Application Note

Genomics

Agilent

Trusted Answers

Correlating NGS Success with Sample Input Quality: a Large Scale Study

Authors

Timothy Butler
Agilent Technologies, Inc.

Dr. Carsten Maus
Next Generation Sequencing Core Facility
German Cancer Research Center (DKFZ)

Abstract

Quality control (QC) steps in any workflow help distinguish between high- and low-quality samples. Knowing the quality of a sample can be extremely beneficial for most workflows, including next generation sequencing (NGS), which requires high-quality samples to perform optimally. The Agilent TapeStation systems are automated electrophoresis instruments that provide objective analysis of nucleic acid samples. The TapeStation system uses the Agilent Genomic DNA ScreenTape assay to assess the sample size and concentration of input genomic DNA (gDNA) and generates a DNA quality metric called the DNA Integrity Number (DIN). The DIN assigns a score from 1 to 10 to each sample analyzed, with 1 indicating low-quality samples, and 10 indicating samples of the highest quality. Using 4,000 samples and their sequenced data collected by the German Cancer Research Center (DKFZ) and their associated input gDNA DIN values as an example, this application note shows users how the DIN can be used to differentiate between samples of different quality. Further, by defining the quality of each input gDNA sample, users can define a threshold to determine if the sample is fit for NGS or not. By incorporating a DIN threshold into the sequencing preparation workflow, the process can be streamlined, saving time and costs.

Introduction

Next-generation sequencing (NGS) is a time-consuming and costly process. To help alleviate these challenges, quality control (QC) steps are used to distinguish between high- and low-quality samples before sequencing. This can save time and costs by identifying samples that are not fit for purpose or need to be optimized prior to sequencing. High-quality samples are necessary in NGS to achieve optimal results. To address this, the Agilent TapeStation systems provide a convenient QC solution that assesses the size, concentration, and integrity of nucleic acid samples at various stages throughout the library preparation process. Analysis of input genomic DNA (gDNA) is an important first step in the NGS workflow. The TapeStation system with the Agilent Genomic DNA ScreenTape assay provides the DNA Integrity Number (DIN) quality metric, which determines the quality of a sample by assessing the degree of fragmentation of the gDNA. The number assigned to each sample can range from 1, the lowest quality or highly degraded, to 10, indicating high quality or a very intact sample. Users can establish a DIN cutoff for their specific workflows to easily distinguish between samples of sufficient and insufficient quality for NGS, allowing for a more streamlined approach to sequencing preparation.

The Next Generation Sequencing Core Facility at the German Cancer Research Center (DKFZ) uses the gDNA ScreenTape assay, which provides the DIN for QC of input DNA in their NGS workflow. The DKFZ has sequenced over 4,000 samples and collected DIN values and sequencing data metrics for each sample. Both high-quality and low-quality samples were sequenced and used in this application note to show a comprehensive comparison between samples of varying quality. To accomplish this comparison, the DIN of each sample was correlated with the sequencing success metric (average coverage per 1 million read pairs) provided by the DKFZ. This data demonstrates how the DIN can be implemented into workflows, and how a DIN threshold can be used for quickly assessing the quality of gDNA samples.

Methods

The DKFZ analyzed genomic DNA extracted from human blood and tumor samples. Using an Agilent 4200 TapeStation system coupled with the Genomic DNA ScreenTape (p/n 5067-5365) and Agilent Genomic DNA Reagents (p/n 5067-5366), the quality of each sample was assessed using the DIN score prior to NGS library preparation. All DNA samples were sequenced using the Agilent SureSelect^{XT} Low Input Reagent kit with Human V5 and UTR kits (p/n 5190-6214). The DKFZ uses several data metrics to determine sequencing success. The more conclusive metric they use as an estimator of overall sequencing success is “average coverage per 1 million read pairs” which is discussed in this application note.

Results

The Researchers from the DKFZ conducted an analysis involving a dataset of 4,000 samples to investigate the applicability of the DIN assigned by the TapeStation software as a tool for determining DNA sample quality prior to NGS workflows. Figure 1 presents representative electropherograms from the sample dataset, illustrating two distinct scenarios. Figure 1A demonstrates a high-quality sample with a DIN of 7.7, while Figure 1B shows a lower quality sample with a DIN of 3.4. The higher-quality sample electropherogram exhibits a greater amount of fragments at a larger size, indicating a more substantial amount of intact DNA of similar size. In contrast, the lower-quality sample displays a more bell-shaped curve, indicating a lack of uniformity in DNA sizes, with a range of fragment lengths from short to long resulting in a smear-like appearance.

