Technical Note

Beyond Nanograms: Amplifiable Genomes to assess NGS library complexity from FFPE samples

The advent of Next-Generation Sequencing (NGS) and targeted library preparation represents a major advance for the study of cancer genomics. However, NGS-enabled cancer genomics presents two unique challenges. First, tumor samples are highly heterogeneous, and key mutations are often found in a very small fraction of tumor cells. Second, tumor samples are often processed for histological analysis and archiving prior to molecular testing. Formalin fixation and paraffin embedding (FFPE) of biopsies is the standard means of preservation and archiving of samples in clincial oncology; however, this harsh chemical treatment causes significant damage to the nucleic acid in the tissue. Here we discuss sample heterogeneity, input nucleic acid damage, and their implications for detection of single nucleotide variants (SNVs) and copy number variations (CNVs). We also introduce the PreSeq™ DNA QC Assay, a simple qPCR assay that facilitates the quantification of the number of sequenceable copies of input genomic DNA, thereby maximizing information recovery from NGS libraries.

Idea in Brief

- The sensitivity of NGS-based variant detection is derived from the number of unique fragments present in a region of interest.
- Traditional amplicon-based enrichment techniques lack the ability to count how many unique fragments are sequenced.
- Mass is not predictive - it is the number of amplifiable genomes that determines the library complexity and ultimately, the sensitivity of NGS assays used in somatic mutation testing.
- The PreSeq DNA QC assay provides a method of quantifying the number of input molecules that are present in a given sample, and the use of this assay can help rescue or avoid libraries that would otherwise fail Archer Analysis QC.

Application	Required quantity of amplified genomes
4 nM library yield	1100
Archer Analysis QC Pass	3750
≥2% allelic frequency mutation detection	6500
CNV calling (≤3-fold change)	1900

Introduction

Somatic variant detection in tumor samples, unlike detection of germline variants, requires the ability to identify mutations present at allele fractions (AF) far below 50%. The reproducible identification of low AF variants requires the interrogation of hundreds (or potentially thousands) of unique input molecules in order to overcome sampling noise. For example, if one wishes to detect a KRAS G12C mutation present at 1% AF, one typically needs to capture and interrogate a minimum of about 1000 distinct copies of the KRAS locus before at least 10 mutant copies of KRAS can be captured with favorable odds (Figure 1).

Unfortunately, damage to input DNA molecules has the effect of decreasing the number of sequenceable input molecules contained in a given mass of DNA. Although 10ng DNA theoretically contains ~3000 copies of every genomic locus, in reality only a small fraction of those molecules may be available for library preparation. Therefore, nucleic acid damage, including that resulting from the FFPE process, can dramatically reduce the ability to detect variants present at low AF.

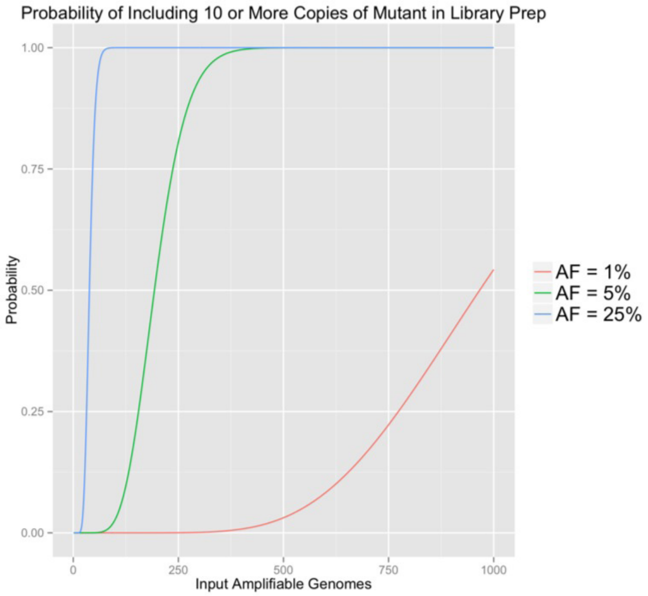

Figure 1. The theoretical likelihood of sampling at least 10 mutant molecules as a function of allelic frequency (AF) and number of molecules integrated. The probability of capturing 10 or more mutant molecules from a population was modeled with a binomial distribution where the number of independent Bernoulli trials (N) is plotted on the x-axis. The probability of success (the probability that any given capture event picks a molecule containing the mutation of interest), p, was set as the allele fraction. The y-axis represents the probability of choosing 10 or more mutant molecules, i.e. Pr (X>= 10) based on each binomial distribution.

