AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim[1†], Can Firtina[1†], Meryem Banu Cavlak[1], Damla Senol Cali[2],
Nastaran Hajinazar[1,3], Mohammed Alser[1], Can Alkan[4] and Onur Mutlu[1,2,4*]

[1] ETH zürich [2] Carnegie Mellon [3] SFU Simon Fraser University [4] Bilkent University

bioRxiv Preprint

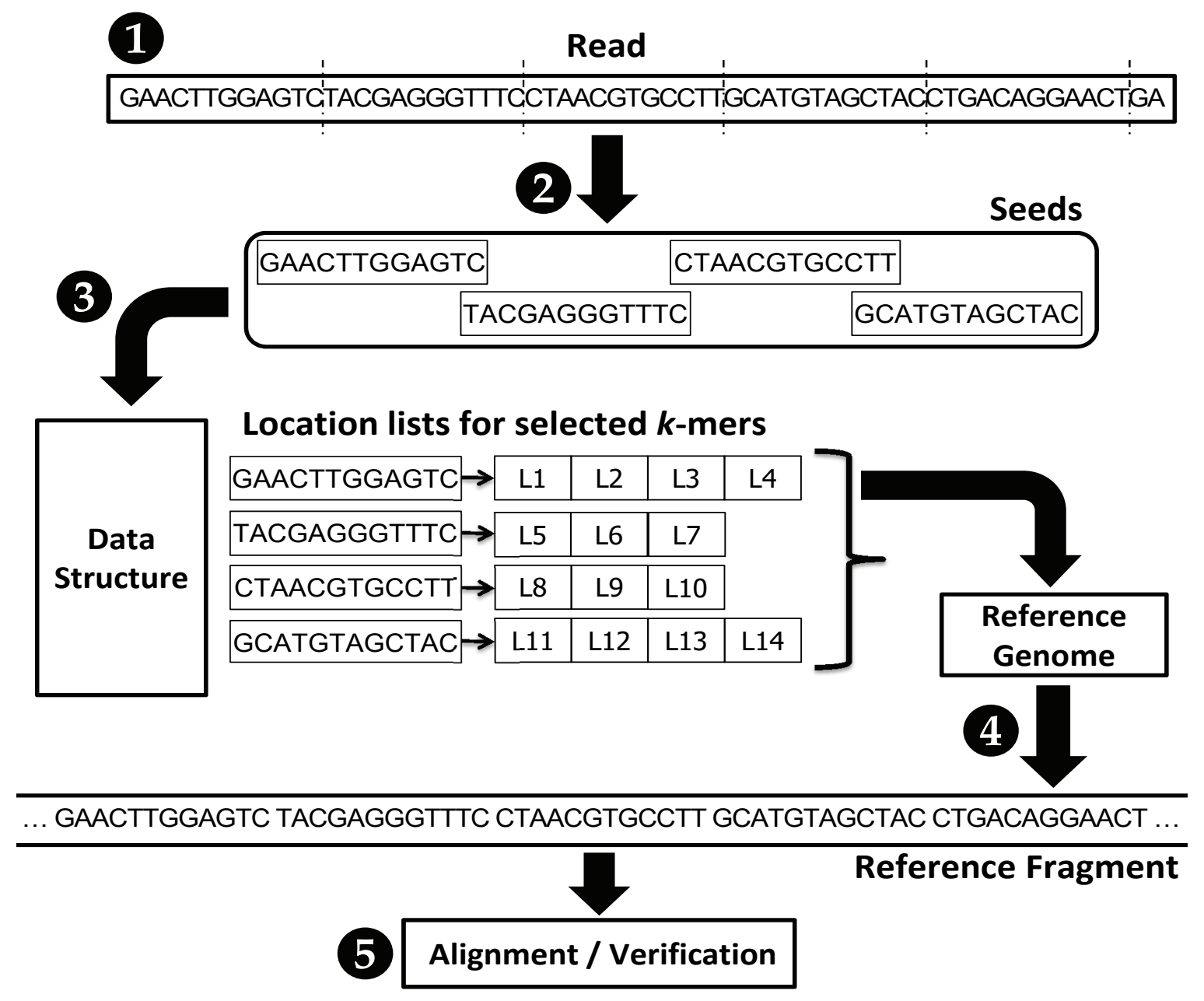
Source Code

Abstract

AirLift is the first read remapping tool that enables users to quickly and comprehensively map a read set, that had been previously mapped to one reference genome, to another similar reference. Users can then quickly run downstream analysis of read sets for each latest reference release. Compared to the state-of-the-art method for remapping reads (i.e., full mapping), **AirLift reduces the overall execution time to remap** read sets between two reference genome versions **by up to 27.4×**. We validate our remapping results with GATK and find that AirLift provides high accuracy in identifying ground truth SNP/INDEL variants.

