For Research Use Only.
Not for use in diagnostic procedures.

# KAPA Universal UMI Adapter

Universal UMI adapter for ligation-based library construction prior to sample barcoding in the KAPA HyperCap Workflow and KAPA HyperPETE Workflow Somatic Tissue DNA, KAPA HyperPETE Workflow Somatic Plasma cfDNA, and KAPA HyperPETE Workflow Tissue RNA Fusion Transcript for KAPA Target Enrichment products.

**Version 02**
Content version: September 2021

Store at -15°C to -25°C

| Product Name and Pack Size | Catalog # |
|---|---|
| KAPA Universal UMI Adapter, 96 Reactions | 09329862001 |

## Contents

| Component | Quantity | Amount |
|---|---|---|
| **Kit for 96 Reactions (Catalog# 09329862001)** | | |
| KAPA Universal UMI Adapter | 1 | 960 µL |

### Storage and Stability

- Store at -15°C to -25 ºC through the expiration date printed on the label. The product can withstand five freeze-thaw cycles without a negative impact to reagent stability and performance.

### Warnings and Precautions

⚠ Do not expose adapter tubes to temperatures above room temperature (~25°C) for extended periods.

### Application

The KAPA Universal UMI Adapter is a truncated adapter used during ligation-based library construction for sequencing on Illumina® instruments. A pair of unique, 8-nucleotide sequencing indices (barcodes) are added by PCR for use in single sample or multiplexed sequencing applications. The KAPA Universal UMI Adapter has been manufactured by duplexing (hybridizing) thirty two oligonucleotides and is designed to ensure high library construction ligation efficiency and low adapter-dimer formation.

See Appendix for analysis guidance.

### Number of Reactions

The supplied KAPA Universal UMI Adapter is sufficient for 96 library preparation reactions.

### How to Use this Product

Instructions for use may be found in the *KAPA HyperCap Workflow, KAPA HyperPETE Somatic Tissue DNA Workflow, KAPA HyperPETE Somatic Plasma cfDNA Workflow* and *KAPA HyperPETE Tissue RNA Fusion Transcript Workflow* Instructions for Use (IFU). To download the latest version of an *IFU*, go to *sequencing.roche.com/support*.

### Changes to Previous Version

Updated to include usage with the *KAPA HyperPETE Workflow.*

### Ordering Information

For a complete overview of Roche Sequencing products, including those used in the *KAPA HyperCap Workflow* and *KAPA HyperPETE Workflow*, please visit: *sequencing.roche.com/products*.

### Conventions

In this document, the following symbols are used to highlight important information:

| Symbol | Description |
|---|---|
| ⚠ | **Important Note:** Information critical to the success of the procedure or use of the product. |

## Notice to Purchaser

### 1. Limited Warranty

A. Products: Roche Sequencing Solutions, Inc. ("Roche") warrants that its Products conform to its published specifications and are free from defects in material or workmanship. Customer's sole and exclusive remedy (and Roche's sole and exclusive liability) under this limited warranty shall be to either (a) replace the defective Products, or (b) provide Customer with a refund, as solely determined by Roche.

B. Under no circumstances shall Roche's liability to Customer exceed the amount paid by Customer for the Services and Products to Roche. Roche will bear all reasonable shipping costs if service is re-performed at Roche or the Products are replaced. This warranty does not apply to any defect or nonconformance caused by (i) the failure by Customer to provide a suitable storage, use, or operating environment for the Materials or Customer's submission of substandard quality Materials or contaminated or degraded Materials to Roche, (ii) Customer's use of non-recommended reagents, (iii) Customer's use of the Products, Materials or Data for a purpose or in a manner other than that for which they were designed, (iv) the failure by Customer to follow Roche's published protocols, or (v) as a result of any other abuse, misuse or neglect of the Products, Materials or Data by Customer. This warranty applies only to Customer and not to third parties.

