Simultaneous digital quantification and fluorescence-based size characterization of massively parallel sequencing libraries

Matthew T. Laurie¹, Jessica A. Bertout¹, Sean D. Taylor¹, Joshua N. Burton², Jay A. Shendure³, and Jason H. Bielas¹³⁴
¹Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, ²Department of Genome Sciences, University of Washington, Seattle, WA, ³Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA, and ⁴Department of Pathology, University of Washington, Seattle, WA

Keywords: droplet digital PCR; absolute quantification; next-generation sequencing; massively parallel sequencing; library preparation; quality control; size determination

Due to the high cost of failed runs and suboptimal data yields, quantification and determination of fragment size range are crucial steps in the library preparation process for massively parallel sequencing (or next-generation sequencing). Current library quality control methods commonly involve quantification using real-time quantitative PCR and size determination using gel or capillary electrophoresis. These methods are laborious and subject to a number of significant limitations that can make library calibration unreliable. Herein, we propose and test an alternative method for quality control of sequencing libraries using droplet digital PCR (ddPCR). By exploiting a correlation we have discovered between droplet fluorescence and amplicon size, we achieve the joint quantification and size determination of target DNA with a single ddPCR assay. We demonstrate the accuracy and precision of applying this method to the preparation of sequencing libraries.

Massively-parallel next-generation sequencing (NGS) technology is rapidly revolutionizing the fields of genomics, molecular diagnostics, and personalized medicine through the increasingly efficient and economical generation of unprecedented volumes of data (1–7). A common characteristic of the various commercially available NGS technologies is the need to load a precise number of viable DNA library molecules onto the instrument to optimize the yield of data from an individual sequencing experiment (8–11). Performing a sequencing run with either too many or too few library molecules results in compromised data yields or completely failed sequencing runs that waste sample, expensive reagents, user time, and instrument time. Similarly, if library molecules are not the appropriate length to fully utilize the capabilities of the sequencing platform, fewer bases can be sequenced in an NGS run and the throughput is wasted. Thus, accurate quantification and size determination of library DNA is essential for achieving optimal data yield and maximizing a laboratory’s efficiency and sequencing throughput.

Protocols for the preparation of NGS libraries include quality control steps to validate the size and concentration of amplifiable library molecules (i.e., molecules properly ligated to NGS adapter sequences) before committing to a sequencing run. Manufacturers typically recommend quantification with quantitative real-time PCR (qPCR) and size determination with gel or capillary electrophoresis. It is also possible to enumerate library DNA using UV spectrophotometry, the Quant-iT PicoGreen assay, or the Agilent BioAnalyzer, but these methods are not ideal because they quantify amplifiable and non-amplifiable molecules equally (12–14). These methods are also only capable of measuring mass per volume, which must be converted to copy number using an estimated average size of library molecules which can introduce further error (15). Although qPCR is widely considered the best option for library quantification, there are considerable drawbacks to the method, including amplification biases due to template size and GC-content as well as the need for a standard curve to estimate the absolute quantity of DNA (16). Creating a standard curve for each sample to be analyzed is a difficult and uncertain process that leads to inaccuracies in measurements of absolute target quantity (15,17).

When intercalating dyes are used for quantification, the concentration reading can include non-amplifiable DNA as dyes measure dsDNA indiscriminately. Because of these potential inaccuracies, some NGS platform manufacturers recommend performing titration runs on their instrument to determine the proper loading amount. The high cost of reagents and the length of NGS runs make this an expensive and time-consuming step.

We have developed a new assay capable of concurrently measuring the absolute concentration and length of unknown amplifiable DNA templates, making it well suited for quality control of NGS libraries. The assay, which we have termed QuantiSize, is based on the previously validated droplet digital PCR (ddPCR) absolute quantification system (18,19) and adds the ability to calculate the size of target DNA by exploiting a linear correlation we have discovered between the fluorescence amplitude of ddPCR droplets and the size of amplicons within them.

