RESEARCH ARTICLE

Impact of Pre-Analytical Variables on Cancer Targeted Gene Sequencing Efficiency

Luiz H. Araujo[1], Cynthia Timmers[1], Konstantin Shilo[1], Weiqiang Zhao[1], Jianying Zhang[2], Lianbo Yu[2], Thanemozhi G. Natarajan[3], Clinton J. Miller[3], Ayse Selen Yilmaz[1,2,4], Tom Liu[1], Joseph Amann[1], José Roberto Lapa e Silva[5], Carlos Gil Ferreira[6], David P. Carbone[1]*

1 James Thoracic Center, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, 43210, United States of America, **2** Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, 43210, United States of America, **3** GenomOncology, Cleveland, Ohio, 44145, United States of America, **4** Biomedical Informatics Shared Resource, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, 43210, United States of America, **5** Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, **6** Brazilian National Cancer Institute, Rio de Janeiro, Brazil

* david.carbone@osumc.edu

OPEN ACCESS

Citation: Araujo LH, Timmers C, Shilo K, Zhao W, Zhang J, Yu L, et al. (2015) Impact of Pre-Analytical Variables on Cancer Targeted Gene Sequencing Efficiency. PLoS ONE 10(11): e0143092. doi:10.1371/journal.pone.0143092

Editor: Sumitra Deb, Virginia Commonwealth University, UNITED STATES

Received: September 4, 2015

Accepted: September 26, 2015

Published: November 25, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: L.H.A. is supported by a Conquer Cancer Foundation of ASCO Long-Term International Fellowship (LIFe) and a Landon Foundation-AACR INNOVATOR Award for International Collaboration in Cancer Research. This work was funded by a NIH/NCI 1RC1 CA146260-01, NCI R01CA60691, NCI R01CA87895, NCI P30CA022453, and the Ohio State Cancer Center Support Grant (CCSG), NCI CA16058. T.G.N and C.J.M. are employed by GenomOncology. This funder provided support in the form of salaries for authors (T.G.N and C.J.M), but did

Abstract

Tumor specimens are often preserved as formalin-fixed paraffin-embedded (FFPE) tissue blocks, the most common clinical source for DNA sequencing. Herein, we evaluated the effect of pre-sequencing parameters to guide proper sample selection for targeted gene sequencing. Data from 113 FFPE lung tumor specimens were collected, and targeted gene sequencing was performed. Libraries were constructed using custom probes and were paired-end sequenced on a next generation sequencing platform. A PCR-based quality control (QC) assay was utilized to determine DNA quality, and a ratio was generated in comparison to control DNA. We observed that FFPE storage time, PCR/QC ratio, and DNA input in the library preparation were significantly correlated to most parameters of sequencing efficiency including depth of coverage, alignment rate, insert size, and read quality. A combined score using the three parameters was generated and proved highly accurate to predict sequencing metrics. We also showed wide read count variability within the genome, with worse coverage in regions of low GC content like in *KRAS*. Sample quality and GC content had independent effects on sequencing depth, and the worst results were observed in regions of low GC content in samples with poor quality. Our data confirm that FFPE samples are a reliable source for targeted gene sequencing in cancer, provided adequate sample quality controls are exercised. Tissue quality should be routinely assessed for pre-analytical factors, and sequencing depth may be limited in genomic regions of low GC content if suboptimal samples are utilized.

Introduction

In the past decade, a better understanding of cancer biology and the identification of somatic mutations in cancer have led to a new era in personalized oncology.[1] Landmark examples included the discovery of mutations in the proto-oncogenes c-*KIT* in gastrointestinal stromal tumors (GIST),[2] epidermal growth factor receptor (*EGFR*) in lung adenocarcinomas,[3] and

not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the author contributions section.

Competing Interests: T.G.N and C.J.M. are employed by GenomOncology. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

v-Raf murine sarcoma viral oncogene homolog B1 (*BRAF*) in melanomas.[4] Tumors harboring these mutations demonstrate outstanding sensitivity to specific kinase inhibitors directed against the respective activated pathways.

These oncogenic mutations are often defined as cancer drivers since they offer a selective advantage to a cell clone, necessary for tumor initiation and maintenance.[5] In the clinic, they may serve as fingerprints that help clinicians to subtype cancers that, otherwise, present similar histological patterns.[6–9] While mutational profiling has become a useful tool to better tailor targeted therapies, new challenges have arisen including the frequent need to obtain optimal tumor specimens for extra genetic testing.[10–12] Moreover, multiple tests may be recommended in a clinical setting where a plethora of candidate driver mutations need to be assessed. In non-small cell lung cancer (NSCLC), alterations in at least 10 proto-oncogenes have been suggested as potentially "druggable", with mutation frequencies varying from 1% to 25% for *MAP2K1* and *KRAS*, respectively, according to the population studied.[13] The best algorithm for testing, including sequential versus multiplex assessment of such alterations is still a matter of debate.

Although Sanger sequencing has been traditionally used for the detection of recurrent point mutations in cancer, newer technologies have enabled a more comprehensive analysis of genetic perturbations. In this scenario, the next generation sequencing (NGS) platforms—also known as massively parallel sequencing—offer a wide range of opportunities to characterize the cancer genome.[14–16] For instance, the availability of hybridization-capture techniques provides a high-throughput and cost-effective strategy to assess hundreds of genes simultaneously.[16–19] As a brief methodological review, genomic DNA (gDNA) is purified from tumor samples and sheared by either sonication or restriction enzymes into millions of small fragments (tens or hundreds nucleotides long). These fragments are then hybridized to a customized probe set that contains baits specific for the genes of interest, and amplified to generate the sequencing library. A unique barcode is ligated to each library—corresponding to each sample—which enables multiple samples to be pooled together for sequencing. Several commercial DNA capture technologies are currently available, and many institutions have designed and implemented customized targeted panels to genotype cancer samples.[10, 20, 21]

Clinical tumor specimens are often preserved as formalin-fixed paraffin-embedded (FFPE) tissue blocks in biorepositories, and this is the most readily available source for obtaining gDNA both in clinical and research settings.[22–24] However, several steps in the FFPE processing are known to cause DNA damage, which directly affect DNA quality and adequacy for sequencing. For instance, formalin fixation may result in various types of crosslinks between two amino acids, two nucleic acids, or between an amino acid and a nucleic acid base.[25–27] These chemical modifications may confound molecular testing through inhibition of enzymatic manipulation of DNA. Formalin fixation may also cause nucleotide oxidation and deamination, the latter being related to the development of artifactual nucleotide transitions (mostly C>T in CpG dinucleotides) among samples stored as FFPE.[23, 28] Lastly, methylene crosslinks caused by formalin may result in DNA fragmentation, which limits the DNA length for sequencing. In addition to formalin fixation, tissue preparation, paraffin embedding, and archival storage *per se* may all ultimately play a role in samples quality.[29] Moreover, FFPE blocks are often obtained from small biopsies, and low tissue quantity may constitute an additional limitation for sequencing. In order to evaluate the quality of gDNA extracted from FFPE samples, PCR-based quality control (QC) assays have been recommended.[30–33] Other variables likely to affect final sequencing results include the amount of DNA used as input for the library preparation, the sequencing depth, and the targeted region of interest (GC content and sequence homology).

