FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD8_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`

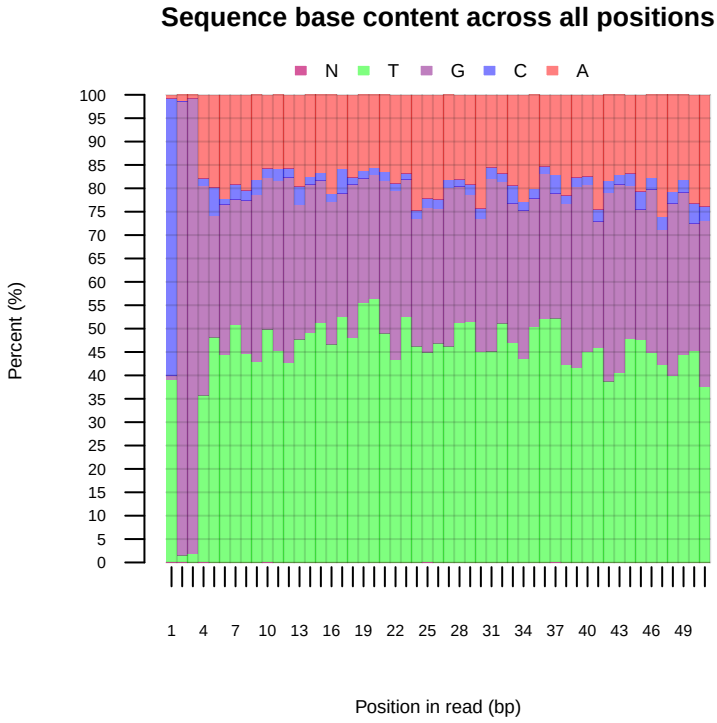
1 Sequence Duplication

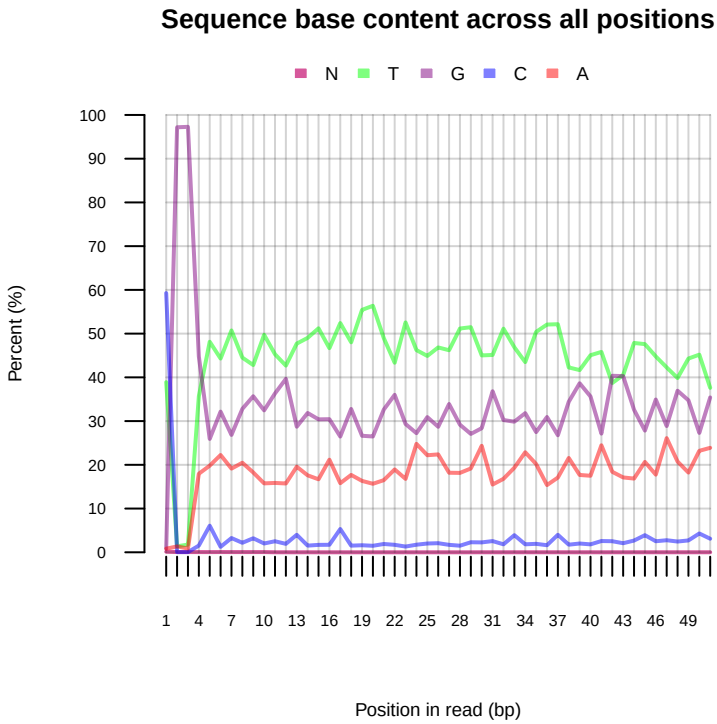
- Estimated Duplication rate 76.1743%

2 Per base sequence quality

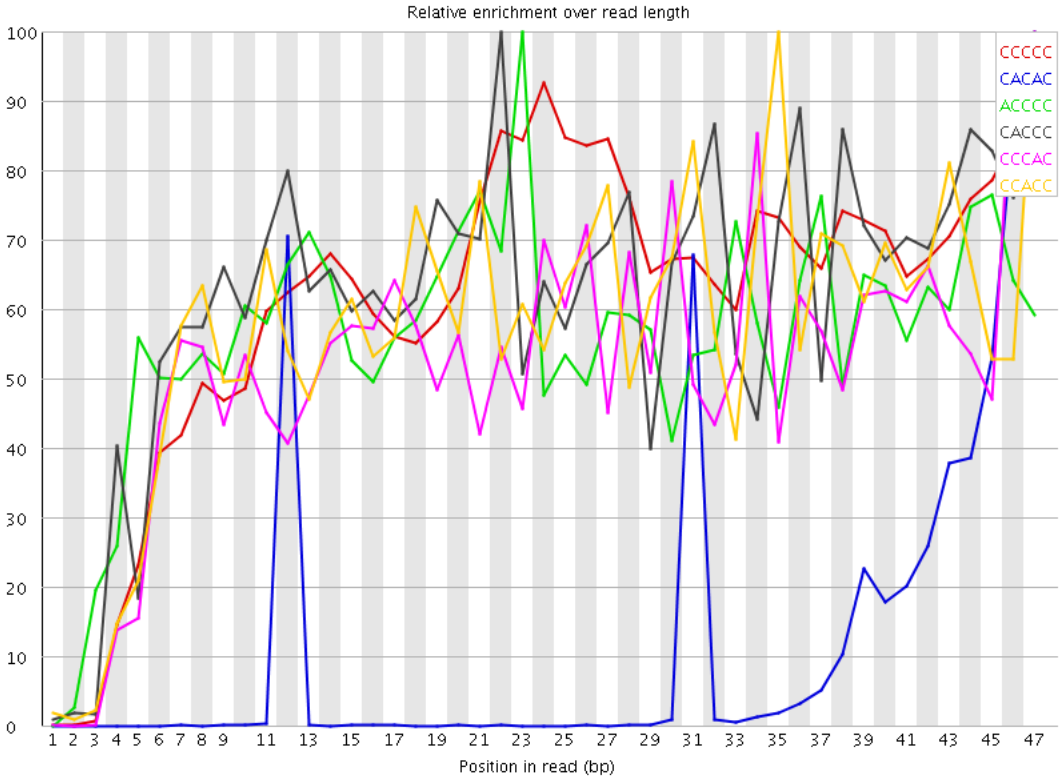
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	124505	699.00854	1126.6104	47
CACAC	1158825	236.24702	1938.3737	47
ACCCC	43825	46.886093	82.95364	23
CACCC	42155	45.099445	72.14458	22
CCCAC	40520	43.350246	83.20899	47
CCACC	39660	42.430176	73.9074	35
CCCCA	39320	42.066425	65.10649	24
AGCAC	1429305	29.609095	205.06874	45
CGGGC	4619750	27.212532	948.48645	1
GCACA	1240530	25.698483	204.88487	46
ACACG	818680	16.959549	172.54884	47
GCCCC	25605	14.607373	34.182804	47
CCCGC	25540	14.57029	26.00574	47
CCGCC	25095	14.316422	24.26276	31
CGCCC	24860	14.182359	24.262787	32
CCCCG	24465	13.957014	27.078161	46
CGGAA	6228680	13.111361	199.34627	1
ACTCC	158400	13.0843	528.34155	23
CTCCA	152365	12.585793	527.12244	24
CGCGG	1901985	11.203602	246.55623	5
GCGCG	1779150	10.480042	244.83675	4
CGCGC	167365	9.702052	60.940033	13
CGGCG	1555615	9.163315	245.2446	1
AGATC	5321095	8.510878	42.311802	43
GGCGC	1309275	7.712255	245.35265	3
CGGGA	6747840	7.5742865	220.77667	1
TCGCG	1380200	6.1775193	22.949831	30
TCCCC	13905	6.0275373	12.63627	5
CGGGT	12921175	5.8765945	227.55638	1
ACGTC	698110	5.859638	56.633648	15
CTCCC	13490	5.847643	10.593084	47
CCTCC	13395	5.806462	10.898358	28
CCCTC	12855	5.572383	9.574125	24
AGACG	2623220	5.5218735	63.916393	27
CCCCT	12695	5.503026	10.796797	47
CACGT	643290	5.399502	56.460377	14
CGGAG	4798395	5.386081	158.47148	1
ACGCG	455370	5.0302525	15.521455	14
CGCGT	1104515	4.943605	20.264956	31
GATCG	5755825	4.909148	23.836775	44
AGAGC	2262115	4.7617483	28.132023	47
CGGTT	13679455	4.72731	161.3609	1
CGGAC	424130	4.6851587	153.95631	1

CGTCG	1029205	4.606531	22.503887	41
