FASTQ QC Report

Report Date	04-28-17
Run ID	170427_D00796_0207_ACB1MMANXX
Project ID	EC-MA-4384
Sample	Sample_Ctrl_Input_R1
FASTX-Toolkit Version	0.0.13.2
FastQC Version	0.10.1
Dupest Version	0.1.0

This report was automatically generated by the WCMC Epigenomics Core QC pipeline and contains information for assessing the quality of FASTQ sequencing data.

The QC Pipeline executes the following analysis:

1. All FASTQ files for the sample are concatenated to a single file. For paired-end sequences, FASTQ files for each read are concatenated and processed separately, with an "R1" or "R2" appended to the sample name.
2. To identify genomic sequencing bias or low sequence diversity k-length oligonucleotide enrichment is calculated and plotted from the combined FASTQ file using FastQC. *Note:* FastQC only analyses the top 2% of the reads in the FASTQ file and the results are extrapolated over the remainder.
3. Duplication level is estimated from the combined FASTQ file as $(N - U)/N$ where N is total reads and U is the number of unique sequences.
4. Sequencing base call quality statistics are calculated from the combined FASTQ file using FASTX-Toolkit FASTQ Quality Filter.

The report contains the following figures:

1. Sequence Duplication - Estimate of duplication level as a percentage of total reads.
2. Base sequence quality - Calculated from FASTX-Toolkit FASTQ Quality Filter.

   Distribution of base quality scores (Q scores) per sequencing cycle. In a reasonably good sequencing run the majority of the signal should be above Q30. Quality scores are divided into three ranges: green indicates calls of very good quality; orange indicates calls of reasonable quality and red indicates calls of poor quality.
   Yellow boxes represent the inter-quartile range. Upper and lower whiskers represent the maximum and minimum excluding outliers. The red line represents the median quality and the blue line represents the mean quality.
3. Sequence base content - Percentage of bases represented at each position in the read; calculated from FASTX-Toolkit FASTQ Quality Filter.
4. K-mer content - calculated and plotted by FastQC. From FastQC Help:

   The k-mer analysis checks if there are short fragments of k-length nucleotides that are over represented at certain positions in the reads. In a diversified library there should not be positional bias in its appearance of k-mers. There may be biological reasons why certain k-mers are enriched or depleted overall, but these biases should affect all positions within a sequence equally. In contrast, if certain k-mers are over represented in certain positions then this could indicate issues with library preparation, quality of the input material or sequencing of the adaptors. This analysis measures the number of each 5-mer at each position in the library and then uses a binomial test to look for significant deviations from an even coverage at all positions. Any k-mer with positionally biased enrichment are reported. The top 6 most biased k-mers are additionally plotted to show their distribution. Note that because of the computational overhead associated with calculating k-mer content this analysis is performed on 2% of the reads.
5. Overrepresented sequences - Calculated and plotted by FastQC. From FastQC Help:

   A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

   This analysis lists all of the reads which make up more than 0.1% of the total. To limit memory use only sequences which appear in the first 200,000 sequences are evaluated for their occurrences in the entire library. It is possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason

could be missed by this analysis. However, this is unlikely since library preparation and sequencing randomize the genomic elements and therefore the first 200,000 reads are sufficient to represent the diversity in the entire library.

For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit doesn't necessarily mean that this is the source of the contamination, but may provide clues about the true source of contamination. It's also worth pointing out that many adapter sequences are similar in sequence so a match to an adaptor sequence may not represent the true source of the adaptor.

