PARAGON GENOMICS

WHAT CONSTITUTES A GOOD TARGETED NEXT-GENERATION SEQUENCING LIBRARY?

Next-generation sequencing (NGS) continues to revolutionize personalized medicine as NGS technology improves and costs continue to fall, resulting in an increased number of scientific and health-related applications. However, the success of a sequencing project rests on the sample quality and NGS library used to generate the data. Poor quality libraries will produce poor quality or biased sequencing data that can slow the advancement and clinical utility of this great new technology. This review will outline important challenges that should be considered when preparing NGS sequencing libraries.

What is an NGS library?

NGS applications can be segregated into several categories including whole genome, epigenome, whole transcriptome and targeted sequcencing. In order to study these categories, researchers will need to create a library generating a collection of DNA fragments. The DNA fragments may represent the entire genome or a subset of specified genes, often referred to as a targeted library. Usually adapters which contain sequencing primer sites, amplification primer sites and various other features are incorporated into the DNA fragments. With some of the new technologies such as multiplex PCR, it is now possible to multiplex and process up to 20,000 amplicons in a single tube. Generating an amplicon library via multiplex PCR is advantageous, as it is easier to target a small portion of the genome, making experiments more cost conscious, which leads to the generation of faster results and easier data analysis.

Which technology should be used to make an NGS library?

There are several methods which can be used to make an NGS library, but the most important factors to consider when selecting a technology are that the resulting library is:

1) Functionally excellent
2) Easy and fast to prepare
3) Inexpensive

Criteria #2 and #3 are straightforward to fulfill, especially when considering a library preparation such as multiplex PCR without a ligation step. Such a method may be the simplest and fastest way to prepare an NGS library, while also allowing researchers to avoid end repair steps that decrease library yield. Therefore, the key to making an excellent library is to find a method which will result in a functionally excellent library.

What problems can occur with libraries?

There are various libraries by applications, each with their own issues, which are depicted in Table 1.
This whitepaper will focus on a very hot and technically difficult frontier, amplicon libraries, which are made with multiplex PCR.