CGCGA	402375	4.4448423	15.905956	5
AAAAA	3135435	4.4230876	14.564388	31
CGGGG	7323995	4.3837943	104.94033	1
GGGCG	7260850	4.345999	101.02054	2
CGGTC	944280	4.2264223	154.91055	1
TCGAG	4879315	4.1615715	44.904102	44
AACCC	20275	4.1334186	8.19114	22
AGGCG	3614440	4.0571213	51.758595	47
ATCGG	4720850	4.0264173	21.882032	45
GAGAC	1892930	3.9846144	61.48991	26
TCGGA	4624295	3.9440649	22.11362	46
CGACG	348980	3.8550136	30.362629	24
ACACC	18500	3.7715533	6.1795716	47
CAACC	18010	3.671658	6.371101	31
ACCAC	17710	3.6104977	5.844219	45
GAGCA	1704365	3.5876858	22.425592	47
ACCCA	17020	3.4698293	7.5687156	33
CCCAA	16795	3.4239588	6.8978643	15
CCAAC	16730	3.4107072	6.8981104	34
CCACA	16710	3.4066303	5.748321	29
TTACG	5208095	3.3751962	37.51744	14
CACCA	16510	3.365857	5.4130955	36
CGTTT	12742415	3.3459435	36.549892	17
CGGTA	3917850	3.3415375	115.60826	1
GGCGG	5407985	3.236962	40.30355	11
GGCGT	7067490	3.2143185	49.26492	3
GACGG	2850825	3.199982	34.167995	28
TACGT	4784005	3.1003573	39.481575	15
ACGTT	4753585	3.0806432	41.168488	16
CGAGG	2738100	3.073451	56.06093	45
CGGAT	3603445	3.0733812	104.859924	1
ACGGG	2715400	3.0479708	33.998623	29
AGAGA	7107295	2.8509037	21.803486	25
TTTCG	10684315	2.8055215	15.725221	30
AAGCG	1329190	2.7979426	52.676655	8
AACTC	171730	2.7031362	103.61988	22
CGAGA	1275410	2.684736	35.890675	25
ATCGC	318130	2.670248	32.326866	29
AGCGA	1259015	2.6502247	53.280186	9
TTTTT	168561100	2.5966837	5.658468	16
TTCGA	4004270	2.5950365	34.80444	31
GAAGA	6334115	2.540763	9.888634	46
GTCGA	2935725	2.5038822	44.155777	43
GGAAG	11695230	2.5015671	12.121956	2
GCGGC	423410	2.494087	8.867741	9
GCGGG	4156950	2.4881523	41.02931	12
CGTTA	3798210	2.4614959	32.73491	9
GGAGG	21546740	2.457594	28.426512	39
TCGTT	9220390	2.421119	5.85188	36
TTCGC	699535	2.3790476	8.27006	33
CGTTC	690170	2.3471982	27.117228	33
TCACA	148320	2.3346484	103.748764	30
ACACT	148005	2.3296897	103.989624	32
GAGGC	2053485	2.304987	43.029724	46
ATTCG	3522475	2.282801	41.9336	34
