FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD16_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`

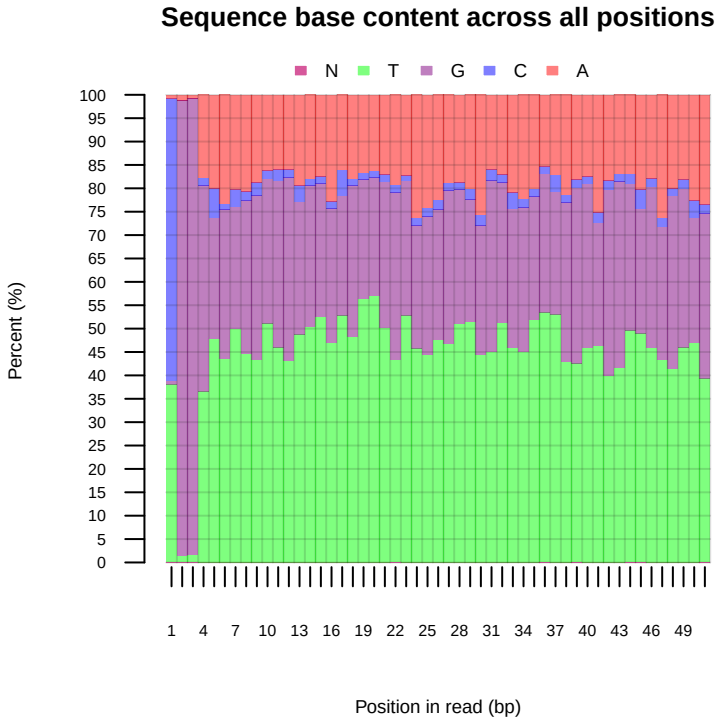
1 Sequence Duplication

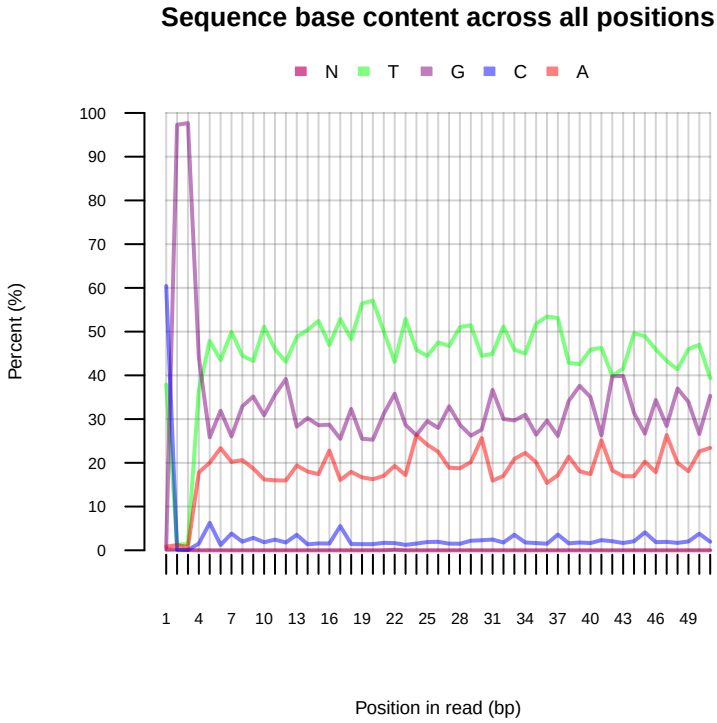
- Estimated Duplication rate 82.0098%

2 Per base sequence quality

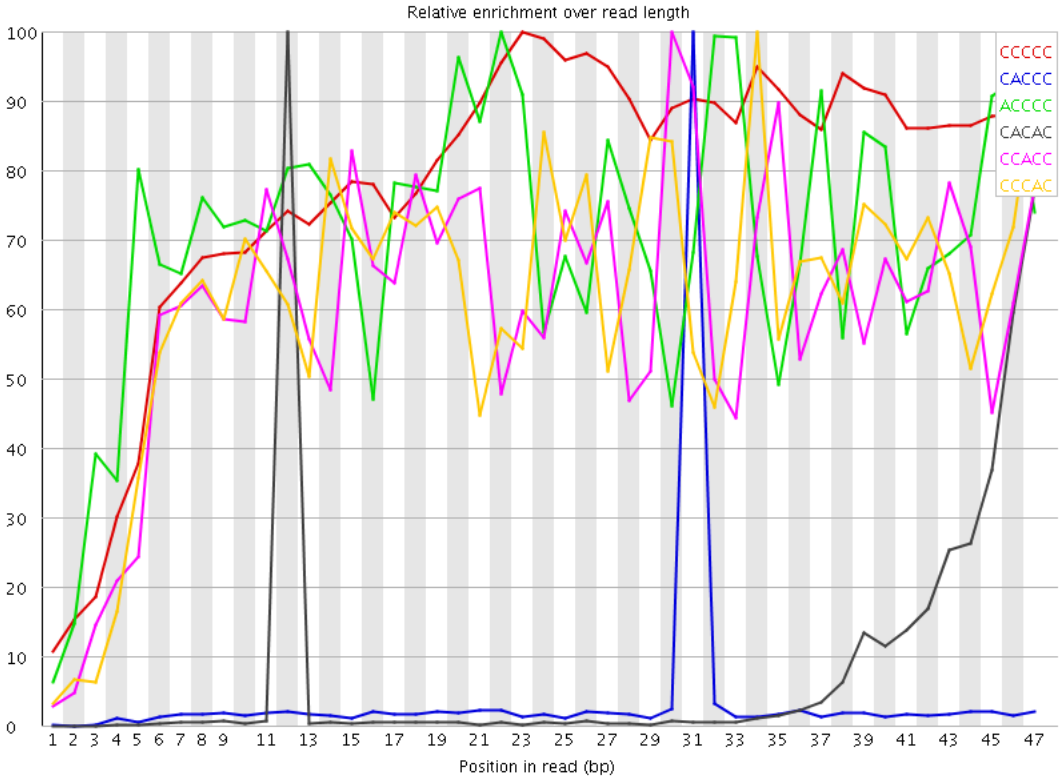
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	288865	2115.6748	2708.4294	23
CACCC	112035	145.50746	3789.755	31
ACCCC	54630	70.95169	100.82631	22
CACAC	276250	63.6226	721.5395	12
CCACC	48870	63.47079	105.5761	30
CCCAC	48710	63.26299	102.8299	34
CCCCA	48150	62.53568	96.727234	33
CGGGC	4607475	32.00404	1173.8589	1
CGCCC	24230	17.435595	32.12143	31
GCCCC	24225	17.431995	31.783312	5
CCCCG	24185	17.40321	30.093307	46
CCCGC	24090	17.33485	35.16521	47
CCGCC	24060	17.313263	30.430826	35
CGCGG	2092835	14.537068	376.74426	5
GCGCG	1980190	13.754622	374.063	4
CGGCG	1606340	11.157819	319.96942	1
GGCGC	1541120	10.704793	375.34772	3
CGCGC	149955	10.601652	59.015457	13
CGGAA	4151520	9.229429	232.51337	1
ACCCG	72295	9.225058	376.7775	32
CCCGA	71810	9.163171	374.73898	38
AGCAC	385600	8.725201	73.95298	10
CGGGA	6424350	7.9131355	239.8074	1
GCACA	325795	7.371958	73.70838	11
ACTCC	70465	6.6083884	279.16278	23
TCGCG	1275310	6.510567	26.812399	30
CTCCA	69240	6.493505	278.4352	24
AGACG	2906845	6.4623365	75.27329	27
TCACC	68570	6.4306707	277.48773	30
TCCCC	11890	6.288201	25.098856	5
CGGGT	12479425	6.2593164	244.08418	1
CGGAG	4715800	5.8086443	173.9456	1
ACGCG	437440	5.4841413	16.836092	6
CGCGT	1043455	5.326928	25.109264	31
CTCCC	9825	5.1960955	10.9341545	4
CCCCT	9800	5.1828732	10.437354	46
CGGTT	14024550	5.169904	182.2083	1
CCTCC	9660	5.1088324	9.691637	28
CGGAC	406520	5.0965	168.8201	1
CGGGG	7418195	5.06255	124.16379	1
CGTCG	986695	5.0371633	24.01735	41
CCCTC	9490	5.018925	8.946303	47
AGGCG	3981600	4.9042997	80.81985	47

AACCC	21265	4.897501	7.372348	2
ACACG	216080	4.889371	71.23095	13
GGGCG	7046430	4.808839	121.02215	2
CGGTC	937305	4.785023	178.56435	1
GAGAC	2143535	4.7653875	72.13527	26
TCGAG	5226330	4.731267	67.91632	44
CGCGA	368445	4.6191583	21.380848	5
AAAAA	3406790	4.375092	11.272148	31
CAACC	18325	4.220395	7.088239	31
ACACC	18065	4.160515	6.2766232	10
CGACG	324355	4.0664062	29.150885	24
