Molecular Therapy
Nucleic Acids
Original Article

An innovative data analysis strategy for accurate next-generation sequencing detection of tumor mitochondrial DNA mutations

Shanshan Guo,[1,4] Kaixiang Zhou,[1,4] Qing Yuan,[2] Liping Su,[1] Yang Liu,[1] Xiaoying Ji,[1] Xiwen Gu,[3] Xu Guo,[1] and Jinliang Xing[1]

[1]State Key Laboratory of Cancer Biology and Department of Physiology and Pathophysiology, Fourth Military Medical University, Xi'an, China; [2]Institute of Medical Research, Northwestern Polytechnical University, Xi'an, China; [3]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

Next-generation sequencing technology has been commonly applied to detect mitochondrial DNA (mtDNA) mutations, which are reported to be strongly associated with cancers. However, several key challenges still exist regarding bioinformatics analysis of mtDNA sequencing data that greatly affect the detection accuracy of mtDNA mutations. Here we comprehensively evaluated several key analysis procedures in three different sample types. We found that a trimming procedure was essential for improving mtDNA mapping performance in plasma but not tissue samples. Mapping with a revised Cambridge reference sequence and human genome 19 reference was strongly suggested for mtDNA mutation detection in plasma samples because of the extreme abundance of nuclear DNA of mitochondrial origin. Moreover, our results showed that a setting of 3 mismatches was most appropriate for mtDNA mutation calling. Importantly, we revealed the presence of a negative logarithmic relationship between mtDNA site sequencing depth and minimum detectable mutation frequency and built an innovative and efficient filtering strategy to increase the accuracy and sensitivity of mutation detection. Finally, we verified that higher sequencing depth was required for a PCR-based compared with a capture-based enrichment strategy. We established an innovative data analysis strategy that is of great significance for improving the accuracy of mtDNA mutation detection for different types of tumor samples.

INTRODUCTION

Human mitochondria possess their own genome, double-stranded, maternally inherited, circular DNA of 16,569 base pairs (bp) with up to 10^3–10^4 copies in each cell.[1] Mitochondrial DNA (mtDNA) exists with plenty of sequence variants that are commonly observed as germline or somatic mutations.[2] Numerous studies have demonstrated that heteroplasmy, the presence of mixed mtDNA mutation genotypes in a cell, is strongly associated with many disease phenotypes, especially when the percentage of mutations exceeds a critical threshold.[3] Thus, it is of great importance to accurately detect mtDNA mutations for a better understanding of mitochondrial biology and cancers.

Traditional techniques, such as Sanger sequencing and high-resolution melt analysis, have been used for mtDNA mutation detection but still face the disadvantages of low throughput and sensitivity. The advent of next-generation sequencing (NGS) technologies provides an opportunity for high-throughput and low-cost detection of mitochondrial genome-wide mutations.[4] Two major methods have been used for NGS-based detection of mtDNA mutations.[5] One is direct data extraction from whole-genome sequencing (WGS) or whole-exome sequencing (WES) for mutation analysis, which has low throughput and is not cost effective.[6] Another is to first enrich mtDNA from total genomic DNA and then detect it by NGS, mainly including PCR-based and capture-based enrichment strategies.[7] This can achieve detection of low minor allele frequency (MAF) and permit customized sequencing strategies, allowing segment-specified or WGS of mtDNA. Combined with NGS technology, such strategies have been proven to be powerful and efficient when depicting the spectrum of mtDNA somatic mutation in many types of cancers.[3]

Although NGS has been applied extensively to effectively detect mtDNA mutations, there are still several big challenges regarding data analysis of mtDNA sequencing.[8] Among them, the most critical is to decrease the number of false positive and false negative mutations, which are greatly affected by mtDNA sequencing depth, especially those with a low heteroplasmy level.[9] To date, the quantitative relationship between sequencing depth and the detection accuracy of mtDNA mutations remains to be determined. Therefore, most previous studies arbitrarily selected 5%, 2%, or 1% as the minimal

Received 24 July 2020; accepted 5 November 2020;
https://doi.org/10.1016/j.omtn.2020.11.002.

[4]These authors contributed equally

Correspondence: Dr. Jinliang Xing, State Key Laboratory of Cancer Biology and Department of Physiology and Pathophysiology, Fourth Military Medical University, 169 Changle West Road, Xi'an 710032, China.
E-mail: xingjl@fmmu.edu.cn

