Forensic Science International: Genetics 7 (2013) 543–549

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic Population Genetics—Original Research

Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM)☆

Walther Parson [a,b,*], Christina Strobl [a], Gabriela Huber [a], Bettina Zimmermann [a], Sibylle M. Gomes [c], Luis Souto [c], Liane Fendt [a,d], Rhena Delport [e], Reina Langit [f], Sharon Wootton [f], Robert Lagacé [f], Jodi Irwin [g]

[a] Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria
[b] Penn State Eberly College of Science, University Park, PA, USA
[c] Department of Biology, University of Aveiro, Campus de Santiago, Aveiro, Portugal
[d] Division of Human Genetics, Innsbruck Medical University, Innsbruck, Austria
[e] Department of Chemical Pathology, School of Medicine, University of Pretoria, South Africa
[f] Life Technologies, Foster City, CA, USA
[g] FBI Laboratory, Quantico, VA, USA

ARTICLE INFO

Article history:
Received 11 April 2013
Accepted 7 June 2013

Keywords:
Next Generation Sequencing
mtDNA genomes
Heteroplasmy
Sanger-type sequencing
PGM
Forensic science

ABSTRACT

Insights into the human mitochondrial phylogeny have been primarily achieved by sequencing full mitochondrial genomes (mtGenomes). In forensic genetics (partial) mtGenome information can be used to assign haplotypes to their phylogenetic backgrounds, which may, in turn, have characteristic geographic distributions that would offer useful information in a forensic case. In addition and perhaps even more relevant in the forensic context, haplogroup-specific patterns of mutations form the basis for quality control of mtDNA sequences. The current method for establishing (partial) mtDNA haplotypes is Sanger-type sequencing (STS), which is laborious, time-consuming, and expensive. With the emergence of Next Generation Sequencing (NGS) technologies, the body of available mtDNA data can potentially be extended much more quickly and cost-efficiently. Customized chemistries, laboratory workflows and data analysis packages could support the community and increase the utility of mtDNA analysis in forensics. We have evaluated the performance of mtGenome sequencing using the Personal Genome Machine (PGM) and compared the resulting haplotypes directly with conventional Sanger-type sequencing. A total of 64 mtGenomes (>1 million bases) were established that yielded high concordance with the corresponding STS haplotypes (<0.02% differences). About two-thirds of the differences were observed in or around homopolymeric sequence stretches. In addition, the sequence alignment algorithm employed to align NGS reads played a significant role in the analysis of the data and the resulting mtDNA haplotypes. Further development of alignment software would be desirable to facilitate the application of NGS in mtDNA forensic genetics.

1. Introduction

Mitochondrial (mt)DNA is present in higher copy number in the cell than nuclear (n)DNA and thus the likelihood of recovering useable DNA data is increased in forensic samples that fail to yield useful nDNA typing results. Due to its maternal mode of inheritance and lack of recombination [1] the discriminatory power of mtDNA is somewhat restricted, however and dependent on mtDNA databases [2]. This limitation is even more evident in routine forensic applications where, to date, analysis has been restricted to the non-coding control region (CR, or its hypervariable segments) for legal and technical reasons. While the use of coding region data outside of the control region would no doubt increase the information content of this genetic marker and increase its utility in practical forensic casework, conventional Sanger-type sequencing (STS) is neither amenable to, nor feasible for, the analysis of the full mtGenome in minute forensic samples.

Yet, outside of evidentiary testing, the forensic field is already making use of full mtGenomes. They are a prerequisite for the phylogenetic assignment of mtDNA haplotypes [3–5] and they form the basis for quality control of novel mtDNA data [6–8]. Most available mtGenomes have been generated with STS. However,

* Corresponding author at: Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria. Tel.: +43 512 9003 70640; fax: +43 512 9003 73640.
E-mail address: walther.parson@i-med.ac.at (W. Parson).

1872-4973/$ – see front matter © 2013 The Authors. Published by Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.06.003

this is a laborious, time-consuming, and expensive endeavor, particularly when high quality data with redundant sequence coverage are required [9].

Next Generation Sequencing (NGS) technologies have the potential to significantly increase both sample throughput and overall process efficiency, thereby facilitating the establishment of larger mtGenome databases in relatively short terms [10]. However, careful validation of these new technologies is required to maintain quality and consistency with the established datasets and technologies [11]. To date, only a few publications are available that describe the application of NGS technology to mtDNA testing in the forensic context. Mikkelsen et al. [12] conducted an early study on pyrosequencing with the FLX (454, Roche). Holland et al. [13] investigated the detection of point heteroplasmy (PHP) with the smaller Roche instrument (454 junior) and Loreille et al. [14] described mtGenome sequencing of highly degraded skeletal remains using the Illumina chemistry. In this study, the Personal Genome Machine (PGM [15]) was used to sequence complete mtGenomes, and the NGS results were evaluated by direct comparison with STS derived consensus haplotypes.

2. Materials and methods

2.1. Samples and DNA extraction

A total of 42 samples were voluntarily provided under informed consent. These included peripheral blood samples from five indigenous Khoe-San individuals from Angola [16] six paraffin-embedded tissue samples from Tyrol, Austria [17] buccal swab samples from eight individuals of the Democratic Republic of Timor-Leste in Dili (present study) and buccal swab samples from 23 individuals from Tyrol, Austria (present study). DNA extraction was performed either as previously described for the pre-existing samples, or using the Chelex protocol as detailed in [18].

