

Tail trimming for better data

For use with:

- xGen™ Methyl-Seq DNA Library Prep Kit
- xGen Adaptase™ Module
- xGen ssDNA & Low-Input DNA Library Prep Kit
- xGen RNA Library Prep Kit
- xGen Broad-Range RNA Library Prep Kit

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

custom oligos • qPCR • next generation sequencing • RNAi • genes & gene fragments • CRISPR genome editing

Table of contents

Introduction	3
xGen Adaptase Technology	4
xGen Library Prep Kits with Adaptase Technology	5
xGen Methyl-Seq DNA Library Prep Kit	5
xGen Adaptase Module	6
xGen ssDNA & Low-Input DNA Library Prep Kit	6
xGen Broad-Range RNA Library Prep Kit and xGen RNA Library Prep Kit	7
Sequencing recommendations	8
References	9

INTRODUCTION

The IDT **xGen Methyl-Seq Library Prep Kit**, **xGen ssDNA & Low-Input DNA Library Prep Kit**, **xGen Broad-Range RNA Library Prep Kit**, **xGen RNA Library Prep Kit**, and **xGen Adaptase Module** utilize the innovative xGen Adaptase technology to deliver high quality NGS data from various sample types and input amounts. This unique technology enables the construction of libraries from single- or double-stranded DNA in a single sample as well as from single cells. The xGen Adaptase technology is a highly efficient, template-independent reaction that performs ligation of adapters to 3' ends of ssDNA fragments. This results in maximum recovery of input DNA, even from samples with denatured or heavily nicked DNA.

The Adaptase technology workflow produces directional libraries as the Adaptase reaction adds a low complexity polynucleotide tail with a median length of 8 bases to the 3' end of each ssDNA fragment during the addition of the R2 (Read 2) Stubby Adapter and the R1 (Read 1) Stubby Adapter is always attached to the 5' ends of the fragments (**Figure 1**). In the case of ssDNA genomes and first strand cDNA substrates, the libraries produced with the Adaptase technology are directional and stranded since a single functional library strand is produced from each fragment. It is therefore normal and expected to observe the Adaptase tail at the beginning of R2. If the sequencing read length is close to library insert size, the tail may also be observed toward the end of Read 1 (R1). These tails, if untrimmed, will affect mapping rates and bisulfite conversion rate calculations.

This technical note describes the nature of these Adaptase tails and provides guidance on when and how they should be trimmed bioinformatically for our specific library prep kits, including:

- **xGen Methyl-Seq DNA Library Prep Kit**
- **xGen Adaptase Module**
- **xGen ssDNA & Low-Input DNA Library Prep Kit**
- **xGen Broad-Range RNA Library Prep Kit**
- **xGen RNA Library Prep Kit**

Note: Quality control software, such as FastQC [1] may raise “Per base sequence content” or “Per base GC content” flags at the beginning of R2. These flags are expected due to the low complexity tail (**Figure 2**).

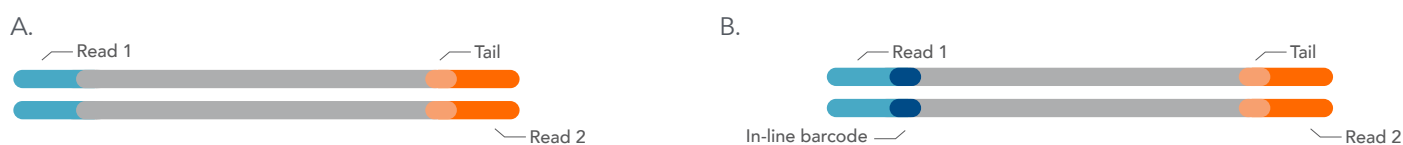


Figure 1. Schematic of Adaptase library fragments following indexing PCR. (A) For the xGen Methyl-Seq DNA Library Prep Kit, xGen ssDNA & Low-Input DNA Library Prep Kit, xGen Broad-Range RNA Library Prep Kit and xGen RNA Library Prep Kit, the Adaptase tail is located at the beginning of Read 2. (B) For single cell sequencing applications like single-nucleus methylcytosine sequencing (snmC-seq) with the xGen Adaptase Module an in-line barcode is located at the beginning of Read 1 and the Adaptase tail is located at the beginning of Read 2.

