CLAIMS

1. A system comprising:

at least one processor; and

a non-transitory computer readable medium comprising instructions that, when executed by the at least one processor, cause the system to:

determine, for one or more genomic coordinates of a genomic sample, an initial structural variant call based on nucleotide reads corresponding to the genomic sample;

identify one or more variant region quality sequencing metrics and other sequencing metrics corresponding to one or more of the initial structural variant call or the one or more genomic coordinates, the one or more variant region quality sequencing metrics indicating a degree to which nucleotide base calls from the nucleotide reads satisfy one or more read quality thresholds or base call quality thresholds at the one or more genomic coordinates;

generate, utilizing a structural variant refinement machine-learning model to process one or more variant region quality sequencing metrics and the other sequencing metrics, a false positive likelihood indicating a likelihood that the initial structural variant call is a false positive; and

determine a modified structural variant call for the one or more genomic coordinates of the genomic sample based on the false positive likelihood.

2. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to determine the initial structural variant call by determining a deletion of more than a threshold number of base pairs, an insertion of more than the threshold number of base pairs, a duplication of more than the threshold number of base pairs, an inversion, a translocation, or a copy number variation (CNV).

3. The system of claim 1 or 2, further comprising instructions that, when executed by the at least one processor, cause the system to determine the initial structural variant call by determining a structural variant call of a number of base pairs within a threshold range of base pairs.

4. The system of any one of claims 1-3, further comprising instructions that, when executed by the at least one processor, cause the system to identify the other sequencing metrics corresponding to the initial structural variant call by identifying one or more of read-based sequencing metrics or reference-based sequencing metrics.

5. The system of claim 4, further comprising instructions that, when executed by the at least one processor, cause the system to identify the read-based sequencing metrics by determining, for the initial structural variant call, one or more of:

one or more base-call quality scores;

a fraction of nucleotide reads supporting an alternate contiguous sequence from a reference genome;

a number of split nucleotide reads from the nucleotide reads corresponding to the initial structural variant call;

a coverage depth of the nucleotide reads corresponding to the initial structural variant call;

an additional structural variant call located within a threshold number of base pairs from the initial structural variant call within the genomic sample;

an alignment of a contiguous sequence corresponding to the nucleotide reads with a reference sequence of a reference genome modified to include a structural variant corresponding to the initial structural variant call;

a deletion length in nucleotide bases based on one or more soft clipped nucleotide reads;

a number of the nucleotide reads exhibiting a mapping quality metric that fails to satisfy a threshold mapping quality metric;

an insert size representing a length of nucleotide-read fragments corresponding to the initial structural variant call; or

a structural-variant likelihood representing a ratio of the initial structural variant call to a reference call for the one or more genomic coordinates based on the insert size.

6. The system of claim 4, further comprising instructions that, when executed by the at least one processor, cause the system to identify the one or more variant region quality sequencing metrics by determining one or more of:

a number of nucleotide reads that comprise at least a threshold number of base calls and correspond to a target genomic region for the initial structural variant call; or

a number of nucleotide bases in an alternate contiguous sequence corresponding to the target genomic region from a reference genome for which base calls for the nucleotide reads fail to satisfy a threshold base call quality score.

7. The system of claim 4, further comprising instructions that, when executed by the at least one processor, cause the system to identify the reference-based sequencing metrics by identifying, within one or more genomic regions of a reference genome corresponding to the one or more genomic coordinates of the genomic sample, one or more of:

a tandem repeat length in nucleotide bases;

a permutation entropy of nucleotide bases;

a cytosine quadruplex (C-quadruplex); or

a guanine quadruplex (G-quadruplex).

8. The system of any one of claims 1-7, further comprising instructions that, when executed by the at least one processor, cause the system to:

generate the false positive likelihood by determining the initial structural variant call is a false positive call or a true positive call based on the one or more variant region quality sequencing metrics or the other sequencing metrics;

determine the modified structural variant call by:

changing the initial structural variant call from a positive structural variant call to a negative structural variant call based on the initial structural variant call being the false positive call; or

changing the initial structural variant call from a negative structural variant call to a positive structural variant call based on the initial structural variant call being the true positive call; and

generate a variant call file comprising the modified structural variant call by modifying data fields for one or more metrics associated with the initial structural variant call.

9. A computer-implemented method comprising:

determining, for one or more genomic coordinates of a genomic sample, an initial structural variant call based on nucleotide reads corresponding to the genomic sample;

identifying one or more variant region quality sequencing metrics and other sequencing metrics corresponding to one or more of the initial structural variant call or the one or more genomic coordinates, the one or more variant region quality sequencing metrics indicating a degree to which nucleotide base calls from the nucleotide reads satisfy one or more read quality thresholds or base call quality thresholds at the one or more genomic coordinates;

generating, utilizing a structural variant refinement machine-learning model to process the one or more variant region quality sequencing metrics and the other sequencing metrics, a false positive likelihood indicating a likelihood that the initial structural variant call is a false positive; and

determining a modified structural variant call for the one or more genomic coordinates of the genomic sample based on the false positive likelihood.

10. The computer-implemented method of claim 9, wherein:

determining the initial structural variant call comprises utilizing a call generation model to determine base calls corresponding to the one or more genomic coordinates of the genomic sample indicate a structural variant in relation to a reference genome; and

determining the modified structural variant call comprises correcting the initial structural variant call for the one or more genomic coordinates based on the false positive likelihood generated by the structural variant refinement machine-learning model.

