Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2025, Volume 5, Issue 1, Page No: 260-275

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

A High-Accuracy Machine Learning Approach Improves Mitochondrial DNA Variant Fidelity in Archival FFPE Samples

Kenji Tanaka[1]*, Yui Kobayashi[1], Takeshi Nakamura[2]

[1]Department of Computational Genomics and Machine Learning, Faculty of Medicine, University of Tokyo, Tokyo, Japan.
[2]Department of Molecular Pathology and FFPE Analysis, Faculty of Medicine, Kyoto University, Kyoto, Japan.

***E-mail** ✉ tanaka.kenji@med.u-tokyo.ac.jp

Received: 26 April 2025; **Revised:** 24 June 2025; **Accepted:** 27 June 2025

ABSTRACT

Archival formalin-fixed paraffin-embedded (FFPE) specimens represent a highly valuable resource for mitochondrial DNA (mtDNA) biomarker research. Nonetheless, artifacts generated during formalin fixation frequently resemble somatic mutations, thereby impeding accurate clinical interpretation. Existing artifact-filtering methods, primarily developed for nuclear DNA, fail to accommodate key mtDNA characteristics such as its elevated copy number, strand-specific GC-content differences, and heteroplasmic properties. These shortcomings call for the creation of specialized computational strategies. To meet this need, we introduced mtFFPECleaner, a machine learning-based system that combines multiple features of mtDNA variants, encompassing variant allele frequency (VAF) patterns, strand orientation bias metrics, surrounding sequence context, and local nucleotide composition. The system employs a random forest classifier trained on 837 verified genuine mutations and 1169 artifacts sourced from 23 matched FFPE–fresh frozen (FF) specimen pairs. Training involved tenfold cross-validation, with performance subsequently confirmed on an independent cohort of 15 additional paired FFPE–FF samples. Our results indicate that formalin-related artifacts in mtDNA next-generation sequencing (NGS) data mainly consist of C > T/G > A transitions, particularly at low VAF levels, and display pronounced strand bias along with clear sequence context preferences. Following optimization with balanced sampling (1:2 ratio of artifacts to genuine mutations), the mtFFPECleaner classifier demonstrated superior performance in the validation cohort, attaining 98.7% specificity and 98.2% sensitivity. It substantially surpassed established nuclear DNA artifact-removal approaches, including SOBDetector and DEEPOMOICS FFPE. After applying mtFFPECleaner for artifact correction, mutational spectra derived from 314 FFPE samples exhibited close alignment with those obtained from FF specimens. Notably, artifact prevalence increased in association with longer FFPE storage times, confirming the tool's capacity to counteract accumulated formalin-induced damage spanning many decades in preserved biospecimens. mtFFPECleaner is the first tool developed specifically to improve the reliability of mtDNA mutation profiles from FFPE material in NGS analyses. Availability of the open-source R package (https://github.com/AlienLemon117/mtFFPECleaner) supports its broad application in extensive archival investigations and enhances the utility of FFPE biobanks for translational research.

Keywords: Artifact removal, FFPE specimens, Machine learning, Mitochondrial DNA mutation, Next-generation sequencing

How to Cite This Article: Tanaka K, Kobayashi Y, Nakamura T. A High-Accuracy Machine Learning Approach Improves Mitochondrial DNA Variant Fidelity in Archival FFPE Samples. Interdiscip Res Med Sci Spec. 2025;5(1):260-75. https://doi.org/10.51847/awQVwVWCFL

Introduction

Mitochondrial DNA (mtDNA) is a circular genome responsible for encoding critical subunits of the oxidative phosphorylation (OXPHOS) machinery. Its susceptibility to mutations arises from the lack of histone protection, constrained DNA repair pathways, and spatial proximity to reactive oxygen species (ROS) [1]. Growing evidence positions somatic mtDNA mutations as important contributors to cancer development through effects on metabolic adaptation, tumor heterogeneity, and treatment resistance in multiple malignancies [2, 3]. Certain

mtDNA alterations also hold promise as clinical biomarkers; pathogenic changes in MT-ND1, for instance, have been linked to metastatic progression in melanoma [4], and MT-Dloop variants are associated with reduced survival in hepatocellular carcinoma [5]. Realizing the clinical value of these insights, however, requires precise separation of authentic mutations from sequencing artifacts—an especially formidable task when working with archival formalin-fixed paraffin-embedded (FFPE) tissues.

FFPE samples provide an irreplaceable archive for retrospective analyses and biomarker validation. Their advantages include exceptional long-term preservation, convenient handling and shipment, and linkage to richly annotated clinical datasets [6, 7]. These features have enabled comprehensive mtDNA mutation surveys across large FFPE collections, identifying signatures connected to disease subtypes and treatment outcomes [8, 9]. Nevertheless, the formalin fixation procedure introduces extensive DNA damage, including crosslinks between proteins and DNA, chemical base alterations, and fragmentation. Such lesions reduce sequencing fidelity and conceal true biological signals [10]. A primary mechanism involves cytosine deamination to uracil, producing U:G mismatches that manifest as artifactual C > T/G > A substitutions upon PCR amplification [11]. Fragmentation and depurination further compound error rates, especially in samples with limited or degraded DNA [7]. Consequently, these artifacts mimic genuine somatic mutations and generate background noise that complicates downstream interpretation of clinically relevant variants.

Both laboratory-based and computational strategies for mitigating formalin artifacts have notable shortcomings. Enzymatic repair performed prior to sequencing provides incomplete restoration of native DNA structure. Although uracil-DNA glycosylase (UDG) treatment effectively excises uracil arising from cytosine deamination and reduces C > T transitions [12], it leaves 20%–30% of residual artifacts stemming from methylated cytosine deamination uncorrected. These persistent lesions continue to produce false-positive calls in nuclear DNA studies [13]. Bioinformatic approaches after sequencing face two central difficulties when applied to mtDNA. First, substantial overlap exists between artifactual and biological mutational signatures, particularly the predominance of C > T/G > A changes, which precludes simple rule-based filtering without inadvertently discarding true mutations [14]. Second, the extensive heterogeneity inherent to mtDNA across organelles and cell populations demands ultra-sensitive yet specific artifact detection that preserves low-frequency genuine variants. While strand orientation bias (SOB) analysis can theoretically separate artifacts (directionally biased) from true mutations (evenly distributed), technical factors such as PCR bias, uneven capture, and coverage variation often distort SOB measurements for authentic variants. Moreover, tools like SOBDetector, trained predominantly on nuclear DNA, inadequately handle mtDNA-specific properties, including its multicopy nature, GC-content asymmetry between strands, and heteroplasmy [15]. The current literature reveals a significant deficiency in detailed characterization of false-positive mutation patterns in mtDNA derived from FFPE samples [15-17]. These combined issues underscore the necessity for mtDNA-tailored analytical solutions.

To resolve these limitations, we created mtFFPECleaner, a sophisticated machine learning framework dedicated to distinguishing formalin-induced artifacts from authentic mtDNA mutations in FFPE-derived sequencing data **(Figure 1)**. The model synthesizes diverse predictive features, including heteroplasmy fractions, VAF interval distributions, local sequence context, and bias-adjusted SOB scores that incorporate mtDNA-specific technical considerations. By accounting for mitochondrial genome attributes such as strand-specific transcription bias and replication-associated patterns, the framework removes artifacts while safeguarding biologically meaningful mutations. This development marks the first dedicated bioinformatics resource for preserving mtDNA mutation accuracy in archival FFPE material and opens new avenues for longitudinal biomarker studies in precision oncology and mitochondrial medicine.

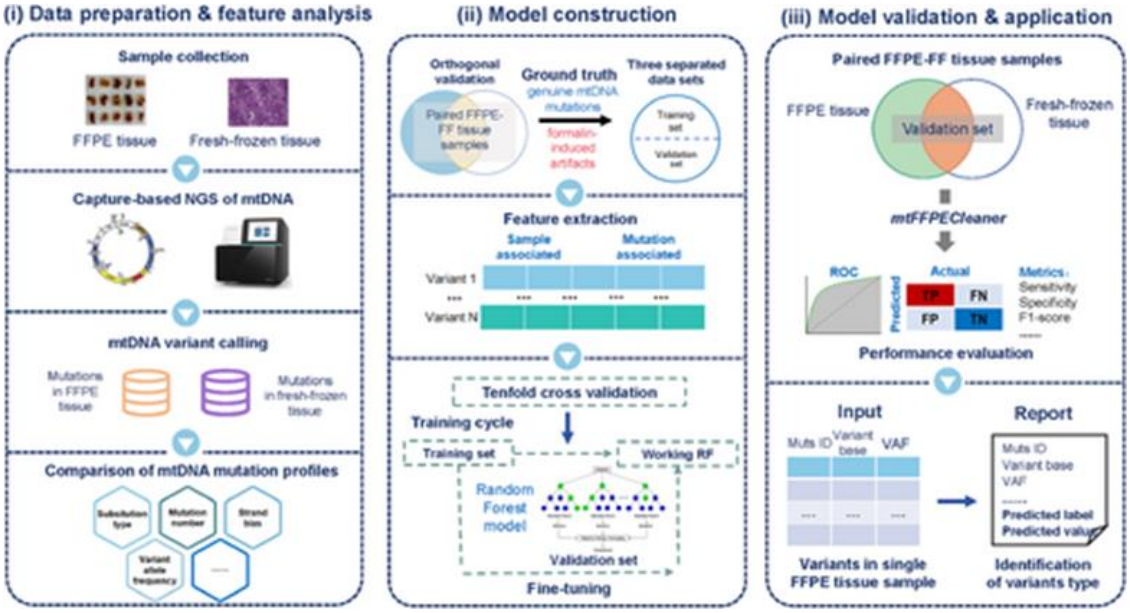


Figure 1. Workflow of mtFFPECleaner.