ACCAC	17380	4.002754	6.7096143	45
CGAGG	3226475	3.9741817	87.38906	45
ACCCA	17120	3.942874	6.6553698	13
CCCAA	16875	3.8864484	6.8176956	14
CCAAC	16825	3.874933	7.304674	30
TTACG	5820700	3.8727288	56.16528	14
AGATC	2353105	3.8447685	20.015614	43
CGGTA	4120640	3.730313	132.27534	1
CCACA	16160	3.7217782	6.6553698	35
GACGG	3015805	3.7146902	41.124054	28
CACCA	16060	3.6987474	5.905599	19
ACGGG	2897450	3.568908	41.322567	29
TCCCG	68670	3.568132	152.25371	37
CCCGT	68505	3.559558	153.73082	33
CCGTC	68215	3.5444894	152.52226	34
TACGT	5302650	3.528051	57.460503	15
CGTCC	67755	3.520588	152.8885	35
ACGTT	5266465	3.5039756	59.068672	16
GTCCC	66870	3.4746025	152.40019	36
GGCGG	5035440	3.436438	37.4704	11
CGTTT	12628870	3.4215217	41.37947	17
CGGAT	3610545	3.2685368	113.49641	1
GGCGT	6497975	3.259195	53.596947	3
AGAGA	7817740	3.0819561	24.665825	25
GTCGA	3402705	3.0803845	67.31147	43
ACGTC	328450	3.0263603	29.542452	15
GAGGC	2399380	2.9554148	66.73893	46
CGAGA	1286040	2.8590527	34.3751	25
AAGCG	1262960	2.8077426	52.92651	8
TTTCG	10320960	2.796243	15.020625	30
ATCGC	297485	2.7410467	39.444054	29
AGCGA	1230885	2.736435	53.596622	9
GCGGC	393140	2.7307947	9.1958885	33
TTCGA	4046270	2.6921344	33.457287	31
GGAGG	21834730	2.6423874	32.675747	39
CGTTC	699570	2.6247957	32.727448	33
CGGTG	5215915	2.6161513	49.21765	1
GATCG	2825775	2.558104	12.404708	44
GCGGG	3714690	2.5350916	38.278603	12
TTTTT	175397835	2.521917	5.439648	16
CGTTA	3715365	2.471971	29.546047	9
GGGAG	19876635	2.4054234	29.177578	38
TCGTT	8853555	2.3986812	6.118368	36
ACGGA	1071660	2.382455	15.08004	30
TTCGC	633080	2.3753242	6.9964643	33
GGTCG	4652105	2.3333607	38.935474	42
GCGGA	1860890	2.2921345	22.40604	7
TTTTA	64352225	2.2722554	12.187522	26
CGTAG	2500090	2.2632697	26.661877	5
AGTAG	14061540	2.2573092	20.368458	35
ATTCG	3389120	2.254908	37.561016	34
TTCGT	8284540	2.2445185	6.0431314	35
TTTAG	46359475	2.2272599	15.728503	27
GAGGT	24647815	2.192241	24.380842	40
GACGC	174335	2.18562	16.323673	3
CGAGT	2410115	2.1818173	41.604965	33
GCGGT	4287210	2.1503398	35.221546	6
TTTAC	4373155	2.1384447	40.370728	13
AGGTC	2360120	2.136558	64.23286	41
GGAAG	9603615	2.0976436	11.932867	2
AGGAG	9464925	2.0673506	8.9416	38
ATTTT	57847980	2.042593	8.518455	25
GCGTT	5480220	2.020187	23.431965	16
