Heteroduplexes Affect Accurate Library Size Determination Without Impacting Targeted Sequencing Performance

E.H. Chilton, K.D. Butcher, R.I. Zeitoun

INTRODUCTION

Genomic DNA (gDNA) libraries are prepared for next-generation sequencing (NGS) by fragmenting the gDNA and ligating adapter duplexes to the ends of the fragments. To ensure efficient downstream target enrichment and sequencing, the libraries should contain a controlled and narrow distribution of gDNA fragment sizes (Head et al. 2014). For this reason, an important quality control step before NGS involves measuring the gDNA library size distribution with size-based separation methods, such as gel electrophoresis.

Interactions between gDNA fragments in a library can sometimes interfere with their separation and lead to inaccurate size determinations. For example, heteroduplexes, which are caused by hybridization between partially homologous molecules, can form if too many PCR cycles are performed on the library. In this scenario, as primers are depleted, the adapter ends of denatured fragments anneal instead to the ends of non-complementary amplicons. This creates a bubble, or space between the two strands, as the non-complementary sequences do not anneal (**Figure 1**). The space affects the mobility of the heteroduplex molecule through a gel matrix to generate an artifactual increase in size (Zischewski et al. 2017). Thus in electropherograms, heteroduplexes often cause the appearance of artificially broad peaks and overestimation of fragment size.

We demonstrate here that the presence of heteroduplexes can alter the apparent size distribution of a gDNA library in an assay-specific manner. We also show that, despite their effects on an electropherogram, heteroduplexes do not actually alter the sequencing insert size or other NGS quality metrics.

