AirLift

A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim*, **Can Firtina***, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu

bioRxiv Preprint

Source Code

SAFARI ETH zürich Carnegie Mellon

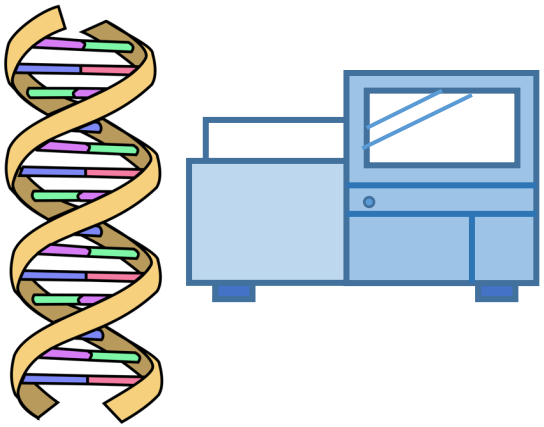
SFU SIMON FRASER UNIVERSITY

Bilkent University

Genome Analysis

- Genome analysis is critical for many applications
  - Personalized medicine
  - Outbreak tracing
  - Evolutionary studies

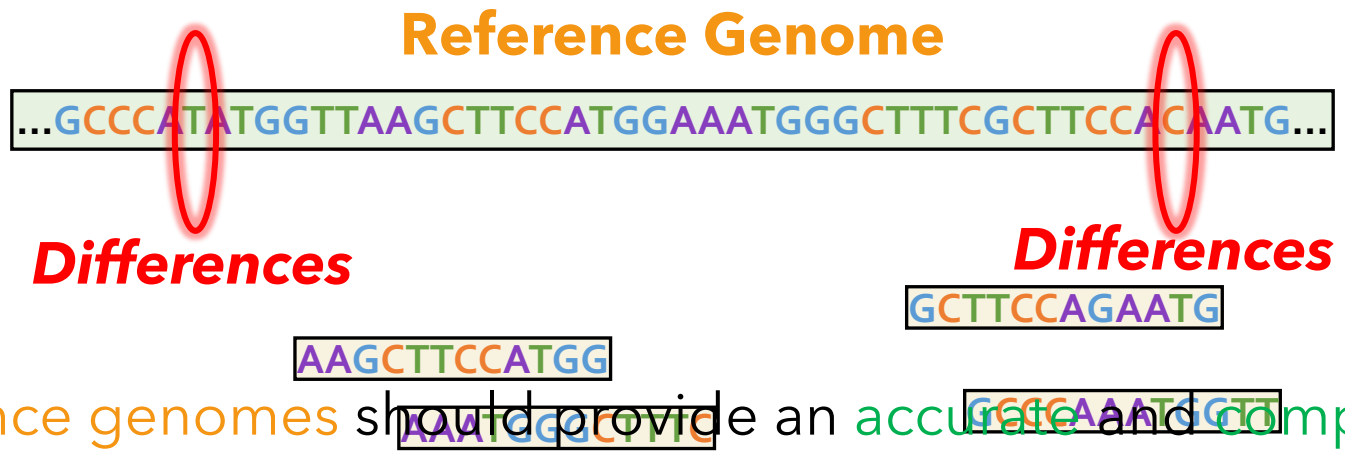
- Genome sequencing machines extract smaller fragments of the original DNA sequence, known as **reads**



Reads

Reference Genomes

- **Reference genomes** play a crucial role in genome analysis for
  - Accurately mapping reads to potential matching locations in the genome
  - Identifying genomic differences in an individual's genome

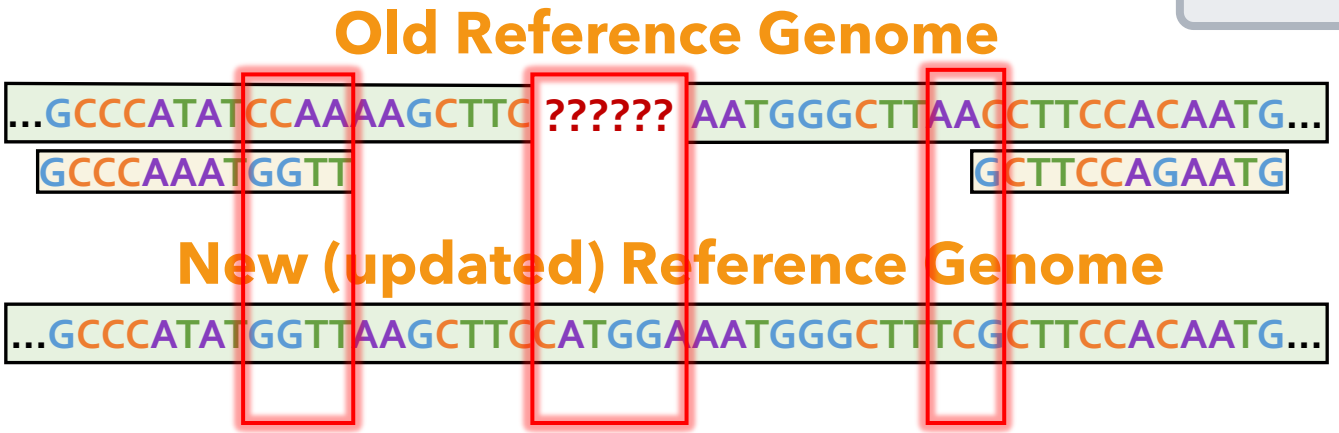


- Reference genomes should provide an accurate and complete representation of a species to **enable accurate analysis in the later steps of genome analysis:**
  - Variant calling
  - Gene annotation and enrichment

