

Lucy, D. Curran, J.M. Pirie, A.A. & Gill, P. (2007) The probability of achieving full allelic representation for LCN-STR profiling of haploid cells. *Science & Justice*, 47:168-171.

## **The probability of achieving full allelic representation for LCN-STR profiling of haploid cells.**

D. Lucy: Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom.

J.M. Curran: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1010, New Zealand.

A.A. Pirie: Lothian & Borders Police Forensic Science Laboratory, 11 Howdenhall Road, Edinburgh, EH16 6TF, United Kingdom.

P. Gill: Forensic Science Service, 2969 Trident Court, Solihull Parkway, Solihull, Birmingham, B37 7YN, United Kingdom.

**Keywords:** Haploid cells, sample size.

# Abstract

Modern forensic techniques allow DNA to be extracted from ever decreasing amounts of cellular material. Low copy number (LCN) profiling enables the production of STR profiles from small numbers of cells. Moreover, methods such as laser micro-dissection enables forensic scientists to potentially isolate individual cells for PCR. The DNA derived from haploid cells (semen) is a common source of forensic evidence in sexual assault cases. Haploid cells contain only half the DNA complement of diploid cells (3pg compared to 6pg). The smaller the number of cells sampled, the smaller the probability that there is a full representation of the alleles comprising the donor profile. This paper investigates the relationship between the number of cells sampled and the probability of full representation of all alleles in the donor sample. It also considers the effect of typing several loci as opposed to just a single locus.

# Introduction

Haploid cells, such as spermatozoa, are formed by the biological process of meiosis and contain half the genetic material of the donor organism. Spermatozoa are frequently tested in cases of sexual assault cases, and any DNA profiles obtained are likely to form a substantial component of the forensic evidence.

A man produces approximately 300 million spermatozoa per ejaculate [1]. Assuming a DNA content of 3pg per cell the expectation is that there will be approximately 0.9mg of DNA per ejaculate [2]. The high cell density of spermatozoa in semen makes it an excellent source of DNA for forensic DNA profiling.

The question now asked by many police officers and forensic scientists is not so much “what materials can you obtain a DNA profile from?” but, “how many cells or how

much material do you need to produce a DNA profile?”. This is especially important in cases of unsolved sexual assaults where the original slides are the only remaining forensic evidence or from cases where only trace amounts of spermatozoa have been isolated from intimate swabs or clothing. It has been widely demonstrated using commercially available kits, such as AMPFISTR (PE Applied Biosystems, Foster City, CA, USA) that accurate DNA profiles can be routinely produced from 100 picograms or less of purified DNA [3, 4, 5].

DNA technology applied to forensic science has given us the ability to generate DNA profiles from limited or trace amounts of biological materials, even from a single cell [6, 7]. Theoretically, it is feasible to amplify DNA from a single haploid or diploid cell, (3 → 6 picograms) but in practice it may prove more difficult to obtain a reliable AMPFISTR™ SGM plus® DNA profile.

With the application of new techniques in forensic science such as low copy number (LCN) and laser capture micro dissection systems, the sensitivity and detection of DNA profiling has increased. These have allowed forensic scientists not only to isolate and recover low numbers of specific cell types for DNA profiling, but have also enabled forensic scientists to attribute the DNA profiles obtained to the cells of origin.

Sampling theory has been applied to plant seeds in cases where multiple trait selection has been desired and linkage taken into account [8], and some work to extend this to forensic cases has been undertaken, (Sjerps, M. personal communication) but to date no formal treatment exists in the forensic literature.