Herein, we evaluated the individual and combined impact of pre-sequencing parameters on targeted gene sequencing efficiency. To this end, we utilized a fully annotated sample set characterized by knowledge of a wide range of pre-analytical variables, which was genotyped for a customized gene panel using a commercially available targeted gene sequencing approach —the Agilent Haloplex Target Enrichment System (Agilent Technologies). This platform differs from other hybridization-capture techniques in that a pool of restriction enzymes is used to digest the sample DNA (as opposed to sonication), and the probes are designed with homology only to the ends of targeted DNA restriction fragments.[34] Subsequently, universal primers are used to amplify the captured regions and will generate a high frequency of similar reads, resembling the results found in amplicon-based platforms (S1 Fig). For this reason, some sequencing metrics like duplication rate and unique reads quantification are not applicable to this technology. We also verified the read depth variability within the genome, disclosed problematic regions based on GC content, and determined the impact of these parameters on variant calling. These data could be very informative to guide the clinical and research community in the adequate selection of clinical samples for targeted gene sequencing, and in the proper interpretation of sequencing results as a function of sample quality and sequencing uniformity.

Materials and Methods

Clinical samples

The studied dataset comprised 113 lung tumor specimens resected from patients at the James Cancer Hospital / The Ohio State University (OSU, Columbus, OH) between 1988 and 2011. All the samples were archived as FFPE tumor blocks, and were selected based on tissue availability. One hundred and ten samples were primary NSCLC (60 adenocarcinomas, 31 squamous cell carcinomas, 10 adenosquamous, and 9 other histological subtypes), while 3 samples were head and neck cancers (all squamous cell carcinomas) metastasizing to the lungs (Table A in S1 File). Each sample was assigned a unique, unidentifiable code, and date of surgery was reviewed and annotated to estimate the tumor block storage time. The Institutional Review Board approved this project, and waived the need for consenting.

Tissue processing

Resected samples with representative tumor tissue were selected for NGS testing. To increase tumor content, a pathologist (K.S.) marked an H&E stained slide to delineate tumor-containing regions, and these areas were macrodissected by manual scraping the marked areas from serial unstained FFPE sections. Tumor cellularity was determined by visual inspection of the number of tumor nuclei compared to stromal background in the areas marked for macrodissection, and most samples (88%) were classified as containing either high or moderate tumor cellularity (Table B in S1 File and S2 Fig). gDNA was extracted from FFPE samples using the Maxwell® 16 FFPE Plus LEV DNA Purification kit (Promega). Two to ten slides containing 10μm thick sections were scraped into a microtube and incubated overnight at 70°C with proteinase K solution and incubation buffer. Subsequently, each sample was treated with lysis buffer, transferred to loading cartridges and run in the automated instrument. Local testing showed that this protocol yielded similar amounts of DNA in comparison to manual systems (data not shown). gDNA quantification was performed using the Quant-iT™ High-Sensitivity DNA Assay Kit (Life Technologies™).

DNA quality assessment

In order to determine overall quality of the gDNA, a PCR-based QC assay was applied, and used as a guide to recommend the amount of DNA input in the library preparation, as recommended by the manufacturer.[35] Briefly, 10 ng of each DNA sample was amplified with 2 independent primer pairs in order to generate amplicons of incremental sizes: 105 base pairs (bp), and 236 bp. As a non-degraded positive control, we used gDNA extracted from a NSCLC cell line (A549). After PCR, products were evaluated for yield and level of fragmentation using the Agilent 2200 TapeStation (Agilent Technologies). The QC ratio was calculated dividing the bands quantification for each sample by the respective band in the positive control, and then averaging each band ratio. A QC ratio above 0.20 indicates favorable quality, while ratios below 0.20 suggests moderate or poor quality.[35]

Hybridization-capture and sequencing

A custom panel was designed using the internet-based Sure Design software (Agilent Technologies) to cover the coding regions of 81 selected genes relevant to NSCLC (Table C in S1 File). The total panel covered 920,980 base pairs, and included 44,234 amplicons. Libraries were constructed and indexed using the Agilent Haloplex Target Enrichment System (Agilent Technologies). The indexed libraries were pooled at equimolar amounts and paired-end sequenced (2 x 100 base pairs) to 1,000X average coverage on an Illumina HiSeq 2500.

Data processing

Sequencing reads were aligned to the human genome (hg19 assembly) and bam files were generated using the SureCall software (Agilent Technologies). Variant calling was performed using the Genome Analysis Tool Kit (GATK) Unified Genotyper. Variant annotation was carried out on GenomOncology's GenomAnalytics platform (GenomOncology, Cleveland, OH), and the Integrative Genomic Viewer (IGV, Broad Institute) was used to confirm true positives. Sequencing performance was assessed by measuring the number of reads, mapped reads, target base coverage, and read quality using Picard (Broad Institute), SAMtools and BEDTools.[36, 37] The depth of coverage in genomic regions was verified with the Depth Of Coverage tool (GATK),[38] and in specific genomic position (hotspots) using IGV.

Statistical methods

Three pre-analytical variables were used to predict final sequencing depth—FFPE storage time, PCR/QC ratio, and DNA input. FFPE storage time in paraffin block was calculated as the interval (in years) from the date of surgery to date of tumor processing (DNA extraction) for sequencing. To test the impact of these variables on the overall sequencing efficiency, we used several parameters like depth of coverage, alignment rate, off-target rate, base quality, among others. The pairwise correlation between pre-analytical variables (storage time, PCR/QC ratio and DNA input) and sequencing performance parameters was evaluated by Pearson's method. Subsequently, we created a training dataset comprising genomic alterations located in genes with the least coverage variability. Ten genes were filtered in: *ALK*, *BCL11A*, *REL*, *VGLL4*, *RAF1*, *FBLN2*, *RET*, *FGFR2*, *MAP2K1*, *U2AF2*. We then excluded genomic regions with overall poor coverage (less than 100 average reads) or with high depth variability (with standard deviation in the upper quartiles of variance) within these genes. These criteria led to a training dataset of 33 genomic regions, which was used to compare the pre-sequencing variables. Multivariable linear regression was performed for the correlation between the median reads and the three pre-analytical variables and for the potential multicollinearity among the three factors. This analysis indicated whether each covariate included in the model was still significantly correlated to sequencing performance after adjusting for other covariates. An equation based on the stepwise model selection procedure of all three individual factors and 2-way interaction terms was built to generate the combined score or the predictive sequencing performance. The final model selected the three individual pre-sequencing factors. The formula is presented here: *Combined score = 202.95–7.86 * Storage time + 249.95 * PCR ratio + 0.08 * DNA input*. In order to evaluate coverage variability within the genome (in GC content analysis), overall coverage was quantile-normalized and then stratified according to the GC content ratio. To compare the effect of GC content and tissue quality altogether, samples were stratified according to baseline quality (defined by the pre-sequencing combined score quartiles), and the overall depth of coverage was quantile-normalized within each group. P-value of ≤ 0.05 was considered statistically significant. Statistical analyses were performed using R version 3.0.1, SAS 9.3, and IBM SPSS version 22.0.