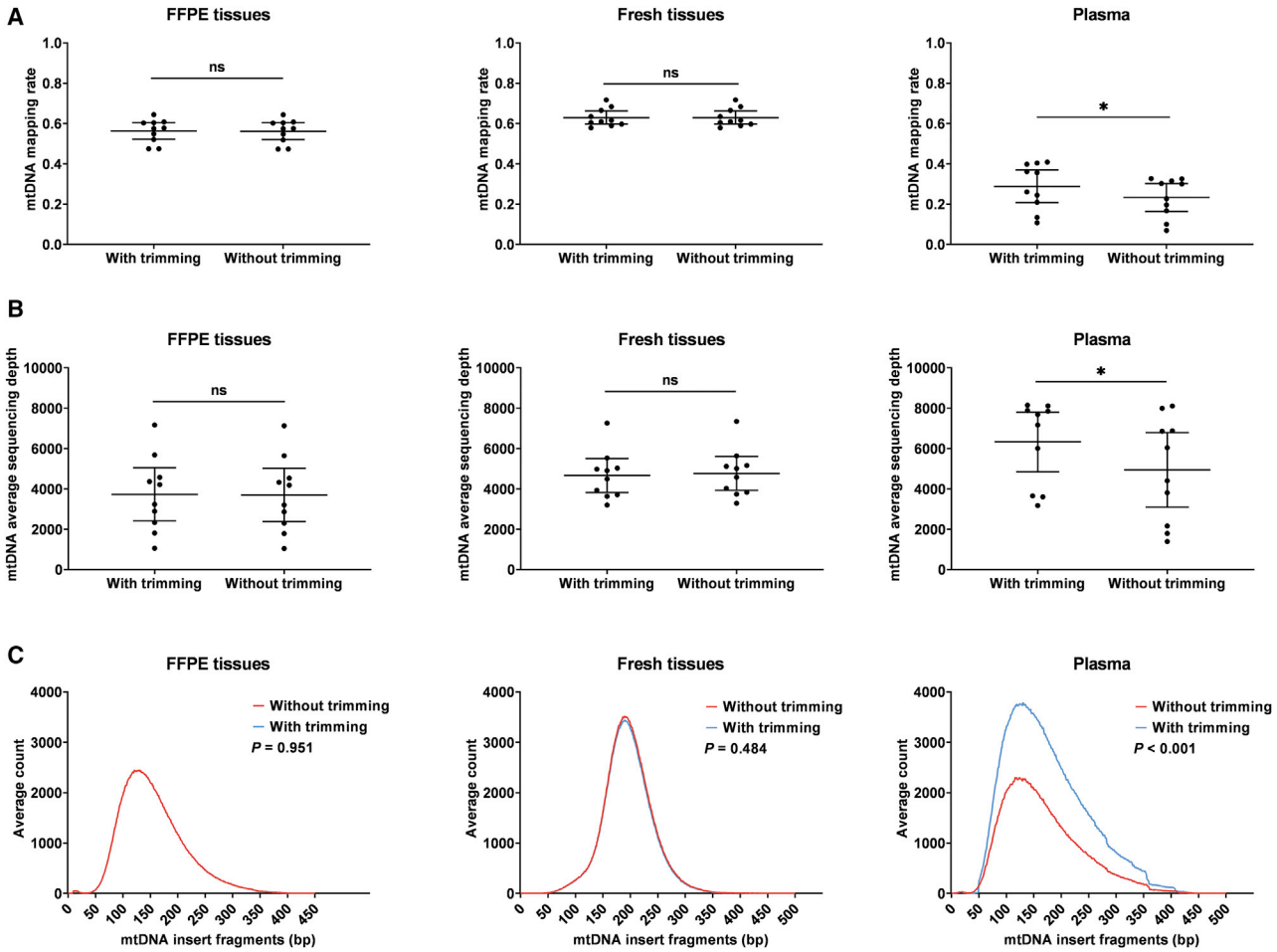


Figure 1. Effect of the trimming procedure on mapped mtDNA sequencing data

(A) Comparison of mtDNA mapping rate between the trimming group and no-trimming group in FFPE tissue, fresh tissue, and plasma samples. (B) Comparison of mtDNA average sequencing depth between the trimming group and no-trimming group in FFPE tissue, fresh tissue, and plasma samples. (C) Distribution of mtDNA insert size between the trimming group and no-trimming group in FFPE tissue, fresh tissue, and plasma samples. ns, no significance; *p < 0.05. (A–C) 10 FFPE tissues, 10 fresh tissues, and 10 plasma samples were used.

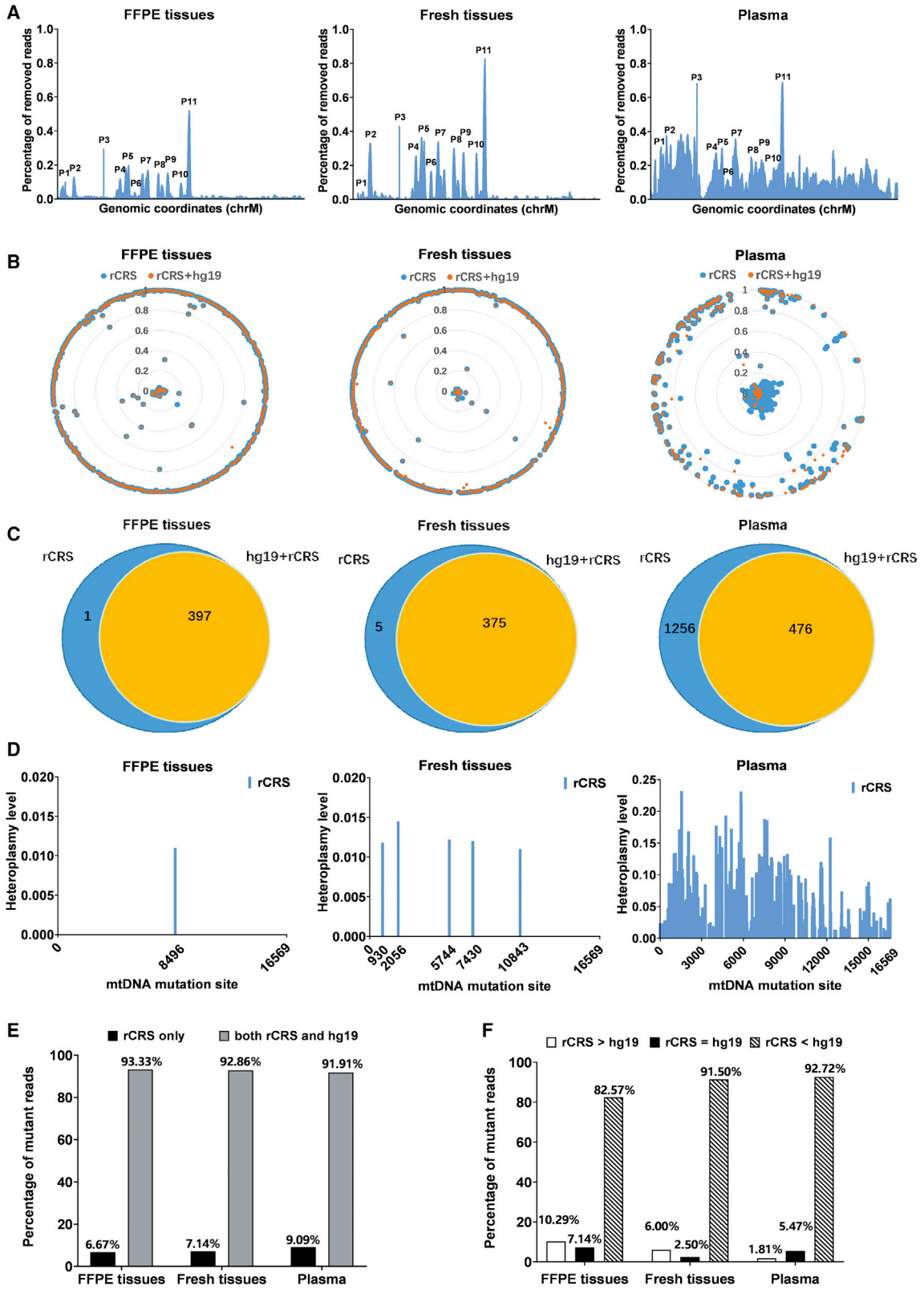
heteroplasmy level to filter mtDNA mutations.[10–12] In addition, because of the presence of homologous sequences in the nuclear genome, referred as nuclear sequences of mitochondrial origin (NUMTs), a suitable mtDNA mapping strategy should be carefully assessed to reduce the effect of NUMTs in different sample types.[13] Considering the great decrease of NGS cost in recent years, a high sequencing depth (usually more than thousands) is becoming practical for tumor mtDNA mutation detection. There is an urgent need to develop an innovative data analysis strategy for more accurate NGS detection of mtDNA mutations, especially those with a low heteroplasmy levels in cancer cells.

In the present study, mainly based on capture-based NGS data, we systematically evaluated several key analysis procedures, including trimming, mapping, and filtering, in mtDNA mutation detection of formalin-fixed and paraffin-embedded (FFPE) tissue, fresh tissue, and plasma samples from individuals with cancer. Finally, an innovative bioinformatics pipeline integrating a newly developed filtering algorithm was established. Furthermore, application of our analysis strategy to different sample types was also evaluated, providing a flexible selection of key analysis procedures.

RESULTS

Effect of the trimming procedure on mapped mtDNA sequencing data

In NGS data analysis, a trimming procedure is commonly used to remove adaptor sequences and bases with low quality from sequencing reads. Therefore, we first evaluated the effect of the trimming procedure on mtDNA mapping of sequencing reads in three types of samples: FFPE tissue, fresh tissue, and plasma. As shown in Figures 1A and 1B, the trimming procedure had no significant influence on the mtDNA mapping rate of sequencing reads and average sequencing



(legend on next page)

depth in FFPE and fresh tissue samples. In contrast, the mtDNA mapping rate and average sequencing depth were increased significantly in plasma samples after trimming (Figures 1A and 1B). Furthermore, our data indicated a very consistent size distribution of mtDNA insert fragments in FFPE and fresh tissue samples (Figure 1C). However, the mapped mtDNA sequencing reads were increased significantly in plasma samples after trimming (Figure 1C; $p < 0.001$). These results indicate that the trimming procedure is essential for improving mtDNA mappability in plasma but not tissue samples.

