



Comparative analysis of human mitochondrial DNA from World War I bone samples by DNA sequencing and ESI-TOF mass spectrometry

Rebecca Howard^{a,*}, Vesela Encheva^a, Jim Thomson^a, Katherine Bache^b, Yuen-Ting Chan^b, Simon Cowen^a, Paul Debenham^a, Alan Dixon^b, Jens-Uwe Krause^c, Elaina Krishan^b, Daniel Moore^b, Victoria Moore^b, Michael Ojo^b, Sid Rodrigues^b, Peter Stokes^a, James Walker^b, Wolfgang Zimmermann^c, Rita Barallon^a

^a LGC Limited, Queens Road, Teddington, Middlesex TW11 0LY, UK

^b LGC Forensics, Queens Road, Teddington, Middlesex, TW11 0LY, UK

^c LGC Genomics GmbH, Ostendstrasse 25, TGS Haus 8, 12459 Berlin, Germany

ARTICLE INFO

Article history:

Received 25 November 2010

Received in revised form 4 May 2011

Accepted 21 May 2011

Keywords:

Mitochondrial DNA
Forensic
Mass Spectrometry
DVI
Skeletal
Teeth

ABSTRACT

Mitochondrial DNA is commonly used in identity testing for the analysis of old or degraded samples or to give evidence of familial links. The Abbott T5000 mass spectrometry platform provides an alternative to the more commonly used Sanger sequencing for the analysis of human mitochondrial DNA. The robustness of the T5000 system has previously been demonstrated using DNA extracted from volunteer buccal swabs but the system has not been tested using more challenging sample types. For mass spectrometry to be considered as a valid alternative to Sanger sequencing it must also be demonstrated to be suitable for use with more limiting sample types such as old teeth, bone fragments, and hair shafts. In 2009 the Commonwealth War Graves Commission launched a project to identify the remains of 250 World War I soldiers discovered in a mass grave in Fromelles, France. This study characterises the performance of both Sanger sequencing and the T5000 platform for the analysis of the mitochondrial DNA extracted from 225 of these remains, both in terms of the ability to amplify and characterise DNA regions of interest and the relative information content and ease-of-use associated with each method.

© 2011 LGC Limited. Published by Elsevier Ltd. All rights reserved.

1. Introduction

The use of mitochondrial DNA in forensic science to identify potential familial relationships is well-established. The mitochondrial genome control region, or D-loop, contains two polymorphic regions known as HyperVariable regions 1 and 2 (HV1 (bases 16024–16365) and HV2 (bases 73–340)). [1,2]. These regions contain a large number of known polymorphisms, inherited as stable haplotypes, that can be used as a basis for familial matching. Mitochondria of individuals with different geographic origins also have patterns of variation so population databases, e.g. the European Mitochondrial POPulation database (EMPOP) [3] have been compiled to provide information on the frequencies of haplotypes in various population groups [4].

The Sanger sequencing method [5] is commonly used for mitochondrial DNA analysis since it is able to provide information on the position and base change of every polymorphic position within the analysed region when compared with a reference

standard, known as the revised Cambridge Reference Sequence (rCRS) [Genbank NC_012920]. DNA sequencing is also able to provide indications of heteroplasmy in some cases.

Despite being commonly used in the scientific community there are a number of issues with current DNA sequencing methods. The use of automation has substantially reduced the amount of hands-on time required both for sample preparation and subsequent sequence analysis but automated base-calling can be prone to errors, particularly in the detection of unbalanced heteroplasmies and mononucleotide repeats. Manual checking of automated results can therefore be time-consuming and labour-intensive. Nonetheless, precise sequence information is currently required for the interrogation of most available databases, e.g. EMPOP, which renders other analysis methods less attractive.

The Abbott T5000 system utilises an Electrospray Ionisation Time-Of-Flight mass spectrometer (ESI TOF MS) to characterise human mitochondrial DNA. The system consists of an automated DNA clean-up station, an Electrospray Ionisation Time-Of-Flight mass spectrometer (ESI TOF MS) and the incorporated IBIS Track software for data analysis. It has been demonstrated that ESI TOF MS has the potential as an analytical tool for the characterisation of nucleic acids such as oligonucleotides and PCR products, e.g. [6–9].

* Corresponding author. Tel.: +44 02089437540.

E-mail address: rebecca.howard@lgc.co.uk (R. Howard).

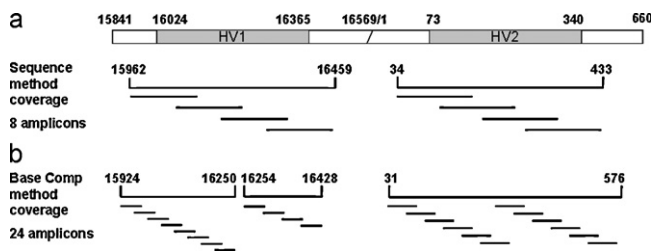


Fig. 1. Coverage map to show regions covered by each analysis method aligned to HV1 and HV2 in the D-loop control region of the mitochondrion. Not to scale.

The T5000 system takes advantage of the fact that the accurate masses of each of the four bases that constitute DNA are known. Therefore, for each measured amplicon mass, there are a limited number of possible base compositions [10]. This is additionally constrained by taking into account the rules of base complementarity and the resultant correlation between complementary DNA strands. Using unmodified nucleotides erroneous assignments could be made when two products have very similar masses, e.g. in cases where a double SNPs $G \rightarrow A$ and $C \rightarrow T$ result in only a 1 Da mass difference. To avoid this difficulty “heavy” ^{13}C -dGTP is used which introduces mass differences of 10 Da or more and allows the discrimination of such potentially ambiguous products. The T5000 approach (which will be referred to during comparisons as the ‘base composition method’) makes use of twenty-four overlapping PCR products of approximately 80–100 bp in length spanning the majority of sequence covering the HV1 and HV2 regions (Fig. 1). Note that a three base pair region of HV1 (positions 16251–16253) is not covered by the assay as it falls between two adjacent amplicons. Following amplification in triplex PCRs, the masses of both strands of each amplicon are determined by mass spectrometry and a base composition is assigned to pairs of complementary strands. Base compositions are then compared *in silico* to those predicted from the rCRS. A minimum number of differences is calculated for each amplicon and for the sample as a whole.

Although base compositions cannot be used to search sequencing databases directly the T5000 system avoids this problem by providing its own database permitting conversion of sequence data into base compositions for profile searching. This provides the facility to convert any known DNA sequence, or group of sequences, spanning the HV1 and HV2 loci into base composition thus permitting the interrogation of existing databases.

The T5000 system has previously been validated by the FBI using DNA extracted from fresh buccal swabs [6,7]. While buccal swabs will no doubt constitute a proportion of samples submitted for mitochondrial DNA analysis a large proportion of mitochondrial analysis in forensic casework is of old or degraded samples. The T5000 system must therefore also be capable of analysis of these sample types if it is considered to be a realistic alternative to the Sanger method.