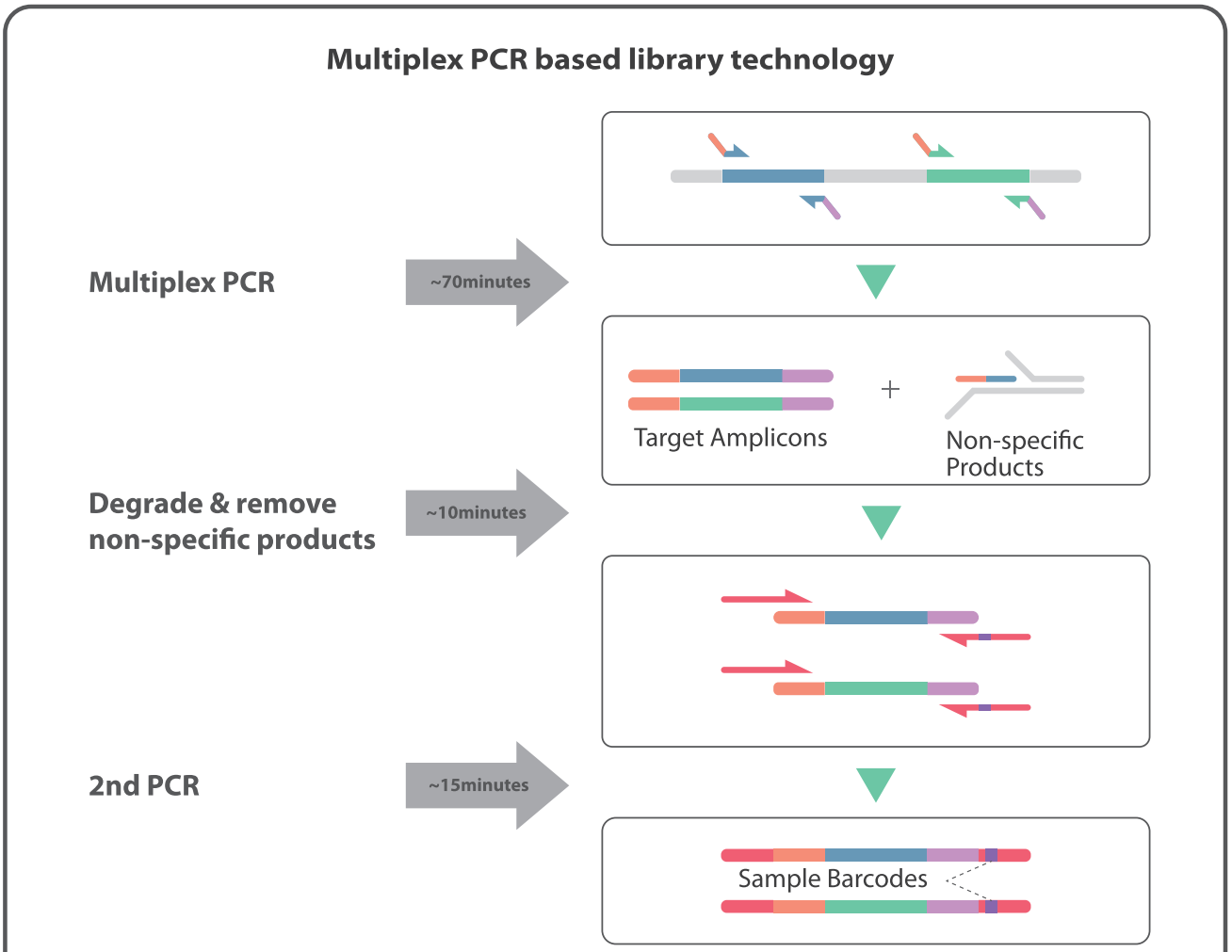

Figure 1. Paragon Genomics workflow provides ligation-free, multiplex PCR based library technology.

	Library Type		
	Fragment libraries	**Targeted libraries by capture method**	**Targeted libraries based on PCR**
Technology-specific bias		Bias caused by capture efficiency (target sequence, size, bait structure, length, etc.) Chemical modification of nucleotides	
Fragmentation	Bias caused by fragmentation technology (sonication or enzymatic) and/or chromatin structure		
PCR amplification	AT/GC bias, amplification bias of certain targets (i.e. targets containing GC-rich or AT-rich regions, repetitive sequences, secondary structures, etc.)		
Ligation efficiency	Bias caused by adapter ligation efficiency		

Table 1: Potential problems with various types of libraries.
Adapted from Van Dijk EL et al (2014) Exp Cell Res. 2014 Mar 10;322(1):12-20.

What defines an excellent library?

In order to define the "excellence" of a library, it is important to understand the intended usage. Libraries are used to find nucleotide sequences, SNPs and/or sequence features such as indels and re-arrangements at a certain sensitivity or limit of detection (LOD), in addition to scarce or "difficult" sources. Any library that has great sensitivity, plus contains amplified and enriched sequences from all target sources including difficult sources, should be regarded as excellent. However, it is practically impossible to achieve these criteria as many of the obstacles are often intertwined. Some of the key issues faced are uniformity, specificity and sensitivity:

1. Uniformity – Every target of the library should be present in approximately equal numbers, despite differences in GC content, genomic positions, or if they contain repetitive sequences, etc. However, this is a daunting task, particularly for libraries made with multiplex PCR and hybridization capture technology, because it is almost impossible to make every target equal in number. An ideal library has 100% of the amplicons with numbers of reads larger than 0.2 times the average number of amplicons, which is > 0.2X mean amplicon reads as depicted in Figure 2.

A multiplex PCR library technology should support a variety of applications, include a wide spectrum of the number of targets to be amplified, and allow low to large amounts of input DNA. These requirements mean that the library technology should produce acceptable uniformities, measured by 0.2X mean reads, when handling a wide range in numbers of amplicons and amounts of input DNA (Figure 3).

When dealing with DNA targets which have a very high GC content the amplicons are usually difficult to amplify by PCR. This is generally a severe issue with some multiplex PCR-based library methods. Excellent library technology should not generate a GC bias (Figure4).

