CLAIMS

1. A method for grouping sequence reads comprising:

under control of a hardware processor:

receiving a plurality of sequence reads each comprising a fragment sequence and a unique molecular identifier (UMI) sequence;

aligning sequence reads of the plurality of sequence reads to a reference sequence using the fragment sequences of the sequence reads;

grouping sequence reads of the plurality of sequence reads into a plurality of families of sequence reads based on the UMI sequences and positions of the fragment sequences of the sequence reads aligned to the reference sequence;

performing UMI statistic estimation of the plurality of families; and

performing probability-based merging of families of the plurality of families using results of the UMI statistic estimation,

wherein performing UMI statistic estimation comprises: performing UMI statistic estimation on a subset of families of the plurality of families, and wherein the subset of families comprises at least 50,000 families of the plurality of families and/or at least 10% of families of the plurality of families and determining fragment size frequency, UMI jumping rate, and/or UMI frequency; and

wherein performing probability-based merging comprises: performing probability-based merging of families of the plurality of families using fragment size frequency, UMI jumping rate, and/or UMI frequency

and determining a relative likelihood of the two families are derived from the same original nucleic acid molecule using the fragment size frequency, the UMI jumping rate, and/or the UMI frequency; and

merging the two families of the plurality of families when the relative likelihood is above a merging threshold; and

wherein determining the relative likelihood of the two families are derived from the same original nucleic acid molecule comprises:

determining a likelihood ratio of unique molecule over non-unique molecule given fragment positions; and

determining a likelihood ratio of UMI transition for unique molecule over non-unique molecule, and

wherein the relative likelihood is a product of (i) the likelihood ratio of unique molecule over non-unique molecule given fragment positions and (ii) the likelihood ratio of UMI transition for unique molecule over non-unique molecule.

2. The method of any one of claims 3-4, wherein determining the relative likelihood of the two families are derived from the same original nucleic acid molecule comprises: determining likelihood of the two families are derived from the same original nucleic acid molecule using a sequencing error rate and/or a mismatch probability, optionally wherein the sequencing error rate is 0.001, optionally wherein the sequencing error rate is predetermined, optionally wherein the mismatch probability is 0.25, optionally wherein the mismatch probability is predetermined.

3. The method of claim 1 or 2, wherein the merging threshold is 1.

4. The method of any one of claims 1-3, wherein merging the two families comprises: merging a smaller family of the two families into a larger family of the two families.

5. The method of any one of claims 1-4, wherein performing probability-based merging comprises: family identification and merging.

6. The method of claim 5, wherein performing probability-based merging comprises: duplex identification and merging.

7. The method of any one of claims 1-6, wherein performing probability-based merging comprises: performing probability-based merging of families of the plurality of families using a probability map.

8. The method of any one of claims 1-7, wherein performing probability-based merging comprises:

(i) for one, one or more, or each pair of families of the plurality of families, determining a relative likelihood of the families of the pair are derived from the same original nucleic acid molecule; and

(ii) for the pair of families with the highest relative likelihood, if the relative likelihood of the families in the pair with the highest relative likelihood are derived from the same original nucleic acid molecule is above a merging threshold, then merging the families.

9. The method of claim 8, wherein performing probability-based merging further comprises: (iii) repeating (i) and (ii) until the relative likelihood of the families in the pair with the highest relative likelihood is not above the merging threshold.

10. The method of any one of claims 1-9, wherein the plurality of families comprises at least 500,000 families.

11. The method of any one of claims 1-10, wherein the plurality of families before probability-based merging is performed comprises at least 10% more families than the plurality of families after probability-based merging is performed.

12. The method of any one of claims 1-11, wherein each family of the plurality of families before or after merging comprises at least 5 sequence reads of the plurality of sequence reads.

13. The method of any one of claims 1-12, wherein one, one or more, or each of the plurality of sequence reads comprises a second UMI sequence.

14. The method of claim 13, wherein the UMI sequence is 5’ to the fragment sequence, and wherein the second UMI sequence is 3’ to the fragment sequence.

15. The method of claim 13, wherein the UMI sequence is 3’ to the fragment sequence, and wherein the second UMI sequence is 5’ to the fragment sequence.

16. The method of any one of claims 1-15, wherein the UMI sequence and/or the second UMI sequence is 4-20 bases in length.

17. The method of any one of claims 1-16, further comprising: subsequent to performing probability-based merging, for one, one or more, or each of the plurality of families, determining a consensus fragment sequence of the family, a position of the consensus fragment sequence aligned to the reference sequence, and/or a consensus UMI sequence of the family, optionally wherein the method further comprises: aligning the consensus fragment sequence to the reference sequence.

18. The method of any one of claims 1-17, further comprising: creating a file or a report and/or generating a user interface (UI) comprising a UI element representing or comprising, for one, one or more, or each of the plurality of families, (i) the family, (ii) sequence reads of the family, fragment sequences of the family, and/or UMI sequences of the family, and/or (iii) a consensus fragment sequence of the family, a position of the consensus fragment sequence aligned to the reference sequence, and/or a consensus UMI sequence of the family.