
ANALYSIS OF DNA FROM FAECES AS A METHOD FOR
ESTIMATING POPULATION SIZE IN CAMBODIAN
ELEPHANTS

FOR

FAUNA & FLORA INTERNATIONAL

BY

SIMON M.C. POULTON

BIOECOSS LTD.
8TH DECEMBER 2009





EXECUTIVE SUMMARY

- Σ A set of 529 faecal samples was collected from nine geographic zones and one independent team in SW Cambodia in 2006/7. The elephant DNA from these samples was extracted and amplified by a commercial laboratory; DNA Solutions in Australia. For each of the samples, DNA was analysed using 18 genetic markers plus a sex marker. Each marker in each sample yielded zero, one or two base-pair values (BP-values), representing individual alleles at a genetic locus.
- Σ This raw data matrix was severely depleted due to the large number of zero values obtained from the amplification process, known as allelic drop-out. Only 52% of all amplifications yielded one or more successful value. Furthermore, only 18% of the amplifications were confirmed values. The major practical consequence of this high rate of allelic drop-out was that markers yielding only a single BP-value could not be interpreted as homozygous loci, as they could easily have been heterozygous, with one BP-value not amplifying.
- Σ 196 faecal samples yielded no BP-values at all, leaving only 333 samples (63%) that could be used in the subsequent analysis. Furthermore, by ignoring unconfirmed values, only 39% of the samples contained one or more BP-value. Only 15 samples contained BP-values at all 18 loci, so 97% of samples had unidentified alleles at one or more locus. Using the stringent criterion that only samples with six or more confirmed BP-values should be included in the analysis would have reduced the sample size to 106 (20%). Furthermore, only 16% of samples had a confirmed identification at the sex marker.
- Σ Instead of taking this sample-based approach and abandoning the vast majority of data, a locus-based approach was used. This included all the samples with BP-values to build a profile of the alleles present at each locus. The polymerase chain reaction (PCR) process creates snippets of DNA and counts the number of base-pairs between the markers. Each different number represents a distinct allele. However, in practice, due to averaging of a large number of readings within the PCR machinery, the BP-values obtained were fractions to one decimal place. So, instead of “clean” distinctions between alleles, frequency histograms of the empirical BP-values showed clusters of similar values, making the interpretation of cut-off points between alleles difficult.
- Σ To overcome this subjectivity, a computer-intensive process called a “genetic algorithm” was developed. For each locus in turn, this modelled the clusters of BP-values and identified the most likely allele to which each individual value belonged. Both confirmed and unconfirmed values were used, but confirmed had a 5× weighting over unconfirmed. Not all BP-values were identified to alleles, mainly because they existed as individual, outlying values. These may have been true, but extremely rare alleles, although they may also have been caused by cross-contamination. Nevertheless, the GA was extremely powerful in assigning values to alleles, with 94% of all BP-values contributing to the subsequent analyses.
- Σ The next stage was to use the alleles to construct the genomes for all 333 samples that had one or more BP-values. This matrix formed the basis for all subsequent analyses and is provided as an embedded MS Excel spreadsheet in Appendix IV.
- Σ Population estimates were made using a pair-wise comparison approach. Each sample was compared with every other sample and the null-hypothesis that there was no difference between the two genomes was tested. If this was rejected at a suitable α -level, the two samples were considered



to have come from different elephants. In addition, a two-tailed test that the two samples were identical was also run. Two different α -levels were used (0.05 and 0.10), providing a conservative test and a more relaxed estimate of population size. Prior to the pair-wise comparisons, an initial run (at each α -level) tested the null-hypothesis that there was no difference between each genome and a genome generated entirely at random. This had the effect of excluding a number of samples (139 and 98 respectively) due to insufficient alleles, but at least this exclusion was based on an objective criterion, rather than a subjective judgement. This resulted in a total of 46,216 pair-wise comparisons at both α -levels.

- Σ The pair-wise comparisons used a Monte Carlo technique to obtain the statistic for hypothesis testing, running between 1,000 and 10,000 randomisations for each pair-wise comparison. This method also carried out three randomisation procedures to complete the genomes;
 - Missing alleles were replaced using an algorithm based on the linkages between loci and the proportions of different alleles within the locus.
 - Single allele values were treated as heterozygous but displaying allelic drop-out, with a probability based on the overall allelic drop-out for the locus. In this case, an allele was selected using the same procedure as for missing values.
 - Unconfirmed alleles were assigned a probability of being correct of 0.9. This meant that in 10% of cases, they were replaced with another allele, based on the above procedure. Confirmed alleles were assigned a probability of being correct of 1, so were always unchanged.
- Σ The tests of difference between samples yielded a very high rate of significant results; 97% for both α -values. Cluster analysis was used on the resultant similarity matrices to identify unique individual samples and clusters of samples that could not be distinguished. By including confirmed sex markers, the final estimates of minimum number alive (MNA) were obtained. Four different estimates were obtained, from each combination of α -value (0.05 & 0.1) and whether geographical zones were considered to hold isolated or free-ranging populations of elephants. In the latter case, estimates were slightly lower because samples from different zones could be clustered together, whereas in the former model, they were always treated as coming from different elephants. The final MNA estimates were;

$\alpha = 0.05$, Amalgamated Zones:	129
$\alpha = 0.05$, Isolated Zones:	136
$\alpha = 0.10$, Amalgamated Zones:	174
$\alpha = 0.10$, Isolated Zones:	175

If a highly conservative estimate of MNA is required, then the lower value of around 135 elephants should be taken. However, if a lower degree of confidence can be accepted, then an estimate of 175 elephants would not be unreasonable.

- Σ The tests of identity returned zero pair-wise comparisons that were significant at either α -level. This was an important result because it precluded the use of capture/mark/recapture techniques (CMR) to estimate the size of the un-sampled population; if no two samples could be considered to have come from the same elephant with any degree of certainty, then no “recaptures” could have occurred. Furthermore, the general paucity of allele identifications precluded the investigation of in-breeding, relatedness or sex-ratios.



Σ Finally, the conclusions from this analysis generated a number of recommendations for future work;

- Design the field survey and arrange laboratory time to minimise the period between sample collection and DNA analysis.
- Pilot the DNA extraction and amplification process on known sources of faeces from either captive or domesticated elephants.
- During fieldwork, collect between 5% and 10% duplicate samples. These should be treated as “true” samples during analysis and not identified as duplicates.
- Review the number of markers used in the DNA extraction / amplification.
- In subsequent surveys, ensure that allelic drop-out is minimised to a level that allows accurate assessments of sex-ratios, rates of in-breeding and relatedness between individuals.
- For future fieldwork and other FFI surveys, ensure that a well-designed database is constructed before field-work begins, which reflects the survey design and models the data structure and analytical procedure.



CONTENTS

Introduction.....	1
The Raw Dataset	2
<i>Data Exploration</i>	2
<i>Restructuring of Raw Data</i>	2
<i>Summary of Raw Data</i>	2
Allele Classification.....	5
<i>The Genetic Algorithm</i>	5
<i>Assignment of BP Data to Alleles</i>	5
Genotyping of Samples	7
Population Estimation	8
<i>The “Absolute” Model</i>	8
<i>The “Probabilistic” Models</i>	9
<i>The Cluster Analysis</i>	12
<i>The Final Population Estimates</i>	14
Conclusions and Recommendations for Future Work	15
Appendices.....	17
<i>Appendix I</i>	17
<i>Appendix II</i>	18
<i>Appendix III</i>	19
<i>Appendix IV</i>	20



Introduction

The scope of this report is limited to the analysis of the dataset provided by FFI to BioEcoSS Ltd. This included the verification and validation of the raw data, the restructuring of the data to enable adequate analysis, and the analysis itself to provide an estimate of elephant populations.

No attempt has been made here to explain the sampling strategy or field methods used to collect the faecal samples. Furthermore, there is no explanation of the polymerase chain reaction (PCR) and DNA amplification procedures used by DNA Solutions. These issues are covered elsewhere. There is also no discussion of the second dataset that was supplied, which required considerable investigation before the decision was made to analyse the original dataset.

The primary aim of this analysis was to interpret the matrix of base-pair values (BP-values) collected from samples of dung into a dataset of individually identified elephants. Secondary objectives included the investigation of rates of in-breeding by analysis of Hardy-Weinberg ratios, the investigation of sex-ratios and the identification of related individuals, all of which would contribute to an understanding of population structure. The final objective was to identify samples that had come from the same elephant. This would enable the use of Capture/Mark/Recapture (CMR) methods to estimate the proportion of the population not sampled, and so provide a more accurate estimate of overall population size.