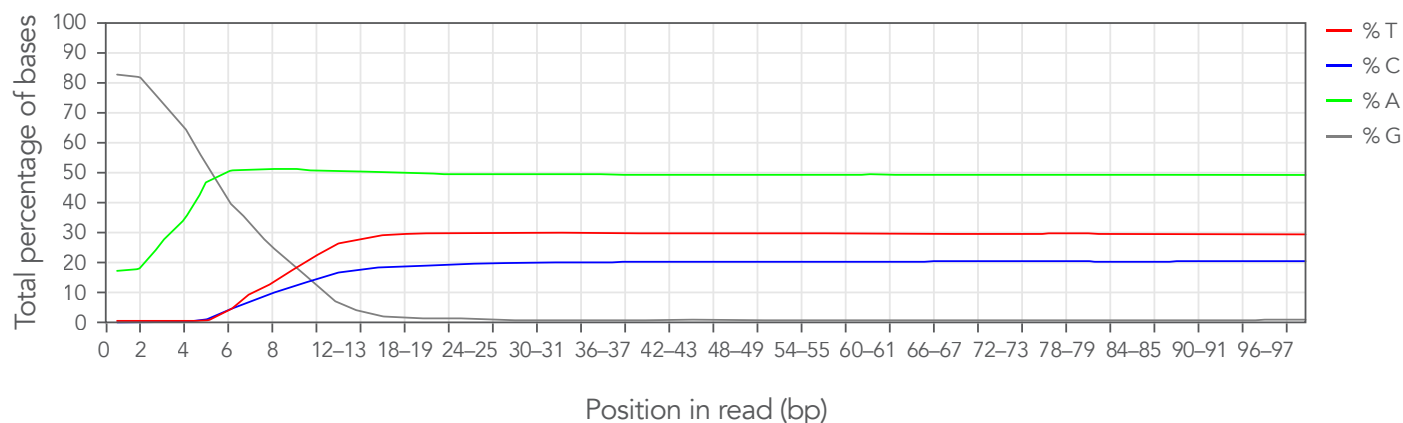


Figure 2. FastQC per base sequence content for Read 2 showing the G-rich Adaptase tail.


For additional tail trimming recommendations, contact IDT Technical Support at custcare@idtdna.com.

xGEN ADAPTASE TECHNOLOGY

The IDT xGen Adaptase technology is integrated into the workflow of many of the NGS library preparation kits sold at IDT including:

- [xGen Methyl-Seq DNA Library Prep Kit](#)
- [xGen Adaptase Module](#)
- [xGen ssDNA & Low-Input DNA Library Prep Kit](#)
- [xGen Broad-Range RNA Library Prep Kit](#)
- [xGen RNA Library Prep Kit](#)

The incorporation of the Adaptase technology into a library preparation introduces additional synthetic sequences that may need to be removed before downstream analyses. Trimming of synthetic sequences is required to obtain improved mapping efficiency (with tools like BSMAP) because some aligners cannot soft clip synthetic sequence content that interfere with mapping. Random primer barcode and tail trimming also remove synthetic unmethylated cytosines and helps achieve accurate methylation and bisulfite conversion information for the xGen Methyl-Seq DNA Library Prep Kit and Adaptase Module. Many informatics pipelines for methylation analysis already include trimming of bases from the beginning of both R1 and R2 to eliminate any synthetic cytosine methylation introduced as a result of filling in overhangs during end repair/polishing steps of conventional dsDNA library preparation. Therefore, this trimming step can be adjusted to remove random primer barcode sequences and Adaptase tails.

