cfSNV: a software package for sensitive detection of somatic mutations from cell-free DNA

Shuo Li, Ran Hu, Colin Small, Ting-Yu Kang, Chun-Chi Liu, Xianghong Jasmine Zhou, Wenyuan Li
UCLA

Presenter: Wenyuan Li
Adjunct Professor, UCLA

ITCR Annual Meeting, September 14 2022

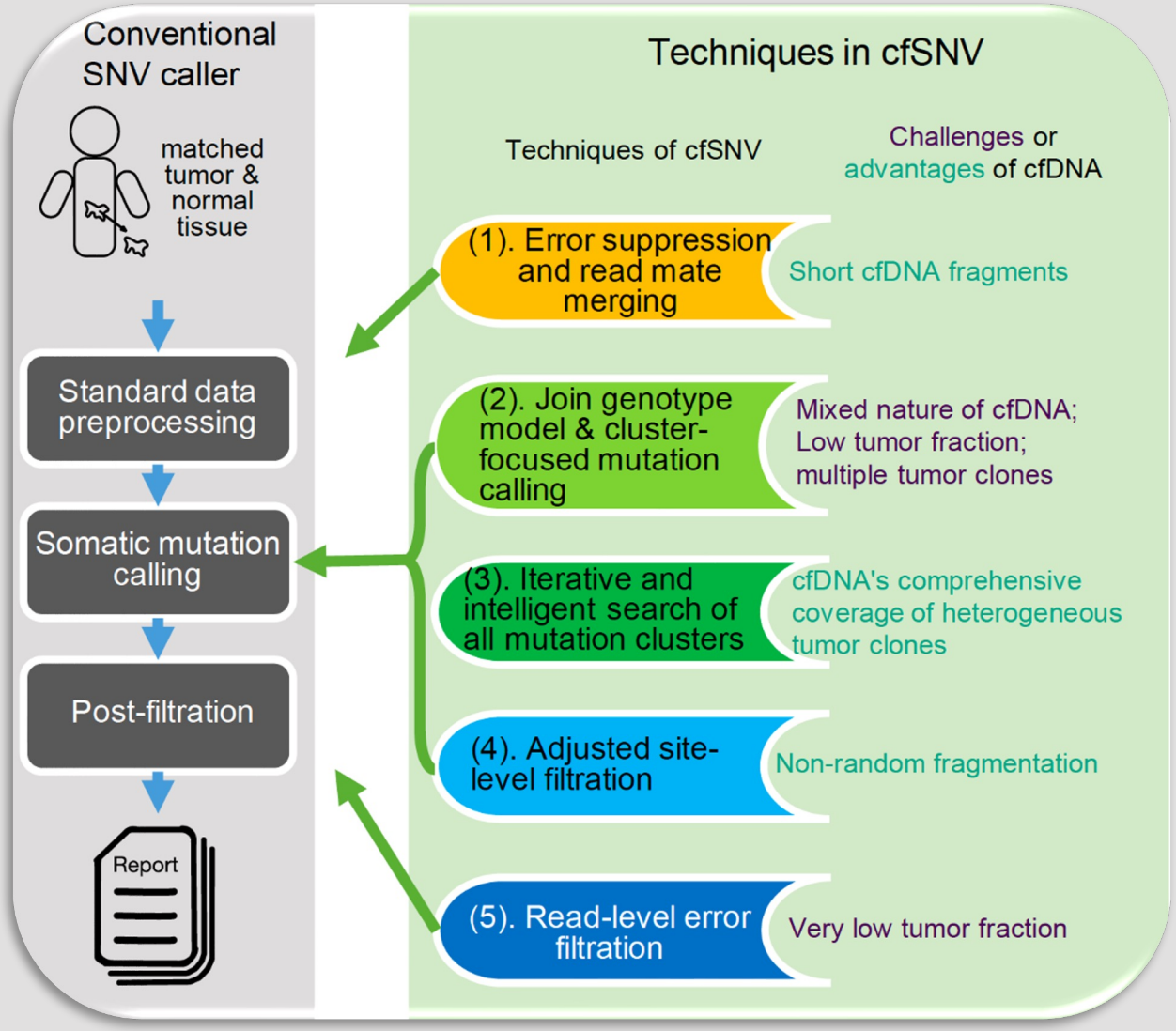
Disclosure

- This work was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers U01CA237711 (to W.L.), R01CA246329 (to X.J.Z. and W.L.), and U01CA230705 (to X.J.Z.).

