# An Innovative Data Analysis Strategy For Accurate NGS Detection of Tumor mtDNA Mutations

**Shanshan Guo**
Fourth Military Medical University
**Kaixiang Zhou**
Fourth Military Medical University
**Qing Yuan**
Fourth Military Medical University
**Liping Su**
Fourth Military Medical University
**Xiaoying Ji**
Fourth Military Medical University
**Xiwen Gu**
Fourth Military Medical University
**Xu Guo**
Fourth Military Medical University
**Jinliang Xing** ( xingjinliang@163.com )
Fourth Military Medical University

**Research**

**Keywords:** mtDNA mutation, Next generation sequencing, Bioinformatics

**Posted Date:** July 23rd, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-46773/v1

# Abstract

**Background:** Next generation sequencing (NGS) technology has been commonly applied to detect mtDNA mutations, which are reported to be strongly associated with cancers. However, several key challenges still exist in the bioinformatic analysis of mtDNA sequencing data, which greatly affect the detection accuracy of mtDNA mutations.

**Methods:** Here, we systematically evaluated several key analysis procedures, including trimming, mapping and filtering, in mtDNA mutation detection of FFPE tissues, fresh tissues and plasma samples from cancer patients. Furthermore, the innovative bioinformatics pipeline integrating a newly-developed filtering algorithm was established.

**Results:** We found that trimming procedure was essential for improving mtDNA mapping performance in plasma but not tissue samples. Mapping with rCRS-hg19 reference was strongly suggested for mtDNA mutation detection in plasma samples due to the extreme abundance of NUMTs. In addition, our results showed that the setting of 3 mismatches was most appropriate for mtDNA mutation calling. More importantly, we revealed the presence of a negative logarithmic relationship between mtDNA site sequencing depth and minimum detectable mutation frequency and thus build up 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 PCR-based than capture-based enrichment strategy.

**Conclusions:** Collectively, we established an innovative data analysis strategy, which is of great significance for improving the accuracy of mtDNA mutation detection for different types of tumor samples.


# Background

Human mitochondria possess their own genome, which is a 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 a plenty of sequence variants, which are commonly observed either as germline or somatic mutations (2). Numerous studies have demonstrated that heteroplasmy, termed as the presence of a mixed mtDNA mutation genotypes within a cell, is strongly associated with lots of 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 better understanding mitochondrial biology and cancers.

Traditional techniques, such as Sanger sequencing and high-resolution melt analysis, have been used for mtDNA mutation detection, which still face the disadvantages of low throughput and sensitivity. Fortunately, 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 the 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 is low-throughput and not cost-effective (6). Another is to first enrich mtDNA from total genomic DNA and then detect by NGS, mainly including PCR-based and capture-based enrichment strategies (7). They can not only achieve the detection of low minor allele frequency (MAF), but also permit customized sequencing strategies, allowing segment-specified or whole-genome sequencing of mtDNA. Combined with NGS technology, such strategies have been proved to be powerful and efficient when depicting spectrum of mtDNA somatic mutation in many types of cancers (3).

Although NGS has been extensively applied to effectively detect mtDNA mutations, there are still several big challenges existing in data analysis of mtDNA sequencing (8). Among them, most critical consideration is to decrease the number of false-positive and false-negative mutations, which are greatly affected by mtDNA sequencing depth, especially for those with 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 of previous studies arbitrarily selected 5%, 2% or 1% as the minimal heteroplasmy level to filter mtDNA mutations (10–12). In addition, owing to the presence of homologous sequences within the nuclear genome, referred as “nuclear sequences of mitochondrial origin” (NUMTs), 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, high sequencing depth (usually more than thousands) is becoming practical for tumor mtDNA mutation detection. It is urgently needed to develop an innovative data analysis strategy for more accurate NGS detection of mtDNA mutations, especially for those with 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 FFPE tissues, fresh tissues and plasma samples from cancer patients. Finally, an innovative bioinformatics pipeline integrating a newly-developed filtering algorithm was established. Furthermore, the application of our analysis strategy in different sample types has also been evaluated, providing a flexible selection of key analysis procedures.

# Materials And Methods

### Sample collection

A total of 50 FFPE (Formalin-Fixed and Paraffin-Embedded) tissue samples, 50 fresh tissue samples and 50 plasma samples were collected from liver cancer patients in Xijing Hospital, Fourth Military Medical University (FMMU) in Xi’an, China. This study was approved by the Ethical Committees of FMMU and written consent was obtained from each patient.

### DNA extraction

Genomic DNA was extracted from fresh tissue, FFPE tissue and plasma using ENZA Tissue DNA Kit (Omega), QIAamp DNA FFPE Kit (Qiagen) and QIAamp Circulating Nucleic Acid Kit (Qiagen) according to

manufacturer's protocols, respectively. DNA quality and concentration were assessed using 2100 Bioanalyzer (Agilent) and Qubit (Invitrogen).