TTTTA	59912990	2.2779064	12.093643	26
AGAAA	3020860	2.272394	5.3328185	22
AAGAG	5647745	2.2654438	9.890323	47
GAGAT	13856180	2.2520003	8.966319	26
TTCGT	8519505	2.2370784	5.7120657	35
GGGAG	19580040	2.2332747	24.636448	38
TTTAG	44425555	2.222932	15.19088	27
GCGGA	1976385	2.218444	25.051905	7
AGTAG	13537730	2.200244	22.047195	35
CGAGT	2565085	2.1877635	43.088123	33
CGGTG	4760185	2.1649485	42.226185	1
CGTAG	2527445	2.15566	22.517998	5
GAGGT	24371990	2.1122296	21.826075	40
GCGTT	5975435	2.0649753	26.178944	16
CACGC	18895	2.0540955	7.765465	47
AGGAG	9464860	2.024499	9.629944	38
ACGGA	941255	1.9813403	10.448678	30
ATTTT	52108905	1.9811931	7.4741054	25
CGCAC	18025	1.9595169	8.097541	47
AAACG	492730	1.9450717	16.307487	7
GGTCG	4261930	1.9383403	25.131413	42
TAAAA	3381905	1.9330186	5.8236694	30
TAGTT	38021455	1.9024887	9.668971	29
GACGC	171985	1.8998352	10.4796295	3
AAAAT	3323680	1.8997381	5.783941	32
TTTAC	3826755	1.8843968	27.654598	13
GCGGT	4116645	1.8722641	24.592455	6
TACGC	221020	1.8551478	10.742666	13
TCGTC	545315	1.8545611	9.207295	40
TAGAG	11393965	1.8518245	9.72993	24
GTCGC	412280	1.8452889	8.456469	3
ACGGC	166395	1.8380853	9.175573	12
AATTT	19568310	1.8362062	16.4384	24
ATCGT	2774695	1.7981892	14.847026	39
TTAGT	35651815	1.7839184	14.607121	28
AGCGC	159645	1.7635212	7.0367703	35
CGAGC	159635	1.7634108	5.9102135	32
AGGTA	10809430	1.7568221	27.738369	47
GGAAA	4370295	1.7530284	12.495074	2
GCGTA	2052985	1.7509925	22.28072	4
TGAGA	10771560	1.7506672	5.9417744	41
GAAAA	2326030	1.749719	5.59055	3
GGAGA	7995835	1.71028	11.059419	2
AAATA	2976410	1.7012469	5.404286	33
AGCGG	1506740	1.6912792	7.164312	6
AGTTT	33590875	1.6807947	8.681199	26
TAGTA	13589805	1.6782621	13.257042	29
AACGC	80765	1.6731056	8.021922	11
GAGCG	1488775	1.6711138	9.424426	28
TATCG	2574960	1.6687474	15.211231	38
GGACG	1481970	1.6634754	16.383896	2

TACGG	1947655	1.6611568	11.813198	5
AGTTA	13430785	1.658624	20.886478	30