 **Note:** If using in-line barcodes for a the single-nucleus methylcytosine sequencing (snmC-seq) workflow with the xGen Adaptase Module, be sure to capture the in-line index information for proper single cell demultiplexing.

xGEN LIBRARY PREP KITS WITH ADAPTASE TECHNOLOGY

xGen Methyl-Seq DNA Library Prep Kit

The IDT **xGen Methyl-Seq DNA Library Prep Kit** can be used for whole genome and targeted methylation sequencing. The synthetic tail sequence added to 3' termini during the Adaptase reaction includes unmethylated cytosines. Therefore, the tail adds both synthetic sequence and methylation information to the beginning of R2. The Adaptase tails can be seen as a G-rich spike at the start of R2, with a median length of 8 bases (**Figure 2**). Trimming of these tails is required for libraries created with the **xGen Methyl-Seq DNA Library Prep Kit** to obtain improved mapping efficiency (with tools like BSMAP [2]) because some aligners cannot soft clip synthetic sequence content which then can interfere with mapping. Tail trimming also removes synthetic unmethylated cytosines and helps with accurate methylation and bisulfite conversion information.

Common analysis workflows for methyl-seq include trimming of bases from the beginning of both R1 and R2 to eliminate any synthetic cytosine methylation introduced as a result of filling in overhangs during end repair/polishing steps of conventional dsDNA library preparation. Adaptase tails can also be removed by an adjustment during this trimming step.

Following standard adapter trimming, trim the recommended number of bases (**Table 1**) from the 5' start of R2 and the 3' end of R1 to eliminate the majority of Adaptase tail sequence. Publicly available tools such as Trimmomatic [3], fastx_trimmer [4], or Trim Galore [5] may be used as part of the sequence data processing pipeline.

! **Important:** Some downstream differential methylation analysis software packages require that both paired end (PE) reads are the same length, therefore the same number of base pairs (bp) will need to be trimmed from the start of R2, as well as the end of R1.

For paired-end (PE) reads, trim 10 bases from the end of R1 (3' end) and 10 bases from the beginning of R2 (5' end) to remove tail sequences that may be encountered at these sites (**Table 1**). If symmetric read lengths are not required, and the R1 length does not approach the library insert size, trimming of Adaptase tails from the end of R1 (3' end) may not be necessary. For example, tails are unlikely to be encountered at the end of R1 with PE75 or PE100 on a 170 bp insert library but are likely to be encountered if performing PE150. These recommendations will help achieve >70% aligned bisulfite reads using BSMAP [2]. To achieve similarly high mapping rates with Bismark [6] we recommend following parameters bowtie2 L,-0.6,-0.6, --directional.

☰ **Note:** Illumina® adapter trimming must be performed before Adaptase tail trimming.

Table 1. Summary of xGen Methyl-Seq DNA Library Prep Kit tail trimming recommendations.

Mapped insert size	Read 1 trimming (optional)	Read 2 trimming
~170 bp	10 bases from END of R1	10 bases from START of R2

* For ease of analysis, trimming 10 bp from 5' of both reads in a single Trimmomatic command yields similar results for most insert sizes.

xGen Adaptase Module

IDT offers the **xGen Adaptase Module** to support the single-cell sequencing approaches like the single-nucleus methylcytosine sequencing (snmC-seq) protocol described in [7] and [8]. Single-cell methyl-seq NGS libraries constructed using the xGen Adaptase Module include a synthetic 8 base random sequence at the beginning of R2 (**Figure 2**). Additionally, the workflow optionally incorporates a 6–8 base in-line barcode 5' to the random primer to enable a three-dimensional indexing strategy, where the barcode is positioned at the beginning of R1 (see the **xGen Adaptase Module protocol**). Both of these synthetic sequences, along with the Adaptase tail sequence added to 3' termini, must be trimmed. The Adaptase tails can be seen as a G-rich spike at the start of R2, with a median length of 8 bases (**Figure 2**).

