Next generation sequencing guide

For Research Use Only. Not for use in diagnostic procedures. | Version 2

Table of contents

INTRODUCTION

Next generation sequencing (NGS) refers to a variety of DNA sequencing technologies that decode millions to billions of DNA or RNA (via cDNA) molecules simultaneously, commonly referred to as massive parallel sequencing. Sequencing can either be **targeted**—i.e., focused on a select region of a genome or a specific set of genes—or whole, meaning that the goal is to obtain sequences from the entire sample (Figure 1). In both cases, to facilitate sequencing, the nucleic acid sample needs to be modified with the addition of platform-specific **adapters** that are recognized by the instrument through a process called NGS **library** preparation. For example, when preparing DNA libraries from genomic DNA for Illumina® platforms, high molecular weight (HMW) DNA is first sheared into short fragments either by sonication or enzymatic fragmentation, DNA ends are repaired, and then adapter sequences are attached to each fragment end to enable library amplification and flow cell binding. Next, cluster generation occurs where clonal copies of each library molecule are generated on the flow cell, and then the sequencer directs the synthesis of a complementary strand of the molecules within each cluster by flowing a sequencing primer with subsequent cycles of DNA polymerase, fluorescently labeled nucleotides, and buffers through microfluidic chambers. The library molecules are copied and upon the incorporation of one base at a time, the signal is detected and recorded that is unique to each base. The process of **sequencing by synthesis** varies across platforms in how detection and recording of the incorporated base occurs, but Illumina sequencing platforms use fluorescently labeled nucleotides in a variety of formats, depending on the instrument. And through this process, each molecule is **assembled** into a short **read**, which is then bioinformatically processed and used for a variety of analyses (Figure 1).

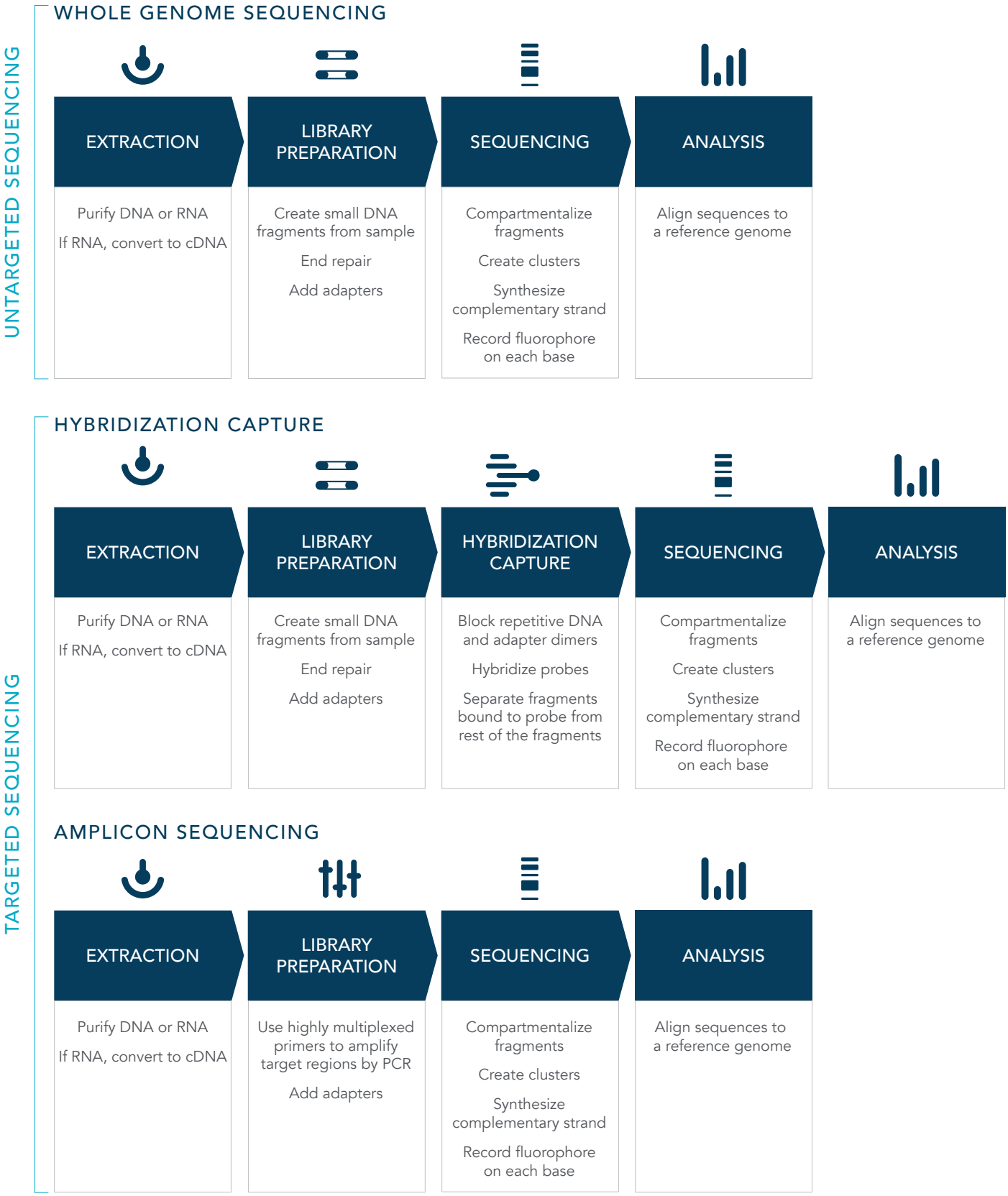


Figure 1. Workflows for whole genome sequencing (WGS) as well as targeted sequencing using hybridization capture and amplicon sequencing. For whole genome sequencing and hybridization capture, samples are extracted and sheared with either enzymatic treatment or Covaris® sonication. Library preparation includes preparing the ends of the fragments with adapters compatible with the sequencing instrument. For hybridization, library preparation is followed by the isolation of desired sequences using hybridization probes, thus reducing the number of fragments for sequencing. In amplicon sequencing, library preparation occurs by using highly multiplexed primers to amplify the genomic regions of interest. A second amplification adds the adapters. In all three methods, individual library fragments are compartmentalized, amplified to create clusters, and then sequenced by synthesis. Sequencing data is then analyzed.

The sequencing technology varies depending on the instrument in use. In general, sequencing can be divided into long-read and short-read technologies. In Table 1, a summary of popular instrumentation is described.

Table 1. Comparisons among representative sequencing technologies

Sequencing technology	Solid support	Compartmentalization	Cluster generation	Sequencing chemistry	Mechanism of base identification	Read length
Illumina (iSeq™ 100, MiniSeq™, MiSeq™, NextSeq™)	Flow cell	Individual binding sites on a flow cell	Bridge amplification	Sequencing by synthesis	Unique fluorophores on each deoxynucleoside triphosphate are identified by their emission wavelength	Varies from 75–250 depending on instrument and application
Thermo Fisher Scientific (Ion Torrent™)	Semiconductor chip	Individual 5 µm beads in wells of a multi-well plate	Emulsion PCR of unique fragments used to make 50,000 copies on each bead	Flood wells with one deoxynucleoside triphosphate at a time	Release of protons during base addition drops pH of well	Up to 600 base pairs (bp)
BGI	Silicon water flow cell	Positively charged material on flow cell has ordered pattern	Combinatorial probe-anchor synthesis (cPAS) uses modified rolling circle replication to create DNA concatamers that fold into DNA nanoballs	Sequencing by synthesis	Unique fluorophores on each base are identified by their emission wavelength	50–800 bp
Pacific Biosciences (Sequel™ system)	SMRT™ Cell with millions of tiny wells called zero-mode waveguides	Individual DNA fragments are immobilized at the bottom of well	N/A	DNA polymerase synthesizes complementary strand	Fluorescent light emitted during individual nucleotide addition is recorded	>10,000 bases
Element Biosciences	Flow cells	Individual binding sites on flow cell	Amplification using PCR	Long fragments are divided into short fragments with identical molecular identifier and sequenced	Unique fluorescent deoxynucleoside triphosphates for each base	

Following sequencing, one common approach for data analysis is to align or map reads to a **reference sequence**, such as a genome or targeted genome reference. Then **sequencing depth** (also called read depth, i.e., the number of short reads that align to the same area of a reference genome) can be determined. This is an important metric for establishing the reliability of sequencing results. Many factors affect the minimum sequencing depth that is necessary to confidently answer a biological question using NGS. These factors include, among others, whether germline vs. low frequency somatic DNA alterations are being tracked and detection of high vs. low expressed transcripts for RNA-seq. For example, detection of germline single nucleotide variants (SNVs) is often achieved from 30X coverage depth whereas rare somatic mutations can require 1000X depth or greater, depending on the desired limit of detection.

In addition, the read count needed to achieve the minimum sequencing depth is influenced by both the size of the target region of interest (whole genome/transcriptome vs. targeted sequencing) as well as the input nucleic acid quantity, as low inputs may be associated with higher PCR duplicate rates that are removed prior to analysis. Approaches that either reduce the number of uninformative reads or increase the number of informative reads, such as WES, have an order of magnitude lower number of reads compared to WGS, and therefore can boost experimental power. More focused targeted sequencing panels enable even deeper sequencing to detect lower abundance alleles and transcripts from a lower total read count due to the reduced target size.

Therefore, design of NGS experiments and choice of the appropriate sequencing method require careful consideration. It is also important to keep in mind that a trade-off is usually needed between sequencing depth and the number of controls and replicates. In addition to having multiple reads for individual regions, sequencing libraries should uniformly cover the entire mappable genome of interest, or the targeted genomic regions if performing a type of target enrichment or capture—this is referred to as breadth of coverage. Both **coverage** and sequencing depth are essential to obtaining interpretable and statistically powerful data for analyses, e.g., **variant calling** in the example of DNA sequencing for identification of SNVs or indels (Figure 2).

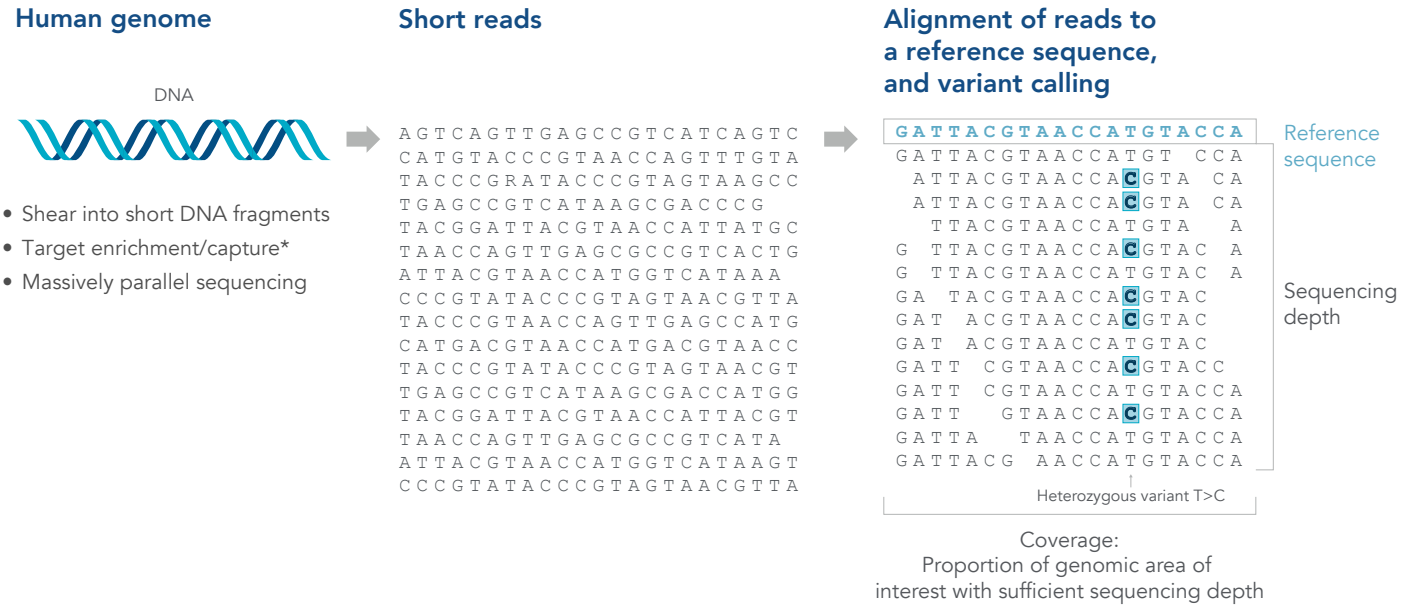


Figure 2. Example next generation sequencing workflow. During an NGS library preparation using high molecular weight DNA samples for short read sequencing, DNA is sheared into similarly sized fragments (between 50–600 nucleotides). The fragments are then end repaired and modified to include platform-specific adapter sequences at each end (not shown), and then sequenced in parallel. DNA sequences, or reads, can then be aligned to reference sequences and variant calling can be completed.

