Int J Legal Med (2016) 130:1169–1180
DOI 10.1007/s00414-016-1355-y

ORIGINAL ARTICLE

Effects of the Ion PGM™ Hi-Q™ sequencing chemistry on sequence data quality

Jennifer D. Churchill[1] · **Jonathan L. King**[1] · **Ranajit Chakraborty**[1] · **Bruce Budowle**[1,2]

Received: 22 January 2016 / Accepted: 4 March 2016 / Published online: 30 March 2016

Abstract Massively parallel sequencing (MPS) offers substantial improvements over current forensic DNA typing methodologies such as increased resolution, scalability, and throughput. The Ion PGM™ is a promising MPS platform for analysis of forensic biological evidence. The system employs a sequencing-by-synthesis chemistry on a semiconductor chip that measures a pH change due to the release of hydrogen ions as nucleotides are incorporated into the growing DNA strands. However, implementation of MPS into forensic laboratories requires a robust chemistry. Ion Torrent's Hi-Q™ Sequencing Chemistry was evaluated to determine if it could improve on the quality of the generated sequence data in association with selected genetic marker targets. The whole mitochondrial genome and the HID-Ion STR 10-plex panel were sequenced on the Ion PGM™ system with the Ion PGM™ Sequencing 400 Kit and the Ion PGM™ Hi-Q™ Sequencing Kit. Concordance, coverage, strand balance, noise, and deletion ratios were assessed in evaluating the performance of the Ion PGM™ Hi-Q™ Sequencing Kit. The results indicate that reliable, accurate data are generated and that sequencing through homopolymeric regions can be improved with the use of Ion Torrent's Hi-Q™ Sequencing Chemistry. Overall, the quality of the generated sequencing data supports the potential for use of the Ion PGM™ in forensic genetic laboratories.

Keywords Massively parallel sequencing (MPS) · Ion PGM™ · Mitochondrial DNA · STRs · Hi-Q™

Electronic supplementary material The online version of this article (doi:10.1007/s00414-016-1355-y) contains supplementary material, which is available to authorized users.

✉ Jennifer D. Churchill
Jennifer.Churchill@unthsc.edu

[1] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., CBH-250, Fort Worth, TX 76107, USA

[2] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

Introduction

Massively parallel sequencing (MPS) technologies have revolutionized genetic analyses by enabling the production of an unprecedented amount of data. These technologies offer the potential to analyze a larger number of markers and different combinations of marker types (e.g., short tandem repeats (STRs), SNPs, insertions/deletions) than was possible previously to analyze with currently used standard capillary electrophoresis (CE) methods [1–3]. MPS also enables detection of length-based and sequence-based genotypes for STRs, thereby providing greater resolution of alleles than previously possible with CE-based systems [2–8].

The advantages of MPS compared to standard CE typing are substantial, and there remains little doubt that MPS technologies will be implemented into forensic genetic laboratories in the not too distant future [2, 3, 8–14]. The Ion Torrent Personal Genome Machine® (Thermo Fisher Scientific, Waltham, MA USA) (Ion PGM™) is a MPS platform that employs a sequencing-by-synthesis chemistry where incorporation of a nucleotide into the growing nascent strand releases a hydrogen ion that is detected by the resulting change in pH in wells in a semiconductor chip [15–17]. The scalability, read length, sequencing time, and cost per analysis make the Ion PGM™ a desirable instrument for forensic genetic analyses [2, 8, 9]. Recent studies have demonstrated the Ion PGM's

potential applicability to genetic analysis in forensic laboratories [7, 8, 12, 13, 18–21].

Since their introduction almost a decade ago, MPS technologies continue to improve. There is every expectation that technical improvements in MPS technologies will continue. In fact, Ion Torrent's Ion PGM™ Hi-Q™ Sequencing Chemistry (Thermo Fisher Scientific) is one potential approach to increase the quality of sequence data generated by the Ion PGM™. In the study herein, the Hi-Q™ Sequencing Chemistry was evaluated to determine what effects the Hi-Q™ Sequencing Chemistry had on the quality of the generated sequencing data in association with selected genetic marker targets. The whole mitochondrial genome and the HID-Ion STR 10-plex (includes amelogenin) panel (Thermo Fisher Scientific) were sequenced on the Ion PGM™ system with the Ion PGM™ Sequencing 400 Kit (Thermo Fisher Scientific) and the Ion PGM™ Hi-Q™ Sequencing Kit (Thermo Fisher Scientific). The results support that sequencing through homopolymeric regions can be improved with the use of the Hi-Q™ Sequencing Chemistry.

Materials and methods

Samples

DNA samples from 31 previously described [11, 18], unrelated African-Americans ($n=24$), Hispanics ($n=3$), and Caucasians ($n=4$) and a negative control were used for this study. The policies and procedures approved by the Institutional Review Board for the University of North Texas Health Science Center in Fort Worth, TX, were followed for the collection and use of these samples. DNA was extracted using the QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocols [22]. The quantity of recovered DNA was determined using the Qubit® dsDNA BR Assay Kit (Thermo Fisher Scientific) and a Qubit® 2.0 Fluorometer (Thermo Fisher Scientific). Samples were normalized to one ng/μl.

Mitochondrial genome

The mitochondrial genomes of these 31 individuals and a negative control were sequenced with two different protocols. One was performed using the Ion PGM™ Sequencing 400 Kit following the manufacturer's protocols [23], and the second was performed using the Ion PGM™ Hi-Q™ Sequencing Kit following manufacturer's protocols [24]. The entire mitochondrial genome was amplified using previously described long PCR primers [25] that generated amplicons approximately eight kb in length. Library preparation, emulsion PCR, enrichment of template beads, and sequencing on the Ion PGM™ were completed as described by Churchill et al. [8]. Sequence data were analyzed using the Torrent Suite software v4.6 with the Alignment (v4.0-r77189), Coverage Analysis (v4.4.2.2), and Variant Caller plugin (v4.6.0.7). Data were aligned to the revised Cambridge Reference Sequence (rCRS) [26], and Integrative Genomic Viewer (IGV) was used for visualization of the aligned binary alignment map (BAM) files [27, 28]. The variant call format (vcf) output files generated by the Variant Caller plugin were used in conjunction with mitoSAVE [29] to generate haplotype calls in standard forensic conventions. A minimum coverage threshold of 10X and point heteroplasmy threshold of 0.20 was set for mitochondrial DNA variant calls.

Mitochondrial genome concordance data—MiSeq

The mitochondrial genomes of the samples included in this study were sequenced previously on the MiSeq desktop sequencer (Illumina, San Diego, CA, USA) as described in King et al. [11]. These data were used to provide concordance information between two different MPS platforms.

STRs

The HID-Ion STR 10-plex panel allows for amplification of amelogenin and nine STRs (CSF1PO, D16S539, D3S1358, D5S818, D7S820, D8S1179, TH01, TPOX, and vWA), with amplicon sizes that range from 75 to 170 base pairs (bp) [7]. Library preparation, emulsion PCR, enrichment of template beads, and sequencing on the Ion PGM™ were completed as described in Churchill et al. [8]. The DNA samples and negative control were sequenced using the Ion PGM™ Sequencing 400 Kit with manufacturer's recommended protocols [23] and the Ion PGM™ Hi-Q™ Sequencing Kit with manufacturer's recommended protocols [24]. Sequence data were analyzed using the Torrent Suite software v4.6 with the Alignment (v4.0-r77189), Coverage Analysis (v4.4.2.2), and HID_STR_Genotyper (v3.1) plugins. Data were aligned to the Hg19 reference genome. The HID_STR_Genotyper plugin makes genotyping calls on Ion PGM™ data using BAM files generated with the Torrent Suite software and BED files that specify the targeted areas of interest within the Hg19 reference genome. Additionally, FASTQ files generated with the Torrent Suite software were analyzed with the STR Allele Identification Tool—Razor (STRait Razor) [30, 31]. A minimum coverage threshold of 10X was set for genotype calls.