- W.L. and X.J.Z. are co-founders of EarlyDiagnostics Inc.

Challenges in single nucleotide variant (SNV) detection from cell-free DNA (cfDNA)

- Low tumor cfDNA fraction
- High tumor heterogeneity
- Existing methods cannot deal with low prevalence SNVs
  - Recall: lack of modelling for tumor content and clonal hierarchy
  - Precision: insufficient site-level statistics for error control

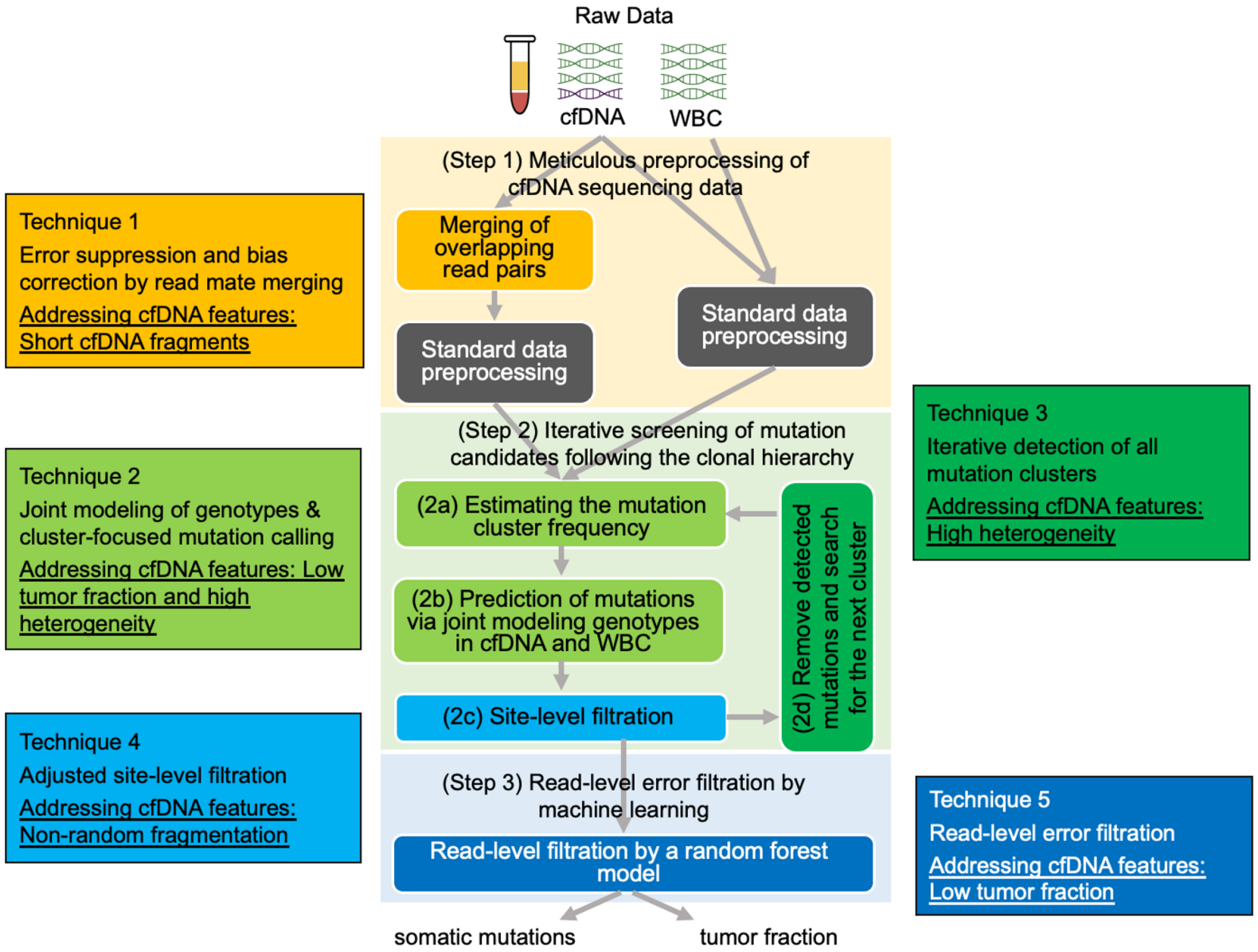


cfSNV

- ✓ cfSNV is an ultra-sensitive and accurate somatic SNV caller designed for cfDNA sequencing.
- ✓ Provide hierarchical mutation profiling and multi-layer error suppression
- ✓ Statistical model and machine learning approach

