FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD11_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.

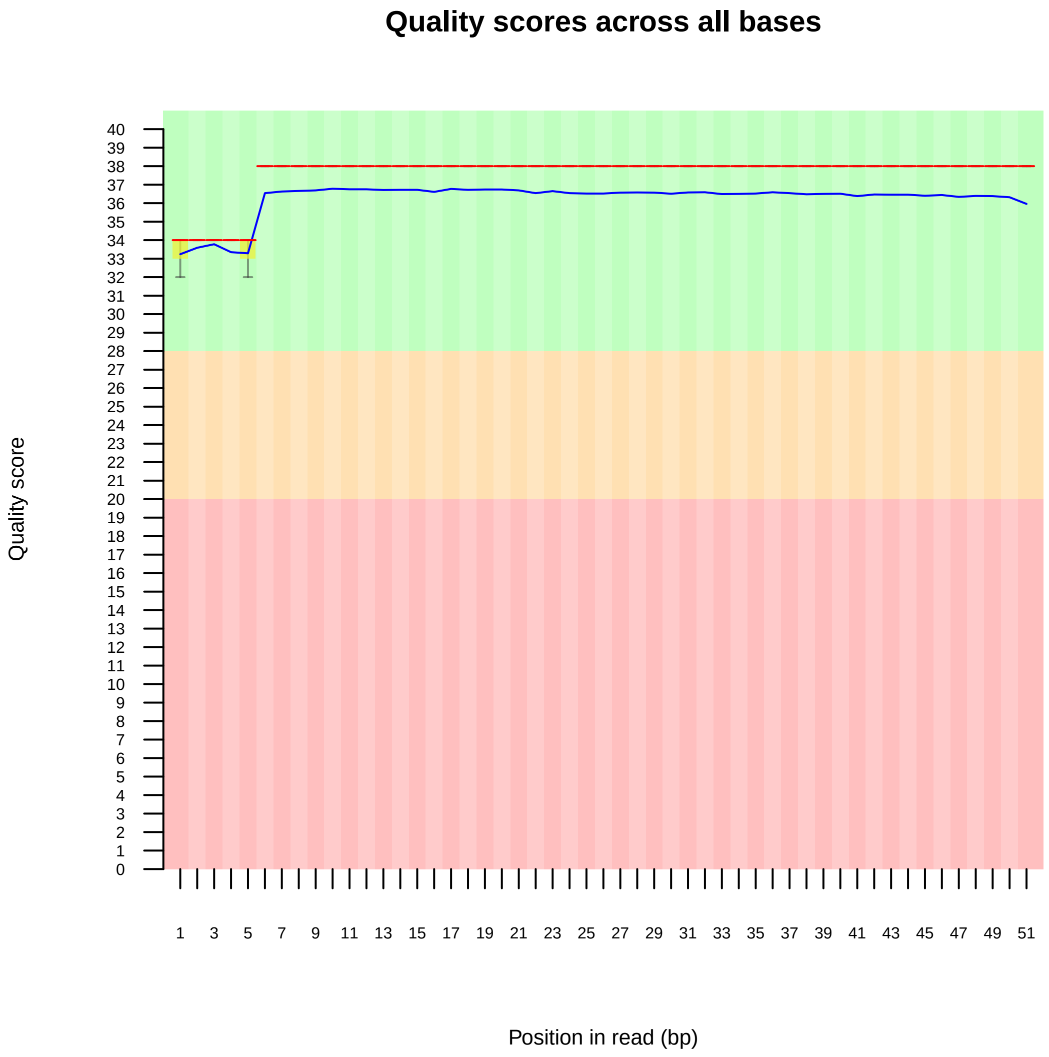
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

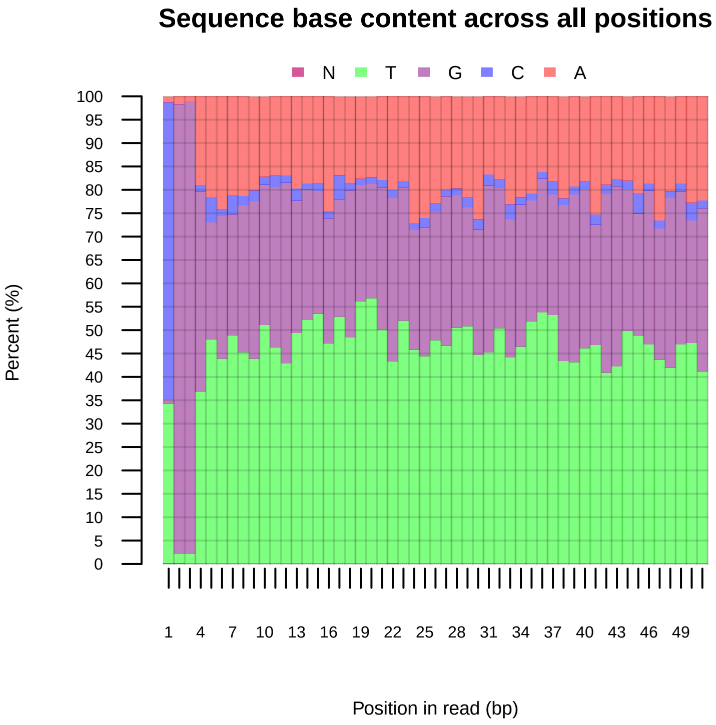
- Estimated Duplication rate 80.5260%

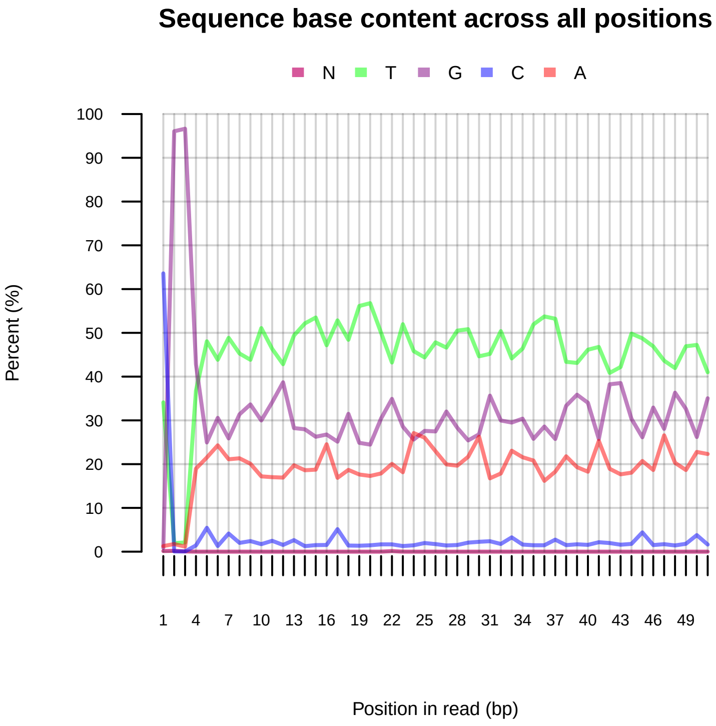
2 Per base sequence quality



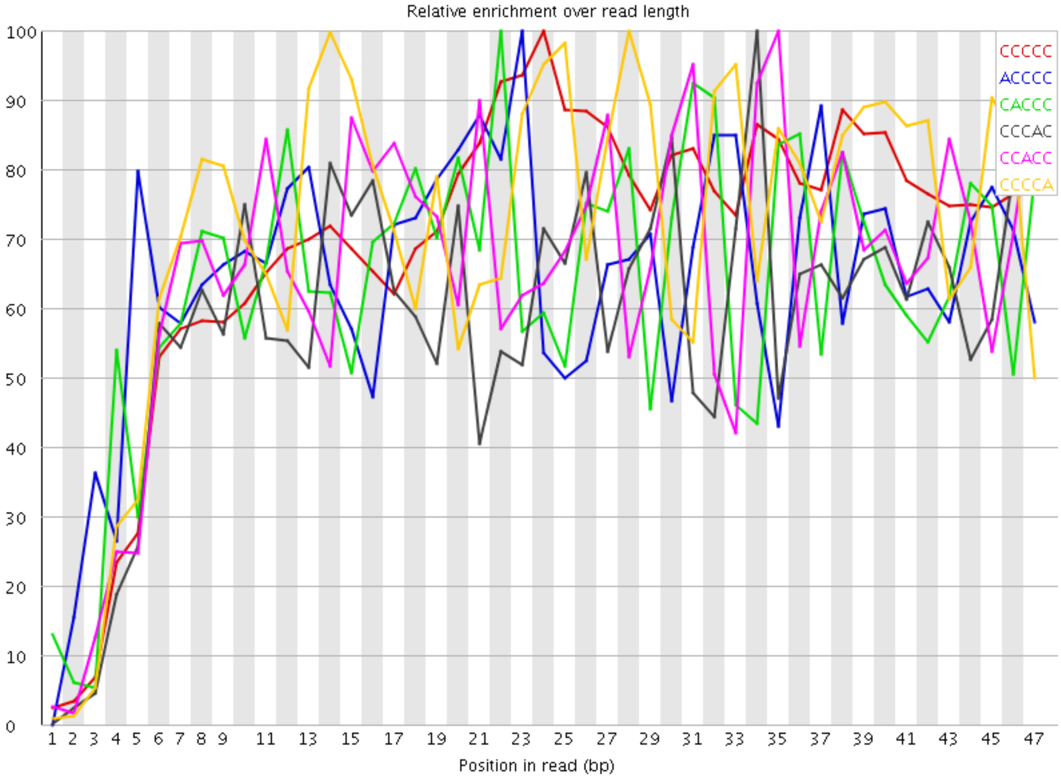
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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
CCCCC	193720	1654.041	2371.0728	24
ACCCC	54025	76.732506	119.12685	23
CACCC	50935	72.343735	113.94053	22
CCCAC	49575	70.41212	120.12758	34
CCACC	49265	69.97182	107.44743	35
CCCCA	49140	69.79428	98.43788	28
CGGGC	3498715	28.824144	1145.454	1
CCCGC	23110	19.498316	37.06838	47
GCCCC	22845	19.274729	32.707367	45
CCCCG	22785	19.224108	36.671146	25
CGCCC	22365	18.869743	33.500214	44
CCGCC	22005	18.566008	35.08536	27
CACAC	70455	16.645979	144.93362	47
CGCGG	1958440	16.134596	434.29968	5
GCGCG	1852920	15.265272	431.92288	4
GGCGC	1462205	12.04637	433.50388	3
CGGCG	1457580	12.0082655	356.1564	1
CGCGC	127155	10.601221	60.936573	27
CGGAA	3594835	8.293297	282.83093	1
CGGGA	5958095	8.165226	283.14026	1
CGGGT	11779425	6.834247	271.93018	1
CGGAG	4849835	6.646419	205.20576	1
TCGCG	1131545	6.6437473	39.869404	30
AGACG	2738415	6.317534	72.53988	27
TCCCC	10345	6.220439	24.170534	3
CGGGG	7244775	5.8979096	153.49843	1
CGGTT	14038280	5.804621	211.41104	1
CGCGT	971345	5.7031484	37.823715	31
ACGCG	405250	5.620285	20.338303	4
AGGCG	4003540	5.486621	98.750824	47
AACCC	22160	5.2356105	7.948273	20
CTCCC	8675	5.2162695	9.747554	11
TCGAG	5339340	5.2148485	84.2869	44