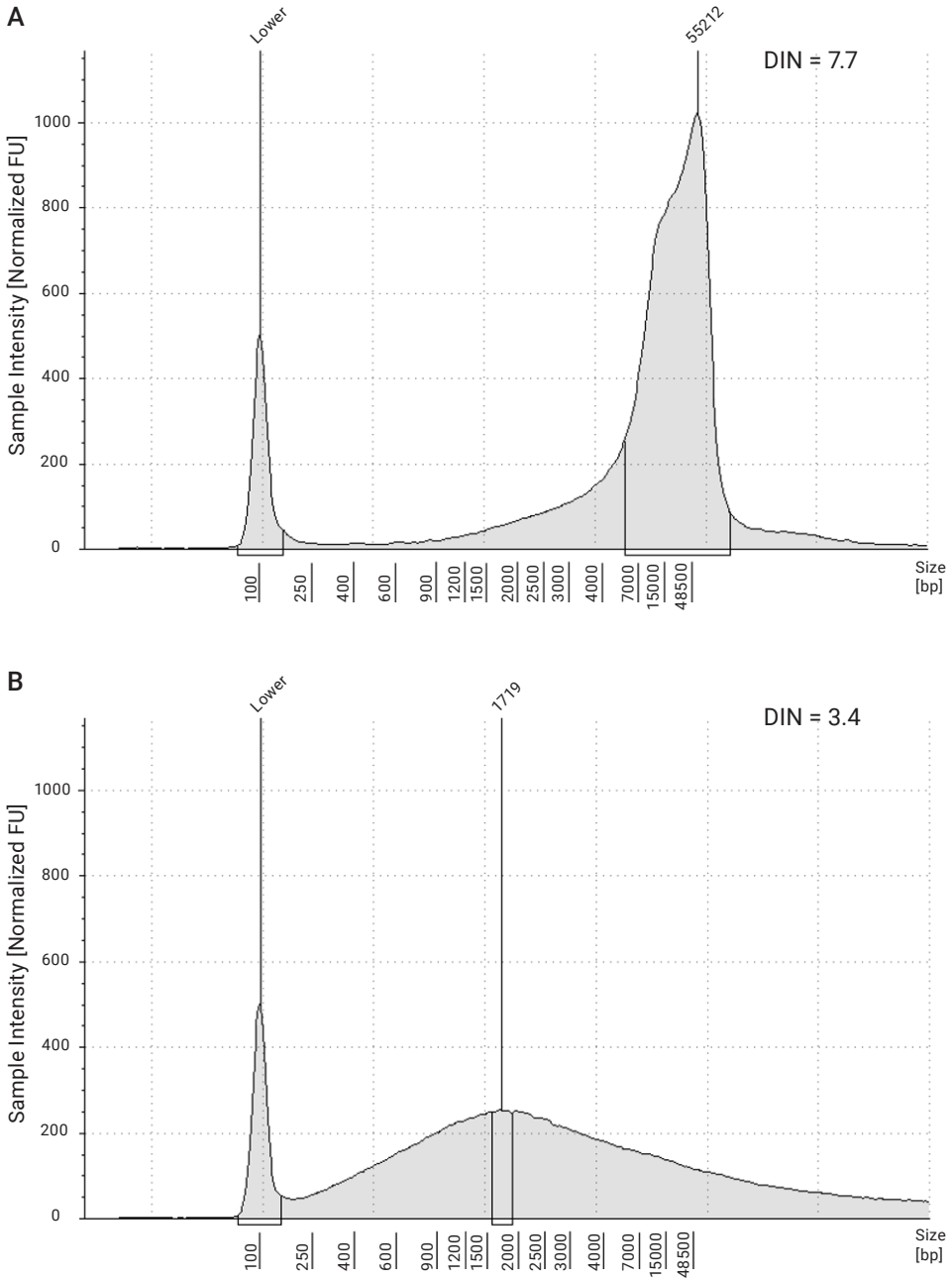

Figure 1. Two representative electropherograms from the Agilent 4200 TapeStation system using the Agilent Genomic DNA ScreenTape assay demonstrate the appearance of **A)** high-quality (DIN = 7.7) and **B)** low-quality (DIN = 3.4) samples.

The correlation between DNA sample quality, as represented by the DIN, and sequencing success is depicted in Figure 2. The DKFZ uses a sequencing quality threshold for assessing whether a sample was successfully sequenced, quantified as an "average coverage of 0.845 per 1 million read pairs". The threshold is shown as a solid line across the plot. Samples that have a sequencing metric above the line are considered passing, while samples below the line are considered failing. As shown in Figure 2, there is a direct correlation between the DIN score and the sequencing success of the sample. The majority of samples with a lower DIN score denote diminishing sample quality, which is concurrent with sequencing metrics falling below the established sequencing success threshold. Conversely, the majority of samples with a higher DIN score indicates better DNA integrity that successfully meets or exceeds the designated threshold for sequencing success.