1: Read Mapping

To identify a sample's full genome sequence for insight into the sample's health, one must first sequence small DNA fragments (reads) from the samples genome.

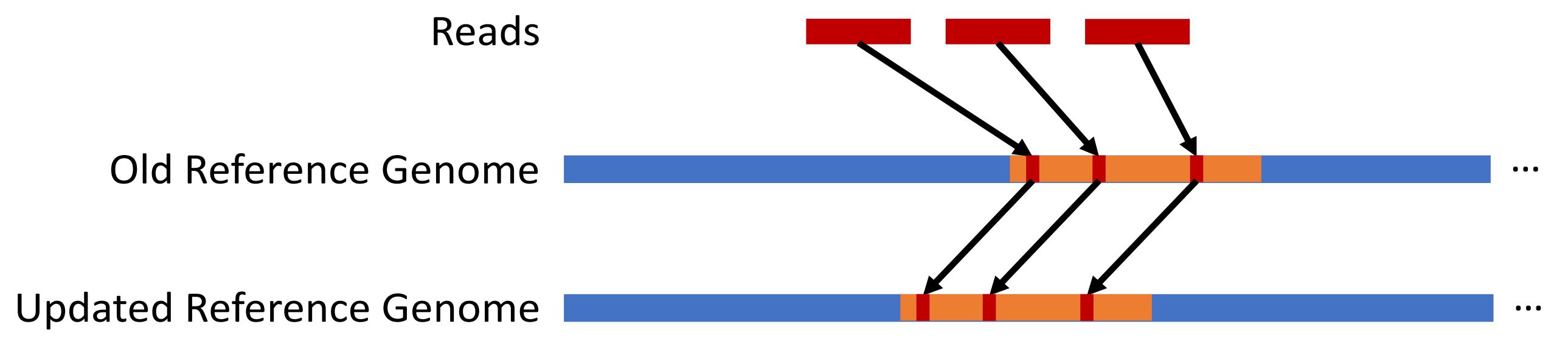

1
Read
GAACTTGGAGTCTACGAGGGTTTCCTAACGTGCCTTGCATGTAGCTACCTGACAGGAACTGA
2
Seeds
GAACTTGGAGTC
CTAACGTGCCTT
TACGAGGGTTTC
GCATGTAGCTAC
3
Data Structure
Location lists for selected k-mers
GAACTTGGAGTC → L1 L2 L3 L4
TACGAGGGTTTC → L5 L6 L7
CTAACGTGCCTT → L8 L9 L10
GCATGTAGCTAC → L11 L12 L13 L14
Reference Genome
4
… GAACTTGGAGTC TACGAGGGTTTC CTAACGTGCCTT GCATGTAGCTAC CTGACAGGAACT …
Reference Fragment
5
Alignment / Verification

Read Mapping: Identifying the original location of billions of reads within the genome using a reference genome, which represents the average genome of a species, to identify genomic variants

- Requires **costly approximate string matching**

2: Genome Updates and Current Remapping Tools

Reference genomes are updated frequently as better sequencing tools and more information become available to researchers. These updates come in form of significant updates or smaller patches increasing the reference's similarity to the average human genome and annotation accuracy.



Current remapping tools simply update previously mapped read coordinates from an old reference to the updated reference in regions based on similar regions.

