OXFORD

BIOLOGY Methods & Protocols

Biology Methods and Protocols, 2025, **10(1)**, bpaf003
https://doi.org/10.1093/biomethods/bpaf003
Advance Access Publication Date: 10 January 2025
Methods Article

A cost-effective and scalable approach for DNA extraction from FFPE tissues

Christoph Geisenberger[1,2,*], Edgar Chimal[1], Philipp Jurmeister[1,2,3], Frederick Klauschen[1,2,3]

[1]Institute of Pathology, Ludwig-Maximilians-Universität München, Thalkirchnerstr. 36, Munich, 80337, Germany
[2]German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and University Hospital Munich, Pettenkoferstr. 8a, Munich, 80336, Germany
[3]BIFOLD—Berlin Institute for the Foundations of Learning and Data, Einsteinufer 17, Berlin, 10587, Germany

*Corresponding author. Institute of Pathology, LMU Munich, Thalkirchnerstr. 36, Room 202, Munich, 80337, Germany. E-mail: Christoph.Geisenberger@med.uni-muenchen.de

Abstract

Genomic profiling of cancer plays an increasingly vital role for diagnosis and therapy planning. In addition, research of novel diagnostic applications such as DNA methylation profiling requires large training and validation cohorts. Currently, most diagnostic cases processed in pathology departments are stored as formalin-fixed and paraffin embedded tissue blocks (FFPE). Consequently, there is a growing demand for high-throughput extraction of nucleic acids from FFPE tissue samples. While proprietary kits are available, they are expensive and offer little flexibility. Here, we present ht-HiTE, a high-throughput implementation of a recently published and highly efficient DNA extraction protocol. This approach enables manual and automated processing of 96-well plates with a liquid handler, offers two options for purification and utilizes off-the-shelf reagents. Finally, we show that NGS and DNA methylation microarray data obtained from DNA processed with ht-HiTE are of equivalent quality as compared to a manual, kit-based approach.

Keywords: FFPE; NGS; DNA extraction; DNA methylation; molecular pathology; exome sequencing

Introduction

The introduction of tissue preservation with formalin in the late 19th century allowed researchers to conserve a variety of samples for extended periods without significant degradation [1]. Formalin fixation followed by paraffin embedding (FFPE) is still the gold standard for the storage of tissue specimens in diagnostic pathology. FFPE tissue blocks can be kept at room temperature and processing them as tissue sections for stainings or immunohistochemistry is straightforward. While biomolecules such as nucleic acids and proteins are maintained, formalin fixation and extended storage lead to cross-linking, fragmentation, and other types of damage (reviewed in Steiert *et al.* [2]). Sequencing RNA from FFPE tissue is especially challenging [3]. However, technological progress in sequencing library generation has resulted in much improved sensitivity. It is now possible to generate whole genome and whole exome sequencing libraries from a few nanograms of DNA and multiple research groups have achieved single-cell RNA sequencing (scRNA-seq) in FFPE tissues [4, 5]. At the same time, the demand for high-throughput analysis of nucleic acids in FFPE tissues has grown rapidly. This is mostly due to the increased use of next generation sequencing in cancer. Profiling of single-nucleotide variants (SNVs), gene fusions, and somatic copy number aberrations is increasingly important for the diagnostic workup and therapy planning of tumor patients. Furthermore, new diagnostic applications in molecular pathology, such as DNA methylation classification, are showing promise [6–10]. These methods use machine learning and artificial intelligence and therefore require large training and validation cohorts. To scale up mutation and DNA methylation profiling, high-throughput DNA extraction is necessary, ideally from FFPE tissues where most tumor samples are stored. Commercial applications are offered by a number of companies such as Maxwell (HT DNA FFPE) and Covaris (truXTRAC® FFPE SMART). However, they use proprietary reagents and are expensive. This precludes researchers from making informed changes to the protocol which may be necessary for specific applications. In addition, early prototyping and testing can benefit from more cost-efficient protocols. Recently, Oba and colleagues provided an interesting new approach for DNA extraction from FFPE tissues [11]. Termed HiTE (highly concentrated Tris-mediated DNA extraction), their approach utilizes high concentrations of the formalin scavenger Tris (tris[hydroxymethyl]aminomethane). This resulted in higher yields and DNA quality, presumably due to more efficient de-crosslinking. This was also reflected in improved data quality in NGS experiments. This article extends HiTE to a high-throughput format (ht-HiTE) that can be performed manually or using liquid handlers. Also, we demonstrate that both whole exome and DNA methylation profiling of DNA extracted with (ht-)HiTE yield high quality data with equal performance compared to a kit-based manual workflow. Lastly, we offer the research community a detailed online version of the protocol.

Materials and methods

Ethics statement

The research project has been approved by the ethics committee of LMU University Munich. All analyses were retrospective and conducted with leftover material of diagnostic cases.

Received: 12 July 2024; **Revised:** 17 December 2024; **Editorial decision:** 18 December 2024; **Accepted:** 08 January 2025

Code and data availability

Raw DNA microarray data generated in this study are available at figshare (http://dx.doi.org/10.6084/m9.figshare.26198033). Code to reproduce results and generate plots is available on github (https://github.com/cgeisenberger/ht-hite).

Experimental methods availability

A detailed step-by-step version of the protocol has been published at protocols.io and is accessible at dx.doi.org/10.17504/protocols.io.6qpvr3jr3vmk/v2.

Study design

The main research objective of this study was to establish a high-throughput implementation of a FFPE DNA extraction protocol using high concentrations of Tris (ht-HiTE). The workflow was tested manually and compared to the kit-based reference method used in our lab. Next, an automated version was set up to process replicates deposited in two 96-well plates. Finally, DNA microarray profiling of DNAs extracted with the automated workflow was compared to previously generated data to assess the quality of methylation data attainable with ht-HiTE.

Patients and samples

Samples were selected from the archives of the Institute of Pathology at LMU Munich. Manual testing was performed from replicates of a colorectal cancer sample processed in 2023. Plate-based processing was performed for different subtypes of sarcoma with tissue block ages ranging from 1 to 7 years old ($n = 89$), lung cancer samples between 9 and 12 years old ($n = 3$) and empty controls ($n = 4$). Methylation profiles were generated for lung cancer samples processed with ht-HiTE and cleaned with beads or columns (three samples with two replicates, $n = 6$ total). These data were compared to previous array experiments of DNA extracted using the reference method ($n = 3$).