2.2. PCR amplification

The entire mtDNA molecule was amplified with two overlapping 8.5 kilo base pair (kbp) fragments according to the protocol described in Ref. [9]. Both amplicons were purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and quality-controlled/quantified using non-denaturating polyacrylamide gel electrophoresis with silver staining [19] and the Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) following the manufacturer's recommendations.

2.3. Sanger-type sequencing

Earlier published mtGenomes from five Khoe-San individuals [16] and six individuals from Tyrol [17] were Sanger-type sequenced following the protocol outlined in the respective publications. For the remaining 31 novel mtGenomes generated in this study, some sequencing primers were replaced (Table S1). All sequences were imported into Sequencher 5.0 (Gene Codes Corporation, Ann Arbor, MI, USA) and aligned relative to the revised Cambridge Reference Sequence for human mitochondrial DNA (rCRS [20]) using the phylogenetic alignment rules detailed in Ref. [21]. STS data were analyzed by at least two independent scientists and the final consensus haplotypes were based on redundant sequence coverage over all positions (at least two independent sequence reads).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

2.4. Library construction for the PGM

The construction of the library involved the following three steps: enzymatic shearing, ligation of the adapters and size selection. The quantity of amplified DNA was determined with a Nanodrop spectrophotometer (Nanodrop Products, Wilmington, DE, USA). Both 8.5 kbp fragments were normalized to a quantity of 100 ng and then pooled. The amplicons were enzymatically sheared into suitable sized fragments using the Ion Xpress Plus Fragment Library Kit (Life Technologies (LT), Foster City, CA, USA) following the manufacturer's recommendations. For the 100 bp sequencing kit, incubation times were set to 25 min to yield fragments with sizes of approx. 130 bp. For the 200 bp sequencing kit shearing times were reduced to 7 min to yield fragments around 260 bp. Size and quality of fragmented DNA were determined with the Agilent DNA High Sensitivity Kit on the Bioanalyzer (Agilent) following the manufacturer's recommendations. Specific Ion Torrent compatible adapters were ligated onto the 5′ and 3′ ends of each fragment and linked by nick translation. For the barcoded libraries, the Ion P1 Adapter and the Ion Xpress barcode *X* adapter (LT, *X* = number of the used barcode) were applied to allow for sequencing multiple samples simultaneously. The fragmented and adapter ligated libraries were size selected using the E-Gel SizeSelect Agarose Gel (Invitrogen Corporation, Carlsbad, CA, USA) following the manufacturer's recommendations, and then batches of 4, 12, 15 and 31 samples were subsequently loaded onto 316 chips.

2.5. Template preparation

The quantity of the size selected library was determined by a real-time PCR approach using the Ion Library Quantitation Kit (LT) following the manufacturer's recommendations, with the template dilution factor calculated for a final concentration of ~26 pM per target. Targets were then subjected to emulsion PCR using the Ion One Touch (LT) following the manufacturer's recommendations. For clonal amplification, DNA was localized to Ion Sphere particles (LT), which were automatically enriched with the Ion OneTouch ES system (LT). Quality was assessed using the Qubit 2.0 Fluorometer (Invitrogen Corporation) following the manufacturer's recommendations.

2.6. PGM sequencing

Next Generation Sequencing (NGS) was performed using the Personal Genome Machine (PGM, LT). Before initializing the PGM Sequencer, a cleaning protocol was performed that started with a chlorite cleaning solution and was followed by a wash with 18 MΩ water (Elga Purelab Flex 3 Water Purification, Veolia Water Systems, Austria). After initialization, the chip was washed with 100% isopropanol and annealing buffer (from sequencing kit) and then tested for its functionality on the PGM. Sequencing primer and Control Ion Spheres of the Ion PGM sequencing kit were added to the library. After the annealing step sequencing polymerase was added and the sample was loaded onto the chip.

2.7. PGM data analysis

All PGM sequences were analyzed with the Ion Torrent Software Suite (Vs. 3.2) using the plug-in variant caller (Vs. 3.2.43647) that employed a TMAP Smith–Waterman alignment optimization [22]. The output of the variant caller was presented in tabular format, as a list of differences to the rCRS without a graphical display of the aligned reads. At this time, graphical displays of the TMAP alignment could only be visualized with separate tools for alignment and assembly viewing, such as the IGV

package (Integrative Genomics Viewer [23]) which accepts BAM (binary alignment map), BAI (binary alignment index files) and other file formats.

The FASTQ files of all 200 bp chemistry experiments (provided via the Ion Torrent server) were turned into "converted-FASTA"-files and aligned to the rCRS using NextGENe software (Vs. 2.3.1; SoftGenetics, State College, PA, USA), which employed a modified Burrows–Wheeler transform (BWT) alignment method. NextGENe analyzed data were presented in both a tabular summary of the haplotype, as well as a viewer, which permitted the visualization of the sequence reads and alignments. NextGENe analysis parameters used in this study are given in Table S2.

