Characterizing cancer genome structural variants from short-read and long-read sequencing.

Katherine Feldmann
Gavin Ha Lab
Winter Rotation 2023

Prostate cancer is an important medical concern and understanding malignancy is essential.

Second most common cancer in the United States with an estimated 268,490 new cases in 2022.

NIH National Cancer Institute

Prostate cancer is an important medical concern and understanding malignancy is essential.

Second most common cancer in the United States with an estimated 268,490 new cases in 2022.

Fifth leading cause of cancer death with an estimated 34,500 deaths.

NIH National Cancer Institute

Prostate cancer is an important medical concern and understanding malignancy is essential.

Second most common cancer in the United States with an estimated 268,490 new cases in 2022.

Fifth leading cause of cancer death with an estimated 34,500 deaths.

No curative therapies for patients with metastatic castration-resistant prostate cancer.

Prostate cancer is an important medical concern and understanding malignancy is essential.

Second most common cancer in the United States with an estimated 268,490 new cases in 2022.

Fifth leading cause of cancer death with an estimated 34,500 deaths.

No curative therapies for patients with metastatic castration-resistant prostate cancer.

A goal of the lab is to study the role of genomic structural alterations in prostate cancer.

Understanding genomic structural variants is important for treating and diagnosing cancer.

Structural variants are larger than 50 base pairs and there are five basic types:

Deletion Insertion Duplication Inversion Translocation

Figure adapted from PacBio

Understanding genomic structural variants is important for treating and diagnosing cancer.

Structural variants are larger than 50 base pairs and there are five basic types:

Deletion Insertion Duplication Inversion Translocation

Chromothripsis results in complex structural variants, containing multiple types of basic variants.

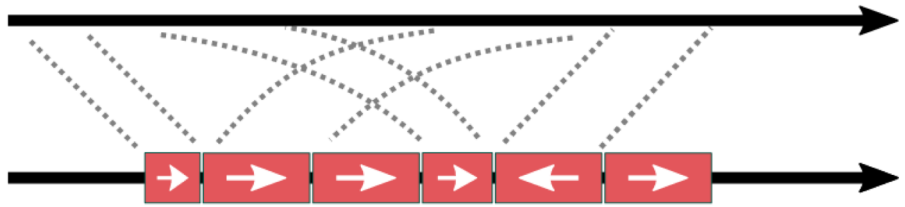


Figure from IAEM van Belzen et al. 2021

Understanding genomic structural variants is important for treating and diagnosing cancer.

Structural variants are larger than 50 base pairs and there are five basic types:

Deletion Insertion Duplication Inversion Translocation

Chromothripsis results in complex structural variants, containing multiple types of basic variants.

Greater than 30% of cancer genomes have a pathogenic structural variant.

Understanding genomic structural variants is important for treating and diagnosing cancer.

Structural variants are larger than 50 base pairs and there are five basic types:

Deletion Insertion Duplication Inversion Translocation

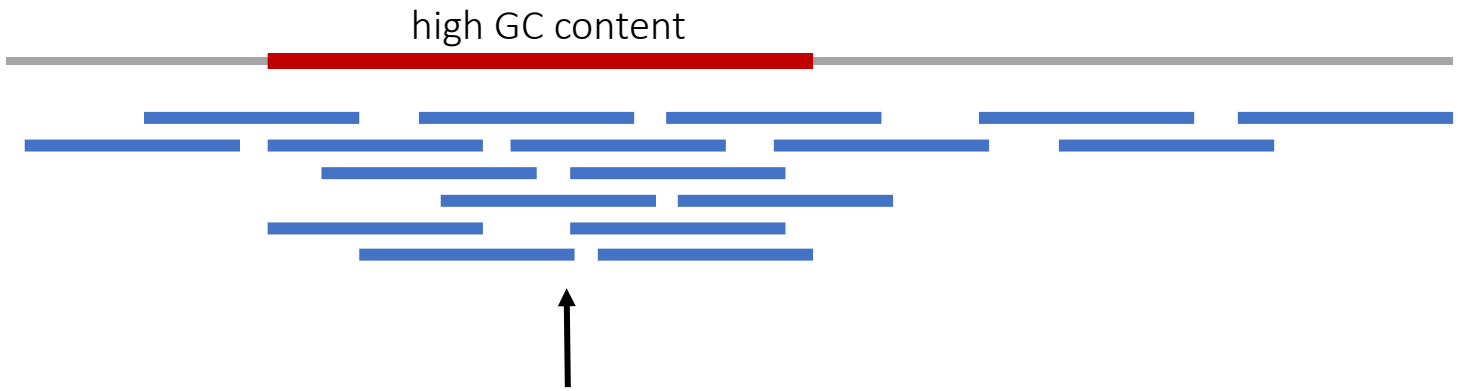
Chromothripsis results in complex structural variants, containing multiple types of basic variants.

Greater than 30% of cancer genomes have a pathogenic structural variant.

Majority of cancer genomes available for identifying structural variants are paired-end, short-read.

What are limitations with paired-end, short-read sequencing for detecting structural variants?

Illumina – the most popular sequencing platform – exhibits biases in high or low GC content regions.

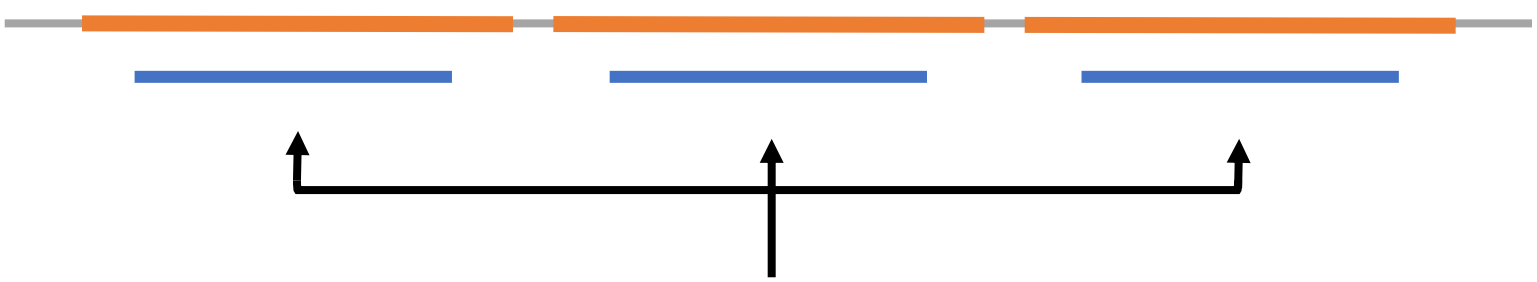


read coverage varies by GC content

What are limitations with paired-end, short-read sequencing for detecting structural variants?

Illumina – the most popular sequencing platform – exhibits biases in high or low GC content regions.

Reads may be too short to uniquely map to highly repetitive regions in the reference genome.



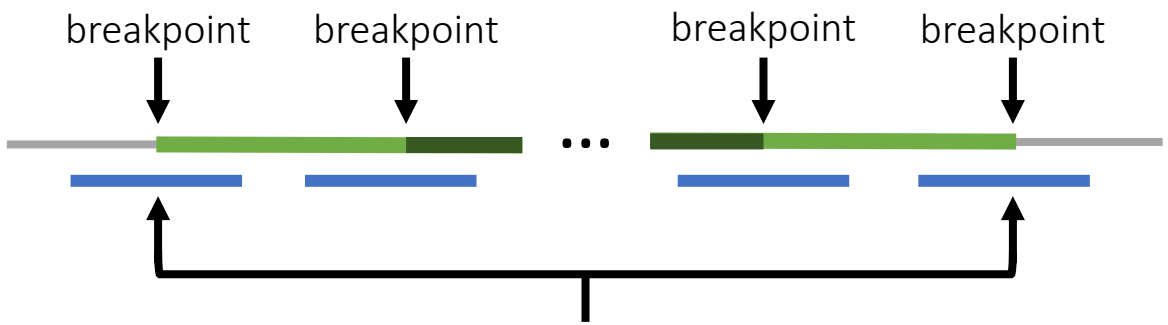
uncertainty where to align read

What are limitations with paired-end, short-read sequencing for detecting structural variants?

Illumina – the most popular sequencing platform – exhibits biases in high or low GC content regions.

Reads may be too short to uniquely map to highly repetitive regions in the reference genome.

Short-reads may be unable to determine if structural rearrangements are on the same chromosome.



*short-reads can only span **one** breakpoint*

What are limitations with paired-end, short-read sequencing for detecting structural variants?

Illumina – the most popular sequencing platform – exhibits biases in high or low GC content regions.

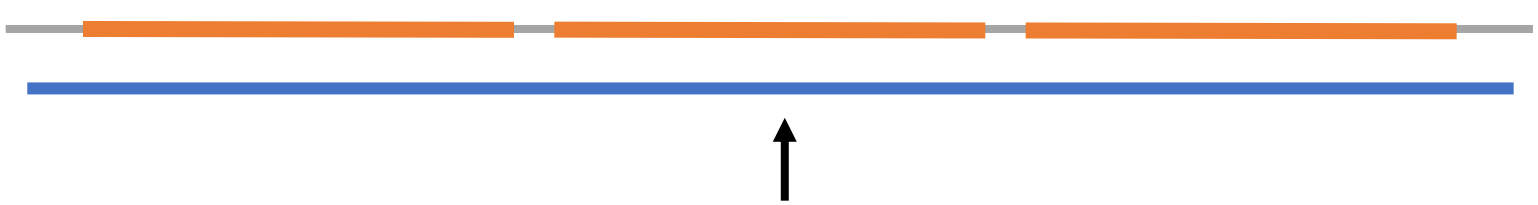
Reads may be too short to uniquely map to highly repetitive regions in the reference genome.

Short-reads may be unable to determine if structural rearrangements are on the same chromosome.

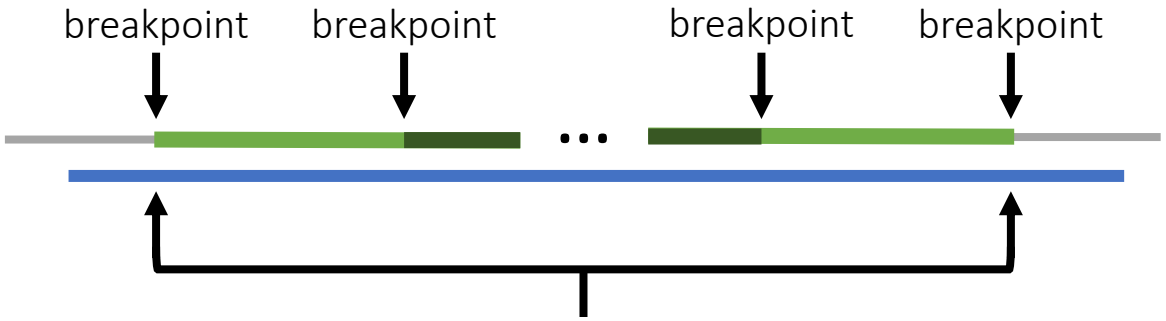
In the GRCh38 reference genome, roughly 55Mb are inaccessible to Illumina short-read sequencing due to repetitive regions and GC bias.

Long-read sequencing improves structural variant detection by reducing alignment ambiguity.

Long-reads resolve ambiguous regions by aligning to larger regions of the reference genome.



long-reads can span repeats



*long-reads can span **multiple** breakpoints*

Long-read sequencing improves structural variant detection by reducing alignment ambiguity.

Long-reads resolve ambiguous regions by aligning to larger regions of the reference genome.

Long-reads have enabled detection of previously unknown genomic rearrangements.

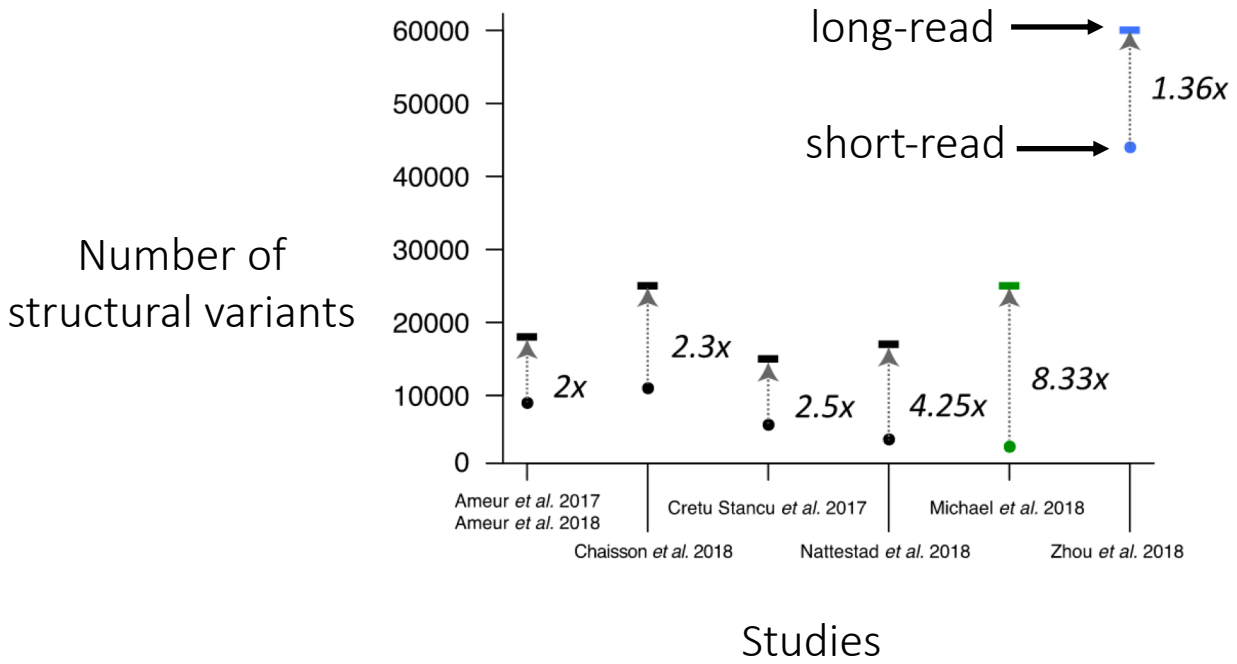


Figure from Mahmoud et al. 2019

Long-read sequencing improves structural variant detection by reducing alignment ambiguity.