DNA extraction: Reference method

Tissue sections (2 µm) were stained with hematoxylin and eosin (H&E) to select regions with high tumor cell content. Macroscopic dissection of tumor areas was performed with sterile scalpels and tissues were extracted using the Maxwell RSC FFPE Plus DNA Purification Kit (Promega) according to the manufacturer's instructions.

DNA extraction: Automated HiTE

Tissue sections (2 µm) and unstained sections (10 µm) were placed on glass slides (TOMO, TOM-14). After staining 2 µm sections with H&E, areas of interest were macroscopically dissected with a scalpel blade (for example Ruck, 2009010) and placed in 1 ml DNA lo-bind, deep-well 96-well plates (Eppendorf, 0030503244) preloaded with 500 µl of mineral oil (Sigma Aldrich, M5904-500ML). Plates were incubated for 15 min at 56°C in a thermocycler to melt paraffin. Next, 100 µl of the following mix were added to each sample: 80 µl Tris-HCl pH 8.0 (Merck Millipore, 648314-100ML), 10 µl SDS 10% (Sigma Aldrich, 71736-100ML), 5 µl Proteinase K (NEB, NEB, P8107S), 5 µl H_2O (Promega, P1199). Samples were incubated 1 h at 56°C followed by 16–24 hours at 80°C and a holding step at 4°C. Next, the aqueous phase of samples was transferred to fresh plates (Eppendorf, 0030503104) with pipette tips positioned close to the bottom of the well during aspiration as to minimize carryover of oil.

For bead-based purification, samples were mixed with 80 µl of Ampure XP beads (Beckman Coulter, A63882) and incubated for 10 min at room temperature. Next, samples were placed on a magnetic rack to pellet beads. Then, supernatant was aspirated and 200 µl of 70% ethanol was added to each well. After 30 s to 1 min, ethanol was removed. After repeating the wash step once, the beads were dried at room temperature for 10 minutes. Next, the plate was removed from the magnet and 60 µl of nuclease free water was added to each sample. Liquid and beads were mixed by pipetting and incubated for 10 min at room temperature off the magnet. Finally, the plate was placed back on the magnetic rack and the purified DNA sample was transferred to a fresh DNA lo-bind plate.

For column-based purification, sample volume was adjusted to 200 µl with nuclease-free water. Next, purification was performed using the DNAeasy Blood and Tissue spin column kit (Qiagen, 69581) according to the manufacturer's instructions. Briefly, 400 µl of Buffer AL was added to each sample and the full volume was transferred to a spin plate. After centrifugation and two cleaning steps with Buffer AW1 and AW2, samples were eluted in 60 µl of buffer AE. The steps outlined above were automated on a Biomek i5 liquid handling station (Beckman Coulter). Of note, a number of samples ($n = 8$) evaporated during DNA extraction due to imperfect sealing and were excluded from the analysis.

DNA quantification

Nucleic acids were quantified using the dye-based Qubit^{TM} HS DNA Assay (Thermo Fisher) or spectrometrically with the NanoDrop^{TM} One platform (Thermo Fisher).

DNA methylation analysis

After DNA extraction, DNA from FFPE tissues was restored using the Illumina Infinium HD FFPE Restore Kit. Subsequently, DNA was bisulfite converted with the EpiTect Bisulfite Kit (Qiagen). Bisulfite-converted DNA was analyzed on Illumina HumanMethylation Epic microarrays according to the manufacturer's specifications. Arrays were scanned on the Illumina NextSeq 550 platform. The resulting data were processed using the software package *minfi* [12] and normalized using single-sample normal-exponential out-of-band (Noob) normalization [13]. Downstream analysis was performed using the software package *tidyverse*. Copy-number plots were generated using *conumee2.0* [14]. Of note, manually extracted DNA samples were assayed on a different microarray platform (EPIC v1) as ht-HiTE samples (EPIC v2). Data were combined by selecting probes available on both platforms ($n = 718{,}960$). Normal control samples for copy number plots ($n = 5$ for each array type) were downloaded from the NCBI GEO database available under accession GSE235717 (EPIC v1) and GSE246337 (EPIC v2).

Next generation sequencing

DNA was extracted with manual HiTE or the reference method as outlined above. For each sample, 250 ng of DNA were processed using the Exome 2.0 kit from Twist Biosciences according to the manufacturer's recommendations. Processing included eight cycles of PCR for library generation and an additional nine cycles during hybridization capture. DNA was pooled in equimolar ratios after the first PCR and before hybridization. Enriched material was sequenced 2x100 bp on an Illumina NovaSeq platform. Raw sequencing data were adapter- and quality trimmed with *trim_galore* (https://github.com/FelixKrueger/TrimGalore) with an additional hard-trimming of 3 bases from the 3' and 5' end. Mapped was performed with bwa mem [15] in paired-end mode using hg38 as the reference (UCSC version GRCh38). After sorting, PCR duplicates were removed using Picard (http://broadinstitute.github.io/picard) with the command MarkDuplicates