Results

Sample and sequencing parameters

We observed a large variation in the sample and pre-sequencing QC parameter across the selected samples (Table 1), including a range in FFPE storage time of 0.32 years to 24.22 years. The PCR-based QC assay indicated a median ratio of 0.19 (range 0.03–0.58), suggesting that the gDNA had a favorable quality in approximately half of the samples, while the other half had lower quality. The amount of gDNA used as input in the library preparation varied from 77 ng to 2,337 ng, with a median of 899 ng.

The median number of paired end reads and mapped reads per sample were 5.0 millions (range 1.4–7.7) and 4.9 millions (range 1.1–7.6), respectively, and 98.1% (range 78.4–98.9) of reads mapped to the target region. The median actual target coverage was 881X (range 204–1,373), with the median percentages of target reads covered at least 20-fold (20x), 50-fold (50x), and 100-fold (100x) being 95.4% (range 78.9–98.8), 90.8% (range 66.7–97.7), and 84.6% (range 52.1–95.1), respectively. The median insert size was 89.4 base pairs (range 73.5–120.7),

Table 1. Pre-analytical parameters and sequencing metrics.

Feature	Median (range)
Pre-analytical parameters	
FFPE storage time (years)	9.17 (0.32–24.22)
PCR/QC ratio	0.19 (0.03–0.58)
DNA input (ng)	899 (77–2,337)
Sequencing metrics	
Sequence reads and coverage	5,041,132 (1,464,386–7,738,458)
Reads mapped to genome	4,918,758 (1,148,705–7,597,557)
Bases mapped to genome	390,673,221 (86,893,349–608,314,077)
On-target bases mapped to genome	277,117,362 (63,725,143)
Alignment rate (%)	98.1 (78.4–98.9)
Target coverage	881x (204–1,373)
Fraction of bases covered ≥ 20x (%)	95.4 (78.9–98.8)
Fraction of bases covered ≥ 50x (%)	90.8 (66.7–97.7)
Fraction of bases covered ≥ 100x (%)	84.6 (52.1–95.1)
Insert size (in base pairs)	89.4 (73.5–120.7)
Mean base quality score (Phred score)	35.1 (32.4–35.7)
Mean quality score last 20 bases	31.6 (25.0–32.6)
Mean quality score last 10 bases	31.2 (23.8–32.2)
Fraction of bases ≥ Q30 (%)	98.5 (74.5–100)

Abbreviations: FFPE, formalin-fixed paraffin-embedded tissue blocks; PCR/QC, PCR-based quality control; Q30, Phred scale quality of 30.

doi:10.1371/journal.pone.0143092.t001

and 98.5% of base calls had a Phred quality score of at least 30. These parameters are summarized in Table 1 and S3 Fig.

Correlation between pre-analytical variables and sequencing efficiency

The pre-analytical variables (FFPE storage time, PCR/QC ratio, and DNA input) were significantly correlated to most parameters of sequencing efficiency (Fig 1 and Table D in S1 File). FFPE storage time was negatively correlated to total number of reads (r = -0.356), mean target coverage (r = -0.405), alignment rate (r = -0.354), insert size (r = -0.764, p<0.01 in all cases), and mean base quality (r = -0.188, p = 0.046), and positively correlated to off-target rate (r = 0.285, p<0.01), compatible with better results if more recent samples are selected. The QC ratio was correlated to insert size (r = 0.601, p<0.01), and insignificantly correlated to target coverage (r = 0.183, p = 0.058), alignment rate (r = 0.169, p = 0.08), and base quality (r = 0.162, p = 0.094). DNA input was correlated to total number of reads (r = 0.548), mean target coverage (r = 0.549), alignment rate (r = 0.449), mean base quality (r = 0.477), and off-target rate (r = -0.336, p<0.01 in all cases), but not to insert size (r = 0.081, p = 0.395). These data suggest that better QC ratio and higher DNA input in the library preparation may predict greater sequencing efficiency.

The combined effect of pre-analytical variables

The combined effect of pre-analytical variables was assessed in a training dataset of 33 genomic regions, with median depth of coverage of 267 (range 43–464). FFPE storage time was negatively correlated to the depth of coverage (r = -0.558, p<0.01; Fig 2A), while QC ratio and DNA input were positively correlated (r = 0.37 and 0.47, respectively; p<0.01 in both; Fig 2B and 2C). Using a multivariate analysis, we showed that each of these variables was still significantly correlated to sequencing performance (Table E in S1 File). In order to generate a unique score that predicts sample quality in this cohort, we merged all the three variables into a combined score, as described in methods. As expected, the combined score was highly correlated to the depth of coverage in the training dataset (r = 0.751; p<0.01; Fig 2D). To confirm its accuracy, we compared it to the mean target coverage in all studied genomic regions (independently of bias caused by copy number alterations) and to the read depth in bases harboring frequent germline or somatic single nucleotide variations (SNV), located in genes not used in the training dataset. There was strong positive correlation between the combined score and the depth of coverage in all these instances. In addition, we demonstrated strong correlation between the combined score and the 20x, 50x, and 100x target base coverage (r = 0.779, 0.790, and 0.792, respectively; p<0.01), as well as to other sequencing efficiency parameters (Table 2 and Fig 3A).

We next sought to define a pre-analytical cut-off value that could predict adequate sequencing results. For this end, we plotted the pre-sequencing variables (including the combined score) against the 50x coverage in our dataset, and defined 90% 50x coverage as a parameter for favorable results. According to this analysis, an FFPE storage time of 8.6 years, a PCR/QC ratio of 0.22, a DNA input of 960 ng, or a combined score of 266 were thresholds associated with sequencing efficiency (Fig 3B).