Updating the Reference Genomes

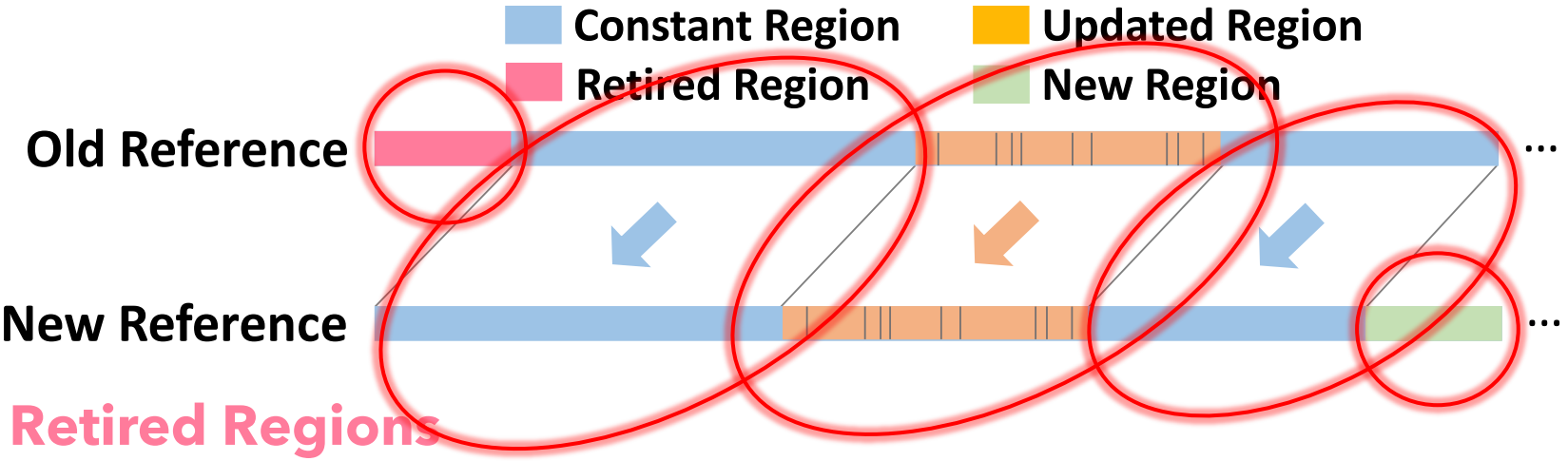
- Reference genomes are updated **regularly** to
  - **Correct the errors** in the older versions
  - **Fill in** the missing genomic sequences

Unmapped Reads



- **Remapping the reads** to the updated reference genome can generate **novel information** due to
  - More **accurately** identified genomic differences
  - **New reads mapped** to updated or completed regions

Changes between Reference Genomes



1. **Retired Regions**
   - **Removed** from the new reference genome
2. **New Regions**
   - **Added** to the new reference genome
3. **Constant Regions**
   - Exactly the **same sequences**
   - **Positions may change**
4. **Updated Regions**
   - Mostly the same sequences with **small changes**

