforts from Balding and Buckleton, Haned and Gill and Rudin and Lohmueller." D. Taylor et. al., "The interpretation of single source and mixed DNA profiles." *Forensic Sci. Int. Gen.* 7 (2013) 517, 516-528.

82. The scientific community is definitively moving away from CPI and other so-called binary models towards seek out more sophisticated models that are referred to as semi or partially continuous and continuous. There are a number of semi-continuous models in use today including those referenced by Taylor et. al., Forensism and LabRetriever (created by Dr.'s Haned and Gill, and Drs. Rudin and Lohmueller respectively), the OCME's FSS and LikeLTD, created by Dr. Balding. These systems account for the drop-out/drop-in of alleles by using a predetermined rate to calculate the probability with which an allele either failed to appear or appeared erroneously. The drawback of this type of system is that it still depends on the use of a stochastic threshold and does not take into account relative peak heights. As Kelly, et. al. note "the shortcomings [in binary and semi-continuous models]

leads into the concept of the continuous model. This model seeks [to] move away from very discreet all or nothing nature of the binary model by making better use of the available information."[32]

83. TrueAllele and STRmix, which are both based on Markov Chain Monte Carlo algorithms, are examples of continuous models. In the last three to four years the scientific community has roundly endorsed the use of this type of model. "Coming finally to the continuous model; this approach is undoubtedly the premier choice in terms of accuracy as defined here." Kelly, H., et al. p. 69. Taylor et. al. (some of the creators of STRmix) write that:

> "As a general principle ignoring information that can be properly evaluated tends to weaken the evidence for a true hypothesis and will more often include a false hypothesis.... Relevant information that can be effectively evaluated should not be ignored. In the LR framework including relevant and properly evaluated information tends to increase LRs if H1 is true and decrease if H2 is true. Certainly the first part of this principle has been elegantly reinforced by Perlin et. al." and "we present a trial that supports Perlin et. al.'s conclusion" Taylor et. al. p. 516.

84. Drs. Michael Coble and John Butler from the National Institute of Standards (NIST) gave a presentation at 24th Congress for the International Society for Forensic Genetics in 2011 entitled "Exploring the Capabilities of Mixture Interpretation Using True Allele Software". They explained how the system worked and described the results of their own use of TrueAllele to evaluate known DNA profiles. They concluded that TrueAllele makes a better use of the data than RMNE (for Random Man Not Excluded, the same as CPI/CPE) and that that TrueAllele performed better than RMNE and classic LR's with low level contributors.[33]

DNA Subcommittee Presentation

85. As stated previously several times Dr. Perlin and other members of Cybergenetics conducted extensive presentations before both the DNA subcommittee and the full Commission on Forensics. On the issue of using TrueAllele for forensic casework, including mixture analysis, the subcommittee was presented to on March 5, 2010 by Dr. Perlin and given several published articles including the validation study contained in *PlosOne* from 2009 as well as a document entitled New York

---

[32] Kelly, H., et. al. "A comparison of statistical models for the analysis of complex forensic DNA profiles" *Science and Justice* 54 (2014) 66-70.

[33] http://www.cstl.nist.gov/strbase/pub_pres/ISFG2011-Coble-TrueAllele.pdf

State TrueAllele Casework Developmental Validation. They were then given a presentation by Ms. Bellrose entitled "New York State Police Validation of TrueAllele: a Statistical Tool for Genotype Inference and Match that solve Casework Mixture Problems"

86. This presentation took the subcommittee through an initial validation, including its analysis of two and three person DNA mixtures. The subcommittee was presented to again on May 19, 2010 and given the documents related to how the next phase of validation would be implemented. They were also given a document entitled "TrueAllele Mixture Interpretation and Least-Square Deconvolution. This paper was prepared in response to the subcommittee's request for an explanation of how TrueAllele compared with the least-squared deviation method for de-convoluting mixture samples.

87. On May 20, 2011 the subcommittee was given a final presentation on the results of the validation that became the basis for the first of the two studies published in the *Journal of Forensic Science* in 2011.[34] At the conclusion of that meeting the subcommittee, on a motion seconded by Dr. Chrakraborty, voted unanimously to approve TrueAllele for forensic casework.

88. In addition to the papers and validation studies cited herein Dr. Perlin has been invited to present on TrueAllele at numerous scientific conferences and workshops.[35]

---

[34] The People are in possession of every document and presentation given to the subcommittee and will be happy to provide the same to the Court or defense counsel if it is requested.

[35] Perlin, M.W. DNA mapping the crime scene: do computers dream of electric peaks? in the Proceedings of Promega's Twenty Third International Symposium on Human Identification. Nashville, TN, 2012.

Perlin, M.W. Combining DNA evidence for greater match information. Forensic Science International: Genetics Supplement Series, DOI 10.1016/j.fsigss.2011.09.112, 2011.

Perlin, M.W. Investigative DNA databases that preserve identification information. Forensic Science International: Genetics Supplement Series, DOI 10.1016/j.fsigss.2011.09.103, 2011.

Perlin, M.W. Explaining the likelihood ratio in DNA mixture interpretation in the Proceedings of Promega's Twenty First International Symposium on Human Identification. San Antonio, TX, 2010.

Perlin, M.W. Scientific validation of mixture interpretation method in the Proceedings of Promega's Seventeenth International Symposium on Human Identification. Nashville, TN, 2006.

Perlin, M.W. Real-time DNA investigation in the Proceedings of Promega's Sixteenth International Symposium on Human Identification. Dallas, TX, 2005.