Evaluation of the mtDNA mapping strategy in mtDNA mutation calling

Next we evaluated application of two commonly used mapping strategies in three different sample types. As shown in Figure 2A, compared with the first mapping strategy (revised Cambridge reference sequence [rCRS] alone), the second mapping strategy (combined rCRS-human genome 19 [hg19]) clearly removed the sequencing reads mapped to hg19 and rCRS, leading to 11 more remarkable peaks. Among them, the most significant peak occurred around mtDNA site 8,500, with 45%–80% removed reads in three different sample types. Compared with tissue samples, plasma DNA exhibited a clearly different distribution pattern of removed reads. The mtDNA mutations detected by the two mapping strategies are depicted in Figure 2B, with mutation site and heteroplasmy level (MAF $\geq$ 1%). The Venn diagram shows that all mtDNA mutations detected by the second mapping strategy were included in those detected by the first one, whereas 1, 5, and 1,256 mtDNA mutations only detected by mapping to the rCRS were observed in FFPE tissue, fresh tissue, and plasma samples, respectively (Figure 2C). The site and heteroplasmy level of mtDNA mutations only detected by mapping to the rCRS are shown in Figure 2D. To explore a potential reason for the extremely high number of mtDNA mutations only detected by mapping to the rCRS in plasma samples, we further analyzed the mtDNA copy numbers and percentages of mtDNA sequencing reads mapped to hg19 and the rCRS in the three sample types. Our results showed that mtDNA copy numbers in plasma were significantly lower than in tissue (Figure S1) and that the percentage of mtDNA sequencing reads mapped to hg19 and the rCRS was significantly higher in plasma (17.18%) compared with FFPE and fresh tissue (3.46% and 4.03%, respectively) (Figure S2). To investigate the possibility of NUMT derivation of those mutations, we determined the percentages of mutant reads mapped to the rCRS only or to hg19 and the rCRS. As shown in Figure 2E, 93.33%, 92.86%, and 91.91% of the mutant reads were mapped to hg19 and the rCRS in the three sample types, respectively, whereas 6.67%, 7.14%, and 9.09% were only mapped to the rCRS. Further analysis showed that 82.57%, 91.50%, and 92.72% of the mutant reads had higher alignment scores when mapped to hg19 than to the rCRS in the three sample types, respectively (Figure 2F). These data suggest that mtDNA mutations only detected by mapping to the rCRS may be introduced by NUMTs. To confirm this hypothesis, the PCR-based enrichment approach, which is not affected by NUMTs, was further leveraged to detect mtDNA mutations of 6 plasma samples. As expected, our results showed no significant difference between the two mapping strategies for detecting mtDNA mutations (Figure S3), which further verified our hypothesis that the increased mtDNA mutations were mainly introduced by NUMTs during the capture process in plasma samples when mapped to the rCRS alone. When mtDNA mutations are detected in tissue samples, both mapping strategies can be used, although rare mutations may be introduced by NUMTs. In comparison, the second mapping strategy is strongly suggested for mtDNA mutation detection in plasma samples because of the extreme abundance of NUMTs.

Effect of mismatch number selection on mtDNA mutation calling

We next investigated the effect of mismatch number selection in the second mapping strategy on mtDNA mutation calling by using the repeated sequencing data in the three sample types. As shown in Figure 3A, the total repeated mtDNA mutation numbers in two repeated experiments of 10 samples increased gradually when the mismatch number changed from 1 to 4 in the three sample types. We found that the average number of repeatable mutations in the group with 3 or 4 mismatches was significantly higher than in the group with 1 mismatch in all three sample types. The average number of repeatable mutations varied from 35.5–42.6, 32.3–36.6, and 37.7–49.2 under different mismatches in the three sample types, respectively. Furthermore, the assumed false positive (AFP) mutations were defined as those only detected in one mismatch group. The assumed false negative (AFN) mutations were defined as those not detected in this group but detected in at least two other mismatch groups. We found no significant difference in AFP mutation number among the groups with 1, 2, or 3 mismatches in three sample types, whereas it was increased significantly in the group with 4 mismatches compared with the groups with 1, 2, or 3 mismatches (Figure 3B). In contrast, the AFN mutation numbers in groups with 1 or 2 mismatches were significantly higher than in the groups with 3 or 4 mismatches among three sample types (Figure 3C). These results demonstrate the effects, to some extent, of mismatch number selection on mtDNA mutation detection and strongly suggest a setting of 3 mismatches for mtDNA mutation calling.

Figure 2. Evaluation of the mtDNA mapping strategy in mtDNA mutation calling

(A) Distribution of removed reads after combined rCRS-hg19 mapping compared with rCRS-alone mapping (calculated as the percentage of removed reads to total reads). (B) Distribution of mtDNA mutations detected by the two mapping strategies (rCRS-alone and combined hg19-rCRS) in the three sample types. Inner circles represent heteroplasmy levels. Each color-coded dot corresponds to the mutations detected by rCRS-alone mapping (orange) or combined hg19-rCRS mapping (blue). (C) Venn diagram of mtDNA mutations detected by the two mapping strategies in the three sample types. (D) Heteroplasmy levels and distribution of mtDNA mutations only detected by rCRS-alone mapping. (E) The majority of mutant reads only detected by rCRS-alone mapping can be mapped to hg19 and the rCRS. (F) The majority of mutant reads with dual hg19 and rCRS mapping showed a greater hg19 alignment score compared with the rCRS alignment score. (A–F) 10 FFPE tissues, 10 fresh tissues, and 10 plasma samples were used.