Existing Solutions for Remapping Reads

1. **Map all the reads** from scratch

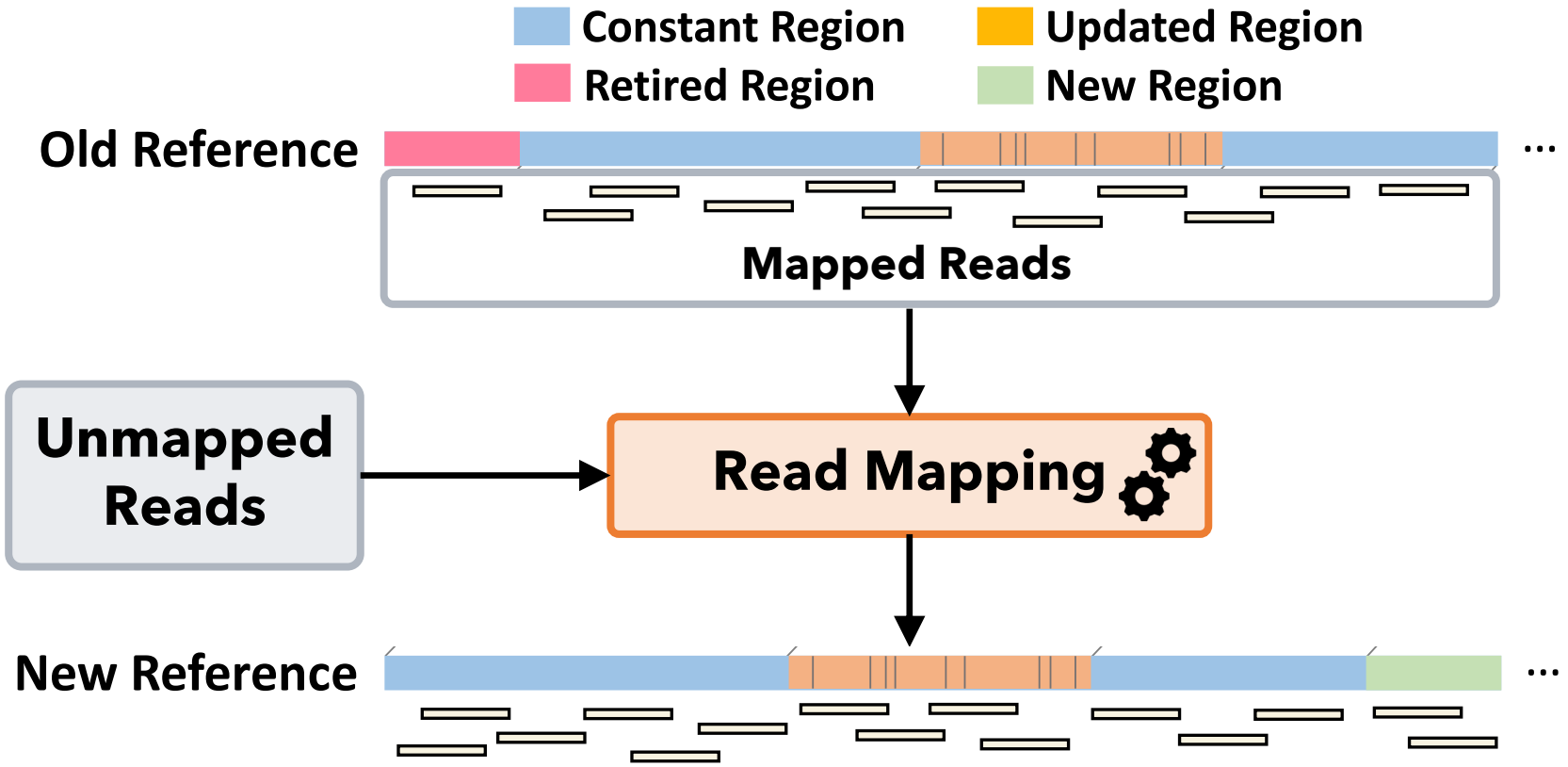
2. **Move** the mapping locations

Existing Solutions for Remapping Reads

1 **Map all the reads** from scratch

2 **Move** the mapping locations

Mapping Reads from Scratch

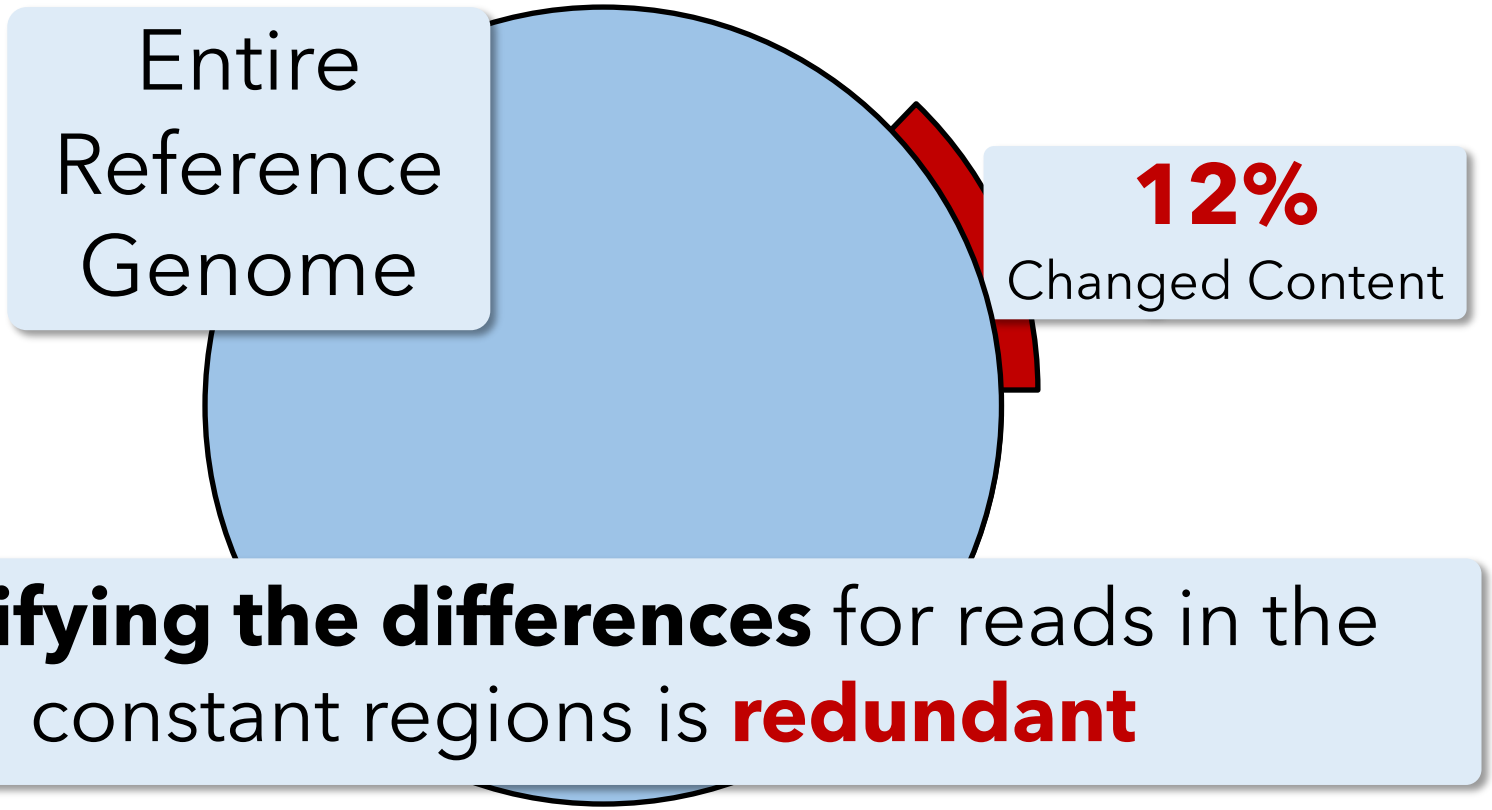


✓ Accurate mapping

✗ Significant computation overhead

Mapping Reads from Scratch

A large portion of the reference genome **remains unchanged (constant regions)**



Identifying the differences for reads in the constant regions is **redundant**

Existing Solutions for Remapping Reads

1 **Map all the reads** from scratch

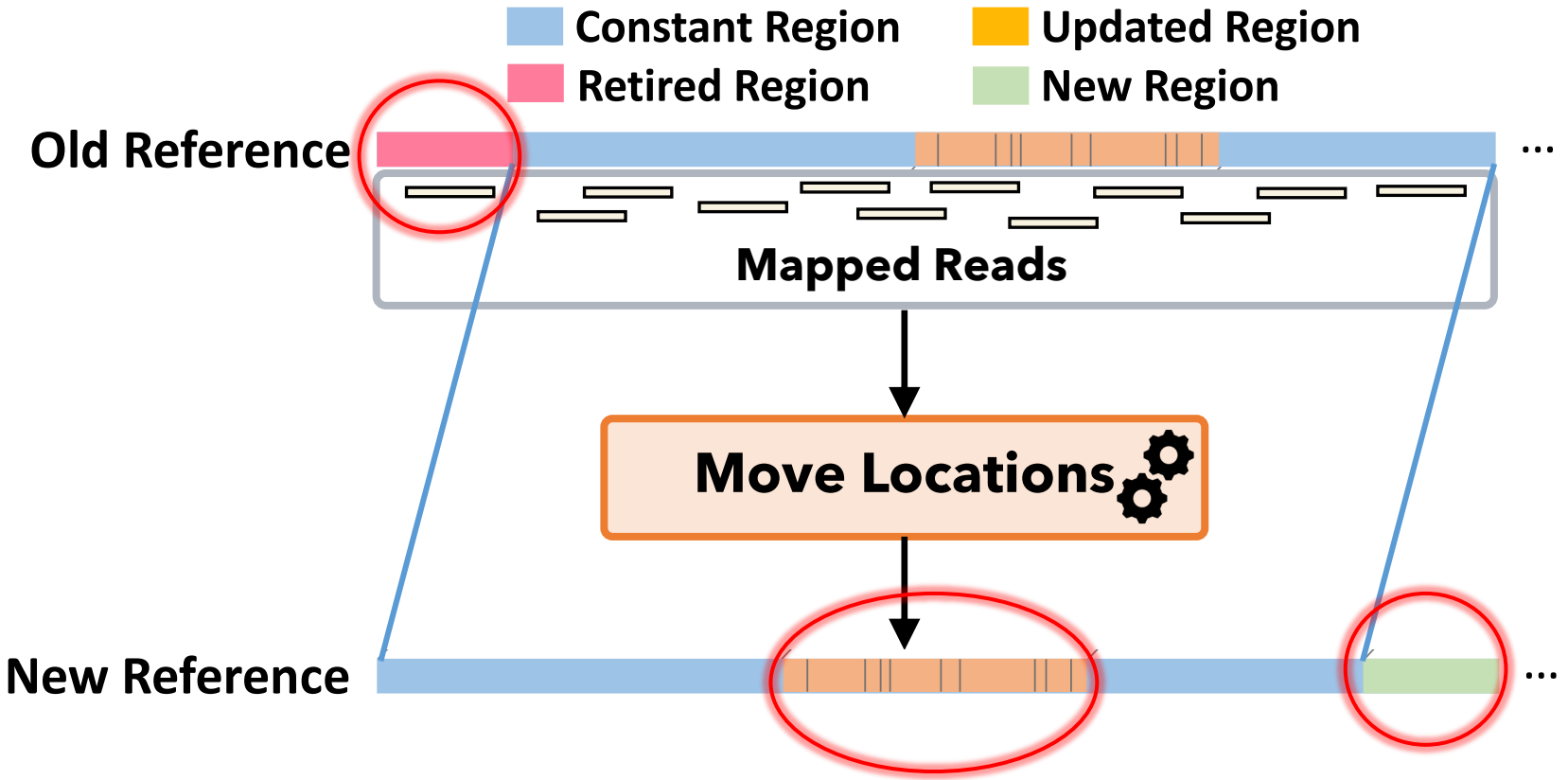
2 **Move** the mapping locations

