FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD12_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`

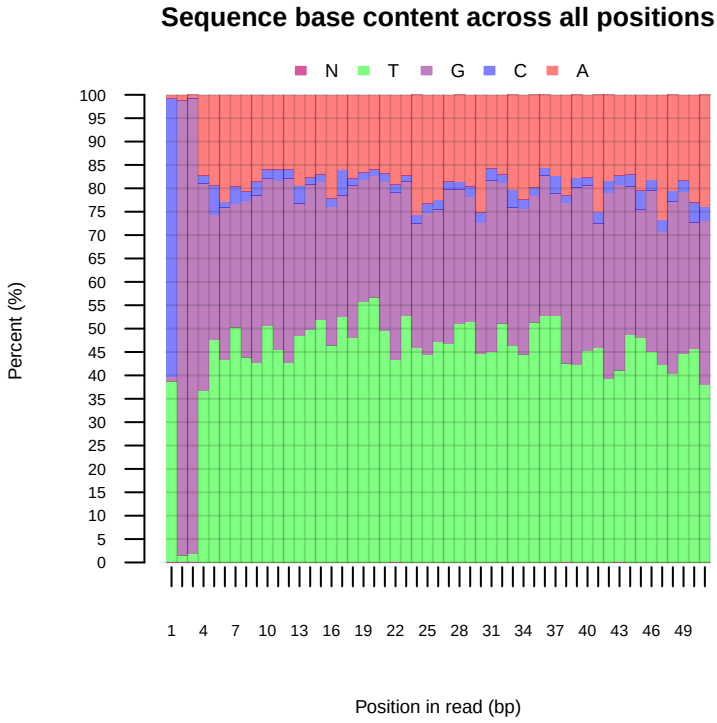
1 Sequence Duplication

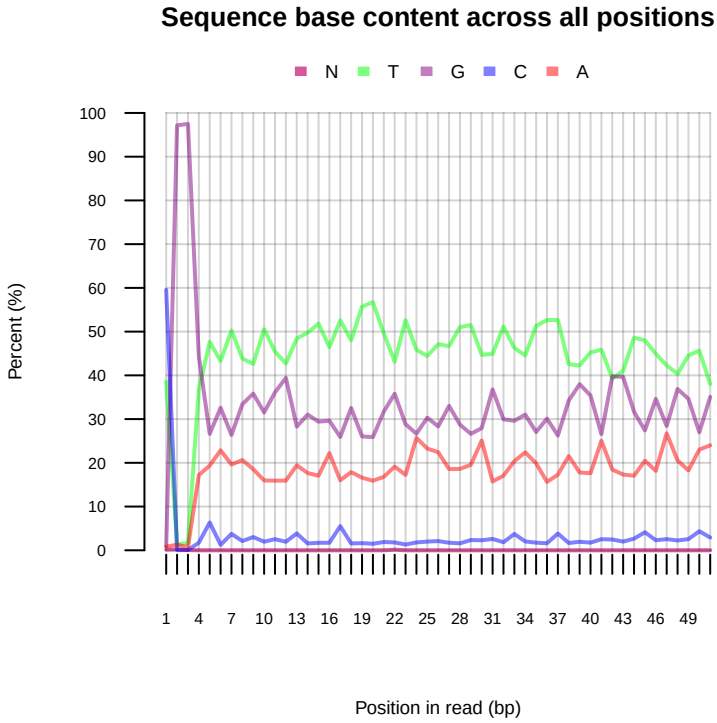
- Estimated Duplication rate 79.2284%

2 Per base sequence quality

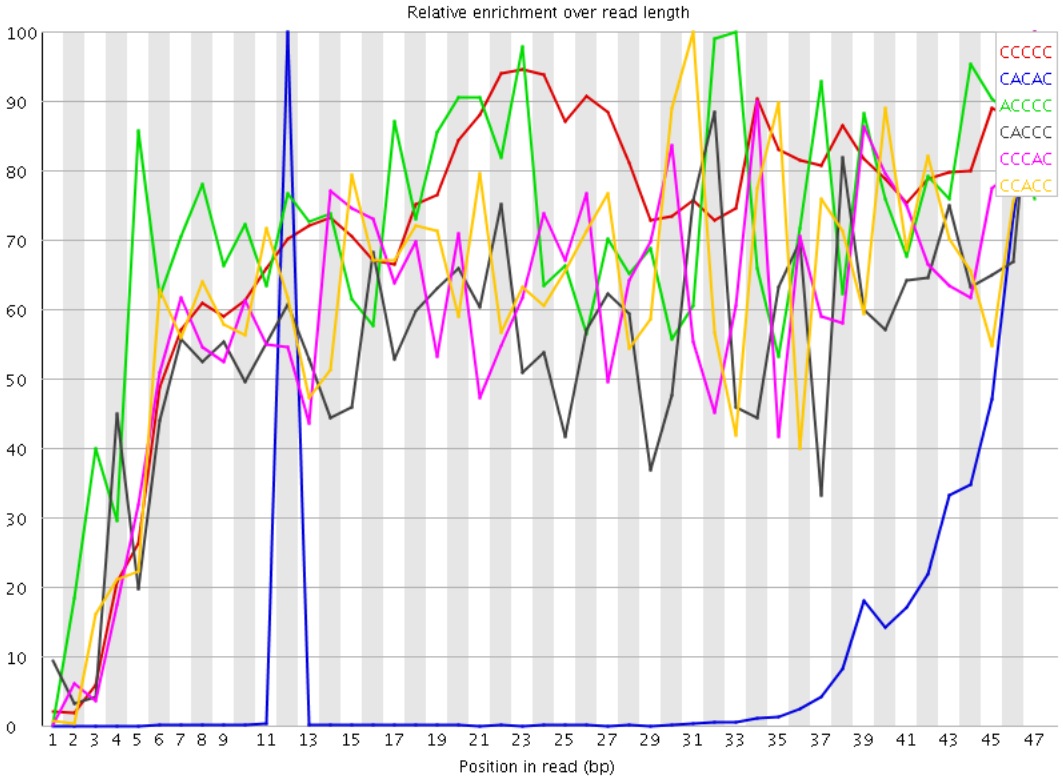
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	145920	1025.1584	1445.9221	47
CACAC	593880	148.05615	1450.6287	12
ACCCC	40280	53.307877	75.24409	33
CACCC	39415	52.16311	95.14545	47
CCCAC	38080	50.396324	84.57373	47
CCACC	37395	49.489773	80.52983	31
CCCCA	36245	47.967827	70.5802	38
CGGGC	3709165	28.336214	1011.6218	1
AGCAC	853415	21.878511	155.99596	10
GCACA	723955	18.559622	155.43568	11
CACCG	125950	17.140757	756.6791	31
ACTCC	128795	13.119225	566.8832	23
CCCGC	18080	13.061842	24.951002	47
CGCGG	1676045	12.804167	295.39313	5
CGCGC	172280	12.798841	59.569942	13
GCCCC	17700	12.787312	27.327288	47
CTCCA	125435	12.776971	564.3692	24
CCGCC	17510	12.650048	27.156956	31
TCACC	123415	12.571211	567.31274	30
CCCCG	17365	12.545292	26.818066	46
CGCCC	17330	12.520006	21.21637	24
GCGCG	1604490	12.257521	292.41196	4
ACACG	475425	12.188199	150.03285	13
CGGAA	4456370	11.748128	189.24646	1
CGGCG	1409080	10.764684	294.85745	1
GGCGC	1172275	8.95561	293.5098	3
AGATC	3779885	7.45836	39.93321	43
CGGGA	5151245	7.413165	216.49843	1
TCGCG	1191820	6.8148093	23.088633	30
CGGGT	9997275	5.8783374	227.91127	1
AGACG	2199575	5.79864	66.158066	27
CGGAG	3756175	5.4055176	159.45068	1
CGTCG	937675	5.361612	23.390259	41
CGCGT	936840	5.3568373	21.311344	31
ACGCG	376400	5.26758	17.928875	6
CGGTC	860850	4.9223275	177.13837	1
CGGAC	349975	4.897773	155.53824	1
CGCGA	349495	4.891055	19.490585	5
ACGTC	465660	4.8776197	61.997467	15
CGGTT	10928205	4.809491	165.61882	1
CGGGG	5920565	4.6511316	114.15244	1
GGGCG	5734810	4.505204	109.21419	2
GATCG	4140225	4.4595675	22.786238	44

TCCCC	8125	4.3934608	15.253187	3
AGGCG	3042630	4.3786535	63.64115	47
TCGAG	4044345	4.3562922	55.27069	44
CGACG	309410	4.3300805	32.185112	24
AACCC	16800	4.1882925	6.865709	2
GAGAC	1586990	4.18371	63.494205	26
AAAAA	2485680	4.142383	12.051549	31
CACGT	387050	4.0542088	61.460995	14
AGAGC	1442725	3.8033905	24.597765	47
CTCCC	6775	3.6634705	9.782035	24
ACACC	14520	3.6198814	5.974376	47