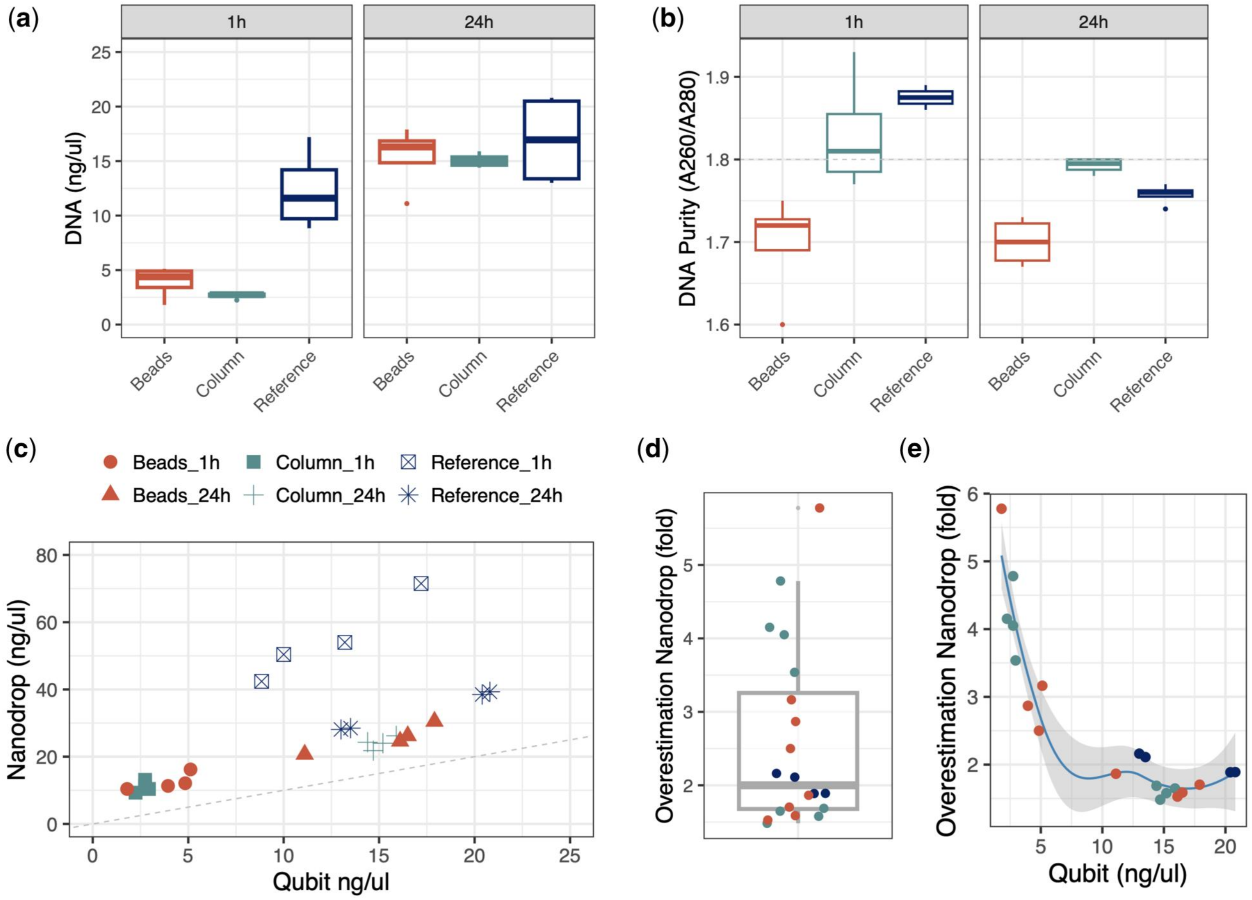

Figure 1. Manual validation of highly concentrated Tris-mediated DNA extraction (HiTE). (**a**) DNA concentrations obtained after 1 (left panel) and 24 h (right panel) of incubation. Samples processed with off-the-shelf reagents and purified with beads or columns yielded roughly equal amounts compared to the reference method. (**b**) DNA purity as measured by the absorption ratio at 260 and 280 nm based on spectrometer readings. A ratio of 1.8 (dashed gray line) is considered pure for DNA samples. (**c**) Dye-based DNA yields (*x*-axis, Qubit) compared to spectrometry-based readings (*y*-axis, Nanodrop) approach. There is a strong linear relationship between both measurements with overestimation of yield by spectrometry. This overestimation is roughly two-fold (**d**) and depends on the amount of DNA (**e**). In general, deviation is larger for smaller DNA concentrations. Reference samples after 1 h were omitted in (d) and (e)

and the flag REMOVE_DUPLICATES set to true. Variants were identified using Strelka2 [16]. Downstream analysis were carried out with custom scripts written in Python and R and included the software packages *Rsamtools*, *DescTools*, *vcfR*, *BSgenome* and *MutationalPatterns*.

Results

Overview of the protocol and manual testing

The high-throughput implementation presented here is based on HiTE, a method published by Oba and colleagues [11]. The protocol includes four basic steps: (i) deparaffinization, (ii) tissue lysis, (iii) reversal of interstrand crosslinks, and (iv) DNA purification. Single Eppendorf tubes or 96-well deep well plates are pre-loaded with 500 µl mineral oil per reaction chamber. Tissue scrolls or fragments scratched from glass slides are placed in each tube or well. Paraffin is removed by immersion in mineral oil and melting, followed by Proteinase K digestion. Formalin-induced crosslinks are reversed by incubation at 80°C overnight in the presence of high concentrations of Tris (800 µM). Finally, DNA can be purified using silica columns or paramagnetic SPRI beads. First, we validated the performance of HiTE. Tissue scrolls (10 µm) from an FFPE sample were processed manually in replicates of four per condition. As a reference, we extracted DNA with the Maxwell® CSC DNA FFPE Kit according to the manufacturer's instructions (from here on *reference method* or simply *reference*). HiTE samples were processed as outlined above and purified with Ampure XP SPRI beads or Qiagen silica columns. Also, we assessed two different incubation times: 24 h as suggested by Oba *et al.* and 1 h. DNA yields were quantified with a dye-based method (Qubit^{TM}) or spectrometrically (Nanodrop^{TM}). Longer incubation yielded significantly more DNA as measured by Qubit (Fig. 1a; $P = 2 \times 10^{-5}$, Student's *t*-test). At 24 h, the average yield was 15.8 ng/µl with similar results for HiTE and the reference method (Fig 1a; $P > .65$, ANOVA). Purity as measured by the A260/A280 ratio was 1.8 for column-based and 1.7 for bead-based purification (Fig. 1b).