Direct comparison between two methods is difficult, since one method is invariably favoured over the other. For example, there is a fivefold difference in recommended input DNA volume (5 μl in aT5000 reaction vs. 1 μl in a Sanger sequencing reaction) which may favour successful amplification in the T5000 PCRs. The T5000 conditions also use one additional cycle of PCR and the assay targets shorter amplicons. This is partially offset by the fact that amplification of T5000 products takes place in less efficient triplex reactions, while the Sanger sequencing is carried out in singleplex reactions. What can be compared, however, is the results obtained from each method when used under recommended conditions. It is this comparison that is the subject of the study described here.

The opportunity to test the two systems in parallel on appropriate sample types was presented during a Commonwealth War Graves Commission project to identify the remains of 250 soldiers buried during the First World War. The T5000 was tested under manufacturer-recommended conditions for the purpose of this comparison.

On 19th July 1916 the Australian Army’s 5th Division and the British Army’s 61st Division engaged German troops near Fromelles in northern France. By the time the battle, now known as the Battle of Fromelles, was over it is believed that 1708 Australian and 503 British soldiers had been killed. Many of these soldiers could not be accounted for in the aftermath of the conflict. In 2007, work carried out by the Glasgow University Archaeology Research Division (GUARD) led to the discovery of 250 sets of remains in a mass grave at Pheasant Wood, an area just outside Fromelles. In 2009, the Commonwealth War Graves Commission announced a project to identify as many sets of these remains as possible and, in 2010, the remains were reburied in a designated military cemetery. The last of these was interred at a commemorative ceremony on 19th July 2010, the 94th anniversary of the battle. Headstones on at least ninety-six of the graves carry the names of Australian soldiers identified as a result of the Commission-sponsored project (Commonwealth War Graves Commission, <http://www.cwgc.org/fromelles/>).

Mitochondrial DNA analysis of the remains was carried out in the forensic laboratories at LGC. Of the 250 sets of remains discovered by the archaeological team directly comparable DNA data was available from both the Sanger sequencing and base composition methods for 225. These 225 samples were used as an example dataset. This paper will evaluate both methods using several criteria including coverage within regions of interest, quality of information available and the discriminatory power when profiles generated by each method are searched against a standard sequence database of potential relatives.

2. Materials and methods

All reagents were obtained from Abbott Molecular unless otherwise specified.

2.1. Bone decalcification and DNA extraction

Individual teeth or bone fragments (0.9–1.5 g) were ground to a fine powder and resuspended in 16 ml 0.5 M EDTA to decalcify. The suspension was mixed by inversion and then on a Luckham R100 Rotatest Shaker for 24 h. The suspension was then centrifuged and the supernatant was then aspirated from the pellet. This process was repeated 5–6 times, before a final centrifugation step to pellet the decalcified material to be used for DNA extraction.

The ground and decalcified tooth or bone sample was mixed with 15 ml ATL buffer and 750 μl proteinase K (Qiagen, UK) and incubated at 56 °C for a minimum of 12 h. The samples were centrifuged and 15 ml AL buffer added (Qiagen, UK). This was vortexed gently until all powder was suspended and incubated at 70 °C for 1 h. Samples were then washed with absolute ethanol and purified on a Qiagen Maxi Column (Qiagen, UK). The DNA was eluted in 3 ml Qiagen AE buffer prior to downstream analysis.

2.2. DNA yield

Extracted DNA was concentrated using Amicon columns (Millipore, Billerica, MA). Prior to amplification, human genomic DNA was quantified using the Quantifiler Human DNA Quantification kit (Life Technologies, Carlsbad, CA) under manufacturer-recommended conditions.

Table 1

List of Sanger sequencing amplicons and primer sequences. The assay uses eight pairs of primer sequences for amplification and sequencing. Primer pairs were designed to overlap to ensure full double-strand sequence at all positions including the problematic 309–315 poly-C tract.

Region	Amplicon ID	Start position (excluding primer)	End position (excluding primer)	Primer ID	Primer sequence	Length (bp)
HV1	HV1-1	15942 (15962)	16098 (16078)	mito I-1F	TTT CCA AGG ACA AAT CAG AG	157
				mito I-1R	TAC GAA ATA CAT AGC GGT TG	
	HV1-2	16020 (16040)	16210 (16190)	mito I-2F	TCT GTT CTT TCA TGG GGA AG	191
				mito I-2R	TAC TTG CTT GTA AGC ATG GG	
	HV1-3	16132 (16152)	16293 (16273)	mito I-3F	ACC ATA AAT ACT TGA CCA CC	162
				mito I-3R	TGG GTA GGT TTG TTG GTA TC	
	HV1-4	16217 (16237)	16436 (16416)	mito I-4F	TCA ACC CTC AAC TAT CAC AC	220
				mito I-4R3	AGC TAC CCC CAA GTG TTA TG	
HV2	HV2-1	14 (34)	201 (181)	mito II-1F	TCA CCC TAT TAA CCA CTC AC	188
				mito II-1R	TTA GTA AGT ATG TTC GCC TG	
	HV2-2	126 (146)	281 (261)	mito II-2F	ATC TGT CTT TGA TTC CTG CC	156
				mito II-2R	TGA TGT CTG TGT GGA AAG TG	
	HV2-3	184 (204)	346 (326)	mito II-3F	GGC GAA CAT ACT TAC TAA AG	163
				mito II-3R	AGA GAT GTG TTT AAG TGC TG	
	HV2-4	243 (263)	453 (433)	mito II-4F	AAT TGA ATG TCT GCA CAG CC	211
				mito II-4R	AAT AAT GTG TTA GTT GGG GG	

2.3. DNA amplification for sequencing

PCR mix (per sample) comprised of 2 μ l 10 \times AmpliTaq Gold Buffer (Life Technologies, Carlsbad, CA), 4 μ l 5 M Betaine (Sigma–Aldrich, Dorset, UK), 2 μ l 2 mM dNTPs (GE Healthcare, Bucks, UK), 1.2 μ l 25 mM MgCl₂ (Qiagen, UK), 100 pmol each primer (sequences in Table 1) (Eurofins, Germany) and 0.3 μ l AmpliTaq Gold (5 U/ μ l) (Life Technologies, Carlsbad, CA). This was made up to a final volume of 19 μ l with tissue culture water (Sigma–Aldrich, Dorset, UK). 1 μ l extracted DNA was added to each 19 μ l mix. Amplification was carried out on a 9700 Thermal Cycler (Life Technologies, Carlsbad, CA). The amplification thermal protocol comprised a 95 °C incubation for 10 min; 35 cycles of 95 °C, 30 s; 50 °C, 30 s; 72 °C, 1 min; and hold steps 72 °C, 10 min, 8 °C, ∞ .