The mtFFPECleaner framework includes data preparation with feature extraction, model development, and performance assessment with practical application. (i) Targeted NGS of mtDNA was carried out on FFPE and FF tissues to detect mtDNA mutations. Mutation patterns were contrasted between FFPE and FF samples utilizing both internal and external datasets. (ii) Orthogonal validation of paired FFPE–FF tissues yielded ground-truth genuine mutations and formalin-induced artifacts, which were partitioned into three separate datasets. A random forest classifier was chosen as the best-performing algorithm through tenfold cross-validation. (iii) Independent evaluation of mtFFPECleaner was performed on a held-out test set. mtDNA mutations identified in FFPE samples were submitted to the mtFFPECleaner package for classification. FF, fresh-frozen; FN, false negative; FFPE, formalin-fixed paraffin-embedded; FP, false positive; NGS, next-generation sequencing; RF, random forest; TN, true negative; TP, true positive; VAF, variant allele frequency.

Materials and Methods

Private and public mtDNA NGS data

Capture-based mitochondrial DNA next-generation sequencing (NGS) datasets from 314 FFPE specimens and 864 fresh frozen (FF) tissues, previously described in our publications, formed the private FFPE cohort 1 and private FF cohort, respectively [8, 18]. FFPE samples were obtained from 157 ovarian cancer (OV) cases treated at Xijing Hospital, part of the Fourth Military Medical University (FMMU) in Xi'an, China. FF samples came from 432 colorectal cancer (CRC) patients recruited at Xijing and Tangdu Hospitals. Matched tumor and adjacent non-tumor tissues were collected from every participant, with FF specimens stored in liquid nitrogen. To limit batch-related variation, an additional set of capture-based mtDNA NGS data from 640 FFPE samples—originating from 160 thyroid cancer (TC) and 160 gastric cancer (GC) patients at Fujian Medical University Union Hospital—was included as private FFPE cohort 2. All study procedures were approved by the institutional review boards of FMMU (KY20183331-1), Xijing Hospital (XJLL-KY20212217), and Fujian Medical University Union Hospital (2024KJCX105). Every patient gave written informed consent prior to participation.

Publicly available datasets were integrated to confirm the validity of our observations. These consisted of ultradeep whole-exome sequencing results from 1015 CRC FFPE specimens and full mitochondrial genome sequences from fresh frozen tissues across 12 organs obtained from 152 autopsy individuals, serving as the public FFPE cohort and public FF cohort, respectively [9, 19]. All public mtDNA data underwent uniform reprocessing with our standardized laboratory pipeline to ensure consistency in read mapping, variant identification, and quality filtering steps adapted specifically for the mitochondrial genome.

Additionally, 38 paired FFPE and FF samples from hepatocellular carcinoma (HCC) patients were processed to generate reliable ground-truth sets of genuine mtDNA mutations and formalin-induced artifacts for model

development, tuning, and testing. Twenty-three of these pairs were assigned to the training phase, while the other 15 pairs were reserved for independent validation.

DNA extraction, library construction, and capture-based mtDNA sequencing

DNA isolation, library preparation, and targeted mtDNA capture sequencing were conducted according to a previously reported protocol with certain refinements [18]. The current work utilized an enhanced capture strategy that combined custom mtDNA probes with reference nuclear DNA probes for NGS library enrichment. Genomic DNA from fresh frozen tissues was purified using the ENZA DNA Kit (Omega), while FFPE-derived DNA was obtained with the FastPure FFPE DNA Isolation Kit (Vazyme) omitting UDG treatment. DNA yield was quantified on a Qubit 4.0 fluorometer (Thermo Fisher Scientific). Shearing produced random fragments, from which 300–500 bp segments were selected. In-house synthesized biotinylated mtDNA probes were hybridized to the libraries to selectively enrich mitochondrial sequences. Libraries were sequenced on the Illumina HiSeq X Ten or NovaSeq instruments, yielding paired-end 150 bp reads (PE150).

mtDNA variant calling pipeline

Variant detection relied on a custom pipeline established within our group for high-accuracy mtDNA analysis [20]. Raw reads were mapped to the revised Cambridge Reference Sequence (rCRS) and hg19 using BWA [21] to exclude nuclear mitochondrial DNA (NUMT) interference. Duplicate reads were eliminated with Picard, followed by local realignment around indels via GATK IndelRealigner to reduce artifactual calls [22]. High-quality alignments were retained using SAMtools [21]. Candidate variants underwent rigorous filtering: (i) VAF $\geq$ 1%; (ii) minimum depth of 100× at each site; (iii) base quality $\geq$ 30; (iv) minimum of three supporting reads on both strands; (v) exclusion of heteroplasmic positions in rCRS repeat regions (66–71, 303–311, 514–523, 12,418-12,425, 16,184-16,193); and (vi) elimination of C > A/G > T changes showing strong sequence context bias at low VAF (1%–2%), which typically arise from guanine oxidation [23].

Feature selection and tenfold cross-validation

Thirteen distinct features were derived for every variant to help separate true mtDNA mutations from formalin-related artifacts. These comprised eight characteristics at the sample level and five at the mutation level. The Weka-InfoGain tool (version 3.8) facilitated selection of the most informative, non-redundant variables. Five established supervised learning algorithms [24]—K-Nearest Neighbors, Logistic Regression, random forest (RF), Support Vector Machine (SVM), and XGBoost—were compared via tenfold cross-validation using their default settings. The training data included 837 verified genuine mutations and 1169 artifacts from the 23 matched FFPE-FF pairs. Algorithm selection was based on six performance indicators: area under the curve (AUC), accuracy, sensitivity, specificity, precision, and F1-score.

Random Forest model for identification of formalin-induced artifacts

Given its superior results in cross-validation, the random forest (RF) algorithm was selected and refined to build mtFFPECleaner. The composition of the training set was systematically adjusted to examine the influence of sample size and the proportion of genuine mutations relative to artifacts. Implementation was carried out with the R package “Random Forest” (version 4.6-14). The best-performing configuration employed a 1:2 ratio of genuine mtDNA mutations to artifacts. Iterative training runs were performed to optimize key parameters such as mtry (variables sampled per split) and ntree (total number of trees). Importance rankings for individual features were determined from the finalized model.

Evaluation and application of mtFFPECleaner

The capacity of mtFFPECleaner to accurately differentiate formalin-induced artifacts from true mutations was systematically tested on capture-based mtDNA NGS data generated from an independent set of 15 paired FFPE-FF tissue samples. This assessment was subsequently broadened to encompass both the internal private FFPE collections and external public FFPE datasets. Mutational spectra obtained from the private FF cohort functioned as the benchmark for these comparative evaluations.

Code implementation and output of mtFFPECleaner

The mtFFPECleaner tool was coded in the R language. Its output consists of a tab-delimited text file that lists, for each mtDNA variant, the sample identifier, chromosomal position, and the full set of 13 primary predictive features. The report also includes the assigned category (genuine mtDNA mutation or artifact) and the estimated probability (ranging between 0 and 1) that the variant is authentic. This format allows users to examine individual predictions and judge the confidence associated with variants classified as genuine mtDNA mutations.