The analysis proceeded in five stages;

- Σ Data exploration, validation and restructuring,
- Σ The identification of the alleles at each of 18 genetic loci,
- Σ The use of these alleles to create a genotype of each sample,
- Σ The application of statistical models to undertake pair-wise comparisons and create similarity matrices between the samples,
- Σ The use of cluster analyses on the similarity matrices to identify individually unique samples or clusters of indistinguishable faecal samples.

These stages form the major sections of the main part of this report. As explained in the “Conclusions and Recommendations for Future Work”, the secondary objectives could not be achieved due to the paucity of raw data. However, conservative population estimates have been derived and a number of recommendations suggested for future work.



The Raw Dataset

Data Exploration

The original data were supplied in an Excel workbook (Summary FINAL(2).xls) with 529 data rows and 37 data columns. The rows represented the faeces samples and were coded with alpha-numeric combinations of the geographic zone and a sample number (AU001, AU002, etc.). The columns represented pairs of columns for 18 genetic markers (EMU01, EMU02, etc.) plus one sex marker (SRY). All data stored in the cells of this table were base-pair (BP) lengths rounded to one decimal place, plus two optional qualifiers for non-confirmed (NC) or approximate values (\approx). The only exception to this was the sex marker which recorded Male or Female, plus the NC qualifier where relevant.

Restructuring of Raw Data

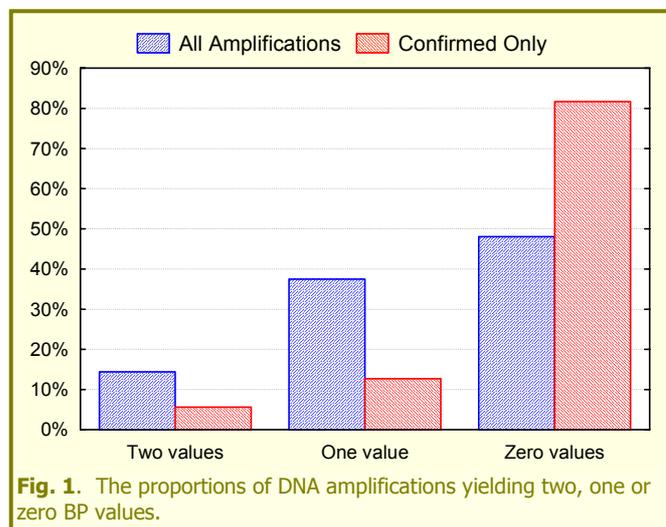
Ignoring the sex marker for now, this raw dataset actually represents a three-dimensional matrix with 529 rows (faecal samples), 18 columns (markers) and 2 planes (alleles). Multiplying these dimensions gives a total of 19,044 cells, or potential BP values. The most efficient way to store these values is to use a relational database. This allows the process of data normalisation, which removes repeating values from the actual data table and stores them once in separate, linked tables. This also has the huge advantage that data ambiguities are reduced or removed altogether.

This process was undertaken by building a MS Access database; the basic relational structure is shown in Appendix I. The process of transforming the “flat-file” structure of the Excel workbook into the fully relational database highlighted many data anomalies, which can easily creep into an un-structured and un-validated repository such as Excel. These were mostly removed using Visual Basic for Applications (VBA) programmes, although six cases each contained two different values, which had to be simplified arbitrarily.

Summary of Raw Data

The 18 DNA amplifications on each of the 529 faecal samples yielded a total of 9,522 possible amplifications. However, only 4,945 (52%) of these were successful amplifications, yielding either one or two BP values (Fig. 1). Furthermore, only 1,744 (18%) were confirmed values, meaning that 82% of all possible amplifications were either missing or unconfirmed.

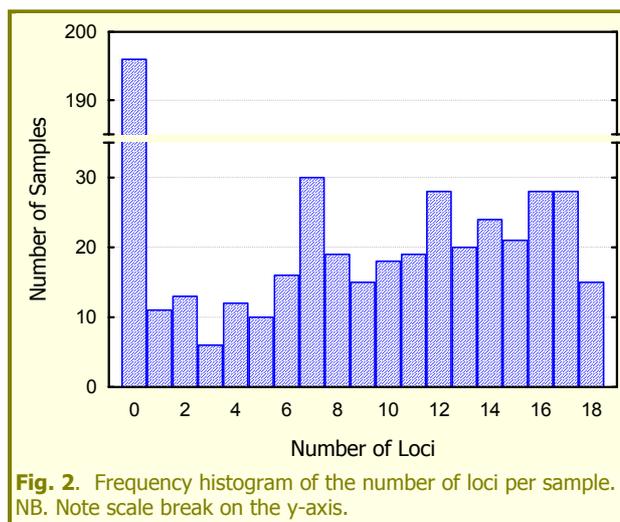
The 48% of DNA amplifications that yielded no BP value (allelic drop-out) caused huge problems throughout this analysis. Primarily, it meant that at this raw data stage, it became very difficult to interpret the amplifications that only yielded a single-value. When success rates are 100%, or at least very high, then a single BP value at a locus can be interpreted as the homozygous condition, with both alleles having the same BP value. Indeed, this was the advice from DNA Solutions, which was reflected in the original dataset, where all such results were displayed as two identical values at each locus. However, this is clearly a misinterpretation when, at best, half the amplifications failed. If we apply the known failure rate (48%) to the 3,572 single-value amplifications this yields a further 1,717 single failures and 1,855 true homozygous amplifications. By incorporating these values, we can now estimate an average failure rate of 57%.



Summary by Sample

Summarising the row dimension of the raw data matrix gave a frequency histogram of the number of loci with BP values in each sample (Fig. 2). Most importantly, 196 faecal samples (37%) yielded no BP values at all, leaving only 333 samples that could be used in the subsequent analysis. A further 52 samples (9.8%) had five or fewer loci with BP values.

These data included both confirmed and unconfirmed values. If only confirmed values were used, then 323 samples (61%) had no BP values at all and another 100 (19%) had five or fewer. Using the stringent criterion that the analysis should only use samples with six or more confirmed BP-values, would have reduced the sample size to 106 (20%).



Summary by Geographic Zone

The raw data were collected from nine geographic zones and two specialist teams (Table 1). The latter only collected five samples in total, all of which were unconfirmed. The number of samples collected from the geographic zones varied from 7 in Aural (with only 2 samples with confirmed BP values) to 137 in Chiphat (CP), of which 45% had at least one confirmed value.

Table 1. Total numbers of samples and proportions of confirmed and unconfirmed by geographic zone.

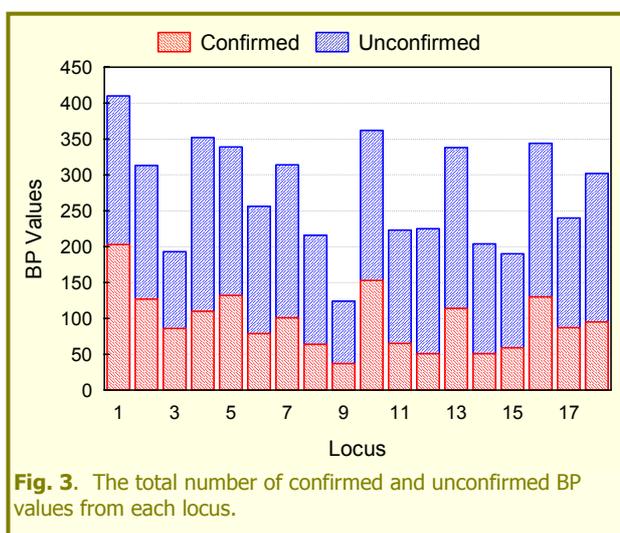
Zone		Faecal Samples	Confirmed	Unconfirmed	Samples with BP Values
Areng	AR	128	45 35%	87 68%	87 68%
Aural	AU	7	2 29%	7 100%	7 100%
Bokor and Kirirom	BK	12	0 0%	3 25%	3 25%
Botom Sakor	BS	77	47 61%	39 51%	58 75%
Conservation International	CI	3	0 0%	0 0%	0 0%
Chiphat	CP	137	62 45%	87 64%	89 65%
Dong Peng	DP	17	14 82%	17 100%	17 100%
Human Elephant Conflict Team	HEC	2	0 0%	2 100%	2 100%
Samkos	SK	45	20 44%	23 51%	34 76%
Viel Taphou	SW	80	0 0%	15 19%	15 19%
Chiphat	SWEC	21	14 67%	21 100%	21 100%

Note also that although Viel Taphou (SW) had 80 samples, none of them contained a single confirmed BP value, and only 19% of these had at least one BP value.

Summary by Locus (Marker)

The rate of allelic drop-out was significantly different between loci ($\chi^2_{(17)} = 365, p = 0$). The largest number of BP values obtained for a single locus was Locus 1 (Fig. 3), where the 410 values represented between 39% and 51% of all possible BP values, depending on the rate of heterozygosity. In contrast, Locus 9 had the lowest number of BP values (124) representing between 12% and 14% of all possible values.