TGGCG	3623315	1.6478962	34.853367	10
TAGCG	1916875	1.6349043	5.2229548	10
GTAGA	10020840	1.6286551	9.417794	23
AGTCG	1893055	1.6145884	14.776512	22
TATTT	42093895	1.60042	5.2107115	32
TCGTA	2468185	1.5995501	6.0070395	43
AGGTC	1871780	1.5964428	42.06483	41
TTGAG	24159055	1.5909321	13.739463	44
TGGGA	18258095	1.5823611	13.1510515	37
GCGAC	142365	1.5726373	19.393744	23
GCGTG	3457220	1.5723555	33.47066	4
CGTGG	3429960	1.5599575	33.221012	5
ATAAA	2722555	1.5561494	5.11112	37
CGTCT	453880	1.5436	22.392916	16
CGATT	2373645	1.5382818	18.532408	11
AACGG	727400	1.5311757	9.184909	29
GGGAA	7118945	1.5227164	14.2446165	2
TAGGA	9354235	1.520314	7.544363	37
GTCGT	4396175	1.5192187	9.842032	3
TTCGG	4381860	1.5142717	22.88111	35
ACGCC	13835	1.5040176	6.4113116	23
GCGTC	335865	1.5032696	9.937302	40
GCACC	13810	1.5012999	5.849643	47
GGTTT	55576070	1.4828812	9.486552	2
AGCGT	1735540	1.4802437	7.5759315	29
AGGTT	22142840	1.4581596	13.985499	41
GTAGT	22075580	1.4537303	9.568018	36
TTATT	38217400	1.4530348	7.3373976	32
TAATT	15385410	1.4437009	16.09584	23
CGAAA	364490	1.4388391	5.648832	32
CGTAC	171255	1.4374416	8.778338	13
GTACG	1660480	1.4162247	11.51492	4
AAGTA	4637875	1.4135748	11.471971	34
ACGGT	1641435	1.3999814	11.170962	6
TTTAA	14910550	1.399142	8.213539	5
TATAG	11326415	1.3987465	16.945814	47
AAAAC	188660	1.3966295	27.937353	6
TCGAC	164310	1.3791482	7.344395	23
GGAAT	8446545	1.3727899	10.282663	2
TTATA	14587815	1.3688579	13.129333	46
GAACG	645860	1.3595341	8.987502	28
TTAAG	11004310	1.3589684	9.901939	6
GCGAT	1582690	1.3498776	22.444069	10
GGCGA	1191555	1.3374915	8.654497	2
GTTTA	26663995	1.3341925	8.293338	4
ATGCC	157250	1.3198897	59.71255	47
TCGGG	2876475	1.3082306	28.772362	36
CCCAG	12025	1.3072505	5.747295	27
AGATA	4288670	1.3071408	5.162867	26
GGAGT	14992065	1.2993064	10.632455	2
TGGAA	7945475	1.2913526	9.616135	1
CGCCA	11870	1.2904004	5.849404	24
TCCAG	153305	1.286777	55.566563	25
GGGTT	36551510	1.2835188	14.464054	2
