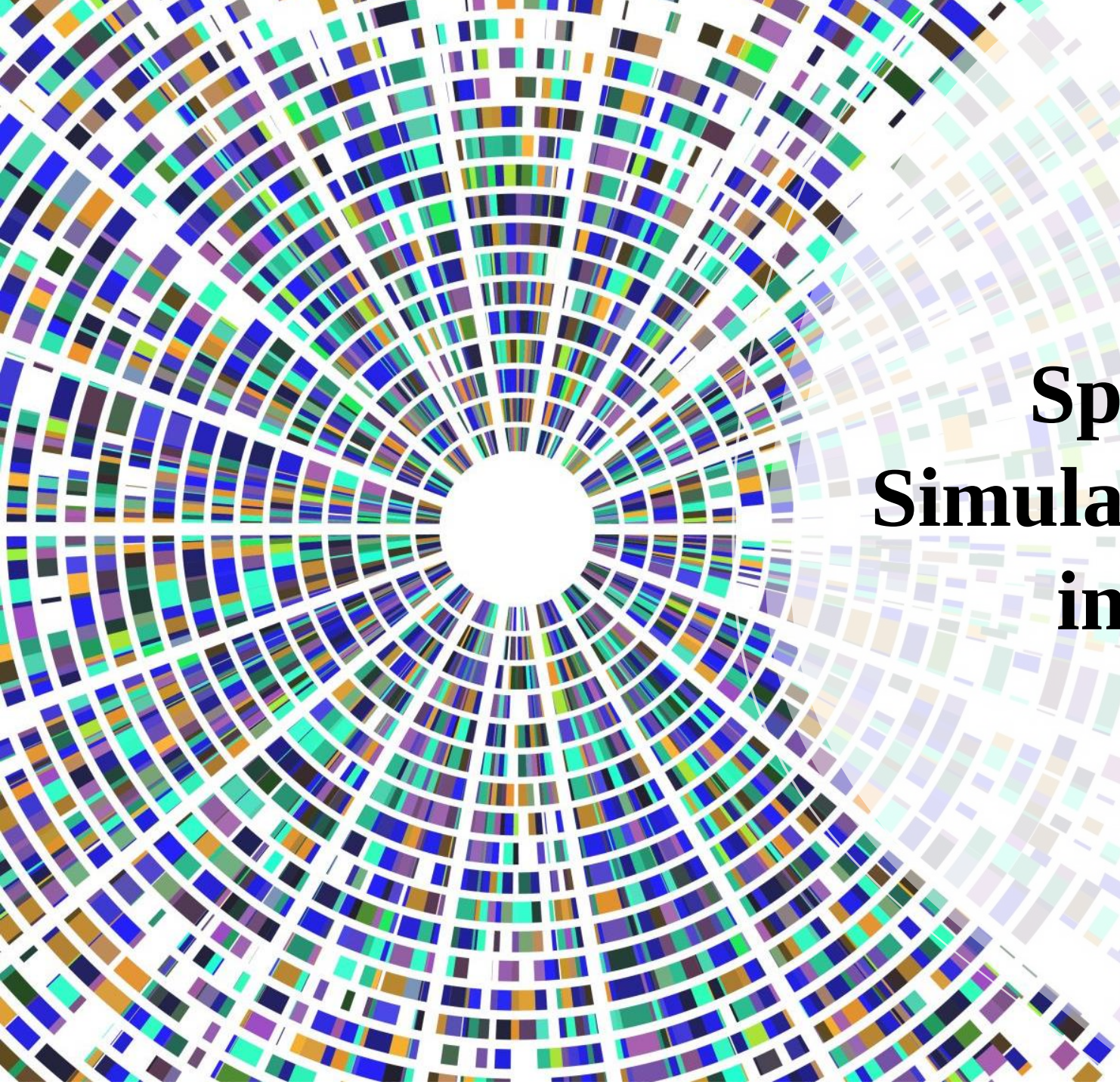


Group 2 :
SpikeVar & TykeVar
Simulation of Mosaic Variants
in Sequencing Data

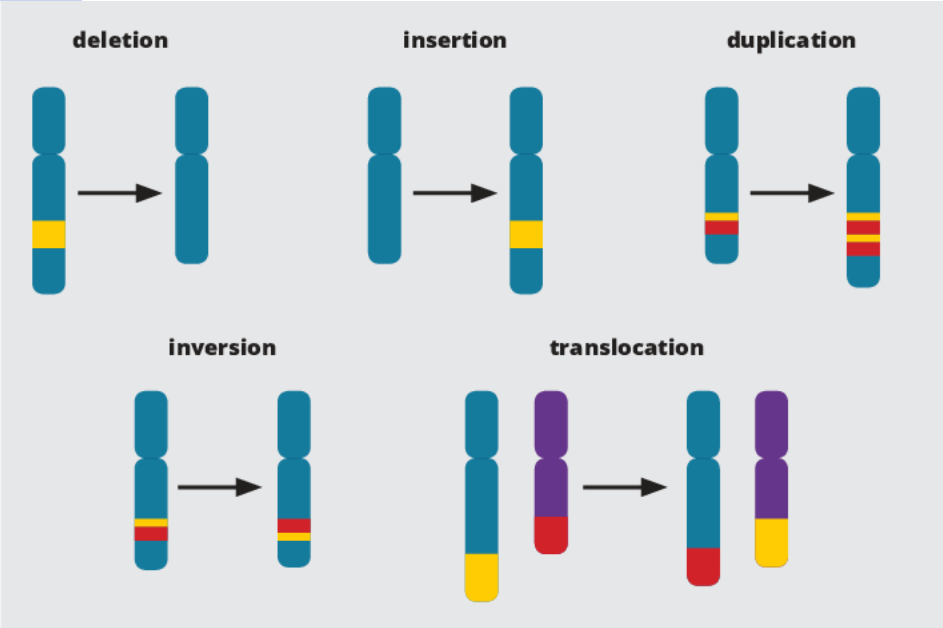
Farhang Jaryani,

Fritz Sedlazeck, Lab

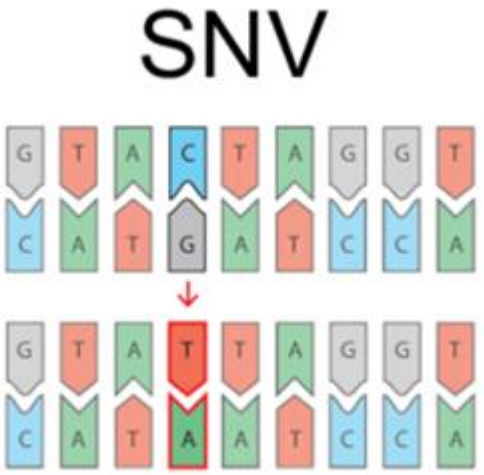
Human Genome Sequencing Center

Baylor college of Medicine

Structural Variants

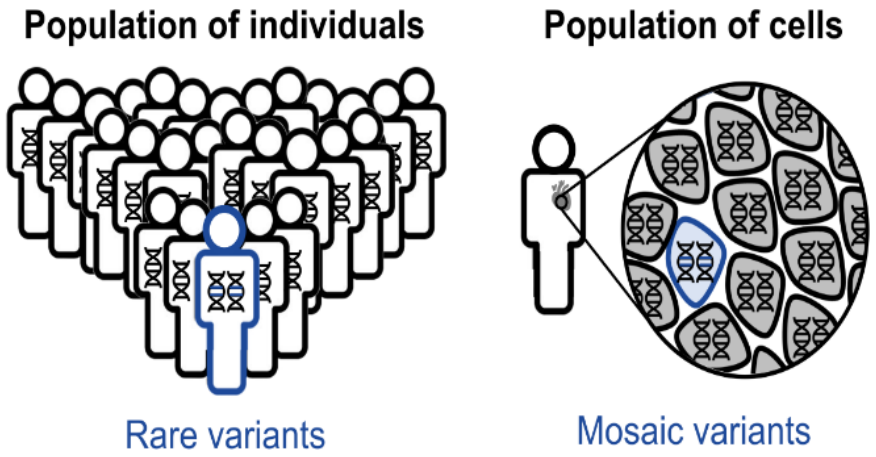


Reference	1 2 3
Insertion	1 2 5 3
Deletion	1 3
Inversion	1 3 2
Copy Number Variation	1 1 1 1 2 3
Tandem Duplication	1 1 2 3
Dispersed Duplication	1 2 1 3
Mobile Element Insertion	1 2 Mobile Element 3
Translocation	1 / 10 11 12 2 3



