FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD7_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`

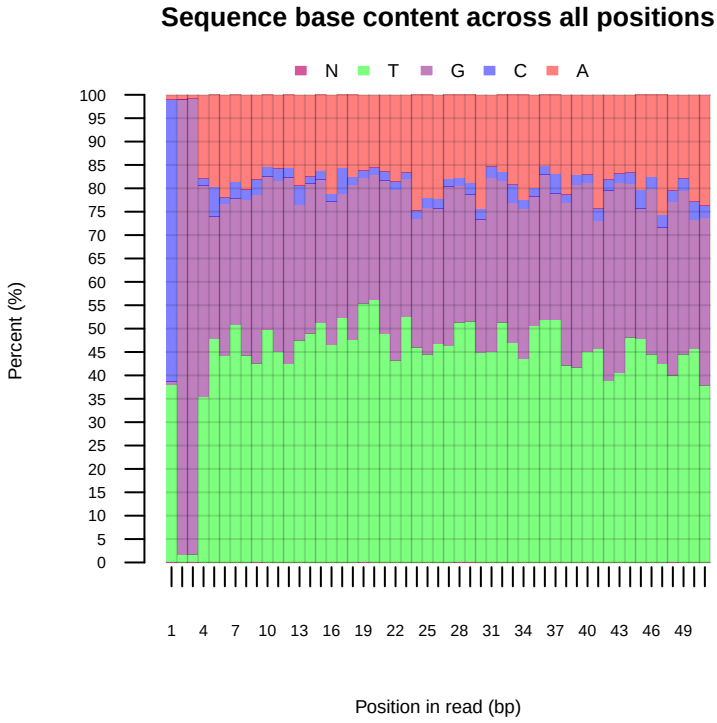
1 Sequence Duplication

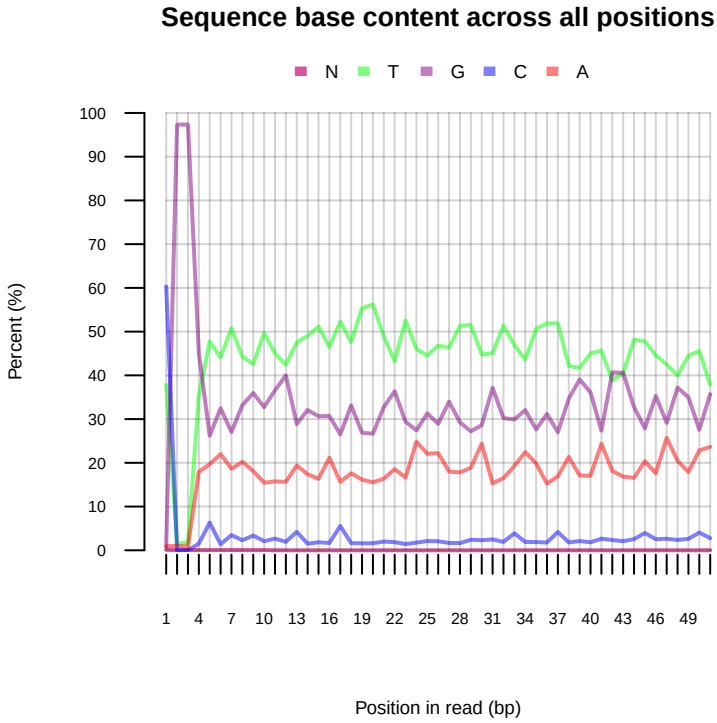
- Estimated Duplication rate 75.5559%

2 Per base sequence quality

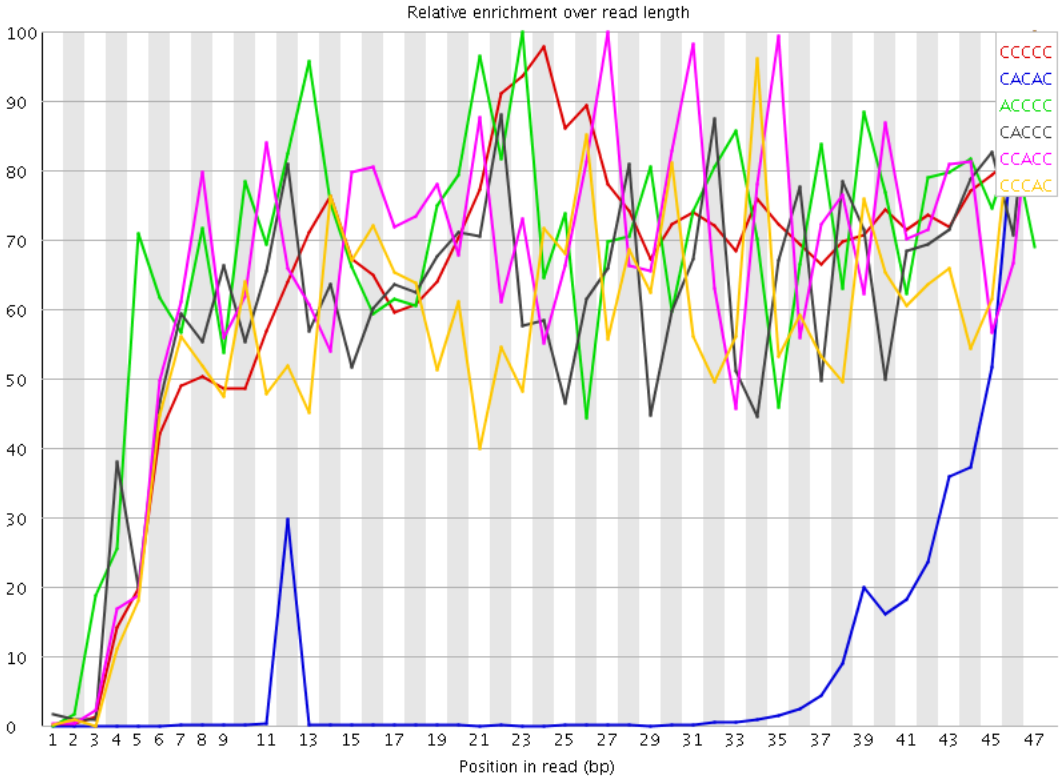
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	123095	650.65845	1008.53046	47
CACAC	691440	140.78903	1485.2458	47
ACCCC	42725	44.324677	65.57151	23
CACCC	40660	42.182358	70.45011	47
CCACC	38825	40.278652	61.672417	27
CCCAC	38660	40.107475	71.18143	47
CCCCA	38475	39.91555	59.23544	38
CGGGC	4686310	26.855158	926.48083	1
AGCAC	1003775	20.996424	157.64998	45
GCACA	896525	18.753025	157.51814	46
GCCCC	25515	13.854884	29.47404	47
CCCCG	24415	13.257573	25.51753	25
CCCGC	24265	13.176122	23.477156	47
CGCCC	24180	13.129966	23.476059	22
CCGCC	23810	12.929052	20.797058	27
CGCGC	216695	12.0879	61.59054	13
CGCGG	2018665	11.568069	243.29532	5
CGGAA	5372570	11.54478	195.8616	1
ACACG	534205	11.1742115	131.69426	47
GCGCG	1904150	10.911836	240.88637	4
CGGCG	1681360	9.635125	250.54187	1
GGCGC	1355105	7.765504	241.27406	3
CGGGA	6760240	7.6034307	218.70854	1
AGATC	4491855	7.394794	37.585976	43
TCGCG	1525490	6.6973467	21.805655	30
CGCAC	58925	6.2799726	218.36906	37
CGGGT	12893120	5.8149524	223.21837	1
CCGCA	54025	5.757751	217.91837	36
TCCCC	13610	5.6619	13.007371	5
AGACG	2539125	5.456167	59.795013	27
CGGAG	4812190	5.4124045	159.72025	1
CCTCC	12950	5.3873334	8.504159	28
ACGCG	491000	5.3756914	16.84339	6
CTCCC	12835	5.3394923	10.361683	47
CGCGT	1194925	5.24607	19.518688	31
CCCCT	12475	5.1897287	9.090841	34
CGTCG	1180760	5.183881	24.667263	41
CCCTC	12430	5.171008	9.286371	45
CGGAC	446920	4.8930836	151.94524	1
CGCGA	445050	4.8726096	16.333778	5
ACTCC	59315	4.8430557	171.47258	23
CGGTT	13354825	4.6144834	154.88258	1
ACGTC	546650	4.5852065	23.674742	47

CTCCA	55775	4.554015	170.84032	24
CGGTC	1031715	4.5295305	158.63824	1
CGACG	404425	4.427829	32.85332	24
CGGGG	7473420	4.399575	106.641014	1
GGGCG	7397715	4.3550076	99.54894	2
TCGAG	4948140	4.2636957	46.119305	44
GATCG	4911205	4.23187	20.746126	44
AGGCG	3640585	4.0946674	52.748863	47
AAAAA	2645175	4.072125	12.6036215	31
AACCC	19660	4.003113	6.650323	32
