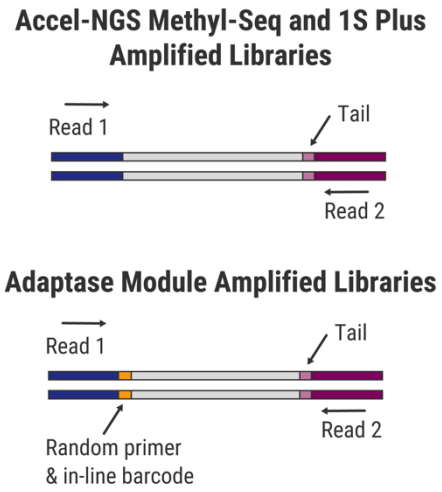
TAIL TRIMMING FOR BETTER DATA: ACCEL-NGS® METHYL-SEQ, ADAPTASE MODULE AND 1S PLUS DNA LIBRARY KITS

INTRODUCTION

Swift Biosciences' Accel-NGS Methyl-Seq, Adaptase Module and 1S Plus kits utilize Swift's innovative 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-stranded as well as double-stranded DNA in a single sample. The Adaptase step is a highly efficient, template-independent reaction that performs ligation of adapters to 3' ends of ssDNA fragments, resulting in maximum recovery of input DNA, suitable for working with denatured or heavily nicked samples.

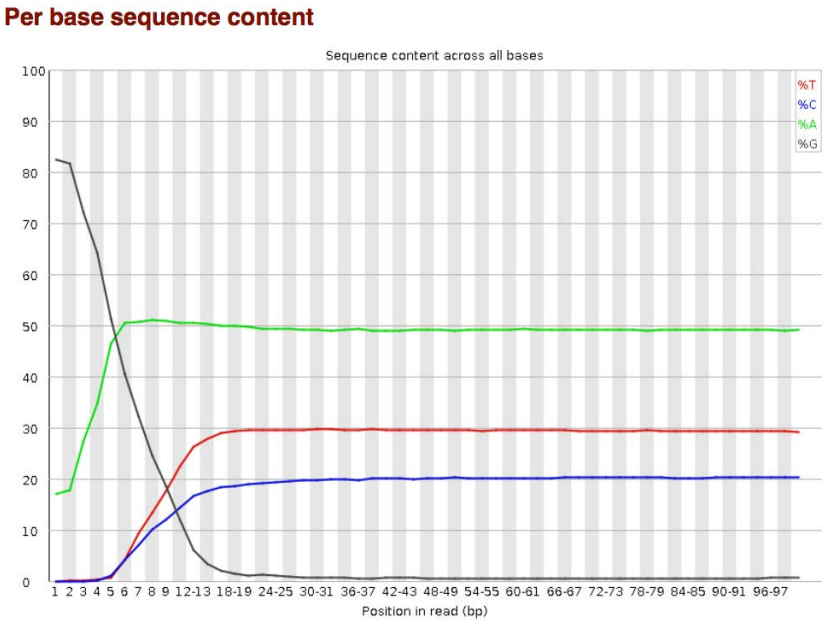
The Adaptase reaction adds a low complexity polynucleotide tail with a median length of 8 bases to the 3' end of each fragment during the addition of the Read 2 Illumina adapter (see Figure 1). It is normal and expected to observe the Adaptase tail at the beginning of Read 2 (R2). When sequencing read 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 the tails and provides guidance for when and how they should be trimmed bioinformatically for the three product lines; Accel-NGS Methyl-Seq, Adaptase Module and Accel-NGS 1S Plus. Please note: Quality control software, such as FastQC FastQC (Babraham Bioinformatics) 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 (see Figure 2).

Figure 1: Schematic of three Swift Adaptase library types



The Adaptase Step simultaneously performs end repair, tailing of 3' ends, and ligation of the Read 2 adapter to 3' ends. The read 1 adapter is located at 5' ends and if using the Adaptase module recommended workflow, an in-line barcode and random primer sequence are present at the beginning of read 1.

Figure 2. FastQC per base sequence content for Read 2 depicts an Adaptase tail.



For additional tail trimming recommendations, please contact Swift Biosciences' Technical Support at technicalsupport@swiftbiosci.com.

ACCEL-NGS METHYL-SEQ LIBRARIES

The synthetic tail sequence added to 3' termini during the Adaptase step 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 Read 2, with a median length of 8 bases (Figure 2). Trimming of these tails is required for Accel-NGS Methyl-Seq libraries to obtain improved mapping efficiency (with tools like BSMAP) because some aligners cannot soft clip synthetic sequence content which can interfere with mapping. Tail trimming also removes synthetic unmethylated cytosines and helps achieve accurate methylation and bisulfite conversion information. Many informatics pipelines for Methyl-Seq analysis 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, an adjustment of the trimming step can be done to also remove the Adaptase tails.