While FFPE provides an inexpensive, long-lasting medium for preservation of tissue, formalin fixation also induces DNA fragmentation and several chemical modifications, many of which are irreversible, including: cross-linking of proteins to nucleic acids, strand cleavage, and base modifications (1, 2). Together, these chemical modifications of nucleic acids can inhibit down-stream enzymatic manipulation and interfere with NGS library preparation (2). In addition to the chemical and physical damage related to the fixation process, both storage conditions and time can cause additional nucleic acid degradation. To complicate matters further, the fixation procedure and processing of samples can vary across procurement organizations, tissue types, and biopsy sizes (1).

Several methods for measuring DNA mass and assessing fragment length and integrity are available. For example, DNA quantification by UV absorption or fluorescence spectroscopy are quick and straightforward techniques to determine DNA mass in a sample. Despite accurate quantification by these methods, neither provides information about the degree of fragmentation or chemical damage of the sample. DNA fragment length and crude mass can be approximated by capillary or gel electrophoresis; however, capillary electrophoresis is not amenable to the high throughput analysis of samples, and does not scale with the rapid growth of NGS and automated workflows. Furthermore, electrophoretic techniques provide no information about the quantity of DNA that can be amplified by PCR and therefore made into a NGS library. As a result, we and others have found that none of these aforementioned methods are sufficient in determining input DNA suitability for NGS applications (3).

In high-throughput sequencing laboratories, an upfront DNA QC assay that can assess sample quality, guide input recommendations, and approximate AF sensitivity can save time and money. In order to streamline the DNA QC testing process, we have developed the PreSeq DNA QC assay, a qPCR assay that facilitates rapid estimation of the quantity of DNA in a sample that is available for library generation. By measuring the quantity of DNA fragments amenable

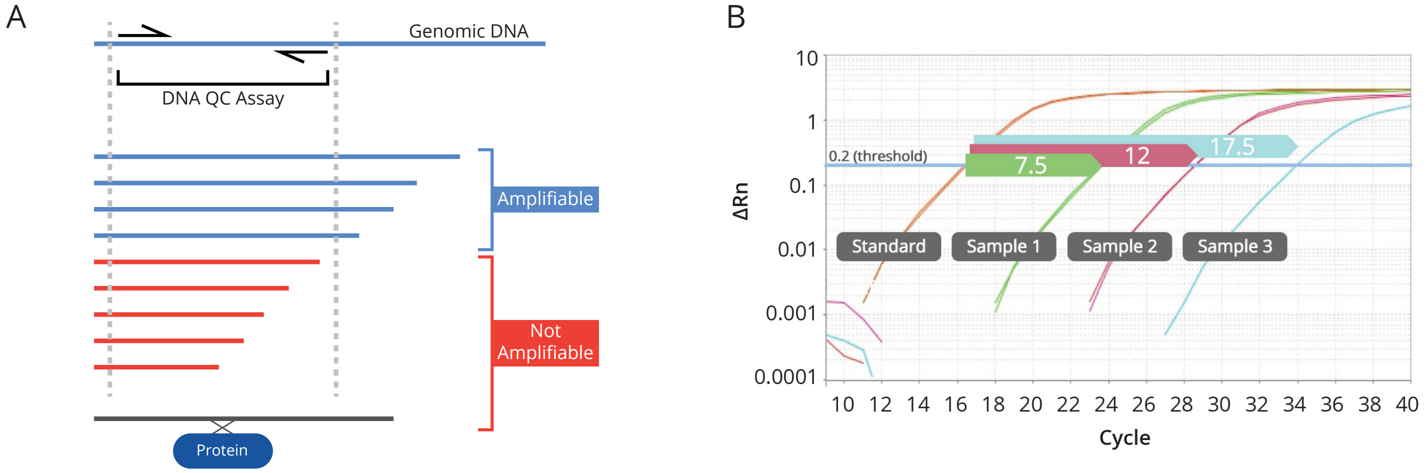

Figure 2. PreSeq DNA Assay is a simple assay for quantifying the concentration of available genomic DNA molecules in a sample. (A) The PreSeq DNA QC assay is a qPCR-based assay for the detection of amplifiable genomic DNA copies. Genomic DNA may become unavailable for library preparation, and therefore PCR amplification, due to over-fragmentation or chemical damage. (B) When 5 ng of human genomic DNA is input into the PreSeq DNA QC assay, the Cq value of a sample can be compared to a known standard (or other high quality human genomic input) to quickly establish DNA QC score and the concentration of genomic copies available for library preparation.

to amplification in a sample prior to library preparation, the PreSeq assay can rescue samples of borderline quality. Here we demonstrate the utility and advantages of the PreSeq DNA QC assay in screening samples to identify those of sufficient quality to yield high-complexity targeted libraries with the Archer™ VariantPlex™ Solid Tumor panel.

