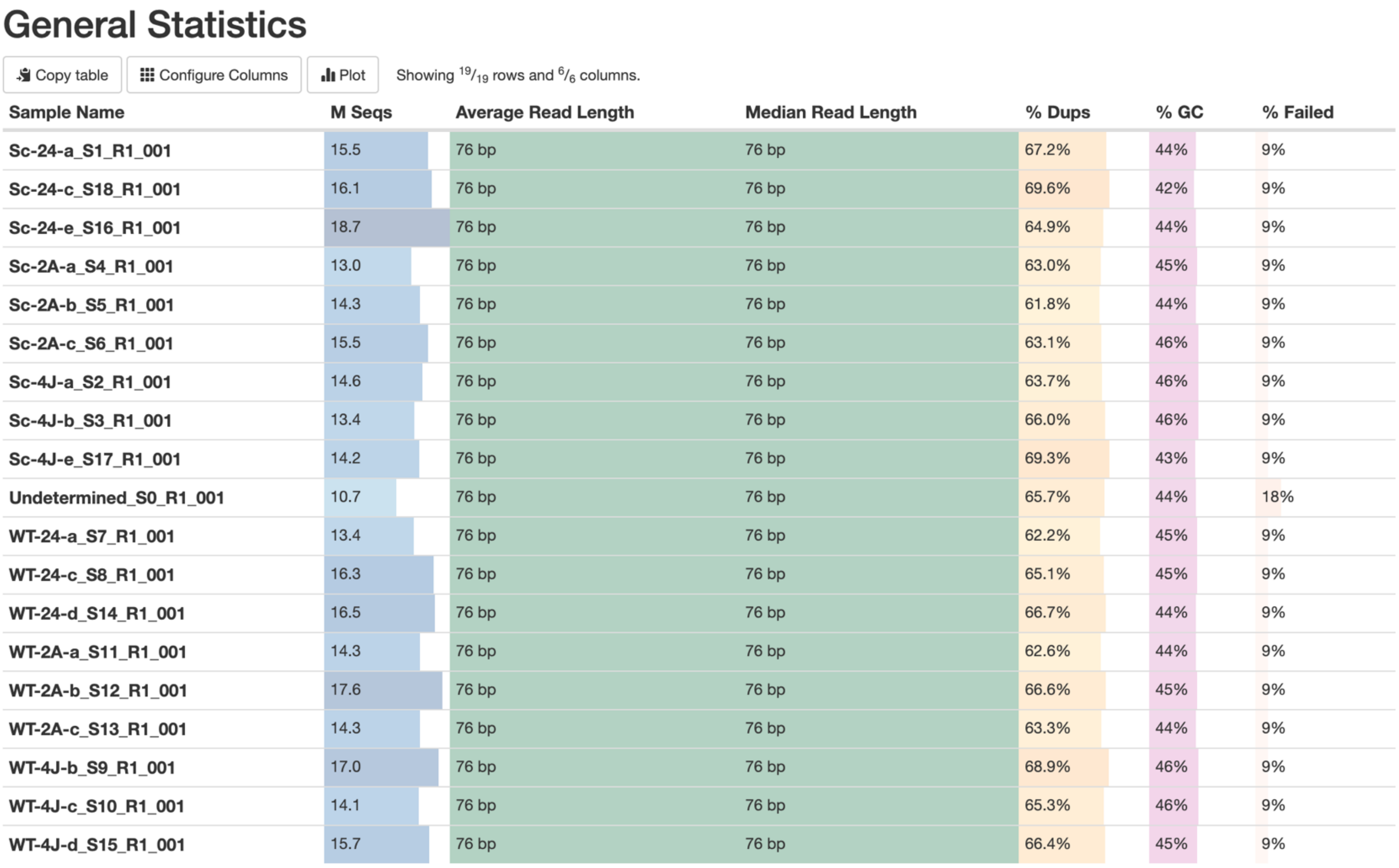
Initial QC Report – FASTqc and MultiQC Aggregation

I decided to learn how to use MultiQC to aggregate and summarize the FASTqc results for each of the 18 RNAseq samples in the DAP CURE dataset. Below I show overall summary statistics and flags, and then highlight modules with unusual results or "failed" sample reports. The full summary can be seen here! (Download .html file and open).