CGTCG	870075	5.108554	15.70934	41
CCTCC	8415	5.0599318	9.182467	31
CCCCT	8285	4.981763	8.899932	38
CCCTC	8255	4.9637237	9.619194	22
GAGAC	2047885	4.724478	69.38323	26
CGAGG	3414695	4.6796427	111.79407	45
CGGTC	792720	4.6543713	172.66168	1
CGGAC	333705	4.628051	153.87863	1
CAACC	19570	4.6236863	8.492676	31
GGGCG	5595560	4.555298	118.6603	2

ACACC	18875	4.459483	6.4475503	21
ACCCA	18465	4.3626146	7.049475	33
ACCAC	18290	4.321268	6.938604	45
CCAAC	18080	4.271653	7.7710757	30
CCCAA	18075	4.2704716	7.3270144	14
CGCGA	302065	4.1892447	22.804834	5
CCACA	17585	4.154702	6.8274446	35
TTACG	5903250	4.109024	68.67253	14
CGGTA	4183740	4.0861926	146.66188	1
CACCA	17195	4.062559	6.8274446	16
AAAAA	3729510	4.055907	8.036257	31
GACGG	2757025	3.778344	42.421246	28
TACGT	5335860	3.7140853	68.87187	15
ACGTT	5274345	3.6712673	69.97015	16
ACGGG	2661530	3.6474733	42.670128	29
CGACG	262895	3.6460083	15.939599	24
GTCGA	3718715	3.6320097	83.67789	43
CGGAT	3614190	3.529922	125.640564	1
GAGGC	2430770	3.3312302	78.66842	46
CGTTT	11195965	3.2992473	38.714367	17
GGCGG	4024580	3.276377	20.083687	11
CGGTG	5245800	3.043535	56.45578	1
CGAGA	1307560	3.0165453	38.758827	25
AAGCG	1306385	3.0138352	57.987698	8
ATCGC	302690	2.9917605	57.001026	29
AGAGA	7781635	2.98629	23.311111	25
GGCGT	5001955	2.9020603	46.130554	3
AGCGA	1235170	2.8495414	58.77794	9
TTTCG	9314335	2.744765	13.778061	30
GGTCG	4686720	2.7191658	51.21678	42
GGAGG	19997495	2.7080815	32.142403	39
TTCGA	3886020	2.704908	29.208796	31
GCGGC	320420	2.6397784	8.394461	33
AGGTC	2619350	2.5582776	79.82342	41
TTTTT	170261115	2.5181463	5.23771	16
GGGAG	18519290	2.5079012	30.106833	38
ACGGA	1083295	2.4991653	11.146887	30
