Khan Inan
Professor Truong
BI-GY 7683

1. The most likely reason for there being no reads mapping to a specific gene of interest is that there is a lack of standardization for depth. When there is too much depth and not enough breadth for a sequencing that is performed, then you will have multiple reads for a specific part of the sequence but you can be missing other parts (like a gene of interest.) To fix this issue, it would be better to use a sequencing method that allows for lower depth but high breadth, so you can collect better data on the entirety of the genome. Another reason for why desired reads can be missing is due to errors in sequencing. Since certains mutations and errors occur essentially randomly in nature and at varying rates, it can be possible to find a genes that are available in certain cells but missing in other identical cells. In order to understand the rate of detecting and receiving these eros, the poisonn process would need to be used in order to get a visualization of the distribution of errors. Additionally a Gamma distribution would allow us to understand how our specific case of sequencing(errors and all) is related to all the other possible errors and reads that could occur.
2. Counting UMI's is preferable over counting reads because reads do not provide unique codes and tags to each molecule in a group of samples. If someone needs to assess the expression of certain genes that may be rarer and individualized to specific cases, then it is more useful to use the identifier method that corrects for errors and outliers and also provides more accuracy. Essentially Read counts are the count of all the mapped "segments" that are collected, while UMI count is the number of distinct identifiers/barcodes, and consequently represents the original number of RNA fragments (before any kind of PCR or alteration.)
3. Since phred scores are essentially the numerical form of representing how confident we are in the accuracy of a sequence of nucleotides, having a consensus read would mean that we have a higher phred score and thus a higher probability that a decision is correct. In terms of the equation for phred score, $Q = -10 \log_{10}(P)$, the Q value would go down and the P value would go up with a consensus read
4. The main reason for noise in DNA sequencing is the deamination of cytosine, or in other words, the deterioration of DNA over time. cytosine and 5-methylcytosine become less stable over time and transition mutations can occur. One of the best early experiment that highlights this is most likely the ones involved in the Human Genome Experiment. Scientists that contributed the sequencing of the human genome most likely discovered that using DNA from older samples and ancient human bodies did not provide a reliable map to the current genome of humans.