Existing Solutions for Remapping Reads

1 **Map all the reads** from scratch

2 **Move** the mapping locations

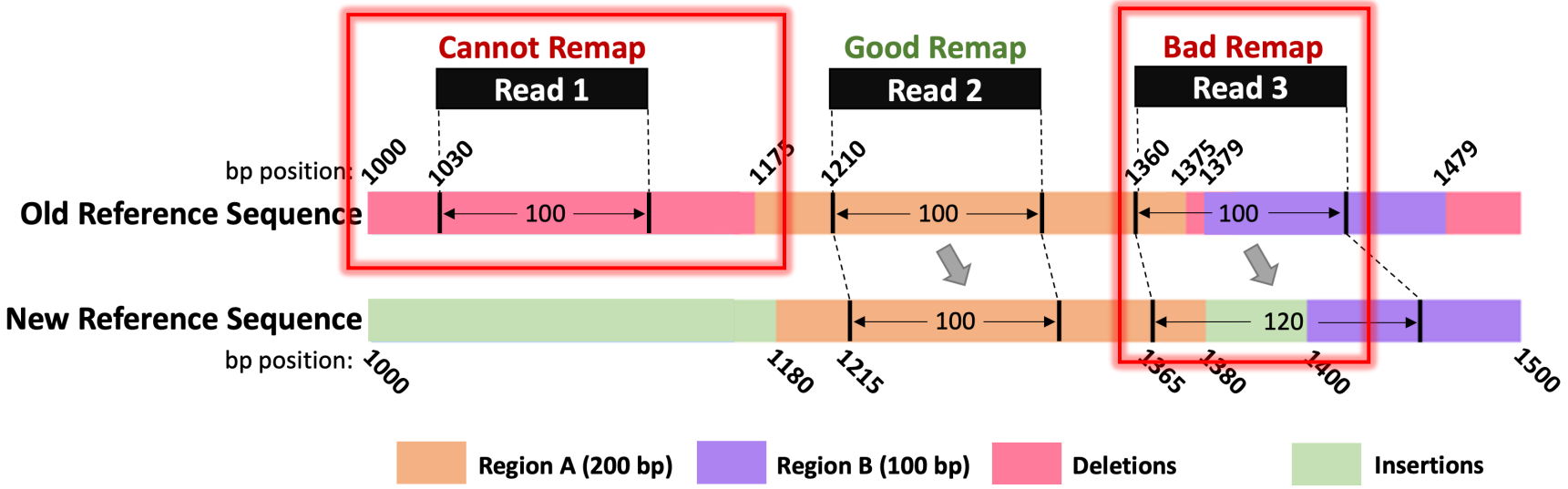
Moving the Mapping Locations



✓ Minimal computation overhead

✗ Inaccurate Mapping

Moving the Mapping Locations



- **Cannot Remap:** Reads in the **deleted regions** are not remapped
- **Bad Remap:** Reads in the **updated regions** may map other regions better

A large portion of the **mapping information is lost or inaccurate**

Outline

Background

Goal and Key Idea

AirLift

Evaluation

Conclusions

Our Goal

Accurately and quickly remap **all reads**
by either **mapping or moving** them
from the old reference genome
to the new reference genome

