FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD2_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`

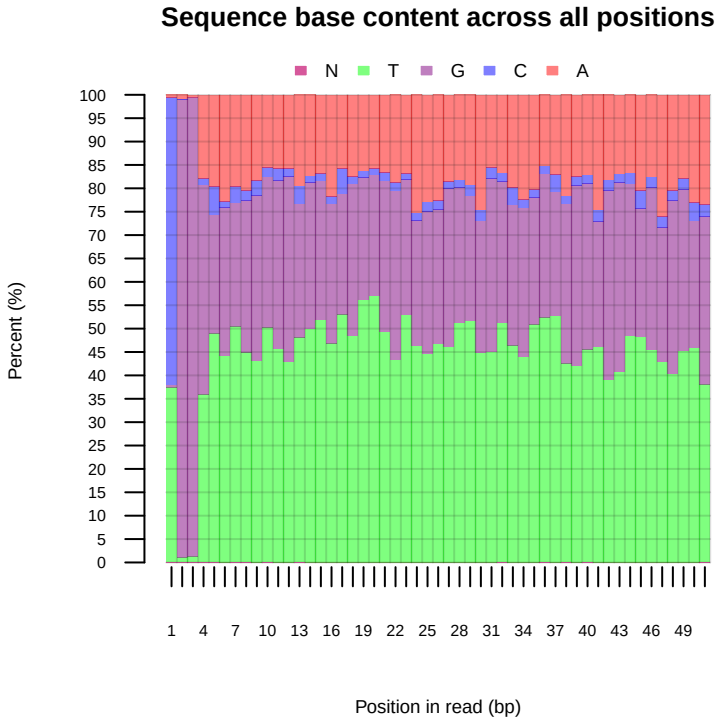
1 Sequence Duplication

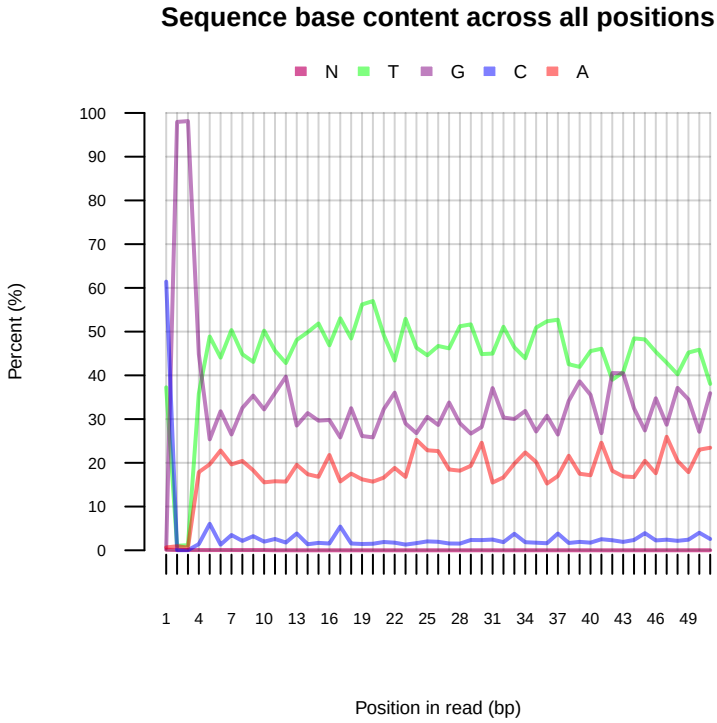
- Estimated Duplication rate 76.2165%

2 Per base sequence quality

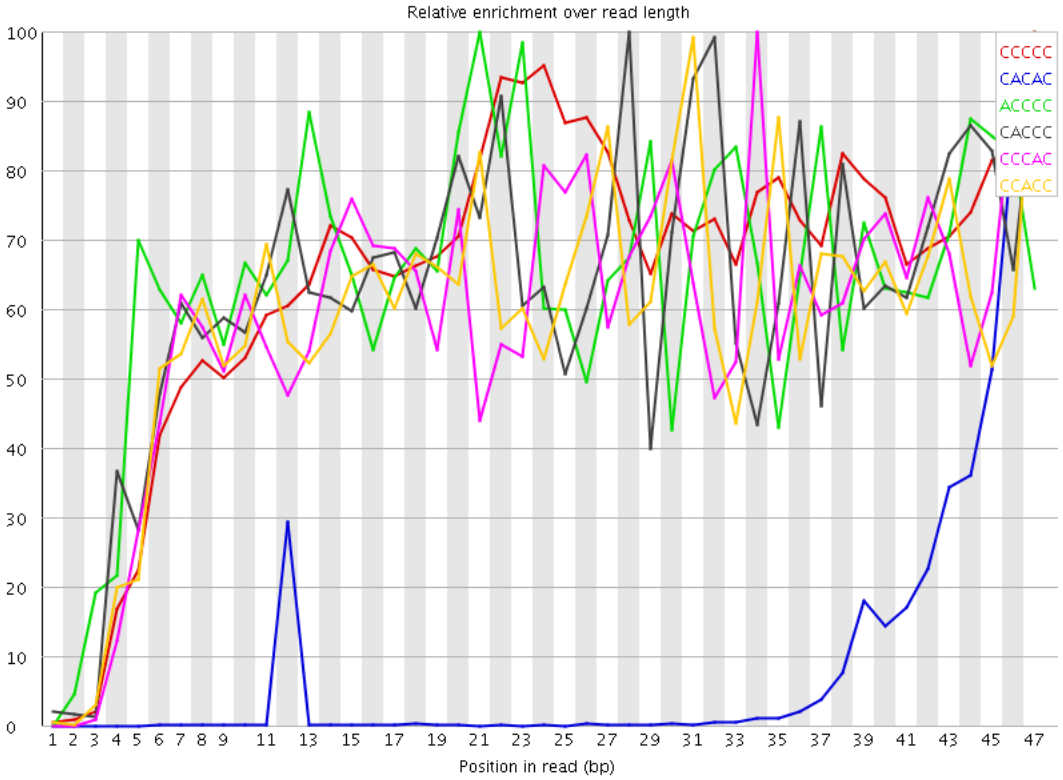
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	122690	758.3444	1157.5286	47
CACAC	576985	126.58294	1367.3575	47
ACCCC	42400	49.37418	76.61046	21
CACCC	40495	47.155834	75.244064	28
CCCAC	39440	45.9273	77.43458	34
CCACC	38625	44.978245	76.34037	47
CCCCA	38485	44.815216	67.03468	25
CGGGC	4427585	28.082066	991.7146	1
AGCAC	848210	18.769339	144.17131	45
GCACA	719740	15.926531	144.10959	46
GCCCC	24655	15.370827	29.737494	47
CCCGC	24285	15.140157	27.540043	34
CCCCG	23790	14.831555	27.979658	46
CGCCC	23765	14.8159685	27.393513	36
CCGCC	23420	14.600884	27.54003	35
CGCGG	1900300	12.052698	275.70447	5
GCGCG	1804380	11.444324	273.30338	4
CGGAA	5059630	11.292739	214.32541	1
CGCGC	179385	11.280108	61.62695	13
ACACG	441475	9.769036	118.90739	47
CGGCG	1532750	9.721503	260.98105	1
GGCGC	1324510	8.400737	273.9207	3
CGGGA	6500700	7.767816	230.88977	1
AGATC	3914665	6.5707707	33.4335	43
TCGCG	1328990	6.339062	24.050373	30
CGGGT	12891240	6.202017	241.65007	1
AGACG	2730605	6.0945187	71.65087	27
TCCCC	12910	6.0528464	15.653475	3
CTCCC	12215	5.726996	11.457097	29
CCTCC	12050	5.649636	11.126584	40
CGGAG	4683740	5.596694	167.22047	1
CCCTC	11830	5.546489	10.245268	39
CCCCT	11535	5.408179	11.67748	38
ACGCG	452975	5.366339	19.877365	6
CACCG	45485	5.342409	200.61066	31
CGCGT	1075575	5.130315	21.960224	31
CGGTT	13929625	5.039864	175.91681	1
CGTCG	1021900	4.874294	22.528399	41
CGGAC	408735	4.8422327	157.22818	1
AAAAA	3233980	4.744364	15.184032	31
CGGGG	7398500	4.7330465	111.20239	1
CGCGA	389125	4.6099157	16.657804	5
GAGAC	2004725	4.4744053	69.07231	26

GGGCG	6993155	4.473735	104.808044	2
TCGAG	4887600	4.392134	51.91364	44
AACCC	19910	4.367993	7.3200507	12
AGGCG	3644870	4.3553276	60.550034	47
ACTCC	49235	4.348947	152.12848	23
CGGTC	906470	4.323712	154.18361	1
CGACG	350550	4.152922	35.537937	24
CTCCA	45935	4.0574555	151.52766	24
GATCG	4322435	3.8842611	19.066956	44
ACACC	17495	3.8381736	5.721871	27
CAACC	17255	3.7855206	6.186002	11