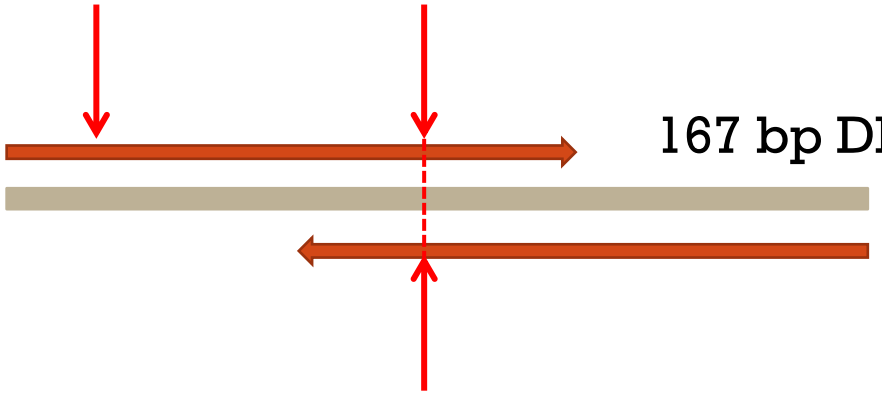
- Li S, *et al*. Sensitive detection of tumor mutations from blood and its application to immunotherapy prognosis. *Nature Communication*. 2021 Jul 7;12(1):4172.
- Li S, *et al*. cfSNV: a software tool for the sensitive detection of somatic mutations from cell-free DNA. *Nature Protocol.* 2022 Conditionally Accepted.

cfSNV workflow



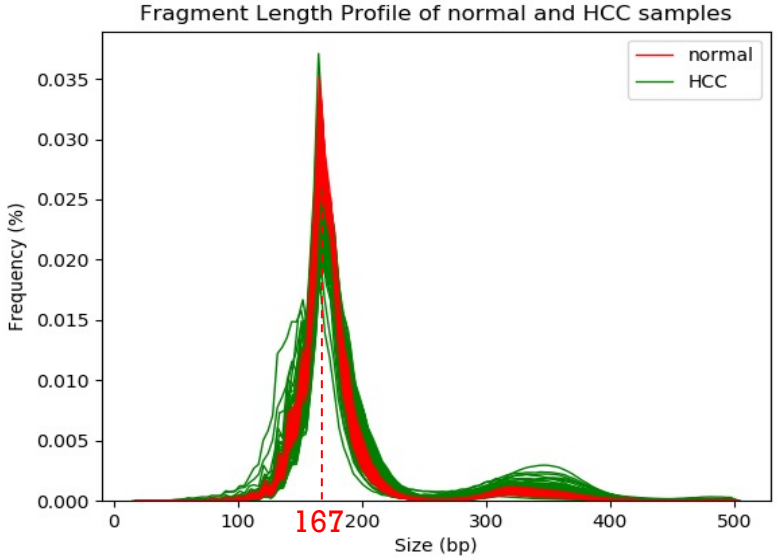
Utilize overlapping read pairs for error suppression

