

CLAIMS

What is claimed is:

1. A computer-implemented method for sequencing data read realignment, the method comprising:

obtaining from a sequence alignment dataset an initial alignment of a read sequence to a reference sequence, the initial alignment comprising an aligned read;

performing realignment processing on the initial alignment, the realignment processing realigning the read sequence to the reference sequence to produce one or more candidate realignments that are different from the initial alignment, and the realignment processing comprising:

identifying one or more candidate indels, the one or more candidate indels comprising at least one indels in the aligned read and at least one indels aligned proximal to the aligned read as indicated by the sequence alignment dataset;

creating a flattened aligned read based at least on removing from the aligned read any indels indicated by the initial alignment; and

determining the one or more candidate realignments of the read sequence to the reference sequence based on introducing, for each candidate realignment of the one or more candidate realignments, a respective at least one candidate indel of the one or more candidate indels into the flattened aligned read; and

providing the initial alignment or a selected candidate realignment of the one or more candidate realignments based on one or more selection criteria, wherein the one or more candidate indels comprise a plurality of candidate indels, and wherein the determining the one or more candidate realignments comprises commencing iteratively introducing the plurality of candidate indels into the flattened aligned read, wherein each iteration of the iteratively introducing provides a candidate realignment

of the one or more candidate realignments by introducing into the flattened aligned read the respective at least one candidate indel for the candidate realignment.

2. The method of claim 1, wherein the iteratively introducing introduces permutations of one or more candidate indels of the plurality of candidate indels into the flattened aligned read to obtain, for each permutation of the permutations, a different candidate realignment of the one or more candidate alignments.

3. The method of claim 2, wherein the realignment processing further comprises:

checking a provided candidate realignment of the one or more candidate realignments to determine whether an aligned read of the provided candidate realignment, the aligned read of the provided candidate realignment having the introduced respective one or more candidate indels, aligns with the reference sequence with no bases mismatched as between the aligned read of the provided candidate realignment and the reference sequence;

halting the iteratively introducing based on determining that the aligned read of the provided candidate realignment aligns with the reference sequence with no bases mismatched; and

selecting the provided candidate realignment as the selected candidate realignment, wherein the providing outputs the selected candidate realignment based on the aligned read of the provided candidate realignment aligning with the reference sequence.

4. The method of claim 1, wherein the realignment processing further comprises prioritizing the plurality of indels for the iteratively introducing, wherein the iteratively introducing introduces the plurality of indels in order of priority based on the prioritizing.

5. The method of claim 4, wherein the prioritizing prioritizes an indel indicated by a reference indel dataset to be a prior known indel over an indel not indicated by the reference indel dataset to be a prior known indel.

6. The method of claim 4, wherein the prioritizing prioritizes an indel of longer length over an indel of shorter length.

7. The method of claim 4, wherein the prioritizing prioritizes an indel indicated in a greater number of aligned reads of the sequence alignment dataset over an indel indicated in a lesser number of aligned reads of the sequence alignment dataset, or the prioritizing prioritizes an indel indicated in a greater proportion of aligned reads of the sequence alignment dataset that correspond to a location of the indel relative to the reference sequence over an indel indicated in a lesser proportion of aligned reads of the sequence alignment dataset.

8. The method of claim 4, wherein the prioritizing prioritizes, as between different indels indicated in a same number of aligned reads of the sequence alignment dataset, an indel whose location relative to a reference genomic sequence indicated by the sequence alignment dataset is upstream from a location, relative to the reference genomic sequence, indicated for another indel.

9. The method of claim 1, wherein the selection criteria is based at least in part on one or more of: number of bases mismatched, number of indels, location of indels relative to a reference genomic sequence indicated by the sequence alignment dataset, and number of softclipped bases.

10. The method of claim 1, 2, 3, 4 or 9, wherein the selection criteria prioritizes one or more of:

an alignment having no indels and only a single base mismatched over an alignment having one or more indels for the providing;

an alignment having a lesser number of bases mismatched over an alignment having a greater number of bases mismatched for the providing;

as between different alignments having a same number of bases mismatched, an alignment having a lesser number of softclips of a specified type over an alignment having a greater number of softclips of the specified type for the providing; and

as between different alignments having a same number of bases mismatched, an alignment having a lesser number of indels over an alignment having a greater number of indels for the providing.

11. The method of claim 1, 2, 3, 4 or 9, wherein the realignment processing further comprises selecting a best candidate realignment of the one or more candidate realignments based on a first criteria of the one or more selection criteria, wherein the selected candidate realignment is the selected best candidate realignment, and wherein the outputting selects between the initial alignment and the best realignment candidate based on a second criteria of the one or more selection criteria.

12. The method of claim 1, 2, 3, 4 or 9, further comprising determining whether the obtained initial alignment is eligible for realignment, the determining being based at least in part on one or more of:

identifying whether there are one or more bases mismatched as between the aligned read of the initial alignment and the reference sequence;

identifying whether the aligned read comprises a softclip;

identifying whether the initial alignment is not a secondary alignment; and

identifying whether there are candidate indels around the aligned read in a region of bases of a reference genomic sequence of the sequence alignment dataset.