### Library construction and mtDNA enrichment

Whole genome sequencing (WGS) library for Illumina platform was constructed as previously described (14). In brief, 1μg genomic DNA from FFPE and fresh tissues were randomly sonicated 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 were used to construct the sequencing library using NEB ultra v2 kit (NEB). Then, WGS libraries were mixed with home-made biotinylated mtDNA capture probes for hybridization. Furthermore, to examine the effect of different enrichment strategy on mtDNA mutation detection, we also constructed PCR-based mtDNA enrichment library using QIAseq Targeted DNA Human Mitochondrial Panel (Qiagen) for 6 plasma samples following the manufacture's protocol.

### Sequencing platforms and next generation sequencing

The capture- and PCR-based mtDNA libraries were sequenced on Illumina XTen platform using paired-end runs (2 x 150 cycles). To further evaluate the suitability of the optimized mtDNA bioinformatics pipeline in different sequencing platforms, 10 capture-based mtDNA libraries from fresh tissue were also sequenced on MGISEQ-2000 platform (BGI) using paired-end runs (2 x 100 cycles). A summarization of the mtDNA sequencing data used in this study was 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 encounter two options, trimming or without trimming for quality control. The mtDNA reads were then mapped to either rCRS (revised Cambridge Reference Sequence) or combined rCRS and human genome reference (hg19) using 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) (15) was used for trimming of mtDNA sequencing data with three parameters: first, all sequencing adaptors were initially removed; second, a sliding window (4 bp in length) approach was used to scan reads from front (5') to tail (3'). If average base quality in the window was below Q30, these bases and downstream part were dropped; third, the reads with 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 rCRS reference (rCRS-alone); the second strategy was to map sequencing reads to both rCRS and hg19 reference (combined rCRS-hg19), but keep only the reads uniquely mapped to rCRS. In addition, sequencing reads with mapping quality below Q20 were removed from subsequent analysis. Different mismatch filter ranging from 1-4 were evaluated in mtDNA mutation calling. We used two error-evaluating parameters (assumed false positive and assumed false negative mutants) to identify the optimum mismatch number selection. 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 absent in this group but presented 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 were 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, 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 depth. Moreover, the relationship between site sequencing depth and the minimum detectable mutation threshold were 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 reads number of the major and minor allele and calculated site-specific minor allele frequency (MAF). Then, mtDNA mutations were initially called using the default settings as we previously described (14, 16, 17): 1) at least 3 reads on each strand have the mutation site; 2) minimum MAF cutoff 1%; 3) remove heterogeneity sites in rCRS repeat regions (66-71, 303-311, 514-523, 12418-12425, 16184-16193).

In addition to the default settings, we further explored the impact of three extra filtering strategies on mtDNA mutation calling: filter 1 removes C > A/G > T mutations with 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 (18, 19); filter 2 removes mtDNA mutations if the mutant-rate and mutant base-quality does not pass binominal test ($P > 0.0001$) (20, 21); filter 3 removes mtDNA mutations if the MAF is smaller than the site-specific threshold determined by site sequencing depth.

### Statistical analysis

Graphpad Prism 7.0 (GraphPad Software, 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.

## Results

### Effect of trimming procedure on mapped mtDNA sequencing data

In NGS data analysis, trimming procedure is commonly used to remove the adaptor sequences and bases with low quality from sequencing reads. Therefore, we first evaluated the effect of trimming procedure on mtDNA mapping of sequencing reads in three types of samples, including FFPE tissues, fresh tissues and plasma. As shown in Fig.1A and B, the trimming procedure had no significant influence on both the mtDNA mapping rate of sequencing reads and average sequencing depth in FFPE and fresh tissue samples. In contrast, the mtDNA mapping rate and average sequencing depth were significantly increased in plasma samples after trimming (Fig. 1A and B). Furthermore, our data indicated the very consistent size distribution of mtDNA insert fragments in FFPE and fresh tissue samples (Fig. 1C). However, the mapped mtDNA sequencing reads were significantly increased in plasma samples after trimming (Fig. 1C). These results indicate that trimming procedure is essential for improving the mtDNA mappability in plasma but not tissue samples.

### Evaluation of mtDNA mapping strategy in mtDNA mutation calling

Next, we evaluated the application of two commonly used mapping strategies in three different sample types. As shown in Fig. 2A, when compared with the first mapping strategy (rCRS-alone), the second mapping strategy (combined rCRS-hg19) clearly removed the sequencing reads mapped to both hg19 and rCRS, leading to 11 more remarkable peaks. Among them, the most significant peak occurred around mtDNA site 8500, with 45% - 80% removed reads in three different sample types. Compared to tissue samples, plasma DNA exhibited a clearly different distribution pattern of removed reads. Then, the mtDNA mutations detected by the two mapping strategies were depicted in Fig. 2B, with mutation site and heteroplasmy level (MAF≥1%). The venn diagram showed that all mtDNA mutations detected by the second mapping strategy were included in those detected by the first one, while 1, 5 and 1256 of mtDNA mutations only detected by mapping to rCRS were observed in FFPE tissue, fresh tissue and plasma samples, respectively (Fig. 2C). The site and heteroplasmy level of mtDNA mutations only detected by mapping to rCRS were shown in Fig. 2D. To explore the potential reason explaining extremely high number of mtDNA mutations only detected by mapping to rCRS in plasma samples, we further analyzed the mtDNA copy number and percentage of mtDNA sequencing reads mapped to both hg19 and rCRS in three sample types. Our results showed that the mtDNA copy number in plasma was significantly lower than tissues (Fig. S1), and the percentage of mtDNA sequencing reads mapped to both hg19 and rCRS was significantly higher in plasma (17.18%) compared to FFPE and fresh tissues (3.46% and 4.03%, respectively) (Fig. S2). To investigate the NUMTs-derived possibility of those mutations, we determined

