# How to evaluate a fastqc report (RADseq data)

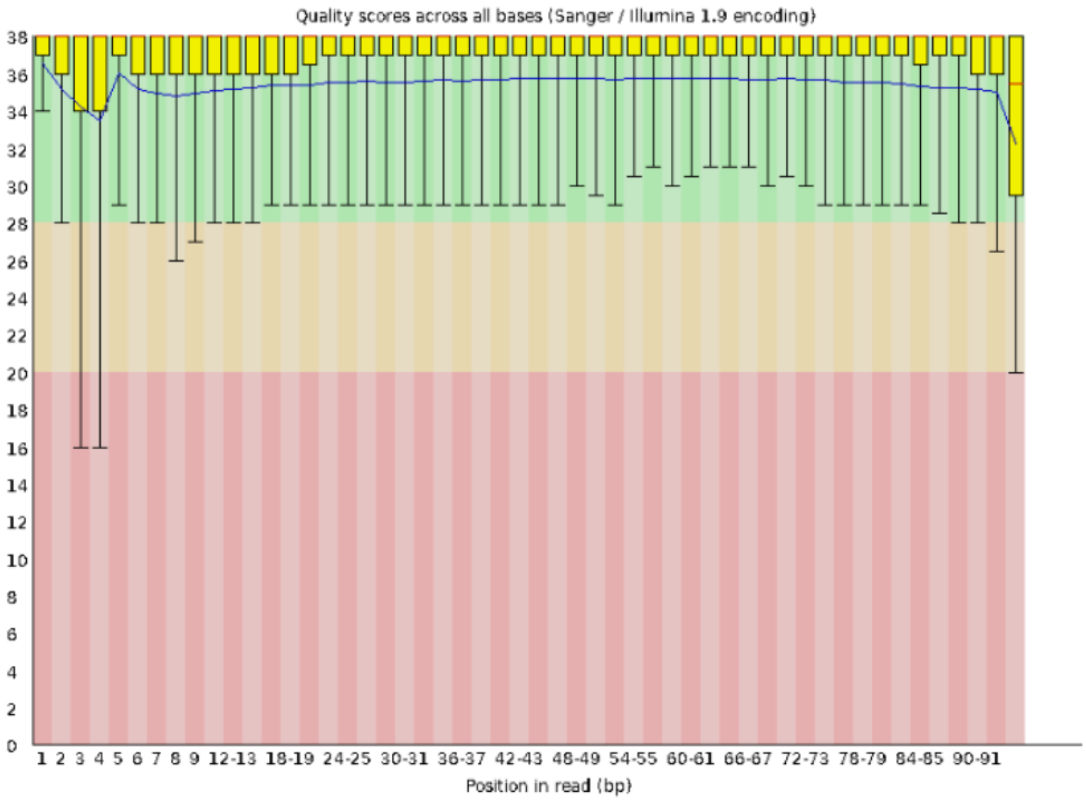
*Sophie Breitbart, 2022*

1. **Basic Statistics**
    a. filename, file type, encoding, total sequences, seqs flagged as poor quality, seq length range, %GC

## Basic Statistics

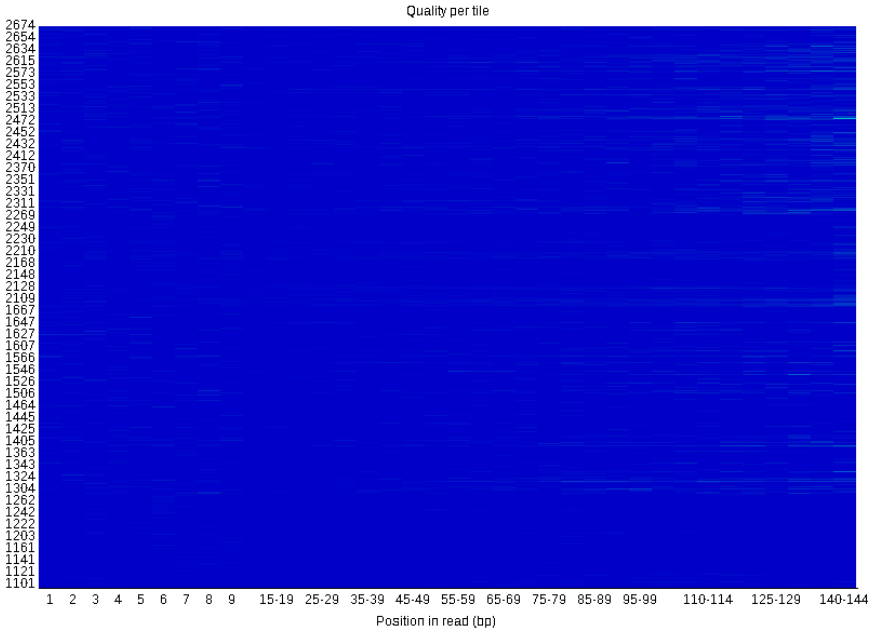
| Measure | Value |
|---|---|
| Filename | 40559-1.2_R1_2G3_R1_val_1.fq |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 2533372 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 20-144 |
| %GC | 42 |