4 μ l PCR product in 1 \times gel loading dye (final volume 20 μ l) was run on a 1.5% agarose gel (containing 1 \times TBE and 1 \times Gel Red (Cambridge Biosciences)) at 150 V for 30 min. Samples for which PCR products were able to be visualised on the gel were then used as described in the DNA sequencing method.

2.4. DNA sequencing method

1 μ l Exonuclease 1 (EXO1) (1 U/ μ l) (GE Healthcare, Bucks, UK) was added to 2 μ l EXO1 buffer, 20 μ l Shrimp Alkaline Phosphatase (SAP) (New England Biolabs, Ipswich, MA) and 17 μ l tissue culture water (Merck, Germany) and mixed. 3 μ l was added to 7 μ l PCR product and incubated at 37 °C for 15 min and 80 °C for 15 min on a 9700 thermal cycler (Life Technologies, Carlsbad, CA). Product concentration was estimated using a 2% agarose gel and a 100 bp DNA ladder (Fermentas, UK) of known concentration for comparison.

10–100 ng PCR product was mixed with 0.4 μ l BigDye 3.1 (Big Dye Terminator 3.1 sequencing kit, Life Technologies, Carlsbad, CA), 1 μ l 10 \times sequencing buffer, 5 pmol sequencing primer and 2.5 μ mol betaine in a 10 μ l reaction. The sequencing thermal protocol comprised a 95 °C incubation for 5 min, 30 cycles of 95 °C, 30 s; 55 °C, 30 s; 60 °C, 4 min. The reaction was held 4 °C prior to cleanup.

Ten microlitres of tissue culture water (Merck, Germany) was mixed with the sequencing reaction and 18 μ l pipetted onto a 96-well filter plate with base (Performa DTR V3 96-well short plate kit, EDGE Biosystems, Gaithersburg, MD, USA) and centrifuged at 850 \times g in a Beckmann Allegra 6R centrifuge (Beckmann Coulter, Brea, CA) for 5 min.

Plates were run on a 3730XL DNA Analyzer (Life Technologies, Carlsbad, CA) using a 50 cm capillary array, POP7 polymer using

a standard module with 10 s injection times and a 6.5 kV run voltage.

2.5. Abbott method

2.5.1. Abbott amplification

Five microlitres of extracted DNA was added to each of eight wells (one column) in an Abbott T5000 mitochondrial typing plate, which contains pre-aliquoted PCR mix including primers [7], buffer, dNTPs and polymerase enzyme (Human Forensics (Mitochondrial) kit (Catalogue MG-00105)) and centrifuged (2000 rpm, 1 min). Amplification was performed using an Eppendorf Mastercycler (Eppendorf, Germany). Reaction conditions were 95 °C, 10 min, followed by 36 cycles comprising: 95 °C, 5 s; 50 °C, 1 min 30 s (ramp rate 5%); 72 °C, 20 s. The plates were incubated at 72 °C for 4 min and 95 °C for 10 min. Plates were held at 4 °C prior to T5000 analysis.

2.6. Sample clean-up

PCR products were desalted using magnetic beads coated with weak-anion exchange resin as previously described [11]. An eight-channel, fixed-needle liquid-handling robot based on a mini auto-sampler (LEAP Technologies) was used to perform a series of washing steps with the clean-up reagents and to transfer the sample aliquots. Initially, the PCR products were transferred to the desalting plate and bound to the magnetic beads. Residual dNTPs, salts, enzymes or any other interfering substances were removed by rinsing the magnetic beads with the NH₄HCO₃/methanol based solvents. The purified/desalted PCR products were eluted in 25 μ l high-pH buffer in a 96-well plate and transferred to the spraying platform using a Thermo Catalyst Express Robotic Arm. All liquid containing plates were automatically sealed using a heat sealer (KBioscience) in order to minimise solvent evaporation.

2.7. Mass spectrometry detection

As part of the IBIS T5000 system, the ESI-TOF was coupled to Tecan Cavro programmable pumps and switching valves. Analyte solutions were electrosprayed at 3 μ l/min against a counter current of dry gas (nitrogen). The ESI source is equipped with an off-axis sprayer and glass capillary operating at 3750 V relative to the spraying needle. The ionisation conditions used allowed for the two strands to be dissociated while avoiding fragmentation of the DNA backbone.

Spectral calibration and data processing were automatically initiated as the next sample is injected into the mass spectrometer. The data was acquired as 72 μs ESI TOF scan including a 34.72 μs delay followed by 36.28 μs comprising 75,200 data points. For each spectrum 660,000 scans were co added.

The mass spectrometry data was internally calibrated using two peptide standards (726.4169 Da and 1346.7048 Da), added to the elution buffer, which bracket the m/z range of the charge state envelope of the expected PCR products. All aspects of the data acquisition were performed using Bruker MicroTOF software package controlled via the IBIS T5000 Controller interface.

2.8. Data processing and deconvolution

For the T5000 mitochondrial assay the processing software performs background subtraction, mass calibration using the two internal standards and the deconvolution of the signal. The raw m/z data is transformed into “mass” data using the MassCollapse algorithm [12]. The calculated masses are imported directly into the database for storage and retrieval during sample analysis.

2.9. Base composition determination

For a product to be assigned for a given primer pair there must be a mass consistent with base composition possible for the forward strand and simultaneously there must be a mass consistent with base composition possible for the reverse strand. The algorithm is also constrained by Watson and Crick base pairing rules and mass errors less than 70 ppm. Additionally, the combined error of calculated and predicted masses for both strands should not exceed 80 ppm. The signal intensities of the two strands must be within a defined ratio of each other where the more intense peak cannot be less than 2.5 times the signal of the less intense peak. Once products have been assigned for each primer pair a composite base composition profile is assembled for each sample ordered by product coordinates relative to rCRS.

3. Results and discussion

3.1. DNA yield

Of the 225 samples tested no nuclear DNA was detected using the Quantifiler kit in eleven. This implies that nuclear DNA concentration in these samples was below the limit of Quantifiler detection, since no indication of inhibition was detected in Quantifiler internal positive controls. Of samples for which a value was generated, concentrations ranged from 0.7 pg/ μl to 3.18 ng/ μl . The median quantification value was 50 pg/ μl , indicating that a large proportion of samples had low copy number nuclear DNA.

3.2. Sample detection and HV1 and HV2 coverage

DNA samples were used to generate two sample datasets: (1) a single set of sequence data using a previously-validated Sanger sequencing method. This was considered the ‘standard method’ and utilised as a reference; (2) the T5000 base composition data, with each sample amplified in duplicate, for evaluation of the ESI-TOF system.