The information obtained from NGS has a broad range of applications including:

- Assessing genetic variation in populations and disease
  - Detecting both germline (inherited) and somatic (acquired) variants
- Characterizing taxonomic profile and functional potential of microbial communities
- Discovering differential gene expression patterns
  - Identifying novel splice variants and fusion transcripts
- Characterizing DNA methylation signatures

Most of these applications are supported by short-read sequencing due to the lower cost per base, high-throughput capability, and sequence fidelity. Long-read sequencing is available to support ***de novo* assembly** and identification of structural variants. These two types of NGS platforms are often combined to build out complete pictures of genomic structure and variations.

IDT offers a diverse line of predesigned and custom products to support:

- **Whole genome sequencing** from high-quality and low-quality samples (e.g., FFPE, cfDNA, ssDNA, etc.) from nanogram down to picogram input ranges
- **Whole exome sequencing** from high-quality and low-quality samples using hybridization capture probes (smaller targeted panels for a subset of loci are also available)
- **Targeted DNA sequencing** from high-quality and low-quality samples using multiplexed PCR panels
- **Metagenomic and metatranscriptomic analysis** from complex, highly diverse samples for microbial populations (e.g., soil, gut, skin, etc.)
- **Whole transcriptome sequencing** from varying quality of RNA samples ranging from nanogram down to picogram input ranges
- **Targeted RNA sequencing** using hybridization capture technology without the need for ribodepletion or poly(A) mRNA selection
- **Methylome analysis** from single-stranded, bisulfite-converted DNA down to picogram inputs, as well as single-cell methylation sequencing (methyl-seq) applications
- **Targeted methylome analysis** using hybridization capture technology
- **Chromatin studies** for chromatin immunoprecipitation (ChIP) or Hi-C (chromosome conformation capture) technologies

There are IDT products available for tackling a variety of workflows, from high-throughput applications focused on cost and turnaround time to workflows that are designed for low abundance or challenging samples.

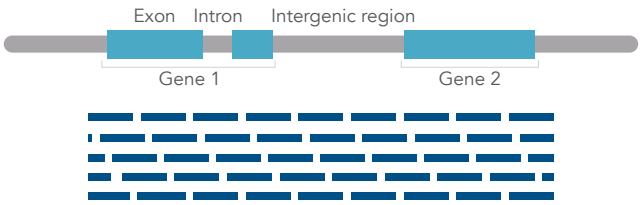
In this handbook, steps of the various NGS applications, from sample input evaluation to sequencing, will be discussed. We also provide links to the innovative solutions that IDT offers to support these applications along the way.

NGS WORKFLOWS

The NGS workflows discussed in the following section are for whole sequencing of a sample or targeted sequencing of a subset of regions in a sample (Figure 3). Besides the objective of the study, the choice to perform whole or targeted sequencing depends on the following:

- Specific genes of interest
- Complexity of the sample
- Desired sequencing depth

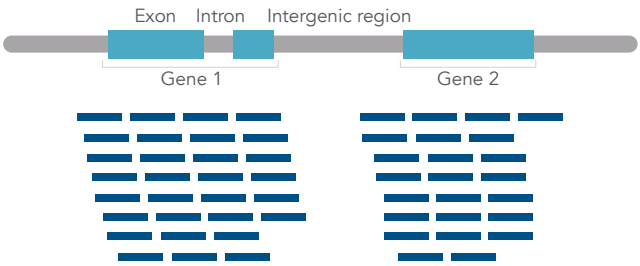
A. Whole Genome Sequencing (WGS)



Sequence entire DNA or RNA sample including non-coding or repetitive regions

- Moderate sequencing depth
- May identify copy number variant and repeat expansions
- Least cost effective
- Largest data set

B. Whole Exome Sequencing (WES)

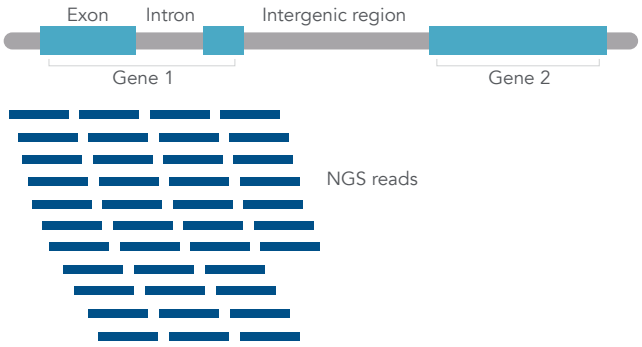


Select exonic sequences from whole genome sample (1.5–2% of human genome)

- Variable sequencing depth depending upon application
- Can achieve greater sequencing depth with same number of reads as WGS
- More cost effective
- Large data set

C. Targeted Sequencing (TS)

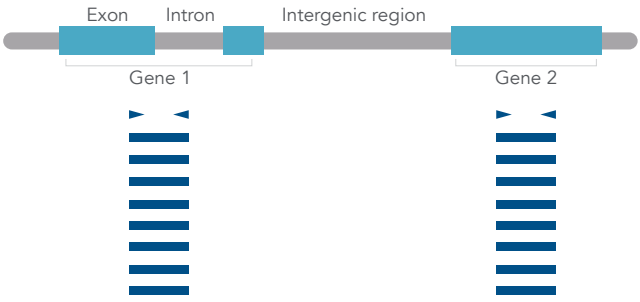
Hybridization capture



Select genes of interest with predesigned or custom hybridization capture probes

- Can achieve greater sequencing depth with lower number of reads
- More cost effective
- Variable size data sets

Amplicon sequencing



Amplify genes of interest with multiplexed PCR primers

- Can achieve greater sequencing depth with lower number of reads
- Most cost effective
- Variable size data sets

Figure 3. Sequencing options. (**A**) Whole genome sequencing analyzes coding and non-coding regions of extracted DNA. (**B**) Whole exome sequencing (WES) focuses on all protein-coding regions of the genome of interest. IDT offers predesigned hybridization capture panels for human genome sequencing. (**C**) Targeted sequencing (TS) focuses on specific regions or genes of interest by either hybridization capture or multiplexed PCR.

Whole genome sequencing (WGS), whole transcriptome sequencing, and whole genome bisulfite sequencing (WGBS) provide comprehensive coverage for discovery research. These applications generate enough reads to cover the whole genome and provide comprehensive data for downstream analyses. However, the number of created library molecules and size of the genome can require a high number of reads per sequencing flow cell, and therefore, reduces the number of samples that can be sequenced at one time. In contrast, targeted sequencing strategies have a smaller target space, meaning that fewer library molecules are needed to obtain sufficient read coverage and sequencing depth for each region, which enables higher multiplexed sequencing. It is also worth noting that sequencing depth for each NGS application and project varies substantially depending on the analyses planned. For applications such as detection of rare somatic mutations, a sequencing depth of 1000X or greater may be needed, whereas germline variant analysis may only need depths of 15–30X [1].

Whole exome sequencing uses targeted hybridization capture probes to enrich for genomic fragments for the protein-coding exons from a prepared whole genome library. By focusing the sequencing on regions of interest, exome sequencing uses fewer sequencing reagents, which, therefore, lowers costs. Whole exome sequencing can also be applied to RNA-seq where deeper sequencing can be achieved relative to whole transcriptome sequencing from an equivalent or lower number of reads. Deeper sequencing that can be achieved in this approach can facilitate RNA fusion discovery from FFPE and other oncology samples commonly used for research. For more information, see the application notes, Whole-exome sequencing using the xGen DNA Library Preparation Kit family enables customizable workflows returning quality in consistent results, RNA-seq for biomarker identification using the xGen Broad-Range RNA Library Kit and xGen Custom Hyb Panels, and A hybridization capture-based approach for RNA sequencing and gene fusion discovery using the xGen Exome Hyb Panel v2.

Other targeted sequencing workflows can be used for small panel designs in the kilobase (kb) size range. Easy workflows include amplicon sequencing, where highly multiplexed primers are designed to amplify the targets of interest. Hybridization capture targeted sequencing can be used for small or large target regions in the kb to megabase (Mb) size range, up to and including the whole exome (Figure 3). The choice between targeted sequencing methods depends on the panel size (kb vs Mb) and the desired workflow; amplicon sequencing workflows are simpler and completed significantly faster.

IDT offers a variety of predesigned amplicon sequencing panels and hybridization capture panels for targeting oncology genes, the 16S rRNA gene, SARS-CoV-2, among others. They are suitable for germline or somatic variant calling, and some are suitable for analysis of differential gene expression. IDT also offers custom amplicon sequencing panels and custom hybridization panels that are supported by our technical experts who can help guide the design process and answer any questions that may arise. Whether performing target discovery or analyzing a known set of targets, IDT has workflows available for many applications.

LIBRARY PREPARATION FOR THE DIFFERENT NGS WORKFLOWS

The choice of a library preparation method depends on both the sample type and the NGS workflow used for the research study (Table 2).

Table 2. Steps for library preparation by NGS workflow for whole genome, transcriptome, or methylome sequencing.

Application	Sample type	Sample input range	Fragmentation	End repair	dA-tailing	Adapter addition strategy	Indexing PCR	Normalase™ compatibility
Whole genome sequencing (high-quality DNA)	Whole-genome dsDNA	0.1 ng to 1 µg	Enzymatic (EZ) or mechanical (MC)	√	√	xGen Stubby Adapter	√	√
Whole genome sequencing (high-quality DNA)	Whole-genome dsDNA	0.1 ng to 1 µg	Enzymatic (EZ UNI) or mechanical (MC UNI)	√	√	Full-length indexed adapter (with UMI)	N/A	√
Whole genome sequencing (FFPE or cfDNA)	FFPE or cfDNA (dsDNA)	1–250 ng	MC (optional)	√		Addition of Ligation 1 adapter followed by priming of gap fill by Ligation 2 adapter	√	
Whole genome sequencing (single-stranded or low-input)	ssDNA/ dsDNA	10 pg to 250 ng	MC (optional)	√		Adaptase™ technology adds R2 Stubby Adapter on ssDNA fragment followed by R1 Stubby Adapter ligation	√	√
Whole transcriptome sequencing (broader input range)	Total RNA Ribo-depeted RNA Poly(A)-enriched RNA	10 ng to 1 µg total RNA 100 pg to 100 ng mRNA	Heat			Adaptase technology adds R2 Stubby Adapter after single-stranded cDNA synthesis, followed by ligation of R1 Stubby Adapter	√	√
Whole transcriptome sequencing (faster workflow)	Total RNA Ribo-depeted RNA Poly(A)-enriched RNA	100 ng to 1 µg total RNA 5 ng to 100 ng mRNA	Heat			Random prime RNA with R1 Stubby Adapter tailed primer for first-strand cDNA synthesis; Adaptase™ technology adds R2 Stubby Adapter	√	√
Whole methylome sequencing	Bisulfite-converted ssDNA from gDNA, FFPE, cfDNA samples	100 pg to 100 ng	Mechanical	√	√	Adaptase™ technology adds R2 Stubby Adapter on ssDNA fragment followed by R1 Stubby Adapter ligation	√	√
Single-cell methylation sequencing	Bisulfite-converted ssDNA (single-cell)	Single-cell	None	√	√	Random prime RNA with R1 Stubby Adapter tailed primer for extension product synthesis; Adaptase™ technology adds R2 Stubby Adapter	√	√

WHOLE GENOME SEQUENCING

NGS approaches like whole genome sequencing can be used for a variety of applications including understanding single nucleotide variation (SNV), insertions/deletions (indels), copy number variation (CNV), and other chromosomal aberrations. Whole genome information provides a comprehensive picture for pinpointing the origin of inherited disease, cancer, or population genetics. Whole genome sequencing is also critical for *de novo* genome assemblies for other organisms for which no reference sequences are currently available.

Shotgun metagenomic sequencing (or metagenomics) is a type of whole genome sequencing that researchers use to capture genomic information from mixtures of species within a sample. This approach is particularly useful for understanding the taxonomic and functional diversity of complex communities of microorganisms that may be found in, e.g., stool, seawater, or soil samples. There is a world of hidden biodiversity of microorganisms (viruses, bacteria, archaea, and micro-eukaryotes) that are largely only accessible through culture-independent methods like NGS. Creating NGS libraries from degraded, high host-contaminated samples, or samples with low biomass but high biodiversity can be challenging and requires special considerations. For example, many viral genomes are single-stranded, dying bacteria release partially degraded DNA, and the sample matrix could include enzyme inhibitors or nucleases that disrupt library construction and lower the conversion rates.

