CLAIMS

1. A computer-implemented method comprising:

identifying one or more paired-end nucleotide reads corresponding to a genomic region of a genomic sample;

determining candidate split groups comprising fragment alignments of the one or more paired-end nucleotide reads;

identifying, from the candidate split groups, candidate pairs of split groups comprising different fragment alignments for mates of a paired-end nucleotide read of the one or more paired-end nucleotide reads;

generating split group scores for split alignments of the candidate split groups, wherein a split group score of the split group scores measures an accuracy of fragment alignments in a split group with respect to a reference genome;

generating, for the candidate pairs of split groups and based on the split group scores, pair scores evaluating pair alignments of the candidate pairs of split groups with the reference genome; and

selecting, for nucleobase calling of the genomic region, a predicted split group from the candidate split groups based on the pair scores.

2. The computer-implemented method of claim 1, further comprising determining a candidate split group of the candidate split groups by grouping, into the candidate split group, one or more fragment alignments of a paired-end nucleotide read from a pair of paired-end nucleotide reads of the one or more paired-end nucleotide reads.

3. The computer-implemented method of claim 1 or 2, further comprising:

generating fragment alignment scores for individual fragment alignments of a candidate split group with the reference genome, wherein a fragment alignment score of the fragment alignment scores measures an accuracy of a fragment alignment with respect to the reference genome; and

generating a split group score for the candidate split group based on the fragment alignment scores.

4. The computer-implemented method of any one of claims 1-3, further comprising:

generating, for a candidate split group of the candidate split groups, a penalty for relative geometries of a first fragment alignment of a first alignment orientation with respect to the reference genome and a second fragment alignment of a second alignment orientation with respect to the reference genome; and

generating a split group score for the candidate split group based on the penalty for relative geometries of the first fragment alignment and the second fragment alignment.

5. The computer-implemented method of any one of claims 1-4, further comprising:

generating, for a candidate split group of the candidate split groups, an overlap penalty for an overlap within a nucleotide read between a first fragment alignment and a second fragment alignment; and

generating a split group score for the candidate split group based on the overlap penalty.

6. The computer-implemented method of any one of claims 1-5, further comprising generating a split group score for a candidate split group of the candidate split groups by:

generating fragment alignment scores, a penalty for relative geometries, and an overlap penalty for fragment alignments of the candidate split group; and

combining the fragment alignment scores and subtracting the penalty for relative geometries and the overlap penalty from the combined fragment alignment scores.

7. The computer-implemented method of any one of claims 1-6, further comprising:

determining the candidate split groups by iteratively grouping possible fragment alignment sequences following an order of outermost fragment alignments to innermost fragment alignments of a nucleotide read; and

generating the split group scores by iteratively scoring groupings of possible fragment alignment sequences following the order in which the possible fragment alignment sequences were grouped.

8. The computer-implemented method of any one of claims 1-7, further comprising selecting the predicted split group from the candidate split groups by:

selecting, from the candidate pairs of split groups, a pair of candidate split groups having a highest pair score; and

selecting, for each mate of a nucleotide-read pair, the predicted split group from the pair of candidate split groups.

9. The computer-implemented method of claim 8, further comprising:

determining sums of split group scores for respective candidate pairs of split groups;

generating pairing penalties based on an estimated insert size between innermost fragment alignments of the candidate pairs of split groups; and

generating the pair scores for the candidate pairs of split groups based on the sums of split group scores and the pairing penalties.

10. The computer-implemented method of claim 8, further comprising:

determining an alt-contig fragment alignment score for an inner fragment alignment and an outer fragment alignment corresponding to a nucleotide read with an alternate contiguous sequence within the reference genome;

determining a split group score for the inner fragment alignment and the outer fragment alignment with a primary-assembly region of the reference genome; and

selecting the alt-contig fragment alignment score as a replacement split group score based on determining that the alt-contig fragment alignment score exceeds the split group score.

