Project 06: Structual Variant Calling Pipeline Report

Author: Tianze (Vincent) Luo
Date : 2025-04-21

Table of Contents

0. Background

- Interactive session command: `crc-interactive --teach -a hugen2072-2025s -t 4:00:00`

- Start `.cram` file: `p6/NA12778.final.cram`

  - Alternative format for `.bam`

  - **NA12778**: Female resident of Utah with Northern and Western European–associated ancestry [1000 Genome Project]

- Reference sequence: `p6/GRCh38_full_analysis_set_plus_decoy_hla.fa`

Part I. SV Calling

1. Extract chr22 alignments (`.BAM`) && index

Will call SVs on chromosome 22 only

2. Manta caller (used single tool in this pipeline)

manta SV discovery caller based on **discordant pairs (PE)** and **split reads (SR)**.

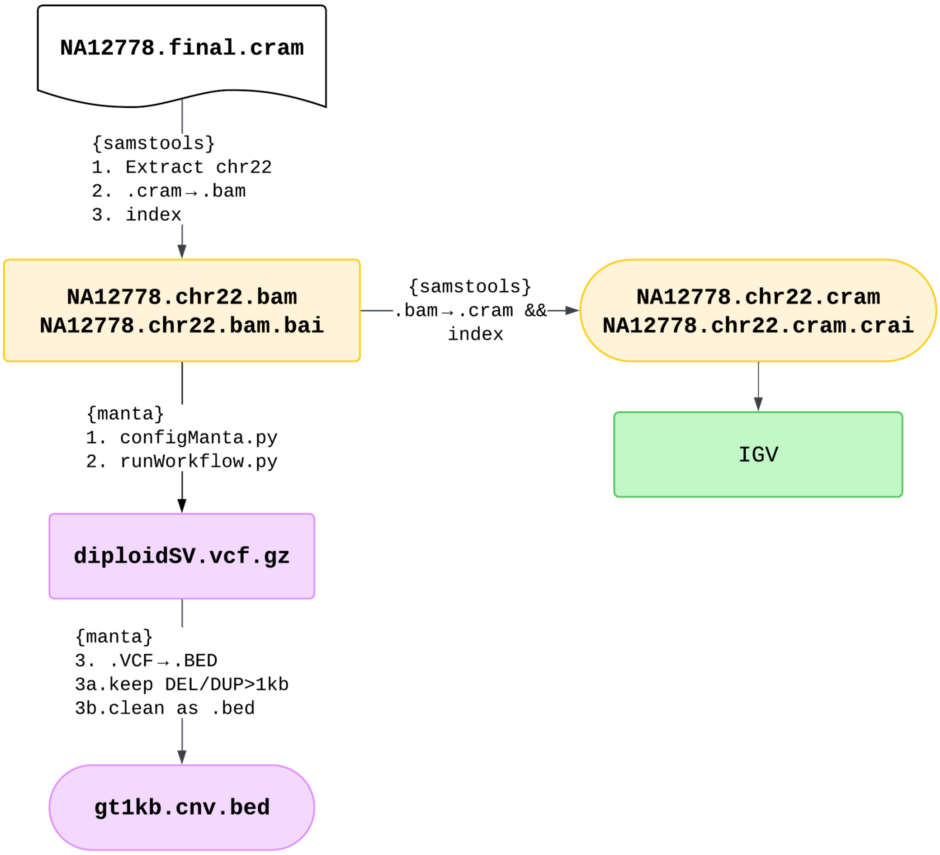
`manta` workflow:

- `configManta.py`
- `manta_test/runWorkflow.py`

- variant `.VCF` -> `.BED`
  - Extract DEL/DUP > 1 kb (`.bed`)

3. Convert chr22 alignments `.bam` to indexed `.cram`

4. Part I Summary



Summary of the SVs called

We have called **17 SVs in chr22**, more specifically DEL/DUP (CNVs) >1kb.