For the reference method, purity showed slightly divergent results depending on the incubation time with a ratio of 1.87 after 1 h and 1.76 after 24 h. Concentrations measured by Nanodrop and Qubit showed a strong linear relationship (Fig. 1c). Omitting samples for the reference method revealed a correlation of 0.95 (0.63 when including all samples, Pearson's *r*). However, Nanodrop measurements overestimate DNA content roughly two-fold (Fig. 1d) with a concentration-dependent effect (Fig. 1e). Overestimation plateaus for concentrations of >10 ng/µl (Qubit) or >20 ng/µl (Nanodrop). Size distribution of the extracted molecules was evaluated electrophoretically with Agilent's

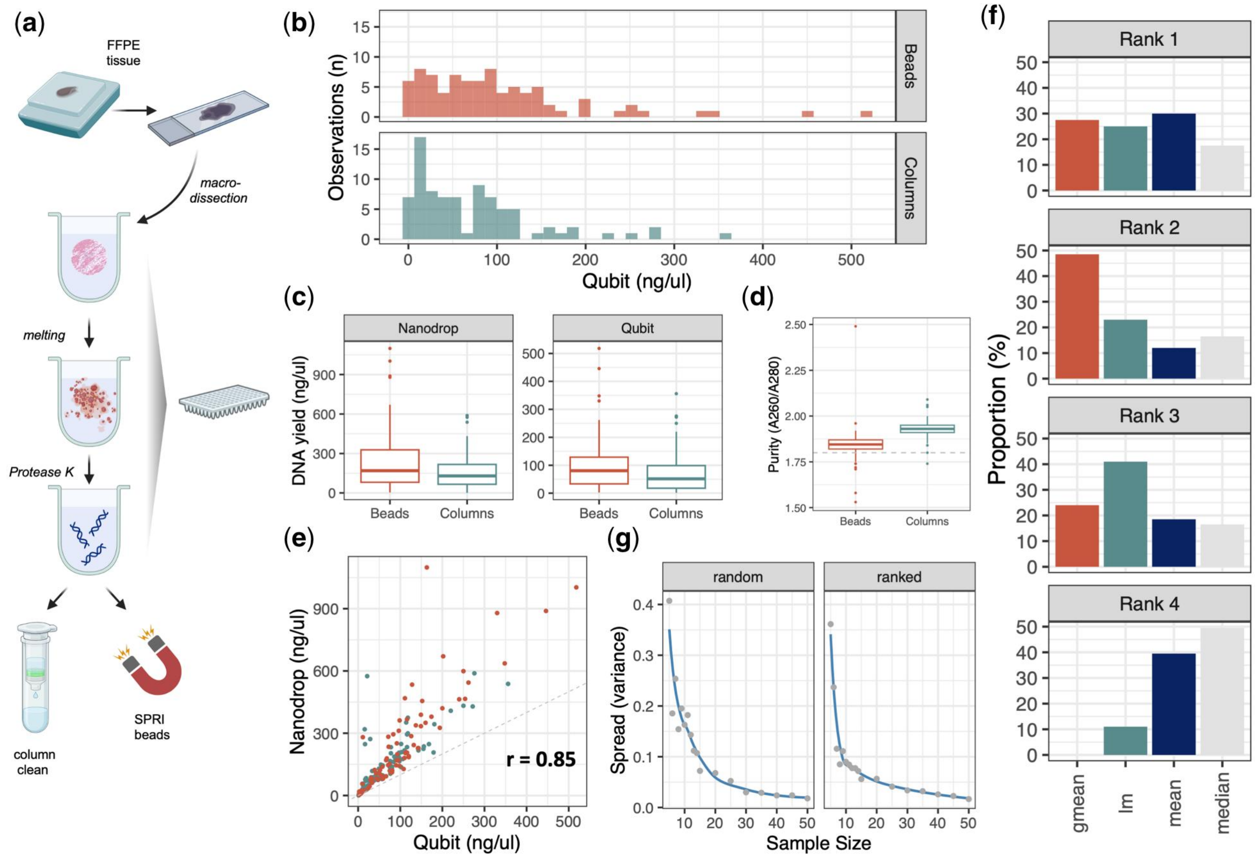

Figure 2. Plate-based processing and correction of spectrometry-based readings. (**a**) Outline of automated sample processing (created with biorender.om). (**b**) Histogram of DNA yield for two replicate 96-well plates purified with beads or columns. (**c**) Same data as in b but represented as a box plot and including Nanodrop measurements. Bead-based purification achieved slightly but significantly higher DNA yields. (**d**) Comparison of UV absorption ratios (260/280 nm ratio) for purification with beads or columns. (**e**) Scatterplot for dye- and spectrometry-based DNA concentration measurements. Values show high correlation (Pearson's $r = 0.85$) with roughly two-fold overestimation by Nanodrop. (**f**) Comparison of four approaches for correcting Nanodrop overestimation. Panels indicate the performance measured by the proportion each approach achieved Rank 1 (best), 2, 3, or 4 (worst) across 200 iterations. Overall the geometric mean (*gmean*) performed best. (**g**) Provides an estimate of the variance observed (y-axis) when estimating the geometric mean for a limited number of samples (x-axis). Samples were either picked randomly (left panel) or evenly spaced according to their DNA concentrations (right panel). The elbow indicates that approximately 20 samples are sufficient for a reliable estimation of the geometric mean. gmean, geometric mean; lm, linear model

Bioanalyzer® platform (Supplementary Fig. S2). Visual inspection of the traces revealed similar profiles regardless of extraction and purification modality. We note that electrophoretic measurements tend to inflate signals for larger fragments due to dye incorporation. However, similar size distributions were later corroborated by NGS. Taken together, our findings validate HiTE as an useful approach for high-quality DNA extraction with off-the shelf reagents.

Plate-based processing

Next, we established an automated version of the protocol on a liquid handler (Beckman Coulter Biomek i5). A total of 92 samples and 4 empty controls were processed in duplicate in two 96-well plates. Processing was identical to the manual samples and DNA extraction was performed with either SPRI beads or columns for one plate each (Fig. 2a). Yield as measured by DNA concentration showed much larger variation than observed for manual processing, owing to the larger differences in tissue size between the processed blocks. Mean and median Qubit readings were 86.6 and 72.0 ng/µl across non-control samples (Fig. 2b, Supplementary Fig. S1a).

Stratifying for the manner of purification revealed small but significant differences between approaches (Fig. 2c). Bead-based purification yielded on average 97.0 ng/µl compared to 68.6 ng/µl for columns as measured by Qubit ($P = 0.023$, *Welch's two-sided t-test*). Nanodrop based measurements were 234.8 and 157.6 ng/µl, respectively ($P = .009$, Welch's two-sided *t*-test). DNA purity as measured by the A260/A280 ratio was 1.84 for bead-based and 1.9 for column-based purification (Fig. 2d). Controls had significantly lower readings with zero measurements for the majority of empty wells ($P = 6.7 \times 10^{-28}$, Welch's two-sided *t*-test, Supplementary Fig. S1b). These results highlight that HiTE can successfully be implemented in an automated setting.