13. The method of claim 1, 2, 3, 4 or 9, further comprising:

determining whether the obtained initial alignment is eligible for realignment, and performing the realignment processing and the providing the initial alignment or selected candidate realignment based on determining that the obtained initial alignment is eligible for realignment;

repeating, for each additional initial alignment of one or more additional initial alignments of the sequence alignment dataset, the obtaining and the determining whether the obtained additional initial alignment is eligible for realignment; and

performing processing for each additional initial alignment of the one or more additional initial alignments, the performing processing comprising (i) providing the additional initial alignment as is, absent performing the realignment processing, or (ii) performing the realignment processing and the providing the additional initial alignment or selected candidate realignment.

14. A computer system for sequencing data read realignment, the computer system comprising memory and at least one processor, the computer system configured to execute program instructions to perform a method comprising:

obtaining from a sequence alignment dataset an initial alignment of a read sequence to a reference sequence, the initial alignment comprising an aligned read;

performing realignment processing on the initial alignment, the realignment processing realigning the read sequence to the reference sequence to produce one or more candidate realignments, and the realignment processing comprising:

identifying one or more candidate indels, the one or more candidate indels comprising zero or more indels in the aligned read and zero or more indels aligned proximal to the aligned read as indicated by the sequence alignment dataset;

creating a flattened aligned read based at least on removing from the aligned read any indels indicated by the initial alignment; and

determining the one or more candidate realignments of the read sequence to the reference sequence based on introducing, for each candidate realignment of the one or more candidate realignments, a respective at least one candidate indel of the one or more candidate indels into the flattened aligned read; and

providing the initial alignment or a selected candidate realignment of the one or more candidate realignments based on one or more selection criteria, wherein the one or more candidate indels comprise a plurality of candidate indels, and wherein the determining the one or more candidate realignments comprises commencing iteratively introducing the plurality of candidate indels into the flattened aligned read, wherein each iteration of the iteratively introducing provides a candidate realignment of the one or more candidate realignments by introducing into the flattened aligned read the respective at least one candidate indel for the candidate realignment.

15. The computer system of claim 14, wherein the iteratively introducing introduces permutations of one or more candidate indels of the plurality of candidate indels

into the flattened aligned read to obtain, for each permutation of the permutations, a different candidate realignment of the one or more candidate alignments.

16. The computer system of claim 15, wherein the realignment processing further comprises:

checking a provided candidate realignment of the one or more candidate realignments to determine whether an aligned read of the provided candidate realignment, the aligned read of the provided candidate realignment having the introduced respective one or more candidate indels, aligns with the reference sequence with no bases mismatched as between the aligned read of the provided candidate realignment and the reference sequence;

halting the iteratively introducing based on determining that the aligned read of the provided candidate realignment aligns with the reference sequence with no bases mismatched; and

selecting the provided candidate realignment as the selected candidate realignment, wherein the providing outputs the selected candidate realignment based on the aligned read of the provided candidate realignment aligning with the reference sequence.

17. The computer system of claim 15 or 16, wherein the realignment processing further comprises prioritizing the plurality of indels for the iteratively introducing, wherein the iteratively introducing introduces the plurality of indels in order of priority based on the prioritizing.

18. A computer program product for sequencing data read realignment, the computer program product comprising:

a tangible storage medium storing program instructions for execution to perform a method comprising:

obtaining from a sequence alignment dataset an initial alignment of a read sequence to a reference sequence, the initial alignment comprising an aligned read;

performing realignment processing on the initial alignment, the realignment processing realigning the read sequence to the reference sequence to produce one or more candidate realignments, and the realignment processing comprising:

identifying one or more candidate indels, the one or more candidate indels comprising zero or more indels in the aligned read and zero or more indels aligned proximal to the aligned read as indicated by the sequence alignment dataset;

creating a flattened aligned read based at least on removing from the aligned read any indels indicated by the initial alignment; and

determining one or more candidate realignments of the read sequence to the reference sequence based on introducing, for each candidate realignment of the one or more candidate realignments, a respective at least one candidate indel of the one or more candidate indels into the flattened aligned read; and

providing the initial alignment or a selected candidate realignment of the one or more candidate realignments based on one or more selection criteria, wherein the one or more candidate indels comprise a plurality of candidate indels, and wherein the determining the one or more candidate realignments comprises commencing iteratively introducing the plurality of candidate indels into the flattened aligned read, wherein each iteration of the iteratively introducing provides a candidate realignment of the one or more candidate realignments by introducing into the flattened aligned read the respective at least one candidate indel for the candidate realignment.

19. The computer program product of claim 18, wherein the iteratively introducing introduces permutations of one or more candidate indels of the plurality of candidate indels into the flattened aligned read to obtain, for each permutation of the permutations, a different candidate realignment of the one or more candidate alignments.

20. The computer program product of claim 19, wherein the realignment processing further comprises:

checking a provided candidate realignment of the one or more candidate realignments to determine whether an aligned read of the provided candidate realignment, the aligned read of the provided candidate realignment having the introduced respective one or more candidate indels, aligns with the reference sequence with no bases mismatched as between the aligned read of the provided candidate realignment and the reference sequence;

halting the iteratively introducing based on determining that the aligned read of the provided candidate realignment aligns with the reference sequence with no bases mismatched; and

selecting the provided candidate realignment as the selected candidate realignment, wherein the providing outputs the selected candidate realignment based on the aligned read of the provided candidate realignment aligning with the reference sequence.

21. The computer program product of claim 19 or 20, wherein the realignment processing further comprises prioritizing the plurality of indels for the iteratively introducing, wherein the iteratively introducing introduces the plurality of indels in order of priority based on the prioritizing.