GGTAG	14805875	1.28317	7.141668	2
GACGT	1497195	1.276959	5.8885126	3
GAGTA	7851000	1.275998	15.769137	34
CCAGC	11690	1.2708323	5.3641562	28
GGTTA	19282250	1.2697828	16.453459	2
TAAGC	793125	1.2685717	38.11513	7
TTGTA	25249840	1.2634321	13.961677	20
CAGTC	150445	1.2627714	55.823536	27
CCAGT	149850	1.257777	55.535187	26
ATTAT	13364630	1.2540796	13.030923	45
GTCAC	148425	1.2458163	55.80827	29
GTTAA	9893920	1.2218417	20.235239	3
TATTC	2456880	1.2098335	30.464975	33
CGTAT	1864055	1.2080332	5.5386596	44
TTTGT	59423090	1.2047455	6.7468944	19
ACGAC	57935	1.2001657	5.349264	18
GGGAT	13819580	1.1976916	11.55257	42
GGGGA	10371345	1.1829425	9.930594	2
GATTA	9504110	1.1737025	16.501154	44
CGTAA	731980	1.1707726	10.422812	21
TCGTG	3369865	1.1645491	6.5581965	40
GGTGG	24924430	1.1518624	10.811504	8
TGAGG	13285060	1.1513667	16.008698	45
GTAAT	9252900	1.1426795	19.952438	22
CGAAC	53870	1.1159563	6.9263554	20
GGATT	16910830	1.1136191	9.054304	43
TTTTC	5494060	1.0961801	11.10501	29
GGGGT	23642815	1.0926336	8.066536	2
GTGGC	2378880	1.0819228	32.811073	9
CGTGA	1267670	1.0811968	8.248357	26
TAGGC	1258215	1.0731328	9.167681	13
TCGAT	1635260	1.0597585	6.606186	11
TGGAG	12190730	1.0565251	10.21866	1
GTATT	21030755	1.052321	5.499936	31
AGTAA	3422960	1.0432818	6.951256	9
GTTAT	20825725	1.0420617	8.781669	31
GGGTA	12012255	1.0410573	14.290672	2
AGTAT	8411325	1.0387498	12.426032	30
TGTAG	15608915	1.0278848	8.228363	21
GTTGA	15590490	1.0266714	12.765703	43
ATCTC	160445	1.0232805	45.159615	40
TTAAT	10844910	1.0176399	13.656855	4
TGTAA	8206000	1.0133933	19.18269	21
AGTTG	15383690	1.0130532	9.563092	38
ATTTC	2024245	0.99679244	5.774925	22
CGTGT	2858395	0.98779666	6.250961	41
GGTTG	27878430	0.9789606	6.9384484	42
TAAGT	7872495	0.9722075	6.336613	7
CACTT	147305	0.93947667	41.671505	33
TGGGG	20303415	0.9383059	8.467	1
TTATC	1886745	0.92908376	11.200332	37

TTGGG	26421430	0.9277975	5.9551835	36
TGCGG	2035510	0.9257569	5.5304556	5
GTTTG	34613490	0.92355746	6.634677	18