AirLift

Avoids redundant read mapping
for the constant regions

Quickly **identifies and maps the reads**
that cannot be accurately moved

Outline

Background

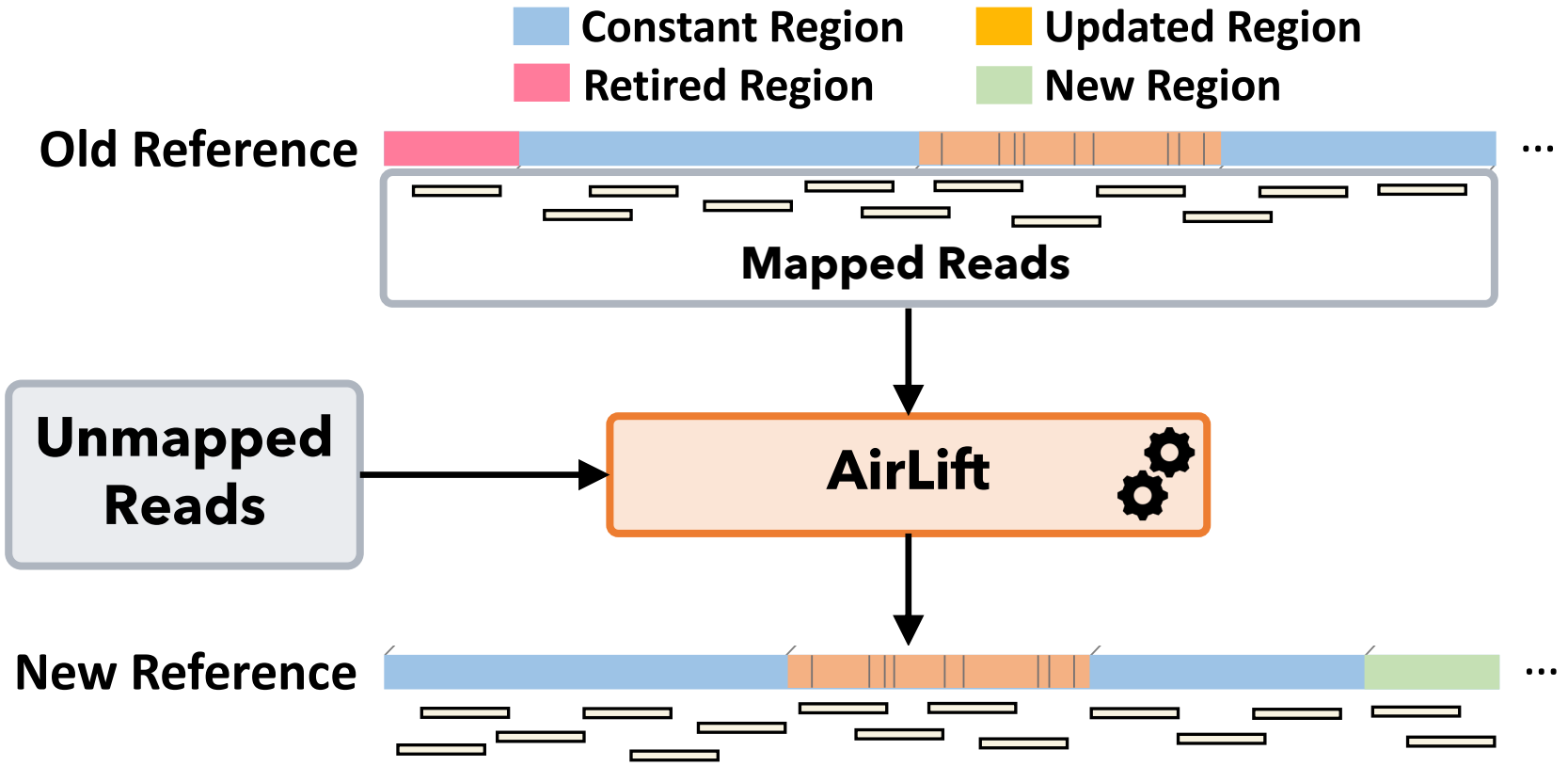
Goal and Key Idea

AirLift

Evaluation

Conclusions

AirLift Overview



✓ Low computation overhead

✓ Accurate Mapping

AirLift

AirLift Indexing (Offline)