Overall Summary Statistics

Results are uniform across all samples with the exception of the undetermined genotype sample, which has double the number of failed sequence reads and the lowest number of sequence reads total. Interestingly the read length of every single sample is 76 base pairs, likely due to library preparation method. Percentage duplication is high, but this is expected with RNASeq data and is consistent across all samples, which may also be a result of PCR bias. Percent GC content is also consistent across all samples. Because exactly 9% of all sequences failed for all 18 samples, it is possible that a specific set of transcripts present in all samples failed sequencing. This could be due to errors introduced in library preparation, high regions of repeats, high kmer content, or poor primer annealing for these transcripts. For some reason FASTqc did not give a report on Kmer content, which is another important metric to establish for each sample and may help determine the source of some biases present in the data.

Some difference in the sequence of the Undetermined genotype sample differentiates it from the other 17 samples in almost every metric/feature analyzed. It is currently unclear what the source of this difference is.

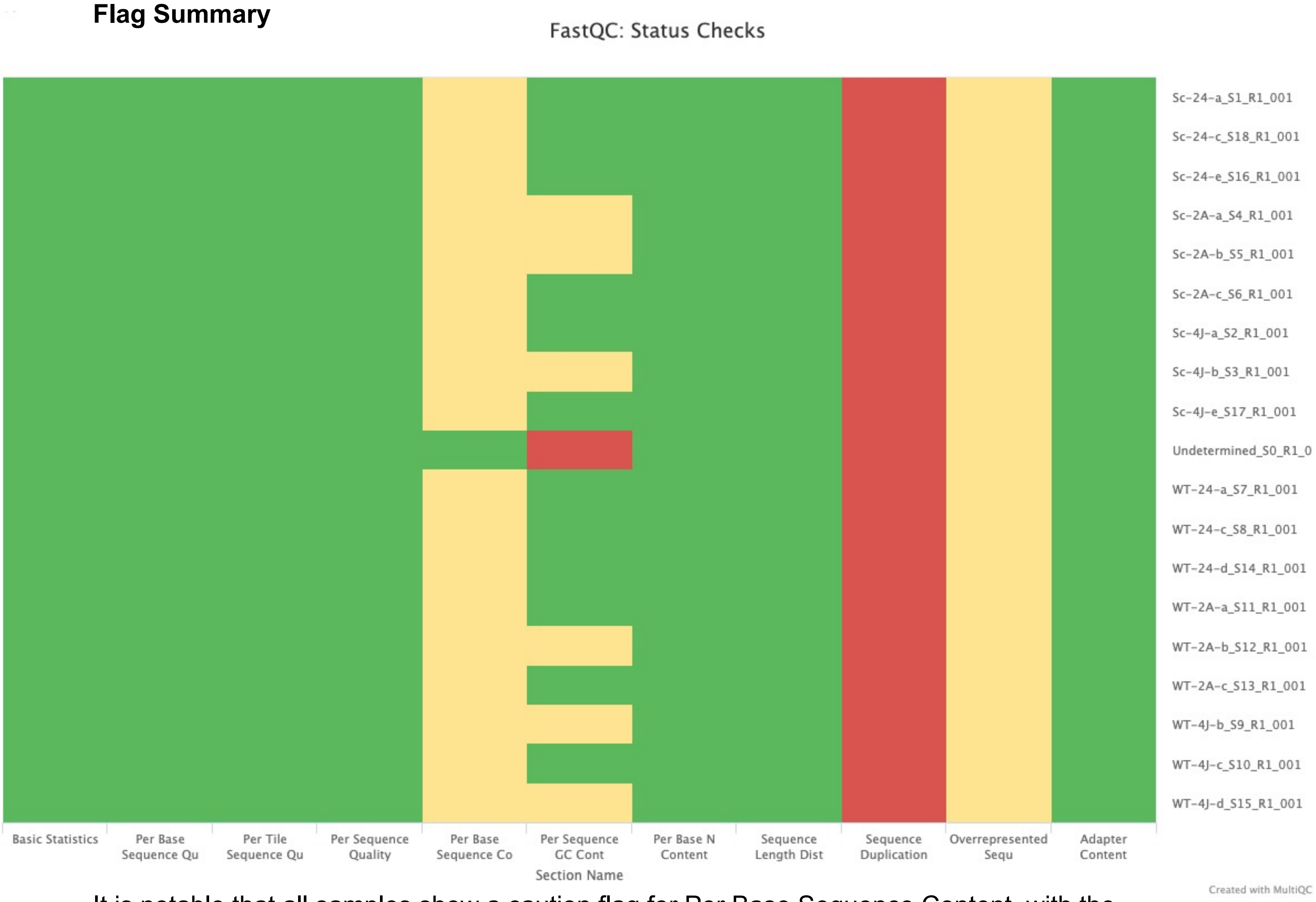


General Statistics

Showing $^{19}/_{19}$ rows and $^{6}/_{6}$ columns.

Sample Name	M Seqs	Average Read Length	Median Read Length	% Dups	% GC	% Failed
Sc-24-a_S1_R1_001	15.5	76 bp	76 bp	67.2%	44%	9%
Sc-24-c_S18_R1_001	16.1	76 bp	76 bp	69.6%	42%	9%
Sc-24-e_S16_R1_001	18.7	76 bp	76 bp	64.9%	44%	9%
Sc-2A-a_S4_R1_001	13.0	76 bp	76 bp	63.0%	45%	9%
Sc-2A-b_S5_R1_001	14.3	76 bp	76 bp	61.8%	44%	9%
Sc-2A-c_S6_R1_001	15.5	76 bp	76 bp	63.1%	46%	9%
Sc-4J-a_S2_R1_001	14.6	76 bp	76 bp	63.7%	46%	9%
Sc-4J-b_S3_R1_001	13.4	76 bp	76 bp	66.0%	46%	9%
Sc-4J-e_S17_R1_001	14.2	76 bp	76 bp	69.3%	43%	9%
Undetermined_S0_R1_001	10.7	76 bp	76 bp	65.7%	44%	18%
WT-24-a_S7_R1_001	13.4	76 bp	76 bp	62.2%	45%	9%
WT-24-c_S8_R1_001	16.3	76 bp	76 bp	65.1%	45%	9%
WT-24-d_S14_R1_001	16.5	76 bp	76 bp	66.7%	44%	9%
WT-2A-a_S11_R1_001	14.3	76 bp	76 bp	62.6%	44%	9%
WT-2A-b_S12_R1_001	17.6	76 bp	76 bp	66.6%	45%	9%
WT-2A-c_S13_R1_001	14.3	76 bp	76 bp	63.3%	44%	9%
WT-4J-b_S9_R1_001	17.0	76 bp	76 bp	68.9%	46%	9%
WT-4J-c_S10_R1_001	14.1	76 bp	76 bp	65.3%	46%	9%
WT-4J-d_S15_R1_001	15.7	76 bp	76 bp	66.4%	45%	9%

