# Structural Variant Calling in Genomes Using Deep Learning

1805010 - Anwarul Bashir Shuaib
1805036 - Abu Humayed Azim Fahmid

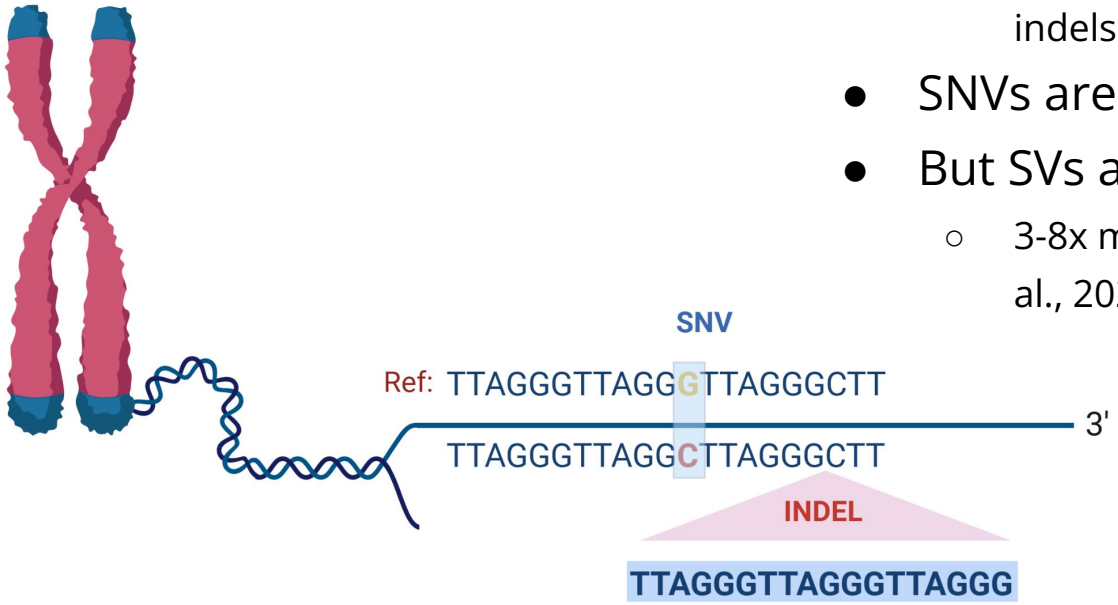
Supervisor:
**Dr. Atif Hasan Rahman**
CSE, BUET

# Introduction

# The Human Genome

- Approximately 3 billion base pairs
- 20,000-25,000 genes

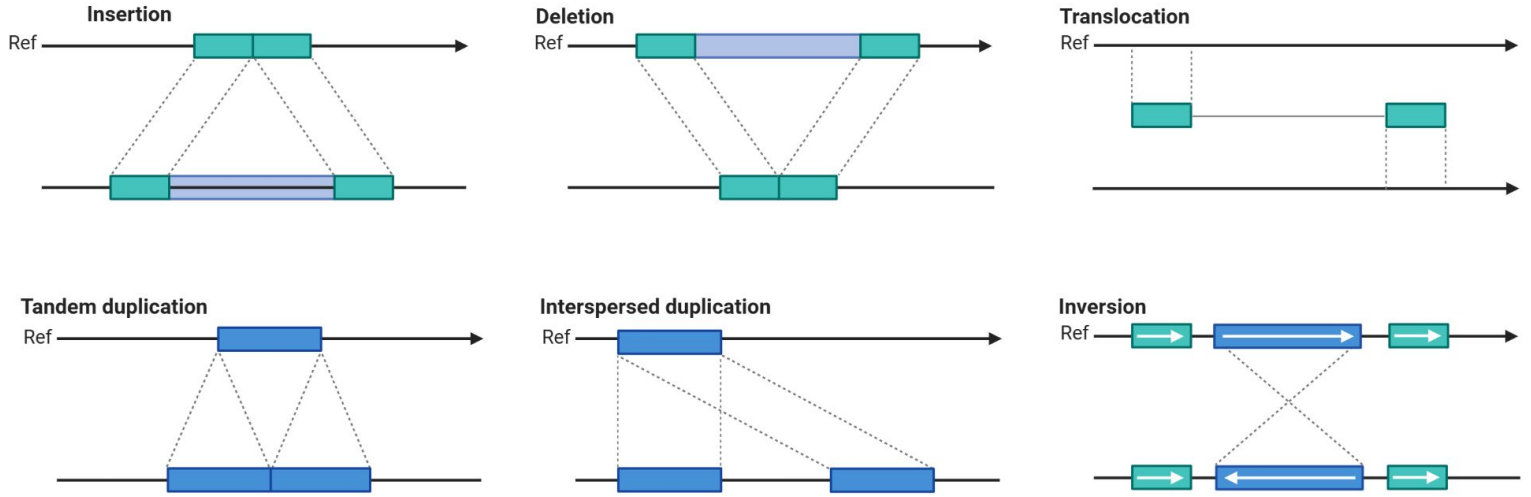
# The Human Genome




- Millions of genetic variations
  - Single Nucleotide Variations (SNV), small indels, structural variations
- SNVs are the most common
- But SVs affect more base pairs
  - 3-8x more (Catanach et al., 2019; Hämälä et al., 2021; Mérot et al., 2020)

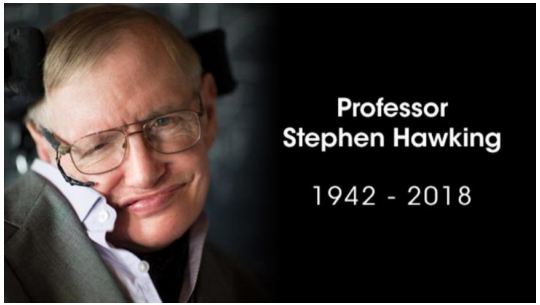
# Structural Variations

- Affect large genomic regions
- Alterations in the sequence (50bp to several kbp)



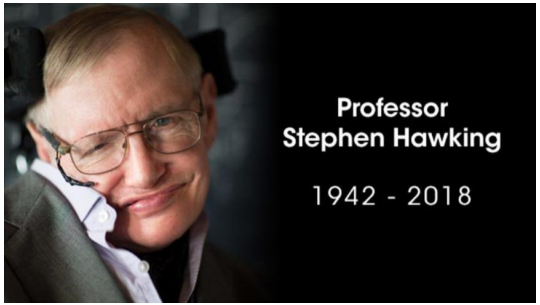
# Significance of SVs

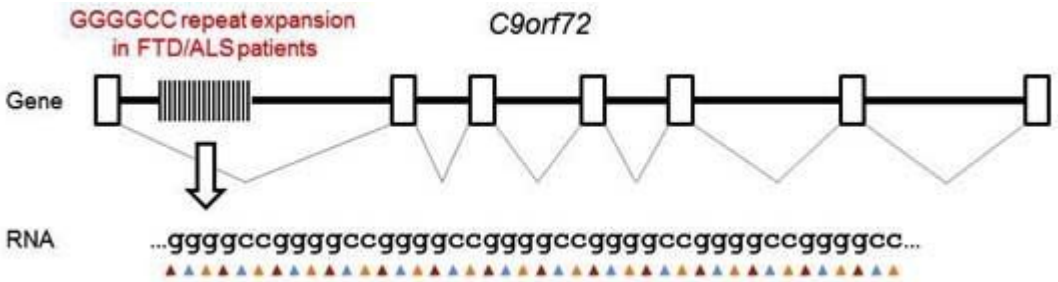
- Amyotrophic Lateral Sclerosis (ALS) / Lou Gehrig’s disease
- Affects nerve cells in the brain and the spinal cord



# Significance of SVs

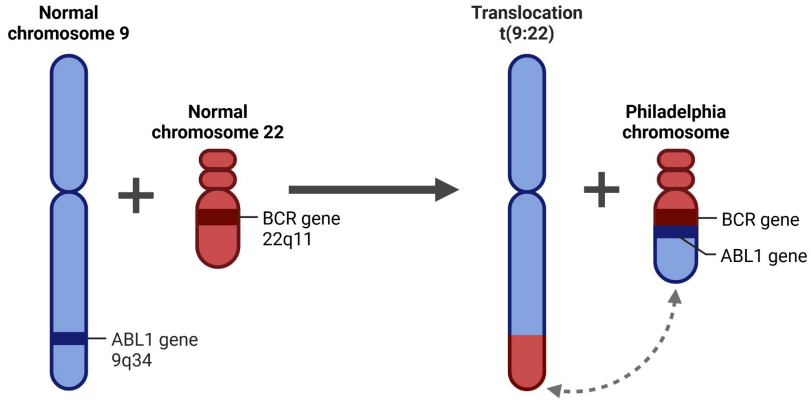
- Amyotrophic Lateral Sclerosis (ALS) / Lou Gehrig's disease
- Affects nerve cells in the brain and the spinal cord
  - Repeat expansion in C9orf72 (Ahmad et al. 2022)
  - Insertion in ERBB4
  - Inversion in VCP gene



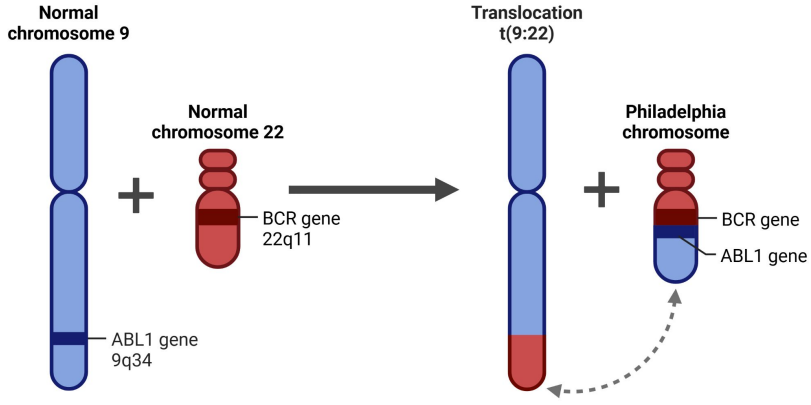
# Significance of SVs

- Chronic Myeloid Leukemia
  - Philadelphia chromosome ( BCR-ABL1 fusion )



# Significance of SVs

- Chronic Myeloid Leukemia
  - Philadelphia chromosome ( BCR-ABL1 fusion )
- DiGeorge Syndrome
  - Deletion in chr22
  - Heart defects, immune system problems
- Schizophrenia
  - Deletion in chr1q and chr15q
- And many more (Stankiewicz, Paweł, Lupski 2010)

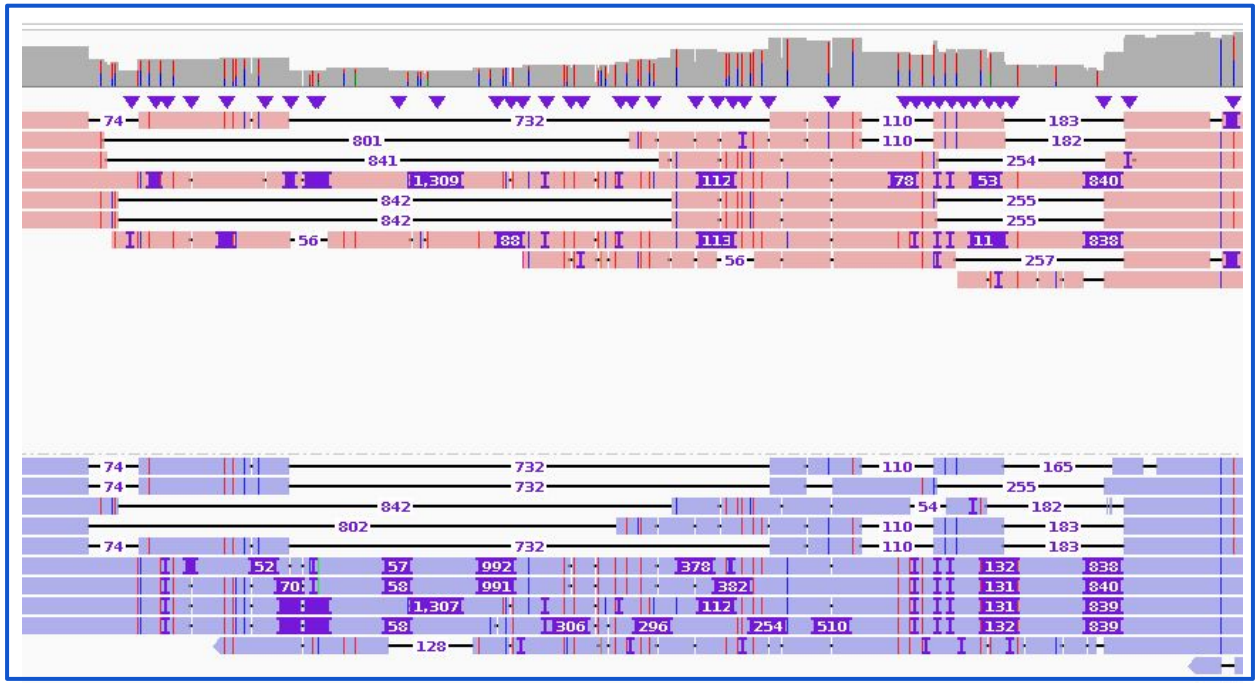


# Difficulties In Finding SVs

- Poorly mapped regions of the genome

# Difficulties In Finding SVs

- Poorly mapped regions of the genome



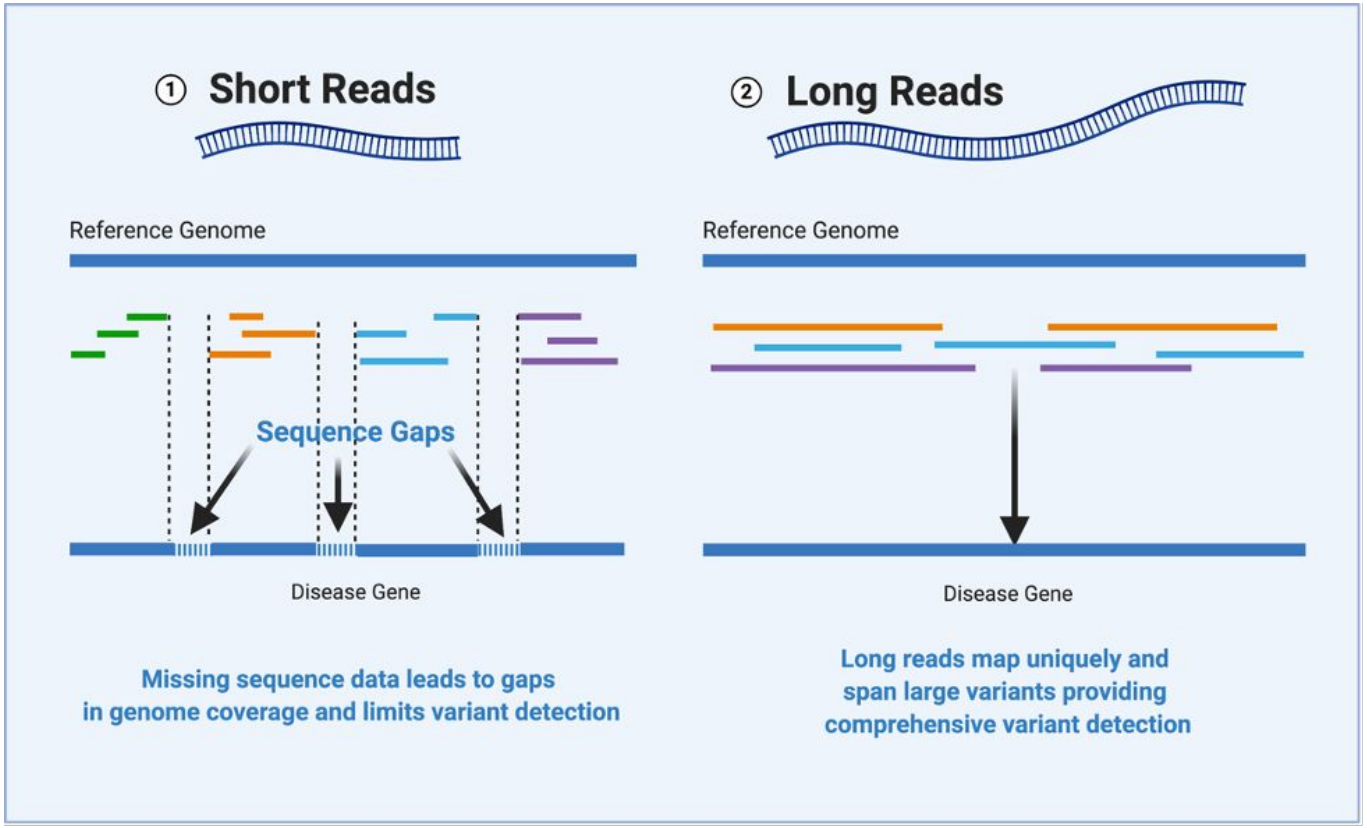
# Difficulties In Finding SVs

- Poorly mapped regions of the genome
  - Challenging to map uniquely with short-read sequences.
  - Long-read sequencing showing hope
  - Pacific Biosciences, Oxford Nanopore

# Difficulties In Finding SVs

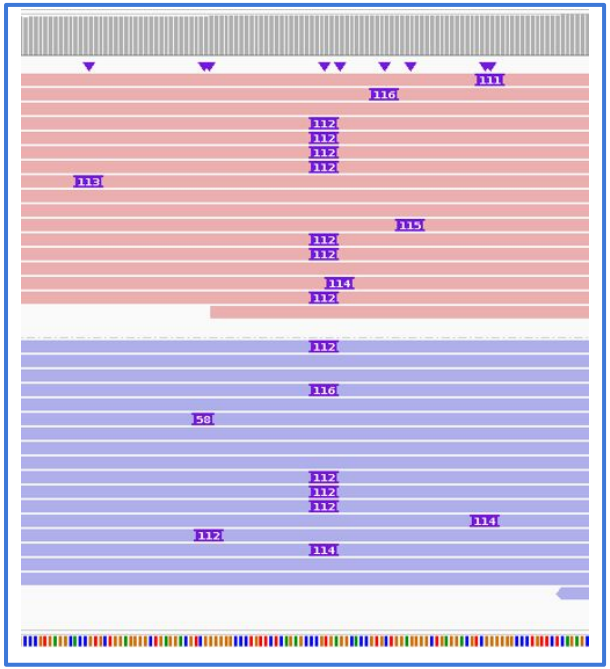
- Poorly mapped regions of the genome
  - Challenging to map uniquely with short-read sequences.
  - Long-read sequencing showing hope
  - Pacific Biosciences, Oxford Nanopore
- SVs spanning multiple reads

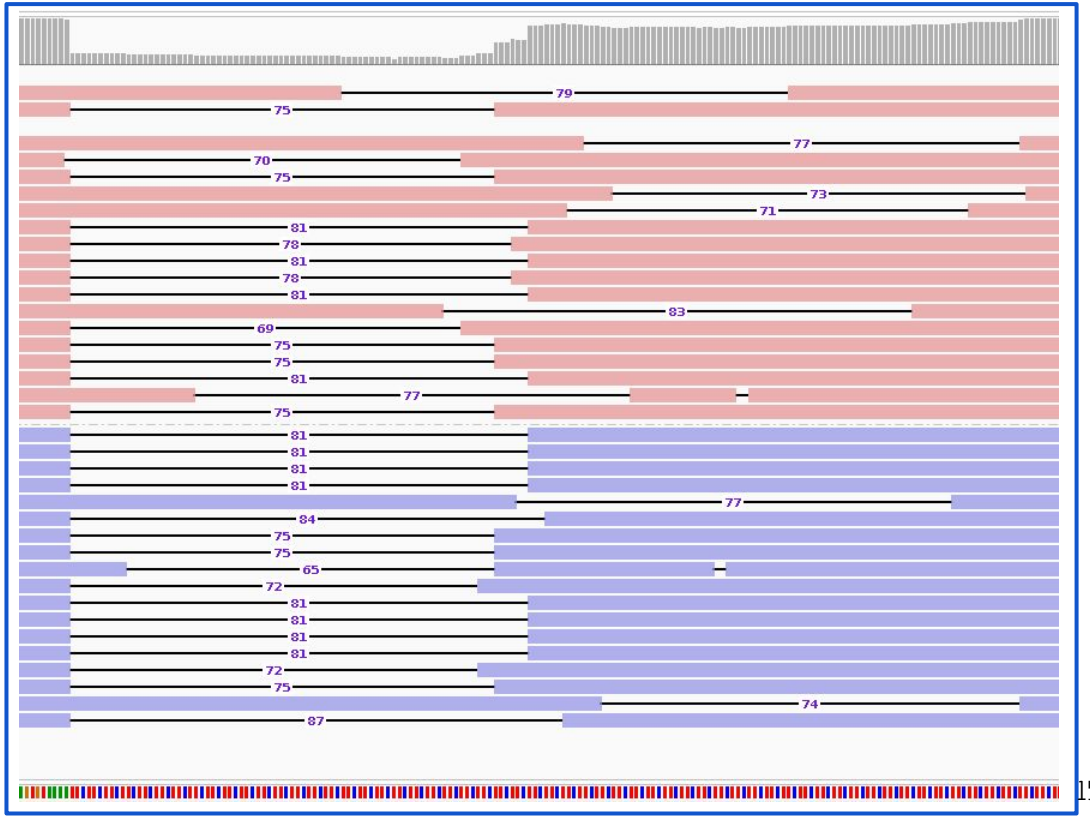
# Short Read vs Long Read Sequencing



# Difficulties In Finding SVs

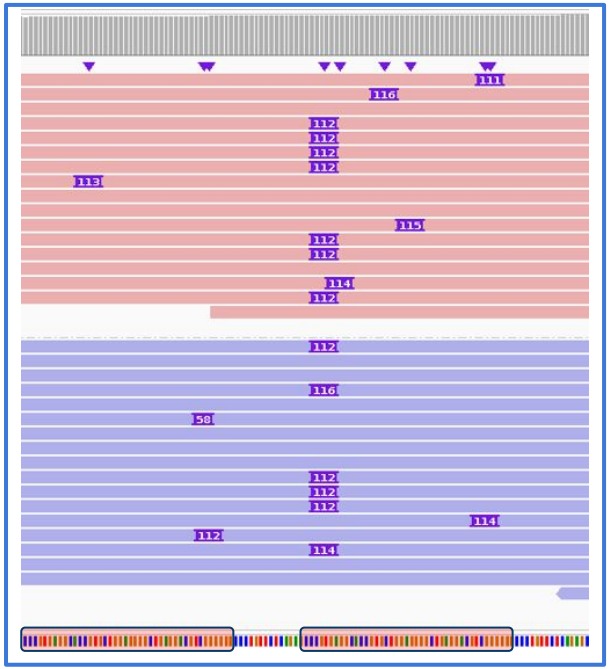
- Repetitive regions

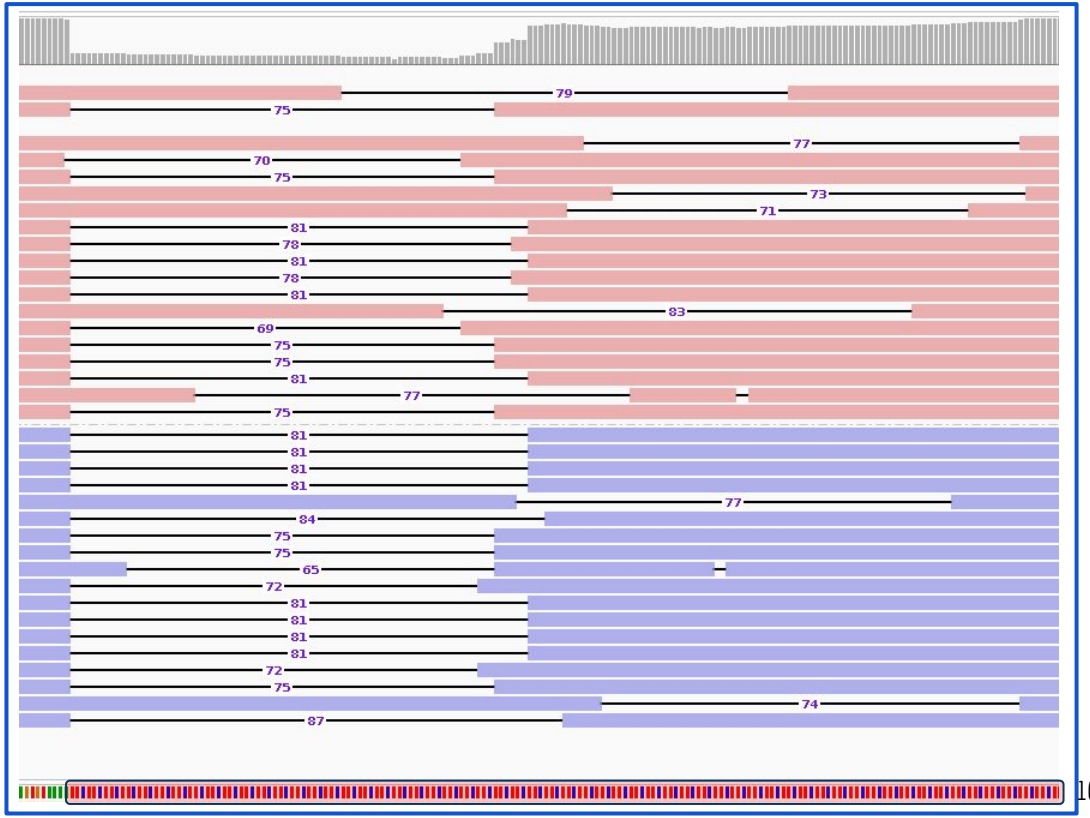


# Difficulties In Finding SVs

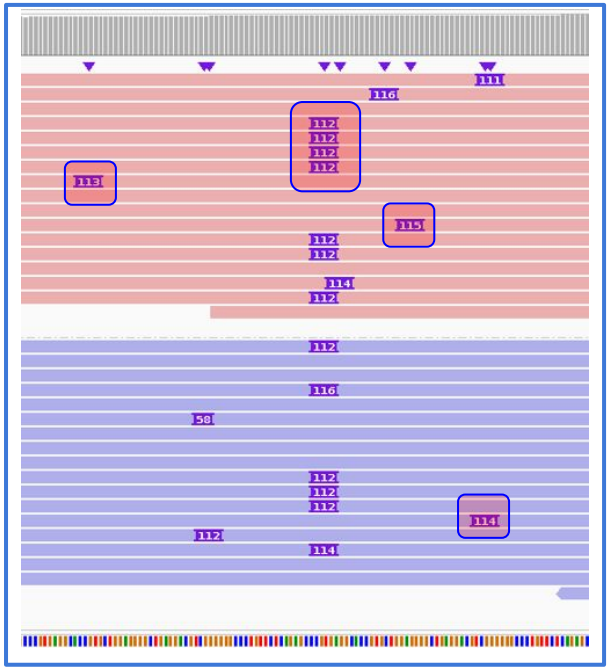
- Repetitive regions

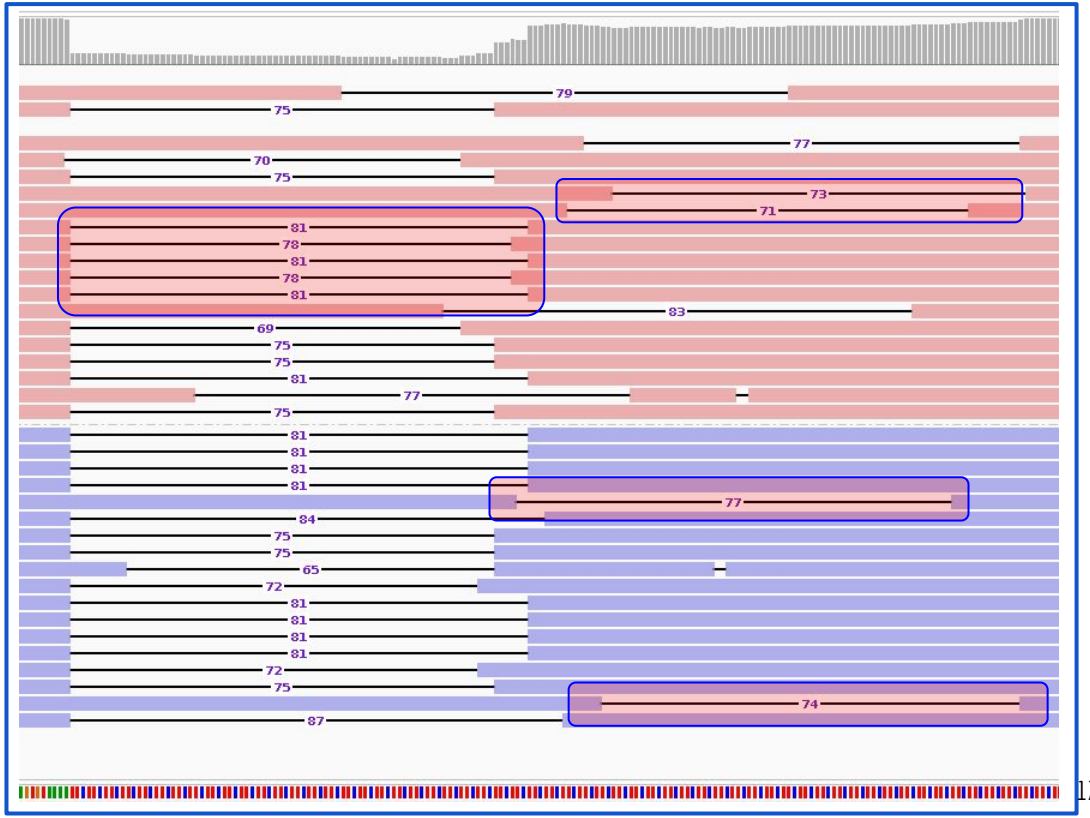


# Difficulties In Finding SVs

- Repetitive regions



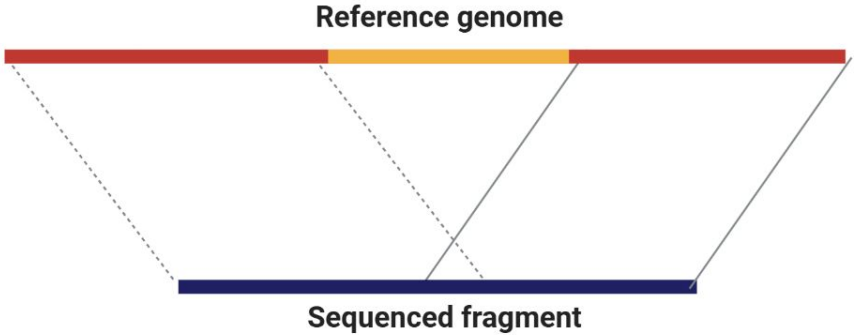



# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions

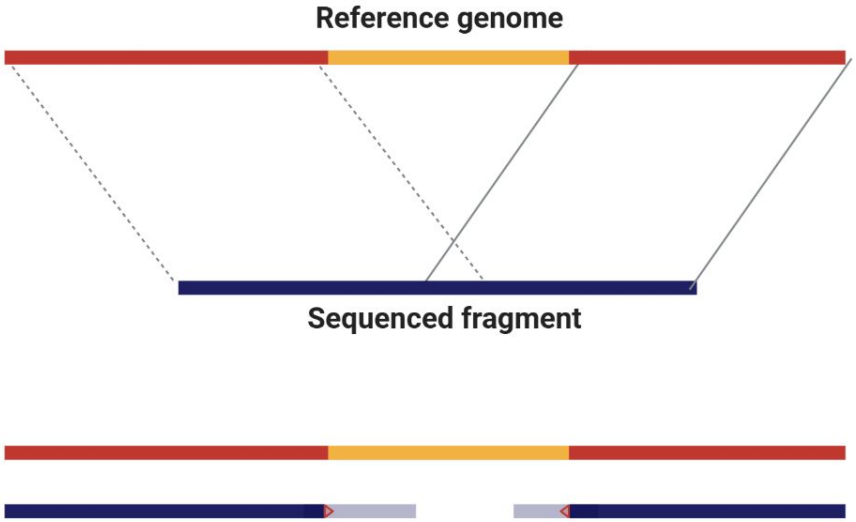
# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions



