Experimental and Molecular Pathology 126 (2022) 104771

Contents lists available at ScienceDirect

Experimental and Molecular Pathology

journal homepage: www.elsevier.com/locate/yexmp

Fragmentation assessment of FFPE DNA helps in evaluating NGS library complexity and interpretation of NGS results

Anine Larsen Ottestad [a,b], Elisabeth F. Emdal [a,c], Bjørn H. Grønberg [a,b,*], Tarje O. Halvorsen [a,b], Hong Yan Dai [a,c]

[a] Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] Department of Oncology, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway
[c] Department of Pathology, Clinic of Laboratory Medicine, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway

ARTICLE INFO

Keywords:
Next-generation sequencing
Formalin-fixed paraffin-embedded tissue
Library preparation
Quality assessment
DNA fragmentation

ABSTRACT

Formalin-fixed paraffin-embedded (FFPE) tissue remains the most common source for DNA extraction from human tissue both in research and routine clinical practice. FFPE DNA can be considerably fragmented, and the amount of DNA measured in nanograms may not represent the amount of amplifiable DNA available for next-generation sequencing (NGS). Two samples with similar input DNA amounts in nanograms can yield NGS analyses of considerably different quality. Nevertheless, many protocols for NGS library preparation from FFPE DNA describe input DNA in nanograms without indication of a minimum requirement of amplifiable genome equivalent DNA.

An important NGS quality metric is the library complexity, reflecting the number of DNA fragments from the original specimen represented in the final library. Aiming to illustrate the relationship between DNA fragmentation degree and library complexity, we assessed the fragmentation degree of 116 lung cancer FFPE DNA samples to calculate the amount of amplifiable input DNA used for library preparation. Mean unique coverage, coverage uniformity, and mean number of PCR duplicates with the same unique molecular identifier were used to evaluate library complexity.

We showed that the amount of amplifiable input DNA predicted library complexity better than the input measured in nanograms. The frequent discrepancy between DNA amount in nanograms and the amount of amplifiable DNA indicate that the fragmentation degree should be considered when performing NGS of FFPE DNA. Importantly, the fragmentation assessment may help when interpreting NGS data and be a useful tool for evaluating library complexity in the absence of unique molecular identifiers.

1. Introduction

Formalin-fixed paraffin-embedded (FFPE) human tumor tissue samples are collected for routine histopathological diagnostic procedures. They also represent a vast and valuable resource for molecular analyses and retrospective cancer genetic studies. However, the quality of DNA from FFPE samples varies largely when compared with DNA isolated from fresh-frozen tumor tissue. When preparing FFPE samples, formalin functions as a cross-linking agent for tissue fixation and stabilizes the tissue structure by creating covalent linkage between macromolecules, such as DNA-DNA, DNA-protein, and protein-protein. Reversing the formalin-formed cross-linking during DNA extraction causes fragmentation of FFPE DNA. In addition, formalin causes the release of purine bases from nucleic acids and induces DNA fragmentation (Do and Dobrovic, 2015; Srinivasan et al., 2002). Since the extent of fixation may vary among samples, the extent of fragmentation may also vary. The quality of FFPE DNA directly affects the quality metrics of downstream NGS analyses, such as library size, average read depth and uniformity (Robbe et al., 2018; Spencer et al., 2013). Although optimizing tissue fixation conditions and DNA extraction methods improve FFPE DNA quality (Einaga et al., 2017; Heydt et al., 2014; McDonough et al., 2019), the quality is still influenced by many stochastic factors, such as time until fixation, perioperative ischemic time, fixation time and size of tissue samples, storage time and extent of necrosis in tissue

* Corresponding author at: Department of Clinical and Molecular Medicine, NTNU, Kunnskapssenteret, Olav Kyrres gate 17, 7030 Trondheim, Norway.
E-mail address: bjorn.h.gronberg@ntnu.no (B.H. Grønberg).

https://doi.org/10.1016/j.yexmp.2022.104771
Received 10 September 2021; Received in revised form 13 March 2022; Accepted 9 April 2022
Available online 12 April 2022

samples (Bass et al., 2014). Many studies have demonstrated the feasibility of using FFPE DNA for next generation sequencing, including gene specific targeted sequencing, whole-exome and whole-genome sequencing (Carrick et al., 2015; Hedegaard et al., 2014; Kerick et al., 2011; Robbe et al., 2018), though in many clinical pathology departments, gene specific targeted NGS analyses of FFPE DNA are still the most common routine diagnostic molecular analyses.

In library preparation for targeted next-generation sequencing (NGS), the genomic regions of interest are enriched from input DNA by either hybridization-capture using gene-specific probes or PCR amplicon-based enrichment using gene-specific primers (Chang and Li, 2013). For both approaches, the final sequencing-ready library is generated by PCR enrichment. Thus, the final library includes PCR duplicates that provide no additional value in the data analysis unless they are identified by a common characteristic unique to each set of duplicates. One option is to utilize an inherent characteristic, such as fragment length, but this is not possible for amplicon-based libraries generated using opposing primer pairs. An alternative is to introduce a synthetic characteristic to the input DNA before PCR, such as a unique molecular identifier (UMI), that will be common for all subsequent PCR duplicates (Kinde et al., 2011)

Several NGS library preparation protocols used in routine diagnostics omit the incorporation of UMIs, possibly because this adds another step in the laboratory and requires specialized algorithms for data analysis. One consequence is that these unidentified duplicates contribute to the mean coverage that is used as a quality metric for NGS. However, the mean unique coverage is arguably a more precise quality metric because it reflects the number of unique human genome equivalents (hGEs) from the input DNA represented in the final library, i.e., the library complexity.