Despite indications that several samples were of poor quality, at least partial product data was available for each of the 225 sequenced samples and the 450 (2×225) sets of mass spectrometry analytes. Mass spectrometry analysis can be considered in terms of number of base pairs analysed or of pass/fail rates of individual amplicons. Analysis of amplicon pass/fail rates shows that 58% of samples analysed had full profiles and that almost 99% of samples generated products for at least 75% of amplicons. This

information is given in Fig. 2. Since sequencing amplicons cannot be considered discrete products (as reads of acceptable quality sequence can begin and end at any point within an amplicon), the pass/fail rate, and therefore any comparison, is best described in terms of number of bases for which data is available. This is highlighted in Fig. 1.

A large number of mitochondrial haplotypes derived from sequence data are already deposited in databases such as EMPOP [3]. Any new analysis method, therefore, must be as compatible as possible with existing databases. Since databases such as EMPOP are generated from DNA sequence information, and the T5000 system has the ability to add previously-sequenced haplotypes to its in-built database of base compositions, both Sanger sequencing and the T5000 are compatible with existing databases. The number and position of the base pairs interrogated by each method is also comparable.

Both the Sanger sequencing and T5000 base composition methods were designed to span regions larger than the designated HV1 and HV2 [1,2]. The Sanger sequencing primers are designed to interrogate bases 34–433 and 15962–16459, while the T5000 amplicons span an even greater range of 31–576 and 15924–16250, 16254–16428 (Fig. 1). A degree of overlap also exists between the amplicons within both the Sanger sequencing and T5000 methods. In order to make a direct comparison between the two methods the datasets were simplified to denote coverage within the designated HV1 (16024–16365) and HV2 (73–340) regions. As amplicon overlap is positioned differently between methods multiple coverage of individual base pairs was ignored for the purposes of this calculation. Calculations were therefore based solely on the number of rCRS bases within the defined HV1 and HV2 regions reported at least once by each method. This data is presented graphically in Fig. 3. Data is presented as percentages with the two base composition datasets combined prior to comparison with the sequencing dataset.

It can be noted that coverage using the base composition method in the HV1 region is slightly reduced when compared with that of the Sanger sequencing. This is due to the inherent absence of coverage on the three base pairs between positions 16250 and 16254. If this is discounted the difference between the mean sequencing coverage (98.23%) and the mean base composition coverage (96.89%) is less than 2%. Even when this three base pair region is included in coverage calculations the difference between coverage remains below 3%. Since SNPs were not detected at rCRS positions 16251–16253 within our dataset the T5000 assay design did not result in loss of coverage at any informative loci in this study. However, variants have been previously characterised in this region [13] which, if present in the example dataset, could not have been detected using the base composition method.

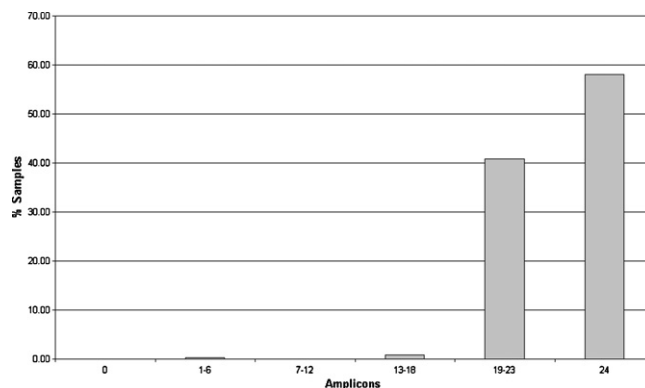


Fig. 2. Number of products generated during amplification of each of the T5000 sample replicates (maximum 24).

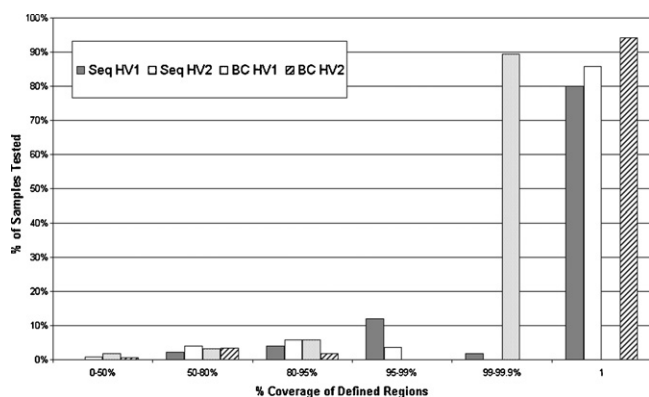


Fig. 3. Coverage statistics for the HV1 (16024–16365) and HVII (73–340) regions. SEQ: Sequencing data coverage; BC: base composition data coverage. No HV1 BC results demonstrate 100% coverage because of the 3 bp gap (positions 16251–16253) in the assay coverage.

When HV2 is considered the data is even more comparable. Mean coverage within the EMPOP-designated region is almost 0.5% higher using the base composition dataset (98.43%) than when using the Sanger sequencing method (98.01%). When both amplicons are considered there is very little net difference (0.46%) between the two methods (98.12% of maximum method coverage using the sequencing method compared with 97.66% for the base composition method).

The data presented therefore demonstrates that the Sanger method has a slight advantage over the base composition method with respect to coverage due to the 3 bp gap within the HV1 region of the latter. Neither method has a substantially greater success rate in the regions covered by both.

3.3. Reproducibility of the T5000

Amplification and analysis of each sample using the base composition method was carried out in duplicate. These duplicates were compared directly in order to ascertain the reproducibility of the method.

Of 10,800 PCR products expected to be generated in the duplicate amplification of the 225 samples, base compositions were obtained for 10,435 (96.6%). For 171 of these products no product was present in the second replicate for comparison. Performing duplicate analysis of the same sample can therefore help to increase coverage as it improves the chances for detection of low intensity products.

For the amplicons whose replicates generated only single, non-heteroplasmic products, base compositions were identical between replicates with the exception of two amplicons (sample 49247, amplicon 2896; and sample 49899, and amplicon 2908). Amplicon ID numbers are as described in [7]. In the first of these, replicate B from amplicon 2896 matched the sequencing data and had an assigned base composition (including primers) of A45 G13 C41 T24. The mass spectrum showed peaks corresponding to both strands with signal-to-noise ratios of over 45 and peak heights over 800 units. In contrast, replicate A, which had an assigned composition of A41 G14 C45 T21, had an assigned signal to noise ratio of 29.4, although the peak heights (of approximately 60 units) were similar to the background peaks so this ratio was apparently erroneous. It was obvious upon reviewing this sample that peaks identified in replicate A were not valid and should have been disregarded, although the T5000 software, using default analysis settings, had assigned them as valid peaks.