AAACG	498110	1.9986708	9.888955	7
GTCGC	391150	1.9968545	10.876059	3
TCGTC	526630	1.9759225	9.822825	40
TAGAG	12240825	1.9650285	10.977946	24
CGAGC	154070	1.9315599	5.84667	32
ATCGT	2899835	1.9293684	16.758959	39
ACGGC	153840	1.9286765	10.1175375	12
TACGC	208870	1.9245422	10.133231	13
AATTT	22073990	1.9140863	18.675392	24
TACGG	2094995	1.896547	16.674997	5
AGCGC	149630	1.875896	9.266309	35
AGAGC	836465	1.8595825	10.6160965	47
TAGTA	15713235	1.8538946	18.22338	29
GCGTA	2041495	1.8481147	26.32413	4
AACGG	814800	1.8114182	14.015058	29
TCGGA	2000030	1.8105775	10.4926405	46
TTAGT	37413405	1.7974615	15.103627	28
GAAAA	2520295	1.7932668	5.4427814	3
TATCG	2691985	1.7910781	17.231354	38
TAGTT	37263800	1.790274	8.095945	29
GAGAT	11118225	1.7848167	9.312104	26
GGAAA	4520135	1.7819546	12.486711	2
GGACG	1436260	1.7691002	17.386751	2
GAGCG	1431995	1.7638468	10.306479	28
ATCGG	1927255	1.7446961	10.149998	45
GTAGA	10807670	1.7349632	10.6784315	23
AGGTA	10789390	1.7320286	25.920156	47
GGAGA	7922405	1.7304298	11.224272	2
AGCGG	1404630	1.7301402	5.1076674	11
TGGGA	18654100	1.6591445	16.755285	37
GGAAT	10328965	1.6581163	11.626692	2
AACGC	72580	1.642311	8.08586	11
CGTGG	3273755	1.6420203	34.56971	5

GAACG	737600	1.6397914	13.757563	28
CACGT	177875	1.6389521	29.025074	14
CGATT	2459635	1.6364869	18.615856	11
AGTCG	1789225	1.619741	13.868518	22
TATTT	45771775	1.6161861	6.6092124	32
GCGTG	3200220	1.6051373	34.685074	4
AGTTT	33384990	1.6039234	7.7665715	26
AGTTA	13583410	1.6026114	17.194145	30
CGTAC	173655	1.6000688	9.083324	13
GCGAC	127420	1.5974516	19.492865	23
GCGTC	309040	1.5776759	10.157929	40
TCGTA	2369270	1.5763637	5.0451584	45
AGCGT	1731620	1.5675926	8.033357	29
GTACG	1731610	1.5675836	16.37915	4
TGGCG	3123065	1.5664387	31.849216	10
ACGGT	1723810	1.5605224	16.07079	6
ACGAG	699795	1.5557454	5.8947544	32
TAATT	17900940	1.5522316	18.424765	23
GTCGT	4193900	1.5460078	10.375897	3
TGGAA	9625005	1.545109	9.300059	1
GGGAA	7029230	1.5353404	13.784224	2
TAGGA	9517955	1.5279243	6.8023467	37
GGTTT	56084050	1.4928751	9.305136	2
CGATC	159580	1.4703808	30.203537	40
AGGTT	22471470	1.4689353	13.984118	41
GGCGA	1191860	1.4680629	10.713679	2
TTCGG	3938530	1.45187	20.799032	35
TCGAC	156795	1.4447197	5.8968115	23
TTATT	39781310	1.4046649	6.043688	32
