FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD6_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`

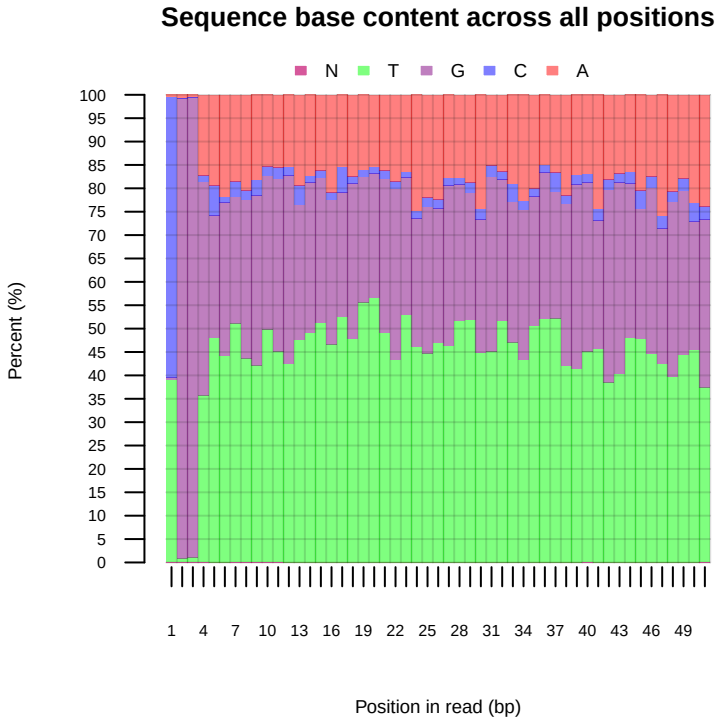
1 Sequence Duplication

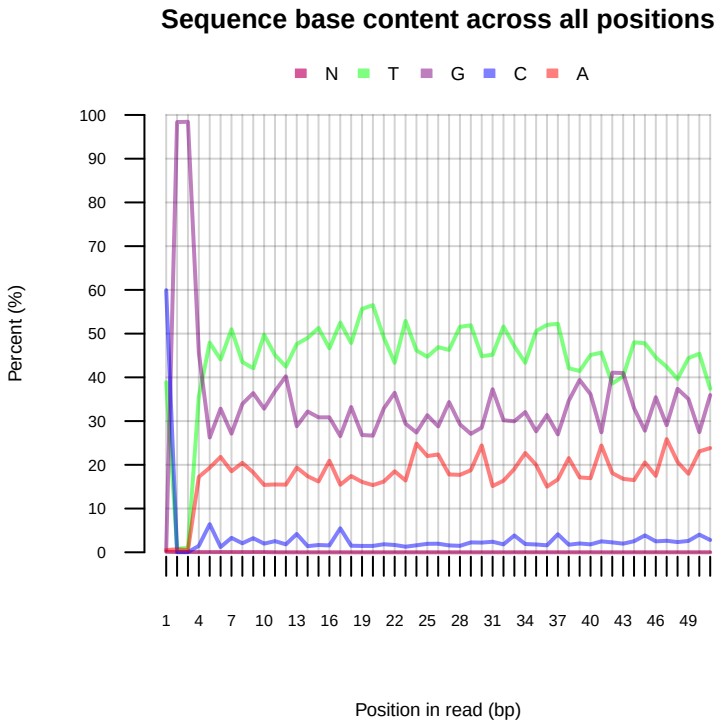
- Estimated Duplication rate 76.7124%

2 Per base sequence quality

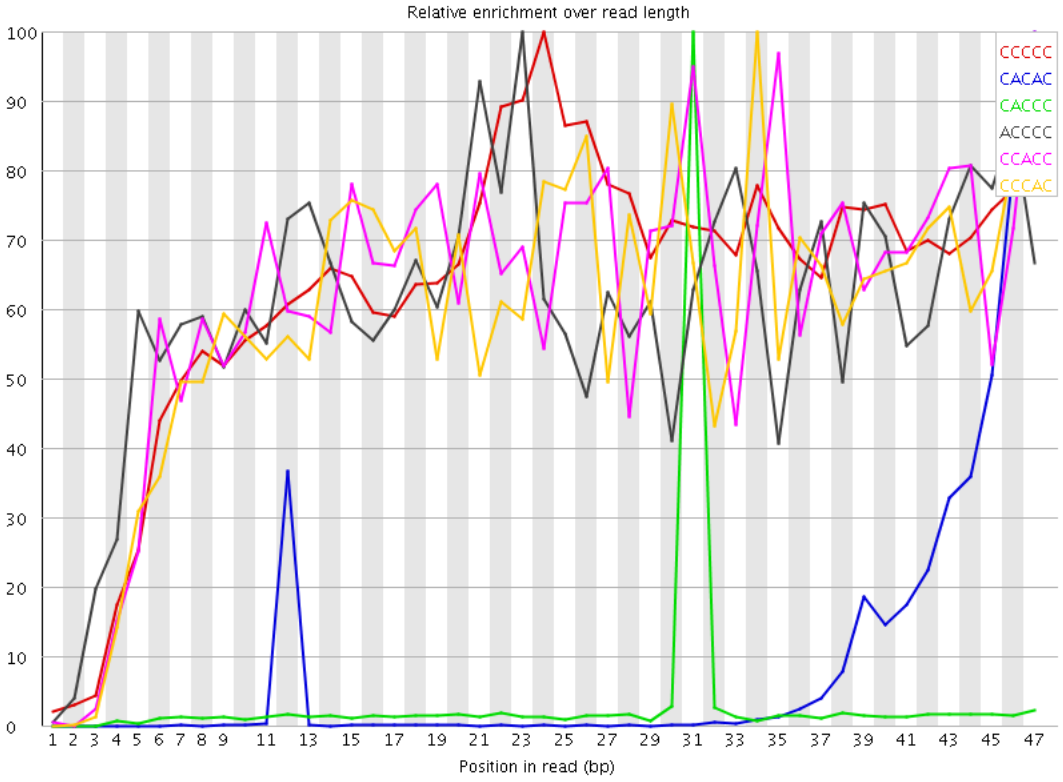
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	125220	792.77905	1242.1317	24
CACAC	748100	173.80475	1876.354	47
CACCC	98030	118.89106	3330.769	31
ACCCC	40685	49.34288	81.49997	23
CCACC	38185	46.31088	74.664154	47
CCCAC	38090	46.195663	78.08342	34
CCCCA	36530	44.30369	66.11212	25
CGGGC	4581100	28.48061	992.7428	1
AGCAC	1099260	25.384548	194.50435	45
GCACA	929855	21.47258	193.91377	46
GCCCC	24500	15.417413	27.35394	24
CGCCC	24380	15.341899	28.241837	31
CCCGC	24060	15.140531	30.311321	26
CCCCG	24035	15.124799	32.972603	25
CCGCC	24000	15.102774	29.57219	27
ACACG	574280	13.2615	152.38722	47
CGGAA	5433655	12.47177	194.94466	1
CGCGG	1779505	11.0631485	254.64938	5
GCGCG	1684880	10.474866	252.73677	4
CGCGC	149250	9.335269	65.12117	13
CGGCG	1490495	9.266378	255.25838	1
ACCCG	71075	8.567894	338.54218	32
CCCGA	69775	8.411183	333.83832	38
AGATC	4716515	8.326353	42.770782	43
GGCGC	1260395	7.835851	253.17876	3
CGGGA	6401310	7.6236014	218.14561	1
ACTCC	74855	6.940257	260.85263	23
CTCCA	70820	6.5661488	260.28717	24
TCCCC	12895	6.241146	14.336521	5
TCACC	66375	6.154026	258.8989	30
CCCTC	12345	5.974948	10.917343	24
TCGCG	1242385	5.9406404	21.681017	30
CTCCC	12270	5.9386477	12.282518	47
CGGGT	12430260	5.90778	229.23506	1
CCTCC	12085	5.849108	10.917521	40
AGACG	2448235	5.61939	63.374596	27
CCCCT	11580	5.6046896	9.780339	38
CGGAG	4386010	5.2234917	153.70522	1
ACGCG	409730	4.9093275	16.0446	14
CGGAC	409060	4.9013	164.25449	1
CGCGT	998935	4.7765493	20.063177	31
CGGTT	12728355	4.6527953	158.42448	1
GATCG	5011415	4.5903864	23.318274	44

ACGTC	491250	4.5271444	27.782358	47
CGCGA	375400	4.4979906	17.55987	5
CGGTC	938340	4.486806	167.37787	1