The ratio between confirmed and unconfirmed values was also significantly different between loci ($\chi^2_{(17)} = 102, p \approx 0$). The largest proportion of confirmed values (50%) was found in Locus 1, with the smallest proportion (23%) in Locus 12.



Expected Base-Pair Ranges

The original dataset had the expected ranges of BP values for the 18 markers from a previous study of African elephants (Table 2). This provided the rationale for the use of the markers in this survey, but as Indian elephants belong to a different genus, there was no guarantee that they would be generally successful. Indeed, the proportions of actual BP values that fell outside the expected ranges were between 2% (EMU15) and 100% (EMU03).

This can be seen in the example of marker EMU01 (Fig. 4). There are clearly three main clusters of values, around 75.4, 77.7 and 79.0, with three or four other minor clusters. The two vertical red lines show the expected BP ranges and illustrate the 11% of observed values that fall within them. It is not clear at this stage what the consequences of this effect were. However, as there was no correlation between the proportion of BP values falling outside the expected ranges and the rate of allelic drop-out, this was not likely to be the cause of the small number of BP values yielded by some markers.

Table 2. Expected ranges of BP values for 18 markers from African elephants, plus actual BP values from this study.

Marker	BP Range		Actual BP Values	
	Lower	Upper	Mean	Outside Range
EMU01	78	82	77.3	89%
EMU02	108	116	109.8	69%
EMU03	137	143	131.4	100%
EMU04	97	107	100.6	20%
EMU05	112	118	113.5	22%
EMU06	146	160	149.0	68%
EMU07	102	120	108.4	40%
EMU08	115	127	120.2	10%
EMU09	163	169	163.6	63%
EMU10	94	104	97.1	51%
EMU11	122	136	124.6	9%
EMU12	120	152	131.7	3%
EMU13	100	110	100.1	44%
EMU14	130	138	137.4	78%
EMU15	142	154	150.7	2%
EMU17	119	137	123.6	11%
EMU18	138	142	138.4	33%
EMU19	99	115	108.2	53%

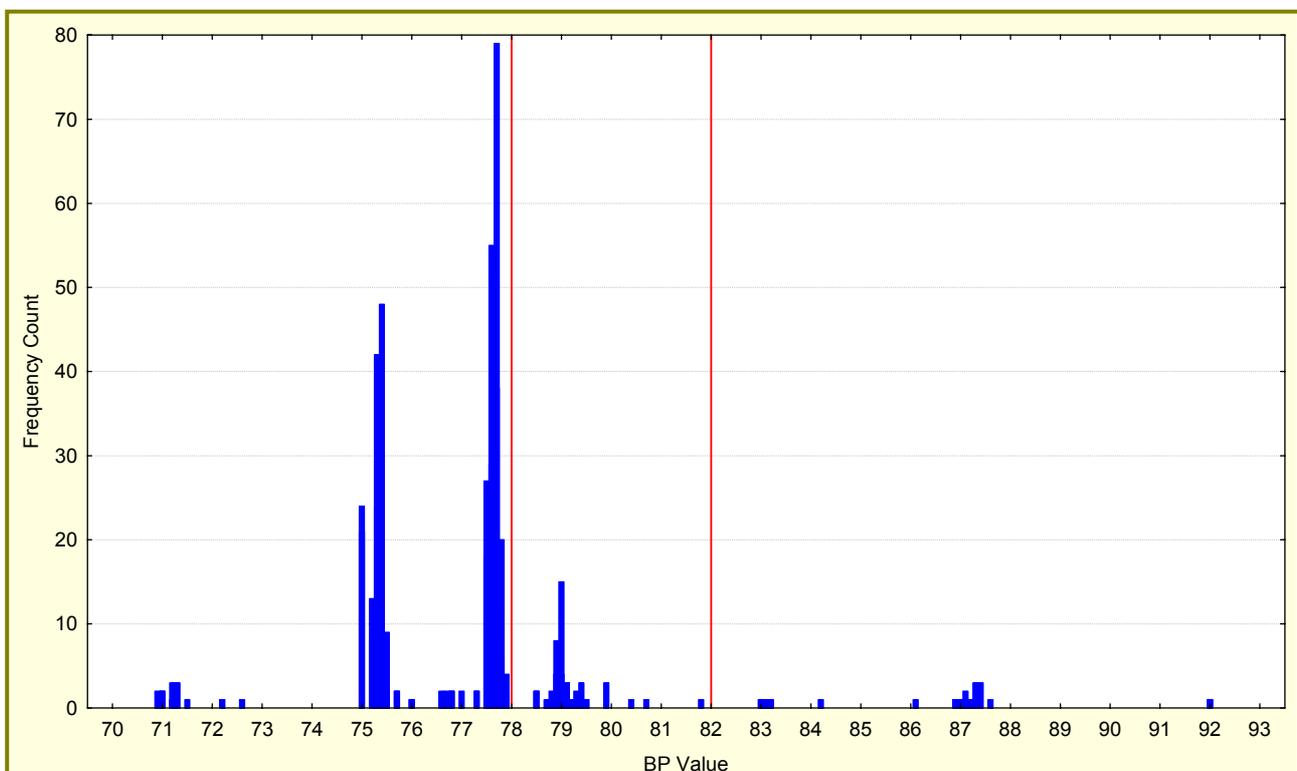


Fig. 4. Frequency Histogram of BP values for marker EMU01. The resolution of the x-axis is 0.1 and is truncated at 93.5, thereby excluding the single value at 101.4. The two vertical red lines represent the lower and upper ranges of BP values from African elephants.



Allele Classification

The PCR process extracts snippets of DNA between the two regions of a chromosome identified by each marker. The number of base-pairs in this snippet gives the BP value which, being a count, should obviously be an integer. But it is quite clear from the raw data and the frequency histograms (e.g. Fig. 4) that fractional values were usually obtained; caused by the averaging of a large number of readings within the PCR machine.

Different BP values represent different alleles. The problem is how to distinguish between similar non-integer values – should they be assigned to the same allele or two different alleles? This is essentially a one-dimensional clustering problem as illustrated in Fig. 4. For example, the largest cluster is focused around BP value 77.7, but how far either side of this value should the cluster extend? For example, it could just include values 77.6 to 77.8, but this would exclude the 27 values of 77.5. What about the scattered values around 77.0 – should any of those be included in this cluster or do they belong to their own cluster, and thereby generate records of a different allele?

The Genetic Algorithm

In many cases it was clearly adequate to define the clusters by eye. However, to remove any subjective element in this process, clustering was carried out using a custom-written genetic algorithm (GA). This is a type of computer-intensive programme that attempts to explain patterns in empirical data. It proceeds from a random starting point, through thousands of generations, “breeding” and “evolving” better and better functions that describe the data; hence the name. A full explanation of the process followed by this GA is given in Appendix II, but a brief résumé is given here.

For each of the 18 loci, the GA attempted to fit a series of functions that best described the observed frequency distributions. Each function defined a cluster, using three parameters;

- Σ the “location” of the cluster as defined by the mean BP value,
- Σ the “spread” of the cluster as defined by the coefficient-of-variation (V) and
- Σ the “size” of the cluster as defined by the proportion of all the BP values that fell into the cluster.

All functions were assumed individually to follow a normal distribution. The observed frequency distributions were modified to provide the counts of confirmed and unconfirmed BP values separately. In this way, confirmed values could be given a higher weighting than unconfirmed; in the final run of the GA these were set at 5:1 respectively. Two further constraints were applied to the functions within each locus;

- Σ no two clusters should have mean values within 0.75 of each other and
- Σ each cluster should contain at least 0.5% of the total number of BP values.

The first constraint was imposed to force clusters (alleles) to be about one whole BP apart. However, because there were many examples of clusters with means slightly less than one integer apart, the constraint was relaxed slightly to prevent two clearly separate clusters being merged into one, or one of the clusters being dropped entirely. The second constraint was required because, as the GA “evolved” more functions, the clusters became smaller so that, in some cases, the functions would define clusters that were too small to contain any actual data points.

Assignment of BP Data to Alleles

After the GA had been run successfully on all loci, the raw BP values could now be assigned to the alleles (clusters) defined by the functions. The alleles were given letter codes, within each locus, in ascending order of their mean BP value. Each raw BP value was tested sequentially against the functions and assigned to the allele to which it had the highest likelihood of belonging. However, if this likelihood was less than 0.5% (defined by the probability density function) then the BP value remained unassigned.



The final classifications of the BP values for each locus are shown as histograms in Appendix III. The minimum number of alleles defined for a locus was four (Locus 3) and the maximum was 18 (Loci 6 & 18). In summary, 94% of all values were classified successfully (Table 3a). The poorest classification rate was 86% (Loci 9 & 12), but three loci (3, 5 & 17) achieved 98%. The effect of the weightings applied to the confirmed BP values can be seen in Table 3b; the overall classification rate rose to 96%, with all except two loci (1 & 7) achieving 95% or greater. In absolute terms only 76 confirmed BP values were not classified, with nearly 40% of these in Locus 1.