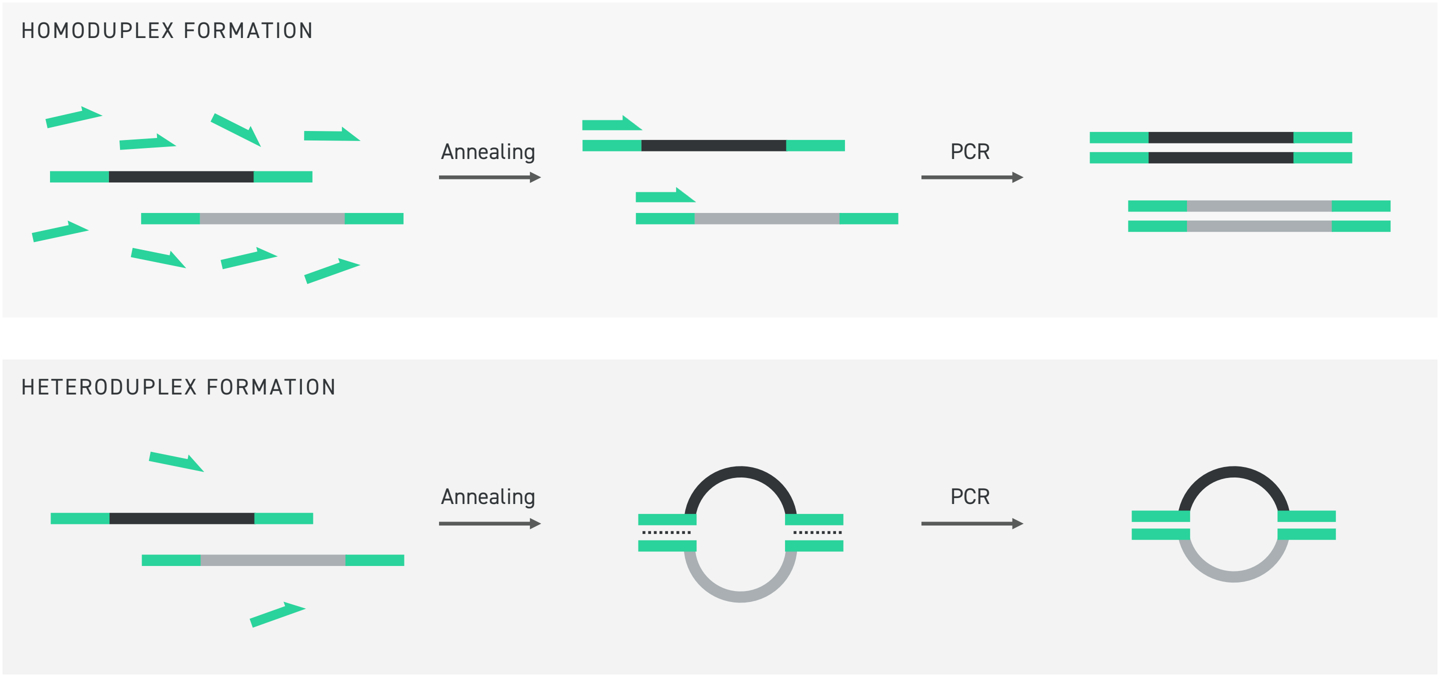

Figure 1. Heteroduplexes form when PCR amplification is performed in excess. For PCR amplification of a gDNA library, primers are designed to anneal to the adapter sequences on fragments. When primers are abundant in the reaction mix, amplification yields homoduplexes of amplified fragments (top panel). If, however, primers are depleted (for example, if PCR is performed too many times), complementary adapter sequences may anneal, forming heteroduplexes of non-complementary fragment sequences (bottom panel).

METHODS

A gDNA library was prepared from 50 ng of NA12878 template gDNA (Coriell Institute) using Twist Bioscience Library Preparation EF Kits 1 and 2 (PN 100253, PN 100401) and full-length combinatorial dual index TruSeq Y-adapters (Illumina). For the reconditioning reaction aimed at eliminating heteroduplexes, a sample of the library (20 ng) was subjected to a single cycle of PCR in the presence of a high concentration of primers (1 µM each). To generate a sample enriched for heteroduplexes, another aliquot of the gDNA library was denatured at 95°C in the absence of primers for 5 minutes, then reannealed at 65°C for another 5 minutes. The same libraries were also subjected to target enrichment using the Twist Human Core Exome Target Enrichment Kit (PN 100252), following the 16-hour custom panel hybridization protocol using a custom 806 kb target size panel.

All samples were separated and analyzed using a Bioanalyzer 2100 system (Agilent) and either the Agilent High Sensitivity DNA Assay or Agilent 7500 DNA Assay, according to manufacturer's instructions (Agilent 2018). Sequencing for both whole genomes and target-enriched libraries was performed using a 2 x 76 cycle NextSeq 550 High-Output Kit and platform (Illumina). Sequenced libraries were first downsampled to 150x raw sequencing coverage and then analyzed by aligning reads to a hg38 reference genome with BWA-MEM (Li 2013). Insert size was calculated using Picard metrics with a mapping quality score threshold of 20.

RESULTS

We prepared a human gDNA library and analyzed its fragment size distribution using a Bioanalyzer System and two different size-based separation methods. The two assays produced different results (**Figure 2**): whereas the Agilent 7500 DNA Assay detected only a single peak (**Figure 2A**), the Agilent High Sensitivity Assay detected a shoulder, or secondary peak, to the right of the standard peak, at around 400 bp (**Figure 2B**).

To determine whether the secondary peak was the result of differences in the sensitivity of the two assays to heteroduplexes, we prepared and analyzed samples that were either devoid of or enriched for heteroduplexes. To prepare a sample lacking heteroduplexes, we performed a single cycle of reconditioning PCR on the same library tested previously. This reaction, performed in the presence of a high concentration of primers, causes each template molecule to replicate its reverse complement to produce a mostly homoduplexed system (Thompson et al. 2002). **Figure 2C** shows that the Agilent High Sensitivity analysis of the reconditioned sample no longer yielded the secondary peak; only a single peak of 350–400 bp was observed. This demonstrated that the analysis of the original sample with the Agilent High Sensitivity Assay produced an artifactual peak that could be removed by removing heteroduplexes from the sample.