Low depth of coverage was characteristic of regions with low GC content

In addition to sample quality, we evaluated the effect of base composition on sequencing depth. To assess depth of coverage uniformity, we evaluated the mean normalized coverage across the genomic regions spanned by the designed probes. We demonstrated a wide variability, and observed that poor coverage was significantly associated with regions presenting lower

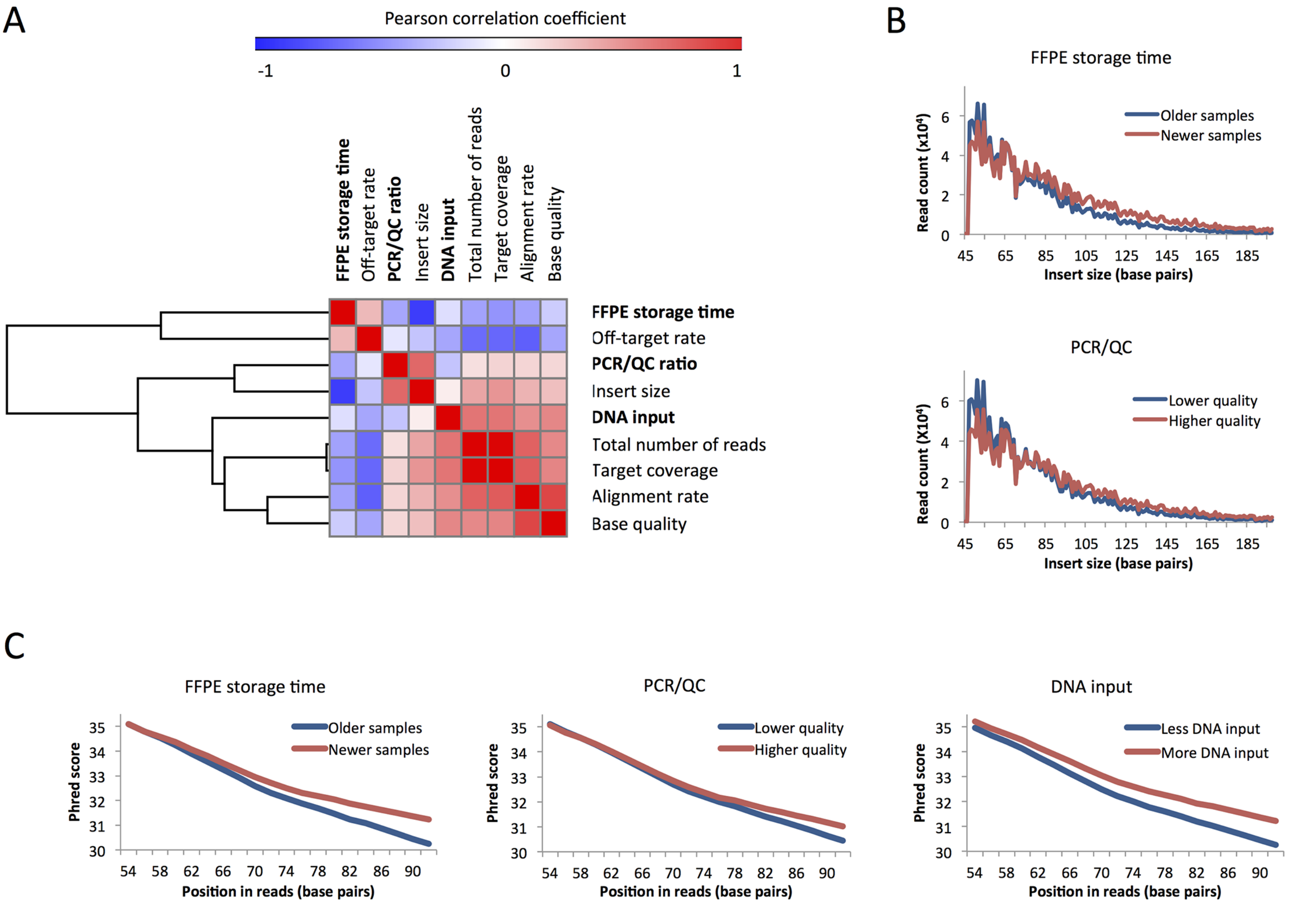

Fig 1. Three pre-analytical variables (FFPE storage time, PCR/QC ratio, and DNA input in the library preparation) were significantly correlated to most post-sequencing parameters (A). The pre-analytical variables were classified as below or above median values to illustrate the impact on insert size (B) and on read quality/Phred score (C). Abbreviations: FFPE, formalin-fixed paraffin-embedded tissue blocks; PCR/QC, PCR-based quality control.

doi:10.1371/journal.pone.0143092.g001

GC content (Fig 4A). The best coverage was observed in regions with 0.5–0.7 GC content ratios, with a marked deterioration below 0.4 ($p<0.01$). Subsequently, we stratified samples according to baseline quality (measured by the pre-sequencing combined score), and re-assessed the GC content effect. Notably, regions with low GC content (below 0.4) had worse coverage in every stratum, with a pronounced lower coverage in samples with poor pre-sequencing quality (Fig 4B).

Impact of coverage variability on gene hotspots

As sequencing metrics are ultimately a surrogate for optimal variant calling, we interrogated if sample quality and GC content would have an impact on target coverage at hotspot positions in *KRAS* and *EGFR*. As shown in Table 3, these genes exemplify the opposite extremes of the coverage spectrum observed herein. While hotspot positions in *EGFR* presented an ideal GC content (0.51–0.55) and an optimal coverage, *KRAS* showed lower GC content (0.33–0.36) and a dramatically worse coverage. The median number of reads in the *KRAS* codon 12 was only 51 (range 3–183), and the 20x and 50x target coverage were 87.9% and 51.4%, respectively. On the