IDT has several DNA library preparation kits for whole genome sequencing and shotgun metagenomic sequencing. The xGen DNA Library Prep EZ Kit (enzymatic fragmentation) and xGen DNA Library Prep MC Kit (mechanical shearing) offer short workflows for NGS library construction from DNA samples and can easily be automated to support low cost, high-throughput studies for whole genome analysis. These kits include an end repair/A-tailing step followed by adapter ligation using xGen Stubby Y Adapters plus indexing primers, or xGen UDI/UMI full length indexed adapters. Conveniently, the workflows only have two bead-based purification steps after the ligation and amplification reactions. The library amplification polymerase used in the PCR is compatible with both direct sequencing and hybridization capture. The xGen Normalase Module is also available to expedite library normalization for high-throughput workflows where multiplexed sequencing is being used (see the Library quantification and the Normalase workflow section for more information).

The xGen cfDNA & FFPE DNA Library Prep Kit includes end repair steps and sequential single-stranded adapter ligations to double-stranded DNA substrates that facilitate working with damaged FFPE and cfDNA samples. The highly efficient chemistry is combined with relaxed bead-based purification steps that preserve sample complexity from difficult samples. The final libraries are suitable for direct whole genome sequencing and include UMIs for error correction. For heavily damaged samples not amenable to dsDNA library prep, the xGen ssDNA & Low Input DNA Library Kit includes the xGen Adaptase technology for library construction from single-stranded DNA (ssDNA) molecules. Since the sample is denatured prior to Adaptase enzymatic attachment of single-stranded adapters, both ssDNA and dsDNA are captured into the library with inputs as low as 10 pg. As a result, this workflow accurately represents the relative abundance of both ssDNA and dsDNA viral genomes in metagenomic samples. This kit is also suited for heavily damaged samples that can't be supported by workflows that require intact dsDNA.

WHOLE TRANSCRIPTOME SEQUENCING

Similar to whole genome sequencing, whole transcriptome sequencing (also called total RNA sequencing or RNA-seq) provides a comprehensive view of the transcripts present in a sample, thus providing insights into differential gene expression and identification of novel splice variants and fusion transcripts associated with disease. Since approximately 80% of cellular RNA is ribosomal, typical whole transcriptome workflows include either ribodepletion to remove rRNA or poly(A) selection to enrich for mRNA to eliminate most ribosomal transcripts in the dataset. This enables detection of rare and less abundant transcripts from samples during sequencing.

Metatranscriptomics is a type of NGS sequencing that captures all expressed genes (RNAs) from every organism in a sample. While metagenomics provides a snapshot of a taxonomic profile of a microbial community, i.e., it helps address the question "what specific microorganisms are present in a sample under given conditions," metatranscriptomics goes one step further and answers the question "what is the functional potential of a microbial community under these conditions." Since metatranscriptomics provides sequencing information about all expressed genes in a sample, it may help identify differences between functional potentials of the same microbial community under different conditions. For example, it may demonstrate how different environmental conditions influence gene expression of a microbial community in a natural ecosystem, or it may help demonstrate what functional changes in a microbial community are caused by its response to different drugs or drug concentrations in a laboratory setting.

Two kits are available for RNA-seq library construction from IDT: the **xGen RNA Library Prep Kit** and the **xGen Broad-Range RNA Library Prep Kit**. Both kits include Adaptase technology for adapter ligation to first strand cDNA, producing stranded libraries with no requirement for second strand synthesis. The xGen RNA Library Prep Kit is designed for more abundant samples or samples of higher quality and offers a shorter workflow that can support higher throughput studies. In contrast, the xGen Broad-Range RNA Library Prep Kit supports both high and low input quantities (100 pg–100 ng mRNA; 10 ng–1 µg total RNA). Both kits include reagents for RNA fragmentation, cDNA synthesis, and library preparation.

Both kits are also compatible with poly(A) mRNA selection (suitable for eukaryotic samples with high integrity RNA scores) and ribodepletion (suitable for low integrity RNA samples or studies involving non-adenylated transcripts).

DNA METHYLATION AND EPIGENETICS

For methylome analysis, researchers use an NGS approach called **methylation sequencing** (methyl-seq) or whole genome bisulfite sequencing (WGBS).

Traditionally, to start this workflow, DNA is subjected to a bisulfite conversion which changes all unmethylated cytosines to uracil and any methylated cytosines are protected. Therefore, any cytosine present in the sequencing read indicates it was methylated *in vivo*. Bisulfite conversion leads to both DNA denaturation and fragmentation, so if using a traditional dsDNA-based library prep, library preparation must precede bisulfite conversion, and methylated adapters are also required. However, bisulfite conversion of complete libraries leads to significant loss of library molecules (**Figure 4A**). For the xGen Methyl-Seq DNA Library Prep Kit, the steps are reversed, and the bisulfite conversion occurs before library construction and uses proprietary Adaptase technology that is designed for fragmented, single-stranded DNA (**Figure 4B**), thus preserving library complexity. Another alternative is random priming which performs bisulfite conversion of unfragmented double-stranded DNA. However, this approach is inefficient and biased, leading to lower complexity libraries (**Figure 4C**).

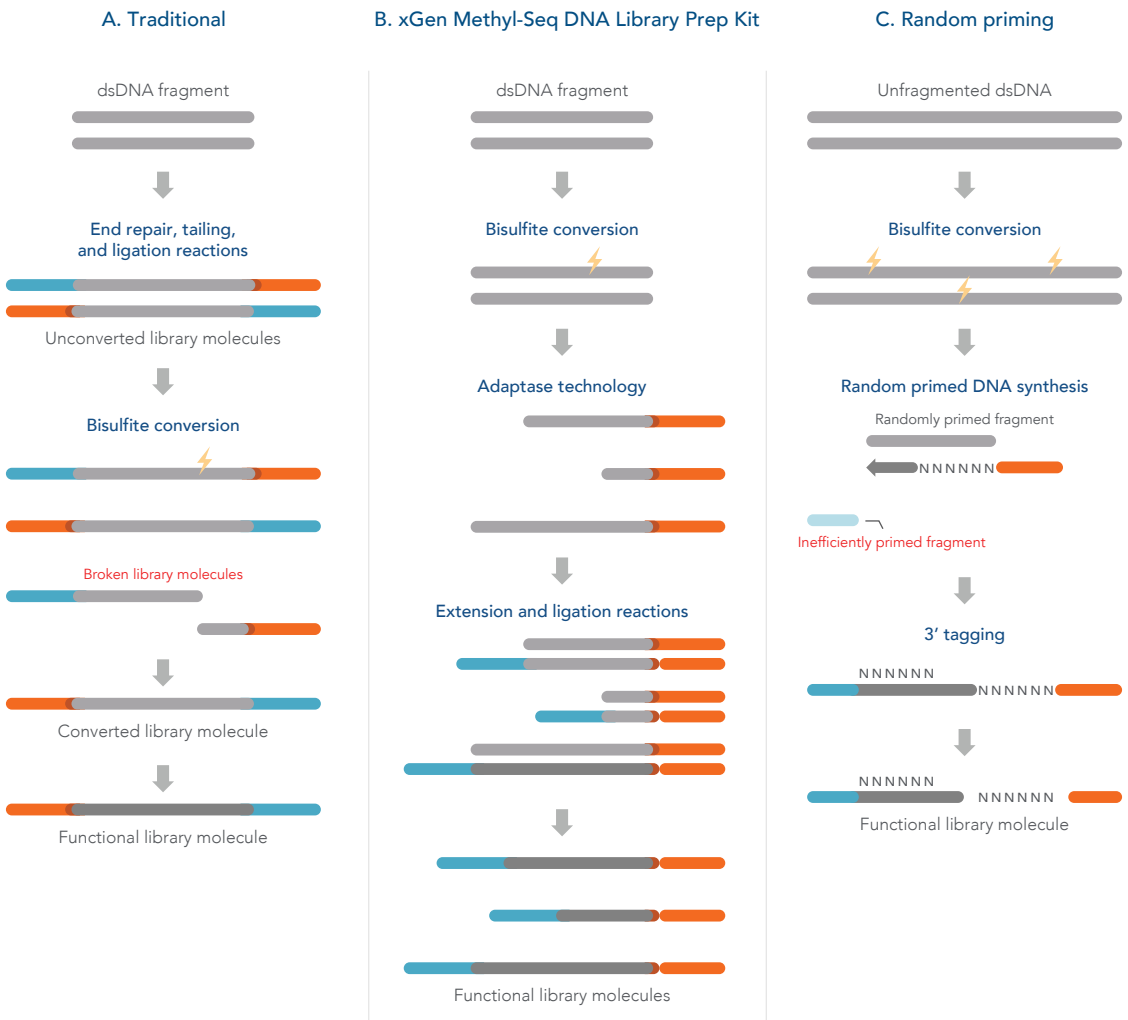


Figure 4. Methyl-seq library preparation workflow comparisons. (**A**) The traditional method of methylation sequencing builds the library before bisulfite conversion, which leads to broken library molecules. (**B**) Before using the **xGen Methyl-Seq DNA Library Prep Kit**, double-stranded DNA (dsDNA) fragments undergo bisulfite conversion. The kit then directly converts these fragments into library molecules. (**C**) The random priming method converts unfragmented dsDNA with bisulfite treatment first, but due to inefficient random priming, the samples are poorly represented in the final library.

For single cell methylation studies, the **xGen Adaptase Module** is available to support a range of research studies [2,3]. For further information please see the **xGen Adaptase Module protocol** which includes considerations for experimental design, procedure, and data analysis.

TARGETED SEQUENCING—SELECTING FOR SPECIFIC GENOME TARGETS

Whole genome, transcriptome, methylation, shotgun metagenomic, and virome sequencing are experimental techniques essential for fundamental target discovery. These approaches can reveal novel genetic variants, gene expression signatures indicative of various disease states, methylation patterns common to various cancers, as well as discovering new viral or microbial species. Following target discovery approaches, implementation of targeted sequencing enables a higher depth of sequencing to enable detection of low frequency somatic variants, as well as limit sequencing to targets of interest to reduce data output and overall cost.

There are two different strategies for targeted sequencing: hybridization capture and amplicon sequencing. The choice of which to use is often based on the size of the desired target; hybridization capture can support kb- to Mb-sized panels, up to and including the exome. Amplicon sequencing is suited for smaller panels in the kb size range and offers a faster workflow.

HYBRIDIZATION CAPTURE

The IDT xGen Hybridization Capture workflow uses a series of 5′ biotinylated oligonucleotide probes and streptavidin magnetic beads to select library fragments of interest. Target enrichment is performed following full genome/transcriptome library preparation. Custom or predesigned panels of hybridization capture probes can be purchased from IDT to target a variety of different relevant gene sets, and optimized hybridization reagents and blockers are also supplied to achieve high on-target rates with uniform target coverage. IDT synthesizes these probes individually and then pools them into predesigned or custom panels, which ensures equal representation of probe in the panel. The IDT xGen Hybridization Capture panels include:

- xGen Custom Hyb Panel can be created to your specifications and if needed, also include a full functional test report and/or quality documentation.
  - IDT also supports custom xGen MRD Hybridization Panels for oncology research.
  - xGen Custom Hyb Panels can also be designed for targeted methylation sequencing, where specific design parameters are required to address targeting the bisulfite converted genome (see the Cost-effective targeted methyl-seq using an xGen Custom Hyb Capture Panel and the xGen Methyl-Seq DNA Library Prep Kit identifies DNA methylation in low input samples consistent with WGBS application note).
- xGen Exome Hyb Panel targets 19,433 genes to cover the human coding regions, which ensures complete coverage of the human exome. Individual synthesis of each probe ensures the panel provides comprehensive coverage for exome studies.
- xGen AML Cancer Hyb Panel targets >260 genes associated with acute myeloid leukemia.
- xGen Inherited Disease Hyb Panel targets genes and single nucleotide polymorphisms (SNPs) associated with inherited diseases including hereditary cancer syndromes, ciliopathies, neuromuscular disorders, short stature, and many others.
- xGen Pan-Cancer Hyb Panel captures 137 genes implicated in 12 tumor types.
- xGen CNV Backbone Hyb Panel is a spike-in panel that can identify copy number variations (CNVs) in the human genome.
- xGen Human ID Hyb Panel is another spike-in panel that identifies a set of SNPs that distinguishes one individual from over 5 million biological samples.
- xGen Human mtDNA Hyb Panel is also a predesigned spike-in panel that identifies key mitochondrial DNA variants.
- xGen SARS-CoV-2 Hyb Panel captures the viral SARS-CoV-2 genome for tracking viral evolution and variant development.