Given that the random primer and barcode are ~15 bases in length and the median Adaptase tail length is 8 bases, trim 15 bases from the beginning of R1 (5' end), and 15 bases from the beginning of R2 (5' end) in order to remove the synthetic sequences, present at the beginning of both reads (see **Table 2**). Because library insert size is >400 bp when following the recommended workflow, it is not necessary to trim reads at the end of R1 (3' end) as Adaptase tails will not be encountered from a PE150 read length. If performing single end sequencing, trim 15 bases from the beginning of R1 (5' end) to remove the random primer and barcode sequences.

 **Note:** Illumina adapter trimming must be performed before random primer/barcode and Adaptase tail trimming.

Table 2. Tail trimming recommendations for xGen Adaptase Module used in snmC-seq.

Mapped insert size	Read 1 trimming (required)	Read 2 trimming
>400 bp	15 bases from START of R1	15 bases from START of R2

xGen ssDNA & Low-Input DNA Library Prep Kit


Whole genome and other DNA sequencing applications involve paired end alignment to reference genomes using common aligners such as BWA-MEM, BWA-ALN [9], Novoalign [10], SOAP [11], or Bowtie [12], etc. Trimming of Adaptase tails from the beginning of R2 can increase mapping efficiency, as some aligners cannot soft clip synthetic sequence content which can interfere with alignment. Similarly, for metagenomics applications requiring sequence assembly, trimming synthetic Adaptase tails from the beginning of R2 will significantly enhance read assembly into contigs.

Many applications require symmetric lengths for both reads, so make sure to trim the same number of bases from both R1 and R2. For PE reads, trim 10 bases from the end of R1 (3' end), and 10 bases from the beginning of R2 (5' end), to remove tail sequences that may be found at the end of R1 and the tail sequence encountered at the beginning of each R2 (see **Table 3**). If symmetric read lengths are not required, and the R1 length does not approach the library insert size, trimming of Adaptase tails from the end of R1 (3' end) may not be necessary. For example, tails are unlikely to be encountered at the end of R1 with PE150 on a 350 bp insert library but are likely to be encountered on a 170 bp insert cell-free DNA library. If performing single read sequencing, trim 10 bases from the end of R1 if the read length is approaching the library insert size as viewed in the FastQC report figure, "Per Base Sequence Content" (**Figure 2**).

xGen Broad-Range RNA Library Prep Kit and xGen RNA Library Prep Kit

For these library kits, a low-complexity polynucleotide tail with a median length of 8 bases is added to the 3' end of each fragment during the Adaptase step. Therefore, it is normal and expected to observe this tail at the beginning of R2. When read length is close to fragment size, the tail may also be observed toward the end of R1 data.

A commonly used RNAseq aligner, STAR [13], is typically able to soft clip the synthetic Adaptase tail sequence, as well as the synthetic random primer sequence at the beginning of R1 if any mismatches were introduced during the priming step. STAR provides efficient mapping without additional processing of the sequencing data.

 **Note:** If you find that soft-clipping is not sufficient for your particular analysis, we recommend implementing STAR with the following argument: `--clip5pNbases 10`.

For a reciprocal trim, we recommend trimming 10 bases from the beginning of both R1 and R2 if insert size is significantly larger than read length (i.e., 2 x 75 bp for a 250 bp insert library), see Table 3. If insert size is approaching the read length, you may encounter tails at the end of R1 (i.e., 2 x 125 bp for a 250 bp library). In this case, we recommend that you trim 10 bases from the end of R1 and the beginning of R2. Tail and random primer trimming (beginning of R1) can be performed using publicly available tools like Trimmomatic [3], or Cutadapt [14].


 **Note:** Illumina adapter trimming must be performed before adaptase tail trimming.

Table 3. Summary of xGen ssDNA & Low-input DNA Library Prep, xGen Broad-Range RNA Library Prep, and xGen RNA Library Prep kits tail trimming recommendations.