High library complexity is desirable to achieve high analytic sensitivity and specificity. McNulty *et al.* prepared NGS libraries from cell culture DNA which exhibits the similar quality as DNA from fresh frozen tissue. By varying the amount of input DNA measured in nanograms, they demonstrated that library complexity was enhanced by increasing the amount of input DNA (McNulty et al., 2020). However, if input FFPE DNA for library preparation is only measured in nanograms, the true amount of DNA available for subsequent PCR in NGS libraries may vary widely because the fragmentation degree of FFPE DNA may differ greatly. A study by McDonough et al. showed that NGS quality metrics such as raw base coverage varied widely among specimens although the same amounts of input DNA in nanogram were used for targeted library preparation (McDonough et al., 2019). Commercial kits for DNA quality assessment, such as the KAPA Human Genomic DNA Quantification and QC kit and a multiplex PCR assay (Life Science Innovations, Qualitative Multiplex PCR Assay) have been used prior to NGS library preparation in many studies (McDonough et al., 2019; Pel et al., 2018). However, publications rarely include descriptions of how the results of DNA quality assessments have been interpreted and used for NGS library preparation and eventually data interpretation.

The aim of this study was to demonstrate how a DNA fragmentation assay might be applied and how FFPE DNA fragmentation might affect subsequent NGS library complexity. We performed NGS analysis of 116 DNA samples extracted from FFPE lung cancer samples using the QIAseq Human Actionable Solid Tumor panel. The NGS libraries were prepared with UMIs. Thus, the number of UMIs for each sequenced region reflected the complexity of each library. We also assessed the fragmentation degree of input DNA and calculated the true quantity of DNA fragments as potential templates for amplification. Using this approach, we demonstrated that in the case of FFPE DNA samples, the quantity of input DNA according to the amount of amplifiable DNA fragments rather than amount in nanograms better reflected the number of potential hGE templates and thus provided better prediction of the complexity of NGS.

2. Material and methods

2.1. Samples and approval

In total, 116 tumor samples from 114 lung cancer patients diagnosed between 2007 and 2018 at St. Olav's University Hospital, Trondheim, Norway, were retrieved from Biobank1, our regional lung cancer biobank. The biobank was approved by the Norwegian Regional Committee for Medical and Health Research Ethics (REC) Central, the Norwegian Health Department, and the Norwegian Data Protection Authority. The REC Central also approved the present study. Patients had a median age of 68 (range 46–86), 48% were female, 92% had a performance status of 0–1 and 29% had stage IV disease. Most tumors were adenocarcinomas (n = 103), and the rest were adenosquamous carcinoma (n = 1), large cell neuroendocrine carcinoma (n = 2), small cell carcinoma (n = 1) and non-small cell lung cancer not otherwise specified (n = 9).

2.2. DNA extraction from FFPE tissue

DNA was extracted from archival FFPE tumor blocks. At the Department of Pathology at St. Olav's University Hospital, phosphate-buffered 4% formaldehyde solution (HistoLab Products AB, Gothenburg, Sweden) with a pH of 7.2–7.4 is used as fixation solution. For the large surgically resected samples, the fixation time is between 3 and 5 days; for small biopsies it is overnight for approximately 12–16 h. Fixation is then performed for another two hours in a tissue processor before paraffin embedding. Fixation of tiny needle biopsies are carried out directly in the tissue processor. Two to five tissue sections of 10 μm were cut from the areas with highest tumor cell density identified by an experienced lung cancer pathologist by regular light microscopy. The number of sections was determined empirically according to the size of the defined area and the tumor cell density. DNA was extracted using the QIAcube Connect (Qiagen, Valencia, CA) and the QIAamp DNA FFPE Tissue Kit (Qiagen, Valencia, CA) and then eluted in 200 uL of the supplied buffer. DNA concentration was measured fluorometrically by Qubit® (Thermo Fisher Scientific, Waltham, MA) using either the dsDNA BR or the HS Assay Kit depending on the yield.

2.3. Fragmentation assessment of FFPE DNA

Quantitative real-time PCR (qPCR) of *FCGR3b* with a fragment length of 300 bp was performed to assess the fragmentation degree in FFPE DNA. The *FCGR3b* gene was chosen for practical reasons, since it is already in routine use and validated at our pathology department. This gene is not known to be amplified in lung cancer, and differential expression or epigenetics would not influence the quantification by qPCR. In this study, we used quantification of the amplifiable copy number of *FCGR3b* not as an absolute number, but rather as a condition to determine the amount of amplifiable input DNA within a certain range. Only amplification with extremely high copy numbers will influence this range assessment, and *FCGR3b* is not known for this kind of amplification. The abundance of this fragment was measured relative to an unfragmented DNA control sample extracted from leukocytes from a healthy person. We assumed that DNA fragmentation caused by formalin fixation occurs randomly and, therefore, also within this gene. We assumed that the number of *FCGR3b* fragments at ≥300 bp present in the DNA sample was proportional to the number of hGEs with a fragment length of at least 300 bp. For comparison in a subset of samples, we also quantified the fragments of the gene *ALB* of ≥150 bp by qPCR.

qPCR was performed according to a protocol developed, validated, and used in diagnostic routine at our pathology department. Specifically, qPCR was performed using 10 ng DNA, 0.6 uM primer solution (for 150 bp or 300 bp), molecular grade water and iQ™ SYBR® Green Supermix (Bio-Rad, Hercules, CA) in a 25 μL reaction. Each assay contained triplicates of FFPE DNA, DNA isolated from peripheral blood as

control of non-fragmented DNA, and a non-template control. All runs were processed on a Bio-Rad® CFX96 using the following run program: 95 °C/10 min – 40 cycles of 94 °C/30 s, 56 °C/30 s, 72 °C/30 s– melting curve program: 95 °C to 60 °C, and increment of 0.5 °C for 1 s.

The difference in mean threshold value (ΔCt) between FFPE DNA and the blood DNA control was used to calculate the number of fragments with at least 150 bp/300 bp in the tumor DNA relative to the control. The fragmentation degree of tumor DNA was defined as $2^{\Delta Ct}$.