The second case is more ambiguous. Replicate B of sample 49899 amplicon 2908, which had an assigned base composition of

A42 G16 C39 T32 (including primers), matched the sequencing data, while replicate A (A39 G17 C39 T34) did not. The product peak heights for replicate A (100–150 units) are lower than those for replicate B (~200 units), but the difference is not substantial. The signal-to-noise ratios of the peaks are also similar – replicate A has ratios of 30.6 and 22.3, while replicate B has ratios of 35.5 and 28.8. The T5000 settings can be altered from the manufacturer's recommended settings to reduce detection sensitivity and prevent calling of these low intensity products, but in the case of sample 49899, the genuine product would be difficult to separate from the false one. It is possible that this false product has arisen as a result of contamination within the reaction. This will be discussed in greater detail below.

3.4. Detection of heteroplasmity

Point heteroplasmity can be detected by the base composition method by the observation of extra peaks in addition to the six expected in each triplex reaction. The software can assign base compositions for these additional peaks and so present two alternative base compositions for any given amplicon containing a point heteroplasmity.

Similarly, length heteroplasmity, frequently encountered in (C)_n stretches in both HV1 and HV2, is characterised by additional peaks representing products with an increased mass corresponding to one or more additional cytosine residues.

Additional peaks, assigned by the T5000 analysis software as evidence of heteroplasmity, were indicated in at least one replicate of 108 of the 225 of teeth/bone samples tested (48%) and in both replicates in 40 of the 225 (17.7%). Of these 40 samples, 12 had identical reported heteroplasmity profiles in both replicates, and a further 14 had identical indications of heteroplasmity in at least one amplicon with additional evidence for the same heteroplasmity in at least one other adjacent amplicon. Since some instances of heteroplasmity are expected to be detected in more than one amplicon, this provides further evidence that the heteroplasmity detected is genuine. For the remaining 14/40, each replicate had additional product peaks and associated assigned heteroplasmity not seen in the corresponding amplicon of the other. Only four of these had supporting evidence of heteroplasmity in overlapping adjacent amplicons.

In 60/108 samples with some reported heteroplasmity, the heteroplasmity was detected in a single amplicon only once in one of two replicates and not at all in the other; and in the remaining 8 samples, heteroplasmity was detected in only one of two replicates but in multiple amplicons. It is the results from these eight samples that would be most indicative of contamination, particularly where three targets are amplified in the same tube, since potential contaminating peaks would appear several times. Similar peaks were not detected in the DNA extraction, PCR or PCR cleanup negative controls, but the possibility exists that this contamination could have been present in the original bone and tooth samples themselves or as low-level laboratory contamination that would not have been detected had additional DNA not been present, i.e. the 'carrier effect' [14]. The carrier effect is often seen in ancient DNA samples, and despite these samples being much more recent, they nonetheless contain low-level DNA, and a similar effect could have occurred here. As mentioned previously, the T5000 settings can be modified (default minor:major peak ratio threshold is 0.15 for point heteroplasmity) to reduce the sensitivity of detection, and this would need to be considered as part of the validation of the T5000 system in a forensic setting. This will be discussed in greater detail below.

Across the 108 samples there were a total of 202 amplicons for which the T5000 software indicated heteroplasmity. Fifty of these indicated the presence of length heteroplasmity while 152 indicated that point heteroplasmity had occurred.

Table 2

Statistical summary of peak height ratios of potential heteroplasmy identified by the T5000 system broken down into length and point heteroplasmy. This is further broken down into heteroplasms confirmed using Sanger sequencing results and those for which there was no evidence in the sequencing traces.

	Length—confirmed	Length—no evidence	Point—confirmed	Point—no evidence
Mean	0.60	0.17	0.67	0.22
Standard error	0.02	0.01	0.06	0.01
Standard deviation	0.16	0.02	0.24	0.07
Median	0.59	0.16	0.67	0.21
Minimum	0.29	0.14	0.18	0.00
Maximum	0.97	0.20	0.99	0.51
Count	45.00	5.00	18.00	98.00

Of the length heteroplasms indicated, 45/50 (90%) were confirmed in the Sanger dataset, while five were not. The mean minor:major peak height ratio for those heteroplasms that were confirmed was 0.6, while that for the unconfirmed heteroplasms was 0.17 (Table 2). This smaller ratio may indicate the presence of minor species that were not able to be detected by the Sanger method, since it has previously been reported that the detection threshold for minor species is approximately 20% [15]. No unconfirmed length heteroplasms had a minor:major peak ratio greater than 0.2.

Where amplicons contained indicated point heteroplasms, only 18/152 (11.8%) could be confirmed in the sequencing dataset (Table 3). In 35 of the remaining 134, insufficient sequence data was available to confirm the heteroplasmy. This number is high since the T5000 assay spans a slightly greater range than the Sanger sequencing method, thus generating amplicons that are not entirely covered in the sequencing range. A single amplicon indicated point heteroplasmy in a region where length heteroplasmy was indicated in the sequencing. The remaining 98 T5000-assigned point heteroplasms were not confirmed with the sequencing dataset. The mean minor:major peak height ratio for those heteroplasms that were confirmed was 0.67, compared to 0.22 for those that were not (Table 2). This is comparable with the results obtained for length heteroplasmy detection, and suggests that in most cases where a high minor:major peak height ratio is observed, the heteroplasmy is also detected by Sanger sequencing. Although over half of the samples in this category had minor:major peak height ratios of 0.2 or greater, only 8% had ratios over 0.3. Differences in individual sample aliquots (as mentioned previously, a greater sample volume was added per T5000 reaction than was added for sequencing) as well as variations in amplification efficiency and stochastic effects could all contribute to the discrepancy between T5000 and sequencing detection. It must also be considered that the T5000 system has the potential to indicate more than one incidence of heteroplasmy while appearing to generate only two products, i.e. two distinct instances of C/T heteroplasmy would not be differentiated unless they occurred within the same molecule. Two low-level heteroplasms may therefore appear to give a minor:major peak ratio greater than 0.2,

but individually neither would be detectable by sequencing. As discussed previously, low-level contamination within the bone samples or due to the carrier effect can also not be ruled out. The T5000 is reported to be sensitive to 4–10 pg total DNA (nuclear plus mitochondrial) per reaction (1.3–3.3 pg/amplicon) (Abbott, unpublished), depending on mitochondrial copy number. The system should therefore be easily capable of detection 50–100 copies of mitochondrial DNA and would be extremely sensitive to low-level contamination.