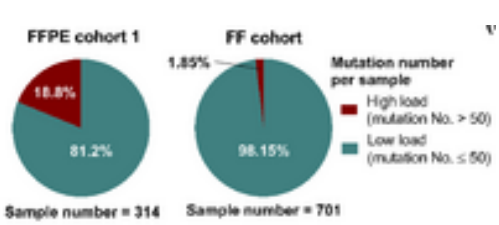
Statistical analysis

Statistical computations were carried out using GraphPad Prism version 8.3.0 (San Diego, CA, USA). Sample size requirements were determined in advance with G × Power software, yielding a statistical power of 90% for a significance level of 0.05. These calculations indicated that over 100 patients would be needed to achieve reliable sensitivity estimates at a minimum specificity of 0.95. Categorical variables were compared using the Chi-square test. Continuous data are reported as mean ± standard error of the mean (SEM). Group comparisons employed the Mann-Whitney U test, Student's t-test, or one-way ANOVA with Bonferroni correction, as appropriate. Variable associations were assessed via Spearman's rank correlation. Similarity between mutational spectra was measured with cosine similarity. All tests were two-tailed, and p values below 0.05 were considered statistically significant.

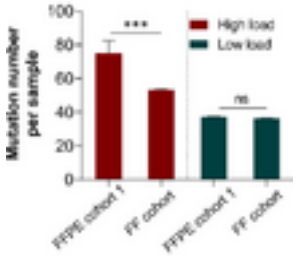
Results and Discussion

FFPE tissue samples with high mutation load showed increased prevalence of mtDNA C > T/G > A mutations

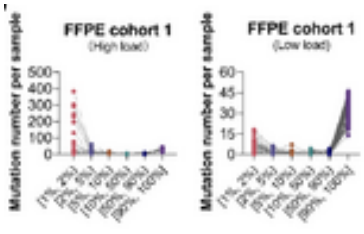
The process of formalin fixation combined with paraffin embedding generates chemical changes capable of producing spurious mutations, particularly C > T/G > A transitions resulting from cytosine deamination, as previously established in nuclear DNA research [11, 13]. To investigate how these artifacts affect mtDNA variant detection in preserved tissues, we conducted a comprehensive comparison of mtDNA mutation landscapes between FFPE and fresh frozen (FF) samples, leveraging both laboratory-generated and publicly accessible resources. Samples were divided into high-load (>50 mutations) and low-load (≤50 mutations) categories to clarify the signature of artifactual events. Within the private datasets, private FFPE cohort 1 contained a considerably larger percentage of high-load cases than the private FF cohort **(Figure 2a)**, (18.8% versus 1.85%, $p < 0.001$). While mutation counts in low-load groups were comparable between FFPE and FF specimens, high-load FFPE samples accumulated far more mutations than high-load FF samples **(Figure 2b)**. Detailed stratification by VAF ranges uncovered pronounced differences in mutation distribution between high- and low-load FFPE groups, most evident at lower VAF thresholds **(Figure 2c)**. Conversely, high- and low-load FF samples showed largely consistent distributions **(Figure 2d)**. High-load FFPE samples displayed elevated mutation frequencies across the 1%–2%, 2%–5%, and 5%–10% VAF intervals when compared with low-load FFPE and both FF groups **(Figure 2e)**. High-load FF samples, however, presented more mutations at very high VAF levels (>90%), reflecting greater sequence divergence from the reference **(Figure 2f)**. Equivalent trends were confirmed in private FFPE cohort 2 and the public FFPE cohort.



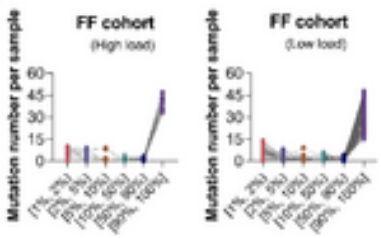
a)



b)



c)



d)

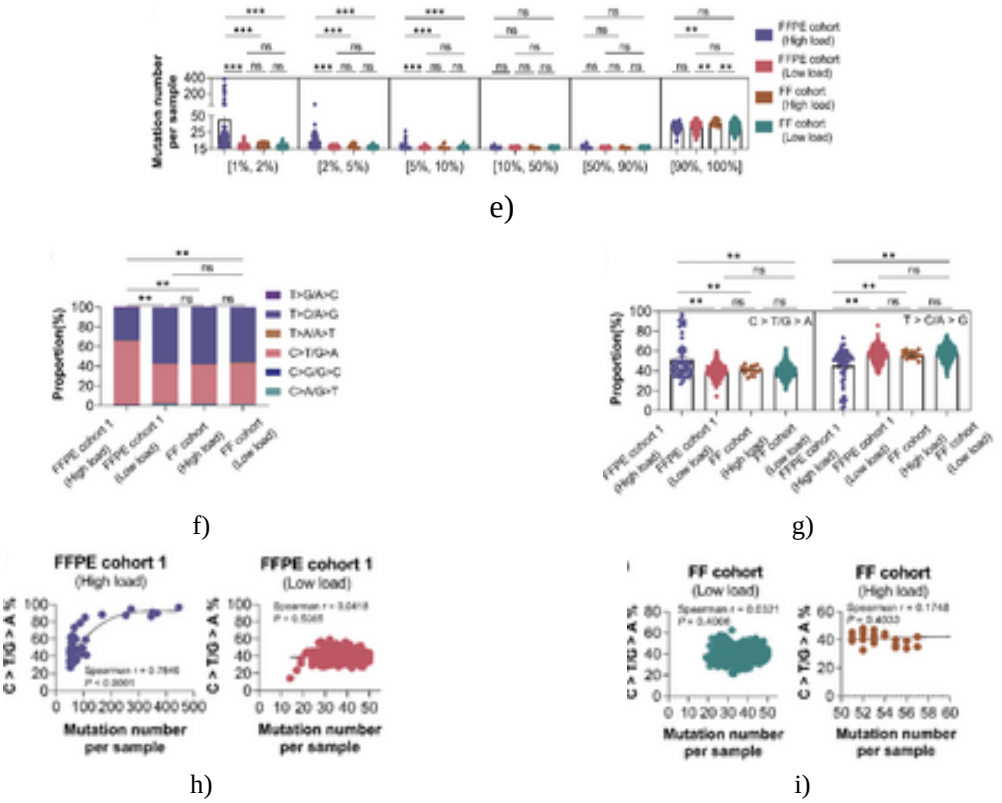


Figure 2. FFPE tissue samples with high mutation load showed increased prevalence of mtDNA C > T/G > A mutations.

(a) Pie charts depicting the relative frequencies of high-load and low-load samples in private FFPE cohort 1 and private FF cohort. A cutoff of 50 mutations distinguished high- from low-mutation load groups. (b) Total mutation counts compared between high-load and low-load samples in private FFPE versus FF cohorts. (c, d) Mutation count distributions across six variant allele frequency (VAF) bins in private FFPE cohort 1 (c) and private FF cohort (d). (e) Mutation counts across six VAF intervals compared between private FFPE cohort 1 and private FF cohort. (f) Mutational spectra by substitution class compared between private FFPE cohort 1 and private FF cohort. (g) Relative frequencies of C > T/G > A versus T > C/A > G mutations compared between private FFPE cohort 1 and private FF cohort. (h–i) Association between overall mtDNA mutation burden and the fraction of C > T/G > A mutations below 90% VAF in private FFPE cohort 1 (h) and private FF cohort (i). Values represent mean ± SEM. Mann-Whitney U test was applied for panel (b). One-way ANOVA with Bonferroni post hoc correction was used for panels (e, g). Chi-square test was performed for panel (f). Spearman correlation analysis provided coefficients and p-values for panels (h, i). VAF, variant allele frequency. FF, fresh-frozen; FFPE, formalin-fixed paraffin-embedded. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Analysis of base changes highlighted unique mutational profiles in high-load FFPE samples that diverged from those in low-load FFPE and FF specimens **(Figure 2f)**. High-load FFPE samples were characterized by strong overrepresentation of C > T/G > A mutations and underrepresentation of T > C/A > G mutations **(Figure 2g)**. When examined by VAF interval, C > T/G > A transitions comprised over 90% of variants in the 1%–2% range among high-load FFPE samples, with declining proportions at higher VAFs. In contrast, proportions remained relatively constant across VAF ranges in FF samples and low-load FFPE samples. A robust positive correlation existed between total mutation number and the percentage of low-VAF (<90%) C > T/G > A mutations specifically in high-load FFPE samples (Spearman's r = 0.7846, p < 0.0001); **(Figure 2h)**, (left panel), but this association was absent in low-load FFPE samples **(Figure 2h)**,(right panel) and FF samples **(Figure 2i)**. Taken together, an unusually elevated mutation burden dominated by C > T/G > A transitions represents a defining feature of artifact contamination in FFPE-derived mtDNA profiles.