2.4. Next-generation sequencing and mutation detection

NGS libraries were prepared following the manufacturers' instructions using 1.7–250.8 ng FFPE DNA without considering the DNA fragmentation degree. Libraries were made using QIAseq Targeted DNA Human Actionable Solid Tumor Panel (Qiagen, Valencia, CA), which included UMIs. In brief, DNA was enzymatically fragmented, end-repaired, and A-tailed followed by ligation to a 5′ sequencing adapter that contained the UMI. The regions of interest were then selected by targeted PCR using an adapter primer and gene-specific primers that contained a universal primer sequence. The library was then amplified in a universal PCR using primers for the 5'adapter and a 3'primer complementary to the primer seat added in the targeted PCR. The 3′ primer also contained the 3′ sequencing adapter sequence. Libraries were quantified by KAPA Library Quantification kit (Roche, Switzerland) and pooled together in equimolar amounts before sequencing. 151 bp pair end sequencing was performed on the Illumina MiSeq or NextSeq platform (Illumina, San Diego, CA).

Data analysis was performed using the CLC Genomic Workbench version 12.0.2 (Qiagen, Valencia, CA) and a panel-specific workflow that utilized the UMIs. All reads passing the quality filters were used for downstream analyses. Mean read coverage was defined as the mean number of reads that covered each target position, without using the UMI information. Duplicates with the same UMI sequence were then grouped into a "UMI family", and the mean unique coverage was defined as the mean number of UMI families that covered each target position. Variants were called if they were present in 75% of the duplicates in a family. Variants below 5% allele frequency were discarded to avoid erroneously calling mutations that spontaneously arise in DNA over time.

2.5. Statistical analysis

Linear regression was used to explore the relation between mean unique coverage, total number of reads, amount of input DNA (Fig. 2A) and number of genome equivalents (Fig. 2B). R version 1.1.463 was used for statistical analyses, and figures were made using the ggplot2 package. The level of statistical significance was defined as $p \leq 0.05$.

3. Results

3.1. High variation in fragmentation degree of FFPE DNA

The concentration of DNA extracted from the 116 samples ranged from 0.103 ng/μL to 136.0 ng/μL (median 4.9 ng/μL). qPCR with a 300 bp fragment of the *FCGR3b* gene was then used to estimate the fragmentation degree in the FFPE samples compared to a control sample of DNA extracted from whole blood. Such DNA is minimally fragmented and considered to be of high quality compared to FFPE DNA. The quality of FFPE DNA was therefore defined as the fragmentation degree relative to the whole blood DNA control sample. The relative fragmentation degree was calculated using the difference in Ct value (ΔCt) between each FFPE sample and the control and the formula $2^{\Delta Ct}$. Thereby, a ΔCt value of 3.3 implied a 10-fold fragmentation degree of FFPE relative to the control. As shown in Fig. 1, the fragmentation degree of FFPE DNA ranged from 1 to >500-fold compared to the control. Sixteen samples were > 500-fold fragmented and the sample with the worst quality was 6339-fold fragmented.

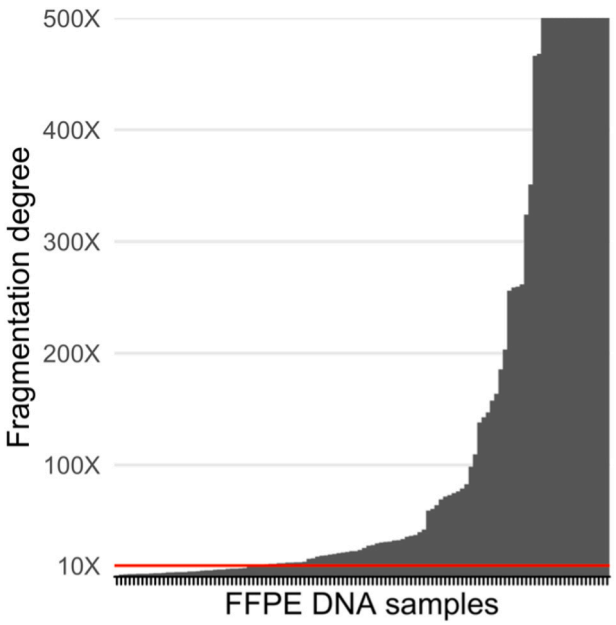

Fig. 1. The fragmentation degree of 116 lung tumor FFPE samples relative to a control of genomic DNA extracted from leukocytes from a healthy person. The y-axis shows the values of $2^{\Delta Ct}$ and is capped at 500-fold. Seventy-two percent of the samples were more than 10-fold fragmented than the control, as indicated by the red line. Based on experience, FFPE derived DNA is usually 10-fold fragmented. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

An additional qPCR assay was performed with a 150 bp fragment in the gene *ALB* in a subset of the FFPE DNA samples (n = 41), and the fragmentation degree varied from 1- to 19-fold (median 3-fold). The samples that were more than 10-fold fragmented at 300 bp (10- to 520-fold) varied between 2- and 19-fold fragmented at 150 bp.

3.2. The number of amplifiable hGEs in input FFPE DNA can predict the library complexity

NGS libraries were prepared using 1.7–250.8 ng input DNA and QIAseq Targeted DNA Human Actionable Solid Tumor Panel. To evaluate the impact of DNA fragmentation on the library complexity, we used the fragmentation degree to calculate the number of potential hGE templates present in the input for each library. The UMIs were used to evaluate the library complexity.

