**Part 1**

1. **The cells of some bacterial species are known to be harder to lyse than others.**

**What would be the result of a bias of cell lysis?** **How would it manifest in the resulting read files?**

Those species would not expose their DNA because their membrane will stay intact.

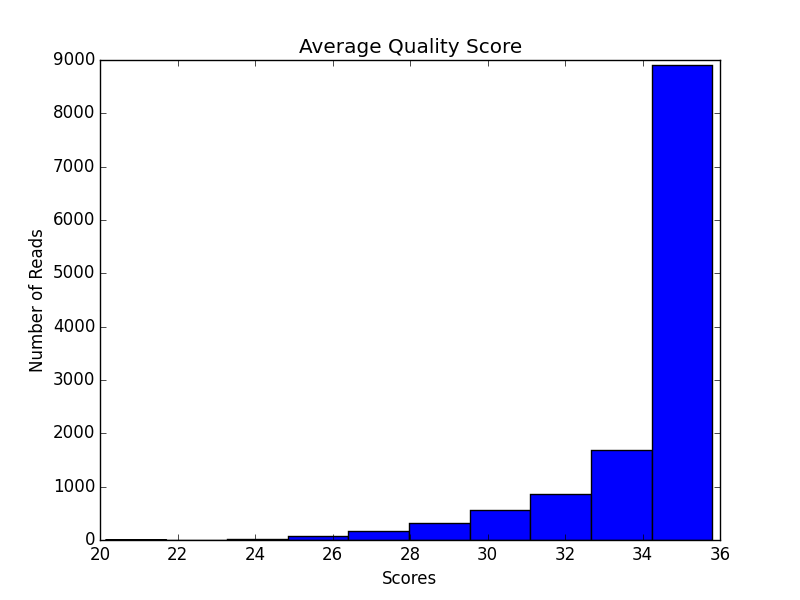
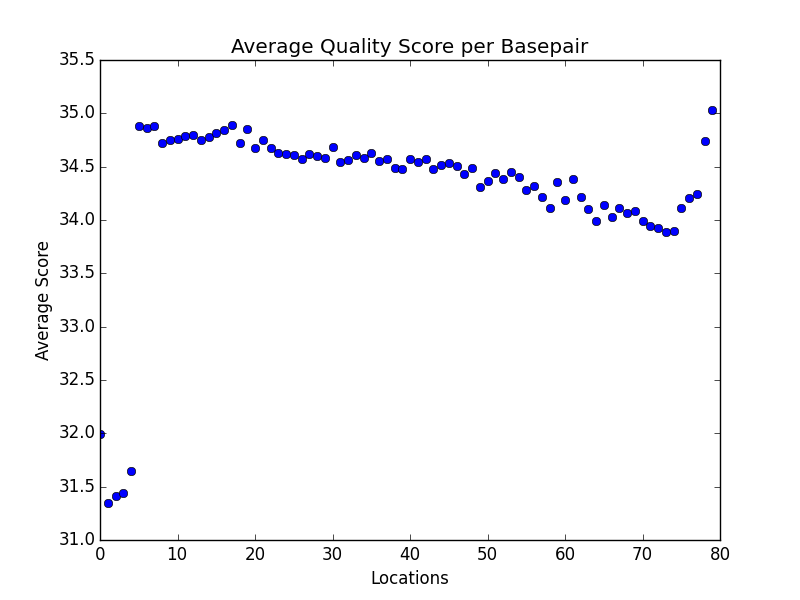
We won’t see their DNA at all or maybe just small amount of it which so we can miss it by mistake.