CGTCG	927575	4.4353313	23.589876	41
GGGCG	7135445	4.4092793	103.355446	2
AAAAA	2662605	4.3485456	14.467478	31
CGGGG	6980820	4.3137307	104.785164	1
AGAGC	1866940	4.285154	27.723316	47
AACCC	18300	4.251607	7.096545	22
TCGAG	4592840	4.206977	47.19746	32
AGGCG	3417485	4.0700326	53.344482	47
GAGAC	1742760	4.0001254	60.910637	26
CACGT	429415	3.9573002	37.849403	47
ACACC	16900	3.9263468	7.041943	21
ACCAC	16560	3.8473554	8.843752	45
CCACA	16095	3.739323	8.843792	46
ATCGG	4077425	3.7348638	21.817715	45
CGACG	309460	3.7079065	31.897518	24
TCGGA	3988350	3.6532729	22.056295	46
ACCCA	15655	3.6370983	6.714646	33
CAACC	15630	3.63129	6.9875684	31
CACCA	15240	3.540682	6.441745	45
CCAAC	15075	3.5023482	6.605416	30
TTACG	4943370	3.4826472	38.716217	14
CCCAA	14855	3.451236	6.004797	15
CGGTA	3763715	3.4475105	119.51694	1
CGTTT	12061045	3.3909714	38.52385	17
GGCGT	6888235	3.2737994	51.50065	3
GGCGG	5234425	3.2345626	42.255043	11
TACGT	4544370	3.201548	40.863655	15
ACGTT	4509490	3.176975	42.691505	16
GACGG	2667485	3.176825	32.865368	28
CCCGT	65610	3.1563134	133.21579	33
CCGTC	65445	3.1483753	132.01819	34
TCCCG	64995	3.1267276	131.11353	37
CGTCC	64760	3.115422	131.65634	35
GAGCA	1344825	3.0867524	22.188683	47
GTCCC	63925	3.0752523	131.29443	36
CGAGG	2571410	3.062405	56.691532	45
CGGAT	3341310	3.0605934	104.73927	1
ACGGG	2559735	3.0485005	32.82722	29
AGAGA	6502670	2.859167	21.554806	25
AAGCG	1245125	2.8579133	56.74157	8
TTTCG	10000980	2.8117826	16.68291	30
AGCGA	1199005	2.7520547	57.33067	9
ATCGC	289775	2.6704392	32.656033	29
CGAGA	1160800	2.6643634	34.924675	25
TTCGA	3758365	2.647801	37.7785	31
TTTTT	156086470	2.5802784	5.7636447	16
CGTTA	3613585	2.545802	35.33334	9
GCGGG	4037375	2.494857	43.064728	12
GTCGA	2709005	2.481411	45.301292	43
GGAGG	20790975	2.4611223	28.921204	39
GAAGA	5585175	2.4557526	10.201464	46
GCGGC	392305	2.4389527	9.343251	9
GGAAG	10642620	2.4280164	11.580245	2
TCGTT	8624470	2.4247758	6.4343324	4
CGTTC	646470	2.3775134	29.244665	33