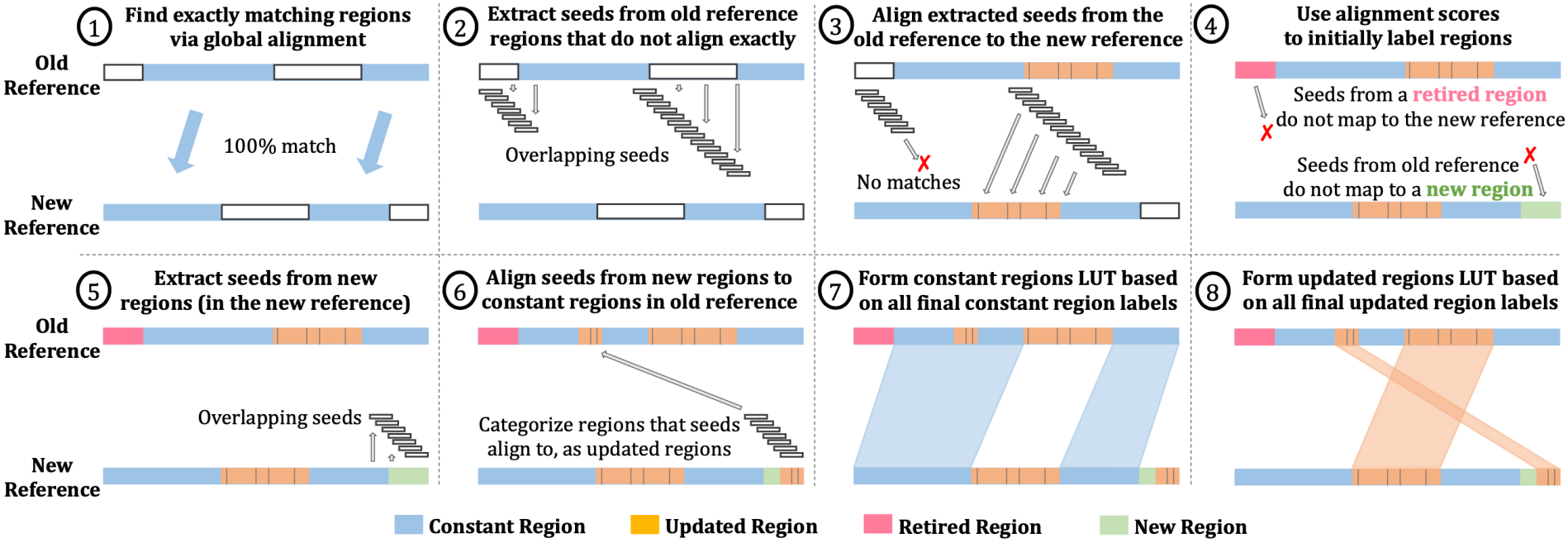
AirLift Mapping

AirLift

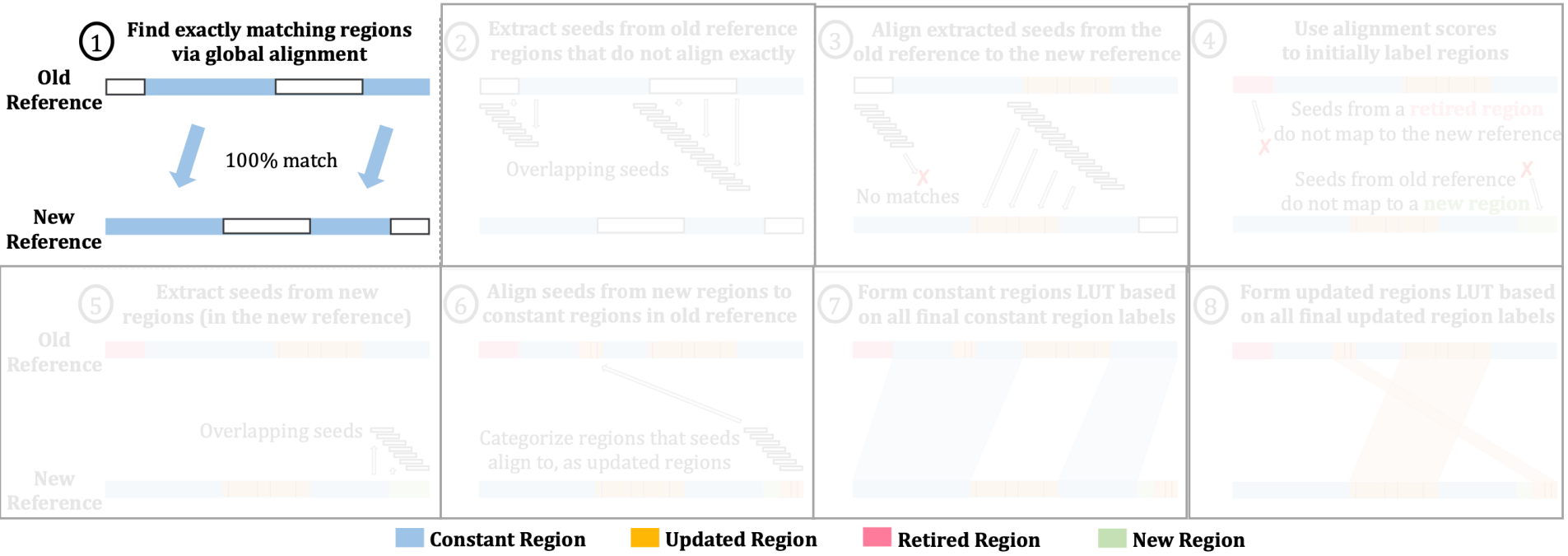
AirLift Indexing (Offline)

AirLift Mapping

AirLift Indexing (Offline)



AirLift Indexing (Offline)

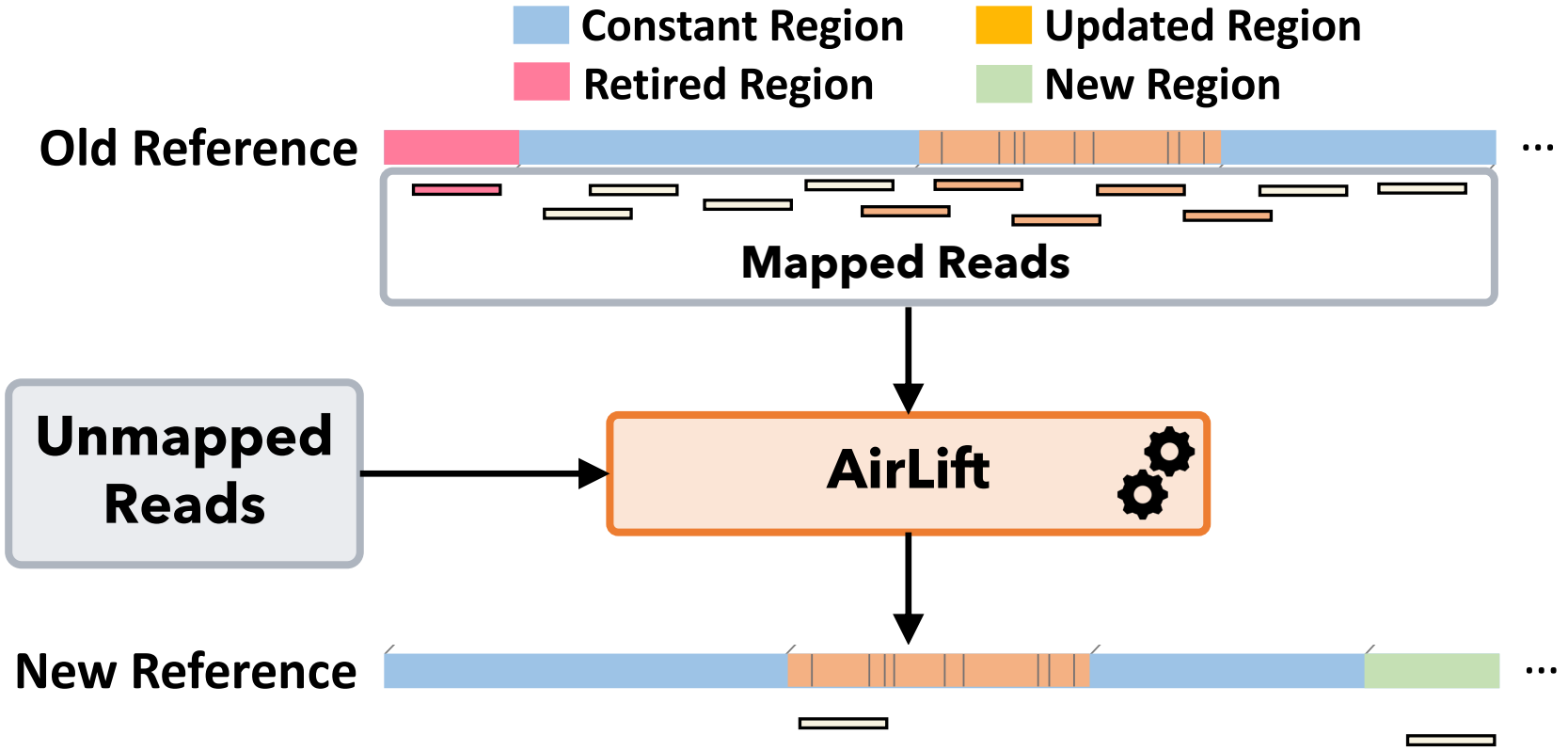


AirLift

AirLift Indexing (Offline)

AirLift Remapping

AirLift Remapping



Quickly **move** reads in the **constant** regions

Remap reads in the **updated** regions

Remap retired and **unmapped** reads

AirLift Remapping

✓ **AirLift fully utilizes all reads by either moving or remapping them**

✓ **AirLift generates an accurate alignment file (BAM) that can easily be used in downstream analysis**

