

How to interpret MultiQC reports

☰ With whom?	
📅 Date	@March 15, 2023
🔍 Revisit: Action Items?	No
☰ Property	

General Stats

% BP Trimmed

- percent of read that was trimmed when adapters, regions with error rates >10% were removed
 - *CutAdapt removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads*

% Dups

- % duplicate reads. This is fine because GBS involves generating lots of duplicate sequences.

% GC

- GC content

Read Length

- Mean read length for each sample. Higher is better

% Failed

- % of modules failed in FastQC report (should not care about these numbers because we expect to fail some FastQC modules due to high duplication, etc.)

M Seqs

- Millions of sequences read per sample
 - for plates 1-2, the vertical line at 0 is for the blank wells

Cutadapt

Filtered Reads

- This plot shows the number of reads (SE) / pairs (PE) removed by Cutadapt. For my plates, makes sense % is 100% for all b/c I didn't indicate PHRED score >0.

Trimmed Sequence Lengths (3')

- This plot shows the number of reads with certain lengths of adapter trimmed for the 3' end.
- Longer is better, though I'll trim to 120 bp later because the longer the read, the higher the chance of sequencing error

FastQC

Sequence Counts

- Sequence counts for each sample. Duplicate read counts are an estimate only.
 - It's good to see duplicates for GBS.

Sequence Quality Histograms

- The mean quality value across each base position in the read.

Per Sequence Quality Scores

- The number of reads with average quality scores. Shows if a subset of reads has poor quality.


Per Base Sequence Content

- The proportion of each base position for which each of the four normal DNA bases has been called.
- 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.

Per Sequence GC Content

- The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.
- Choppy distribution is fine for RAD data (source: RADseq workshop)

Per Base N Content

- The percentage of base calls at each position for which an  was called.
 - Goal: low N content

Sequence Length Distribution

- The distribution of fragment sizes (read lengths) found.
 - For some sequencing platforms it is entirely normal to have different read lengths so warnings here can be ignored. This is the case with RADseq data.
 - In any case, this graph shows a warning. This module will raise a warning if all sequences are not the same length.

Sequence Duplication Levels

- The relative level of duplication found for every sequence.
 - again, high duplication is fine for GBS— This is expected for RADseq datasets because the same sequences are targeted and amplified

Overrepresented sequences

- The total amount of overrepresented sequences found in each library.
 - 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.

Adapter Content

- The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.
 - Goal: Low adapter content since that should've been trimmed.
 - The samples that go into the yellow/red territory are blanks and will be removed from the dataset anyway (except for A1 and A10 for plate 3- see explanation in pipeline google doc)

Status Checks

- Status for each FastQC section showing whether results seem entirely normal (green), slightly abnormal (orange) or very unusual (red). Sort by highlight
 - Don't pay attention to this- again, it's concerned with the modules "passing", and we know our samples shouldn't pass certain modules b/c duplication is expected.