AMPLICON SEQUENCING

Amplicon sequencing uses highly multiplexed sets of PCR primers to amplify the regions of interest generally followed by a second round of PCR to uniquely index each library for multiplexed sequencing analysis. As with hybridization capture panels, amplicon panels are available as a custom product as well as predesigned panels. Amplicon sequencing panels have a short workflow from sample to sequencer in under three hours and include the IDT Normalase reagents to streamline equimolar pooling for multiplexed sequencing. IDT provides three different technologies for amplicon sequencing, including xGen Amplicon Technology, standard highly multiplexed primer panels, and xGen HS Amplicon Technology.

xGen™ Amplicon Technology

xGen Amplicon Technology from IDT enables single-tube multiplexed PCR even when contiguous target coverage of overlapping amplicons is needed, unlike other methods that require multiple tubes for target enrichment. In addition, this workflow can produce super amplicons that prevent amplicon dropout when a novel variant overlaps with a primer, and therefore maintains complete coverage (Figure 5).

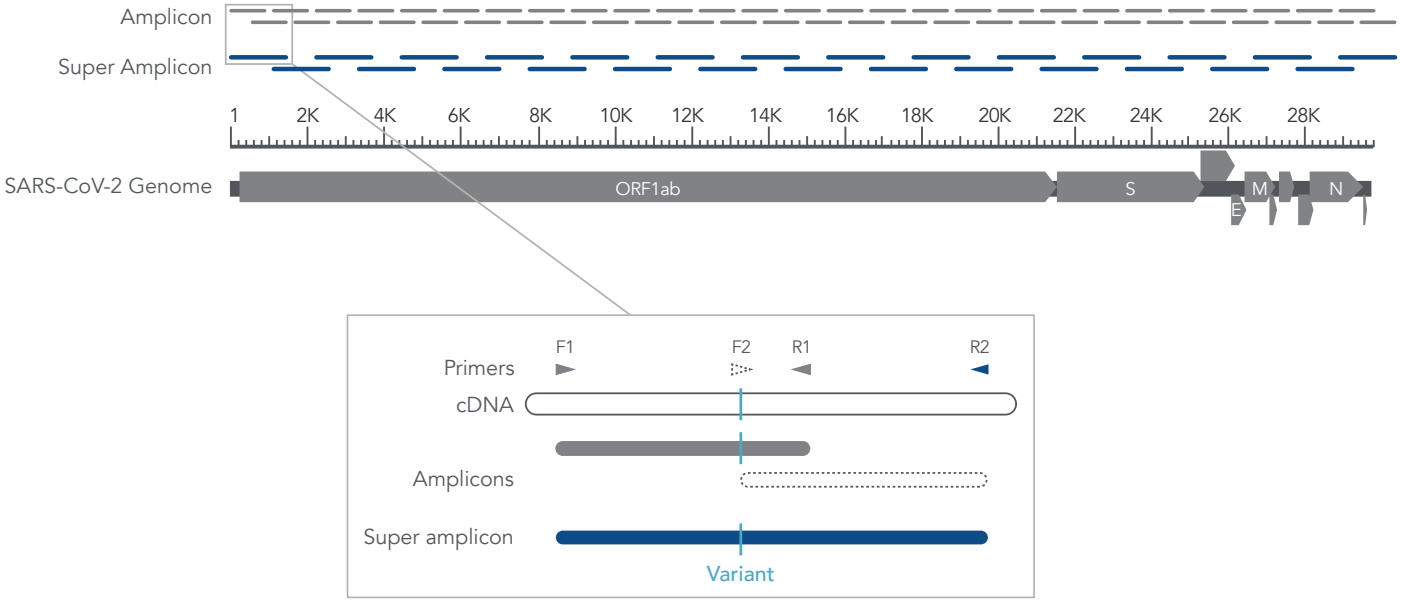
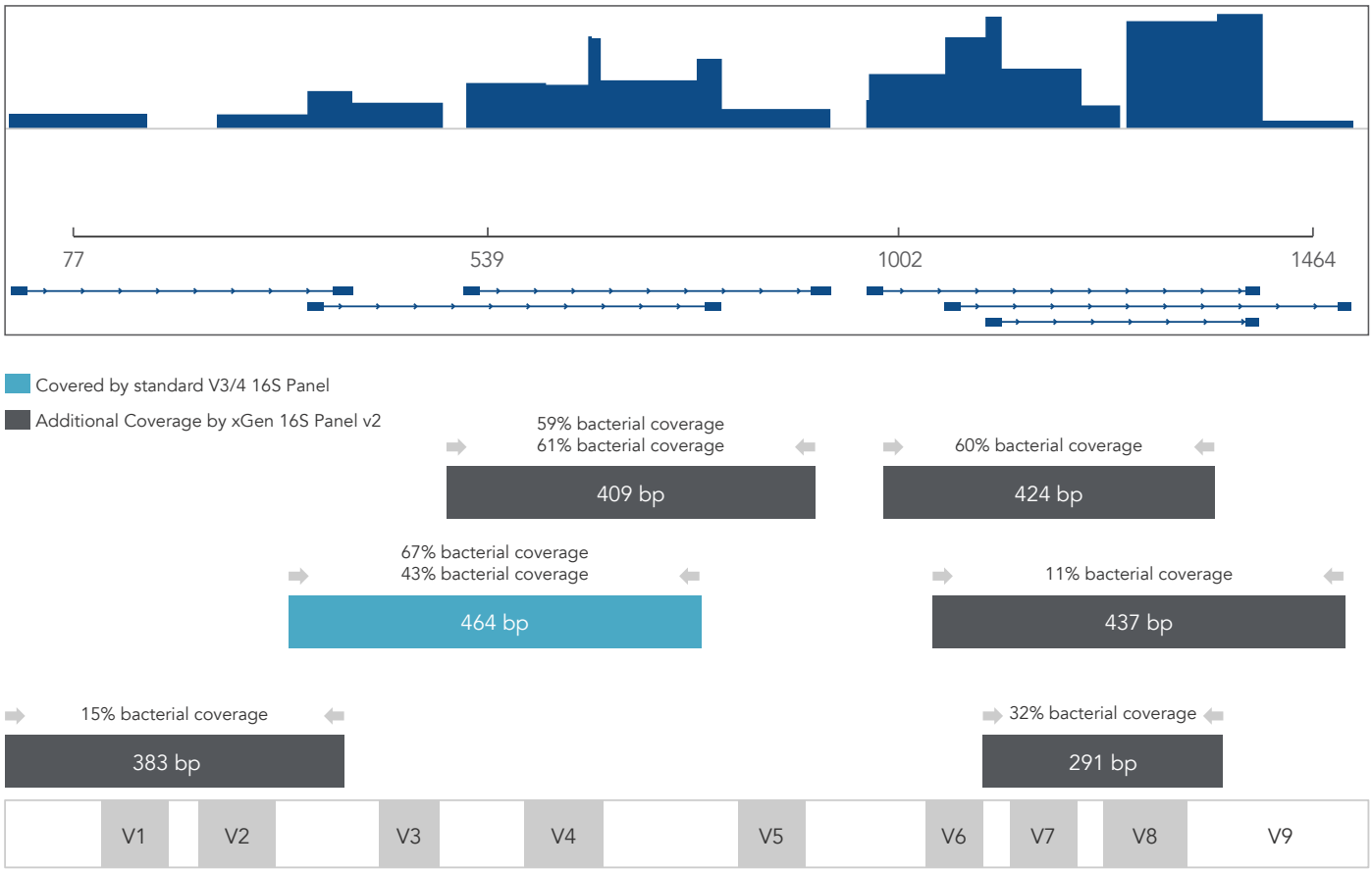


Figure 5. xGen amplicon one-tube solution results in super amplicons. With the xGen amplicon single-tube workflow, there is high overlapping amplicon coverage. For example, using the xGen SARS-CoV-2 Amplicon Panel, when an individual primer (such as primer F2) anneals less efficiently due to a genomic mutation, the nearby primer pair, F1 and R2 create a super amplicon that covers and identifies the variant.

- Custom amplicon panels created to target your specific gene set or small genome to create multiple overlapping amplicons for contiguous coverage in a single-tube workflow.
- xGen SARS-CoV-2 Amplicon Panel creates 345 overlapping amplicons of SARS-CoV-2 for complete genome coverage with the ability to create super amplicons that prevent coverage gaps as new variants arise.
- xGen Oncology & Inherited Disease Amplicon Panels provide both germline and somatic variant identification down to 1% allele frequency for a variety of known cancer-associated genes including *BRCA1* or *BRCA2*; genes associated with colorectal cancer, myeloid cancer, or lung cancer; genes in the *EGFR* pathway; *TP53* gene and pan cancer panels; as well as inherited disease such as cystic fibrosis (CFTR) and Lynch syndrome.
- xGen 16S and ITS1 Amplicon Panels amplify 16S rRNA or ITS1 variable regions of a genome. The 16S rRNA Amplicon Panel primers amplify all 9 variable regions of the 16S rRNA gene in comparison to a panel that only targets the V4 region which provides additional data for discriminating among the various bacterial species (Figure 6).

A. xGen 16S Amplicon Panel v2 (primers)



B. xGen ITS1 Amplicon Panel (primers)

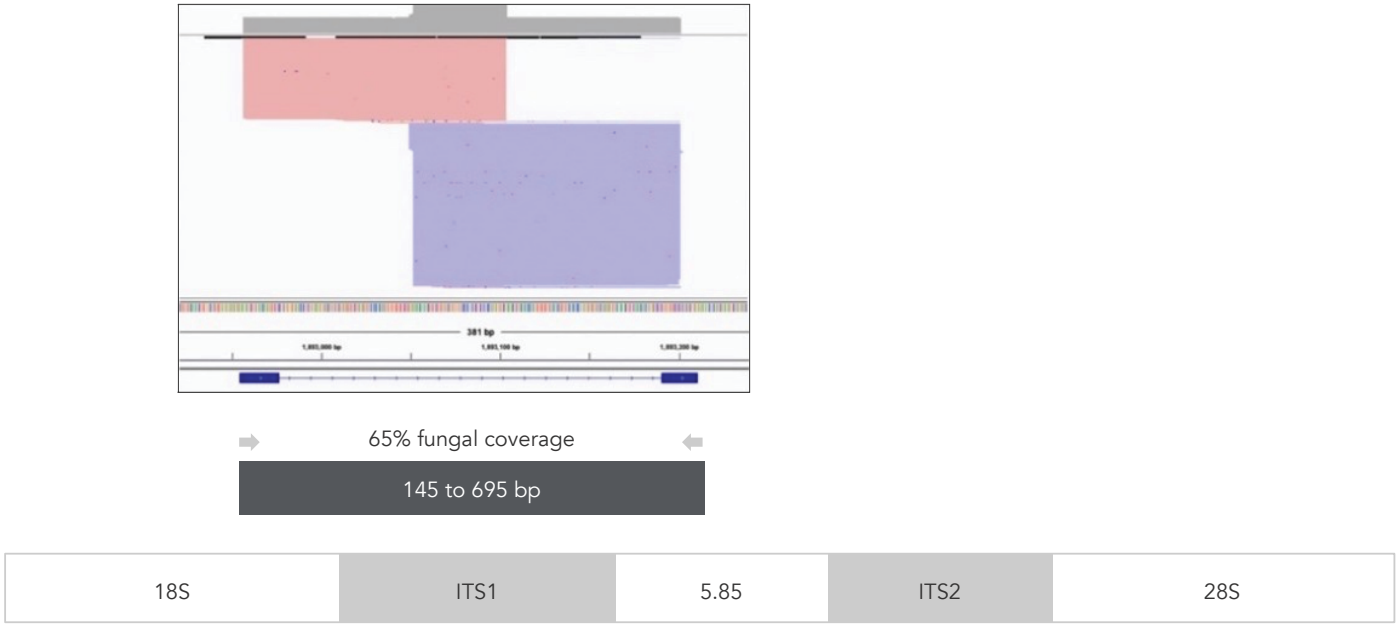


Figure 6. (**A**) **xGen 16S Amplicon Panel v2 (primers).** Sequencing read coverage for an *E. coli* DNA sample (*n* = 1) observed in Integrative Genomics Viewer (IGV) Sashimi plot and illustration of multiplexed primer coverage of all nine variable regions of 16S rRNA compared to a standard V3/V4 16S sequencing read coverage. (**B**) **xGen ITS1 Amplicon Panel (primers).** Sequencing read coverage observed in the IGV Sashimi plot and illustration of multiplexed primer coverage of ITS1 region for *Candida albicans* (*n* = 1). Reads originating from the forward primer are shown in red; reads from the reverse primer are in blue.

Multiplexed amplification primers

A two-tube workflow amplifies alternating segments of the SARS-CoV-2 genome separately and then combines the amplicons for final analysis.

- SARS-CoV-2 ARTIC Amplicon Panel, created by the ARTIC Network, has 98 primers in two pools that amplify the SARS-CoV-2 genomic material.
- SARS-CoV-2 Midnight Amplicon Panel is another set of amplicon primers to sequence SARS-CoV-2 that creates 29 tiled amplicons approximately 1200 bp in length.

xGen HS Amplicon Technology

Single-tube workflow with three PCR steps with incorporation of unique molecular identifiers (UMIs) in PCR I, target amplification in PCR II, and the addition of combinatorial dual indexed adapters in PCR III.