Method summary:
QuantiSize allows for the determination of absolute quantity and size distribution of target DNA molecules in a single experiment. This assay exploits a correlation, reported herein, between the length of an amplified DNA molecule and the fluorescence amplitude produced in droplet digital PCR (ddPCR), to allow the user to calculate the size of unknown DNA. As ddPCR simultaneously measures the concentration of target DNA, the user can accurately determine the target population size and quantity in a single step.
As a quantification method, ddPCR has demonstrated greater precision and sensitivity than real-time PCR (18). We demonstrate that QuantiSize accurately measures the size and concentration of target DNA simultaneously while avoiding the limitations of other quantification systems and, we highlight the utility of this assay for preparation of NGS libraries.

Materials and methods

Purification of DNA size standards
An exACTGene 50 bp DNA Ladder (Fisher, Waltham, MA) and 1 kb Plus DNA Ladder (Fisher) were run on a 1.0% UltraPure Low-Melting Point Agarose (Invitrogen, Carlsbad, CA) electrophoresis gel and the 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, and 1000 bp ladder bands were manually excised. The DNA in these gel slices was purified using the QIAcube automated gel extraction protocol with the QIAquick Gel Extraction Kit (QIAGEN, Hilden, Germany). The size and purity of all DNA fragments were verified by gel electrophoresis.

Ligation of DNA size standards to adapter sequences
The DNA fragments were ligated to the following sequences containing the Nextera version 1 adapters (Epicenter Biotechnologies, Madison, WI):

Adapter 1:
5′-CAAGCAGAAGACGGCATACGAGATTCGCTTACGGTCTGCCTTGCCAGCCCGCTCAGAGATGTGTATAAGAGACAG[INDEX 1]CCC-3′

Adapter 2:
5′-GGG[INDEX 2]CTGTCTCTTATACACATCTCTGATGGCGCGAGGGAGGCGCGATCTAGTGTAGATCTCGGTGGTGCCTATCATT-3′

All size standard fragments were ligated to adapters with the following 7 bp indices:

Index 1: 5′-TACCTCT-3′

Index 2: 5′-ACACATT-3′

Ligations were carried out in 20 µL reactions containing 1 µL T4 DNA Ligase and 2 µL 10× T4 DNA Ligase Buffer (New England BioLabs, Ipswich, MA). The ligation reactions were incubated at room temperature for 2 h. The ligated DNA was purified with a phenol/chloroform/isoamyl alcohol extraction. All samples were sent to the Fred Hutchinson Cancer Research Center ABI capillary sequencing facility to verify that the correct insert had been ligated to the adapters in each sample.

Library preparation for Illumina MiSeq
Samples of the plasmid pET-23a were sheared to an average size of 150 bp using the Covaris S220 Ultrasonicator (Covaris, Woburn, MA). The sheared DNA was run on a 1.0% UltraPure Low-Melting Point Agarose gel and a gel slice corresponding to ~100–200 bp was manually excised and purified using the QIAcube gel extraction protocol. The sheared DNA was blunted and phosphorylated using the Quick Blunting Kit (New England BioLabs) and purified with a phenol/chloroform/isoamyl alcohol extraction.

Eight samples of sheared DNA were ligated to adapter sequences with unique 7 bp indices in separate 20 µL reactions containing 1 µL T4 DNA Ligase and 2 µL 10× T4 DNA Ligase Buffer. The ligation reactions were incubated at room temperature for 2 h. All ligations were purified with a phenol/chloroform/isoamyl alcohol extraction.