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

2.8. Comparing PGM to STS data

The PGM variant calls were listed with total coverage (CV) and variant frequency (VF) values. In this study variant calls exceeding 20% of the total coverage value were compared to STS consensus haplotypes with the following considerations: (i) Length heteroplasmy (LHP) in C tracts was represented by the dominant variant (major molecule) in STS data [24]. This threshold was not applied when investigating point heteroplasmy, which was present in lower mixture ratios. Differences to PGM results were counted as a single difference even when the dominant type differed by two insertions (this was the case in a total of six samples for all C tracts) and (ii) The Ion Torrent variant caller employed a 5′ indel alignment (i.e. indels in regions such as the dimeric repeat between 513 and 525 were recorded at positions 513 and 514), which adheres to the convention in medical genetics. In forensic genetics however, a 3′ alignment of indels is recommended [25]. Under these guidelines, indels in the dimeric repeat region would be recorded at positions 523 and 524. These conflicting practices were not regarded as differences when comparing haplotypes between technologies in this study.

3. Results

In total, 64 mtGenomes were generated from 42 DNA extracts with the PGM. Over the course of the study, a new sequencing chemistry ("200 bp chemistry") was released and used to produce a total of 33 mtGenomes. A total of 31 extracts was analyzed with the earlier version, the "100 bp chemistry", and 22 extracts (of the 33 and 31 sequenced with the separate chemistries) were sequenced with both. The 42 mtGenomes represented a total of 695,910 bp that were redundantly sequenced with STS and recorded relative to the rCRS (Table S3). A total of 1,060,437 bp (representing 64 mtGenomes) were generated with the PGM and the variant calls among these bases were directly compared with the STS consensus haplotypes. This comparison revealed 176 (0.017%) differences overall, of which 95 (0.018%) were found with the 100 bp chemistry (31 mtGenomes; 513,651 bp) and 81 (0.015%) were found with the 200 bp chemistry (33 mtGenomes; 546,786 bp; Tables 1 and S4). Fifty (28.4%) of the discrepancies could be identified as false positives (not present in STS data but reported with the variant caller) and 126 (71.6%) as false negatives (present in the STS calls but not reported by the variant caller). The relative quantity of false positives was higher in the 100 bp chemistry data (n = 38; 40.0%) when compared to the 200 bp chemistry output (n = 12; 14.8%, Table 1). The majority of the discrepancies was observed in or around the three hypervariable C tracts, with 55 (57.9%) and 53 (65.4%) occurrences in the 100 and 200 bp chemistry versions, respectively. False negatives were dominant here with 52 (91.2%) occurrences for the 100 bp chemistry and 53 (76.8%) for the 200 bp chemistry. In the following paragraphs the individual differences are described in more detail.

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

3.1. Homoplymeric C-tracts in HVS-1 (16183–16194), HVS-2 (302–310 and 310–316) and HVS-3 (567–574)

In the rCRS, T16189 is flanked by 5 and 4 Cs between positions 16183 and 16194, respectively. This sequence stretch was consistently and reproducibly represented in both the PGM variant caller and the STS analyses when no length variation was present. However, in three samples (WGS01, WGS02 and WGS04, all 100 bp chemistry), a combination of the T16189C transition and a deletion at position 16189 was observed with relatively low variant frequencies (VF) and coverage values (CV) for the deleted variant (VF 21% CV 270, VF 29% CV 49, and VF 22% CV 185, respectively). The transitional variant (16189C) in these samples was observed exclusively (100%) but with low total CVs of 126, 18, and 84, respectively. Samples WGS01 and WGS02 were also analyzed with the 200 bp chemistry, for which the STS haplotypes and PGM variant calls were concordant in this region. With this chemistry, the transitional variants were represented with VF of 100% CV 117 and VF 98% CV 104. The deleted variants were also present in the 200 bp reads, but they did not exceed the defined 20% threshold (VF 16% CV 198 and VF 16% CV 186).

In samples WGS28 (16193.1C), WGS34 (16193.1C) and WGS42 (16193.1C, 16193.2C), LHP was present in the STS data and thus the dominant types were determined for these analyses. For the NGS data, the dominant types apparent in the STS data were not reported by the variant caller. In addition, samples WGS11 and WGS34 showed variation with respect to the rCRS at the 5′ end of the HVS-1C tract (16184A, 16183C), but this variation was not present in the output of the Ion Torrent variant caller (Table S4).

Table 1
Summary of observed differences in sequence outputs of STS and PGM mtGenome typing.

Location	100 bp chemistry			200 bp chemistry			200 bp chemistry_NextGENe		
	# of differences	False positives	False negatives	# of differences	False positives	False negatives	# of differences	False positives	False negatives
HVS-1 (16183–16189)	2		2	1		1	12	6	6
HVS-1 (16190–16194)	5	3	2	3		3	2		2
HVS-2 (302–310)	16		16	14		14	1		1
HVS-2 (311–316)	31		31	33		33	1		1
HVS-3 (567–574)	1		1	2		2	1		1
AC-stretch	0			0			2		2
Indels	11	9	2	13	11	2	40		40
Substitutions	29	26	3	15	1	14	6	2	4
Pointheteroplasmy	0			0			1		1
Total	95			81			66		

The majority of differences was observed in the HVS-2C tract between positions 310 and 316, where all PGM variant calls reported 5 Cs and the STS data reflected 6 Cs (315.1C) (Tables 1, S3 and S4). Within the HVS-2 tract between positions 302 and 310, 16 of 31 (100 bp chemistry) and 14 of 33 (200 bp chemistry) haplotypes showed differences with respect to the STS data. These were observed when single or multiple insertions of Cs (≥8) as well as uninterrupted HVS-1 and HVS-3C tracts (e.g. samples WGS07 and WGS42) were present (Table S4). In haplotypes that harbored the rCRS variant in the HVS-3C tract (6 Cs between positions 567 and 574) no differences were observed between STS haplotypes and PGM variant calls.