CE concordance data

Conventional STR typing by CE was performed to provide concordance data using the GlobalFiler™ PCR Amplification Kit (Thermo Fisher Scientific) with 1 ng of DNA for each reaction following the recommended manufacturer's protocols [32]. The GeneAmp® PCR System 9700 thermal cycler (Thermo Fisher Scientific) was used for PCR amplification,

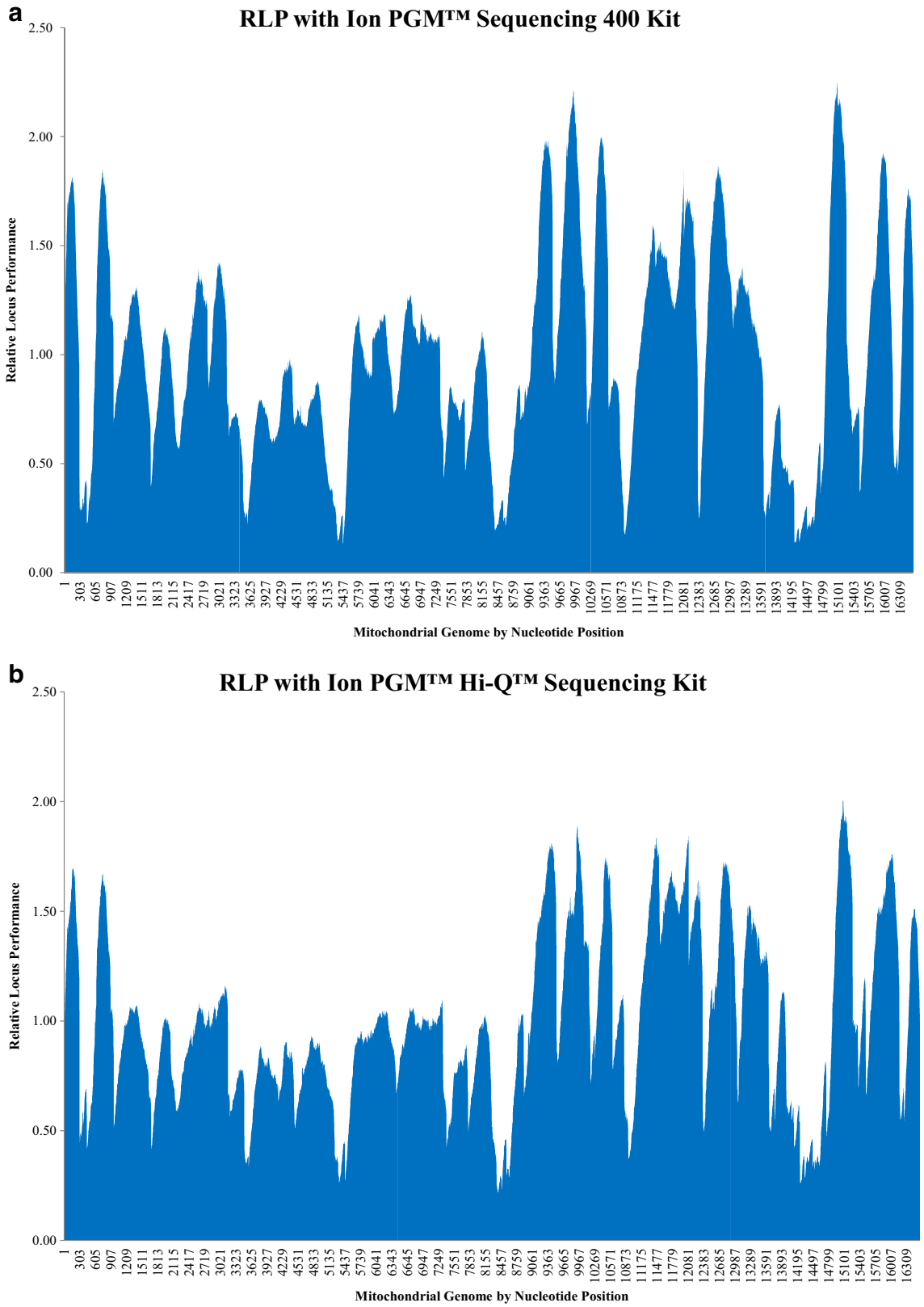


Fig. 1 Average RLP across the mitochondrial genome ($N = 31$) for the Ion PGM™ Sequencing 400 Kit **(a)** and for the Ion PGM™ Hi-Q™ Sequencing Kit **(b)**

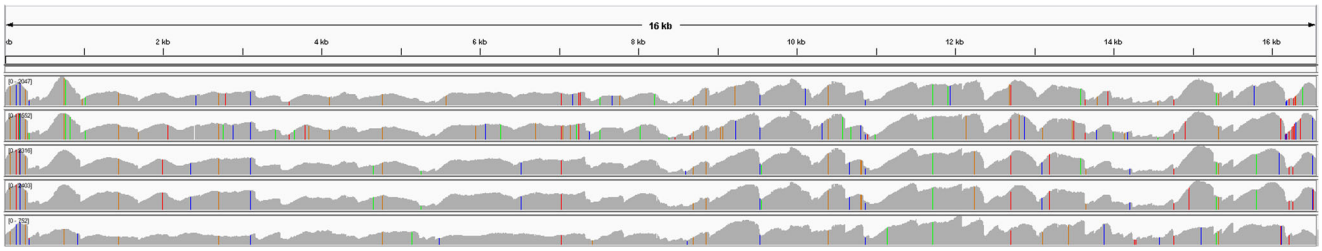


Fig. 2 View of coverage plots with IGV illustrating areas of consistently high and low coverage across samples

and electrophoresis was completed on an ABI Prism® 3500xL Genetic Analyzer (Thermo Fisher Scientific). Raw data were analyzed with GeneMapper® ID-X software v1.2 (Thermo Fisher Scientific). A minimum peak height of 50 relative fluorescence units (RFUs) was set for data interpretation.