HUNTING STRUCTURAL VARIANTS , Luke Hickey
https://api.semanticscholar.org/CorpusID:204805198

Mosaic Variants



Mosaic Variants:

- ✓ **Potentially associated with disease**
- ✓ **Mixed in sequence data**
- ✓ **Requires mosaic variant callers to detect**

THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

Home Profiles Research units **Research output** Projects

Disease-associated mosaic variation in clinical exome sequencing: a two-year pediatric tertiary care experience

Cecelia R. Miller, Kristy Lee, Ruthann B. Pfau, Shalini C. Reshmi, Donald J. Corsmeier, Sayaka Hashimoto, Ashita Dave-Wala, Vijayakumar Jayaraman, Daniel Koboldt, Theodora Matthews, Danielle Mouhlas, Maggie Stein, Aimee McKinney, Tom Grossman, Benjamin J. Kelly, Peter White, Vincent Magrini, Richard K. Wilson, Elaine R. Mardis, Catherine E. Cottrell

Comprehensive Cancer Center, Comprehensive Cancer Center - Innate Immunity, Pathology, Pediatrics, Center for Clinical and Translational Science, Comprehensive Cancer Center - Molecular Biology and Cancer Genetics, Comprehensive Cancer Center - Experimental Therapeutics

Research output: Contribution to journal › Article › peer-review

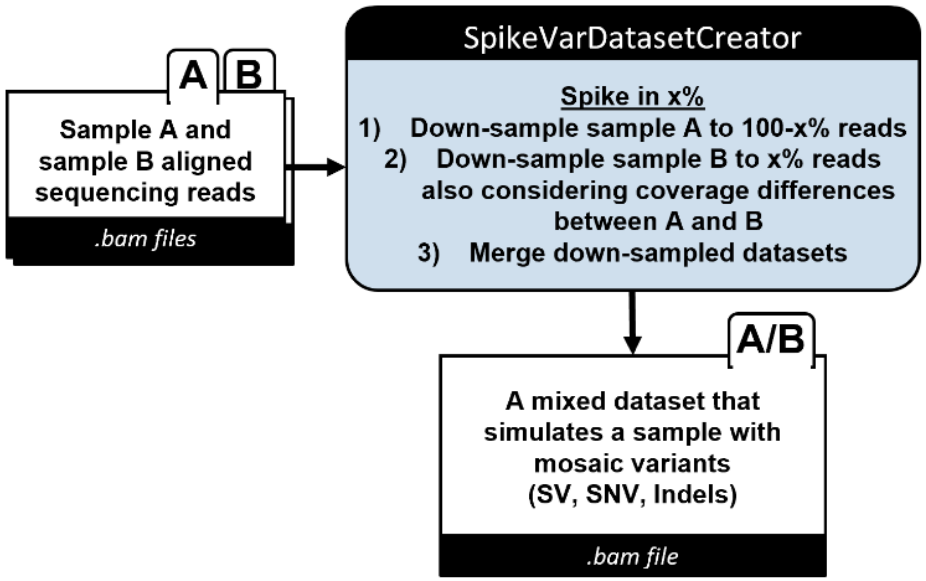
Disease-associated mosaic variation in clinical exome sequencing: a two-year pediatric tertiary care experience. Cold Spring Harb Mol Case Stud. 2020;6(3). Epub 20200612. doi: 10.1101/mcs.a005231. PubMed PMID: 32371413; PMCID: PMC7304353.