Setting a threshold for DIN is a quick and efficient way to distinguish between samples that are and are not fit for NGS. For example, using the plot from Figure 2, if the DIN cutoff threshold was set at seven, the resulting data set shows that 99% of the samples above the threshold resulted in successful sequencing. However, as the DIN value decreased below seven, the percentage of samples resulting in successful sequencing decreases. As another example, of the samples that had a DIN of four or less, only 28% were passing sequencing. The data shown indicates that the DIN of input gDNA samples correlates with sequencing success. Setting a DIN threshold that samples must pass to move forward into the NGS workflow is a quick and easy way to help ensure successful sequencing results.

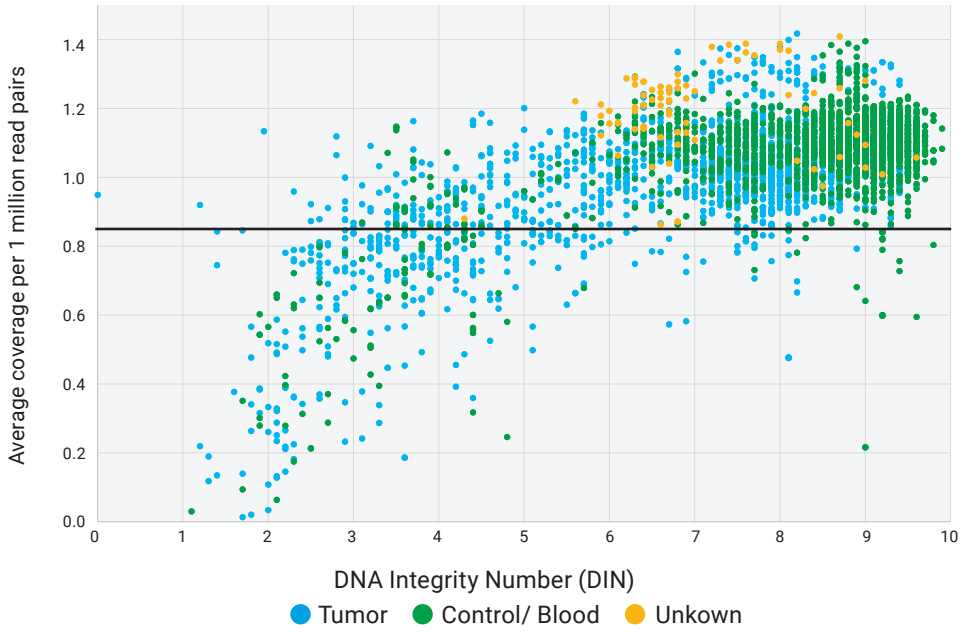

Figure 2. Sample metrics were plotted using the DNA Integrity Number (DIN) from the Agilent 4200 TapeStation system and the sequencing metric "Average coverage per 1 million read pairs". A horizontal line was added to the plot to show where the DKFZ's sequencing data threshold is set, to distinguish between passing and failing sequencing sample results.

DIN threshold for different sample types

High- or low-quality samples, in terms of a DIN value can depend on the type of sample being analyzed. Thus, a user may have to define different DIN thresholds for their specific sample types. For instance, the 4,000 samples from the dataset were parsed by sample type, and the gDNA extracted from blood or tumor were plotted individually to show their DIN distributions and sequencing metric (Figure 3A and B). Closer examination of the blood samples revealed a clear distinction between samples with lower DIN and samples with higher DIN and thus, successful and failed sequencing. This distinction is most prevalent at a DIN of approximately five, with most of the samples above this DIN passing sequencing, and most below failing sequencing. Thus, with this type of sample a clear DIN threshold can be set to distinguish between samples of sufficient or insufficient quality before further downstream steps. If a lower DIN is used, the proportion of sample results below the sequencing threshold rises. This increase in samples below the threshold results in more sequencing failures, as illustrated in Figure 3A.

In contrast, the tumor samples shown in Figure 3B exhibit no clear distinction where a DIN threshold may be placed, as seen by numerous samples both above and below the sequencing threshold. This makes it challenging to determine which DIN value is associated with a consistent onset of sequencing failures and underscores the need to recognize that not all samples can adhere to a singular DIN threshold. The example with tumor samples shows that users may need to evaluate the trade-offs between the number of sequencing successes versus sequencing failures to determine an appropriate placement of a DIN threshold.