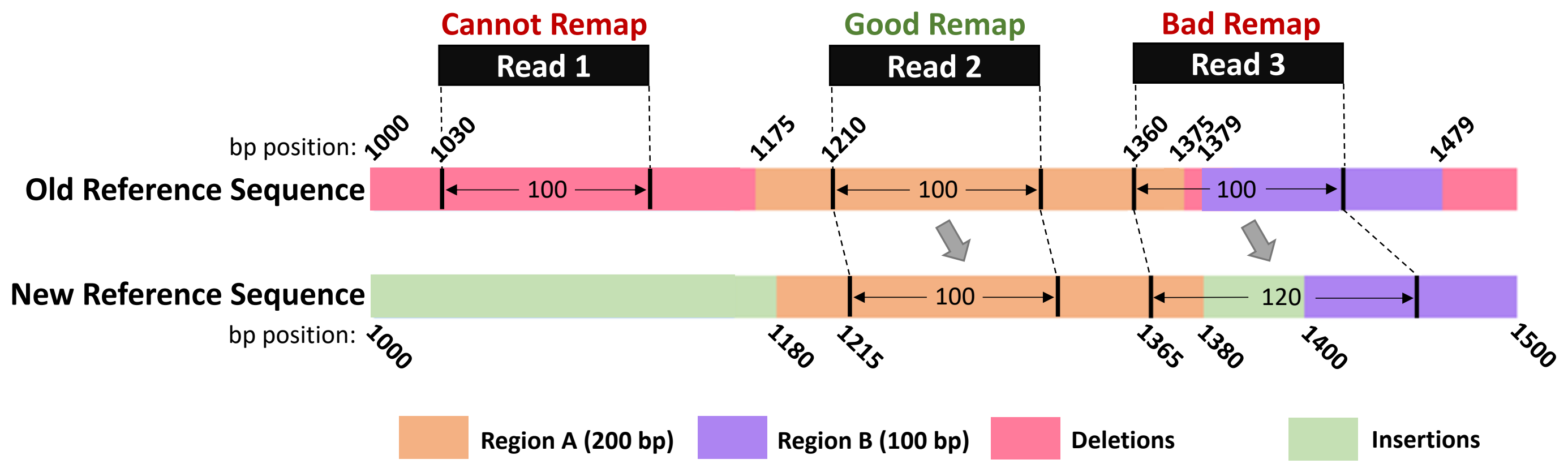
3: Problem

1. We want to utilize the latest information/annotations provided by the latest reference genome, but ***read mapping is a computationally expensive workload***. We do **not** want to read map the *entire read set* to updated reference.

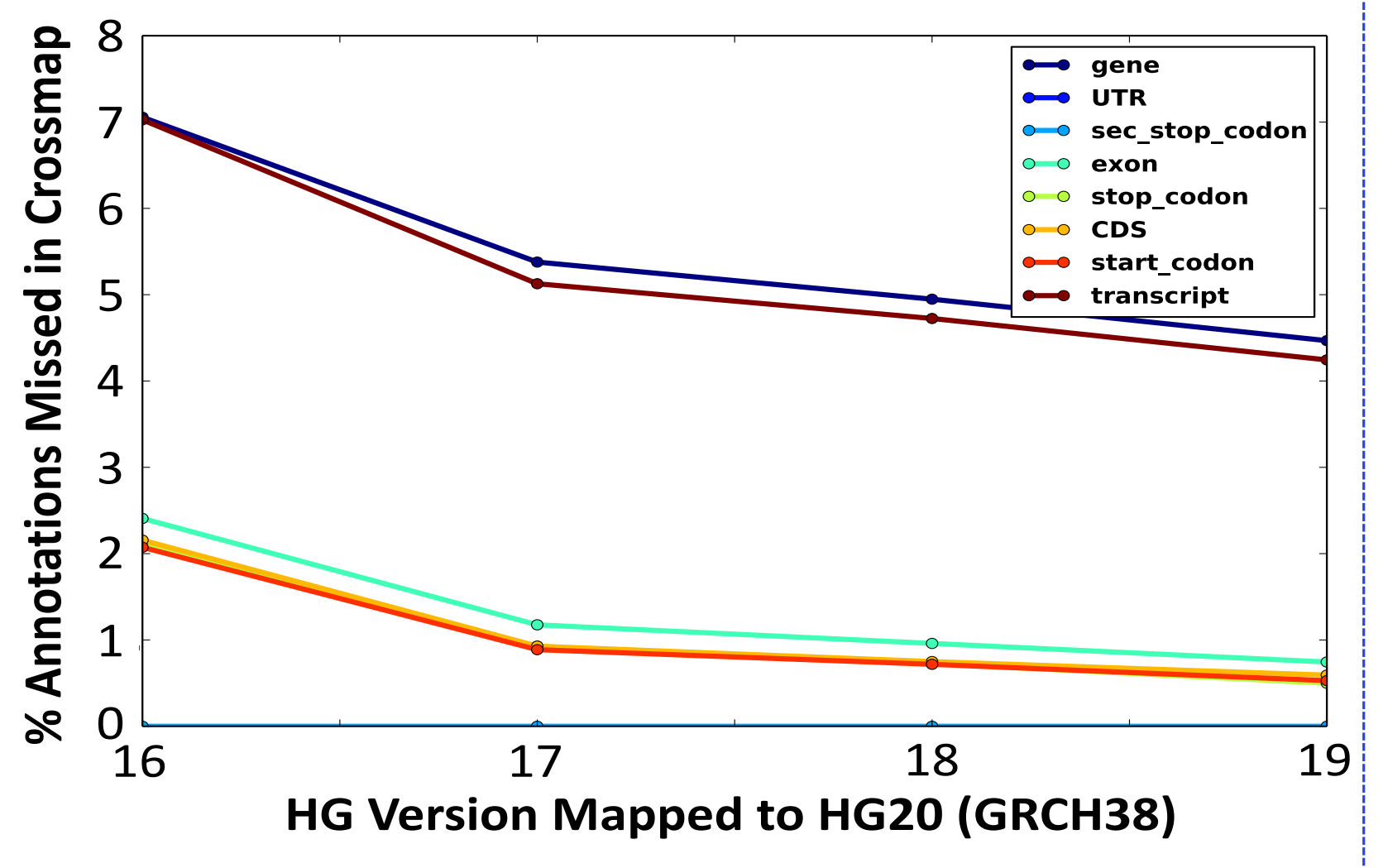
2. More and more samples are being sequenced, resulting in ***significantly many read data sets that must be mapped to the latest reference genome***.