the percentage of mutant reads mapped to rCRS only or both hg19 and rCRS. As shown in Fig. 2E, 93.33%, 92.86% and 91.91% of the mutant reads were mapped to both hg19 and rCRS in three sample types, respectively, while 6.67%, 7.14% and 9.09% were only mapped to 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 rCRS in three sample types, respectively (Fig. 2F). These data suggest that these mutations may be introduced by NUMTs. Collectively, when mtDNA mutations are detected in tissue samples, both the two 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 due to the extreme abundance of NUMTs.

### Effect of mismatch number selection on mtDNA mutation calling

Then, we investigated the effect of mismatch number selection in the second mapping strategy on mtDNA mutation calling by using the repeated sequencing data in three sample types. As shown in Fig. 3A, the total repeated mtDNA mutation numbers in two repeat experiments of 10 samples gradually increased when the mismatch number changed from 1 to 4 in three sample types. We then found that the average number of repeatable mutations in the group with 3 or 4 mismatches was significantly higher than that 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 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 of AFP mutation number among the groups with 1, 2, or 3 mismatches in three sample types, while it was significantly increased in the group with 4 mismatches when compared to the groups with 1, 2, or 3 mismatches (Fig. 3B). In contrast, the AFN mutation number in groups with 1 or 2 mismatches were significantly higher than that in the groups with 3 or 4 mismatches among three sample types (Fig. 3C). Collectively, these results demonstrate the impacts to some extent of mismatch number selection on mtDNA mutation detection, and strongly suggests the setting of 3 mismatches for mtDNA mutation calling.

### Relationship between the 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 three sample types. As shown in Fig. 4A, a very consistent trend was observed in all three sample types, indicating a faster rise of the consistency when the site sequencing depth was relatively low. To achieve the consistency of 90%, the site sequencing depth of greater than 2700X, 2300X, and 3200X was needed in FFPE tissues, fresh tissues and plasma, respectively, when the heteroplasmy level was set above 0.5%. In comparison, the consistency of mtDNA mutations was notably decreased 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 the heteroplasmy level of 0.1%. Furthermore, the negative logarithmic relationship was established between mtDNA site sequencing depth and minimum detectable mutation frequency when the consistency was set at 90% (Fig. 4B). When the minimum detectable mutation frequency was 1% and 0.5%, the site sequencing depth was required to be higher than 1000X and 4000X, 1200X and 3600X, 1700X and 4700X in FFPE tissue, fresh tissue and plasma, respectively.

Commonly, the average sequencing depth was calculated and provided based on mtDNA sequencing data. Considering the accuracy of mtDNA mutation calling was actually affected by site sequencing depth, we thus determined the relationship between them. As shown in Fig. 4C, with the increasing average sequencing depth, the percentage of mtDNA sites with site depth greater than average sequencing depth were gradually increased. When the average sequencing depth was greater than 8000X, there were 80% of sites with site depth higher than the average.

## Assessment of different mtDNA mutation filtering strategies

To reduce false positive mutations, several filtering strategies has commonly been applied (18). First, we assessed the effect of two filtering strategies previously reported 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 (18, 19). As shown in Fig. 5A, no difference of mtDNA mutation numbers was observed between two groups filtered with and without filter 1 in FFPE tissue, fresh tissue and plasma. Furthermore, second filtering strategy (filter 2) is to remove mtDNA mutations where the quality of mutant bases does not fit well with binominal distribution ($P > 0.0001$) (20, 21). Those may be false positive mutations introduced by sequencing errors. Very similarly with filter 1, filter 2 exhibited no significant effect on mtDNA mutation numbers in all three sample types (Fig. 5B). Considering the great influence of site sequencing depth on mtDNA mutation detection, we built up a novel filtering strategy (filter 3) based on negative logarithmic functions presented in Fig. 4B to remove mtDNA mutations with heteroplasmy level lower than minimum detectable mutation frequency. As shown in Fig 5C, the number of mtDNA mutations was significantly decreased after filtering in all three sample types. Moreover, the repeated mtDNA sequencing data were used to analyze whether those filtered mutations were repeatable or not. We found that 91.84%, 89.4%, and 93.3% of these filtered mutations were unrepeatable in FFPE tissue, fresh tissue and plasma, respectively. In comparison, only 9.4%, 5.7% and 15% of the unaffected mtDNA mutations was unrepeatable (Fig. S3). These findings indicate that both filter 1 and filter 2 are not necessary in our analysis pipeline, whereas filter 3 greatly contribute to improve the detection accuracy.