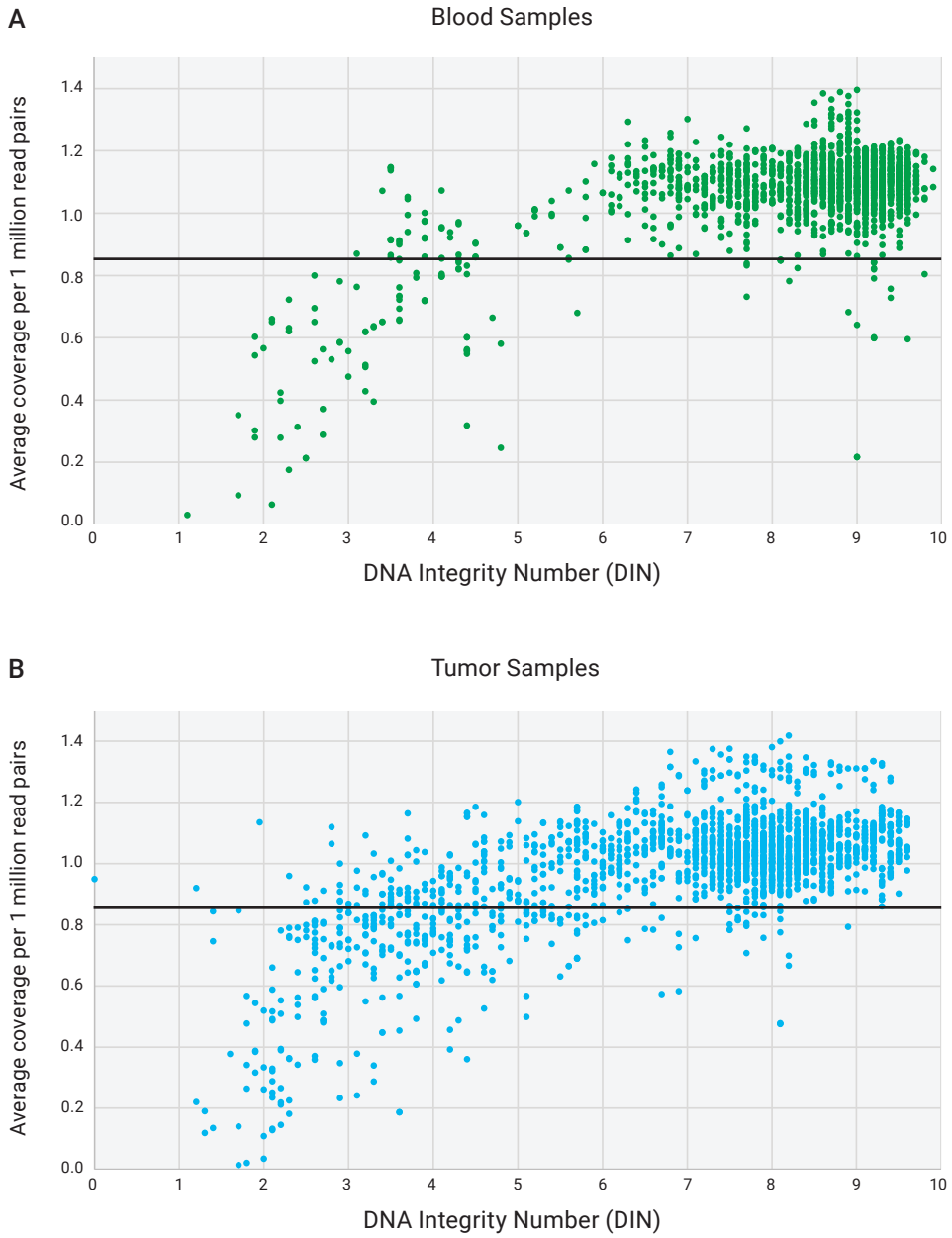

Figure 3. Sample metrics plotted using the DNA Integrity Number (DIN) from the Agilent 4200 TapeStation system and the sequencing metric "Average coverage per 1 million read pairs" separated by sample type, where plot **A)** shows blood samples, and **B)** shows tumor samples. Horizontal line added to plot to show where the DKFZ's sequencing data threshold is set to distinguish between passing and failing sequencing samples.

Conclusion

This application note examined how the DIN metric provided by the Agilent TapeStation systems can be used for quality control of input DNA before entering the NGS workflow. The data supplied by the DKFZ demonstrated that the higher the DIN, the more likely samples are to pass sequencing. Further, for different sample types, the location of a DIN threshold can vary, indicating that users may need to determine different DIN thresholds for each sample type processed. From the data presented in this application note, users can feel confident using the DIN from the TapeStation system in their workflows to achieve optimal sequencing results.

www.agilent.com/genomics/automated-electrophoresis

PR7001-2120

Published in the USA, March 1, 2024
5994-7119EN

Agilent
Trusted Answers