11. 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:

identify one or more paired-end nucleotide reads corresponding to a genomic region of a genomic sample;

determine candidate split groups comprising fragment alignments of the one or more paired-end nucleotide reads;

identify, from the candidate split groups, candidate pairs of split groups comprising different fragment alignments for mates of a paired-end nucleotide read of the one or more paired-end nucleotide reads;

generate split group scores for split alignments of the candidate split groups with a reference genome, wherein a split group score of the split group scores measures an accuracy of fragment alignments in a split group with respect to a reference genome;

generate, for the candidate pairs of split groups and based on the split group scores, pair scores evaluating pair and alignments of the candidate pairs of split groups with the reference genome; and

select, for nucleobase calling of the genomic region, a predicted split group from the candidate split groups based on the pair scores.

12. The system of claim 11, further comprising instructions that, when executed by the at least one processor, cause the system to:

generate, for a candidate split group of the candidate split groups, an overlap penalty for an overlap within a nucleotide read between a first fragment alignment and a second fragment alignment; and

generate a split group score for the candidate split group based on the overlap penalty.

13. The system of claim 11 or 12, further comprising instructions that, when executed by the at least one processor, cause the system to generate a split group score for a candidate split group of the candidate split groups by:

generating fragment alignment scores, a penalty for relative geometries, and an overlap penalty for fragment alignments of the candidate split group; and

combining the fragment alignment scores and subtracting the penalty for relative geometries and the overlap penalty from the combined fragment alignment scores.

14. The system of any one of claims 11-13, further comprising instructions that, when executed by the at least one processor, cause the system to select the predicted split group from the candidate split groups by:

selecting, from the candidate pairs of split groups, a pair of candidate split groups having a highest pair score; and

selecting, for each mate of a nucleotide-read pair, the predicted split group from the pair of candidate split groups.

15. The system of any one of claims 14, further comprising instructions that, when executed by the at least one processor, cause the system to:

determine an alt-contig fragment alignment score for an inner fragment alignment and an outer fragment alignment corresponding to a nucleotide read with an alternate contiguous sequence within the reference genome;

determine a split group score for the inner fragment alignment and the outer fragment alignment with a primary-assembly region of the reference genome; and

select the alt-contig fragment alignment score as a replacement split group score based on determining that the alt-contig fragment alignment score exceeds the split group score.

16. The system of any one of claims 11-15, further comprising instructions that, when executed by the at least one processor, cause the system to:

determine the candidate split groups by iteratively grouping possible fragment alignment sequences following an order of outermost fragment alignments to innermost fragment alignments of a nucleotide read; and

generate the split group scores by iteratively scoring groupings of possible fragment alignment sequences following the order in which the possible fragment alignment sequences were grouped.

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

identify one or more paired-end nucleotide reads corresponding to a genomic region of a genomic sample;

determine candidate split groups comprising fragment alignments of the one or more paired-end nucleotide reads;

identify, from the candidate split groups, candidate pairs of split groups comprising different fragment alignments for mates of a paired-end nucleotide read of the one or more paired-end nucleotide reads;

generate split group scores for split alignments of the candidate split groups, wherein a split group score of the split group scores measures an accuracy of fragment alignments in a split group with respect to a reference genome;

generate, for the candidate pairs of split groups and based on the split group scores, pair scores evaluating pair alignments of the candidate pairs of split groups with the reference genome; and

select, for nucleobase calling of the genomic region, a predicted split group from the candidate split groups based on the pair scores.

18. The non-transitory computer-readable medium of claim 17, further comprising instructions that, when executed by the at least one processor, cause the computing device to determine a candidate split group of the candidate split groups by grouping, into the candidate split group, one or more fragment alignments of a paired-end nucleotide read from a pair of paired-end nucleotide reads of the one or more paired-end nucleotide reads.

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

generate fragment alignment scores for individual fragment alignments of a candidate split group with the reference genome, wherein a fragment alignment score of the fragment alignment scores measures an accuracy of a fragment alignment with respect to the reference genome; and

generate a split group score for the candidate split group based on the fragment alignment scores.

20. The non-transitory computer-readable medium of any one of claims 17-19, further comprising instructions that, when executed by the at least one processor, cause the computing device to:

generate, for a candidate split group of the candidate split groups, a penalty for relative geometries of a first fragment alignment of a first alignment orientation with respect to the reference genome and a second fragment alignment of a second alignment orientation with respect to the reference genome; and

generate a split group score for the candidate split group based on the penalty for relative geometries of the first fragment alignment and the second fragment alignment.