Formalin-induced artifacts showed differential characteristics compared to genuine mtDNA mutations

To better understand the properties of sequencing artifacts in FFPE-derived mtDNA data, ultra-deep targeted NGS was performed on 38 FFPE samples, each accompanied by a matched fresh-frozen (FF) specimen. Mean sequencing depth surpassed 10,000× across samples, enabling sensitive detection of low-frequency variants. This paired analysis confirmed markedly higher mtDNA mutation counts in FFPE tissues compared with matched FF tissues ($p < 0.001$) **(Figure 3a)**. Certain FFPE samples contained more than 400 unique mutations not present in the corresponding FF tissue, underscoring the extensive artifact accumulation caused by formalin exposure and embedding **(Figure 3a)**.

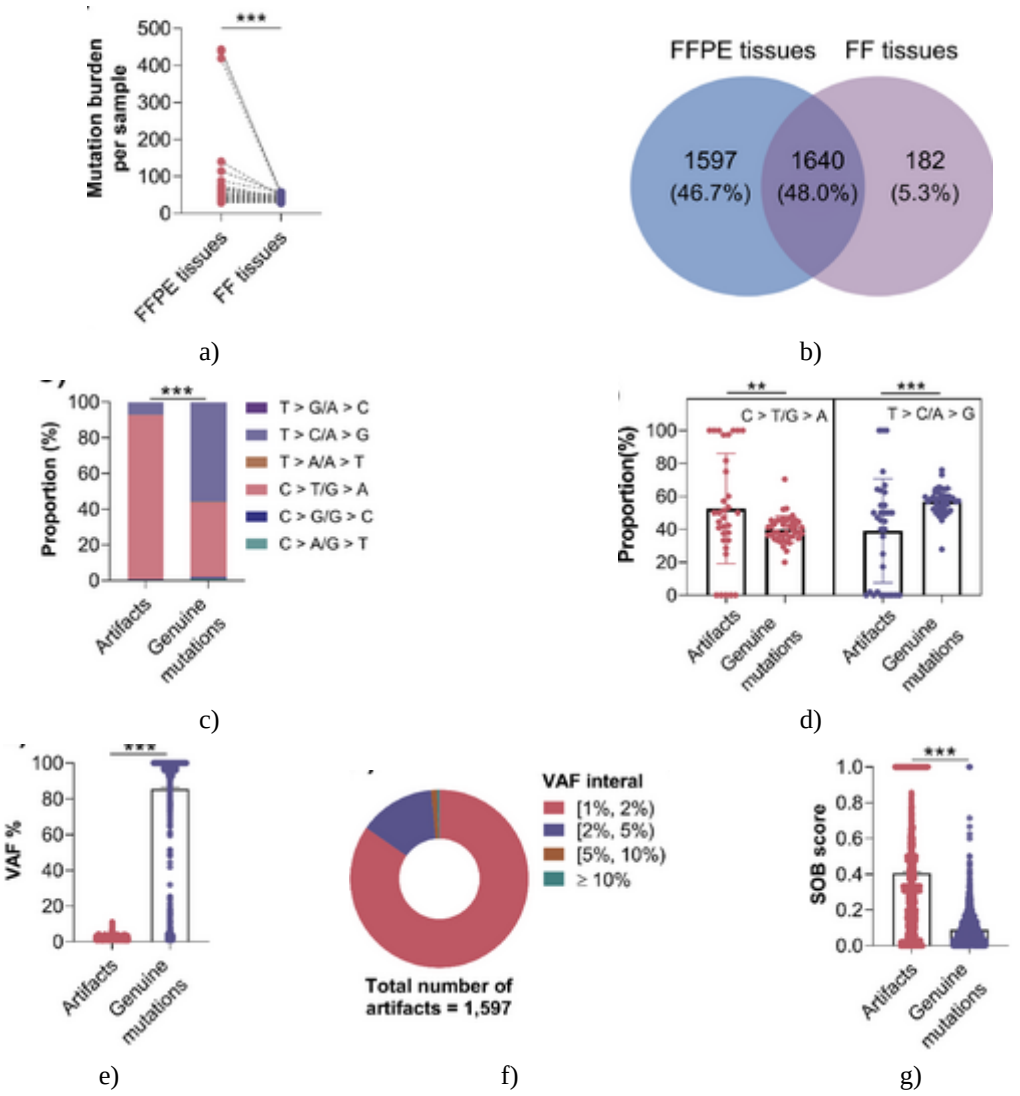


Figure 3. Formalin-induced artifacts showed differential characteristics compared to genuine mtDNA mutations.

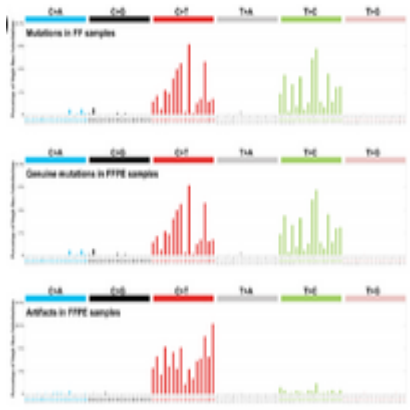
(a) Mutation counts compared between 38 matched FFPE and FF tissue pairs. (b) Venn diagram illustrating shared and unique mtDNA mutations identified in the 38 paired FFPE and FF samples. (c) Substitution type distributions for formalin-induced artifacts versus genuine mtDNA mutations. (d) Relative abundance of C > T/G > A and T > C/A > G mutations compared between artifacts and genuine mutations. (e) Variant allele frequency (VAF) distributions compared between formalin-induced artifacts and genuine mtDNA mutations. (f) Distribution of formalin-induced artifacts across four VAF ranges shown as a pie chart. (g) Strand orientation bias (SOB) scores compared between formalin-induced artifacts and genuine mtDNA mutations. Results are displayed as mean ± SEM. Paired t-test was used for panel (a). Chi-square test was applied for panel (c). Mann-Whitney U test was performed for panels (d, e, and g). FF, fresh-frozen; FFPE, formalin-

fixed paraffin-embedded; SOB, strand orientation bias; VAF, variant allele frequency. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

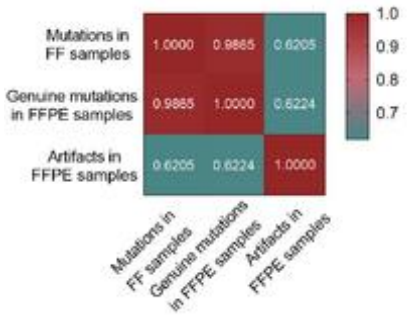
Mutations identified in FFPE samples were categorized as either exclusive to FFPE tissue (termed formalin-induced artifacts) or present in both FFPE and FF tissues (termed genuine mtDNA mutations). This classification yielded 2257 artifacts and 1352 genuine mutations across the 38 pairs **(Figure 3b)**. Substitution profiling demonstrated that artifacts were overwhelmingly dominated by C > T/G > A transitions (91.9%), far exceeding the corresponding proportion among genuine mutations **(Figure 3c)**. Artifacts further showed enrichment of C > T/G > A changes and depletion of T > C/A > G changes relative to genuine variants **(Figure 3d)**. VAF values for artifacts were consistently lower than those of genuine mutations, with maximum artifact frequencies remaining below 20% **(Figure 3e)**. The bulk of artifacts (84.7%) clustered in the narrow 1%–2% VAF window **(Figure 3f)**. Since authentic mutations generally involve both mtDNA strands whereas artifacts primarily affect one strand, strand orientation bias (SOB) scores were computed [15]. Consistent with expectations, SOB scores were substantially elevated for artifacts compared with genuine mutations **(Figure 3g)**. These observations collectively underscore the distinctive molecular features that distinguish formalin-induced artifacts from true mtDNA mutations in sequencing data from FFPE specimens.

The mutational spectrum of formalin-induced artifacts differs from that of genuine mtDNA mutations

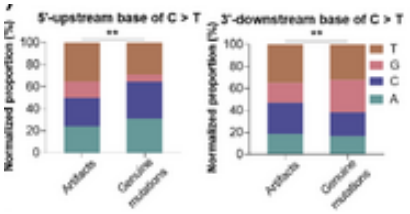
The nucleotides immediately flanking a mutation site—collectively termed sequence context—provide essential insights into the processes driving mutagenesis and their connection to particular damaging agents or local genomic properties. For example, oxidative lesions such as 8-oxoguanine (8-oxoG) characteristically produce C > A and G > T transversions, with pronounced preferences for specific motifs including CCN > CAG and NGG > NTG [23]. Applying this analytical lens, we conducted a comprehensive side-by-side examination of formalin-induced artifacts and genuine mtDNA mutations, which uncovered markedly different mutational signatures and contextual patterns. Both groups generated 96-component spectra dominated by transition substitutions, chiefly C > T (CL > TL and CH > TH) and T > C (TL > CL and TH > CH) changes **(Figure 4a)**. Pairwise cosine similarity analysis of median values confirmed that genuine mtDNA mutations from FFPE samples closely mirrored the spectra observed in paired FF samples, yet diverged substantially from the artifactual mutations within the same FFPE specimens **(Figure 4b)**.