ATTCG	3371685	2.3753815	45.66205	34
GAGGC	1941150	2.3118007	44.524532	46
TTTTA	55654125	2.3053977	12.863291	26
TTCGC	624370	2.2962365	9.020893	33
TTTAG	41863410	2.2546835	16.007105	27
AGAAA	2656430	2.25108	5.7546916	22
AGTAG	12782110	2.2428653	23.563051	35
GGGAG	18883540	2.2353306	25.011301	38
TTCGT	7943380	2.2332869	5.8321548	35
CGAGT	2433470	2.2290246	46.04438	33
GAGAT	12599050	2.210744	8.618494	26
CGTAG	2389260	2.1885285	23.944418	5
GCGGA	1834130	2.184346	23.684975	7
AAGAG	4962375	2.181913	10.221589	47
GAGGT	23396025	2.13009	22.461142	40
CGGTG	4480650	2.1295366	41.859406	1
GCGTT	5694705	2.0816748	27.932789	16
AGGAG	9020720	2.0579948	10.166375	38
ATTTT	48018035	1.9890828	7.7987947	25
TTTAC	3656375	1.9812291	28.926792	13
ACGGA	855690	1.9640498	9.785474	30
TAGTT	36287490	1.9543749	10.356567	29
AAACG	441515	1.953108	15.549825	7
GGTCG	4020640	1.910906	25.245049	42
TAAAA	2898535	1.8891594	5.827627	30
TACGC	202350	1.8647689	11.335651	13
GACGC	155525	1.8634789	10.84851	3
CACGC	15435	1.8606464	7.109705	47
TAGAG	10554405	1.8519719	9.492536	24
GCGGT	3896070	1.851701	24.869427	6
AATTT	17797130	1.8473351	17.413973	24
AAAAT	2832160	1.8458989	5.650926	32
TTAGT	33682135	1.8140554	15.395499	28
AGGTA	10311315	1.8093171	29.800926	47
ACGGC	150470	1.8029106	8.406837	12
CGAGC	149755	1.7943436	6.2417293	32
GCGTA	1956760	1.7923648	23.522364	4
ATCGT	2535955	1.7866024	14.48865	39
TCGTC	484140	1.7805147	9.620477	40
AGCGC	148510	1.7794261	6.528941	35
GTCGC	371335	1.7755911	9.292761	3
CGCAC	14600	1.7599895	6.8547893	46
GGAAA	3972120	1.7465063	12.257201	2
GAAAA	2042020	1.7304242	5.5652213	3
TGAGA	9858205	1.7298102	5.962508	41
AGTTA	12703185	1.7143946	22.777716	30
GAGCG	1435335	1.7094034	10.396386	28
GGAGA	7490265	1.7088355	10.67624	2
AGCGG	1427335	1.699876	6.702629	6

AGTTT	31503515	1.6967192	9.178893	26
TAGTA	12571025	1.6965586	13.690422	29
TACGG	1834580	1.6804495	11.940231	5
GGACG	1403190	1.6711205	17.130674	2
TATCG	2371520	1.6707565	14.894982	38