3. Many remapping tools update **only the coordinates** between regions that exist both in the old and updated reference genome. ***While existing remappers remap reads quickly, a lot of important information is lost***.

4: Motivation



Percentage of missing annotations when remapping from one reference genome (x-axis) to the latest (GRCh38).



Limitations of Existing Remapping Tools. Existing remapping tools correctly remap reads that mapped completely within a region indicated by the chain file (e.g., Read 2). However, these tools:

1) Cannot remap reads that mapped within a region in the old reference that does not appear in the new reference (e.g., Read 1)

2) May incorrectly remap reads that align to multiple constant regions in the old reference (e.g., Read 3).

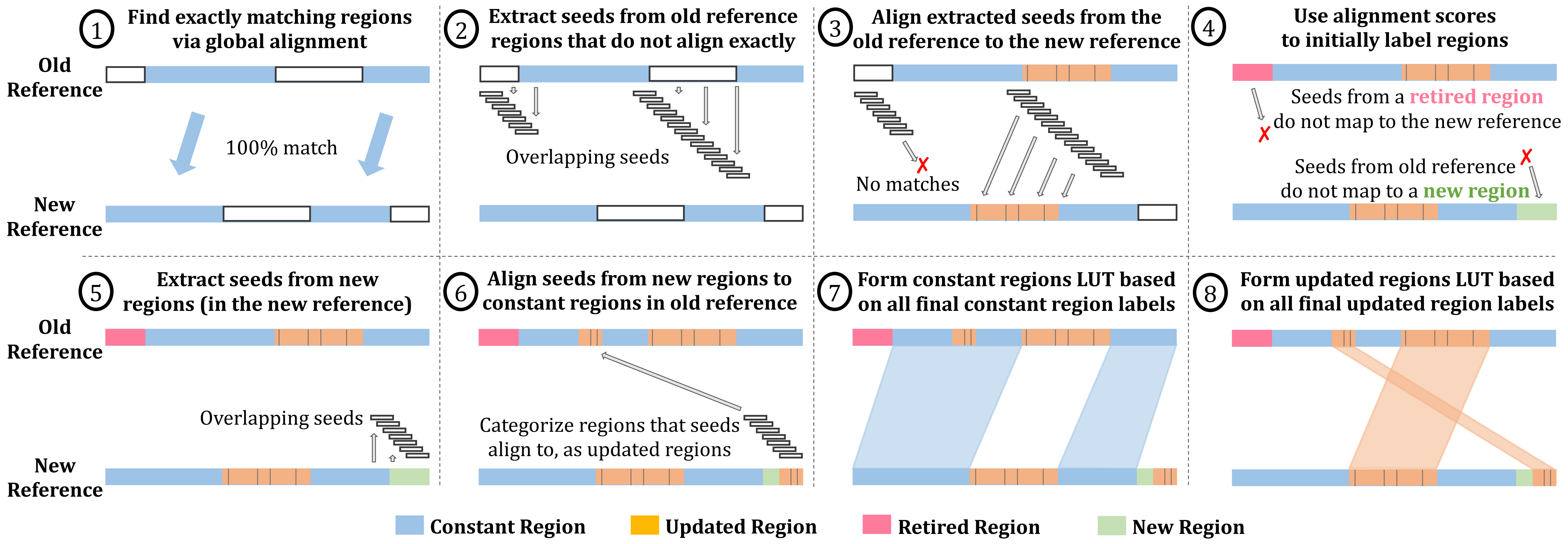
5: Our Goal

Our **goal** is to provide a set of tools and methodology for reducing the execution time of mapping a set of reads to one reference genome, when it had already been mapped to another similar reference genome.

We want to provide:

1. Similar ***execution time*** as existing remapping tools on a read data set.