1. Short fragment size
2. Non-random fragmentation



167 bp DNA fragment

2 x 100 bp paired-end sequencing

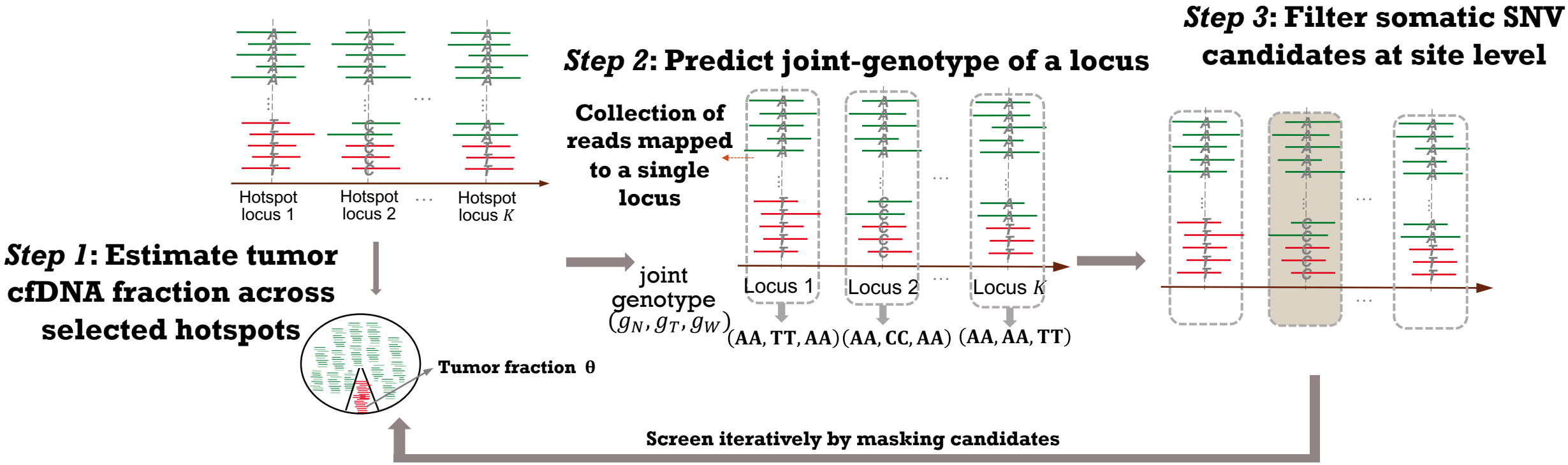


nucleosome + linker histone: 167 bp

Joint genotype modeling and iterative SNV calling

Mixed nature of cfDNA → Joint genotype model

cfDNA's coverage of all clones → iterative SNV calling



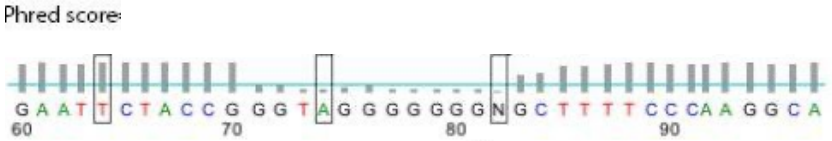
Read-level machine learning to distinguish true variants from sequencing errors for each read