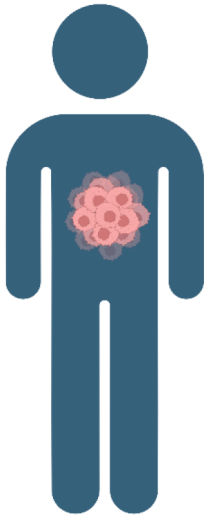
Long-reads resolve ambiguous regions by aligning to larger regions of the reference genome.

Long-reads have enabled detection of previously unknown genomic rearrangements.

Sequencing cost, nucleotide accuracy and sample requirements have limited the use of long-reads.

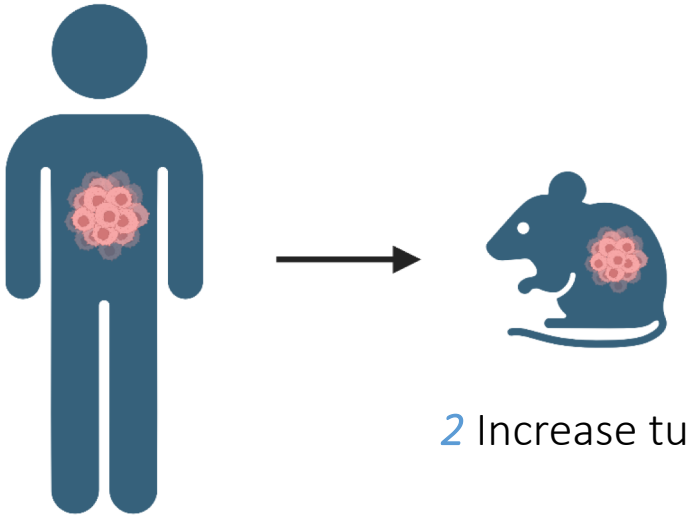
The goal of this rotation project is to compare genomic rearrangements between short-read and long-read sequencing data.

Tumor tissue from *two* prostate cancer PDX models was short-read and long-read sequenced.



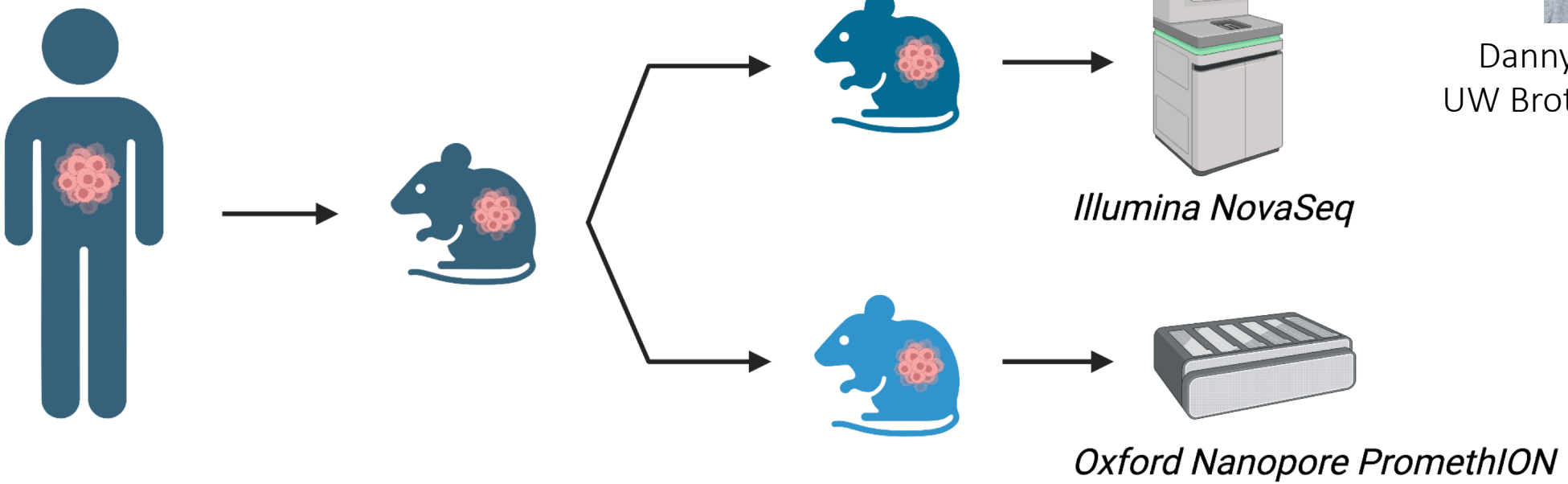
1 Tumor tissue from patient

Tumor tissue from *two* prostate cancer PDX models was short-read and long-read sequenced.



2 Increase tumor material through passaging

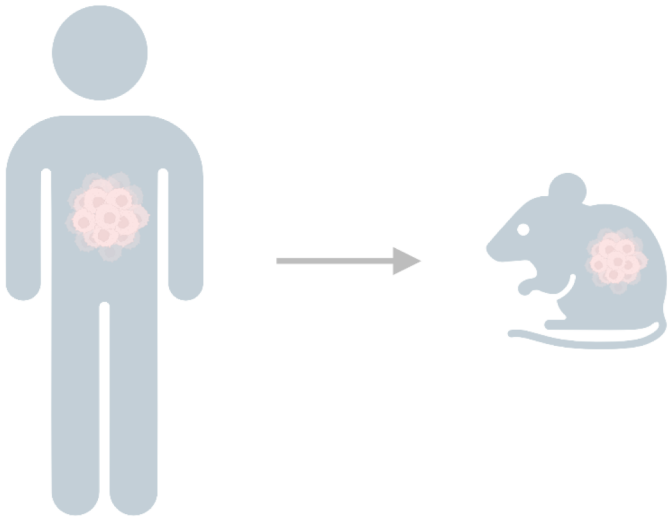
Tumor tissue from *two* prostate cancer PDX models was short-read and long-read sequenced.



Danny Miller, MD, PhD
UW Brotman-Baty Institute

3 Sequence tumor tissue

Genomic data came from different samples due to different sample requirements.

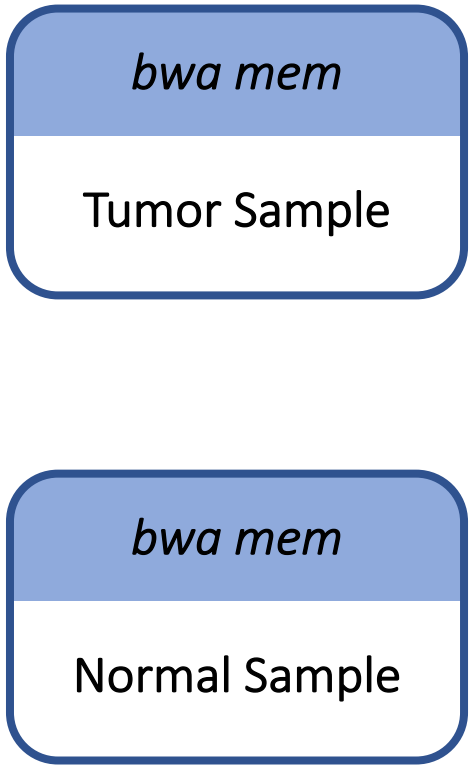


Illumina NovaSeq

Oxford Nanopore PromethION

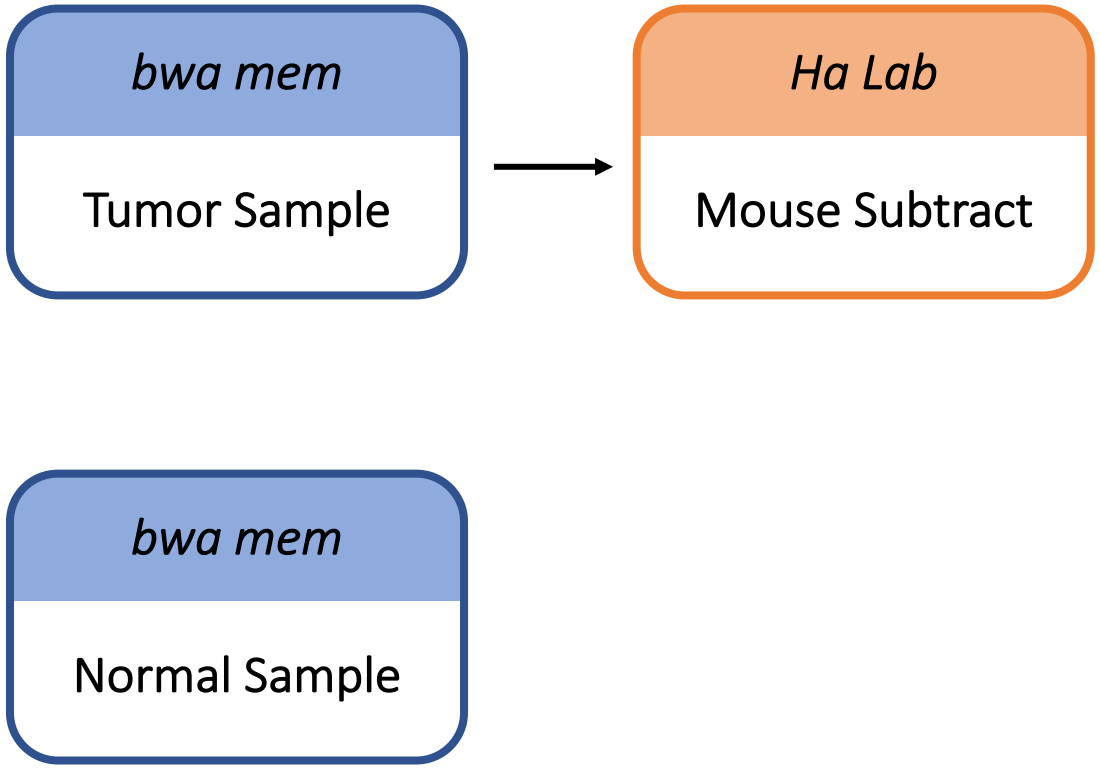
Danny Miller, MD, PhD
UW Brotman-Baty Institute

Bioinformatics pipeline for identifying structural variants using short-read sequencing data.



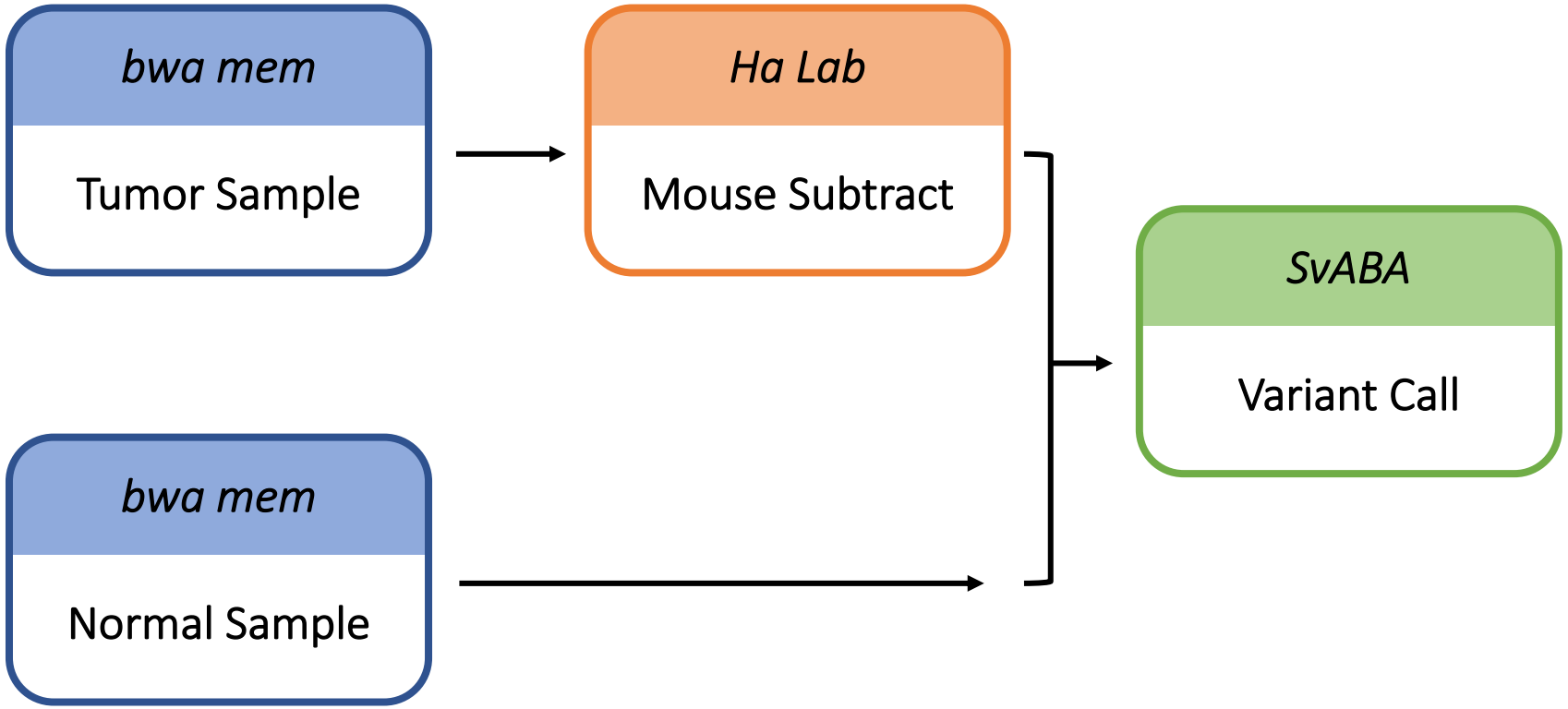
Align reads to GRCh38 reference genome.