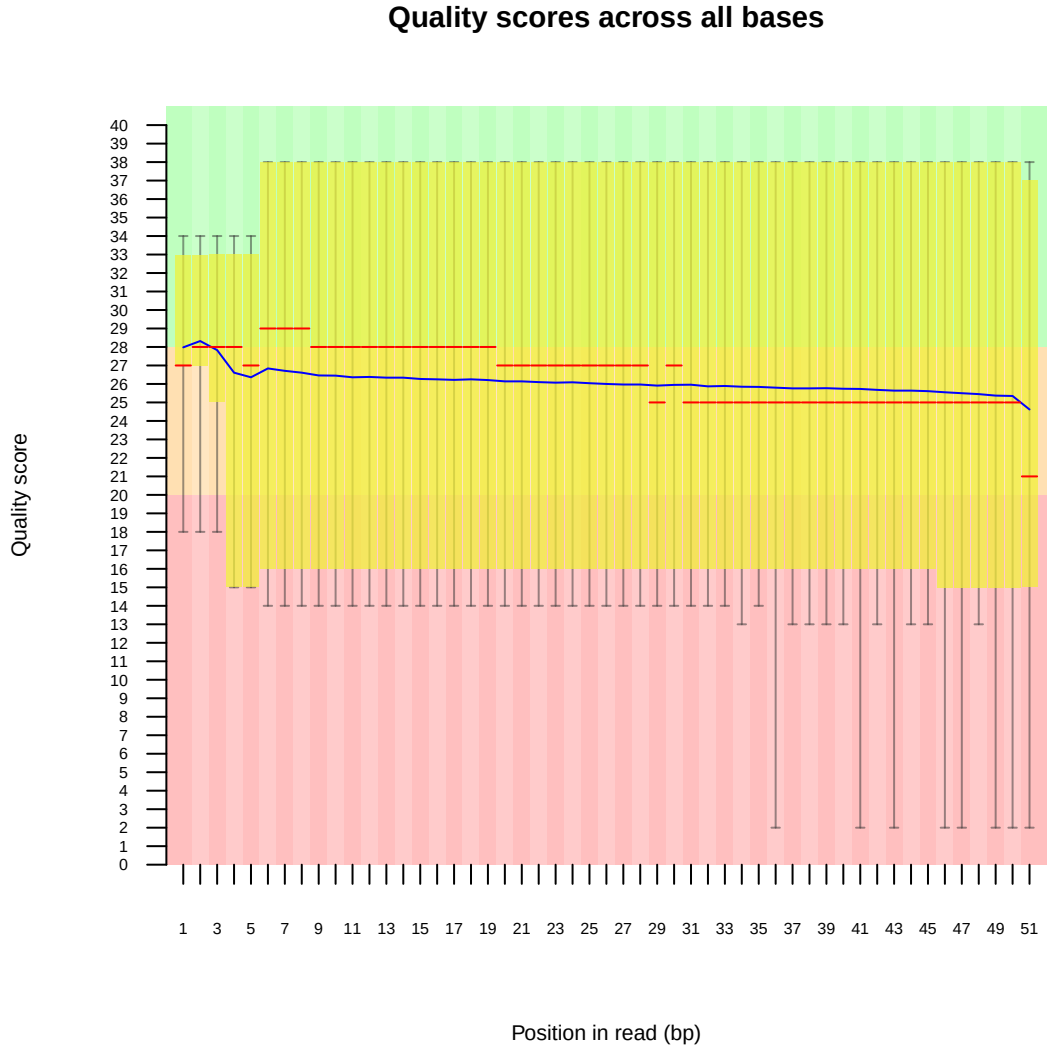
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

`FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc`
`FASTX-Toolkit:  http://hannonlab.cshl.edu/fastx_toolkit`