2. Similar ***accuracy*** as if we run read mapper on the entire read data set to the updated reference genome.

6: Constructing AirLift Data Structures

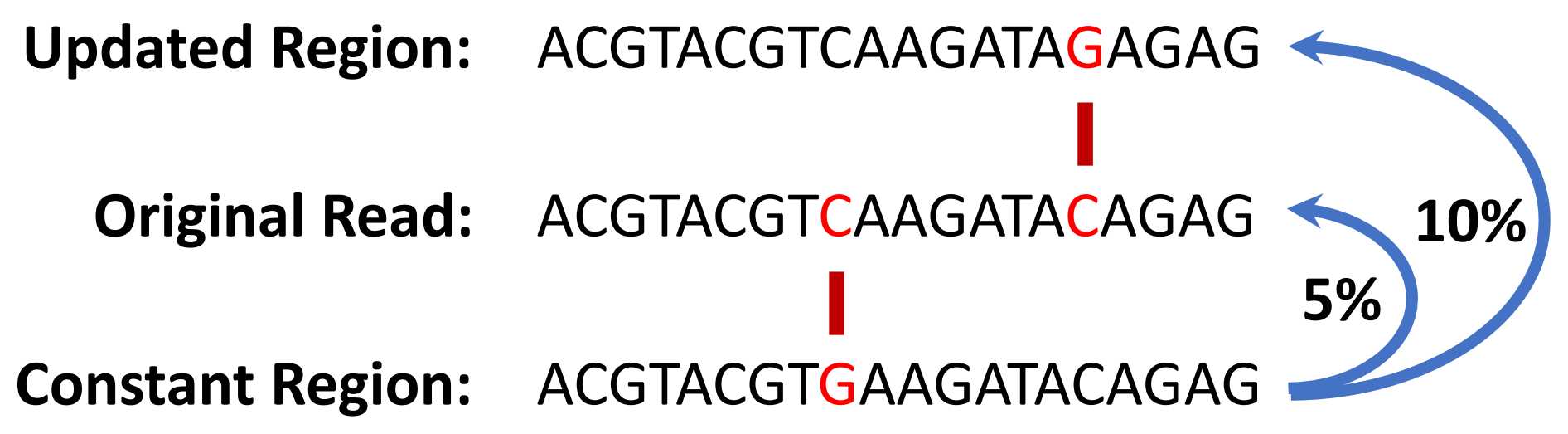

1 Find exactly matching regions via global alignment
Old Reference
100% match
New Reference
2 Extract seeds from old reference regions that do not align exactly
Overlapping seeds
3 Align extracted seeds from the old reference to the new reference
No matches
4 Use alignment scores to initially label regions
Seeds from a retired region do not map to the new reference
Seeds from old reference do not map to a new region
5 Extract seeds from new regions (in the new reference)
Old Reference
Overlapping seeds
New Reference
6 Align seeds from new regions to constant regions in old reference
Categorize regions that seeds align to, as updated regions
7 Form constant regions LUT based on all final constant region labels
8 Form updated regions LUT based on all final updated region labels
Constant Region
Updated Region
Retired Region
New Region

7: Comprehensive Remapping

For an assumed acceptable error rate of *e* (e.g., 5%), we consider reads with less than an *e* (e.g., 5%) difference from a sequence in any region a match.