C. TO THE FULLEST EXTENT PERMITTED BY APPLICABLE LAW, ROCHE DISCLAIMS ALL OTHER REPRESENTATIONS, AND WARRANTIES, EXPRESS OR IMPLIED, WITH RESPECT TO THE PRODUCTS, SERVICES AND DATA, INCLUDING BUT NOT LIMITED TO, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT. CUSTOMER'S SOLE REMEDY FOR BREACH OF WARRANTY IS STATED ABOVE.

D. Any action by Customer against Roche for Roche's breach of this warranty must be commenced within 12 months following the date of such breach. Notwithstanding such 12-month period, within twenty (20) days of the delivery of Data and/or Products to Customer, Customer must notify Roche in writing of any nonconformity of the Services and Products, describing the nonconformity in detail; otherwise all Services and Products shall be conclusively deemed accepted without qualification.

### 2. Further Liability Limitation

To the fullest extent permitted under applicable law, Roche shall not have any liability for incidental, compensatory, punitive, consequential, indirect, special or other similar damages, however caused and regardless of form of action whether in contract, tort (including negligence), strict product liability or otherwise, even if Roche has been advised of the possibility of such damages. Customer understands that any risks of loss hereunder are reflected in the price of the services and products and that these terms would have been different if there had been a different allocation of risk.

Evidence of original purchase is required. It is important to save your sales receipt or packaging slip to verify purchase.

## Restrictions and Liabilities for Roche Products

This document is provided "as is" and Roche Sequencing Solutions, Inc. ("Roche") assumes no responsibility for any typographical, technical or other inaccuracies in this document. Roche reserves the right to periodically change information that is contained in this document; however, Roche makes no commitment to provide any such changes, updates, enhancements or other additions to this document to you in a timely manner or at all.

OTHER THAN THE LIMITED WARRANTY CONTAINED IN THIS DOCUMENT, ROCHE MAKES NO REPRESENTATIONS, WARRANTIES, CONDITIONS OR COVENANTS, EITHER EXPRESS OR IMPLIED (INCLUDING WITHOUT LIMITATION, ANY EXPRESS OR IMPLIED WARRANTIES OR CONDITIONS OF FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, MERCHANTABILITY, DURABILITY, TITLE, OR RELATED TO THE PERFORMANCE OR NON-PERFORMANCE OF ANY PRODUCT REFERENCED HEREIN OR PERFORMANCE OF ANY SERVICES REFERENCED HEREIN).

This document might contain references to third party sources of information, hardware or software, products or services and/or third party web sites (collectively the "Third-Party Information"). Roche does not control, and is not responsible for, any Third-Party Information, including, without limitation the content, accuracy, copyright compliance, compatibility, performance, trustworthiness, legality, decency, links, or any other aspect of Third-Party Information. The inclusion of Third-Party Information in this document does not imply endorsement by Roche of the Third-Party Information or the third party in any way.

Roche does not in any way guarantee or represent that you will obtain satisfactory results from using Roche products as described herein. You assume all risk in connection with your use of Roche products.

Roche is not responsible nor will be liable in any way for your use of any software or equipment that is not supplied by Roche in connection with your use of Roche products.

## Trademarks

KAPA is a trademark of Roche.
All other product names and trademarks are the property of their respective owners.

## Patent License Limitations

For patent license limitations for individual products please refer to: *www.technical-support.roche.com*.

## Regulatory Disclaimer

For Research Use Only. Not for use in diagnostic procedures.

## Contact and Support

If you have questions, contact your local Roche Technical Support.
Go to *sequencing.roche.com/support* for contact information.



Roche Sequencing Solutions, Inc.
4300 Hacienda Drive
Pleasanton, CA 94588 USA

# Appendix – Analysis guidelines for deduplication of sequencing data with unique molecular identifiers (UMIs)

## Overview

Many approaches exist for the analysis of next-generation sequencing (NGS) data. The most appropriate set of tools depends on the specific biological question that is asked, as well as the availability of computational resources.

Although multiple analysis objectives are possible, this appendix only describes the recommended methodology to obtain consensus reads from PCR and optical duplicates, irrespective of application. It is specific for sequencing data from libraries constructed using KAPA Universal Adapter with UMI and KAPA UDI Primer Mixes and also provides an example command line script.