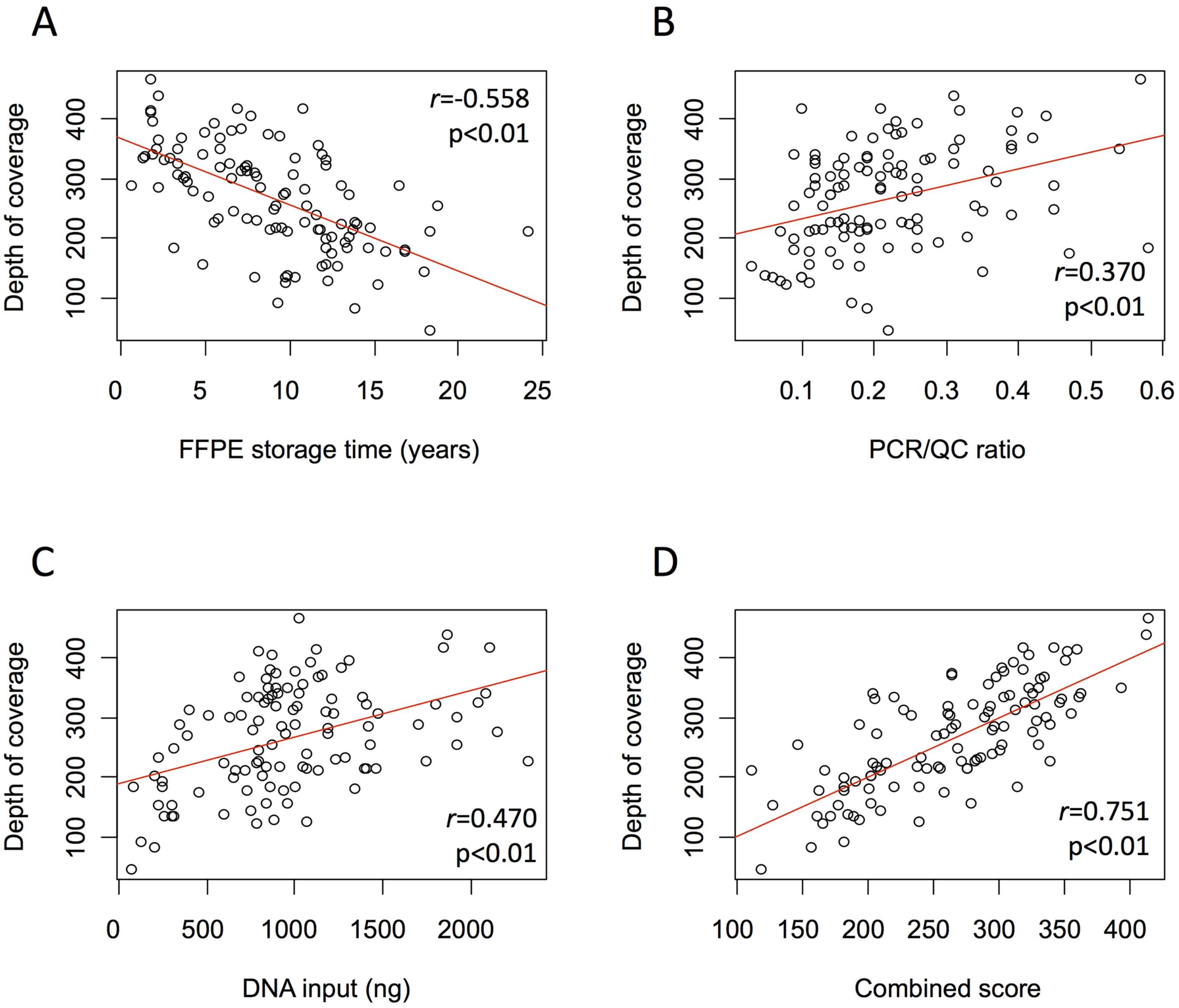

Fig 2. FFPE storage time (A), PCR/QC ratio (B), and DNA input (C) were correlated to sequencing depth of coverage. A combined score (D) was constructed based on these three parameters, and was highly correlated to sequencing depth. Abbreviations: FFPE, formalin-fixed paraffin-embedded tissue blocks; PCR/QC, PCR-based quality control.

doi:10.1371/journal.pone.0143092.g002

other hand, all *EGFR* hotspot positions presented satisfactory coverage (Table 3). As NGS variant calling pipelines will often include filters bases on a minimal coverage (eg. 20x or 50x), recurrent *KRAS* mutations could be easily missed due to low coverage. In fact, 12 out of 22 *KRAS* mutant cases were detected among samples with 50 reads or less, and 3 cases were found with less than 20 reads (Table F in S1 File), all of which were confirmed by visual read inspection. Poor coverage in these sites could also impair the sensitivity to detect low allele frequency mutations. Using a minimum combined score threshold of 266, the median number of reads in the *KRAS* codon 12 was 72.5 (range 18–183), and the 20x and 50x coverage was 98.2% and 80.4%. In line with recent reports, we observed a negative correlation between the PCR/QC

Table 2. Summary of correlation analyses between the pre-sequencing combined score and final sequencing parameters.

Parameter	Pearson correlation coefficient (*r*)	p-value
Mean coverage		
Overall dataset	0.628	< 0.01
20x coverage	0.779	< 0.01
50x coverage	0.790	< 0.01
100x coverage	0.792	< 0.01
Target base coverage		
ATM (rs659243)	0.650	< 0.01
EYS (rs9294631)	0.796	< 0.01
KRAS G12	0.703	< 0.01
EGFR L858	0.666	< 0.01
Other parameters		
Insert size	0.664	<0.01
Alignment rate	0.570	<0.01
Off-target rate	-0.423	<0.01
Base quality	0.476	<0.01

doi:10.1371/journal.pone.0143092.t002

ratio and the dinucleotide CpG to TpG transitions (*r* = -0.186; p = 0.049). Similar correlations were not seen to other pre-analytical variables or other dinucleotide changes.

Discussion

In the present study, we confirmed that clinical FFPE samples are a reliable source of DNA for targeted gene sequencing in cancer, provided adequate sample quality controls are exercised. We demonstrated that three pre-analytical variables—FFPE storage time, PCR/QC ratio, and DNA input in the library preparation—were significantly correlated to most parameters of sequencing efficiency. The combined examination of these features may be particularly useful to define sample adequacy for sequencing, as demonstrated by a pooled model derived from them, which was highly correlated to sequencing efficiency. We also showed a significant variability in depth of coverage within the genome, dependent on the GC content ratio. Genomic regions with lower GC content presented worse depth of coverage, and this effect was additive to sample quality.

It has been shown that NGS data from FFPE samples have smaller library insert sizes and greater coverage variability.[23] Herein, we went further to demonstrate that FFPE storage time, PCR/QC ratio, and DNA input may all predict sequencing quality within this group. For instance, FFPE storage time (or tumor age) was negatively correlated to several post-sequencing parameters, including depth of coverage, insert size, and base quality. These results are in line with the findings from Hedegaard et al,[12] who also showed better results when more recently obtained FFPE samples were used. In this sense, several factors may have negatively affected the results in older samples, including non-standardized methods used in the past for tumor fixation, processing, embedding, as well as storage time *per se*. On the other hand, Schweiger et al[39] did not find an influence of tumor age on sequencing depth, however that study was limited by a very small sample size (only 7 FFPE samples). Although older tumors may be an exception in the clinical setting, pathologists may need to utilize ancient FFPE samples in specific research settings. Some possible scenarios include retrospective analysis of unique samples acquired in clinical trials, study of rare diseases, and when tissue bank samples