3.2. Deletions outside the hypervariable segments

Comparing STS and PGM variant calls, a total of 11 differences with regard to deletions outside the hypervariable segments were observed for the 100 bp chemistry and 13 were found with the 200 bp chemistry (Tables 1, S4 and S5). A total of 13 (54.2%) were reported as a deletion in conjunction with a substitution, with only the substitution confirmed by STS. Therefore, the deleted variants represented false positives and were called with VF 26% CV 90 to VF 37% CV 1000. However, higher VFs for the deleted variant were also observed (e.g. WGS34 100 bp, 9545del VF 49% CV 636, Table S5).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

Two false negative deletion events were reported with the Ion Torrent variant caller. The tandem deletion at positions 290 and 291, which was present in sample WGS34, and the STS confirmed deletion at position 498 in samples WGS02 (200 bp chemistry) and WGS05 (100 bp chemistry, Tables S3 and S4) were not reported by the variant caller. The latter deletion (498) was however reported by the variant caller in samples WGS02 and WGS04 (100 bp chemistry).

3.3. Substitutions

A total of 29 substitutions (transitions and transversions) differed between the Ion Torrent variant calls and STS consensus haplotypes with the 100 bp chemistry, and 15 differed with the 200 bp chemistry. Three of these 44 differences were associated with deletions at the same site, similar to the parallel substitutions/deletions described previously (100 bp chemistry: WGS08 position 299, WGS34 position 494; 200 bp chemistry WGS02 position 8251; Table S4). The majority of differences in the 100 bp chemistry batch was caused by 7 samples (WGS02, WGS05, WGS06, WGS10, WGS11, WGS12, and WGS13) with 25 occurrences among them (86.2%). These were associated with low overall coverage values and also resulted in only weak STS signals (data not shown). Eleven substitutions (in 11 samples) were false negatives (Table 2) with only one of these (10664T) present in the 100 bp chemistry data and the remaining found in the 200 bp chemistry data. On the other hand, nearly all of the false positives were reported with the 100 bp chemistry data (Table 2). When considering the false negatives, three positions were hit in multiple samples: 456T was missing in all 4 haplogroup H5 samples (where this transition constitutes a signature mutation) and 10664T was lacking in 3 samples (Tables 2 and S4). Position 10,664 as well as the false positive 10651C/G that was also found in three samples, resided close to the region of the forward PCR primer of fragment B. Similarly, position 2689, which was also observed in three samples (Table 2), resided close to the binding site of the reverse PCR primer of fragment B. The latter showed differences with CV 750 VF 32%, CV 1205 VF 22%, and CV 369 VF 26%, respectively.

Table 2
Summary of false positive and false negative substitutions between sequence outputs of STS and PGM mtGenome typing.

Substitutions		
False negatives	100 bp chemistry	200 bp chemistry
10664T	WGS02	WGS02
10664T	WGS04	
10664T	WGS05	
13651G		WGS34
14374C		WGS01
16166C		WGS01
16172C		WGS01
295T		WGS18
456T		WGS15
456T		WGS17
456T		WGS23
456T		WGS27
493G		WGS34
5442C		WGS03
8251A		WGS02
961C		WGS03
		17

Substitutions				
False positives	100 bp chemistry		200 bp chemistry	
10651G	WGS02	CV 207 VF 23		
10651G	WGS05	CV 402 VF 26		
10651C	WGS05	CV 402 VF 20		
11147C	WGS12	CV 151 VF 33		
11604C	WGS10	CV 98 VF 40		
12797C	WGS12	CV 384 VF 36		
12959T	WGS13	CV 87 VF 55		
13159G	WGS12	CV 213 VF 23		
13507C	WGS11	CV 272 VF 20		
14912A	WGS06	CV 163 VF 35		
14955T	WGS11	CV 415 VF 29		
15618C	WGS13	CV 72 VF 25		
1902T	WGS06	CV 133 VF 75		
2664C	WGS05	CV 200 VF 33		
2689A	WGS02	CV 369 VF 26		
2689A	WGS05	CV 750 VF 32		
2689A	WGS40	CV 1205 VF 22		
2933A	WGS13	CV 220 VF 22		
299A	WGS08	CV 82 VF 20		
3226T	WGS06	CV 145 VF 24		
3229A	WGS06	CV 252 VF 38		
3573T	WGS10	CV 59 VF 29		
392C	WGS13	CV 83 VF 58		
494G	WGS34	CV 42 VF 60		
534T	WGS13	CV 29 VF 45		
86T	WGS10	CV 280 VF 60		
948C	WGS12	CV 350 VF 41		
962T			WGS07	CV 84 VF 21
				27