TAGCG	1811890	1.659666	5.532438	10
AAATA	2541155	1.6562324	5.3164697	33
TGGCG	3481940	1.6548759	36.641335	10
GTAGA	9315865	1.6346464	9.193511	23
TATTT	38867470	1.6100329	5.322997	32
AGGTC	1748950	1.6020138	43.18743	41
AGTCG	1748430	1.6015375	14.091138	22
GCGTG	3356880	1.595438	35.54545	4
TGGGA	17498780	1.5931756	13.259392	37
TTGAG	22740255	1.5923876	14.464815	44
CGTGG	3330545	1.5829216	35.34157	5
CGATT	2231620	1.572196	19.855076	11
TCGTA	2200900	1.5505532	6.4172025	45
GCGAC	129295	1.5491947	21.919966	23
GCACC	12715	1.5327581	6.3165913	47
TAGGA	8719245	1.529958	7.9975414	37
ATAAA	2340715	1.5255927	5.056283	37
GGGAA	6684745	1.5250635	13.907848	2
TTCGG	4160855	1.5209825	24.34105	35
AGCGT	1648465	1.509971	8.414585	29
GTCGT	4110320	1.5025098	10.643093	3
AACGG	653395	1.4997258	8.74078	29
GTAGT	21167775	1.4822745	10.044716	36
GCGTC	308100	1.4732239	10.554494	40
TTATT	35388340	1.4659146	7.854638	32
GGTTT	52384750	1.463899	9.28249	2
AGGTT	20894465	1.463136	14.60218	41
TAATT	14049975	1.4583818	17.075195	23
AACTC	82105	1.4582641	51.559765	22
AAGTA	4295425	1.4526185	12.498454	34
GTACG	1572835	1.440695	11.717576	4
TCGAA	815720	1.4400403	5.2639303	32
CGTAC	156090	1.438457	9.243915	13
TATAG	10625975	1.4340588	18.336117	47
CGAAA	321555	1.4224468	5.158697	32
AAAAC	165280	1.409113	27.4424	6
TTTAA	13574175	1.408994	8.851029	5
TTATA	13520800	1.4034536	14.39379	46
ACGGT	1529595	1.4010876	11.48435	6
CGATC	151530	1.3964342	28.33167	40
TTAAG	10246175	1.3828018	10.613351	6
GCGAT	1494680	1.369106	23.770985	10
ACGTA	775130	1.3683844	5.2186027	26
GGAAT	7738795	1.3579193	9.974735	2
GTTTA	25069950	1.3502197	8.764495	4
TAAGC	762490	1.3460703	41.618866	7
GAACG	581920	1.3356706	8.590797	28
TCGAC	143965	1.3267183	6.73849	23
AGATA	3877495	1.3112837	5.1418047	26
GAGTA	7442025	1.3058454	16.874287	34
TCGGG	2746870	1.305516	30.324482	36
GGCGA	1094775	1.3038157	8.682112	2
TATTC	2396105	1.2983439	33.733948	33
GGTAG	14187895	1.2917362	7.376424	2