TTACG	4436755	3.5769403	46.690914	14
CCTCC	6595	3.5661385	6.352164	41
ATCGG	3290110	3.543882	21.052273	45
CAACC	14130	3.5226533	6.442813	31
CGGTA	3256145	3.507297	122.61007	1
ACCCA	14025	3.4964762	6.5013833	33
CGAGG	2422370	3.4860365	69.76727	45
ACCAC	13955	3.4790251	5.8572264	45
TCGGA	3220000	3.4683642	21.389612	46
CCCCT	6395	3.4579918	6.987322	47
CGTTT	10343535	3.4071925	38.7195	17
GACGG	2366185	3.4051802	35.88063	28
CCCTC	6230	3.3687706	6.352145	40
GGCGG	4252575	3.3407767	37.590324	11
CCCAA	13340	3.3257036	5.6229725	40
CCACA	13280	3.3107457	6.091387	35
ACGTT	4072845	3.2835534	49.962955	16
ACCGA	127720	3.2742848	146.44917	32
TACGT	4058115	3.2716782	48.4367	15
CCAAC	13010	3.243434	6.208529	34
GGCGT	5496055	3.2316473	51.04454	3
ACGGG	2233765	3.2146144	36.008434	29
CACCA	12845	3.2022984	4.920115	41
CGGAT	2813405	3.0304077	102.912544	1
GCGGC	386080	2.9494631	10.180239	9
AGAGA	5872230	2.9161925	22.752407	25
GTCGA	2624380	2.826803	54.781406	43
TTTCG	8447395	2.7825983	14.389132	30
CGAGA	1040600	2.7432868	33.631195	25
GAGCA	1039020	2.7391217	19.399992	47
AAGCG	1036090	2.7313972	50.001583	8
CGTTC	628270	2.6888514	25.495075	33
AACTC	137405	2.6365588	109.85159	22
ATCGC	251025	2.629396	33.596214	29
AGCGA	984465	2.5953004	50.594933	9
TTCGA	3196950	2.5774014	31.84122	31
GGAGG	17258475	2.5540164	30.831272	39
TTTTT	133875565	2.5404673	5.471046	16
TTCGC	592315	2.5349722	8.59388	33
GAGGC	1744345	2.5102894	52.08929	46
GCGGG	3151800	2.4760199	38.530823	12
TCGTT	7388365	2.4337504	6.311434	4
CGTTA	3016810	2.4321716	28.433811	9
CGGTG	4122560	2.4240406	47.34294	1
GGAAG	8792840	2.3836713	11.176711	2
GGGAG	15686950	2.3214524	27.270592	38
GAAGA	4655040	2.311727	9.287907	46
TTTTA	49283430	2.288926	12.493142	26
TTCGT	6816750	2.2454586	5.5166426	35
GCGGA	1554355	2.2368748	23.15017	7
TTTAG	36018965	2.2350368	15.864881	27
ATTCG	2771360	2.2342877	38.103844	34
GTCGC	388150	2.219436	11.679084	3
GGTCG	3734570	2.1959045	31.313103	42
AGTAG	10808685	2.1931455	20.110373	35
TCGTC	512280	2.1924407	9.038552	40
CGTAG	2026025	2.1822958	24.488638	5
GAGGT	19554885	2.1659775	23.411413	40
GAGAT	10672510	2.165515	8.977261	26
CGAGT	1935330	2.0846052	38.80147	33
AAGAG	4150735	2.0612855	9.269595	47
GCGTT	4636545	2.0405383	24.441046	16
ACGGA	773905	2.0402107	9.501586	30