Outline

Background

Goal and Key Idea

AirLift

Evaluation

Conclusions

Evaluation Methodology

Remapping

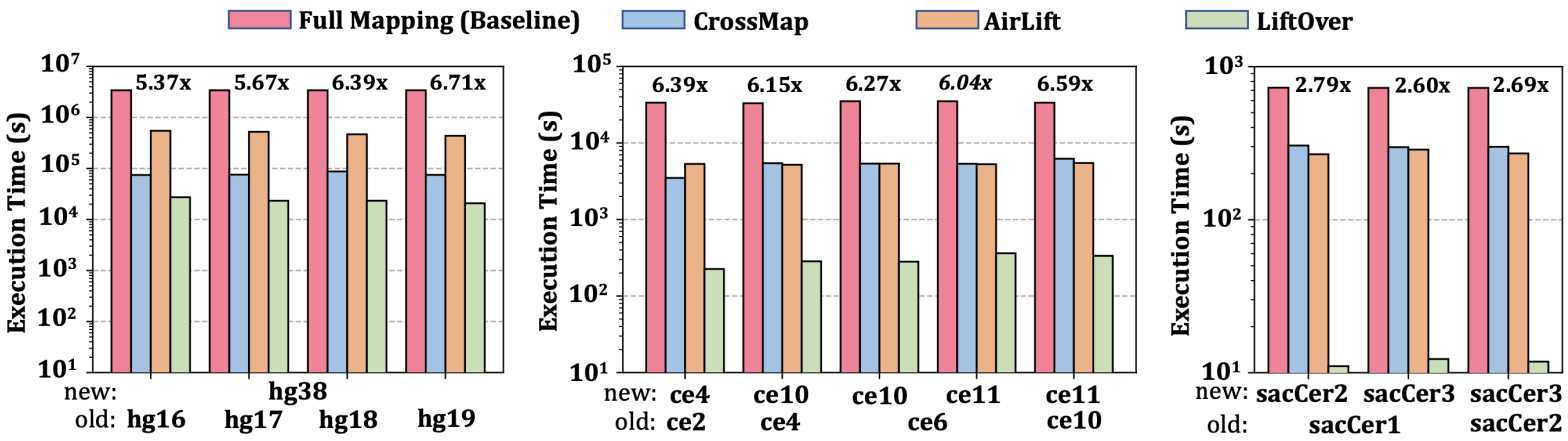
- **Baseline:** Fully mapping all reads
  - CrossMap remapper that can generate alignment files (BAM)
  - LiftOver remapper that generates only the updated positions

Accuracy: **Variant calling** using AirLift and full mapping

Datasets

- **Human (hg):** Oldest: HG16 Newest: HG38 (5 versions)
- **Worm (ce):** Oldest: ce2 Newest: ce11 (5 versions)
- **Yeast (sacCer):** Oldest: sacCer1 Newest: sacCer3 (3 versions)

Performance

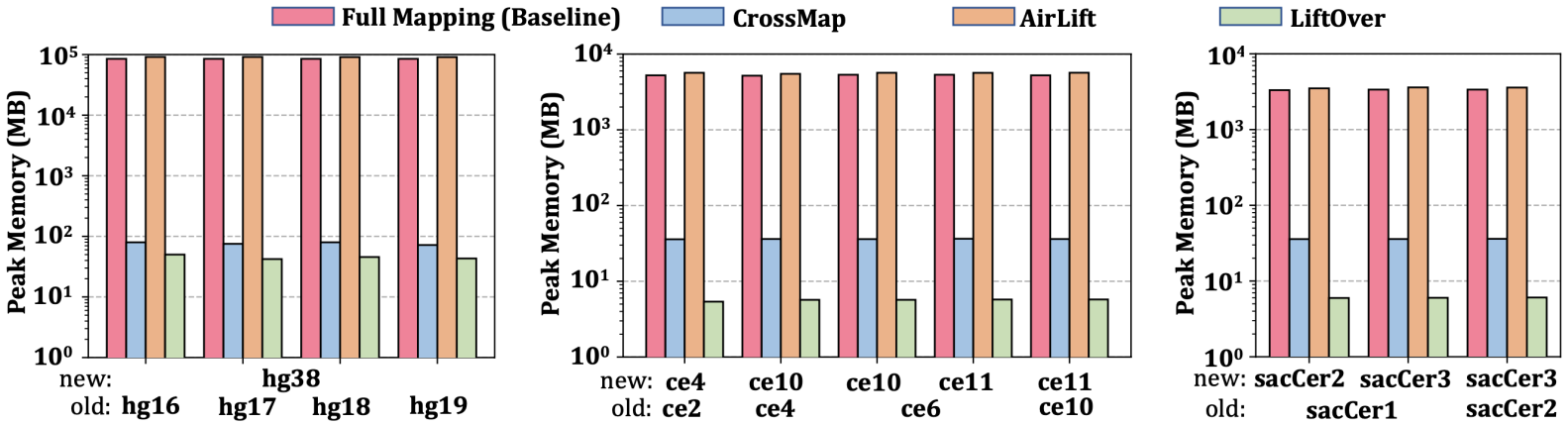


2.6× – 6.7× speedup compared to the full mapping

More comprehensive mapping:

Longer execution times than CrossMap and LiftOver

Peak Memory Usage



Peak memory usage similar to full mapping

Accuracy – Variant Calling

Precision/Recall values compard to
- Ground truth
- Full mapping

	Remap Technique	Read Sets from	Read Sets to	vs. Full Mapping SNP (%)	vs. Full Mapping Indel (%)	vs. Ground Truth SNP (%)	vs. Ground Truth Indel (%)
Baseline:	Full Mapping	-	hg38	-	-	99.54/88.00	81.31/92.38

Comparable accuracy to full mapping without the significant performance cost

Outline

Background

Goal and Key Idea

AirLift

Evaluation

Conclusions

AirLift Summary

Problem	Remapping to a new reference genome is either **costly (full mapping)** or **inaccurate (moving mapping positions)**
Goal	**Accurately and quickly** remap **all reads** by either **mapping or moving** them from the old reference genome to the new reference genome
AirLift	• **AirLift Indexing: Accurately categorize and label each region** in the old reference genome compared to the new reference genome • **AirLift Remapping:** 1. Remap a read to a new reference genome or 2. Quickly move its position based on **AirLift index**
Key Results	AirLift **consistently outperforms full mapping** • **2.6x – 6.7x speedup** over full mapping AirLift **identifies SNPs and INDELs with precision and recall similar to full mapping**