First, we examined the relationship between input DNA amount and the mean unique coverage in the panel target region. Overall, the mean unique coverage ranged from 16× to 3098× (median 326×) in the 116 libraries. A high total number of reads per library did not result in a high mean unique coverage $R^2 = 0.00137$, $p = 0.694$) (point colors in Fig. 2). Furthermore, a higher input DNA amount in nanograms did not systematically increase the mean unique coverage (Fig. 2A), though the correlation between input DNA and mean unique coverage was statistically significant ($R^2 = 0.277$, $p < 0.0001$), mainly because the highest input DNA amounts resulted in high mean unique coverage; 10 out of 13 libraries made with more than 200 total ng input DNA had mean unique coverage >1000×. Fig. 2A shows that high fragmentation degree decreased the mean unique coverage among libraries made from the approximately same input amount in ng. In line with this observation, a higher number of amplifiable hGEs in the input DNA generally increased the mean unique coverage ($R^2 = 0.410$, $p < 0.0001$) (Fig. 2B).

Second, we examined the relationship between input DNA amount and the number of duplicates/UMI family. The libraries made from lower amounts of DNA in nanograms contained a high number of

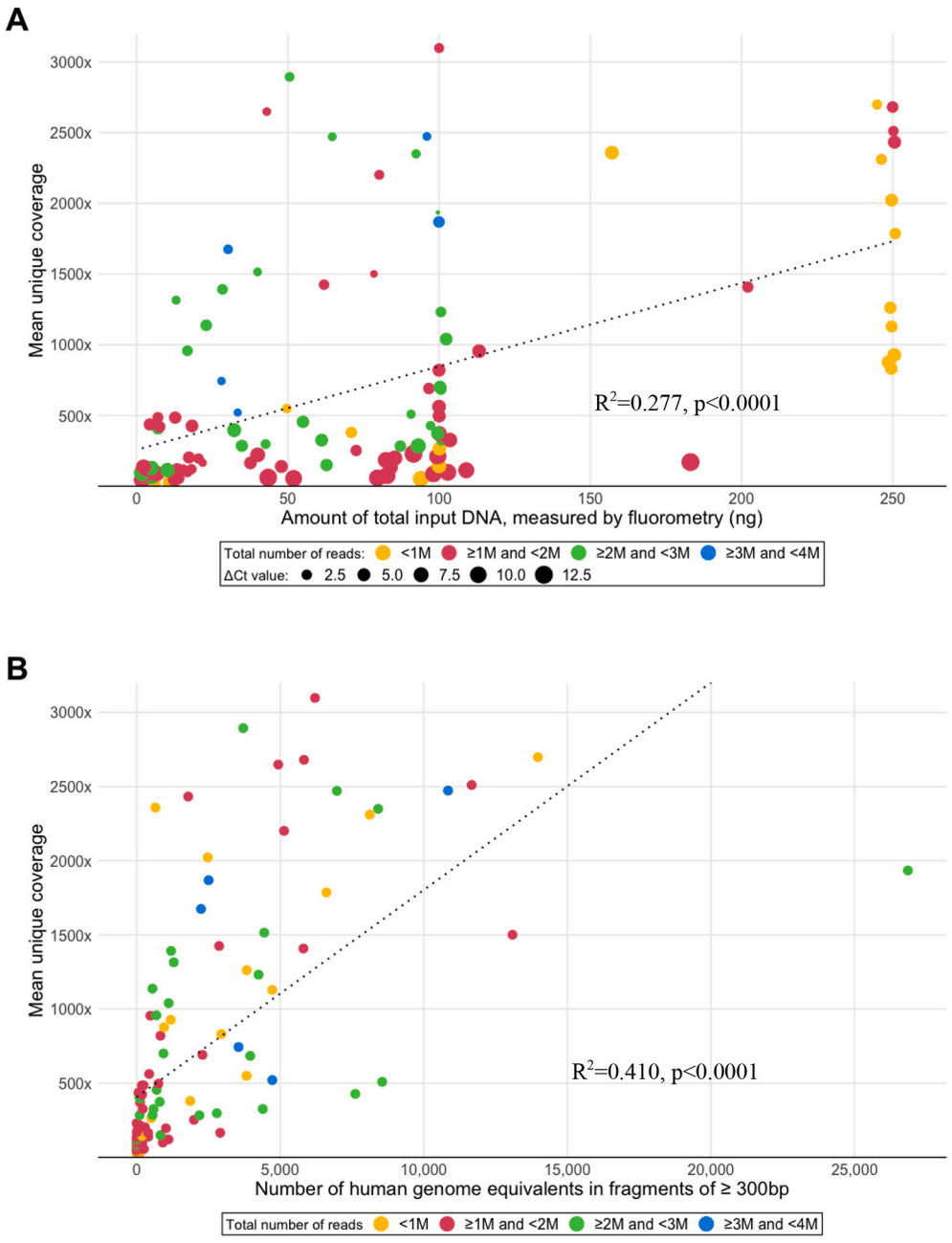

Fig. 2. Mean unique coverage for each of the 116 libraries versus A) the amount of input DNA measured by fluorometry and B) the amount of amplifiable hGE templates measured by qPCR. All reads passing the quality filters were analyzed, and the color shows the number of reads analyzed in each sample. In fig. A, the point size increases with increased fragmentation degree of the input DNA.

duplicates/UMI family (Fig. 3A). However, Fig. 3A shows that the number of duplicates/UMI family varied among samples with the same amount of DNA in nanograms, while it was more uniform among samples with the same input DNA calculated as the number of amplifiable hGEs (Fig. 3B).

We observed that all samples with >200 ng total DNA input had few duplicates per UMI family (1.1–2.8). These libraries did not have sufficient duplicates per UMI family for error correction, which indicates that an input DNA of more than 200 ng potentially causes ineffective PCR reactions.

Third, we examined the relationship between input DNA amount and uniformity of the mean unique coverage of each gene in the target panel. To exemplify, in Fig. 4 we present the coverage uniformity in two libraries that were prepared using 100 ng DNA input, and both libraries generated approximately 1.7 million reads. The input DNA used to prepare the first library (Fig. 4A) was 5-fold fragmented and the input DNA for the second library (Fig. 4B) was 520-fold fragmented relative to the control. This corresponds to an estimated input of 6200 and 58 potential hGE templates, respectively. The coverage uniformity was superior in the library made from the highest number of potential templates (Fig. 4A).