So overall, the GA has worked extremely well in providing an objective means of classifying the numerical BP values into discrete alleles. Most importantly, it used unconfirmed BP values, with low weightings, to provide added information for the clustering algorithm. Furthermore, by taking the locus-based approach to defining alleles, it was possible to include data from samples that may only have had a few BP values. The sample-based approach favoured by DNA Solutions, meant that many of these would have been excluded, even if they were confirmed values, located directly within a clearly defined cluster.

Table 3. Summary of classifications of BP values into alleles, by locus, for a) all values and b) confirmed values only.

a)																						
Locus	?	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	Alleles	Classified Values	
1	38	11	2	112	8	185	36	3	3	12										9	372 91%	
2	12	2	200	7	2	3	79	8													7	301 96%
3	3	79	32	1	78																4	190 98%
4	27	16	6	4	7	8	20	31	24	150	41	3	4	11							13	325 92%
5	7	7	12	37	2	209	13	24	8	7	5	5	3								12	332 98%
6	11	1	144	2	2	1	2	4	2	20	7	21	20	3	2	1	4	2	7		18	245 96%
7	29	2	6	32	45	4	21	12	39	13	26	58	8	16	3						14	285 91%
8	15	13	56	33	12	6	4	14	63												8	201 93%
9	17	2	3	3	11	7	45	12	24												8	107 86%
10	16	17	19	23	47	8	5	8	11	2	91	4	17	29	8	20	28	9			17	346 96%
11	6	5	6	6	125	11	21	30	7	3	3										10	217 97%
12	31	5	7	73	5	1	29	22	47	5											9	194 86%
13	21	20	2	6	50	5	44	5	124	9	4	32	4	2	7	3					15	317 94%
14	15	62	2	23	14	2	4	5	2	55	19	1									11	189 93%
15	6	2	32	15	101	34															5	184 97%
16	18	9	5	10	44	49	3	41	6	130	8	2	4	8	2	5					15	326 95%
17	5	4	5	2	1	59	135	4	23	2											9	235 98%
18	22	4	5	2	5	2	5	3	55	4	4	6	5	5	2	6	42	7	118		18	280 93%
	299																					4646 94%
b)																						
Locus	?	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	Alleles	Classified Values	
1	29	3		63	3	93	10		1	1											7	174 86%
2	1		93				31	2													3	126 99%
3	2	29	18		37																3	84 98%
4	4	1	1			1	13	9	10	60	10			1							9	106 96%
5	1	1	3	14		99	1	11	1		1										8	131 99%
6	2		51				1			10	2	5	7						1		7	77 97%
7	8		1	7	17		16	4	13	2	7	14	1	11							11	93 92%
8	2	1	11	16	5	4		1	24												7	62 97%
9	1	1			7	1	21	5	1												6	36 97%
10	8	11	6	13	25		1		3		47		3	11	4	6	8	7			13	145 95%
11	0	1	2	2	46		6	7		1											7	65 100%
12	1		2	21			7	6	14												5	50 98%
13	6	3			25		12		54	2		10	1		1						8	108 95%
14	2	23		2	2					9	13										5	49 96%
15	1		8	6	32	12															4	58 98%
16	2	4		1	17	23		20	1	60			1	1							9	128 98%
17	1					33	44		9												3	86 99%
18	5	1	1						18						1	18		51			6	90 95%
	76																					1668 96%



Genotyping of Samples

The process of genotyping the samples using the alleles defined in the previous stage, was now straightforward. Table 3a shows that 4,646 BP values were assigned to alleles. These can be displayed as a two-dimensional matrix of letter codes, with the 333 faeces samples in rows and the 18 loci in columns. The third dimension of the raw data matrix, the two alleles, can be embedded within the cells of the matrix by using two codes where these were detected (Table 4). Furthermore, confirmed values are displayed with upper case letters and unconfirmed with lower case. The full matrix of these data is provided in Appendix IV as an embedded Excel Worksheet.

Table 4. The first ten faeces samples from the full matrix of 333 samples x 18 loci. Letters represent allele codes; upper case letters are confirmed BP values, lower case are unconfirmed.

Sample Code	Locus																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
AR001	E	cg		i	h	b	jk	cd	e	cp	ch		d	b	bd	ei	f	q
AR002	E	b		J	f	B	H	ah	eg	j	eh		h		d	IM	f	pr
AR003	F	b		dJ	f					gj	h	g	g		c	m		k
AR004	c	cg		f	f			a		h	e	g			d		b	fq
AR007				g	dj	i	k			p	e			i				q
AR008	e	c		jm	fh		k	c			ch	c				EI	f	q
AR009	cg	g		gj	f		c	a		g		g			d	im	e	iq
AR013	ef	B	ad	ij	E	b	cj	ah	eg	CM	BD	gh	h	i	d	di	f	R
AR014	CE	Bf	aD	cJ	cE	B	HJ	C	dh	Cl	D	cf			d	D	f	R
AR015	c	BF			E	b	k			Jo	G		hi			gi	h	a

This small sample shows a good range of results. Sample AR007 shows a very poor result with only nine alleles detected in eight loci. Furthermore, all of them were unconfirmed. At the other extreme, Sample AR013 had a total of 28 alleles detected in all 18 loci, although only seven of these were confirmed. It could be argued that Sample AR014 had a better result with 14 confirmed alleles.



Population Estimation

Having obtained a robust dataset of the genotypes of 333 faecal samples, the final stage was to estimate the population of elephants that produced the samples. To do this, pair-wise comparisons were made between all possible combinations of samples, to determine which samples differed from which others. With 333 samples, there were;

$$\frac{N(N-1)}{2} \text{ possible comparisons} = 55,278.$$

For each of these pair-wise comparisons the alleles at each locus were compared in turn. At first sight, this should result in 36 comparisons (2 x 18). However, bear in mind that the two alleles at a locus have no intrinsic order; AB is exactly the same genetically as BA. Because of this, it was necessary to compare each of the alleles in the first sample with each in the second, which gives up to four comparisons per locus.

However, this may not be necessary if a match is found. For example if the samples were AB and AB, then the first comparison would yield a match; A:A. The second comparison, would then only need to be made between the second alleles at each locus which, in this case, would also yield a match; B:B, which means that only two comparisons would be necessary at this locus. Imagine a comparison of AB to BC. In this case, the first allele in sample 1 would need to be compared with each of the alleles in sample 2, neither of which would give a match. The second allele would then be compared and would immediately yield a match, so requiring three comparisons. Consequently, the total number of comparisons between two samples could be anything between 36 and 72 or anywhere between 1.99 million and 3.98 million comparisons overall.

The "Absolute" Model

The simplest approach to undertaking the pair-wise comparisons is to use an "absolute" model, where we assume that both the confirmed and unconfirmed alleles are absolutely correct. Furthermore, with this model we are making two implicit assumptions;

- Σ that all loci with a single allele were homozygous, *i.e.* there was no allelic drop-out, and
- Σ that all null loci should be ignored in the comparison.

This model attempts to identify groups of identical samples. Firstly, of the 55,278 pair-wise comparisons, 51,641 (93.4%) had at least one different locus, and so by our rigorous criterion, must have come from different elephants. Of the remaining 3,637 comparisons where no alleles were different, 2,113 (58.1%) cases were due to having no identified loci in common, so no comparisons were actually made. This left 1,524 comparisons with at least one identical locus, and no differences. However, the vast majority of these (96.7%) had only one or two identical loci. Indeed, only 50 out of more than 50,000 pair-wise comparisons had three or more identical loci (Fig. 5) and there were only 16 comparisons with six or more identical loci, which is the traditional number used to match samples. However, this was a selection of six out of 18 loci tested, which is entirely different from only testing six loci and them all being identical. Furthermore, these 16 comparisons only comprised four homogenous groups, one in the BS geographical zone, two in the CP zone and one large group in CP but with a single record from AR (Table 5). This implied that we had 321 (333 - 16 + 4) animals producing 333 faeces samples - an extremely unlikely situation.

There are two main problems with this model, one producing false positives (apparently identical samples

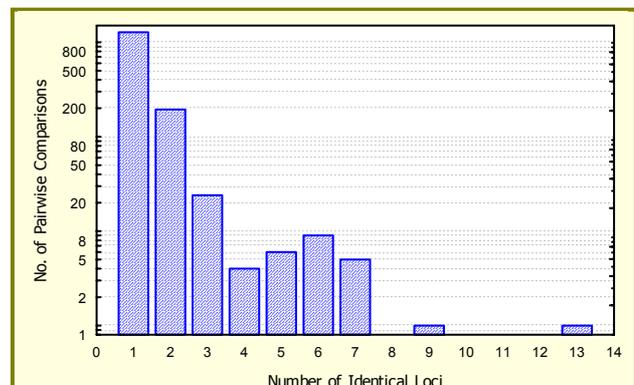


Fig. 5. Frequency histogram of number of identical loci from pairwise comparisons with no different loci. Note the logarithmic scale on the y-axis.



that in reality are not) and the other producing false negatives (apparently different samples that in reality cannot be distinguished). The figures given above suggest that the latter are far more prevalent.