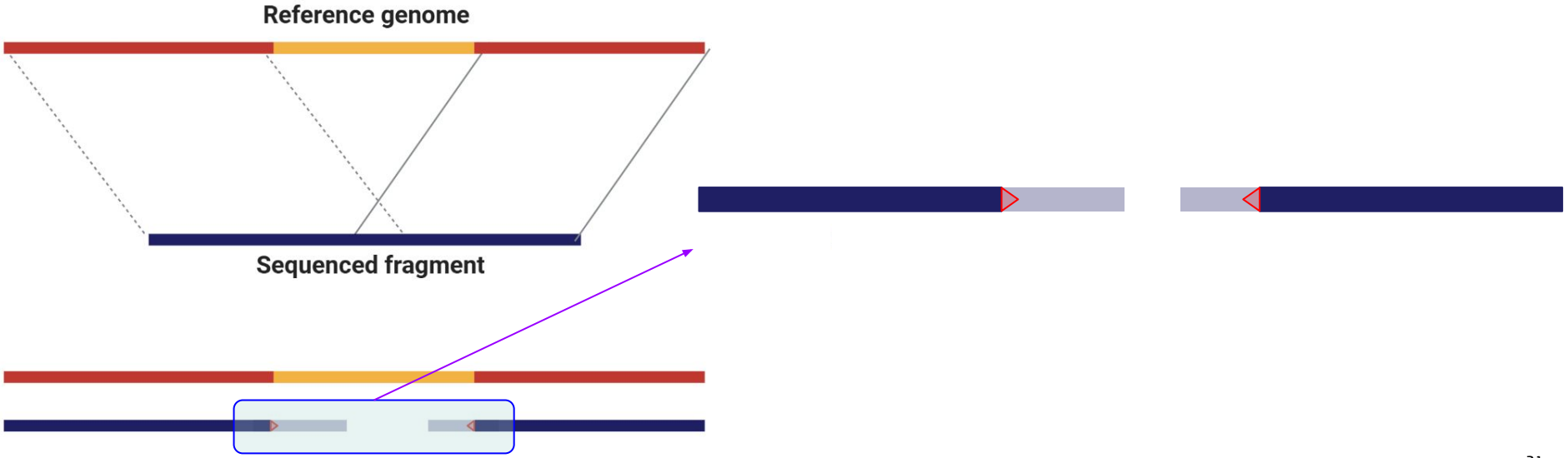
# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions



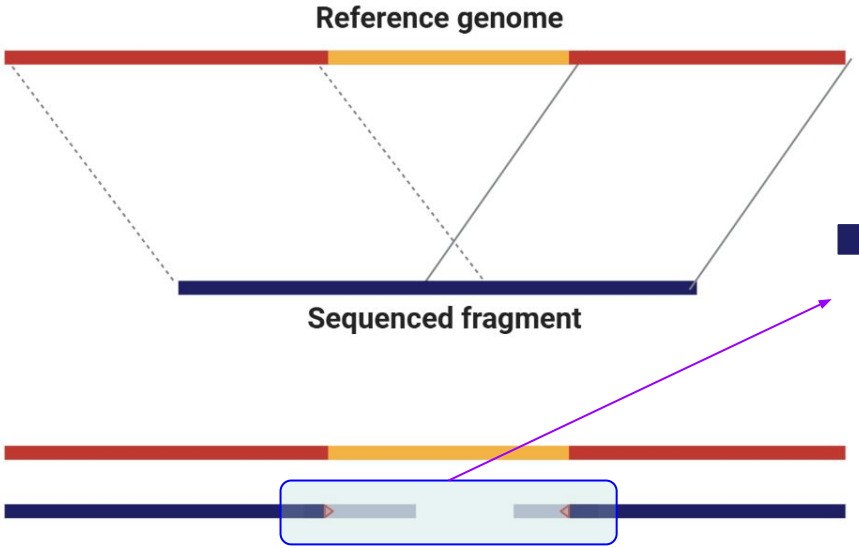
# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions



# Difficulties In Finding SVs

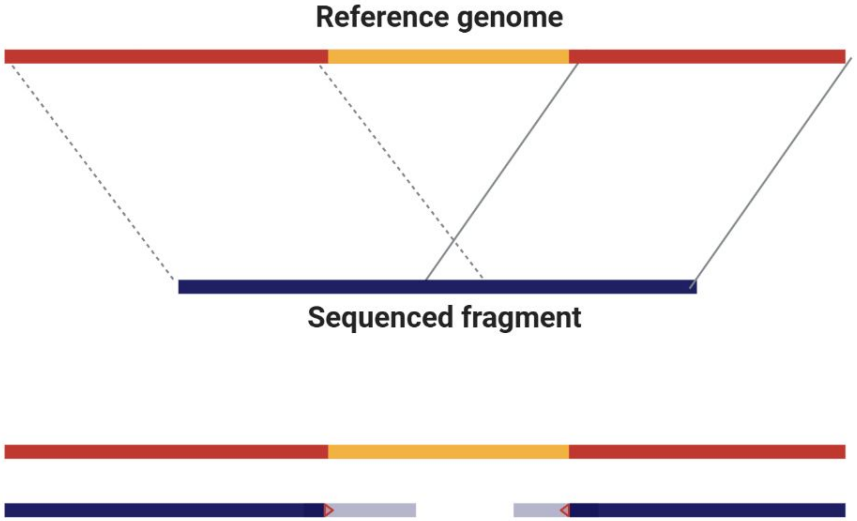
- Chimeric reads
  - Single sequencing read aligns to multiple positions




- Soft clipped read
  - “Soft” → bases are present, but not aligned to reference

# Difficulties In Finding SVs

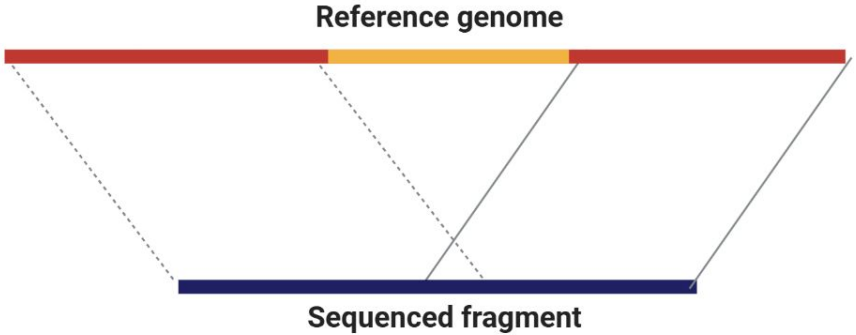
- Chimeric reads
  - Single sequencing read aligns to multiple positions