PreSeq DNA QC Assay Product Concept and Workflow

The PreSeq DNA QC assay is used to evaluate the quantity of amplifiable DNA in a sample relative to a control template (Figure 2). The assay is comprised of two separate reactions: one that targets a 100 bp genomic DNA sequence in the sample and one that targets a synthetic control template (the Synthetic Standard). A comparison of the two quantification cycles (Cq) results in a DNA QC score, a ΔCq value that provides a quantitative measure of DNA quality, and can be used to estimate the concentration of sequenceable copies of genomic DNA present in the sample. More specifically, we define an amplifiable genome as one complete haploid genome of sufficient quality and fragment length to be detectable by PCR. Thus for each amplifiable genome present in a sample, there is on average a single copy of each genomic locus that is available for library generation.

The DNA QC score is defined as the ΔCq between the sample and Synthetic Standard template, when a uniform fluorescence intensity threshold is applied to both (Equation 1). The PreSeq DNA QC score may be used to estimate the number of amplifiable genomes present in a given DNA sample.

Equation 1

DNA QC Score = Sample Cq – Synthetic Standard Cq

The number of amplifiable genomes as a function of input DNA mass and DNA QC score in a sample may be estimated in either of two ways. The first method requires a normal, high quality genomic DNA sample (for example, Genome in a Bottle, provided by NIST). In this method, the reference genomic DNA is assumed to be fully amplifiable, and therefore contain 333 amplifiable genomes per nanogram of DNA (Equation 2.1). The second method utilizes a non-human control sequence at a known concentration (the Synthetic Standard) allowing the number of amplifiable genomes in a given sample to be determined using Equation 2.2. The remainder of this paper illustrates how this number can then be used to determine if the sample can yield informative libraries of sufficient quality and complexity to support confident variant and copy number calling.

Equation 2.1: Calculation of Amplifiable Genomes from Reference DNA

Amplifiable Genomes per ng = $2^{-(\text{DNAQC ScoreSample- DNAQC ScoreReferenceDNA})}$ x 333copies/ng

Equation 2.2: Calculation of Amplifiable Genomes from Synthetic Standard

Amplifiable Genomes per ng = $2^{-(\text{DNAQC ScoreSample})}$ x 499427.9 Synthetic Copies x 0.647 (length correction)/5 ng sample input

The Archer VariantPlex Assay Overview

Archer VariantPlex assays used in these studies are targeted DNA sequencing assays that are used to enrich and identify a number of variants, including single nucleotide variants (SNVs), insertions/deletions (indels), and copy number variants (CNVs). The underlying enrichment chemistry, Anchored Multiplex PCR (AMP™), utilizes ligated adapters with universal priming binding sites combined with gene specific primers (GSPs) in order to selectively amplify genomic areas of interest (4). Two rounds of nested PCR enrichment utilize universal primers combined with GSP1 and GSP2 primers, respectively. Following library preparation, VariantPlex libraries are sequenced, and results are analyzed via the Archer Analysis software. As part of the variant calling process, Archer Analysis uses molecular barcodes and unique start sites to deduplicate the reads, such that each unique read represents information captured from a unique copy of genomic DNA.

Crude input mass does not predict library yield or complexity

To assess the utility of spectrophotometric and dye-based DNA quantification methods for predicting successful NGS library generation, we screened 202 archived FFPE samples, each characterized by multiple methods, with the Archer VariantPlex Solid Tumor Panel (AK0051-8) and Archer DNA Universal Reagent Kit (AK0037-8). Input DNA mass was measured with the Qubit dsDNA HS Assay kit (Life Technologies, Cat # Q32854) and library concentration was measured by KAPA BioSystems Library Quantification Kit (Cat# KK4824). The same samples were then assayed with the PreSeq DNA QC Assay and used input into the Archer VariantPlex Solid Tumor target enrichment preparation. Although we normalized mass input into the VariantPlex library preparation, the total library yields (measured by KAPA) varied from 0-66nM (in 20µl) with numerous samples failing to produce adequate library concentrations for sequencing (Figure 3A). DNA mass is therefore a poor predictor of successful library preparation.

Archer Analysis is a free software tool that is used in conjunction with Archer assays for variant calling and reporting. Archer Analysis provides a simple QC metric to assess DNA library complexity. This met-