False positives occurred in two ways. Firstly, null alleles were ignored, so clearly did not discriminate between two samples when they could, in fact, be different. The spreadsheet embedded in Appendix IV provides an example; CP013 versus CP018. These had nine identical loci, but also nine which were null in one or both samples. This left a very large “opportunity” for there to be differences between these samples. Secondly, single allele values were assumed to be homozygotes, but they could, in fact, have been heterozygous with allelic drop-out where the missing allele was different. Locus 1, for example, had allele E in both samples. By treating these as homozygous, they were considered to be two identical alleles, but it is possible that one or other was actually heterozygous and therefore not identical.

False negatives, also occurred for two reasons. Firstly, for example, CP013 is considered different from CP014 simply because of Locus 16. The unconfirmed Allele m was clearly different from the confirmed allele DE in CP013. However, although we don't know the likelihood, there was a possibility that the allele m was incorrect and could have been either D or E. The other cause of false negatives was the assumption of homozygosity when the two single alleles were different; as in the case of Locus 1 in samples CP001 and CP002. If we did not make this assumption, then it was possible that the missing allele in CP001 was C and the missing allele in CP002 was E, giving two identical diploid values (CE).

Table 5. The 16 pair-wise comparisons with ≥ 6 identical loci.

Sample1	Sample2	Identical Loci
AR22	CP66	7
BS17	BS19	13
CP13	CP18	9
CP49	CP50	7
CP49	CP65	6
CP49	CP79	6
CP50	CP65	6
CP50	CP79	6
CP51	CP62	6
CP51	CP63	6
CP62	CP63	7
CP64	CP65	6
CP64	CP67	6
CP65	CP66	6
CP65	CP67	7
CP66	CP67	7

The “Probabilistic” Models

The absolute model was clearly inadequate to identify both similarities and differences between samples. The alternative approach was to develop a “probabilistic” model which, rather than stating explicitly that two samples are either identical or not, calculates a likelihood that they are the same or different. This allows us to do two things;

- Σ to replace the missing alleles with a value before the pair-wise comparisons are made, and
- Σ to set a probability that confirmed and unconfirmed alleles are actually correct.

Four different models were developed to replace missing alleles; a simple random model, a model using an estimate of the proportion of homozygotes, a model using known allele frequencies and a model that used the linkage between loci.

The Simple Random Model

This model makes no assumptions about the missing allele and simply assigns one of the alleles belonging to the locus in a random way. So, for example, in Locus 1 there were nine alleles identified. A missing allele in a sample would, therefore, have an 11.1% probability of being replaced by any one of these alleles.

The “Homozygous” Model

This model makes the single assumption that a certain proportion of the samples that had only a single BP value at a given locus, were actually homozygous at this locus. This proportion was estimated separately for each locus, based on the lower and upper ranges of possible missing values. For example, Locus 1 had 119 samples with no BP values, 152 with a single value and 129 with two different BP values (Table 6). If we assume that all 152 single-value samples were homozygous, then the lower rate of allelic drop-out would be 22% (119 / 529). But if we assume that none of these were homozygous, then the allelic drop-out rate would be



37% ($119 + [152/2] / 529$). For simplicity, the arithmetic mean (30%) was taken as the likely allelic drop-out rate, so the rate of homozygosity was 70%. Overall across all loci, the homozygosity rate was 46%.

The “Allele Frequency” Model

A second assumption can be made, based on the considerable amount of information about the relative occurrence of different alleles within a locus, as shown in Table 3. In Locus 1, for example, just under half of all the identified alleles were E, whereas B, G and H, only occurred a few times each. By exploiting these frequency distributions the model used the likelihood that a missing allele would take any particular value. This process is highly intuitive because, when making the pair-wise comparisons, we can be fairly certain that any locus with a rare allele is likely to be different in another sample.

The “Linkage” Model

The fourth model takes this one stage further by utilising the known linkages between loci. For example, take the 185 samples containing Allele E at Locus 1 (Table 3). Now also assume that in these samples the allele at Locus 2 was usually B (of which there were 200 samples). During the pair-wise comparisons, if a sample was missing an allele in Locus 1, we could check to see what alleles were present in Locus 2 and if it was B, we could be fairly confident that the allele in Locus 1 should be E.

Setting the Confirmed and Unconfirmed Probabilities

By setting these probabilities at less than 1, we are able to accommodate the first source of false negatives; that an incorrect assessment of an allele has occurred. In this case, we can apply the procedures described above to replace it with another value. For example, if we set this probability at 0.9 then, when carrying out the pair-wise comparisons, we would apply a 10% probability that the allele is incorrect and replace it.

The values for the probabilities for confirmed and unconfirmed alleles were explored extensively, using simulated data. This indicated that as the probabilities declined, the genotypes of the samples approached a random or average composition. This made them less likely to be distinguished from each other. Consequently, confirmed probabilities were set at 1 and unconfirmed at 0.9, to allow a small variation from recorded allele values.

The Pair-wise Algorithm

The first stage in undertaking this part of the analysis was to set the correct hypothesis. The initial assumption was that all samples came from one elephant – the most conservative situation. Each pair-wise comparison then tested the null-hypothesis that there was no difference between the two samples. If this was rejected at an appropriate α -level (0.05 or 0.1), it would indicate that the samples had come from two different elephants.

To obtain an estimate of the probability that the two samples were different, Monte Carlo methods were used. As the process of completing the genomes was in itself a partial randomisation process, this was repeated thousands of times for each comparison. The test statistic was derived in the following way.

- Σ The two sample genomes were completed using the models described above.
- Σ The samples were compared allele by allele and the number of identical alleles was tallied.
- Σ One of the samples was compared with itself and the number of identical alleles was tallied.

Table 6. The Rate of Allelic Drop-out for each of the 18 Loci.

Locus	Samples with Missing Values:			Total	Proportion Missing		
	0	1	2		Lower	Upper	Mean
1	119	152	129	529	22%	37%	30%
2	216	159	77	529	41%	56%	48%
3	336	81	56	529	64%	71%	67%
4	177	156	98	529	33%	48%	41%
5	190	177	81	529	36%	53%	44%
6	273	148	54	529	52%	66%	59%
7	215	118	98	529	41%	52%	46%
8	313	96	60	529	59%	68%	64%
9	405	84	20	529	77%	84%	81%
10	167	104	129	529	32%	41%	36%
11	306	125	49	529	58%	70%	64%
12	304	125	50	529	57%	69%	63%
13	191	124	107	529	36%	48%	42%
14	325	84	60	529	61%	69%	65%
15	339	108	41	529	64%	74%	69%
16	185	108	118	529	35%	45%	40%
17	289	148	46	529	55%	69%	62%
18	227	110	96	529	43%	53%	48%
	4577	2207	1369	9522	48%	60%	54%



Σ These two counts were compared and the results stored as;

1 = between-sample count < within-sample count

2 = between-sample count = within-sample count

3 = between-sample count > within-sample count

This process was completed 1,000 times and two probabilities calculated;

Σ The one-tailed probability that the between-sample count of identical alleles was less than the within-sample count. This was a test of difference between samples.

Σ The two-tailed probability that the two counts were identical. This was a test of the similarity between samples.

To provide added precision to the Monte Carlo estimates, at the end of each cycle of 1,000 randomisations, the two probabilities were tested to see if either lay within a critical range of $\alpha / 2$ to $\alpha \times 2$. If so, a further cycle was run, up to a total of ten cycles. This meant that when the probabilities were close to the critical α , and remained so during the randomisation process, they were calculated from up to 10,000 randomisations. For example, for a critical α of 0.05, if the p -values after the first cycle of 1,000 were 0.001 and 0.999, respectively, then the process would end. However, if one of the p -values was 0.049, then further cycles would be carried out, until either the p -value “shifted” outside the critical range, or 10 cycles were completed.

To reduce the overall number of pair-wise comparisons, an initial pass was made through the 333 samples using a variation on the Monte Carlo algorithm described above. This tested the number of identical alleles derived from a comparison of a sample with itself against a the number of identical alleles from a purely random genome. In this way, it was possible to identify samples that had so few alleles that their full genomes could not be distinguished from random.