Correction of spectrometry-based DNA measurements

Dye-based measurements are more specific and sensitive than UV-based approaches, but carry a higher price point. Measurements are usually well-correlated with a Pearsons' *r* of 0.85 for samples processed in plates (Fig. 2e). Similar to manual processing, spectrometry-based measurements exhibited a roughly 2-fold overestimation with larger deviations for low DNA concentrations (Supplementary Fig. S1c). Overestimation did not depend on the age of the sample (Supplementary Fig. S1d) and did not depend on purification (Supplementary Fig. S1e). We therefore investigated the following approach to calculate a

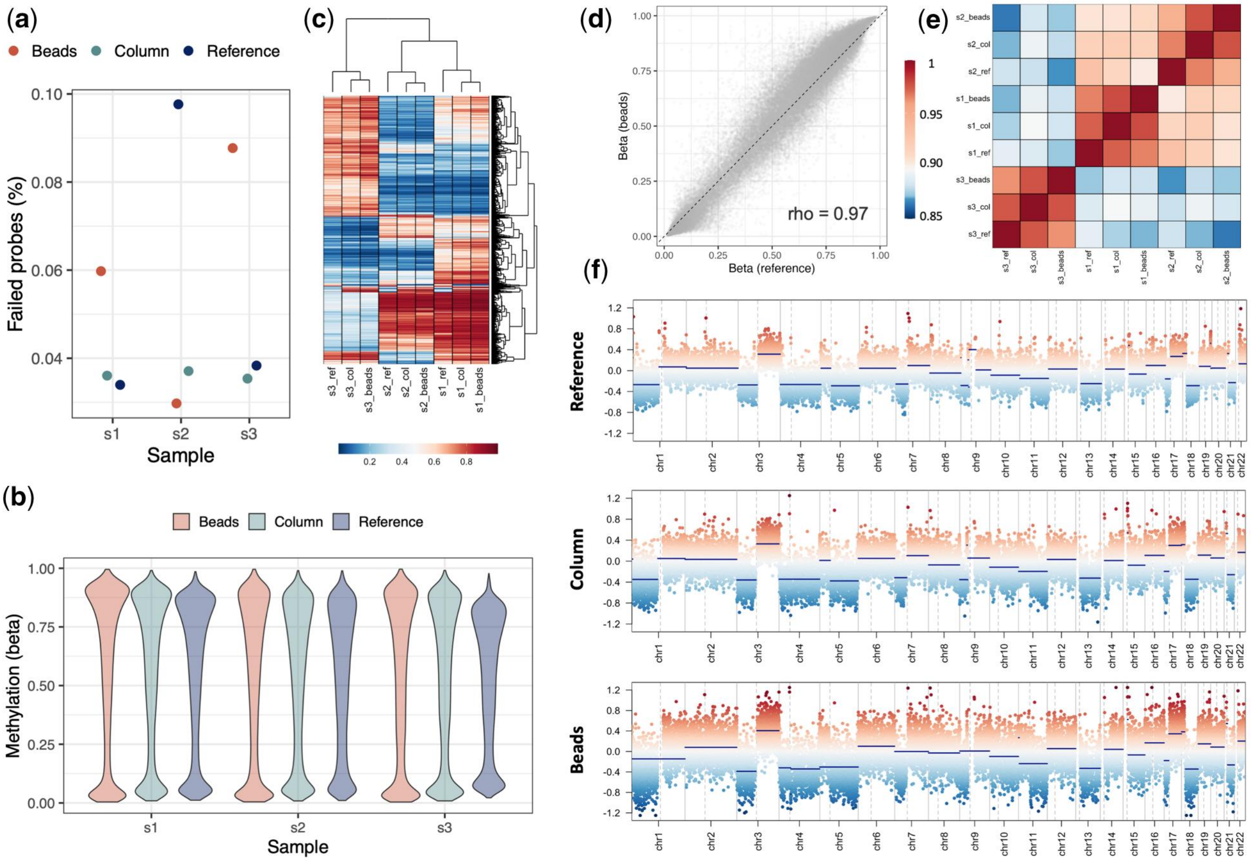

Figure 3. Comparison of methylation data for manual and plate-based processing. (**a**) Proportion of failed probes (detection P-value > .05) on the methylation array stratified by sample and color-coded by purification. All samples exhibited detection rates of >99.9%. (**b**) Distribution of beta values as violin plots for all samples. Values are centered around 1 (full methylation) and 0 (methylation absent) without major differences between purification approaches. (**c**) Heatmap of methylation data for the 5000 most variable probes (rows) on the array. DNA extracted with different methods from the same sample cluster together. (**d**) Example scatter plot comparing beta values of sample 1 for the reference method (x-axis) and ht-HiTE with column purification (y-axis). Reproducibility is high (Spearman's *rho* = 0.97). (**e**) Heatmap for all pairwise correlations (dendrogram not shown). Within-sample correlations are higher than between-sample correlations. (**f**) Genome wide copy-number plots extracted from methylation array data. Copy-number alterations are readily detected regardless of extraction and purification method

plate-specific correction factor while saving reagents by limiting Qubit measurements: (i) obtain Nanodrop measurements for all samples (ii) obtain Qubit measurements for a selected few samples and (iii) calculate a correction factor. First, we identified the most accurate metric to calculate the correction factor. To this end, we performed 200 random splits of the data into a training and test cohort. Then, concentrations in the test cohort were corrected using either a linear model or a correction factor estimated by the median, mean or geometric mean. For each iteration, the four approaches were ranked by their mean squared error (MSE), resulting in a rank between 1 and 4 (1 representing the best performance). Then, rankings were summarized across the 200 random splits. Figure 2f shows how many times each approach was ranked first (Rank 1, top panel), second, third or last (Rank 4, bottom panel). Overall, a simple correction factor based on the geometric mean performed best and was selected for subsequent analyses. Having determined the appropriate way to calculate the correction factor, we assessed the number of samples for which duplicate (Nanodrop & Qubit) measurements are needed to arrive at a reliable estimate. In addition to randomly picking samples, we also investigated a more educated approach. Here, samples were first ranked by their Nanodrop values and then picked in even intervals to cover the full dynamic range. Subsets with increasing numbers of samples were selected and used to estimate Nanodrop overestimation. Variance of the estimate decreased with sample size and showed an elbow for sample sizes of 15–20 (Fig. 2g, Supplementary Fig. S3). While the general shape of the curve was similar to random or concentration-based selection, lower variances were observed for the latter. These results (i) indicate that picking samples based on their concentration to represent the dynamic range of the measurements is superior and (ii) that samples size of 20 are sufficient to obtain reliable estimates. As a validation, we applied this approach for the two plates processed in this study. As outlined above, $n = 20$ samples were selected evenly spaced based on their Nanodrop readings, the geometric mean was used to estimate overestimation and Nanodrop measurements were corrected for the remaining samples. DNA concentrations after *in silico* correction deviated less than two-fold in either direction (i.e. 0.5× or 2×) in 94.5% and 90.4% of cases (bead- and column-based purification, respectively). In summary, correcting Nanodrop readings through Qubit measurements for a subset of samples is a potential way of decreasing costs with acceptable error margins.