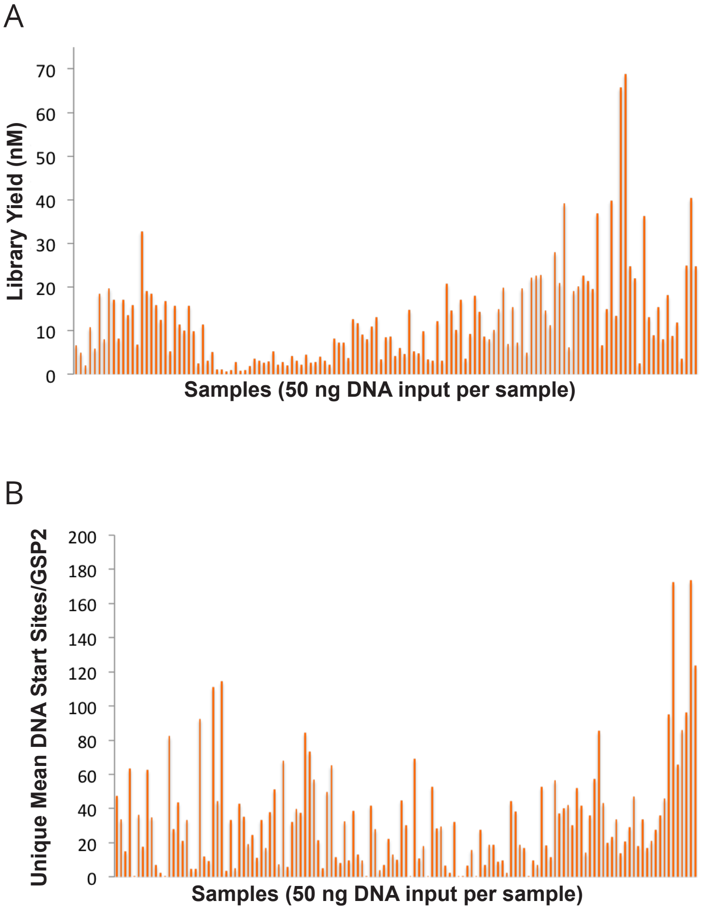

Figure 3: DNA mass is not an accurate predictor of library yield or complexity. The distribution of library yield from 202 FFPE samples (a) and library complexity from 120 unique FFPE samples (b). In each case, 50ng DNA, as measured by Qubit, was used as input for the Archer VariantPlex Solid Tumor Panel library preparation.

ric, the mean number of unique DNA start sites per GSP2, is useful for determining the theoretical sensitivity of variant calling in a given sequencing run. When we examined the mean number of unique start sites per GSP2 in each library produced by the 120 sequenceable FFPE samples and 4 BioChain samples, we found that despite normalization of the input mass to 50ng, the libraries varied in complexity from extremely low (< 1) to very high (> 173) mean start sites per GSP2 (Figure 3B). These data further support that input mass alone does not correspond with molecular complexity needed for sensitive, targeted sequencing applications.

PreSeq DNA QC predicts successful library preparation yield.

We next examine whether the PreSeq DNA QC Assay could more accurately predict whether a given FFPE input would successfully produce targeted NGS libraries. For each of the 202 FFPE samples in our archive, we used PreSeq DNA QC to estimate the number of amplifiable genomes that were input into the Archer VariantPlex Solid Tumor assay. We then used Receiver-Operating Characteristics (ROC) analysis to determine a predictive threshold for the number of amplifiable genomes of input that would likely result in production of at least 4nM library, a concentration high enough to permit sequencing. The ROC analysis shows that number of amplifiable genomes in the input held significant value for generating Illumina libraries concentrated enough to sequence (Figure 4A). Specifically, the ROC analysis revealed that input of at least 1108 amplifiable genomes is highly predictive of a 4nM (in 20µl) library preparation, with 82% specificity, 86% sensitivity, and an area under the curve (AUC) of 0.90. Thus, unlike crude mass, a measurement of amplifiable genomes of input is predictive of successful generation of library yield.

The established cut-off obtained by PreSeq DNA QC for amplifiable material can be used to adjust DNA input and increase the likelihood of producing sufficient library yield. For example, if a sample has a DNA QC score of 13, then 50ng of input would contain ~320 amplifiable genomes, and most likely fail to generate adequate library yield. However, by increasing the DNA input amount to 200ng, or ~1281 amplifiable genomes, the sample would be much more likely to generate sequenceable yield (Figure 4B). However, we also found that library yield from heavily degraded samples - those with DNA QC scores above 15 - could not be rescued by using more input (data not shown).