Spike in Known Exogenous Variants

❖ *Introducing User-Defined Spiking in Sequencing Data*

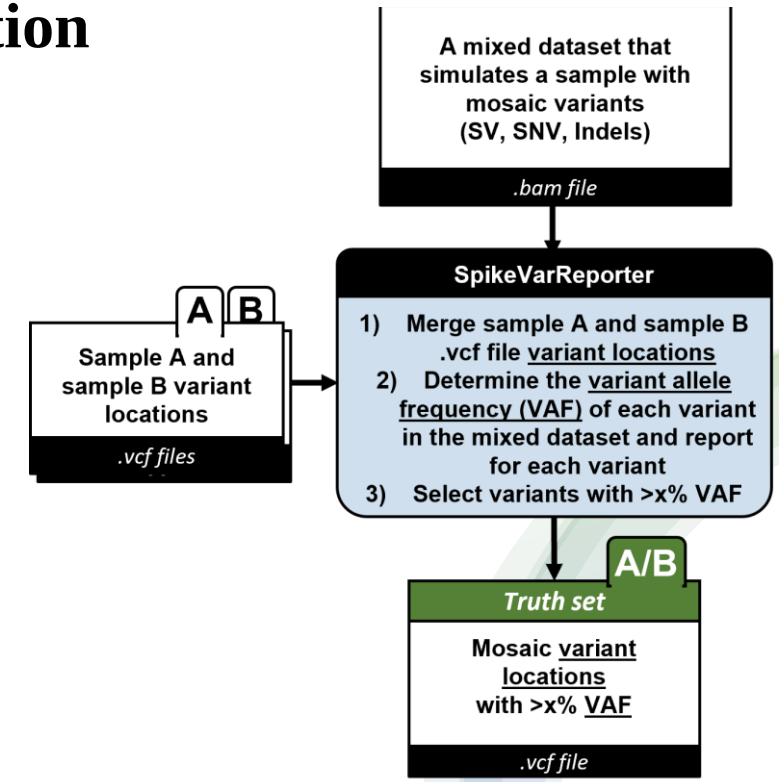
1) SpikeVar DatabaseCreator –

Generate Spiked-in Dataset

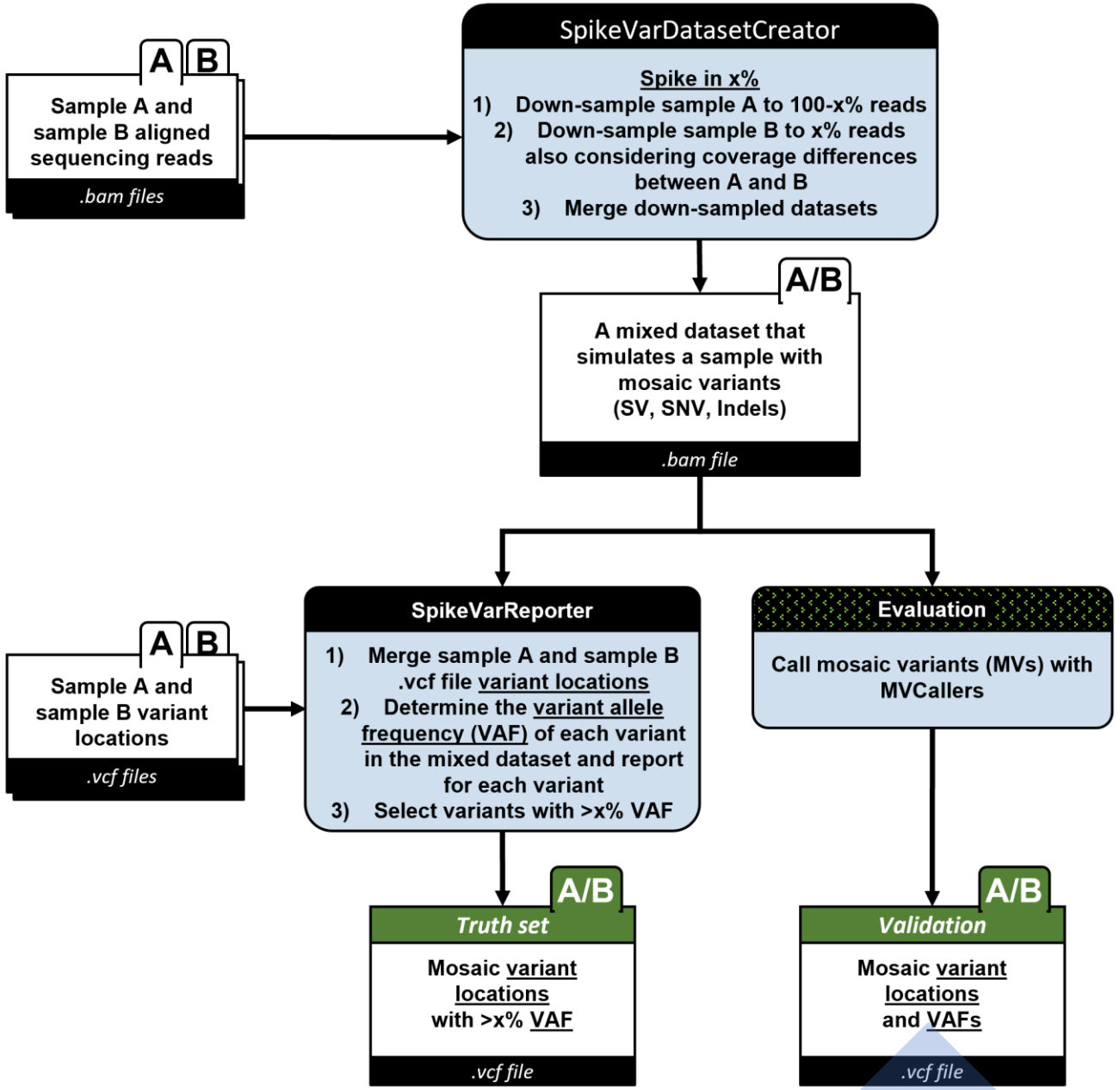


2) SpikeVarReporter –

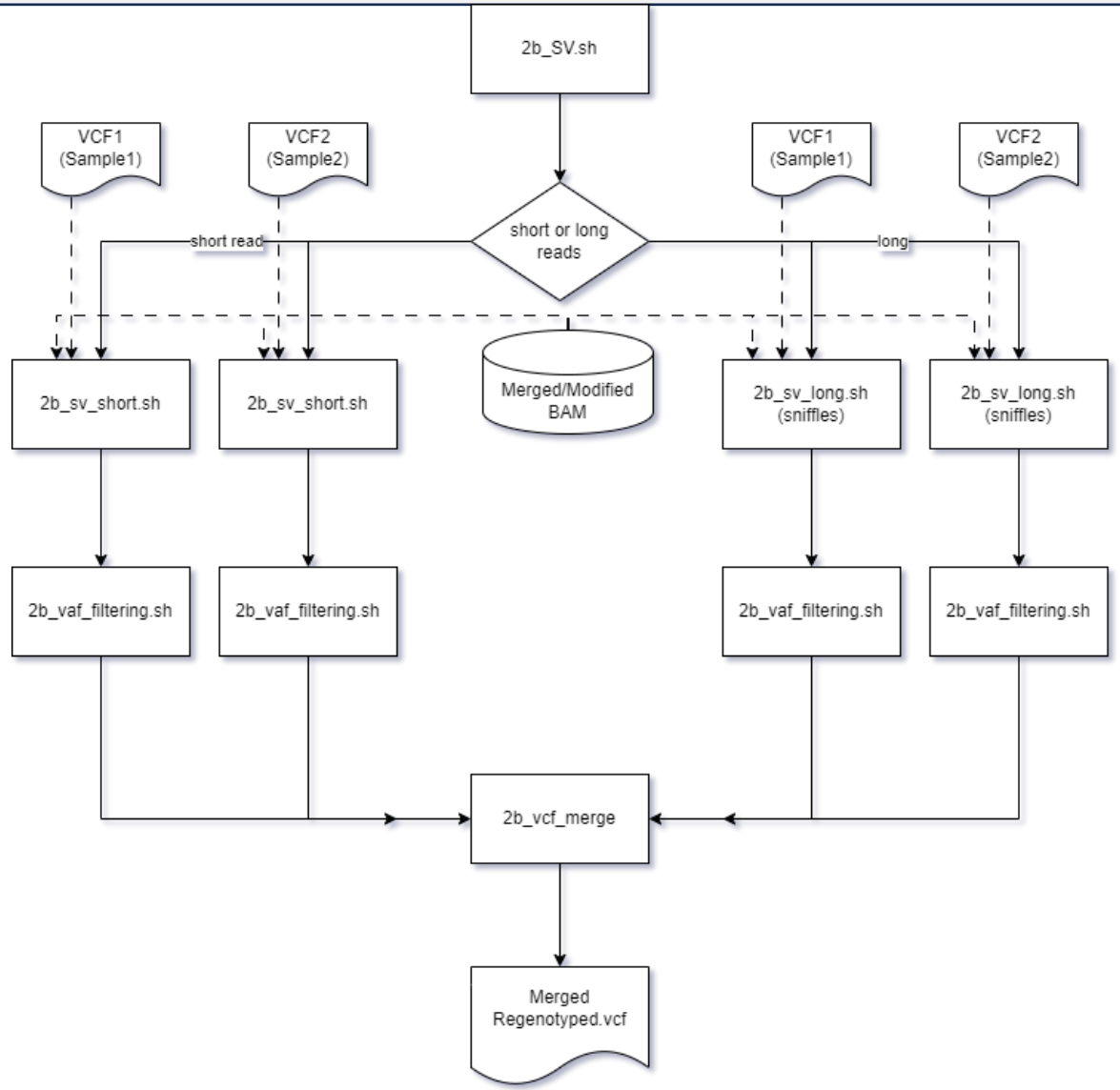
Filter Reads After Variant Allele Frequency Recalculation