Statistical analysis

Coverage (or read depth) was used to calculate normalized relative locus performance (RLP) at each nucleotide position of the mitochondrial genome (i.e., coverage of one nucleotide position divided by the total coverage across the entire mitochondrial genome for that sample all multiplied by the length of the rCRS (i.e., 16,569)). Strand balance for the mitochondrial genome data was calculated by dividing the coverage of one strand by the total coverage of that nucleotide position (e.g., 275X/500X = 0.55; 0.5 indicating equal coverage). Noise for the mitochondrial genome data was calculated by dividing the number of reads not attributed to nominal allele calls at a nucleotide position by the total coverage at that nucleotide position. False deletions were measured as a ratio of the number of reads indicating a deletion divided by the total number of reads at that position. The deletion ratios from the two data sets were compared by calculating a delta (i.e., deletion ratio from data generated with the Sequencing 400 Kit minus the deletion ratio from the data generated with the Hi-Q™ Sequencing Kit = delta). A positive delta indicates that less false deletions were observed with the Hi-Q™ Sequencing Kit, and a negative delta indicates that less false deletions were observed with the Sequencing 400 Kit. Statistical significance from delta = 0 was assessed with a paired *t* test [33] for each nucleotide position analyzed using one-sided *p* values. These calculations allowed classification of each nucleotide position as non-significant (NS), +ve (when delta >0 and is significant with $p < 0.05$), and –ve (when delta <0 and is significant with $p < 0.05$). To examine whether or not the significance of the delta values is dependent on the nucleotide type (A, T, G, or C) and the length of homopolymers, the trinomial (NS, +ve, −ve) distribution was tabulated grouping the site by nucleotide type and length of homopolymers. A chi-square test of heterogeneity of contingency table analysis was performed, with *p* values determined by shuffling to test if the differences of three classifications of the delta values depended upon nucleotide types and/or length of homopolymers.

Coverage for the HID-Ion STR 10-plex panel also was used to calculate RLP at each locus in the panel (i.e., coverage of one marker divided by total coverage for that sample). Allele coverage ratios (ACRs; i.e., heterozygote balance) were calculated for each STR locus by dividing the lower coverage allele by the higher coverage allele at that locus (e.g., 400X/500X = 0.8; 1.0 indicating equal coverage). Sequence coverage ratios (SCRs) analyze noise levels for STRs by dividing the number of reads used to make nominal repeat length allele calls and the number of reads attributed to stutter by the total number of reads at that locus (e.g., 360X/400X = 0.9 indicating that 10 % of reads are attributable to noise).