The methodology, illustrated below, consists of three steps:

1. Extraction of the UMI from insert reads
2. Grouping of these UMIs into families/groups based on alignment coordinates and UMI sequence composition
3. Consensus calling of all reads within a particular UMI grouping

## Installation of required software packages

Several open source software packages are required to execute the methodology described in this appendix.

**NOTE:** While UMI-tools[1] may be able to perform some parts of the required analysis, it does not have consensus calling capabilities and thus the recommended methodology makes use of fgbio (Fulcrum genomics) exclusively.

- fgbio v1.3.0 fulcrumgenomics.github.io/fgbio
- GATK 4.1.8.1 github.com/broadinstitute/gatk
- bwa v0.7.17 bio-bwa.sourceforge.net
- fastp v0.20.1 github.com/OpenGene/fastp
- samtools v1.9 htslib.org

## How to use the provided command line script

Refer to the KAPA UDI Primer Mixes IFUs and the Illumina Systems documentation for demultiplexing guidance.

While we recommend typing the code manually, users may elect to copy and paste the code from the respective text boxes.

**NOTE:** All code indicated in *italics* require user input and cannot simply be copied and pasted. This is indicated where relevant.

[1] Smith T, Heger A, Sudbery I. (2017) UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res, 27:491–499.

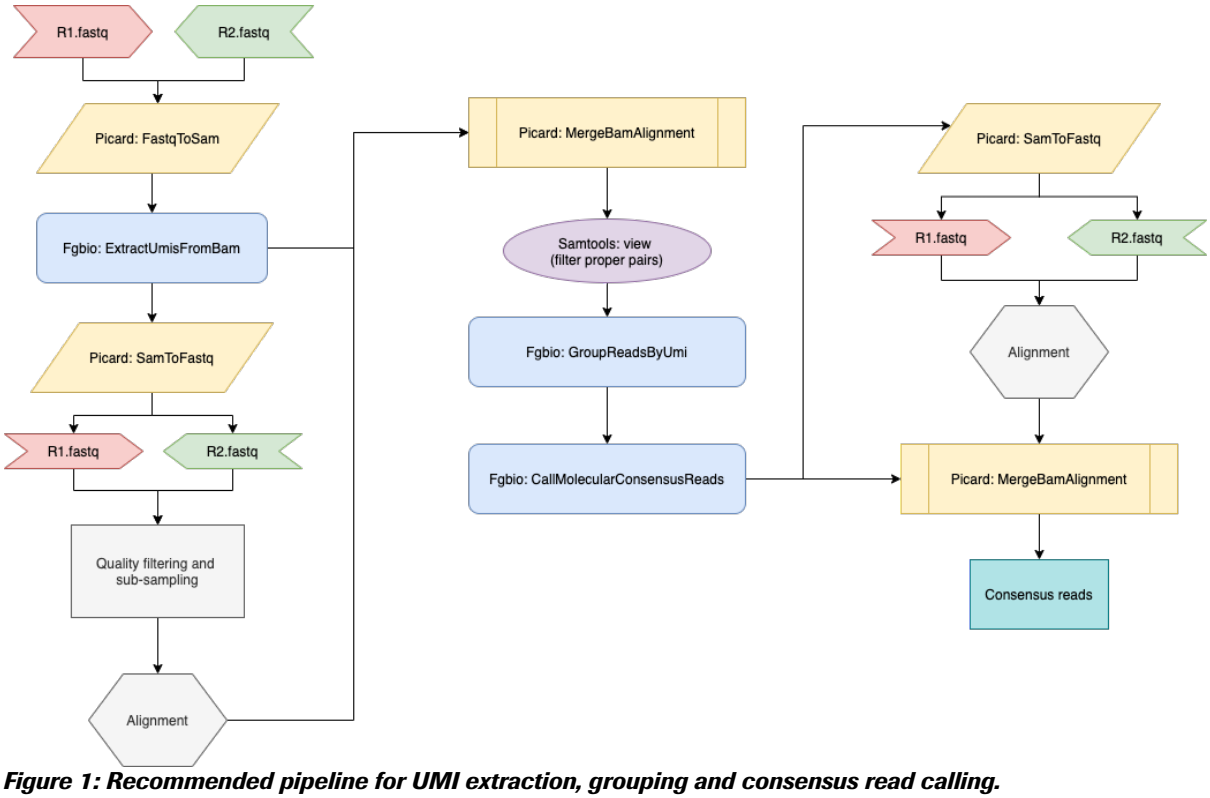


***Figure 1: Recommended pipeline for UMI extraction, grouping and consensus read calling.***

# Sequences of Unique Molecular Identifiers used by KAPA Universal Adapter with UMI

| Unique Molecular Identifier | | | |
|---|---|---|---|
| 1 | AATCC*T | 9 | ACC*T |
| 2 | AGAAG*T | 10 | ATG*T |
| 3 | CCAGG*T | 11 | CAG*T |
| 4 | CTTAC*T | 12 | CGC*T |
| 5 | GAAGC*T | 13 | GCG*T |
| 6 | GGTCG*T | 14 | GTC*T |
| 7 | TCTAG*T | 15 | TAC*T |
| 8 | TTACC*T | 16 | TGG*T |

"*T" denotes the 3' T-overhang of the KAPA Universal Adapter