- xGen HS EGFR Pathway Amplicon Panels amplify variants of the EGFR pathway genes with a minor allele frequency of ≤1% from ≥10 ng of cell-free DNA (cfDNA) or formalin-fixed, paraffin-embedded (FFPE) DNA; custom panels also available.

For more information on amplicon sequencing and hybridization capture technologies, review the IDT Targeted Sequencing guide.

TIPS FOR KEY STEPS OF NGS LIBRARY PREP

NGS workflows vary depending upon many factors, including:

- Number, type, and mass of input samples
- Target region as well as the sequencing depth and coverage needed for analysis
- Type of planned data analysis
- NGS instrumentation that will be used for sequencing
- Number of libraries pooled into a single sequencing run

Figure 7 shows the steps of a typical NGS workflow. The steps in blue represent those for which IDT offers products or services.

Figure 7. General steps in a typical NGS workflow. If the input material is RNA, it needs to be converted to cDNA by reverse transcription (RT) prior to adapter and index addition.

The remaining sections of this guide will provide tips and tricks on these various factors for specific NGS library prep steps.

SAMPLE PREPARATION

Sample quantity and quality

NGS workflows require nucleic acid extraction prior to library preparation. A number of kits are available on the market that extract nucleic acids of sufficient quality for NGS from different types of samples including, among others, tissue, cell culture, FFPE, cell-free (cfDNA or cfRNA), viral, and environmental samples (soil, sediments, marine, and fresh water). Then, extracted samples need to be quantified to determine their concentration and purity. Quantification methods generally require the use of a portion of your extracted nucleic acid sample which can't be used for library construction. It is important to account for this sample loss during experimental design, especially when using low abundance samples.

Recommendations for quantification and quality assessment based on sample type:

- **Double-stranded DNA (dsDNA)**: Assess quantity using Qubit™ (Thermo Fisher Scientific) or other fluorometric dsDNA quantification method to directly measure the usable quantity of material in the sample for dsDNA-based library prep workflows. Quantification is typically performed before fragmentation by sonication or a similar mechanical method (see xGen DNA Library Prep Kit for more information).
- **DNA samples of lower quality (such as cfDNA, ctDNA, or FFPE):** The integrity of the sample may be critical for determining usable amount. For human samples, including cfDNA, the Input Quantification Primers from IDT can be used to assess usable quantity and integrity with a pair of ALU repeat qPCR assays where cross-links and nicks can be assessed as well as dsDNA breaks. The obtained integrity score determines the type of library prep workflow that will best support the sample. Alternatively, a DNA Integrity Number (DIN) can be measured by the Bioanalyzer®/TapeStation™ electrophoretic chip assay (Agilent). Remember, a DIN will not assess cross-links or nicks, only dsDNA breaks. Using this method to quantify cfDNA can also indicate the presence of genomic DNA (gDNA) contamination (see xGen cfDNA & FFPE Library Prep Kit for more information on input sample quality).
- **Precious DNA samples** of extremely limited input that lack sufficient material for quantification (such as low cell number, chromatin immunoprecipitation (ChIP) samples, etc.): Use a kit compatible with low input and the number of PCR cycles recommended for the lowest supported input during library amplification (see the xGen ssDNA & Low-Input DNA Library Prep Kit).
- **ssDNA, or mixtures of ssDNA and dsDNA:** Use a method that supports ssDNA library synthesis. NanoDrop™ (Thermo Fisher Scientific) or another spectrophotometric method can be used to determine quantity, although RNA should be removed for accurate DNA quantification (see xGen ssDNA & Low-Input DNA Library Prep Kit).
- **dsDNA for methyl-seq** requires a bisulfite treatment to convert the unmethylated cytosines to uracil, which produces ssDNA. It is important to quantify the dsDNA input by Qubit prior to conversion to obtain accurate library input amounts (see the xGen Methyl-Seq DNA Library Prep Kit and Adaptase Module).
- **RNA samples** can be assessed by a Bioanalyzer/TapeStation to obtain an RNA Integrity Number (RIN) and DV_{200} score, where RIN scores of 3–7 indicate usable damaged samples and scores >7 indicate high-quality samples; a DV_{200} score further refines the degree of fragmentation by calculating the percent of fragments >200 bases. This information not only determines input quantity and integrity, but also defines compatible upstream and downstream modules for mRNA enrichment or rRNA depletion. The NEBNext® Poly(A) mRNA Magnetic Isolation Module requires intact RNA samples to capture the entire transcript, whereas ribodepletion methods can support FFPE and other low-quality samples. Alternatively, targeted RNA sequencing with hybridization capture can be used that omits the rRNA fraction. Another advantage of this approach is that it provides deeper sequencing of targeted regions while omitting the undesired reads that often result in a high number of sequencing reads not relevant to the research study (see A streamlined workflow for targeted RNA-seq using the xGen RNA Library Prep Kit with xGen Custom Hyb Panels, xGen RNA Library Prep Kit protocol, and xGen Broad-Range RNA protocol).

Input amount for sufficient library complexity and sequencing coverage

Normally, the required nucleic acid input quantity for an NGS analysis is determined by the experimental objectives and the sequencing chemistry/instrument that will be used. It is important to check the input requirements for your library construction and sequencing needs before starting the extraction process.

In general, using a larger input quantity leads to higher complexity libraries and the ability to identify less abundant species in the sample (Figure 8). Here, the term complexity refers to the number of unique fragments that can be identified after sequencing. In other words, if a high input quantity is used, deep sequencing will result in high genome coverage, whereas, if a low input is used, even if deep sequencing is performed, PCR duplicates will be encountered due to the limited number of unique fragments, and the library may saturate (all unique molecules sequenced) before high genome coverage can be achieved.

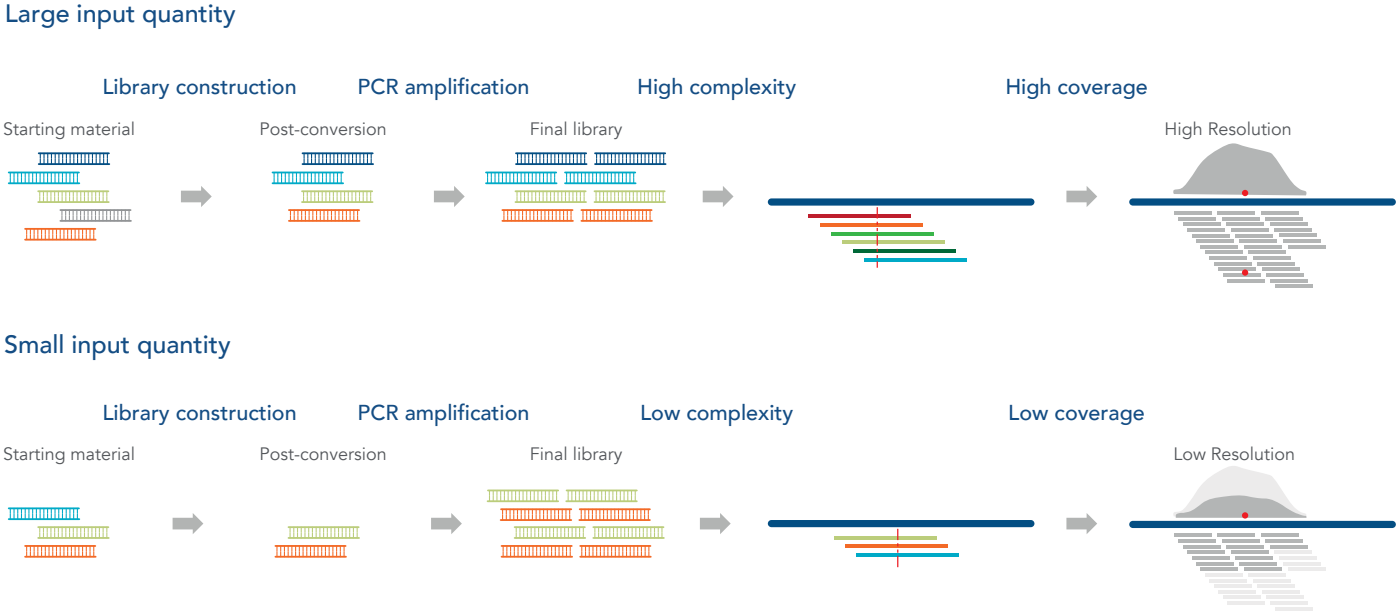


Figure 8. Input quantity directly relates to library complexity and depth of unique sequencing reads. There is a correlation between the amount of the input material and the resulting library complexity, which in turn determines the sequencing coverage and therefore the ability to identify rare microbial species, mutations, or variations. For a larger input quantity, the complexity of the library is usually higher, the depth of coverage is also higher, and the resolution, assuming sufficient sequencing depth, is generally higher as well so less abundant copies can be detected.

It is recommended to have some prior knowledge of the relative abundance for the genome targets of interest. As with the other sequencing technologies, it is best to use sufficient inputs in the nanogram range to ensure the detection of targets that may be rare in the sample. When processing sequencing data, it is important to have enough 'genome equivalents' to reliably detect low frequency variants (Table 3). A minimum of 10 target copies in a human DNA input sample is recommended and may also require a higher sequencing depth to distinguish between a sequencing error and a true rare variant.

Table 3. Relationship of human DNA input quantity and variant detection

Quantity of human DNA (ng)	Number of genome equivalents	Number of copies in the sample		
		10% variant	1% variant	0.1% variant
1	330	33	3.3	0.3
10	3300	330	33	3.3
20	6600	660	66	6.6
40	13,200	1320	132	13.2

For RNA samples, using the highest available input leads to higher complexity libraries and detection of low-abundance transcripts, especially since the bulk of the sample is rRNA and will be selectively removed during mRNA enrichment, ribodepletion, or hybridization capture targeted enrichment.

Similarly, for methylation sequencing, bisulfite conversion reduces sample recovery, so using the highest input quantity available produces higher complexity libraries and lower PCR duplicate rates. For ChIP and chromatin studies, where the copy number input into enrichment determines the output copy number for the enriched genome targets of interest, estimating the number of targets can help determine the required input material.

Fragmentation

For many workflows, fragmentation of the nucleic acid sample is required to support short read sequencing instruments. Typically, samples are fragmented to produce as little size variation as possible. The average length for a short read sequencing library can vary between 150–550 bases. Depending on the sequencing instrument, different **read lengths** are available to support different insert sizes. Libraries can also be prepared for either **single-end** (SE) or **paired-end** (PE) sequencing. Single-end sequencing means that DNA is sequenced in a single direction, while paired-end sequencing involves sequencing the library fragments from both directions and aligning/merging the reads bioinformatically post-sequencing. PE sequencing generates longer assembled reads and is particularly useful for identifying single nucleotide variants by verifying the sequence alteration from both strands.

The average fragment length depends upon the type of workflow:

- For **DNA library preps followed by direct sequencing or hybridization capture**, fragmentation of the input high molecular weight DNA is required. This can be achieved using Covaris® Adapative Focused Acoustics® Technology or other mechanical means of shearing the DNA prior to the library preparation. Alternatively, there are DNA library preps with integrated enzymatic DNA fragmentation, suitable for lower cost workflows or automated higher throughput workflows since separate consumables and instrumentation are not required (see the xGen DNA Library Prep EZ and EZ UNI Kits). Regardless of the fragmentation method, when handling damaged samples, fragmentation times may need to be reduced to achieve the desired fragment size. Samples that are already fragmented, such as cfDNA, DNA from FFPE tissue, or DNA samples from ChIP, may not require fragmentation.
- For **amplicon sequencing of DNA or RNA**, prior fragmentation is not recommended as the insert size of NGS library molecules is determined by the primer pair distance in the panel design. The fragment size for oncology panels is usually shorter (typically ≤150 bp) for compatibility with FFPE that may be cross-linked and cfDNA that has a fragment size ~165 bp. For genotyping or other applications, panels are designed with an average insert size of ~300 bp or greater. For targeted RNA-seq using amplicon panels, RNA fragmentation is also not required.
- For **RNA-seq workflows including direct sequencing and hybridization capture**, the RNA is typically fragmented before cDNA synthesis. Conveniently, fragmentation modules are integrated into the library prep kits with the reverse transcription reagents. If RNA is severely damaged according to the RIN and DV_{200} score, fragmentation may not be required.
- For **methyl-seq workflows**, bisulfite conversion denatures and fragments DNA, but high molecular weight DNA is not reduced to a size sufficient for Illumina and other short read sequencing. Before bisulfite treatment, fragmentation using Covaris or other mechanical means of shearing the DNA is required. Note, enzymatic fragmentation modules are not compatible with methylation applications.