Results and discussion

Mitochondrial genome data

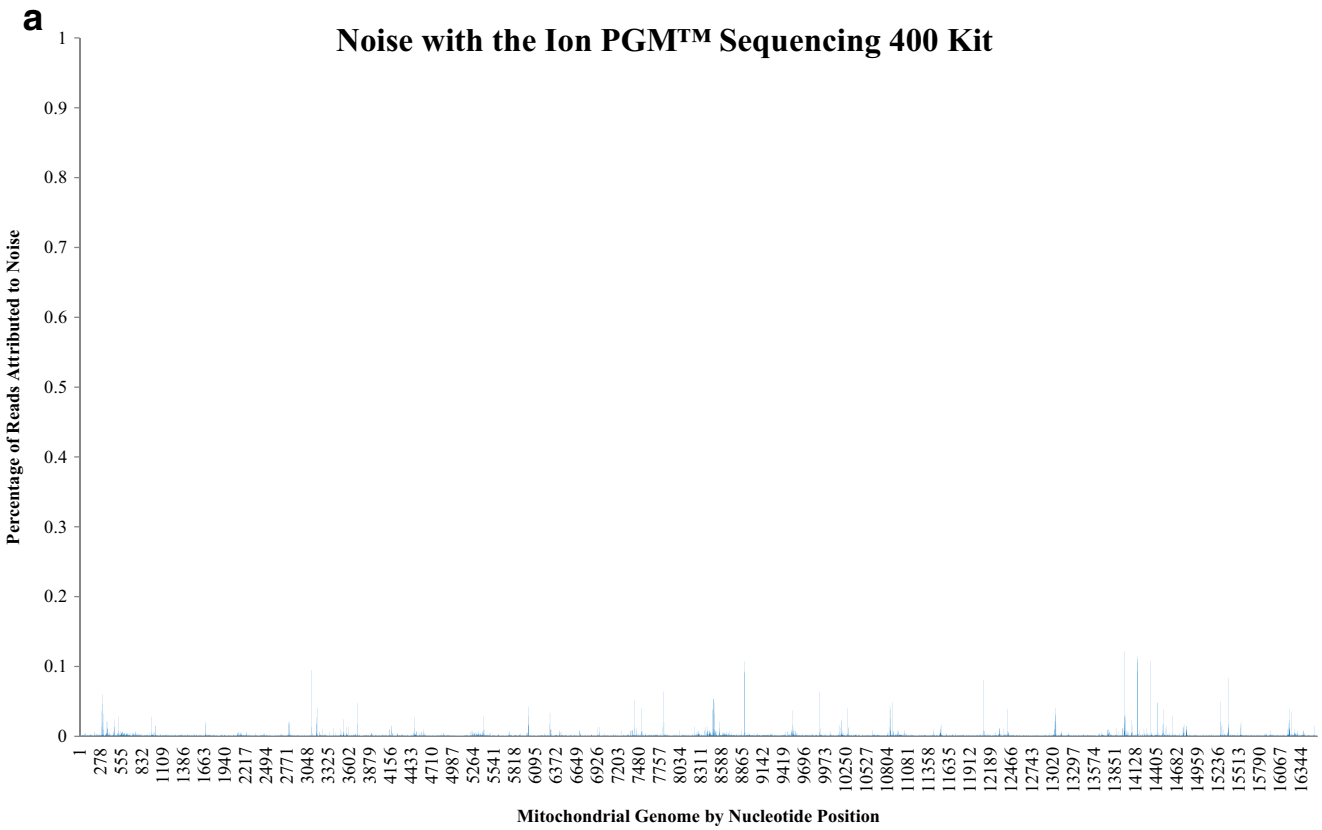
Run information

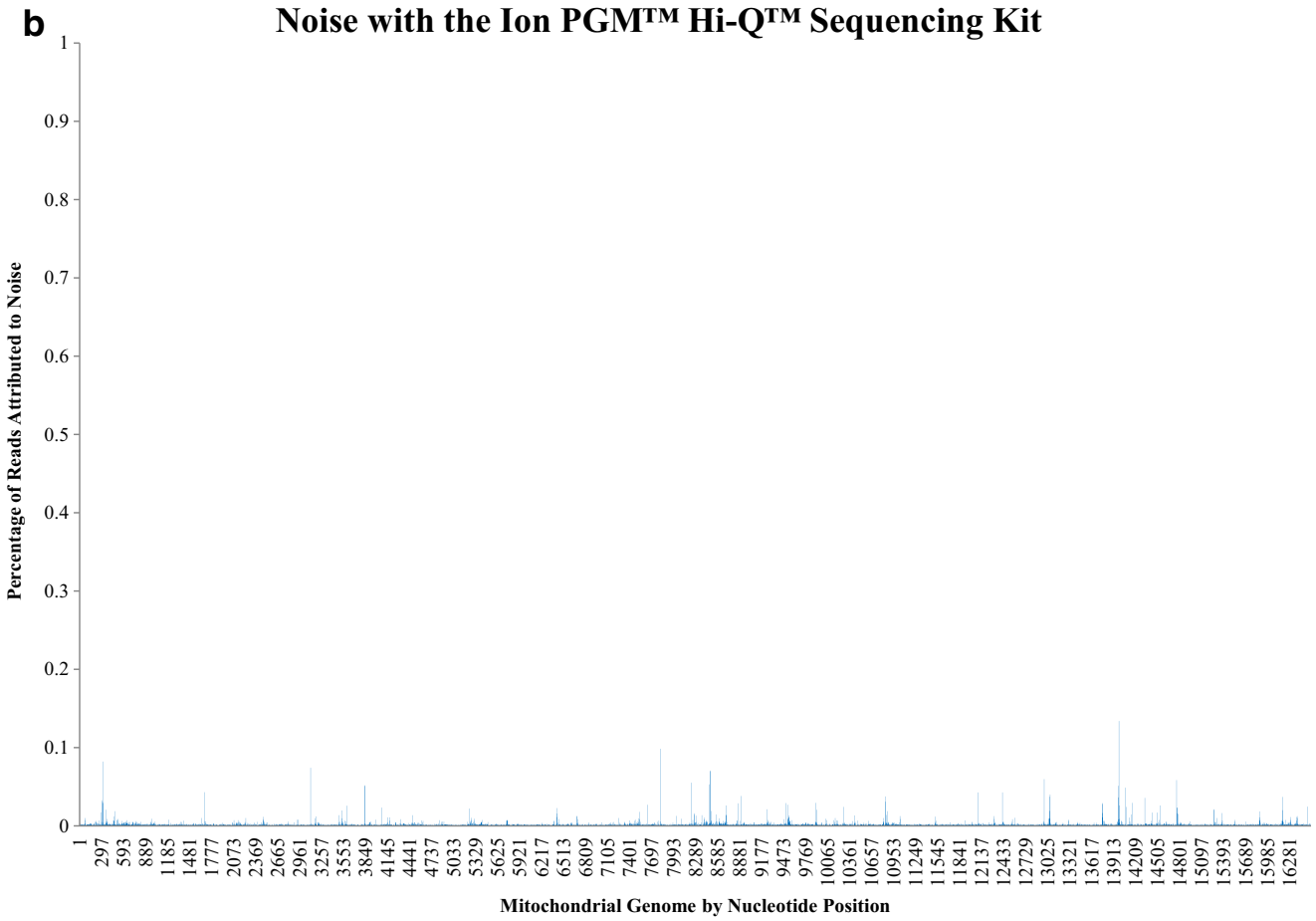
All 32 samples per sequencing chemistry were run on separate semi-conductor chips. The mitochondrial genome sequencing runs generated 674 and 606 megabases (Mb) of sequence data with the Sequencing 400 Kit and Hi-Q™ Sequencing Kit, respectively, and the mean read lengths were 199 and 187 bp, respectively.

Concordance

The two mitochondrial genome data sets generated using the Ion PGM™ Sequencing 400 Kit and the Ion PGM™ Hi-Q™ Sequencing Kit were compared to evaluate the effect on the quality and accuracy of the generated sequence data. Data analysis showed negative controls were clean with a read depth across the entire mitochondrial genome of 2X or lower. Haplotype calls for data generated with both the Sequencing 400 Kit and the Hi-Q™ Sequencing Kit were concordant.

Fig. 3 Average noise across the mitochondrial genome ($N = 31$) for the ▶ Ion PGM™ Sequencing 400 Kit **(a)** and for the Ion PGM™ Hi-Q™ Sequencing Kit **(b)**


a
Noise with the Ion PGM™ Sequencing 400 Kit
Percentage of Reads Attributed to Noise
Mitochondrial Genome by Nucleotide Position


b
Noise with the Ion PGM™ Hi-Q™ Sequencing Kit
Percentage of Reads Attributed to Noise
Mitochondrial Genome by Nucleotide Position

These samples were sequenced previously on the MiSeq [11] allowing for an evaluation of concordance between MPS platforms as the lower throughput of Sanger sequencing makes it impractical for concordance testing. Thus, other positive control samples were not needed. Haplotypes were concordant between the PGM and MiSeq data, excluding the number of Cs in homopolymers around nucleotide positions 310 and 16,189. Parson et al. [12] and Seo et al. [18] reported similar results when evaluating concordance.

Coverage

Average coverage across the mitochondrial genome for the 31 samples ranged from 145 reads (X) (±65X) to 2713X (±1255X) for the Sequencing 400 Kit and 222X (±91X) to 2224X (± 1006X) for the Hi-Q™ Sequencing Kit. However, since the two data sets were sequenced on different Ion 318™ Chips v2, a RLP calculation was performed to account for variability between the two runs that would not be attributed to chemistry performance (e.g., chip loading). Average RLP across the mitochondrial genome for the 31 samples ranged from 0.13 (± 1.56E−06) to 2.25 (±1.94E−05) for the Sequencing 400 Kit (Fig. 1a) and 0.22 (±2.80E−06) to 2.01 (±1.86E−05) for the Hi-Q™ Sequencing Kit (Fig. 1b). The RLP plots (Fig. 1) illustrate that both sequencing chemistries yield similar high- and low-coverage areas across the mitochondrial genome (Fig. 2) as was observed by Seo et al. [18]. Some of the coverage variation likely is attributable to the Ion PGM™ Chemistry's difficulty sequencing through homopolymeric stretches [2, 12, 16–18, 34] which is supported by the drop in coverage around the homopolymeric C stretches at nucleotide positions 310 and 16,189. An increase in the minimum RLP was observed with the Hi-Q™ Sequencing Kit, which may be indicative of better sequencing performance through homopolymeric regions.

Strand balance

Coverage on a per strand basis was analyzed to assess strand balance. Strand balance ratios for the Sequencing 400 Kit (Supplementary Fig. 1A) and the Hi-Q™ Sequencing Kit (Supplementary Fig. 1B) were highly concordant. The average positive strand balance across the entire mitochondrial genome for all samples was 50.31 % (±15.1 %) with the Sequencing 400 Kit and 44.77 % (±15.5 %) with the Hi-Q™ Sequencing Kit, with 50 % indicating equal coverage. For the Sequencing 400 Kit, 84.6 % of nucleotide positions fell within the positive strand balance range of 30 to 70 %. For the Hi-Q™ Sequencing Kit, 81.3 % of nucleotide positions fell within the positive strand balance range of 30 to 70 %. No pattern for strand bias between the two kits could be discerned. However, the 546 additional nucleotide positions (3.3 %) that fell within the positive strand balance range of 30 to 70 % for the Sequencing 400 Kit but not the Hi-Q™ Sequencing Kit had a positive strand balance below 30 % indicating a larger number of reads were generated for the negative strand with the Hi-Q™ Sequencing Kit at these nucleotide positions.

Noise

Noise evaluates the quality of the generated data. No substantial difference in the noise level of the generated data was observed between the two sequencing kits. However, 15

