Nick Wagner

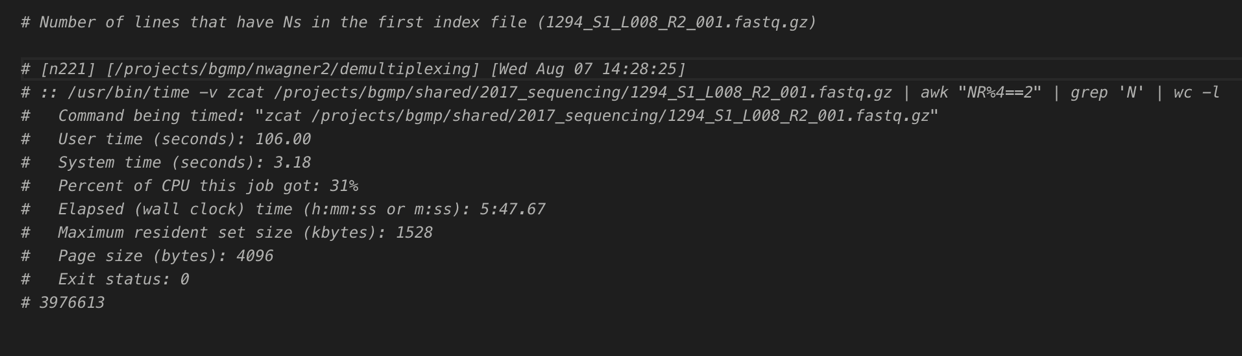
Bi622

August 7, 2019

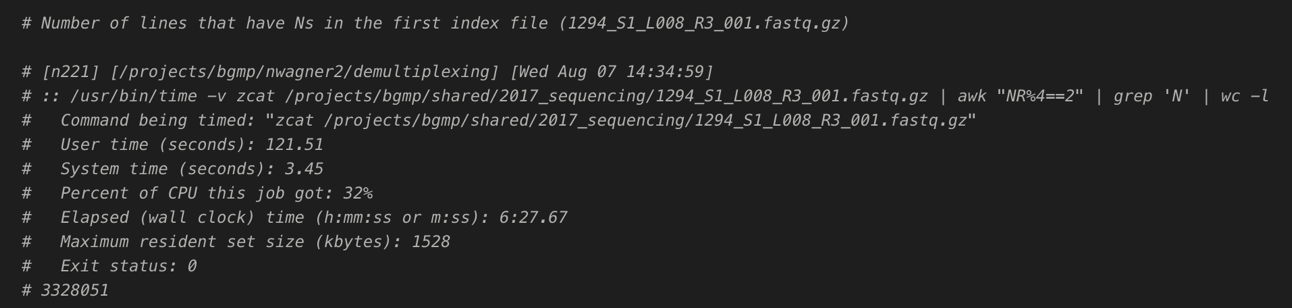
Demultiplexing Part 1 answers

1. Determine which files contain the indexes, and which contain the paired end reads containing the biological data of interest.

|  |  |
| --- | --- |
| ***File*** | **Type** |
| *1294\_S1\_L008\_R1\_001.fastq.gz* | Biological sequences |
| *1294\_S1\_L008\_R2\_001.fastq.gz* | Indexes |
| *1294\_S1\_L008\_R3\_001.fastq.gz* | Indexes |
| *1294\_S1\_L008\_R4\_001.fastq.gz* | Biological sequences |

1. Generate a per base distribution of quality scores for read1, read2, index1, and index2.
   1. Histograms included in github repo
   2. What is a good quality score cutoff for index reads and biological read pairs to utilize for sample identification and downstream analysis, respectively?
      1. The quality score cutoff would depend on if we analyzing the index or the biological reads. It is a lot more important that we are really confident in our index reads, because that is how we determine if they are dual matched or not. The change of one base could mean a completely different index. The four lowest averages from both index files occurred at base one and base 0, and all had average phred scores of around 31. I think 31 is a reasonable cutoff for indexes because we don’t want to lose too many of our reads, but we also don’t want to allow bad phred scores.
      2. For analyzing biological reads, we don’t need to be AS confident because coverage here is a lot more important than a few bad bases. I would choose a biological read cutoff of 38. This is higher than what I chose for indexes because per base, the average phred scores for biological reads are closer to the high 30s/low 40s.
   3. How many indexes have undetermined (N) base calls?

3,976,613 lines with Ns



3,328,015 lines with Ns

7,304,628 total lines with Ns