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

To reduce sequencing cost, both 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 the coverage across the whole

mitochondrial genome and the coefficient of variation (CV) significantly decreased when compared to PCR-based approach (Fig. 6A). Then, six plasma DNA samples were sequenced three times, twice by capture-based approach and one time by PCR-based approach. The number of mtDNA mutation sites with more than 1% heteroplasmy level was depicted in Fig. 6B. Moreover, we found that the consistency of the mtDNA mutations between two capture-based sequencing data was significantly higher than those between each capture-based sequencing data and PCR-based sequencing data (Fig. 6C). With the increasing of mtDNA site sequencing depth, the mtDNA mutation consistency between capture-based and PCR-based sequencing data was gradually elevated (Fig. 6D). To get more accurate mtDNA mutations, our results suggest higher sequencing depth for PCR-based sequencing approach when compared with capture-based sequencing.

Additionally, to test the robust application of our innovative bioinformatics pipeline, we evaluated the consistence of mtDNA mutation profiling between Illumina and BGI sequencing platforms, which were most widely used for next-generation sequencing. The mtDNA sequencing data of 10 fresh tissue samples from two platforms were analyzed. As shown in Fig. 6E, all samples exhibited the 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 ($r = 0.9974$, $P < 0.01$), indicating the widespread applicability of our innovative data analysis pipeline.

# Discussion

Here, based on systematical 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 trimming procedure, mapping strategy with rCRS and hg19 as reference genomes and a new-developed filtering approach. Therefore, we for the first time reported several key findings about the application of data analysis pipeline in three different sample types. First, we demonstrated that trimming procedure is 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 impact of NUMTs in plasma samples and provided the optimal choice for reference genomes in different sample types. Third, we demonstrated that the setting of 3 mismatches is most suitable for mtDNA mutation calling. More importantly, we found the negative logarithmic relationship between mtDNA site sequencing depth and minimum detectable mutation frequency and thus innovatively built up an efficient filtering strategy to increase accuracy of mutation detection. All these efforts greatly contribute to the establishment of an innovative and versatile bioinformatics pipeline for the accuracy of mtDNA mutation detection, laying a foundation for translational application of mitochondrial mutations in clinical practices.

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 eliminate adaptors and poor-quality bases of the sequencing reads, is an initial step of the analysis that is heterogeneously applied in previous mtDNA mutation analyses (22).

Whether the trimming procedure is indispensable to mtDNA mutation analysis remains to be confirmed, especially for different sample types. In the present study, we demonstrated that trimming procedure is essential for improving mtDNA mapping performance in plasma but not tumor tissue samples.

Since the true origin of homologous reads are difficult to discriminate, NUMTs and mtDNA cross-mapping occurs, leading to the detection of false positive (NUMTs reads aligning to chrM) or false negative (mtDNA reads aligning to NUMTs loci) mtDNA mutations. To address this concerning source of error, Li et al. have previously created a database of NUMTs containing mismatches from the mitochondrial genome, which may appear as false heteroplasmies if aligned to chrM (23). Complementing this approach, two main mapping strategies have been commonly used at present. The first is mapping to the human reference mtDNA sequence rCRS, and the second is mapping to rCRS and hg19 (human genome 19) that contributes to remove NUMTs during analysis. Therefore, the accuracy of mtDNA mutation calling can be largely affected by mapping strategies with different selection of reference genomes and mismatch number, owing to existence of sequence similarity between NUMTs and mtDNA (a known source of confounding in mtDNA NGS studies) (24). However, there are few studies focusing on the applicability of the two 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 both the two mapping strategies can be effectively used in both FFPE and fresh tissue samples, although only mapping to rCRS may introduce a very small amount of NUMTs. However, the mapping strategy with both hg19 and rCRS as reference genomes is strongly suggested for mtDNA mutation detection in plasma samples due to the extreme abundance of NUMTs. Furthermore, our study strongly suggests the setting of 3 mismatches for mtDNA mutation calling in 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 the somatic mtDNA mutations at a very low heteroplasmic level (MAF > 1%), with the NGS depth for the mitochondrial genome 9959 × (18). 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 in order to obtain confident and reproducible detections 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 up to now. Since there are no consensus criteria for determining mtDNA heteroplasmy threshold, it is very common to arbitrarily select a constant heteroplasmy level (mainly 2% or 5%) in previous NGS analyses for mtDNA mutations (12, 17). Here, we innovatively revealed the relationship of site sequencing depth and the minimum mutation threshold via establishing logarithmic functions in different sample types, which contributes to the discrimination of false-positive and false-negative mutations with ultra-low heteroplasmy level (up to 0.5%). By providing a standardized option of proper mutation frequency threshold, it can largely increase the sensitivity and accuracy of mtDNA mutation detection during NGS data analysis.

One unique aspect of our study is the integrative analysis of mtDNA mutation calling based on the assessment of key procedures. We found that the filtering strategy based on site sequencing depth greatly contributes to improve the detection consistency of repeated data. Previous studies on 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. Besides, 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 (19, 21). However, those filtering strategies may not be necessary for the mutation analysis in ultra-deep sequencing (with 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 600X when detecting mutations at the threshold greater than 1%. Moreover, we also observed that the filtering strategies considering DNA oxidation and binominal distribution exhibited no significant impact in our analysis system, which may at least partially be due to the strict removing 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 indeterminate, but can be applied for providing a reference for the range of sequencing depth and heteroplasmy level in NGS data.