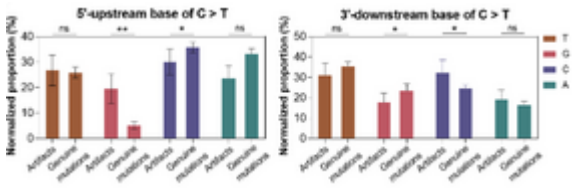
a)



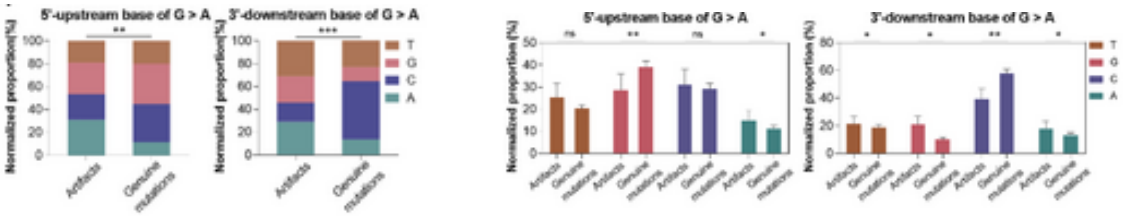
b)



c)



d)



e) f)

Figure 4. The mutational spectrum of formalin-induced artifacts differs from that of genuine mtDNA mutations.

(a, b) Ninety-six-component mutational spectra (a) and pairwise similarity comparisons (b) across the three mutation groups. (c, e) Flanking nucleotide composition at 5′-upstream (left panels) and 3′-downstream (right panels) positions for C > T (c) and G > A (e) mutations in artifacts versus genuine mutations. (d, f) Relative frequencies of flanking bases at 5′-upstream (left) and 3′-downstream (right) positions for C > T (d) and G > A (f) mutations compared between formalin-induced artifacts and genuine mtDNA mutations. Results shown as mean ± SEM. Cosine similarity was used for spectrum comparisons in panel (b). Chi-square tests were applied in panels (c, e). Mann-Whitney U tests were performed in panels (d, f). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Further investigation of the dominant C > T and G > A substitutions highlighted clear context-specific distinctions. Examination of base frequencies at positions immediately upstream and downstream of mutation sites, normalized to the mitochondrial reference genome (rCRS) composition, revealed systematic differences. For C > T mutations, artifacts displayed altered flanking nucleotide profiles compared with genuine mutations **(Figure 4c)**, including higher G and lower C content upstream and the reverse pattern (higher C, lower G) downstream **(Figure 4d)**. G > A mutations exhibited parallel shifts **(Figure 4e)**, with artifacts showing elevated A and reduced G upstream, and increased T, G, and A together with decreased C downstream **(Figure 4f)**. These contextual biases were independently validated in a separate large dataset of candidate low-VAF (1%–2%) artifacts from the highest-mutation-load FFPE samples in private cohort 1 and genuine mutations from the private FF cohort. Overall, these results establish sequence context and local base composition as powerful discriminative elements for distinguishing artifactual from authentic mtDNA variants.

Construction and optimization of mtFFPEcleaner for identification of formalin-induced artifacts

Despite the existence of several differentiating characteristics between formalin-induced artifacts and genuine mtDNA mutations, single-feature approaches are insufficient for robust classification. Reliable separation therefore demands the integration of multiple complementary attributes. Accordingly, we designed mtFFPEcleaner as a machine learning system following extensive benchmarking of five supervised algorithms—k-nearest neighbors, logistic regression, RF, SVM, and XGBoost—across six key performance measures (accuracy, precision, F1-score, AUC, specificity, and sensitivity) using tenfold cross-validation. Training relied on a ground-truth collection of 1169 artifacts and 837 genuine mutations sourced from 23 matched FFPE-FF sample pairs.

Random forest (RF) consistently outperformed the other methods **(Figure 5a)**, benefiting from its ensemble architecture, capacity to process high-dimensional inputs, and inherent resistance to overfitting [25]. Experiments varying training set size showed progressive performance gains up to roughly 250 artifacts, beyond which additional data yielded diminishing returns **(Figure 5b)**. Statistical assessments (Kolmogorov-Smirnov test and Cramér's V) on multiple bootstrapped subsets confirmed that feature diversity saturated at this point. Systematic exploration of class imbalance demonstrated that raising the artifact proportion initially boosted accuracy, precision, specificity, and F1-score, but excessive artifact representation led to deterioration in these metrics and reduced sensitivity, while AUC remained largely unaffected. Balancing the need to control false positives with preservation of sensitivity for low-level variants, a 1:2 ratio of genuine mutations to artifacts provided the most favorable trade-off, delivering peak values for accuracy, precision, specificity, and F1-score alongside adequate sensitivity **(Figure 5c)**. Ranking of predictor contributions highlighted VAF, the fraction of mutations in the 1%–90% VAF range, and SOB score as the top influential features **(Figure 5d)**. To safeguard against overfitting, substitution-type features were intentionally excluded from the final model. A dedicated sensitivity analysis

verified this choice: models incorporating substitution categories suffered greater overfitting, shown by a steeper validation AUC-ROC decline ($\Delta = -0.023$ versus -0.007) and exaggerated dependence on mutation type (mean decrease in accuracy of 85 versus 0). The retained model successfully captured context effects through the trinucleotide feature.