GAGAC	1807165	3.8833044	57.43925	26
CACGT	450780	3.7810655	32.16548	47
AGAGC	1727510	3.7121387	23.71213	47
ACACC	17145	3.4910157	6.1240783	37
TTACG	5287105	3.4902692	38.623962	14
ACCAC	17050	3.471672	5.932701	37
CGTTT	12889145	3.411974	36.980225	17
CGGTA	3930175	3.3865392	115.57796	1
CAACC	16600	3.3800442	5.836967	31
ATCGG	3879745	3.3430848	19.142002	45
CCACA	16360	3.3311758	6.937499	46
CCAAC	16240	3.3067417	5.5498857	30
TCGGA	3826480	3.2971876	19.426579	46
CCCAA	16165	3.2914703	5.645443	15
CACCA	16155	3.2894342	5.696121	9
GGCGG	5547130	3.2655752	40.146477	11
GGCGT	7202835	3.2485652	49.05088	3
ACCCA	15860	3.2293673	6.6503463	33
TACGT	4839415	3.1947277	40.571156	15
GACGG	2836650	3.1904595	31.338482	28
ACGTT	4823065	3.1839342	42.29042	16
CGAGG	2827285	3.1799264	57.11983	45
CGGAT	3603295	3.1048744	104.93011	1
ACGGG	2680100	3.0143833	31.395996	29
TTTCG	10911730	2.888519	15.555955	30
AAGCG	1337320	2.8736835	53.426456	8
AGAGA	6798660	2.8673303	21.018839	25
CGAGA	1296225	2.785377	35.31368	25
GCGGC	485780	2.7837892	9.741213	9
AGCGA	1266005	2.7204392	54.010284	9
ATCGC	322130	2.7019713	31.124119	29
TTCGA	4076930	2.6913753	34.75447	31
CGTTC	795080	2.6742437	26.984045	33
GTCGA	3057975	2.634985	45.333755	43
GAGCA	1221320	2.6244183	18.528097	47
TTCGC	774990	2.6066716	8.800263	33
TTTTT	161246040	2.5737057	5.601985	16
GCGGG	4317470	2.541679	41.0363	12
CGTTA	3829620	2.528114	32.259304	9
TCGTT	9387625	2.4850628	5.982677	4
GGAGG	21272875	2.4579234	28.599606	39
GGAAG	10819605	2.3884144	12.247374	2
GAAGA	5577755	2.3524144	9.034977	46
ATTCG	3559230	2.3496168	42.30849	34
GAGGC	2056360	2.3128455	43.712563	46
TTCGT	8736360	2.312662	5.614024	35
TTTTA	57713855	2.2972553	12.517538	26
AGAAA	2837680	2.2865145	5.3838673	22
GCGGA	2007035	2.2573683	24.572157	7
TTTAG	43380960	2.253884	15.641144	27
GGGAG	19419655	2.2437975	24.919395	38
CGTAG	2574940	2.218765	23.02527	5
CGAGT	2563245	2.2086878	42.35296	33
AGTAG	13046500	2.2064245	22.16833	35
TCCGC	51310	2.1928089	87.17363	35
CGGTG	4860885	2.1923175	42.302696	1
CGTCC	51095	2.1836207	88.05703	33
GAGAT	12717940	2.1508582	8.629526	26
TCGTC	637015	2.142594	9.92397	40
GTCGC	484000	2.1249013	9.561403	3
GAGGT	23998625	2.124346	22.126604	40
GCGTT	6086185	2.102955	26.34364	16
AAGAG	4946140	2.0860312	9.059646	47