3.3. *DNA fragmentation assessment helps to evaluate the NGS analytic sensitivity*

We examined the impact of input DNA amount on the detection of mutations in six representative libraries made from variable DNA amounts (Table 1; results from all 116 samples are presented in the

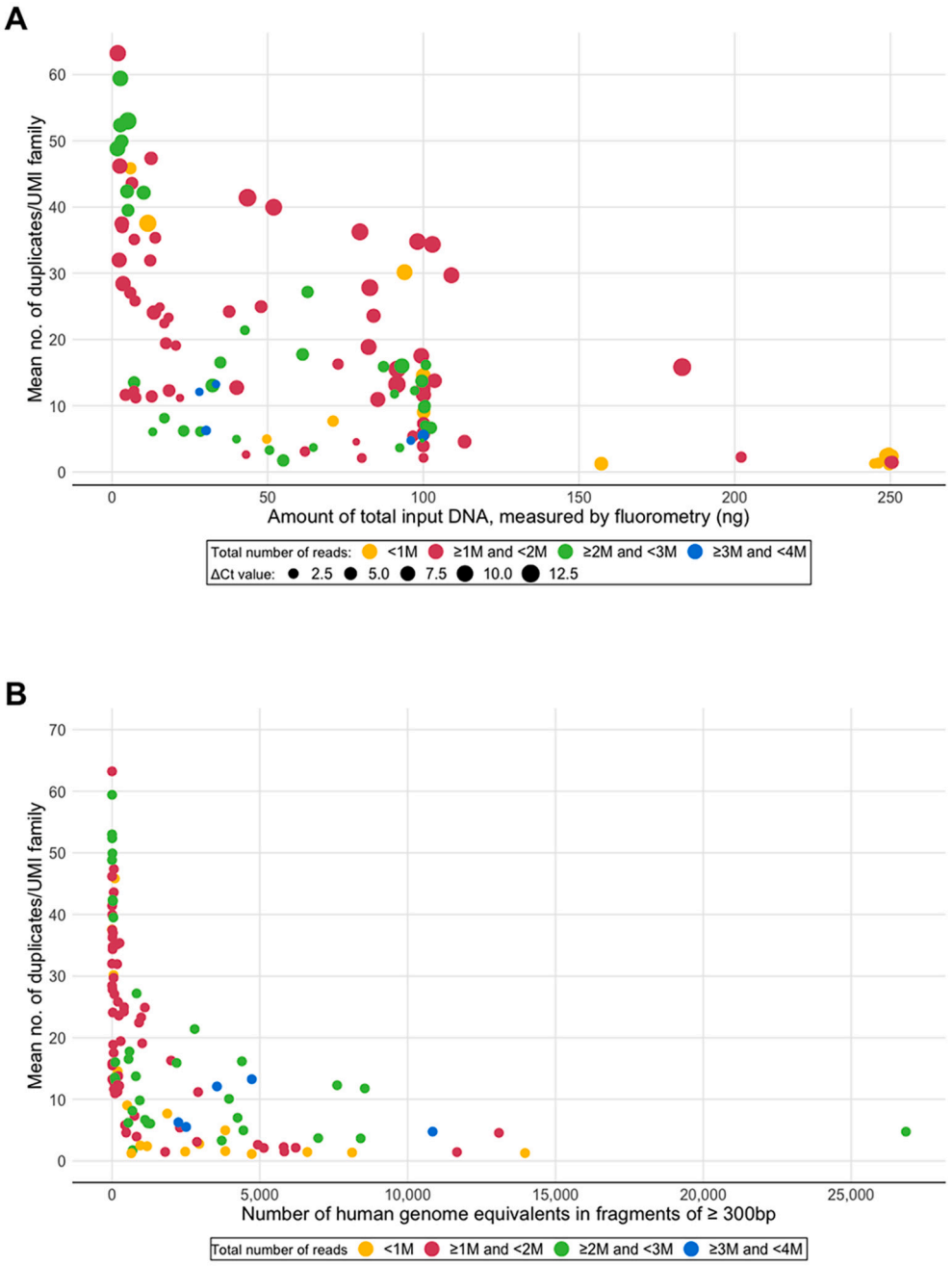

Fig. 3. Mean number of duplicates per UMI family in each of the 116 libraries versus A) the amount of total input DNA and B) the amount of amplifiable input DNA. All reads passing the quality filters were analyzed, and the color shows the number of reads analyzed in each sample. In fig. A, the point size increases with increased fragmentation degree of the input DNA.

supplementary table). The libraries 237 and 841 were prepared from approximately 100 ng DNA. The mean read coverage in the panel target region (calculated without using the UMIs) was comparable between the two samples. This similarity indicates comparable analytical sensitivity. However, the estimated number of input hGEs for each library differed considerably (21 in 237 vs. 26,855 in 841). As a result, the mean number of duplicates per UMI family in library 237 was much higher (34.4 compared to 4.7 in library 841). Therefore, the validity of a low frequency variant in a library such as library 237 in our cohort should be carefully evaluated, especially if UMIs are not incorporated.

Fragmentation assessment further enabled evaluation of whether a mutation-negative sample was likely a true negative. Data analysis of libraries 29 and 837 shown in Table 1 resulted in no detected mutations. Considering the number of input hGEs, rather than the mean read coverage, it is likely that sample 837 is a true negative sample, while we cannot rule out that the analysis of sample 29 represents a false negative result. Without UMIs or awareness of the DNA fragmentation degree, it would not be possible to accurately evaluate an apparently negative result.

