FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD17_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`

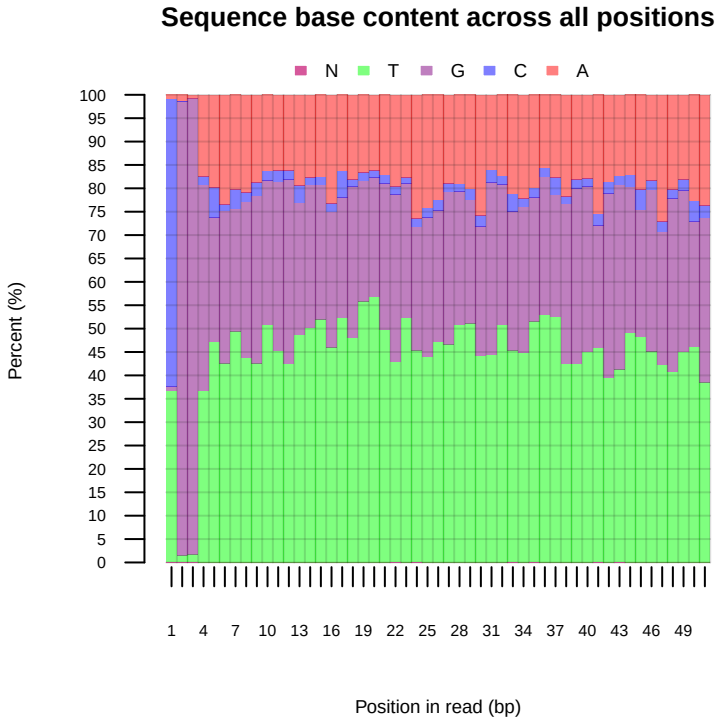
1 Sequence Duplication

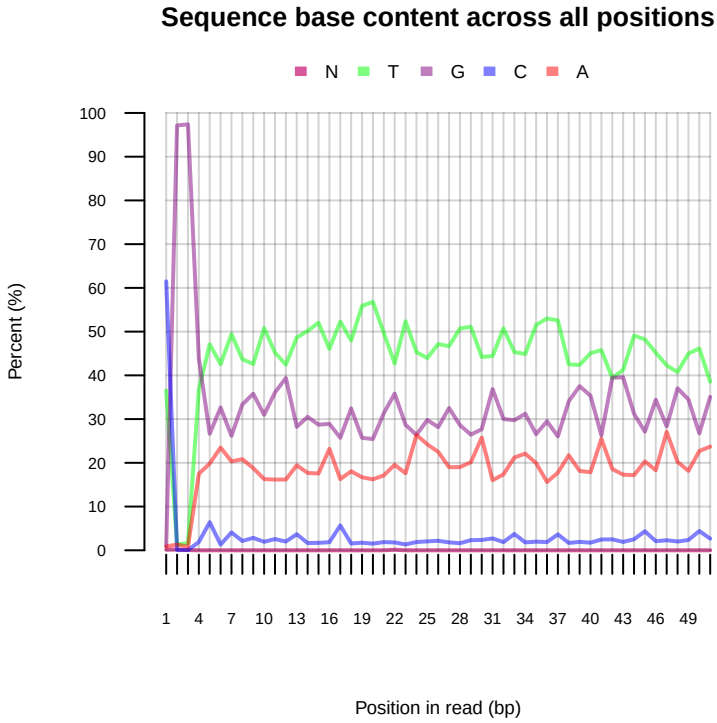
- Estimated Duplication rate 82.2662%

2 Per base sequence quality

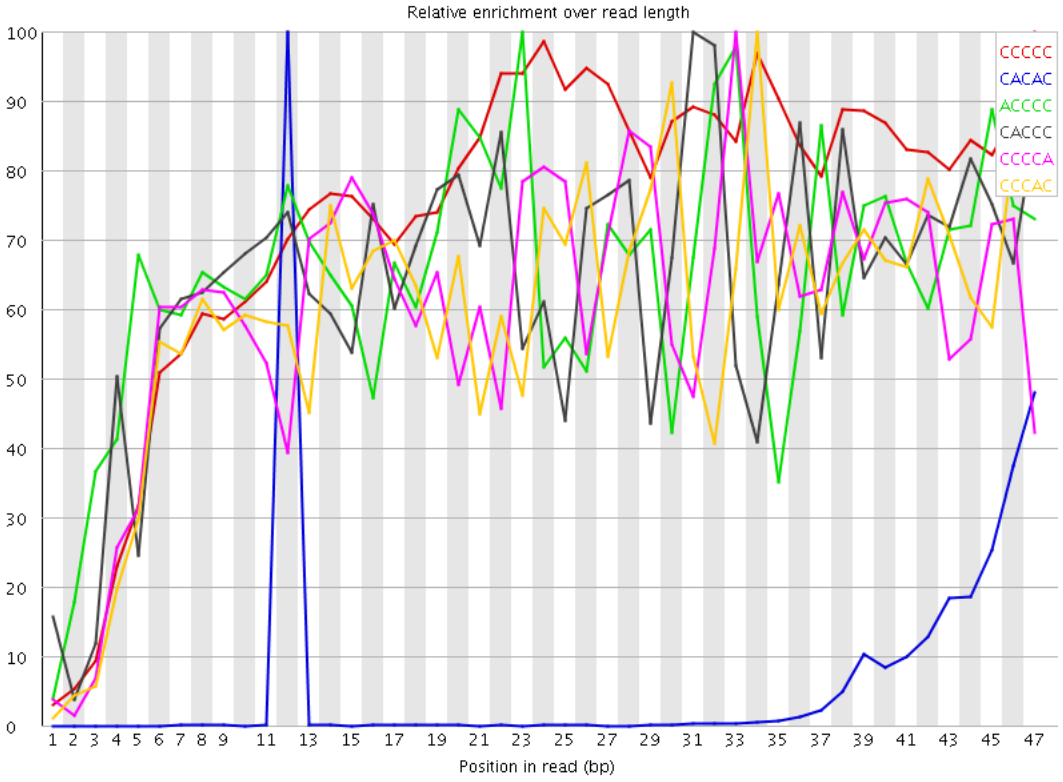
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	227745	1108.4519	1500.2717	47
CACAC	695875	119.79283	1825.4543	12
ACCCC	60260	55.15856	85.1604	23
CACCC	56990	52.16539	81.71942	31
CCCCA	55025	50.36674	83.22478	33
CCCAC	54930	50.279785	84.084984	34
CCACC	54420	49.812958	82.36458	35
CGGGC	5025285	28.084572	1029.2185	1
CACGC	232575	22.292555	973.1261	31
ACGCC	228840	21.934553	969.6133	32
CGCCA	227485	21.804676	969.5006	33
AGCAC	931105	16.78461	199.07878	10
ACTCC	230570	16.476936	724.609	23
CTCCA	227465	16.255045	722.52576	24
GCACA	814205	14.677307	198.60847	11
CGCGG	2396175	13.391392	328.02057	5
GCCCC	25755	13.126308	24.068256	45
CCCGC	25175	12.830705	24.666424	34
GCGCG	2284295	12.766132	325.56787	4
CGCCC	24700	12.588616	23.229544	36
CCCCG	24545	12.509619	21.67342	46
CCGCC	24240	12.354173	24.187464	35
CGCGC	216145	11.535577	50.305206	13
CGGCG	1924620	10.756034	304.52417	1
ACACG	578865	10.43494	192.56459	13
CGGAA	5398165	10.189944	202.15327	1
GGCGC	1708245	9.546789	326.56754	3
CGGGA	7039410	7.3987646	223.56194	1
TCGCG	1551105	6.4628654	26.516655	30
AGACG	3168475	5.98103	68.91521	27
CGGGT	13472250	5.8781013	227.35149	1
CGGAG	5305820	5.5766764	164.97754	1
AGATC	3754255	5.2835546	27.20295	43
ACGCG	525540	5.2749176	20.034723	6
CGCGT	1263515	5.2645874	24.167278	31
TCCCC	13695	5.2037916	15.452547	3
CGTCG	1230650	5.127651	20.194464	41
CGGTT	15370560	4.999924	174.85527	1
CGGGG	8359510	4.892164	121.657196	1
CTCCC	12855	4.8846107	14.551352	24
CCTCC	12695	4.8238134	9.284297	24
ACGTC	639480	4.7853494	80.17658	15
CGGAC	466410	4.6814213	147.05603	1

CGGTC	1122655	4.677677	169.86523	1
AGGCG	4405605	4.630507	73.52113	47
GGGCG	7771915	4.548291	113.17905	2
TCGAG	5752495	4.50771	64.08986	44
GAGAC	2329425	4.397182	66.167786	26
CGCGA	435855	4.374737	20.206823	5
AACCC	24865	4.2804365	7.1680474	22
CCCCT	11230	4.267147	7.141931	46
CCCTC	10960	4.1645527	7.588127	24
GCCAA	230440	4.1540384	187.44427	34