ADAPTER AND INDEXING OPTIONS

For DNA library prep, after nucleic acid extraction and mechanical or enzymatic DNA fragmentation, the dsDNA fragments are end repaired and polished to prepare the termini for adapter ligation. End polishing blunts the termini of each template molecule by using DNA polymerases to fill in 5′ **overhangs** and digest 3′ overhangs. For workflows utilizing T-overhang adapters, an additional polymerase is used to A-tail the 3′ ends of template DNA for compatibility with these adapters and to avoid chimera formation (ligation of two insert fragments into a single library molecule). Other dsDNA workflows use alternative adapters with proprietary chemistries and sequential adapter ligation to avoid chimera formation and adapter dimers (e.g., xGen cfDNA and FFPE DNA Library Kit). Alternatively, the Adaptase technology offered by IDT streamlines this process by simultaneously performing end repair, tailing, and ligation of the first NGS adapter to the 3′ single-stranded **DNA fragments**. An extension step then converts the single-stranded fragment into a double-stranded library fragment, followed by ligation of the second adapter. After this step, the termini are compatible for ligation of adapters onto each molecule, which allows them to be clustered and sequenced on the instrument. Each NGS instrument supplier has unique requirements for the adapters, and there are also different ways to index the samples. For this guide, the information will focus on Illumina sequencers, but IDT has experience with other sequencing instruments. For help with your specific library and instrument, please contact us.

In the schematic of a typical Illumina-based library molecule, the original fragment is shown as the insert (Figure 9).

- The SP1 and SP2 sequences are used as priming sites for sequencing by synthesis. Since SP1 and SP2 flank the insert, sequencing can be done from either side (paired-end sequencing) which results in read 1 and read 2 from each insert.
- The i5 and i7 portions of the adapter contain the **indexes** (also called barcodes) that uniquely tag all library molecules from each sample and enable multiplexed sequencing.
- Each index consists of 8–10 bases and are either combinatorial or unique dual indexes (see description below).
- The P5 and P7 sequences of the 5′ and 3′ adapter are used for PCR amplification of libraries, as well as annealing to the flow cell and bridge amplification to create clusters.

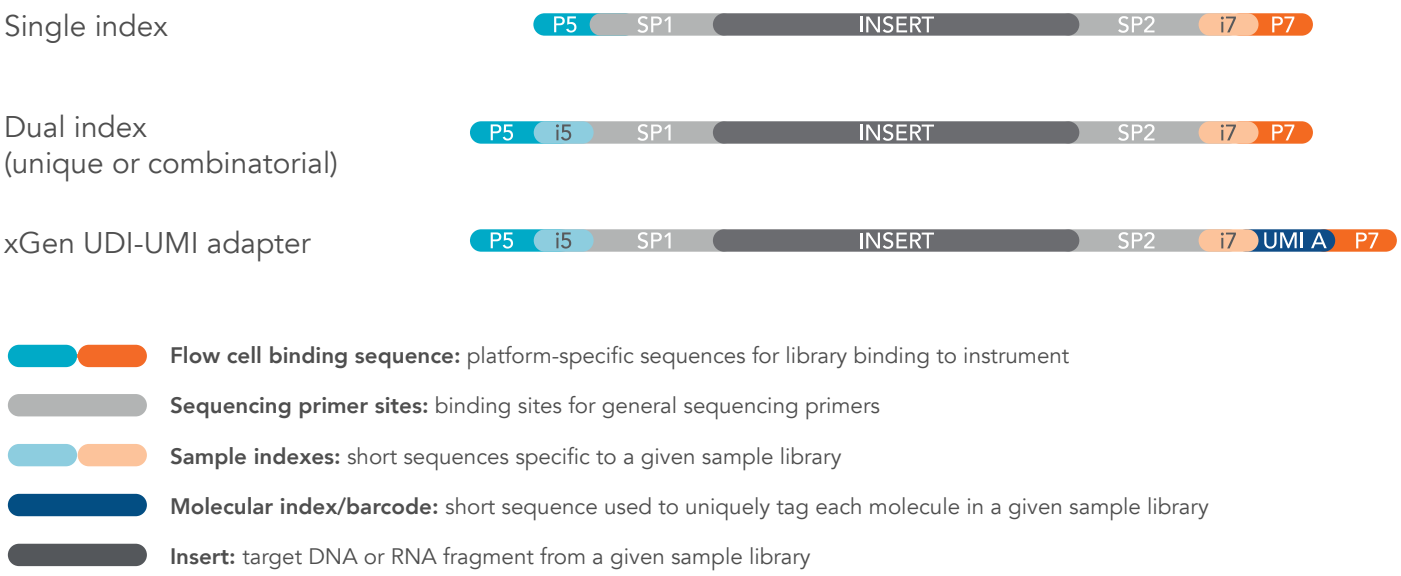


Figure 9. Examples of adapter designs. A variety of NGS adapter designs are available from IDT. When selecting or designing adapters, consideration must be given to the application, multiplexing needs, specification requirements, and analysis methods.

If you're using one of the xGen DNA Library Prep kits, adapters can be added to the insert using one of two strategies (Figure 10). One option is indexing by ligation where full-length indexed adapters are ligated to each fragment end, followed by PCR-free sequencing or an optional library amplification step using terminal P5 and P7 primers (Figure 10A). A second option is indexing by PCR, where first the stubby adapters containing only SP1 and SP2 are attached to each fragment end, followed by an indexing PCR during which library amplification incorporates the indexes and P5 and P7 sequences with 5′ tailed indexing primers (Figure 10B).

Most workflows utilize indexing by PCR (Figure 10B) since it is generally easier and results in lower costs to enable high multiplexing up to 1536-plex. Additionally, stubby adapter ligation followed by indexing PCR makes it easier to remove any adapter dimers that may form since size selection by bead cleanup removes the stubby adapter dimers more effectively than the longer full-length adapter dimers. For these reasons, most indexing options available from IDT are offered as primers.

Full-length indexed adapters with both unique dual indexes (UDIs) and unique molecular identifiers (UMIs) are available for the xGen DNA Library Prep Kit EZ UNI and xGen DNA Library Prep Kit MC UNI. Full-length adapters are required for incorporation of UMIs located in the index position, as they uniquely tag each molecule within a sample for later error correction during data analysis.

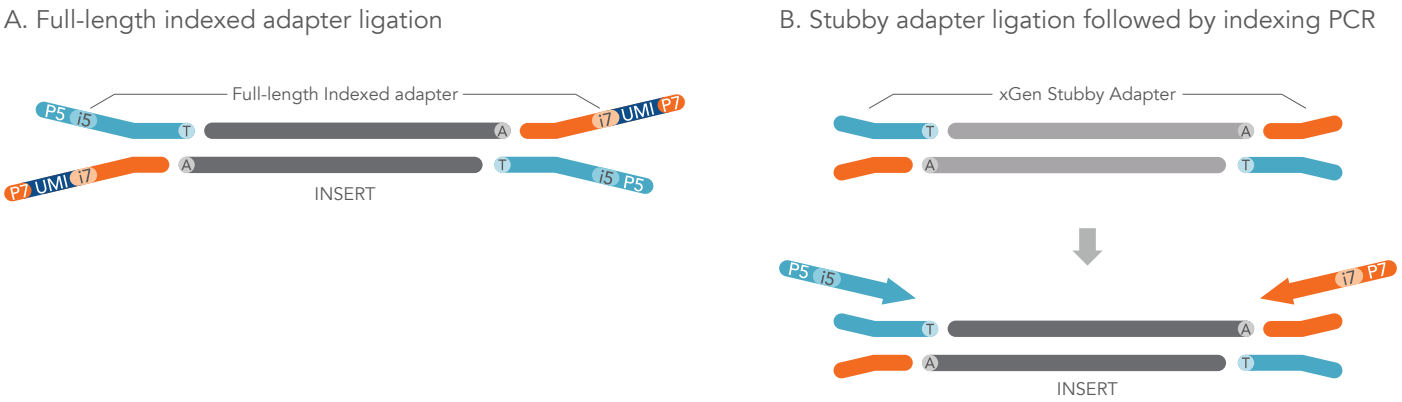


Figure 10. DNA library prep kit adapter ligation options. Adapters can be attached to double-stranded DNA by direct ligation of (**A**) full-length adapters or (**B**) stubby adapters. Indexing by PCR is done to add the remaining P5, P7, i5, and i7 sequences to the stubby adapters.

Single-stranded ligations are used to add adapters to the single-stranded input DNA or first strand cDNA from RNA fragments in the xGen ssDNA & Low-Input DNA Library Prep Kit, xGen RNA Library Prep Kit, xGen Broad-Range RNA Library Prep Kit, and the xGen Methyl-Seq DNA Library Prep Kit. IDT offers a unique single-stranded method called Adaptase technology that attaches the 5′ end of the first adapter to the 3′ tail of the ssDNA, bisulfite-converted DNA, or cDNA fragment. An extension step then converts the single-stranded fragment into a double-stranded library fragment. The second adapter is ligated as a single-stranded R1 Stubby Adapter onto the extension product. The full-length primers are then generated with indexing primers during the indexing PCR amplification (Figure 11A).

Since cfDNA and FFPE DNA samples typically have lower input amounts that are often degraded, the adapter ligation strategy for the xGen cfDNA & FFPE DNA Library Prep Kit (formally called Prism DNA Library Prep Kit) includes unique sequential adapter ligation chemistries to each dsDNA strand. The blend of a novel ligase and proprietary adapter modifications provides a high library conversion efficiency that enables identification of low-frequency, somatic variants, inherited germline SNPs, or perform whole genome sequencing from degraded samples. The Ligation 1 enzyme and adapters are modified to directly attach the 5′ end of the adapter to the 3′ end of the insert. Modified Ligation 1 enzyme is unable to catalyze the ligation of inserts, which reduces chimera formation (when two inserts ligate together and then Read 1 and 2 map to different genome positions). In addition, the 3′ ends of the adapters are blocked to prevent the formation of adapter dimers. In Ligation 2, the Ligation 2 Adapter primes the gap-fill process across the UMI followed by a 5′ ligations. Finally, the PCR step using the **xGen UDI Primer Pairs** creates the final double-stranded library inserts (Figure 11B).