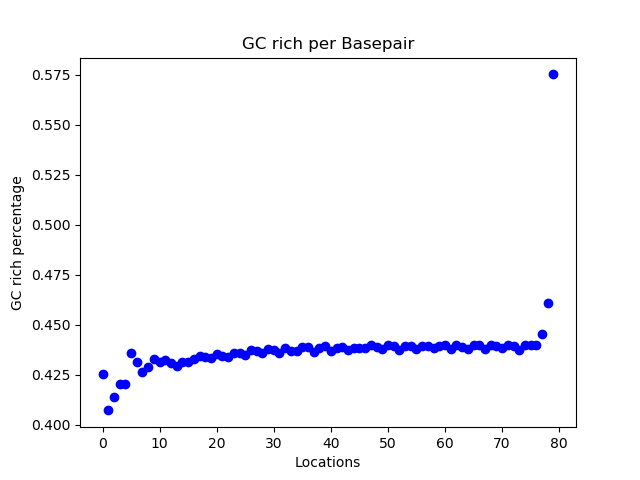
**What would be the causes and results of possible sequence biases in DNA fragmentation or tagging?**

* Pollution of the sample - some DNA which isn’t coming from the stool sample itself. It will result in bad alignment or just noise.
* Sample size - small sample which may be missing some of the species.
* Repetitive sequences- harder to put back together because they are not very distinct between the different bacteria, the same apply to cg rich areas.
* Cells in Replication stage - some parts of the DNA would be in double amount - it will be harder to understand the correct amount of bacteria from the specific species.
* Methylation - difficulties slicing and tagging the DNA.
* Very big differences in sizes of bacterial population – we may miss the small species.
* The tagging enzyme cannot tag some sequences (because of specific sequence maybe) - we will have the reads but we will not know from which sample it came.

**How would that be manifested in read files?**

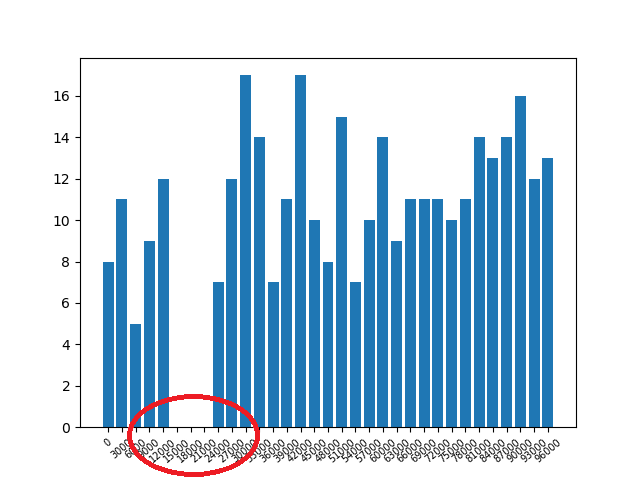
* missing reads
* short reads
* very similar reads
* wrong reads (not from the sample)

1. **Sample**: <https://www.ebi.ac.uk/ena/data/view/ERX1189850> **:** [SAMEA3646703](https://www.ebi.ac.uk/ena/data/view/SAMEA3646703)
   1. **Average read quality -** the graph shows the amount of reads for each average score range. Most of the reads had a good score of 34-36.
   2. **Average quality per read basepair** (location) - this graph shows the average score between the different locations on the read. The first and the last location looks different from the general trend, maybe because of the way the reads are cut. 
   3. **Probabilities for an error** 
      1. **First basepair location average probability** of **error : 0.000632354069012**
      2. **Tenth basepair location average probability** of **error: 0.000335173628737**
      3. From those numbers, as well as the graph, we see that the first position is with worse score then the tenth- the extreme positions values are very different from the rest.
   4. Accession : [SAMEA3646703](https://www.ebi.ac.uk/ena/data/view/SAMEA3646703)
   5. url for the downloaded fastq<https://www.ebi.ac.uk/ena/data/view/ERX1189850>
2. GC rich regions – in this graph we can see percentage of G/C of the general amount of nucleotides per read location. It looks like, again, the differences are in the first and last position.