# Methodology

## 1. Extraction of UMIs from the insert reads

Ultimately, this step will yield a mapped BAM file containing the UMI information in the RX tag. This workflow utilizes the *ExtractUmisFromBam* tool in fgbio.

1.1 fgbio processes unmapped BAM files. Convert the demultiplexed, raw sequencing FASTQ files to BAM files using the *FastqToSam* tool in GATK.

**IMPORTANT!** Provide the names of the corresponding raw FASTQ files where indicated by code in *italics*.

```
gatk FastqToSam \
   -F1
demultiplexed_R1.fastq.gz \
   -F2
demultiplexed_R2.fastq.gz \
   -O unmapped.bam \
```

*For information on the functionality of the tool please refer to github.com/broadinstitute/gatk*

1.2 Use the *ExtractUmisFromBam* tool in fgbio. The read structure is defined as 3M3S+T. Extract the first three (3) bases and store them as the UMI in the RX tag of the BAM file (3M). Trim the subsequent three (3) bases off the start of the read (3S). These bases constitute a punctuation sequence that increases the sequence diversity to ensure optimal sequencing performance. Maintain the remaining sequence as part of the insert read (+T). The UMIs extracted from read 1 and read 2 are stored in the RX tag of the unmapped BAM file as UMI1-UMI2 (hereafter referred to as "the UMI" and considered as a single sequence).

```
fgbio ExtractUmisFromBam \
  -i unmapped.bam \
  -o unmapped_umi_extracted.bam \
  -r 3M3S+T 3M3S+T \
  -t RX \
  -a true
```

*For information on the functionality of the tool please refer to fulcrumgenomics.github.io/fgbio/tools/latest/ExtractUmisFromBam.html*

1.3 Convert the BAM file with UMI extracted reads to a FASTQ file for adapter and quality trimming. Adapter and quality trimming should only take place after UMI extraction, to avoid any bias and ensure that only the template/insert is trimmed. In this workflow, the unmapped BAM file is first converted to FASTQ using GATK then adapter and quality trimmed using fastp. Set the parameters so that the tool automatically detects adapter sequences or specify adapter sequences (available in Illumina Adapter Sequences Document #1000000002694 v.11 or later).