2. **Per base sequence quality**
    a. Example A: the sequence quality per base is uniformly quite high, with dips only in the first and last 5 bases. This is typical for Illumina reads.
    b. Based on information from this plot we can see that this data doesn't need any trimming, which is good.

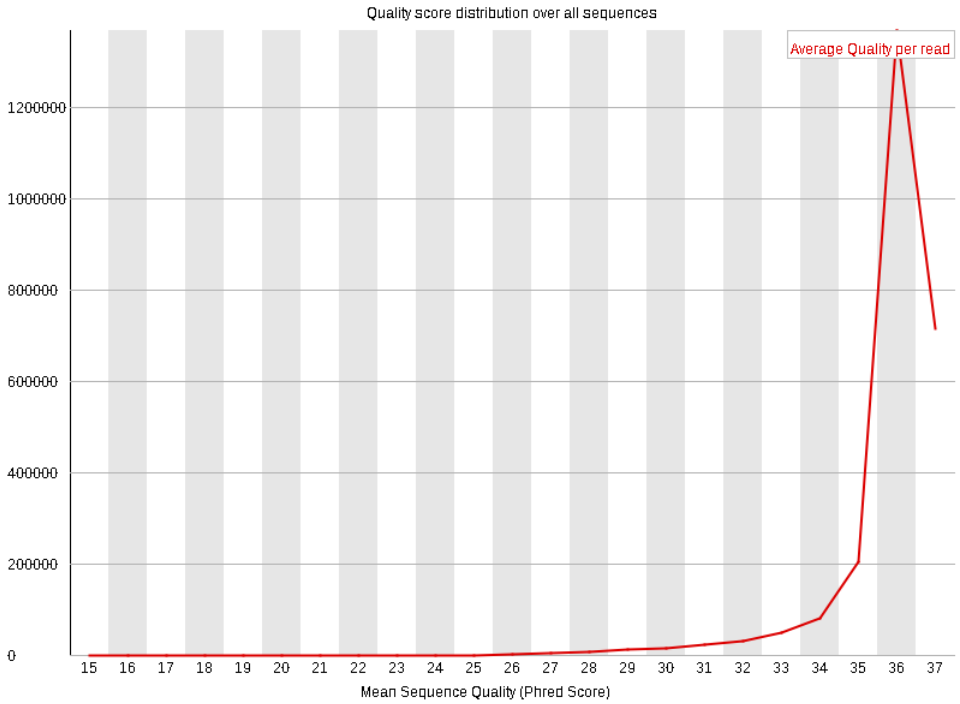



3. Per tile sequence quality

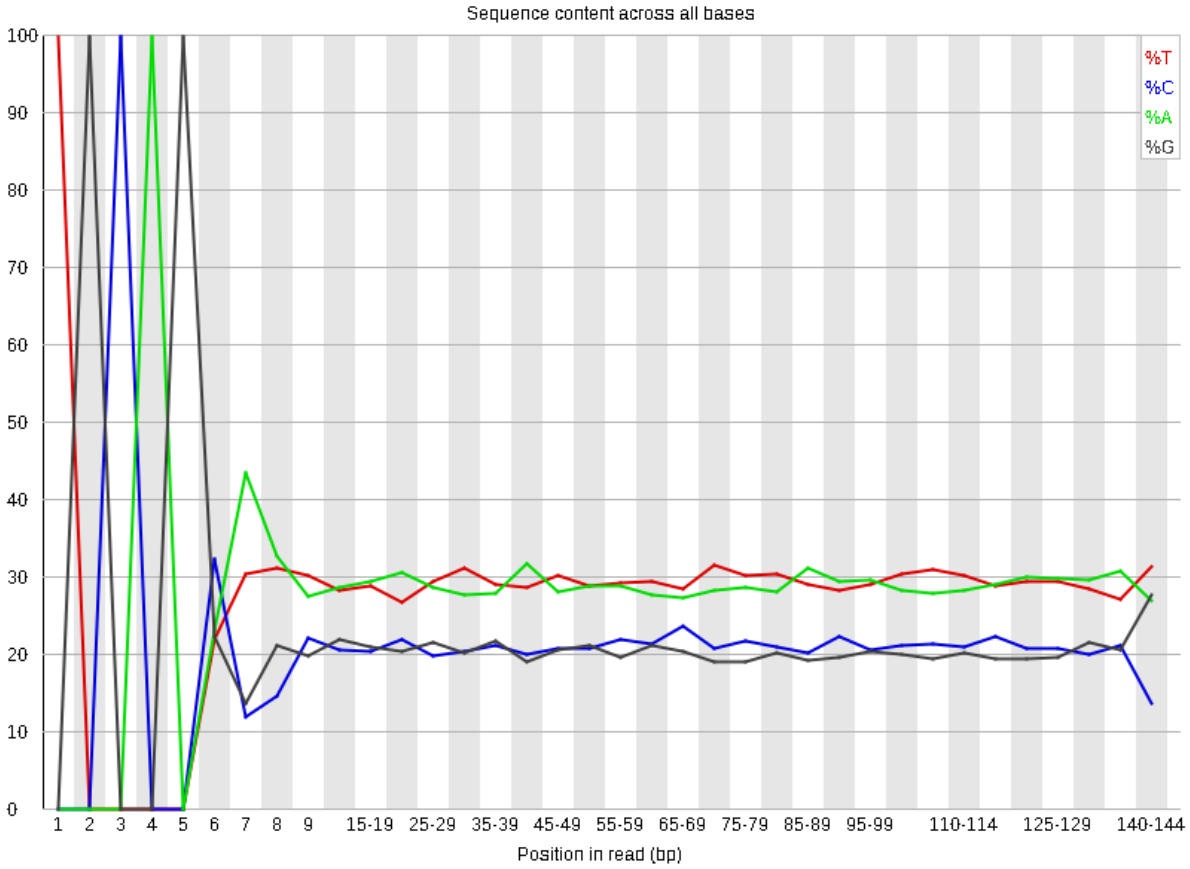
    a. All blue is fine



4. Per sequence quality scores
    a. Higher the peak Phred score, the better




5. **Per base sequence content**
    a. The squiggles indicate base composition per base position averaged across the reads. It looks like the signal FastQC is concerned about here is related to the *extreme* base composition bias of the first 5 positions. We know this is a result of the restriction enzyme overhang present in all reads (`TGCAG` in this case for the PstI enzyme used), and so it is in fact of no concern.