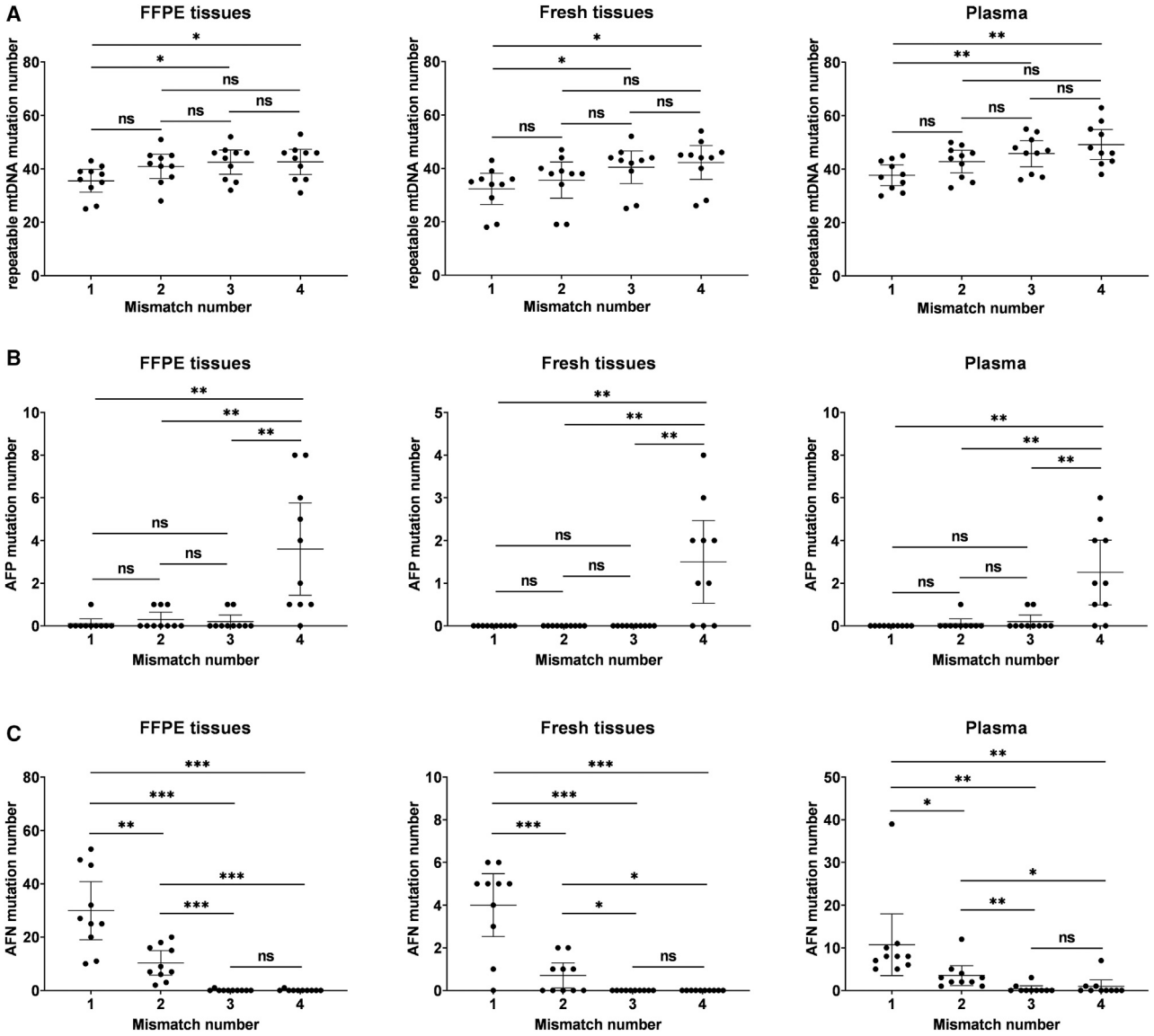


Figure 3. Effect of mismatch number selection on mtDNA mutation calling

(A) The number of repeatable mtDNA mutations detected in two repeated experiments when the mismatches were set to 1–4. (B) The number of assumed false positive (AFP) mtDNA mutations detected in the three sample types when the mismatches were set to 1–4. (C) The number of assumed false negative (AFN) mtDNA mutations detected in three sample types when the mismatches were set to 1–4. (A–C) 10 FFPE tissues, 10 fresh tissues, and 10 plasma samples were used. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Relationship between site sequencing depth and consistency of mtDNA mutations at different heteroplasmy levels

Next we comprehensively investigated the effect of mtDNA site sequencing depth on the detection accuracy of mutations at low heteroplasmy levels by using the repeated sequencing data in the three sample types. As shown in Figure 4A, a very consistent trend was observed in all three sample types, indicating a faster rise of consistency when the site sequencing depth was relatively low. To achieve a consistency of 90%, a site sequencing depth of more than 2,700×, 2,300×, and 3,200× was needed in FFPE tissue, fresh tissue, and plasma, respectively, when the heteroplasmy level was set above 0.5%. In comparison, the consistency of mtDNA mutations was decreased notably in all three sample types when the heteroplasmy level was set at 0.1%, suggesting that the system used in this study may not be suitable for mtDNA mutation detection at a heteroplasmy level of 0.1%. Furthermore, a negative logarithmic relationship was established between mtDNA site sequencing depth and minimum detectable mutation frequency when the consistency was set at 95%

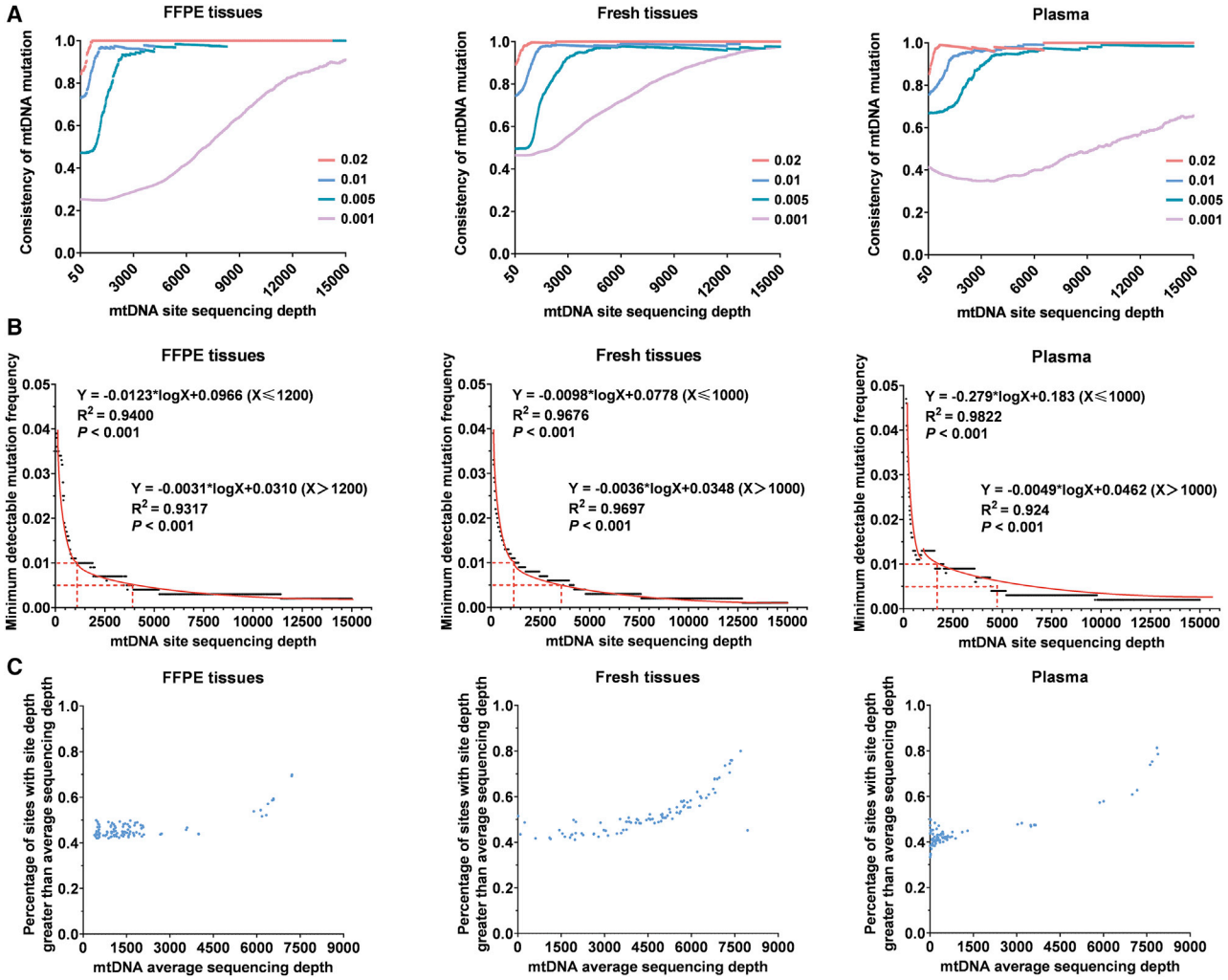


Figure 4. Relationship between site sequencing depth and consistency of mtDNA mutations at different heteroplasmy levels

(A) Consistency of mtDNA mutations at different site sequencing depths of two repeated experiment under the given heteroplasmy level. (B) The negative logarithmic relationship between mtDNA site sequencing depth and minimum detectable mutation frequency, with the consistency of mtDNA mutations in two repeated experiments higher than 95%. (C) Relationship between the average mtDNA sequencing depth and percentage of mtDNA sites with a site sequencing depth greater than the average sequencing depth. (A–C) 50 FFPE tissues, 50 fresh tissues, and 50 plasma samples were used.

(Figure 4B). When the minimum detectable mutation frequency was 1% and 0.5%, the site sequencing depth was required to be higher than 1,000× and 4,000×, 1,200× and 3,600×, and 1,700× and 4,700× in FFPE tissue, fresh tissue, and plasma, respectively.

Commonly, the average sequencing depth was calculated and provided based on mtDNA sequencing data. Considering that the accuracy of mtDNA mutation calling was affected by site sequencing depth, we determined the relationship between them. As shown in Figure 4C, with increasing average sequencing depth, the percentage of mtDNA sites with a site depth greater than the average sequencing depth increased gradually. When the average sequencing depth was greater than 8,000×, there were 80% of sites with a site depth higher than the average.