**NOTE:** BAM to FASTQ conversion does not retain extracted UMI information. Thus, it is important to retain the output file from the UMI extraction, *unmapped_umi_extracted.bam*, to preserve the UMI information stored in the RX tag that is used downstream after genomic alignment

```
   gatk SamToFastq \
-I unmapped_umi_extracted.bam \
-F umi_extracted_R1.fastq \
-F2 umi_extracted_R2.fastq \
--CLIPPING_ATTRIBUTE XT \
 --CLIPPING_ACTION 2
```

*For information on the functionality of the tool please refer to github.com/broadinstitute/gatk*

```
fastp \
   -i umi_extracted_R1.fastq \
   -o umi_extracted_trimmed_R1.fastq \
   -I umi_extracted_R2.fastq \
   -O umi_extracted_trimmed_R2.fastq \
   -g -W 5 -q 20 -u 40 -x -3 -l 75 -c \
   -j fastp.json \
   -h fastp.html \
   -w 12
```

*For information on the functionality of the tool please refer to github.com/OpenGene/fastp*

1.4 To group reads into UMI families/groups requires alignment coordinate data. Align quality and adapter trimmed reads to the reference genome. BWA-mem is recommended for this step.

**NOTE:** RNA-Seq data will require a splice-aware aligner if aligning to a genome e.g., STAR or HISAT2 aligner.
**IMPORTANT!** Replace code indicated in *italics* with the file name of the reference sequence.

```
bwa mem \
   -t 10 \
       -M path_to_reference.fa \
   umi_extracted_trimmed_R1.fastq \
   umi_extracted_trimmed_R2.fastq \
   | \
   samtools view -Sb > umi_extracted_aligned.bam
```

*For information on the functionality of the tool please refer to bio-bwa.sourceforge.net*

1.5 Use the tool *MergeBamAlignment* in gatk to merge the two BAM files containing:

*i.* the UMI information *(unmapped_umi_extracted.bam)*
*ii.* the alignment coordinate information *(umi_extracted_aligned.bam).*

```
    gatk MergeBamAlignment \
      --ATTRIBUTES_TO_RETAIN X0 \
      --ATTRIBUTES_TO_REMOVE NM \
      --ATTRIBUTES_TO_REMOVE MD \
 --ALIGNED_BAM umi_extracted_aligned.bam \
 --UNMAPPED_BAM unmapped_umi_extracted.bam \
 --OUTPUT umi_extracted_aligned_merged.bam \
 --REFERENCE_SEQUENCE reference_genome.fa
 --SORT_ORDER 'queryname' \
--ALIGNED_READS_ONLY true \
--MAX_INSERTIONS_OR_DELETIONS -1 \
--PRIMARY_ALIGNMENT_STRATEGY MostDistant \
--ALIGNER_PROPER_PAIR_FLAGS true \
--CLIP_OVERLAPPING_READS false
```

*For information on the functionality of the tool please refer to github.com/broadinstitute/gatk*

The UMI information is now stored in the RX tag of the new *umi_extracted_aligned_merged.bam* file.

1.6 The final step of UMI extraction is to ensure that the BAM file only contains reads that are aligned in proper pairs. Use the tool *samtools view* and the flag *-f2* for this purpose.

```
samtools view \
    -f 2 \
    -bh umi_extracted_aligned_merged.bam \
    > \
    umi_extracted_aligned_merged_filtered.bam
```

*For information on the functionality of the tool please refer to htslib.org*

## 2. Group reads by UMI

Identify and group reads originating from the same source molecule using the *GroupReadsByUmi* tool in fgbio. This tool utilizes the UMI (UMI1-UMI2) and the genomic alignment start site to assign unique source molecules to each applicable read. *GroupReadsByUmi* implements the adjacency strategy introduced by UMI-tools[22]. The user can control how many errors/mismatches are allowed in the UMI sequence when assigning source molecules (*--edits=n*). UMI group statistics are output to a *umi_group_data.tsv* file using the *-f* flag.

**NOTE:** The parameter *--edits=1* will account for a single mismatch in the entire UMI sequence (UMI1+UMI2). Altering this parameter to >1 will have significant impact on the outcome of the UMI grouping algorithm and the resultant UMI groups. All read pairs with >1 error in the UMI will be discarded.

```
fgbio GroupReadsByUmi \
    --input= umi_extracted_aligned_merged_filtered.bam \
    --output=umi_grouped.bam \
    --strategy=adjacency \
    --edits=1 \
    -t RX \
    -f family_size_counts.txt
```

*For information on the functionality of the tool please refer to fulcrumgenomics.github.io/fgbio/tools/latest/GroupReadsByUmi.html*

## 3. Generate consensus reads

3.1 Calculate the consensus sequence using the *CallMolecularConsensusReads* tool in fgbio for each group of reads identified as originating from the same unique source molecule. The consensus of a group of reads can only be calculated if there are a minimum of two (2) reads in a group. Reads that occur as singletons are discarded by default but this can be changed by setting the *–min-reads* flag to 1, in so doing the single read will be considered the consensus.