Following adapter trimming, **trim the recommended number of bases (Table 1) from the 5' start of Read 2 (R2) and the 3' end of Read 1 (R1)** to eliminate the majority of Adaptase tail sequence. For longer sequencing chemistries (PE125 or PE150), we find best results when reads are trimmed to a final length of 90-100bp also considering different sample types such as FFPE or cfDNA, which yield smaller fragments/inserts. Trimming length can also be determined/inferred from fastqc plots (Per_Base_Sequence_Content, Figure 2). Publicly available tools such as Trimmomatic (Bolger, et al. 2014 Bioinformatics), fastx_trimmer (http://hannonlab.cshl.edu/fastx_toolkit/) or trimgalore (https://github.com/FelixKrueger/TrimGalore) may be used as part of the sequence data processing pipeline. It is important to note that some downstream differential methylation analysis packages require that both PE reads be of same length, hence trim the same number of base pairs (bp) from start of Read2 and end of Read1.

For paired end reads, trim 15 bases from the end of R1 (3' end) and 15 bases from the beginning of R2 (5' end) to remove tail sequences that may be encountered at at these sites (see 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 should help achieve >70% aligned bisulfite reads using BSMAP (https://code.google.com/archive/p/bsmap/).

Note: Illumina adapter trimming must be performed prior to adaptase tail trimming.

Table 1. Summary of Accel-NGS Methyl-Seq Tail Trimming Recommendations

Mapped Insert Size	Read 1 trimming (optional)	Read 2 trimming
~170 bp	15 bases from END of Read1	15 bases from START of Read2

ADAPTASE MODULE LIBRARIES

The Accel-NGS Adaptase Module constructs NGS libraries from single cells that have undergone lysis, bisulfite conversion and random primer extension that introduces a synthetic 8 base random sequence at the beginning of Read 1. Additionally, the workflow optionally incorporates a 6-8 base in-line barcode 5' to the random primer to enable a three-dimensional indexing strategy (see 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 Read 2, with a median length of 8 bases (Figure 2). Trimming of synthetic sequences is required to obtain improved mapping efficiency (with tools like BSMAP) because some aligners cannot soft clip synthetic sequence content which can interfere with mapping. Random primer barcode and tail trimming also removes synthetic unmethylated cytosines and helps achieve accurate methylation and bisulfite conversion information. 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 (if using in-line barcodes, be sure to capture the index information for proper single cell demultiplexing).

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 utilizing 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 read 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 prior to random primer/barcode and adaptase tail trimming.

Table 2: Summary of Adaptase Module trimming Recommendations

Mapped Insert Size	Read 1 Trimming (Required)	Read 2 Trimming
>400 bp	15 bases from START of Read 1	15 bases from START of Read 2

ACCEL-NGS 1S PLUS LIBRARIES

Whole genome and other DNA sequencing applications involve paired end alignment to reference genomes using common aligners such as BWA-MEM, BWA-ALN (Li H. and Durbin R., 2009), Novoalign, SOAP or Bowtie (Langmead B. and Salzberg S., 2012). 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 Adaptase synthetic tails from the beginning of R2 should significantly enhance read assembly into Contigs.

Many applications require symmetric lengths for both reads, so ensure to trim the same number of bases from both R1 and R2. For PE reads trim 15 bases from the end of R1 (3' end) and 15 bases from the beginning of R2 (5' end) in order to remove tail sequences that may be encountered 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 (cfDNA) library. If performing single read sequencing, trim 15 bases from the end of R1 if the read length is approaching the library insert size as viewed in Fastqc report figure "Per Base Sequence Content"(Figure 2).

Note: Illumina adapter trimming must be performed prior to adaptase tail trimming.

Table 3: Summary of 1S Plus tail trimming Recommendations

Mapped Insert Size	Read 1 trimming (optional)	Read 2 trimming
200 bp, 350 bp*	15 bases from END of Read1	15 bases from START of Read2

**For Covaris sheared DNA. Mapped insert size will be smaller for cfDNA and heavily nicked DNA*

SEQUENCING RECOMMENDATIONS

lllumina® sequencing chemistry is sensitive to low complexity base composition such as that introduced by bisulfite conversion. The following recommendations by Illumina lead to successful sequencing runs, where PhiX or any balanced, high sequence complexity library can be spiked in with the bisulfite converted libraries

These recommendations are subject to change, depending on the version of sequencer software. Please contact Illumina technical support for the most up to date recommendations.

Platform	PhiX Aligned (%)†
iSeq 100	minimum 5%
MiniSeq	10-50%*
MiSeq (MCS 2.2 or higher)	minimum 5%
NextSeq	10-50%*
HiSeq 2500 (HCS 2.2.38 or higher)	minimum 10%
HiSeq 3000/4000 (HCS 3.3.76 or lower)	10-50%*
HiSeq 3000/4000 (HCS 3.4.0 or higher)	5-20%*
NovaSeq	minimum 10%

Source: Illumina webpage (https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html)

REFERENCES

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. 2014. *Bioinformatics.* 30(15):2114-20. doi: 10.1093/bioinformatics/btu170

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. 2012.*Nature Methods*. 9(4):357-9. doi: 10.1038/nmeth.1923

Li H, Durbin R. Fast and accurate alignment with Burrows-Wheeler transform. 2009. *Bioinformatics*. 25(14):1754-60. doi: 10.1093/bioinformatics/btp324

Swift Biosciences
674 S Wagner Rd, Suite 100 • Ann Arbor, MI 48103 • 734.330.2568 • www.swiftbiosci.com