Mapped insert size (bp)	Read 1 trimming (optional)	Read 2 trimming
200, 300, 350	10 bases from END of R1	10 bases from START of R2

SEQUENCING RECOMMENDATIONS

Illumina sequencing chemistry is sensitive to low complexity base composition such as that introduced by bisulfite conversion when using the [xGen Methyl-Seq DNA Library Prep Kit](#) or the [xGen Adaptase Module](#). Follow the recommendations by Illumina for successful sequencing runs on the instrument of choice, where PhiX or any balanced, high sequence complexity library can be spiked in with the bisulfite converted libraries. Since recommendations are specific to the sequencing instrument chemistry and version of sequencer software, contact Illumina technical support for the most up to date recommendations before sequencing. For best results, the insert size should always exceed sequence read length to avoid sequencing into adapters which lowers data quality because of low sequence diversity. When generating read counts for RNA-Seq with a program like HTSeq, we also recommend that you include the flag 'strandedness = reverse' in your command.

REFERENCES

1. Andrews S. **FastQC: A Quality Control Tool for High Throughput Sequence Data**. [Online]. 2010.
2. Xi Y, Li W. **BSMAP: whole genome bisulfite sequence MAPping program**. BMC Bioinformatics. Jul 27 2009;10:232. doi:10.1186/1471-2105-10-232
3. Bolger AM, Lohse M, Usadel B. **Trimmomatic: a flexible trimmer for Illumina sequence data**. Bioinformatics. Aug 1 2014;30(15):2114-20. doi:10.1093/bioinformatics/btu170
4. Hannon G. **FASTX-Toolkit: FASTQ/a short-reads pre-processing tools**. [Online]. 2010;
5. Krueger F. **TrimGalore**. GitHub repository. 2017;doi:https://zenodo.org/badge/latestdoi/62039322
6. Krueger F, Andrews SR. **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**. Bioinformatics. Jun 1 2011;27(11):1571-2. doi:10.1093/bioinformatics/btr167
7. Luo C, Keown CL, Kurihara L, et al. **Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex**. Science. Aug 11 2017;357(6351):600-604. doi:10.1126/science.aan3351
8. Luo C, Rivkin A, Zhou J, et al. **Robust single-cell DNA methylome profiling with snmC-seq2**. Nat Commun. Sep 20 2018;9(1):3824. doi:10.1038/s41467-018-06355-2
9. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. Bioinformatics. Jul 15 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324
10. **Novoalign**. [Online].
11. Li R, Li Y, Kristiansen K, Wang J. **SOAP: short oligonucleotide alignment program**. Bioinformatics. Mar 1 2008;24(5):713-4. doi:10.1093/bioinformatics/btn025
12. Langmead B, Salzberg SL. **Fast gapped-read alignment with Bowtie 2**. Nat Methods. Mar 4 2012;9(4):357-9. doi:10.1038/nmeth.1923
13. Dobin A, Davis CA, Schlesinger F, et al. **STAR: ultrafast universal RNA-seq aligner**. Bioinformatics. Jan 1 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
14. Martin M. **Cutadapt removes adapter sequences from high-throughput sequencing reads**. EMBnetjournal. 2011;17(1):10-12. doi:doi:https://doi.org/10.14806/ej.17.1.200

Tail trimming for better data

Technical support: applicationsupport@idtdna.com

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.

For Research Use Only. Not for use in diagnostic procedures. Unless otherwise agreed to in writing, IDT does not intend these products to be used in clinical applications and does not warrant their fitness or suitability for any clinical diagnostic use. Purchaser is solely responsible for all decisions regarding the use of these products and any associated regulatory or legal obligations.

© 2022 Integrated DNA Technologies, Inc. All rights reserved. Illumina, iSeq 100, MiniSeq, MiSeq, HiSeq 2500, HiSeq 3000/4000 and NovaSeq are registered trademarks of Illumina, Inc. Trademarks contained herein are the property of Integrated DNA Technologies, Inc. or their respective owners. For specific trademark and licensing information, see www.idtdna.com/trademarks. Doc ID: RUO22-0691_001 04/22