All ligated libraries were amplified with 20 cycles of standard PCR using the following primer sequences:

Primer 1:
5′-AATGATACGACGGGACACCCAGA-3′

Primer 2:
5′-CAAGCAGAAGACGGCATACGA-3′

Figure 1. ddPCR amplification of 10 size standards designed for use with the QuantiSize assay. All size standards were amplified in parallel with standard reagent and thermal cycling conditions. (A) Scatter plot of fluorescence amplitude of individual droplets for each size standard. Droplets whose fluorescence amplitude is above a specified threshold (“positives”) are shown in black and droplets with fluorescence amplitude below the threshold (“negatives”) are shown in gray. (B) Box-and-whisker plots showing distribution of fluorescence amplitudes of positive droplets. Horizontal bars mark the mean fluorescence amplitude, boxes mark the interquartile range, and whiskers mark the 95% confidence interval. (C) Plot of mean fluorescence amplitude ± SEM vs. amplicon size showing a linear correlation (R² = 0.9943).
The amplified libraries were quantified using the ddPCR system (Bio-Rad, Hercules, CA) and the Quant-iT Picogreen assay (Invitrogen). In the ddPCR experiment, the libraries were run in parallel with adapter-ligated size standards to allow for the estimation of library size distribution. The measured concentrations of the 8 differently indexed libraries were used to dilute and combine the libraries in a molar ratio of 100:50:10:1 with 2 libraries at each concentration. The combination of libraries was denatured and diluted in preparation for loading onto the MiSeq flow cell (Illumina, San Diego, CA) as per the Illumina protocol.

Genomic DNA was purified from the human colon cancer cell line HCT 116 using the QIAcube automated purification protocol with the DNeasy Kit (QIAGEN). The Nextera XT DNA Sample Preparation Kit (Illumina) was used to generate a MiSeq compatible library from the HCT 116 DNA. The optional bead-based normalization step in the Nextera XT protocol was omitted and the library was instead normalized by quantifying the libraries was denatured and diluted in preparation for loading onto the MiSeq flow cell (Illumina, San Diego, CA) as per the Illumina protocol.

TaqMan probe and primer design

The following primers (Invitrogen) and TaqMan probe (Applied Biosystems, Foster City, CA) were designed to hybridize to the Nextera adapter sequences:

**Primer 1:**
5′-GCGACCACCGAGATCTACAC-3′

**Primer 2:**
5′-AGCAGAAGACCGCATACGAG-3′

**Probe:**
5′-FAM-CTGT+CT+CT+TA+TA+CA+CAC-1BFQ-3′

The ‘+’ indicates that the previous base is a locked nucleic acid (LNA) base.

Droplet digital PCR

All quantified DNA libraries and size standards were prepared for droplet PCR in 25 µL reactions containing 2× ddPCR Master Mix (Bio-Rad), 250 nM TaqMan probe, 900 nM each of the appropriate flanking primers, and ~10,000 copies of target DNA. Emulsified 1 nL reaction droplets were made by adding 20 µL of each reaction mixture to the sample wells of a droplet generator DG8 cartridge (Bio-Rad) and 70 µL ddPCR Droplet Generation Oil (Bio-Rad) to the oil wells of the cartridge for use in the QX100 Droplet Generator (Bio-Rad). Forty microliters of the generated droplet emulsions were transferred to Twin.tec semi-skirted 96-well PCR plates (Eppendorf, Hamburg, Germany), which were then heat sealed with pierced foil sheets. To amplify the target DNA, the droplet emulsions were thermally cycled using the following protocol: initial denaturation at 95°C for 10 min, followed by 40 cycles of 94°C for 30 s and 60°C for 1 min. The fluorescence of each thermally cycled droplet was measured using the QX100 Droplet Reader. All measurements were performed in triplicate.

Data analysis

The equation of the line fitting the correlation between amplicon size and fluorescence amplitude for the size standards was generated using Microsoft Excel (Redmond, WA) and applied to the measured fluorescence amplitude of each sequencing library to calculate amplicon size. The .fastq files produced by the MiSeq were imported to Sequencer (Gene Codes, Ann Arbor, MI) and aligned to the pET-23a plasmid sequence to generate a sequence alignment/map file (SAM). A perl script was used to count the length of each read pair by retrieving the number corresponding to the “TLEN” field of the SAM file. Only library molecules for which both paired-end reads passed the quality filter were included in the analysis.