*Sp*ike *i*n *K*nown *E*xogenous *Var*iants

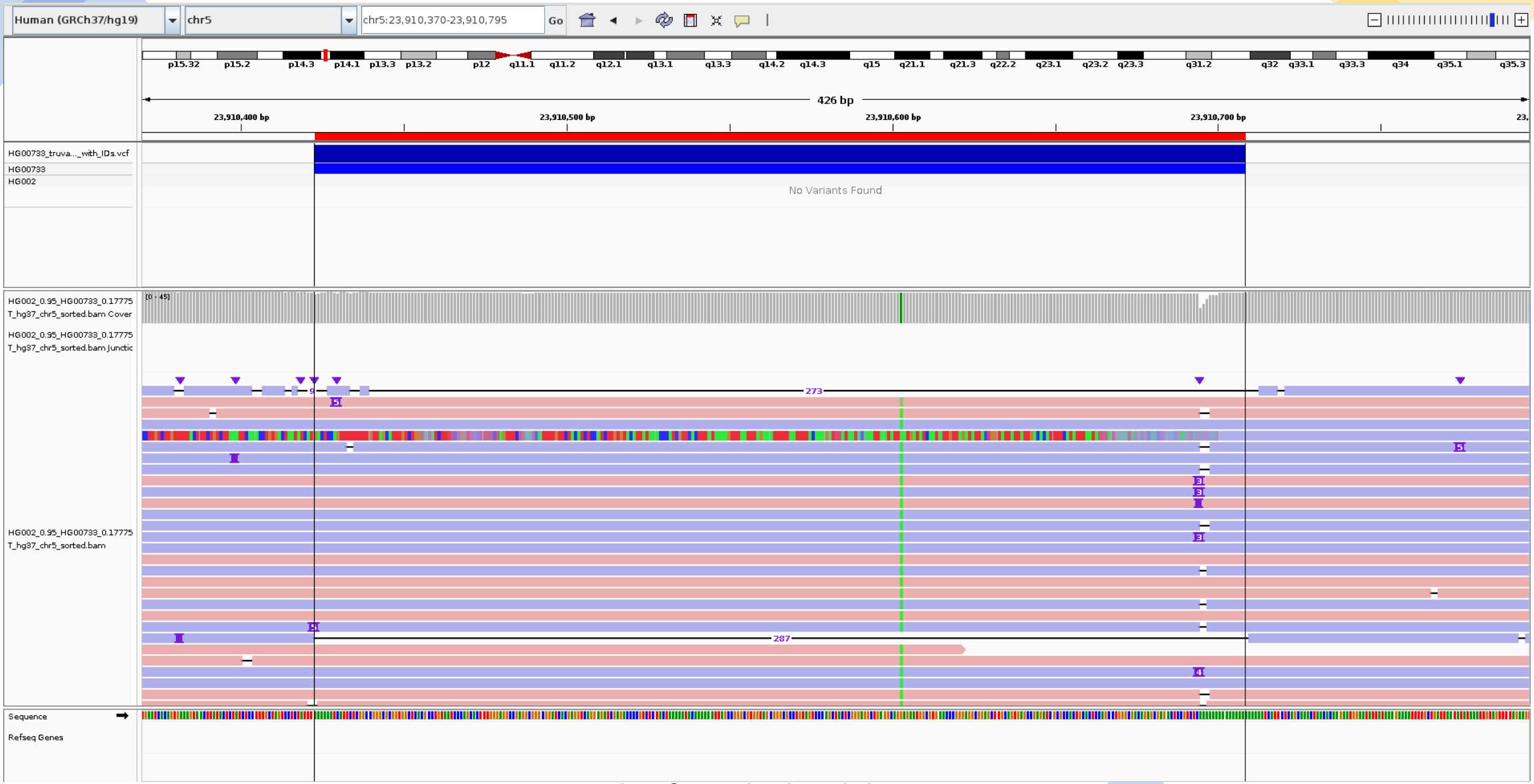


*Sp*ike *i*n *K*nown *E*xogenous *Var*iants



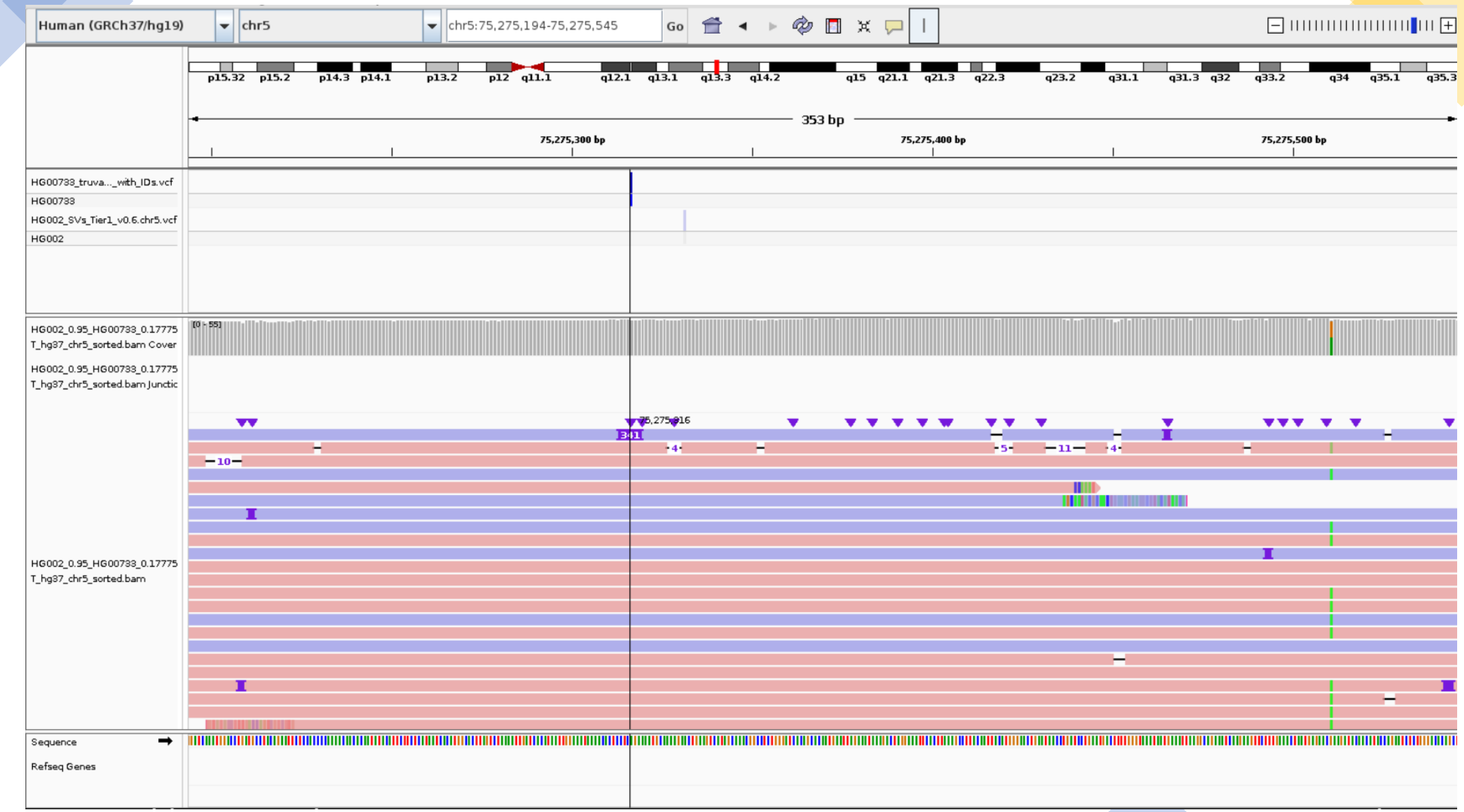
Flowchart of re-genotototyping both short and long reads SVs based on the modified generated bam file

Spike in Known Exogenous Variants



Example of a spiked in deletion.

*Sp*ike *i*n *K*nown *E*xogenous *Var*iants

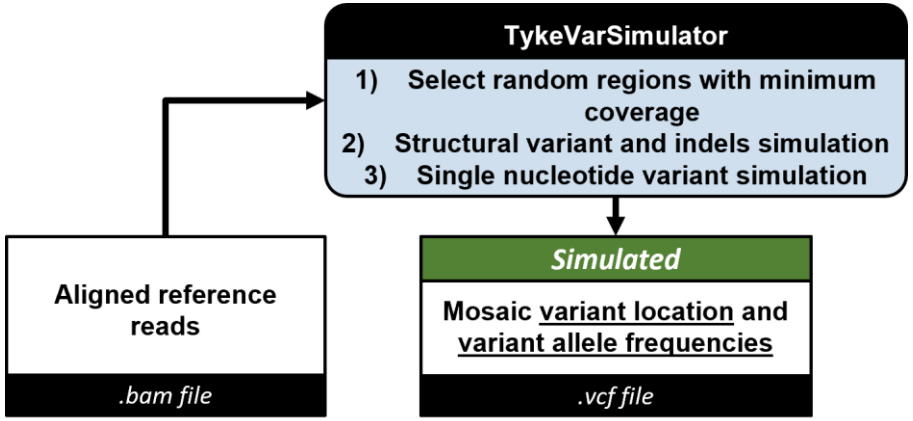


Example of a spiked in insertion.

*T*rack in *Y*our *K*ey *E*ndogenous *Var*iants

- creates a list of random mutations
- modifies a fraction of existing reads to match the user-defined MV frequency

1) TykeVarSimulator – Generate Simulated VCF



#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

chr22 34898635 HackDel1 N <DEL> 60 PASS AF=0.12;END=34902269;PRECISE;SVLEN=-3634;SVTYPE=DEL GT:GQ 0/0:60

chr22 32562272 HackIns1 N
CGGCCTTGGAACTTTAACCATCTGTCCACTACGTTGTACATATTCATCTTCGTCTAGCCAAAGGCGAACCCGTCCTGGAATCTATATTACCTTTTTATGCCGCG>

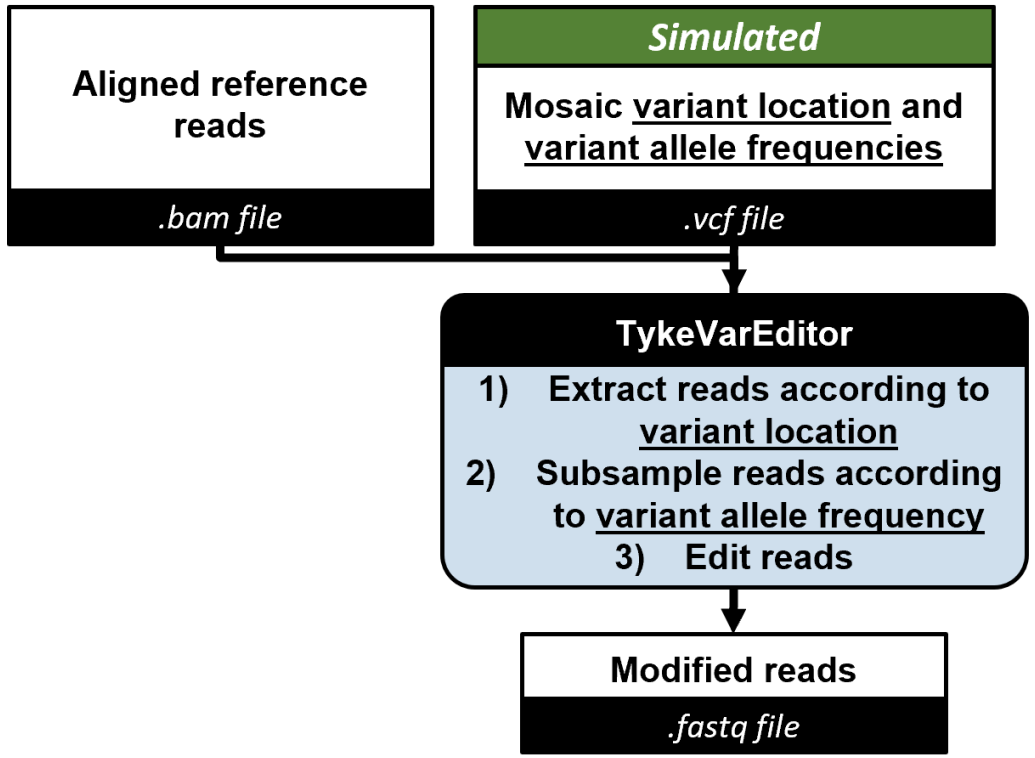
chr22 42134824 HackIns2 N
CCCCAGTGCTCAATCCATTCTCGTTTTTTGGTGGACCTCGGCGTTCGAGTAGGCTTTGATGTACGGTCTTCTCAGTTATAGATGCCTCGTGAGGCCCGCCGTTT>

chr22 43938337 HackIns3 N
CTCTAGAGAGTCGGGACCCGGTCGTGCTATTATCTAAACGCGTAATGCTTACACTGCAATGTACGTGGCTTAGGCATTTCGCGATACTTGAGTACCTAGACCTA>