There is a correlation between individual amplicons and the probability of apparent heteroplasmy detection. Amplicons 2892, 2895, 2904 and 2905 each have over twenty incidences of heteroplasmy within the sample dataset and account for 100 of the 202 (49.5%) T5000 heteroplasmy assignments. 28% of these are confirmed by sequencing. Irwin et al. studied the incidence of heteroplasmy in 5015 individuals from various background populations [16,17] and identified 48 loci with at least two incidences of point heteroplasmy within their sample dataset. When compared with data presented in this paper, it is evident that the four identified amplicons coincide with many of these heteroplasmy 'hotspots'. Amplicon 2895 spans the region from 16157 to 16201 and includes seven of the loci identified as sources of point heteroplasmy (in order from most frequently occurring: 19189, 16183, 16192, 16168, 16169, 16173 and 16190) in addition to the $[C]_n$ stretch starting at position 16184 which is prone to length heteroplasmy. Amplicon 2892 spans the region from 16254 to 16301, which contains several of the loci where point heteroplasmy was reported (16294, 16256, 16261, 16290, 16291, 16301, 16278 and 16266). Amplicons 2904 and 2905 overlap, spanning bases 103–162 and 138–217 respectively. Amplicon 2904 covers five of the identified heteroplasmy hotspots including two of the most frequently identified (positions 152, 146, 150, 151 and 153). Amplicon 2905 covers fifteen heteroplasmic loci in total including three of the five SNPs most frequently identified in the population (positions 152, 146, 204, 195, 150, 215, 214, 151, 153, 189, 194, 199, 183, 198 and 207) [16,17]. Heteroplasmy detection by the T5000 therefore appears to some degree concordant with what is currently known about heteroplasmy localisation, though it must be noted that the frequency of

Table 3

Summary of T5000-assigned heteroplasms separated by heteroplasmy type, concordance between replicates and degree of supporting evidence. The left-hand category highlights amplicons for which heteroplasmy is confirmed in both T5000 replicates. This is further broken down into those for which supporting evidence was available in the corresponding sequencing trace, those for which there was no evidence of heteroplasmy and those for which insufficient corresponding sequence data was available to draw conclusions. The right-hand side shows the same figures for all indications of heteroplasmy identified by the T5000. This section contains an additional category to indicate samples for which point heteroplasmy could not be detected, but for which evidence of length heteroplasmy was apparent.

	Amplicons with heteroplasmy in both replicates			All heteroplasmy indications		
	Total	Length	Point	Total	Length	Point
Supporting evidence in sequencing	25	18	7	63	45	18
No evidence	4	1	3	103	5	98
Complete seq data not available	3	0	3	35	0	35
No evidence but evidence of length heteroplasmy	N/A	N/A	N/A	1	N/A	1
Total	32	19	13	202	50	152

heteroplasmy observed in this study is much greater than that observed by Irwin et al. [17], although their studies were carried out in saliva and blood samples rather than the bone and teeth samples employed here. There is evidence, however, that heteroplasmy occurs more frequently at levels not commonly detected by DNA sequencing [18]. Point heteroplasmy has previously been reported in approximately 10% of saliva samples using the T5000 system [7]. This figure remains substantially lower than that observed in this study.

It is clear from these figures that in almost all cases the base composition method is reproducible when used for the identification of homoplasmic amplified products. The base composition method can also be used to give an indication of heteroplasmy which, if the minor species is present at sufficient concentration, is mirrored by similar indications within a Sanger sequence trace. The veracity of the low intensity heteroplasmic products cannot be confirmed since they are difficult to identify at low level using Sanger sequencing and are often not reproducible between replicates. The base composition method therefore appears to have an increased ability to distinguish heteroplasmic loci but this ability may be prone to variation in poor quality samples. This variation could, as previously discussed, be due to system artefacts, sporadic contamination, copy number variation in the input DNA itself or due to stochastic effects caused by primer competition within the reaction mix.

3.5. Loss of information – insertion/deletion/SNP detection

It is important that as much information as possible is obtained from the method of choice. Sanger sequencing has an obvious advantage in this respect as it reports information for each individual base within the coverage region. The base composition approach is more limited since it contains information only on base composition within a specific amplicon. While the base composition method is able to indicate the presence of SNPs, insertions and deletions within an amplicon it is unable to pinpoint their positions exactly. Reciprocal SNPs within an amplicon (e.g. G > A and A > G within the same amplicon) are also undetectable using the T5000 system. To determine exactly how much information would be lost, all SNP combinations observed in the Sanger dataset that would not be detected using the base composition method were identified.

The results of this analysis, displayed in Table 4, indicate that the base composition method was able to detect over 94% of the nucleotide changes within the test dataset and nearly 93% of discriminatory changes, i.e. changes that do not result from expansion of a pre-existing poly-C tract (these regions are often polymorphic between relatives and are therefore frequently discounted during familial searching [19]). It is worth noting that while the loss of information is relatively small in terms of absolute changes nearly one in five samples is affected by some loss of discriminatory power. Several commonly observed SNP combinations are not observed by the base composition method, e.g. reciprocal changes of C194T, T195C or changes of T16186C,

C16189T within the HVI poly-C region. Undetected changes can also be a result of several combinations within a polymorphic amplicon, e.g. T146C, C150T, T152C or C16294T, C16295T, C16296T, T16298T, and T16304C. The base composition method may therefore identify a greater number of false positive matches than would be encountered using sequencing alone. The Sanger sequencing method consequently has an advantage over the base composition method in terms of information content.

3.6. Discrepancies between sequencing and the T5000

When adopting a new analysis method, it is also essential that it be as error-free as possible. The base composition data was therefore combined into a single dataset and the Sanger sequencing data compared directly to the base composition data. When heteroplasmy was excluded, discrepancies were identified between the base composition dataset and sequence data in only one of the 225 samples.

This discrepancy was shown to exist in only one of the two replicates of the sample in question. The T5000 peaks in the discrepant sample had a signal to noise ratio below 30, indicating that low intensity products displaying poor peak morphology could lead to erroneous mass assignments.

Care must be taken when selecting thresholds above which a peak detected should be considered genuine. For example, manually filtering peaks with a height below 100 units or a signal-to-noise ratio below 30 results in loss of information from 403 additional amplicons compared with the Abbott default analysis (data not shown), which includes neither a minimum product peak height nor a minimum signal-to-noise ratio (although both figures are calculated). It has already been shown that the majority of these amplicons represent genuine products, thus routine use of such raised thresholds would result in substantial data loss. Additional validation studies would therefore need to be carried out to ascertain more appropriate detection thresholds prior to implementation for use on old or degraded samples. Current manual filtering should easily be able to be automated through minor changes to the T5000 software.

Since the base composition method permits more accurate 'high throughput' analysis than can be carried out using Sanger sequencing, primarily as a result of a decrease in manual input and thus a lower expected incidence of human analysis error, the base composition method may be a better method for the initial screening of large numbers of samples and in that respect represents an advantage over Sanger sequencing. The T5000 method is particularly suited to approaches such as the one described in this study or in missing persons cases, where a sample of known source is screened against a database of individuals, the majority of whom would be expected to be excluded.