- 16 of them are DELs and 1 of them is DUP:TANDEM

The table below is generated from results of `luo_script_SV-summary.R`

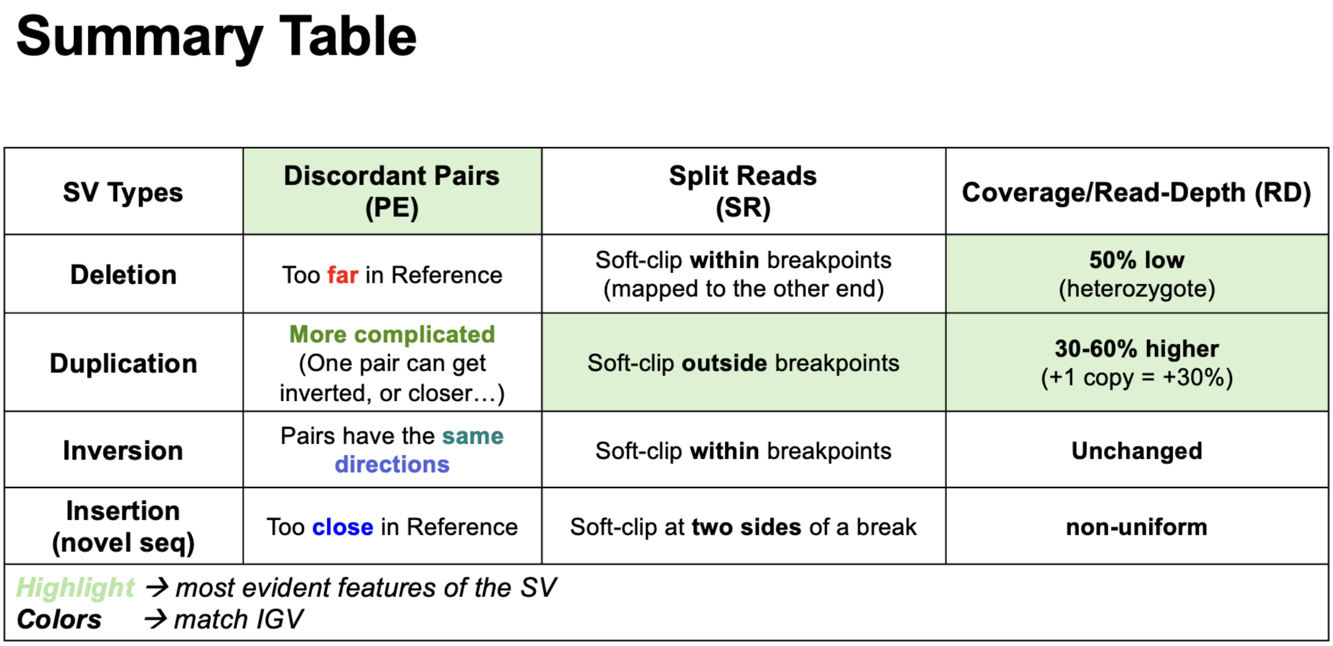
sv_type	count	avg_length
DEL	16	18638
DUP:TANDEM	1	58303
ALL	17	20971

```
wc -l gt1kb.cnv.bed
# 17 gt1kb.cnv.bed
```

```
cut -f4 gt1kb.cnv.bed | sort | uniq -c
#      16 <DEL>
#       1 <DUP:TANDEM>

## Summary
module load gcc/12.2.0 r/4.4.0
Rscript --vanilla luo_script_SV-summary.R
```