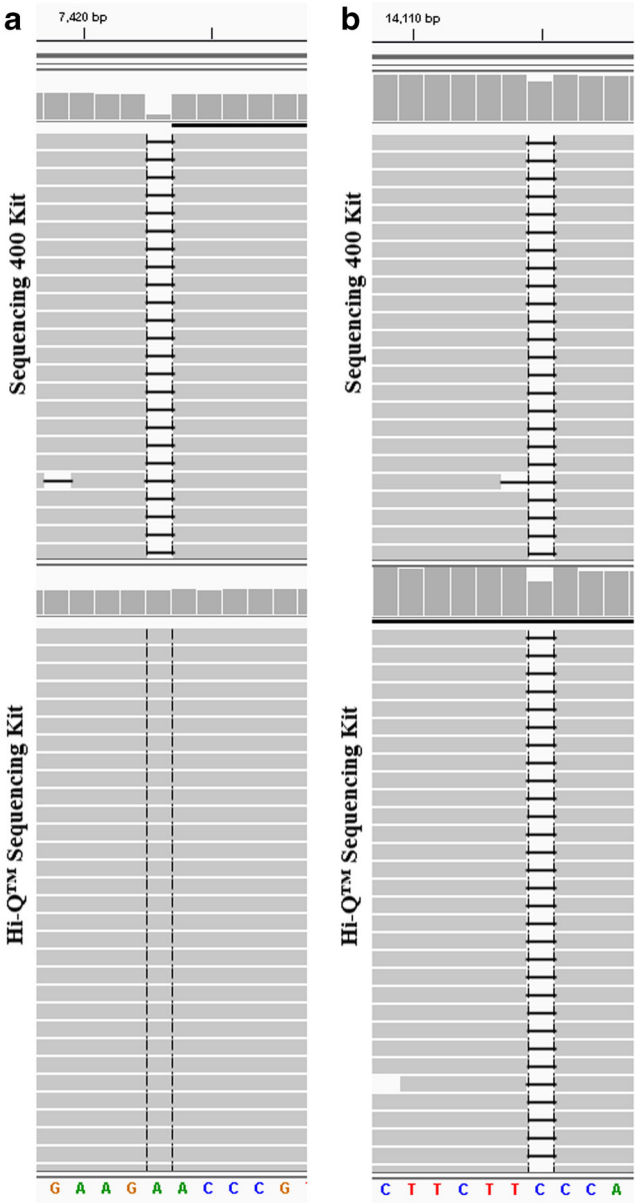


Fig. 4 **a** An example of data generated with the Sequencing 400 Kit (*top*) and the Hi-Q™ Sequencing Kit (*bottom*) in IGV where a decrease in false deletions can be seen at one nucleotide position. **b** An example of a comparison of data generated with the Sequencing 400 Kit (*top*) and the Hi-Q™ Sequencing Kit (*bottom*) in IGV where the false deletion rate at one nucleotide position remains unchanged

nucleotide positions exhibited a change greater than five percent. Of these 15, only four were lower with the Sequencing 400 Kit (Supplementary Table 1). The average percentage of reads attributed to noise across the entire mitochondrial genome for all samples was 0.14 % (±0.4 %) with the Sequencing 400 Kit and 0.12 % (±0.3 %) with the Hi-Q™ Sequencing Kit. Noise for the Sequencing 400 Kit ranged from 0 (±0 %) to 12.16 % (±0.6 %) with only 23 nucleotide positions at a noise level above five percent (Fig. 3a). Noise for the Hi-Q™ Sequencing Kit ranged from 0 (±0 %) to 13.38 % (±0.05 %) with only 19 nucleotide positions at a noise level above five percent (Fig. 3b).

Positions at or above a five percent noise level (Supplementary Table 1) were further investigated to determine their potential cause. One cluster of apparent noise was found to be the result of a mispriming of one of the long PCR primers (L644). In this instance, the last seven bases of the forward primer are complementary with another region (nucleotide positions 8486 to 8492; Supplementary Table 2). This homology contributes to seven of the positions with noise levels greater than five percent.

The remaining 27 of 34 nucleotide positions were found to be the result of a combination of false deletions and false insertions (discussed in greater detail below). Reads with these errors do not align well with the reference sequence, result in false substitutions, and subsequently were called "noise." For example, the nucleotide position 13,984 displayed the highest percentage of noise with minimal strand bias (average positive strand balance of 62 (±7.4 %) and 40 % (±7.1 %) with the Sequencing 400 Kit and Hi-Q™ Sequencing Kit, respectively) in both data sets. This nucleotide position falls at the end of a cytosine homopolymer four nucleotides long and appears to be more refractory to the Ion PGM™ Chemistry. Three apparent sequence species (or types) were detected between nucleotide positions 13,983 to 13,987 (Supplementary Table 3). Both major and minor noises contained a false deletion within the cytosine homopolymer 5′ to this nucleotide position. The major noise type was found to have a false insertion at nucleotide position 13,984 causing a shift in alignment for this position and creating an apparent substitution. The minor noise type contained a similar insertion at nucleotide position 13,985. Thus, interpretation of apparent substitutions should include considerations for possible false insertions and deletions associated with nearby homopolymers.

Deletion ratios

False deletions (termed "false" due to their partial presence in Ion PGM™ MPS reads) were observed throughout the Ion

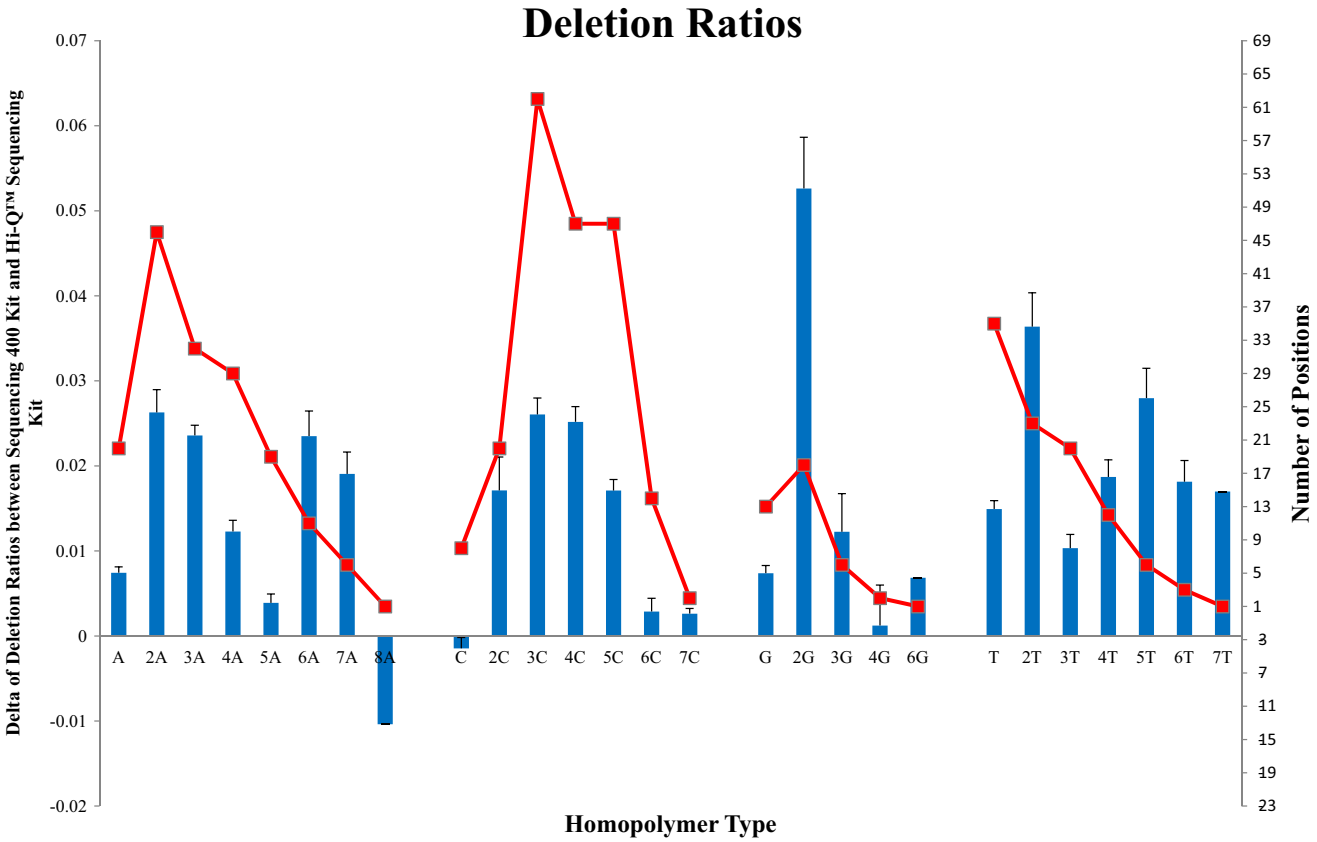


Fig. 5 Average delta of deletion ratios between Sequencing 400 Kit and Hi-Q™ Sequencing Kit generated sequences for each type of nucleotide and number of consecutive identical nucleotides. The *trend line* indicates the number of positions with that homopolymer type (e.g., data include one instance of an 8A homopolymer)