We also analyzed a predominantly heteroduplexed sample, which we prepared by denaturing and reannealing the gDNA library. Under these conditions, DNA strands anneal at primer regions and are more likely to hybridize to more abundant, non-complementary strands than to less-abundant complementary strands. Agilent High Sensitivity analysis of samples prepared in this manner yielded only a single peak at ~1,200 bp (**Figure 2D**).

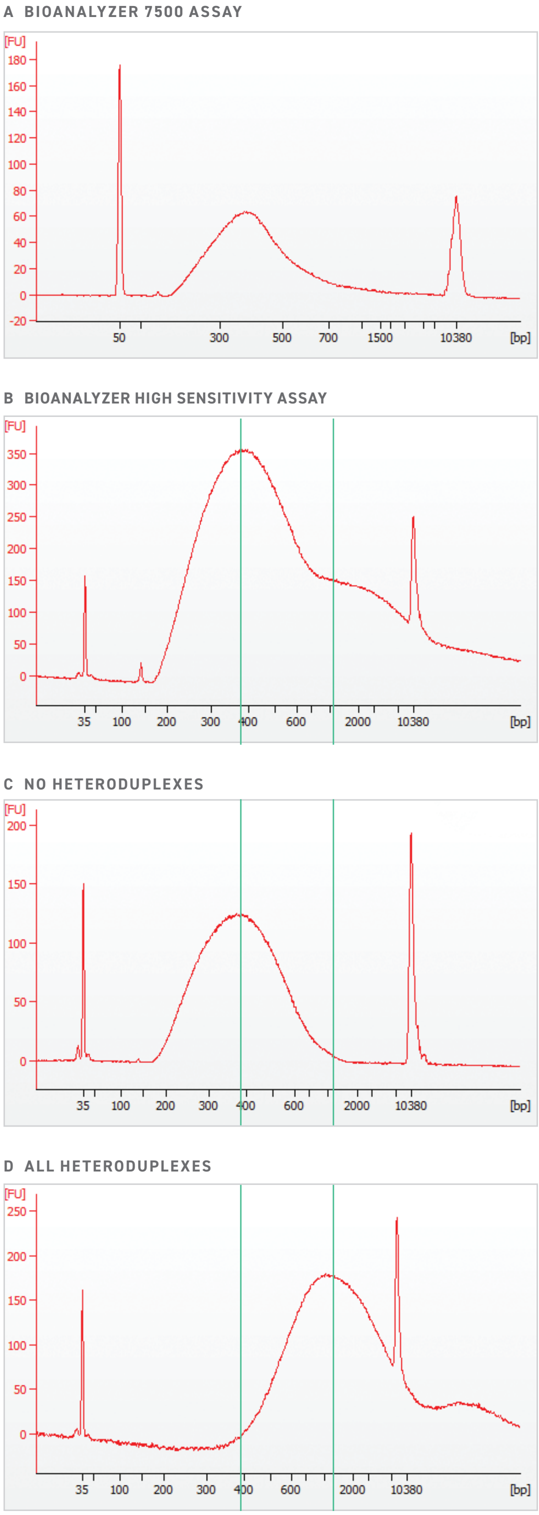

Figure 2. Heteroduplexes increase the apparent fragment size of libraries in Agilent High Sensitivity electropherograms. A Mixed heteroduplex state analyzed using the Agilent 7500 DNA Assay. Note the single primary peak at around 400 bp. **B** The same mixed heteroduplex state analyzed using the Agilent High Sensitivity Assay. Note the presence of two peaks: a primary peak at 400 bp and a secondary (shoulder) peak at 1,200 bp. **C** Reconditioned library with no heteroduplexes, analyzed using Agilent High Sensitivity Assay. Note the presence of a single peak at 400 bp and the lack of the secondary peak. **D** Denatured and reannealed gDNA library (heteroduplexes) analyzed using the Agilent High Sensitivity Assay.