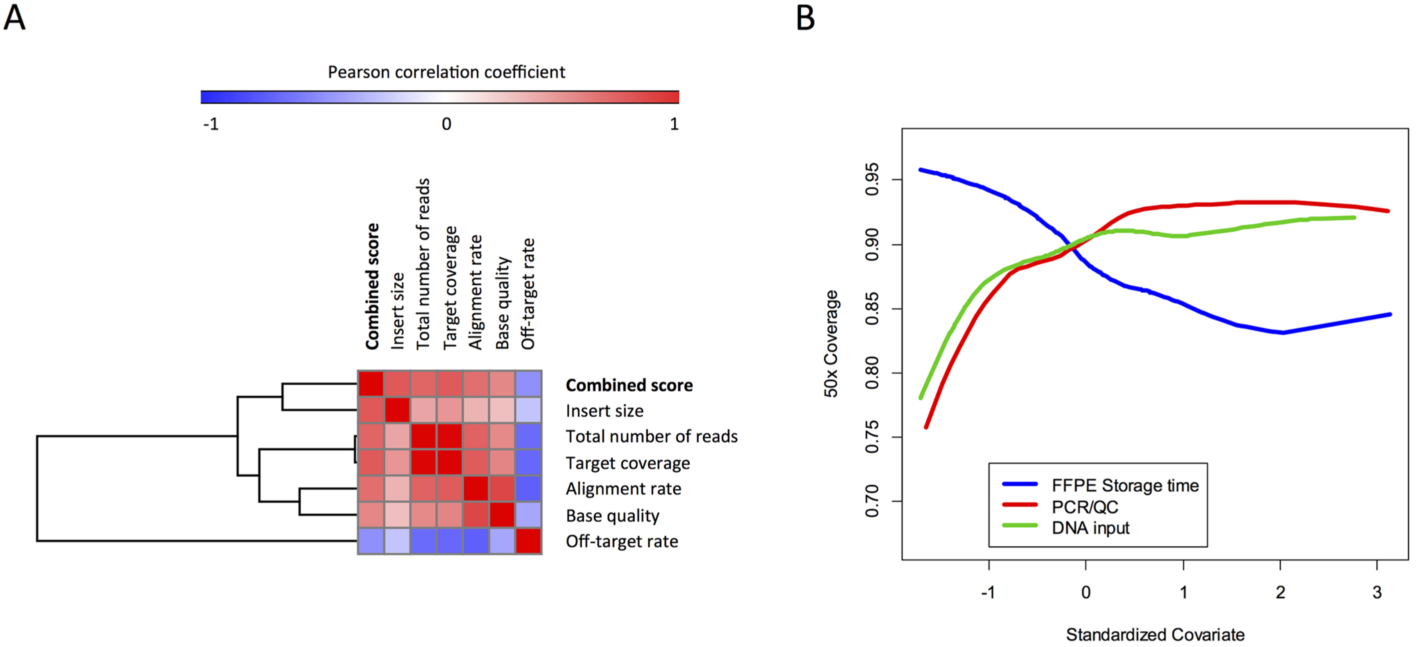

Fig 3. The combined score was strongly correlated to post-sequencing parameters (A). Correlation to the 50x coverage was used to define pre-analytical thresholds that could predict sequencing efficiency, and are illustrated by smoothing curves (B).

doi:10.1371/journal.pone.0143092.g003

are the only available source. If older samples are to be included, it may be essential to select for those with better DNA quality (using estimations of DNA fragmentation). When alternative sources are not an option, increasing the DNA input or the sequencing depth may help overcome the intrinsic limitations related to longer sample storage and older handling methods.

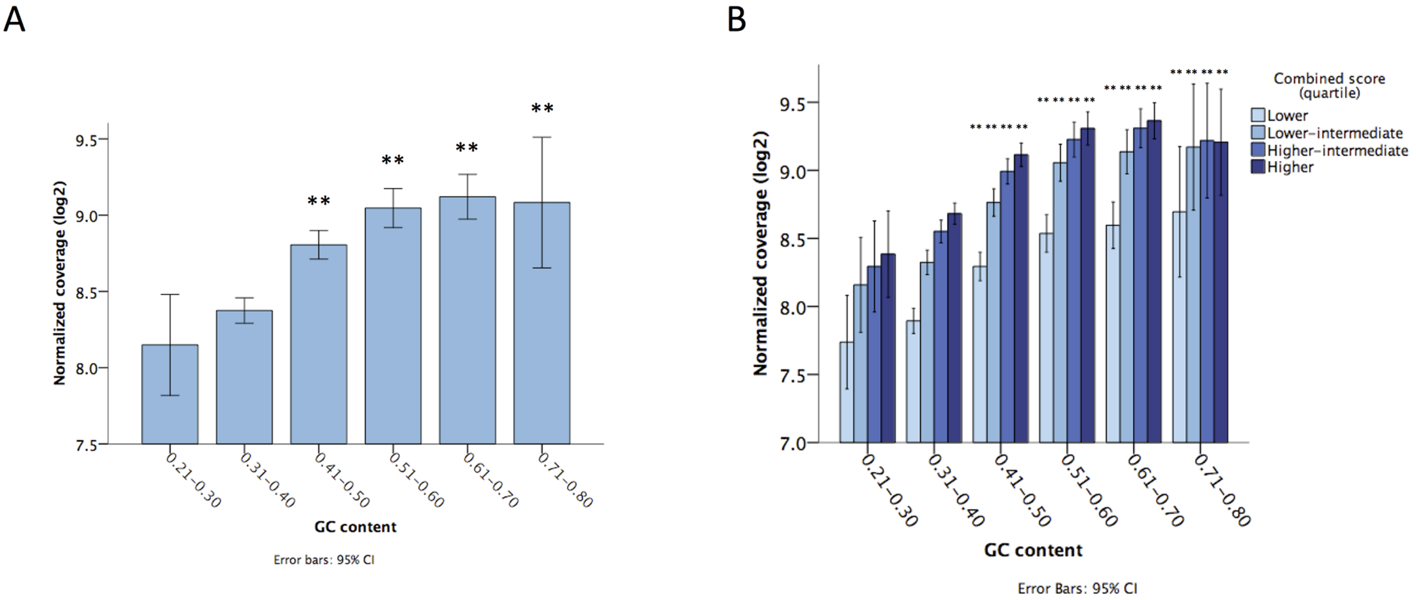

Fig 4. Normalized depth of coverage presents wide variability within the genome, with worse coverage observed in regions with lower GC content (A). The GC content effect was additive to sample quality to predict depth of coverage, as observed after stratifying sample quality using the combined score (B). Abbreviations: St. Dev., standard deviation. Obs: ** indicates significance at p<0.01.

doi:10.1371/journal.pone.0143092.g004

Table 3. GC content and depth of coverage at hotspot positions for recurrent somatic mutations in non-small cell lung cancer.