Bioinformatics pipeline for identifying structural variants using short-read sequencing data.



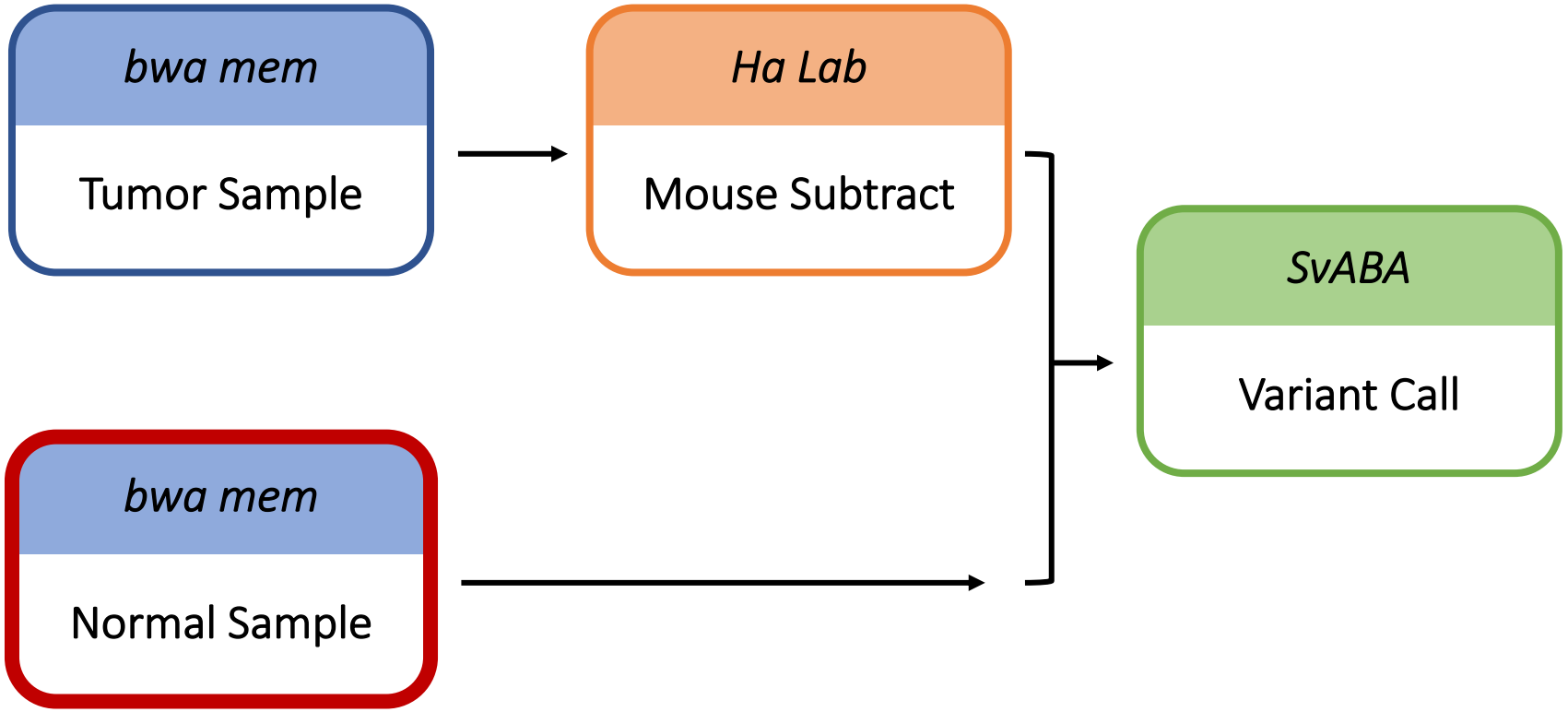
Remove mouse reads by aligning to a concatenated human and mouse genome.

Bioinformatics pipeline for identifying structural variants using short-read sequencing data.

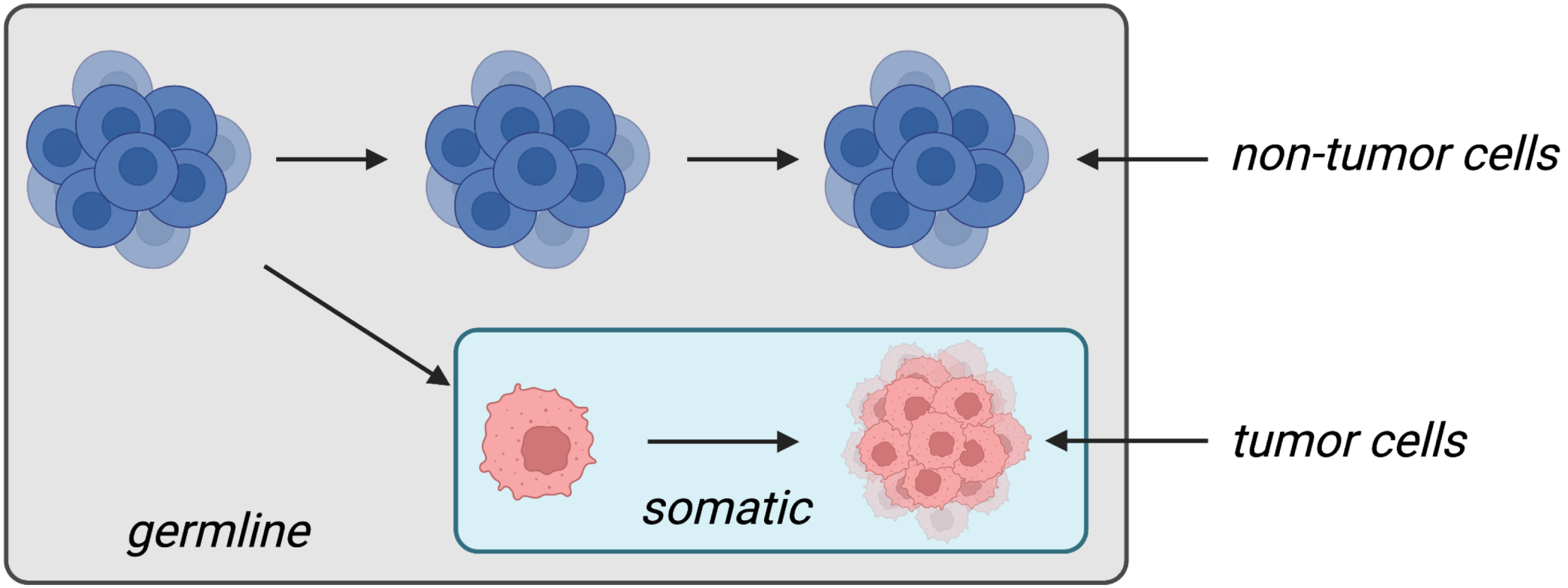


Call structural variants using an algorithm specific to short-read sequencing.

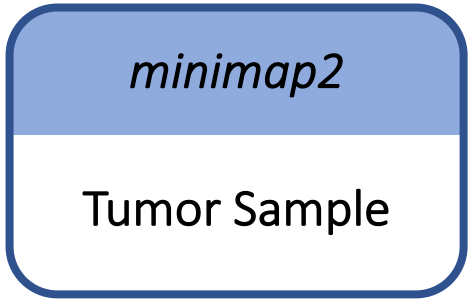
Short-read tumor sample was compared to a normal sample to remove germline variants.



Short-read tumor sample was compared to a normal sample to remove germline variants.

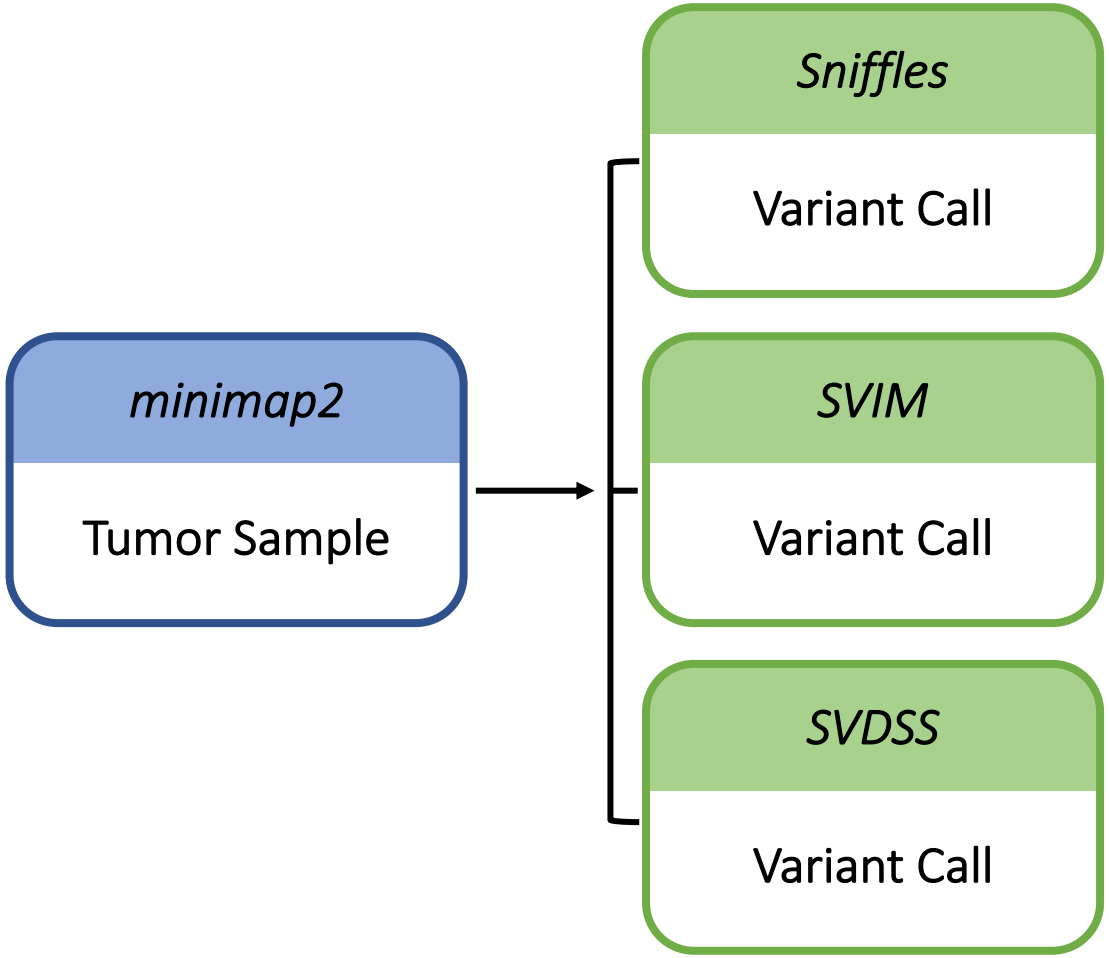


Bioinformatics pipeline for identifying structural variants using long-read sequencing data.