3.4. Point heteroplasmy

Point heteroplasmy (PHP) was detected at 14 positions in 12 samples with STS (Tables 3 and S3), some of which were very low but confirmed by multiple sequences (Figure S1). Twelve of those positions (present in 10 samples) were sequenced with the PGM using the 100 bp chemistry and 11 positions were sequenced (9 samples) using the 200 bp chemistry. Some of the heteroplasmic mixture ratios were below the defined 20% threshold, ranging from 10 to 19% (samples WGS14, WGS18 and WGS20, all 100 bp chemistry; samples WGS18 and WGS20, 200 bp chemistry). In all but three samples (WGS03, WGS05, WGS11; all 100 bp chemistry), PHP was confirmed in the variant caller output (Table 3). In order to gain further clarification on the unreported PHPs, the NGS data were re-analyzed in NextGENe. For two of the three false negatives, the NextGENe viewer showed the mixed bases (Figure S1). In

Table 3
Summary of point heteroplasmy reported with STS and PGM.

Sample	STS	PGM – variant caller	
		100 bp chemistry	200 bp chemistry
WGS01	6367Y	6367C Var. Freq. = 25, cv = 260	6367C Var. Freq. = 22, cv = 245
WGS03	966M	Not found	966C Var. Freq. = 66, cv = 161
WGS05	204Y	Not found	nd
WGS09	8473Y	8473C Var. Freq. = 41, cv = 79	nd
WGS11	16245Y	Not found	nd
WGS14	15623R	15623A Var. Freq. = 19, cv = 643	15623A Var. Freq. = 26, cv = 689
	16391R	16391A Var. Freq. = 48, cv = 519	16391A Var. Freq. = 39, cv = 593
WGS15	9966R	9966A Var. Freq. = 72, cv = 563	9966A Var. Freq. = 68, cv = 1248
WGS18	152Y	152C Var. Freq. = 72, cv = 438	152C Var. Freq. = 75, cv = 622
	1578R	1578R Var. Freq. = 16, cv = 575	1578R Var. Freq. = 16, cv = 897
WGS20	16201Y	16201Y Var. Freq. = 10, cv = 145	16201Y Var. Freq. = 18, cv = 272
WGS29	195Y	nd	195C Var. Freq. = 67, cv = 3743
WGS36	8252Y	8252T Var. Freq. = 31, cv = 216	8252T Var. Freq. = 44, cv = 371
WGS42	234R	nd	234G Var. Freq. = 47, cv = 2762

nd: not determined.

sample WGS03 PHP was present at position 966, which is located adjacent to a stretch of 10 Cs with extensive LHP due to the transition T961C. In sample WGS05 PHP at position 204 was very low in STS but visible in the NextGENe viewer (5 in 44), whereas low-level PHP at position 16245 in sample WGS11 was not observed in the NextGENe viewer (0 in 137; Figure S1).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

3.5. Comparison to other alignment algorithms

In order to gain further insight into the source of the STS–NGS discrepancies, the raw NGS reads were re-assessed both with alternative settings in the Ion Torrent variant caller (e.g. using a modified rCRS) and with completely separate alignment software based on a modified Burrow–Wheeler transform (BWT) alignment method (NextGENe). When a modified rCRS with 6 Cs instead of 5 between positions 310 and 316 was employed in the Ion Torrent variant caller (TMAP algorithm) to align the PGM sequences, the variant caller confirmed the results obtained by STS (tested for samples WGS08, WGS10, WGS21, WGS25, and WGS41; data not shown). This suggests that the omission of 315.1C in the previous variant reports may represent an artifact of the software, perhaps with the TMAP alignment algorithm itself, and not a technical or chemical issue with the PGM. The same confirmatory results in this C tract were found when using the modified BWT alignment method on sample WGS28 (Figure S2). According to STS, this sample also harbored a C insertion between 302 and 310 (which was reported with NextGENe), as well as the T16189C transition that resulted in an uninterrupted C stretch and extensive LHP in the HVS-1C tract. There, the dominant STS type reflected 11–12 Cs (16193.1C/16193.2C) with minor contributions of 3 additional length variants (9, 10, 13 Cs; Figure S2). Although the C insertion between 302 and 310 was reported in the tabular output of NextGENe, length variation at 16193.1C was not, even though the graphical view of the aligned reads indicated the presence of the 16193.1C insertion in some of the sequences (Figure S2). This phenomenon was also observed in other C tracts, e.g. sample WGS42. There, two C insertions between 567 and 574 were recorded as the dominant STS type, and these variants were also apparent in the NextGENe views of the aligned reads (Figure S3).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

Using the BWT alignment in NextGENe, both deletion events at positions 290/291 and 498 that had been missing from the Ion Torrent variant reporter were visible in the NextGENe alignment for samples WGS34 and WGS05, albeit at low coverage and sometimes together with the other substitutions at these positions that were described earlier (Tables S5 and S6). While the original variant caller report indicated false positive deletions in samples WGS34 (nps 492, 6297 and 9545) and WGS06 (np 8286) that were not reported with the BWT method, the combined call of a deletion and a substitution at a single position was observed regardless of the analysis method employed (Table S6).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