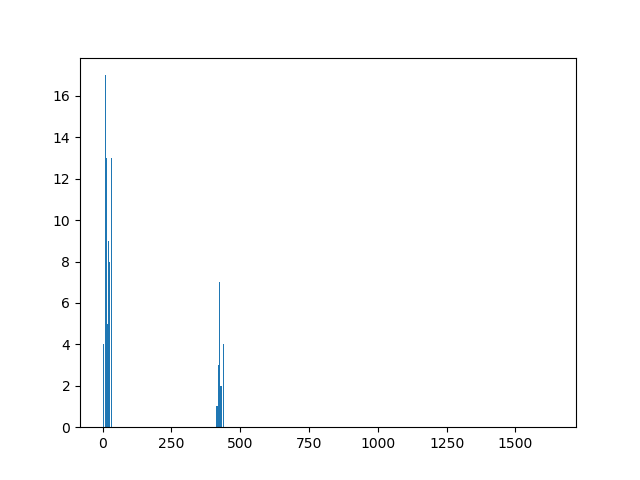
**Part 2:**

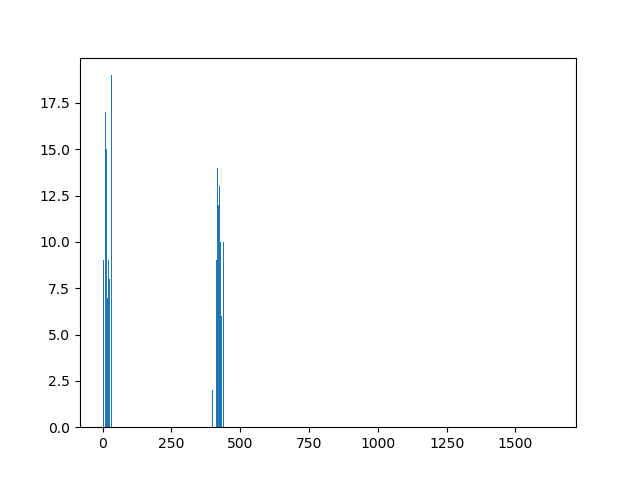
1. We can see the deletion in this file in the area where no reads were mapped:



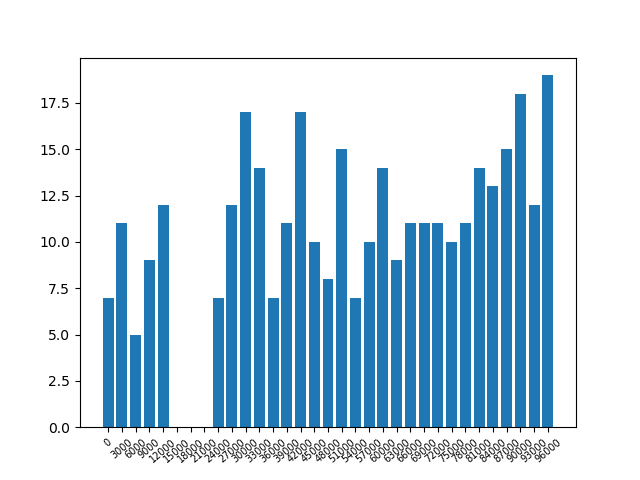
2. After I created this graph I run the same code on the second and third sam files:

Second: it looks like each read can map only to one of the references each time. So there were less reads for the kraken genome. Also there were two peaks because of different positions on the different genomes.

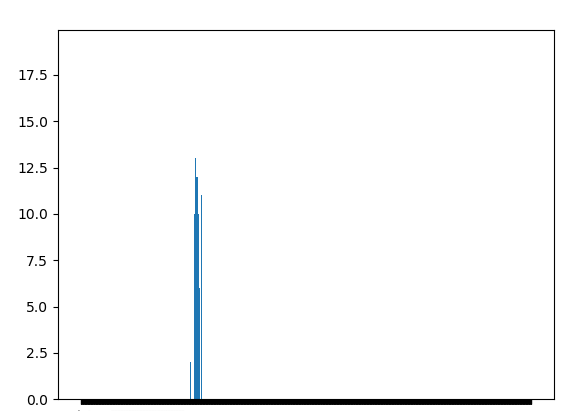


Third file (with the same code), I think this time reads can map twice: on each genome so the graph still had two peaks but more reads per each: 

After that I decided to try and fix my code to separate the alignments:



And (only one peak because of the separation):



3. (Bonus:) I think that we need to perform a multiple alignment between the reference genomes before we try to find the locations of the aligned reads. The problem is that it is a hard work to multiple align the full database (it sounds hard). Also there will be deletions and changes between the reference genomes so it will hard to understand the real amount of aligned reads.

We can also make a separate alignment like we did in 2a for each reference genome and align only the relevant areas and calculate the average between them.