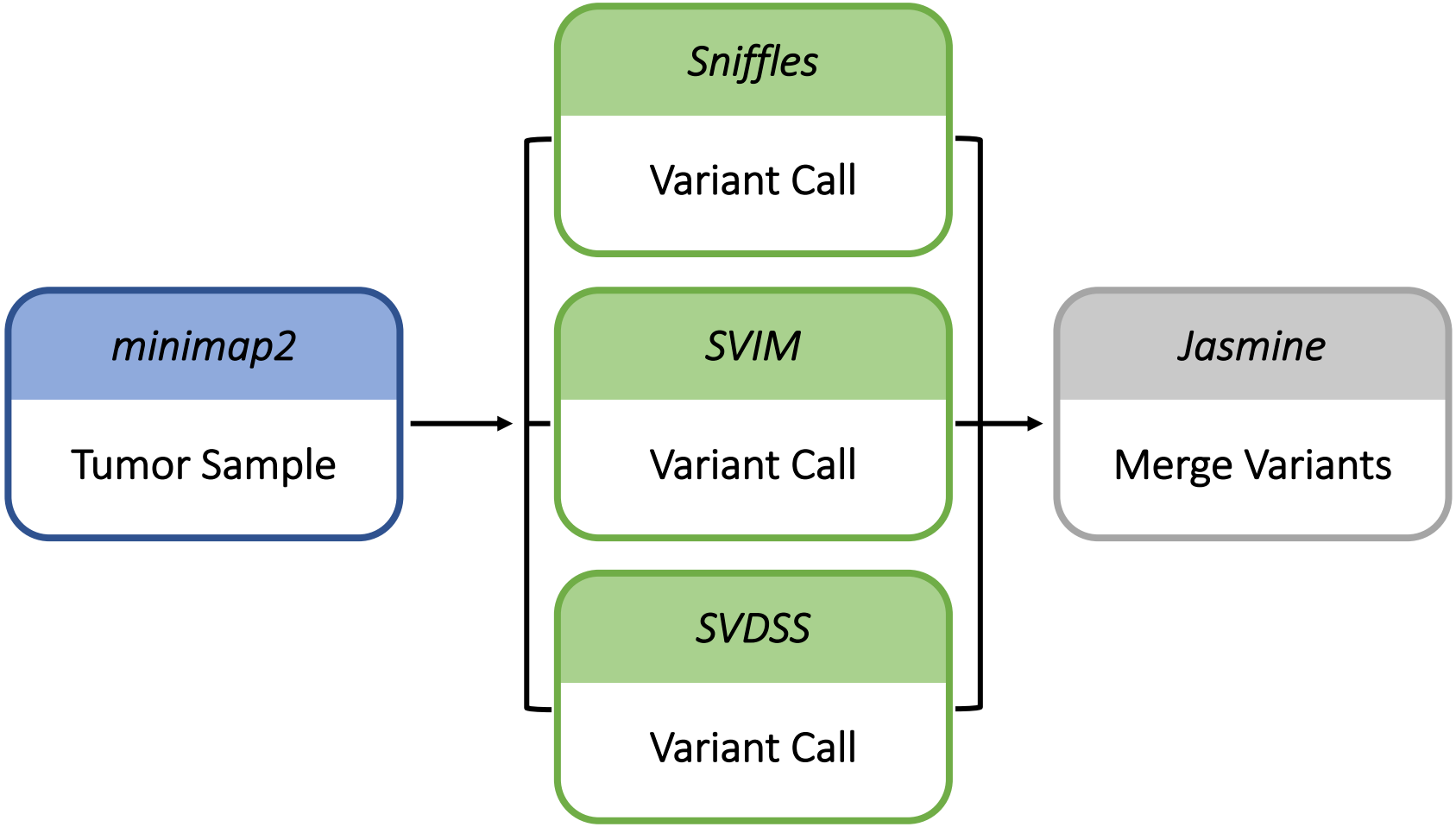
Align reads to GRCh38 reference genome.

Bioinformatics pipeline for identifying structural variants using long-read sequencing data.



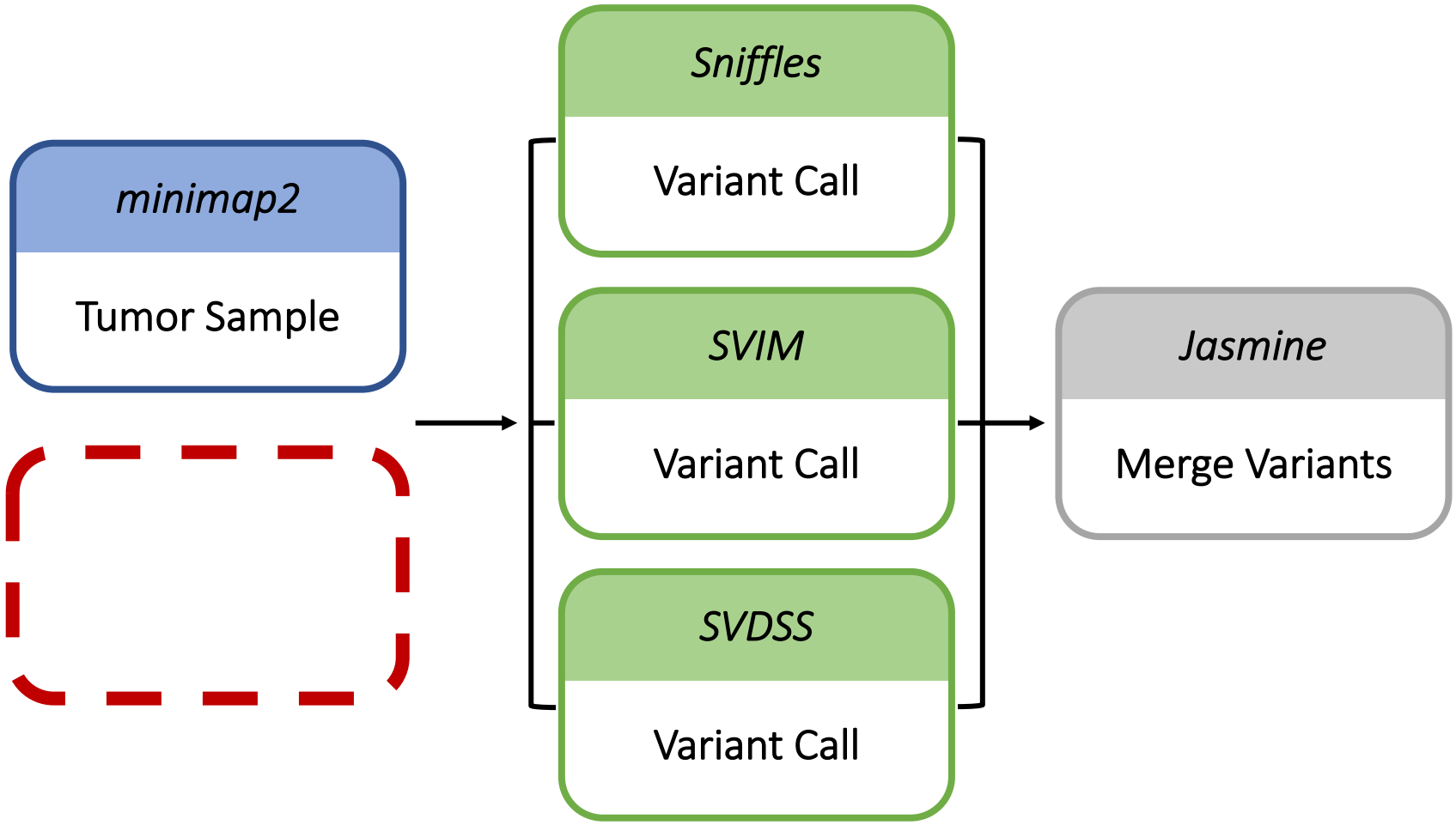
Call structural variants using algorithms specific to long-read sequencing.

Bioinformatics pipeline for identifying structural variants using long-read sequencing data.

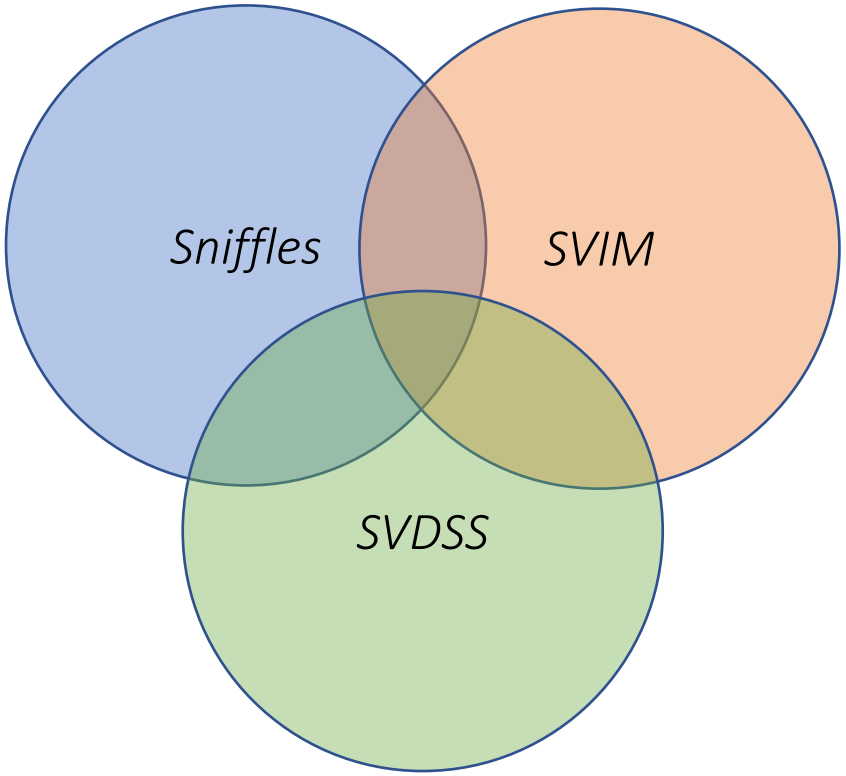


Find structural variants called by multiple algorithms.

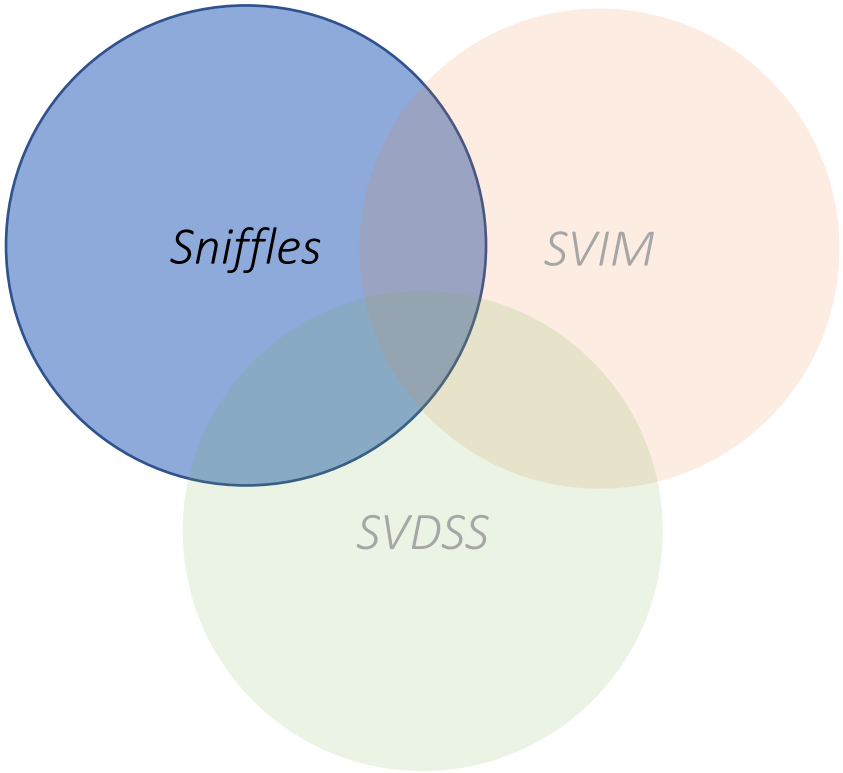
This project did *not* have a long-read normal sample to remove germline variants.



Multiple variant calling tools were used for long-read sequencing to improve detection.

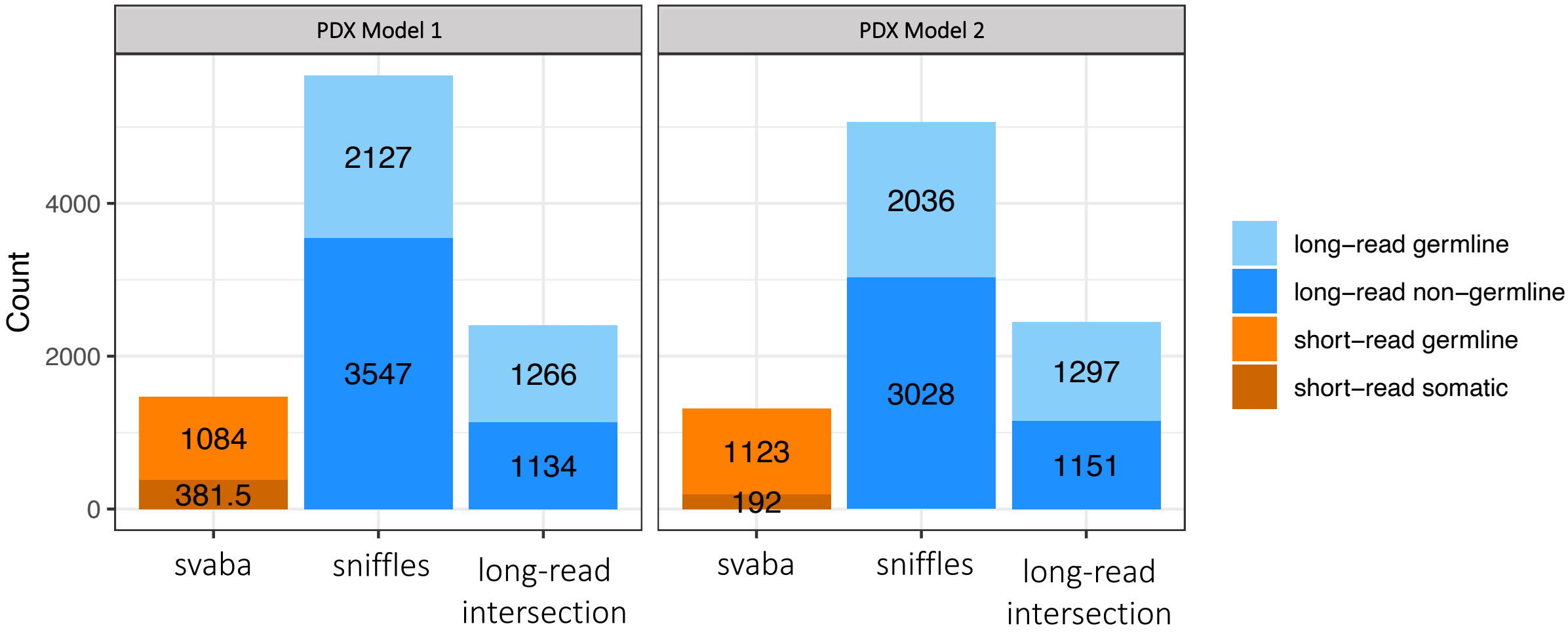


Multiple variant calling tools were used for long-read sequencing to improve detection.