Since we already have mappings of reads to constant regions in an updated reference genome, we can quickly determine candidate matches in the updated region (***constant to updated map***).

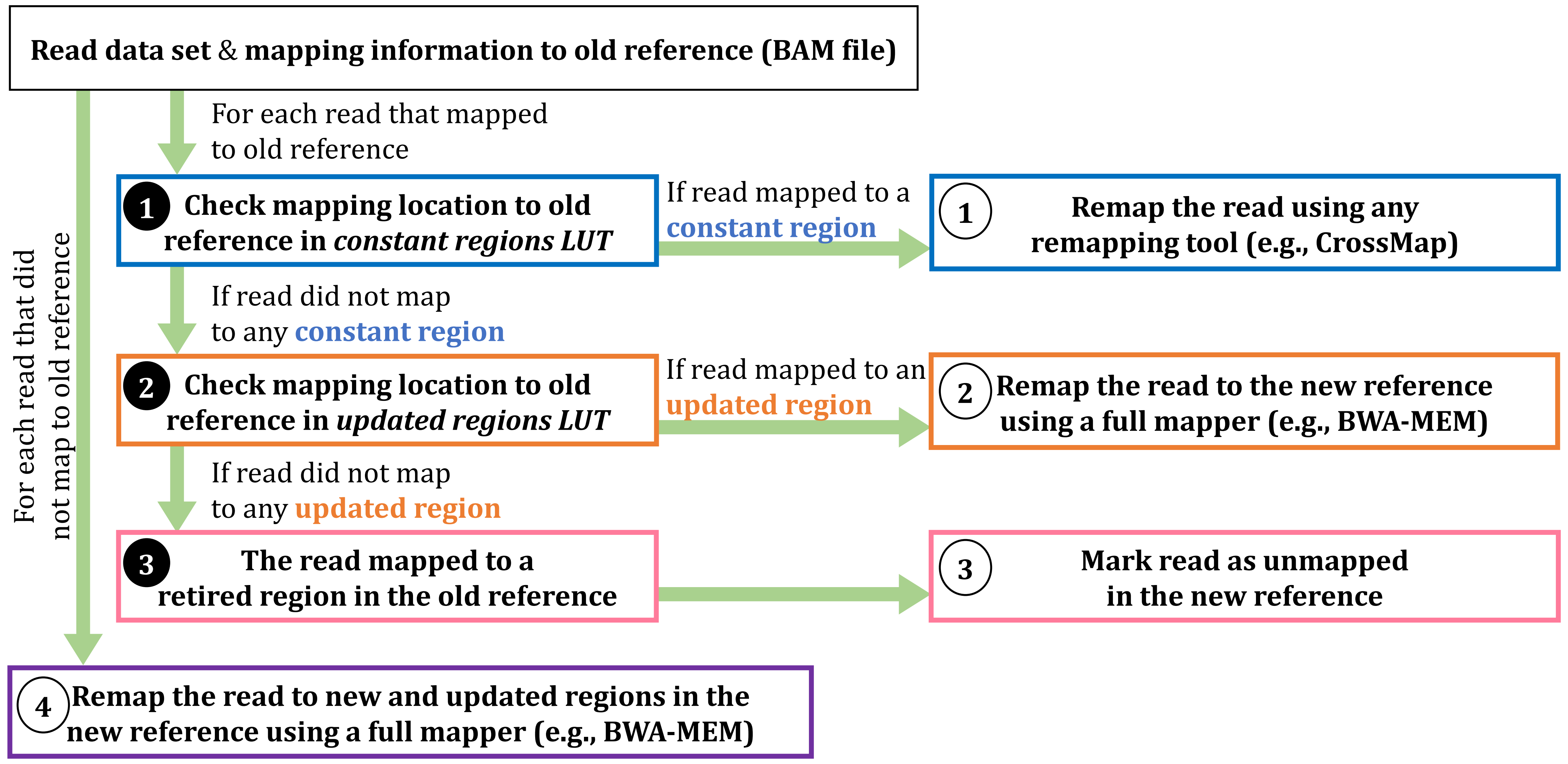
The ***constant to updated map*** contains the map from every location in the constant region to each location in the updated region that matches within a *2e* (e.g., 10%) error acceptance rate.