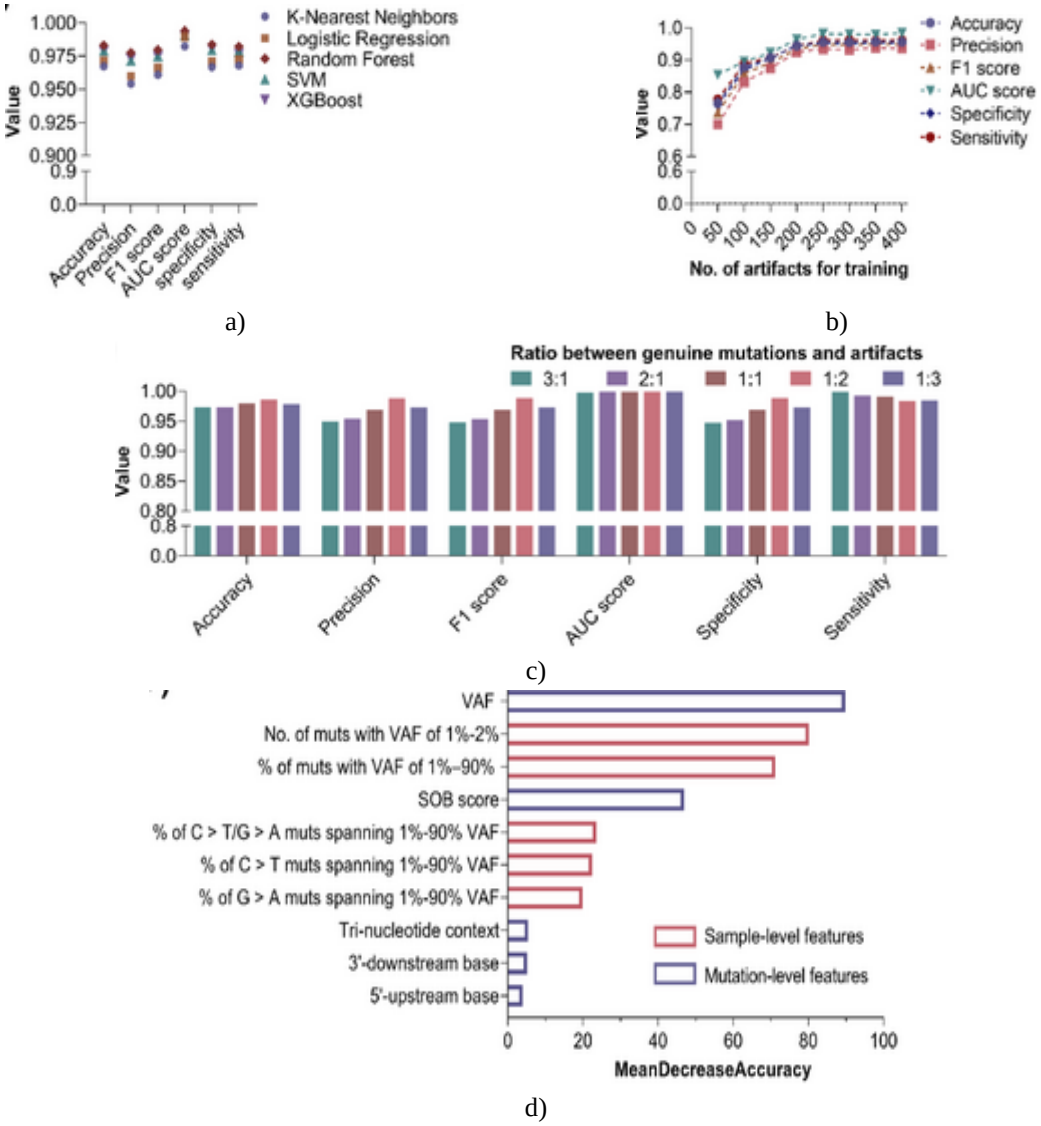


Figure 5. Construction and optimization of mtFFPEcleaner for identification of formalin-induced artifacts. (a) Comparative evaluation of machine learning classifiers trained on 1169 ground-truth artifacts and 837 genuine mtDNA mutations from 23 FFPE samples. (b) Relationship between training set size and model performance. Models were built with progressively larger numbers of artifacts (x-axis) and matched genuine mutations across 10 independent iterations. (c) Influence of artifact-to-genuine mutation ratio on performance metrics. Five training sets of fixed size (n = 1200) were tested at ratios of 3:1, 2:1, 1:1, 1:2, and 1:3 (10 independent runs per condition). (d) Feature importance measured by mean decrease in accuracy in the optimized mtFFPEcleaner model. AUC, area under the curve; SOB, strand orientation bias; SVM, support vector machine; VAF, variant allele frequency

mtFFPECleaner outperforms state-of-the-art software in independent validation set

A dedicated independent validation cohort was assembled to rigorously test mtFFPECleaner's effectiveness at recognizing and eliminating formalin-induced artifacts in FFPE specimens. This cohort consisted of mtDNA NGS data from 15 paired FFPE-FF tissue samples and included 1064 artifacts together with 509 genuine mtDNA mutations detected in the FFPE material. Upon classification with mtFFPECleaner, 1046 of the 1064 artifacts

were accurately identified (specificity 0.9830), while only 4 of the 509 genuine mutations were erroneously excluded (sensitivity 0.9921) **(Figure 6a)**. These figures indicate highly reliable discrimination. Performance was then directly compared against two prominent existing tools developed for nuclear DNA artifact removal: SOBDetector [15] and DEEPOMICS FFPE [17]. Both tools showed inferior results on the same validation data. SOBDetector preserved almost all genuine mutations (508 of 509, sensitivity 0.998) but filtered merely 2.2% of artifacts (23 of 1064) **(Figure 6b)**, leading to a situation in which only 32.5% of retained calls represented true mutations. DEEPOMICS FFPE correctly classified 444 genuine variants out of 633 predicted mutations (positive predictive value 70.1%), leaving 189 artifacts unfiltered **(Figure 6c)**. Precision-recall curve evaluation further highlighted the advantage of mtFFPECleaner, which reached an AUC-PRC of 0.9989. In comparison, SOBDetector attained 0.7306 and DEEPOMICS FFPE reached 0.8859 **(Figure 6d)**. Additional performance measures also favored mtFFPECleaner significantly over the competing tools **(Figure 6e)**. Overall, these findings position mtFFPECleaner as a specialized, high-accuracy resource that markedly exceeds current nuclear-DNA-oriented solutions when applied to mtDNA sequencing data from FFPE samples.

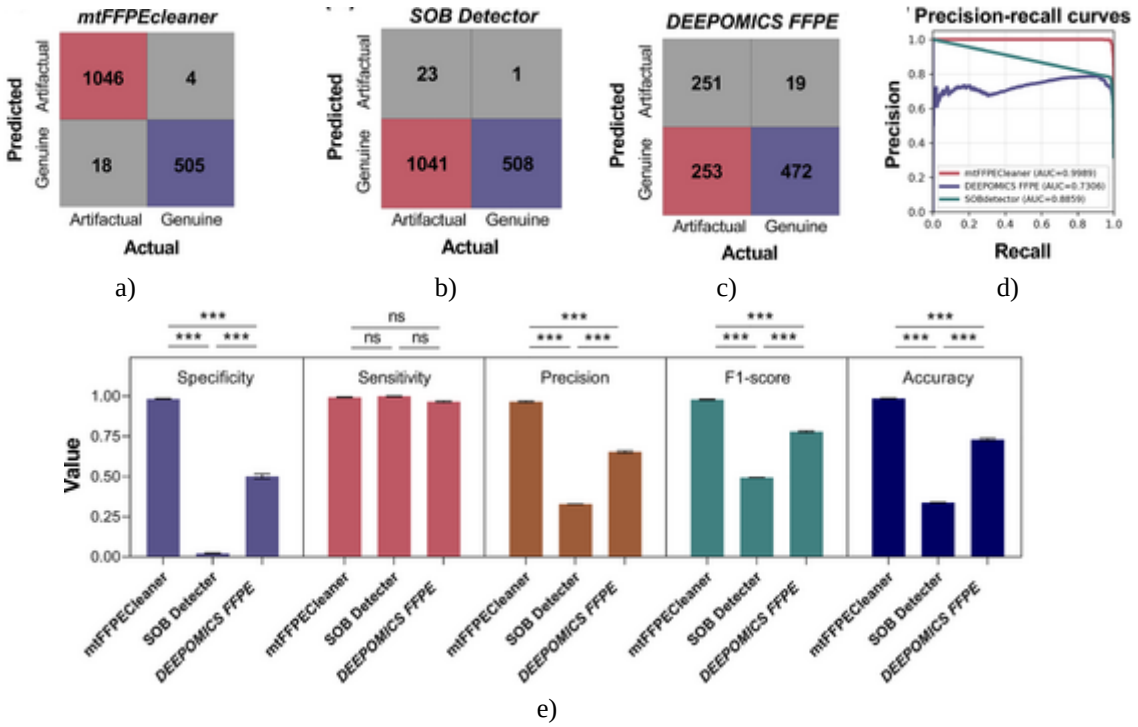


Figure 6. mtFFPECleaner outperforms state-of-the-art software in an independent validation set. Classification performance was assessed using 1064 formalin-induced artifacts and 509 genuine mtDNA mutations identified in mtDNA next-generation sequencing data from 15 paired FFPE-FF samples. (a–c) Confusion matrices for predictions of artifacts and genuine mutations in FFPE tissues by mtFFPECleaner (a), SOB Detector (b), and DEEPOMICS FFPE (c). (d) Precision-recall curves generated for mtFFPECleaner, SOB Detector, and DEEPOMICS FFPE on the validation cohort. (e) Model comparison based on six standard performance metrics. Results shown as mean ± SEM. One-way ANOVA with Bonferroni's post hoc test was applied in panel (e). FFPE, formalin-fixed paraffin-embedded; PRC, precision-recall curve. ***$p < 0.001$.

Mutational signatures of mtDNA in FFPE tissues match those in fresh-frozen tissues after filtering artifacts by mtFFPECleaner

Additional evaluation of mtFFPECleaner's practical value involved applying the tool to private FFPE cohort 1 (314 ovarian cancer samples) to quantify artifact reduction and its effect on mutation profiles. This process flagged 4293 artifacts among 15,088 total mutations detected. The mutational spectra of the classified artifacts differed clearly from those of retained genuine mutations **(Figure 7a)**. Cosine similarity measurements confirmed that genuine mutations retained in FFPE samples exhibited profiles highly consistent with mutations observed in 864 fresh-frozen colorectal cancer samples, while standing apart from the artifact group **(Figure 7b)**. Samples were again stratified by mutation load to examine artifact patterns and the impact of filtering.