If heteroduplexes form during library preparation as a result of PCR amplification, then increasing the number of amplification cycles to deplete primers should increase heteroduplex formation and result in a more pronounced additional peak (Michu et al. 2010). To test this, libraries were prepared using a range of 4–12 cycles of amplification and analyzed with the Agilent High Sensitivity Assay. The resulting electropherograms contained a shoulder to the right of the primary peak that grew in size as the number of amplification cycles was increased (**Figure 3**). The electropherograms of the same samples analyzed with the Agilent 7500 DNA Assay did not exhibit shoulders or double peaks, but the peak shape appeared more skewed as the number of cycles was increased.

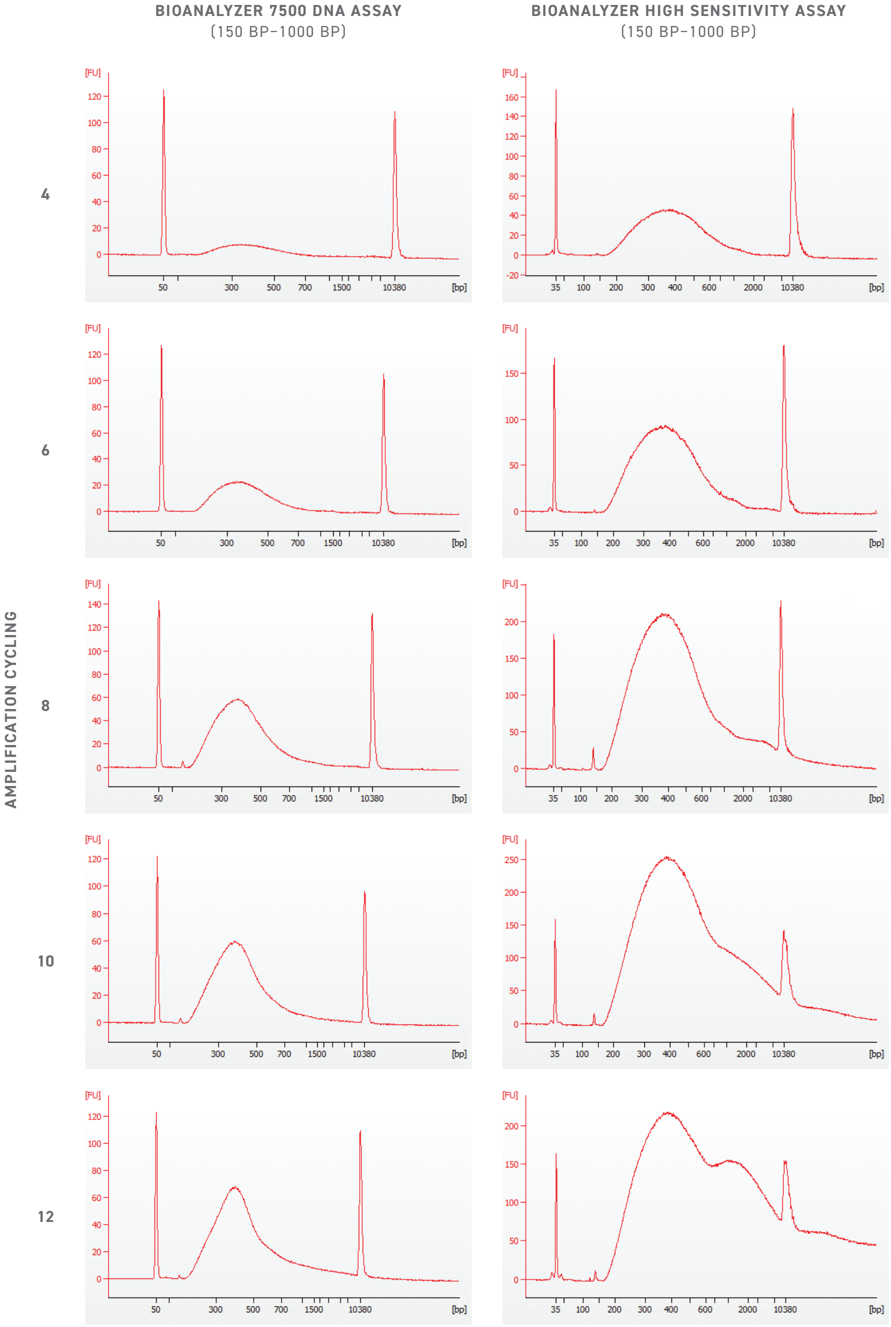

Figure 3. Heteroduplex state increases with increased numbers of amplification cycles. Samples of the gDNA library were subjected to the indicated number of rounds of amplification and analyzed with either the Agilent 7500 DNA Assay (left panels) or Agilent High Sensitivity Assay (right panels). Note that, as the number of rounds of amplification increased, so did the size of the shoulder peak at 1,500 bp in the Agilent High Sensitivity Assay.

Importantly, heteroduplex formation did not affect the actual insert size of the gDNA libraries. Whole genome sequencing of libraries subjected to various amounts of amplification revealed insert size histograms displaying only single peaks, even for samples that had undergone as many as 12 cycles of amplification (**Figure 4**).