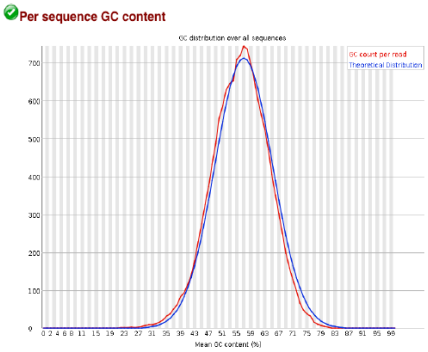

6. **Per sequence GC content**

    a. Choppy distribution is fine for RAD data

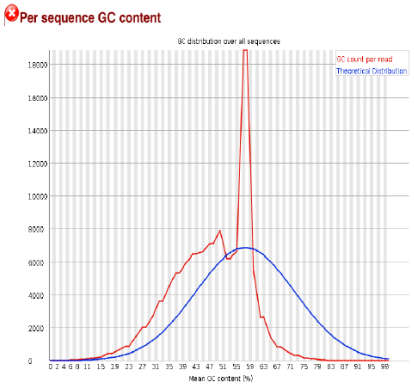
# FASTQC: Per sequence GC content

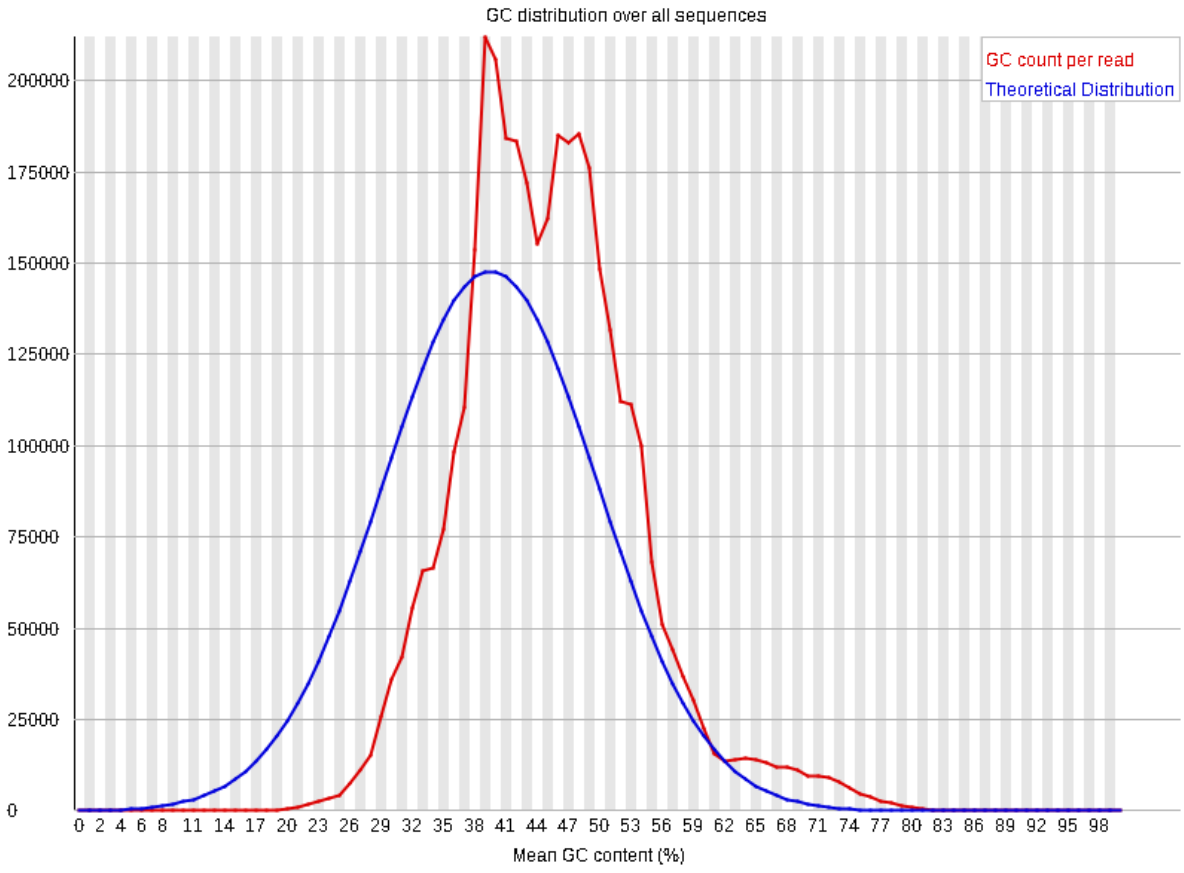
RAD data has lots of identical sequences, so the distribution should not be smooth