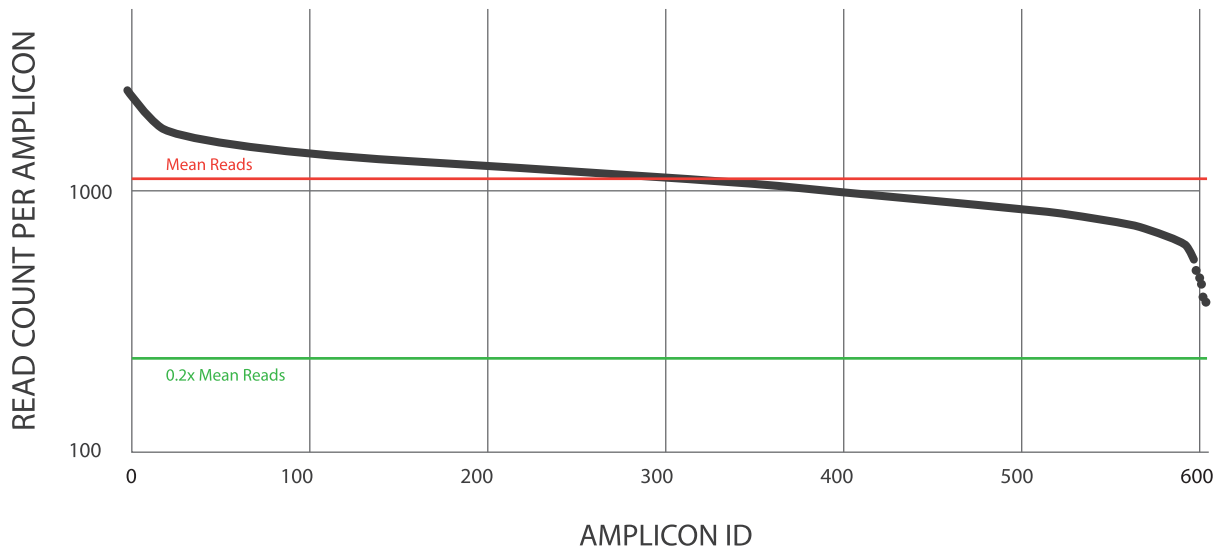

Figure 2. Measurement of uniformity by plotting the read count per amplicon. 100% of the library targets, made with Paragon Genomics' CleanPlex™ OncoZoom™ and 10ng of genomic DNA, has reads larger than 0.2 X mean reads. When sequenced at a mean read depth of about 1000, the amplicon with the lowest read depth has about 400 reads which is sufficient to support most applications.

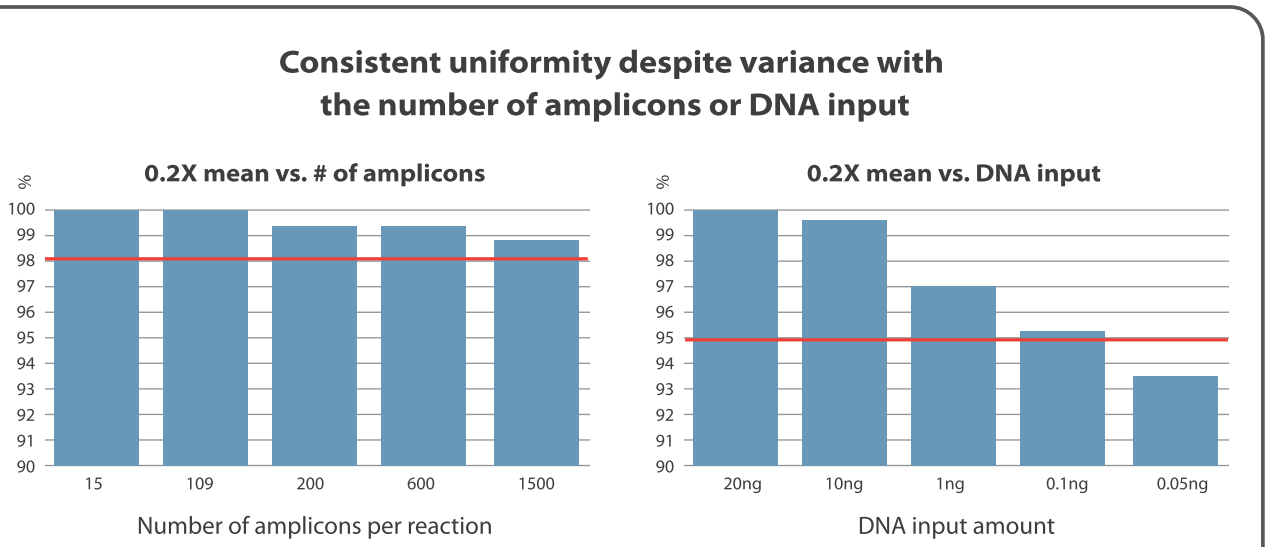

Figure 3. Paragon Genomics' CleanPlex™ technology produces > 98% uniformity at 0.2X mean reads with up to 1500 amplicons in a single reaction. CleanPlex™ technology also produces > 95% uniformity at 0.2X mean reads down to 0.1ng of input genomic DNA with 600 amplicons in a single reaction.