Methylation profiling of samples processed in 96-well format

To assess the quality of DNA generated with ht-HiTE, methylation data were generated for three lung cancer samples preserved as FFPE tissue (9–12 years old). After macrodissection of

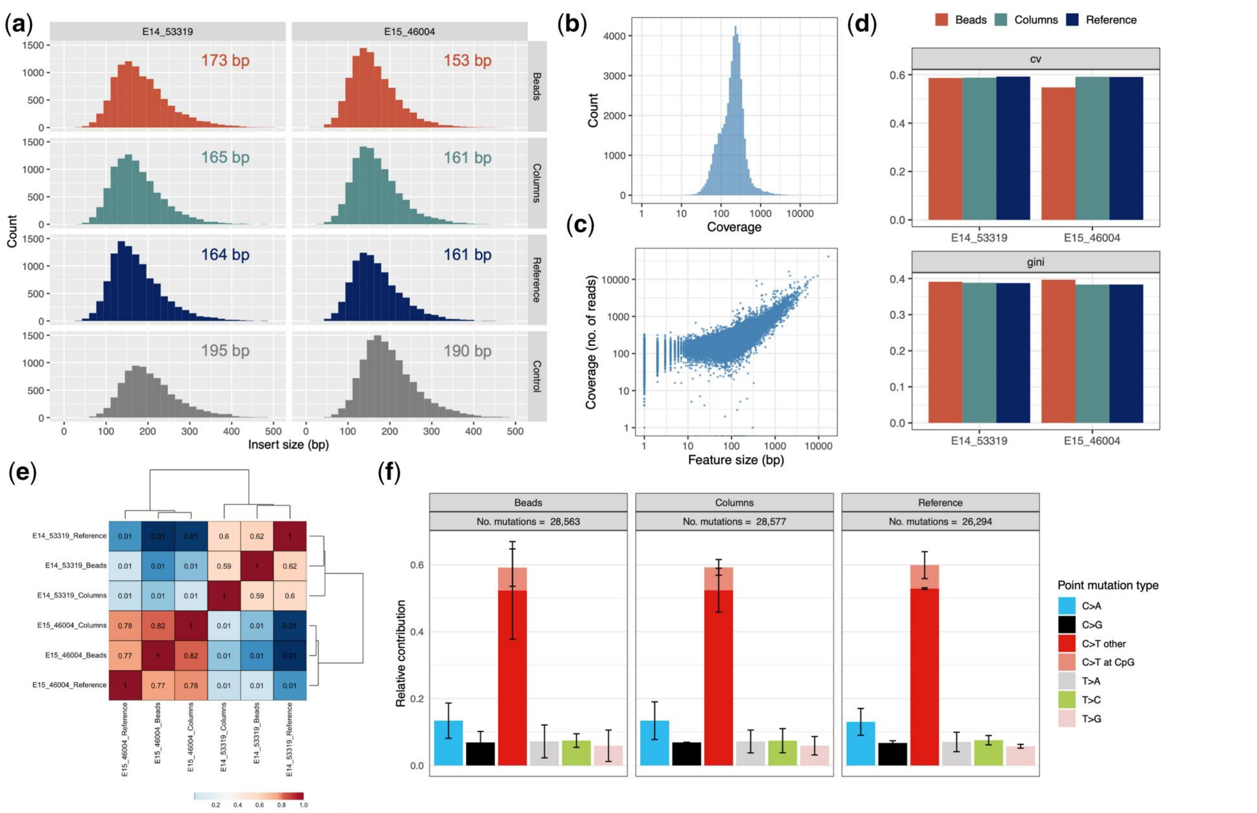

Figure 4. Comparison of NGS results between HiTE and the reference method. (**a**) Distribution of insert sizes/read lengths of exome sequencing data for two lung cancer samples processed in triplicate and matched normal controls. While insert sizes were similar between extraction methods, controls showed larger fragments. (**b**) Example histogram for coverage of target regions in one representative sample (E14_53319 Beads). (**c**) Same data as in (b) but plotted against the size of the target region. (**d**) Comparison of coverage uniformity between the different DNA extraction methods. Coverage uniformity was assessed by calculating the coefficient of variation (CV, upper panel) or Gini coefficient (lower panel). There were no discernible differences between the extraction methods. (**e**) Heatmap showing the pairwise overlaps of SNVs. Overlap was measured by Jaccard's index with variants identified in both (intersection) divided by variants identified in either sample (union). (**f**) Mutational signatures detected in exome sequencing stratified by extraction method (panels). The strongest signature corresponded to C-to-T transitions without differences with respect to extraction modality. Of note, C-to-T transitions are common technical artifacts in observed in FFPE material

Table 1. Cost comparison of plate-based purification.

Reagent	Price per sample (columns), €	Price per sample (beads), €
Glass slides	0.64	0.64
Water	0.03	0.03
96-well plates	0.10	0.10
Mineral oil	0.10	0.10
Tris-HCl	0.16	0.16
SDS 10%	0.01	0.01
Proteinase K	0.24	0.24
RNAse A	0.32	0.32
Spin columns (96w)	3.63	
SPRI beads		1.01
Qubit dsDNA Kit	0.80	0.80
Total:	6.02	3.41
Without RNAse:	5.70	3.09
No RNAse, diluted beads, Nanodrop correction	**5.06**	**1.69**

Using the prices available in Germany as of May 2024, processing costs per sample were calculated for column-based (left) and bead-based (right) purification. Lowest prices can be achieved for bead purification without RNAse treatment, dilution of bead-binding buffer and obtaining dye-based measurements for only a subset of samples.

tumor areas, technical replicates were generated for each sample using DNA from (i) manual processing, (ii) ht-HiTE + columns or (iii) ht-HiTE + beads. First, basic quality measures were assessed. Detection rate, that is, the proportion of probes on the array which yielded usable data, was >99.9% for all samples (Fig. 3a) with a trend toward fewer failed probes for column purification.