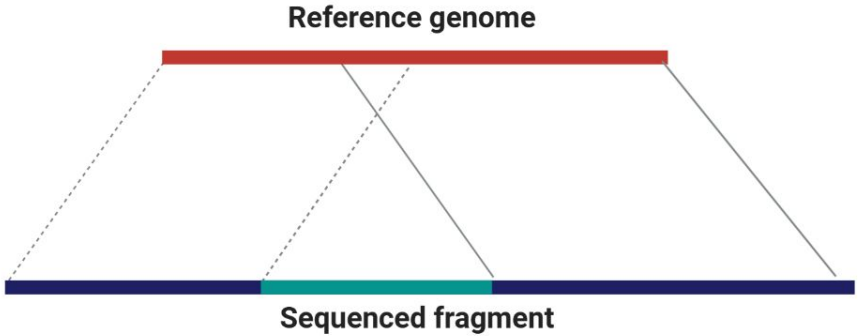



**Deletion** of reference

# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions

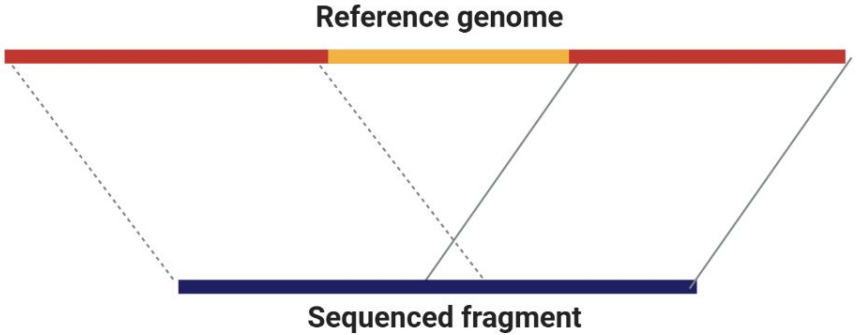


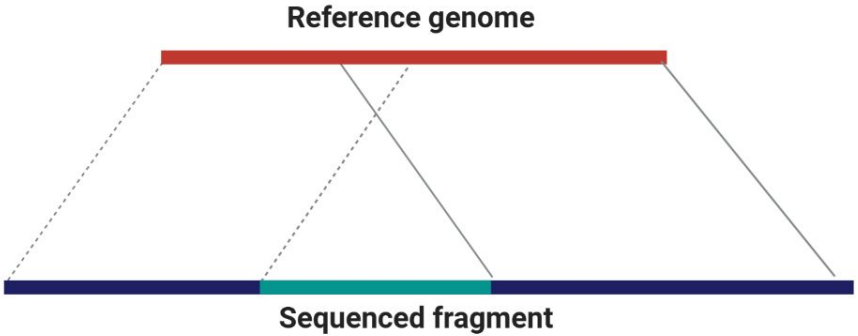




**Deletion** of reference

# Difficulties In Finding SVs

- Chimeric reads
  - Single sequencing read aligns to multiple positions

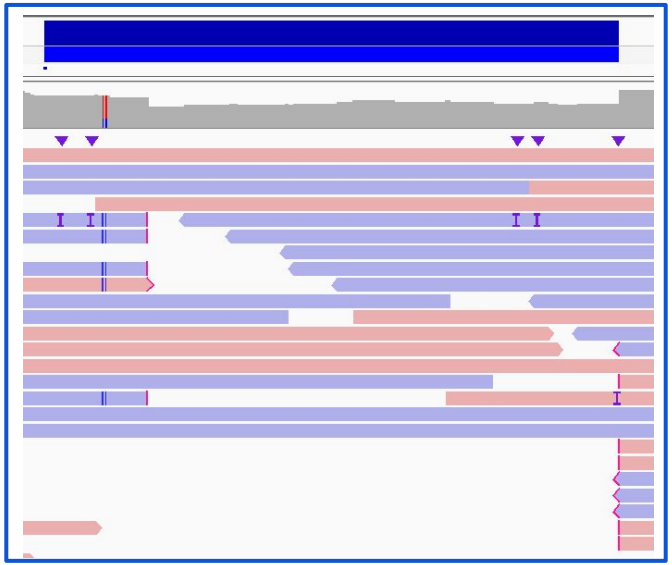


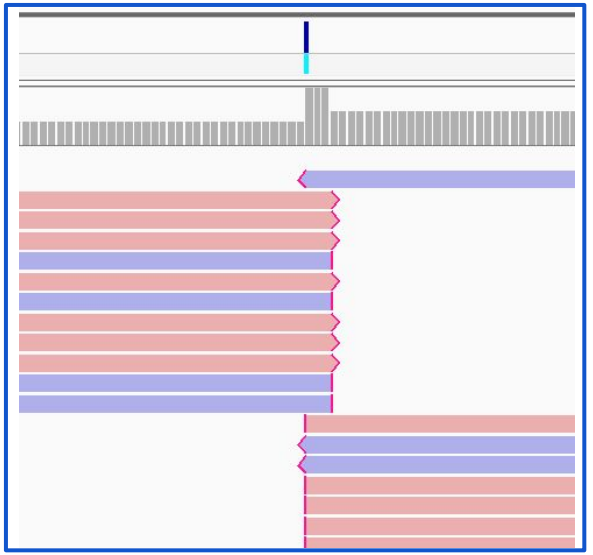




**Deletion** of reference

**Insertion** of new sequence

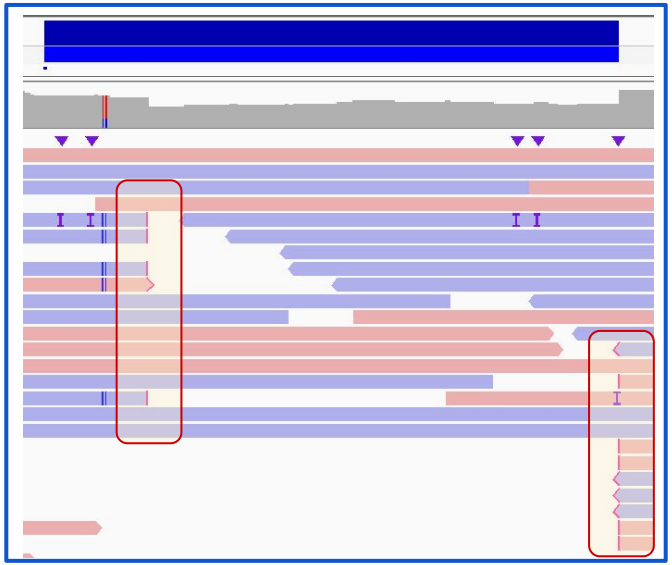
# Difficulties In Finding SVs



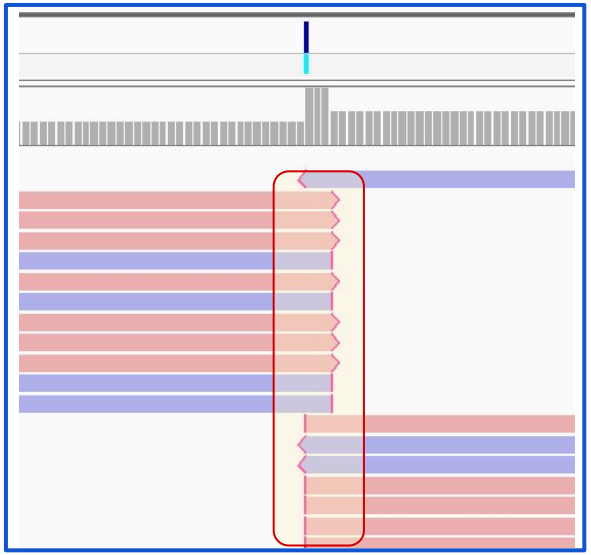


# Difficulties In Finding SVs

Deletion



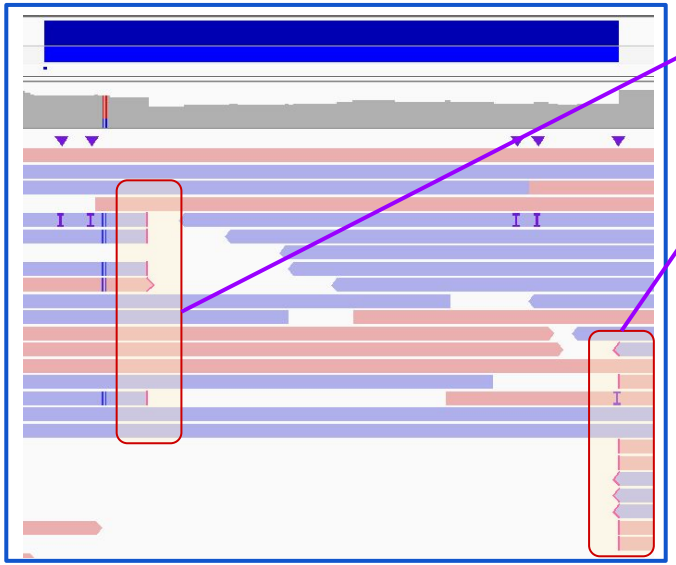
Insertion



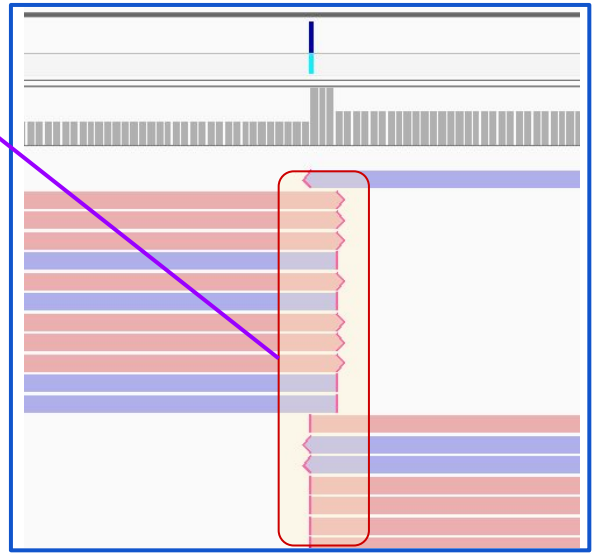
# Difficulties In Finding SVs

Soft clipped reads

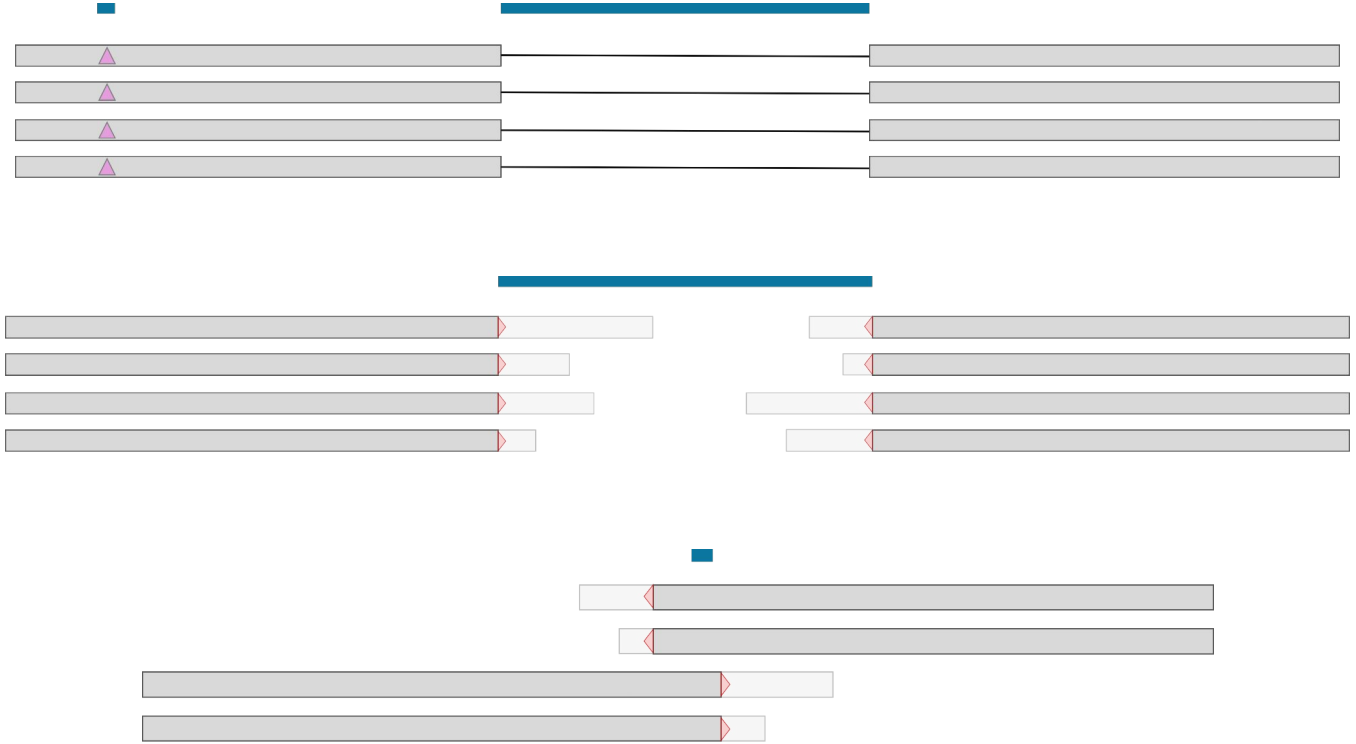
Deletion



Insertion

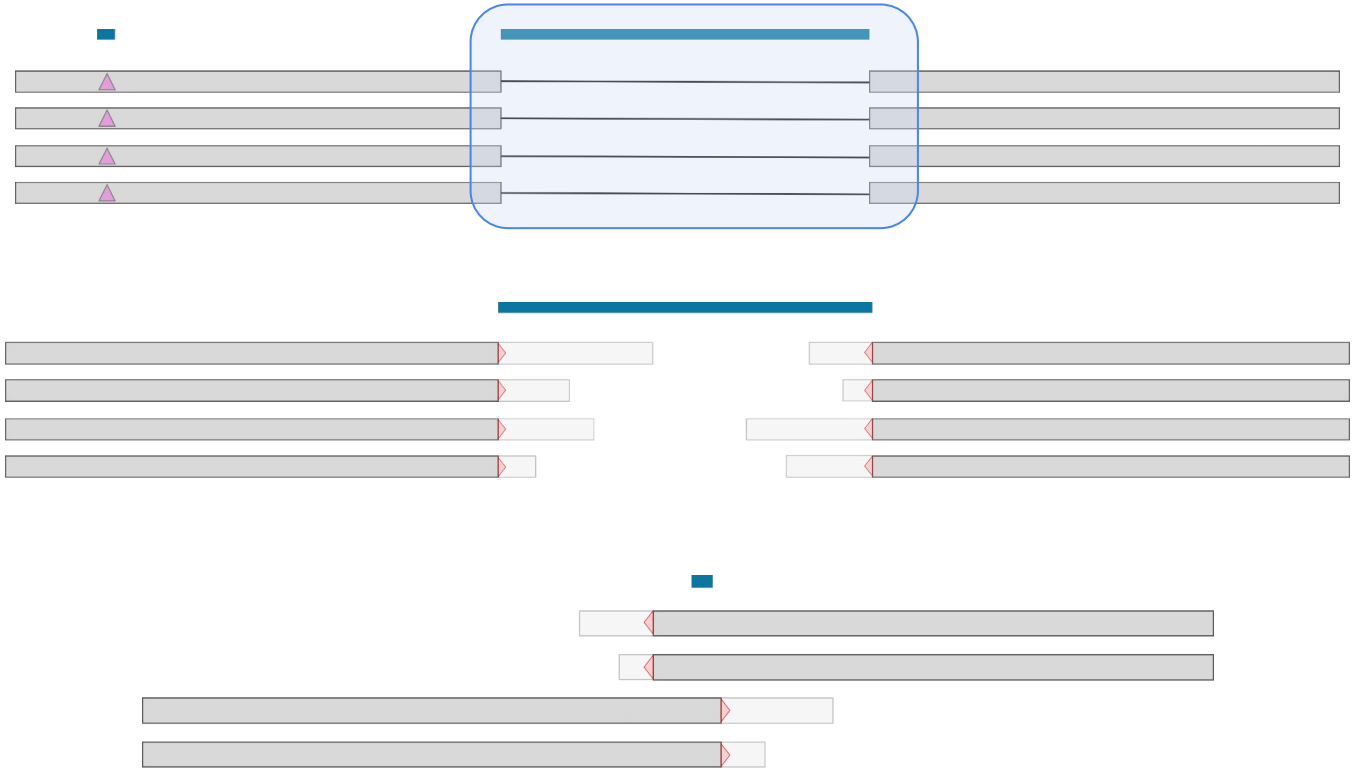


# SVs Under Consideration



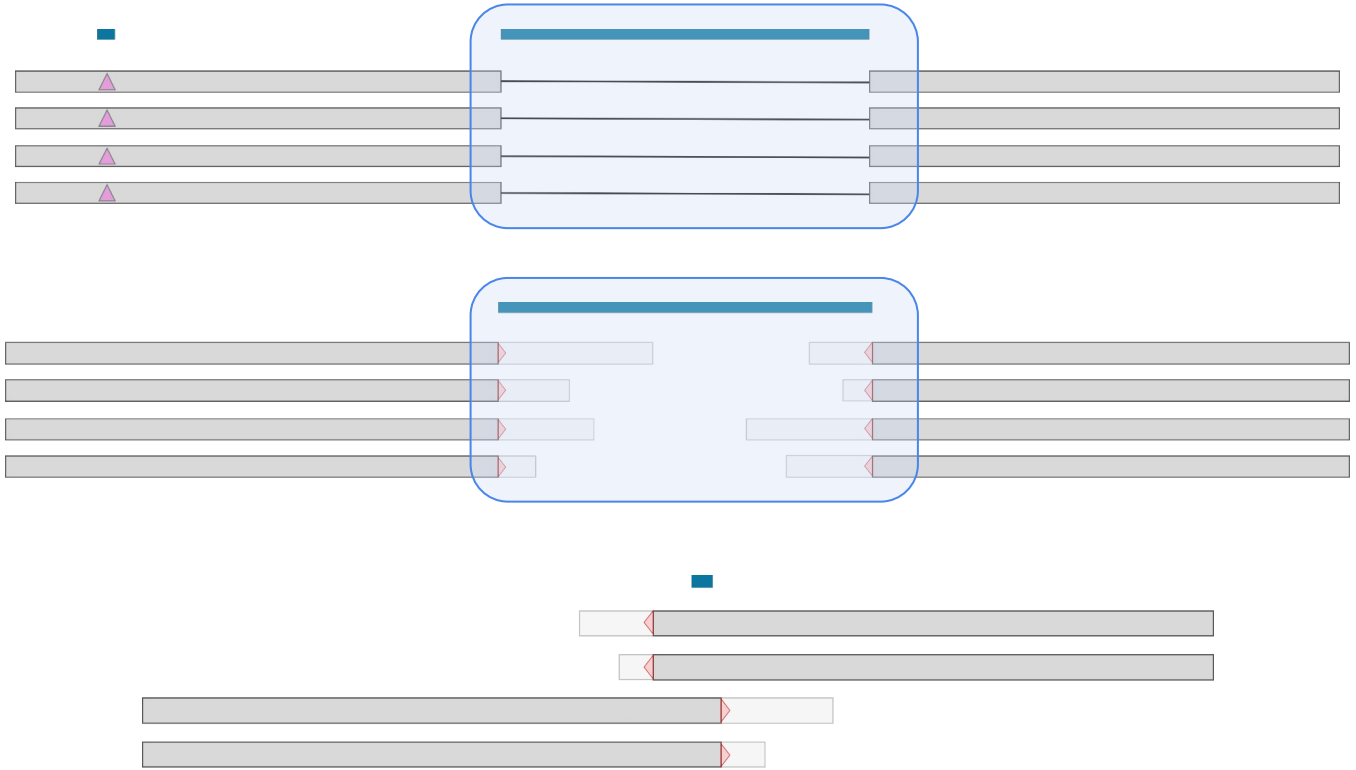
# SVs Under Consideration

Deletion



# SVs Under Consideration

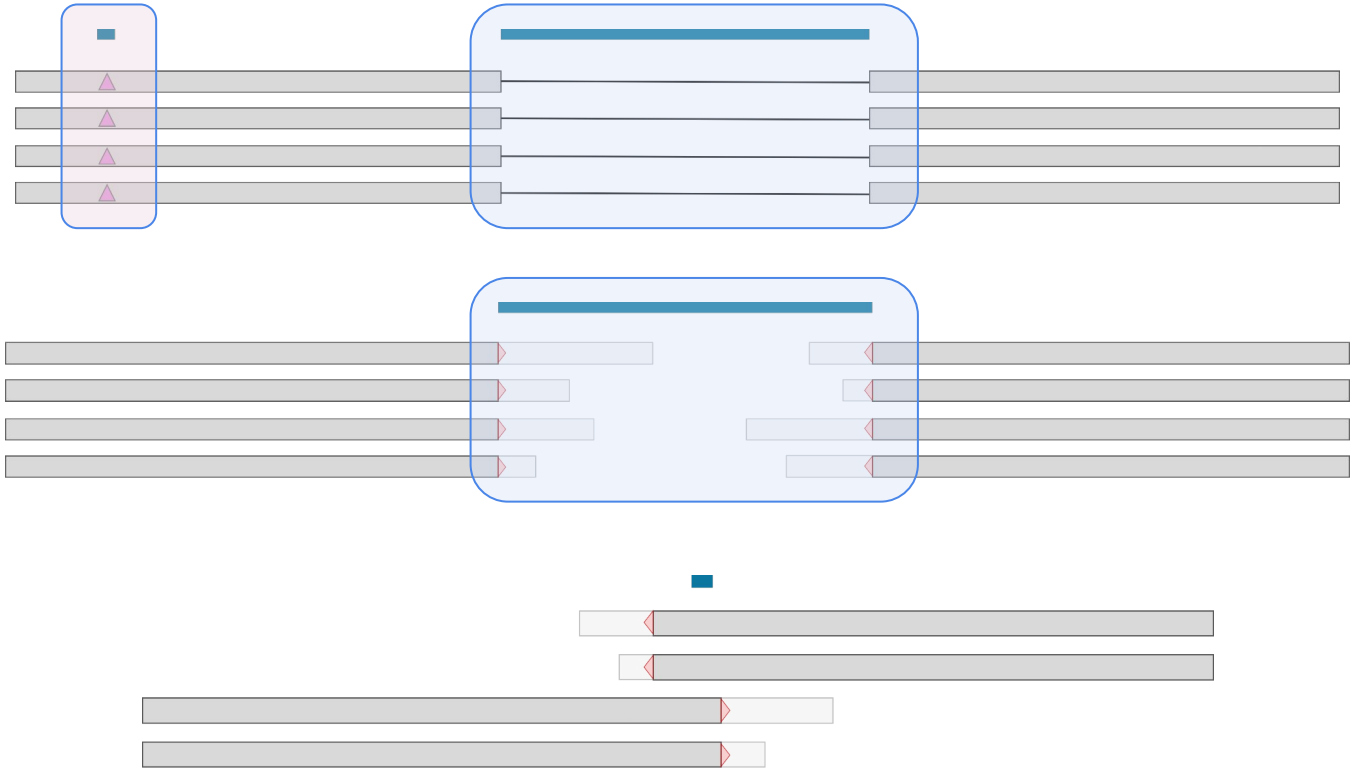
Deletion



# SVs Under Consideration

Deletion

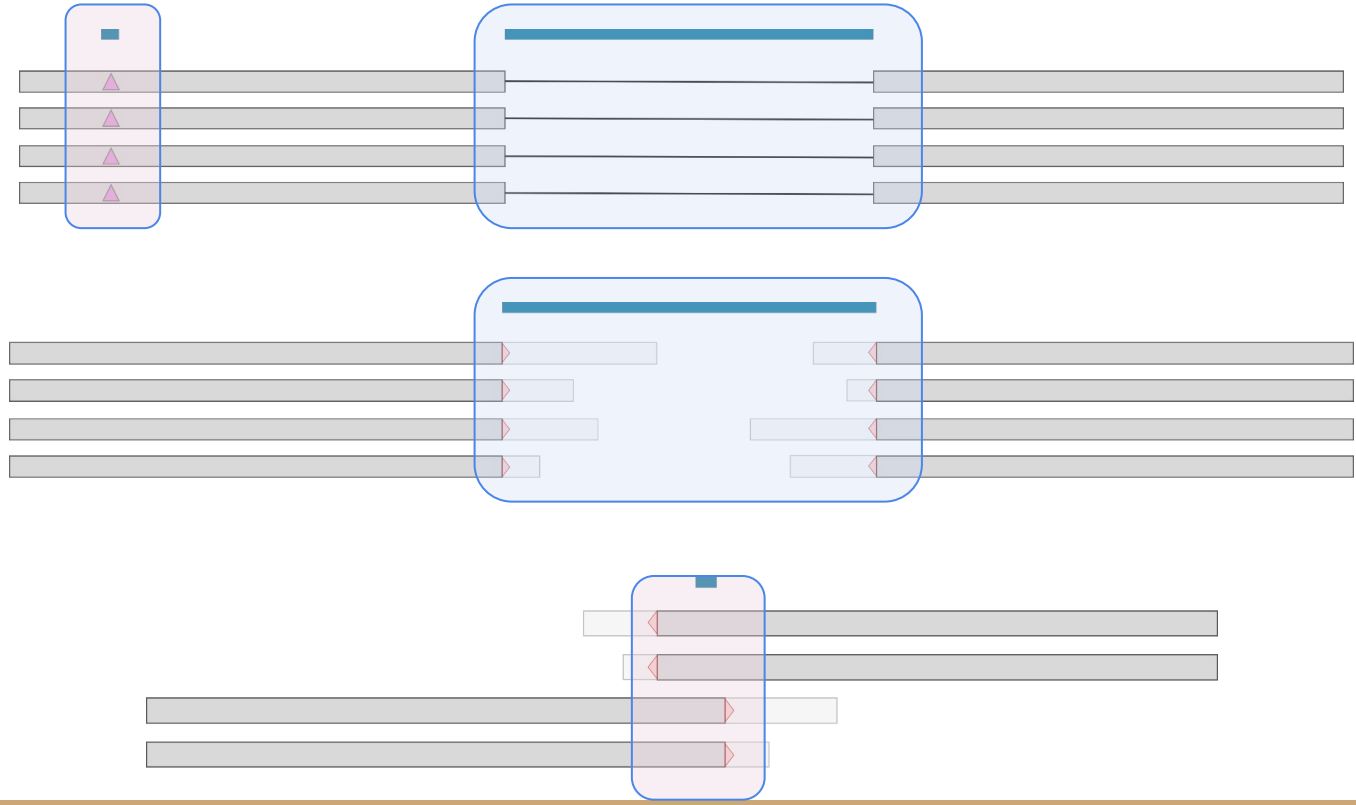
Insertion



# SVs Under Consideration

Deletion

Insertion

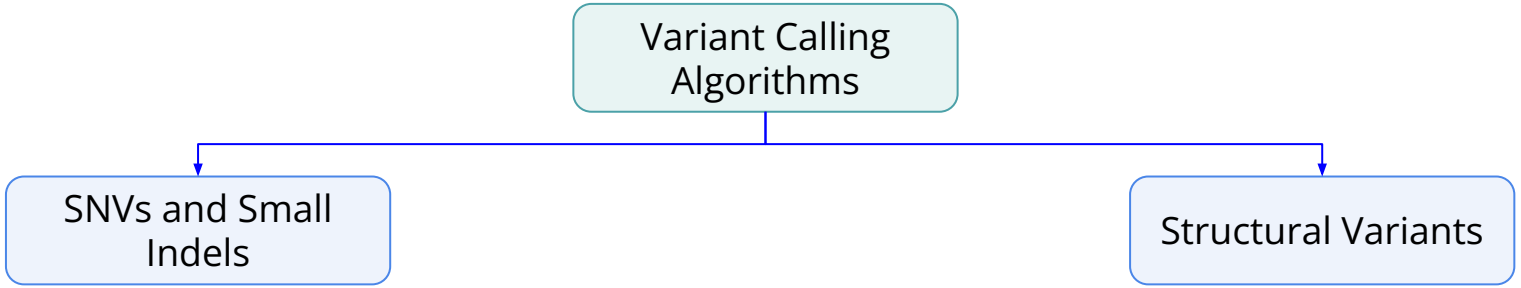


# Related Works

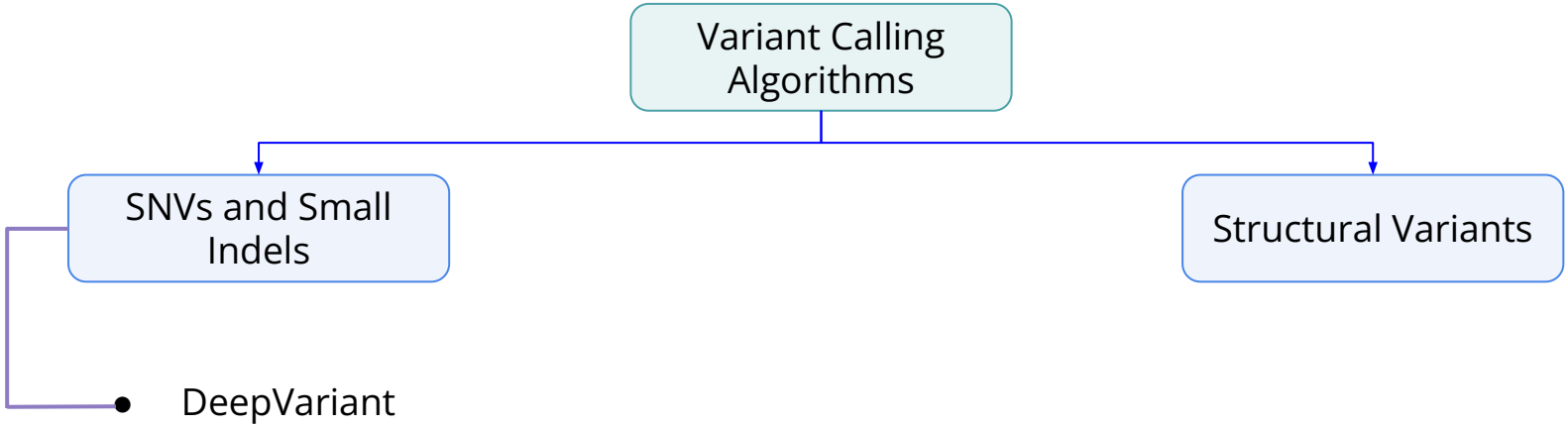
# Related Works

Variant Calling Algorithms

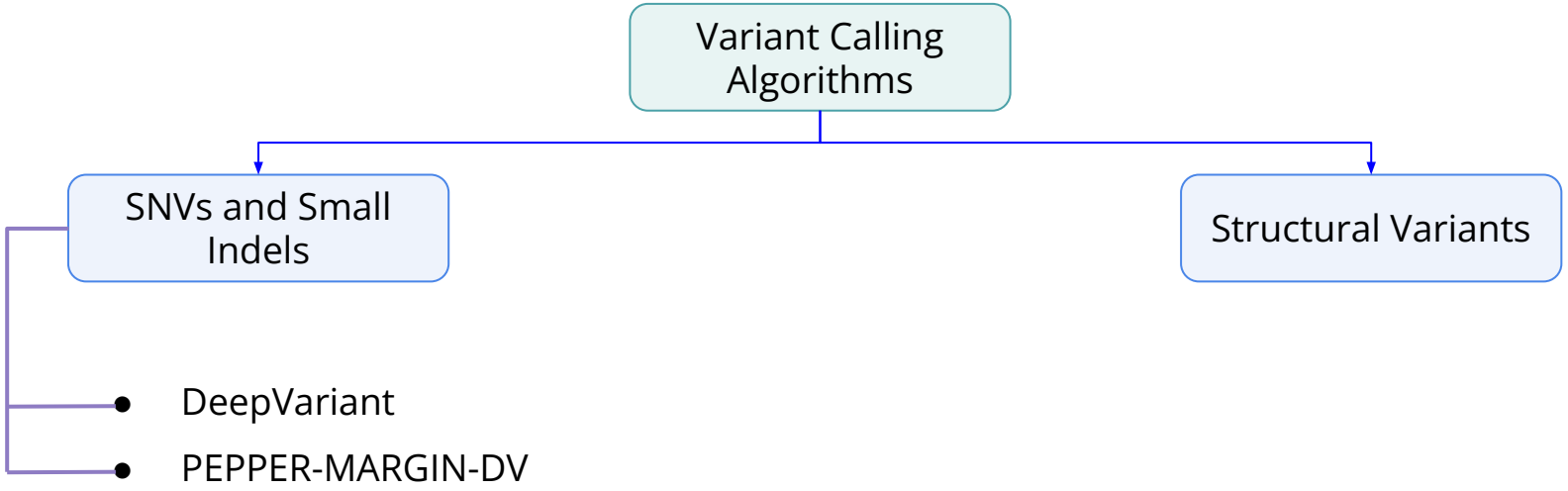
# Related Works



# Related Works



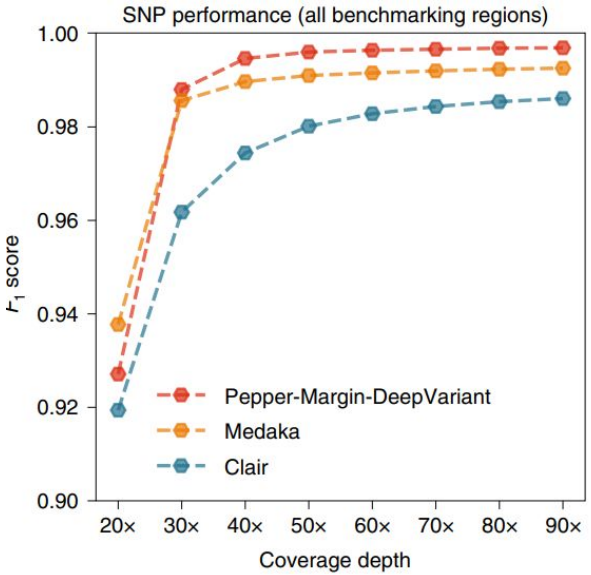
# Related Works



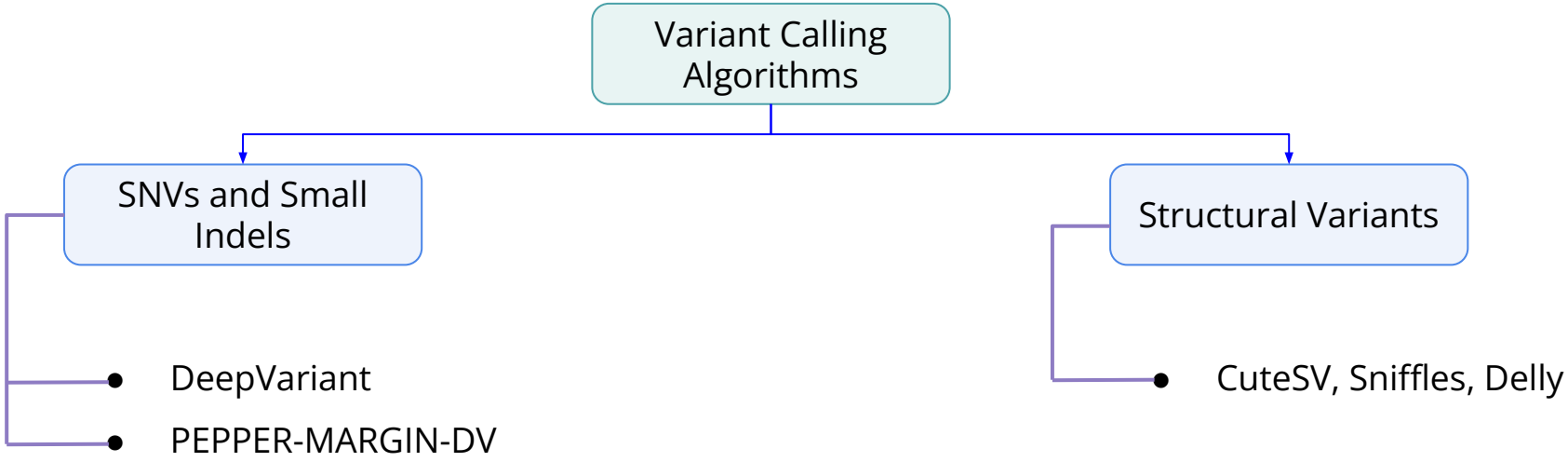
# Related Works

Variant Calling Algorithms

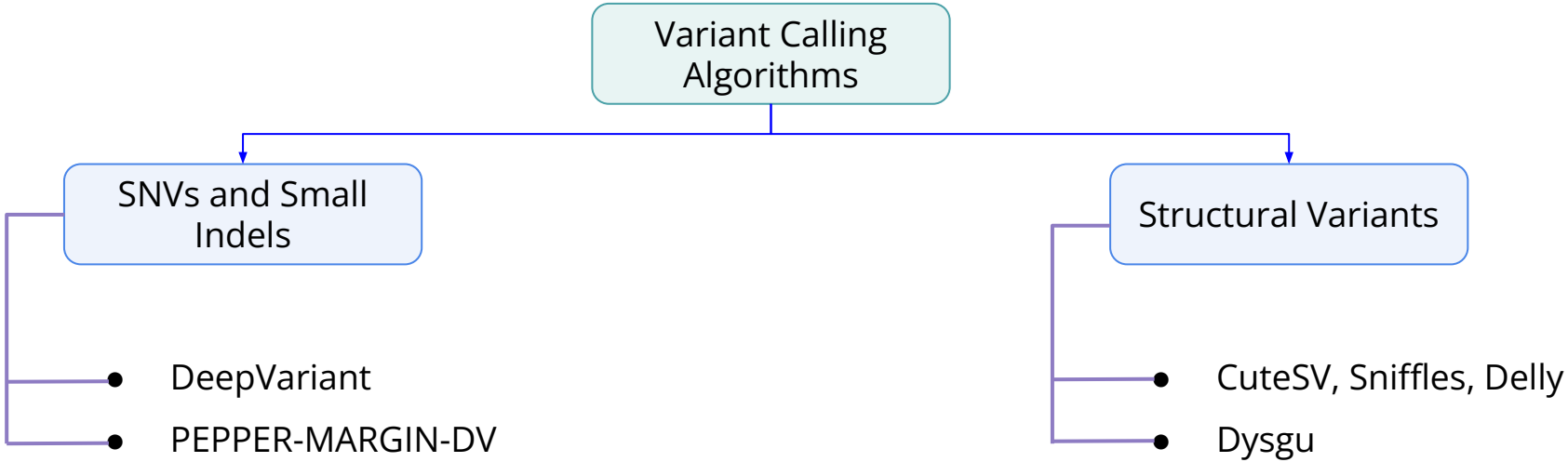
- SNVs and Small Indels
  - DeepVariant
  - PEPPER-MARGIN-DV
- Structural Variants



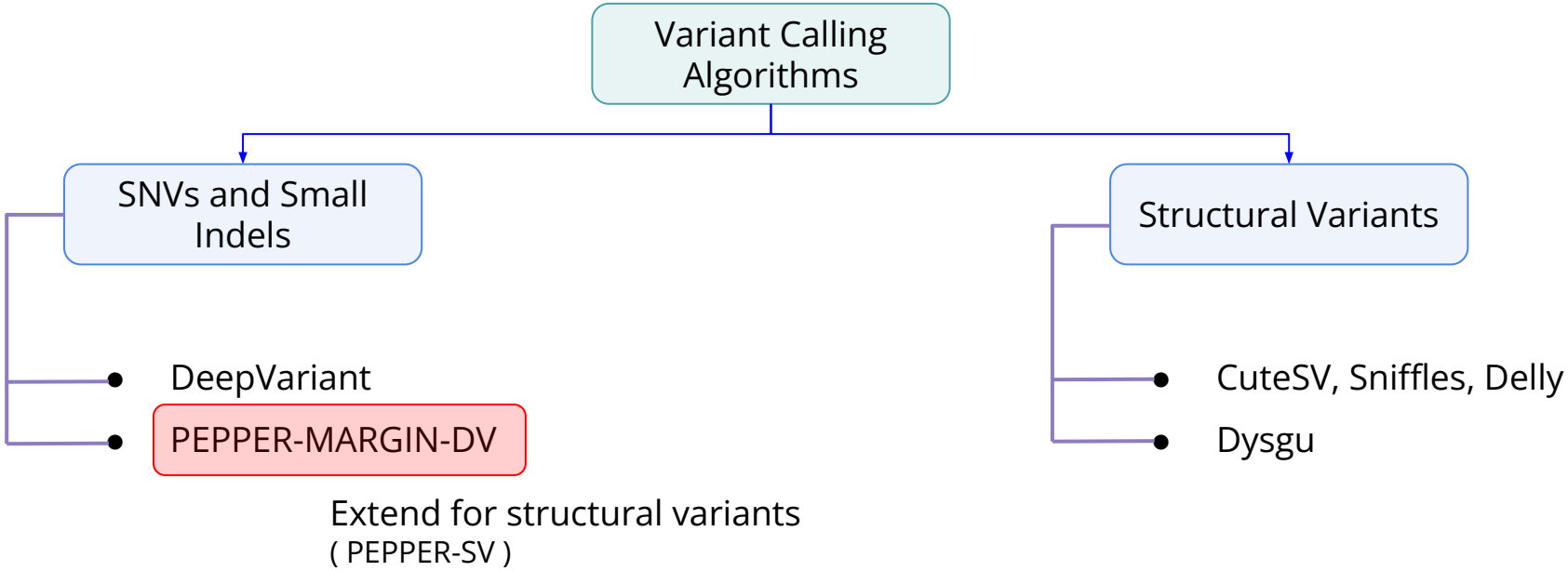
# Related Works



# Related Works

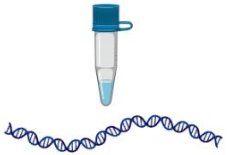


# Related Works



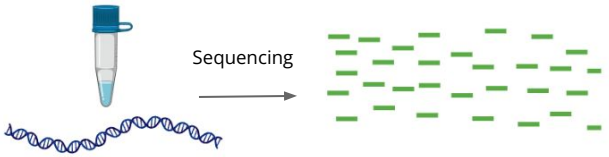
# Methodology

# Overview



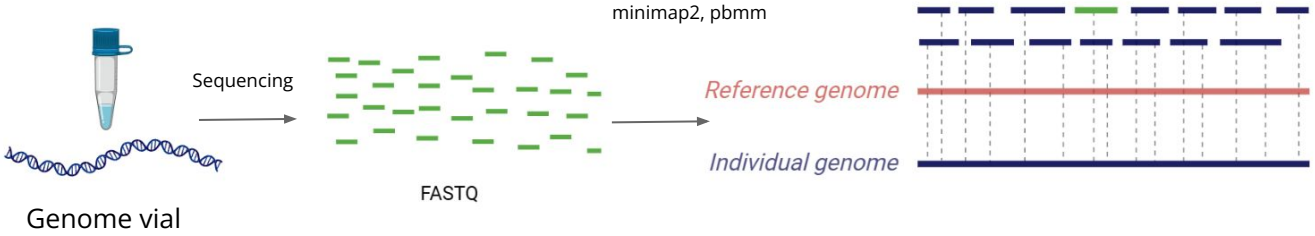
Genome vial

# Overview



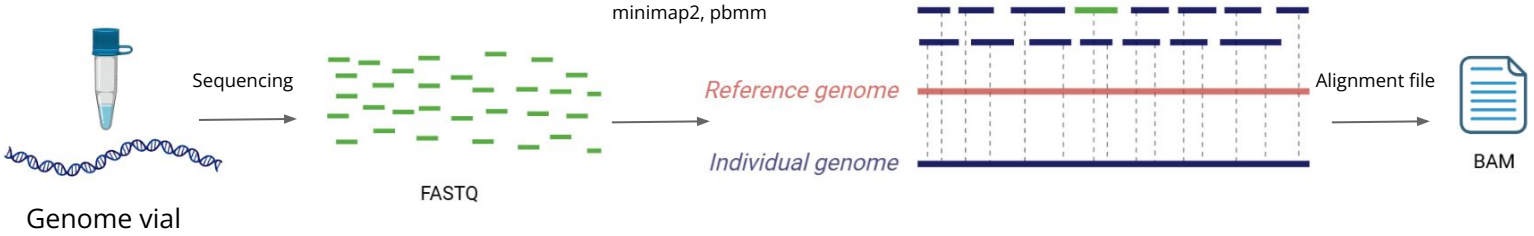

Genome vial

# Overview

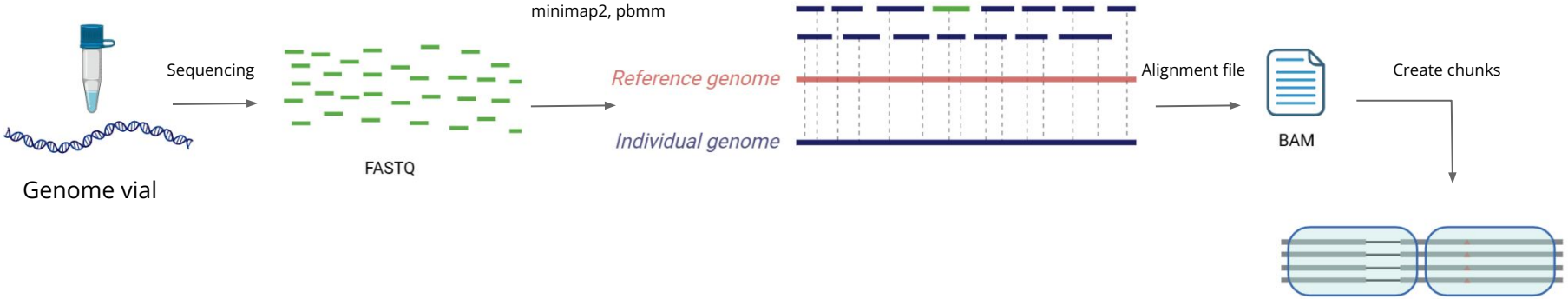



Genome vial

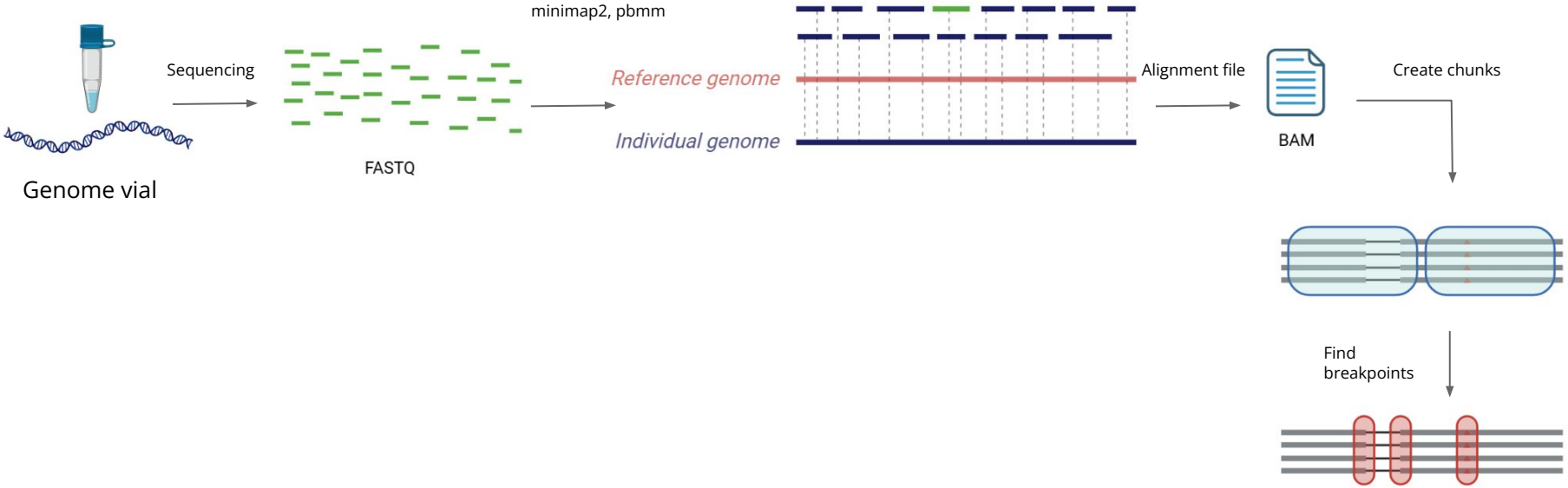
# Overview


Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Genome vial

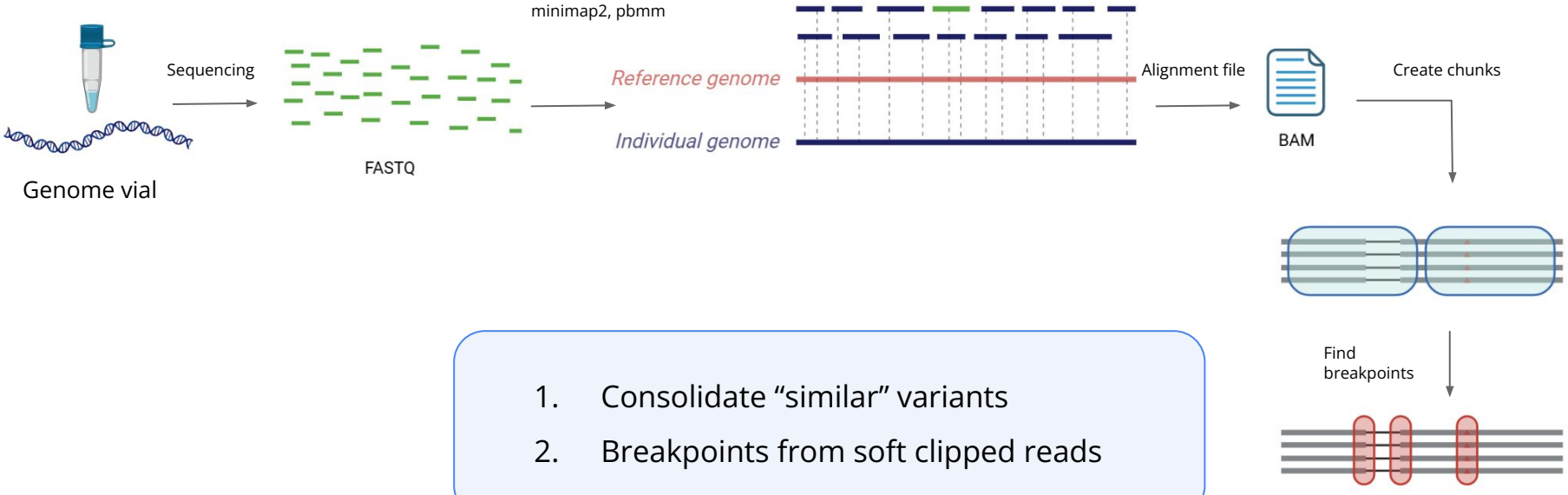
# Overview


Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Genome vial

# Overview

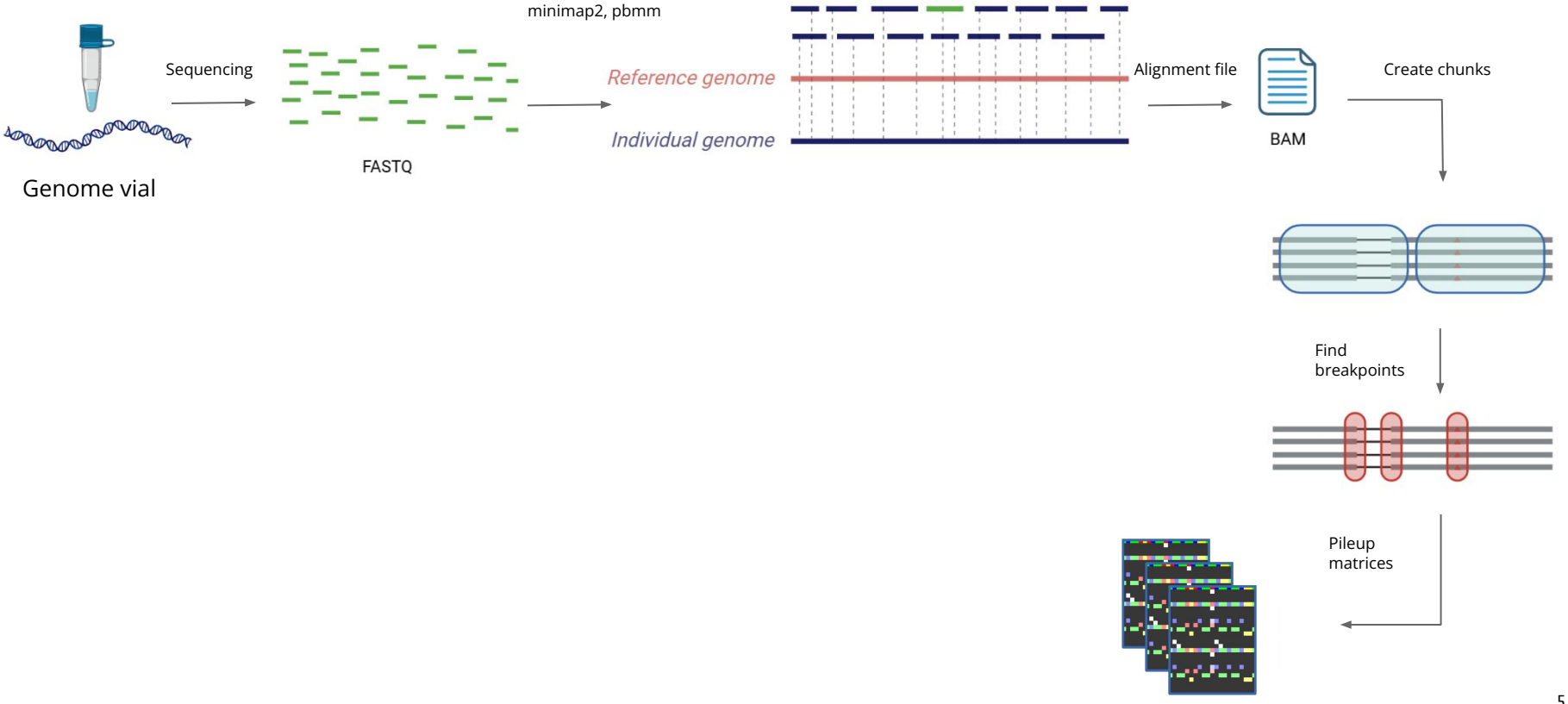

Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find
breakpoints
Genome vial

