

CLAIMS

1. A computer-implemented method of generating base calls by a base caller, comprising:

receiving, from a plurality of clusters within a region of a flow cell, a plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels, wherein the plurality of sensor data is within a first range;

identifying a second range, such that at least a threshold percentage of the plurality of sensor data are within the second range;

mapping at least a subset of the plurality of sensor data, that are within the second range and that represent, from the region of the flow cell, a subset of the plurality of clusters incorporating different nucleotide bases with different labels, to a third range, thereby generating a plurality of normalized sensor data for a target cluster of the plurality of clusters;

processing the plurality of normalized sensor data in a base caller, to determine, for the target cluster, a base call and a plurality of quality scores corresponding to the base call; and

remapping each of at least a subset of the plurality of quality scores to a remapped quality score for the base call.

2. The computer-implemented method of claim 1, wherein the second range is fully encompassed within the first range.

3. The computer-implemented method of claim 1 or 2, wherein one or more outlier sensor data within the first range are absent from the second range of sensor data.

4. The computer-implemented method of any of claims 1-3, wherein identifying the second range comprises:

identifying, within the first range, a low value, such that a lower threshold percentage of the plurality of sensor data have a value that is lower than the low value; and

identifying, within the first range, a high value, such that an upper threshold percentage of the plurality of sensor data have a value that is higher than the high value,

wherein the second range is defined by the low value and the high value.

5. The computer-implemented method of claim 4, wherein at least one of the lower threshold percentage or the upper threshold percentage is 0.5% or less.

6. The computer-implemented method of claim 4, wherein at least one of the lower threshold percentage or the upper threshold percentage is 1.0% or less.

7. The computer-implemented method of any of claims 4-6, wherein each of the lower threshold percentage and the upper threshold percentage is 0.5% or less.

8. The computer-implemented method of any of claims 4-6, wherein each of the lower threshold percentage and the upper threshold percentage is 1% or less.

9. The computer-implemented method of any of claims 4-8, further comprising:

identifying (i) a first outlier sensor data of the plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels that is lower than the low value and (ii) a second outlier sensor data of the plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels that is higher than the high value; and

prior to the mapping, assigning the low value to the first outlier sensor data, and assigning the high value to the second outlier sensor data, such that the first outlier sensor data and the second outlier sensor data are within the second range subsequent to the assignment.

10. The computer-implemented method of any of claims 4-9, further comprising:

identifying (i) a first outlier sensor data of the plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels that is lower than the low value and (ii) a second outlier sensor data of the plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels that is higher than the high value; and

excluding the first outlier sensor data and the second outlier sensor data from the subset of the plurality of sensor data representing the subset of the plurality of clusters incorporating different nucleotide bases with different labels during the mapping, for being outside the second range, such that the first outlier sensor data and the second outlier sensor data are not mapped to the third range.

11. The computer-implemented method of any of claims 1-10, wherein mapping at least a subset of the plurality of sensor data representing the subset of the plurality of clusters incorporating different nucleotide bases with different labels comprises:

mapping a first sensor data within the subset from a first value that is within the second range to a second value that is within the third range; and

mapping a second sensor data within the subset from a third value that is within the second range to a fourth value that is within the third range.

12. The computer-implemented method of any of claims 1-11, wherein individual sensor data of the plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels comprises corresponding intensity of a corresponding section of an image generated from the flow cell.

13. The computer-implemented method of any of claims 1-12, further comprising:
processing the plurality of normalized sensor data in a base caller, to assign, the corresponding base called for the target cluster, a first quality score indicating a first probability of the corresponding base being an A, a second quality score indicating a second probability of the corresponding base being a C, a third quality score indicating a third probability of the corresponding base being a T, and a fourth quality score indicating a fourth probability of the corresponding base being a G.

14. The computer-implemented method of claim 13, wherein the plurality of quality scores corresponding to the base call comprise the first quality score, the second quality score, the third quality score, and the fourth quality score.

15. The computer-implemented method of claim 14, further comprising:
quantizing each of a plurality of remapped quality scores to a corresponding one of a plurality of quantized remapped quality score.

16. A non-transitory computer readable storage medium comprising computer program instructions that, when executed on a processor, cause a computing device to:

receive, from a plurality of clusters within a region of a flow cell, a plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels, wherein the plurality of sensor data is within a first range;

identify a second range, such that at least a threshold percentage of the plurality of sensor data are within the second range;

map at least a subset of the plurality of sensor data, that are within the second range and that represent, from the region of the flow cell, a subset of the plurality of clusters incorporating different nucleotide bases with different labels, to a third range, thereby generating a plurality of normalized sensor data for a target cluster of the plurality of clusters;

process the plurality of normalized sensor data in a base caller, to determine, for the target cluster, a base call and a plurality of quality scores corresponding to the base call; and

remap each of at least a subset of the plurality of quality scores to a remapped quality score for the base call.

17. The non-transitory computer readable storage medium of claim 16, wherein identifying the second range comprises:

identifying, within the first range, a low value, such that a lower threshold percentage of the plurality of sensor data have a value that is lower than the low value; and

identifying, within the first range, a high value, such that an upper threshold percentage of the plurality of sensor data have a value that is higher than the high value, wherein the second range is defined by the low value and the high value.

18. The non-transitory computer readable storage medium of claim 17, further comprising computer program instructions that, when executed on the processor, cause the computing device to quantize each of a plurality of remapped quality scores to a corresponding one of a plurality of quantized remapped quality score.

19. A system comprising:

at least one processor; and

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

receive, from a plurality of clusters within a region of a flow cell, a plurality of sensor data representing the plurality of clusters incorporating different nucleotide bases with different labels, wherein the plurality of sensor data is within a first range;

identify a second range, such that at least a threshold percentage of the plurality of sensor data are within the second range;

map at least a subset of the plurality of sensor data, that are within the second range and that represent, from the region of the flow cell, a subset of the plurality of clusters incorporating different nucleotide bases with different labels, to a third range, thereby generating a plurality of normalized sensor data for a target cluster of the plurality of clusters;

process the plurality of normalized sensor data in a base caller, to determine, for the target cluster, a base call and a plurality of quality scores corresponding to the base call; and

remap each of at least a subset of the plurality of quality scores to a remapped quality score for the base call.

20. The system of claim 19, further comprising instructions that, when executed by the at least one processor, cause the system to quantize each of a plurality of remapped quality scores to a corresponding one of a plurality of quantized remapped quality score.