Genes	Codon	Exon	Chr position	GC content	Depth Median (range)
KRAS	G12	2	Chr12:25,398,285	0.33	51 (3–183)
	Q61	3	Chr12:25,380,276	0.36	741(161–1,513)
PIK3CA	E545	10	Chr3:178,936,091	0.30	159 (23–312)
	H1047	20	Chr3:178,952,085	0.37	1,067 (210–1864)
BRAF	V600	15	Chr7:140,453,136	0.34	459 (95–867)
NRAS	Q61	3	Chr1:115,256,529	0.40	648 (69–1,512)
MEK1	Q56	2	Chr15:66,727,451	0.50	808 (118–1,525)
	D67	2	Chr15:66,727,483	0.50	1,204 (188–2,185)
EGFR	G719	18	Chr7:55,241,708	0.51	1,409 (289–2,940)
	Dels	19	Chr7:55,242,470	0.53	503 (125–1,224)
	T790	20	Chr7:55,249,071	0.55	1,534 (344–3,019)
	L858	21	Chr7:55,259,524	0.54	918 (190–1,826)
ERBB2	G776	20	Chr17:37,880,996	0.58	930 (111–1593)

Abbreviations: Chr, chromosome; Dels, deletions.

doi:10.1371/journal.pone.0143092.t003

Different methods have been reported to assess quality of gDNA derived from clinical FFPE samples. These include verifying the A260/280 ratio using Nanodrop spectrophotometer (a ratio of 1.8 or greater suggests reasonable purity), calculating the estimated double-stranded DNA amount dividing Qubit® DNA estimation by Nanodrop (ratios of 0.4 or higher are ideal), running an aliquot of gDNA on an agarose gel or TapeStation (fragments of 200 bp or less indicate poor quality), or using a PCR-based approach.[30–33] In this study, we utilized a standard protocol recommended by the manufacturer, based on the PCR amplification of genomic regions of different sizes.[35] Low quality DNA will generate less abundant amplicons, simulating the expected results during target enrichment. This kind of analysis was independently correlated to depth of coverage, with higher ratios giving the best coverage. This PCR-based test is relatively straightforward, inexpensive, uses low amount of DNA as input, and are therefore easily applicable in most laboratories.

DNA input in the library preparation is an important predictor of sequencing success.[21] For the platform used in the current study, the manufacturer recommends a minimum of 225 ng of gDNA (estimated by fluorescence methods like PicoGreen® or Qubit®), which may be increased in the case of low quality DNA. Our data show that DNA input was correlated to sequencing depth, alignment rate, base quality, and off-target rate. More importantly, DNA input was interchangeable with other pre-sequencing parameters (tumor age, PCR/QC) to predict sequencing depth. It means that higher DNA input may often compensate for low-quality DNA, while high-quality DNA could be used at substantially lower input, as shown by the combined score analysis presented herein.

We generated a unique score that takes into account data from three pre-analytical variables that demonstrated independent impact on sequencing depth. Furthermore, we speculated on potential cut-off values for each of these variables that could help define tissue adequacy for sequencing. Though it may be attractive to consider these values in the routine of sequencing laboratories, some limitations need to be discussed. For instance, it is still uncertain if this evaluation could be applied to other settings, especially if distinct QC or NGS assays are employed. In addition, due to the retrospective nature of the current study, some specific data on histological processing could not be retrieved, like the amount of time spent between specimen collection and formalin fixation (ischemic time) and time spent in formalin fixation prior to processing. These and other non-assessable pre-analytical variables could have an additional impact on downstream sequencing analysis. Lastly, cut-off values may require local validation before implementation, and different metrics—like depth of coverage at specific hotspots—may be preferred as surrogates for optimal sequencing. Nonetheless, the concept of interchangeability between the pre-sequencing variables is likely universal, emphasizing the need to assess them in combination.

It has been demonstrated that hybridization-capture techniques may present significant coverage variability across the genome, with some nuances according to different platforms and base composition.[34, 40] In the current study, we confirmed a wide variability in depth of coverage within the target regions, with the best results achieved in genomic regions with 0.5–0.7 GC content ratios. On the other hand, genomic regions characterized by low GC content (below 0.4) presented significantly worse coverage, which was likely due to impaired probe hybridization and PCR amplification in these regions during library preparation.[34, 40] This effect was additive to sample quality in predicting sequencing depth. According to our analysis, regions with low GC content may be the most affected if samples with poor quality are used, achieving worse results in terms of depth of coverage. As different hybridization-capture platforms are currently available,[41] choosing those with the highest sequencing uniformity may be essential, especially when dealing with low-quality DNA.

In summary, we confirmed that FFPE samples are a reliable source for targeted gene sequencing in cancer, and validated sample storage time, DNA input, and DNA quality assessment as pre-analytical predictors of final sequencing metrics. We also showed that these variables are interchangeable (one variable could compensate for others), as demonstrated by a combined score analysis. Lastly, samples with poor quality yielded significantly worse sequencing results in genomic regions characterized by low GC content like in *KRAS*, emphasizing the relevance of sample quality controls.

Supporting Information

S1 Fig. The high rate of similar reads generated with Haloplex is exemplified by repetitive reads in a targeted exonic region in *EGFR*.
(TIFF)

S2 Fig. Representative images of the tumor cellularity among samples assessed with targeted gene sequencing. Tissue microarray (TMA) images were selected for convenience, although TMAs were not used in the DNA extraction process prior to sequencing.
(TIFF)

S3 Fig. Median number of reads (A) and target base coverage (B) across 113 lung tumor samples assessed with targeted gene sequencing. Each point represents a unique case. Abbreviation: PF, passing filter; PCT, percentage.
(TIFF)

S1 File. Supplemental tables A-F.
(DOCX)

Author Contributions

Conceived and designed the experiments: LHA CT KS WZ TL JA JRLS CGF DPC. Performed the experiments: LHA CT TL. Analyzed the data: LHA CT JZ LY TGN CJM ASY. Contributed reagents/materials/analysis tools: KS WZ JA DPC. Wrote the paper: LHA CT DPC.