Assessment of different mtDNA mutation filtering strategies

To reduce false positive mutations, several filtering strategies have commonly been applied.[14] First, we assessed the effect of two filtering strategies reported previously on mtDNA mutation calling. One filtering strategy (filter 1) is to remove C > A/G > T mutations with low MAF (1%) and strong sequence context bias (at CpCpN > CpApN; most frequently at CpCpG > CpApG), which is known to arise from artificial guanine oxidation during sequencing library preparation steps.[14,15] As shown in Figure 5A, no difference in mtDNA mutation numbers was observed between two groups filtered with and without filter 1 in FFPE tissue, fresh tissue, and plasma. The second filtering strategy (filter 2) is to remove mtDNA mutations where the quality of the mutant bases does not fit well with binominal distribution ($p > 0.0001$).[16,17] These may be false positive mutations

introduced by sequencing errors. Very similarly to filter 1, filter 2 exhibited no significant effect on mtDNA mutation numbers in all three sample types (Figure 5B). Considering the great influence of site sequencing depth on mtDNA mutation detection, we built a novel filtering strategy (filter 3) based on negative logarithmic functions, presented in Figure 4B, to remove mtDNA mutations with a heteroplasmy level lower than the minimum detectable mutation frequency. As shown in Figure 5C, the number of mtDNA mutations was decreased significantly after filtering in all three sample types. Moreover, the repeated mtDNA sequencing data were used to analyze whether those filtered mutations were repeatable. We found that 91.84%, 89.4%, and 93.3% of these filtered mutations were not repeatable in FFPE tissue, fresh tissue, and plasma, respectively. In comparison, only 9.4%, 5.7%, and 15% of the unaffected mtDNA mutations were not repeatable (Figure S4). These findings indicate that filter 1 and filter 2 are not necessary in our analysis pipeline, whereas filter 3 greatly contributes to improving detection accuracy.

Comparison of PCR-based and capture-based enrichment strategies or two sequencing platforms

To reduce sequencing cost, capture-based and PCR-based strategies are commonly used for mtDNA enrichment from total genomic DNA. Therefore, we systematically compared the performance of two enrichment strategies in NGS-based mtDNA mutation detection of plasma samples. We found that capture-based mtDNA sequencing exhibited a more uniform distribution of coverage across the whole mitochondrial genome and that the coefficient of variation (CV) decreased significantly compared with the PCR-based approach (Figure 6A). Six plasma DNA samples were sequenced three times: twice by capture-based approach and once by PCR-based approach. The number of mtDNA mutation sites with more than 1% heteroplasmy level is depicted in Figure 6B. Moreover, we found that the consistency of the mtDNA mutations between two capture-based sequencing datasets was significantly higher than between each capture-based sequencing dataset and PCR-based sequencing dataset (Figure 6C). With the increase in mtDNA site sequencing depth, the mtDNA mutation consistency between capture-based and PCR-based sequencing data was elevated gradually (Figure 6D). To get more accurate mtDNA mutations, our results suggest a higher sequencing depth for the PCR-based sequencing approach compared with capture-based sequencing.

Additionally, to test the robust application of our innovative bioinformatics pipeline, we evaluated the consistency of mtDNA mutation profiling between the Illumina and MGISEQ-2000 platform (BGI) sequencing platforms, which are used most widely for NGS. The mtDNA sequencing data of 10 fresh tissue samples from two platforms were analyzed. As shown in Figure 6E, all samples exhibited good consistency of mtDNA mutation numbers. We further compared the heteroplasmy level of mtDNA mutations. Our results revealed a remarkable correlation between two sequencing platforms ($r = 0.9974$, $p < 0.001$), indicating widespread applicability of our innovative data analysis pipeline.

DISCUSSION

Here, based on systematic evaluation of key analysis procedures, we established an innovative data analysis strategy for improving the accuracy of NGS-based mtDNA mutation detection, which is mainly integrated with a trimming procedure, a mapping strategy with the rCRS and hg19 as reference genomes, and a newly developed filtering approach. For the first time, we report several key findings regarding application of the data analysis pipeline to three different sample types. First, we demonstrated that the trimming procedure was essential for improving mtDNA mapping performance in plasma but not tumor tissue samples. Second, our systematic analysis of mtDNA reference genomes clearly showed the great effect of NUMTs in plasma samples and provided an optimal choice for reference genomes in different sample types. Third, we demonstrated that a setting of 3 mismatches was most suitable for mtDNA mutation calling. More importantly, we found a negative logarithmic relationship between mtDNA site sequencing depth and minimum detectable mutation frequency and innovatively built an efficient filtering strategy to increase the accuracy of mutation detection. All of these efforts greatly contribute to establishment of an innovative and versatile bioinformatics pipeline for accurate mtDNA mutation detection, laying a foundation for translational application of mitochondrial mutations in clinical practice.

Quality control and preprocessing of sequencing data are critical to obtain highly accurate mtDNA mutations in downstream data analysis, especially important for detecting low-MAF mutations. The trimming procedure, which refers to elimination of adaptors and poor-quality bases of the sequencing reads, is an initial step of the analysis that has been applied heterogeneously in previous mtDNA mutation analyses.[18] Whether the trimming procedure is indispensable for mtDNA mutation analysis remains to be confirmed, especially for different sample types. In the present study, we demonstrate that the trimming procedure is essential for improving mtDNA mapping performance in plasma but not tumor tissue samples.

Because the true origin of homologous reads is difficult to discriminate, NUMTs and mtDNA cross-mapping occurs, leading to detection of false positive (NUMT reads aligning to chromosome Mitochondrion [chrM]) or false negative (mtDNA reads aligning to NUMTs loci) mtDNA mutations. To address this concerning source of error, Li et al.[19] have created a database of NUMTs containing mismatches from the mitochondrial genome, which may appear as false heteroplasmies when aligned to chrM. Complementing this approach, two main mapping strategies are used commonly at present. The first is mapping to the human reference mtDNA sequence rCRS, and the second is mapping to the rCRS and hg19, which contributes to removal of NUMTs during analysis. Therefore, the accuracy of mtDNA mutation calling can be greatly affected by mapping strategies with different selection of reference genomes and mismatch numbers because of sequence similarities between NUMTs and mtDNA (a known source of confounding in mtDNA NGS studies).[20] However, a few studies are focusing on the applicability of the two

