Joshua Schaaf – Computational Genomics Week 2 Assignment (Dr. Jody Hey)

For this week’s assignment, we created a program that samples reads from a fasta file, creating a fastq file. The fastq files created with my program are labeled chr21\_CM000482.1.16741801.fastq and chr21\_CM000511.1.16741801.fastq. These files hold reads from each strand of their respective chromosomes, and are 10,000 reads, making them 10 fold coverage. ART is a program that can do this similarly, and these files are named similarly, with .art.fq instead of .fastq (also around 10,000 reads, so 10 fold coverage). These fastq files were then submitted to fastqc, a program that can give quality analysis of the fastq file reads.

Fastqc returns an html file, when displayed in the browser, gives graphs of quality analysis. Per base sequence content is a graph that shows the average A,T,C and G content for each position in the reads. You would expect no difference between reads, so if there are large differences between reads, there is a problem with the fastq data. Per sequence GC content measures GC content across the entire length of the sequence and compares it to a normal distribution of GC content. In a normal library, you expect a normal distribution of GC around the average GC content, so if the sum of the deviations from the normal distribution represents more than 15% of the reads, a warning is raised. Per base N content does this, but with N content instead of GC. It is expected that a good library has little to no N content, and if there is N content, it is spread along all bases. If not, there may be some sequencing error causing issues. Sequence length distribution returns a graph with all the lengths of sequences in the fastq file. Sequence Duplication levels are also graphed, showing any sequence placing duplicated sequences at their sequence duplication level. Good sequence information will have lower amounts of duplicate sequences, as high levels of non-unique sequences indicate poor sequencing coverage. Overrepresented sequences as well as Kmer content describe sequences that are overrepresented. A high-throughput library will contain a diverse set of sequences, decreasing likelihood of overrepresented sequences as well as overrepresented k-mers. The fastqc module lists all the sequences which make up more than 0.1% of the total sequence.

Everything seemed similar between my program’s fastq file and the ART created fastq file. Besides quality, all the graphs have extremely similar characteristics, and the specifics are described in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Chr21 – 482.1.fq | Chr21 – 482.1.ART.fq | Chr21 – 511.1.fq | Chr21 – 511.1.ART.fq |
| Per base seq content | 33% T, 31% A, 18% C & G for all positions | 32% A & T, 18% C & G for all positions | 33% T, 31% A, 18% C & G for all positions | 32% A & T, 18% C & G for all positions |
| Per seq GC content | Theoretical peak at 36%, GC count per read is less gaussian, with higher peak centered at 34% | Theoretical peak at 36%, GC count per read is less gaussian, with higher peak centered at 34% | Theoretical peak at 36%, GC count per read shifted slightly right, however similar peak, w/ highest 33% | Theoretical peak at 36%, GC count per read is shifted slightly right, with peak at 33% |
| Per base N content | No per base N content | No per base N content | No per base N content | No per base N content |
| Seq length distribution | All sequences at 100 bp | All sequences at 100 bp | All sequences at 100 bp | All sequences at 100 bp |
| Seq duplication levels | 95.26% seqs remaining if deduplicated, no sequences duplication level over 2 | 98.44% seqs remaining if deduplicated, no sequences duplication level over 2 | 94.83% seqs remaining if deduplicated, no sequences duplication level over 2 | 98.44% seqs remaining if deduplicated, no sequences duplication level over 2 |
| Overrepresented sequences | No overrepresented sequences | No overrepresented sequences | No overrepresented sequences | No overrepresented sequences |
| Kmer content | No overrepresented k-mers | No overrepresented k-mers | No overrepresented k-mers | No overrepresented k-mers |