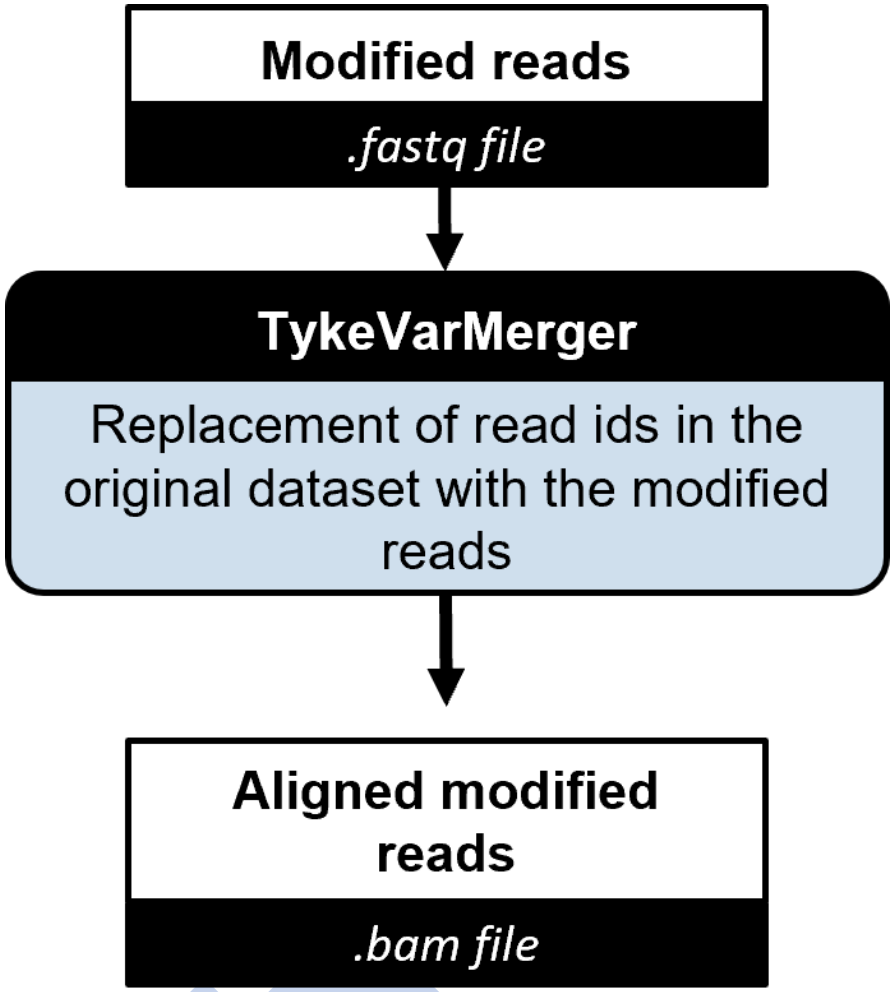
chr22 43366574 HackDel2 N <DEL> 60 PASS AF=0.07;END=43375935;PRECISE;SVLEN=-9361;SVTYPE=DEL GT:GQ 0/0:60

*T*rack in *Y*our *K*ey *E*ndogenous *Var*iants

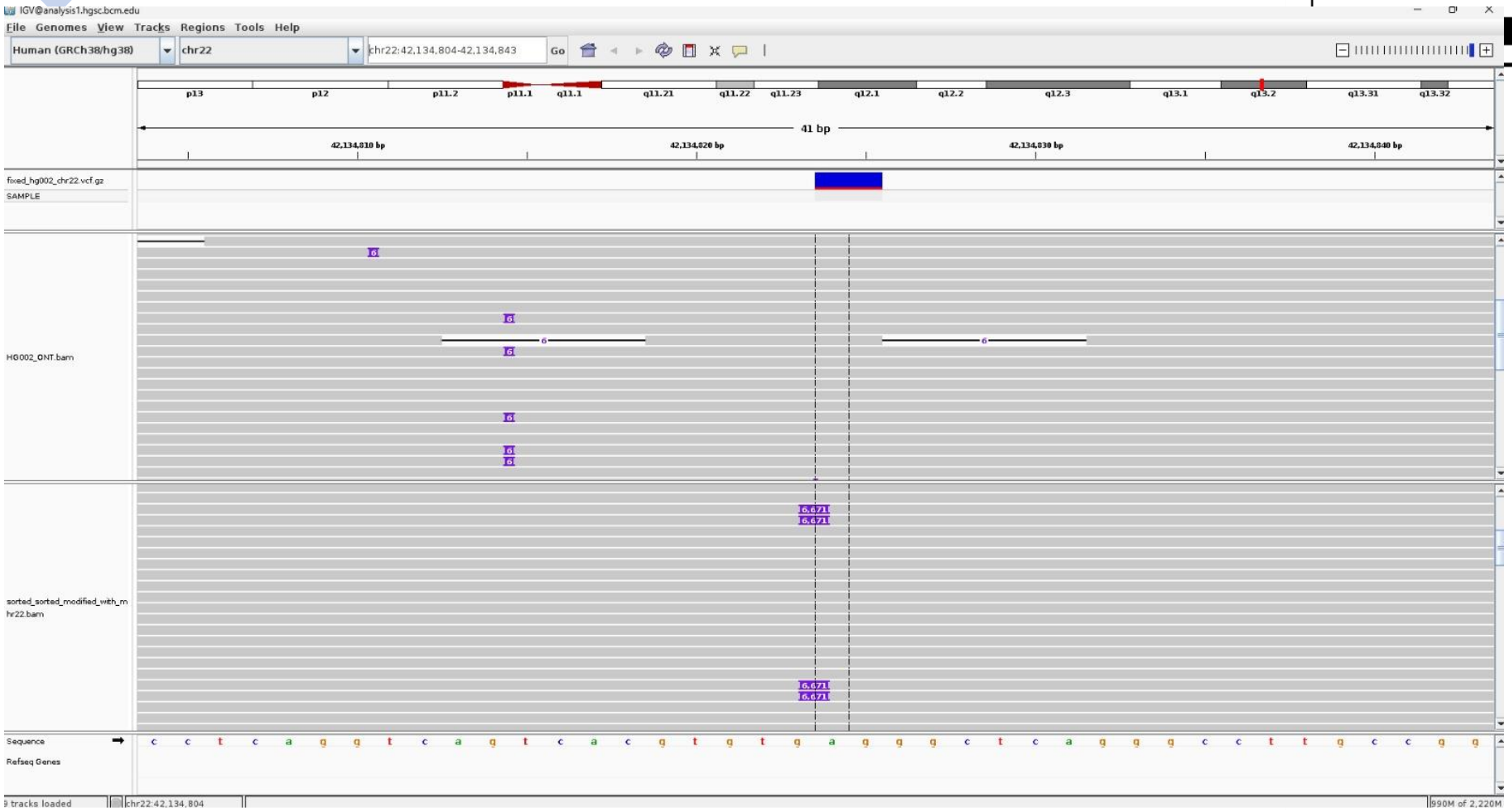
2) TykeVarEditor – Generate Edited Reads Based on Simulated VCF