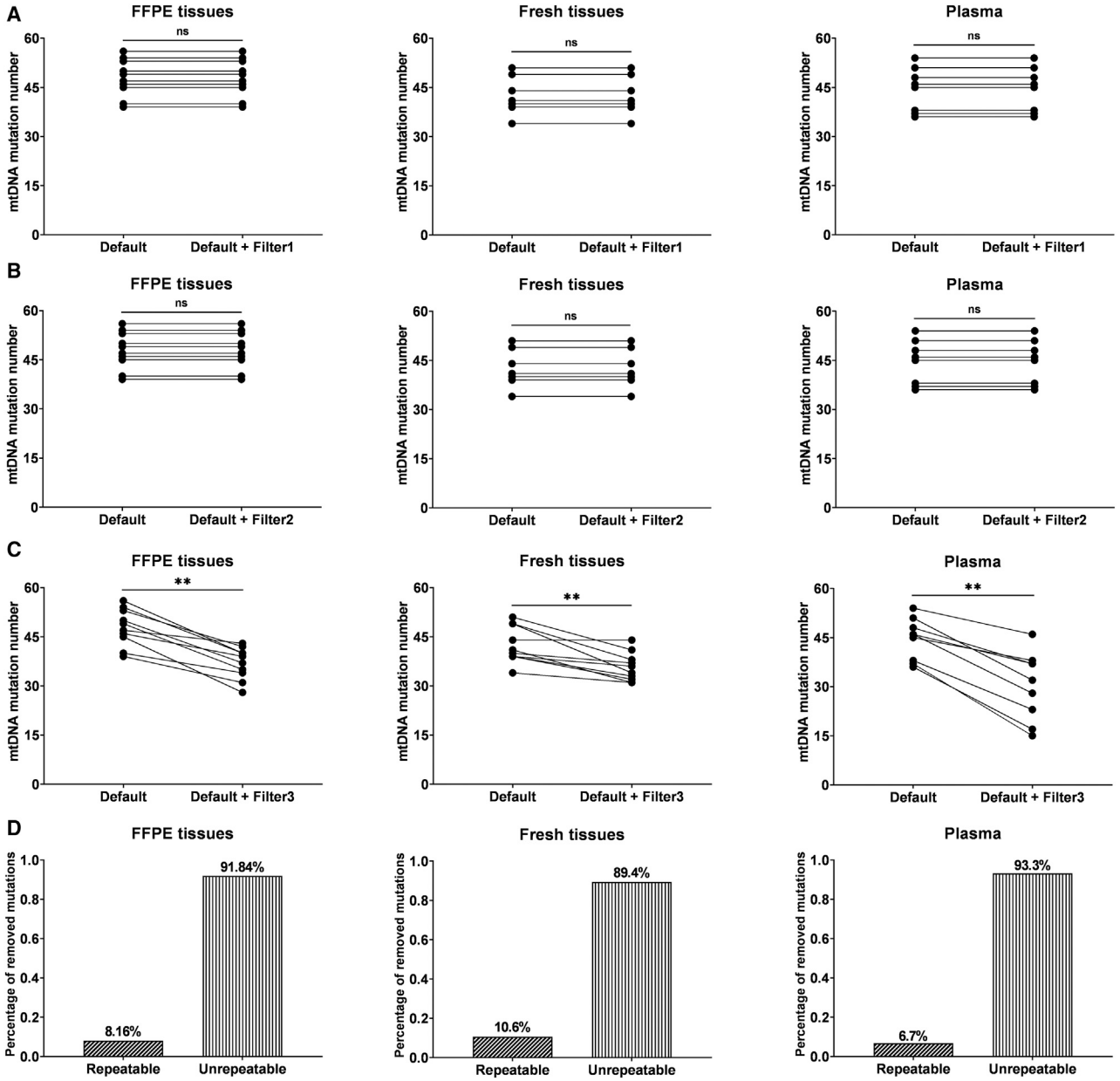


Figure 5. Assessment of different mtDNA mutation filtering strategies

(A) Comparison of mtDNA mutation calling with or without filter 1 in FFPE tissue, fresh tissue, and plasma samples. Filter 1 removes C > A/G > T mutations with a low MAF (1%), which is known to arise from artificial guanine oxidation during sequencing library preparation. (B) Comparison of mtDNA mutation calling with or without filter 2 in FFPE tissue, fresh tissue, and plasma samples. Filter 2 removes mtDNA mutations where mutation rate and mutation base quality do not pass a binominal test ($p > 0.0001$). (C) Comparison of mtDNA mutation calling with or without filter 3 in FFPE tissue, fresh tissue, and plasma samples. Filter 3 is the novel filter that removes mtDNA mutations when the MAF is smaller than the site-specific threshold determined by site sequencing depth. (D) Percentages of the repeatable and unrepeatable mtDNA mutations that were removed by filter 3 in FFPE tissue, fresh tissue, and plasma samples. (A–D) 10 FFPE tissues, 10 fresh tissues, and 10 plasma samples were used. $^{**}p < 0.01$.

mapping strategies under different circumstances. In the present study, we performed a systematic performance comparison of mtDNA mutation calling pipelines with different mapping strategies (rCRS alone or combined rCRS-hg19) or mismatch settings (changed from 1 to 4) in three different sample types. Our data indicate that the two mapping strategies can be used effectively in FFPE and fresh tissue samples, although only mapping to the rCRS may introduce a very small amount of NUMTs. However, the mapping strategy with hg19 and the rCRS as reference genomes is strongly suggested for mtDNA mutation detection in plasma samples because of the

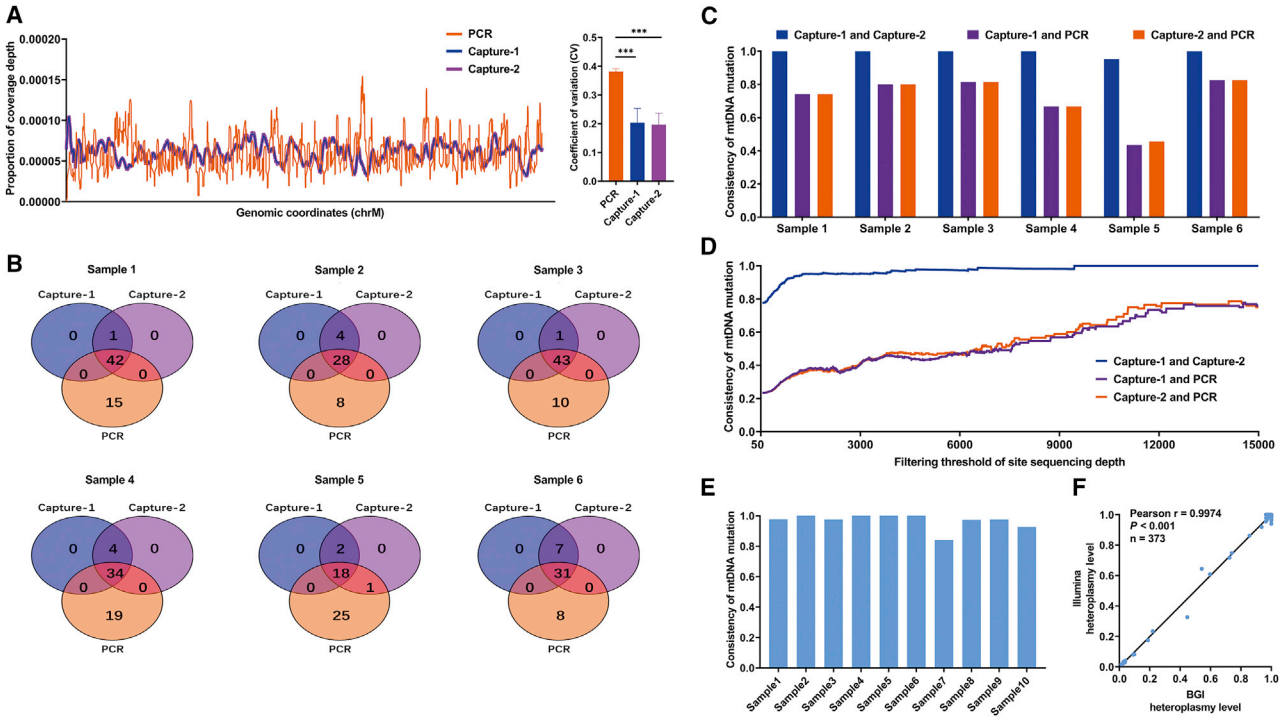


Figure 6. Performance comparison of PCR-based and capture-based enrichment strategies or of two sequencing platforms

(A) Distribution of coverage depth across the mitochondrial genome (left panel) and comparison of the coefficient of variation (CV, right panel) between capture-based and PCR-based mtDNA sequencing. (B) Venn diagram of mtDNA mutations detected by capture-based and PCR-based mtDNA sequencing in six plasma samples. (C) Consistency of mtDNA mutations detected by capture-based and PCR-based mtDNA sequencing in six plasma samples. (D) Consistency of mtDNA mutations between capture-based and PCR-based mtDNA sequencing at different sequencing depth thresholds. (E) Consistency of mtDNA mutations between the Illumina and BGI sequencing platforms in 10 fresh tissue samples. (F) Consistency of heteroplasmy levels of mtDNA mutations detected by the Illumina and BGI sequencing platforms in 10 fresh tissue samples. (A–F) 6 plasma samples were used for capture-based and PCR-based experiments. ***$p < 0.001$.

extreme abundance of NUMTs. Furthermore, our study strongly suggests a setting of 3 mismatches for mtDNA mutation calling for all three sample types.