Part II. Investigate with IGV

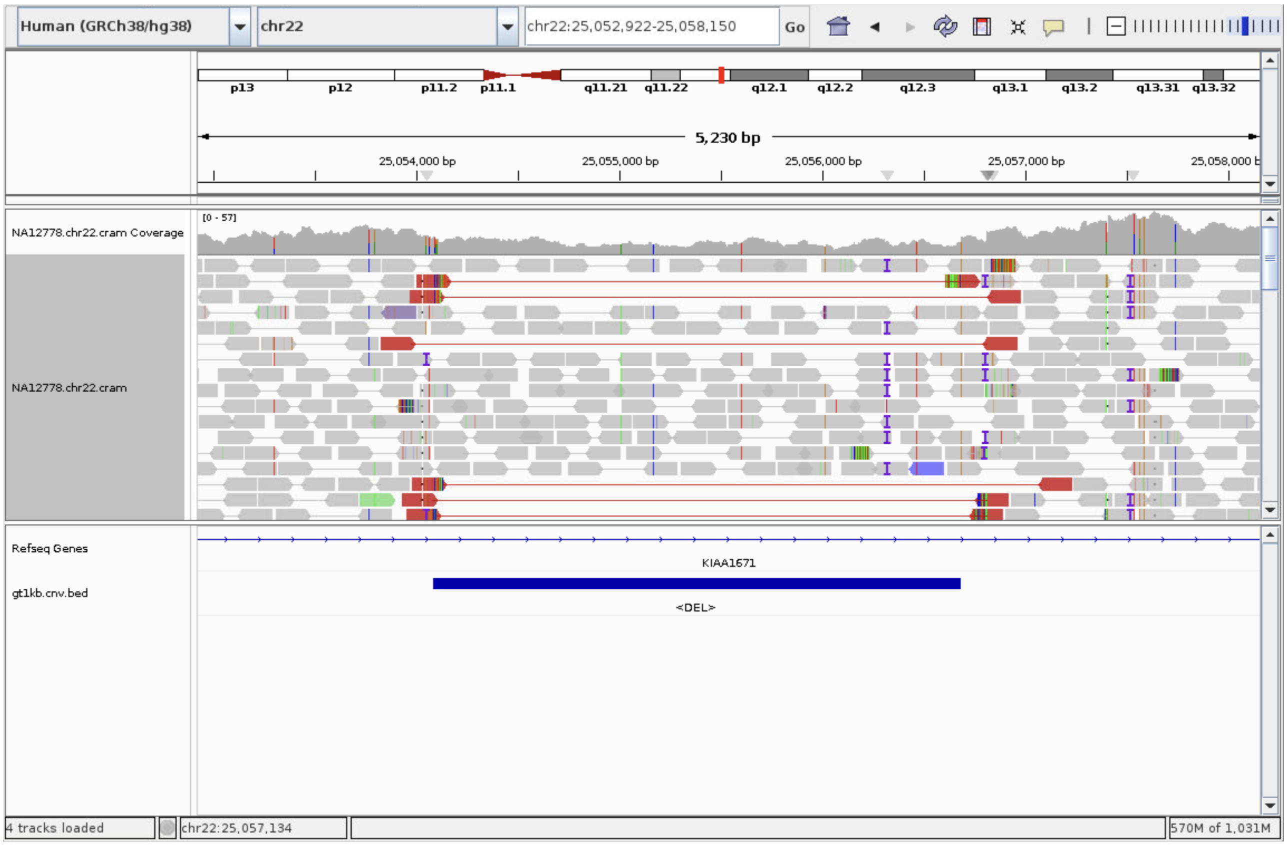


Summary Table

SV Types	Discordant Pairs (PE)	Split Reads (SR)	Coverage/Read-Depth (RD)
Deletion	Too **far** in Reference	Soft-clip **within** breakpoints (mapped to the other end)	**50% low** (heterozygote)
Duplication	**More complicated** (One pair can get inverted, or closer…)	Soft-clip **outside** breakpoints	**30-60% higher** (+1 copy = +30%)
Inversion	Pairs have the **same directions**	Soft-clip **within** breakpoints	**Unchanged**
Insertion (novel seq)	Too **close** in Reference	Soft-clip at **two sides** of a break	**non-uniform**

Highlight* → *most evident features of the SV
Colors* → *match IGV

1. one SV on *q arm*: `chr22 25054080 25056685 <DEL>`

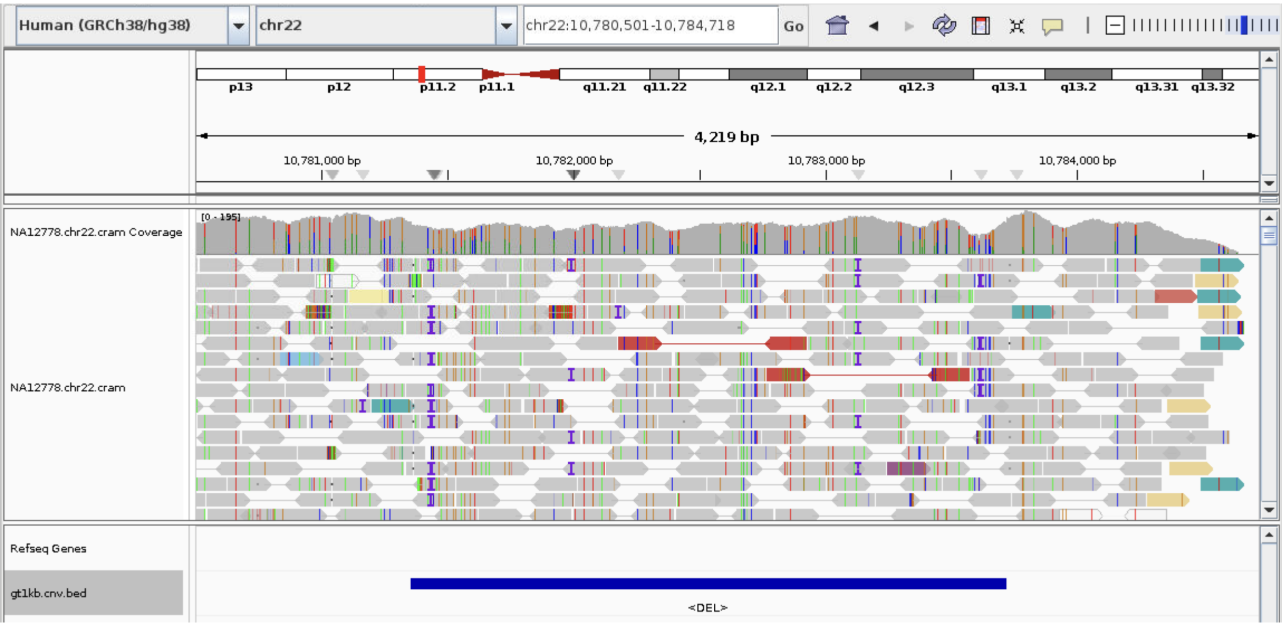


  - intron of KIAA1671

  - As shown in the summary table above (summarized from Dr. Brand's slide), the evidences to call this SV as a **DEL** include:

    1. Red paired reads: TLEN > insert size.
    2. Soft-clips within breakpoints: meaning that those bases mapped to the other end/pair.
    3. The coverage within breakpoints is ~50% lower.

2. one SV on *p arm*: `chr22 10781349 10783722 <DEL>`



- intergenic region
- **Seems more complicated than simply DEL**. The only evidences to call it as a DEL (by `manta`) is the 2x red paired reads (meaning TLEN > insert size). I am less confident in this call because:

    1. The signals are noisy at this region (a lot of softclips)
    2. Minial decrease in coverage --> not likely DEL