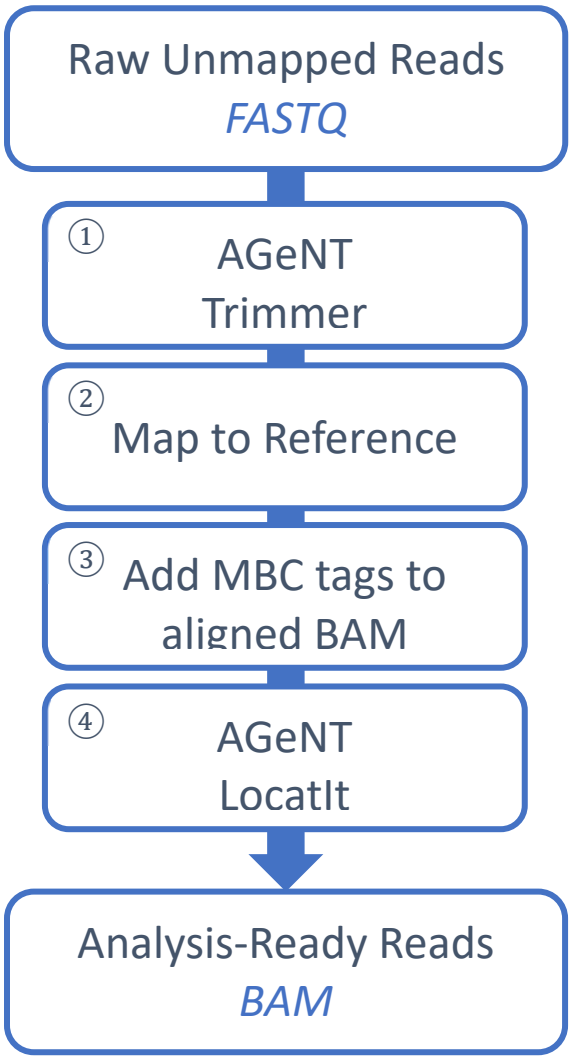
Best practices for processing SureSelect XT HS2 DNA and RNA data prior to variant discovery

Purpose:
Prior to variant discovery, raw sequencing data must be pre-processed to correct for technical biases and align to a reference genome. For SureSelect XT HS2 data, this process involves: 1) pre-processing the raw reads to extract the molecular barcode sequences and remove sequencing adaptors, 2) aligning the processed reads to the reference, 3) annotating the aligned file with the molecular barcode information (DNA only) and 4) PCR de-duplication leveraging the molecular barcodes.

Basic workflow using the **A**gilent **Ge**nomics **N**extGen **T**oolkit (AGeNT)



NOTE: Step 3 is only applicable to duplex-consensus mode (and hybrid mode) for the DNA workflow and is not relevant for RNA, which should only be run using single-consensus mode.

Input:
This workflow operates on raw Illumina sequencing data in FASTQ format. The data is expected to be demultiplexed, **but not adaptor trimmed**.

Steps:

1. Extract MBCs and trim sequencing adaptors.

   Use **AGeNT Trimmer** to extract molecular barcodes (MBCs) from the beginning of each read and trim embedded MBCs and sequencing- adaptors from the end of each read**.** This tool processes the reads in pairs, using the knowledge from the MBC extraction on one read to inform the adaptor trimming on the opposite strand read.

   Example invocation (on linux/mac):

```
agent.sh trim -v2 -fq1 /path/to/fastq_input_dir/sample_R1.fastq.gz -fq2
  /path/to/fastq_input_dir/sample_R2.fastq.gz
```

   The output files have SAM style tags injected into the read name headers.
   For example:

```
@D00266:1113:HTWK5BCX2:1:1102:9976:2206 BC:Z:CTACCGAA+AAGTGTCT
  ZA:Z:TTAGT ZB:Z:TCCT RX:Z:TTA-TCC QX:Z:DDD DDA
```

   List of tags

Tag	Type	Description
BC	Z	Sample barcode
RX	Z	Two MBC sequences (concatenated with "-")
QX	Z	Base quality representation of the MBCs (concatenated with space)
ZA	Z	3 MBC bases for read 1 followed by 1 or 2 dark bases
ZB	Z	3 MBC bases for read 2 followed by 1 or 2 dark bases

   NOTE: If the first 5 bases are not recognized as a valid molecular barcode, they are masked with "**N**" and the corresponding base qualities are marked as "**$**". This allows downstream filtering of these reads.