PGM™ data. The presence of false deletions in Ion PGM™-generated sequence data has been well-reported [2, 12, 16–18, 34]. These false deletions can be confirmed to be false with concordance testing as previous studies have illustrated a lack of these deletions in Sanger sequencing and MiSeq data [11, 12, 18]. A total of 504 nucleotide positions were found to have some level of false deletions, which was measured by a deletion ratio.

Most false deletions were located in reads of one of the two strands. With the Sequencing 400 Kit, the average deletion ratios at each of these 504 nucleotide positions ranged from 0 (±0) to 0.66 (±0.22) with 70 nucleotide positions displaying a deletion ratio greater than 0.10. With the Hi-Q™ Sequencing Kit, the average deletion ratios across these 504 nucleotide positions ranged from 0 (±0) to 0.44 (±0.45) with 38 nucleotide positions displaying a deletion ratio greater than 0.10. An overall decrease in systematic false deletions was observed with the Hi-Q™ Sequencing Kit. A total of 312 of the 504 nucleotide positions (61.9 %) had a positive delta value (Fig. 4a) while 127 of the 504 nucleotide positions (25.2 %) had a negative delta value. The remaining 65 nucleotide positions showed relatively no change in the deletion ratio between data sets (Fig. 4b).

Each of the 504 nucleotide positions were grouped by type of nucleotide (defined by the forward strand) and number of consecutive identical nucleotides indicating that these false deletions were associated largely (84.9 % of the 504 nucleotide positions analyzed) with homopolymeric regions (Supplementary Table 4). The deletion ratios were averaged across the nucleotide positions categorized by type of nucleotide and number of consecutive identical nucleotides (e.g., A, 2A, 3A) (Fig. 5). All but two categories, positions with an adenosine homopolymer eight nucleotides long or single cytosine residue, produced a positive average delta value. Thus, an overall reduction in the number of false deletions was observed when using the Hi-Q™ Sequencing Kit. Significant

Table 1 Observed distribution of non-significant, positive, and negative delta values for sequence sites classified by nucleotide type and number of consecutive identical nucleotides (total $n = 504$)

Nucleotide and homopolymeric length of the site	Number of sites with delta classification			
	NS	+ve	−ve	Total
A	10	8	2	20
2A	3	33	10	46
3A	2	24	6	32
4A	1	21	7	29
5A	5	11	3	19
6A	0	9	2	11
7A	1	3	2	6
8A	0	0	1	1
C	2	3	3	8
2C	1	13	6	20
3C	9	36	17	62
4C	4	32	11	47
5C	6	29	12	47
6C	3	7	4	14
7C	0	1	1	2
G	6	5	2	13
2G	1	10	7	18
3G	1	3	2	6
4G	0	1	1	2
6G	0	1	0	1
T	11	19	5	35
2T	3	12	8	23
3T	1	13	6	20
4T	0	8	4	12
5T	1	4	1	6
6T	1	2	0	3
7T	0	1	0	1

differences (by the permutation test) were observed across length of homopolymers only for the sites with an adenosine nucleotide ($\chi^2 = 38.52$, $p = 0.0001$) (Table 1). The other three nucleotides did not show any effect of length of homopolymers ($\chi^2 = 6.49$, $p = 0.9079$ for nucleotide C; $\chi^2 = 9.91$, $p = 0.2599$ for nucleotide G; and $\chi^2 = 14.12$, $p = 0.2961$ for nucleotide T). Combined data on all four types of nucleotide sites showed a significant effect of length of homopolymers ($\chi^2 = 51.05$, $p < 0.0001$). When data on all four nucleotide types were pooled over all lengths of homopolymers, there was no significant difference across the four nucleotide types ($\chi^2 = 5.92$, $p = 0.4287$). However, overall, delta values showed a significant effect of length of homopolymers over all nucleotides ($\chi^2 = 51.05$, $p < 0.0001$) (Fig. 6).

While the Hi-Q™ Sequencing Chemistry does not completely eliminate false deletions, it does improve sequencing through homopolymers. Moreover, the change in deletion ratios for the single occurrence adenine, guanine, cytosine, and thymine residues indicate homopolymers are not the sole cause of false deletions. A total of 76 of the 504 nucleotide positions analyzed (15.1 %) were single occurrence adenosine, guanosine, cytosine, or thymine residues. Positive average delta values of the deletion ratios between the two enzymes were seen for the single adenosine, guanosine, and thymine residues, while a negative average delta value was seen for the single cytosine residues.

There were no incorrect typings with either sequencing chemistry, and the results also were concordant with those obtained previously on the MiSeq platform. Long homopolymeric stretches of mitochondrial DNA often are not considered in evidence and reference profile comparisons [12, 14, 18, 35]. However, the noise that results from sequencing through these homopolymers can make interpretation difficult for analysts. As an example, the thymine residue at nucleotide position 16,189 is flanked by five cytosine residues and four cytosine residues between nucleotide positions 16,183 and 16,194, respectively. The T16189C transition creates a long cytosine homopolymer that generates length heteroplasmy and false deletions during sequencing (Fig. 7a) [12]. These length heteroplasmies and false deletions create noise and thus less confidence in variant calls. Thus, base calling for the positions surrounding homopolymers can be affected (Fig. 7b).