GGAGT	14121560	1.2856967	10.299874	2
ATTAT	12329035	1.2797489	14.277482	45
GACGT	1396055	1.278767	6.191114	3
TGGAA	7275865	1.2766894	9.288772	1
TTGTA	23682685	1.2755042	14.718657	20
GGGTT	34857500	1.2664992	14.329066	2
GGTTA	18007605	1.2609835	16.21047	2
GTTAA	9032255	1.2189739	20.027328	3
TTTGT	56344205	1.2110248	6.9847226	19
GGGAT	13280300	1.2091043	12.038468	42
GATTA	8904525	1.2017356	17.904062	44
GGGGA	10065495	1.1914984	9.8013315	2
CGTAA	668195	1.1796055	10.358283	21
TGAGG	12837460	1.1687859	16.707094	45
CGTAT	1655925	1.1666136	5.515849	13
GTAAT	8630255	1.1647208	20.921541	22
GGTGG	24537560	1.1591576	10.918264	8
TCGTG	3118555	1.1399742	6.3520384	40
GGATT	15957145	1.1173999	9.540731	43
TTTTC	5136580	1.1107358	11.978177	29
GTGGC	2323140	1.1041281	34.61323	9
CGTGA	1192590	1.0923958	8.536071	26
TAGGC	1190425	1.0904127	9.783145	13
GGGGT	22880355	1.0808711	8.050555	2
GTTAT	19715265	1.0618266	9.390783	31
TCGAT	1498125	1.0554421	7.0545597	11
GTATT	19568915	1.0539445	5.587373	31
TGGAG	11547755	1.0513648	10.00662	1
GGGTA	11523540	1.0491601	14.485657	2
AGTAA	3101805	1.0489622	7.5566545	9
GTTGA	14949460	1.0468367	13.4852085	43
AGTAT	7754805	1.0465719	12.811623	30
TGTAG	14844985	1.0395209	8.748624	21
TGTAA	7657230	1.0334034	20.161001	21
AGTTG	14657815	1.0264143	10.064453	38
ATTTC	1890005	1.0241106	6.0899415	22
CGTCT	277755	1.0214956	10.0644655	16
TTAAT	9695505	1.0063895	13.634158	4
TAAGT	7330680	0.98933285	6.7860203	7
GGTTG	26993975	0.9807889	7.1268964	42
CGTGT	2638095	0.9643442	6.0527368	41
TAGAC	533615	0.9420232	12.280503	25
TTATC	1734735	0.9399766	11.082293	37
AAGAC	211820	0.93701756	10.038827	32
TGGGG	19827335	0.93664604	8.538383	1
GTTTG	33021410	0.92278785	6.809454	18
TTGGG	25280125	0.9185184	5.8959126	36
AAGGC	396635	0.91038924	13.39705	46

GGATA	5156830	0.90486425	7.3573575	2
GTGGT	24861810	0.9033196	7.4529767	9
TGCGG	1892130	0.89928037	5.400303	5
GGAGC	751645	0.895167	9.519087	27
AGTGA	5098275	0.89458966	5.384408	18
TGGTT	31723845	0.8865272	7.4843	1