Both capture-based and PCR-based methods are commonly used for mtDNA enrichment before sequencing (16). Therefore, selecting the right enriching approach and sequencing platform is of great importance to accurately identify mtDNA mutations, especially those with ultra-low heteroplasmy level. A previous study has compared the detection performance of different enrichment strategies in fresh tissue samples and illustrated that DNA quality was a great challenge for PCR-based approach, which would not work with highly fragmented DNA samples such as FFPE tissues and plasma (25). In the present study, we explored different enriching approaches and sequencing platforms in three different sample types. Our results suggest higher sequencing depth for PCR-based sequencing approach when compared with capture-based sequencing, and the innovative data analysis pipeline is applicable for both Illumina and BGI sequencing platforms in the detection of mtDNA mutations.

# Conclusions

Taken together, 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 mitochondrial genome in cancer cells.

# Abbreviations

NGS
Next generation sequencing; mtDNA:Mitochondrial DNA; WGS:Whole genome sequencing; WES:Whole exome sequencing; NUMTs:Nuclear sequences of mitochondrial origin; rCRS:revised Cambridge Reference Sequence; AFP:Assumed false positive; AFN:Assumed false negative; CV:Coefficient of

variation; hg19:Human genome 19; FFPE:Formalin-Fixed and Paraffin-Embedded; WGS:Whole genome sequencing; MAF:Minor allele frequency;

# Declarations


### Acknowledgements

The authors thank Deyang Li, Xiaohong Du and Xiangxu Wang in Department of Physiology and Pathophysiology for ongoing support and discussion.


### Authors' contributions

SG and KZ carried out the sample collection, performed the data analysis, and drafted the manuscript. QY and LS participated in the bioinformatics analyses. YL and XJ performed the laboratory experiments. XG and XG participated in the design of the study and performed the draft revision. JX conceived of the study, and participated in its design and coordination and helped to revise the manuscript. All authors read and approved the final manuscript.


### Funding

This work was supported by the National Natural Science Foundation of China (grants 81830070); and Autonomous Project of State Key Laboratory of Cancer Biology, China (grants CBSKL2019ZZ06).


### Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Ethics approval and consent to participate

This study was approved by the Ethical Committees of FMMU and written consent was obtained from each patient.

### Consent for publication

Not applicable.

### Competing interests

The authors have declared no competing interests.

### Author details


[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.

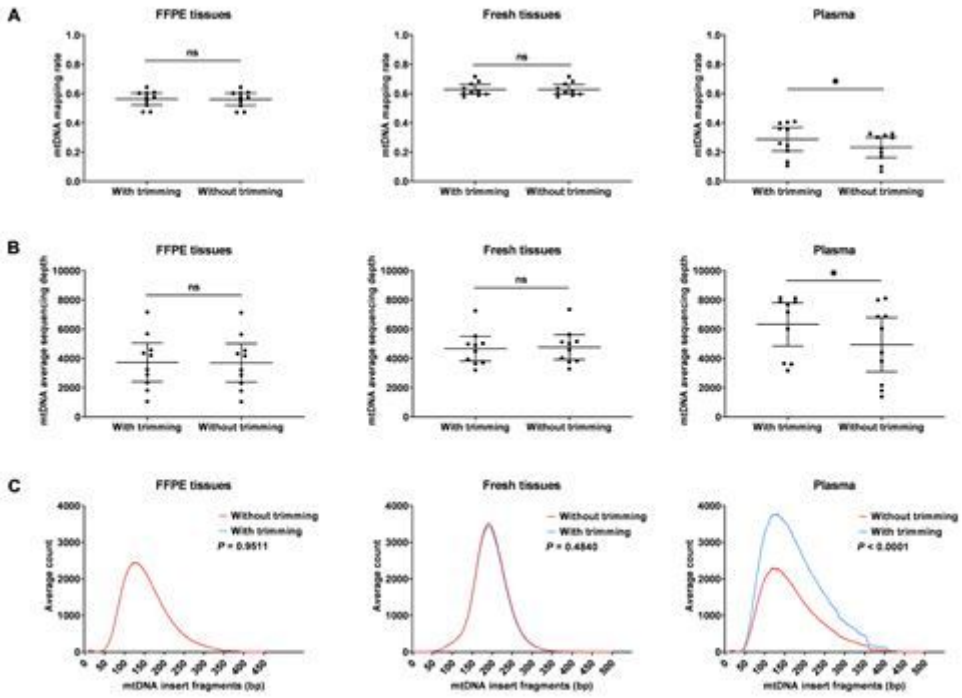
[3]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