CGACG	405325	4.0683026	26.066772	24
AAAAA	3547300	4.062047	10.41479	31
CACGT	505685	3.784136	80.11153	14
CGAGG	3586115	3.7691824	81.28975	45
TTACG	6354295	3.712307	54.050972	14
ACACC	21555	3.71063	5.9857593	37
CAACC	21535	3.707187	6.228425	31
CCCAA	21160	3.6426313	6.4306464	16
CGGTA	4616375	3.6174355	127.48829	1
ACCAC	20815	3.583241	5.9454465	45
CCAAC	20740	3.57033	6.875534	34
GACGG	3365690	3.537505	38.090057	28
CCACA	20540	3.5359	5.621888	46
ACCCA	20540	3.5359	6.18798	33
CACCA	20185	3.474788	5.2982054	36
GATCG	4426135	3.4683616	16.272526	44
CGTTT	13957530	3.385003	40.030464	17
TACGT	5765505	3.3683233	55.21353	15
ACGTT	5750825	3.3597476	56.703186	16
ACGGG	3169430	3.3312263	38.074753	29
GGCGG	5656585	3.310355	34.05688	11
AACTC	241610	3.2471647	140.66614	22
GGCGT	7281705	3.177094	50.920868	3
CGGAT	3924070	3.0749385	104.80503	1
CCAAT	227930	3.06331	139.83118	35
GTCGA	3906805	3.0614097	63.51791	43
AGAGA	8565340	3.040788	23.937555	25
AGAGC	1568930	2.9616196	21.820196	8
CGAGA	1462800	2.761281	33.389896	25
TTTCG	11307600	2.7423377	13.548039	30
GCGGC	487190	2.7227356	8.920552	9
AAGCG	1440480	2.7191489	49.68873	8
GAGGC	2572670	2.7040021	59.688267	46
TCGGA	3371870	2.6422296	14.720919	46
ATCGG	3329880	2.6093256	14.300593	45
CGGTG	5973935	2.6064982	49.8111	1
GGAGG	23637845	2.6016202	31.98673	39
AGCGA	1371830	2.5895598	50.479904	9
ATCGC	345170	2.5829718	37.554855	29
TTCGA	4366630	2.5510728	29.519444	31
TTTTT	180050745	2.5416281	5.4518423	16
CGTTC	799255	2.4828246	23.760902	33
GCGGG	4126140	2.4147055	34.861423	12
TCGTT	9807765	2.378595	5.8411746	4
GGGAG	21585785	2.3757672	28.8652	38
TTCGC	756065	2.348658	6.965447	33
GGTCG	5354145	2.3360765	36.65226	42
CGTTA	3994600	2.3337255	23.608671	9
TTTTA	67960975	2.3110147	12.726416	26
TTTAG	49238590	2.2458043	16.169933	27
GCGGA	2127515	2.2361224	22.517721	7
GGAAG	11306425	2.2349308	11.236018	2
TTCGT	9141495	2.2170103	5.38022	35
AGTAG	14964745	2.2053874	18.819017	35
CGTAG	2804330	2.1974998	25.565462	5
GAGGT	26619365	2.184291	24.305794	40
ACGGA	1147875	2.1668074	9.150082	30
GCGGT	4879385	2.1289332	33.42884	6
ATTCG	3639460	2.1262455	34.370926	34
ATTTT	61620795	2.0954163	9.0014305	25
TCGTC	673585	2.0924404	8.003465	40
GAGCA	1104500	2.0849297	21.399323	9