TACGC	244010	2.046714	10.889011	13
AGGAG	9266990	2.0456767	9.640526	38
GGTCG	4478790	2.019988	25.383316	42
GACGC	183950	2.0139682	14.79203	5
ATTTT	50105965	1.9944292	7.7173586	25
CGAGC	181655	1.9888415	8.409717	32
ACGGA	923100	1.9835919	8.7834425	30
AAACG	477665	1.9610252	12.998225	7
TTTAC	3851375	1.9478413	28.724321	13
TAGTT	37043870	1.9246368	9.868249	29
GCGGT	4236740	1.9108207	24.664606	6
TAAAA	3027930	1.8691887	5.2789664	30
AATTT	18670885	1.853333	17.157137	24
ACGGC	169140	1.8518217	9.086359	12
TAGAG	10946580	1.8512859	9.353606	24
AAAAT	2991250	1.8465456	5.0612664	32
ATCGT	2773975	1.8312327	13.828689	39
AGCGC	167065	1.8291036	9.335903	35
GCGTA	2103405	1.8124545	22.706892	4
TTAGT	34659735	1.8007677	15.035208	28
AGGTA	10553575	1.7848209	28.44921	47
GGAAA	4201120	1.7718195	12.706906	2
GGACG	1574435	1.7708111	16.306923	2
GCGAC	161720	1.7705841	18.67107	23
GAAAA	2190965	1.7654117	5.6116014	3
GGAGA	7873855	1.7381438	11.313633	2
AGCGG	1538810	1.7307426	7.2229376	6
TACGG	2004105	1.7268901	12.102828	5
GAGCG	1530700	1.721621	9.373822	28
AACGC	82155	1.7184739	7.613069	23
GCGTC	389605	1.7104797	11.180261	40
TAGTA	13162765	1.7054498	13.962729	29
AGTTA	13135520	1.7019198	21.539803	30
AGTTT	32750205	1.7015568	8.728254	26

TATCG	2566890	1.6945261	14.203539	38
TAGCG	1939080	1.6708598	5.253893	10
TGGCG	3691160	1.6647574	34.71807	10
AGTCG	1913825	1.6490982	14.318947	22
TGAGA	9747895	1.6485643	5.647495	41
AGGTC	1900195	1.6373533	43.059227	41
CACGC	15335	1.6343381	6.3357644	47
TCGAC	193620	1.6240515	9.24713	23
GTAGA	9596200	1.6229097	9.028154	23
CGTAC	192515	1.6147828	9.048225	13
CGATT	2441480	1.6117369	19.096825	11
TATTT	40456675	1.6103467	5.4215612	32
GCGTG	3556175	1.6038777	33.2349	4
TCGTA	2426395	1.6017786	5.948891	45
CGTGG	3539125	1.5961877	32.982018	5
GTCGT	4617080	1.5953364	10.081605	3
TGGGA	17866070	1.5814954	13.393955	37
CGAAA	383435	1.5741696	7.2918496	32
TTCGG	4487465	1.5505506	22.880259	35
TTGAG	22852805	1.5497983	13.889113	44
TAGGA	9115970	1.5416932	7.6237526	37
GGGAA	6965400	1.5376034	14.283341	2
TCGAA	931540	1.5335639	5.1946664	32
AGCGT	1760800	1.5172398	7.6277475	29
GTACG	1739035	1.4984857	11.867694	4
AACGG	690080	1.4828699	7.3211946	29
GGTTT	54289940	1.4763715	9.2630415	2
AGGTT	21742245	1.474484	14.205754	41