**WGS should like this**



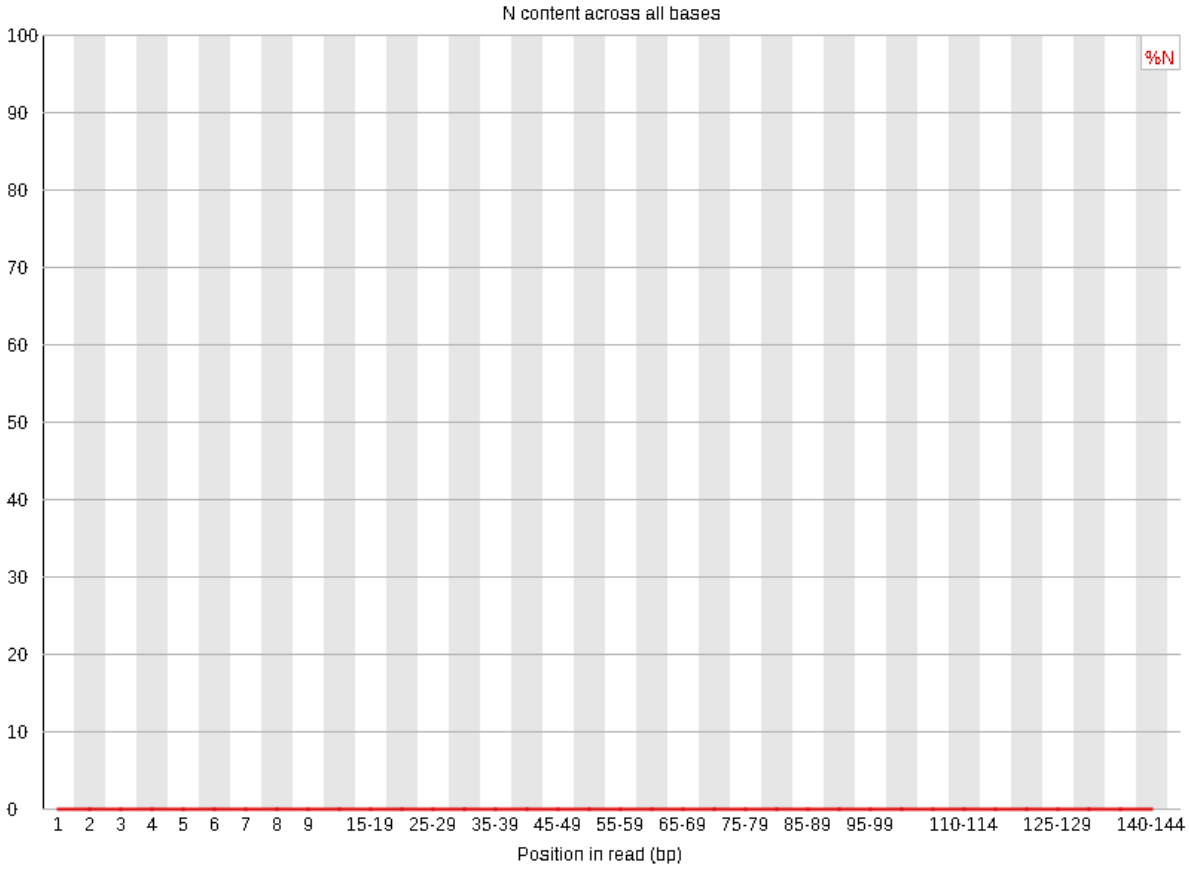

**RAD data looks choppy**




7. **Per base N content**

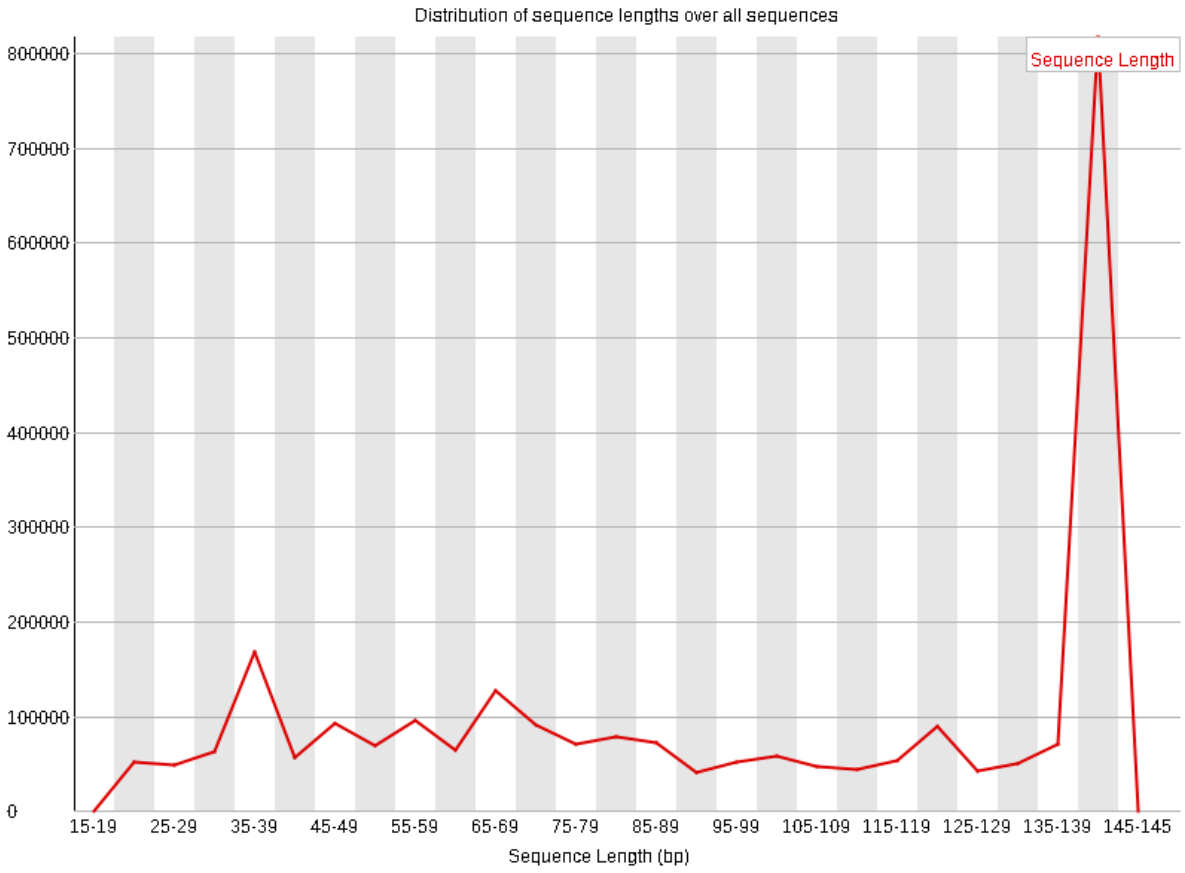
a. Goal: low N content. More info here: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/6 Per Base N Content.html