ATTAC	650180	0.8828049	5.0254884	29
GGGGG	14345635	0.88111657	5.5950084	2
GGGTG	18376095	0.86808926	8.093699	2
CACTC	9295	0.8617954	10.2610035	31
GAAGC	363645	0.83466786	11.160273	4
GGTAT	11681500	0.81799763	6.047492	2
GGAAC	347950	0.79864335	7.757534	27
TACCC	8405	0.77927804	9.237081	31
GGTAC	846445	0.7753318	11.756612	3
TGGTG	21065165	0.7653737	6.1020656	7
TGGGT	21049545	0.7648062	8.991622	1
GGTAA	4303570	0.7551435	6.133002	2
CACCT	7940	0.7361652	10.1738615	31
GTTGG	20238335	0.73533213	5.3525167	39
GTGCG	1514895	0.7199903	5.0570273	4
CATCC	7405	0.68656206	9.4985075	31
GAGTC	725810	0.6648318	12.952582	21
TGGTA	9460465	0.66246957	5.2619514	1
TCCAG	70245	0.64734715	26.87395	25
CAGTC	68550	0.63172674	26.967373	27
CCAGT	67860	0.62536806	26.811317	26
CACAT	34850	0.61896956	6.2099786	47
GTCAC	66940	0.6168897	26.874483	29
CCGAT	66510	0.612927	26.406635	39
ATCTC	73595	0.5216361	21.922508	42
TGGGC	1077430	0.51207453	5.8517175	13
GATTC	723265	0.50954646	5.935864	29
TGGAT	7242065	0.507126	5.13919	1
TGAAC	220420	0.38912094	5.334704	20
GGTGC	776625	0.36910972	5.0938697	3
CTCGT	83110	0.30565247	11.482072	44
TCTCG	82925	0.3049721	11.456104	43
CTGAA	158630	0.28003925	5.095768	19
GAACT	104260	0.18405657	5.3711925	21
AGTCA	75275	0.13288757	5.3206854	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
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	251333	0.34263188890216245	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	176891	0.24114818770234078	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	158138	0.21558299804327394	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	94589	0.12894927343152968	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	84581	0.11530578075793392	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	80482	0.1097177835088263	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	76454	0.10422657762460931	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	73969	0.10083887985343769	No Hit