GTCGC	500365	2.084831	11.341974	3
GACGC	207285	2.080548	17.735502	5
GAAGA	5782775	2.0529475	6.5460515	46
AGGTC	2610370	2.0455108	60.41982	41
TTTAC	4672250	2.0350697	39.41539	13
CGAGT	2592280	2.031335	35.82654	33
AGGAG	10138335	2.0040355	8.294689	38
GCGTT	6123155	1.9918149	22.185753	16
GAGAT	13350690	1.9675202	9.459193	26
TACGC	261570	1.9573773	8.176941	13
AATTT	23790465	1.9488211	19.41857	24
CGAGC	193310	1.940279	6.7301035	32
TAGAG	13131320	1.9351913	10.844522	24
ATGCC	253670	1.8982601	84.489426	47
ATCGT	3222305	1.8825351	15.752837	39
AAACG	555220	1.8823205	8.408684	7
TAGTA	17028945	1.8710269	18.903782	29
AAGAG	5230150	1.8567595	6.5423856	47
TACGG	2350630	1.8419759	16.56289	5
AGCGC	183365	1.8404597	13.028669	35
ACGGC	182160	1.8283653	11.5831375	12
TTAGT	39766935	1.8137959	15.526152	28
GCGTA	2313015	1.8125006	25.339018	4
TAGTT	39369290	1.7956591	8.068507	29
TCACG	237090	1.7741892	78.98987	30
GGACG	1672365	1.7577376	16.16232	2
CAGTC	234100	1.7518145	79.23424	27
TATCG	2997945	1.7514595	16.138538	38
TCCAG	233790	1.7494947	78.27783	25
GTCAC	232725	1.7415249	79.27818	29
GAAAA	2729535	1.7403352	5.126157	3
CCAGT	231415	1.731722	78.25673	26
GGAAA	4848585	1.7212999	12.237635	2
GTAGA	11604110	1.7101232	10.514059	23
AGGTA	11601855	1.7097907	25.102976	47

GGAGA	8602990	1.7005452	10.844038	2
TCGTA	2896770	1.6923512	8.068875	43
GAGCG	1600485	1.6821883	9.446213	28
AGCGG	1586275	1.6672527	5.11278	6
TATTT	48736110	1.6572725	7.0670404	32
TGGGA	19886300	1.6317993	16.99037	37
CGTGG	3699375	1.6140811	32.47689	5
AGTTT	35379690	1.6136907	7.5377016	26
GCGTC	387095	1.6128778	8.935702	40
AGTTA	14590320	1.603087	16.795835	30
GTCGT	4902085	1.5946105	10.510442	3
GCGTG	3653930	1.5942527	32.60953	4
GCGAC	158500	1.5908862	15.231192	23
AGTCG	2028515	1.5895637	13.901524	22
CGATT	2710895	1.5837591	18.040398	11
TAATT	19293025	1.5804083	19.177671	23
CGAAA	462670	1.568555	5.4193926	15
AACGG	825670	1.5585912	7.7118444	29
CGTAC	207630	1.5537344	7.837628	13
GTACG	1964345	1.5392795	16.194504	4
ACGGT	1949990	1.528031	15.776995	6
TGGCG	3478815	1.5178483	29.374989	10
GGGAA	7648650	1.5119016	13.369178	2
GGTTT	59414260	1.5088784	9.275479	2
AACGC	83690	1.5086422	8.86422	11
AGCGT	1922815	1.506736	7.5172367	29