In terms of false positive substitutions, those originally reported with the variant caller for samples WGS05 (10651, 2664, and 2689), WGS34 (494) and WGS06 (3229, 1902, 3226, and 14912; all 100 bp chemistry) were not observed with NextGENe (Table S6). However, views of the NextGene alignments revealed that some positions showed mixtures of the same false positive substitution or deletion recorded by the variant caller. The false negative substitutions reported by the variant caller, such as the omitted C456T in four samples (WGS15, WGS17, WGS23, WGS27), were successfully captured with the BWT alignment method (Figure S4). The transition C456T prolongs a short stretch of four Ts between positions 452 and 455 by another T. Also, the false positive and false negative substitutions close to the PCR primer binding sites (e.g. positions 10651, 10664) reported by the variant caller were not observed with the NextGENe aligner (Figure S5). When considering the 200 bp chemistry, longer deletions were found in the variant caller output at positions 11798 (9 bp deletion, WGS42) and 2554 (>100 bp del, WGS02). These same long deletions were not reported in the BWT alignment method (data not shown).

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.06.003.

Overall, the application of the BWT alignment method as implemented in NextGENe led to a reduction in the number of discrepancies between the STS and PGM generated data in the HVS-2 and HVS-3C tracts and when considering substitutions across the genome (Table 1). On the flip side, the variant caller/TMAP alignment algorithm yielded fewer discrepancies with STS in the HVS-1C tract, the AC repeat region and when comparing deletions outside the C tracts (Table S4). It is worth noting that the option to view the NextGENe alignments proved extremely useful from the standpoint of evaluating discrepancies and identifying parameters worthy of further investigation.

Finally, all PGM runs using the 200 bp chemistry were analyzed with NextGENe under the defined settings and compared to the STS consensus haplotypes. This resulted in a total of 66 discrepancies to STS, interestingly at different positions than the variant caller (Table 1). NextGENe showed more discrepancies to STS in the HVS-1C tract (14) and in deletions outside the

homopolymeric regions ($n = 40$, 30 of which were observed as deletion at 9548) compared to the variant caller (4 and 13, respectively). NextGENe resulted in fewer differences to STS than the variant caller in the HVS-2C tract (2 versus 47) and in terms of substitutions (6 versus 15).

4. Discussion

In this study, the PGM was used to sequence 64 mtGenomes from 42 different samples that were also sequenced with conventional STS protocols. The use of barcoded adaptors allowed simultaneous sequencing of multiple samples with the PGM. In this early phase of NGS assessment, 32 different barcodes were available and these were used to generate up to 32 individual mtGenomes on a single 316 PGM chip. All things considered, but particularly given the advantage of sample barcoding, the PGM clearly outperformed conventional STS in terms of throughput and analysis time.

When assessing the performance and validity of a new technology, it is useful to compare the workflows, costs and, most importantly, the results to established methods. In this study, the standard and well-established technology for forensic mtDNA analysis – Sanger-type sequencing – was used as the reference for the PGM data. Although STS data are not a perfect reflection of the mtDNA composition in vivo (this pertains, for example, to the detection of mixtures such as length (LHP) and point heteroplasmy (PHP), where it has been described that mixtures may be displayed differently with STS depending on the adjacent nucleotide sequence and the primer used [26]), control data are nevertheless needed to characterize the NGS data and assess their utility for forensic applications. STS, as the standard, most-reliable and best-established technology in use for forensic mtDNA typing logically serve as this frame of reference.

The PGM pipeline is supported by an alignment and analysis software, the Ion Torrent variant caller that is available through the Ion Community website. This tool assembled the BAM (binary alignment map) and BAI (binary alignment index) files and aligned them to the rCRS using TMAP Smith–Waterman alignment algorithm [22]. Variant calls were reported with coverage and variant frequency values. After initial tests with varying parameter settings (data not shown) we defined the relatively high variant call threshold of 20% (of total coverage) for comparing the PGM output to STS consensus haplotypes. This particular threshold value yielded the lowest number of differences between STS and PGM variant calls and was required to compensate for the relatively high background noise that manifested in data produced by both STS and NGS technologies (due to low quality and quantity DNA extract), but was not adequately accommodated with the variant caller. When raw data and reads can be more directly evaluated in integrated viewers, and the distinction between signal and noise (as a result of either a chemistry artifact in the raw data or alignment artifacts at the analysis/interpretation stage) can be more easily assessed, it is likely that this threshold value can be reduced. Optimized chemistries, laboratory assays and data analyses that produce more robust raw and final data will allow for more sensitive detection and interpretation of PHP. For the current study, however, a relatively high threshold was required to differentiate between signal and noise.

When analyzing PGM generated FASTQ files with the alternative software NextGENe analysis settings were again adapted to yield minimal differences to STS data; but because of inherent differences in the software, the parameters and settings used to with the variant caller were not the same as those used with NextGENe. Thus, the comparison of the respective NGS data analyses (variant caller versus NextGENe) to STS consensus sequences were not performed with perfectly comparable settings. Instead, the comparisons were performed with those settings that optimized the results from each NGS software package. At this stage of development, this data evaluation strategy was helpful in identifying pros and cons of the individual software versions and aligners, and provided insight into the particular parameters likely to improve overall sequence output for mtDNA applications.