**Results and discussion**

The ability of the Quantsize assay to combine quantification and size determination in a single ddPCR experiment is derived from a correlation that exists between the fluorescence amplitude of droplets and the size of amplicons within them. With standard ddPCR reagent concentrations, DNA amplification is eventually limited by the availability of dNTPs and inhibited by the presence of pyrophosphate (19,20); thus, long DNA templates, which consume more dNTPs and generate more pyrophosphate, will produce fewer products than short templates at the end point of a standard reaction. Because the final number of products generated within a droplet determines its level of fluorescence, the measured fluorescence amplitude of droplets containing short templates will be greater than that of droplets containing long templates. The Quantsize assay exploits this fact to generate an equation relating fluorescence amplitude to amplicon size by using measurements of known size standards. The equation describing the relationship between fluorescence amplitude and amplicon size can be used to calculate the size of any unknown ddPCR template that shares common primer and probe binding sites with the size standards. Creating size standards that have primer and probe binding sites in common with DNA samples can be accomplished in a number of ways including cloning sample DNA into a vector and appending adapter sequences to both the sample DNA and size standards (21).

We created a set of size standards applicable to Illumina NGS libraries containing inserts ranging from 25 to 1000 base pairs flanked by adapter sequences compatible with the Illumina MiSeq platform. A pair of primers and a fluorescent TaqMan probe were designed to hybridize to the adapter sequences such that the length of each amplicon is 160 bp plus the length of the insert. As primers and probes are specific to the MiSeq adapter sequences, only adapter-ligated molecules that will be amplified on the MiSeq flow cell will be quantified.

A ddPCR experiment was performed with the aforementioned size standards in separate wells of a 96-well plate. Droplets containing the target (positive) increased in fluorescence following amplification of the target whereas droplets lacking the target (negative) remained at the background.

![Figure 2. Effect of ddPCR elongation time on the relationship between fluorescence amplitude ± SEM and amplicon size.](image-url)
The size standards used for QuantiSize are highly analogous to the standards used in gel and capillary electrophoresis. The size of unknown DNA can be determined by visually comparing the fluorescence amplitude of the unknown DNA to the size standards. The droplet reader software counts positive and negative droplets by using a threshold of fluorescence between the well-defined populations of high and low fluorescence amplitude droplets. One particular TaqMan probe displayed a fluorescence amplitude for droplets containing amplicons larger than 660 bp that is too low to reliably discriminate between positive and negative droplets when templates are amplified with a 1 min elongation time. When this is the case, the average fluorescence amplitude for these amplicons cannot be calculated. Increasing the elongation time to two minutes increases the fluorescence amplitude for all droplets containing amplifyable template (Figure 2). This enables the acquisition of accurate concentration and fluorescence amplitude data for longer templates, but the slope of the relationship between amplicon size and fluorescence amplitude is decreased (from m = -11.66 to m = -9.12), which decreases the ability to resolve small differences in amplicon size (Figure 2). Decreasing the elongation time to 30 s increases the resolution of the relationship between amplicon size and fluorescence amplitude, but prevents targets longer than 460 bp from amplifying to the point that they fluoresce detectably above the background fluorescence (Figure 2). This is likely due to the fact that longer products require more time for complete polymerization of nascent strands to occur. Thus, there is a tradeoff between the resolution and range of QuantiSize, though the assay can be easily adjusted to fit particular experimental needs.

To validate the use of the QuantiSize assay for the sizing and quantification steps in library preparation for NGS, we compared the quantity and size distribution of library DNA predicted by the QuantiSize assay to the quantity and size distribution of reads generated by the Illumina MiSeq platform. Eight uniquely indexed test libraries were generated by ligating DNA sheared to an average size of 150 bp onto MiSeq compatible adapter sequences similar to those used to create the aforementioned size standards. The libraries were run in individual wells of a ddPCR experiment alongside the set of size standards. Using the concentrations measured by ddPCR, the 8 uniquely indexed libraries were diluted and combined in a molar ratio of 100:50:10:1 with 2 libraries at each concentration. The observed number of MiSeq reads containing each index was compared with the expected number of copies of each uniquely indexed library loaded onto the MiSeq. The observed ratio of the number of reads containing each index very closely matched the expected ratio of 100:50:10:1 that was measured by ddPCR and the correlation between the expected and actual number of library molecules gave an R² value of 0.9693 (Figure 3A).