   Trimmer also creates a FASTQ-like txt file containing just the MBC sequences. The format of the sequences matches the format specified above.

2. Align the trimmed reads.

   BWA-MEM is strongly recommended because it contains an option that will easily propagate the SAM tags from the FASTQ read names to the final aligned BAM file. For RNAseq data, any aligner designed for RNA data should work. LocatIt was tested with output from STAR.

3. Add MBC tags to the aligned file (DNA duplex/enhanced hybrid mode only – not relevant for RNA or DNA single consensus mode).

If using BWA-MEM, the "-C" option will append the FASTQ comment from the read header to the SAM output. If using a different aligner, it is necessary to annotate the aligned BAM file with the RX and QX MBC tags listed in step 1. Convert the FASTQ files from step 1 into an unaligned BAM (uBAM). Sort the uBAM and aligned BAM by read name and merge the two together using a tool such as the picard MergeBamAlignment tool.

Example invocation:

```
bwa mem -C -t 2 /hg38.fa trimmed_dir/sample_R1.cut.fastq.gz
  trimmed_dir/sample_R2.cut.fastq.gz | samtools view -b - >
  aligned_dir/sample.bam"
```

4. Generate consensus reads

The AGeNT LocatIt tool is used to generate consensus reads using the MBCs. The tool has been tested with datasets containing up to 70 M read pairs. By default, this tool generates a file containing all of the input reads, with duplicate reads flagged as *read is PCR or optical duplicate* (SAM flag 0x400) and filtered reads flagged as *read fails platform/vendor quality checks* (SAM flag 0x200). If desired, the tool can instead generate files with duplicate(SAM flag 0x400), filtered(SAM flag 0x200), secondary(SAM flag 0x100) and supplementary(SAM flag 0x800) reads all removed (using the "-R" parameter). Alternately, a third-party tool such as "samtools view" can be used to remove the marked reads.

NOTE: The below options for duplex consensus and enhanced hybrid modes are relevant for DNA workflows only. Unlike DNA, where both strands are present and the MBCs in the strands can be matched to form a duplex consensus read, single-stranded RNA stops at single consensus generation.

Example invocation (on linux/mac):

```
agent.sh -Xmx12G locatit -S -v2Only -d 1 -m 3 -q 25 -Q 25 -l
  Covered.bed -o deduped_dir/sample.bam aligned_dir/sample.bam
```

For SureSelect XT HS2 data, LocatIt can run in 3 consensus generation modes:

Mode	Command-line option	Description
Duplex consensus	`-v2Only`	Requires at least 1 read for each strand (reads where there is only 1 strand support for the MBC family are flagged as *read fails platform/vendor quality checks* (SAM flag 0x200)).

Single-strand consensus	`-i`	Ignores strand information at treats the duplex MBC as a single MBC. This mode requires input of the FASTQ-like MBC file generated by AGeNT Trimmer in addition to the aligned and annotated BAM file (see usage example of SureSelect non-XTHSv2 in LocatIt manual).
Enhanced Hybrid	`-v2Hybrid`	The approach is the same as for -v2Only mode but different from the legacy "-v2" mode. With this option, the single-strand consensus reads are not flagged as failing the vendor quality test. Each duplex consensus read is written *twice* to the output file to ensure that these reads are weighted properly when compared with the retained single consensus reads. The read names for the two duplex consensus reads shall match the read names for the two single consensus reads that were used to generate it.

The consensus reads in the output BAM files may have the following annotations:

Tag	Purpose	Example
`fi:Z`	If a read has been marked as filtered, the justifications for filtering	`fi:Z:BAD_READ2_QUALITY; BELOW_MBC_NREADS_LIMIT;`
`Xl:i`	Total number of duplicates collapsed for single-strand consensus	`XI:i:2`
`XJ:i`	Total number of duplicates collapsed for single-strand consensus on the other strand (only present when a duplex consensus has been formed)	`XJ:i:1`

Output:

Once LocatIt consensus generation is complete, the resulting BAM files are ready for analysis with any downstream analysis tools, such as variant discovery.