ACGTC	423925	3.7768877	20.80882	47
TCACC	42335	3.7394664	149.91212	30
ACCAC	16810	3.6878934	6.443732	46
CCCAA	16315	3.579297	5.9795437	15
TTACG	5285410	3.571904	43.14858	14
CCAAC	16255	3.5661337	6.1859503	34
ACCCA	16220	3.5584552	6.185918	33
CCACA	16140	3.5409043	6.3405952	35
GACGG	2953190	3.5288255	38.33072	28
CACCA	15740	3.4531496	5.3611593	45
CGGTA	3841840	3.4523852	119.64578	1
CGAGG	2827235	3.378319	65.61412	45
AGAGC	1505030	3.3591213	21.169584	47
GGCGG	5248580	3.357677	38.92658	11
ACGGG	2809565	3.3572052	38.19335	29
CGTTT	12198325	3.3191037	37.370144	17
TACGT	4827575	3.2624974	44.90586	15
ACGTT	4808555	3.2496436	46.56047	16
CGGAT	3578215	3.2154841	110.393654	1
GGCGT	6638270	3.193693	49.528088	3
AGAGA	7263810	3.0543704	24.39489	25
ATCGG	3354940	3.014843	17.292906	45
TCGGA	3312770	2.976948	17.645824	46
CACGT	328920	2.9304569	28.99837	47
AAGCG	1300930	2.9035847	56.586254	8
CGAGA	1262565	2.8179562	35.92912	25
TTTCG	10297505	2.8019	15.88592	30
AGCGA	1244570	2.7777932	57.245113	9
GTCGA	3038220	2.7302296	51.374683	43
ATCGC	302645	2.6963642	35.06869	29
GCGGC	422605	2.6803825	9.919752	9
TTCGA	3917655	2.6475692	35.637768	31
TTTTT	167383720	2.5980716	5.7323623	16
GCGGG	3966000	2.5371714	39.75065	12
GGAGG	20859260	2.5140424	29.748877	39
CGTTC	698965	2.507263	26.04361	33
GAGGC	2091745	2.4994678	49.728573	46
CGTTA	3653900	2.4693224	32.542645	9
TCGTT	8793515	2.3926718	5.690736	36
TTCGC	664160	2.3824139	7.8209214	33
CGGTG	4871440	2.3436654	44.543	1
AGAAA	2961990	2.32639	5.1155076	22
GAGCA	1040500	2.3223228	16.485802	47
GGAAG	10288045	2.316052	12.557051	2
GGGAG	19038520	2.2945995	26.119576	38
TTTTA	59505760	2.2940214	12.112007	26
GCGGA	1909505	2.2817056	24.306067	7
AGTAG	13364370	2.2625859	22.650057	35
TTCGT	8208380	2.2334595	5.559995	35
ATTCG	3296965	2.2281044	39.174465	34
TTTAG	43415755	2.2255874	15.356809	27
CGAGT	2465020	2.2151358	44.184723	33
GAAGA	5217115	2.1937525	7.760254	46
CGTAG	2397855	2.1547797	23.486137	5
GAGGT	23468055	2.1271203	22.56653	40
GACGC	177750	2.1057825	17.707796	5
GAGAT	12344585	2.0899365	9.347917	26
GGTCG	4299785	2.0686402	28.956774	42
AAACG	494205	2.060295	16.454975	7
ACGGA	917420	2.0476172	10.611368	30
ATTTT	52542225	2.0255687	7.9025555	25
GCGTT	5589535	2.022344	24.901817	16
AGGAG	8952190	2.0153232	9.387664	38