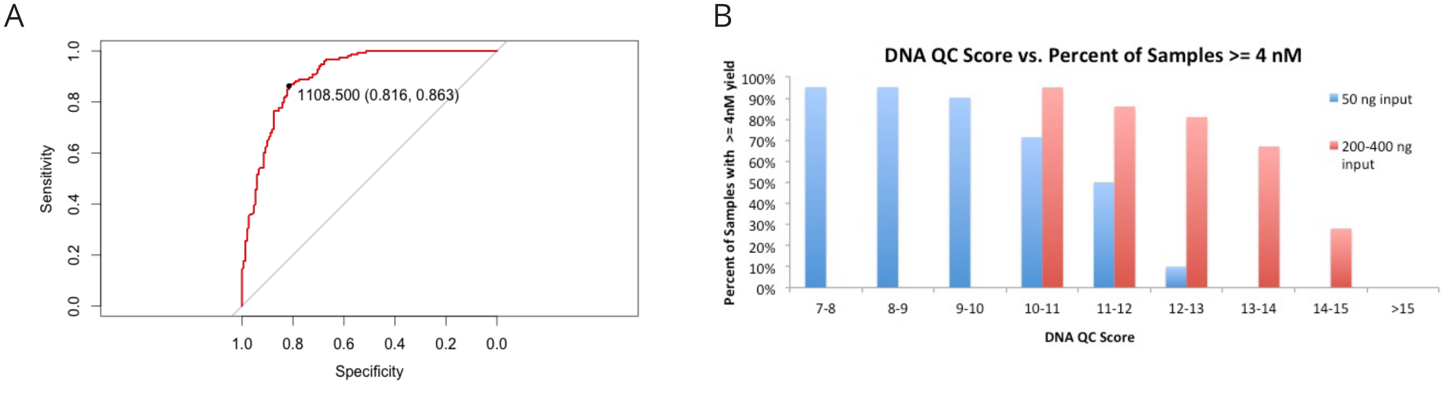

Figure 4: qPCR quantification of amplifiable genomes identifies samples that will yield sequenceable library with Archer VariantPlex Solid Tumor panel. (A) The 202 unique archived FFPEs used as input for the Archer VariantPlex Solid Tumor panel were characterized by the PreSeq DNA QC Assay to determine the number of amplifiable genomes present in the 50ng input. We considered libraries that produced 20μL of at least 4nM library to be successful. We then carried out ROC analysis to determine if the number of amplifiable genomes of input was predictive of successful library preparation. Libraries in which at least 1108 genomes were used as input were successful 86% of the time. (B) The library yield from low quality samples can be rescued by increasing the number of amplifiable genomes of input above the threshold of 1108. The percent of samples producing a library concentration of at least 4nM in 20μL was plotted for samples that were binned by DNA QC score. Blue bars represent input quantities of 50ng, and red bars indicate input quantities of 200 – 400ng. For poor quality samples (those with DNA QC scores ≥ 11), increasing input quantity such that the number of amplifiable genomes (indicated above each bar in graph) approaches 1108 restores library yield.

Amplifiable genomes of input DNA drive library complexity, coverage and sensitivity of variant calling

In some instances, we found that although inputs with greater than 1100 amplifiable genomes produced sequencable library yields, many of these libraries were of low complexity, and therefore failed Archer Analysis QC requirement of 50 mean unique DNA start sites per GSP2. A ROC analysis was conducted to determine if the amplifiable genomes of input could predict Analysis QC passes. To assess this, we created libraries using the Archer VariantPlex Solid Tumor assay (as previously described) from a set of 82 unique FFPE samples, which were sequenced and filtered to a read depth of at least 1.9 million reads. We found that Analysis QC passes can be predicted with 79% specificity and 92% sensitivity at an input threshold of 3750 amplifiable genomes (Figure 5). Overall, the ROC plot has an AUC of 0.90, indicating that the number of amplifiable genomes is a robust predictor of passing Analysis QC pass. Notably, the QC score needed to produce libraries that pass Analysis QC was found to be significantly greater than the cut-off to simply generate a sequenceable library, meaning that although low input quantities (below 3750 amplifiable genomes) can produce a sequenceable library, these libraries may not be complex enough to pass Archer Analysis QC.

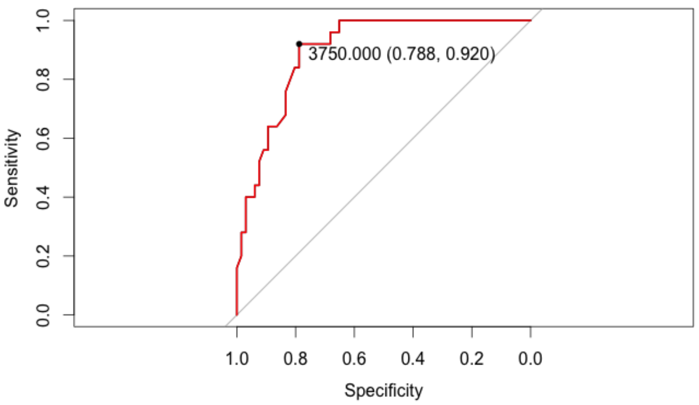

Figure 5. Input complexity is predictive of Archer Analysis QC pass. Archer Analysis uses a simple measure of library complexity, the average number of DNA start sites per GSP2 (default is ≥ 50), as a post-sequencing library QC. We carried out ROC analysis to determine if the total number of amplifiable genomes of input could serve as a predictor of Archer Analysis QC pass. For the 82 FFPE samples analyzed, we found that 64% of samples passed Analysis QC if at least 3750 amplifiable genomes of input were used for library preparation.