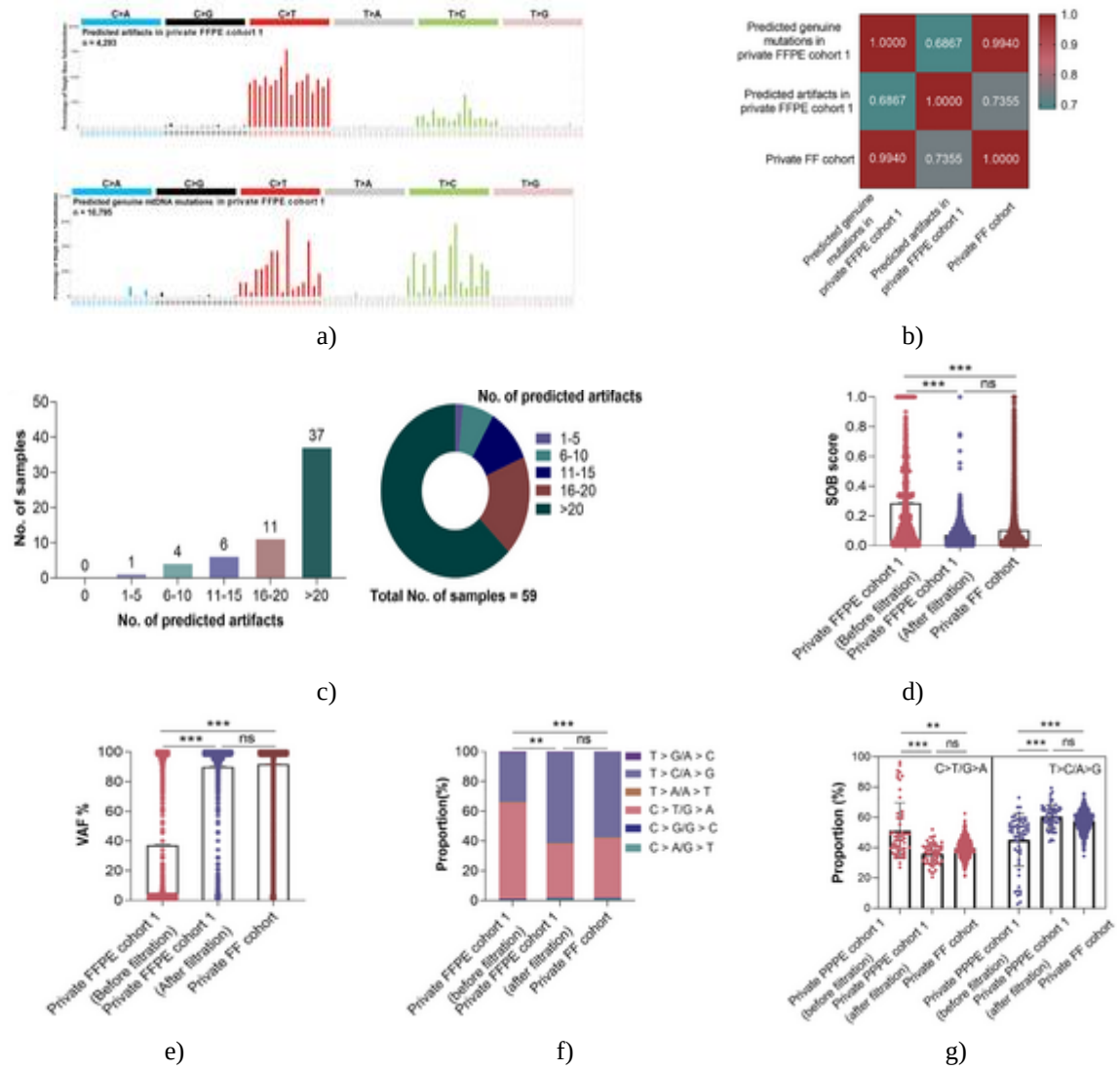


Figure 7. Mutational signatures of mtDNA in FFPE tissues match those in fresh-frozen tissues after filtering artifacts by mtFFPECleaner.

(a) Ninety-six-component mutational spectra of 4293 formalin-induced artifacts and 10,795 genuine mtDNA mutations identified in 314 private FFPE cohort 1 samples following mtFFPECleaner classification. (b) Pairwise mutational spectrum similarity across the three mutation categories. (c) Distribution of artifacts within high-load private FFPE samples (n = 59), shown as proportions of samples containing different artifact counts. (d–f) SOB scores (d), VAF values (e), and substitution type distributions (f) in high-load FFPE samples before versus after artifact removal, compared with private FF samples. (g) Proportions of C > T/G > A and T > C/A > G mutations in high-load FFPE samples before and after filtration relative to private FF samples. FF, fresh-frozen; FFPE, formalin-fixed paraffin-embedded; VAF, variant allele frequency; SOB, strand orientation bias. Results presented as mean ± SEM. Cosine similarity was used in panel (b). One-way ANOVA with Bonferroni's post hoc test was performed in panels (d, e, g). Chi-square test was applied in panel (f). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

All high-load FFPE samples contained artifacts, and 62.7% (37 of 59) harbored over 20 artifacts each **(Figure 7c)**, highlighting the critical need for correction in this group. Following removal, SOB scores in high-load samples decreased to levels comparable with FF samples **(Figure 7d)**. VAF distributions similarly shifted toward higher values, aligning with FF patterns **(Figure 7e)**. Pre-correction substitution profiles in high-load samples were strongly biased **(Figure 7f)**, featuring excess C > T/G > A and reduced T > C/A > G mutations **(Figure 7g)**. After filtering, substitution patterns in high-load FFPE samples became nearly identical to those in FF samples **(Figure 7f)**.

Low-load FFPE samples showed limited artifact presence, with 67.8% (173 of 255) free of detectable artifacts and few exceeding 10 (. Consequently, VAF, SOB, and substitution characteristics in low-load samples changed little after filtering and remained consistent with FF samples. Parallel benefits were documented in private FFPE cohort 2 (gastric and thyroid cancer cases) and the public CRC FFPE cohort. Analysis of somatic mutations in 1015 tumor samples initially detected 3375 variants. mtFFPECleaner removed 54.9% (1846 of 3357) of these. Absent correction, mutation burden estimates would be inflated, heteroplasmy levels understated, and detection of functional changes potentially confounded by excess synonymous variants. These outcomes illustrate how mtFFPECleaner produces more faithful representations of somatic mtDNA alterations in FFPE material, facilitating reliable biomarker exploration. Finally, artifact quantity correlated positively with FFPE storage duration, both in total count (Spearman's $r = 0.2293$, $p < 0.0001$) and maximum VAF (Spearman's $r = 0.3893$, $p < 0.0001$). In summary, mtFFPECleaner excels at removing formalin-induced artifacts, particularly in heavily affected samples, thereby enabling precise and trustworthy analysis of mtDNA mutations in archival FFPE collections.

The extensive adoption of formalin-fixed paraffin-embedded (FFPE) specimens within clinical biobanks has transformed retrospective biomarker research by linking molecular changes to detailed longitudinal clinical data [6]. Nevertheless, formaldehyde fixation causes widespread molecular injury, such as protein-DNA crosslinks, cytosine deamination through hydrolysis, and DNA strand fragmentation. These processes substantially hinder accurate identification of biologically relevant variants [7]. In the present work, we performed an in-depth characterization of formalin-induced artifacts in mtDNA next-generation sequencing data from FFPE tissues and developed mtFFPECleaner—the first dedicated computational platform designed to separate such artifacts from authentic mtDNA mutations. The framework incorporates multiple features grounded in the unique biology of the mitochondrial genome.

Our detailed examination identified three primary differences between formalin-induced artifacts and genuine mtDNA mutations that diverge from the patterns typically seen in nuclear DNA damage. First, the strong strand-biased C > T/G > A transitions at low variant allele frequencies (1%–2%) displayed sequence context preferences distinct from those in nuclear DNA. Whereas nuclear artifacts often arise at CpG dinucleotides due to methylation-associated deamination, mtDNA artifacts showed enrichment at TCN trinucleotides, implying that formaldehyde-related instability may be heightened in thymine-rich mitochondrial sequences. Second, the inverse relationship between strand orientation bias (SOB) scores and VAF values pointed to a hierarchy of damage effects: low-VAF artifacts (<5%) exhibited pronounced strand bias (SOB >0.8), consistent with single-strand lesions, while higher-VAF artifacts (>20%) showed diminished bias, possibly resulting from clonal propagation of damaged mtDNA copies during sample processing. Third, discrepancies in allele frequencies of near-homoplasmic mutations between matched FFPE and fresh frozen samples indicate that formalin damage can alter even high-confidence calls. This observation emphasizes the requirement for specialized analytical corrections to differentiate technical distortions from true shifts in heteroplasmy. Furthermore, artifact accumulation increased by approximately 2.3% per year of storage for C > T/G > A mutations, demonstrating that formaldehyde-induced damage continues as a progressive process over decades, contrary to the assumption that FFPE samples remain molecularly stable.