Perlin, M.W. Simple reporting of complex DNA evidence: automated computer interpretation in the Proceedings of Promega's Fourteenth International Symposium on Human Identification. Phoenix, AZ, 2003.

Perlin, M.W., Coffman, D., Crouse, C.A., Konotop, F. and Ban, J.D. Automated STR data analysis: validation studies in the Proceedings of Promega's Twelfth International Symposium on Human Identification. Biloxi, MS, 2001.

Perlin, M.W. An expert system for scoring DNA database profiles in the Proceedings of Promega's Eleventh International Symposium on Human Identification. Biloxi, MS, 2000.

Perlin, M.W., Computer automation of STR scoring for forensic databases. In First International Conference on Forensic Human Identification in The Millennium, London, UK. The Forensic Science Service, 1999.

Frye Standand

89. New York State adheres to the standard set forth under Frye v. United States, 293 F 1013 (D.C., Cir., 1923) that testimony based on novel scientific principles or procedures is admissible only if it has gained general acceptance in its particular field. Id., at 1014. The Court of Appeals has noted that does not mean that principle need by "'unanimously indorsed' by the scientific community but must be 'generally accepted as reliable'" People v. Wesley, 83 N.Y.2d 417 (1994).

90. The first question that must be addressed by the Court is whether TrueAllele even constitutes a novel scientific principle. The application of a generally accepted technique, even though its application in a specific case was unique or modified, does not require a Frye hearing. Parker v. Mobil Oil Corp., 7 N.Y.3d 434 (2006), Peole v. Magri, 3 N.Y.2d 562 (1955), People v. Byrd, 855 N.Y.S.2d 505 (1st Dept., 2008) Nonnon v. City of New York, 819 N.Y.S.2d 705 (1st Dept., 2006) The principles utilized by the TrueAllele system, likelihood ratios, Bayesian statistics and Markov Chain Monte Carlo algorithms have been used and accepted in the scientific community for years.

91. TrueAllele is also not the only system that uses MCMC to deconvolute mixtures, and Dr. Perlin is not the only member of the scientific community to propose such methods.[36]

92. The decision to use these tools to deconvolute complex DNA mixture profiles may be an emerging approach, but that does not make those methods novel and should not subject them to scrutiny under Frye.

93. Even if the Court does determine that TrueAllele is a novel scientific method there is ample evidence that it is gained general acceptance in the scientific community. The particular procedure need not be unanimously endorsed by the scientific community but must be generally acceptable as reliable" People vMiddleton, 54 NY2d 42, 44 (1981). The Frye Court placed the line at where a technique passes from experimental to demonstrable. The Court found that there was no precise way to

---

[36] Curran J.M., "A MCMC method for resolving two person mixtures" *Science and Justice* 48 (2008) 168-177.

measure when that line was crossed, but the People would submit that repeated studies published in peer reviewed journals that validate a method clearly show that it is well past experimental.

94. The defense points to the fact that TrueAllele "abandons the human element in analysis" and "because it analyzes data that falls below the thresholds incorporated in standard practice in DNA laboratories." These are factors that ought to be embraced by the legal system, not shunned. For years defense experts have argued against forensic DNA analysis because they claim it is much too subjective.[37] TrueAllele provides a completely objective inference based on the data. It also makes uses of all available data. As Drs. Buckleton and Gill wrote in 2010;

> "Biologists and applied scientists in many fields use thresholds to delineate between two states. Thresholds are always applied for 'convenience,' which means that the transition between two states is gradual. For this reason any attempt to apply a strict threshold will always fail... The purpose of the ISFG DNA commission document was to provide a way forward to demonstrate the use of probabilistic models to circumvent the requirement for a threshold and to safeguard the legitimate interests of defendants."[38]

95. The People do not dispute the contribution that Dr. Charkraborty has made to the field of forensic science during his long career. However we cannot help but to question his opinion as to whether TrueAllele is accepted within the scientific community. During the year and a half that TrueAllele was presented to him and the rest of the DNA subcommittee its intended use for de-convoluting DNA mixture samples was a made readily apparent. Each of the scientific papers and presentations made to the subcommittee contained data from mixture samples. Each of the validation studies addressed the effectiveness of TrueAllele in dealing with mixture samples. Nonetheless, he now states in his affidavit that "On May 20, 2011 our DNA subcommittee approved TrueAllele for use the New York State Police for their forensic casework without any mention of the type of forensic casework." That statement is not supported by the record made to the subcommittee. Every journal in which an article describing TrueAllele has been published allows for letters to the editor that raise legitimate

---

[37] Thompson, William C. "Painting the target around the matching profile: the Texas Sharpshooter fallacy in forensic DNA interpretation" *Law, Probability and Risk*, 8 (3): 257-258.

[38] Buckleton, J., Gill, P. "Commentary on Budowle et. al." *Journal of Forensic Science* 55 : 265-268 (2010)

scientific critiques. Despite what is now apparently his opinion of TrueAllele, Dr. Chakraborty has never once published any article or letter, nor presented at any scientific conference criticizing its methods or approach.

96. The People have provided the Court with numerous citations to members of the scientific community working both with, and independently of, Cybergenetics that support TrueAllele's reliability and acceptance in the scientific community.

97. The TrueAllele system for forensic casework clearly passes the Frye standard and the defendant's motion to preclude must be denied.

Respectfully submitted,
ROBERT M. CARNEY
District Attorney

Peter H. Willis
Assistant District Attorney

Dated: May 2, 2014