## Theory

Let  $A_{ij}$  represent the  $j^{\text{th}}$  ( $j = \{1, 2\}$ ) allele of an individual at the  $i^{\text{th}}$  locus ( $i = \{1, 2, \dots, k\}$ ). Assume for the moment that the individual is heterozygous at locus  $i$ , such that  $A_{i1} \neq A_{i2}$ . Then the probability that the full genotype of the individual is represented in a sample of  $n$  cells is given by:

$$\Pr\left(\bigcap_{i=1}^k n_{A_{i1}} \geq 1 \cap n_{A_{i2}} \geq 1\right)$$

where  $n_{A_{ij}}$  is the number of copies of allele  $A_{ij}$  observed in a sample of  $n$  cells.

For a single locus the probability of observing any particular allele in one haploid cell is  $\pi_{A_{i1}}$  (in the following we assume  $\pi_{A_{i1}} = \pi_{A_{i2}} = 0.5$ , but will initially maintain generality). In this particular situation a fixed number  $n$  of haploid cells is sampled, and each cell has probability  $\pi_{A_{i1}}$  of being allele  $A_{i1}$ , and  $1 - \pi_{A_{i1}}$  of being allele  $A_{i2}$ , so that the probability of observing  $n_{A_{i1}}$  copies of allele  $A_{i1}$  (and  $n - n_{A_{i1}}$  of allele  $A_{i2}$ ) is given by the binomial expression:

$$\Pr(n_{A_{i1}} | n, \pi_{A_{i1}}) = \binom{n_{A_{i1}}}{n} \pi_{A_{i1}} (1 - \pi_{A_{i1}})^{n - n_{A_{i1}}}.$$

Therefore the probability of interest (at a single locus) is by definition:

$$\begin{aligned}
\Pr(n_{A_{i1}} \geq 1 \cap n_{A_{i2}} \geq 1) &= 1 - \Pr(\overline{n_{A_{i1}} \geq 1 \cap n_{A_{i2}} \geq 1}); \quad n \geq 2 \\
&= 1 - \Pr(n_{A_{i1}} = 0 \cup n_{A_{i2}} = 0) \\
&= 1 - \Pr(n_{A_{i1}} = 0) - \Pr(n_{A_{i2}} = 0); \quad \text{since } \Pr(n_{A_{i1}} = n_{A_{i2}} = 0) = 0 \\
&= 1 - (1 - \pi_{A_{i1}})^n - \pi_{A_{i1}}^n
\end{aligned}$$

If  $\pi_{A_{i1}} = \pi_{A_{i2}} = 0.5$  this simplifies to:

$$1 - 0.5^n - 0.5^n = 1 - 0.5^{n-1} \quad (1)$$

In the  $k$  loci case, if  $\pi_{A_{i1}} = \pi_{A_{i2}} = 0.5$ , and independence between loci is assumed, Equation 1 can be expanded to:

$$\begin{aligned}
\Pr\left(\bigcap_{i=1}^k n_{A_{i1}} \geq 1 \cap n_{A_{i2}} \geq 1\right) &= \prod_{i=1}^k \Pr(n_{A_{i1}} \geq 1 \cap n_{A_{i2}}) \\
&= (1 - 0.5^{n-1})^k.
\end{aligned}$$

A graph showing the relationship between the probability of having all alleles represented in a sample of  $n$  haploid cells for  $k$  loci is given in Figure 1.

If one wishes to be  $100\alpha\%$  sure that sufficient cells have been sampled to obtain full representation of the genotype of the donor, the number of cells to sample is given by:

$$n = \frac{\log(1 - \alpha^{1/k})}{\log(0.5)} + 1 \quad (2)$$

This is unlikely to yield an integer solution, so we recommend that the result be rounded up. Table 1 gives such a tabulation of haploid cell sample sizes for a given probability and given number of heterozygous loci observed.

## Homozygotic loci

Equation 2 applies only to where the number of heterozygous loci may be known. A more general, and more realistic situation, is where the number of heterozygous loci are not known.