AGGAG	7517040	2.0378118	9.176351	38
ATTTT	43807715	2.034611	8.305327	25
GACGC	144765	2.0259335	15.729244	5
GCGGT	3439720	2.0225346	29.549477	6
CGAGC	142445	1.993466	7.812016	32
TTTAC	3279605	1.9789972	34.1803	13
TACGC	187200	1.9608523	10.074927	13
AAACG	395385	1.909429	11.47185	7
AATTT	16628785	1.8902093	18.099596	24
AGCGC	134840	1.8870367	11.487872	35
TAGAG	9295985	1.8862098	10.227992	24
TAAAA	2746325	1.8699863	5.162748	30
ATCGT	2310605	1.8628244	15.283857	39
TAGTT	29760340	1.8466787	8.962686	29
AGGTC	1708900	1.8407104	52.054554	41
GCGTA	1679410	1.8089458	24.248737	4
TTAGT	29028400	1.8012607	15.239938	28
TAGTA	11771580	1.7877498	16.389015	29
ACGGC	127210	1.7802576	9.051549	12
TACGG	1620915	1.745939	14.402417	5
GCGTC	304510	1.7411836	10.85479	40
AGGTA	8559165	1.7367047	26.819489	47
TATCG	2153045	1.7357984	15.680247	38
GGACG	1198295	1.7244682	16.780037	2
GAAAA	1886200	1.7159193	5.026923	3
GCGAC	122540	1.7149028	19.316368	23
GGAAA	3437030	1.7068543	11.6319275	2
GAGCG	1182865	1.7022629	10.1339035	28
GGAGA	6257465	1.6963507	10.739803	2
AGCGG	1157920	1.6663643	6.2724633	6
GTAGA	8212280	1.6663198	9.923399	23
TCGTA	2044105	1.6479704	6.3248515	43
TGAGA	8114295	1.6464381	5.723733	41

AGTTA	10817155	1.6428012	19.139315	30
AGTTT	26340720	1.6344856	7.9606714	26
TATTT	35192430	1.6344817	6.1944904	32
TGGGA	14590015	1.6160486	15.328175	37
AGTCG	1479875	1.5940202	13.727763	22
GTCGT	3608265	1.5879934	10.949959	3
GCGTG	2698005	1.5864108	33.627666	4
CGTGG	2693275	1.5836295	33.490353	5
CGATT	1963110	1.5826718	17.955648	11
TGGCG	2681540	1.5767293	32.578403	10
CGTAC	149830	1.5694151	8.721432	13
TCGAC	149130	1.5620828	7.9044294	23
CGAAA	318925	1.5401815	6.0870686	32
AACGG	583855	1.5391905	8.233758	29
TTGAG	18524805	1.5357826	13.395458	44
GGGAA	5601565	1.5185412	13.579733	2
TAATT	13324360	1.5145923	17.841389	23
TAGGA	7448330	1.5113097	7.1004725	37
AGCGT	1398115	1.5059539	8.048833	29
GTACG	1389245	1.4963999	14.193643	4
GGTTT	43727635	1.4811993	9.2955065	2
AACGC	57750	1.4805037	6.932483	23
ACGGT	1374330	1.4803343	13.900626	6
AGGTT	17750680	1.4716042	14.191961	41
ACGAG	554580	1.4620142	5.127673	32
TTCGG	3318865	1.4606289	21.2523	35
TTATT	31253760	1.451554	6.7540236	32
GTAGT	17215620	1.4272457	8.796601	36
ATGCC	135890	1.4233985	65.22022	47
GGAAT	7010615	1.4224949	9.896779	2