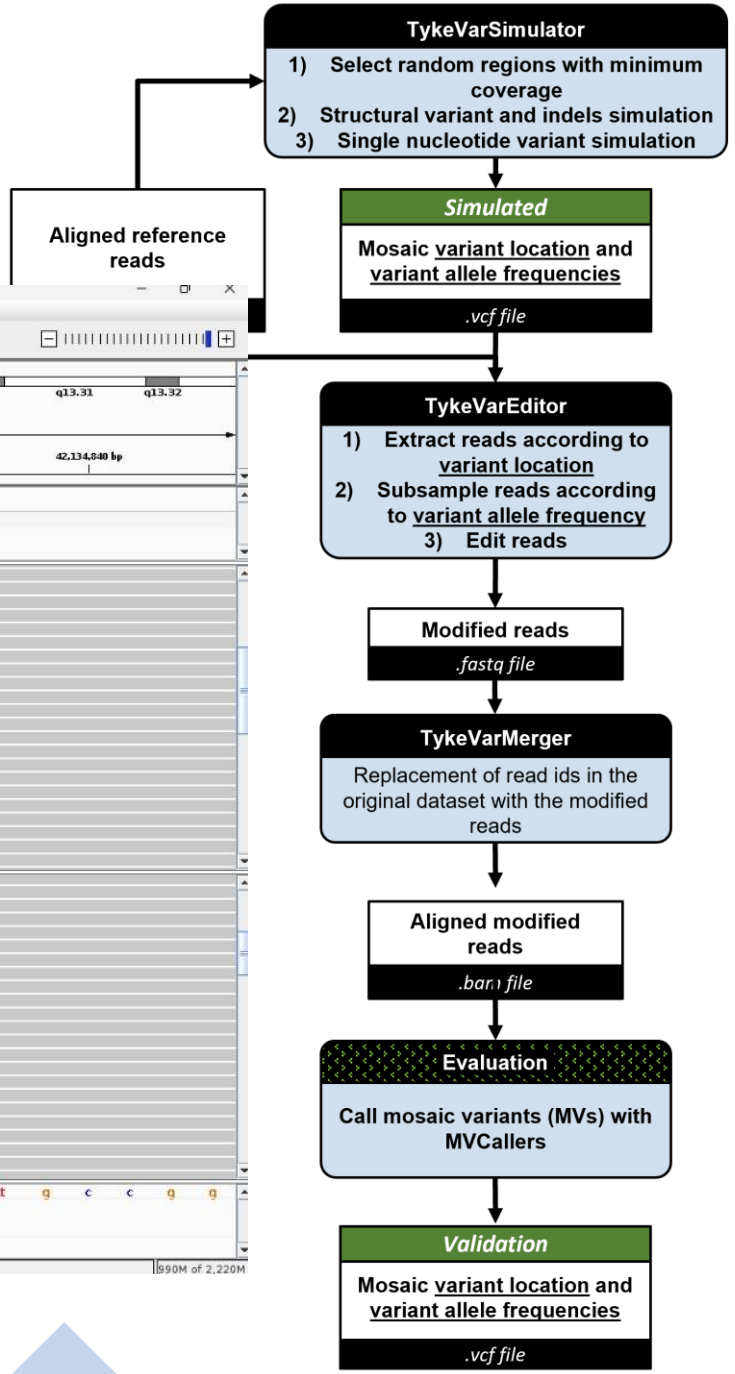
3) TykeVarMerger – Re-Align Modified Reads and Merge Them



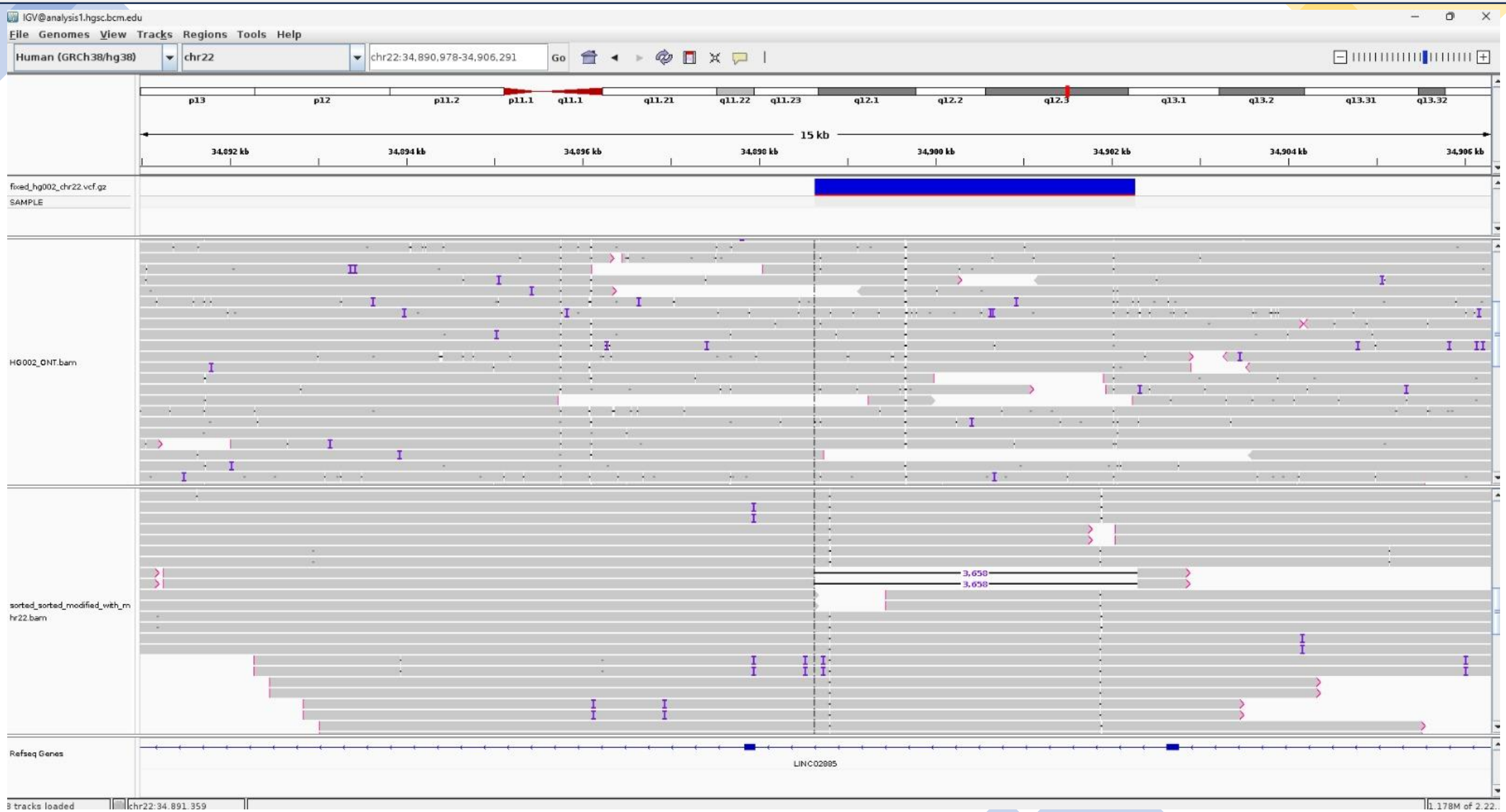
TykeVar – Creation of Sequencing Data With a Subset of Modified Reads