# Overview

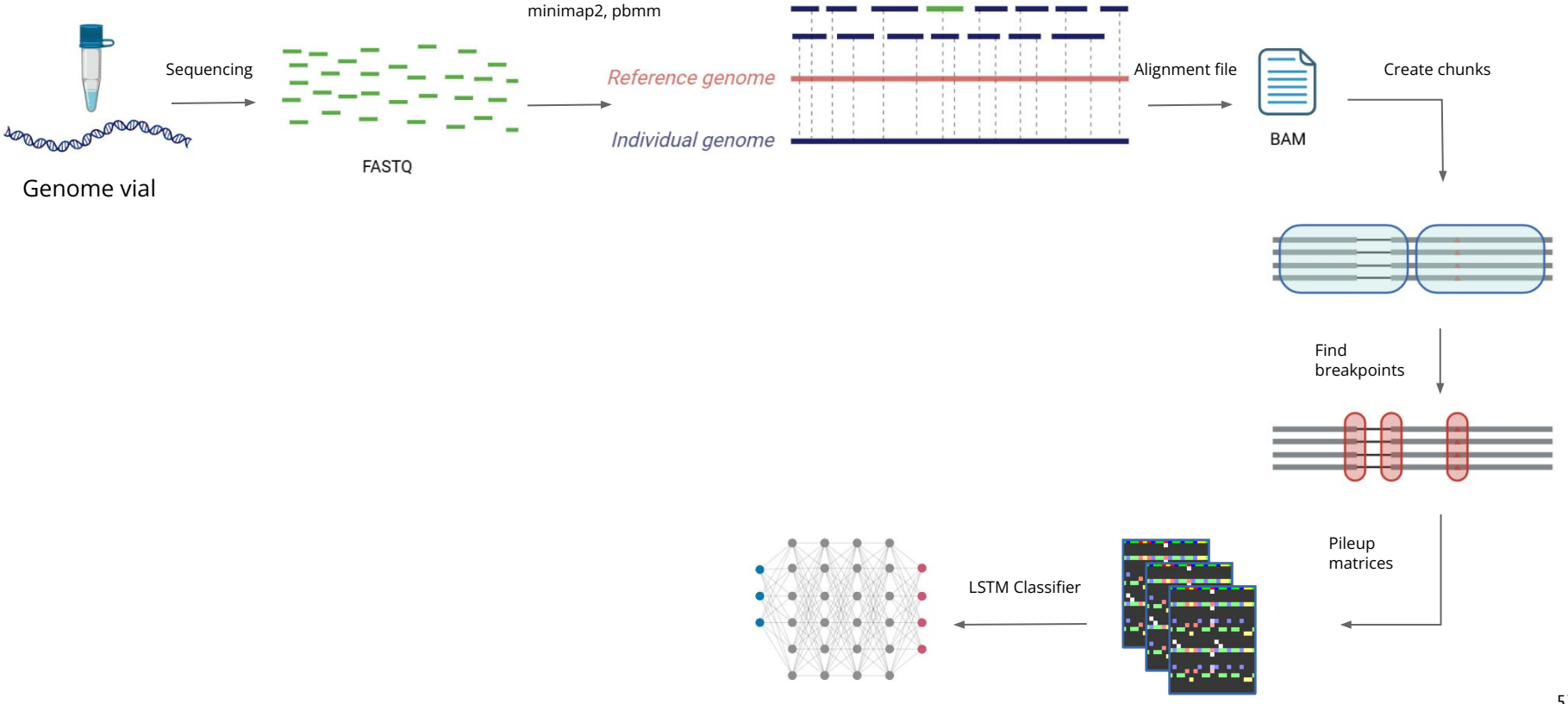



1. Consolidate “similar” variants
2. Breakpoints from soft clipped reads

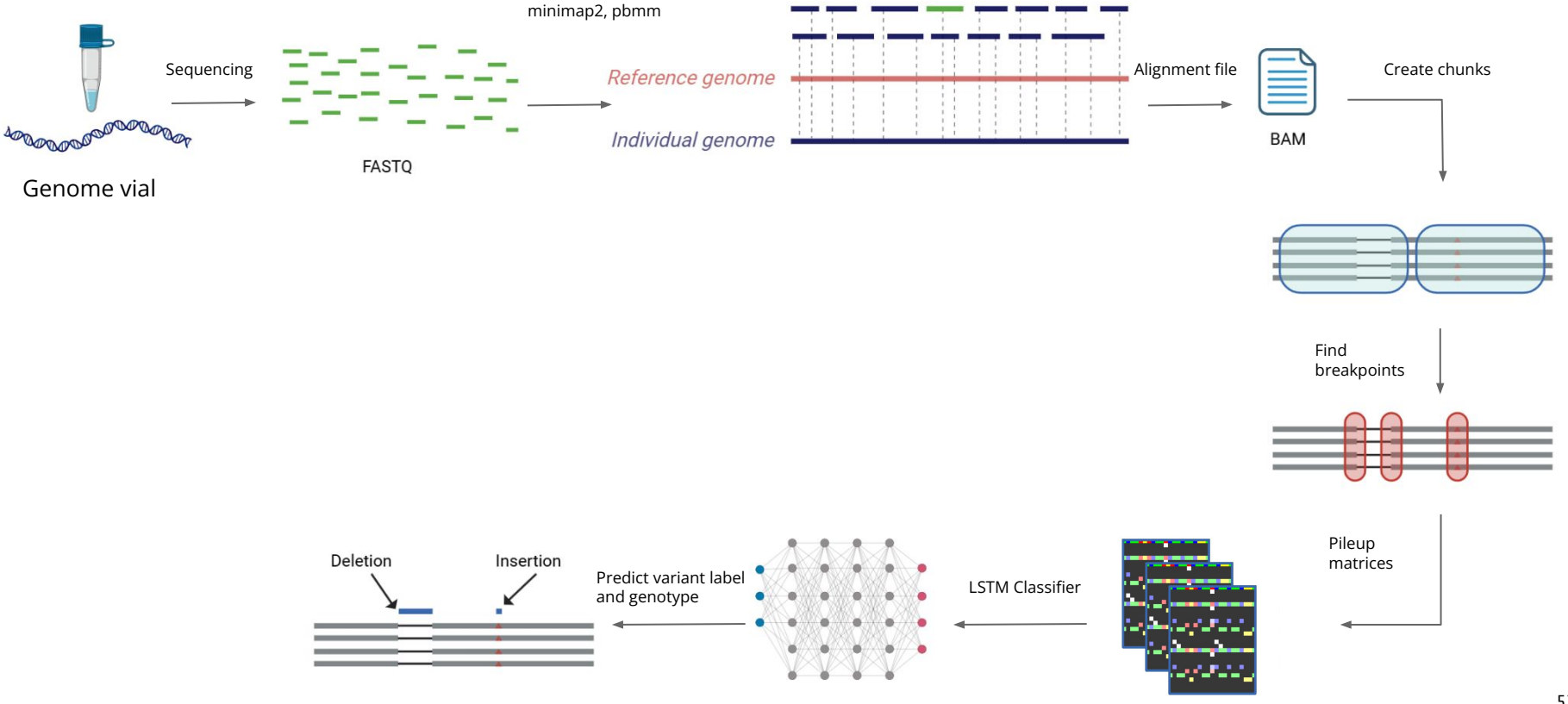
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find
breakpoints
Pileup
matrices

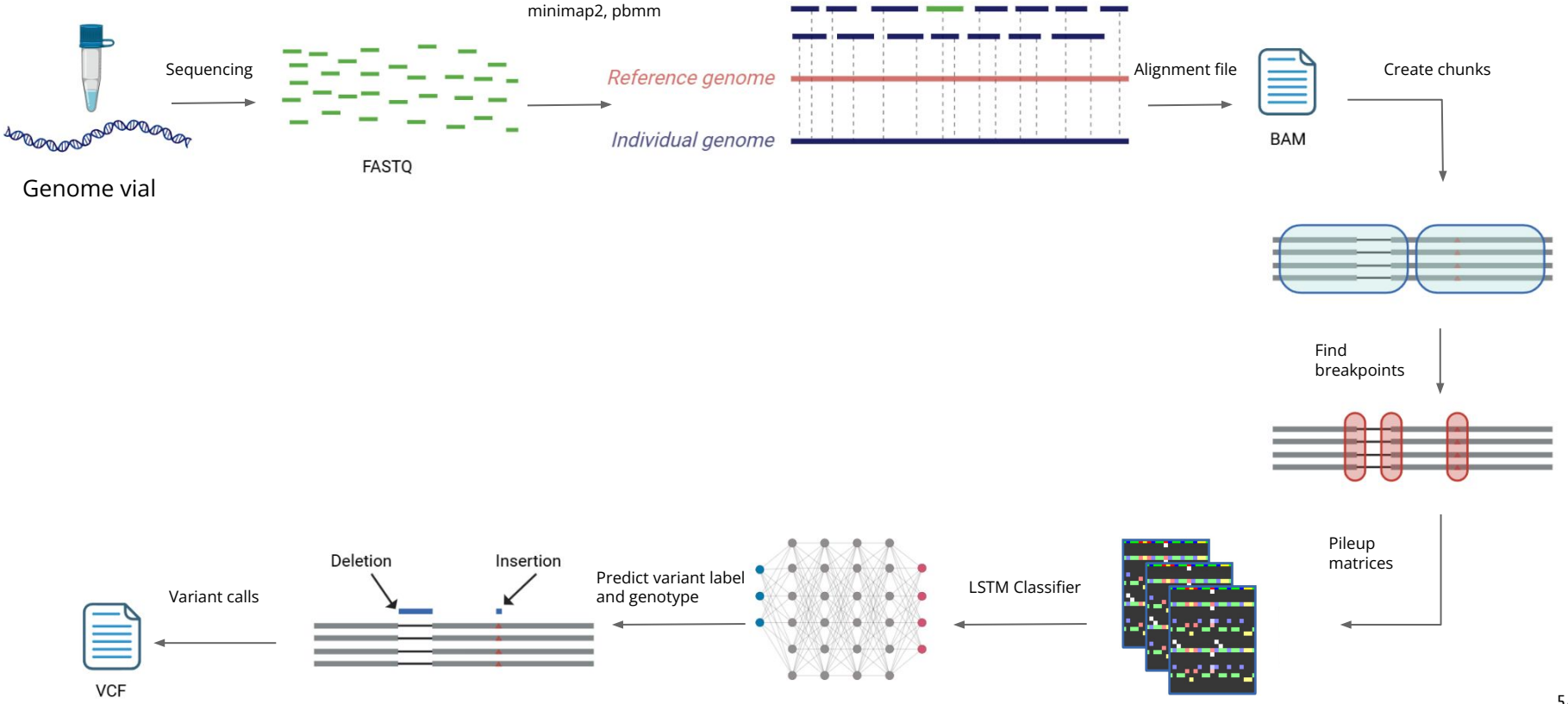
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find
breakpoints
Pileup
matrices
LSTM Classifier

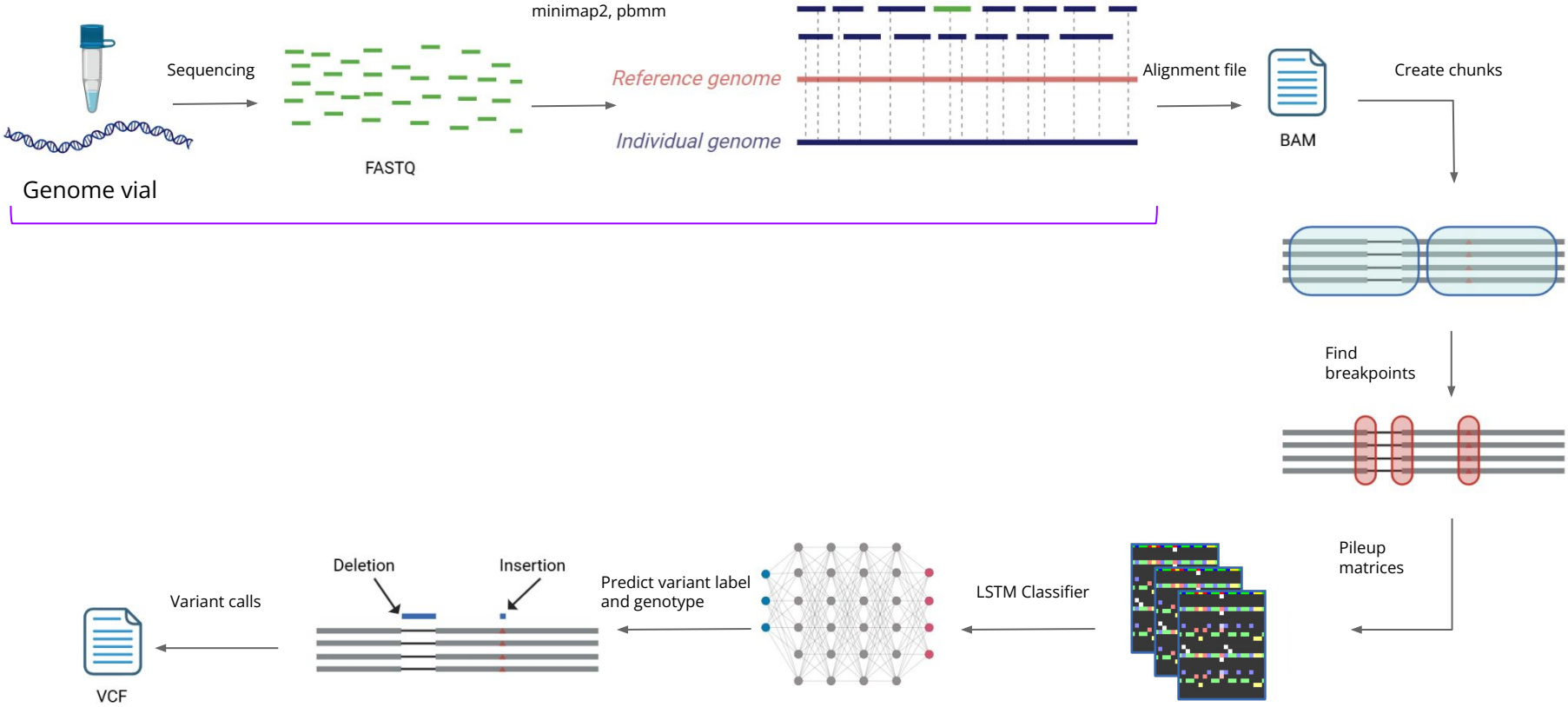
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion

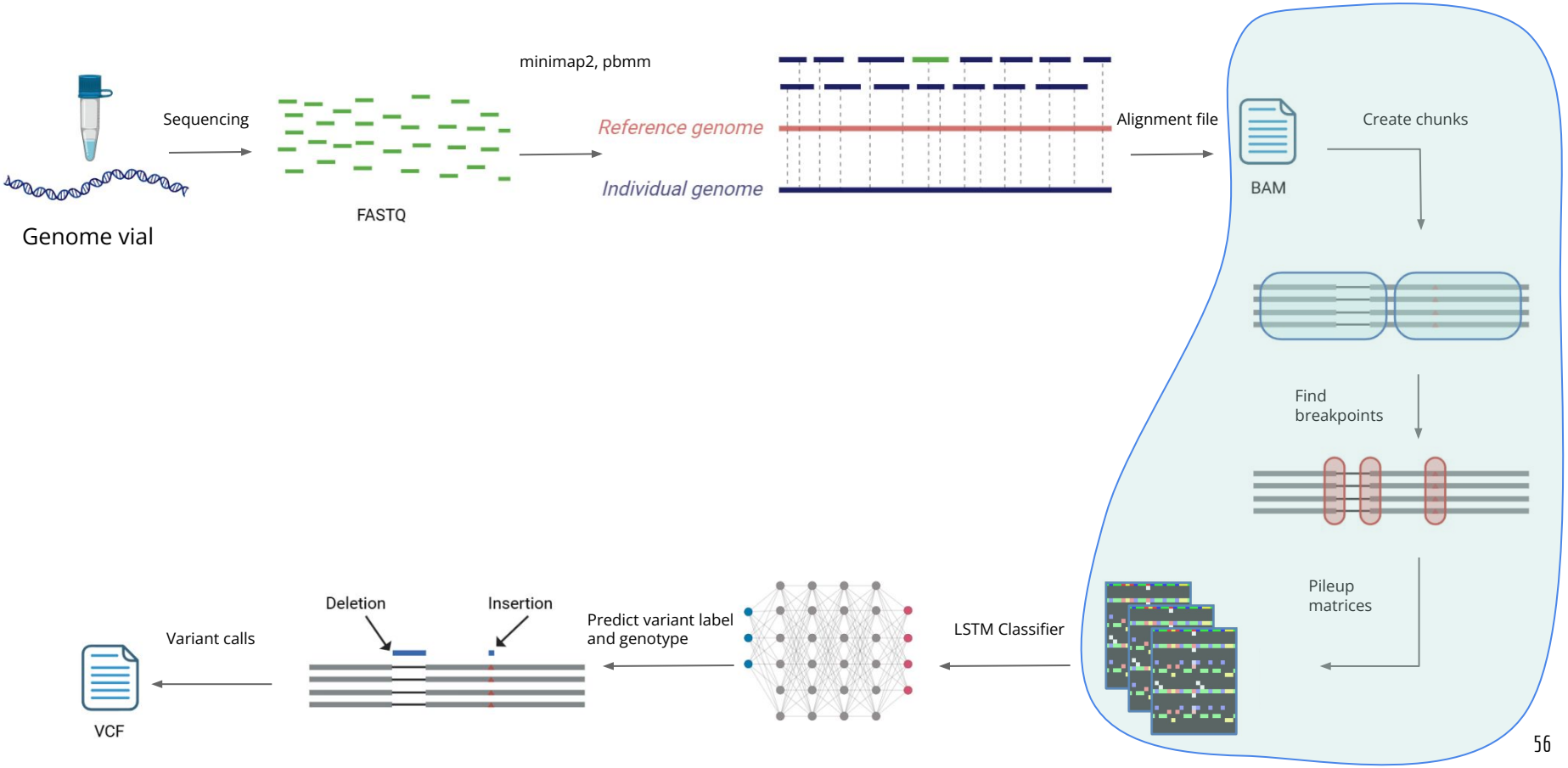
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion
Variant calls
VCF