AirLift

- Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu,
  "AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes"
  *Preprint in **arXiv** and **bioRxiv**, 2022.*
  [bioRxiv preprint]
  [arXiv preprint]
  [AirLift Source Code and Data]

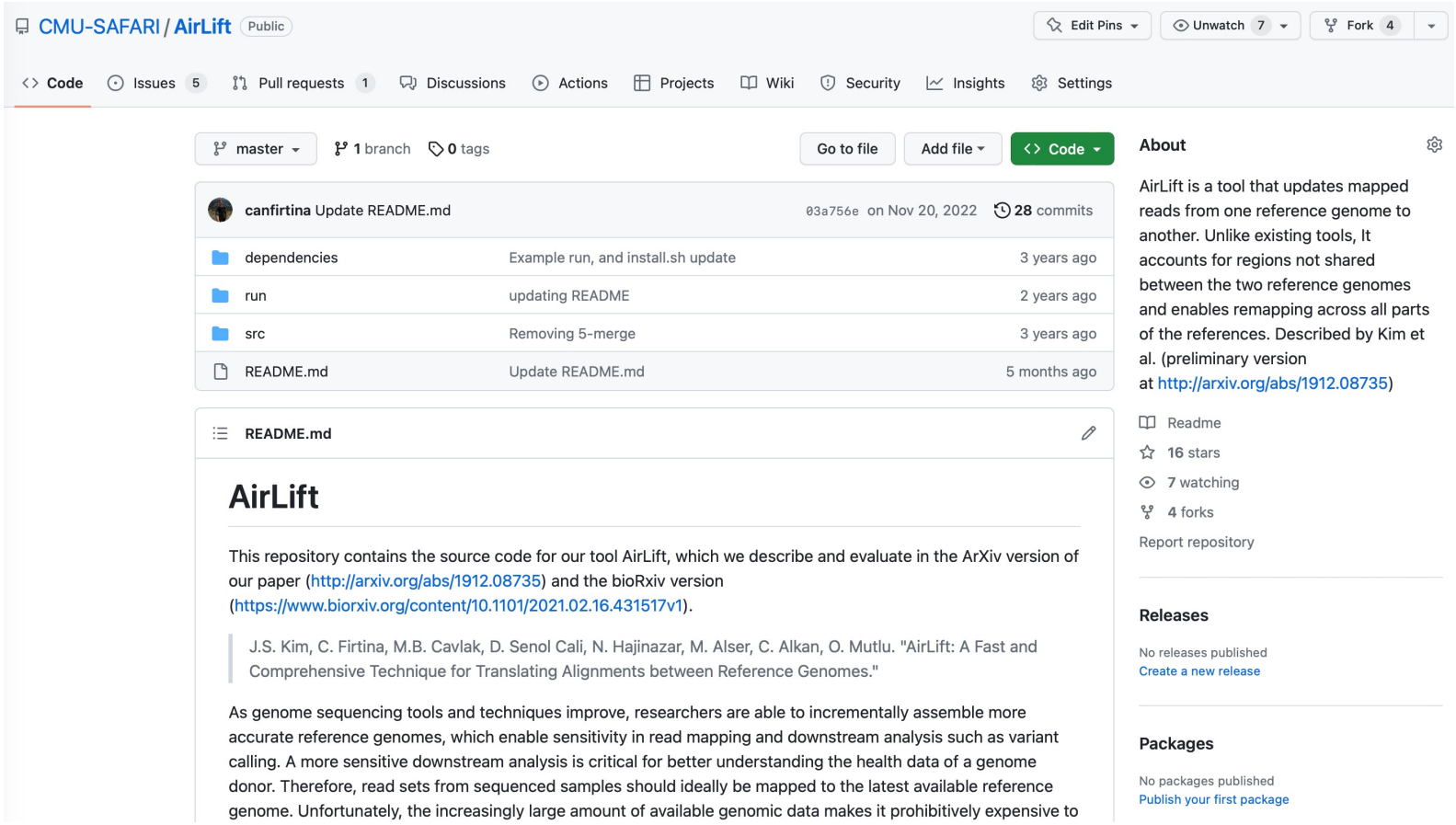
bioRxiv Preprint

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] Onur Mutlu[1,2,4]
[1]*ETH Zurich* [2]*Carnegie Mellon University* [3]*Simon Fraser University* [4]*Bilkent University*

AirLift Source Code

Source Code



CMU-SAFARI / AirLift Public

Edit Pins | Unwatch 7 | Fork 4

Code | Issues 5 | Pull requests 1 | Discussions | Actions | Projects | Wiki | Security | Insights | Settings

master | 1 branch | 0 tags | Go to file | Add file | Code

canfirtina Update README.md — 03a756e on Nov 20, 2022 — 28 commits

dependencies	Example run, and install.sh update	3 years ago
run	updating README	2 years ago
src	Removing 5-merge	3 years ago
README.md	Update README.md	5 months ago

README.md

AirLift

This repository contains the source code for our tool AirLift, which we describe and evaluate in the ArXiv version of our paper (http://arxiv.org/abs/1912.08735) and the bioRxiv version (https://www.biorxiv.org/content/10.1101/2021.02.16.431517v1).

> J.S. Kim, C. Firtina, M.B. Cavlak, D. Senol Cali, N. Hajinazar, M. Alser, C. Alkan, O. Mutlu. "AirLift: A Fast and Comprehensive Technique for Translating Alignments between Reference Genomes."

As genome sequencing tools and techniques improve, researchers are able to incrementally assemble more accurate reference genomes, which enable sensitivity in read mapping and downstream analysis such as variant calling. A more sensitive downstream analysis is critical for better understanding the health data of a genome donor. Therefore, read sets from sequenced samples should ideally be mapped to the latest available reference genome. Unfortunately, the increasingly large amount of available genomic data makes it prohibitively expensive to

About

AirLift is a tool that updates mapped reads from one reference genome to another. Unlike existing tools, It accounts for regions not shared between the two reference genomes and enables remapping across all parts of the references. Described by Kim et al. (preliminary version at http://arxiv.org/abs/1912.08735)

Readme
16 stars
7 watching
4 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

https://github.com/CMU-SAFARI/AirLift

AirLift

A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim*, **Can Firtina***, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu

bioRxiv Preprint

Source Code

SAFARI ETH zürich Carnegie Mellon

SFU SIMON FRASER UNIVERSITY

Bilkent University

Backup Slides

AirLift Remapping