TAGGA	10161875	1.4975777	6.387379	37
ACGAG	791490	1.4940708	5.5804324	32
AGGTT	24190625	1.4799145	14.110634	41
TCGAC	197175	1.4754976	6.7458572	23
GGAAT	9833415	1.4491718	10.276037	2
TTATT	42094415	1.4314216	6.0688357	32
GAACG	752255	1.420008	7.514934	28
TTCGG	4355925	1.4169488	19.29661	35
TTTAA	17264335	1.414226	7.8410583	5
TTGAG	23044815	1.4098171	12.523411	44
GGCGA	1340955	1.4094095	9.9353285	2
ATCTC	251735	1.4044523	62.974197	40
GTAGT	22872825	1.3992952	8.352107	36
TTAAG	12665410	1.3915907	9.570553	6
GTTTA	30509825	1.3915731	9.067842	12
GCGAT	1762775	1.3813272	22.001488	10
GGTTA	22500115	1.3764938	18.843712	2
AAGTA	5198315	1.3758811	9.545003	34
TATAG	12430870	1.365821	14.351126	47
TTATA	16673145	1.3657985	11.091514	46
AGATA	5150145	1.3631313	5.708707	26
CGAAC	75590	1.3626269	7.6910834	9
GTTAA	12276410	1.34885	22.773148	3
TGGAA	9086105	1.3390392	8.4567375	1
GGTAG	15987685	1.3118927	7.965857	2
TTGTA	28623570	1.3055397	14.878536	20
CGTAT	2232325	1.304169	7.4603443	44
GACGT	1659715	1.3005685	6.194182	3
GGAGT	15836795	1.2995113	10.492719	2
GGGTT	37713040	1.2846308	13.689826	2
GTAAT	11409125	1.2535585	23.933096	22
TCGTG	3849505	1.2522142	7.8817687	40
ATTAT	15252925	1.2494594	11.003339	45
CGTCT	401520	1.2472912	33.0253	16
TCGGG	2813865	1.2277226	24.530106	36
GAGTA	8302680	1.223584	13.532463	34
TAAGC	868100	1.2217214	35.88056	7
GGGGA	11046345	1.215779	10.784947	2
TTTGT	62868010	1.1903372	7.2181144	19
AGTAT	10790760	1.1856167	17.982931	30
GGTGG	25883880	1.1826041	11.963312	8
AAAAC	193895	1.1805866	13.140642	6
GGGAT	14328325	1.1757315	10.208688	42
CGTAA	820260	1.1543938	7.3760886	21
TCGAT	1974415	1.1534929	5.784979	11
GTATT	25258325	1.1520487	7.9826317	31
GATTA	10372370	1.1396468	14.038009	44
AAGGC	593145	1.1196611	20.439138	46
TGTAA	10171010	1.1175227	23.473577	21
TGAGG	13585315	1.1147628	14.877558	45
GGGGT	24387130	1.1142193	8.756478	2
TTAAT	13430175	1.100147	15.430542	4
GGATT	17804050	1.0892018	7.958356	43
TAGGC	1384505	1.0849113	7.331663	13
GGGTA	13152740	1.0792673	15.350447	2
CGTGT	3280995	1.0672823	7.5726438	41
TTTTC	5895305	1.0659419	9.333677	29
TATTC	2425540	1.056481	23.99356	33
AGTAA	3978620	1.0530542	6.5942197	9
TGGAG	12664270	1.0391852	9.470925	1
CGTGA	1289620	1.0105585	6.926087	26