TAAAA	3389815	2.00224	6.2676263	30
TTTAC	3888730	1.9763784	31.52593	13
GCGGT	4088685	1.9670795	27.283823	6
AAAAT	3329995	1.9669064	5.9258223	32
TACGC	218590	1.9474905	9.954028	13
AAGAG	4629935	1.9468483	7.762316	47
GTCGC	407710	1.9447092	8.9404545	3
TCGTC	541295	1.9416838	8.677166	40
CGAGC	162495	1.9250581	7.7469044	32
TAGAG	11313505	1.9153749	10.758327	24
AATTT	19850305	1.9006677	17.489637	24
ATCGT	2771445	1.8729553	16.13706	39
TAGTT	36375290	1.8646775	9.063466	29
AGCGC	156255	1.8511338	11.794421	35
CACGC	15500	1.8205416	7.175602	47
GAAAA	2310385	1.8146099	5.7436795	3
ACGGC	152890	1.811269	9.389344	12
TTAGT	35090255	1.7988038	14.764167	28
GGAAA	4229545	1.7784874	13.225426	2
GCGTA	1977480	1.7770189	23.250498	4
AAATA	2998660	1.771199	5.5749807	33
TAGTA	13811690	1.7585093	15.146935	29
AGGTC	1954220	1.7561169	48.814938	41
TACGG	1941675	1.7448436	13.207427	5
TATCG	2580825	1.7441336	16.536213	38
GGACG	1458015	1.7422111	16.622314	2
AGGTA	10269125	1.7385616	27.070864	47
GGAGA	7680220	1.7289765	11.495463	2
CGCAC	14695	1.7259909	6.844431	46
AGCGG	1427955	1.7062919	6.5119376	6
GAGCG	1427895	1.7062201	8.84216	28
GTAGA	9940915	1.6829957	10.420926	23
AACGC	74605	1.6508725	6.623865	23

AGTTT	32133780	1.6472484	8.716479	26
GCGAC	138285	1.6382452	20.59833	23
AGTTA	12863945	1.637842	19.536165	30
TAGCG	1817945	1.6336563	5.0168104	10
CGAAA	389890	1.6254153	7.8277063	32
TGGCG	3363165	1.6180295	33.404026	10
TATTT	41849250	1.6133411	5.7193594	32
TGGGA	17788355	1.6123179	14.423194	37
AGTCG	1785730	1.604707	14.484828	22
GCGTC	336140	1.6033322	10.512674	40
ATAAA	2711425	1.6015397	5.23219	37
CGTGG	3317720	1.5961658	32.82337	5
CGATT	2360720	1.5953853	19.716574	11
AACGG	713770	1.5930847	9.466505	29
GCGTG	3302475	1.5888313	33.024162	4
TGAGA	9063570	1.5344613	5.1594014	41
GGGAA	6808570	1.53275	14.342226	2
GTCGT	4227065	1.5293902	9.432727	3
TCGAC	171635	1.5291529	9.771749	23
TCGAA	902360	1.5146126	5.1851287	32
TAATT	15747535	1.5078272	17.131104	23
TAGGA	8890605	1.5051783	7.288026	37
TTCGG	4139935	1.497866	21.63257	35
ACGAG	669485	1.4942435	5.370739	32
GGTTT	54439140	1.4940559	9.590219	2
GTACG	1656845	1.4888873	13.006217	4
ACGGT	1656220	1.488326	12.786146	6
TTGAG	21711630	1.479955	13.141506	44
AGCGT	1641115	1.4747522	7.00316	29