# References

1. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–65.
2. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nature reviews Genetics. 2012;13(12):878–90.
3. He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, et al. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature. 2010;464(7288):610–4.
4. Mardis ER. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome medicine. 2009;1(4):40.
5. Ye F, Samuels DC, Clark T, Guo Y. High-throughput sequencing in mitochondrial DNA research. Mitochondrion. 2014;17:157–63.
6. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. eLife. 2014;3.
7. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nature methods. 2010;7(2):111–8.
8. Wong LJ. Challenges of bringing next generation sequencing technologies to clinical molecular diagnostic laboratories. Neurotherapeutics: the journal of the American Society for Experimental NeuroTherapeutics. 2013;10(2):262–72.
9. González MDM, Ramos A, Aluja MP, Santos C. Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing. Mitochondrion. 2020;50:88–93.
10. Li M, Schröder R, Ni S, Madea B, Stoneking M. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc Natl Acad Sci USA. 2015;112(8):2491–6.
11. Stewart JB, Chinnery PF. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nature reviews Genetics. 2015;16(9):530–42.
12. Hodgkinson A, Idaghdour Y, Gbeha E, Grenier JC, Hip-Ki E, Bruat V, et al. High-resolution genomic analysis of human mitochondrial RNA sequence variation. Science. 2014;344(6182):413–5.
13. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010;6(2):e1000834.
14. Yin C, Li DY, Guo X, Cao HY, Chen YB, Zhou F, et al. NGS-based profiling reveals a critical contributing role of somatic D-loop mtDNA mutations in HBV-related hepatocarcinogenesis. Annals of oncology:

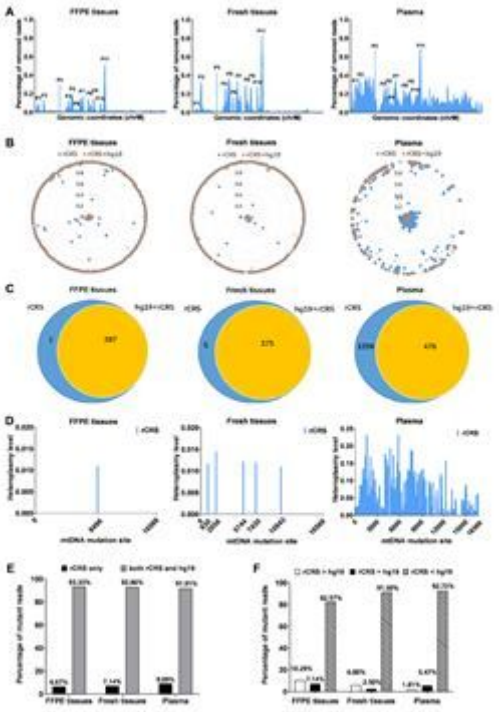
official journal of the European Society for Medical Oncology. 2019;30(6):953–62.
15. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.
16. Liu Y, Guo S, Yin C, Guo X, Liu M, Yuan Z, et al. Optimized PCR-Based Enrichment Improves Coverage Uniformity and Mutation Detection in Mitochondrial DNA Next-Generation Sequencing. The Journal of molecular diagnostics: JMD. 2020;22(4):503–12.
17. Li X, Guo X, Li D, Du X, Yin C, Chen C, et al. Multi-regional sequencing reveals intratumor heterogeneity and positive selection of somatic mtDNA mutations in hepatocellular carcinoma and colorectal cancer. International journal of cancer. 2018;143(5):1143–52.
18. Yuan Y, Ju YS, Kim Y, Li J, Wang Y, Yoon CJ, et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat Genet. 2020;52(3):342–52.
19. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research. 2013;41(6):e67.
20. Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, et al. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Veterinary research. 2013;44(1):12.
21. Campo DS, Nayak V, Srinivasamoorthy G, Khudyakov Y. Entropy of mitochondrial DNA circulating in blood is associated with hepatocellular carcinoma. BMC Med Genom. 2019;12(Suppl 4):74.
22. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinform. 2016;17:103.
23. Li M, Schroeder R, Ko A, Stoneking M. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. Nucleic acids research. 2012;40(18):e137.
24. Maude H, Davidson M, Charitakis N, Diaz L, Bowers WHT, Gradovich E, et al. NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele. Frontiers in cell developmental biology. 2019;7:201.
25. Kaneva K, Merkurjev D, Ostrow D, Ryutov A, Triska P, Stachelek K, et al. Detection of mitochondrial DNA variants at low level heteroplasmy in pediatric CNS and extra-CNS solid tumors with three different enrichment methods. Mitochondrion. 2020;51:97–103.