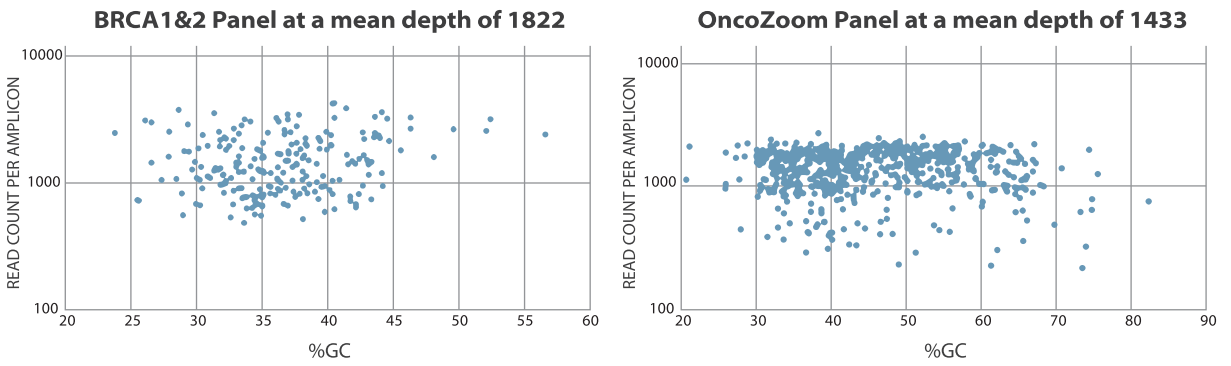

Figure 4. There is no evidence of GC bias with Paragon Genomics' CleanPlex™ BRCA1&2 panel which has 109 amplicons per multiplex PCR reaction (Left plot) and CleanPlex™ OncoZoom™ panel which has 601 amplicons per reaction (Right plot).

2. Specificity – Specificity covers two aspects. First, non-specific amplification products which are sequences unrelated to the targets. Having very few or no sequences which are unrelated to the targets are key goals for excellent specificity. Secondly, random errors may be introduced into the targets during enrichment. When making a library there should be very few or no errors introduced to the targets.

Library quality is usually checked by running a gel which immediately confirms the presence of non-specific products, or backgrounds, and primer-dimers from PCR, in addition to the library shape and quantity. This provides a first glance at the specificity of the library chemistry as seen in Figure 5.