Reads with variant or error from selected loci

Extract features

Read context cttcccatggattgacgccatcatgacgcatctctgccaggagggctgga

Base quality

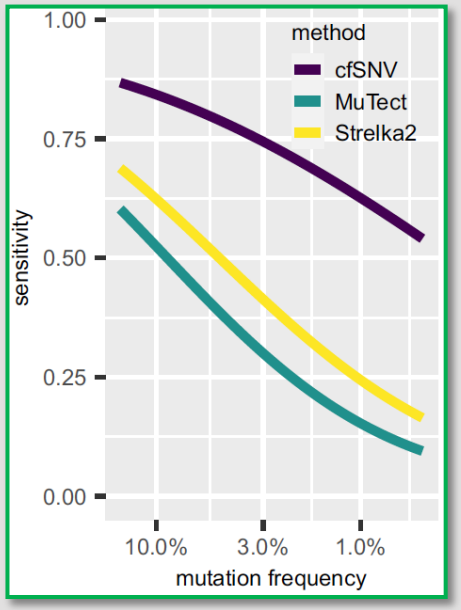


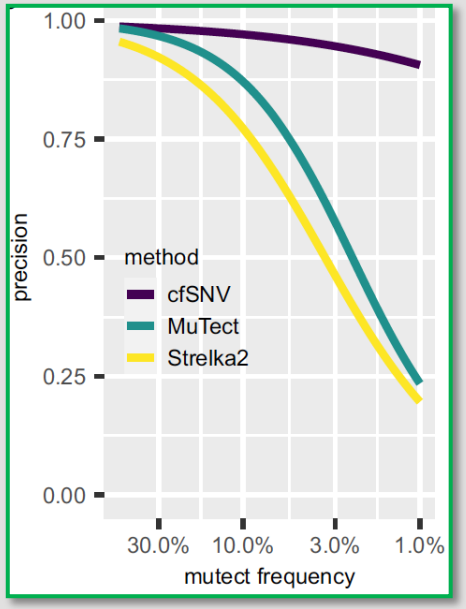
Mapping quality

Cigar string CIGAR string -> 8M2I4M1D3M

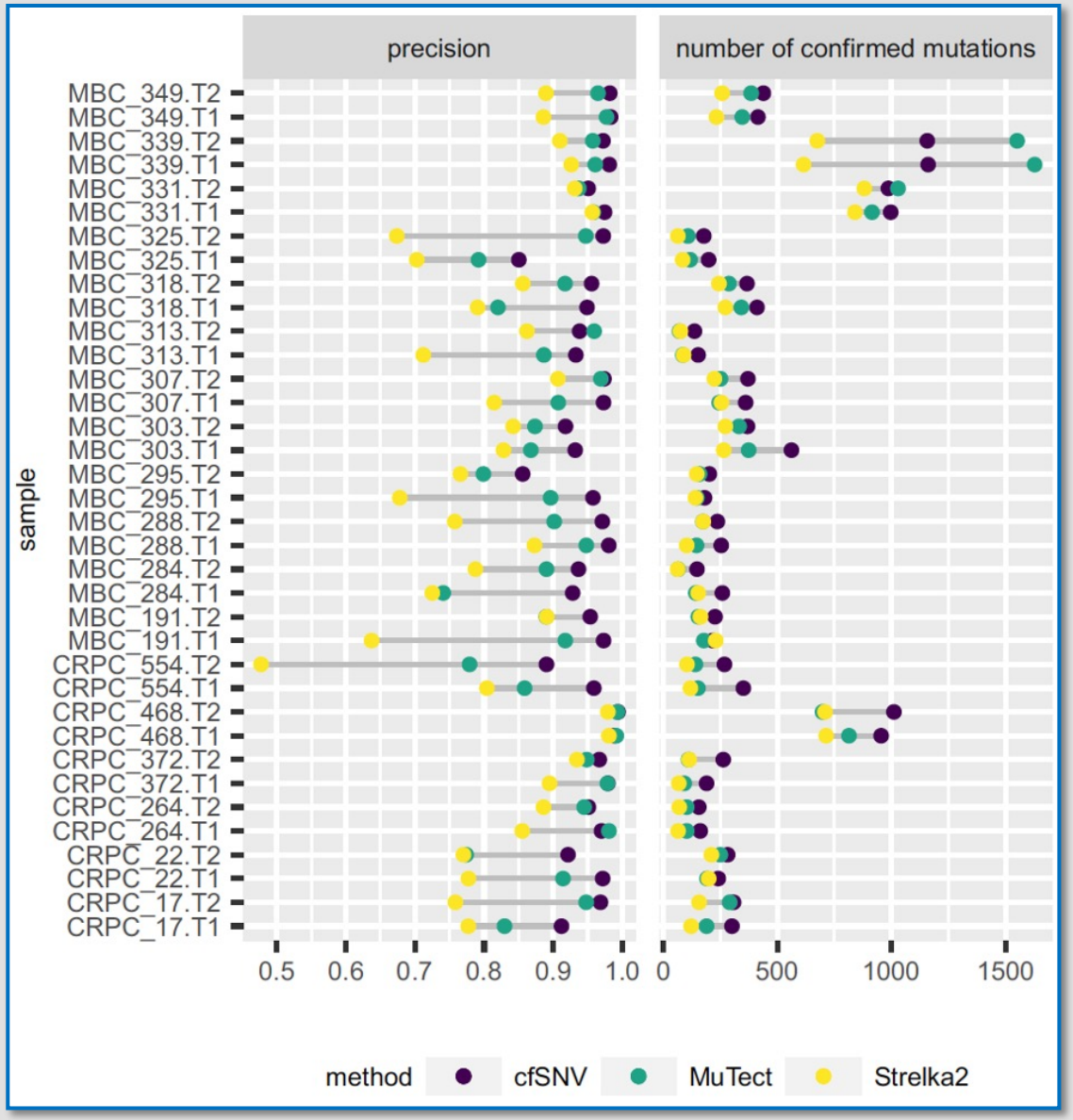
Insert size

Data ready for training





Simulation data



Real data from cancer patients

cfSNV

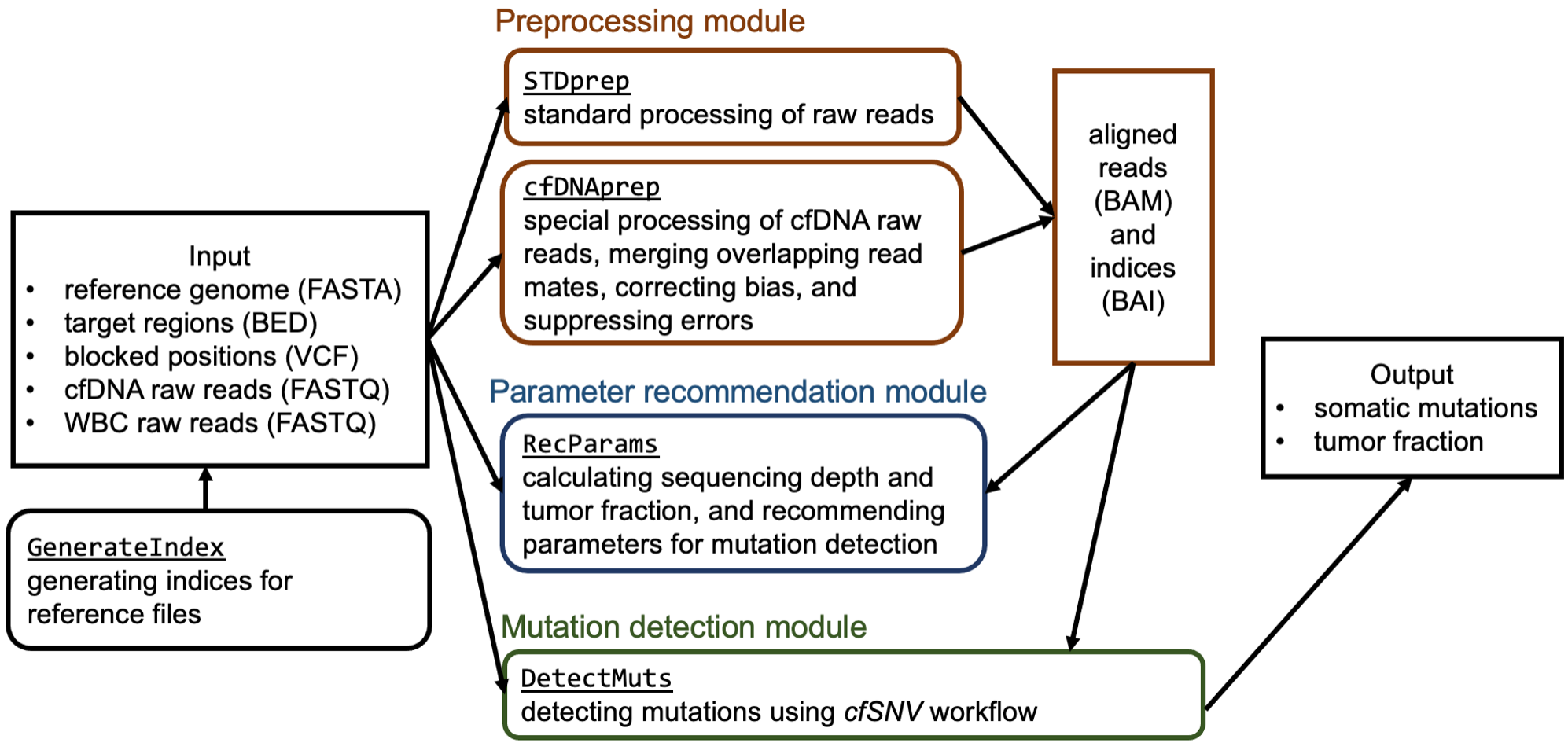
- ✓ Outperform existing tools in sensitivity while maintaining high precision.
- ✓ Improve mutation detection performance in medium-depth sequencing data, such as Whole-Exome Sequencing
- ✓ Wide-spectrum applications:
  - ❑ Cancer detection
  - ❑ Cancer monitoring
  - ❑ Therapy response prediction

Efficient implementation of cfSNV

- We implemented cfSNV by C++ and python, wrapped into an R package
- We built a Docker image, which is designed to enable researchers and clinicians with a limited computational background to easily carry out analyses on both high-performance computing platforms and local computers.
- Mutation calling from a standard preprocessed WES dataset (~250x and ~70 M target size) can be carried out in 3 hours on an Amazon Web Services cloud server with 8 vCPUs and 32 GB of RAM.
- It can automatically detect the statistics of the users' input data and recommend parameter settings that are tailored to the specific experimental protocol, sequencing coverage, and the tumor fraction of the dataset.
- cfSNV R package: https://github.com/jasminezhoulab/cfSNV
- cfSNV Docker image: https://github.com/jasminezhoulab/cfSNV_docker

Modules of cfSNV package

Three modules: (1) data preprocessing, (2) parameter recommendation, (3) mutation detection.



Limitations and future work

- Does not support detection of somatic indels
- Does not consider haplotypes of somatic mutations

  Due to short size of cfDNA fragments, it might be difficult to resolve haplotypes from cfDNA
- The module, "error suppression in the overlapping read mates", does not support single-end sequencing data
- Does not provide data preprocessing of UMI-tagged sequencing data, due to the often customized UMI design