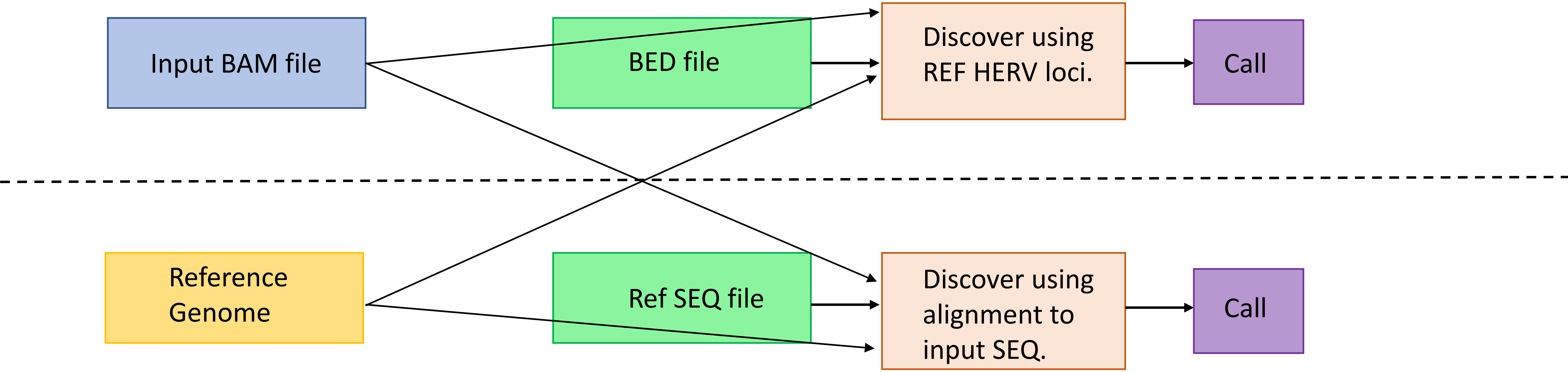
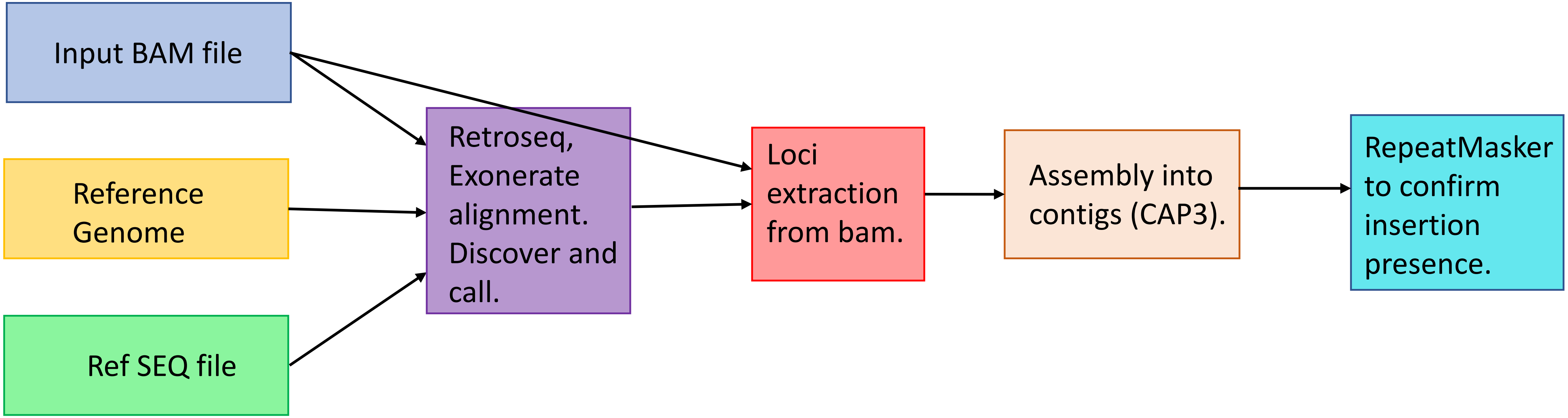
Flow diagrams for tested tools