This algorithm was run twice, firstly using a critical α of 0.05 and then with an α of 0.1. On the first run, 139 of the 333 samples were not significantly different from random, leaving only 194 samples for the pair-wise comparisons. On the second run an additional 41 samples were included, due to the less stringent alpha level, giving 235 non random samples. This resulted in 18,721 and 27,495 comparisons respectively.

These two runs provided two separate sets of pair-wise comparisons. The first important result is that not a single pair-wise comparison showed any significant similarity. In fact, the lowest p -values were 0.870 ($\alpha = 0.05$) and 0.840 ($\alpha = 0.1$). This is not surprising, when the requirements for two samples to be identical are considered. For example, take two samples, one of which has 36 confirmed alleles and the other has 35 confirmed and one unconfirmed, such that the alleles at Locus 1 are AB and Ab. During the Monte Carlo process, all the confirmed alleles are compared “as found” because the confirmed probability was set at 1. This resulted in 35 identical comparisons. Furthermore, in 9 out of 10 randomisation the unconfirmed Allele b was also treated as a “B” and so the 36th comparison was also identical. However, in 1 out of 10 randomisations, this allele was replaced by another which, unless Allele B was either very common or strongly linked to another locus, it would have been a different allele. It would only take one or two unconfirmed or missing alleles for the test statistic to rise above 0.05 or even 0.1.

The other tail of the test statistic showed a completely different result. The vast majority of comparisons showed a significant difference between samples; 97% for both α -values. However, for a sample to be considered unique, it must be significantly different from all other samples – in the case of $\alpha = 0.05$, this would mean all the 193 comparisons in which it took part. In fact, only 54 samples met this criterion, with the remainder having from between 1 and 65 non-significant comparisons. Clearly, these 54 samples should be considered unique, but what about the 21 samples that were significantly different from all others except one? This is a common problem with pair-wise comparisons that has been addressed using cluster analysis.



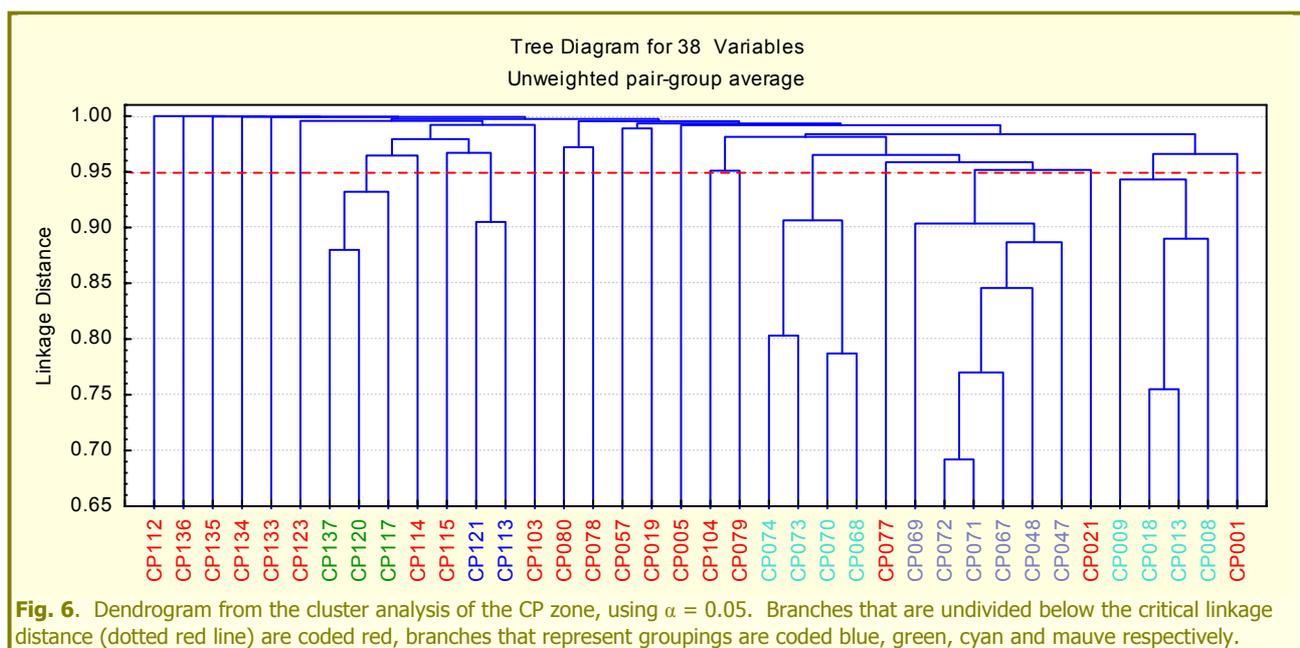
The Cluster Analysis

Cluster analysis is a standard method for grouping data points (records or samples) with similar values in multi-variate space. It produces a tree diagram (dendrogram) showing the “relatedness” between the different records, in much the same way as a family tree shows the relationships between people. A number of different clustering algorithms are available – for this analysis the “Joining (Tree Clustering)” algorithm was used, with “Unweighted Pair-Group Averaging” as the test statistic.

At this point it became necessary to consider the geographical zones from which the samples had been collected. Two zones, Aural (AU) and Samkos (SK) were considered to hold totally isolated elephant populations. Of the other seven groupings, one was derived from the Human Elephant Conflict team and the others were assumed to represent geographic zones through which elephants could roam freely. So the cluster analyses were run twice on each of the two sets of pair-wise comparisons. In both cases the samples from AU and SK were clustered separately, but in one run the remainder were treated as a single population (“The Rest”), whereas in the second run, they were clustered separately, as if they were seven different populations.

The cluster analyses allowed the individual samples to be classified into unique groups. The test statistic used $1 - p$ as the Linkage Distance. In other words, if the pair-wise comparison between two samples gave a p -value of 0.001, then the Linkage Distance would be 0.999. These values could then be tested against the critical α -values (0.05 or 0.1) as the cut-off points

As an example of this process, the cluster analysis of the CP zone using $\alpha = 0.05$ is shown in Fig. 6. Of the original 89 samples collected in this zone, only 38 had non-random alleles. The critical linkage distance of 0.95 is shown as a dotted red line in the figure. All branches that remained undivided below this level were considered to be significantly unique samples – these are coded in red. If they had a single branch below this level they represented a pair of samples (coded in blue) which were not significantly different from each other, but were significantly different from all other samples. Green, cyan and mauve colour-codes are used to signify homogenous groups of three, four and five or more, respectively. In this example, there were 19 unique individuals, one pair, one triple, two homogenous groups of four and a single group of six. This means that the original 38 samples are 95% likely to have come from 24 elephants.



There were two further considerations to be made before the final population estimates could be derived. Firstly, the sex of the elephant producing the faeces sample was derived from the SRY marker. Overall, there were 243 successful extractions at this marker, of which only 84 were confirmed. These data were not included within the pair-wise analyses or the subsequent cluster analyses. However, they could now be used within apparently homogeneous groups to identify different animals, if both confirmed males and females were present. The second consideration arose from the way that the pair-group averages were calculated in the cluster analyses. As their name suggests, when two different groups were compared, the unweighted averages of the linkage distances were used. This meant that, although it would be possible for two samples to be included within a group because most of their alleles were identical, at least one locus could contain three different confirmed alleles. If we are assuming the probability is 1 that confirmed alleles really were different, this would force the two samples to be significantly different. Using the example of the 38 samples from CP given above, we can explore these two “irreconcilable conflicts” (Table 7).

Table 7. The empirical genomes for the 38 samples from zone CP, based on the results of the cluster analysis with $\alpha = 0.05$. Colour coding of sample names is the same as in Fig. 6. Colour-coding of sex codes and alleles identifies irreconcilable conflicts.