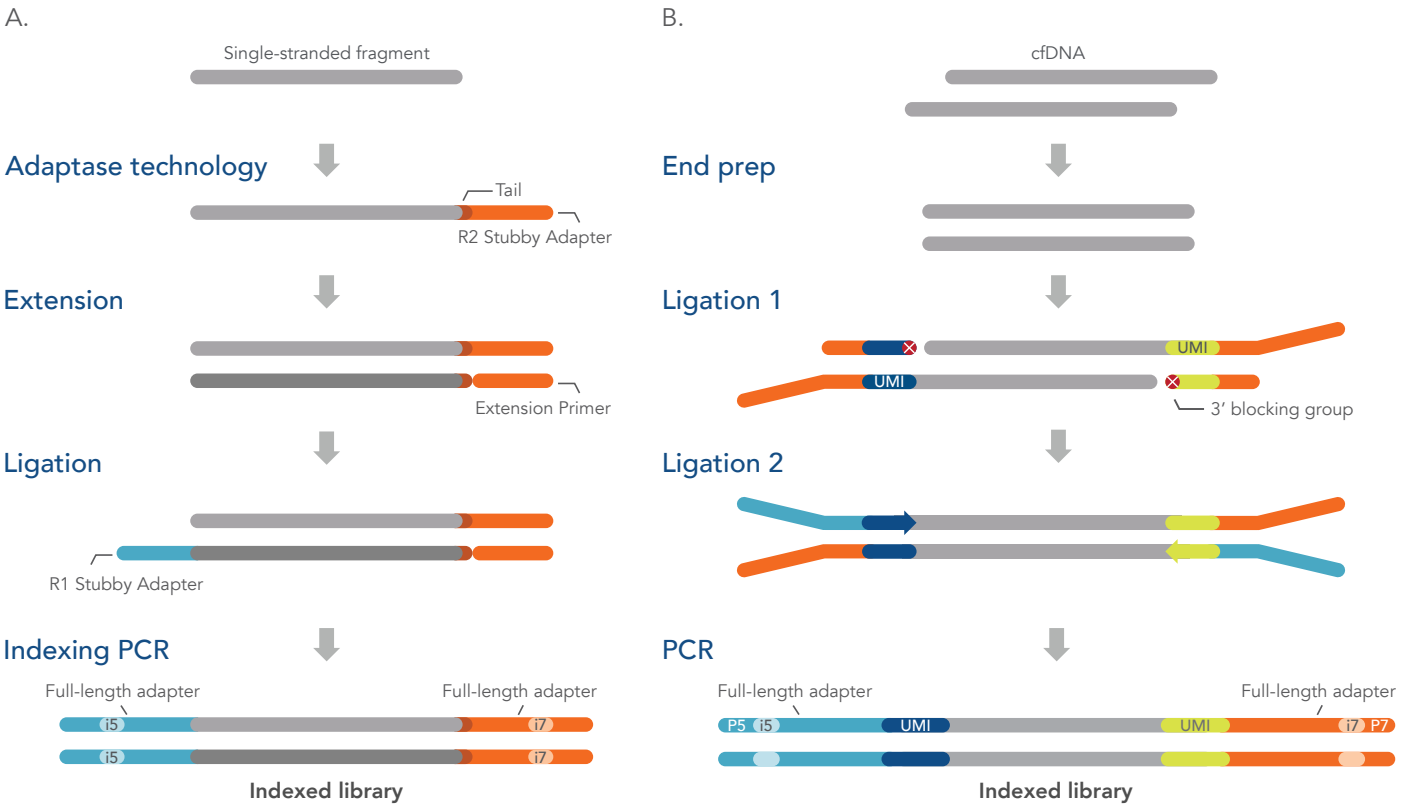


Figure 11. Adaptase technology includes a proprietary single-stranded adapter ligation. (**A**) Adaptase technology ligates the R2 Stubby Adapter to single-stranded DNA, bisulfite-converted ssDNA, or cDNA fragments with a propriety methodology. Extension creates the second strand, where the R2 Stubby Adapter is added (note: the xGen ssDNA & Low-Input DNA Library Prep Kit ligates the R1 Stubby Adapter onto the original single-stranded fragment instead of the extension product). During indexing PCR, both the R2 Stubby Adapter and R1 Stubby Adapter are converted into full-length double-stranded adapters. (**B**) The workflow for the xGen cfDNA & FFPE DNA Library Prep Kit begins with an end repair that creates blunt-ended DNA. Ligation 1 enzyme then catalyzes the single-stranded addition of Ligation 1 Adapter to the 3′ ends of the insert. This novel enzyme is unable to ligate inserts together, which minimizes chimera formation. The 3′ end of the Ligation 1 Adapter include a blocking group to prevent adapter-dimer formation. The Ligation 2 Adapter acts as a primer to gap-fill bases complementary to the Ligation 1 Adapter, followed by ligation to the 5′ end of the DNA insert, to create a double-stranded product. In the final step, PCR incorporates sample index sequences for sequencing on Illumina platforms.

The choice of indexing is also dependent on two additional factors: the number of samples to be co-sequenced, and whether there is concern for index hopping and other events that may reduce fidelity of sample demultiplexing. For the highest fidelity and multiplexed capacity, IDT offers 1536-plex unique dual indexing primers for both standard library quantification workflows and workflows that support Normalase enzymatic library normalization. For convenience, these primers are provided in ready-to-use, single-use 96-well plates. UDIs are unique dual indexes where each i5 and i7 index is used only once, such that if index read misassignment occurs due to index hopping on patterned flow cells, the read is discarded instead of being misassigned to the incorrect sample. For less sensitive sequencing applications, combinatorial dual indexes (CDIs) can be used. These sets utilize a small set of indexes, such as eight i5 by twelve i7, in a combinatorial fashion to produce 96 distinct combinations, but since each index is used more than once in multiplexed sequencing, not all index hopping errors can be filtered out (Figure 12 and 13).

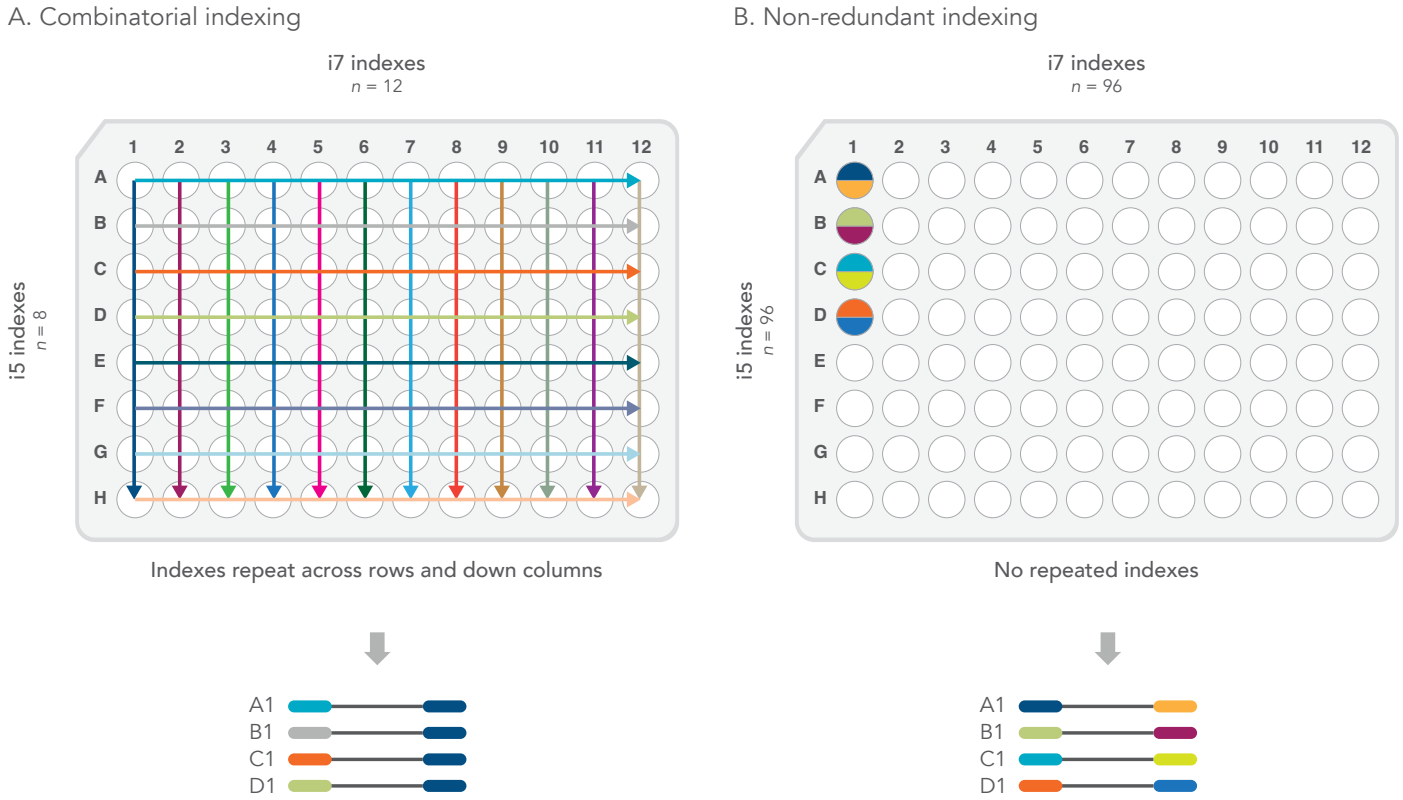


Figure 12. Comparison of combinatorial vs. unique dual indexing. (**A**) Combinatorial indexing (CDIs) will have indexes shared such that i5 will repeat across rows and i7 down columns. (**B**) With non-redundant or unique dual indexes (UDIs) there will not be any indexes repeated or shared on a plate, which prevents index hopping.

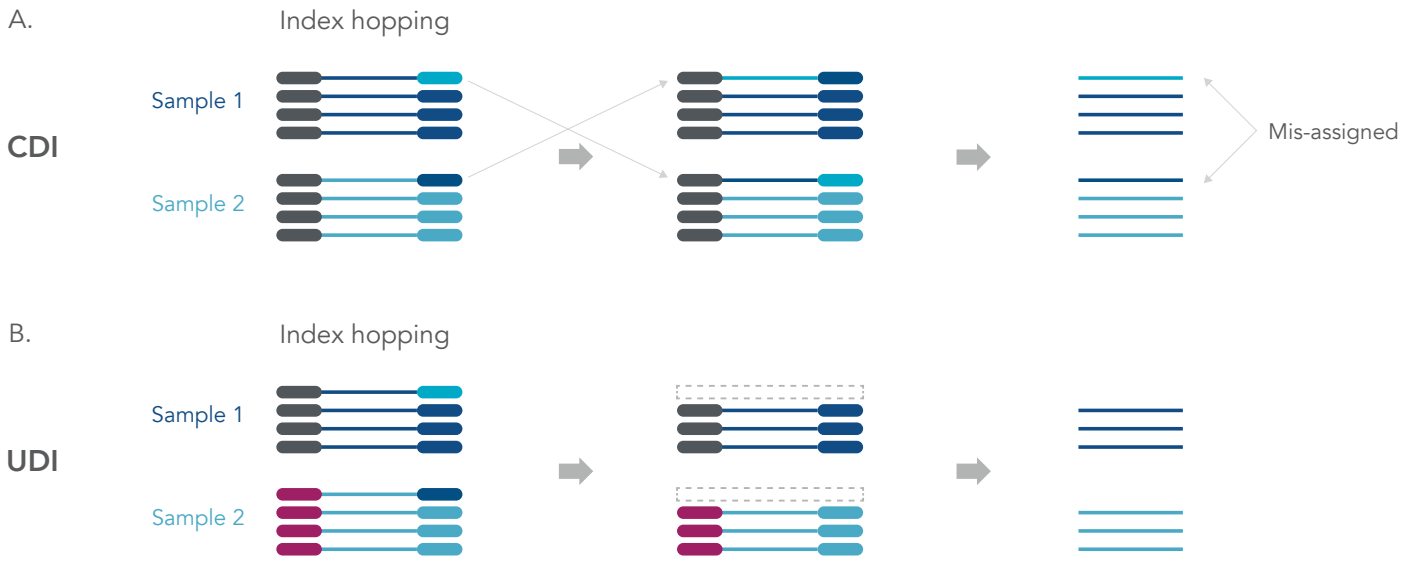


Figure 13. Index hopping. (**A**) When index hopping occurs for libraries constructed with combinatorial dual indexing strategy, the reads will be misassigned to another sample resulting in incorrect sequencing data. (**B**) In contrast, use of unique dual indexes flags misassigned reads as undetermined if index hopping occurs and excludes them from analysis, providing a more reliable sequencing platform.

Finally, adapters may incorporate Unique Molecular Identifiers (UMIs) to achieve error correction during deep sequencing for low frequency variant calling. They also improve fidelity by removing PCR duplicates for better data retention (refer to this application note for more details).

LIBRARY QUANTIFICATION AND THE NORMALASE™ WORKFLOW

Once library construction has been completed, many applications mix different libraries and use the index sequences to demultiplex the data after the sequencing run. Multiplexed libraries can be pooled directly, but if the libraries have different concentrations, the data will be non-uniform and may miss essential information from the library with the lowest concentration in the pool. Normalization, that is, ensuring each library has equal representation in the pool, is a key step to get equal distribution of data (Figure 14).