GCGAT	1546095	1.3996414	22.82826	10
AAGTA	4822105	1.3971503	10.387213	34
TTTAA	16095575	1.3956842	7.62708	5
GTAGT	21317335	1.3934907	8.845542	36
TTAAG	11733940	1.3844054	9.405739	6
TATAG	11606715	1.369395	15.329184	47
GGTTA	20899055	1.3661482	18.60949	2
AGATA	4708465	1.3642244	5.7985916	26
GTTTA	28342245	1.3616536	8.543363	12
TTATA	15483320	1.3425943	11.639175	46
GTTAA	11339875	1.3379124	22.935726	3
TTGAG	20306445	1.3274099	12.40261	44
GACGT	1459715	1.3214438	6.3040667	3
GGTAG	14759125	1.312715	7.7093873	2
GGAGT	14691820	1.3067288	10.605437	2
GGGTT	35479470	1.2849905	14.129603	2
TTGTA	26731235	1.2842554	14.531435	20
TAAGC	778535	1.2720584	37.599617	7
TCGTG	3442725	1.2691002	8.434427	40
TCGGG	2503740	1.2558032	26.835907	36
GTAAT	10604240	1.25112	23.56807	22
GAGTA	7758965	1.2455523	14.504427	34
ATTAT	14235470	1.2343903	11.498687	45
AACTC	74015	1.2308904	51.172398	22
GGGGA	10140270	1.2271514	10.833967	2
AAAAC	167690	1.2144271	15.435697	6
GGGAT	13526855	1.2031138	11.024195	42
GGTGG	24102575	1.1877501	12.006675	8
TTTGT	60182565	1.1773783	6.945629	19
TCGAT	1758105	1.1697328	7.481388	11
CGTAA	711405	1.1623739	7.444612	21
AGTAT	9830205	1.1597971	17.331715	30
GATTA	9764745	1.1520739	14.988548	44
AAGGC	513440	1.1414514	20.961111	46
CGAAC	50400	1.1404308	6.5707636	9
TGAGG	12750425	1.1340561	15.334775	45
GGGGT	22700130	1.1186389	8.708017	2
GTATT	23270780	1.1180038	7.539266	31
TGTAA	9457825	1.1158626	23.140247	21
TATTC	2252275	1.1013479	25.829231	33
GGATT	16762890	1.0957717	8.442607	43
TTAAT	12548220	1.0880849	15.376534	4
TAGGC	1201135	1.0873578	8.028678	13
GGGTA	12159720	1.0815172	15.294338	2
CGTGA	1182930	1.0708773	9.297564	26
CGTGT	2885900	1.0638366	8.068786	41
AGTAA	3667670	1.0626658	6.6143727	9
GAGCA	477860	1.0623517	7.867663	47
TTTTC	5314320	1.0581903	10.294875	29
TGGAG	11890355	1.0575593	9.922805	1
GGAAC	472885	1.0512916	12.847179	27
GTGGC	1987290	0.9967669	30.088694	9
GTTAT	20684960	0.9937727	7.225671	31
TTATC	2013800	0.98473525	12.382357	37
ATTTC	1994940	0.9755128	6.0766296	22
TGTAG	14867770	0.9718897	6.952357	21
TGCGG	1934525	0.9703015	7.6478925	5
AGTTG	14663660	0.95854723	8.851485	38
TAAGT	8032195	0.9476625	6.048547	7
GAATA	3268405	0.94698334	5.212295	3
ATTAC	788355	0.94669914	5.967598	29
TTGGG	26118485	0.9459558	7.357601	36
