|  |  |  |  |
| --- | --- | --- | --- |
| R1 | R2 | R3 | R4 |
| Forward Read 1 | Index 1 | Index 2 | Reverse Read 2 |

Severiano Villarruel

Sequencing Files Table

1. What is a good quality score cutoff for index reads and biological read pairs to utilize for sample identification and downstream analysis, respectively?

For my index sequences, average quality scores at each basepair are never below 30. This signifies that in general the basecalls can be trusted. Due to this general high quality I believe I can be more selective when picking the indexes I want to keep. This high quality is fortunate because for demultiplexing I want to ensure I have the correct basecall at each position. Taking this information into account I believe I should keep the index only if all quality scores are above 33, then if this condition is met I should select only the reads that have an average quality score greater than 35. This 33 quality score signifies that all basecalls in this index were made with at least 99.9% accuracy, then further eliminating indexes whose average quality score is above 35 ensures that there are not too many basecalls in the read with a score of 33.

For the quality scores of the sequence reads these are also no basepairs that average quality scores below 20, with this general high quality I am going to again be selective and say that the average quality score of the basecalls in the sequence must be above 20 (or 99% accuracy) or else they must be thrown out. The reason I am willing to keep low quality scores, as long as they are part of a sequence with an average of quality score of 30 or more, is because the sequences are long and one bad quality score should not dictate the selection of a read with a lot of quality scores.

1. How many indexes have undetermined (N) base calls? (Utilize your command line tool knowledge. Submit the command you used. CHALLENGE: use a one-line command)

zcat /home/svillarr/shared\_files1/2017\_sequencing/1294\_S1\_L008\_R2\_001.fastq.gz | awk 'NR % 4 == 2' | grep "N" | wc -l

zcat /home/svillarr/shared\_files1/2017\_sequencing/1294\_S1\_L008\_R3\_001.fastq.gz | awk 'NR % 4 == 2' | grep "N" | wc -l

R2

3976613 lines

R3

3328051 lines

Total 7304664 lines in indexes that have an “N”

Test Files

index1\_paired.fq: dual paired

index1\_hopped.fq: first read is hopped, second read is not

index1\_lowqual3.fq: first read quality scores are 3, second read is above cutoff

index1\_lowqual34.fq: first read quality scores are 34, second read is above cutoff

index1\_unknown.fq: first index is not in index reference, second index is in index reference

Part II

Given four input FASTQ files (2 with biological reads, 2 with index reads) and the 24 known indexes above, demultiplex reads by index-pair, outputting one forward FASTQ file and one reverse FASTQ file per matching index-pair, another two FASTQ files for non-matching index-pairs (index-hopping), and two additional FASTQ files when one or both index reads are unknown or low quality (do not match the 24 known indexes or do not meet a quality score cutoff).

Outputting the correct reads for each case is necessary.

index1\_paired.fq and index2.fq: output forward read to dual matched file and output reverse read to dual matched file

index1\_hopped.fq and index2.fq: output forward read to hopped file and output reverese read to hopped file

index1\_lowqual3.fq and index2.fq: output forward read to lowquality\_and unknown file and output reverese read to lowquality\_and unknown file

index1\_lowqual34.fq and index2.fq: output forward read to lowquality\_and unknown file and output reverese read to lowquality\_and unknown file

index1\_unknown.fq and index2.fq: output forward read to lowquality\_and unknown file and output reverese read to lowquality\_and unknown file