It is important to note and explain the discrepancy between the size measured by the BioAnalyzer system and the insert size profile predicted through sequencing. First, the insert size profile does not account for adapter sequence length, which would add 135 bp to each insert (Illumina 2018). Second, NGS is biased toward representing short sequences, and this bias influences the sequencing size profile. Regardless, the presence of a single peak suggests heteroduplexes do not affect the actual length of gDNA fragments.

To demonstrate this conclusion also applies to target-enriched samples, we performed a target enrichment capture on the same libraries that had undergone whole-genome sequencing and then sequenced them again. The resulting insert size histogram also had only one peak, indicating the length of the library was constant, regardless of the amount of amplification performed (heteroduplex content, **Figure 5**). Again, we observed a difference in the predicted size relative to the BioAnalyzer measurement, which was introduced not only from adapter trimming and sequencing bias towards shorter samples but also from the preference of target enrichment for the capture of larger molecules.

FIGURE 4 INSERT SIZE HISTOGRAM, GDNA, NORMALIZED TO MAX

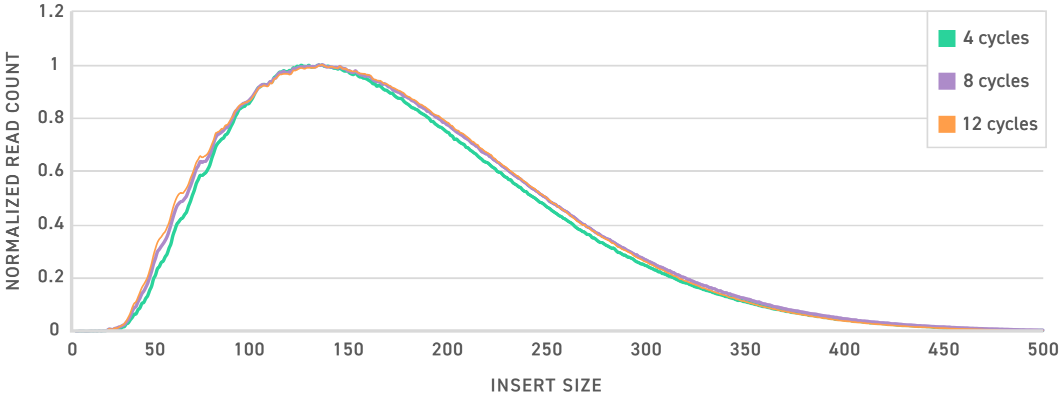

Figure 4. Heteroduplex state does not affect insert size, as determined by whole genome sequencing. Samples of the gDNA library that had undergone 4–12 rounds of PCR amplification were subjected to whole genome sequencing. Note that size determinations are not affected by the number of PCR cycles performed. The average size of the libraries differed from those observed in the BioAnalyzer electropherograms because the sequencing pipeline removes the adapter sequences from consideration when calculating insert size.