Suppose there are  $m$  alleles possible for the  $i^{\text{th}}$  loci indexed by  $l$ , such that  $l = \{1, 2, \dots, m\}$ , and that this allele is observed amongst the population with probability  $p_{il}$ . Assuming Hardy-Weinberg equilibrium, the probability of the  $i^{\text{th}}$  loci being heterozygous is  $p_i = \sum^m p_{il}^2$ , and, assuming linkage equilibrium, an expectation for the numbers of homozygous loci ( $q$ ) can be written:

$$E(Q) = \sum_{q=0}^k q \Pr(Q = q).$$

Which for any specific system of loci and alleles can be shown to equal to:

$$q_h = \sum_i^k \sum_l^m p_{il}^2.$$

Thus Equation 2 can be rewritten<sup>§</sup>:

---

<sup>§</sup>The authors understand that the marginal distribution of  $n$  is the sum of the joint probability function of  $N \leq n$  and  $Q = q$  over  $q$ . However, this requires specification of  $\Pr(Q = q)$ , and solving the equation for  $n$  and  $\alpha$ , and is consequently much more complicated than the formula given above. Trial calculations using the SGM+ loci from the UK Caucasian data suggest that the approximation given above is conservative because  $q_h$  is likely to be low for most multiplexes.

$$n = \frac{\log(1 - \alpha^{1/(k-q_h)})}{\log(0.5)} + 1$$

As any value for  $q_h$  calculated directly is unlikely to be an integer it is suggested that  $q_h$  be rounded down to give a conservative estimate for  $n$ .

## Discussion

The analysis above assumes either that all loci of interest are heterozygous, or, that the allelic frequencies for the population of interest are known. Of these two options the second should usually be satisfied. A conservative estimate for the number of cells required for a given probability  $p$  of full allelic representation can be made by assuming that all loci are heterozygous, then the number of cells given by Table 1 will be an over-estimate, therefore safe to use for practical purposes.

The number of cells given in Table 1 are theoretical quantities only. They will not necessarily reflect the probabilities of what is observed after extraction, PCR, and subsequent electrophoresis. The calculations presented here do not model either the extraction or PCR processes. The efficiency of the processes will govern the actual numbers of cells required. Recent work [9] suggests an empirical estimate of several hundred haploid cells required in the sample required to stand a high probability of obtaining a complete STR profile. The expected discrepancy between the theoretical and empirical results must be reflected by the relative efficiency of the extraction and PCR processes [10]. Other effects such as template degradation and damage to the DNA molecule will also play a part.

## Acknowledgements

The authors should like to acknowledge the helpful comments made by two anonymous reviewers on an earlier draft of this paper, and to Professor Chris Triggs of The University of Auckland for assistance proving some of the results presented in this paper.

## References

1. Saferstein, R. (2003) *Criminalistics: An introduction to forensic science*. 7<sup>th</sup> Edition; Prentice Hall, New Jersey.
2. Butler, J.M. (2003) *Forensic DNA typing: Biology & Technology behind STR Markers*. Academic Press.
3. van Oorschot, R.A.H. & Jones, M.K. (1997) DNA fingerprints from fingerprints. *Nature*. **387**(6635); 767-767.
4. Gill, P. Whitaker, J. Flaxman, C. Brown, N. & Buckleton, J. (2000) 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.
5. Wickenheiser, R.A. (2002) Trace DNA: A review, discussion of theory, and application of the transfer of trace quantities of DNA through skin contact. *Journal of Forensic Sciences*. **47**(3); 442-450.

6. Findlay, I. Frazer, R. Taylor, A. & Urquart, A. (1997) Single cell DNA fingerprinting for forensic applications. *Nature*. **389**: 555-556.
7. Li, H. Gyllensten, U.B. Cui, X. Saiki, R.K., Erhlich, H.A. & Arnheim, N. (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**:414-417.
8. Schwager, S.J. Mutschler, M.A. Federer, W.T. & Scully, B.T. (1993) The effect of linkage on sample size determination for multiple trait selection. *Theoretical and Applied Genetics*. **86**(8): 964-974.
9. Elliott, K. Hill, D.S. Lambert, C. Burroughes, T.R. & Gill, P. (2003) Use of laser microdissection greatly improves the recovery of DNA from sperm on microscope slides. *Forensic Science International*. **137**(1): 28-36.
10. Taberlet, P. Griffin, S. Goossens, B. Questiau, S. Manceau, V. Escaravage, N. Waits, L.P. & Bovet, J. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*. **24**(16): 3189-3194.
11. Foreman, L. & Evett, I. (2001) Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system *International Journal of Legal Medicine*. **114**(3): 147-155.