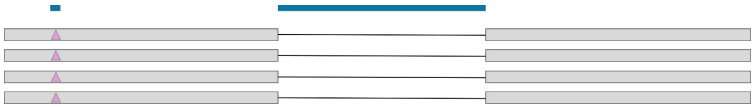
# Overview

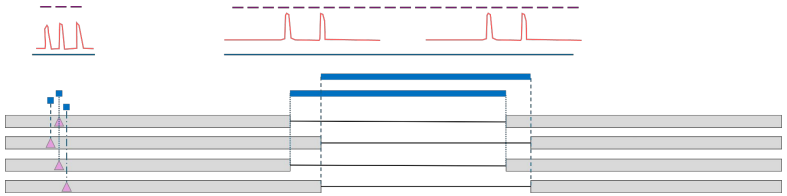

Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion
Variant calls
VCF

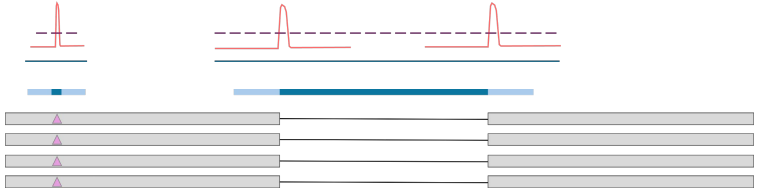
# Overview

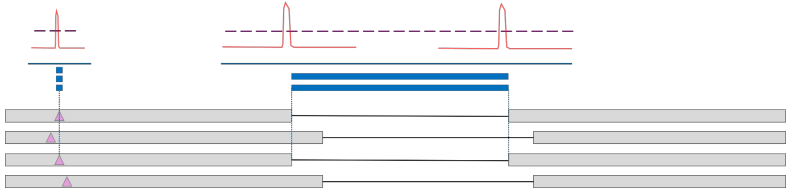

Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion
Variant calls
VCF

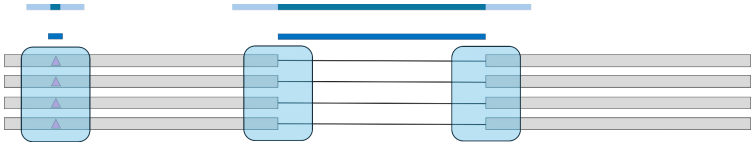
# Candidate Variant Matrix Generation

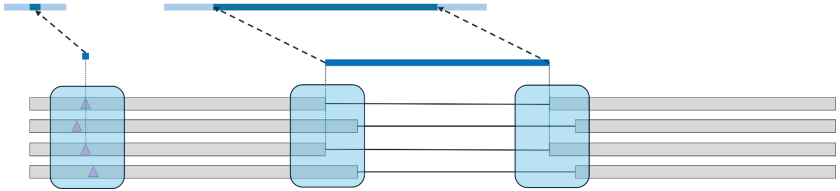




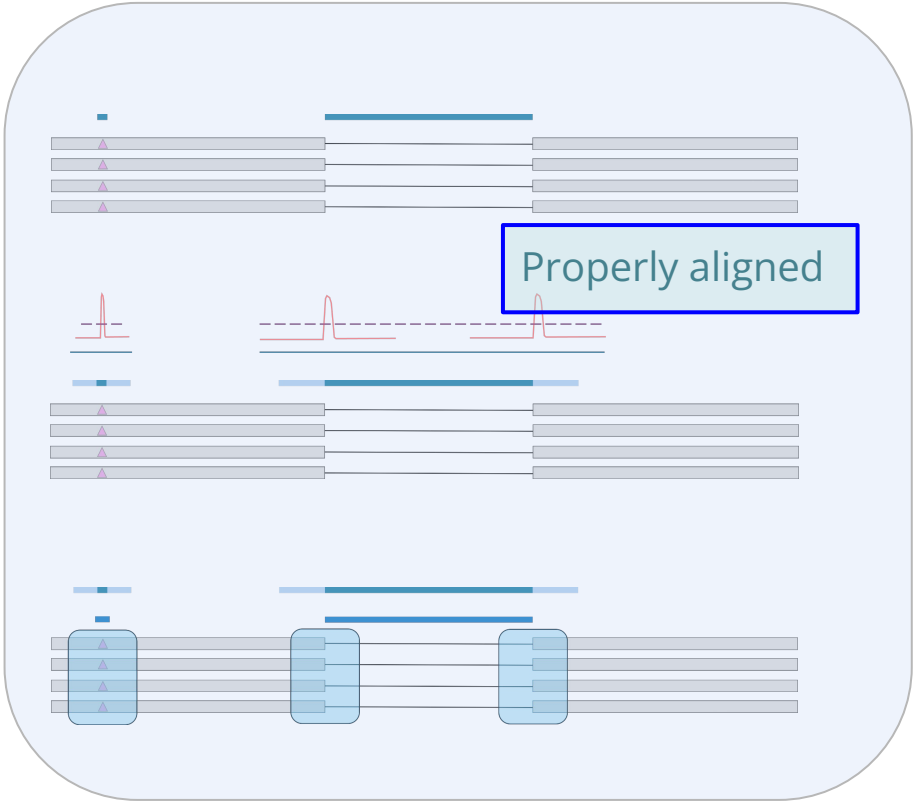


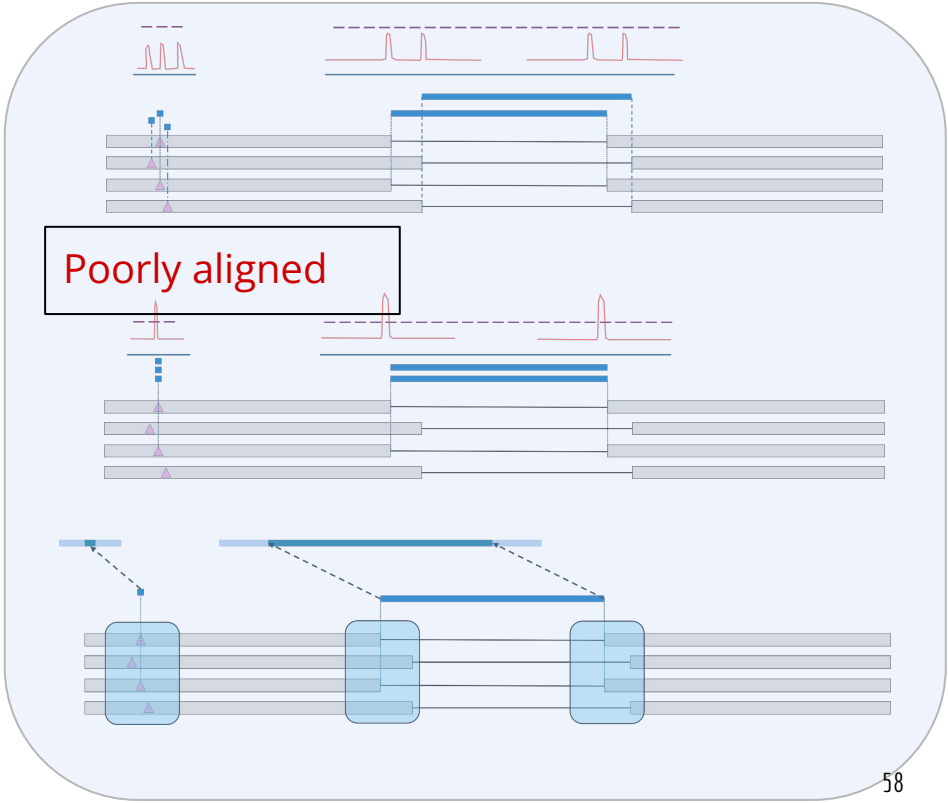




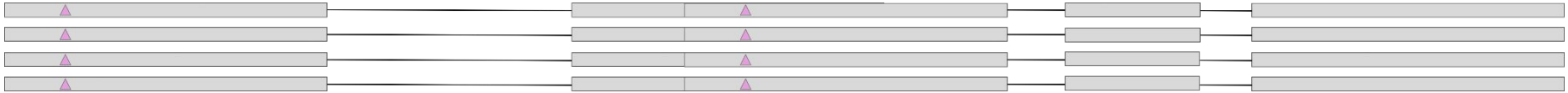


# Candidate Variant Matrix Generation

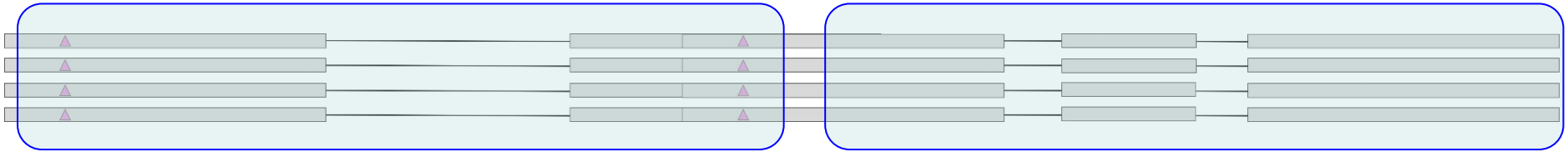


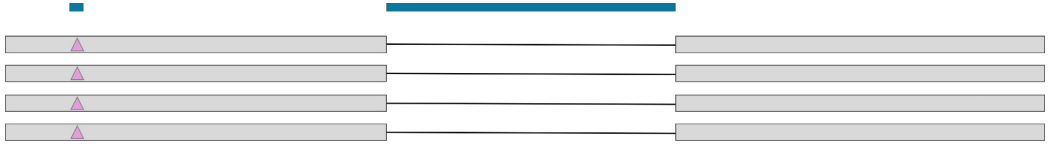
# Candidate Variant Matrix Generation



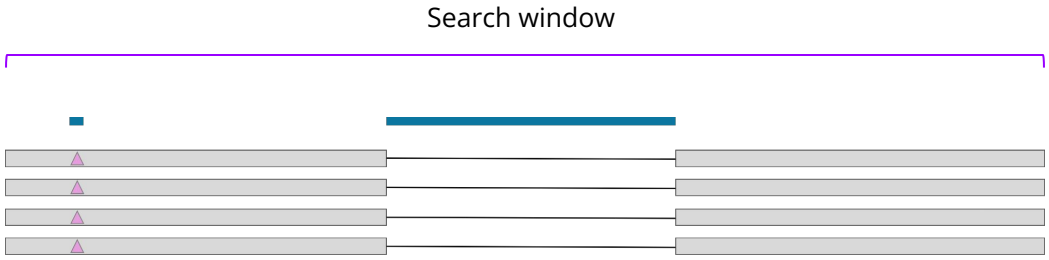
# Candidate Variant Matrix Generation



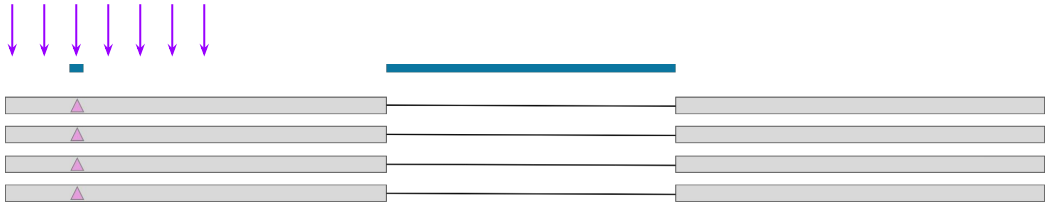
# Candidate Variant Matrix Generation



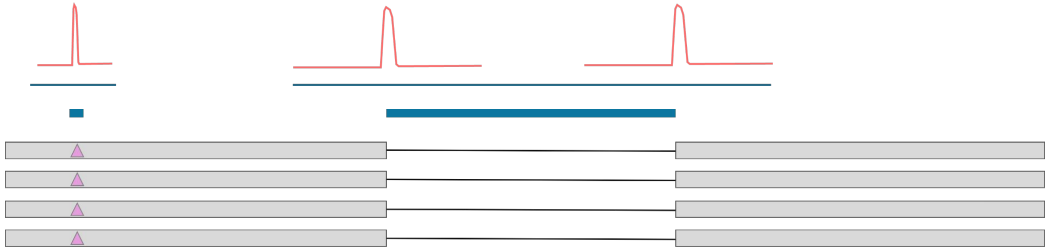
# Candidate Variant Matrix Generation



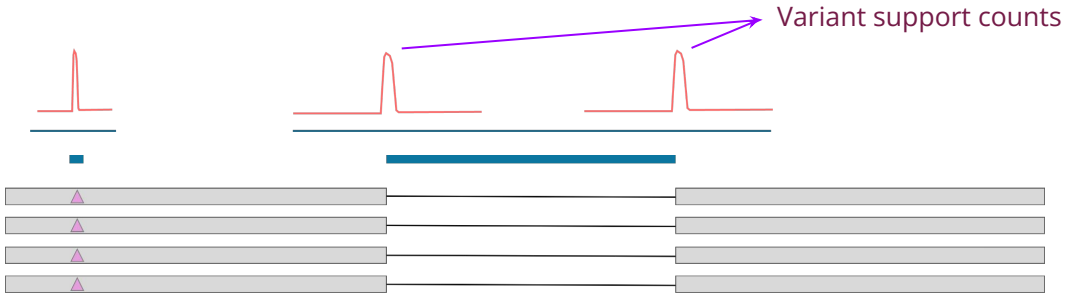
# Candidate Variant Matrix Generation



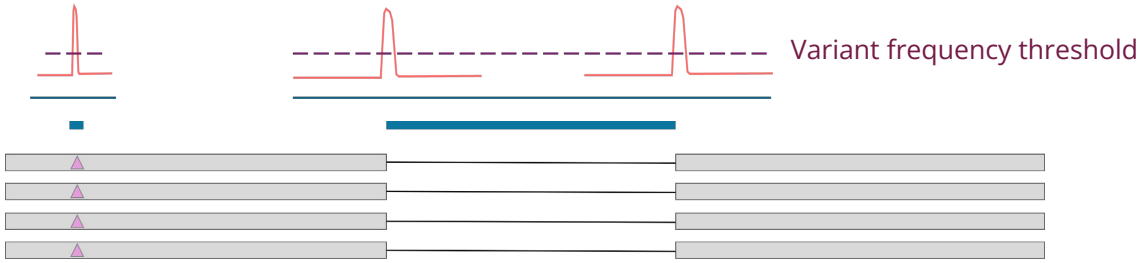
# Candidate Variant Matrix Generation



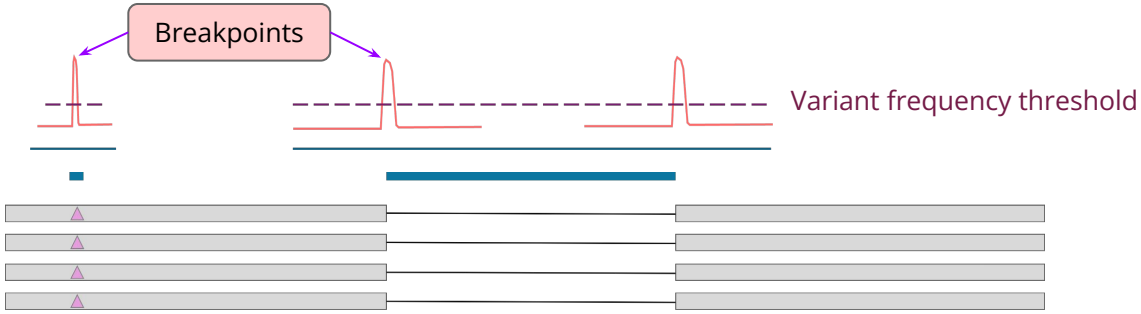
# Candidate Variant Matrix Generation



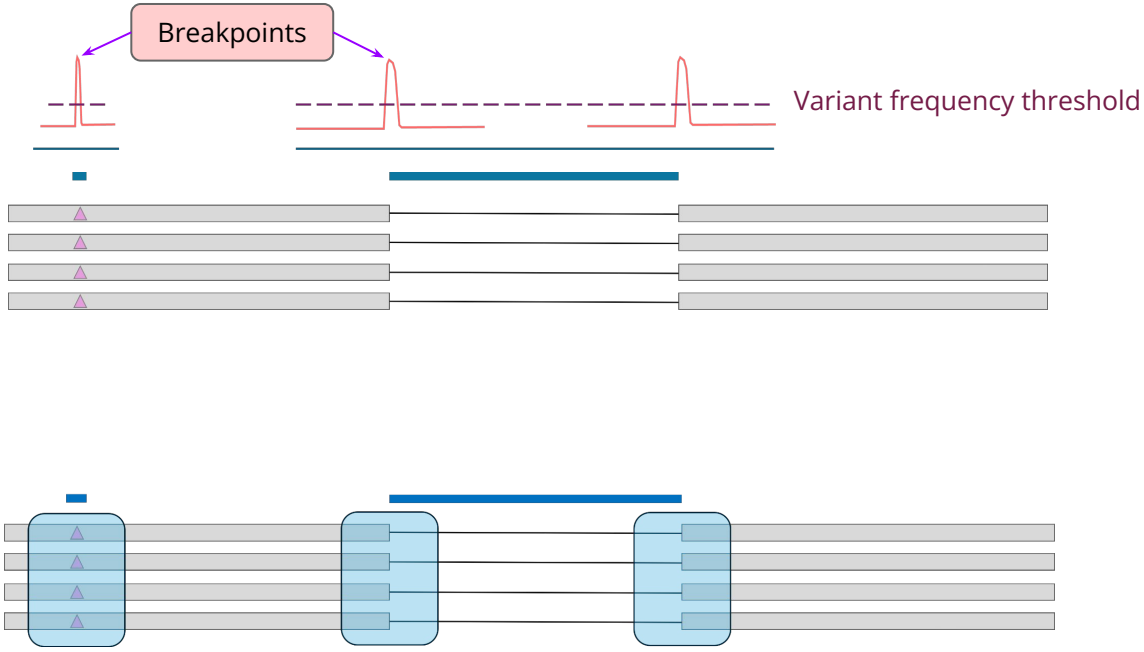
# Candidate Variant Matrix Generation



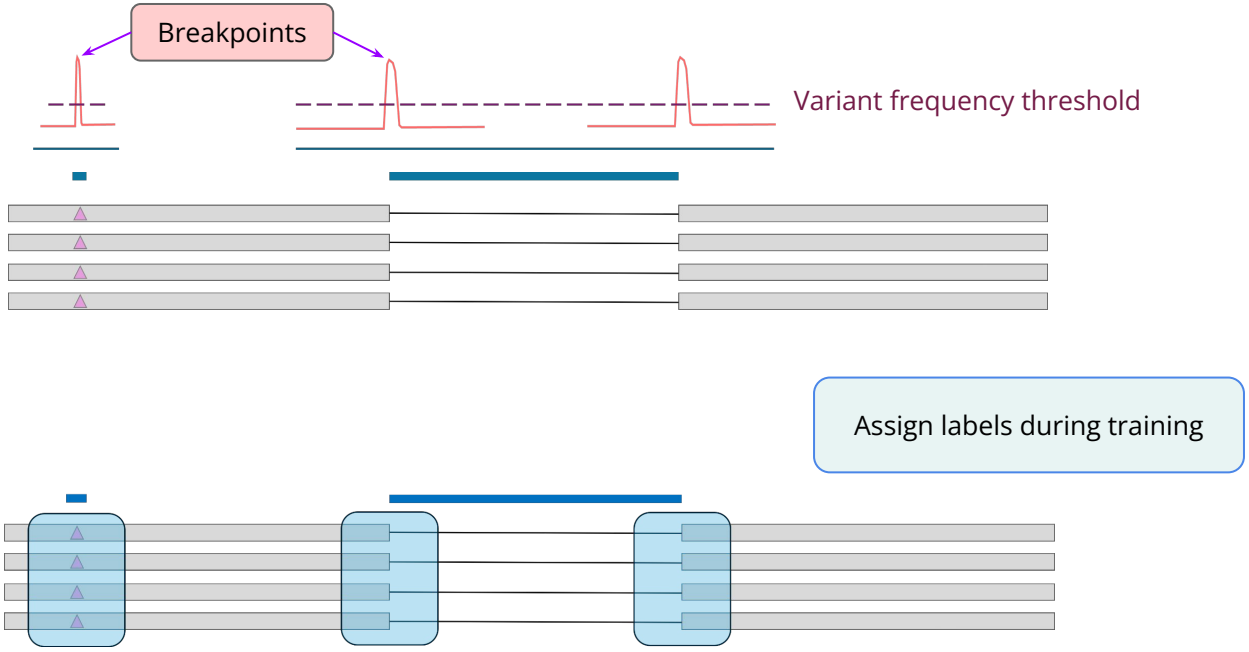
# Candidate Variant Matrix Generation



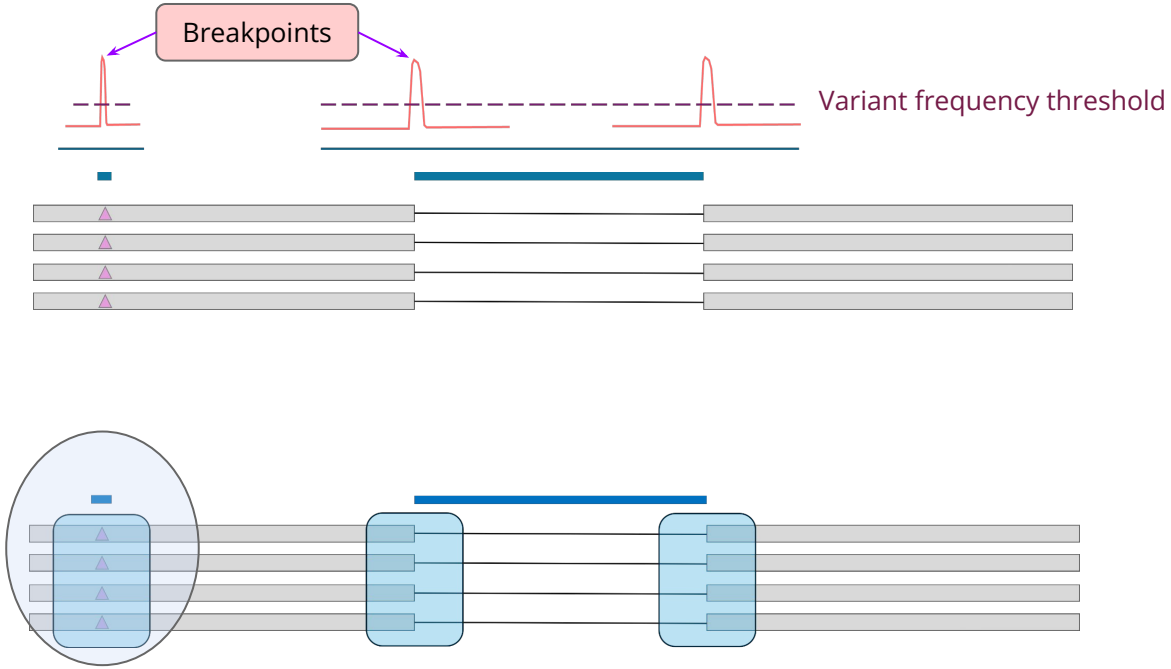
# Candidate Variant Matrix Generation



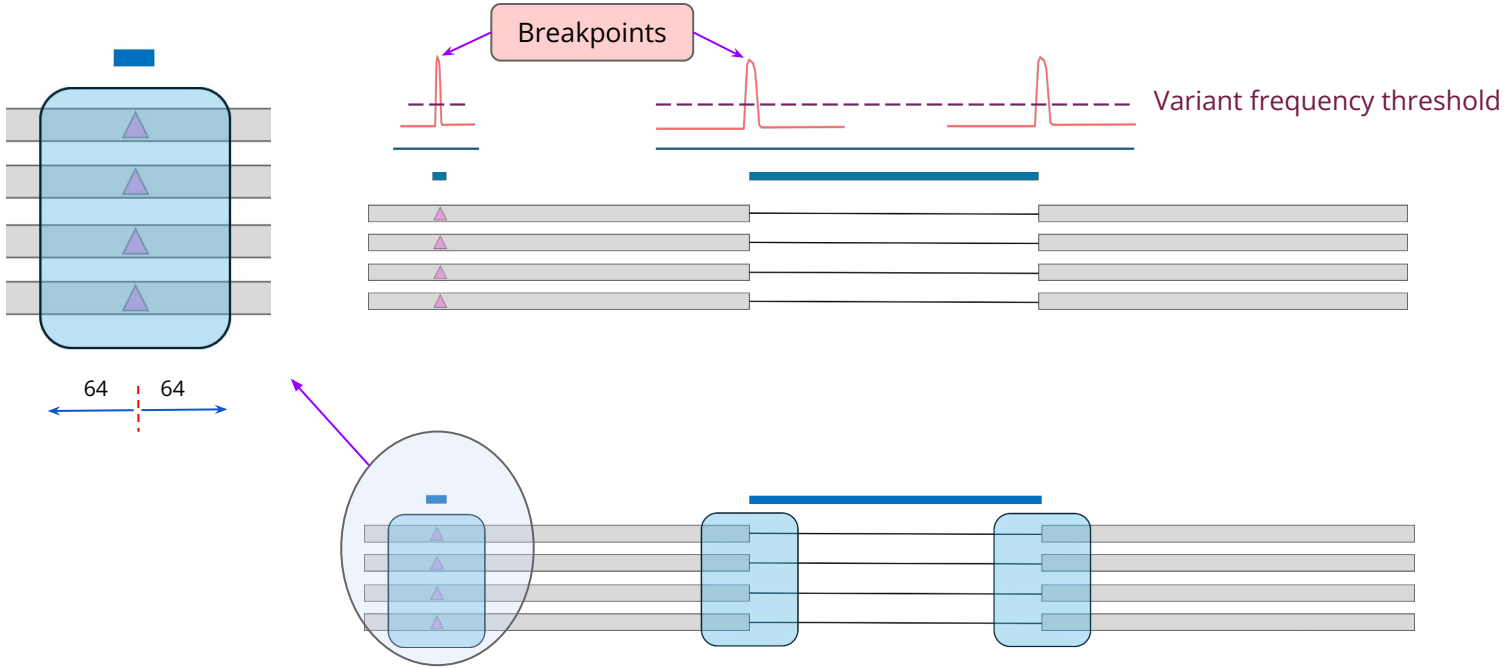
# Candidate Variant Matrix Generation



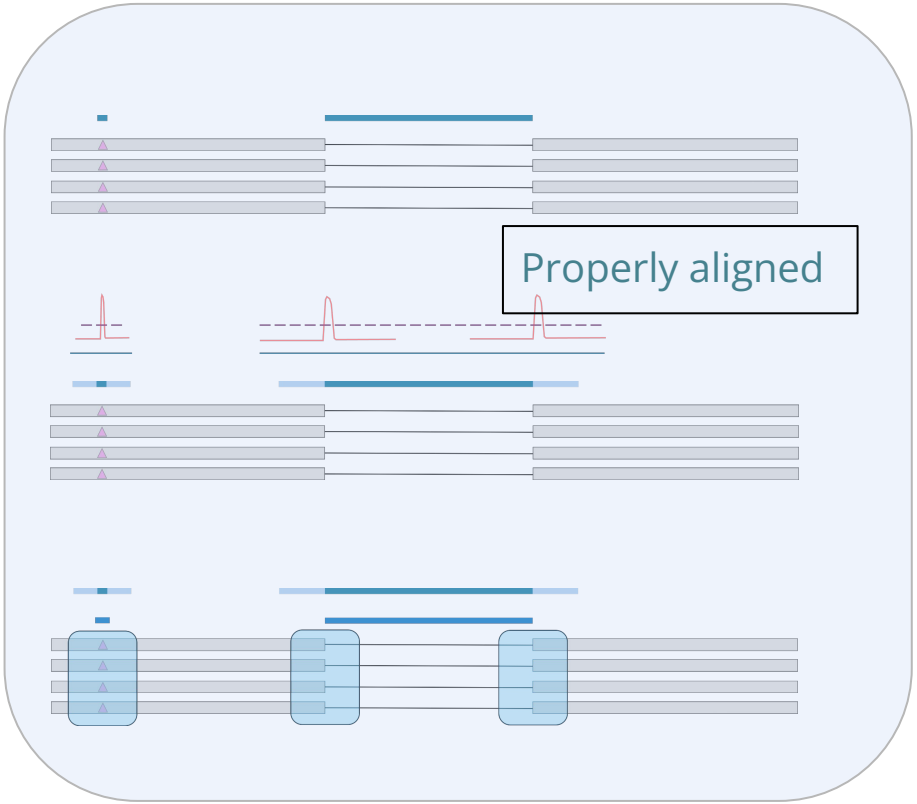
# Candidate Variant Matrix Generation

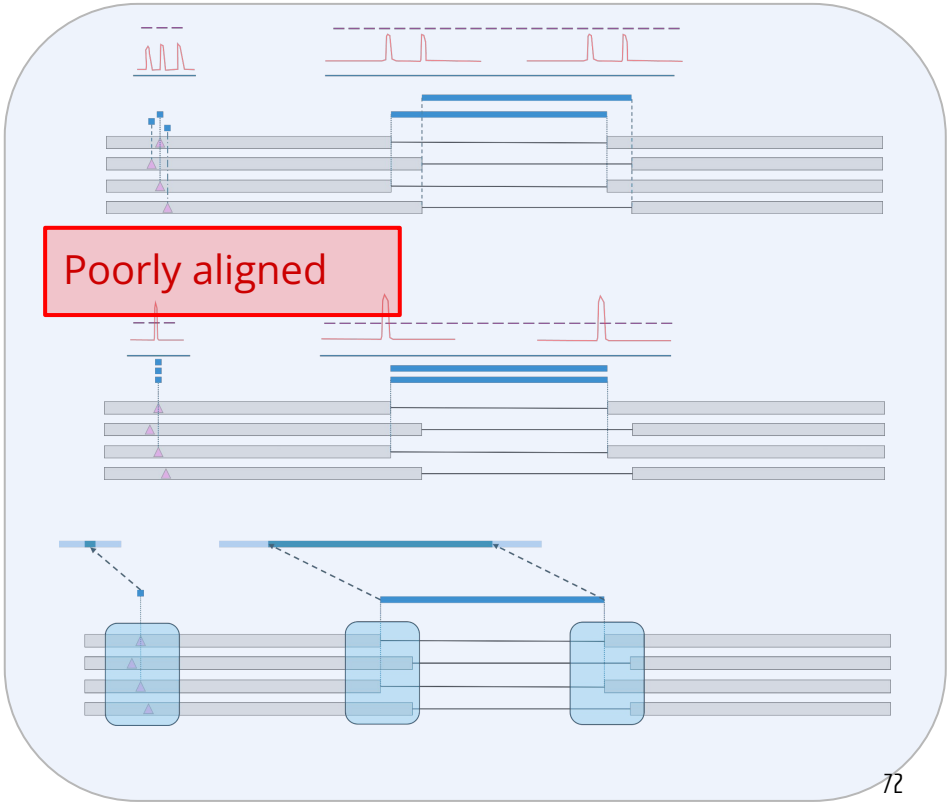


# Candidate Variant Matrix Generation

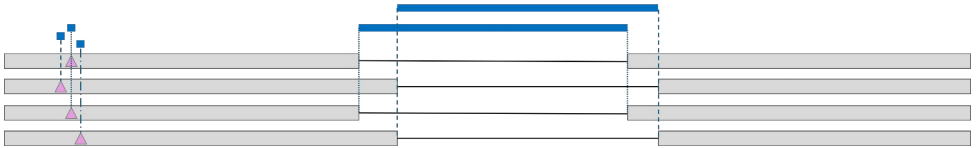


# Candidate Variant Matrix Generation

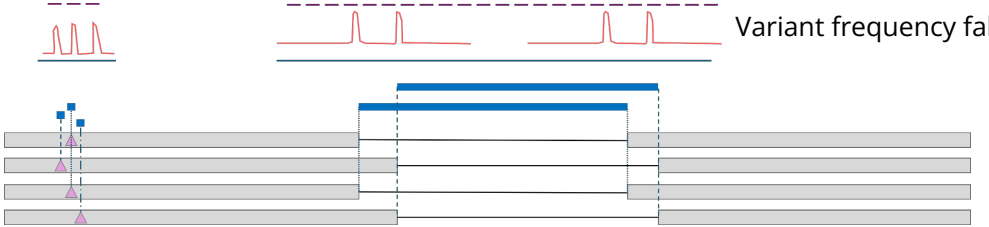


# Candidate Variant Matrix Generation

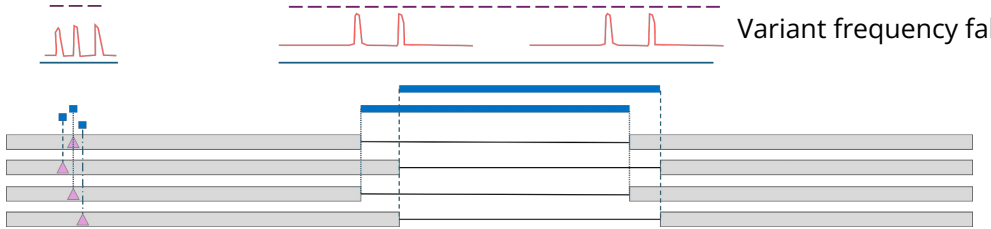


# Candidate Variant Matrix Generation

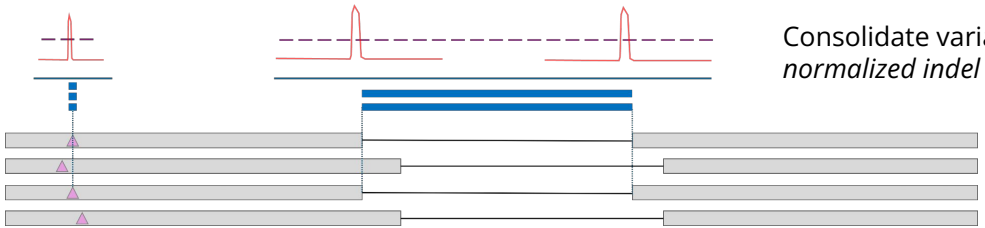


Variant frequency falls below threshold

# Candidate Variant Matrix Generation

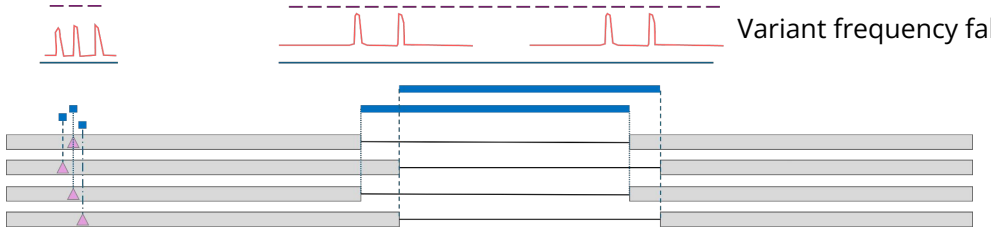


Variant frequency falls below threshold

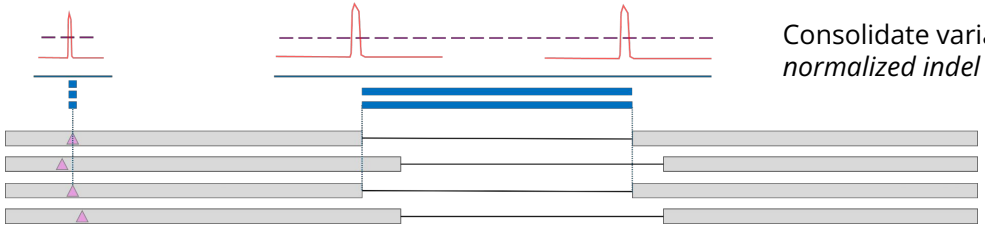


Consolidate variants based on
*normalized indel similarity*

# Candidate Variant Matrix Generation



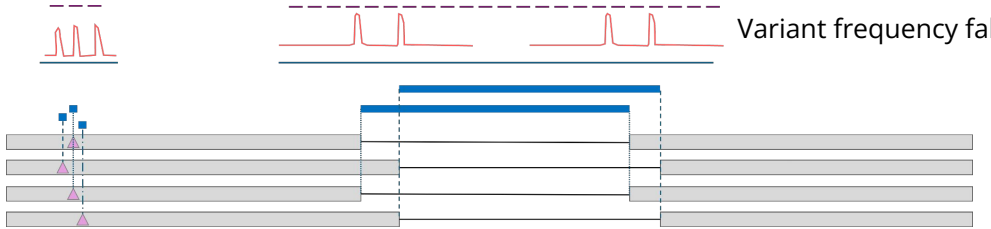
Variant frequency falls below threshold



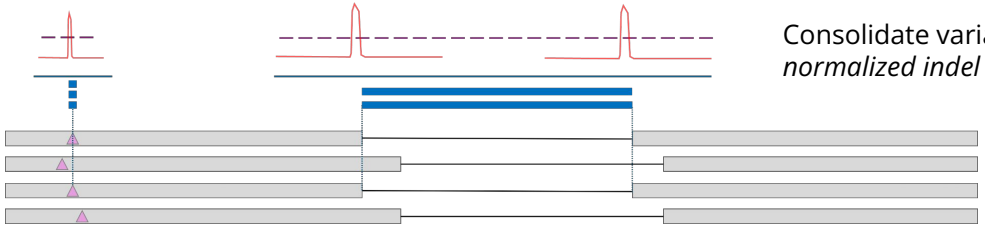
Consolidate variants based on *normalized indel similarity*

$$D = \alpha \cdot n_{\text{mis}} + \beta \cdot n_{\text{ins}} + \gamma \cdot n_{\text{del}}$$

# Candidate Variant Matrix Generation



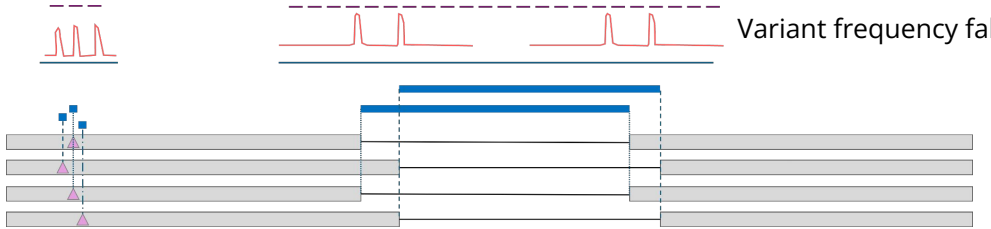
Variant frequency falls below threshold



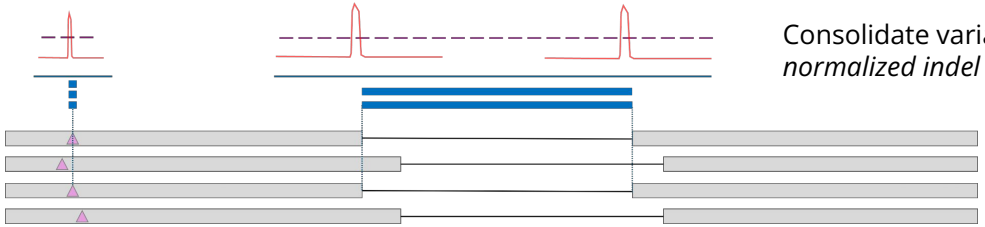
Consolidate variants based on *normalized indel similarity*

$$D = \alpha \cdot n_{\text{mis}} + \beta \cdot n_{\text{ins}} + \gamma \cdot n_{\text{del}} \qquad (\alpha = 2, \beta = 1, \gamma = 1)$$

# Candidate Variant Matrix Generation



Variant frequency falls below threshold

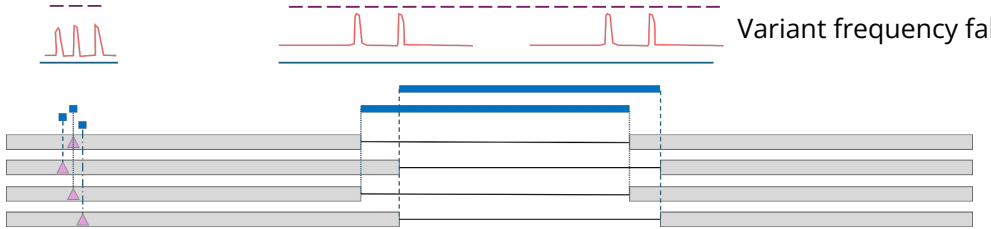


Consolidate variants based on *normalized indel similarity*

$$D = \alpha \cdot n_{\text{mis}} + \beta \cdot n_{\text{ins}} + \gamma \cdot n_{\text{del}} \qquad (\alpha = 2, \beta = 1, \gamma = 1)$$

$$D_{\text{norm}} = \frac{D}{L_1 + L_2}$$

# Candidate Variant Matrix Generation



Variant frequency falls below threshold

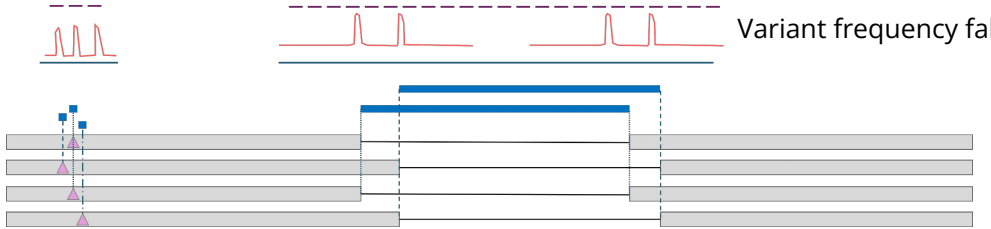
Consolidate variants based on *normalized indel similarity*

$$D = \alpha \cdot n_{\text{mis}} + \beta \cdot n_{\text{ins}} + \gamma \cdot n_{\text{del}} \qquad (\alpha = 2, \beta = 1, \gamma = 1)$$

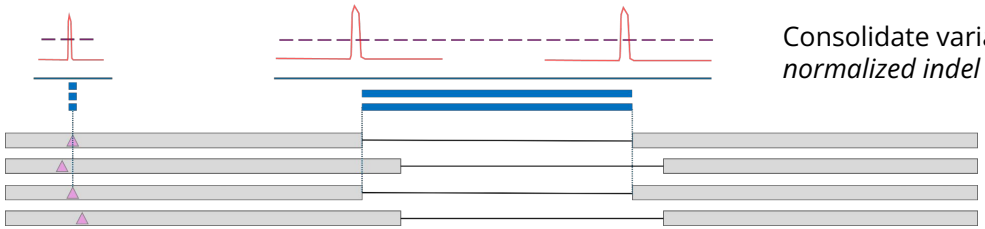
$$D_{\text{norm}} = \frac{D}{L_1 + L_2}$$

$$S_{\text{norm}} = 1 - D_{\text{norm}}$$

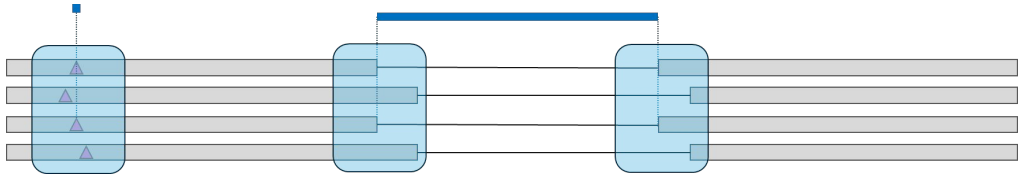
# Candidate Variant Matrix Generation



Variant frequency falls below threshold



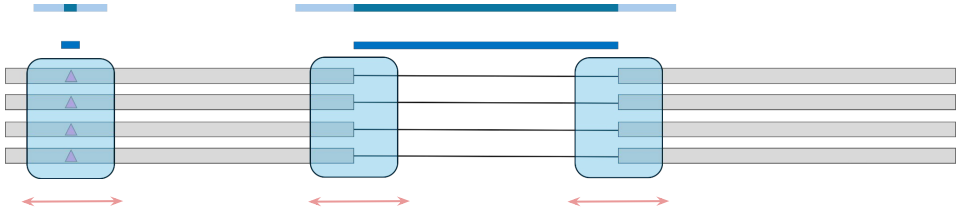
Consolidate variants based on *normalized indel similarity*

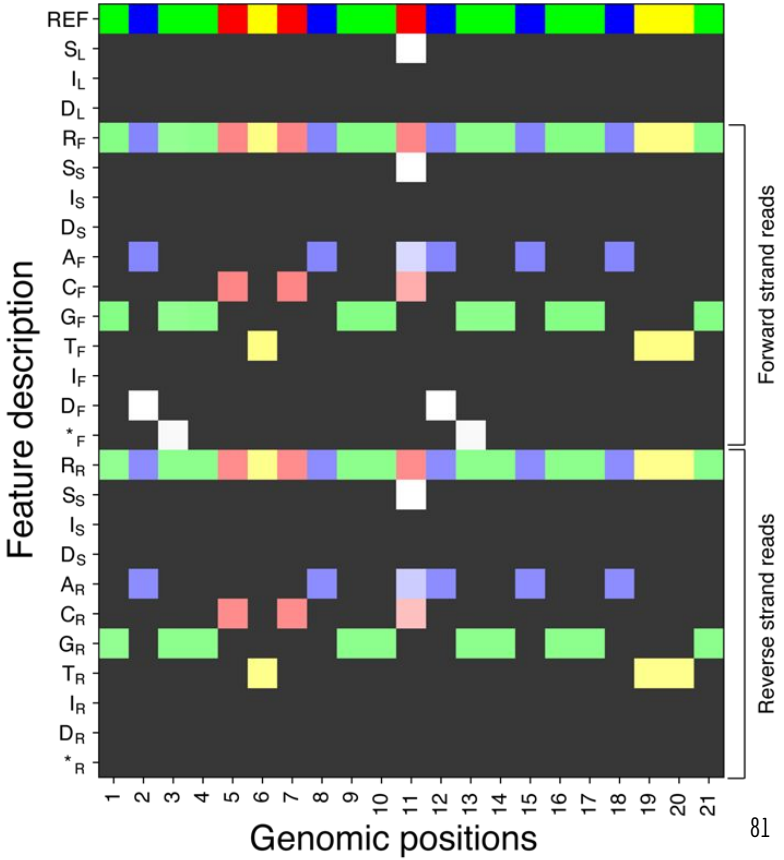


Candidate windows on breakpoints

# Matrix Features

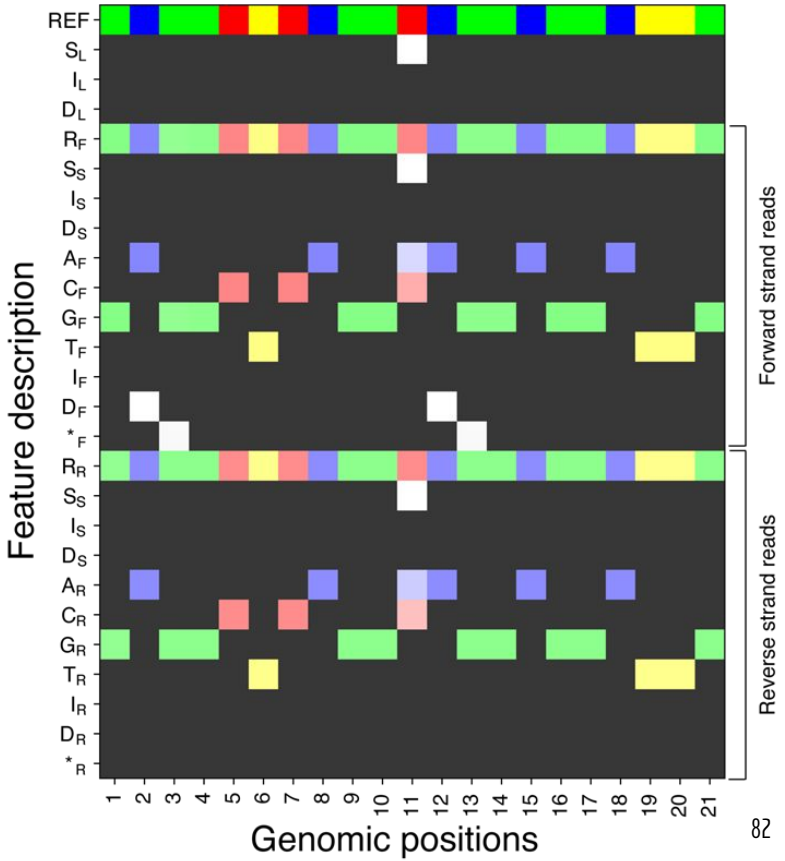
- 26 Initial features, 128 width





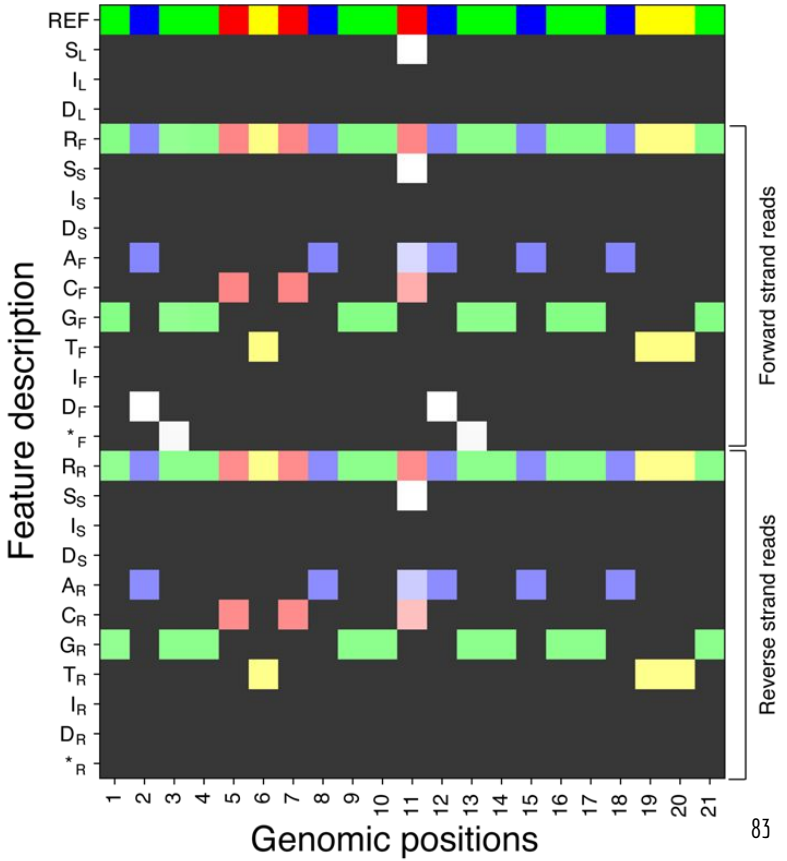
# Matrix Features

- 26 Initial features, 128 width
- $S_L$,$I_L$,$D_L$: Variant length
- $R_F$: Reference support
- $S_S$,$I_S$,$D_S$: Variant support
- $A_F$,$C_F$,$T_F$,$G_F$: Nucleotide count
- $I_F$, $D_F$: Indel breakpoint count
- $*_F$: Total deletes observed