GCGGT	4116400	2.388274	41.436035	6
CGTTA	3346250	2.329195	19.822754	9
GCGGA	1692315	2.31922	24.614393	7
GACGC	166155	2.3043518	20.317135	3
CGTAG	2350900	2.2960868	26.35686	5
CGTTC	540880	2.2632656	20.179743	33
GCGGG	2771905	2.2565844	20.4243	12
TTTTA	63714235	2.2258508	10.75349	26
AGTAG	13585425	2.2071927	17.86308	35
TTTAG	44598545	2.186185	14.187988	27
TTTAC	4393350	2.1793988	47.967896	13
ACGGC	157065	2.1782856	16.271944	12
CGAGT	2230020	2.178025	36.75421	33
TTCGC	518145	2.168133	5.3302755	13
TACGG	2188965	2.1379273	21.486998	5
GAGGT	22094500	2.1323755	23.332521	40
ATTTT	59753545	2.0874846	8.995969	25
GTCGC	351075	2.0612996	10.744827	3
AGCAC	86750	2.0253139	15.742352	47
GGAAT	12399580	2.0145314	13.880831	2
AGCGC	145215	2.0139415	14.848067	35
TCGTC	479300	2.0055895	7.099046	40
TAGAG	12215525	1.9846281	10.756775	24
ATCGT	2847630	1.9821249	16.316557	39
AAACG	498615	1.9364268	5.8476667	10
AATTT	23461675	1.936033	18.929024	24
TAGTA	16692910	1.9328246	21.70485	29
AGGAG	8467735	1.9303715	6.180502	38
TACGC	194810	1.9254844	7.1265807	25
GGAAG	8413530	1.9180146	12.391631	2
CGAGC	137255	1.9035468	8.344523	7
ATTCG	2695820	1.876456	21.445438	34
GCGTT	4527975	1.8722507	12.900417	16
GCGTA	1893260	1.8491169	25.922718	4
TGGAA	11243720	1.8267412	9.140102	1
AACGG	791430	1.825832	10.152308	29
TATCG	2607680	1.8151052	16.695486	38
ACGTC	183300	1.8117206	6.2107186	21
GAAAA	2801830	1.8100494	5.309996	3
GGACG	1320255	1.8093333	15.881226	2
GGAAA	4647975	1.7837129	13.44568	2
GAGCG	1293495	1.7726604	8.832299	28
AGATC	1076650	1.7701749	20.97476	27
TTAGT	35716565	1.7507975	13.546047	28
GAACG	755565	1.7430913	10.41724	3
GTAGA	10690730	1.7368982	10.432168	23