CACGC	10435	1.4201174	5.9791307	47
TTTAA	12303480	1.398548	7.8584824	5
GAACG	527465	1.3905321	8.124132	28
CGAAC	53655	1.3755226	5.2038693	12
AAGTA	3696395	1.3739437	10.340146	34
TATAG	9039835	1.3728796	15.313737	47
TTAAG	8996120	1.3662407	9.550607	6
GTTTA	21996370	1.3649114	7.8400354	12
TTATA	11980290	1.3618108	11.816551	46
GCGAT	1261205	1.3584839	21.751083	10
CAGTC	129030	1.3515425	61.118923	27
TCCAG	128595	1.346986	60.225624	25
GGCGA	934825	1.3453082	9.123216	2
GTCAC	127975	1.3404918	60.93682	29
CCGAT	127605	1.3366164	60.05829	33
AGATA	3594265	1.3359822	5.4332747	26
TGGAA	6551655	1.3293692	8.896728	1
CCAGT	126825	1.3284461	60.230553	26
GGTTA	15946605	1.322039	17.664955	2
GACGT	1207210	1.3003241	6.297352	3
GGAGT	11725500	1.2987633	10.5498085	2
GGTAG	11719980	1.298152	7.661005	2
GTTAA	8454085	1.283922	21.366615	3
TTGTA	20661170	1.28206	14.6066475	20
GGGTT	28096840	1.2715621	14.032614	2
TCGGG	2155140	1.2672094	26.998198	36
CGTAT	1561970	1.2592702	5.87256	44
AAAAC	141865	1.2550312	18.955275	6
ATTAT	11003230	1.2507473	11.781323	45
ACGAC	48520	1.2438794	5.119547	28
CGTCT	289445	1.2387581	25.040596	16
GAGTA	6079275	1.2335205	14.381903	34
TAAGC	618285	1.2199821	36.15011	7
GTAAT	7999130	1.2148279	22.338104	22
GGGGA	8166610	1.2085457	10.535154	2
TCGTG	2719600	1.196893	7.093394	40
TTTGT	47090420	1.1938987	7.0309496	19
GGGAT	10682405	1.183226	10.711602	42
GGTGG	19493520	1.178672	11.55362	8
CGTAA	593170	1.170426	8.834648	21
TATTC	1907340	1.1509376	27.248266	33
GATTA	7575855	1.150545	14.979153	44
TGAGG	10329285	1.1441132	15.64458	45
TCGAT	1417575	1.142858	6.1200123	11
AGTAT	7377325	1.1203943	15.467717	30
GTATT	17795100	1.1042157	6.8217816	31
ATCTC	139600	1.0944649	48.814144	40
GGATT	13194355	1.0938662	8.36456	43
GGGGT	18080490	1.0932335	8.634112	2
TAGGC	1012200	1.0902728	8.3846445	13
TGTAA	7133605	1.0833806	21.779207	21
TTTTC	4385040	1.0811318	10.039657	29
GGGTA	9600870	1.0634309	15.266479	2
TTAAT	9291685	1.0561945	14.454399	4
TGGAG	9466825	1.0485834	9.962742	1
CGTGC	182820	1.045362	5.193498	13
AGTAA	2791615	1.0376385	6.6094894	9
CGTGA	959785	1.0338149	7.3195143	26
CGATC	98345	1.0301282	7.1515164	44
GTTAT	16572820	1.0283712	8.058799	31
GTGGC	1748865	1.0283221	30.80161	9
AAGGC	389365	1.0264653	17.191296	46
CGTGT	2300670	1.0125223	6.798726	41