Updated Region: ACGTACGTCAAGATAGAGAG
Original Read: ACGTACGTCAAGATACAGAG
Constant Region: ACGTACGTGAAGATACAGAG
10%
5%

If a read maps to a location in the constant region, we know the ***exact*** candidate locations in the updated regions that we need to check for a better match.

8: AirLift Steps for Remapping Reads

AirLift remaps each read differently depending on the label of the region in the old reference that the read had originally mapped to: constant, updated, retired, or unmapped. This set of cases ensures that AirLift will comprehensively remap each read to the new reference genome as quickly and correctly as possible



Read data set & mapping information to old reference (BAM file)

For each read that mapped to old reference

1 Check mapping location to old reference in *constant regions LUT*

If read mapped to a constant region → 1 Remap the read using any remapping tool (e.g., CrossMap)

If read did not map to any constant region

2 Check mapping location to old reference in *updated regions LUT*

If read mapped to an updated region → 2 Remap the read to the new reference using a full mapper (e.g., BWA-MEM)

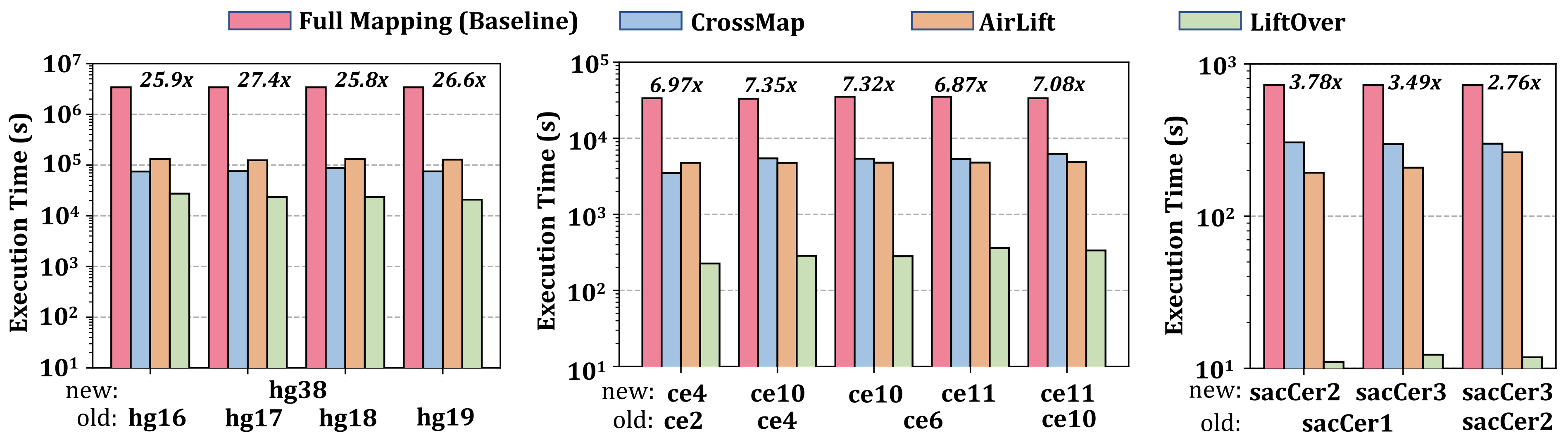
If read did not map to any updated region

3 The read mapped to a retired region in the old reference → 3 Mark read as unmapped in the new reference

For each read that did not map to old reference

4 Remap the read to new and updated regions in the new reference using a full mapper (e.g., BWA-MEM)

9: Results



Conclusion

We compare AirLift against the only comprehensive and accurate method of fully mapping a read data set to the new reference using BWA-MEM, and find **that AirLift significantly reduces the execution time by up to 27.4×, 7.35×, and 3.78×** for large (human), medium (C. elegans), and small (yeast) reference genomes, respectively. We validate our results against the ground truth and show that AirLift identifies similar rates of SNPs and Indels as the full mapping baseline. We conclude that AirLift is the first comprehensive and accurate remapping tool that substantially reduces the execution time of remapping a read data set, while providing end-to-end BAM-to-BAM results on which downstream analysis can be performed. We look forward to future works that take advantage of as well as improve AirLift for various genomic analysis studies.

SAFARI