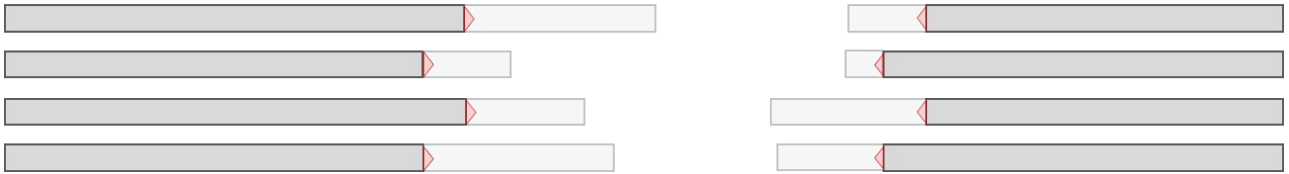
# Matrix Features

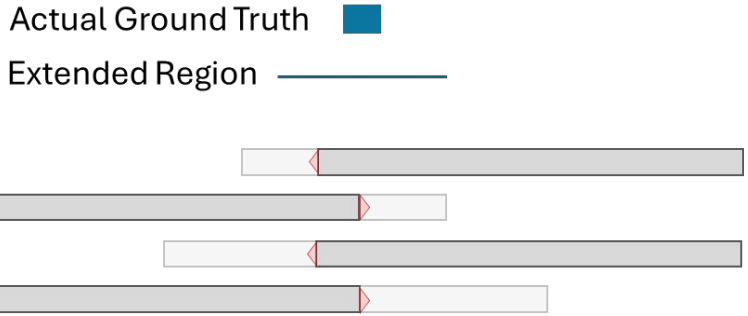
- Two newly added features:
  a. $SC_F$: Soft clip count for forward strand
  b. $SC_R$: Soft clip count for reverse strand



# Variants from Soft Clipped Regions for Model Training

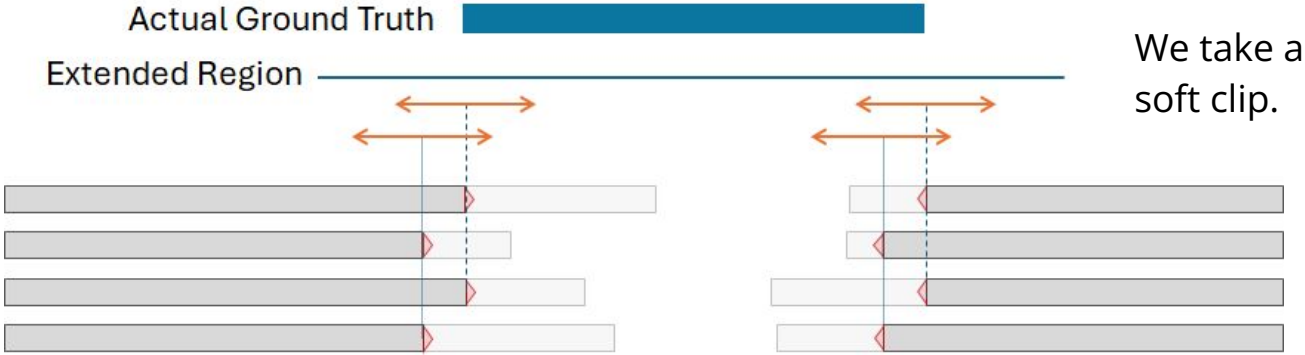

This alignment misrepresented a deletion as soft clipped region.



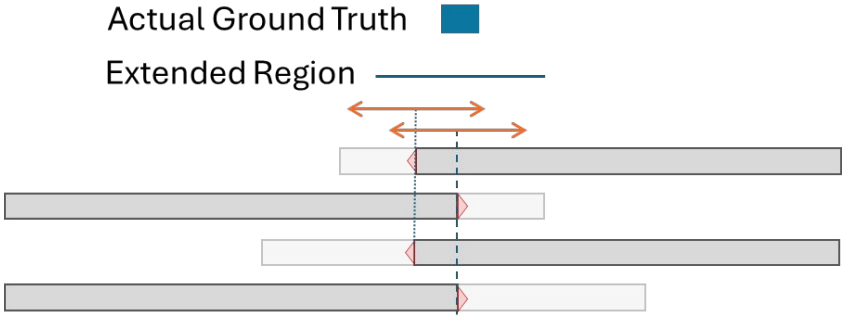



This alignment misrepresented an insertion as soft clipped region.

# Variants from Soft Clipped Regions for Model Training

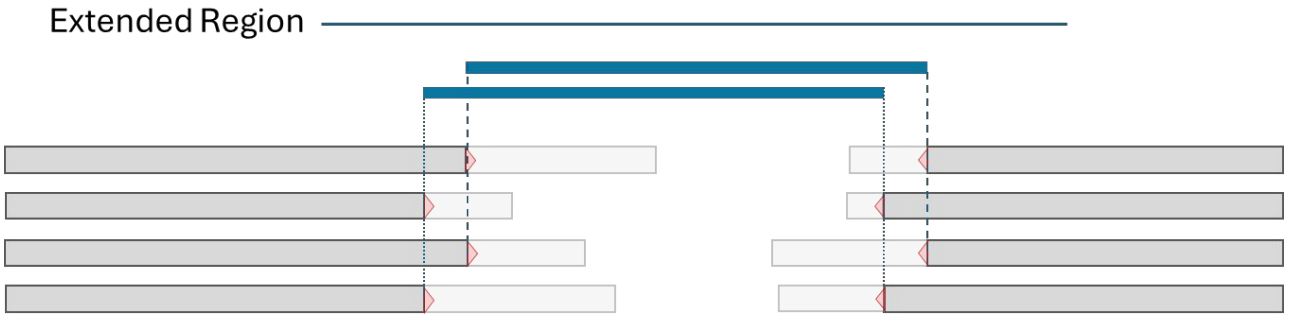


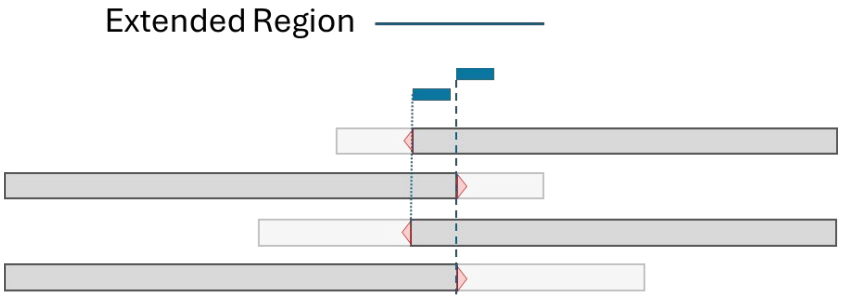

We take a small window around the soft clip.



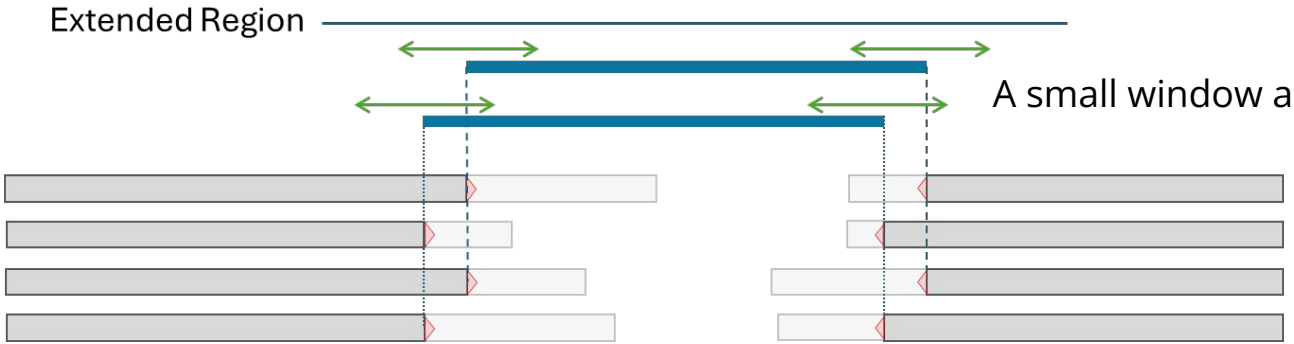

Then the category information is taken from the truth VCF.

# Variants from Soft Clipped Regions for Model Training

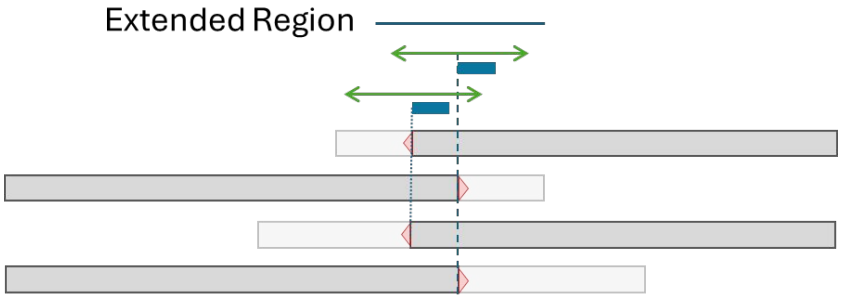


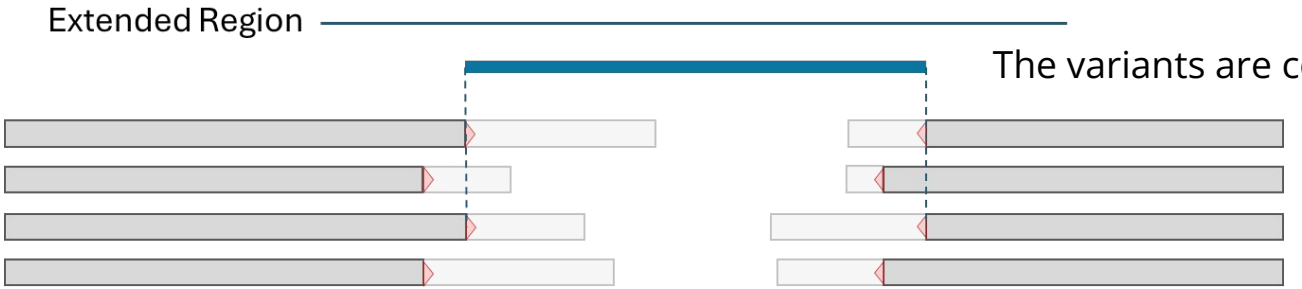
# Variants from Soft Clipped Regions for Model Training



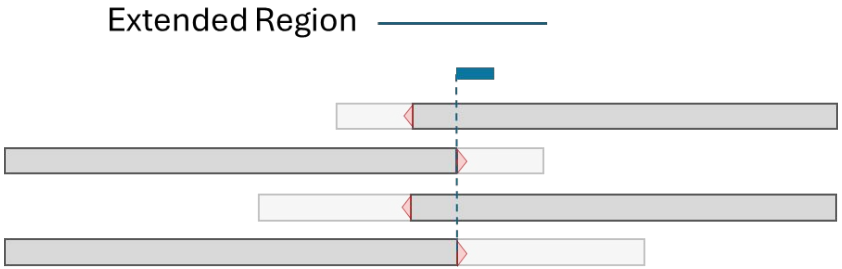

A small window around the soft clip is taken.



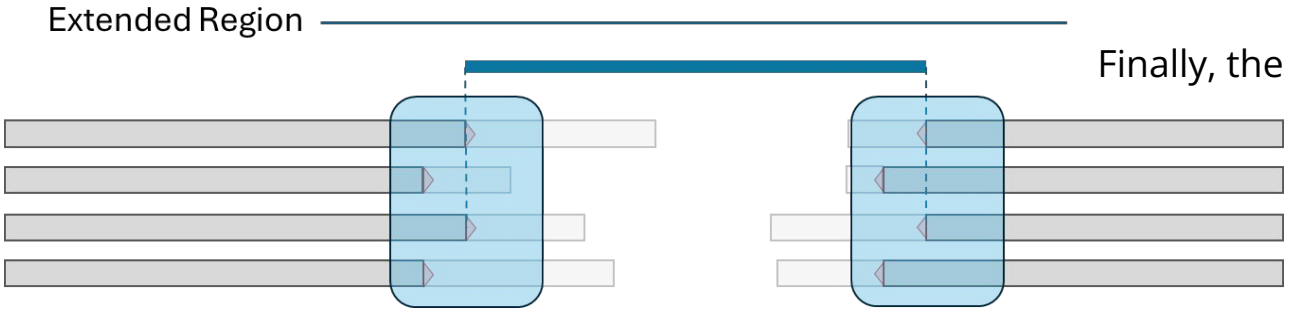
# Variants from Soft Clipped Regions for Model Training



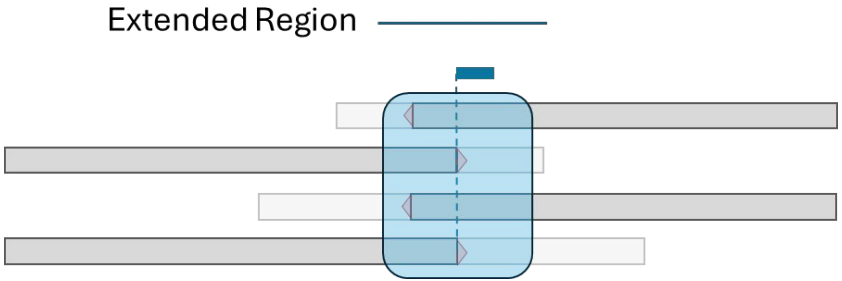

The variants are consolidated to the first one.



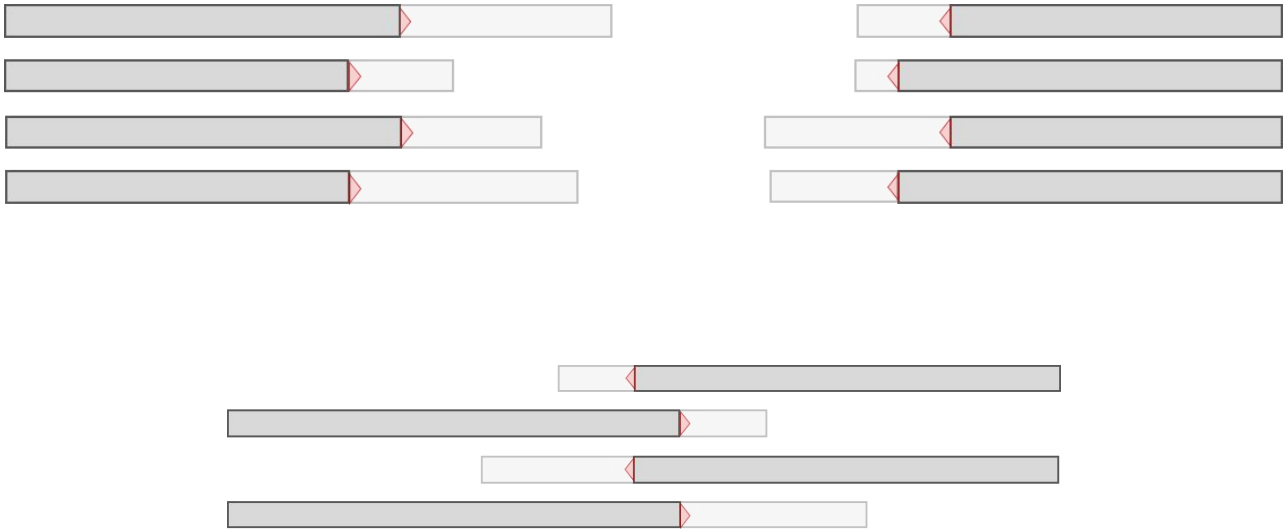
# Variants from Soft Clipped Regions for Model Training




Finally, the variant matrix is generated.

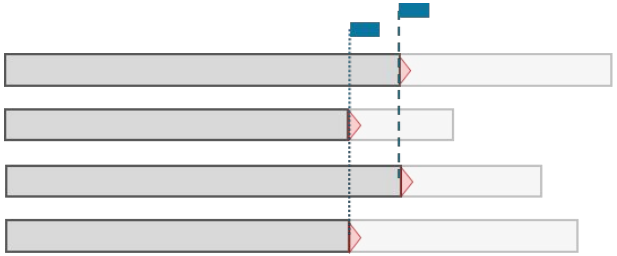


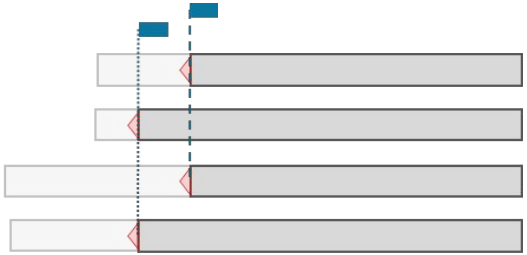
# Variants from Soft Clipped Regions for Prediction

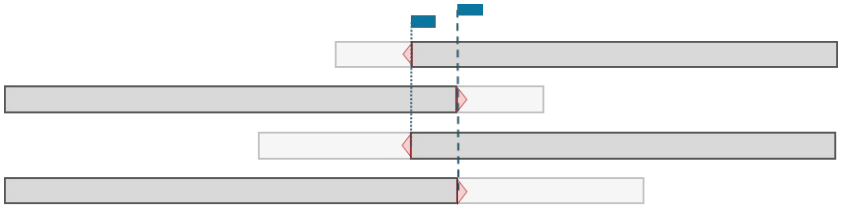


For soft clipped regions, it is not possible to categorize regions as insertion or deletion accurately from the aligned reads.

# Variants from Soft Clipped Regions for Prediction

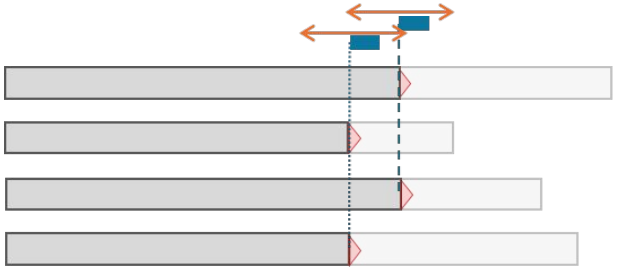


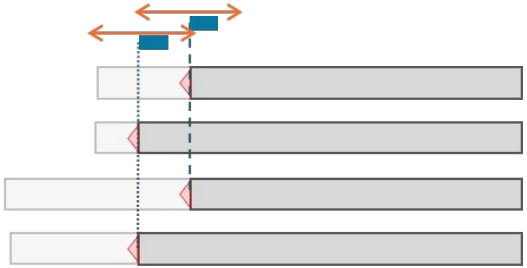




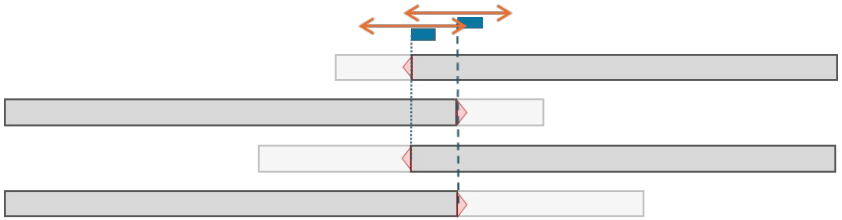
The variant is given a generic category (insertion) to pass to our model.

# Variants from Soft Clipped Regions for Prediction

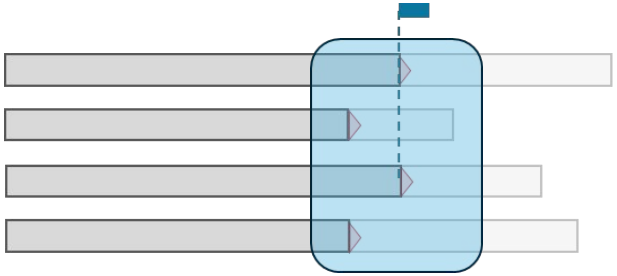


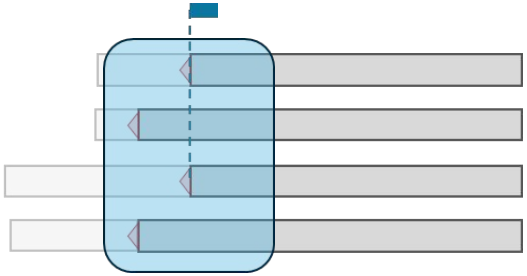


Then, we consolidate them.

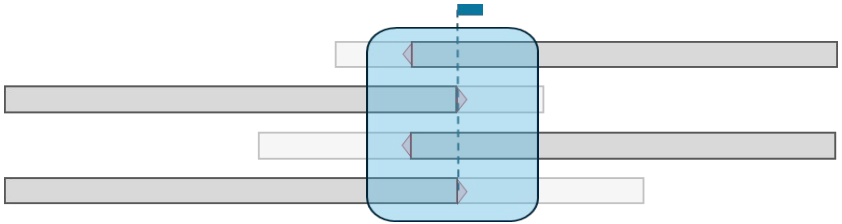


# Variants from Soft Clipped Regions for Prediction





Finally, the variant matrix is generated.

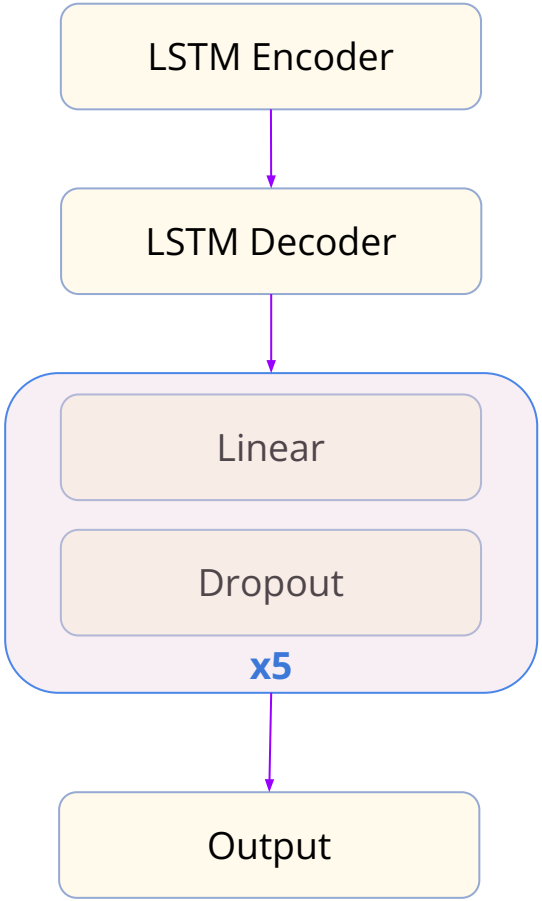


# Overview of the Model

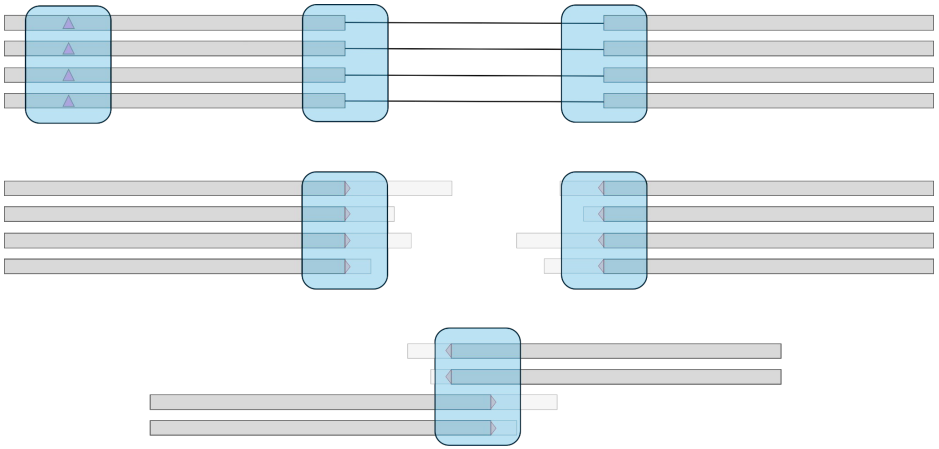
- Parameters: 2.7M
- Learning rate: 0.0001

| | |
|---|---|
| Train dataset | chr 1 - 14 |
| Test dataset | chr 15 - 22 |

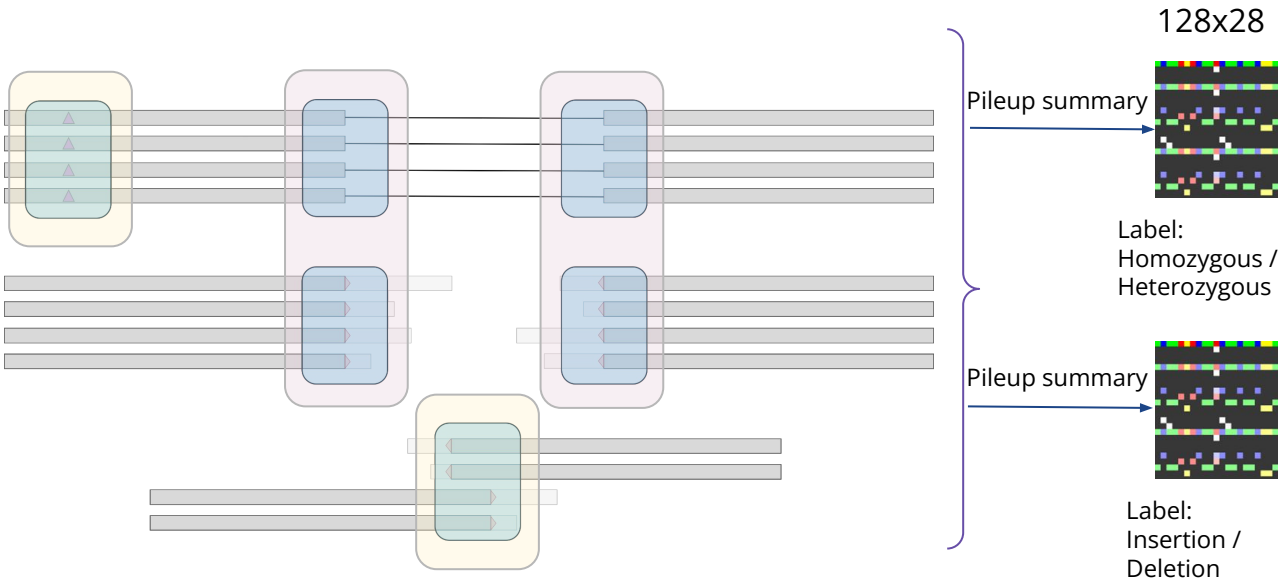
- Train-test ratio ≈ 80 / 20
- Epochs: 100