FIGURE 5 INSERT SIZE HISTOGRAM, TE, NORMALIZED TO MAX

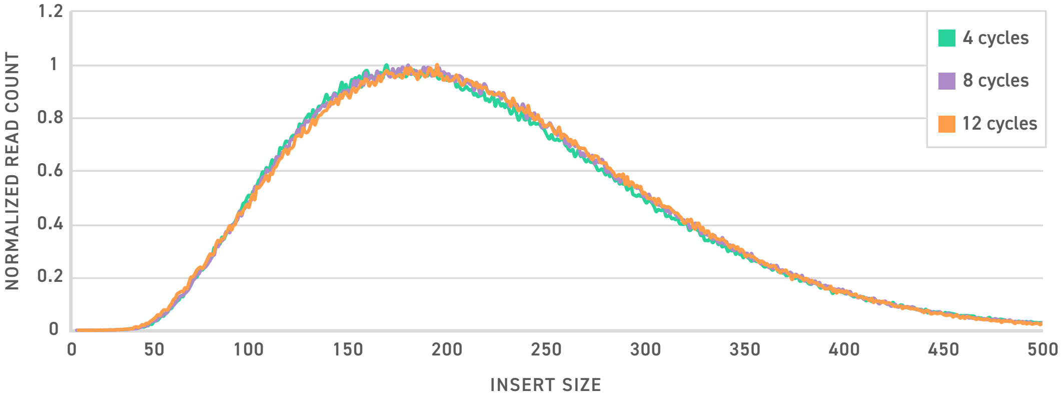

Figure 5. Heteroduplex state does not affect insert size, as determined by target enrichment. Samples of the gDNA library that had undergone 4–12 rounds of PCR amplification were carried through target enrichment. Note that size determinations are not affected by the number of PCR cycles performed. The average size of the libraries differed from those observed in the BioAnalyzer electropherograms because the sequencing pipeline removes the adapter sequences from consideration when calculating insert size. It is also different from those observed in whole genome sequencing (Figure 4) because of the bias toward larger molecules introduced by target enrichment.