The Nextera XT DNA Sample Preparation Kit was used to prepare a sequencing library from genomic DNA extracted from the human colon cancer cell line HCT 116. In lieu of the optional bead normalization step in the Nextera XT protocol, the concentration and size distribution of the library were measured with QuantiSize and the library concentration was adjusted to the proper concentration based on this measurement. The process of quantifying, normalizing, denaturing, and loading the library onto the MiSeq was repeated three times to demonstrate the precision of QuantiSize for predicting cluster density. The target cluster density for each MiSeq run was 1.0x10^6 clusters/mm² (1.028x10^6 clusters/mm² including the phi X control DNA). The mean cluster density ± SEM obtained was 1.039 ± 0.053x10^6 clusters/mm² (Figure 3B). The mean number of reads ± SEM obtained was 1.813 ± 0.070x10^6 (Figure 3C).

There are several potential sources of error in the MiSeq sample loading process including the pipetting of small volumes, variability in the efficiency of the denaturation reaction, variability of flow cell surface area, and user error. These factors may account for some of the observed variance in cluster density. However, even with the potential error caused by these factors, the target cluster density was achieved with high precision using the QuantiSize method.

The equation relating amplicon size and fluorescence amplitude can be applied either to the average (mean or median) fluorescence amplitude of a sample or to the fluorescence amplitude of individual droplets. Applying
the equation to individual droplets allows for a more detailed analysis of the distribution of product sizes present in a sample. We applied the equation generated by the adapter-ligated size standards to the fluorescence amplitude of individual droplets containing library DNA to calculate the expected amplicon size within each droplet and compared the distribution of sizes to the distribution of read sizes measured by the MiSeq (Figure 4). The frequency distribution shows a high degree of overlap and a common center point for the estimation made using the QuantiSize assay and the observations from the MiSeq. As depicted in Figure 2A, there is an inherent variance in droplet amplitude that occurs even within a completely homogeneous sample of amplicon lengths. This variance likely accounts for the wider distribution of product sizes estimated by ddPCR than were observed in the MiSeq data. Size determination with QuantiSize provides a detailed calculation of the distribution of fragment sizes present in a sample, whereas gel or capillary electrophoresis provide only a range of sizes.

The QuantiSize assay demonstrates accuracy, reliability, and flexibility through the strength of the correlation between fluorescence amplitude and amplicon size in ddPCR experiments, and the ease with which the assay can be adjusted to fit specific experimental needs. Applying the QuantiSize assay to NGS library preparation avoids the limitations of other independent quantification and size determination methods and has the potential to increase the average yield of usable data generated from sequencing runs, thereby increasing the efficiency and throughput. The ability to determine the absolute quantity and the detailed size distribution of target DNA with a single experiment will be useful for a broad range of applications that require the quantification and sizing of target DNA.

Acknowledgments

We would like to thank Nolan Ericson and Mariola Kulawiec for helpful discussion and critical reading of our manuscript. This work was supported by an Ellison Medical Foundation New Scholar award (AG-NS-0577-09), an Outstanding New Environmental Scientist Award (ONES) (R01) from the National Institute of Environmental Health Sciences (R01ES019319), and a grant (W81XWH-10-1-0563) from the Congestionally Directed Medical Research Programs/U.S. Department of Defense. This paper is subject to the NIH Public Access Policy.

Competing interests

The authors declare no competing interests.

References


To purchase reprints of this article, contact: biotechniques@fosterprinting.com