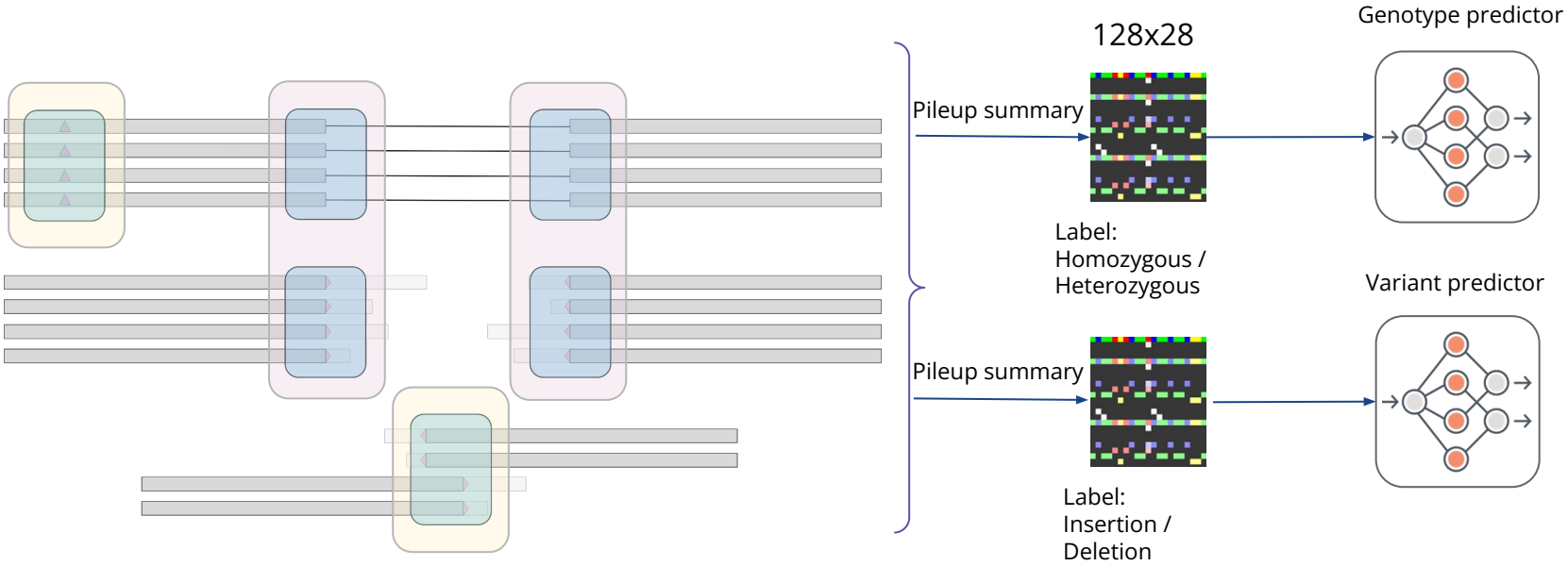
# Model Training



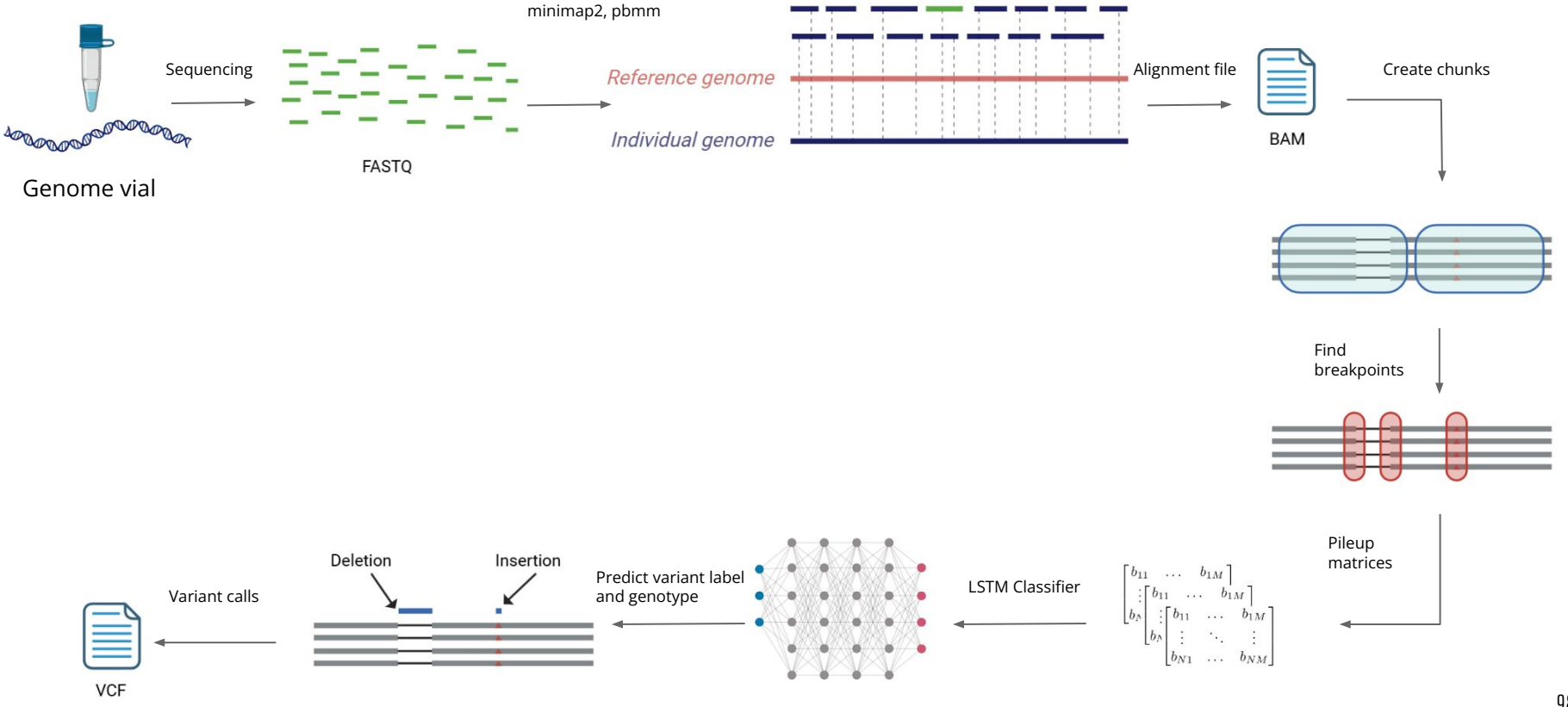
# Model Training



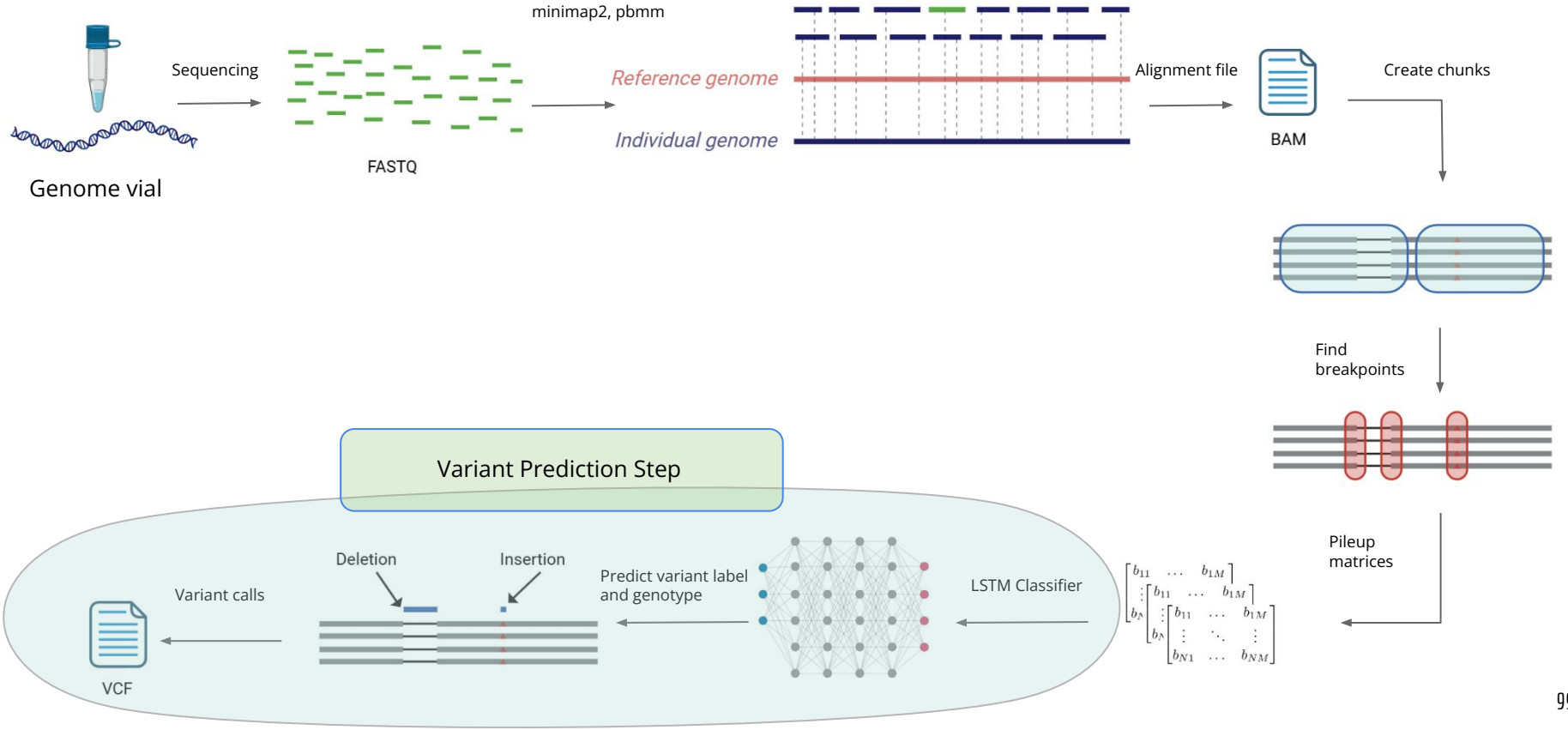
# Model Training



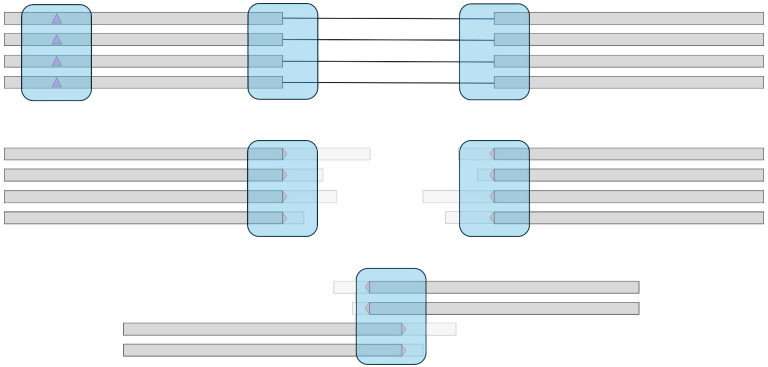
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion
Variant calls
VCF

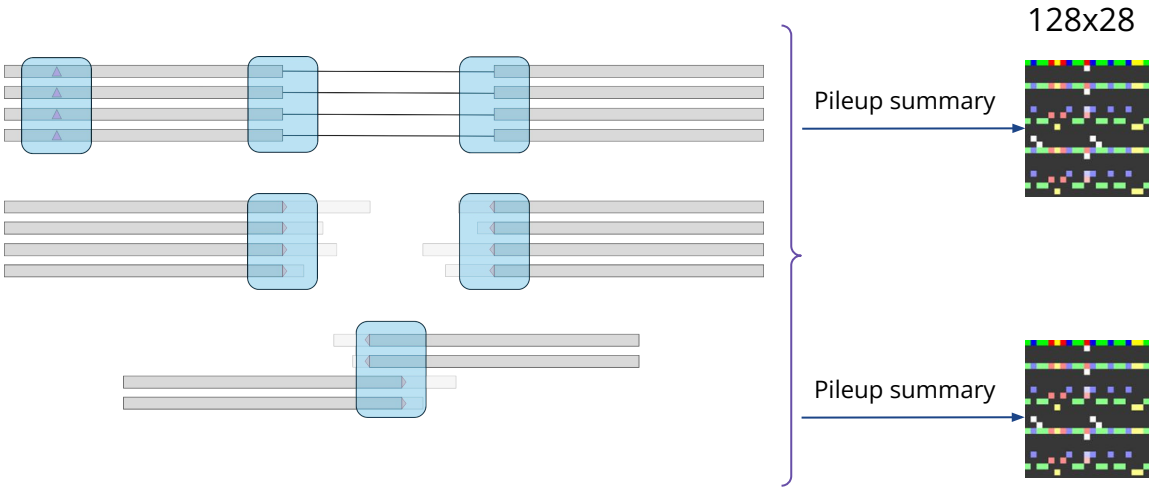
# Overview


Genome vial
Sequencing
FASTQ
minimap2, pbmm
Reference genome
Individual genome
Alignment file
BAM
Create chunks
Find breakpoints
Pileup matrices
Variant Prediction Step
LSTM Classifier
Predict variant label and genotype
Deletion
Insertion
Variant calls
VCF

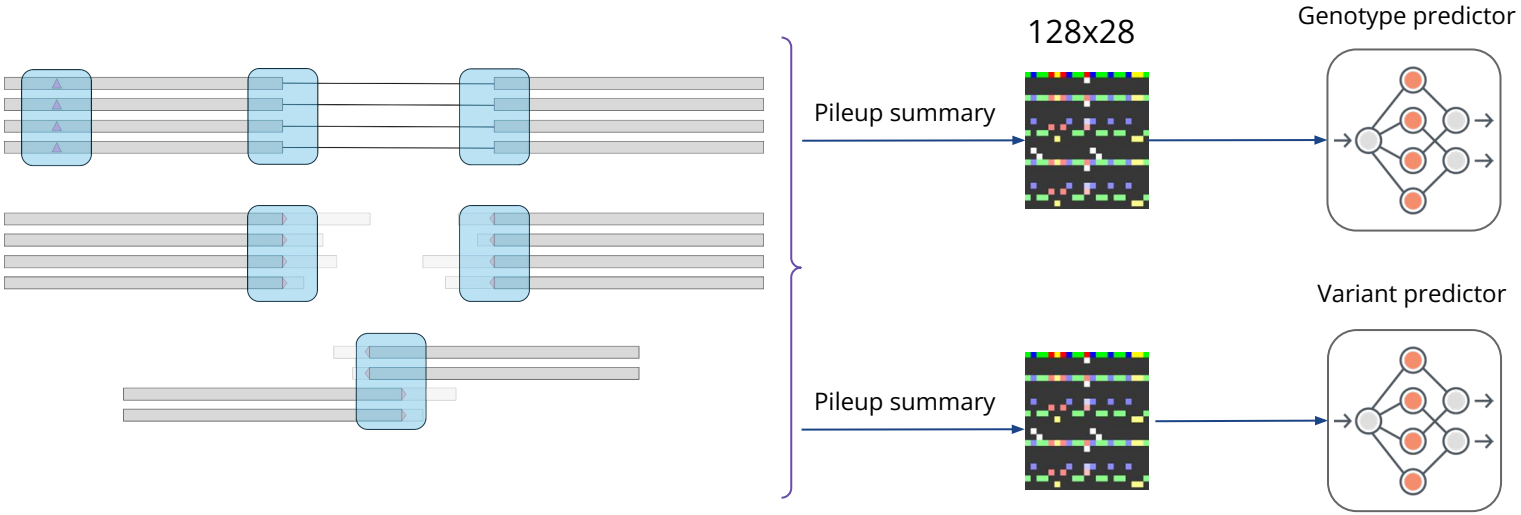
# Variant Prediction



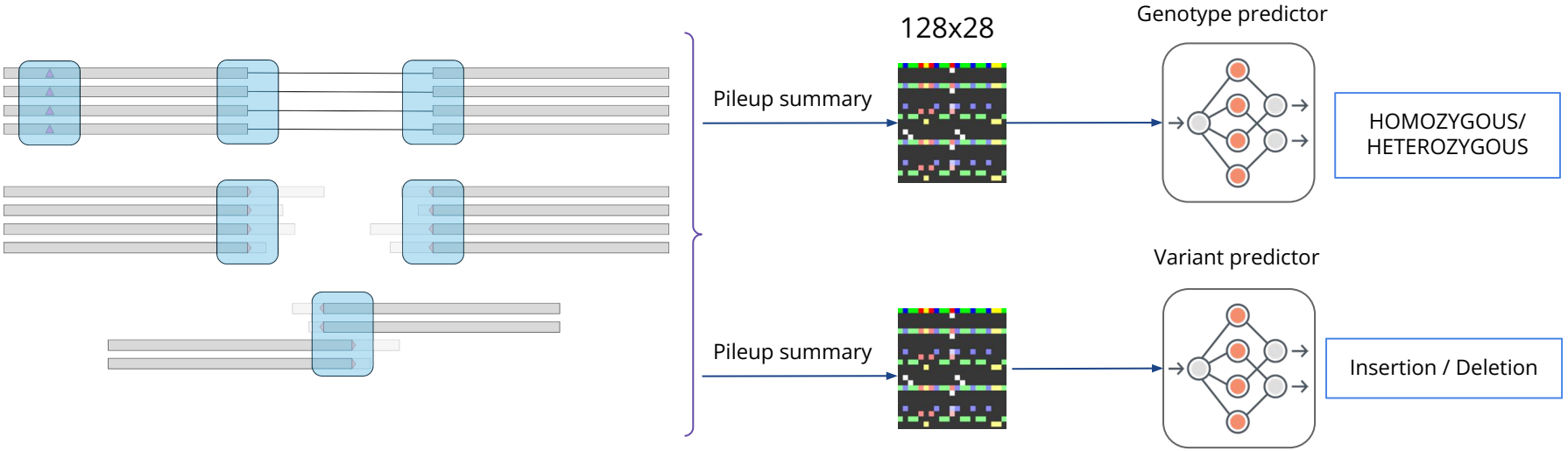
# Variant Prediction



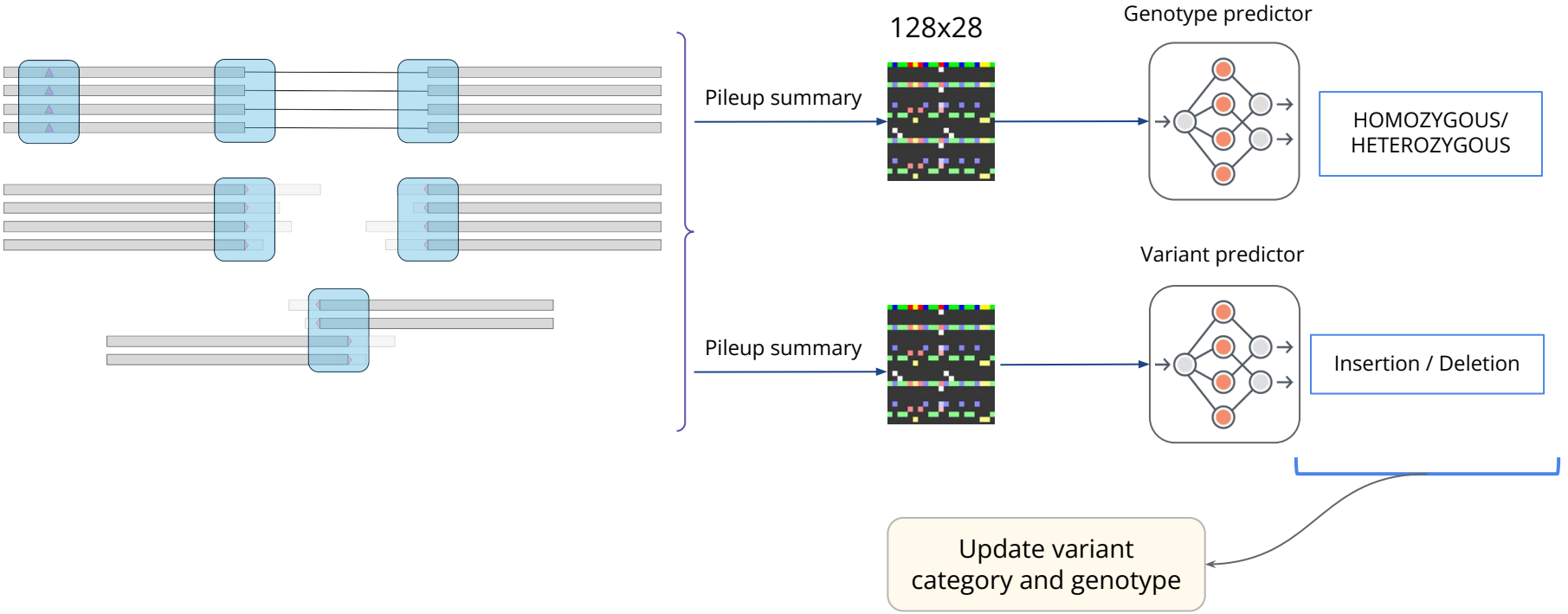
# Variant Prediction



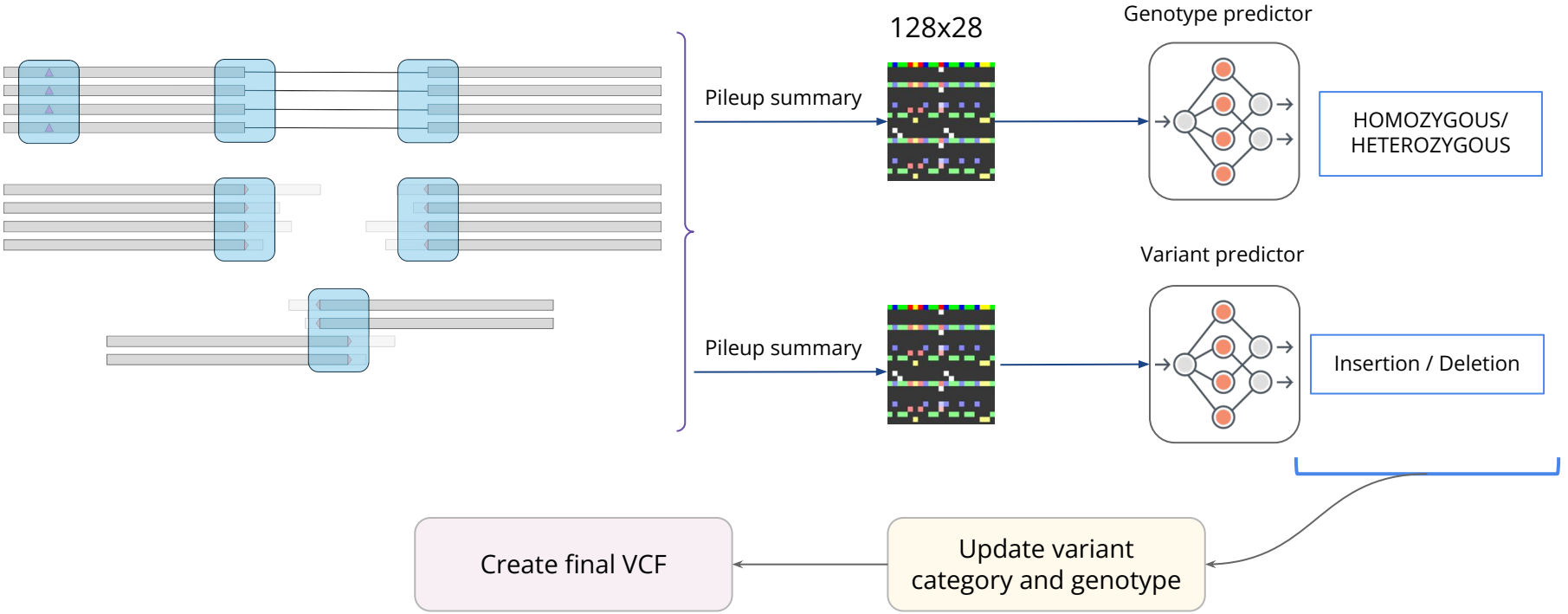
# Variant Prediction



# Variant Prediction



# Variant Prediction



# Dataset

- HG002_35x_HiFi_2_GRCh37
  - HG002 is a **human genome sample** from the Genome in a Bottle (GIAB) project.
  - The reads are generated by **PacBio HiFi (High-Fidelity)** sequencing technology.
  - Genome Reference Consortium Human Build 37 (GRCh37) is a specific version of the **human reference genome**, released in February 2009.
  - The reads are aligned by **minimap2 with 35x coverage**.

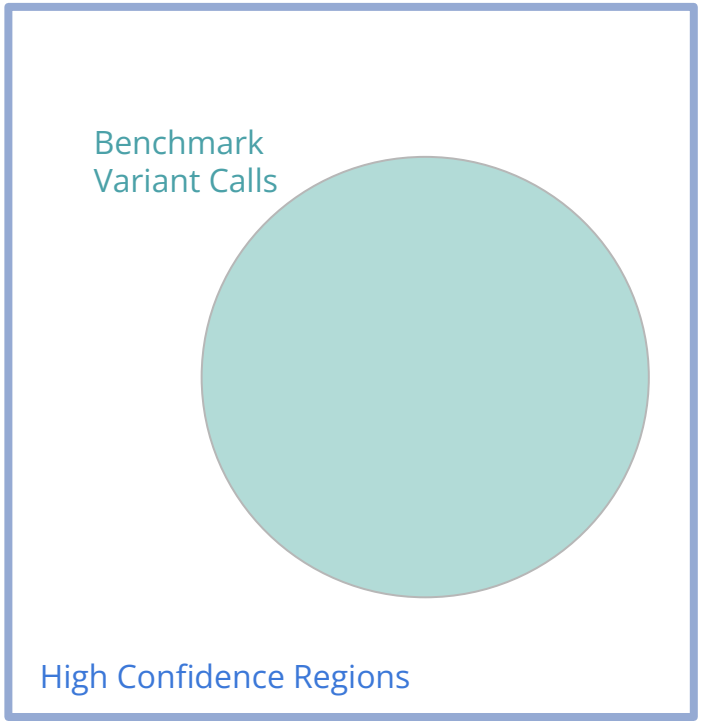
# HG002 Tier 1 Benchmark

- **Tier 1** refers to the highest confidence benchmark regions for structural variant (SV) calls within a genomic dataset (Zook et al., 2020).
- **HG002 Tier 1** refers to the high-confidence structural variant (SV) benchmark set for the HG002 genome.
- It  spans 2.51 Gbp and includes 5,262 insertions and 4,095 deletions.
- We are using the version 0.6

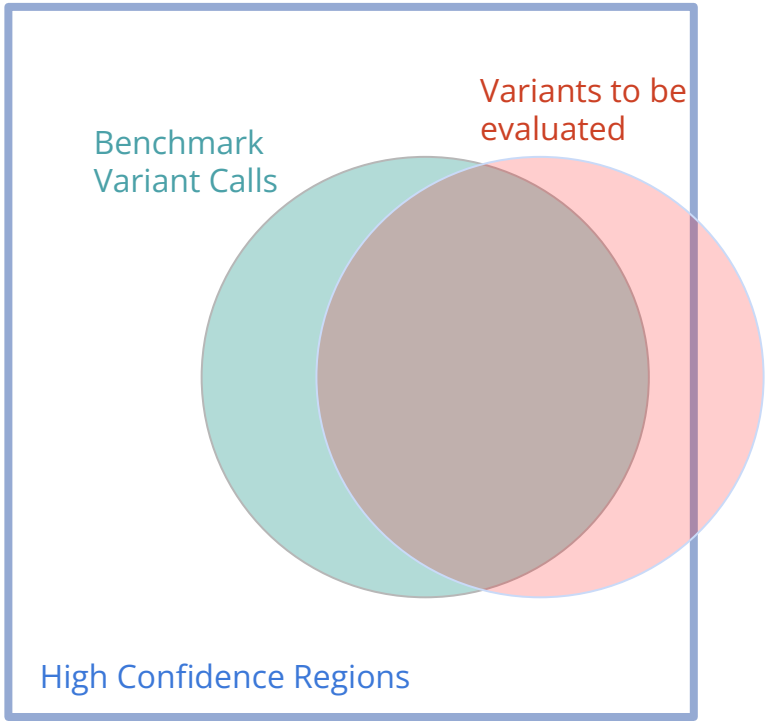
# CMRG (Curated Medically Relevant Genes) Benchmark

- **CMRG benchmark** is a standard set of genetic variants for challenging, medically relevant genes (Wagner et al., 2022).
- It characterizes 273 of the 395 challenging medically relevant genes (repetitive and complex).
- It reports over 17,000 SNVs, 3,600 indels and 200 other SVs for human genome **reference GRCh37 across HG002**.
- We are using the version 1.0

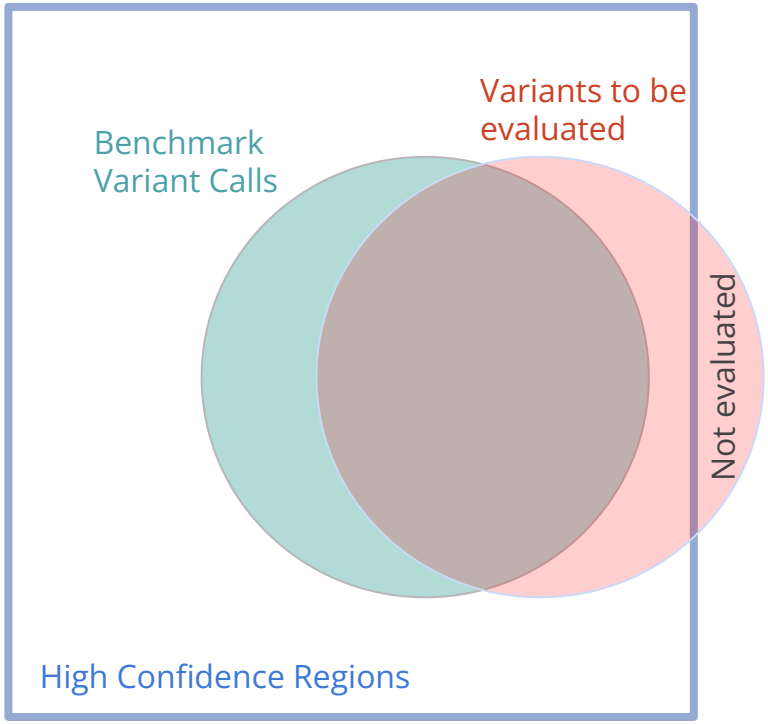
# Benchmarking Process



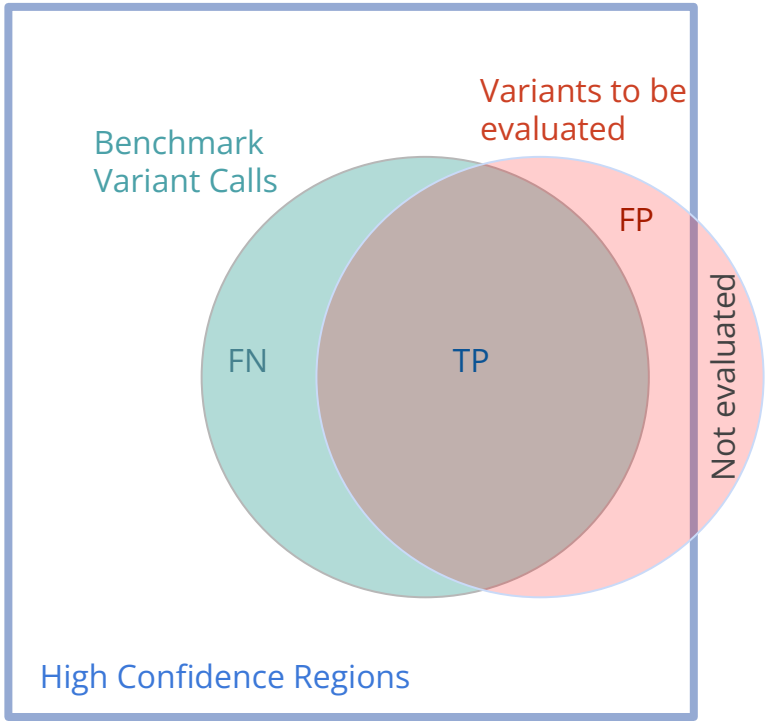
# Benchmarking Process



# Benchmarking Process

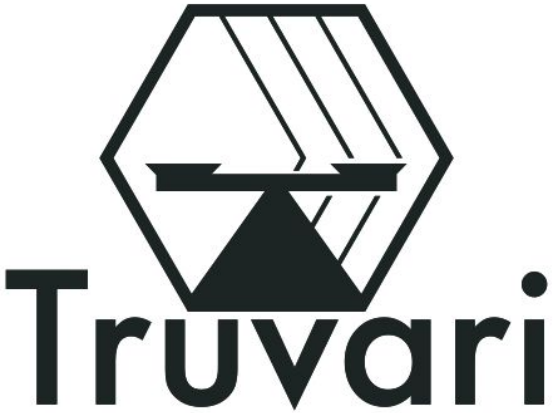


# Benchmarking Process



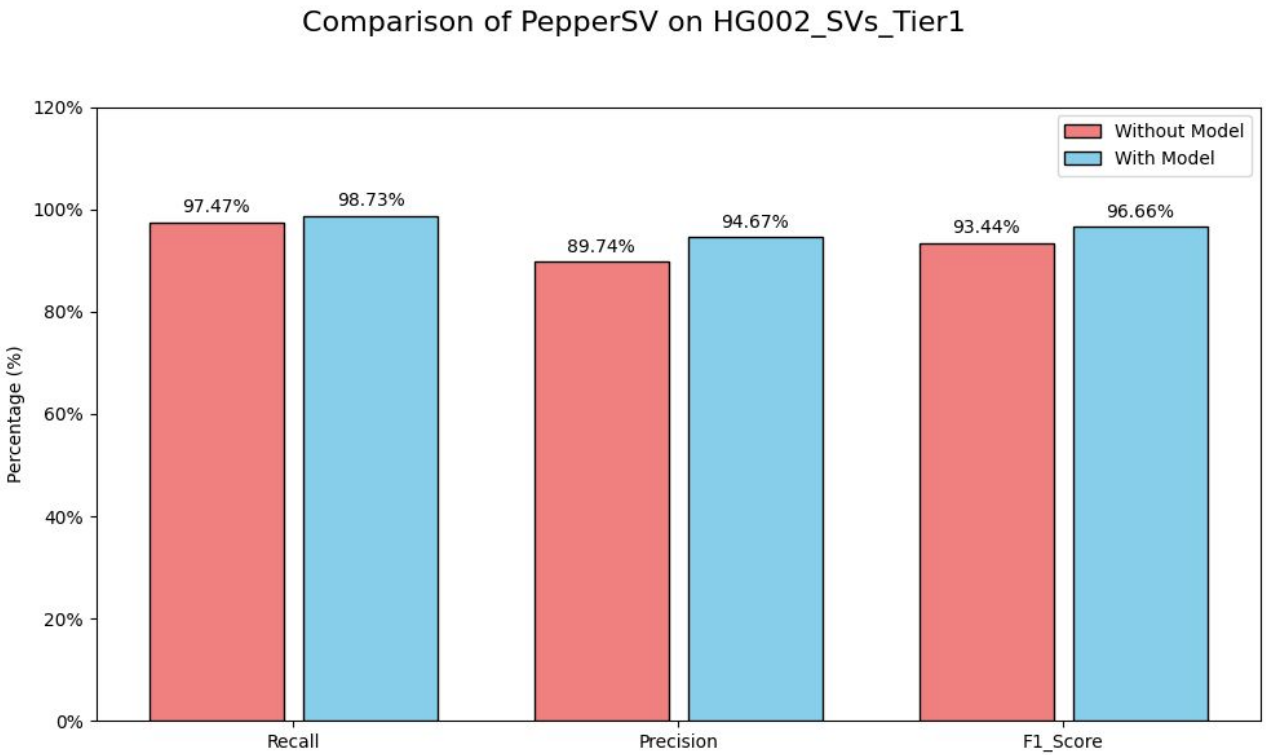
# Truvari

- Toolkit for Benchmarking SVs in Variant Call Format (VCF) files (English et al., 2022).
- It compares two VCF files to provide performance metrics:
  - True Positive, False Positive and False negative count
  - Precision
  - Recall
  - F1 Score
- We are using the version 4.2.0