Sample	Group	Sex	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CP112	1	m	ac	b		g	Ej	h	G	D		p	e	i	AK	ai		Ej		k
CP136	2	F	eF	b	a	ah	ek	kr	dh			l	e	Ce	ck	Ae	d		i	im
CP135	3	f	ef	b	a	j	b	kR	b	ag		cN		C	f	a		go	fh	
CP134	4	m	cd	BF	a	ej	G	kr	dk			b			k	a		i	f	eR
CP133	5	F	ei	B	a	j	gi	r	di			kn	e	bc	dm	a		d	ef	
CP123	6	f	f	ab		c	k			Eg			g	c	af	ci			cj	
CP137	7	f	ef	b		j	ce	br	di	bh	fg	d	d	c	df	a	bd	dh	f	hr
CP120	7	f	F	B			cE					dl	e	c	d	ai	b	d		
CP117	7	F	f	B		j	ce	b		b		dk	d		D	ai		cn	f	
CP114	8	F	F	B		J	cE			g		D	d	c	fj	ai	b	j	F	bh
CP115	9	m	F	b	a	gj	k	b	K	b		jp		c	D	i			f	R
CP121	10	m	f	B			CE	b	l	G		j			ad	i	d		f	r
CP113	10	M	F	B		J	CE		L			kn	d	ch	D	ci	d	i	F	R
CP103	11	f		b		j	ce	l	i			hM		c	f	aI	b	jn	fi	hR
CP080	12	f	CE	bf	a	G	e		c			mo		cg	fh	i		gi	E	h
CP078	13	F	C	B	AD	G	E	B		c		CO		gh	fH					hr
CP057	14	f	E	bf	A	I	e	B		B		bf			ei			b	f	d
CP019	15	m	e	b	a	i	e	b	k		a	pq		a	e	j	e	f		
CP005	16	F	CE	B	d	Fi	E	b	C		g	HJ	A	h	FH		D	im	ef	PR
CP104	17	F		Bg	aD	m	E	b	dh	ch	h		d	c	H	i		I	f	h
CP079	18	m	ce	BF	AD		Ce	B		H		PQ		c	H	I		i	F	
CP074	19	F	E	BF	AD	HI	EG	B	c	H	F	A	Dg	G	DH	D	BC	GI	F	PR
CP073	19	F	E	BF	A	HI	EG	B		H	EF	A	D	G	DH	D	BC	GI	F	PR
CP070	19	F	E	BF	A	Ij	EG	B	C	H	F	A	D	G	DH		BC	I	F	PR
CP068	19	f	E	bf	B	I	eg	B	c	H	f	a	d	g	DH	di	c	I	F	pr
CP077	20	F	CE	B	d	gI	CE	B		b		J		g	H		d	I	F	r
CP069	21	f	E	b	A	Hi	eg	B	C	h				c	e	ai		I	F	p
CP072	21	F	E	B	A	HI	EG	B		bH	F	AD		H	H		CD	GI	F	PR
CP071	21	F	E	B	A	HI	EG	B	c	bH	F	AD		H	H	aI	CD	I	F	PR
CP067	21	f	E		a	i		b	c	bh				H				I	f	
CP048	21	f	E		AD	I		B	d	B				H				GI	f	
CP047	21	f	e	b		ai		br	c			d		f				I	f	
CP021	22	f	E		d	i	eg	b	h		bf		a		h	j	c	gi	f	
CP009	23	F	e	B		i	E		h			J	ag	f	fh		cd	ei	f	pr
CP018	23	M	E	B		i	E		Dh			J	d		H	j	d	de	ef	
CP013	23	M	E	B		I	E	bl	DH		bf	J		f	H			DE	EF	r
CP008	23	F	E	B		i	E		h		bd	J	d	f	h		d	i	ef	pr
CP001	24	f	E	bf			E		h		A	JO	h			k	c	i		



Firstly, the single pair in this table had one confirmed and one unconfirmed male, so there was no conflict. The group of three had one confirmed and two unconfirmed females, so again no conflict. However, one of the two groups of four samples (CP009 – CP008) had both confirmed males and females, meaning that there must have been at least two animals producing these samples. This must, therefore, increase the overall count by one. Secondly, the other group of four samples (CP074 – CP068) had a conflict at Locus 3, with one sample having confirmed alleles “AD” and another “B”. So, although there were no sex conflicts, this group must also have comprised at least two individuals, increasing the overall count in the CP zone by another one to 26.

The Final Population Estimates

The final population estimates are shown in Table 8. Firstly, for a critical $\alpha = 0.05$, 194 non-random samples were found. By assuming the more conservative situation of amalgamated zones, the pair-wise comparisons and cluster analyses suggested a population of 110 elephants. This almost certainly represents the minimum population size as all assumptions take the most conservative position. However, if the confirmed sex markers are taken as 100% certain, we can assume five additional animals. And, if we assume 100% confidence in confirmed alleles, we can be 95% confident that the minimum number of elephants is 129.

If we assume that there is no migration between geographical zones, then a number of samples from different zones which were grouped in the previous cluster analysis, must now come from different animals. This results in a small increase in the estimates to 136. It should also be pointed out that the cluster analyses have worked in a very robust manner, because not one of the 12 pairs or 4 triples had a confirmed sex conflict.

The second run used a critical α of 0.1, resulting in 235 non-random samples. Taking the amalgamated results first, the cluster analyses identified 165 unique groups, only one of which had a sex conflict. In addition there were a further eight allele conflicts, giving an estimated total of 174 elephants. The results

of the assumption of isolated zones are very similar. These suggested only four more individual groups, no sex conflicts and one allele conflict giving an estimated total of 175.

These results suggest that the 90% confidence level has worked very well, for two reasons. Firstly, only one sex conflict was found in the amalgamated analysis and none in the isolated, suggesting that the cluster analyses were not spuriously grouping different animals because the sex was “unknown” to the analysis. Secondly, there was only a difference of one in the final estimate between the amalgamated and isolated analyses. So, although there were four more unique groups identified from the isolated zones, three of the amalgamated zone groups were subsequently split by allele conflicts.

Table 8. Summary of the number of samples and the population estimates from each of the geographic zones, for critical alpha levels of 0.05 and 0.10. Estimates are based on the unique groups as derived directly from the cluster analysis, plus the additional counts of irreconcilable conflicts.

Zone	$\alpha = 0.05$				$\alpha = 0.10$			
	Samples	Unique Groups	Irreconcilable Conflicts		Samples	Unique Groups	Irreconcilable Conflicts	
			Sex	Alleles			Sex	Alleles
<i>Amalgamated Zones</i>								
AU	1	1	1	1	3	3	3	3
SK	18	10	10	14	19	14	14	17
"The Rest"	175	99	104	114	213	148	149	154
Total	194	110	115	129	235	165	166	174
<i>Isolated Zones</i>								
AU	1	1	1	1	3	3	3	3
SK	18	10	10	14	19	14	14	17
AR	70	44	45	47	77	60	60	61
BK	2	1	1	1	3	2	2	2
BS	42	21	23	26	44	30	30	31
CP	38	24	25	26	59	34	34	35
DP	13	11	11	11	13	12	12	12
HEC	1	1	1	1	1	1	1	1
SW / SWEC	9	9	9	9	16	15	15	15
Total	194	122	126	136	235	169	169	175



Conclusions and Recommendations for Future Work

The biggest single problem to overcome in analysing these data was the extremely low success-rate of DNA extraction and amplification. Apart from rendering 196 of the 529 samples entirely useless due to zero amplifications, at best a further 98 samples had so few alleles identified that their genomes were not significantly different from random.

The reasons for this low success-rate are difficult to determine from the raw data. The dates of collection of most of the samples appeared to be between July 2006 and June 2007 and most were described at the time as “fresh” or “reasonably fresh”. If the DNA analysis did not take place until 2008, then much of the DNA degradation probably took place in storage. Presumably, both collection time in the field and laboratory analysis are expensive processes, so it would seem highly cost-effective to maximise the number of successful amplifications from the minimum number of samples.

Recommendation 1: Design the field survey and arrange laboratory time to minimise the period between sample collection and DNA analysis.

The accuracy and rate of confirmed extractions and amplifications was also very low. It was not clear from the dataset how samples became confirmed – presumably through a second amplification. It would be extremely valuable to quantify the causes of non-confirmed alleles. This would best be undertaken by a pilot analysis from known sources of dung. By taking multiple samples, it would be possible to quantify the within-sample and between-sample variances in error-rate. It would only be necessary to do this with, say, 25 unique dung sources with four samples from each. This would represent a relatively small effort to ascertain very important information about the rate of allelic drop-out and mis-identification. It would also be possible to test for storage time during this pilot, if the correct factorial experimental design were constructed.

During fieldwork, this issue should also be addressed by taking a small number of duplicate samples, say 5 – 10%. These should not be identified as duplicates, but treated as “true” samples during DNA analysis. Clearly, these will be identical, so any confirmed discrepancies between these samples can be used as a benchmark to indicate real rates of misidentification.

Recommendation 2: Pilot the DNA extraction and amplification process on known sources of faeces from either captive or domesticated elephants.

Recommendation 3: During fieldwork, collect between 5% and 10% duplicate samples. These should be treated as “true” samples during analysis and not identified as duplicates.

Another source of inaccuracy may have come about through the choice of markers. The empirical data show that several markers had almost entirely different alleles at their loci than the African elephants for which the markers were originally chosen. In particular, markers EMU11, EMU12 and EMU15 all had less than 10% of their BP-values within the expected range. Although this may not be a problem, it might indicate that these markers are not ideal for Indian elephants.

There was also a significant difference between markers in the success-rate for overall amplification, confirmed amplification and the final rate of assignment to alleles. In addition to the three markers mentioned above, EMU03, EMU06, EMU07, EMU08, EMU14, EMU18 and EMU19 all had fewer than 100 confirmed alleles.