GTTAT	21992295	1.003083	7.191667	31
TGTAG	15948830	0.9757046	6.892041	21
TGCGG	2216325	0.9670088	7.4958496	5
TTATC	2204720	0.9602995	11.774619	37
GTGGC	2197335	0.95872325	27.696115	9
GTTGA	15670380	0.9586699	11.879501	43
AGTTG	15630735	0.95624447	8.370018	38
GGTTG	28057115	0.95571804	6.659728	42
TAAGT	8627815	0.9479667	6.1129045	7
TTGGG	27793590	0.9467416	7.5872784	36
TAAGG	6412370	0.94500494	6.2141905	45
ATTAC	896240	0.94038033	5.6076427	29
CGATC	124960	0.93509907	7.0729976	44
TGGGG	20391395	0.93165886	8.809938	1
AAGAC	271455	0.92029345	6.2748356	32
GGATA	6223395	0.91715527	7.6356764	2
TGGTT	36027780	0.9149578	7.250146	1
GTGGT	26857240	0.91484636	8.980982	9
GTTTG	35783315	0.9087494	7.2584	18
TTTGG	35728290	0.9073519	6.0714216	35

TAGAC	625325	0.8800518	8.843398	25
GGAGC	830485	0.8728804	8.530334	27
GGTAC	1108820	0.86888194	16.293093	3
AGTGA	5869680	0.8650276	5.548479	18
GGGTG	18585735	0.8491604	8.390137	2
GGTAT	13483230	0.8248661	6.4007187	2
GGGGG	13250925	0.8120438	6.3748503	2
GTGCG	1859375	0.8112674	6.9683475	4
TCTCG	260015	0.8077167	35.0428	41
GGAAC	427470	0.8069216	7.393223	2
CTCGT	258870	0.8041599	35.043568	42
GGTAA	5221605	0.7695193	6.5171847	2
GAAGC	403645	0.76194793	7.377454	4
AGTGG	9228265	0.75723875	5.6485243	8
TGGGT	22108930	0.7531032	8.7743225	1
TGGTG	21493500	0.7321397	5.8650546	1
GAGTC	867065	0.6794404	12.657943	21
TGGTA	10498775	0.64228565	5.3561654	1
ATATC	587245	0.6161672	12.071177	38
CAATA	240565	0.6080477	27.33762	36
TGAAC	411105	0.5785691	15.772892	20
CACAT	33965	0.45647928	5.873004	12
CTGAA	318045	0.44760096	15.4659815	19
GGTGC	985425	0.4299526	7.0344257	3
GAACT	269405	0.3791474	15.438833	21
AGTCA	248515	0.34974784	15.358404	28
TATGC	592305	0.34603646	6.7555304	46
TCTGA	349050	0.20392202	6.4427714	18

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	245918	0.2900518996849611	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	206028	0.2430030042058457	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCC	144357	0.17026416156125998	TruSeq Adapter, Index 6 (100CGGGTTTA
118813	0.1401358841454033	No Hit	
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	116973	0.13796566685581763	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	98206	0.11583062996796205	No Hit