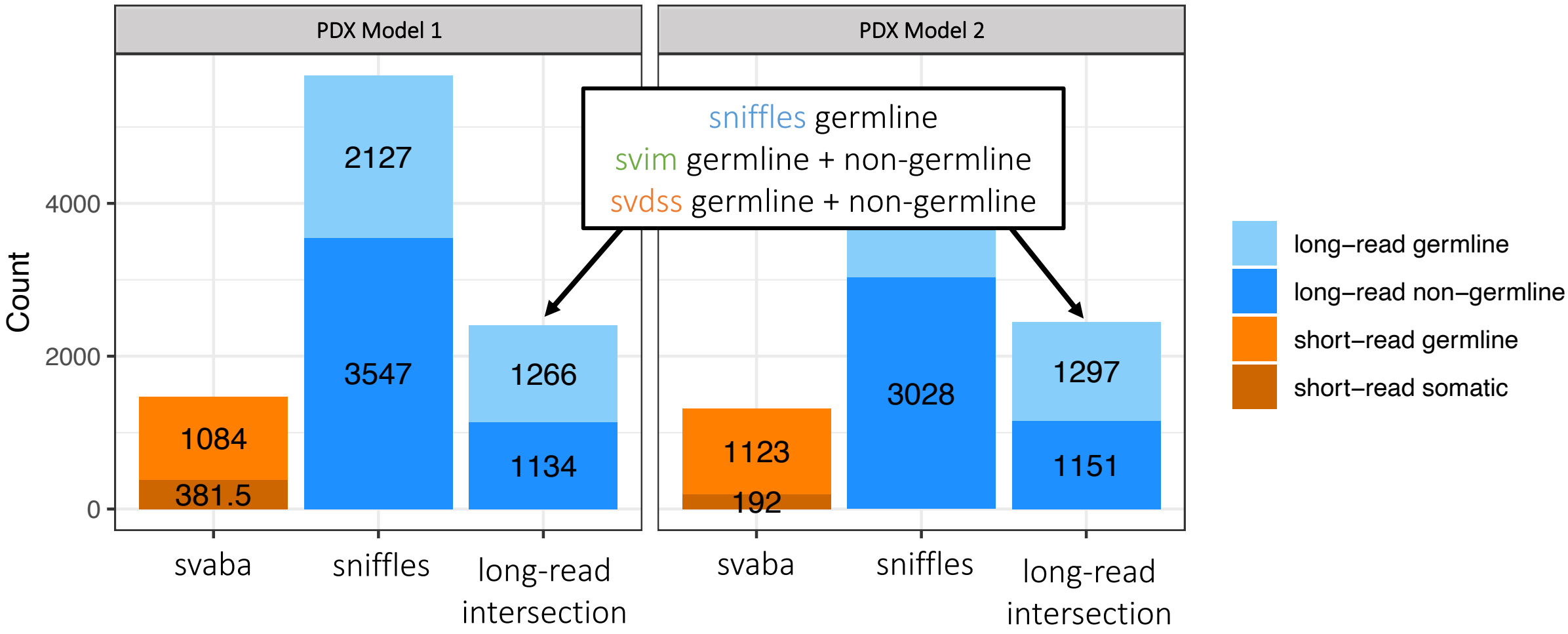


Identify non-germline variants by increasing sensitivity for low-frequency structural variants.

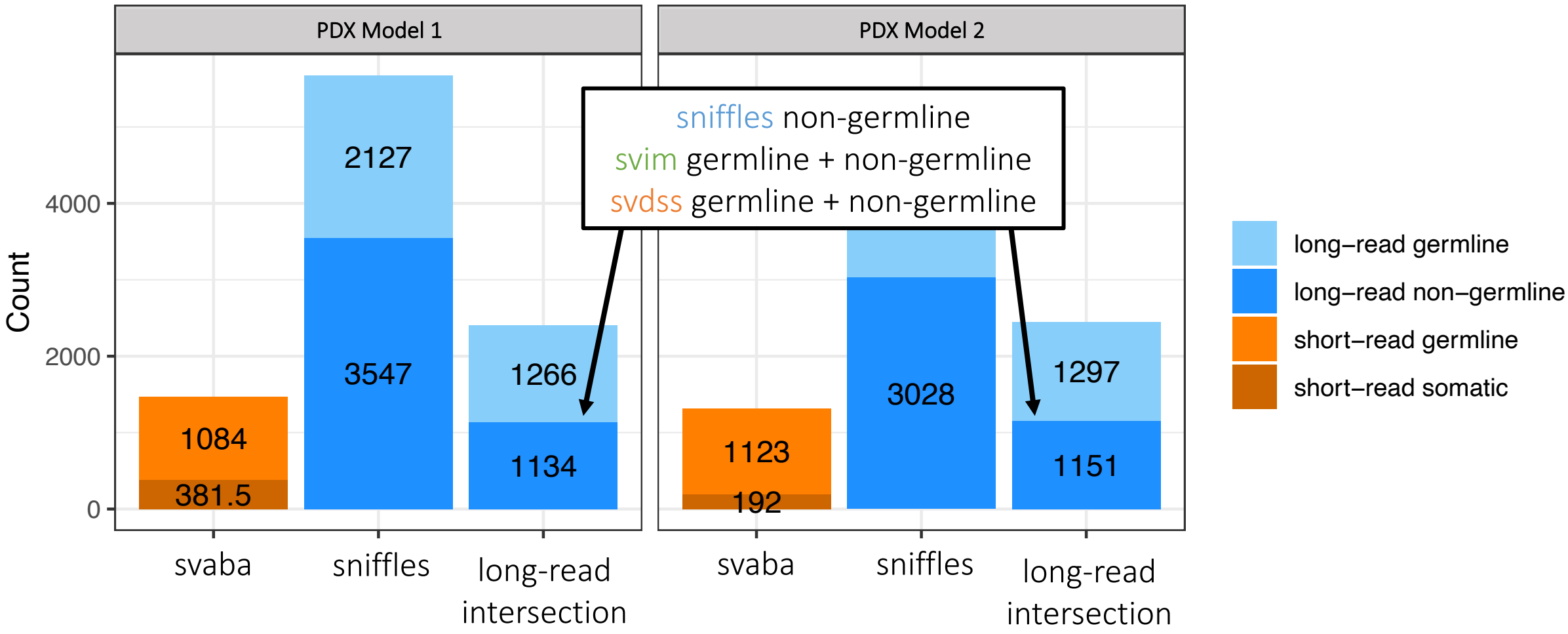
How many structural variants were identified by short-read and long-read technologies?



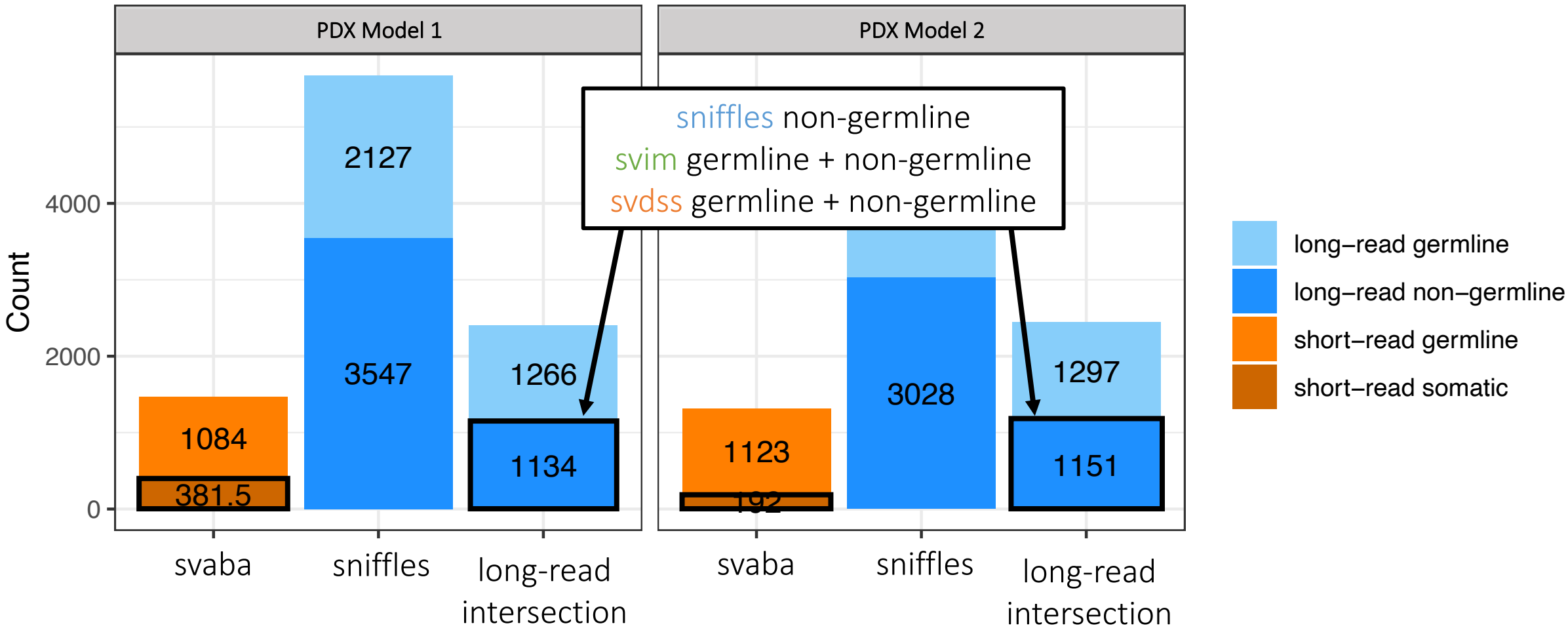
More long-read variants with support from *three* callers than *one* short-read caller.



More long-read variants with support from *three* callers than *one* short-read caller.

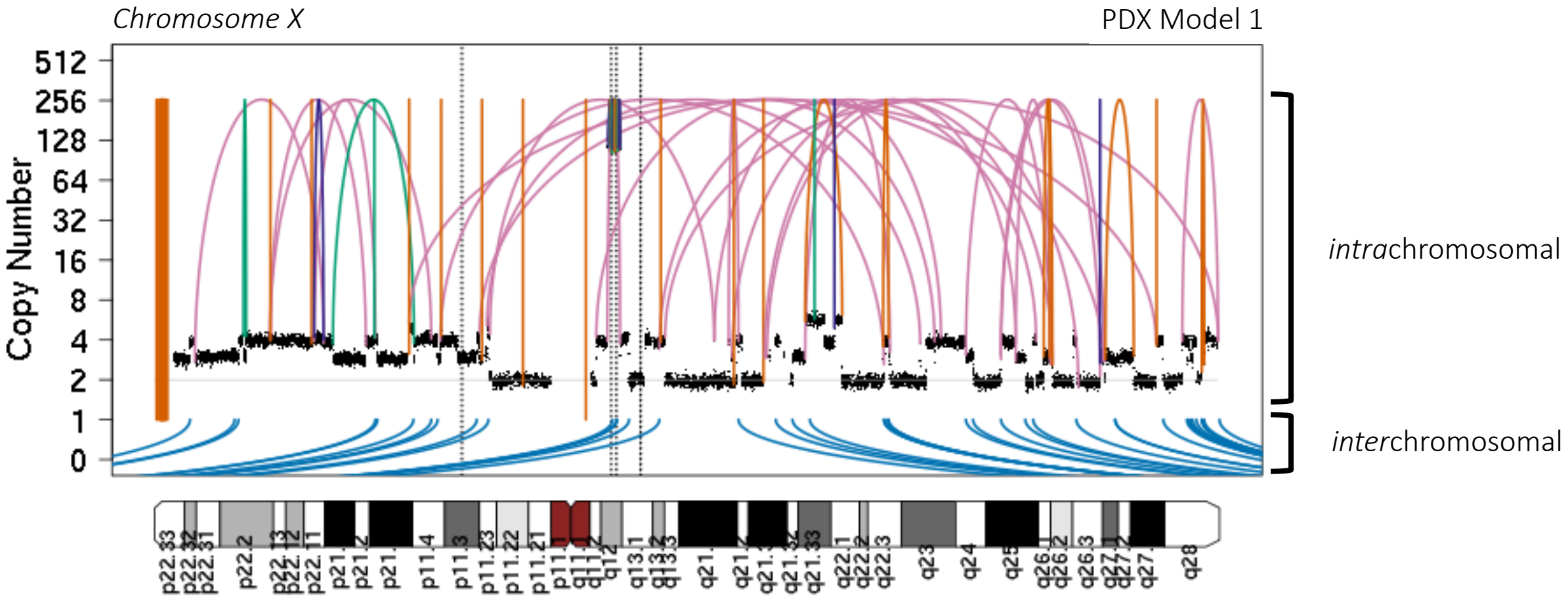


More long-read variants with support from *three* callers than *one* short-read caller.