GTTGA	12163400	1.0083959	12.496549	43
TGTAG	12067335	1.0004318	7.6213098	21
AGTTG	11832335	0.98094916	8.853455	38
ATTTC	1615575	0.9748791	5.0913687	22
GGTTG	21331140	0.96537083	6.8385153	42
TTATC	1578655	0.9526007	11.481237	37
TAAGT	6257350	0.9503038	6.117203	7
TGCGG	1600765	0.9412402	6.518264	5
TGGGG	15398065	0.9310411	9.083747	1
AAGAC	192465	0.92946935	8.015868	32
TTGGG	20321905	0.91969645	6.809018	36
GTGGT	20193215	0.9138724	8.315215	9
GTTTG	26945495	0.91273284	7.0114164	18
TAAGG	4498180	0.9127071	5.4276395	45

TAGAC	459545	0.90676093	10.432752	25
GGATA	4466430	0.9062648	7.5026064	2
TGGTT	26598500	0.9009789	7.588142	1
ATTAC	609425	0.9000413	5.170264	29
GGAGC	620920	0.89356685	9.268025	27
TTTGG	25991430	0.8804155	5.4785395	35
AGTGA	4314940	0.8755266	5.3773084	18
GGGTG	14107350	0.85299826	8.362859	2
GGTAC	778885	0.8389617	14.308113	3
GGTAT	9907610	0.82138157	6.279159	2
GGGGG	10151895	0.8201106	6.1806846	2
GAAGC	304495	0.8027264	9.577767	4
GTGCG	1324280	0.7786686	6.036007	4
GGAAC	294690	0.7768779	6.97584	27
CGATG	714205	0.76929295	6.8513536	34
AGTGG	6865650	0.7604669	5.363802	8
TGGGT	16759555	0.7584773	9.193929	1
GGTAA	3721720	0.7551588	6.3442774	2
TGGTG	16644415	0.75326645	6.1196966	1
GAGTC	629860	0.6784422	12.600381	21
TGGTA	7846260	0.6504872	5.423773	1
TCTCG	148015	0.6334702	26.663252	41
CTCGT	146650	0.6276283	26.682388	42
TGAAC	265055	0.52299887	12.162031	20
TGGAT	6134525	0.5085773	5.1236296	1
TGGGC	854370	0.5023644	5.1411195	13
GATTC	622870	0.50216174	5.2669044	29
GTATC	528530	0.42610425	5.211029	38
GGTGC	693390	0.40770915	6.1045957	3
CTGAA	206045	0.40656206	11.939216	19
GAACT	159340	0.31440508	11.870514	21
TATGC	383045	0.3088133	5.1592402	46
AGTCA	138590	0.2734618	11.84202	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	179379	0.2874267835153267	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	159638	0.2557949195101975	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	112476	0.18022519304193851	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCC	80863	0.12957030642048326	TruSeq Adapter, Index 2 (100CGGGCGT
77163	0.12364163528837352	No Hit	
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	74628	0.11957969439110377	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	63869	0.10234008014505824	No Hit