1 Sequence Duplication

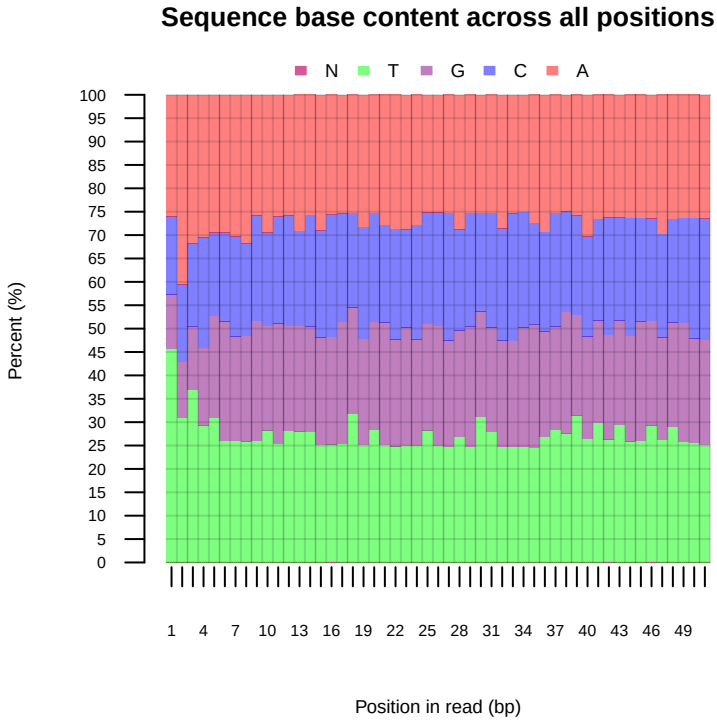
- Estimated Duplication rate 5.5581%

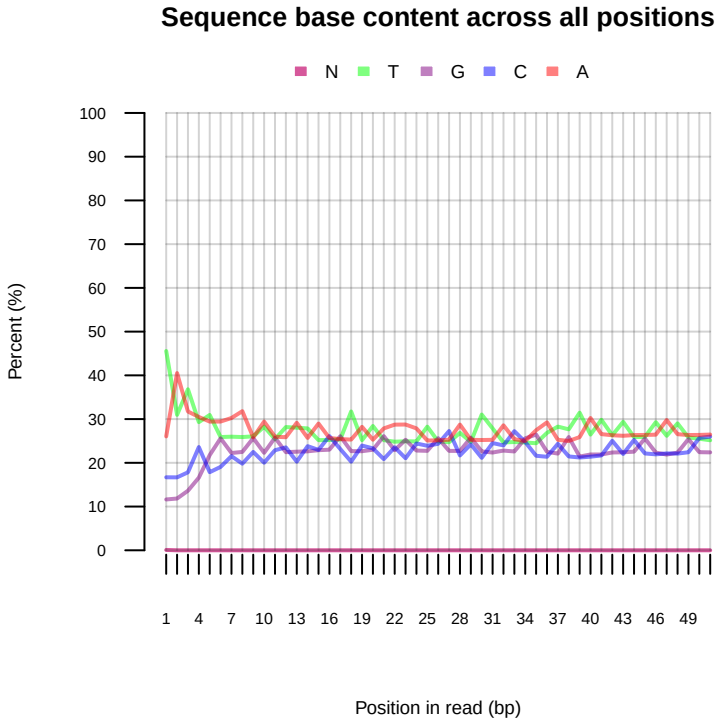
2 Per base sequence quality



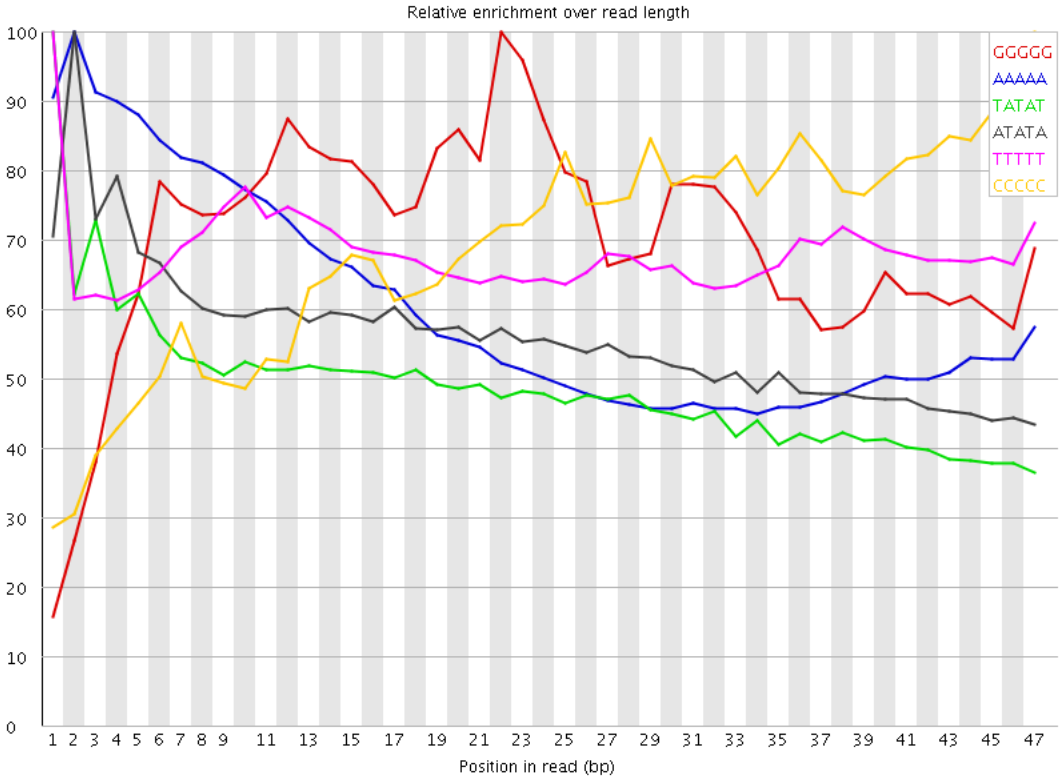
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