We observed that the mean unique coverage was higher than the estimated input of hGE templates in some libraries (Table 1). This discrepancy suggests that a significant number of hGE templates of shorter fragment lengths than 300 bp were available for amplification in these samples. When using the additional assay with a 150 bp fragment to analyze a subset of the FFPE DNA samples, we observed that the number of 150 bp templates was up to a 10-fold higher than the number of 300 bp templates (data not shown). On the other hand, e.g. libraries 841 and 837 had lower unique coverage than the estimated input of hGE

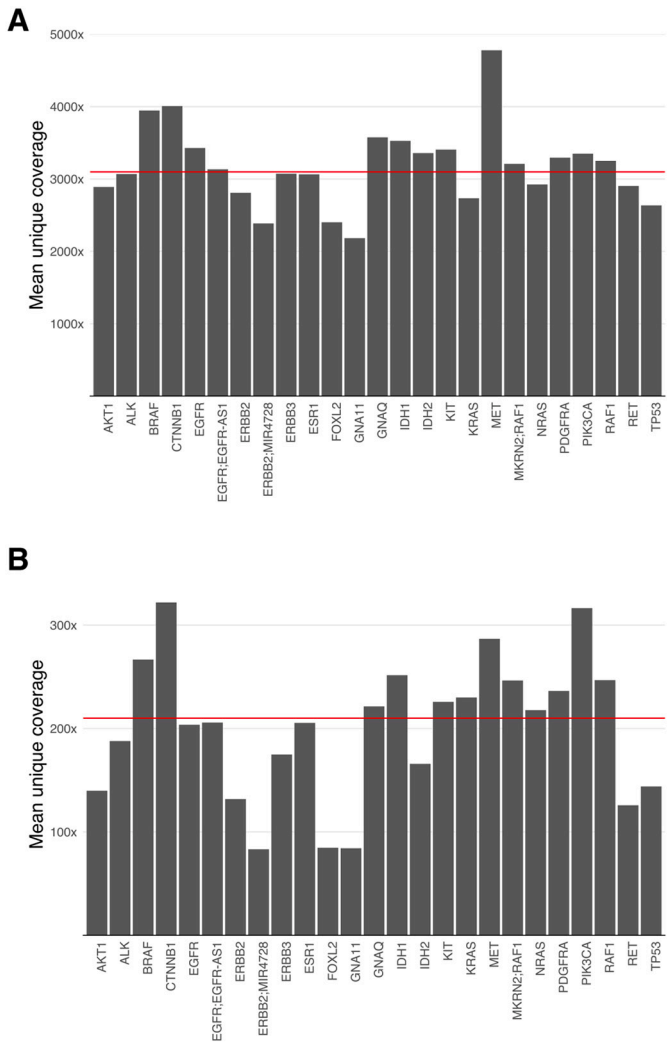

Fig. 4. The coverage uniformity in two libraries made from 100 ng input DNA. The graphs show the mean unique coverage for each gene covered by the QIAseq Human Actionable Solid Tumor panel. The red line indicates the mean unique coverage for the whole target region in each library. The fragmentation degree varied in the input DNA used to prepare the libraries. The library in fig. A) was prepared using 6200 potential hGE templates. The library in fig. B) was prepared using 58 potential hGE templates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

templates; this suggests that too much input DNA may decrease the number of duplicates per UMI family and consequently reduces complexity.

4. Discussion

The challenge of using FFPE DNA for NGS analyses in clinical practice is ensuring a sufficient amount and quality of input DNA. There is often less input DNA available than optimal. It is, however, possible to generate an NGS library from as little as a few nanograms of FFPE DNA by using amplicon-based enrichment. For such samples, ensuring sufficient library complexity is necessary for precisely interpreting whether variants with low frequency are truly present or not.

In this study, we calculated the potentially amplifiable hGEs for NGS library by assessing the fragmentation degree of FFPE DNA using a qPCR method. We incorporated UMIs and constructed NGS libraries from 116 lung cancer FFPE DNA samples. We did not exclude the samples with lower amounts of available input DNA as recommended by the manufacturer. Using UMIs enabled us to trace the DNA fragment in the original input DNA and evaluate the complexity of each library. We then demonstrated that, for libraries prepared with FFPE DNA, increasing input DNA amount would help to increase such complexity, but only when adjusting for amplifiable hGEs in each sample.

By comparing the NGS results of libraries constructed with different amounts of cell line derived DNA, McNulty *et al.* showed that higher input DNA amounts provide higher NGS library complexity (McNulty et al., 2020). In our study, we found similar associations when the input DNA amount from FFPE tissue was measured as the number of potential hGE templates. Libraries made from higher amounts of input hGE had higher library complexity in terms of higher mean unique coverage, higher coverage uniformity and fewer duplicates per UMI family. However, if input FFPE DNA was fragmented to such a degree that few hGEs were available, the library complexity did not necessarily increase with higher input FFPE DNA amount measured in nanograms. Libraries made from a low number of hGEs generated a comparable number of total reads as libraries made from a high number of hGEs, but many of the reads were duplicates belonging to the same UMI families. Duplicates are necessary for error correction. However, Xu *et al.* concluded that four duplicates per UMI family are sufficient (Xu et al., 2017), suggesting that very high numbers of duplicates do not increase the library complexity.

Others have also concluded that the DNA amount from FFPE samples measured in nanograms by either spectrophotometer or by fluorometer may not always represent the amount of amplifiable input DNA (hGEs) available for library preparation. Heydt *et al.* compared and evaluated the impact of five different DNA quantification methods on downstream amplicon-based NGS performance in order to find the best method to assess the quantity of input FFPE DNA (Heydt et al., 2014). Both spectrophotometric and fluorescent dye-based quantification systems and a qPCR method were used, and they found that the DNA concentration varied widely when using different methods.

Heydt *et al.* constructed 24 libraries from two samples that were divided into four sets of different input DNA amounts. Each set consisted of three solutions with the same amount of DNA calculated by three different methods. In contrast to our results, they found that the mean read coverage and number of called variants were comparable independent of the quantification method. Their results are not necessarily comparable to our findings since they did not use molecular barcodes to label their input DNA or assess mean unique coverage; instead, they used read coverage as the quality metric. As we have shown, PCR can generate high amounts of duplicates even from low amounts of input DNA and consequently, high mean coverage.