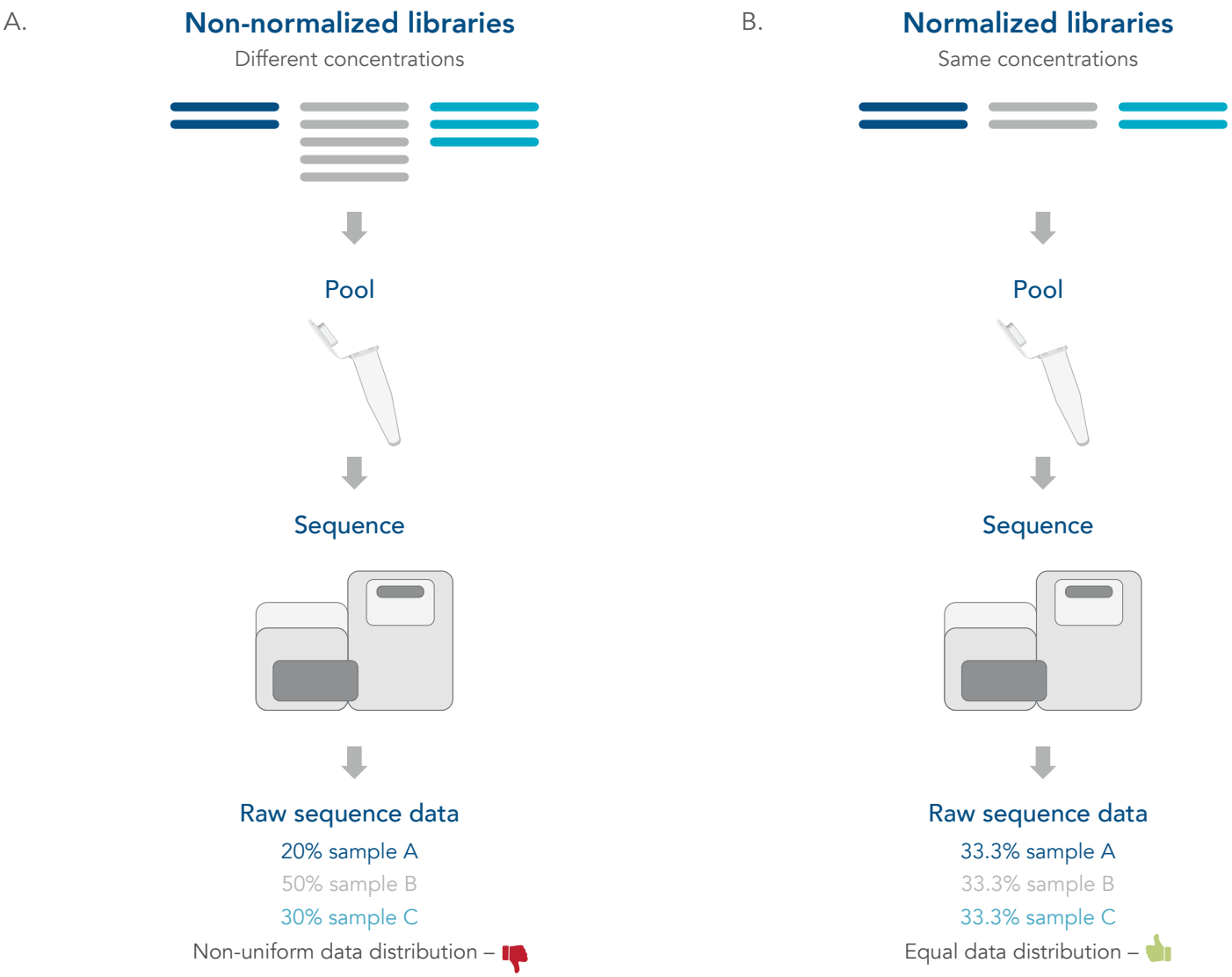


Figure 14. Library normalization. (**A**) Non-normalized libraries lead to non-uniform data, whereas (**B**) normalized libraries result in equal data distribution.

There are two methods for creating equimolar multiplex library pools (Figure 15). The first method includes library quantification followed by equal molar pooling. The second uses the IDT proprietary enzymatic normalization method found in the xGen Normalase Module. The workflow of choice depends mainly on the library prep and sample types used. The Normalase workflow requires library amplification and is suited for consistent sample types when direct sequencing is used, whereas traditional library quantification and pooling is used for PCR-free library prep workflows and workflows suitable for difficult samples with less consistent performance.

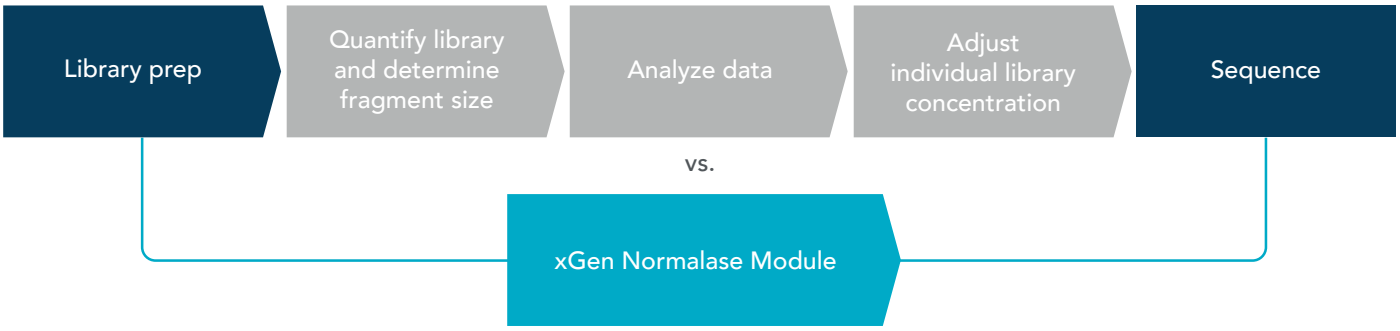


Figure 15. Normalase vs library quantification. Use of the xGen Normalase Module eliminates the need to normalize sample concentrations before pooling libraries, resulting in faster turnaround times.

PCR-free libraries cannot be visualized by a BioAnalyzer or TapeStation because the native electrophoretic conditions produce migration artifacts and appear to run 'larger' than expected due to the ssDNA adapter tails at each terminus. PCR-free sequencing workflows also have a mixture of fully adapter ligated library molecules as well as partially adapter ligated library molecules; therefore, qPCR quantification is the recommended method that can accurately distinguish fully from partially ligated library molecules (qPCR library quantification kits are available from several suppliers).

For workflows that end with a library amplification step, PCR enriches for the fully adapter ligated library molecules so a variety of quantification methods can be used, including qPCR. Qubit or other methods that measure dsDNA content, as well as electrophoretic analysis to determine library insert size including BioAnalayzer and TapeStation are options that provide interpretable data. Following quantification where the molarity of each sample is determined based on library size and nucleic acid quantity, each library concentration needs to be individually adjusted to generate an equimolar library pool before sequencing.

In contrast, the Normalase workflow is best used for direct sequencing applications, including WGS and shotgun metagenomics to investigate the composition of complex samples, as well as whole transcriptome sequencing or RNA-seq. Often these studies have a high volume of samples with specific sequencing depth requirements, where data quality, cost, and turnaround time are also important factors.

Combining the xGen Normalase Module with xGen Library Prep Kits is an efficient solution to support such applications. Normalase technology is an enzymatic library normalization chemistry that enables more balanced multiplexed sequencing pools with a coefficient of variation (CV) <10%, outperforming pooling by qPCR (Figure 16). The process not only streamlines library normalization by omitting individual sample quantification and different per sample pooling volumes; it also performs simple equal volume pooling.

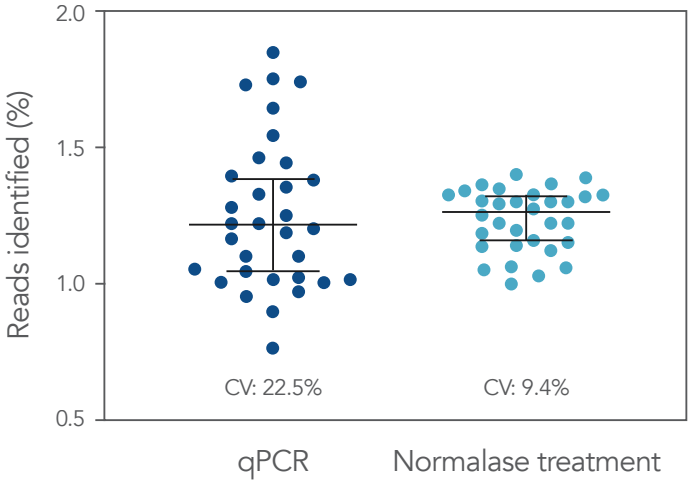


Figure 16. qPCR vs. Normalase enzymatic normalization. Thirty-two xGen DNA libraries were generated with full-length indexed adapters between two users (*n* = 16/user) with 1–250 ng inputs of NA12878 gDNA. Post-Normalase PCR libraries were quantified with qPCR assuring the libraries met the minimum threshold. Libraries were either normalized, pooled, and sequenced based on the qPCR quantification or, using the same libraries, pooled and normalized using the xGen Normalase Module and sequenced to determine percent reads identified of each index (MiSeq). The coefficient of variation (CV) for the qPCR pool was 22.5% across the two users, while the CV for the Normalase pool was 9.4%. Lines are median and 95% confidence interval.

The streamlined workflow, combined with more uniform sample sequencing depth, saves time plus results in fewer discarded reads when subsampling the data. This leads to more efficient use of sequencing data and lower costs. The xGen Normalase Module can be combined with IDT's xGen DNA Library Prep Kits (both enzymatic and mechanical fragmentation permutations), as well as the xGen RNA Library Prep Kit and the xGen Broad-Range RNA Library Prep Kit. Since the Normalase workflow is suitable for sample types and library prep workflows with reproducible performance, it is recommended for routine samples that produce consistent results.

GLOSSARY

16S rRNA: a gene coding for the small subunit (SSU) of the prokaryotic ribosome. Some regions of this gene are highly conserved across all prokaryotes, while other regions are variable and are often used for taxonomic identification of bacteria and archaea.

Adapter: a short nucleotide sequence that is attached to an index to bind it to a surface (chip/bead) in an NGS sequencer. Adapter sequences are typically specific to the sequencing platform.

Amplicon: a replicated DNA fragment and the end-product of the polymerase chain reaction (PCR).

Coverage (sequencing coverage): the average number of sequencing reads that align to (cover) a specific target region.

***De novo* assembly**: a sequence assembly technique where a genome is assembled by overlapping sequencing reads without the use of a known reference sequence.

DNA fragment: a small DNA piece—usually 150–300 base pairs (bp)—produced by a mechanical or chemical shearing (fragmentation) of larger DNA molecules.

Indel: stands for "INsertions and DELetions" and refers to a genetic variation caused by the insertion or deletion of a nucleotide or nucleotides in one DNA molecule when compared to another DNA molecule (typically a reference sequence).

Index: a known short nucleotide sequence added to DNA samples during library preparation that acts as a unique identifier and allows multiple libraries to be pooled and sequenced together.

Library (sequencing library): a collection of DNA fragments prepared for sequencing.

Metagenomics: an untargeted NGS approach that provides access to the full genetic content in each sample (the resulting data include fragmented sequences from all of the DNA present in the sample). This method allows researchers to study microbial diversity in different natural and artificial environments.

Next generation sequencing (NGS): a massively parallel sequencing technology that decodes millions to billions of DNA or RNA (via cDNA) molecules from hundreds of samples simultaneously on a single sequencing run. This makes NGS a high-throughput, fast, and cost-effective method for omics (e.g., genomics, transcriptomics) research.

Overhang: a DNA segment tethered to the end of a PCR primer which facilitates adapter binding.

Paired-end read (PE): a sequencing chemistry that obtains sequence reads from both ends of a DNA fragment template. This technology allows contigs to be joined when they contain read pairs from a single template fragment, even if no reads overlap.

Read length: the length of the DNA segment that is sequenced in a sequencing reaction.

Reference sequence (genome): a curated set of sequencing data (e.g., a genome) that is used as a reference for post-sequencing computational analysis.

Sequence assembly: computational technique of overlapping identical (or nearly identical) nucleotide sequences from a sequencing run and iteratively merging (assembling) them together to form longer sequences (contigs).

Sequence (sequencing) read: the DNA sequence of base pairs from one DNA fragment (a small section of DNA). Sequence reads produced by NGS machines are often referred to as short reads.

Sequence variants: nucleotide differences at specific positions between two aligned sequences. They include single nucleotide polymorphisms (SNPs), insertions and deletions, copy number variants, and structural rearrangements.

Sequencing by synthesis: the term used by Illumina to describe the chemistry used in its sequencing machines (e.g., MiSeq, HiSeq™).

Sequencing depth (read depth): the number of times a specific nucleotide is sequenced. The deeper the sequencing depth is, the more precise the base calling is.

Single-end read (SE): a sequencing chemistry that obtains sequence reads from only one end of the DNA fragment template.

Targeted sequencing: a protocol where only specific (targeted) regions of the genome are sequenced.

Whole genome sequencing (WGS): a protocol where the entire genome is sequenced.

Whole exome sequencing (WES): a protocol where only the exome (protein coding component of the genome) is sequenced.

Variant calling: the process by which sequence variants are identified (called) by aligning sequencing reads to a reference sequence and then identifying differences between the reads and the reference.

REFERENCES

1. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59.
2. Luo C, Rivkin A, Zhou J, *et al*. Robust single-cell DNA methylome profiling with snmc-seq2. *Nat Commun*. 2018;9(1):3824.
3. Luo C, Keown CL, Kurihara L, *et al*. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*. 2017;357(6351):600-604.

RECOMMENDED READING

1. Clark DP, Pazdernik NJ. *Molecular biology*. Second edition. ed. Waltham, MA: Academic Press; 2013.
2. Hu T, Chitnis N, Monos D, *et al*. Next-generation sequencing technologies: An overview. *Hum Immunol*. 2021;82(11):801-811.

Next generation sequencing guide

For more information, go to: **www.idtdna.com/ContactUs**

> For more than 30 years, IDT's innovative tools and solutions for genomics applications have been driving advances that inspire scientists to dream big and achieve their next breakthroughs. IDT develops, manufactures, and markets nucleic acid products that support the life sciences industry in the areas of academic and commercial research, agriculture, medical diagnostics, and pharmaceutical development. We have a global reach with personalized customer service.

> SEE WHAT MORE WE CAN DO FOR **YOU** AT **WWW.IDTDNA.COM**.

Doc ID: RUO22-0782_001 06/23

INTEGRATED DNA TECHNOLOGIES