The number of differences between PGM variant calls and STS consensus haplotypes was surprisingly low. We found 176 incidences (0.017%) over all experiments, which, in total, involved more than one million individual base calls. About two thirds of those were observed in or around homopolymeric stretches, mostly C-tracts. For example the 315.1C insertion was present as a false negative in all PGM results. Interestingly, the same alignment algorithm applied to a modified reference sequence (bearing the 315.1C) gave fully concordant results, suggesting that the TMAP algorithm employed by the variant caller in conjunction with a reference sequence bearing different numbers of C residues was the reason for the discrepant results. The modified BWT method yielded fewer discrepancies to STS in homopolymeric stretches. Insertions and deletions were usually reported consistently with STS calls for stretches up to 9 Cs and in some cases up to 11 Cs. With the exception of 315.1C, which is a stable difference to the rCRS, C insertions and deletions within the homopolymeric stretches have only minor relevance when interpreting forensic evidence due to their high variability even among tissues of the same individual [27] and thus discrepancies in these regions have little or no impact on downstream data interpretation. There are stable deletions outside the described C tracts with a strong phylogenetic signature. These include the paired deletion at positions 290 and 291 (haplogroup C1) and the deletion at position 498 (haplogroups K1c and L0d1′2), both of which were present in our dataset. These deletions are located in short homopolymeric stretches, which seems to be the reason why the variant caller failed to identify them in some of the samples. They were, however, captured by the BWT alignment method. A similar situation was observed for substitutions that prolonged short homopolymeric T stretches such as the haplogroup H5′36 specific C456T transition. This variant was called concordantly with BWT alignment software only. Conclusively, the majority of differences observed between PGM variant calls and STS consensus sequences seem to be attributable to the TMAP alignment algorithm. Though read pre-processing and filtering cannot be ruled out as contributing factors, those steps are invisible to a user and thus cannot be assessed directly. Regardless of the discrepancy source, modified reference sequences and alternative alignment algorithms were effective in decreasing the number of observed differences between the technologies.

An observation new to STS users was the parallel occurrence of deletions and substitutions at single positions. These were observed with all applied alignment algorithms and even with relatively high coverage values. If they were authentic, they would represent heteroplasmic mixtures of deleted and substituted variants that have so far only rarely been observed in non-human mtDNA outside homopolymeric C tracts [28]. The fact that the STS data did not reveal these variants, and the fact that they have not been previously reported with other technologies, suggests that they are likely artifacts of the PGM sequencing chemistry and/or the alignments.

Over the course of this study, two versions of the sequencing chemistry were compared. They generally performed very similar. We observed slightly fewer discrepancies to the STS derived haplotypes with the newer 200 bp chemistry that has meanwhile replaced the earlier 100 bp chemistry version.

5. Conclusions

In this study, PGM generated mtDNA sequence data were evaluated by direct comparison with Sanger-type sequence data. In order to evaluate the data as strictly as possible, differences across

the entire mtGenome were considered, including those in homopolymeric C-tracts, which have little or no relevance for excluding samples in a forensic setting. The goal of this study was to evaluate the current state-of-the-art of Next Generation Sequencing (NGS) with the PGM and highlight issues that may need to be addressed for application of this technology in forensic genetics.

We observed an overall generally high level of consistency between Sanger-type and PGM derived data, with the total number of recorded differences below 0.02%. The majority of discrepancies was related to the alignment algorithm and was either found in homopolymeric C tracts (variant caller) or observed with high reproducibility at single positions (NextGENe). It is encouraging that the algorithms differed at discrepant positions as this suggests that further development of the individual software packages has room for improvement.

As forensic scientists familiar with Sanger-type electropherograms and data analyses, we found it very useful to view a graphical representation of the aligned NGS reads for investigation of NGS–STS discrepancies and increased understanding of other NGS data features. The transition from STS to NGS data interpretation will no doubt require revised analytical guidelines due to the fundamental differences in data production, detection, output and volume. Software that permits visualization and scrutiny of data and alignments will facilitate both the development of data interpretation guidelines and, ultimately, the adoption of NGS in forensic genetics.

In conclusion, our data suggest that aligners, alignment parameters and pre-alignment data filtering tools all have a great impact on final NGS haplotypes. During the course of our experiments we observed software updates that generally improved the interpretation of the NGS data. It is likely that this ongoing software development process will be critical to more accurate and streamlined NGS data analysis and interpretation, ultimately maximizing the reliability of NGS-produced mtGenome haplotypes.

Disclaimer

This is publication 13-13 of the Laboratory Division of the Federal Bureau of Investigation (FBI). Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the author's and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

Acknowledgements

This work leading to these results has received funding from the Austrian Science Fund (FWF) [P22880-B12] and [TRL397] and was financially supported from the European Union Seventh Framework Program (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE). This work was further financially supported by the National Institute of Justice (NIJ) grant 2011-MU-MU-K402 and by the Foundation of Science and Technology Portugal (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE), co-funded by the European Community Fund FEDER with the Project PTDC/CS-ANT/108558/2008 and also by the FCT fellowship SFRH/BD/63165/2009''. We would like to thank Alexander W. Röck for technical help and discussion.