Flag Summary

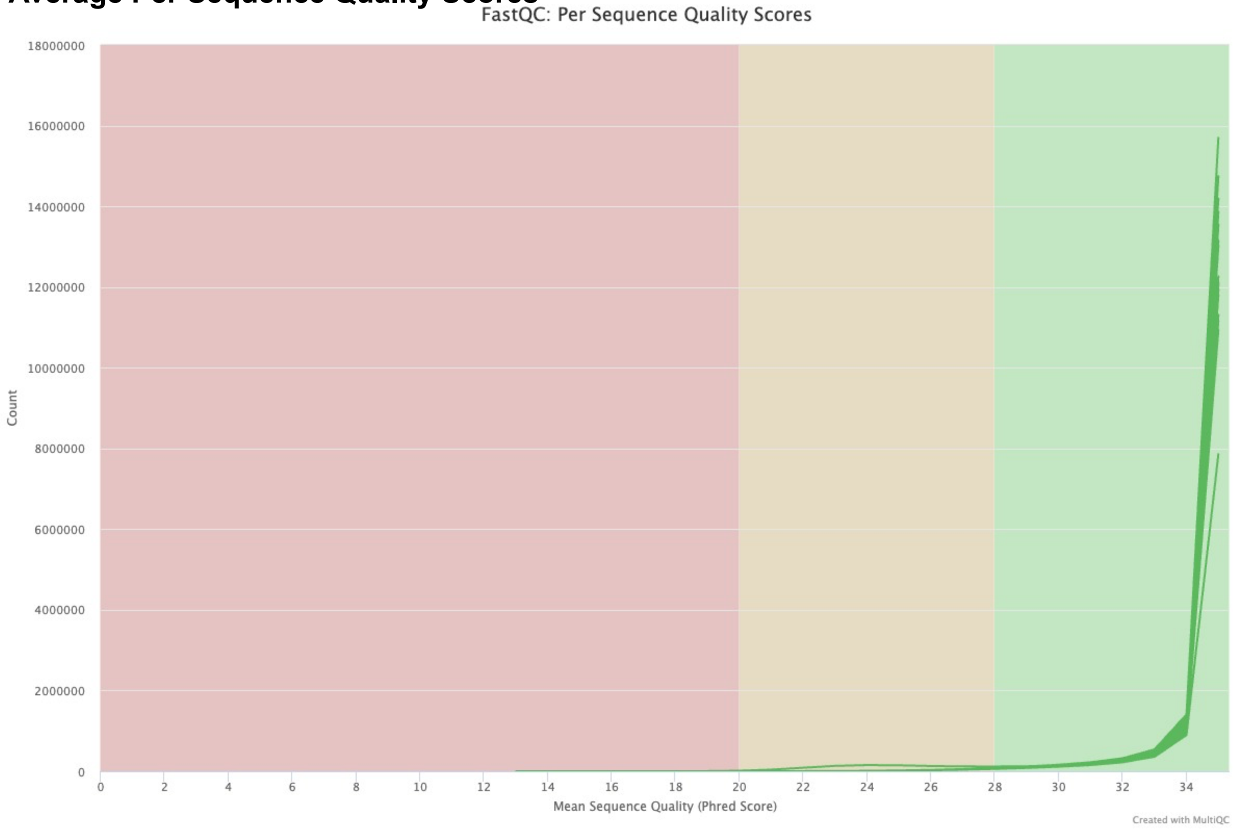


It is notable that all samples show a caution flag for Per Base Sequence Content, with the exception of the Undetermined Sample. “Failure” or flagging of this feature is likely due to library preparation methods. If you look at the summary graph (see below), we see that the sequence bias is in the first 11 bases at the 5’ end of the read. This is typical of Illumina library preparation methods because the “random” priming method is not exactly “random” and is biased toward kmers with better annealing. Per sequence GC content shows flags for 6 samples, however I do not know the expected GC content and the distributions look relatively normal and similar in amplitude across all samples, except for the undetermined sample which has a much higher and narrower peak. Sequence duplication flags are present, but this is almost always the case with RNASeq data. All samples had a moderately high number of overrepresented sequences, which may indicate low library complexity or bias introduced if a nonrandom fragmentation method used. This last scenario makes sense in combination with the fact all samples have an average read length of exactly 76 bp. Another possible source of overrepresented sequence content is the presence of ribosomal RNA due to errors in library preparation, contamination, or errors in filtering adaptors or other library components.