In this study, most FFPE DNA samples (72%) were more than 10-fold fragmented compared to the non-fragmented control DNA. Furthermore, the fragmentation degree of FFPE DNA varied considerably between samples, underscoring the need to evaluate fragmentation of each sample. We believe that unknown and low quantities of amplifiable input DNA might be one reason for failed NGS which is frequently reported in studies involving clinical FFPE tissue samples (Flaherty et al., 2020; Middleton et al., 2020; Stockley et al., 2016).

We observed that the mean unique coverage varied between libraries generated from similar amounts of hGEs in input DNA. One explanation might be the different fragmentation degree of fragments shorter than 300 bp. It may also indicate that fragmentation alone might not affect complexity (Hedegaard et al., 2014), and that other FFPE-related DNA modification factors might impede efficient library generation.

Our study demonstrated several benefits of assessing the available amount of amplifiable input DNA for NGS. First, we show that the number of amplifiable hGE can vary considerably between FFPE DNA samples with the same DNA amount in nanograms. Second, we show that the amount of hGEs better predicts library complexity than the amount of input DNA in nanogram. Third, assessing fragmentation is valuable when interpreting NGS data, especially for the samples with low yields and poor quality, since the risk of both false positive and false

Table 1
NGS quality and mutations detected in six representative samples. Full overview of all 116 samples is listed in the Supplementary Table.

Sample ID	Total DNA concentration (ng/uL)	Fragmentation degree relative to the control	Total DNA input amount for NGS (ng)	No. of potential hGE templates	Mean read coverage in the panel target region	Mean unique coverage in the panel target region	Number of duplicates/ UMI family	Mutations detected	Mutant allele frequency (unique count/ unique coverage)
37	0.103	466×	1.7	1	13,750×	97×	48.8	*TP53* NP_001119586.1:p.Arg273Leu	56.3% (27/48)
29	0.136	260×	2.3	3	11,356×	136×	32.0	No mutation detected	N/A
237	34.3	1520×	102.9	21	9216×	98×	34.4	*ERBB3* NP_001973.2:p.Arg103Cys *KRAS* NP_203524.1:p.Gly13Cys *TP53* NP_001263625.1:p.Glu248*	7.9% (11/139) 13.1% (21/160) 55.4% (62/112)
841	6.6	1×	99.6	26,855	13,236×	1935×	4.7	*RAF1* NP_002871.1:p.Ser259Phe	76.2% (2073/2721)
826	24.0	31×	249.6	2476	3973×	2022×	1.5	*EGFR* NP_005219.2:p.Leu858Arg	17.8% (440/2470)
837	39.4	7×	250.2	11,667	4552×	2511×	1.4	No mutation detected	N/A

hGE; human genome equivalent, NGS; next-generation sequencing, UMI; unique molecular index,

negative variant calling increases. Therefore, a successful NGS library requires input DNA amount that is neither too low nor too large, and with a predefined amount of amplifiable hGE.

The most obvious approach to overcome challenges due to high fragmentation is to add more input DNA, but this is not always possible. For example, biopsies from lung cancer tumors are often small, especially those that are obtained through bronchoscopy. Our results indicate that when more DNA cannot be analyzed, it is important to be aware of how the hGE content in input DNA might influence the library complexity and potentially help in interpreting the variant call and avoid false positives or negatives.

CRediT authorship contribution statement

Anine Larsen Ottestad: Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Elisabeth F. Emdal:** Validation, Investigation, Resources, Data curation. **Bjørn H. Grønberg:** Resources, Writing – original draft, Writing – review & editing, Funding acquisition. **Tarje O. Halvorsen:** Software, Formal analysis, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Hong Yan Dai:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

The library sequencing was provided by the Genomics Core Facility (GCF), Norwegian University of Science and Technology (NTNU). GCF is funded by the Faculty of Medicine and Health Sciences at NTNU and The Central Norway Regional Health Authority. This study was supported by The Central Norway Regional Health Authority. The funding source did not have any role in design, conduct of the study, interpretation, or publication of results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yexmp.2022.104771.

References

Bass, B.P., Engel, K.B., Greytak, S.R., Moore, H.M., 2014. A review of Preanalytical factors affecting molecular, protein, and morphological analysis of formalin-fixed, paraffin-embedded (FFPE) tissue: how well do you know your FFPE specimen? Arch. Pathol. Lab. Med. 138, 1520–1530. https://doi.org/10.5858/arpa.2013-0691-RA.

Carrick, D.M., Mehaffey, M.G., Sachs, M.C., Altekruse, S., Camalier, C., Chuaqui, R., Cozen, W., Das, B., Hernandez, B.Y., Lih, C.-J., Lynch, C.F., Makhlouf, H., McGregor, P., McShane, L.M., Phillips Rohan, J., Walsh, W.D., Williams, P.M., Gillanders, E.M., Mechanic, L.E., Schully, S.D., 2015. Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. PLoS One 10, e0127353. https://doi.org/10.1371/journal.pone.0127353.

Chang, F., Li, M.M., 2013. Clinical application of amplicon-based next-generation sequencing in cancer. Cancer Gene Ther. 206, 413–419. https://doi.org/10.1016/j.cancergen.2013.10.003.

Do, H., Dobrovic, A., 2015. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin. Chem. 61, 64–71. https://doi.org/10.1373/clinchem.2014.223040.

Einaga, N., Yoshida, A., Noda, H., Suemitsu, M., Nakayama, Y., Sakurada, A., Kawaji, Y., Yamaguchi, H., Sasaki, Y., Tokino, T., Esumi, M., 2017. Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation. PLoS One 12, e0176280. https://doi.org/10.1371/journal.pone.0176280.