GGTTG	26079280	0.94453585	6.660232	42
TGGGG	19152145	0.943798	8.978846	1
TAAGG	5873885	0.94293904	5.89568	45
AAGAC	234565	0.941194	7.6707573	32
TAGAC	574855	0.939263	10.624053	25
GGAGC	751355	0.92547476	9.356714	27
GGATA	5765030	0.92546445	7.835564	2
GTTGA	14141645	0.92442375	11.862712	43
GTGGT	25156725	0.911123	8.777023	9
TGGTT	34086030	0.9073201	7.3930397	1
TTTGG	33673275	0.89633316	5.820679	35
GTTTG	33441960	0.8901759	7.0777645	18
GGTAC	977600	0.884997	16.49391	3
AGTGA	5386475	0.8646948	5.433656	18
GGGTG	17537105	0.86421037	8.511903	2
GGGGG	12195630	0.8177198	6.139221	2

GGTAT	12432975	0.8127299	6.2786117	2
GTGCG	1573255	0.789099	7.185373	4
GGTAA	4851520	0.7788181	6.568681	2
GAAGC	349150	0.7762109	7.065766	4
AGTGG	8619740	0.7666622	5.577946	8
TGGGT	21046910	0.76227427	9.167435	1
TGGTG	20254010	0.73355705	5.8186607	1
GTTGG	20015340	0.7249129	5.01519	39
TACCC	7445	0.6982112	9.981096	31
GAGTC	747530	0.6767204	12.673995	21
CAGTC	70955	0.65378416	28.683046	27
CCGAT	70905	0.6533234	28.133966	39
TCCAG	70895	0.65323126	28.447086	25
GTCAC	69660	0.6418519	28.587791	29
TGGTA	9809650	0.641246	5.306005	1
CCAGT	69310	0.638627	28.382143	26
CACTC	6580	0.6170893	11.148863	31
CACCT	6110	0.57301146	10.664129	31
CATCC	5840	0.5476902	10.53193	31
CGTCT	144725	0.5430101	11.716016	16
ATCTC	78415	0.53102106	22.347916	42
TGGGC	1056020	0.5296689	5.055308	13
GATTC	793585	0.5280017	6.487998	29
TGGAT	7817090	0.51099455	5.110066	1
GGTGC	841630	0.42213714	7.2601795	3
TCTCG	90210	0.3384691	12.421769	43
CTCGT	88820	0.3332538	12.433789	44
TGAAC	195205	0.31894797	5.559627	20
CTGAA	104555	0.17083377	5.2794166	19
GAACT	87345	0.14271413	5.2940235	21
AGTCA	79960	0.13064767	5.2659616	28

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	249726	0.3168894798862306	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	226133	0.2869511735066152	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	144436	0.18328187259975978	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	108144	0.1372291868400428	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	97047	0.12314766325700578	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	86986	0.110380770514018	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	80501	0.10215163827683722	No Hit