GTACG	1769315	1.7280618	21.082228	4
CGATT	2471965	1.7206391	20.154863	11
ACGGT	1756415	1.7154626	20.720646	6
TGGGA	17641415	1.7026011	19.230442	37
GGAGA	7460520	1.7007589	11.11729	2
AGTCG	1725010	1.6847899	16.670963	22
GCACA	71500	1.6692789	14.162625	46
TCGAA	1012180	1.6641765	5.1989803	44
TAGTT	33127440	1.6238807	6.818399	25
CGAAA	416900	1.6190776	6.408058	39
ACGAG	701595	1.618582	6.315449	32
CGTGG	2775355	1.6102197	26.131794	5
AGCGT	1646270	1.6078857	6.8042192	29
TATTT	46010650	1.6073778	7.733395	32
TAATT	19324175	1.5946108	18.680769	23
GATCG	1617045	1.5793421	13.62496	28
GTCGT	3802665	1.5723457	9.310353	3
AGGTA	9653575	1.5683941	17.642183	47
GGTTT	53792780	1.5664002	9.730029	2
GAGAT	9581680	1.5567132	9.943523	26
GCGAT	1589630	1.5525664	26.005611	10
GCGTC	263305	1.5459673	9.018654	4
GGCGA	1125695	1.5427	10.73619	2
AGTTT	31434310	1.5408849	6.986529	26

GGGAA	6757895	1.5405829	14.101358	2
AACGC	65070	1.5191607	12.999519	11
CGTAC	153195	1.5141655	7.2287536	13
GCGTG	2568260	1.4900663	26.099081	4
CGAAG	633310	1.4610484	6.077093	45
AGGTT	21163050	1.4556304	11.672953	41
ATGGA	8933260	1.4513661	6.403802	10
TTAAG	12388900	1.4344755	9.676655	6
GGTTA	20785595	1.4296684	19.40518	2
TTTAA	17264610	1.4246576	7.6135817	5
AGTTA	12303045	1.4245347	8.822867	30
GTTTA	28733190	1.4084779	10.519166	12
AATGG	8662805	1.4074259	6.430826	9
GAATG	8630525	1.4021815	6.3198085	18
ACGAA	355785	1.3817306	6.6295915	38
TCGTG	3333890	1.3785142	9.4053545	40
GTTAA	11903675	1.3782926	22.505398	3
AAGGC	594300	1.3710521	28.842894	46
AAGTA	4996965	1.3666612	8.614859	34
AGATA	4975715	1.3608493	5.85438	26
TATAG	11713280	1.3562474	13.176816	47
ACTCC	13525	1.352822	43.02774	23