Figure 1: Probability of all alleles being represented versus haploid cell sample size where 1, 2, 3, 6, 10 and 15 heterozygous loci are observed.

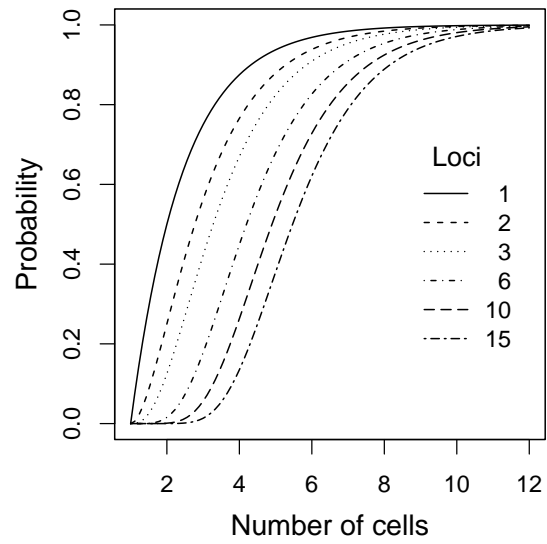


Table 1: Number of cells for a given probability (columns), and number of heterozygous loci (rows) for complete representation of the full profile from haploid cells.

loci	probability				
	0.9	0.95	0.99	0.999	0.9999
1	5	6	8	11	15
2	6	7	9	12	16
3	6	7	10	13	16
6	7	8	11	14	17
10	8	9	11	15	18
15	9	10	12	15	19

## Example calculations

Suppose that a sexual assault had taken place in the United Kingdom, and that semen traces had been recovered from the clothing of a woman who described the offender as being a man of European descent.

Table 2 contains the sums of squared frequencies for each of 10 alleles. These were calculated taken from Foreman & Evett [11]. The sum of the squared allele frequencies is 1.825, and is the expectation of the number of homozygous loci from these 10 loci, and will be rounded down to 1 in subsequent calculations.

Table 2: Sums of squared frequencies for all alleles from each of 10 loci for United Kingdom Caucasians. Calculated from Foreman & Evett [11].

locus	SS frequencies
D18	0.123
D21	0.169
D2	0.122
FGA	0.134
THO1	0.219
D3	0.206
D19	0.238
D16	0.222
VWA	0.194
D8	0.198
$\Sigma$	1.825

The frequencies from Table 2 are from a sample of 437 individuals described as Caucasian, and include specimens from Forensic Science Service staff from the SGM database. A more ideal conditioning might be those allelic frequencies from United Kingdom men alone as only men supply spermatozoa.

Were an investigating scientist to need to require a 95% probability that their sample of

cells represented every allele from the donor of the stain then they would need to sample:

$$\begin{aligned}n &= \frac{\log(1 - \alpha^{1/(k-q_h)})}{\log(0.5)} + 1 \\&= \frac{\log(1 - 0.95^{1/(10-1)})}{\log(0.5)} + 1 \\&= \frac{\log(0.0057)}{\log(0.5)} + 1 \\&= 8.45,\end{aligned}$$

or at least 9 cells to stand a greater than 95% probability that there is at least one copy of each and every allele from the donor organism. As discussed earlier the sample size of 9 cells is a theoretical sample size only, it does not take into account variation due to recovery and the PCR process.