In fact it is when considering sample throughput that the biggest advantage of the base composition method becomes evident. T5000 plates are supplied with pre-aliquoted reagents and require only sample addition prior to amplification, and PCR clean-up is carried out automatically by the T5000 system. Typically the T5000 platform will analyse fifteen plates (150 samples) overnight with little manual input, thus the estimated hands-on time required for amplification, measurement and analysis is approximately 60–90 min, or six to nine minutes per sample. In contrast, the Sanger method requires several hours of manual time per sample, since the PCR mix must be made, PCR products cleaned up and run through a DNA analyzer, and sequence traces examined and compared with automated base calls. Commonly, the availability of thermal cyclers is the rate-limiting step in the T5000 process. The automated software generates very few erroneous calls, so while there may be some requirement for manual intervention in the case of low intensity signals, the T5000

Table 4

Data on number of SNPs detected using the Sanger sequencing method and a summary of how these changes would be interpreted by the T5000.

Total changes (in/del/sub) detected by sequencing	1547
Total excluded as due to poly-C expansion	357
Remaining informative changes	1190
Changes not detectable by T5000	88
% informative changes not detectable by T5000	7.39
% all changes not detectable by T5000	5.69
Samples for which not all nt change data would be detectable	43
% samples for which not all nt change data would be detectable	19.11

provides the capacity for more reliable automated screening, leading to substantially lower operational costs.

3.7. Matches to a database

A reference database was generated comprising mitochondrial DNA sequence from 466 maternal relatives of soldiers thought to have died at the battle of Fromelles. These 'reference' samples were not analysed on the T5000 and the reference sample database was therefore analogous to many pre-existing sequencing databases, e.g. EMPOP. To ascertain how both sequencing and T5000-generated datasets would fare when searched against a standard sequence database three searches were performed. The first of these was a direct comparison searching the sequence-derived SNP information from the 225 sets of remains against SNP information from the reference database. Approximately 55.6% of samples generated at least one match to the reference database using this method.

The second method was a search of the sequencing dataset against the reference database after conversion of both to base composition format. The percentage of samples with at least one match to the database increased to 63.1% using this method.

The third method used a search of the T5000-generated base composition dataset against the reference database also using base composition format. The percentage of samples matching the database was 63.6% using this method.

The results of searches against the database are given in Fig. 4. The figure highlights the number of matches to the reference database identified for each sample using the three methods described above. A match to the database means that, as far as it is possible to determine from the data available, the haplotype of the query and target samples are consistent with coming from the same matrilineage. Where heteroplasmy was indicated, a match to any of the indicated bases would result in a match to the database.

In total 28% of samples analysed by Sanger sequencing has less than the maximum coverage of the targeted HVI and HVII regions, compared with only 14.9% generated using the T5000. Despite this, direct sequence comparison (Method 1 above) proved to be the most specific method of database searching. Using the base composition method (Method 3 above), 97.6% of searches generated an equal or greater number of matches than the same searches carried out using direct sequence searching. This will be discussed in further detail below.

It is to be expected that searches of direct sequence data confer a greater specificity than those using a base composition format. When the full region of interest is covered by each method, sequence data has a higher information content. For example, the presence of reciprocal SNPs C150T and T152C is easily identified

within a direct sequence dataset. Amplicons containing both SNPs, however, are identical in base composition to those containing neither, so samples containing only these two SNPs appear to match the rCRS using base composition searching. A direct sequence search is therefore able to eliminate a greater number of false positives than a base composition search. This SNP information is also lost when data generated using a direct sequencing approach are converted to base composition format prior to the search (Method 2 above), and this conversion therefore also resulted in a higher number of false positive matches.

Just under one third of searches carried out in base composition format (whether using converted or T5000-generated data) resulted in a greater number of matches than those identified using a direct sequence search whereas two thirds resulted in an equal number of matches. When searching the target sequence database of 466 samples that had been converted to base composition format our query dataset contained sufficient information to exclusively ensure accurate matches (or absence of matches) in over 66% of cases. The remaining 30% of base composition searches generated additional false positive matches (e.g. matches in which reciprocal SNPs are not detected as described above). The number of false positives would be expected to decrease when searching a reference database that has been generated using the T5000 instrument since T5000 amplicon coverage extends beyond the defined HVI and HVII regions. There should also be improved accuracy when analysing high quality samples for which a full dataset can be generated.

The Sanger sequencing and base composition methods therefore both identify all matches between a query sample and the reference database, but the T5000 method will also generate a greater number of false positives and further downstream analysis may be necessary to confirm matches. It must be noted at this point that relatedness cannot be inferred solely on the basis that two samples share a haplotype. Instead a probability of relatedness is calculated based on the frequency with which the haplotype occurs in existing databases. In the case of the Fromelles study this probability was combined with that for available Y-STR data. Identification of remains was then made by representatives of the Commonwealth War Graves Commission based on genetic evidence and other available data.

4. Conclusions

The base composition analysis carried out on the T5000 system provides a highly sensitive alternative to high throughput mitochondrial DNA sequencing of old or degraded DNA. The use of short amplicons facilitates the amplification of degraded DNA resulting in usable profiles comparable in amplification efficiency to those generated using eight-amplicon Sanger sequencing. The sensitivity of the T5000 has resulted in the detection of additional heteroplasmic products not evident using Sanger sequencing. It has not been possible to confirm at this stage whether these are the result of genuine heteroplasmy within the sample set or of low-level contamination in the samples or laboratory itself. This potential for increased sensitivity highlights the already difficult issues encountered when working with old or degraded DNA, since the potential to extract additional information comes with an increased risk of reporting error. Care must be taken during the validation of any system to set thresholds that maximise the generation of data while minimising the capacity for error, but this is particularly important for systems such as this. Multiple replicates of all degraded samples would need to be used during validation to identify genuine products and permit accurate threshold settings.

The T5000 system has reduced discrimination power for searching of converted Sanger sequencing databases since reciprocal translocations within amplicons, e.g. C150T, T152C,

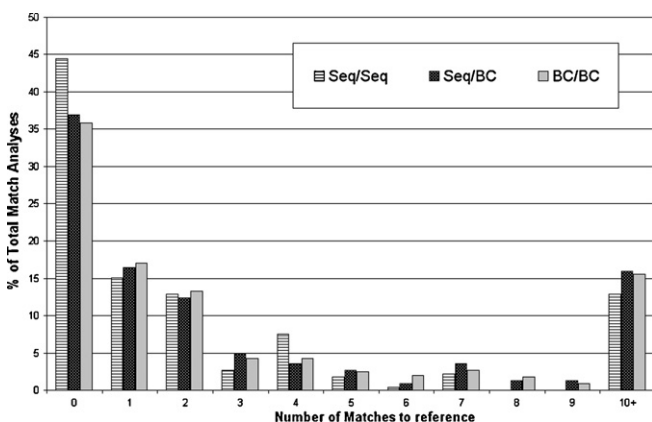


Fig. 4. Variation in the number of matches to a reference database between sequencing and base composition methods.

are not detected using the T5000 system. The overall coverage of the T5000 amplicons has a greater span than that of the sequencing amplicons used in this study which may provide additional discrimination when searching databases also generated using T5000 products. This has limited use in interrogating existing databases due to lack of comparable data within these ranges.