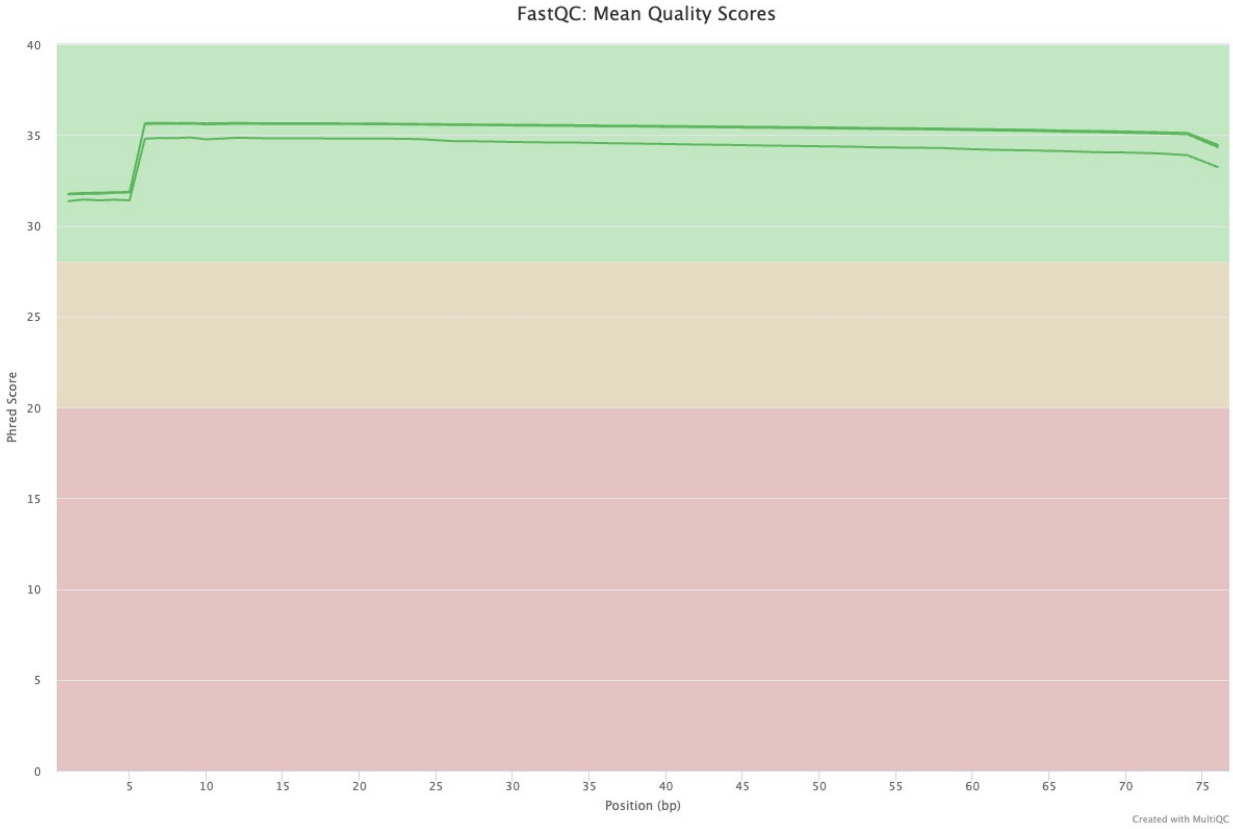
Per Sequence Quality Scores and Per Base Sequence Quality (Below)

Consistent across all samples, indicating high quality reads and reliable base calling.