**NOTE:** Bases with a sequencing quality less than 20 will not be used in the consensus calculation but this can also be altered with the *–min-input-base-quality* flag.

**NOTE:** Here, reads are defined as those grouped into a UMI family/group, i.e. reads that have the same UMI tag and the same 5' start position. Furthermore, the two reads minimum could consist of two reads from the same unique source molecule, or one read that was derived from the forward template of the unique source molecule and one read that was derived from the reverse template of the unique source molecule.

[2] Smith T, Heger A, Sudbery I. (2017) UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res, 27:491–499.

```
fgbio CallMolecularConsensusReads \
   --input=umi_grouped.bam \
   --output=consensus_unmapped.bam \
   --error-rate-post-umi 40 \
   --error-rate-pre-umi 45 \
   --output-per-base-tags false \
   --min-reads 2 \
   --max-reads 50 \
   --min-input-base-quality 20 \
   --read-name-prefix='consensus'
```

*For information on the functionality of the tool please refer to fulcrumgenomics.github.io/fgbio/tools/latest/CallMolecularConsensusReads.html*

3.2 After consensus calling, the collapsing of the UMI groups results in the loss of alignment coordinate information*. To rectify this, convert the *consensus_unmapped.bam* to FASTQ format using *SamToFastq in* gatk.

```
    gatk SamToFastq \
      -I consensus_unmapped.bam \
      --F consensus_unmapped_R1.fastq \
 --F2 consensus_unmapped_R2.fastq \
--CLIPPING_ATTRIBUTE XT \
--CLIPPING_ACTION 2
```

*For information on the functionality of the tool please refer to github.com/broadinstitute/gatk*

*Loss of alignment coordinates is an inherent limitation of consensus calling and is related to alignment quality. When base information is statistically extrapolated from two or more molecules the alignment quality is also statistically averaged. Many downstream variant callers rely on alignment quality and thus, to avoid error, the consensus reads are realigned to ensure correct alignment qualities.

3.3 Align the FASTQ files to the reference with BWA-mem and output to a new *consensus_mapped.bam* file.

```
bwa mem \
   -v 3 \
   -t 8 \
   -Y \
   -M \
   reference_genome.fa  \
   consensus_unmapped_R1.fastq \
   consensus_unmapped_R2.fastq \
    | \
   samtools view -bh - > consensus_mapped.bam
```

*For information on the functionality of the tool please refer to bio-bwa.sourceforge.net/*

3.4 Finally, merge the *consensus_mapped.bam* with the *consensus_unmapped.bam* to retain the UMI group information. This will yield an aligned BAM file with consensus reads and the UMI information retained in the RX flag.

```
    gatk MergeBamAlignment \
      --ATTRIBUTES_TO_RETAIN X0 \
      --ATTRIBUTES_TO_RETAIN RX \
 --ALIGNED_BAM consensus_mapped.bam \
--UNMAPPED_BAM consensus_unmapped.bam \
--OUTPUT consensus.bam \
--REFERENCE_SEQUENCE reference_genome.fa \
 --SORT_ORDER coordinate \
--ADD_MATE_CIGAR true \
--MAX_INSERTIONS_OR_DELETIONS -1 \
--PRIMARY_ALIGNMENT_STRATEGY MostDistant \
--ALIGNER_PROPER_PAIR_FLAGS true \
--CLIP_OVERLAPPING_READS false
```

*For information on the functionality of the tool please refer to github.com/broadinstitute/gatk*

The BAM file can now be used for all downstream applications and analysis described by the particular NGS analysis workflow. Quality filtering will be required if proceeding to variant calling.