The detectable level of mtDNA heteroplasmy is heavily dependent on the depth of NGS coverage. Furthermore, with the sequencing depth becoming deeper, the accuracy of mtDNA mutation detection is increasing. Recent studies have detected somatic mtDNA mutations at a very low heteroplasmic level (MAF > 1%), with an NGS depth for the mitochondrial genome of 9,959×.[14] However, as we lower the detectable threshold of heteroplasmy, it becomes increasingly difficult to distinguish between ultra-low MAF mutations and sequencing errors. Thus, addressing the technical question of correctly standardizing the sequencing depth to obtain confident and reproducible detection of low-MAF mutations is of great importance. However, which sequencing depth can provide sufficient information to detect low frequency mtDNA mutations remains to be investigated. Because there are no consensus criteria for determining an mtDNA heteroplasmy threshold, it has been very common to arbitrarily select a constant heteroplasmy level (mainly 2% or 5%) in previous NGS analyses for mtDNA mutations.[12,21] Here we revealed the relationship of site sequencing depth and the minimum mutation threshold by establishing logarithmic functions in different sample types, which contributes to discrimination of false positive and false negative mutations with an ultra-low heteroplasmy level (up to 0.5%). By providing a standardized option of proper mutation frequency threshold, this can greatly increase the sensitivity and accuracy of mtDNA mutation detection during NGS data analysis.

One unique aspect of our study is integrative analysis of mtDNA mutation calling based on assessment of key procedures. We found that a filtering strategy based on site sequencing depth greatly contributes to improving the detection consistency of repeated data. Previous studies of mtDNA mutation calling have also applied several filtering conditions, such as at least 3 reads per mutation allele and strand bias correction,[3] to reduce false positive mtDNA mutations. Several studies have also focused on developing filter-based methods to remove oxidation-mediated mutations during DNA shearing or artifacts arising from Illumina sequence errors.[15,17] However, those filtering strategies may not be necessary for mutation analysis in ultra-deep sequencing (whose sequencing depth is deep enough). For example, the minimum number of mutant reads must be greater than 3 as long as the site depth is greater than

600× when detecting mutations at a threshold greater than 1%. Moreover, we also observed that filtering strategies considering DNA oxidation and binominal distribution had no significant effect in our analysis system, which may at least be partially due to strict removal of low-quality bases during the trimming procedure. In addition, the negative logarithmic function established in our study is based on capture-based sequencing data, whose applicability to all sequencing data may be indeterminate but can be applied to provide a reference for the range of sequencing depths and heteroplasmy levels in NGS data.

Capture-based and PCR-based methods are commonly used for mtDNA enrichment before sequencing.[14] Therefore, selecting the right enrichment approach and sequencing platform is of great importance to accurately identify mtDNA mutations, especially those with an ultra-low heteroplasmy level. A previous study compared the detection performance of different enrichment strategies in fresh tissue samples and illustrated that DNA quality was a great challenge for the PCR-based approach, which would not work with highly fragmented DNA samples, such as FFPE tissues and plasma.[22] In addition, PCR-based sequencing may further suffer from higher amplification bias and amplification-related errors, requiring a higher sequencing depth to achieve better consistency. In the present study, we explored different enrichment approaches and sequencing platforms for three different sample types. Our results suggest a higher sequencing depth for the PCR-based sequencing approach compared with capture-based sequencing.

Based on systematic evaluation of key analysis procedures, we established an innovative data analysis pipeline for different tumor sample types, which is of great significance for improving the accuracy of mtDNA mutation detection. These efforts lay a foundation for broader biomedical applicability for accurate investigation of the mitochondrial genome in cancer cells.

MATERIALS AND METHODS

Sample collection

In total, 50 FFPE tissue samples, 50 fresh tissue samples, and 50 plasma samples were collected from 150 individuals with liver cancer in Xijing Hospital, Fourth Military Medical University (FMMU) in Xi'an, China. This study was approved by the Ethics Committee of FMMU, and written consent was obtained from each individual.

DNA extraction

Genomic DNA was extracted from fresh tissue, FFPE tissue, and plasma using the ENZA Tissue DNA Kit (Omega), QIAamp DNA FFPE Kit (QIAGEN), and QIAamp Circulating Nucleic Acid Kit (QIAGEN) according to the manufacturers' protocols. DNA quality and concentration were assessed using a 2100 Bioanalyzer (Agilent Technologies) and Qubit (Invitrogen).

Library construction and mtDNA enrichment

The WGS library for the Illumina platform was constructed as described previously.[23] In brief, 1 μg genomic DNA from FFPE and fresh tissue was sonicated randomly by focused ultrasonicator (Scientz98, Ningbo, China) to obtain fragments mainly distributed between 300 and 500 bp in length. DNA fragments were end repaired, ligated with sequencing adapters, and slightly PCR amplified (9 cycles). For plasma samples, 20 ng genomic DNA was used to construct the sequencing library using the NEB Ultra v.2 Kit (New England Biolabs). Then WGS libraries were mixed with homemade biotinylated mtDNA capture probes for hybridization. Furthermore, to examine the effect of different enrichment strategies on mtDNA mutation detection, we also constructed a PCR-based mtDNA enrichment library using the QIAseq Targeted DNA Human Mitochondrial Panel (QIAGEN) for 6 plasma samples, following the manufacturer's protocol.

Sequencing platforms and NGS

The capture- and PCR-based mtDNA libraries were sequenced on an Illumina XTen platform using paired-end runs (2 × 150 cycles). To further evaluate the suitability of the optimized mtDNA bioinformatics pipeline for different sequencing platforms, 10 capture-based mtDNA libraries from fresh tissue were also sequenced on the BGI using paired-end runs (2 × 100 cycles). A summary of the mtDNA sequencing data used in this study is shown in Table S1.

Bioinformatics pipeline for mtDNA mutation calling

We systematically evaluated the analysis pipeline for mtDNA deep-sequencing data. Briefly, raw mtDNA sequencing data first encountered two options: trimming or no trimming for quality control. The mtDNA reads were then mapped to the rCRS or combined rCRS-hg19 using About Burrow-Wheeler Aligner (BWA) software. After sorting and removing duplicated reads with Picard, the Genome Analysis Toolkit 4 (GATK4) was used for local realignment. Finally, we applied a series of filtering conditions (removing false positive mutations) to detect mtDNA mutations and analyze heteroplasmy levels.

Trimming procedure

The FASTQ preprocessor fastp (version 0.20.0)[24] was used for trimming mtDNA sequencing data with three parameters. First, all sequencing adaptors were removed. Second, a sliding window (4 bp in length) approach was used to scan reads from front (5′) to tail (3′). When the average base quality in the window was below Q30, these bases and downstream parts were dropped. Third, reads with a length below 50 bp were discarded to avoid ambiguous mapping of short reads.