The creation of mtFFPECleaner fulfills a significant gap in mtDNA analyses involving FFPE material, where distinguishing overlapping characteristics of artifacts and true mutations presents a major difficulty. Conventional strategies, including enzymatic repair or single-feature bioinformatic filters such as SOB scoring [15, 26], prove insufficient given the multifaceted nature of mtDNA artifacts. The strength of mtFFPECleaner stems from its integration of mitochondrial-specific biology with advanced computational modeling [18]. By employing machine learning, the tool combines VAF patterns, SOB metrics, sequence context, and local nucleotide composition. The random forest ensemble approach facilitates nonlinear feature interactions while controlling overfitting through careful cross-validation—an important consideration given the limited availability of paired FFPE-FF reference datasets [25]. An additional refinement involved optimizing the training composition to a 1:2 ratio of genuine mutations to artifacts, which enhanced specificity while preserving adequate sensitivity. This emphasis on specificity is especially important for clinical translation, where false-positive calls risk misinterpretation. Model performance stabilized once the training set reached approximately 250 artifacts, reflecting saturation of relevant feature space and indicating that relatively compact datasets can support effective modeling. Independent validation across multiple cohorts, including paired HCC samples used for training and a large ovarian cancer FFPE collection (n = 314 samples from 157 patients), confirmed broad applicability. Although trained primarily on hepatocellular carcinoma material, successful transfer to ovarian cancer samples

supports the chemistry-based, rather than tissue- or cancer-specific, origin of these artifacts [10, 13]. The observed decrease in low-frequency C > T/G > A variants and removal of storage-time-dependent artifacts further validate mtFFPECleaner's capacity to restore high-quality mtDNA profiles from long-term archival specimens.

False-positive mutations in FFPE-derived mtDNA data introduce two important analytical distortions in clinical contexts. First, they inflate estimates of mutational burden, particularly in high-load samples dominated by low-VAF artifactual C > T/G > A transitions, which may skew prognostic and predictive assessments. Second, they alter mutational signatures, concealing genuine biological patterns such as replication-associated mutations, strand-specific G > A and T > C changes, and corresponding deficiencies in C > T and A > G substitutions on the light strand [14]. The clinical relevance of these findings is considerable. By generating mtDNA profiles comparable to those from fresh frozen tissues, mtFFPECleaner facilitates reliable identification of clinically meaningful heteroplasmies in archival collections. This capability holds particular value for investigating mitochondrial contributions to neurodegenerative diseases and cancer progression, fields that rely heavily on longitudinal FFPE resources [7, 27]. Additionally, the observed relationship between artifact load and storage duration offers a practical quality metric for biobanks, enabling prioritization of samples based on molecular preservation before large-scale mtDNA investigations [8].

Several limitations should be acknowledged. The current implementation addresses only single-nucleotide variants (SNVs), leaving structural variants such as the common 4977 bp deletion—frequently linked to inherited mitochondrial disorders—for future extension [28]. The 1% VAF cutoff may also exclude extremely rare somatic mutations, especially in ultra-deep sequencing experiments. Additional technical noise from polymerase errors or platform-specific artifacts requires ongoing refinement through methods such as molecular barcoding, orthogonal validation across platforms, and tailored error modeling for archival workflows.

Conclusion

In summary, this study introduces a novel framework for harnessing FFPE specimens in mitochondrial genomics. By addressing the longstanding challenge of formaldehyde's preservative yet mutagenic properties, mtFFPECleaner converts archival samples into dependable sources of high-fidelity mtDNA information. As clinical biobanks expand their focus on mtDNA heteroplasmy for diagnostic purposes, this tool offers a vital resource for realizing the full translational potential of FFPE archives in precision medicine.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Stewart JB, Chinnery PF. Extreme heterogeneity of human mitochondrial DNA from organelles to populations. Nat Rev Genet. 2021;22(2):106-18. doi:10.1038/s41576-020-00287-x
2. Kim M, Mahmood M, Reznik E, Gammage PA. Mitochondrial DNA is a major source of driver mutations in cancer. Trends Cancer. 2022;8(12):1046-59. doi:10.1016/j.trecan.2022.08.001
3. Vodicka P, Vodenkova S, Danesova N, Vodickova L, Zobalova R, Tomasova K, et al. Mitochondrial DNA damage, repair, and replacement in cancer. Trends Cancer. 2025;11(1):62-73. doi:10.1016/j.trecan.2024.09.010
4. Shelton SD, House S, Melo LMN, Ramesh V, Chen Z, Wei T, et al. Pathogenic mitochondrial DNA mutations inhibit melanoma metastasis. Sci Adv. 2024;10(44):eadk8801. doi:10.1126/sciadv.adk8801
5. 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. Ann Oncol. 2019;30(6):953-62. doi:10.1093/annonc/mdz105
6. Wang K, Kumar T, Wang J, Minussi DC, Sei E, Li J, et al. Archival single-cell genomics reveals persistent subclones during DCIS progression. Cell. 2023;186(18):3968-82.e15. doi:10.1016/j.cell.2023.07.024

7. Steiert TA, Parra G, Gut M, Arnold N, Trotta JR, Tonda R, et al. A critical spotlight on the paradigms of FFPE-DNA sequencing. Nucleic Acids Res. 2023;51(14):7143-62. doi:10.1093/nar/gkad519
8. Xie F, Guo W, Wang X, Zhou K, Guo S, Liu Y, et al. Mutational profiling of mitochondrial DNA reveals an epithelial ovarian cancer-specific evolutionary pattern contributing to high oxidative metabolism. Clin Transl Med. 2024;14(1):e1523. doi:10.1002/ctm2.1523
9. Zhao Q, Wang F, Chen YX, Chen S, Yao YC, Zeng ZL, et al. Comprehensive profiling of 1015 patients' exomes reveals genomic-clinical associations in colorectal cancer. Nat Commun. 2022;13(1):2342. doi:10.1038/s41467-022-30062-8
10. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem. 2015;61(1):64-71. doi:10.1373/clinchem.2014.223040
11. Chen G, Mosier S, Gocke CD, Lin MT, Eshleman JR. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. Mol Diagn Ther. 2014;18(6):587-93. doi:10.1007/s40291-014-0113-2
12. Yi Z, Zhang X, Wei X, Li J, Ren J, Zhang X, et al. Programmable DNA pyrimidine base editing via engineered uracil-DNA glycosylase. Nat Commun. 2024;15:6397. doi:10.1038/s41467-024-50714-9
13. Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. The mutational signatures of formalin fixation on the human genome. Nat Commun. 2022;13:4487. doi:10.1038/s41467-022-32153-9
14. 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. doi:10.1038/s41588-020-0541-7
15. Diossy M, Sztupinszki Z, Krzystanek M, Borcsok J, Eklund AC, Csabai I, et al. Strand Orientation Bias Detector to determine the probability of FFPE sequencing artifacts. Brief Bioinform. 2021;22(6):bbab186. doi:10.1093/bib/bbab186
16. Ikegami M, Kohsaka S, Hirose T, Ueno T, Inoue S, Kanomata N, et al. MicroSEC filters sequence errors for formalin-fixed and paraffin-embedded samples. Commun Biol. 2021;4:1396. doi:10.1038/s42003-021-02907-9
17. Heo DH, Kim I, Seo H, Kim SG, Kim M, Park J, et al. DEEPOMICS FFPE, a deep neural network model, identifies DNA sequencing artifacts from formalin fixed paraffin embedded tissue with high accuracy. Sci Rep. 2024;14:2559. doi:10.1038/s41598-024-52827-9
18. Guo W, Liu Y, Ji X, Guo S, Xie F, Chen Y, et al. Mutational signature of mtDNA confers mechanistic insight into oxidative metabolism remodeling in colorectal cancer. Theranostics. 2023;13(1):324-38. doi:10.7150/thno.78718
19. 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 U S A. 2015;112(8):2491-6. doi:10.1073/pnas.1419651112
20. Guo S, Zhou K, Yuan Q, Su L, Liu Y, Ji X, et al. An innovative data analysis strategy for accurate next-generation sequencing detection of tumor mitochondrial DNA mutations. Mol Ther Nucleic Acids. 2021;23:232-43. doi:10.1016/j.omtn.2020.11.032
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303. doi:10.1101/gr.107524.110
23. 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 Res. 2013;41(6):e67. doi:10.1093/nar/gks1443
24. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0
25. Wood DE, White JR, Georgiadis A, Van Emburgh BO, Parpart-Li S, Mitchell J, et al. A machine learning approach for somatic mutation discovery. Sci Transl Med. 2018;10(457):eaar7939. doi:10.1126/scitranslmed.aar7939
26. Cazzato G, Caporusso C, Arezzo F, Cimmino A, Colagrande A, Loizzi V, et al. Formalin-fixed and paraffin-embedded samples for next generation sequencing: problems and solutions. Genes (Basel). 2021;12(10):1472. doi:10.3390/genes12101472

27. Basyuni S, Heskin L, Degasperi A, Black D, Koh GCC, Chmelova L, et al. Large-scale analysis of whole genome sequencing data from formalin-fixed paraffin-embedded cancer specimens demonstrates preservation of clinical utility. Nat Commun. 2024;15:7731. doi:10.1038/s41467-024-51577-2
28. Ng YS, Bindoff LA, Gorman GS, Klopstock T, Kornblum C, Mancuso M, et al. Mitochondrial disease in adults: recent advances and future promise. Lancet Neurol. 2021;20(7):573-84. doi:10.1016/S1474-4422(21)00098-3