Average Per Sequence Quality Scores



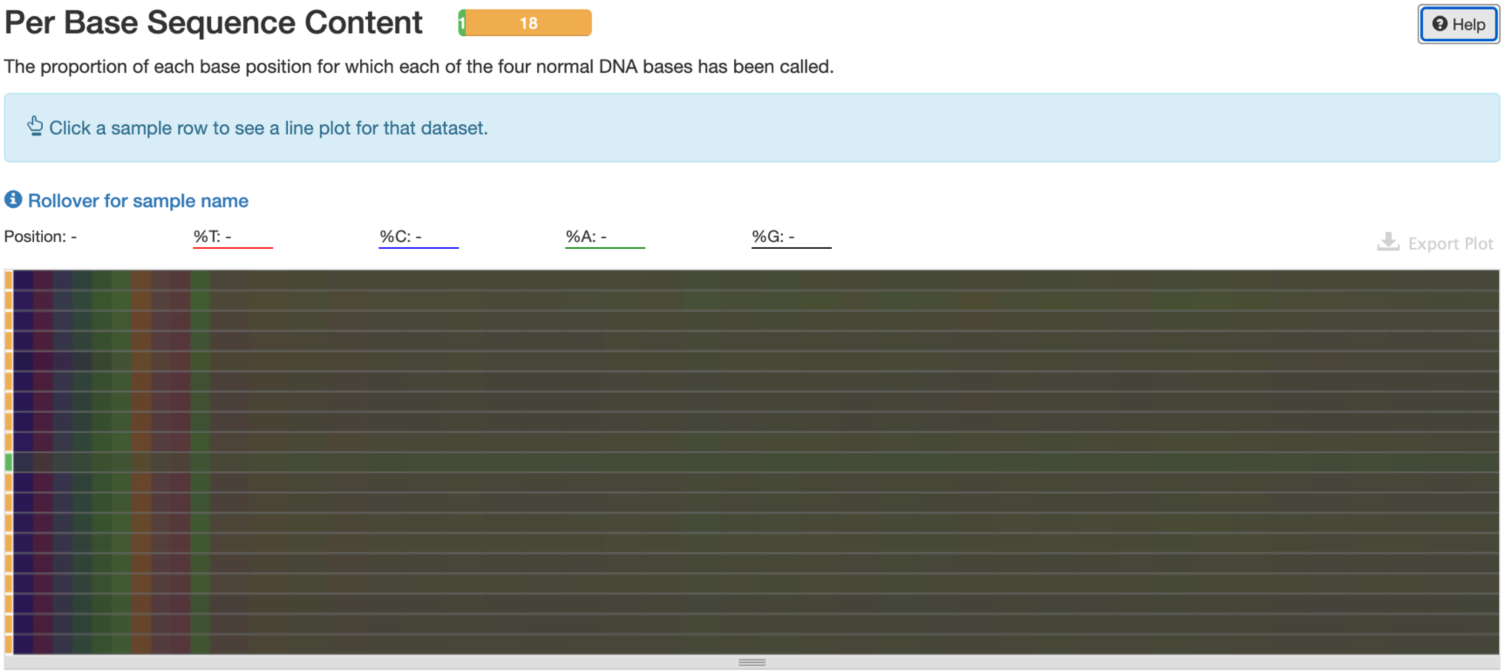
Average Per-Base Sequence Quality



Subset of Features with Flags or Sample Failures

Per Sequence Base Content

Consistent pattern for the first 11 nucleotides across all samples suggests bias due to library preparation method or primer annealing to reads. Will need to be trimmed from sequences in QC filtering. Unlikely it is due to adapters in reads, as adapter content was < 0.1% for all samples.