Mapping strategies and mismatch selection

Two mapping strategies were compared. The first strategy was to only map sequencing reads to the rCRS (rCRS alone). The second strategy was to map sequencing reads to the rCRS and hg19 reference (combined rCRS-hg19) but keep only reads uniquely mapped to the rCRS. In addition, sequencing reads with a mapping quality below Q20 were removed from subsequent analysis. Different mismatch filters ranging from 1–4 were evaluated in mtDNA mutation calling. We used two error-evaluating parameters (AFP and AFN mutants) to

identify the optimum mismatch number selection. AFP mutations were defined as those only detected in one mismatch group. AFN mutations were defined as those absent in this group but present in at least two other mismatch groups.

Identification of minimum detectable mutation frequency based on site sequencing depth

To identify the minimum detectable mutation threshold, genomic DNA from different sample types was used for independent library construction and sequencing. These repeated sequencing datasets enabled us to experimentally analyze the consistency of mtDNA mutation calling. Here, consistency between two repeated experiments was defined as

$$\text{Consistency} = \frac{A \cap B}{A \cup B},$$

where A is the collection of mtDNA mutations in experiment 1, and B is the collection of mtDNA mutations in experiment 2. Setting the consistency level at 95%, a dynamic minimum detectable mutation threshold was identified for sites with different sequencing depths. The relationship between site sequencing depth and the minimum detectable mutation threshold was explored by logistic regression analysis using R software.

Development of novel filtering criteria for mtDNA mutation calling

For each site, we first counted the respective read numbers of the major and minor alleles and calculated site-specific MAF. Then mtDNA mutations were initially called using the default settings, as we described previously:[14,21,23] (1) at least 3 reads on each strand have the mutation site; (2) minimum MAF cutoff 1%; and (3) remove heterogeneity sites in rCRS repeat regions (66–71, 303–311, 514–523, 12,418–12,425, and 16,184–16,193).

In addition to the default settings, we further explored the effect of three extra filtering strategies on mtDNA mutation calling. Filter 1 removes C > A/G > T mutations with a low MAF (1%) and strong sequence context bias (at CpCpN > CpApN, most frequently CpCpG > CpApG), which is known to arise from artificial guanine oxidation during sequencing library preparation.[14,15] Filter 2 removes mtDNA mutations when the mutant rate and mutant base quality do not pass a binominal test ($p > 0.0001$).[16,17] Filter 3 removes mtDNA mutations when the MAF is smaller than the site-specific threshold determined by site sequencing depth.

Statistical analysis

GraphPad Prism 7.0 (GraphPad, USA) was used for statistical analysis. Mann-Whitney *U* test was used to compare the difference between two groups. All p values were two tailed and reported using a significance level of 0.05.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2020.11.002.

ACKNOWLEDGMENTS

The authors thank Deyang Li, Xiaohong Du, and Xiangxu Wang in the Department of Physiology and Pathophysiology for ongoing support and discussions. This work was supported by the National Natural Science Foundation of China, China (grant 81830070) and the Autonomous Project of the State Key Laboratory of Cancer Biology, China (grant CBSKL2019ZZ06).

AUTHOR CONTRIBUTIONS

S.G. and K.Z. carried out sample collection, performed data analysis, and drafted the manuscript. Q.Y. and L.S. participated in the bioinformatics analyses. Y.L. and X.J. performed the laboratory experiments. X. Gu and X. Guo participated in design of the study and performed draft revision. J.X. conceived the study, participated in its design and coordination, and helped to revise the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457–465.
2. Schon, E.A., DiMauro, S., and Hirano, M. (2012). Human mitochondrial DNA: roles of inherited and somatic mutations. Nat. Rev. Genet. *13*, 878–890.
3. He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz, L.A., Jr., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature *464*, 610–614.
4. Mardis, E.R. (2009). New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med. *1*, 40.
5. Ye, F., Samuels, D.C., Clark, T., and Guo, Y. (2014). High-throughput sequencing in mitochondrial DNA research. Mitochondrion *17*, 157–163.
6. Ju, Y.S., Alexandrov, L.B., Gerstung, M., Martincorena, I., Nik-Zainal, S., Ramakrishna, M., Davies, H.R., Papaemmanuil, E., Gundem, G., Shlien, A., et al.; ICGC Breast Cancer Group; ICGC Chronic Myeloid Disorders Group; ICGC Prostate Cancer Group (2014). Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. eLife *3*, e02935.
7. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.
8. Wong, L.J. (2013). Challenges of bringing next generation sequencing technologies to clinical molecular diagnostic laboratories. Neurotherapeutics *10*, 262–272.
9. González, M.D.M., Ramos, A., Aluja, M.P., and Santos, C. (2020). Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing. Mitochondrion *50*, 88–93.
10. Li, M., Schröder, R., Ni, S., Madea, B., and Stoneking, M. (2015). Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc. Natl. Acad. Sci. USA *112*, 2491–2496.
11. Stewart, J.B., and Chinnery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat. Rev. Genet. *16*, 530–542.
12. Hodgkinson, A., Idaghdour, Y., Gbeha, E., Grenier, J.C., Hip-Ki, E., Bruat, V., Goulet, J.P., de Malliard, T., and Awadalla, P. (2014). High-resolution genomic analysis of human mitochondrial RNA sequence variation. Science *344*, 413–415.
13. Hazkani-Covo, E., Zeller, R.M., and Martin, W. (2010). Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. *6*, e1000834.

14. Liu, Y., Guo, S., Yin, C., Guo, X., Liu, M., Yuan, Z., Zhao, Z., Jia, Y., and Xing, J. (2020). Optimized PCR-Based Enrichment Improves Coverage Uniformity and Mutation Detection in Mitochondrial DNA Next-Generation Sequencing. J. Mol. Diagn. *22*, 503–512.

15. Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67.

16. Morelli, M.J., Wright, C.F., Knowles, N.J., Juleff, N., Paton, D.J., King, D.P., and Haydon, D.T. (2013). Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet. Res. (Faisalabad) *44*, 12.

17. Campo, D.S., Nayak, V., Srinivasamoorthy, G., and Khudyakov, Y. (2019). Entropy of mitochondrial DNA circulating in blood is associated with hepatocellular carcinoma. BMC Med. Genomics *12* (*Suppl 4*), 74.

18. Williams, C.R., Baccarella, A., Parrish, J.Z., and Kim, C.C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics *17*, 103.

19. Li, M., Schroeder, R., Ko, A., and Stoneking, M. (2012). Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. Nucleic Acids Res. *40*, e137.

20. Maude, H., Davidson, M., Charitakis, N., Diaz, L., Bowers, W.H.T., Gradovich, E., Andrew, T., and Huntley, D. (2019). NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele. Front. Cell Dev. Biol. *7*, 201.

21. Li, X., Guo, X., Li, D., Du, X., Yin, C., Chen, C., Fang, W., Bian, Z., Zhang, J., Li, B., et al. (2018). Multi-regional sequencing reveals intratumor heterogeneity and positive selection of somatic mtDNA mutations in hepatocellular carcinoma and colorectal cancer. Int. J. Cancer *143*, 1143–1152.

22. Kaneva, K., Merkurjev, D., Ostrow, D., Ryutov, A., Triska, P., Stachelek, K., Cobrinik, D., Biegel, J.A., and Gai, X. (2020). Detection of mitochondrial DNA variants at low level heteroplasmy in pediatric CNS and extra-CNS solid tumors with three different enrichment methods. Mitochondrion *51*, 97–103.

23. Yin, C., Li, D.Y., Guo, X., Cao, H.Y., Chen, Y.B., Zhou, F., Ge, N.J., Liu, Y., Guo, S.S., Zhao, Z., et al. (2019). NGS-based profiling reveals a critical contributing role of somatic D-loop mtDNA mutations in HBV-related hepatocarcinogenesis. Ann. Oncol. *30*, 953–962.

24. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.