CGTAC	164115	1.4621547	8.526341	13
AGGTT	21301600	1.4520056	13.899995	41
GCACC	12355	1.4511478	6.6236324	47
AAAAC	185895	1.4475436	28.255945	6
TTATT	37354185	1.4400505	6.7921853	32
AAGTA	4530220	1.4325761	11.882493	34
GTAGT	20899870	1.424622	9.751843	36
TTTAA	14845065	1.4214157	8.485935	5
GAACG	631435	1.4093186	9.270843	28
TATAG	11046445	1.4064374	17.307898	47
GCGAT	1562995	1.4045511	23.957445	10
GGAAT	8275755	1.4010842	10.678326	2
TTAAG	10930795	1.3917127	10.328518	6
TTATA	14506370	1.3889856	13.295512	46
GGCGA	1137675	1.3594306	8.8444605	2
AGATA	4287320	1.3557645	5.6711655	26
GTTTA	26232695	1.3447458	8.391884	4
TCGGG	2747470	1.3218167	27.513844	36
GGTTA	19370520	1.3203752	17.954939	2
TAAGC	786600	1.3203092	40.574493	7
GACGT	1459440	1.3114939	6.03684	3
TGGAA	7736200	1.3097377	9.656709	1
GGGTT	35663690	1.3014895	14.761836	2
GTTAA	10218045	1.3009651	22.796019	3
GGAGT	14247625	1.29139	10.708142	2
ATTAT	13374035	1.2805645	13.219562	45
GGTAG	14086220	1.2767605	7.0194206	2
GAGTA	7535195	1.2757076	16.1177	34
TTGTA	24725540	1.2674856	14.085132	20
CGAAC	57125	1.2640718	5.225399	29
TCGTG	3394260	1.2280741	7.447499	40
GGGAT	13355970	1.2105713	11.961898	42
GGGGA	10042055	1.2103093	10.150193	2
GATTA	9384695	1.1948627	16.87694	44
TTTGT	57710565	1.1911105	6.7558937	19
GTAAT	9294245	1.1833465	21.430746	22
CGTAA	697500	1.1707548	9.458083	21
GGTGG	23973675	1.1633431	11.115952	8
CGTAT	1721015	1.1630698	5.235957	13
TATTC	2273540	1.1554868	28.006222	33
TGAGG	12499270	1.1329209	15.696677	45
GGGGT	23063360	1.1191692	8.022238	2
GGATT	16374600	1.1161609	9.290767	43
TCGAT	1641075	1.109046	6.2065377	11
TTTTC	5379955	1.1008812	11.100795	29
TAGGC	1215745	1.0925025	8.527561	13
AGTAT	8491795	1.0811783	14.282726	30
GTATT	21072945	1.0802455	6.2307096	31
TTAAT	11158315	1.0684092	15.278552	4
CGTGA	1182575	1.062695	7.5383096	26
GTGGC	2199235	1.058059	31.48319	9
AGTAA	3341920	1.0568038	7.3584933	9
GGGTA	11631915	1.0543047	14.725582	2
CGTGT	2905485	1.0512307	7.1393633	41
TGGAG	11572210	1.0488931	10.025337	1
TGTAA	8205360	1.0447093	20.724533	21
AAGGC	459935	1.0265427	15.987085	46
GTTAT	19895020	1.0198625	8.160371	31
AGTTG	14605695	0.9955849	9.740076	38
TGTAG	14592180	0.99466366	7.690272	21
CGTGC	208115	0.9926741	5.181202	13
GTTGA	14475585	0.9867161	12.370091	43