Low background noise ensures non-specific by-products are not sequenced

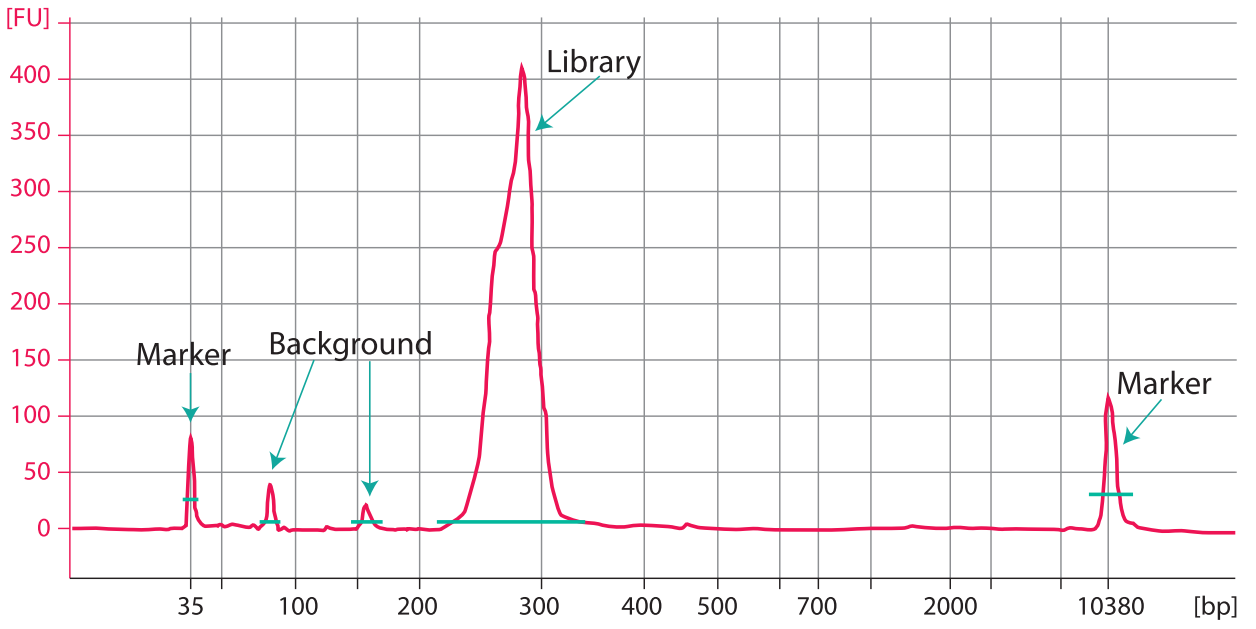

Figure 5. Paragon Genomics' library chemistry generates very low amount of background. The trace of CleanPlex™ OncoZoom library (601 amplicons in a single reaction) made with 10ng of genomics DNA, assayed with Agilent BioAnalyzer 2100 High sensitivity DNA kit.

Instead of eyeing the gel trace, a definitive measure of specificity is made after completing sequencing analysis. Measures of specificity include the mapping rate and on-target rate. Mapping rate measures the percentage of amplicons that match onto the genomic reference. From the sequences that are matched onto the reference genome, it is then possible to further calculate how many are matched onto the intended target regions thereby providing the on-target rate.

Paragon Genomics uses CleanPlex™, a proprietary technology, which removes background products generated during multiplex PCR. Paragon Genomics' unique algorithm achieves high coverage to design primer panels. In combination, these technologies achieve ultra-high levels of mapping and on-target rates which are depicted in Figure 6.

A key goal ,when making libraries is to introduce as few errors as possible. This translates into minimal nucleotide sequence errors when the sequence data is read. Errors seen in sequencing data may include random errors which occurred during both library making and sequencing. The latter is reported around 0.3-0.5% on Illumina sequencers. A library that contains few errors should be regarded as excellent, particularly when the random nucleotide error is on a par with the error level produced by the sequencing instruments, as seen in the example of libraries generated using Paragon Genomics' technology shown in Figure7. Molecular barcoding, or molecular tags, while not directly reducing random errors introduced during library preparation and sequencing, can be utilized to identify them.

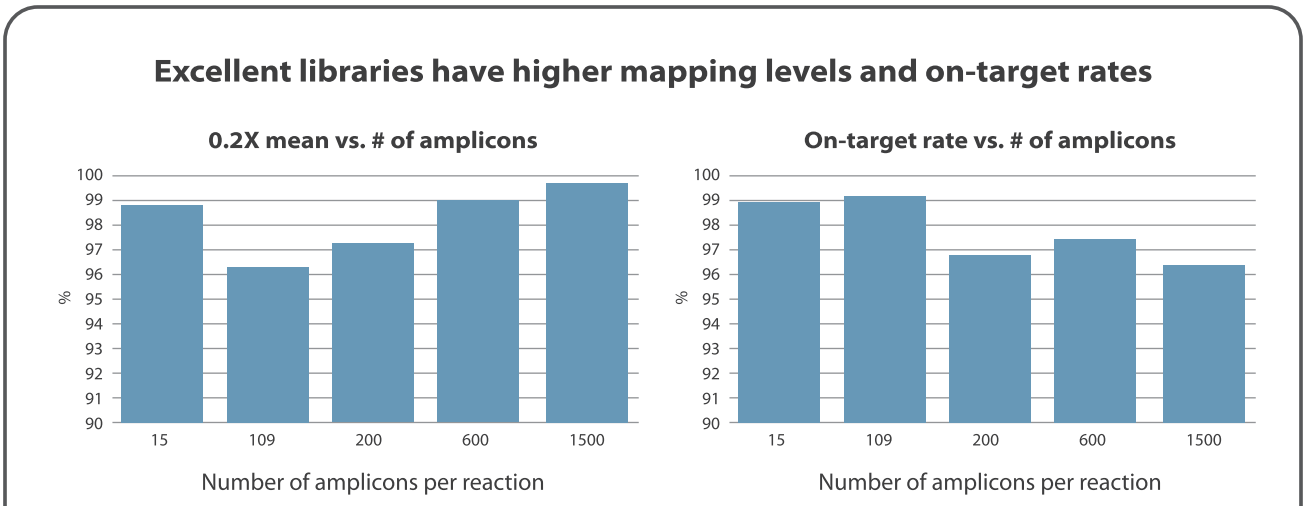

Figure 6. Mapping and on-target rates of Paragon Genomics' libraries. Due to the unique features of CleanPlex™ technology, libraries with larger amplicon numbers per reaction tend to be cleaner, and have a higher mapping rate.