References

[1] W.M. Brown, Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 3605–3609.
[2] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.
[3] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) E386–E394.
[4] H.-J. Bandelt, M. van Oven, A. Salas, Haplogrouping mitochondrial DNA sequences in legal medicine/forensic genetics, Int. J. Legal Med. 126 (2012) 901–916.
[5] A.W. Röck, A. Dür, M. van Oven, W. Parson, Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA), Forensic Sci. Int. Genet. (2013) (submitted for publication).
[6] Y.G. Yao, A. Salas, I. Logan, H.-J. Bandelt, mtDNA data mining in GenBank needs surveying, Am. J. Hum. Genet. 85 (2009) 929–933.
[7] H.-J. Bandelt, Y.G. Yao, C.M. Bravi, A. Salas, T. Kivisild, Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies, J. Hum. Genet. 54 (2009) 174–181.
[8] B. Zimmermann, A.W. Röck, G. Huber, T. Krämer, P.M. Schneider, W. Parson, Application of a west Eurasian-specific filter for quasi-median network analysis: sharpening the blade for mtDNA error detection, Forensic Sci. Int. Genet. 2 (2011) 133–137.
[9] L. Fendt, B. Zimmermann, M. Daniaux, W. Parson, Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences, BMC Genomics 10 (2009) 139.
[10] 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. 5 (2011) 222–225.
[11] H.-J. Bandelt, A. Salas, Current next generation sequencing technology may not meet forensic standards, Forensic Sci. Int. Genet. 6 (2012) 143–145.
[12] M. Mikkelsen, E. Rockenbauer, A. Wächter, L. Fendt, B. Zimmermann, W. Parson, S. Abel Nielsen, T. Gilbert, E. Willerslev, N. Morling, Application of full mitochondrial genome sequencing using 454 GS FLX pyrosequencing, Forensic Sci. Int.: Genet. Suppl. Series 2 (2009) 518–519.
[13] M.M. Holland, M.R. McQuillan, K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy, Croat. Med. J. 52 (2011) 299–313.
[14] O. Loreille, H. Koshinsky, V.Y. Fofanov, J.A. Irwin, Application of next generation sequencing technologies to the identification of highly degraded unknown soldiers' remains, Forensic Sci. Int.: Genet. Suppl. Series 3 (2011) e540–e541.
[15] J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, J. Hoon, J.F. Simons, D. Marran, J.W. Myers, J.F. Davidson, A. Branting, J.R. Nobile, B.P. Puc, D. Light, T.A. Clark, M. Huber, J.T. Branciforte, I.B. Stoner, S.E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J.A. Fidanza, E. Namsaraev, K.J. McKernan, A. Williams, G.T. Roth, J. Bustillo, An integrated semiconductor device enabling non-optical genome sequencing, Nature 475 (2011) 348–352.
[16] L. Fendt, G. Huber, A.W. Röck, B. Zimmermann, M. Bodner, R. Delport, K. Schmidt, W. Parson, Mitochondrial DNA control region data from indigenous Angolan Khoe-San lineages, Forensic Sci. Int. Genet. 6 (2012) 662–663.
[17] L. Fendt, H. Niederstätter, G. Huber, B. Zelger, M. Dünser, C. Seifarth, A. Röck, G. Schäfer, H. Klocker, W. Parson, Accumulation of mutations over the entire mitochondrial genome of breast cancer cells obtained by tissue microdissection, Breast Cancer Res. Treat. 128 (2011) 327–336.
[18] P.S. Walsh, D.A. Metzger, R. Higuchi, Chelex-100 as a medium for simple extraction of DNA for PCR-based typing from forensic material, Biotechniques 10 (1991) 506–513.
[19] W. Parson, K. Pegoraro, H. Niederstätter, M. Föger, M. Steinlechner, Species identification by means of the cytochrome b gene, Int. J. Legal Med. 114 (2000) 23–28.
[20] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[21] H.-J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, Int. J. Legal Med. 122 (2008) 11–21.
[22] H. Li, N. Homer, A survey of sequence alignment algorithms for next-generation sequencing, Brief. Bioinform. 11 (2010) 473–483.
[23] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.
[24] C. Berger, P. Hatzer-Grubwieser, C. Hohoff, W. Parson, Evaluating sequence-derived mtDNA length heteroplasmy by amplicon size analysis, Forensic Sci. Int. Genet. 5 (2011) 142–145.
[25] Á. Carracedo, W. Bär, P. Lincoln, W. Mayr, N. Morling, B. Olaisen, P. Schneider, B. Budowle, B. Brinkmann, P. Gill, M. Holland, G. Tully, M. Wilson, DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing, Forensic Sci. Int. 110 (2000) 79–85.
[26] M. Mayr-Eduardoff, G. Huber, B. Bayer, D. Schmid, K. Anslinger, T. Göbel, B. Zimmermann, P.M. Schneider, A.W. Röck, W. Parson, Mass spectrometric base composition profiling: implications for forensic mtDNA databasing, Forensic Sci. Int. Genet. (2013) (submitted for publication).
[27] G. Tully, W. Bär, B. Brinkmann, Á. Carracedo, P. Gill, N. Morling, W. Parson, P.M. Schneider, Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles, Forensic Sci. Int. 124 (2001) 83–91.
[28] C. Eichmann, W. Parson, Molecular characterization of the canine mitochondrial DNA control region for forensic applications, Int. J. Legal Med. 121 (2007) 411–416.