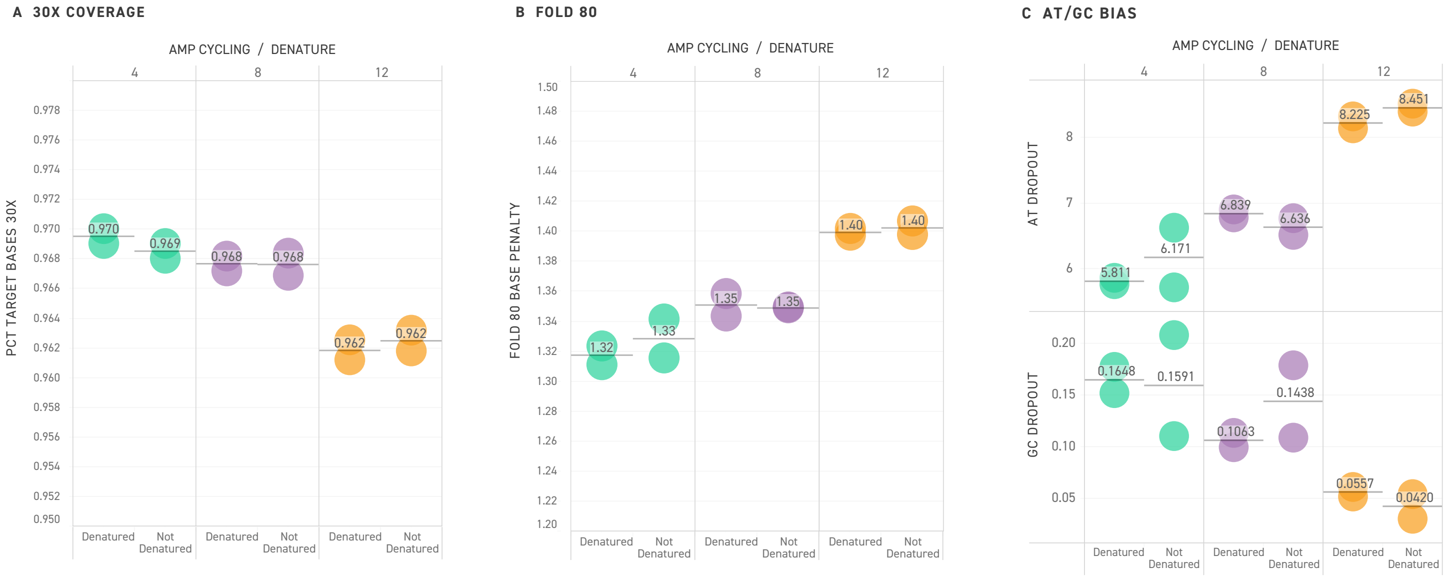

Figure 6. Heteroduplexes do not alter NGS metrics. Target enrichment was performed on gDNA libraries that had and had not been subjected to denaturation and reannealing beforehand (labeled heteroduplex and control, respectively). Both sets of samples yielded similar 30x coverage **A** and fold-80 base penalty scores **B**. AT/GC bias increased with the number of PCR cycles **C**, explaining the increased fold-80 and decreased 30x coverage seen in A and B. Panel is 806 kb downsampled to 150x raw sequencing coverage.

The presence of heteroduplexes also did not affect sequencing efficiency. We performed target enrichment with a set of identical samples, both before and after denaturation and reannealing (enrichment for heteroduplexes). The 30x coverage for these samples did not change (Figure 6A), even though increasing the number of amplification cycles can increase GC/AT biases (Figure 6C) to slightly reduce overall performance (Polz and Cavanaugh 1998). Similarly, the fold-80 base penalty was unchanged between the heteroduplexed and homoduplexed sets (Figure 6B), further demonstrating that the heteroduplex state does not affect key sequencing metrics.

CONCLUSION

In gDNA libraries, the presence of heteroduplexes can affect the results of size-based separation measurements. Depending on the sized-based assay used, the library size distribution may appear broad and multimodal if heteroduplexes are present. Whole genome sequencing and target enrichment show, however, that this apparent increase in size does not reflect the actual size of the library, nor does it affect performance in target enrichment. Taken together, these results illustrate that the separation method and specific assay should be considered when quantifying libraries and performing quality control analysis prior to NGS.

REFERENCES

Agilent (2018). 2100 Expert Software User's Guide.
Document # G2946-90005 Rev. B.

Head S, Komori H, LaMere S, et al. (2014) Library construction for next generation sequencing: overviews and challenges. Biotechniques 56 (2): 61–passim.

Illumina (2018) Illumina adapter sequences.
Document #1000000002694 v07 1: 24.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2

Michu E, Mráčková M, Vyskot B, et al. (2010) Reduction of heteroduplex formation in PCR amplification. Biologia Plantarum 54 (1): 173–176.

Polz MF, Cavanaugh CM. (1998) Bias in template-to-product ratios in multitemplate PCR. App. Env. Microbio. 64(10): 3724–3730.

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Research 30 (9): 2083–2088.

Zischewski J, Fischer R, Bortesi L (2017) Detection of on-target and off-target mutations generated by CRISPR/Cas9 and other sequence-specific nucleases. Biotechnology Advances 35 (1): 95–104.

These Products are subject to certain use restrictions as set forth in Twist's Supply Terms and Conditions www.twistbioscience.com/supply-terms-and-conditions