TAAGC	820495	1.3490174	40.05648	7
GGAGT	13906800	1.3421675	11.12708	2
GGGTT	32792995	1.3398817	14.510244	2
CTCCA	13380	1.3383186	42.863125	24
TTATA	16178115	1.3350012	9.788244	46
GGTAG	13831385	1.334889	7.989849	2
GTAAT	11486640	1.3300054	24.488556	22
TTTTG	63932625	1.3267668	5.1904707	34
GACGT	1353335	1.3217808	5.8632793	3
GCGAC	94925	1.3164852	9.387187	23
GTAGT	19076220	1.3120947	8.060222	36
TTCGG	3097780	1.2808862	11.878076	35
TTGTA	26012505	1.275112	13.007381	20
TGGCG	2186930	1.2688242	17.14053	10
AGTAT	10854395	1.2567996	20.85269	30
CGAAC	53510	1.2492744	11.046862	9
GGGGA	9170880	1.2419301	11.500991	2
GAGTA	7519725	1.2217124	12.834521	34
GGGAT	12527190	1.2090191	10.309757	42
ATTAT	14607940	1.2054319	9.57217	45
CGTGT	2891300	1.1955098	8.995905	41
GTATT	24137750	1.1832132	9.35192	31
TGTAA	10200165	1.1810479	24.23841	21
CGTAA	713285	1.172748	5.11459	21
GGGGT	20377380	1.1682622	9.292628	2
GGTGG	20175015	1.1566604	11.667904	8
TTTGT	55289115	1.1473918	6.4602103	19
TTGAG	16425935	1.1298037	8.695439	44
TTAAT	13552005	1.1182973	14.663333	4
GGGTA	11537925	1.1135435	15.578565	2
GATTA	9591935	1.1106229	12.790794	44
GGATT	16129865	1.1094395	7.7067823	43
AGTAA	4023675	1.100468	6.7194667	9
ACACG	46880	1.0944867	10.805495	13
GGAAC	471090	1.0868064	10.525886	2
TCGGG	1854320	1.0758488	15.603615	36
TGCGG	1846310	1.0712016	10.274236	5
TTTTC	5037250	1.0578893	9.017739	29
AAGGT	6509380	1.0575638	5.3149447	46
TGGAG	10895795	1.0515705	8.976235	1
TTGGG	24966580	1.0201038	8.691578	36
TGAGG	10548085	1.0180125	11.179756	45
TAAGG	6229175	1.0120397	7.4848404	45
GGTAC	1035325	1.0111855	21.224257	3
ATTAC	854515	1.001278	6.818935	29