# Figures



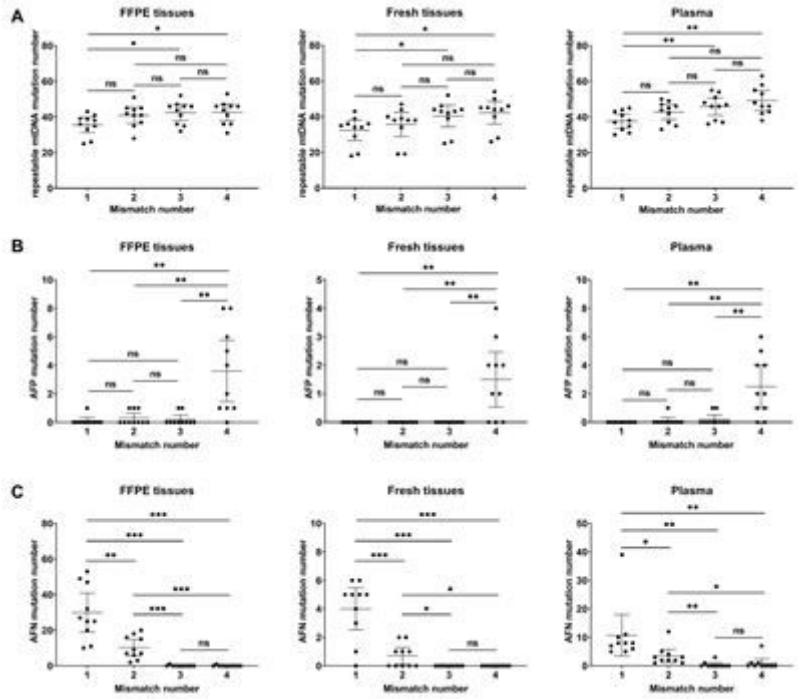

## Figure 1

Effect of trimming procedure on mapped mtDNA sequencing data A. Comparison of mtDNA mapping rate between the trimming group and without trimming group in FFPE tissue, fresh tissue, and plasma samples. B. Comparison of mtDNA average sequencing depth between the trimming group and without trimming group in FFPE tissue, fresh tissue, and plasma samples. C. Distribution of mtDNA insert size between the trimming group and without trimming group in FFPE tissue, fresh tissue, and plasma samples. ns, no significance; *P < 0.05.



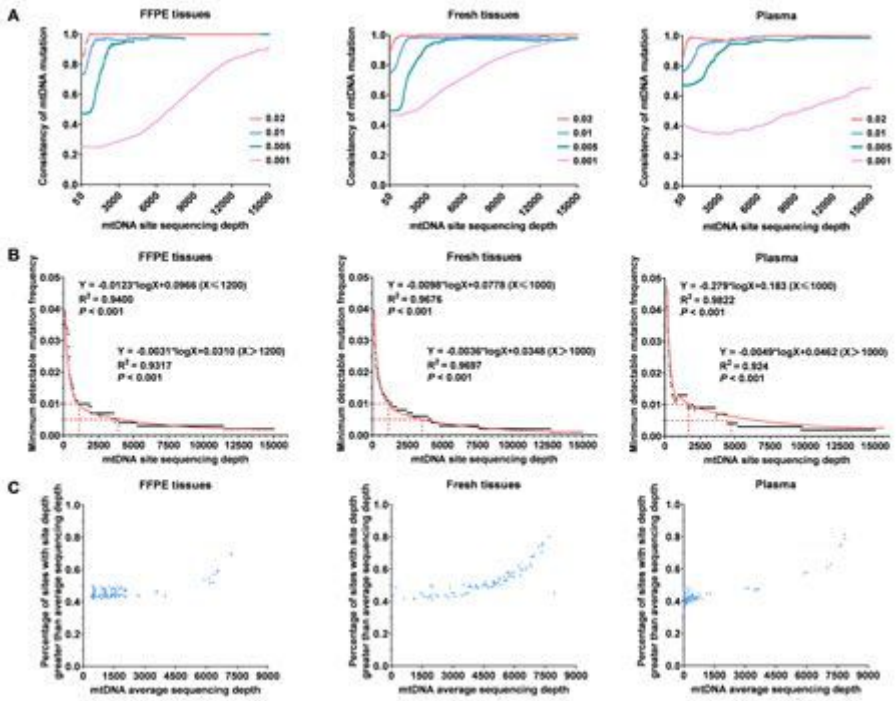
## Figure 2

Evaluation of 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 two mapping strategies (rCRS-alone and combined hg19-rCRS) in FFPE tissue, fresh tissue, and plasma samples. Inner circles represent heteroplasmy level. 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 two mapping strategies (rCRS-alone and combined hg19-rCRS) in FFPE tissue, fresh tissue, and plasma samples. D. Heteroplasmy level 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 both hg19 and rCRS reference. F. The majority of mutant reads with dual hg19 and rCRS mapping showed greater hg19 alignment score compared to rCRS alignment score.