Retroseq


Input BAM file
BED file
Discover using REF HERV loci.
Call
Reference Genome
Ref SEQ file
Discover using alignment to input SEQ.
Call

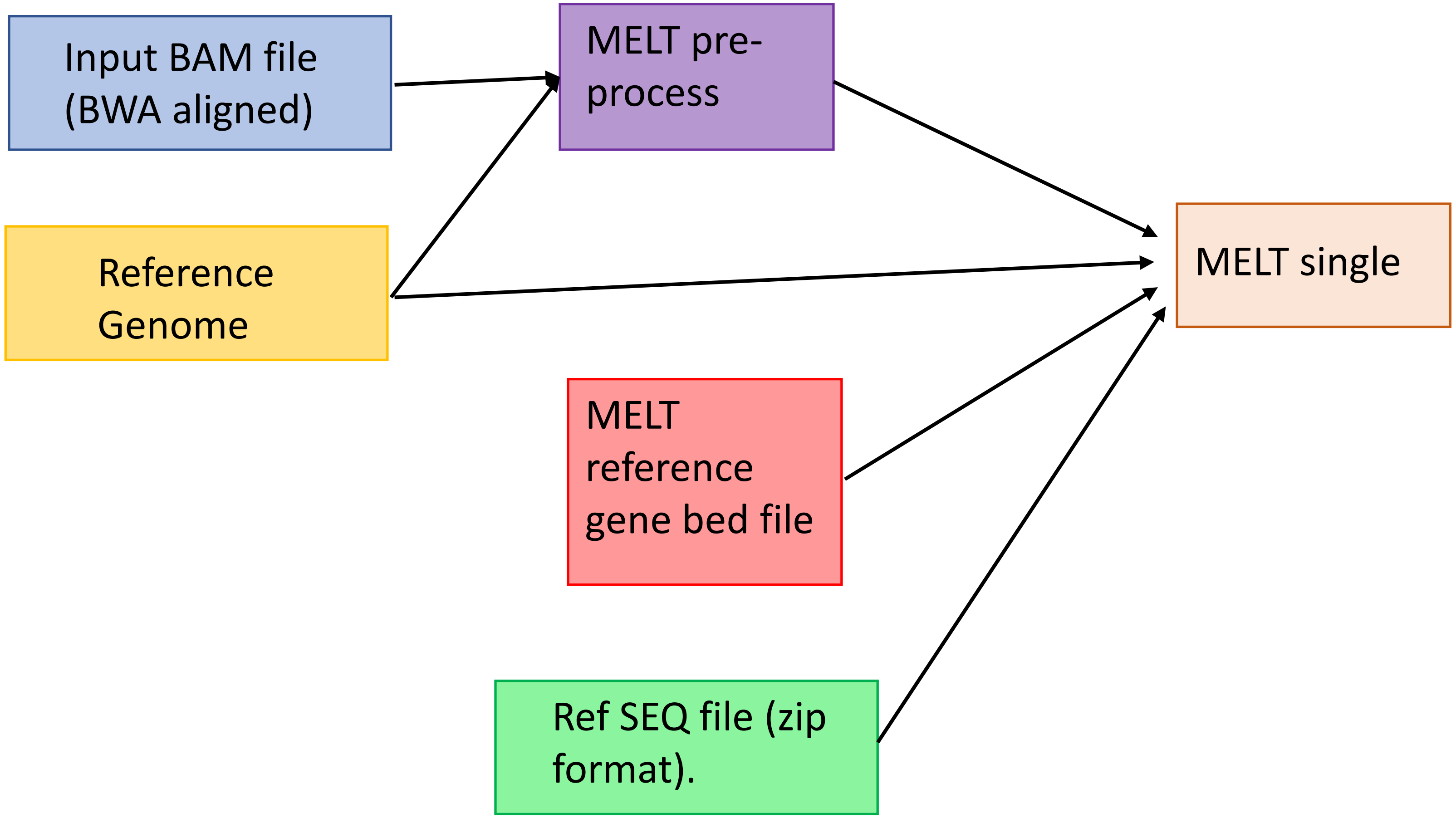
Retroseq can be implemented either using alignment to a reference fasta of a target sequence (e.g. LTR5_Hs) or instead by mapping discordant reads to reference loci – assuming that for a non-reference insertion the reads are aligned to reference copies. Alignment requires exonerate and takes longer to run. More information can be found here: https://github.com/tk2/RetroSeq