GTAGT	21671445	1.4696826	9.543891	36
CGAAC	70040	1.4650588	6.138775	29
TTATT	36729090	1.4619731	7.517707	32
TAATT	14669510	1.4561434	16.796825	23
ACGGT	1686655	1.4533509	11.405769	6
AAGTA	4444825	1.4361699	11.711437	34
TATAG	11007030	1.4261391	17.222279	47
TTTAA	14299405	1.4194057	8.428815	5
GCACC	13305	1.4179895	6.2355943	47
TTATA	14049640	1.3946131	13.424764	46
GCGAT	1609695	1.3870362	22.66707	10
TTAAG	10686740	1.3846406	10.123576	6
GAACG	638385	1.3717856	7.2423897	28
GTTTA	26252830	1.3639817	8.3324175	4
GACGT	1572445	1.3549387	6.212462	3
GGCGA	1202720	1.3527328	8.288868	2
TCGGG	2983060	1.3453959	28.5311	36
AGATA	4123470	1.3323367	5.180238	26
GGAAT	7853910	1.3282534	10.007846	2
AAAAC	168285	1.3199615	22.411375	6
GGAGT	14763020	1.3068149	10.681811	2
TAAGC	793165	1.3057616	38.980362	7
GAGTA	7690415	1.3006033	15.882427	34
GGTAG	14516805	1.2850201	7.157924	2
TTGTA	24474375	1.2715809	14.360858	20
GGGTT	35752090	1.2690564	14.219288	2
ATTAT	12745320	1.2651421	13.331206	45
GGTTA	18637970	1.2639627	15.826857	2
TATTC	2483255	1.2559116	31.0749	33
TGGAA	7340310	1.2413933	9.505952	1
CGTAA	745420	1.2271606	10.564995	21
GTTAA	9422765	1.2208719	19.478	3
TTTGT	57834535	1.2049258	6.8541727	19
GGGAT	13567635	1.2010001	11.436315	42
TCGTG	3469105	1.1986774	6.1447473	40
GGGGA	10332850	1.1938844	10.0465	2
GATTA	9163715	1.187308	16.71477	44
CGTAT	1788280	1.1805284	5.263752	13
CACAT	73295	1.1745731	34.5177	39
GTAAT	8954280	1.1601722	20.672802	22
TGAGG	13031770	1.1535656	16.147646	45
TTTTC	5660235	1.1479232	11.031203	29
TCGAT	1729845	1.1419529	6.5700016	11
GGTGG	24570345	1.1383977	10.705952	8
GGATT	16514495	1.119956	9.024629	43
CGTGA	1291465	1.1128249	7.904203	26
TAGGC	1274465	1.0981765	9.021859	13
GTGGC	2434165	1.0978376	32.701973	9
GGGGT	23203600	1.0750735	7.9731407	2
AGTAT	8255955	1.069693	13.107356	30
AACTC	66750	1.0696876	34.48687	22
GTATT	20407895	1.0603046	5.7334642	31
AGTAA	3280565	1.0599853	7.1086864	9
GTTAT	20362130	1.0579268	8.964868	31
TGGAG	11947240	1.0575635	10.079983	1
GGGTA	11794460	1.0440396	14.197393	2
CGTGC	236600	1.038743	5.191907	13
TGTAG	15293985	1.0371852	8.428474	21
TGTAA	7951875	1.0302943	19.871357	21
ATTTC	2032035	1.0277061	5.6360207	22
GTTGA	15044105	1.0202392	12.9566145	43
CGTGT	2931545	1.0129346	5.860375	41