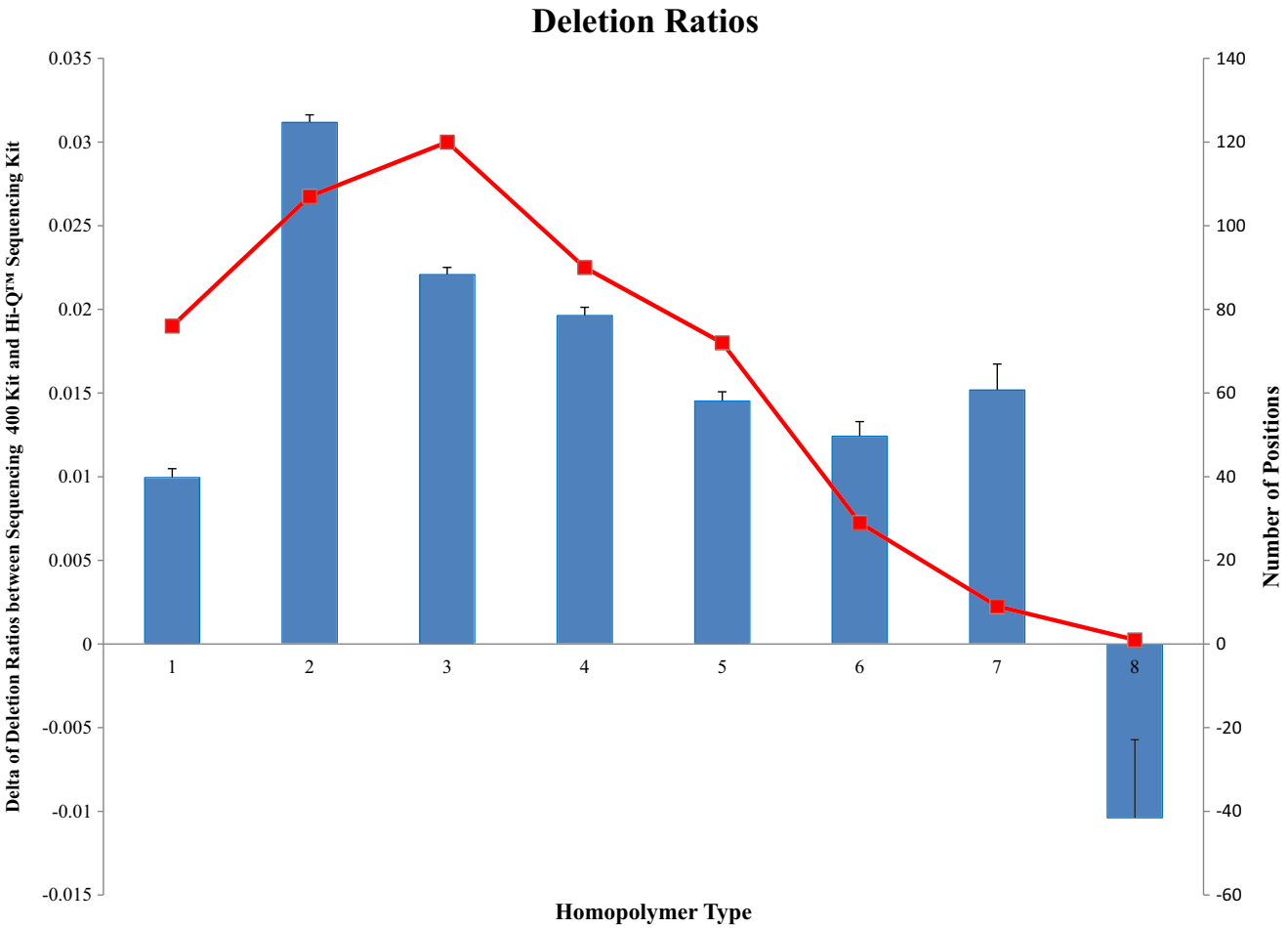


Fig. 6 Average delta of deletion ratios between Sequencing 400 Kit and Hi-Q™ Sequencing Kit generated sequences grouped by number of consecutive identical nucleotides

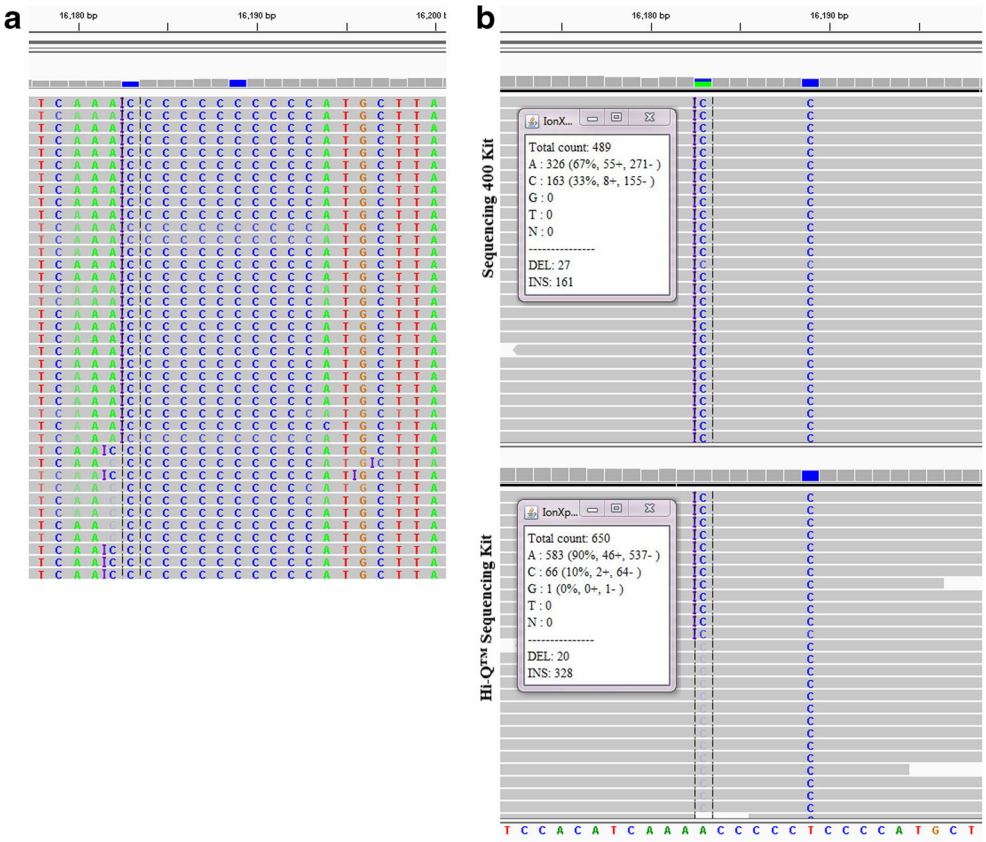


Fig. 7 **a** An illustration in IGV of the T16189C transition that results in an uninterrupted C stretch and noise. **b** A comparison in IGV of the T16189C transition and resulting sequence noise with the Sequencing 400 Kit (*top*) and the Hi-Q™ Sequencing Kit (*bottom*)

STR data

The two STR data sets generated using the Ion PGM™ Sequencing 400 Kit and the Ion PGM™ Hi-Q™ Sequencing Kit also were compared to evaluate the effect on the quality and accuracy of the generated sequence data. The STR sequencing runs generated 237 and 481 Mb of sequence data with the Sequencing 400 Kit and Hi-Q™ Sequencing Kit, respectively, and the mean read length was 106 bp for both sequencing runs. The difference between the two STR sequencing runs in the amount of sequence data generated can be attributed mostly to the difference in ISP loading for the Ion 318™ Chips v2 used in the sequencing runs. The first run had an ISP loading of 34 % while the second run had an ISP loading of 68 %. Differences in ISP density after the chip loading process are a user-induced variability and not a result of the sequencing chemistry. Data analysis showed negative controls were clean with a read depth of 3X or lower, and genotypes were generated for all ten markers on all samples in both data sets. Additionally, genotype calls produced by the HID_STR_Genotyper plugin and STRait Razor were concordant. Genotype calls for data generated with both sequencing kits were concordant. These samples also were typed with the GlobalFiler™ Kit on a CE instrument, and the genotype calls were concordant between the MPS and CE data.