There is little evidence of error in the automated analysis of base composition products. It must be noted, however, that not all data was directly comparable with a sequencing profile. We recommend that caution be taken with all samples known to be of poor quality and that amplification and analysis should be repeated where necessary to ensure accuracy. Analysis methods can be modified to alter detection sensitivity if this is deemed necessary by the user.

The T5000 software is able to accurately identify base compositions within amplified mitochondrial DNA PCR products and this makes the method more desirable for use with high sample numbers since reduced analyst input results in fewer sequence interpretation and compilation errors. As the presented data demonstrates the reduction in discrimination power when interrogating a mitochondrial sequence database results in an increase in potential matches, some of which are proved to be false positives when their sequences are compared directly. We therefore conclude that the T5000 mitochondrial base composition analysis system has great potential as a rapid high-throughput screening tool. It is especially suited to projects such as this where large numbers of samples need to be screened against one another for the purposes of exclusion. Traditional Sanger sequencing can be utilised to provide more specific sequence information on matches identified through the T5000 system if required for the purpose of forensic casework.

Acknowledgements

We would like to acknowledge Abbott Molecular for providing the T5000 platform and supplying the Human Forensics (Mitochondria) kits. We would also like to thank Steffen Koch for installing the T5000, Tom Hall and Steve Hofstadler for their assistance with data analysis.

We would also like to thank the Foreign and Commonwealth War Graves Commission for giving us access to the DNA samples for analysis on the T5000.

References

- [1] B.D. Greenberg, J.E. Newbold, A. Sugino, Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA, *Gene* 21 (1–2) (1983) 33–49.

- [2] B. Budowle, M.R. Wilson, J.A. DiZinno, C. Stauffer, M.A. Fasano, M.M. Holland, K.L. Monson, Mitochondrial DNA regions HVI and HVII population data, *Forensic Sci. Int.* 103 (1) (1999) 23–35.
- [3] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, *Forensic Sci. Int. Genet.* 1 (2) (2007) 88–92.
- [4] D. Comas, F. Calafell, E. Mateu, A. Pérez-Lezaun, J. Bertranpetit, Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations, *Mol. Biol. Evol.* 13 (8) (1996) 1067–1077.
- [5] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, *Proc. Natl. Acad. Sci. U. S. A.* 74 (12) (1977) 5463–5467.
- [6] T.A. Hall, B. Budowle, Y. Jiang, L. Blyn, M. Eshoo, K.A. Sannes-Lowery, R. Sampath, J.J. Drader, J.C. Hannis, P. Harrell, V. Samant, N. White, D.J. Ecker, S.A. Hofstadler, Base composition analysis of human mitochondrial DNA using electrospray ionization mass spectrometry: a novel tool for the identification and differentiation of humans, *Anal. Biochem.* 344 (1) (2005) 53–69.
- [7] T.A. Hall, K.A. Sannes-Lowery, L.D. McCurdy, C. Fisher, T. Anderson, A. Henthorne, L. Gioeni, B. Budowle, S.A. Hofstadler, Base composition profiling of human mitochondrial DNA using polymerase chain reaction and direct automated electrospray ionization mass spectrometry, *Anal. Chem.* 81 (18) (2009) 7515–7526.
- [8] M.T. Krahmer, Y.A. Johnson, J.J. Walters, K.F. Fox, A. Fox, M. Nagpal, Electrospray quadrupole mass spectrometry analysis of model oligonucleotides and polymerase chain reaction products: determination of base substitutions, nucleotide additions/deletions, and chemical modifications, *Anal. Chem.* 71 (14) (1999) 2893–2900.
- [9] J.L. Beck, M.L. Colgrave, S.F. Ralph, M.M. Sheil, Electrospray ionisation mass spectrometry of oligonucleotide complexes with drugs, metals and proteins, *Mass Spec. Rev.* 20 (2001) 61–87.
- [10] D.J. Aaserud, Z. Guan, D.P. Little, F.W. McLafferty, DNA sequencing with blackbody infrared radiative dissociation of electrosprayed ions, *Int. J. Mass Spec. Ion Processes* 167–168 (1997) 705–712.
- [11] Y. Jiang, S.A. Hofstadler, A highly efficient and automated method of purifying and desalting PCR products for analysis by electrospray ionization mass spectrometry, *Anal. Biochem.* 316 (1) (2003) 50–57.
- [12] J.C. Hannis, S.M. Manalili, T.A. Hall, R. Ranken, N. White, R. Sampath, L.B. Blyn, D.J. Ecker, R.E. Mandrell, C.K. Fagerquist, A.H. Bates, W.G. Miller, S.A. Hofstadler, High-resolution genotyping of *Campylobacter* species by use of PCR and high-throughput mass spectrometry, *J. Clin. Microbiol.* 46 (4) (2008) 1220–1225.
- [13] S. Horai, K. Hayasaka, Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA, *Am. J. Hum. Genet.* 46 (4) (1990) 828–842.
- [14] A. Cooper, A. Removal of colourings, inhibitors of PCR, and the carrier effect of PCR contamination from ancient DNA samples, *Anc. DNA Newslett.* 1 (1992) 31–32.
- [15] G.G. Paneto, J.A. Martins, L.V.G. Longo, G.A. Pereira, A. Freschi, V.L.S. Alvarenga, B. Chen, R.N. Oliveira, M.H. Hirata, R.M.B. Cicarelli, Heteroplasmy in hair: differences among hair and blood from the same individuals are still a matter of debate, *Forensic Sci. Int.* 173 (2–3) (2007) 117–121.
- [16] J.A. Irwin, J.L. Saunier, H. Niederstatter, K. Strouss, K.A. Sturk, T.M. Diegoli, A. Brandstätter, W. Parson, T.J. Parsons, Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples, *J. Mol. Evol.* 68 (5) (2009) 16–27.
- [17] J.A. Irwin, W. Parson, M.D. Coble, R.S. Just, mtGenome reference population databases and the future of forensic mtDNA analysis, *Forensic Sci. Int. Genet.* DOI: 10.1016/j.fsigen.2010.02.008.
- [18] L.A. Tully, T.J. Parsons, R.J. Steighner, M.M. Holland, M.A. Marino, V.L. Prenger, A sensitive denaturing gradient-Gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region, *Am. J. Hum. Genet.* 67 (2) (2000) 432–443.
- [19] S. Lutz, H.J. Weisser, J. Heizmann, S. Pollak, Mitochondrial heteroplasmy among maternally related individuals, *Int. J. Legal Med.* 113 (3) (1999) 155–161.