8. **Sequence Length Distribution**

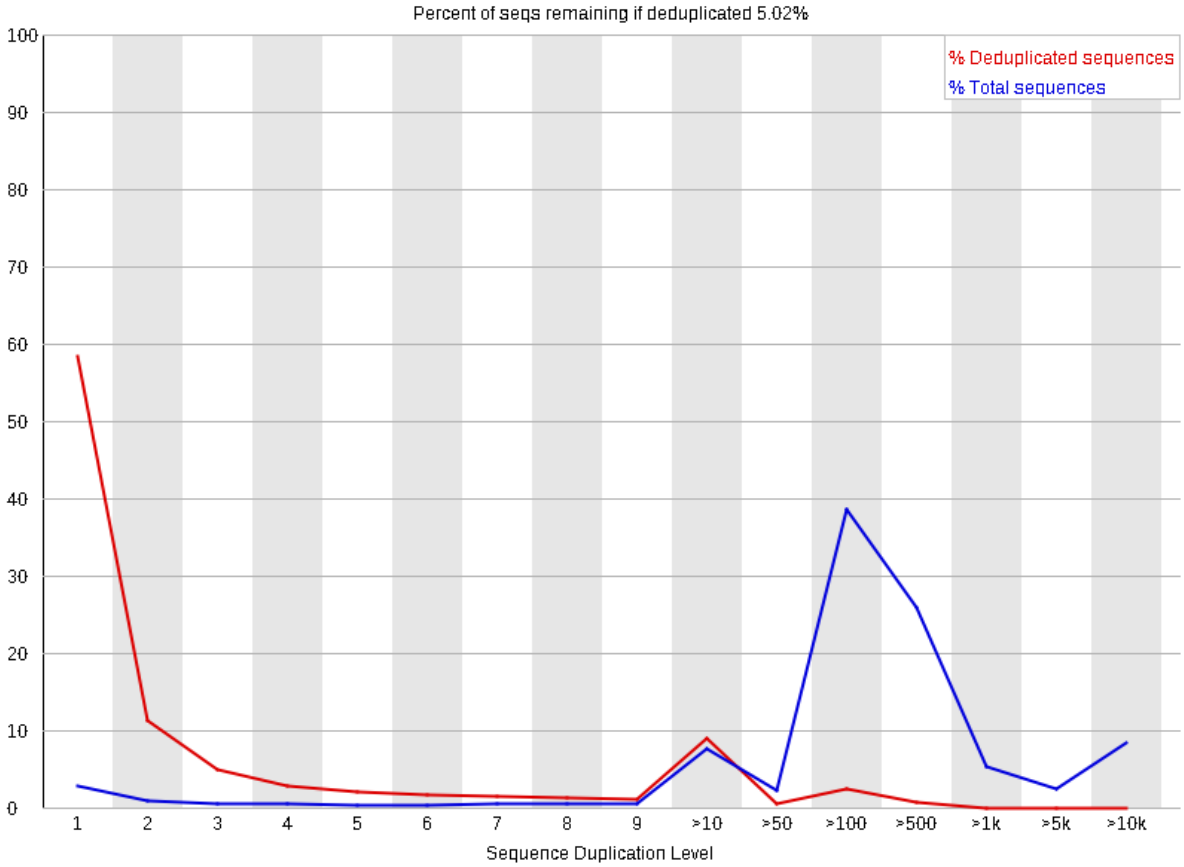
a. For some sequencing platforms it is entirely normal to have different read lengths so warnings here can be ignored. I think this is the case with RADseq data.