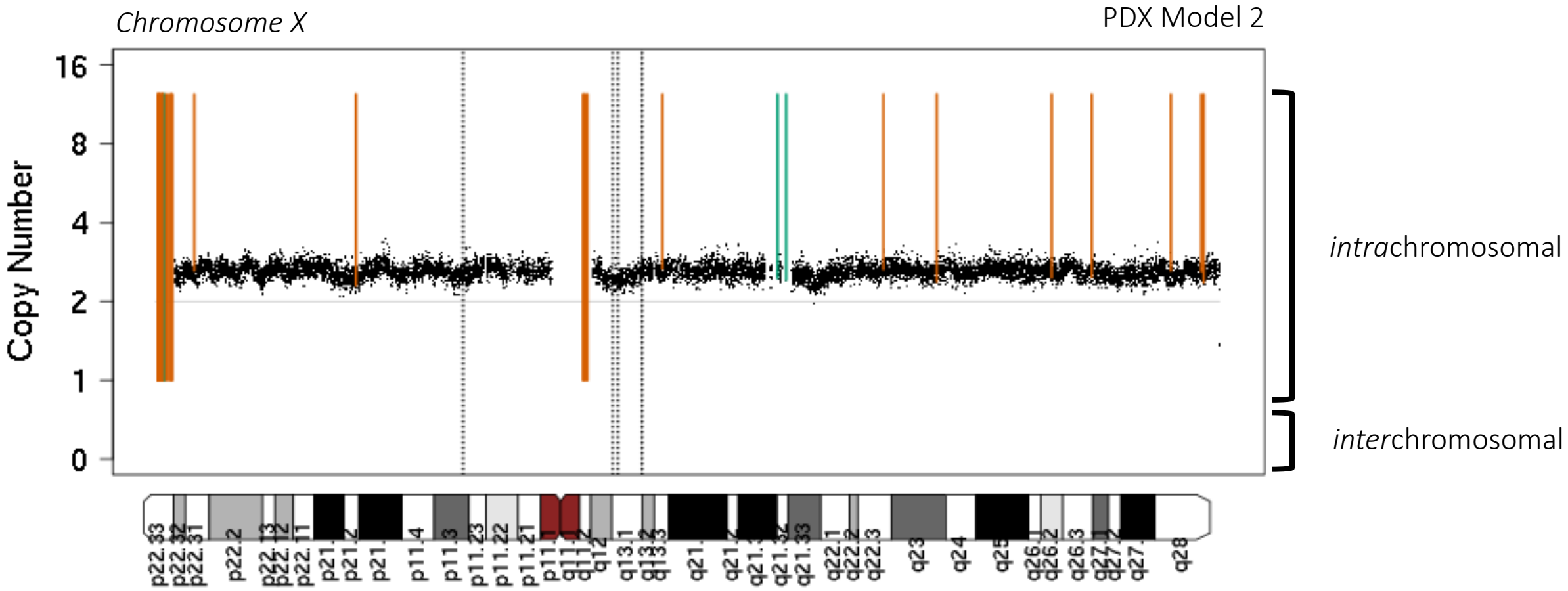


Do long-reads identify unknown genomic rearrangements important for understanding prostate cancer?

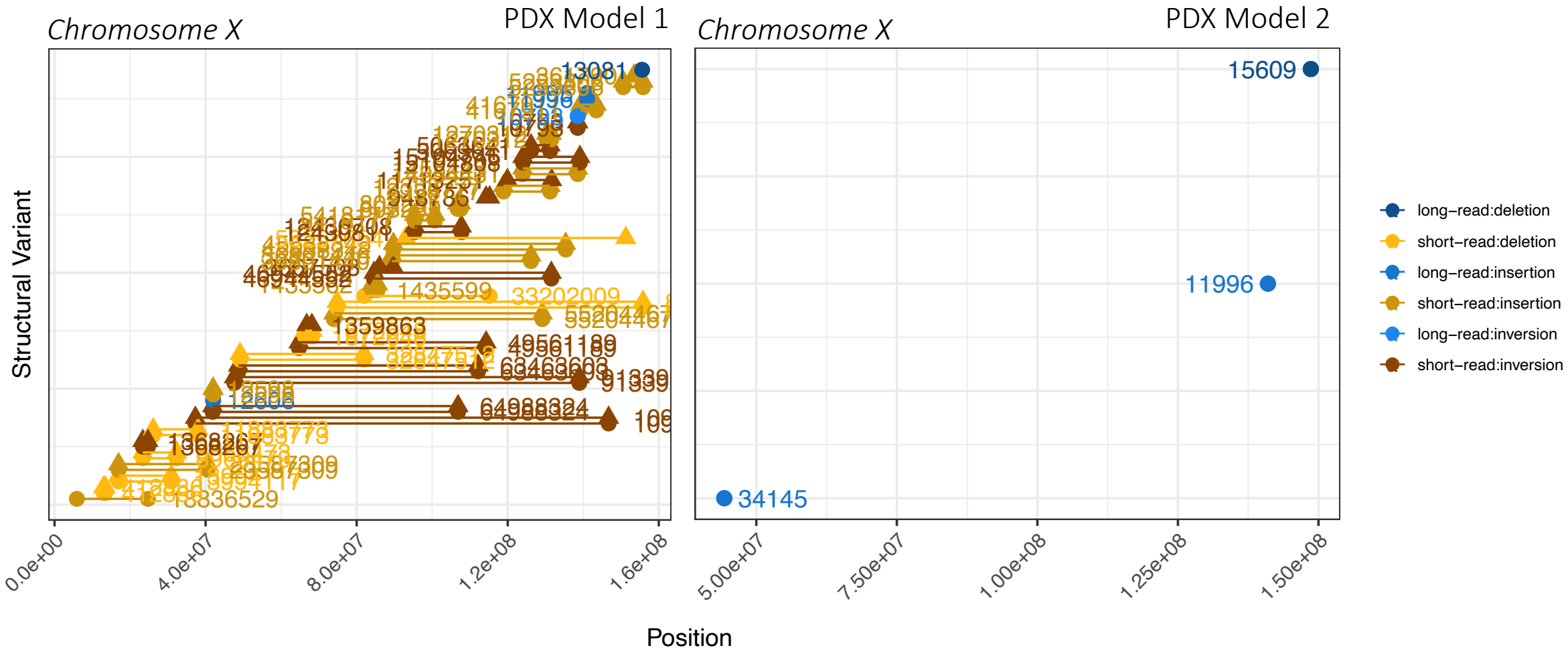
Previous research with short-reads has demonstrated chromothripsis in PDX model 1...



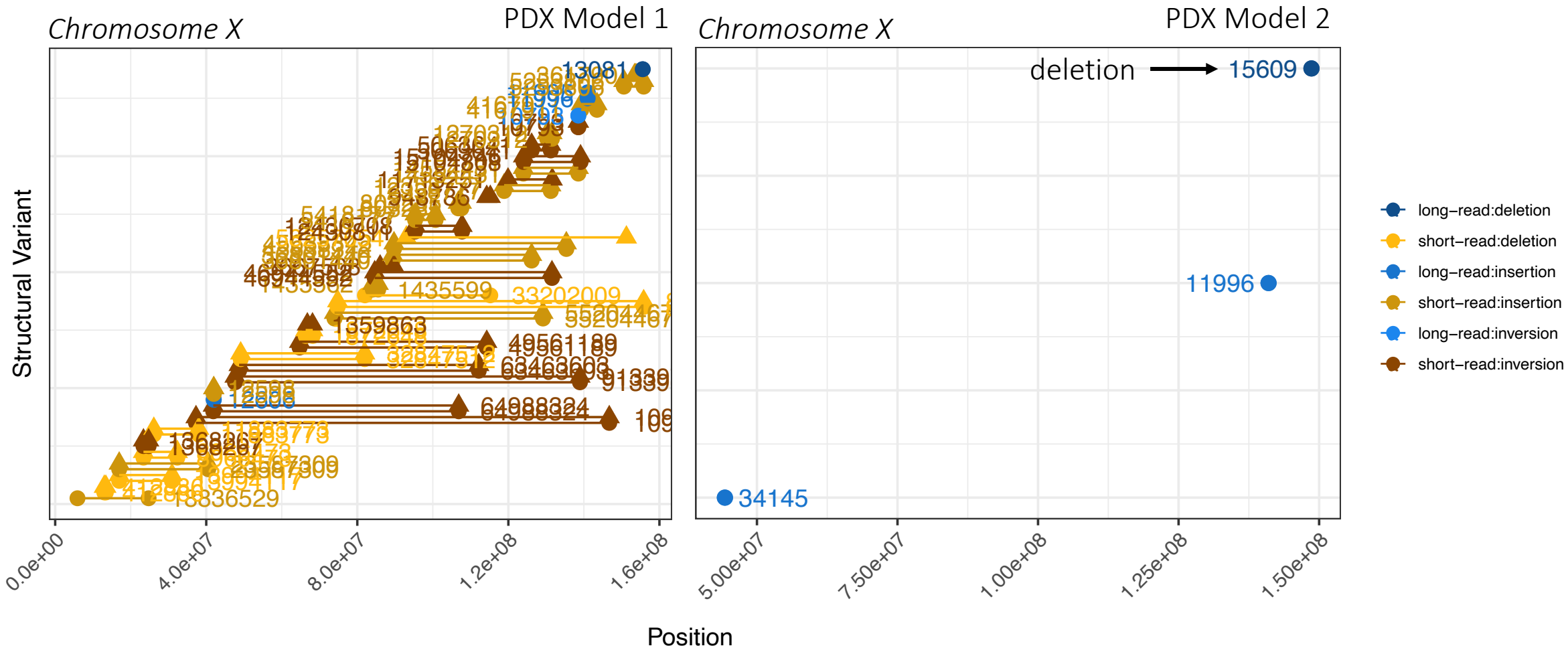
…but limited structural variation in PDX model 2.



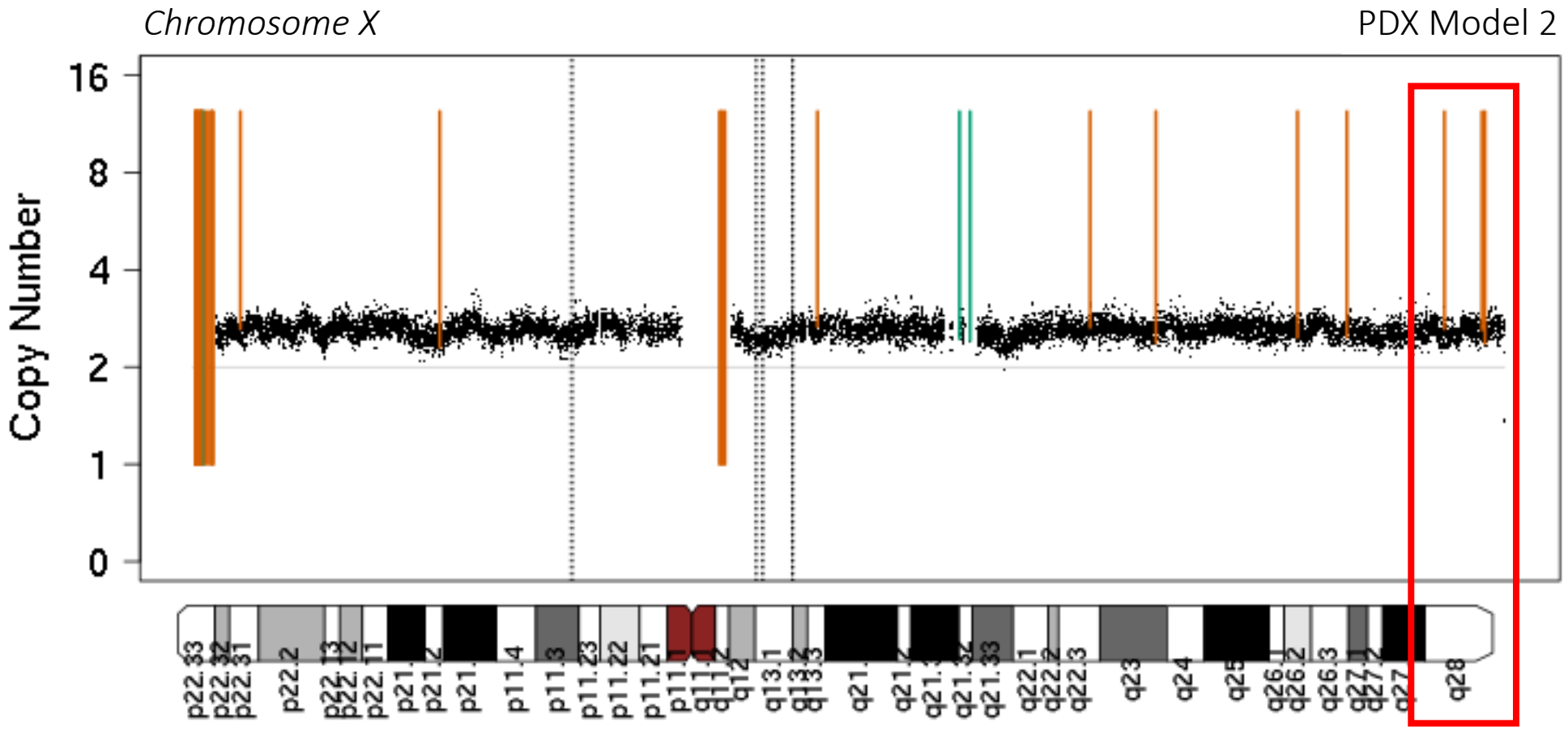
Short-read analyses indicate *many* variants in PDX model 1 and *no* variants in PDX model 2.