Retroseq+


Input BAM file
Reference Genome
Ref SEQ file
Retroseq, Exonerate alignment. Discover and call.
Loci extraction from bam.
Assembly into contigs (CAP3).
RepeatMasker to confirm insertion presence.

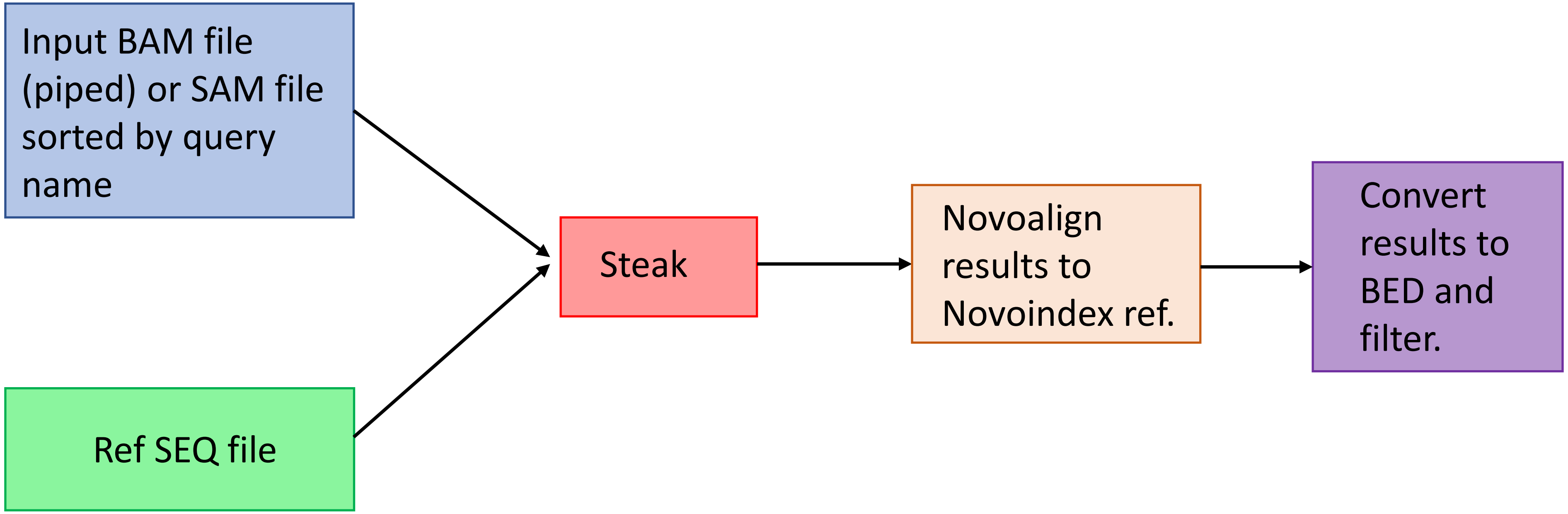
Retroseq+ first applies Retroseq to an input bam file, high confidence loci are filtered. Reads at these loci are extracted from the input bam file and assembled into continuous sequences. These sequences are input to repeatmasker which identifies repetitive elements. If contigs are positive for target sequence in contigs up and downstream from the locus, they are output into the final result. More information can be found here: https://www.pnas.org/doi/10.1073/pnas.1602336113