b. In any case, this graph shows a warning. This module will raise a warning if all sequences are not the same length.




9. **Sequence Duplication Levels**

    a. This is expected for RADseq datasets because the same sequences are targeted and amplified




10. **Overrepresented sequences**

    a. The percentages are out of 100, not 1! Anything >0.1% gets flagged. Since this is RADseq data, we expect to see sequences occurring over and over again. Less than 1% is fine esp. since the sources aren't identified as primers or adapters or contamination.

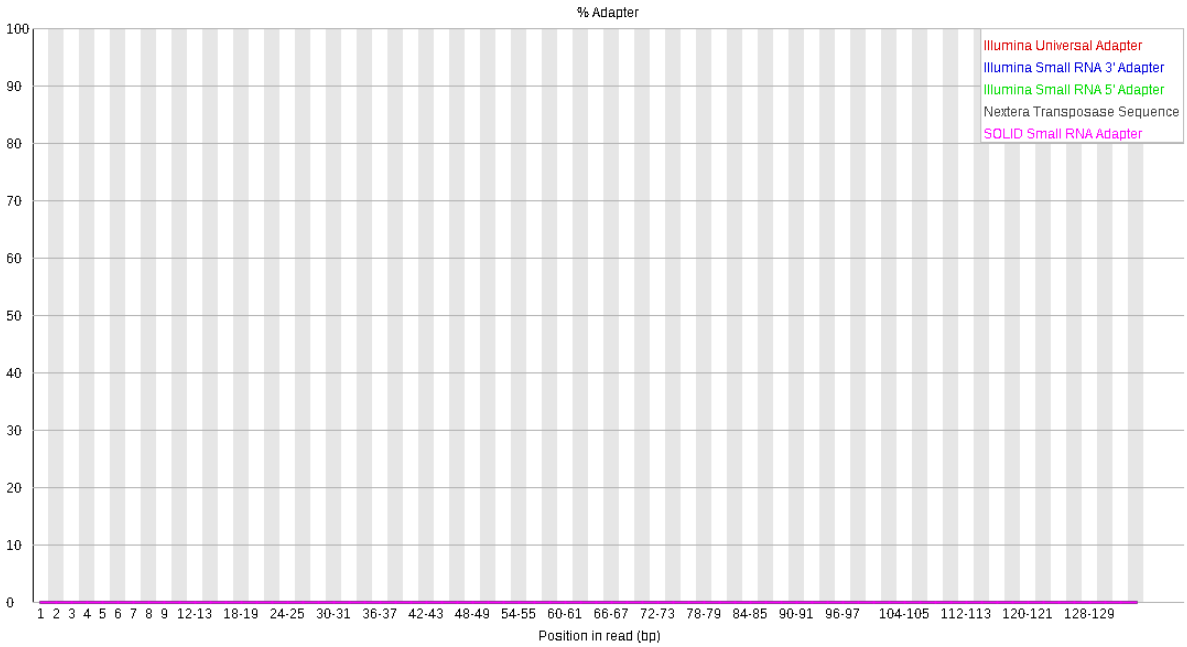
    b. Tip: Take a sequences, BLAST it, and if it comes up in the genome, that's a good sign! That's what happened here.