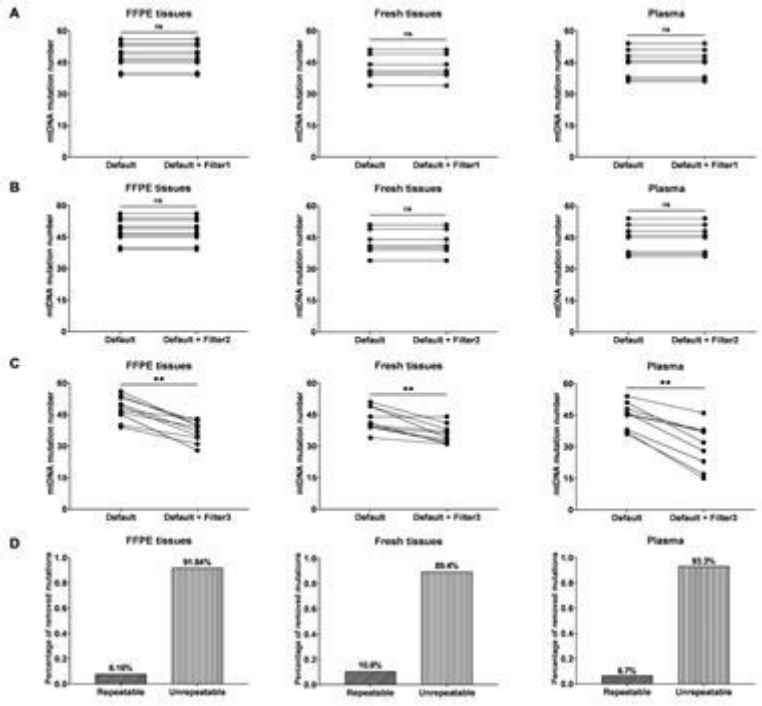

## Figure 3

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 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. ns, no significance; *P < 0.05；**P < 0.01；***P < 0.001.



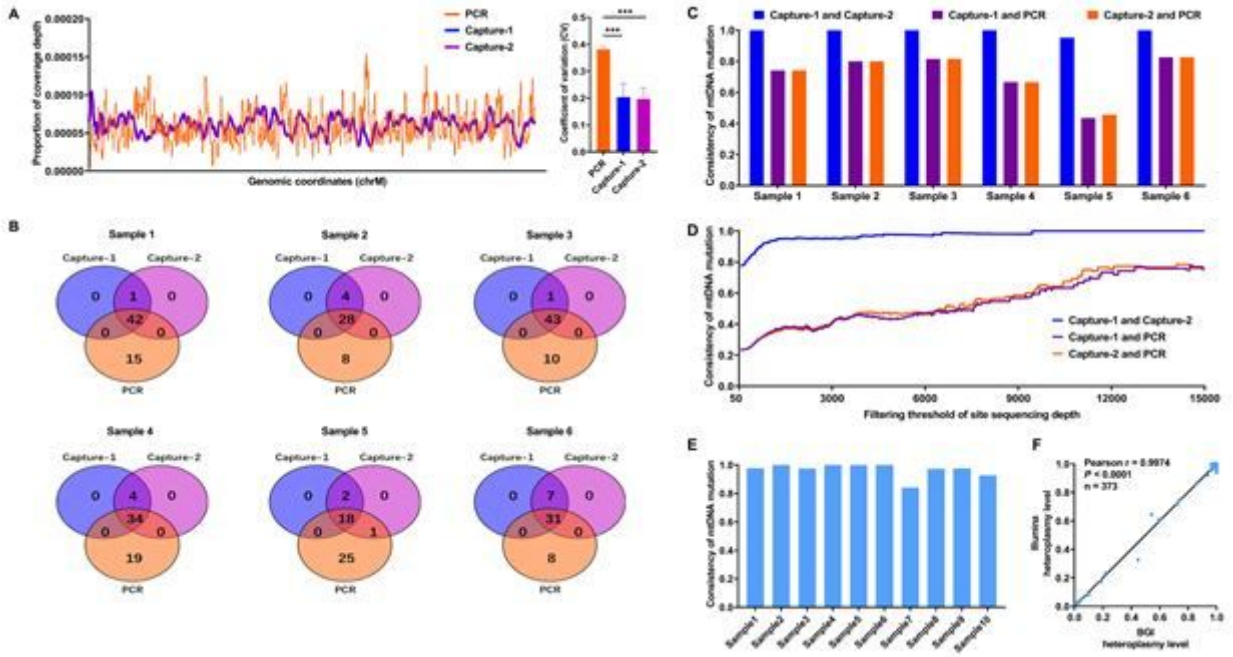

## Figure 4

Relationship between the site sequencing depth and consistency of mtDNA mutations at different heteroplasmy levels A. Consistency of mtDNA mutations at different site sequencing depth of two repeat 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 site sequencing depth greater than average sequencing depth.



## 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 is to remove C > A/G > T mutations with 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 is to remove mtDNA mutations where mutation-rate and mutation base-quality does not pass 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 if the MAF is smaller than the site-specific threshold determined by site sequencing depth. D. Percentage of the repeatable and unrepeatable mtDNA mutations that were removed by filter 3 in FFPE tissue, fresh tissue and plasma samples. ns, no significance; **P < 0.01.




## Figure 6

Performance comparison between PCR-based and capture-based enrichment strategies or between two sequencing platforms A. Distribution of coverage depth across mitochondrial genome (left panel) and comparison of 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 threshold. E. Consistency of the mtDNA mutations between Illumina and BGI sequencing platforms in ten fresh tissue samples. F. Consistency of heteroplasmy level of mtDNA mutations detected by Illumina and BGI sequencing platforms in ten fresh tissue samples. ***P < 0.001.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigure.docx
- Supplementaryfigure.docx
- SupplementaryTable1.doc
- SupplementaryTable1.doc