Flaherty, K.T., Gray, R.J., Chen, A.P., Li, S., McShane, L.M., Patton, D., Hamilton, S.R., Williams, P.M., Iafrate, A.J., Sklar, J., Mitchell, E.P., Harris, L.N., Takebe, N., Sims, D.J., Coffey, B., Fu, T., Routbort, M., Zwiebel, J.A., Rubinstein, L.V., Little, R. F., Arteaga, C.L., Comis, R., Abrams, J.S., O'Dwyer, P.J., Conley, B.A., for the NCI-MATCH team, 2020. Molecular landscape and actionable alterations in a genomically guided cancer clinical trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). J. Clin. Oncol. 38, 3883–3894. https://doi.org/10.1200/JCO.19.03010.

Hedegaard, J., Thorsen, K., Lund, M.K., Hein, A.-M.K., Hamilton-Dutoit, S.J., Vang, S., Nordentoft, I., Birkenkamp-Demtröder, K., Kruhøffer, M., Hager, H., Knudsen, B., Andersen, C.L., Sørensen, K.D., Pedersen, J.S., Ørntoft, T.F., Dyrskjøt, L., 2014. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human Cancer and Normal tissue. PLoS One 9, e98187. https://doi.org/10.1371/journal.pone.0098187.

Heydt, C., Fassunke, J., Künstlinger, H., Ihle, M.A., König, K., Heukamp, L.C., Schildhaus, H.-U., Odenthal, M., Büttner, R., Merkelbach-Bruse, S., 2014. Comparison of pre-analytical FFPE sample preparation methods and their impact on massively parallel sequencing in routine diagnostics. PLoS One 9, e104566. https://doi.org/10.1371/journal.pone.0104566.

Kerick, M., Isau, M., Timmermann, B., Sültmann, H., Herwig, R., Krobitsch, S., Schaefer, G., Verdorfer, I., Bartsch, G., Klocker, H., Lehrach, H., Schweiger, M.R.,

2011. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. BMC Med. Genet. 4, 68. https://doi.org/10.1186/1755-8794-4-68.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci 108, 9530–9535. https://doi.org/10.1073/pnas.1105422108.

McDonough, S.J., Bhagwate, A., Sun, Z., Wang, C., Zschunke, M., Gorman, J.A., Kopp, K. J., Cunningham, J.M., 2019. Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. PLoS One 14, e0211400. https://doi.org/10.1371/journal.pone.0211400.

McNulty, S.N., Mann, P.R., Robinson, J.A., Duncavage, E.J., Pfeifer, J.D., 2020. Impact of reducing DNA input on next-generation sequencing library complexity and variant detection. J. Mol. Diagn. 0 https://doi.org/10.1016/j.jmoldx.2020.02.003.

Middleton, G., Fletcher, P., Popat, S., Savage, J., Summers, Y., Greystoke, A., Gilligan, D., Cave, J., O'Rourke, N., Brewster, A., Toy, E., Spicer, J., Jain, P., Dangoor, A., Mackean, M., Forster, M., Farley, A., Wherton, D., Mehmi, M., Sharpe, R., Mills, T.C., Cerone, M.A., Yap, T.A., Watkins, T.B.K., Lim, E., Swanton, C., Billingham, L., 2020. The National Lung Matrix Trial of personalized therapy in lung cancer. Nature 583, 807–812. https://doi.org/10.1038/s41586-020-2481-8.

Pel, J., Leung, A., Choi, W.W.Y., Despotovic, M., Ung, W.L., Shibahara, G., Gelinas, L., Marziali, A., 2018. Rapid and highly-specific generation of targeted DNA sequencing libraries enabled by linking capture probes with universal primers. PLoS One 13, e0208283. https://doi.org/10.1371/journal.pone.0208283.

Robbe, P., Popitsch, N., Knight, S.J.L., Antoniou, P., Becq, J., He, M., Kanapin, A., Samsonova, A., Vavoulis, D.V., Ross, M.T., Kingsbury, Z., Cabes, M., Ramos, S.D.C., Page, S., Dreau, H., Ridout, K., Jones, L.J., Tuff-Lacey, A., Henderson, S., Mason, J., Buffa, F.M., Verrill, C., Maldonado-Perez, D., Roxanis, I., Collantes, E., Browning, L., Dhar, S., Damato, S., Davies, S., Caulfield, M., Bentley, D.R., Taylor, J.C., Turnbull, C., Schuh, A., 2018. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 genomes project. Genet. Med. 20, 1196–1205. https://doi.org/10.1038/gim.2017.241.

Spencer, D.H., Sehn, J.K., Abel, H.J., Watson, M.A., Pfeifer, J.D., Duncavage, E.J., 2013. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J. Mol. Diagn. 15, 623–633. https://doi.org/10.1016/j.jmoldx.2013.05.004.

Srinivasan, M., Sedmak, D., Jewell, S., 2002. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am. J. Pathol. 161, 1961–1971. https://doi.org/10.1016/S0002-9440(10)64472-0.

Stockley, T.L., Oza, A.M., Berman, H.K., Leighl, N.B., Knox, J.J., Shepherd, F.A., Chen, E. X., Krzyzanowska, M.K., Dhani, N., Joshua, A.M., Tsao, M.-S., Serra, S., Clarke, B., Roehrl, M.H., Zhang, T., Sukhai, M.A., Califaretti, N., Trinkaus, M., Shaw, P., van der Kwast, T., Wang, L., Virtanen, C., Kim, R.H., Razak, A.R.A., Hansen, A.R., Yu, C., Pugh, T.J., Kamel-Reid, S., Siu, L.L., Bedard, P.L., 2016. Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. Genome Med. 8, 109. https://doi.org/10.1186/s13073-016-0364-2.

Xu, C., Nezami Ranjbar, M.R., Wu, Z., DiCarlo, J., Wang, Y., 2017. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. BMC Genomics 18. https://doi.org/10.1186/s12864-016-3425-4.