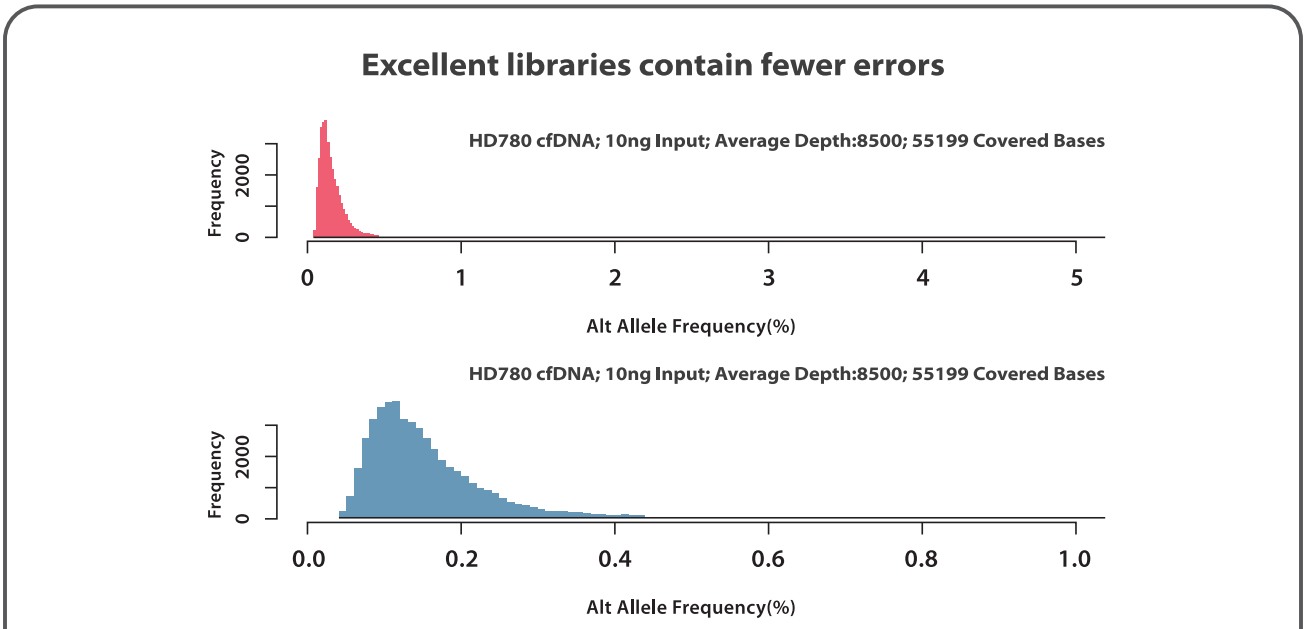

Figure 7. Random errors from sequencing a Paragon Genomics OncoZoom™ library made with Horizon Discovery's standard cfDNA (HD780). Only random error distribution is shown.

3. Sensitivity – Sensitivity is the ability to confidently detect low frequency alleles or variants. It is dependent on the ability of the NGS library preparation technology to produce libraries with low background and few random errors. By having low background, more accurate calls can be made. But when there is poor sensitivity, due to high background, true calls are buried among background noise. With good sensitivity they are well above the noise. However, good sensitivity does not always imply accuracy which is why it is important to have as few errors as possible in the library to ensure fewer false positives occur. Through good library preparations it is possible to achieve sensitivity of 1%.

Benefits of a good NGS library preparation:

By addressing the needs of uniformity, specificity and sensitivity, allele variances can be detected at their true frequencies. A convenient way of verifying the sensitivity and accuracy of an NGS library preparation method is to make and sequence a library with a commercial standard DNA of known variant frequencies as seen in Figure 8.