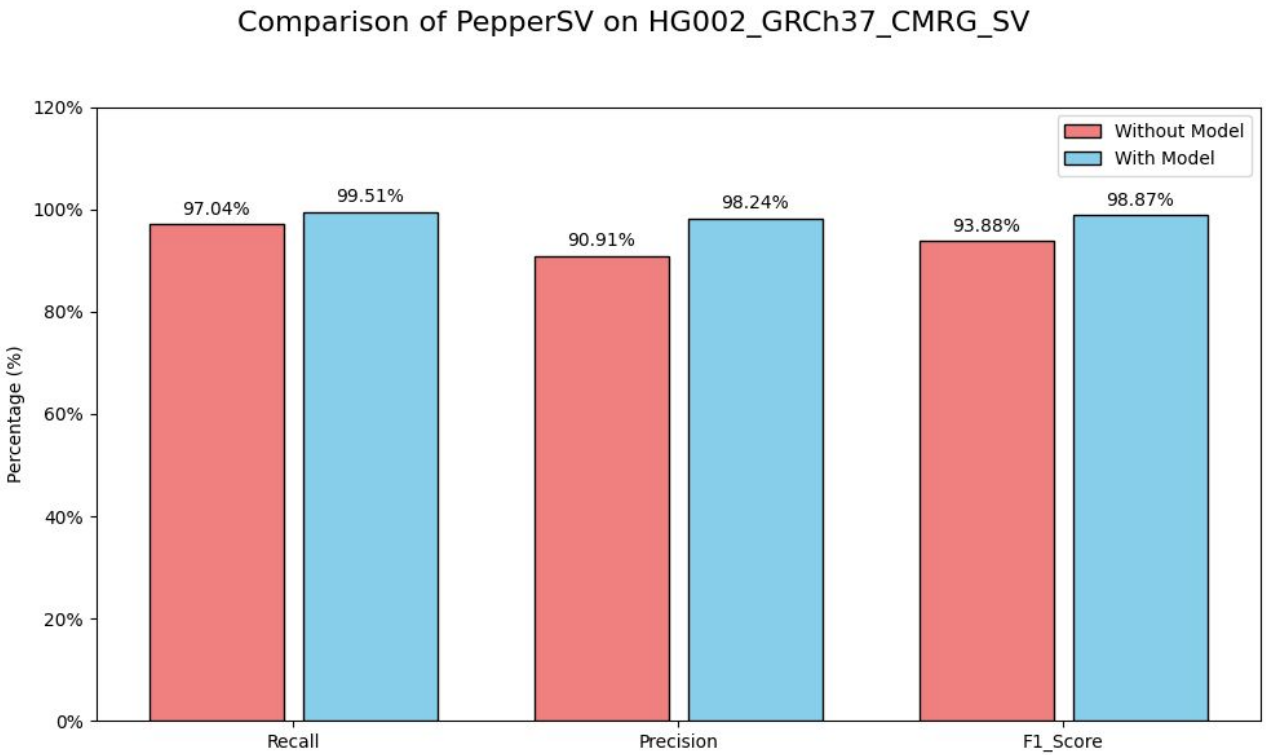


# Results

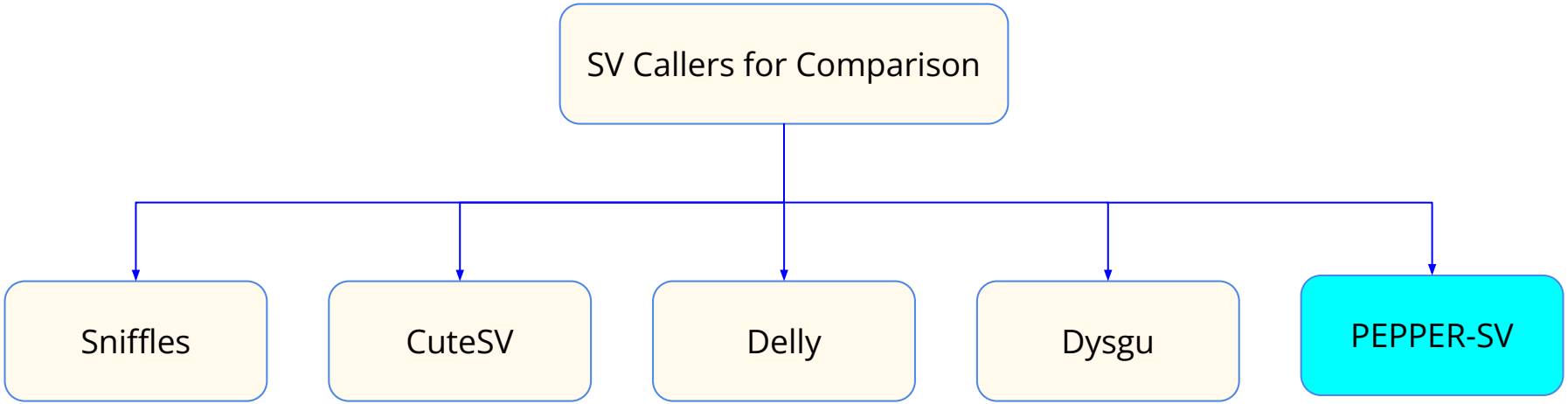
# Impact of the Model on SV Detection



# Impact of the Model on SV Detection

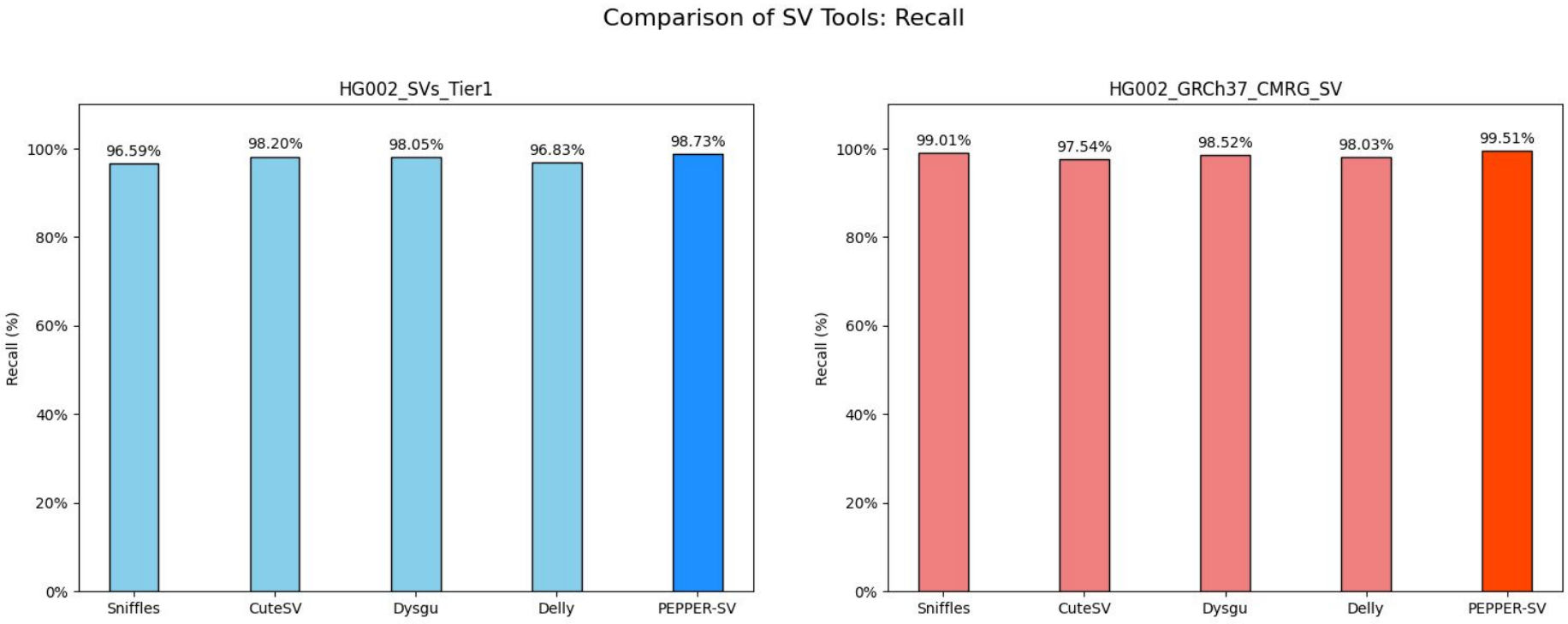


# SV Callers Used for Comparison

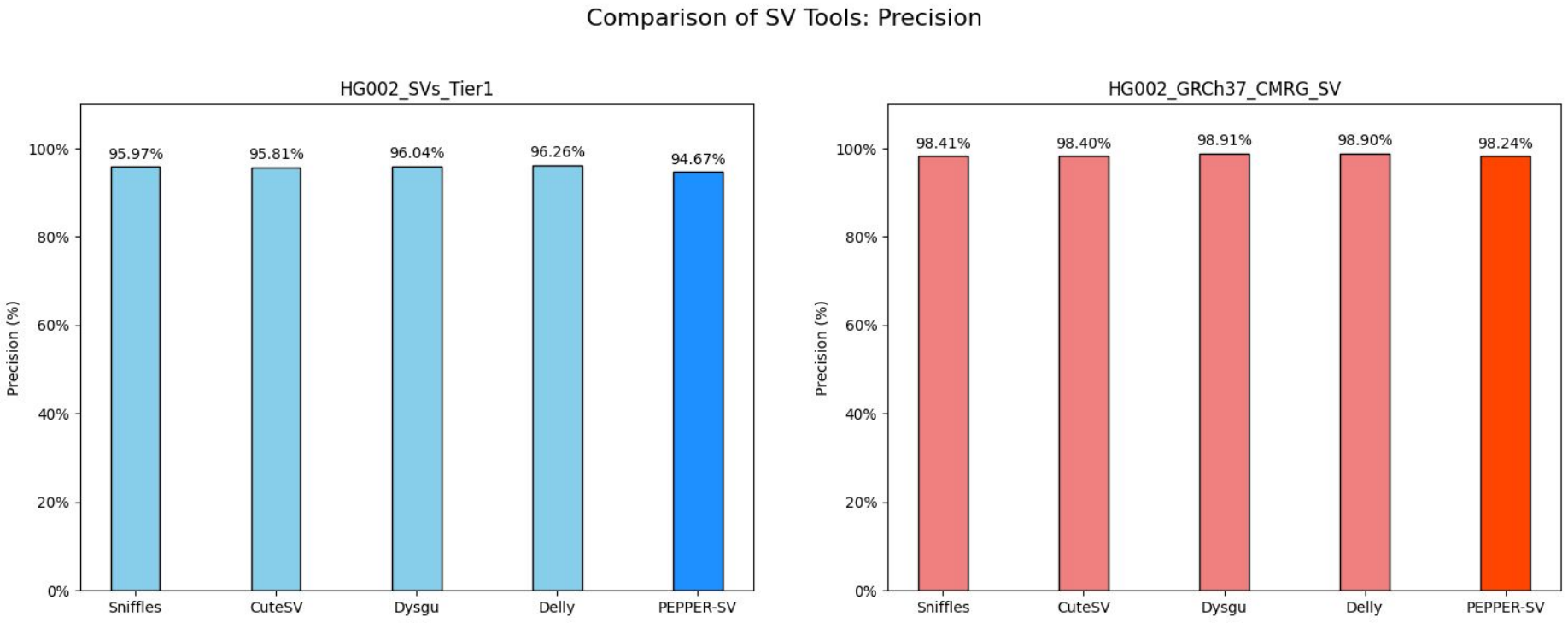



Benchmarking is done on Chromosomes 15-22.

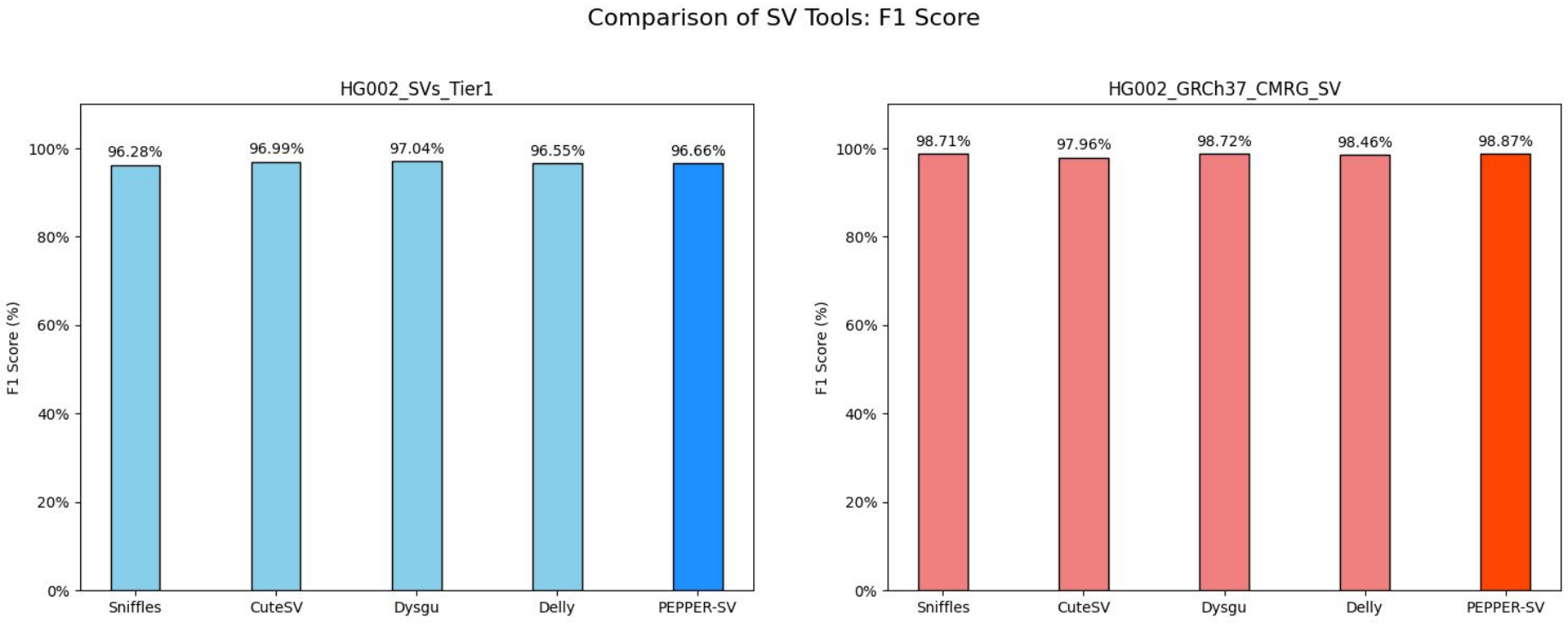
# Comparison of SV Tools



# Comparison of SV Tools



# Comparison of SV Tools



# Discussion and Future Work

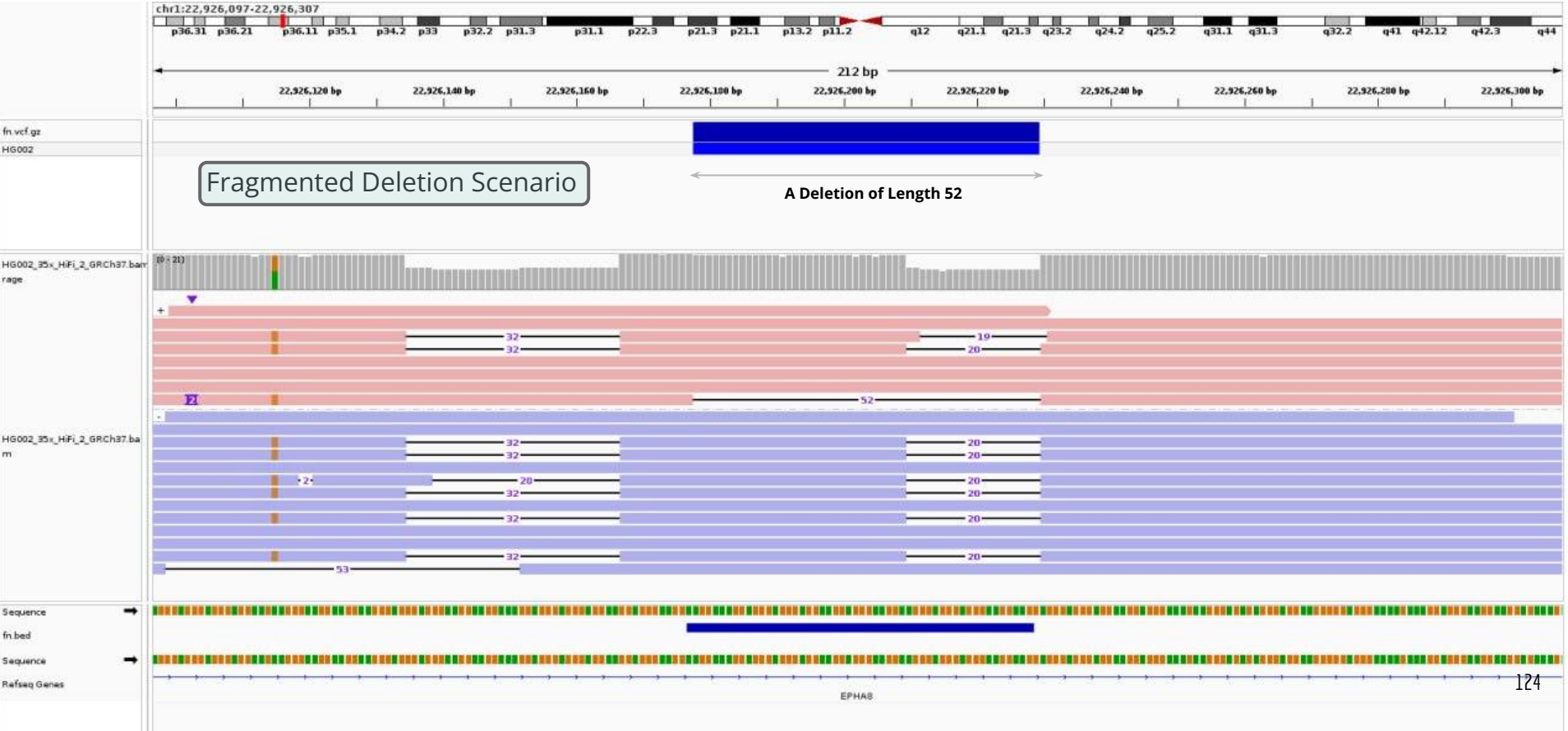
# Discussion

- Our approach is the first-of-its-kind Deep Learning-based SV caller.
- We consolidate variants based on normalized indel similarity.
  - Keep dissimilar variants separate, merge similar variants together.
- We detect additional breakpoints from soft-clipped reads.

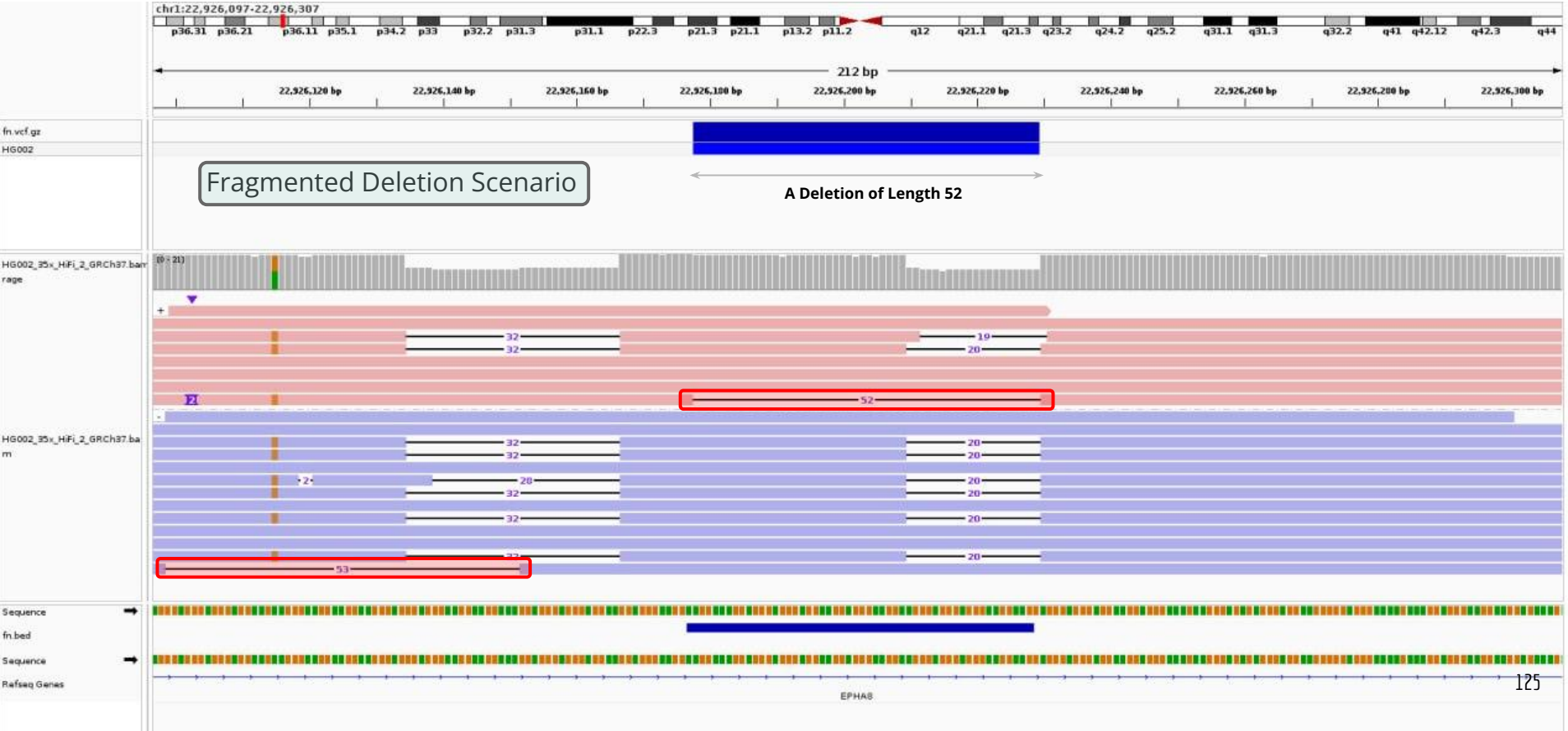
# Discussion

- Our method works very well in challenging and complex regions as well (CMRG Benchmark).

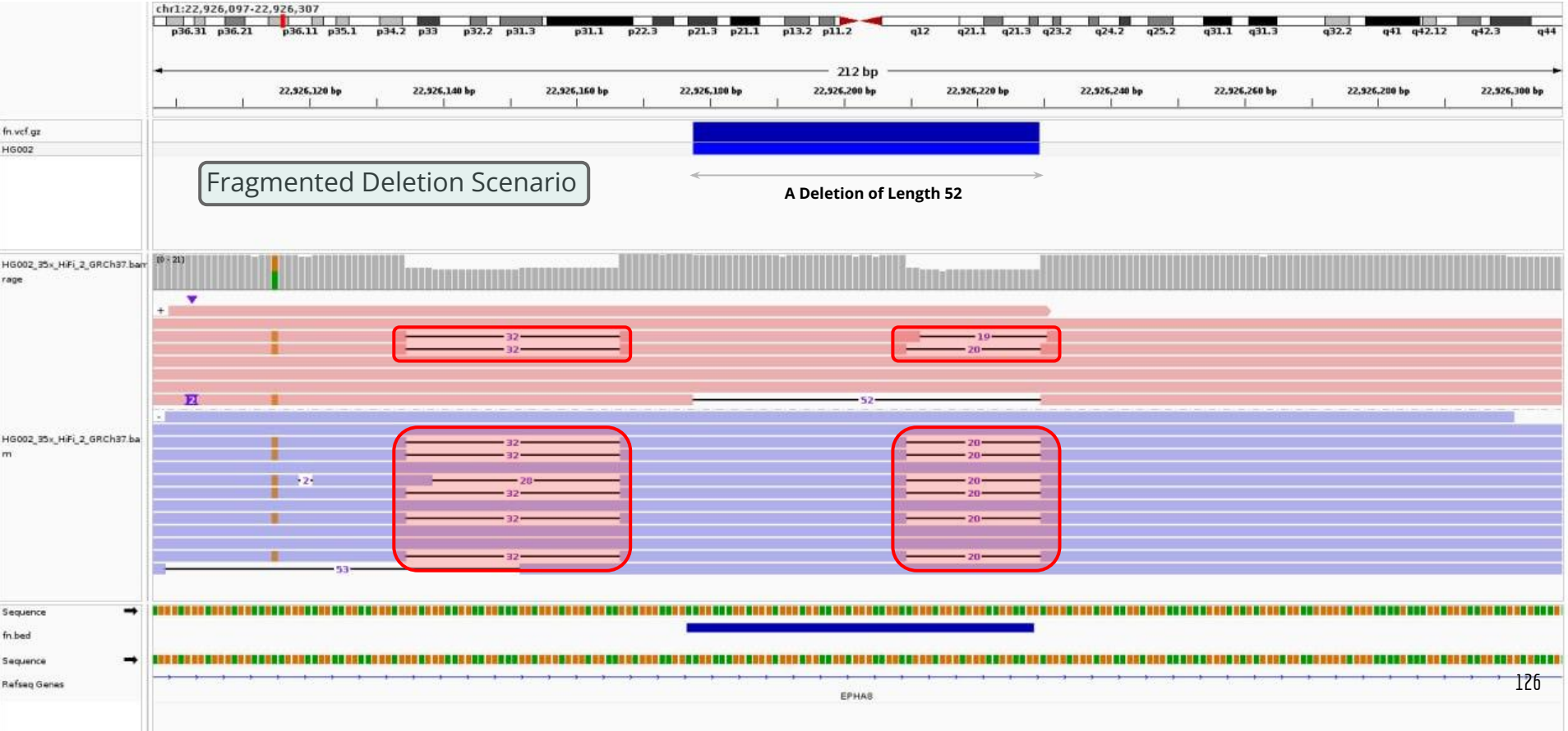
# Limitations and Challenges



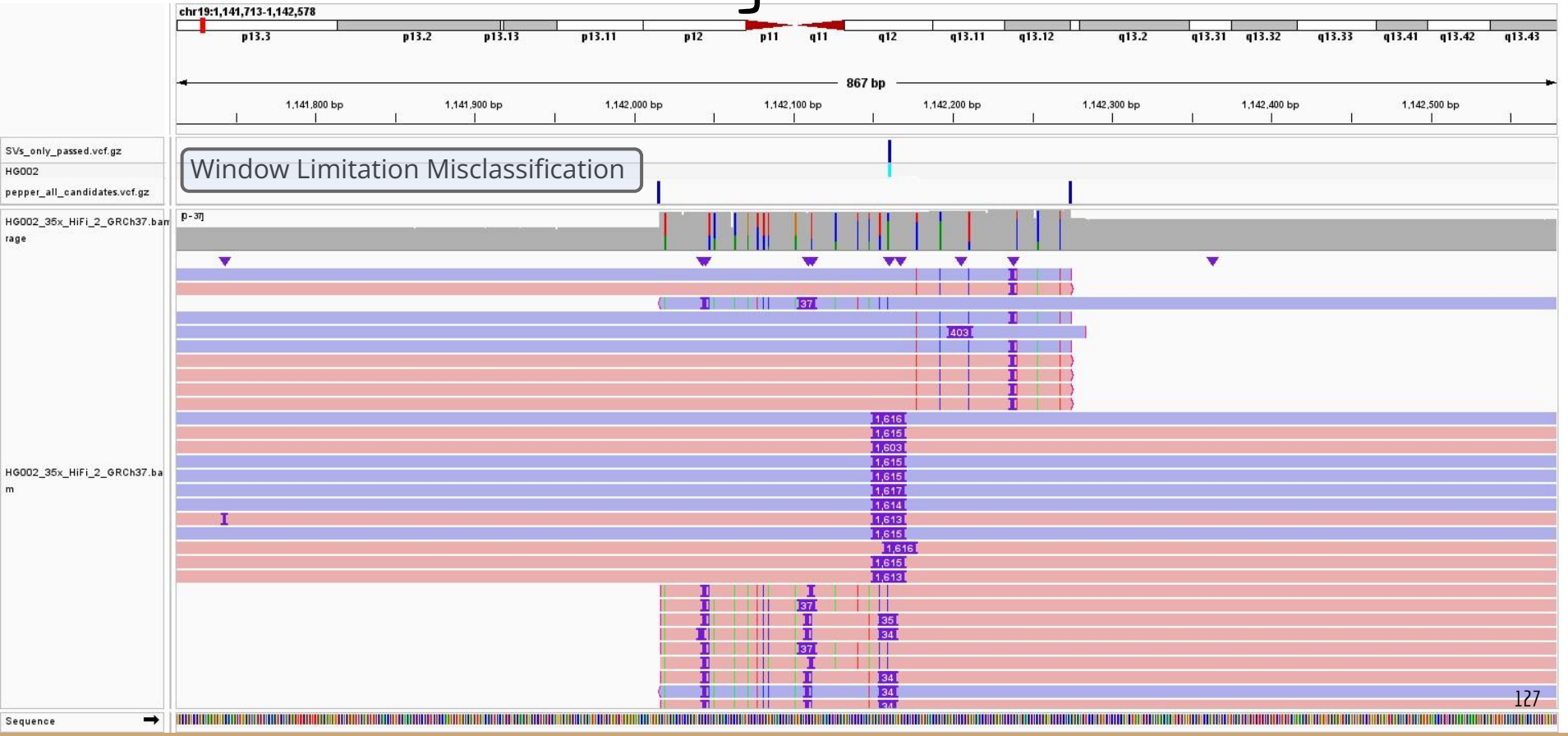
# Limitations and Challenges



# Limitations and Challenges



# Limitations and Challenges



# Limitations and Challenges

# Limitations and Challenges

# Future Work

- Fixing the limitations and challenges.
- Training a CNN based model.
- Training the model with different types of read sequences.

# References


1. Al Khleifat, Ahmad, et al. "Structural variation analysis of 6,500 whole genome sequences in amyotrophic lateral sclerosis." NPJ genomic medicine 7.1 (2022): 8.
2. Stankiewicz, Paweł, and James R. Lupski. "Structural variation in the human genome and its role in disease." Annual review of medicine 61 (2010): 437-455.
3. Zook, Justin M., et al. "A robust benchmark for detection of germline large deletions and insertions." Nature biotechnology 38.11 (2020): 1347-1355.
4. Wagner, Justin, et al. "Curated variation benchmarks for challenging medically relevant autosomal genes." Nature biotechnology 40.5 (2022): 672-680.
5. English, Adam C., et al. "Truvari: refined structural variant comparison preserves allelic diversity." Genome Biology 23.1 (2022): 271.