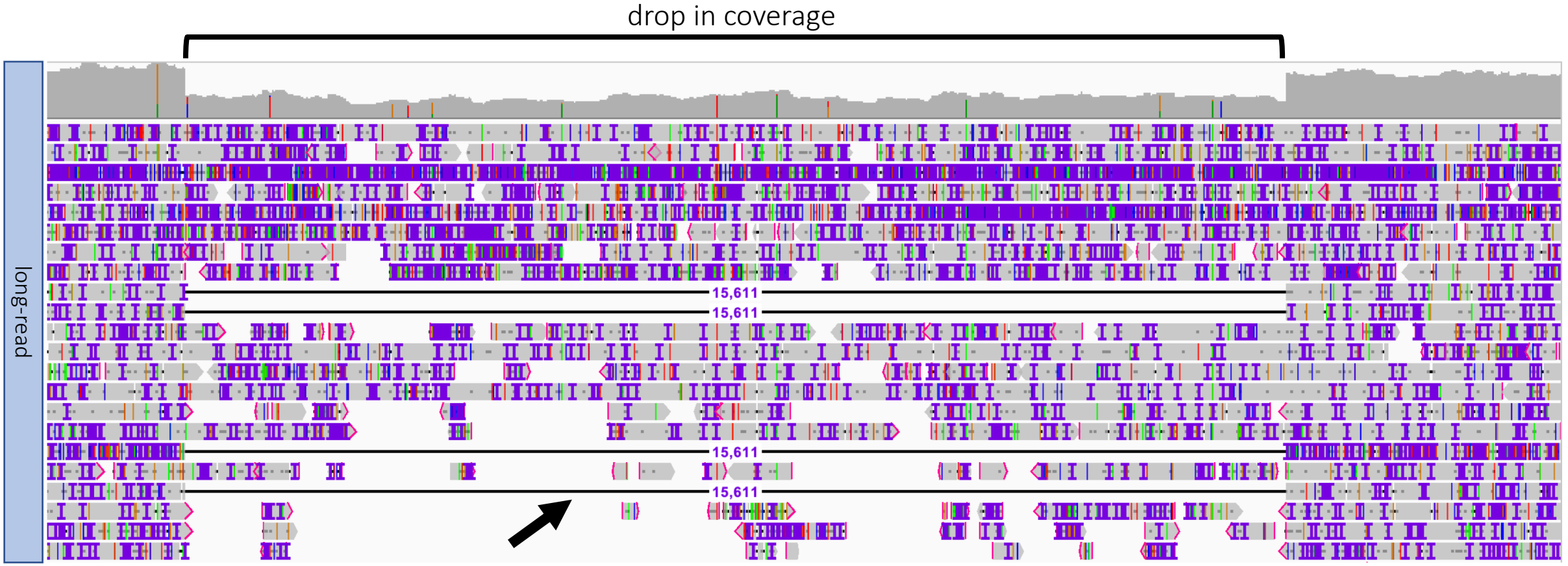
In PDX model 2, there is a long-read 15 Kb deletion that was *not* identified in short-reads.



Located in q28 on chromosome X, this deletion also does *not* appear in prior research.

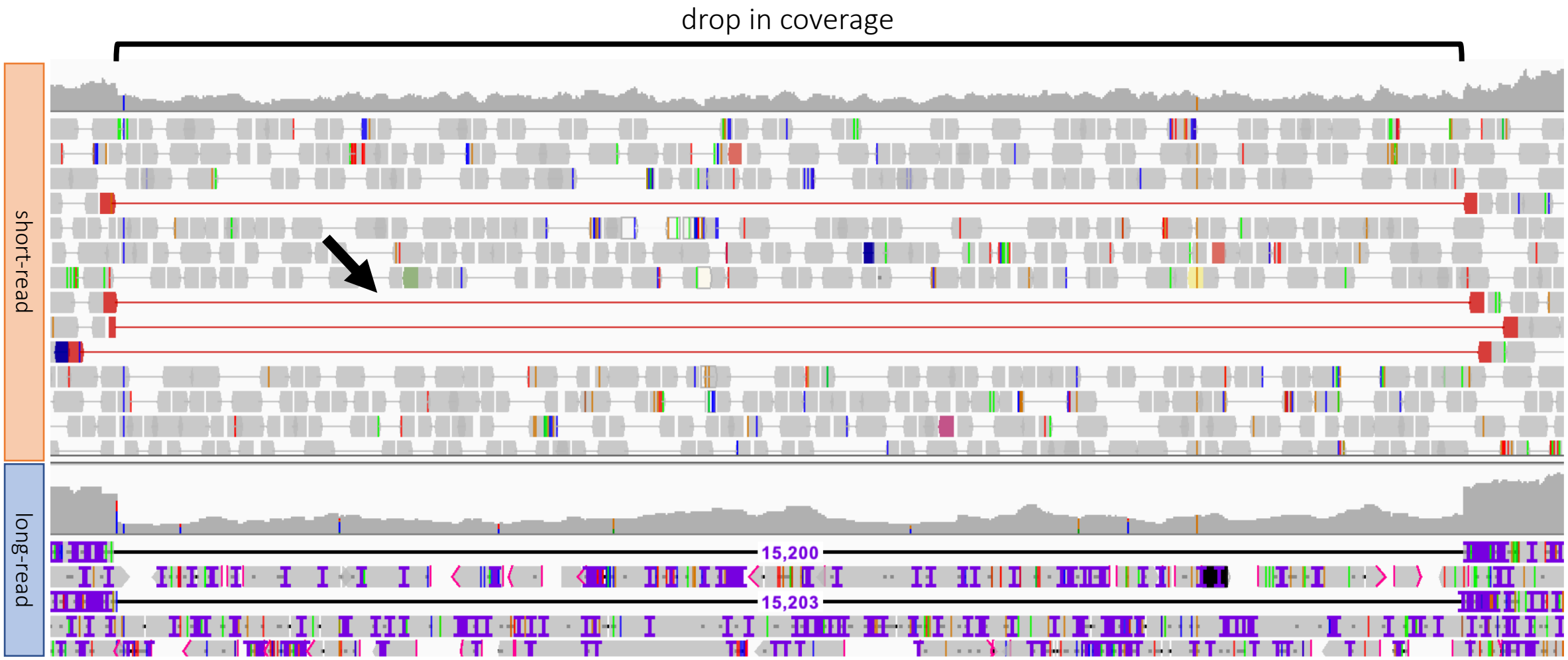


For long-reads, the Integrative Genomics Viewer (IGV) shows a drop in coverage and the deletion.



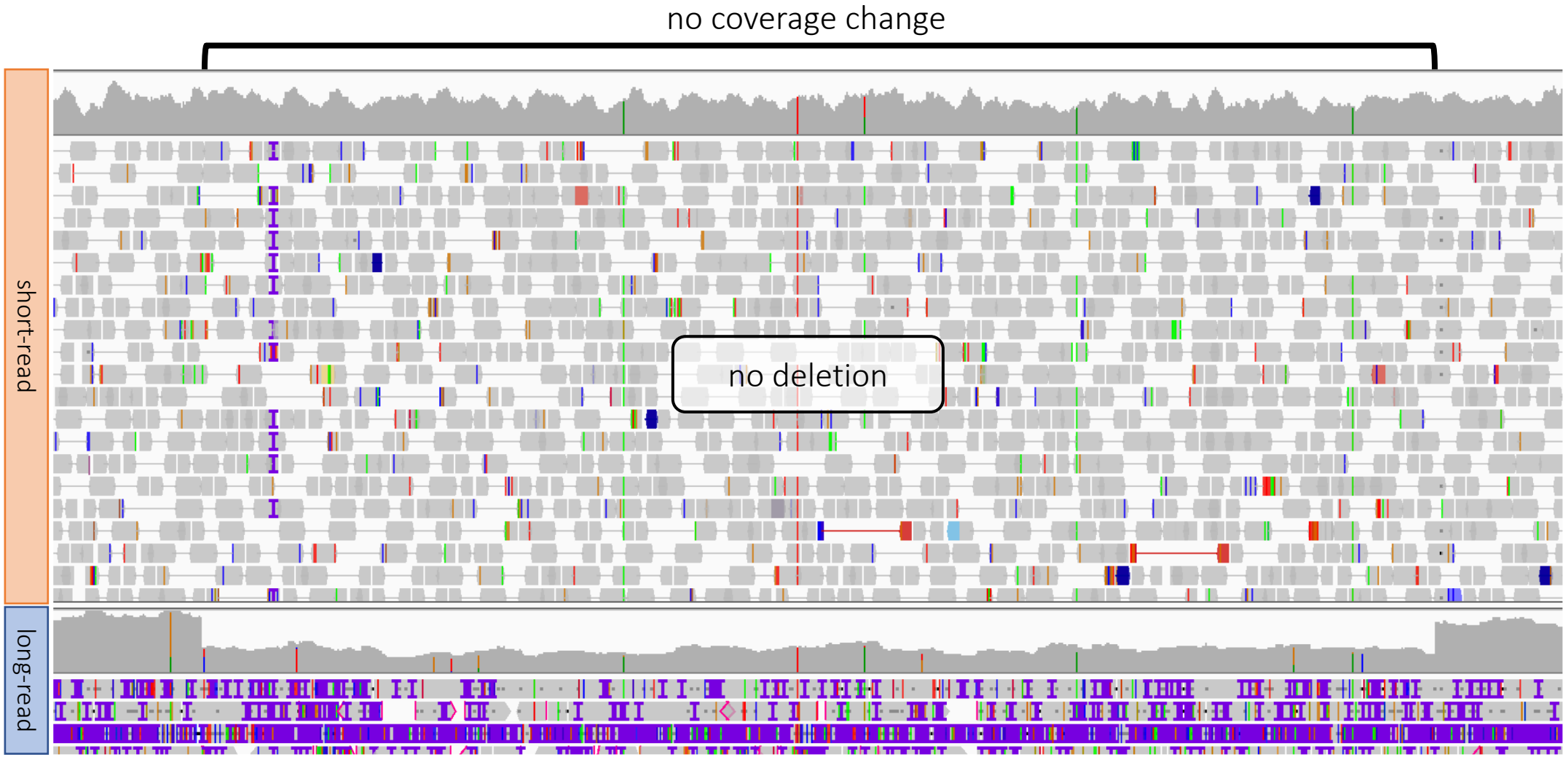
Chromosome X deletion

For short-reads, the IGV *would* also show a drop in coverage and deletion.



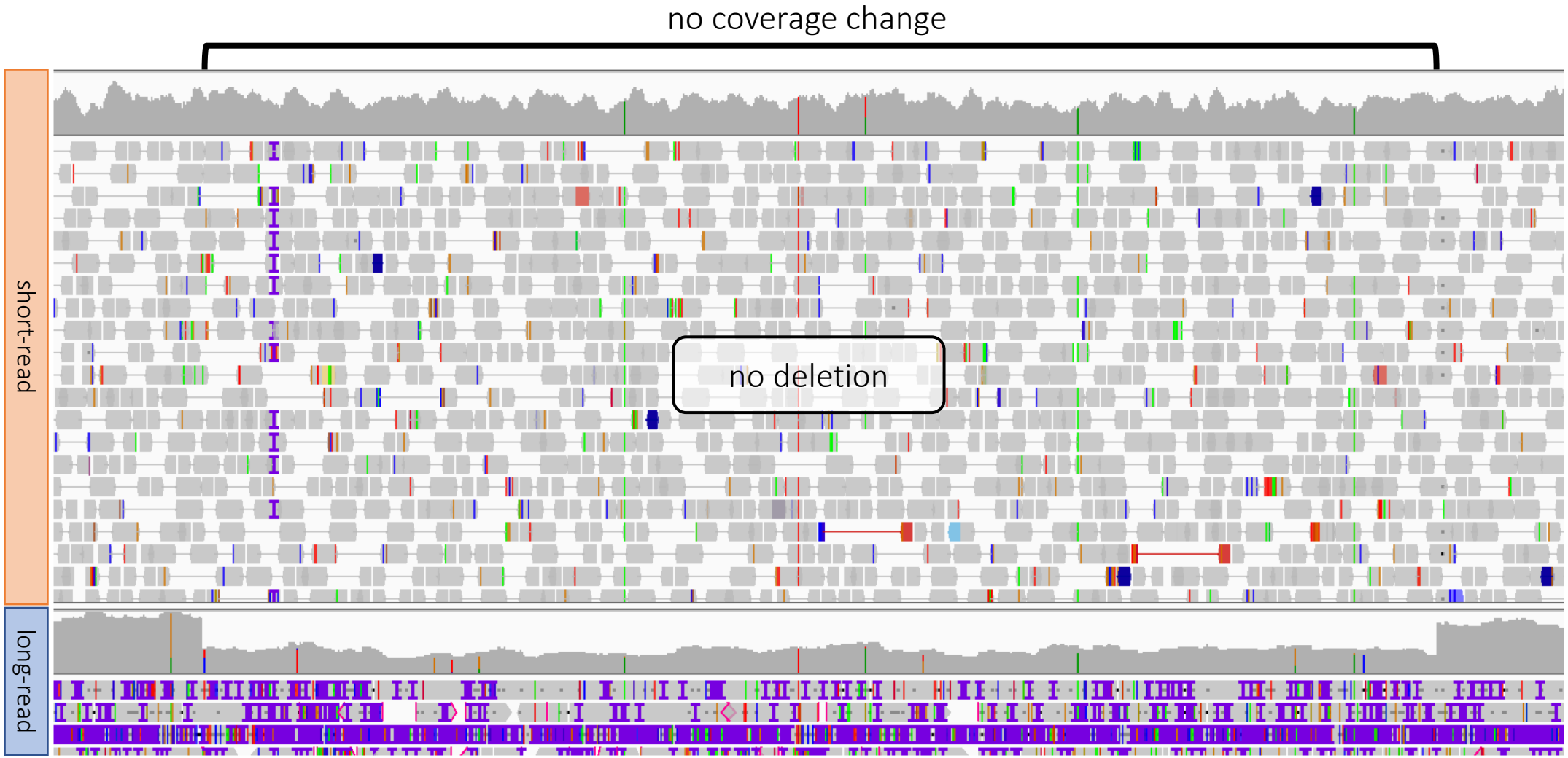
Chromosome 13 deletion

For short-reads, there is no drop in coverage and none of the reads show a large deletion.