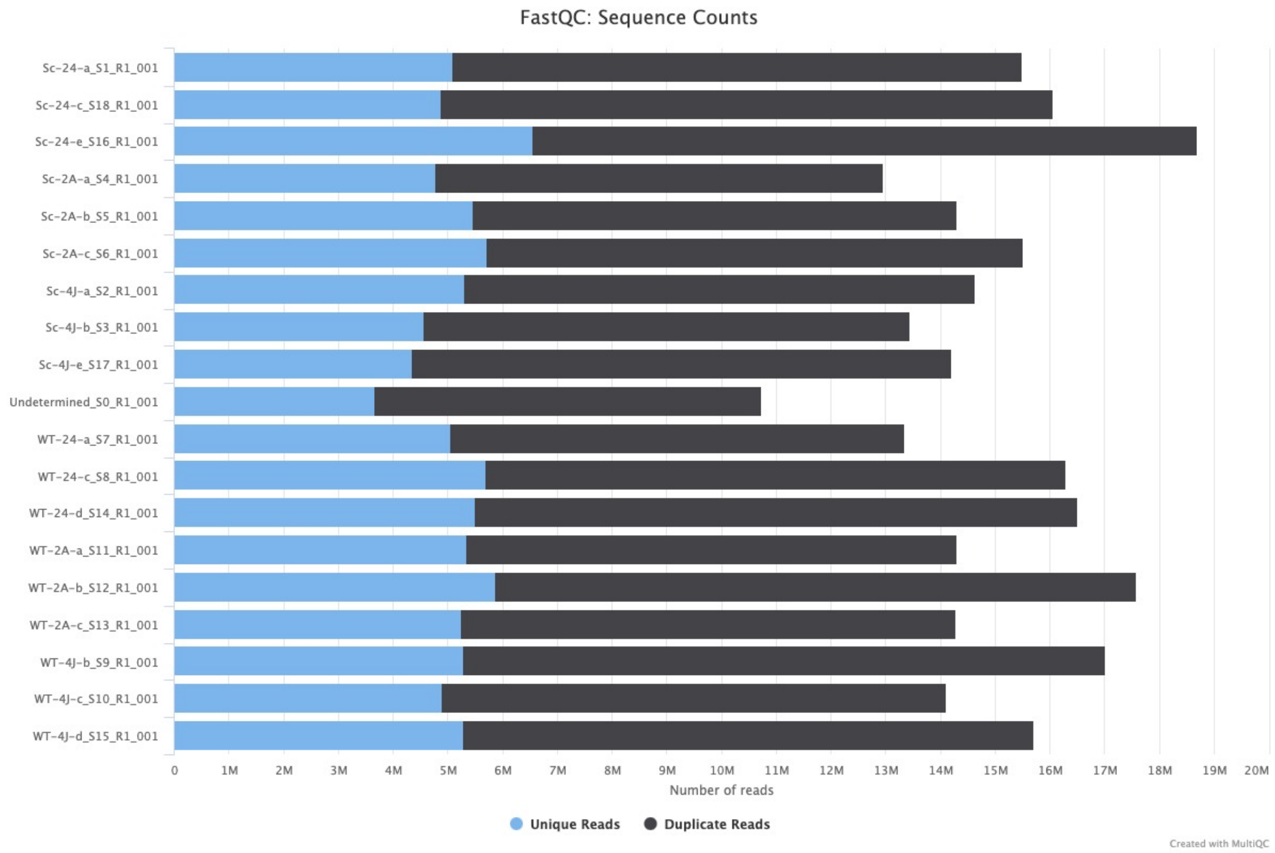
Per Base Sequence Content

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