AAAAC	153120	1.0010487	5.113923	22
TTATC	1930170	0.95749485	11.635344	37
TTTGG	32797890	0.9550467	6.6648226	35
GGATA	5875240	0.95453674	8.218271	2
TAAGT	8225150	0.95236677	6.2318563	7
TGGGG	16532740	0.9478439	8.422199	1
CGTGA	969995	0.9473787	5.806522	26
TGGTT	32053310	0.93336517	6.6905417	1
AGGTG	9553860	0.9220583	5.8218794	47
AGTTG	13391555	0.9210938	8.042416	38
GGTTG	22500605	0.9193472	5.188444	42
GGAGC	666110	0.91286534	7.7714067	27
CGATC	92355	0.91282845	7.8999953	44
GTGCG	1528875	0.88703054	9.699969	4
GTGGT	21603440	0.88269013	9.724959	9
GTTTG	29894660	0.8705071	6.679357	18
TAGAC	527155	0.8667222	6.72583	25
AGTGA	5307165	0.8622428	6.3011184	18
GGGTG	14897400	0.8540877	8.928504	2
GGGGG	10550165	0.84870726	6.650241	2
GGTAT	11841280	0.8144633	6.447531	2
GTTGA	11702405	0.80491114	8.48131	43
GGTAA	4896835	0.7955775	6.755346	2
AGTGG	8123820	0.78404284	6.325549	8
TATTC	1558255	0.7729998	12.463052	33
AAAGC	196395	0.76272184	5.1806917	9
TGGGT	18533480	0.7572553	8.500278	1
GAGTC	769480	0.7515389	15.177688	21
GGGGC	883135	0.7189528	5.0777307	2
GTGGC	1192380	0.6918012	15.991041	9
GGGAC	495240	0.6786979	5.2595487	2
TGGTG	16000345	0.6537545	5.007465	1
AAGTC	362245	0.59558535	5.8653917	41
GGTGC	873415	0.5067425	9.804164	3
AACTC	14090	0.2344376	7.4763374	22
CACGG	16165	0.22418734	5.695506	31
ACATC	12275	0.20423858	6.985483	38
CTACA	11360	0.18901429	6.879937	36

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
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	202945	0.26968149345915576	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	134454	0.1786678928850542	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	108404	0.1440516032272109	No Hit