Chromosome X deletion

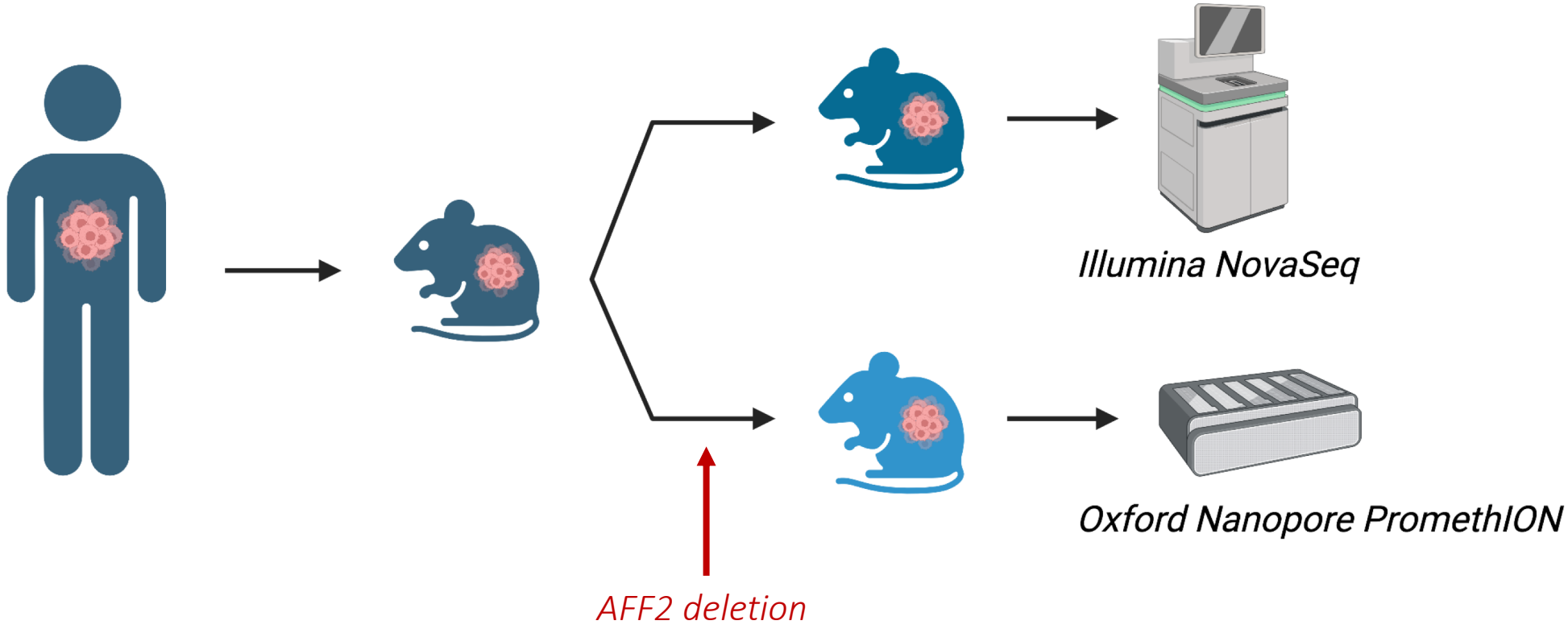
Deletion is in the *AFF2* gene and mutations in this gene have been implicated in breast cancer.



Chromosome X deletion

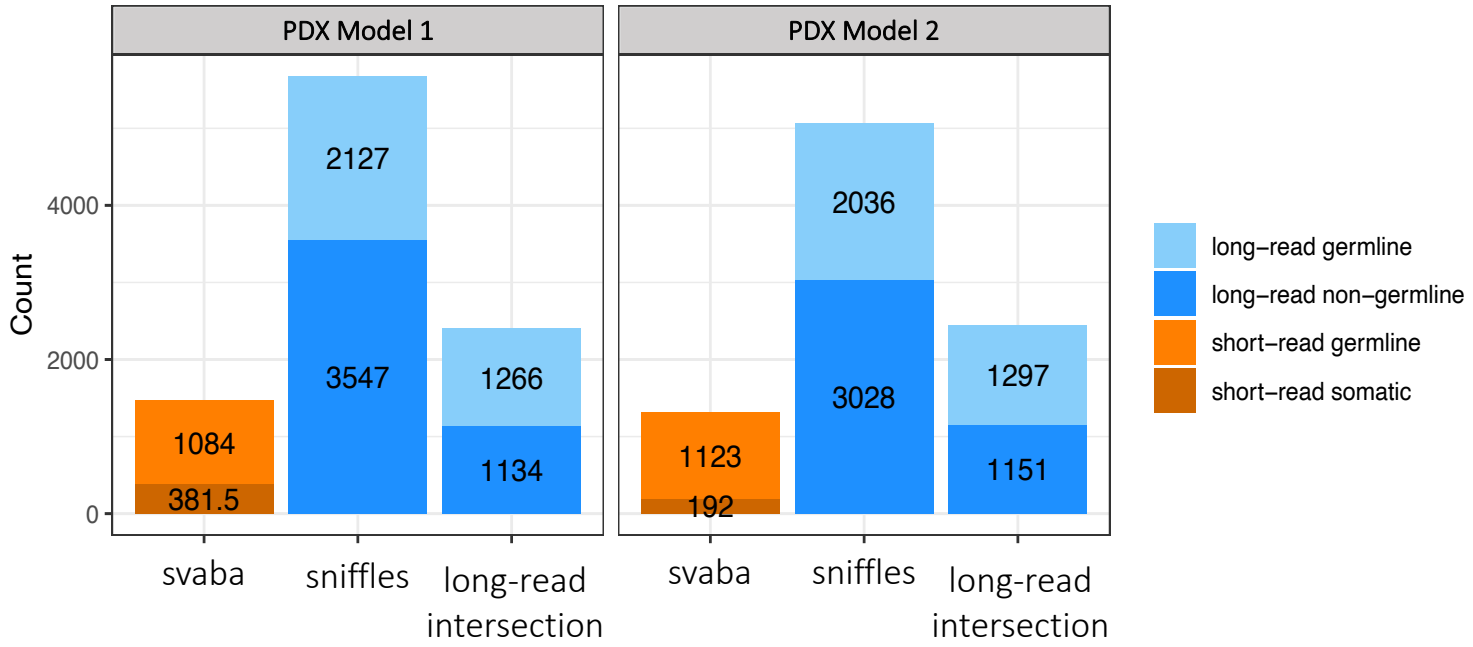
The Cancer Genome Atlas Network, Nature 2012

Short-reads were unable to identify *AFF2* deletion **or** this variant occurred between passages.



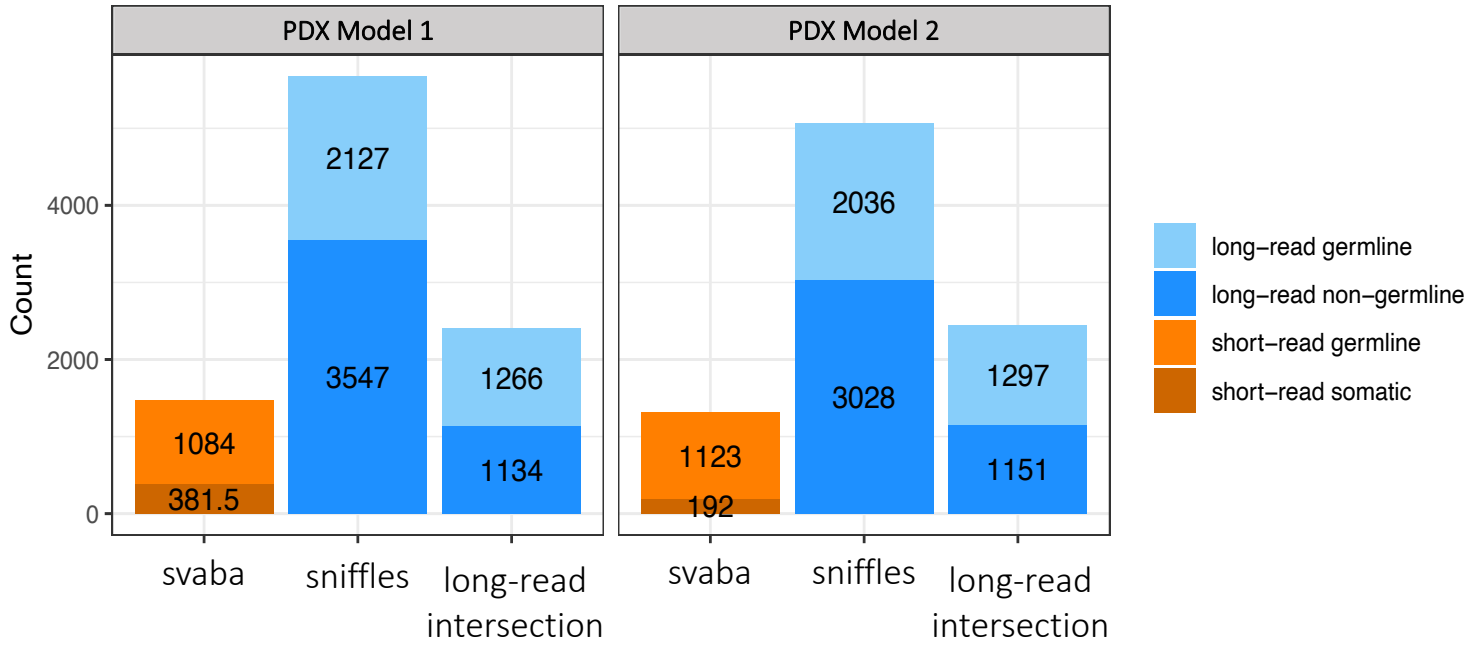
Conclusions

Identified more *high confidence* long-read structural variants than the *total* short-read variants...



Conclusions

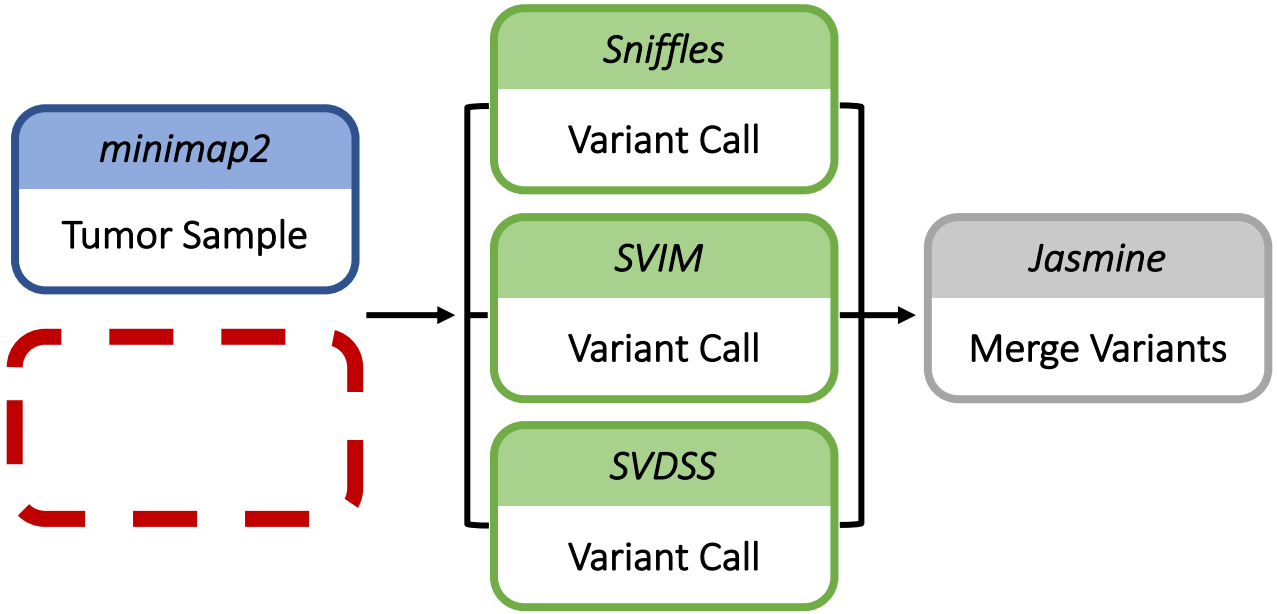
…additional structural variants are germline, unknown genomic rearrangements or variants that occurred between passages.



Conclusions

…additional structural variants are germline, unknown genomic rearrangements or variants that occurred between passages.

A long-read matched-normal comparison would be beneficial for removing germline variants.



Conclusions

…additional structural variants are germline, unknown genomic rearrangements or variants that occurred between passages.

A long-read matched-normal comparison would be beneficial for removing germline variants.

Long-reads can identify unknown genomic rearrangements important for understanding malignancy.

Acknowledgements

Ha Lab

Gavin Ha
Patty Galipeau
Robert Patton
Pushpa Itagi
Anat Zimmer
Sitapriya Moorthi
Eden Cruikshank
Mohamed Adil
Michael Yang
Erin Kawelo

Miller Lab

Danny Miller

Molecular & Cellular Biology in Seattle
AN INTERDISCIPLINARY Ph.D. PROGRAM *offered through the*
UNIVERSITY OF WASHINGTON *and* FRED HUTCH

Fred Hutch
Cancer Center