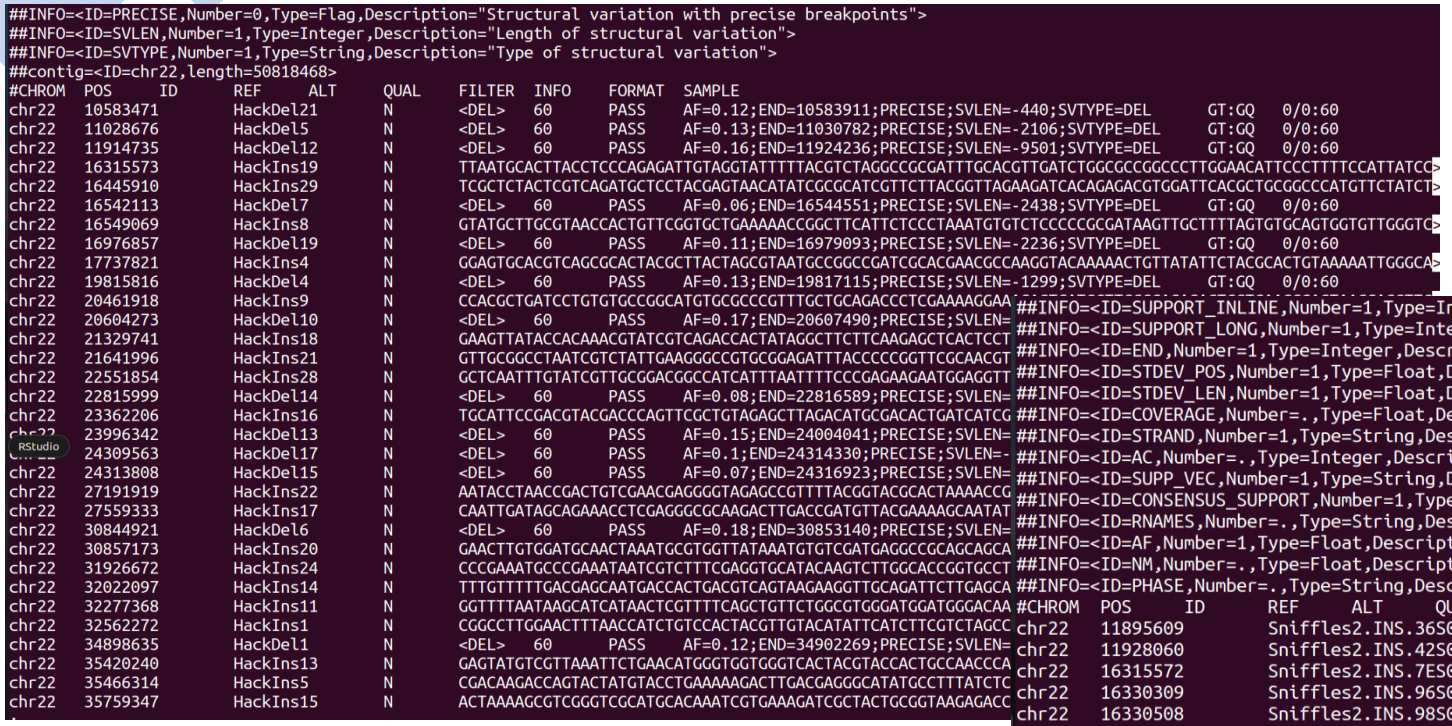
A mosaic SV insertion.