When the NGS library has good uniformity, specificity and sensitivity, in addition to a high mapping and on-target rate, excellence is easily achieved, and the library can be classed as excellent. The success of any NGS sequencing project is in the end, measured by the reliability of the data obtained, as well as the accuracy of the conclusions arising from that data. However, the resources and time spent on analysis to generate data and conclusions should be taken into consideration when choosing a library kit. Clearly every scientist desires reliable, accurate data but there is also great pressure to reduce workloads, cut costs, speed up processes and simplify NGS analysis. Good NGS library preparation will accomplish this and ultimately provide higher quality, more efficient data which will lead to more accurate and cost-efficient results, perfect for today's competitive research and clinical environment.

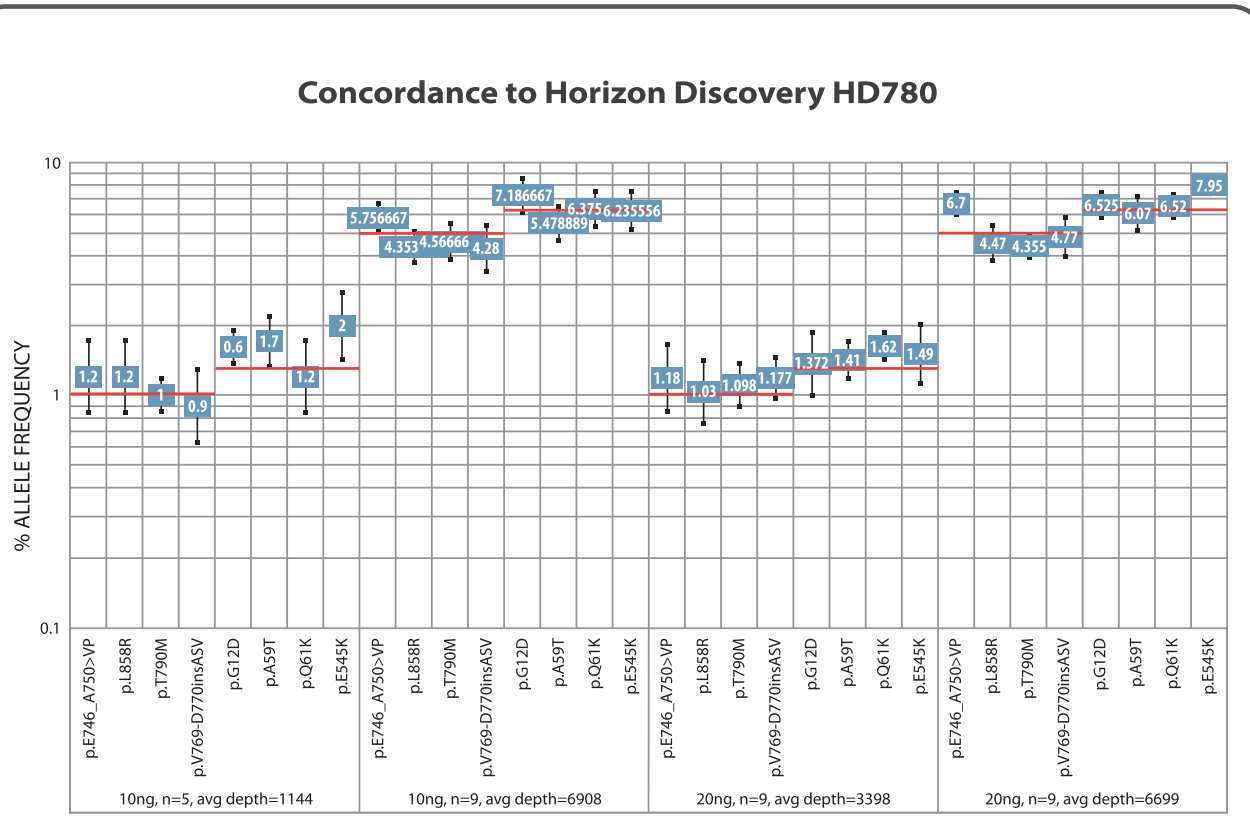

Figure 8. Paragon Genomics OncoZoom™ library made with Horizon Discovery's standard cfDNA (HD780). The red lines represent the known allele frequencies of eight mutations. The numbers highlighted in blue are detected allele frequencies from sequencing the OncoZoom™ libraries at specified read depths.

Paragon Genomics, Inc

3521 Investment Blvd Suite 1, Hayward CA 94545

415-855-0957

contact@paragongenomics.com