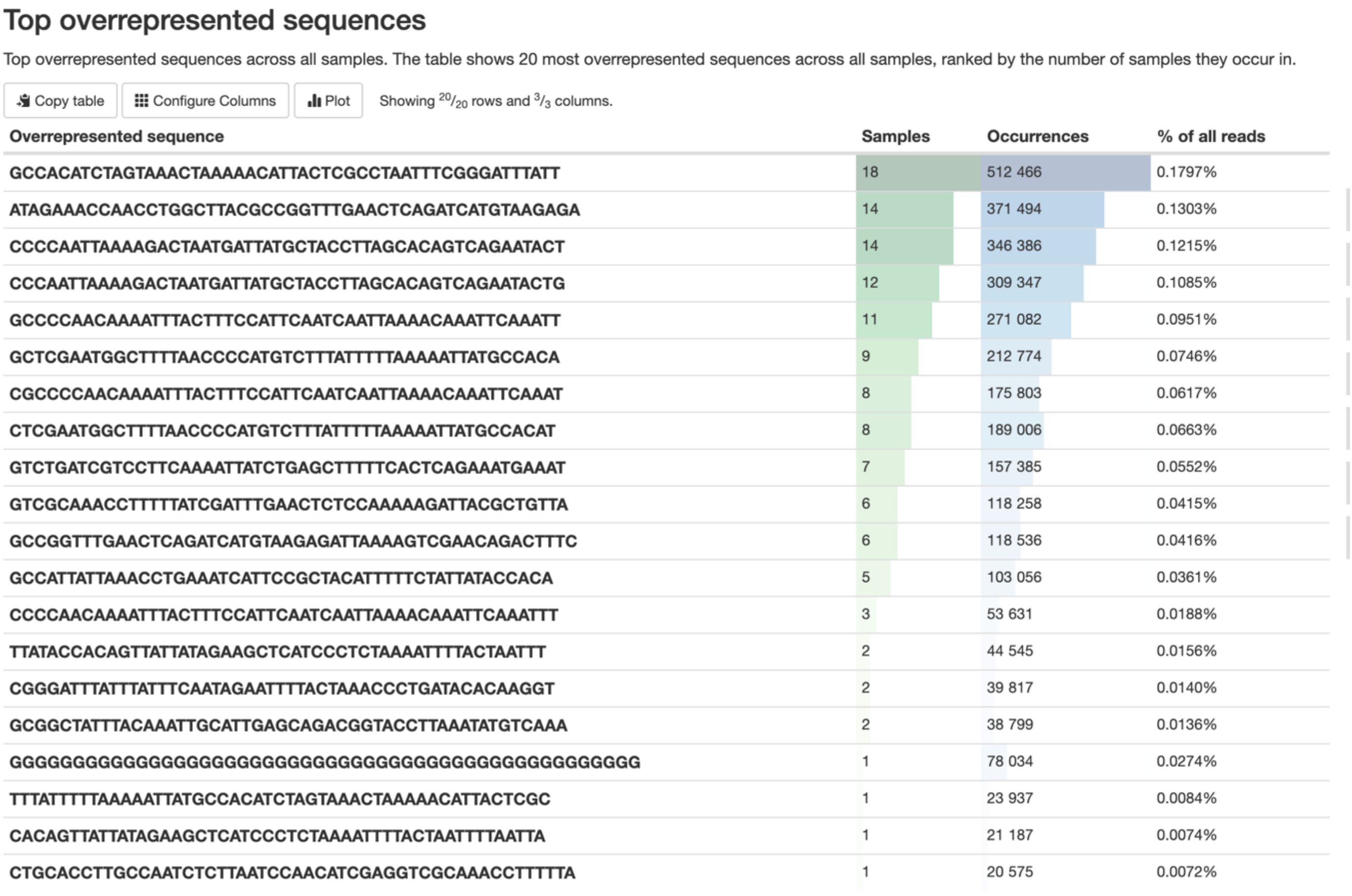
Click a sample row to see a line plot for that dataset.