References

1. Macconaill LE, Garraway LA. Clinical implications of the cancer genome. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2010; 28(35):5219–28. doi: 10.1200/JCO.2009.27.4944 PMID: 20975063; PubMed Central PMCID: PMC3020694.
2. Heinrich MC, Corless CL, Demetri GD, Blanke CD, von Mehren M, Joensuu H, et al. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2003; 21(23):4342–9. doi: 10.1200/JCO.2003.04.190 PMID: 14645423.
3. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine. 2004; 350(21):2129–39. doi: 10.1056/NEJMoa040938 PMID: 15118073.
4. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002; 417(6892):949–54. doi: 10.1038/nature00766 PMID: 12068308.
5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. doi: 10.1126/science.1235122 PMID: 23539594; PubMed Central PMCID: PMC3749880.
6. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446(7132):153–8. doi: 10.1038/nature05610 PMID: 17344846; PubMed Central PMCID: PMC2712719.
7. Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, et al. High-throughput oncogene mutation profiling in human cancer. Nature genetics. 2007; 39(3):347–51. doi: 10.1038/ng1975 PMID: 17293865.
8. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–8. doi: 10.1038/nature12213 PMID: 23770567; PubMed Central PMCID: PMC3919509.
9. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nature genetics. 2013; 45(10):1127–33. doi: 10.1038/ng.2762 PMID: 24071851.
10. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature biotechnology. 2013; 31(11):1023–31. doi: 10.1038/nbt.2696 PMID: 24142049.
11. MacConaill LE, Campbell CD, Kehoe SM, Bass AJ, Hatton C, Niu L, et al. Profiling critical cancer gene mutations in clinical tumor samples. PloS one. 2009; 4(11):e7887. doi: 10.1371/journal.pone.0007887 PMID: 19924296; PubMed Central PMCID: PMC2774511.
12. Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, et al. 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. 2014; 9(5):e98187. doi: 10.1371/journal.pone.0098187 PMID: 24878701; PubMed Central PMCID: PMC4039489.
13. Kris MG, Johnson BE, Berry LD, Kwiatkowski DJ, Iafrate AJ, Wistuba II, et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA: the journal of the American Medical Association. 2014; 311(19):1998–2006. doi: 10.1001/jama.2014.3741 PMID: 24846037.
14. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011; 470(7333):198–203. doi: 10.1038/nature09796 PMID: 21307932.
15. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nature reviews Genetics. 2010; 11(10):685–96. doi: 10.1038/nrg2841 PMID: 20847746.
16. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155(1):27–38. doi: 10.1016/j.cell.2013.09.006 PMID: 24074859; PubMed Central PMCID: PMC3969849.
17. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nature methods. 2010; 7(2):111–8. doi: 10.1038/nmeth.1419 PMID: 20111037.
18. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461(7261):272–6. doi: 10.1038/nature08250 PMID: 19684571; PubMed Central PMCID: PMC2844771.
19. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology. 2009; 27(2):182–9. doi: 10.1038/nbt.1523 PMID: 19182786; PubMed Central PMCID: PMC2663421.

20. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer discovery. 2012; 2(1):82–93. doi: 10.1158/2159-8290.CD-11-0184 PMID: 22585170; PubMed Central PMCID: PMC3353152.

21. Chen K, Meric-Bernstam F, Zhao H, Zhang Q, Ezzeddine N, Tang LY, et al. Clinical Actionability Enhanced through Deep Targeted Sequencing of Solid Tumors. Clinical chemistry. 2015. doi: 10.1373/clinchem.2014.231100 PMID: 25626406.

22. Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, et al. Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. The Journal of molecular diagnostics: JMD. 2013; 15(2):234–47. doi: 10.1016/j.jmoldx.2012.11.006 PMID: 23321017.

23. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. The Journal of molecular diagnostics: JMD. 2013; 15(5):623–33. doi: 10.1016/j.jmoldx.2013.05.004 PMID: 23810758.

24. Cottrell CE, Al-Kateb H, Bredemeyer AJ, Duncavage EJ, Spencer DH, Abel HJ, et al. Validation of a next-generation sequencing assay for clinical molecular oncology. The Journal of molecular diagnostics: JMD. 2014; 16(1):89–105. doi: 10.1016/j.jmoldx.2013.10.002 PMID: 24211365.

25. Auerbach C, Moutschen-Dahmen M, Moutschen J. Genetic and cytogenetical effects of formaldehyde and related compounds. Mutation research. 1977; 39(3–4):317–61. PMID: 331091.

26. Feldman MY. Reactions of nucleic acids and nucleoproteins with formaldehyde. Progress in nucleic acid research and molecular biology. 1973; 13:1–49. PMID: 4573489.

27. Karlsen F, Kalantari M, Chitemerere M, Johansson B, Hagmar B. Modifications of human and viral deoxyribonucleic acid by formaldehyde fixation. Laboratory investigation; a journal of technical methods and pathology. 1994; 71(4):604–11. PMID: 7967515.

28. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Science translational medicine. 2012; 4(136):136ra68. doi: 10.1126/scitranslmed.3003726 PMID: 22649089.

29. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. The American journal of pathology. 2002; 161(6):1961–71. doi: 10.1016/S0002-9440(10)64472-0 PMID: 12466110; PubMed Central PMCID: PMC1850907.

30. van Beers EH, Joosse SA, Ligtenberg MJ, Fles R, Hogervorst FB, Verhoef S, et al. A multiplex PCR predictor for aCGH success of FFPE samples. British journal of cancer. 2006; 94(2):333–7. doi: 10.1038/sj.bjc.6602889 PMID: 16333309; PubMed Central PMCID: PMC2361127.

31. Didelot A, Kotsopoulos SK, Lupo A, Pekin D, Li X, Atochin I, et al. Multiplex picoliter-droplet digital PCR for quantitative assessment of DNA integrity in clinical samples. Clinical chemistry. 2013; 59(5):815–23. doi: 10.1373/clinchem.2012.193409 PMID: 23403697.

32. Wang F, Wang L, Briggs C, Sicinska E, Gaston SM, Mamon H, et al. DNA degradation test predicts success in whole-genome amplification from diverse clinical samples. The Journal of molecular diagnostics: JMD. 2007; 9(4):441–51. doi: 10.2353/jmoldx.2007.070004 PMID: 17690213; PubMed Central PMCID: PMC1975106.

33. Sah S, Chen L, Houghton J, Kemppainen J, Marko AC, Zeigler R, et al. Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies. Genome medicine. 2013; 5(8):77. doi: 10.1186/gm481 PMID: 24001039; PubMed Central PMCID: PMC3978876.

34. Samorodnitsky E, Datta J, Jewell BM, Hagopian R, Miya J, Wing MR, et al. Comparison of custom capture for targeted next-generation DNA sequencing. The Journal of molecular diagnostics: JMD. 2015; 17(1):64–75. doi: 10.1016/j.jmoldx.2014.09.009 PMID: 25528188; PubMed Central PMCID: PMC4279426.

35. Agilent T. FFPE-derived DNA quality assessment: In preparation for HaloPlex target enrichment. 2013.

36. 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 PMID: 19505943; PubMed Central PMCID: PMC2723002.

37. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. doi: 10.1093/bioinformatics/btq033 PMID: 20110278; PubMed Central PMCID: PMC2832824.

38. 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 research. 2010; 20(9):1297–303. doi: 10.1101/gr.107524.110 PMID: 20644199; PubMed Central PMCID: PMC2928508.

39. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. PloS one. 2009; 4(5):e5548. doi: 10.1371/journal.pone.0005548 PMID: 19440246; PubMed Central PMCID: PMC2678265.

40. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nature biotechnology. 2011; 29(10):908–14. doi: 10.1038/nbt.1975 PMID: 21947028; PubMed Central PMCID: PMC4127531.

41. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nature medicine. 2014; 20(6):682–8. doi: 10.1038/nm.3559 PMID: 24836576; PubMed Central PMCID: PMC4048335.