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| TGCAGGAAACTCGATTACATAAGAAATGGTTCGACTGAAAAAAAGAATGGAGCTACATTGCTTCCTGCA | 20935 | 0.8251215315138487 | No Hit |
| TGCAGGAAGCAATGTAGCTCCATTCTTTTTTTCAGTCGAACCATTTCTTATGTAATCGAGTTTCCTGCA | 20014 | 0.7888217020166309 | No Hit |
| TGCAGCATGCCAGACGAGAAGTCCGCTTAGCTTCTAGCTAGAGATGAGAA | 17411 | 0.6862283728295974 | No Hit |
| TGCAGGGATTCTTTAAGTGCTTCCTCTTAGGCGATGCCCGGATCTCAGTC | 13741 | 0.5415808437798804 | No Hit |
| TGCAGCTCAGGAACATGACCGTCCCACCATACGTCTGTACAAAAACATCA | 13685 | 0.5393736880232634 | No Hit |
| TGCAGCAATCATAGGAAAAAAAAGACAGAAATTCGGCTGCAAATGCTGCA | 13477 | 0.5311756809272576 | No Hit |
| TGCAGCATTTGCAGCCGAATTTCTGTCTTTTTTTTCCTATGATTGCTGCA | 11688 | 0.46066493720247736 | No Hit |
| TGCAGTAAGGAAGGAATTCTGTTTAAAGGGGTTTGATGGTAAGGAATTAC | 11595 | 0.45699948210666713 | No Hit |
| TGCAGCAAAACAGCATTCCAGAACAGCCGAATGCTGCA | 11333 | 0.44667314624535215 | No Hit |
| TGCAGCAATCATAGGAAAAAAGACAAGAATTTCAGGAATGTGCTGCTGGAATATGCTGCTGAAATCCTGCA | 10953 | 0.43169601789688017 | No Hit |
| TGCAGCGCAATCGTGTGGAAAATAGGCTGCATGGGCCATACCAAGCATGA | 10761 | 0.4241286267313364 | No Hit |
| TGCAGGTTGGGAGAGTGTAACACTAAGTCAATTTGACTTTCGAACTTTCT | 10590 | 0.41738891897452385 | No Hit |
| TGCAGTTGACCAGAACTTAAAGGGAGATGGGAGCATTTTCAGTCGATGAG | 10123 | 0.3989828165041648 | No Hit |
| TGCAGGATTTCAGCAGCATATTCCAGCTGCACATTCCAGCTACTTTCCTG | 9292 | 0.36623020161579567 | No Hit |
| TGCAGGATTTCAGCAGCATATTCCAGCAGCACATTCCTGAAATTCTTGTCTTTTTTCCTATGATTGCTGCA | 9217 | 0.36327418944175516 | No Hit |
| TGCAGCATTCGGCTGTTCTGGAATGCTGTTTTGCTGCA | 8992 | 0.3544061529196335 | No Hit |

11. Adapter Content

    a. Goal: Low adapter content since that should've been trimmed.




# General things to keep in mind

- Quality Drop - The quality decreases with read length.
- R1 vs. R2 - Average quality is higher for the first (forward) read (R1) compare to the second (reverse) read (R2).

## Helpful tutorials

- #1: From the FastQC creators: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/
- https://training.galaxyproject.org/archive/2021-06-01/topics/sequence-analysis/tutorials/quality-control/tutorial.html#sequence-duplication-levels
- https://medium.com/@shilparaopradeep/analysing-raw-sequencing-reads-with-fastqc-for-quality-control-and-filtering-cacaf06b8988
- https://radcamp.github.io/NYC2019/RADCamp-PartII-Day1-AM.html