*T*rack in *Y*our *K*ey *E*ndogenous *Var*iants

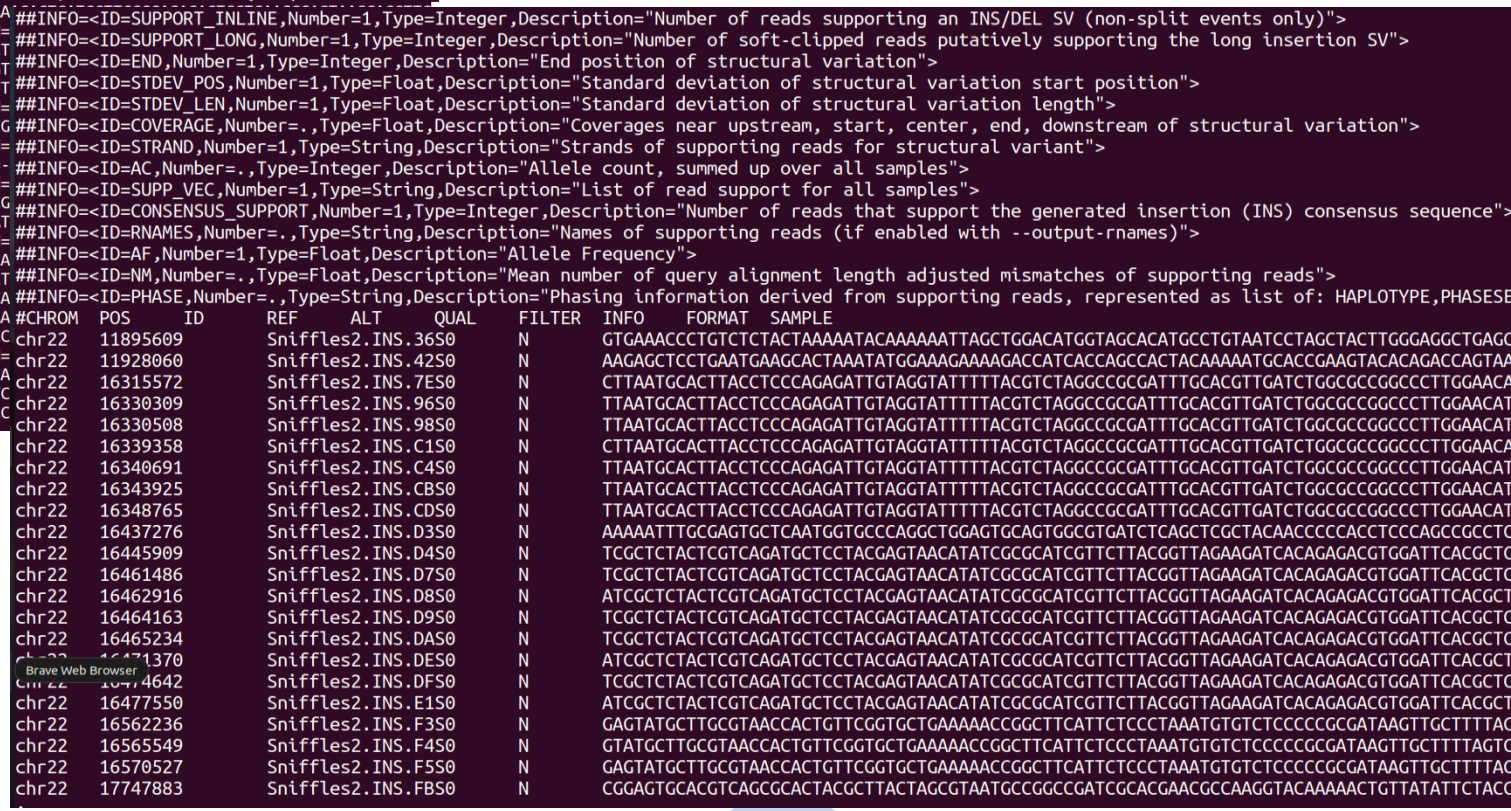


*T*rack in *Y*our *K*ey *E*ndogenous *Var*iants



```
##INFO=<ID=PRECISE,Number=0,Type=Flag,Description="Structural variation with precise breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Length of structural variation">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variation">
##contig=<ID=chr22,length=50818468>
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
chr22   10583471        HackDel21       N       <DEL>   60      PASS    AF=0.12;END=10583911;PRECISE;SVLEN=-440;SVTYPE=DEL      GT:GQ   0/0:60
chr22   11028676        HackDel5        N       <DEL>   60      PASS    AF=0.13;END=11030782;PRECISE;SVLEN=-2106;SVTYPE=DEL     GT:GQ   0/0:60
chr22   11914735        HackDel12       N       <DEL>   60      PASS    AF=0.16;END=11924236;PRECISE;SVLEN=-9501;SVTYPE=DEL     GT:GQ   0/0:60
chr22   16315573        HackIns19       N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACATTCCCTTTTCCATTATCC>
chr22   16445910        HackIns29       N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTGCGGCCCATGTTCTATCT>
chr22   16542113        HackDel7        N       <DEL>   60      PASS    AF=0.06;END=16544551;PRECISE;SVLEN=-2438;SVTYPE=DEL     GT:GQ   0/0:60
chr22   16549069        HackIns8        N       GTATGCTTGCGTAACCACTGTTCGGTGCTGAAAAACCGGCTTCATTCTCCCTAAATGTGTCTCCCCCGCGATAAGTTGCTTTTAGTGTGCAGTGGTGTTGGGTC>
chr22   16976857        HackDel19       N       <DEL>   60      PASS    AF=0.11;END=16979093;PRECISE;SVLEN=-2236;SVTYPE=DEL     GT:GQ   0/0:60
chr22   17737821        HackIns4        N       GGAGTGCACGTCAGCGCACTACGCTTACTAGCGTAATGCCGGCCGATCGCACGAACGCCAAGGTACAAAAACTGTTATATTCTACGCACTGTAAAAATTGGGCA>
chr22   19815816        HackDel4        N       <DEL>   60      PASS    AF=0.13;END=19817115;PRECISE;SVLEN=-1299;SVTYPE=DEL     GT:GQ   0/0:60
chr22   20461918        HackIns9        N       CCACGCTGATCCTGTGTGCCGGCATGTGCGCCCGTTTGCTGCAGACCCTCGAAAAGGAA
chr22   20604273        HackDel10       N       <DEL>   60      PASS    AF=0.17;END=20607490;PRECISE;SVLEN=
chr22   21329741        HackIns18       N       GAAGTTATACCACAAACGTATCGTCAGACCACTATAGGCTTCTTCAAGAGCTCACTCCT
chr22   21641996        HackIns21       N       GTTGCGGCCTAATCGTCTATTGAAGGGCCGTGCGGAGATTTACCCCCGGTTCGCAACGT
chr22   22551854        HackIns28       N       GCTCAATTTGTATCGTTGCGGACGGCCATCATTTAATTTTCCCGAGAAGAATGGAGGTT
chr22   22815999        HackDel14       N       <DEL>   60      PASS    AF=0.08;END=22816589;PRECISE;SVLEN=
chr22   23362206        HackIns16       N       TGCATTCCGACGTACGACCCAGTTCGCTGTAGAGCTTAGACATGCGACACTGATCATCG
chr22   23996342        HackDel13       N       <DEL>   60      PASS    AF=0.15;END=24004041;PRECISE;SVLEN=
chr22   24309563        HackDel17       N       <DEL>   60      PASS    AF=0.1;END=24314330;PRECISE;SVLEN=-
chr22   24313808        HackDel15       N       <DEL>   60      PASS    AF=0.07;END=24316923;PRECISE;SVLEN=
chr22   27191919        HackIns22       N       AATACCTAACCGACTGTCGAACGAGGGGTAGAGCCGTTTTACGGTACGCACTAAAACCG
chr22   27559333        HackIns17       N       CAATTGATAGCAGAAACCTCGAGGGCGCAAGACTTGACCGATGTTACGAAAAGCAATAT
chr22   30844921        HackDel6        N       <DEL>   60      PASS    AF=0.18;END=30853140;PRECISE;SVLEN=
chr22   30857173        HackIns20       N       GAACTTGTGGATGCAACTAAATGCGTGGTTATAAATGTGTCGATGAGGCCGCAGCAGCA
chr22   31926672        HackIns24       N       CCCGAAATGCCCGAAATAATCGTCTTTCGAGGTGCATACAAGTCTTGGCACCGGTGCCT
chr22   32022097        HackIns14       N       TTTGTTTTTGACGAGCAATGACCACTGACGTCAGTAAGAAGGTTGCAGATTCTTGAGCA
chr22   32277368        HackIns11       N       GGTTTTAATAAGCATCATAACTCGTTTTCAGCTGTTCTGGCGTGGGATGGATGGGACAA
chr22   32562272        HackIns1        N       CGGCCTTGGAACTTTAACCATCTGTCCACTACGTTGTACATATTCATCTTCGTCTAGCC
chr22   34898635        HackDel1        N       <DEL>   60      PASS    AF=0.12;END=34902269;PRECISE;SVLEN=
chr22   35420240        HackIns13       N       GAGTATGTCGTTAAATTCTGAACATGGGTGGTGGGTCACTACGTACCACTGCCAACCA
chr22   35466314        HackIns5        N       CGACAAGACCAGTACTATGTACCTGAAAAAGACTTGACGAGGGCATATGCCTTTATCTC
chr22   35759347        HackIns15       N       ACTAAAAGCGTCGGGTCGCATGCACAAATCGTGAAAGATCGCTACTGCGGTAAGAGACC
```

Ground Truth VCF



```
##INFO=<ID=SUPPORT_INLINE,Number=1,Type=Integer,Description="Number of reads supporting an INS/DEL SV (non-split events only)">
##INFO=<ID=SUPPORT_LONG,Number=1,Type=Integer,Description="Number of soft-clipped reads putatively supporting the long insertion SV">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of structural variation">
##INFO=<ID=STDEV_POS,Number=1,Type=Float,Description="Standard deviation of structural variation start position">
##INFO=<ID=STDEV_LEN,Number=1,Type=Float,Description="Standard deviation of structural variation length">
##INFO=<ID=COVERAGE,Number=.,Type=Float,Description="Coverages near upstream, start, center, end, downstream of structural variation">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Strands of supporting reads for structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count, summed up over all samples">
##INFO=<ID=SUPP_VEC,Number=1,Type=String,Description="List of read support for all samples">
##INFO=<ID=CONSENSUS_SUPPORT,Number=1,Type=Integer,Description="Number of reads that support the generated insertion (INS) consensus sequence">
##INFO=<ID=RNAMES,Number=.,Type=String,Description="Names of supporting reads (if enabled with --output-rnames)">
##INFO=<ID=AF,Number=1,Type=Float,Description="Allele Frequency">
##INFO=<ID=NM,Number=.,Type=Float,Description="Mean number of query alignment length adjusted mismatches of supporting reads">
##INFO=<ID=PHASE,Number=.,Type=String,Description="Phasing information derived from supporting reads, represented as list of: HAPLOTYPE,PHASESE
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
chr22   11895609        Sniffles2.INS.36S0      N       GTGAAACCCTGTCTCTACTAAAAATACAAAAAATTAGCTGGACATGGTAGCACATGCCTGTAATCCTAGCTACTTGGGAGGCTGAGG
chr22   11928060        Sniffles2.INS.42S0      N       AAGAGCTCCTGAATGAAGCACTAAATATGGAAAGAAAAGACCATCACCAGCCACTACAAAAATGCACCGAAGTACACAGACCAGTAA
chr22   16315572        Sniffles2.INS.7ES0      N       CTTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACA
chr22   16330309        Sniffles2.INS.96S0      N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACAT
chr22   16330508        Sniffles2.INS.98S0      N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACAT
chr22   16339358        Sniffles2.INS.C1S0      N       CTTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCGGCCCTTGGAACA
chr22   16340691        Sniffles2.INS.C4S0      N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACAT
chr22   16343925        Sniffles2.INS.CBS0      N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACAT
chr22   16348765        Sniffles2.INS.CDS0      N       TTAATGCACTTACCTCCCAGAGATTGTAGGTATTTTTACGTCTAGGCCGCGATTTGCACGTTGATCTGGCGCCCGGCCCTTGGAACAT
chr22   16437276        Sniffles2.INS.D3S0      N       AAAAATTTGCGAGTGCTCAATGGTGCCCAGGCTGGAGTGCAGTGGCGTGATCTCAGCTCGCTACAACCCCCACCTCCCAGCCGCCTG
chr22   16445909        Sniffles2.INS.D4S0      N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTG
chr22   16461486        Sniffles2.INS.D7S0      N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTG
chr22   16462916        Sniffles2.INS.D8S0      N       ATCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCT
chr22   16464163        Sniffles2.INS.D9S0      N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTG
chr22   16465234        Sniffles2.INS.DAS0      N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTG
chr22   16471370        Sniffles2.INS.DES0      N       ATCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCT
chr22   16474642        Sniffles2.INS.DFS0      N       TCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCTG
chr22   16477550        Sniffles2.INS.E1S0      N       ATCGCTCTACTCGTCAGATGCTCCTACGAGTAACATATCGCGCATCGTTCTTACGGTTAGAAGATCACAGAGACGTGGATTCACGCT
chr22   16562236        Sniffles2.INS.F3S0      N       GAGTATGCTTGCGTAACCACTGTTCGGTGCTGAAAAACCGGCTTCATTCTCCCTAAATGTGTCTCCCCCGCGATAAGTTGCTTTTAC
chr22   16565549        Sniffles2.INS.F4S0      N       GTATGCTTGCGTAACCACTGTTCGGTGCTGAAAAACCGGCTTCATTCTCCCTAAATGTGTCTCCCCCGCGATAAGTTGCTTTTAGTG
chr22   16570527        Sniffles2.INS.F5S0      N       GAGTATGCTTGCGTAACCACTGTTCGGTGCTGAAAAACCGGCTTCATTCTCCCTAAATGTGTCTCCCCCGCGATAAGTTGCTTTTAC
chr22   17747883        Sniffles2.INS.FBS0      N       CGGAGTGCACGTCAGCGCACTACGCTTACTAGCGTAATGCCGGCCGATCGCACGAACGCCAAGGTACAAAAACTGTTATATTCTACG
```

Sniffles Mosaic Results

Acknowledgement

Baylor College of Medicine

HGSC
HUMAN GENOME SEQUENCING CENTER

Sedlazeck lab

- Fritz J Sedlazeck
- Luis Paulin
- Xinchang Zheng
- Michal Izydorczyk
- Philippe Sanio
- Ryan Cheng
- Divya Kalra
- Farhang Jaryani