Relative methylation as measured by the beta value (methylated signal divided by total signal, M/[U + M]) is bounded between 0 and 1 and typically shows a bimodal distribution. The same was observed when plotting the distribution of beta values across samples (Fig. 3b) without major differences between processing modalities. Selecting the 5000 most variable probes across all samples, clustering of beta values revealed the highest similarities between DNAs from the same tissue (Fig. 3c). Correlation of beta values was very high for technical replicates (Fig. 3d, 0.96–0.98, Pearson's *r*) and higher than between-sample correlations (0.89–0.94, Pearson's *r*). Unsurprisingly, clustering samples based on their pairwise correlations reproduced the high similarity for technical replicates (Fig. 3e and Fig. S4). Finally, genome-wide copy-number plots generated with the software *conumee* [14] showed highly reproducible profiles between DNAs extracted by different means (Fig. 3f, Supplementary Figs S5 and S6).

Next-generation sequencing

To showcase the broader usefulness of our approach for diagnostic and research laboratories, we further performed Illumina short-read sequencing. Specifically, we performed exome sequencing for two lung cancer samples (*E14_53319* and

E15_46004). Tissues were stored as FFPE and 9 and 10 years old. Again, DNA was extracted in triplicate for each tumor using the reference method or HiTE with bead or column purification ($n = 6$ samples in total). Library preparation and enrichment was performed using the Exome 2.0 kit from Twist Biosciences. We sequenced on average 57 million reads per tumor sample. Normal controls for variant calling were available through an unrelated project. Supplementary Table S4 provides an overview of the basic sequencing statistics. Mapping rates were >99% and base quality was high (~97% of bases with quality scores >30). Duplication rate was 14%, and on-target rate was 60% without major differences between extraction modalities. Insert size distributions were also similar for tumor samples whereas control tissues tended to yield larger fragment sizes (Fig. 4a).

While bead purification resulted in slightly longer fragments for E14/53319, we observed the opposite trend for E15/46004. Next, we obtained annotation for the (exomic) target regions and calculated the coverage, ie number of reads, overlapping each region. Fig. 4b provides a histogram for a representative sample. In the whole dataset, coverage per target region spans three orders of magnitude. However, the majority of regions show a much more narrow distribution with a p90/p10 ratio of 5.2. Furthermore, coverage is proportional to the size of the target region with a roughly linear relationship on a log-log scale and a plateau for small regions (Fig. 4c). This could be explained by a probe design which uses a fixed number of probes up to a given target size (~100 bp) and a constant probe-to-size ratio for larger regions. To more formally compare coverage uniformity between extraction methods, Lorentz curves were plotted for each sample (Supplementary Fig. S7). These curves are essentially identical between samples which is also substantiated by the highly reproducible coefficients of variation and Gini indices between samples (Fig. 4d). This implies that coverage biases are caused by technical effects of exome enrichment rather than DNA purification method. Next, we identified somatic variants for each sample. Here, we focused our analysis on SNVs. SNVs were identified by comparison with normal tissue for each sample. After filtering for coverage (>100 reads), overlap of the identified variants was calculated for all pairwise comparisons. More specifically, we calculated Jaccard's index by dividing the variants identified in both samples (intersection) by the variants identified in either sample (union). As expected, overlap between DNAs from the same source was much greater than between samples (Fig. 4e). Again, there was no obvious impact of the extraction method on the overlap. We note that overlap was slightly lower for DNA extracted with the reference method for *E15/46004*. However, this sample also had the lowest number of sequencing reads. Finally, we extracted the mutational signatures [17] for filtered variants and stratified them by extraction method (Fig. 4f). Here, C-to-T transitions were by far the most common signature observed in the dataset. At the same time, mutational patterns were highly reproducible without any differences regarding the DNA extraction method. ($P = .996$, Chi-square test).

Discussion

In this article, we presented and thoroughly tested ht-HiTE, the high-throughput implementation of HiTE, a technique for DNA extraction published by Oba and colleagues [11]. Manual testing validated that HiTE yields appropriate amounts of DNA with good purity, comparable to our reference method, the Maxwell® CSC DNA FFPE Kit. Extended incubation times (24 hours) significantly increased DNA yield compared to shorter times (1 hour). DNA purity, measured by the A260/A280 ratio, was similar between ht-HiTE and the reference method, confirming the method's reliability. In addition, we provided evidence that purification using SPRI beads is effective and reliable. Automating the ht-HiTE protocol on the Beckman Coulter Biomek i5 platform demonstrated its scalability for high-throughput applications. Although DNA yield varied more in automated processing due to differences in tissue size, bead-based purification consistently yielded higher DNA concentrations than column-based purification. The samples processed in this study encompass biopsies and larger tissue sizes, including calcified and adipose tissue, which underscores the broad applicability in terms of tissue types. Electrophoretic measurements further provided evidence of the quality of the extracted DNA and indicated that size distribution of the extracted molecules does not depend on the purification method.

A significant challenge encountered was the overestimation of DNA content by UV spectrometry (Nanodrop) compared to dye-based methods (Qubit). Nanodrop measurements overestimated DNA content by approximately two-fold, especially for lower concentrations. Applying a correction factor based on the geometric mean of overestimation improved the accuracy of DNA quantification. A sample size of around 20 was sufficient to obtain reliable estimates. While we acknowledge that simple correction is not sufficient for the standards of IVDR-compliant workflows, it can save resources in a more permissive research setting. Assuming no RNAse digest and dilution of SPRI beads with lab-made bead binding buffer, DNA can be extracted for as little as 2 euros per sample based on list prices in Germany as of May 2024 (Table 1).