11. The computer-implemented method of claim 9 or 10, wherein identifying the other sequencing metrics corresponding to the initial structural variant call comprises identifying one or more of read-based sequencing metrics or reference-based sequencing metrics.

12. The computer-implemented method of any one of claims 9-11, wherein identifying the other sequencing metrics comprises determining, for the initial structural variant call, one or more of:

one or more base-call quality scores;

a fraction of nucleotide reads supporting an alternate contiguous sequence from a reference genome;

a number of split nucleotide reads from the nucleotide reads corresponding to the initial structural variant call;

a coverage depth of the nucleotide reads corresponding to the initial structural variant call;

an additional structural variant call located within a threshold number of base pairs from the initial structural variant call within the genomic sample;

an alignment of a contiguous sequence corresponding to the nucleotide reads with a reference sequence of a reference genome modified to include a structural variant corresponding to the initial structural variant call;

a deletion length in nucleotide bases based on one or more soft clipped nucleotide reads;

a number of the nucleotide reads exhibiting a mapping quality metric that fails to satisfy a threshold mapping quality metric;

an insert size representing a length of nucleotide-read fragments corresponding to the initial structural variant call; or

a structural-variant likelihood representing a ratio of the initial structural variant call to a reference call for the one or more genomic coordinates based on the insert size.

13. The computer-implemented method of any one of claims 9-12, wherein identifying the one or more variant region quality sequencing metrics comprises determining one or more of:

a number of nucleotide reads that comprise at least a threshold number of base calls and correspond to a target genomic region for the initial structural variant call; or

a number of nucleotide bases in an alternate contiguous sequence corresponding to the target genomic region from a reference genome for which based calls for the nucleotide reads fail to satisfy a threshold base call quality score.

14. The computer-implemented method of any one of claims 9-13, wherein identifying the other sequencing metrics comprises identifying, within one or more genomic regions of a reference genome corresponding to the one or more genomic coordinates of the genomic sample, one or more of:

a tandem repeat length in nucleotide bases;

a permutation entropy of nucleotide bases;

a cytosine quadruplex (C-quadruplex); or

a guanine quadruplex (G-quadruplex).

15. A non-transitory computer readable medium comprising instructions that, when executed by at least one processor, cause a computing device to:

determine, for one or more genomic coordinates of a genomic sample, an initial structural variant call based on nucleotide reads corresponding to the genomic sample;

identify one or more variant region quality sequencing metrics and other sequencing metrics corresponding to one or more of the initial structural variant call or the one or more genomic coordinates, the one or more variant region quality sequencing metrics indicating a degree to which nucleotide base calls from the nucleotide reads satisfy one or more read quality thresholds or base call quality thresholds at the one or more genomic coordinates;

generate, utilizing a structural variant refinement machine-learning model to process the one or more variant region quality sequencing metrics and the other sequencing metrics, a false positive likelihood indicating a likelihood that the initial structural variant call is a false positive; and

determine a modified structural variant call for the one or more genomic coordinates of the genomic sample based on the false positive likelihood.

16. The non-transitory computer readable medium of claim 15, wherein the structural variant refinement machine-learning model comprises one or more gradient boosted decision trees.

17. The non-transitory computer readable medium of claim 15 or 16, further comprising instructions that, when executed by the at least one processor, cause the computing device to:

generate the false positive likelihood by determining the initial structural variant call is a false positive call or a true positive call based on the one or more variant region quality sequencing metrics or the other sequencing metrics;

determine the modified structural variant call by:

changing the initial structural variant call from a positive structural variant call to a negative structural variant call based on the initial structural variant call being the false positive call; or

changing the initial structural variant call from a negative structural variant call to a positive structural variant call based on the initial structural variant call being the true positive call; and

generate a variant call file comprising the modified structural variant call by modifying data fields for one or more metrics associated with the initial structural variant call.

18. The non-transitory computer readable medium of any one of claims 15-17, further comprising instructions that, when executed by the at least one processor, cause the computing device to:

determine, from a truth dataset, a ground truth structural variant call corresponding to the modified structural variant call is incorrectly labeled as a false positive instead of a true positive based on one or more truth set nucleotide reads for the ground truth structural variant call satisfying structural variant criteria;

change a label for the ground truth structural variant call from false positive to true positive; and

adjust parameters of the structural variant refinement machine-learning model based on a comparison of the modified structural variant call and the ground truth structural variant call.

19. The non-transitory computer readable medium of claim 18, further comprising instructions that, when executed by the at least one processor, cause the computing device to determine the ground truth structural variant call is incorrectly labeled based on the structural variant criteria by:

parsing a Concise Idiosyncratic Gapped Alignment Report (CIGAR) string to identify a truth set nucleotide read of the truth dataset that satisfies a threshold mapping quality metric;

determining a portion of the CIGAR string comprising a starting index of a corresponding structural variant call generated by a call generation model; and

determining that the starting index corresponds to a structural variant and matches a length of the corresponding structural variant call generated by the call generation model.

20. The non-transitory computer readable medium of any one of claims 15-19, further comprising instructions that, when executed by the at least one processor, cause the computing device to generate the false positive likelihood utilizing the structural variant refinement machine-learning model based on the one or more variant region quality sequencing metrics or the other sequencing metrics and the initial structural variant call as inputs.