MELT


Input BAM file (BWA aligned)
Reference Genome
MELT pre-process
MELT reference gene bed file
Ref SEQ file (zip format).
MELT single

MELT has an initial pre-processing step to speed up run-time.
This step produces 3 files related to discordant read pairs.
Following this, MELT 'single' is run to identify HERV-K loci. More information can be found here: https://melt.igs.umaryland.edu/manual.php

Steak



Input BAM file (piped) or SAM file sorted by query name

Ref SEQ file

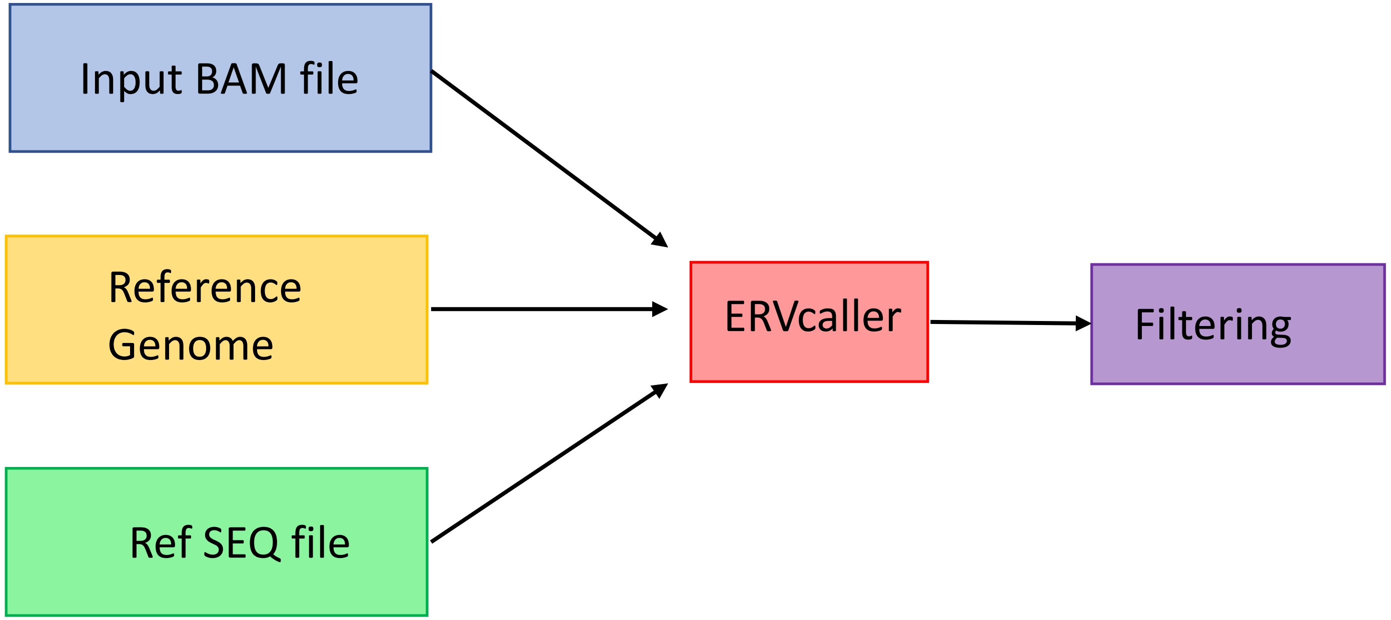
Steak

Novoalign results to Novoindex ref.

Convert results to BED and filter.

For our analysis, inputs were in SAM format. Steak takes in a genome sequence file and a fasta of the target insertion. Following the Steak command, post-processing with Novoalign and filtering is applied as outlined on the Steak github page. More information can be found here: https://github.com/applevir/STEAK

ERVcaller



ERVcaller is the simplest tool to run, with no pre-processing. Recommended filtering options for ERVcaller results are outlined in the ERVcaller github page and we used these recommended filters for our analysis. More information can be found here: https://github.com/xunchen85/ERVcaller