When we further examined the relationship between amplifiable genomes of input and mean unique DNA start sites per GSP2 with a single sample, we found these two values were closely related. We therefore produced Archer VariantPlex Solid Tumor libraries from increasing quantities of the Quantitative Muliplex FFPE Reference Standard (HorizonDx, Cat # HD200), while tracking the number of amplifiable genomes of input with PreSeq. We found the library complexity metric, mean unique DNA start sites per GSP2, to be very tightly correlated with the number of amplifiable genomes of input (Figure 6A). Therefore, the total number of amplifiable genomes of input drives the complexity of the final targeted libraries.

Next, we examined the relationship between input genomes and variant calling sensitivity by examining unique read coverage over the ~51 kbp of targeted sequence in Archer VariantPlex Solid Tumor assay with varying input quantities. As previously mentioned, each unique read represents information captured from a unique copy of genomic DNA. We found that unique read coverage improved dramatically as the number of input genomes increased (Figure 6B). Thus, it is the amplifiable genomes of input material that predicts Archer Analysis QC pass, library complexity, and overall coverage.

Finally, we examined the relationship between library complexity and variant detection sensitivity. The HorizonDx Quantitative Multiplex FFPE Reference Standard (Horizon #HD200) has 16 clinically relevant variants (SNVs and small deletions) ranging in AF from 32.5% to 0.9%, each verified by droplet digital PCR. At input quantities below 1600 amplifiable genomes, low start site complexity was observed and this was correlated with sub-optimal variant detection sensitivity. As number of input genomes increased, both the number of unique start sites and variant detection sensitivity also increased (Figure 6C). In fact, at approximately 6500 amplifiable genomes of input, 100% of variants, including those at allelic frequencies below 2% were detected, consistent with the concept that as more unique molecules encompassing a given genomic locus are interrogated, assay sensitivity goes up, and lower AF variants are more readily detected.