AGTTG	14922440	1.0119884	9.525238	38
TTAAT	10194245	1.011914	13.282841	4
CGATC	120020	1.0067072	6.352243	44
TAAGT	7596385	0.9842348	6.391355	7
GGTTG	27338230	0.970398	6.938573	42
TGCGG	2126870	0.95924395	5.568926	5
AAGAC	231555	0.9506353	8.413747	32
AAGGC	440285	0.946101	13.452638	46
TTATC	1855190	0.9382664	10.529851	37
TTGGG	26026760	0.92384607	6.025634	36
GTTTG	33931005	0.9227265	6.7339525	18
GGATA	5445830	0.92099893	7.4554806	2
TGGGG	19865240	0.9203999	8.110095	1
CGTCT	270965	0.9113881	8.801948	47
GGAGC	810240	0.9112996	8.488216	27
TAGAC	547275	0.900961	10.044002	25
AGTGA	5283775	0.89359224	5.4564705	18

GTGGT	25152475	0.89281243	7.4032087	9
TGGTT	32497790	0.8837514	7.1793957	1
GGGGG	14495210	0.8766182	5.5086803	2
GAAGC	403110	0.866218	11.750702	4
GGGTG	18468890	0.85570395	7.9517765	2
ACATC	51805	0.83018976	34.574215	40
GGTAT	12046665	0.8169632	6.0004206	2
GGTAC	933855	0.8046808	11.903859	3
GTGCG	1742760	0.7860056	5.112776	4
TGGGT	21383770	0.75903845	8.661823	1
GGTAA	4485330	0.75855917	6.0583286	2
GGAAC	350740	0.7536833	6.5710535	2
TGGTG	20953650	0.7437709	5.875643	7
GTTGG	20687645	0.7343288	5.155303	39
GAGTC	830660	0.7157602	13.161656	21
TGGTA	9733475	0.6600907	5.061556	1
GATTC	798270	0.5269759	5.5563436	29
TGGGC	1136470	0.51256156	5.3870244	13
TGGAT	7515435	0.5096708	5.1277194	1
TCACG	58990	0.49479803	18.000097	30
TCCAG	55970	0.46946678	18.15733	25
CAGTC	53325	0.44728097	18.12012	27
CCAGT	53290	0.44698742	18.082552	26
GTCAC	51815	0.43461534	18.086756	29
GGTGC	880375	0.39705974	5.1566653	3
ATCTC	57695	0.37075287	14.368638	42
CATCT	51210	0.32907972	13.873276	41
GTCCG	57010	0.25029054	9.276129	34
TCTCG	69345	0.23324123	7.578428	43
CTCGT	67440	0.2268338	7.604537	44

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
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	241768	0.3174030567334333	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	176869	0.23220095811433114	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	152949	0.20079779013071164	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	96416	0.126578923257051	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82158	0.10786042956514266	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	79562	0.10445229310671975	No Hit