DNA extracted with (ht-)HiTE was also used as input for genetic and epigenetic measurements. We performed microarray DNA methylation profiling for 9 samples which revealed universally high detection rates and consistent beta value distributions. Clustering of the most variable probes and correlation analysis confirmed the high similarity between DNAs extracted by different methods, indicating that the ht-HiTE protocol maintains DNA integrity. Of note, samples using the reference method were assayed on a different array platform (EPIC v1) as HiTE samples (EPIC v2). Nevertheless, correlations between technical replicates are as high as those reported in the characterization of the novel EPIC v2 platform [18]. Exome sequencing for six samples further substantiated our method. Mapping efficiency, base quality, duplication and on-target rates were reproducible between methods and similar to previously published data [19]. Interrogation of read lengths validated our electrophoretic measurements with similar sizes for different extraction modalities. However, we noted that controls extracted from surrounding healthy tissue exhibited slightly larger insert sizes and that numbers in general were lower than reported elsewhere [20]. We assume that shorter fragments in tumor samples might be caused by necrosis, and we also note that the samples used in this study were stored for considerably longer periods than those from Basyuni *et al.* [20]. Lorentz curves and measures such as the coefficient of variation and Gini index revealed no influence of extraction method on coverage uniformity. The overlap between SNVs identified for each sample was high between extraction methods but differed between samples. Overlap as measured by Jaccard's index ranged between 0.6 and 0.8 which is higher than reported elsewhere [21]. We note, however, that this measure is sensitive towards intratumoral heterogeneity. Finally, we investigated the base context of the identified SNVs. SNVs were identical between the extraction methods and exhibit a strong prevalence of C-to-T transitions, which are associated with deamination of cytosine to uracil. While also a byproduct of naturally occurring mutational processes, these variants have been identified as a major technical artifact in sequencing data of FFPE material [21, 22]. Taken together, our data suggest that DNA extracted with (ht-)HiTE is suitable material for NGS and methylation profiling.

To summarize, our findings have significant implications for molecular pathology, particularly in high-throughput DNA extraction and analysis. The ht-HiTE protocol offers a cost-effective, scalable solution for extracting DNA from FFPE tissues, facilitating next-generation sequencing and other high-throughput techniques in cancer diagnostics and research. DNA extraction from FFPE tissues with ht-HiTE is robust and reliable for both manual and automated settings and produces high-quality DNA suitable for different downstream applications. Future studies should aim to optimize the protocol further and explore its applicability to various tissue types and clinical samples.

Author contributions

Christoph Geisenberger (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Visualization [equal]), Edgar Chimal (Data curation [equal]), Philipp Jurmeister (Data curation [equal]), and Frederick Klauschen (Conceptualization [equal], Project administration [equal], Supervision [lead])

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Conflict of interest statement. None declared.

Funding

Christoph Geisenberger is supported by a School of Oncology fellowship awarded by DKTK (Deutsches Konsortium für Translationale Krebsforschung).

References

1. Fox CH, Johnson FB, Whiting J, Roller PP. Formaldehyde fixation. *J Histochem Cytochem* 1985;**33**:845–53.
2. Steiert TA, Parra G, Gut M *et al.* A critical spotlight on the paradigms of FFPE-DNA sequencing. *Nucleic Acids Res* 2023;**51**:7143–62.
3. Liu Y, Bhagwate A, Winham SJ *et al.* Quality control recommendations for RNASeq using FFPE samples based on pre-sequencing lab metrics and post-sequencing bioinformatics metrics. *BMC Med Genomics* 2022;**15**:195.
4. Vallejo AF *et al.* 2022. snPATHO-seq: unlocking the FFPE archives for single nucleus RNA profiling. *bioRxiv*. https://doi.org/10.1101/2022.08.23.505054
5. Xu Z, Zhang T, Chen H *et al.* High-throughput single nucleus total RNA sequencing of formalin-fixed paraffin-embedded tissues by snRandom-seq. *Nat Commun* 2023;**14**:2734.
6. Capper D, Jones DTW, Sill M *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* 2018;**555**:469–74.
7. Jurmeister P, Glöß S, Roller R *et al.* DNA methylation-based classification of sinonasal tumors. *Nat Commun* 2022;**13**:7148.
8. Jurmeister P, Bockmayr M, Seegerer P *et al.* Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci Transl Med* 2019;**11**:eaaw8513. https://www.science.org/doi/10.1126/scitranslmed.aaw8513
9. Hackeng WM, Dreijerink KMA, de Leng WWJ *et al.* Genome methylation accurately predicts neuroendocrine tumor origin: an online tool. *Clin Cancer Res* 2021;**27**:1341–50.
10. Verschuur AVD, Hackeng WM, Westerbeke F *et al.* DNA methylation profiling enables accurate classification of nonductal primary pancreatic neoplasms. *Clin Gastroenterol Hepatol* 2024;**22**: 1245–54.e10.
11. Oba U, Kohashi K, Sangatsuda Y *et al.* An efficient procedure for the recovery of DNA from formalin-fixed paraffin-embedded tissue sections. *Biol Methods Protoc* 2022;**7**:bpac014.
12. Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**:1363–9.
13. Triche TJ, Jr, Weisenberger DJ, Van Den Berg D *et al.* Low-level processing of illumina infinium DNA methylation BeadArrays. *Nucleic Acids Res* 2013;**41**:e90.
14. Daenekas B, Pérez E, Boniolo F *et al.* Conumee 2.0: enhanced copy-number variation analysis from DNA methylation arrays for humans and mice. *Bioinformatics* 2024;**40**:btae029. https://academic.oup.com/bioinformatics/article/40/2/btae029/7582283
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
16. Kim S, Scheffler K, Halpern AL *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;**15**:591–4.
17. Alexandrov LB, Kim J, Haradhvala NJ, PCAWG Consortium *et al.* The repertoire of mutational signatures in human cancer. *Nature* 2020;**578**:94–101.
18. Noguera-Castells A, García-Prieto CA, Álvarez-Errico D, Esteller M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* 2023;**18**:2185742.
19. Zhou J, Zhang M, Li X *et al.* Correction to: performance comparison of four types of target enrichment baits for exome DNA sequencing. *Hereditas* 2023;**160**:35.
20. Basyuni S, Heskin L, Degasperi A, Personalised Breast Cancer Program Group *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.
21. Shi W, Ng CKY, Lim RS *et al.* Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep* 2018;**25**:1446–57.
22. Guo Q, Lakatos E, Bakir IA *et al.* The mutational signatures of formalin fixation on the human genome. *Nat Commun* 2022;**13**:4487.

Biology Methods and Protocols, 2025, 10, 1–8
https://doi.org/10.1093/biomethods/bpaf003
Methods Article