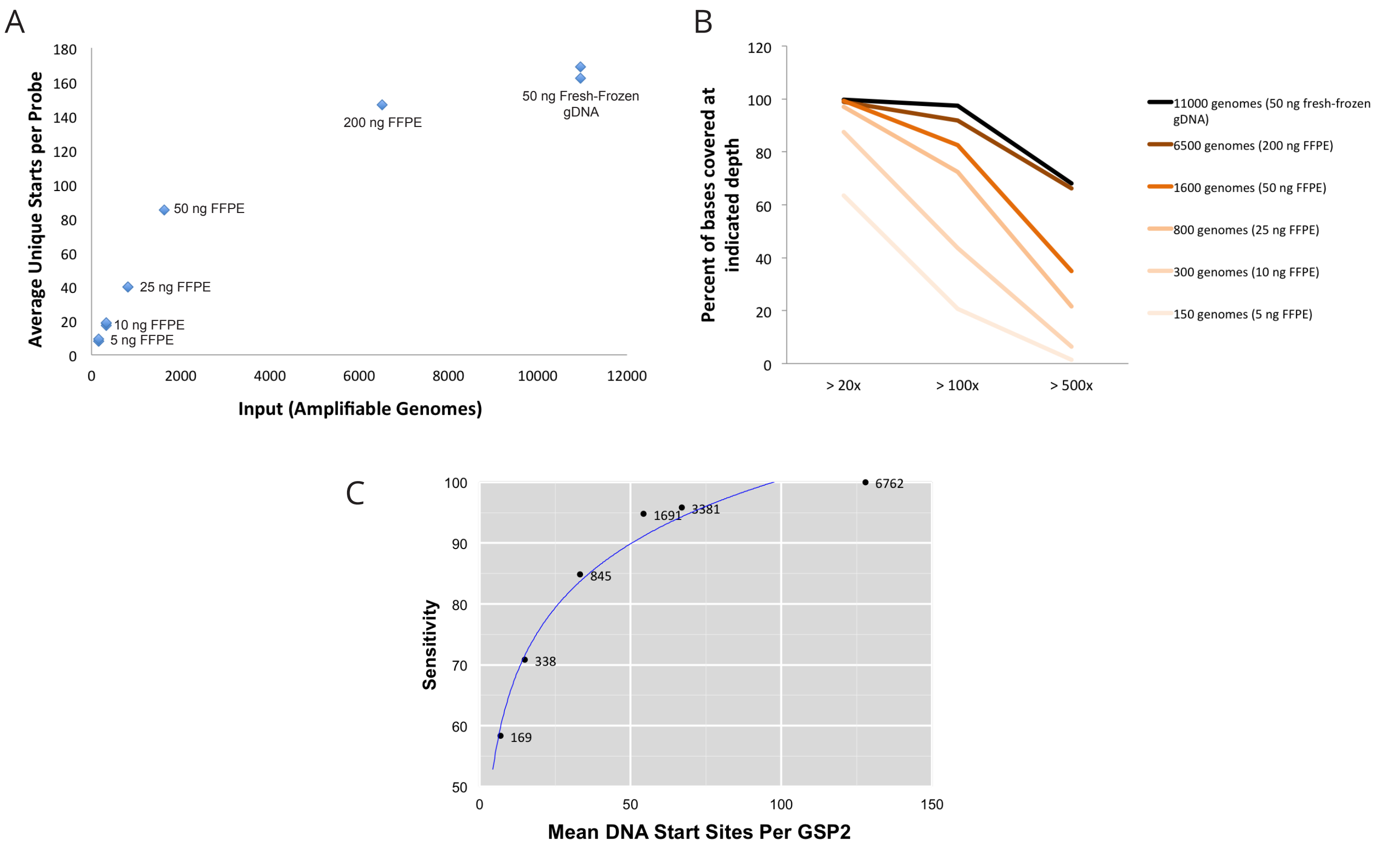

Figure 6: Input complexity drives library complexity, per-base unique coverage, and variant detection sensitivity. (A) Library complexity, as measured by start sites per GSP2, tracks closely with the total number of amplifiable genomes used as input. The graph shows the library complexity as a function of total number of input genomes for an FFPE and a high quality, fresh frozen input. Total mass of input (ng) is shown next to each point. (B) The fraction of bases covered (y-axis) at the indicated depth (x-axis) is shown for several different input amplifiable genome quantities. Note that as the number of amplifiable genomes of input increases, the fraction of bases covered at high depth increases. (C) The same set of libraries was examined for both their complexity (start sites per GSP2), and sensitivity to the known variants. As amplifiable genomes of input increased, so did the library complexity. Both amplifiable genomes and library complexity correlated with increasing sensitivity to known variants, including the EGFR T790M variant present at an AF of 0.9%, in the sample.

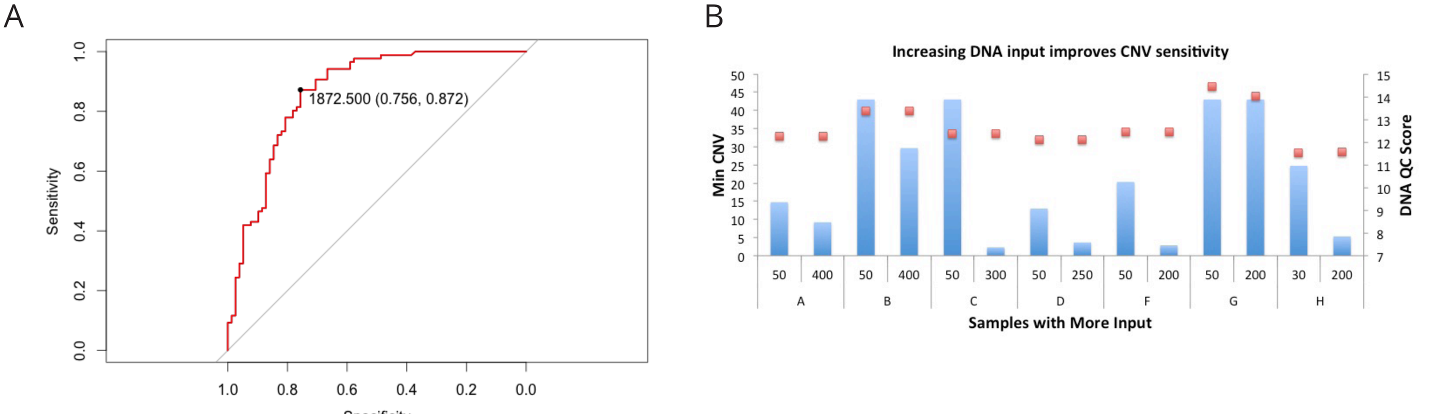

Figure 7: Increasing number of amplifiable genomes of input improves CNV sensitivity. (A) ROC analysis of the number of amplifiable genomes of input as a predictor for high sensitivity (defined ability to detect a CNV of magnitude ≤ 3X) CNV detection. Minimum CNV magnitudes were determined using a bootstrap power analysis (see materials and methods). The data is representative of 139 unique samples. (B) For seven samples of varying DNA QC score (red dots), two different input quantities, low (left bar) and high (right bar) were used to generate VariantPlex Solid Tumor libraries. After sequencing, the minimum CNV detectable in each library was calculated. The high quantity input consistently produced higher sensitivity CNV calling, as reflected by the decrease in the minimum CNV.