3 Sequence base content





4 Sequence K-mer content



Note: FastQC analyses 2% of the sequence data and results are extrapolated to the rest of the sequence.

Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
GGGGG	1083310	4.3108253	6.1680503	22
AAAAA	2539890	3.8181639	6.3096457	2
TATAT	2518550	3.764821	7.693725	1
ATATA	2481910	3.7170212	6.6118965	2
TTTTT	2465500	3.6717083	5.38136	1
CCCCC	899870	3.668275	5.3253937	47
CACAC	1045605	2.8599129	47.39281	12
GGAAG	963040	2.5962396	47.218113	5
CTCCA	848810	2.3172896	46.086266	24
AGAGC	848870	2.2995167	46.716904	8
GCACA	833650	2.2692065	46.606335	11
GAGCA	825445	2.2360604	46.575584	9
TCCAG	821330	2.231478	45.704884	25
AGCAC	809895	2.2045448	46.530853	10
CACCG	649710	2.159028	54.093296	31
AAGAG	971260	2.1551743	38.708763	7
GTCTG	791610	2.1363678	46.187946	17
GAAGA	944055	2.0948079	38.796886	6
ACTCC	767055	2.0940948	45.985546	23
TCACC	754350	2.0594096	45.097645	30
TATTT	1357320	2.0251667	5.839212	1
CCAGT	731130	1.9864129	45.5906	26
CAGTC	710225	1.9296162	45.477596	27
TCTGA	844120	1.8750609	38.058517	18
TAAAA	1248110	1.872739	5.4575624	1
ATGCC	686000	1.8637989	44.85775	47
CTGAA	826450	1.8392596	37.974705	19
GTCAC	673135	1.8288459	45.38757	29
ACACG	668860	1.8206459	45.948124	13
TGTAT	1002420	1.8205268	30.257622	37
CGTCT	650180	1.7631662	45.982403	16
ATGTA	967885	1.7611097	30.192616	36
ATCTC	787450	1.7576365	36.719425	40
CACGT	641010	1.7415652	46.008793	14
CGGAA	629160	1.7043408	46.767193	4
AACTC	757995	1.6950701	37.76209	22
TATCT	926580	1.6909283	29.944717	39
TCGGA	611975	1.654679	46.8917	3
TGAAC	738430	1.6433717	37.667717	20
TCTCG	604025	1.6380025	44.108334	41
AGTCA	734560	1.6347591	37.23678	28
ATCGG	597460	1.6154329	47.447666	2
ACGTC	592385	1.6094557	45.89068	15

GATGT	719725	1.5910466	36.264446	35
TATGC	713150	1.5841346	36.558773	46
GAACT	711645	1.5837619	37.592934	21
GATCG	584495	1.5803776	45.73834	1
GTATG	700595	1.5487572	36.419353	45
CTCGT	564295	1.5302622	44.025997	42
CGATG	555255	1.5013176	43.797997	34
ACCGA	543510	1.4794414	44.04641	32
CCGAT	513950	1.3963548	43.905064	33
GTATC	609620	1.3541613	35.777897	38
CGTAT	536685	1.1921493	36.248657	44
TCGTA	521235	1.1578299	36.15793	43

5 Overrepresented sequences

Note: FastQC tracks sequences that appear in the first 200,000 reads to the end of the file.

Sequence	Count	%	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCC	314111	3.4535347080727425	TruSeq Adapter, Index 2 (100ATATATATAT
10447	0.11486091571207611	No Hit	