Sequence Duplication Levels (below)

Consistent across samples, with duplicate reads ~ 60-70% of all reads total for each sample. Note the total number of reads for the Undetermined genotype is lower than the other 17 samples. Likely due to the nature of the data type, some possibility of PCR amplification bias.



One sequence was overrepresented in all 18 samples. This could be a biologically significant sequence, or contamination in all samples due to library preparation/quality of biological data.



Top overrepresented sequences

Top overrepresented sequences across all samples. The table shows 20 most overrepresented sequences across all samples, ranked by the number of samples they occur in.

Showing $^{20}/_{20}$ rows and $^{3}/_{3}$ columns.

Overrepresented sequence	Samples	Occurrences	% of all reads
GCCACATCTAGTAAACTAAAAACATTACTCGCCTAATTTCGGGATTTATT	18	512 466	0.1797%
ATAGAAACCAACCTGGCTTACGCCGGTTTGAACTCAGATCATGTAAGAGA	14	371 494	0.1303%
CCCCAATTAAAAGACTAATGATTATGCTACCTTAGCACAGTCAGAATACT	14	346 386	0.1215%
CCCAATTAAAAGACTAATGATTATGCTACCTTAGCACAGTCAGAATACTG	12	309 347	0.1085%
GCCCCAACAAAATTTACTTTCCATTCAATCAATTAAAACAAATTCAAATT	11	271 082	0.0951%
GCTCGAATGGCTTTTAACCCCATGTCTTTATTTTTAAAAATTATGCCACA	9	212 774	0.0746%
CGCCCCAACAAAATTTACTTTCCATTCAATCAATTAAAACAAATTCAAAT	8	175 803	0.0617%
CTCGAATGGCTTTTAACCCCATGTCTTTATTTTTAAAAATTATGCCACAT	8	189 006	0.0663%
GTCTGATCGTCCTTCAAAATTATCTGAGCTTTTTCACTCAGAAATGAAAT	7	157 385	0.0552%
GTCGCAAACCTTTTTATCGATTTGAACTCTCCAAAAAGATTACGCTGTTA	6	118 258	0.0415%
GCCGGTTTGAACTCAGATCATGTAAGAGATTAAAAGTCGAACAGACTTTC	6	118 536	0.0416%
GCCATTATTAAACCTGAAATCATTCCGCTACATTTTTCTATTATACCACA	5	103 056	0.0361%
CCCCAACAAAATTTACTTTCCATTCAATCAATTAAAACAAATTCAAATTT	3	53 631	0.0188%
TTATACCACAGTTATTATAGAAGCTCATCCCTCTAAAATTTTACTAATTT	2	44 545	0.0156%
CGGGATTTATTTATTTCAATAGAATTTTACTAAACCCTGATACACAAGGT	2	39 817	0.0140%
GCGGCTATTTACAAATTGCATTGAGCAGACGGTACCTTAAATATGTCAAA	2	38 799	0.0136%
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1	78 034	0.0274%
TTTATTTTTAAAAATTATGCCACATCTAGTAAACTAAAAACATTACTCGC	1	23 937	0.0084%
CACAGTTATTATAGAAGCTCATCCCTCTAAAATTTTACTAATTTTAATTA	1	21 187	0.0074%
CTGCACCTTGCCAATCTCTTAATCCAACATCGAGGTCGCAAACCTTTTTA	1	20 575	0.0072%