Amplifiable genomes of input DNA drive sensitivity of CNV detection

In addition to SNV/InDel detection, Archer VariantPlex Solid Tumor enables the detection of CNVs. To determine how DNA input quality and quantity contribute to the sensitivity of CNV calling, we applied a bootstrap power analysis to copy number calling in libraries generated from our 202 archival FFPE samples. Briefly, the bootstrap power analysis informatically simulates copy number changes in a single gene, while maintaining the read count noise characteristics of that sample. After performing multiple simulations, the minimum magnitude CNV that would be called with 95% confidence is reported. We again used ROC analysis to determine if the number of amplifiable genomes of input could predict high-sensitivity CNV (magnitude of 3X or less) calling in a sample. The total number of amplifiable genomes of input proved to be an effective predictor of high-sensitivity CNV calling, with a minimum input threshold of 1872 genomes maximizing the predictive value (Figure 7A). Consistent with this observation, we found that increasing input genomes generally improved CNV sensitivity, as determined by the minimal detectable CNV at each input dose (Figure 7B). Therefore, for even relatively low quality samples, reliable CNV and SNV/InDel calling may be accomplished with increased sample inputs.

Discussion

Laboratories typically have limited amounts of DNA present in FFPE cancer samples bound for sensitive molecular testing techniques like NGS. The DNA derived from these FFPE samples can be severely compromised through crosslinking and deamination, further limiting the amount of DNA that can be amplified in PCR-based assays. Since mass quantification of nucleic acid does not indicate nucleic acid integrity, we developed a pre-analytical QC test that measures the amount of amplifiable genomes in a sample. We also illustrated that the number of amplifiable genomes is indicative of the number of unique reads present in the final NGS library.

The sensitivity of somatic variant detection is directly related to the number of unique PCR-deduplicated fragments spanning the region of interest. False negatives and inconclusive results occur when too few unique molecules span a region of interest and there are insufficient reads to support a given mutation call. Amplicon-based enrichment techniques lack the ability to count how many unique fragments are sequenced, since libraries consist of identical amplicons that cannot be traced back to unique input molecules. As a result, the scarcity or diversity of input molecules is unknown in the final library.
Unlike amplicon-based enrichment technologies for NGS, Archer VariantPlex assays permit a quantitative assessment of the number of input molecules represented in the final sequencing libraries, allowing users to estimate the sensitivity of the assay at a given position of interest. Since the input complexity is retained in the final library with VariantPlex assays and represented in Archer Analysis, one can distinguish high-confidence negative results from false-negative or inconclusive results stemming from a lack of input.

The PreSeq DNA QC assay provides a method of quantifying the number of input molecules that are present in a given sample prior to library preparation. The use of this assay can help rescue or avoid libraries that would otherwise fail Archer Analysis QC. The studies on 202 libraries discussed in this literature demonstrate the importance and predictive power of the Archer PreSeq DNA QC assay. As well, library yield can be rescued with additional DNA input.

The quantitative nature of sequencing data generated with Archer VariantPlex assays emphasize the necessity of predictive, data-driven metrics on sample quality from a pre-analytical QC assay. The Archer PreSeq DNA QC assay enables the user to derive critical information on the sample input complexity prior to library preparation, ensuring that targeted libraries maximize input molecule representation for confident detection of SNVs, Indels, and CNVs from NGS. Implementation of this QC assay in a high-throughput screening laboratory allows the potential to save significant time and cost associated with failed library prep, sequencing, and data analysis. More importantly, it has the ability to reduce the number of samples that would otherwise result in false negative mutation calls due to insufficient coverage.

References

1. M. Srinivasan, D. Sedmak, S. Jewell, Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am. J. Pathol. 161, 1961–1971 (2002).
2. J. Ben-Ezra, D. A. Johnson, J. Rossi, N. Cook, A. Wu, Effect of fixation on the amplification of nucleic acids from paraffin-embedded material by the polymerase chain reaction. J. Histochem. Cytochem. 39, 351–354 (1991).
3. S. Q. Wong et al., Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. Sci Rep. 3, 3494 (2013).
4. Z. Zheng et al., Anchored multiplex PCR for targeted next-generation sequencing. Nature Medicine. 20, 1479–1484 (2014).1484 (2014).

For more information visit archerdx.com/preseq-dna

Limitations of Use:
For Research Use Only. Not for use in diagnostic procedures. This product was developed, manufactured, and sold for in vitro use only. This product is not suitable for administration to humans or animals. SDS sheets relevant to this product are available upon request. Qubit® is a registered trademark of Thermo Fisher Scientific, Inc. Biochain® is a registered trademark of BioChain Institute, Inc. Archer™, PreSeq™, and VariantPlex™ are trademarks of ArcherDX, Inc.