ATTTC	1930735	0.9812622	5.3438225	22
TAAGT	7662100	0.9755413	6.5865984	7
ACCGA	43835	0.9699885	38.954224	32
GGTTG	26490950	0.966745	6.792692	42
TTATC	1899435	0.9653544	12.049922	37
TGCGG	1999715	0.9620693	6.3234234	5
TGGGG	19619435	0.95204985	8.2538185	1
TTGGG	25880980	0.94448507	6.473068	36
GGGGG	14566985	0.9399434	5.704772	2
AAGAC	224895	0.93756646	8.678948	32
ATTAC	730775	0.92245793	5.4185805	29
CGATC	103340	0.9206902	6.3263483	44
TAGAC	543595	0.912425	10.543508	25
TAAGG	5369335	0.9090277	5.0729814	45
GGATA	5357660	0.9070512	7.669382	2
GTTTG	32996880	0.9055833	6.7340074	18

AACTC	54200	0.90195745	29.63825	22
TTTGG	32520565	0.89251107	5.1945543	35
GTGGT	24417645	0.891083	7.8301353	9
TGGTT	32421860	0.88980216	7.2599497	1
AGTGA	5185770	0.87795025	5.5948644	18
GGAGC	734410	0.87756115	7.938354	27
GGGTG	18019190	0.8743966	8.386981	2
GAAGC	376495	0.8403104	10.186129	4
GGAAC	368695	0.82290137	8.327374	27
GGTAC	908545	0.816444	13.062282	3
GGTAT	11864505	0.808734	6.0199056	2
GTGCG	1646490	0.7921316	5.8969293	4
GGTAA	4567390	0.77325857	6.3773375	2
CGTCT	214530	0.76954234	7.746817	47
TGGGT	20921930	0.76351243	8.887661	1
GTTGG	20482795	0.7474869	5.407999	39
TGGTG	20191640	0.73686165	5.8321843	7
CGGCC	11650	0.73257667	5.209109	1
GAGTC	776865	0.69811267	13.346328	21
TGGTA	9533800	0.6498635	5.0333323	1
TGGGC	1088660	0.5237578	5.363751	13
TGGAT	7403880	0.50467926	5.1219463	1
GATTC	740960	0.50074416	5.1883745	29
GGTGC	862845	0.41511753	5.955601	3
TCCAG	45710	0.4072455	15.903277	25
ATGCC	45235	0.40301356	17.22066	47
CAGTC	43900	0.3911196	16.016502	27
CCAGT	43350	0.38621947	15.859388	26
GTCAC	42715	0.38056204	15.9622	29
CCGAT	42405	0.37780014	15.556237	33
ATCTC	47530	0.31845933	12.887391	40
TCTCG	56255	0.20179278	6.9586263	41
CTCGT	55925	0.20060903	6.9662614	42

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	231877	0.3063235669745685	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	175079	0.23128996744972755	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	151918	0.20069288307008668	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	86894	0.11479223911249564	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	83526	0.1103429070374285	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82461	0.10893597750656552	No Hit