Finally, from the number of different alleles identified at each locus (between 3 and 13 confirmed) eight or nine markers would be sufficient to identify individuals, although this would have to be confirmed if there were high rates of in-breeding and relatedness. Given that the major cost is DNA extraction and amplification, it would be far more cost-effective to undertake two (repeat) extractions / amplifications on nine markers than only one on 18 markers. Indeed, if increasing the number of markers simply increases the number of null BP-values, then this is entirely counter-productive.

Recommendation 4: Review the number of markers used in the DNA extraction / amplification.



The final, and most important, consequence of the low success-rate is that subsequent ecological analyses could not be carried out. The “absolute” pair-wise model was totally unable to identify samples usefully. Consequently, probabilistic models had to be used, which simply indicated the **probability** that two samples were not from the same elephant. However, it cannot be emphasised too strongly, that two samples not being significantly different from one another, **does not make them identical**. Indeed, the true test of identity showed that no single pair-wise comparison (out of a total of over 45,000) had more than a 16% probability of being identical.

So, this meant that no CMR analysis could be carried out because, in effect, there were no recaptures. It would be possible to treat the samples within a homogeneous group as recaptures (as long as they were from entirely discrete time periods), but this is highly unlikely to give any improvement, or a higher degree of confidence, to the analyses carried out here.

Furthermore, the low success-rate for allele identification made the calculation of Hardy-Weinberg ratios virtually impossible, and so rates of in-breeding could not be calculated. There are models that allow for a small amount of allelic drop-out, but given the high rates of complete drop-out and non-confirmed allele identification, there was little point in using them. Similarly, useful sex ratios could not be ascertained with only 84 confirmed alleles at the SRY marker. And finally, potentially important information on relatedness between individuals was entirely unavailable because of allelic drop-out.

Recommendation 5: In subsequent surveys, ensure that allelic drop-out is minimised to a level that allows accurate assessments of sex-ratios, rates of in-breeding and relatedness between individuals.

The final recommendation is a technical data-based issue rather than an ecological or genetic one. A great-deal of time was lost during analysis, due to poor data structure and ambiguity. For a dataset of this importance, it is essential that a well structured database is used to store both the ecological data and the raw genetic data obtained from the DNA extractions and amplifications. This is doubly important when data are being compiled from different sources and datasets are passed from one organisation to another. The correct relational structure and proper validation will ensure that data anomalies do not occur (such as the 69 samples that were collected in the future!). Even simple expedients like not using alpha-numeric sample codes that embed information within them will make data integrity much stronger.

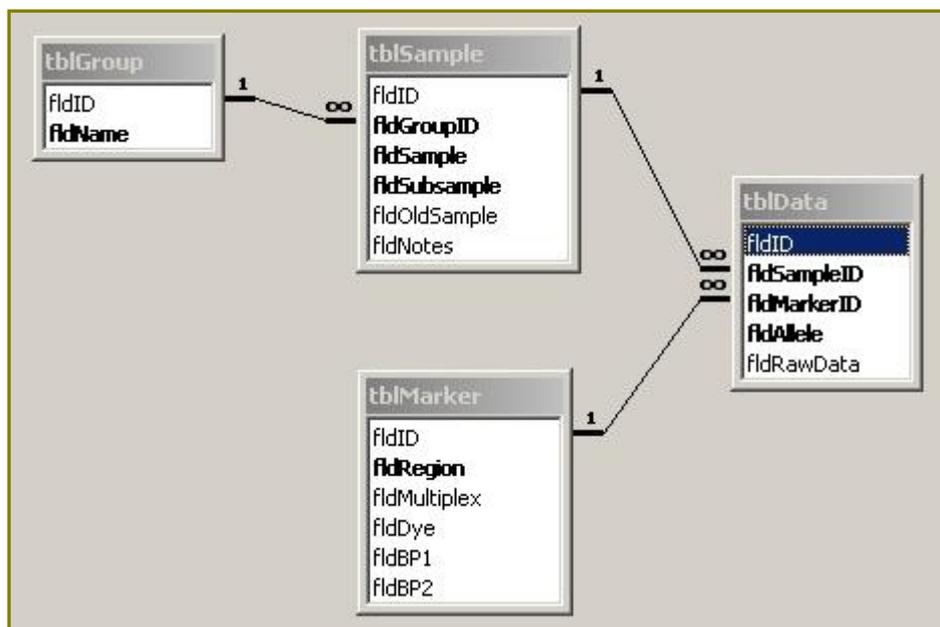
Recommendation 6: For future fieldwork and other FFI surveys, ensure that a well-designed database is constructed before field-work begins, which reflects the survey design and models the data structure and analytical procedure.



Appendices

Appendix I

The basic Entity/Relationship diagram for the normalised data structure



This model was implemented in a MS Access database. tblGroup held 11 records, one for each geographic zone. tblSample held the 529 records representing the faecal samples. Note that the old alpha-numeric sample code was stored as a text field, but was replaced by a numeric fldSample and an alpha fldSubsample. In combination with the Group ID, these formed the unique (Primary) key. tblMarker held 18 records, representing the markers (or loci). In addition, this table also held experimental information on the multiplex and colour of dye used, as well as the expected base-pair range that the marker had yielded in previous studies of African elephants. Finally tblData held the 3-D data matrix with dimensions identified by the Sample ID, the Marker ID and a code (1 or 2) for the Allele. The actual BP value was held in fldRawData. This table could have held the full matrix of 19,044 records (529 x 18 x 2) but actually only held the 4,945 successful DNA extractions. This is where a normalised data model achieves its efficiency.

The final database design was considerably more complex than this, as additional tables were added to store the results of the various analytical techniques used. Nevertheless, the use of Structured Query Language (SQL) then allowed the data from all the tables to be easily combined, summarised and extracted.



Appendix II

The Genetic Algorithm (GA) used to define clusters of BP values

In principle, all genetic algorithms attempt to “evolve” a set of rules or functions that describe the patterns in an empirical dataset. They do this by treating the functions as pseudo-organisms, which “breed” over thousands of generations to become better “adapted”, *i.e.* better fitted to the empirical data. They start with purely randomly generated functions, whose parameters are scored against the data, with only the best-scoring functions “surviving”. With each generation, the surviving functions mutate or recombine to create better-fitting functions with higher scores. Finally, after a specified number of generations, or a period of no improvement in the scores, the best fitting functions are stored. In technical parlance, the GA minimises a loss-function that represents the deviation of the model from the empirical data.

This GA was run separately for each of the 18 loci (markers). Furthermore, as GAs can have a tendency to become stuck in local minima, which may not represent the best overall solution, it was run ten times for each locus and the best scoring function-set was stored. Within each run the following procedure was undertaken;

- Σ Loop through 10,000 generations.
- Σ For the first generation create a breeding population of 10 function-sets, each with a single function that defines a cluster. Each function has three parameters; the mean of the cluster, the coefficient-of-variation (V) and the proportion of the population in the cluster.
- Σ For each subsequent generation, breed 10 new function-sets from the previous generation. Breeding can involve one of four “genetic” processes;
 - randomly create a new function,
 - randomly delete an existing function,
 - mutate a parameter value for an existing function or
 - merge a function from another breeding function-set
- Σ Constrain the 20 function-sets in two ways;
 - If two means are less than 0.75 BP apart, then merge the two functions,
 - If the proportion of the population in the cluster defined by a function is less than 0.5% then delete the function.
- Σ Score the 20 function-sets. Do this by calculating the cumulative (across all clusters) Probability Density Function (PDF) from a normal distribution for every value of BP at a resolution of 0.1. Use this with the total sample size to calculate an expected frequency distribution. Score this against the observed frequency distribution by tallying the positive weighting (5 for confirmed, 1 for unconfirmed) when an observed count lies within the expected frequency and a minus weighting when it lies outside the expected frequency.
- Σ Sort the 20 function-sets in descending order of their scores and “kill” the 10 lowest scoring sets.
- Σ Store the best score, and if there is no improvement over 1000 generations, drop out.

This process ran for an average of around 7,500 generations, each of which created and scored 10 function-sets, which on average held about 8 functions (clusters). Consequently, this required a total of;

$18 \times 10 \times 7500 \times 10 \times 8 \approx 108$ million cluster evaluations.



Appendix III

The Frequency Distributions for the BP Values from the 18 Loci

The genetic algorithm generated expected frequency distributions for the BP values in each locus. The histograms in the embedded Adobe file show these as red traces. These are superimposed on blue bars; dark blue represent the counts of confirmed BP values and light blue stacked on top represent the counts on unconfirmed values.

Note that the numeric coding for loci is 1 to 18, for simplicity, whereas the coding for the original markers was EMU1 to EMU19, with EMU16 excluded.



Appendix IV

Genotypes for the 333 Positive Faecal Samples

This embedded worksheet contains the matrix of faecal samples by 18 loci, with alleles coded as upper case letters for confirmed BP values and lower case for unconfirmed.