There were no notable differences in RLP, strand balance, ACRs, and SCRs for the STR data between the two sequencing kits (data not shown). The lack of difference in performance is not surprising as this HID-Ion STR 10-plex panel was selected because these loci performed particularly well with the original sequencing chemistry. The bias in selection for high performing STRs likely reduced the chances of observing an improved sequencing performance with this panel. Nonetheless, the similar performance demonstrates that the Hi-Q™ Sequencing Chemistry does not have a negative impact on obtaining resultant data.

Conclusions

The mitochondrial genome and STR data produced in this study were accurate and reliable with concordance between different methodologies. While depth of coverage and strand balance variations were identified, these variations did not impact the accuracy of typing calls. Noise levels, generally low, had no impact on the reliability of typing calls. The Ion Torrent's Hi-Q™ Sequencing Chemistry offers an improvement in sequencing performance. The Ion PGM™ Sequencing 400 Kit and the Ion PGM™ Hi-Q™ Sequencing Kit were found to produce highly concordant sequencing results in relation to data accuracy, coverage, strand balance, and noise levels. However, significant differences were observed between the two sequencing kits for deletion

ratios in the mitochondrial genome sequencing data and the Ion PGM™ system's ability to sequence through homopolymeric regions. The decrease in deletion ratios supports that sequencing through homopolymeric regions can be improved with the use of the Hi-Q™ Sequencing Chemistry.

There were no observable differences between the two sequencing kits and the STR resultant data. Unlike the mitochondrial genome where the entire genome was used for study in an unbiased fashion, the STRs selected for the HID-Ion STR 10-plex were those that performed well and were robust with the original sequencing chemistry. Therefore, finding no difference in performance between the kits was expected. It is important to note though that the new chemistry did not have a negative impact on sequencing performance for the selected STRs.

The overall data support that the Ion PGM™ system is robust for sequencing the mitochondrial genome and the HID-Ion STR 10-plex panel. By understanding the limitations of MPS data, interpretation guidelines and bioinformatic tools can be developed that allow for better analysis of MPS data. Thus, the data presented herein also have bioinformatic value. Defining depth of coverage levels, strand balance ratios, and noise levels will contribute to providing reliable allele calls. Observations on the potential causes of noise, showing false deletions largely correlate to homopolymeric regions, and that these false deletions are primarily found in reads of one direction may allow for development of algorithms that better analyze such data. Continued validation studies of MPS technologies will facilitate development of robust interpretation guidelines and bioinformatic tools.

Acknowledgments We thank Thermo Fisher Scientific for providing reagents necessary to complete this study and Robert Lagace, Joseph Chang, Sharon Wootton, and Chien-Wei Chang, specifically, for their necessary technical expertise. We also thank Monika Stoljarova for her technical expertise and work in developing background knowledge over the potential causes of noise seen when sequencing the mitochondrial genome with MPS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46
2. Borsting C, Morling N (2015) Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet 18:78–89
3. Churchill JD, Schmedes SE, King JL, Budowle B (2016) Evaluation of the Illumina Beta Version ForenSeq™ DNA signature. Forensic Sci Int Genet 20:20–29
4. Warshauer DH, Churchill JD, Novroski N, King JL, Budowle B (2015) Novel Y-chromosome short tandem repeat variants detected through the use of massively parallel sequencing. Genomics Proteomics Bioinformatics 13:250–257
5. Gettings KB, Aponte RA, Vallone PM, Butler JM (2015) STR allele sequence variation: current knowledge and future issues. Forensic Sci Int Genet 18:118–130
6. Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM (2016) Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Sci Int Genet 21:15–21
7. Fordyce SL, Mogensen HS, Borsting C, Lagacé RE, Chang CW, Rajagopalan N, Morling N (2015) Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™. Forensic Sci Int Genet 14:132–140
8. Churchill JD, Chang J, Ge J, Rajagopalan N, Wootton SC, Chang CW, Lagace R, Liao W, King JL, Budowle B (2015) Blind study evaluation illustrates utility of the ion PGM™ system for use in human identity DNA typing. Croat Med J 56:218–229
9. Zhao X, Ma K, Li H, Cao Y, Liu W, Zhou H, Ping Y (2015) Multiplex Y-STRs analysis using the ion torrent personal genome machine (PGM). Forensic Sci Int Genet 19:192–196
10. Zeng Z, King J, Hermanson S, Patel J, Storts DR, Budowle B (2015) An evaluation of the PwerSeq™ auto system: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing. Forensic Sci Int Genet 19:172–179
11. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warchauer DH, Davis CP, Parson W, Sajantila A, Budowle B (2014) High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12:128–135
12. Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagace R, Irwin J (2013) Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet 7:632–639
13. Borsting C, Fordyce SL, Olofsson J, Mogensen HS, Morling N (2014) Evaluation of the Ion Torrent™ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing. Forensic Sci Int Genet 12:144–154
14. Just RS, Irwin JA, Parson W (2015) Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int Genet 18:131–139
15. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352
16. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30:434–439
17. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in ion torrent PGM data. PLoS Comput Biol 9:e1003031
18. Seo SB, Zeng X, King JL, Larue BL, Assidi M, Al-Qahtani MH, Sajantila A, Budowle B (2015) Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform Ion Torrent™ PGM™. BMC Genomics 16(Suppl 1):S4
19. Bottino CG, Chang CW, Wootton S, Rajagopalan N, Langit R, Lagace RE, Silva R, Moura-Neta RS (2015) STR genotyping using ion torrent PGM and STR 24-plex system: performance and data interpretation. Forensic Sci Int Genet 5:e325–e326
20. Eduardoff M, Santos C, de la Puente M, Gross TE, Fondevila M, Strobl C, Sobrino B, Ballard D, Schneider PM, Carracedo A, Lareu MV, Parson W, Phillips C (2015) Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™. Forensic Sci Int Genet 17:110–121

21. Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Legal Med 127: 1079–1086
22. Qiagen (2012) QIAamp® DNA mini and blood mini handbook. Qiagen, Valencia
23. Thermo Fisher Scientific (2013) Ion PGM™ sequencing 400 kit. Thermo Fisher Scientific, Waltham
24. Thermo Fisher Scientific (2014) Ion PGM™ Hi-Q™ sequencing kit. Thermo Fisher Scientific, Waltham
25. Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M (2011) High throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res 21:1–11
26. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147
27. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192
28. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
29. King JL, Sajantila A, Budowle B (2014) mitoSAVE: mitochondrial sequence analysis of variants in Excel. Forensic Sci Int Genet 12: 122–125
30. Warshauer DH, King JL, Budowle B (2015) STRait Razor v2.0: the improved STR allele identification tool—Razor. Forensic Sci Int Genet 14:182–186
31. Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, Budowle B (2013) STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet 7:409–417
32. Thermo Fisher Scientific (2015) GlobalFiler™ PCR amplification kit. Thermo Fisher Scientific, Waltham
33. Snedecor GW, Cochran WG (1967) Statistical methods, 6th edn. Iowa Univ Press, Ames, pp 95–98
34. Bertolini F, Ghionda MC, D'Alessandro E, Geraci C, Chiofalo V, Fontanesi L (2015) A next generation semiconductor based sequencing approach for the identification of meat species in DNA mixtures. PLoS ONE 10, e0121701
35. Scientific Working Group on DNA Analysis Methods [SWGDAM] (2013) Interpretation guidelines for mitochondrial DNA analysis by Forensic DNA Testing Laboratories. http://swgdam.org/SWGDAM%20mtDNA_Interpretation_Guidelines_APPROVED_073013.pdf