GGATA	5532000	0.8990983	7.410634	2
GTGGT	25536640	0.89672786	7.4147515	9
AAGGC	425345	0.8953505	13.104974	46
GGGGG	14656040	0.8913954	5.5432053	2
TGGTT	33389515	0.89089936	7.4903674	1
AAGAC	225550	0.8903678	8.646701	32
AGTGA	5425580	0.8818022	5.5072856	18
TAGAC	550740	0.8808865	10.727309	25
GGAGC	780510	0.8761035	8.563536	27
GGGTG	18776235	0.86772853	8.183766	2
ATTAC	708170	0.8606617	5.068793	29
GAAGC	401010	0.8441254	11.7756195	4
GGAAC	387440	0.8155605	8.184238	27
GGTAT	12290845	0.8093818	5.97047	2
TGGGT	21970255	0.7714931	9.097634	1
CACAT	49005	0.77136886	6.261795	47
GGTAA	4673920	0.7596371	6.0697765	2
GGTAC	886015	0.7556828	11.553033	3
TGGTG	21420625	0.7521927	5.9321074	7
GTGCG	1625830	0.7394331	5.104491	4
GTTGG	21052260	0.7392573	5.244425	39
CGGCC	12245	0.7098354	5.9204636	1
GAGTC	805715	0.68719506	13.543031	21
GAATC	425000	0.6797704	11.463071	38
TGGTA	9990520	0.6579	5.147209	1
TCTCG	169145	0.575245	24.158659	41
CTCGT	168425	0.5727964	24.156414	42
TGAAC	345860	0.55318916	11.210908	20
TGGGC	1122760	0.5106351	5.5981784	13
TGGAT	7743905	0.50995487	5.242064	1
GATTC	783570	0.5078061	5.546804	29
CTGAA	290315	0.46434715	10.945928	19
GGTGC	828020	0.37658632	5.1434536	3
GAACT	205655	0.3289369	11.007579	21
AGTCA	160455	0.25664136	10.921754	28
AATCT	165600	0.20125897	8.532686	39

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	247927	0.3166667266977893 4	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	177015	0.22609381239804127	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	157239	0.20083475958340033	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	92223	0.11779255803623738	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCC	89109	0.11381517684364072	TruSeq Adapter, Index 8 (100CGGGTTT
85150	0.10875851270058026	No Hit	
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	81255	0.10378359306501056	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	80315	0.1025829706112402	No Hit