FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD11_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`

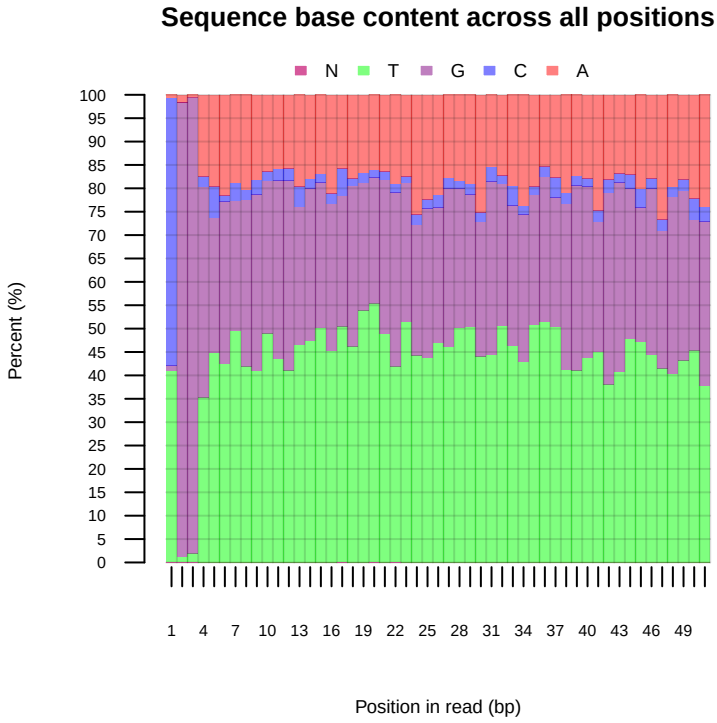
1 Sequence Duplication

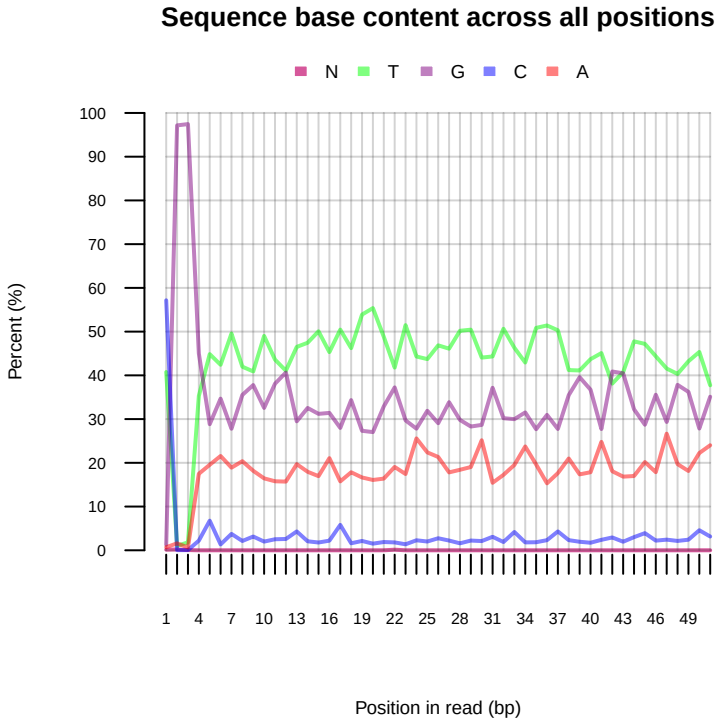
- Estimated Duplication rate 83.1924%

2 Per base sequence quality

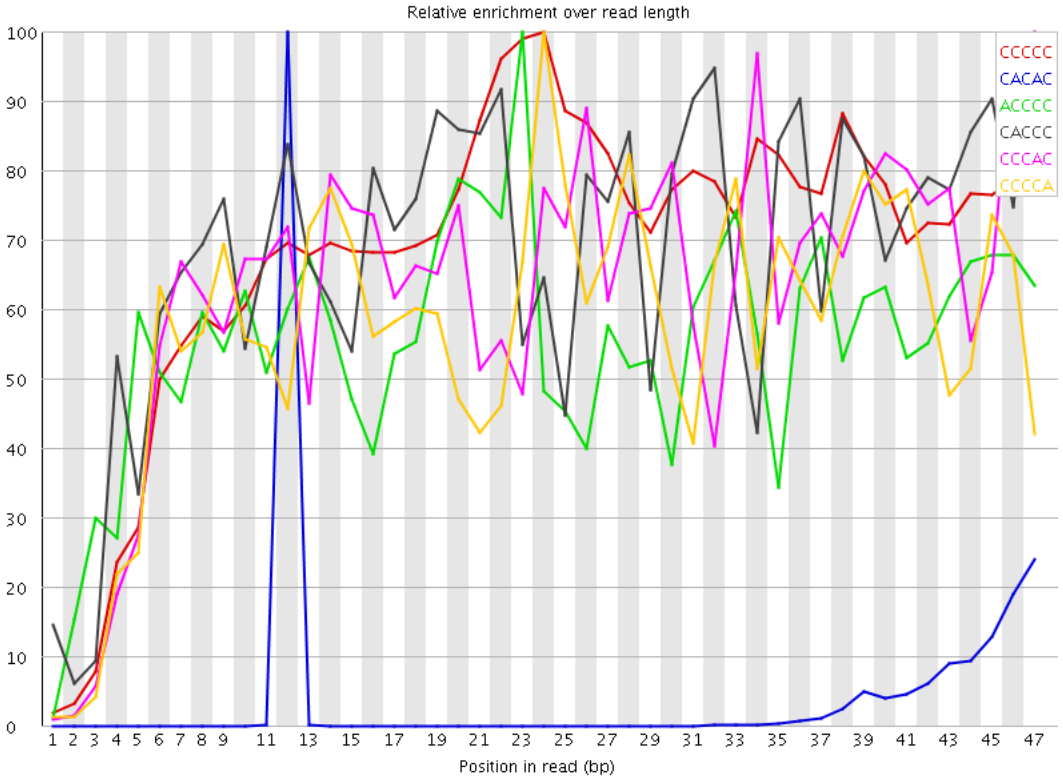
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	184445	886.95	1282.29	24
CACAC	977290	182.95105	4232.3794	12
ACCCC	49975	47.415974	85.15193	23
CACCC	47210	44.792557	65.09116	47
CCCAC	45215	42.899715	67.76614	47
CCCCA	44710	42.420574	73.78329	24
CCACC	44385	42.112217	67.76614	47
ACTCC	478335	37.022892	1678.1232	23
CTCCA	473075	36.615772	1670.918	24
CGGGC	5068655	26.61556	921.40485	1
AGCAC	1235785	23.822714	452.20792	10
GCACA	1112155	21.439451	451.05725	11
ACACG	858135	16.542606	438.7918	13
CGCGG	2196890	11.5358925	258.5908	5
GCCCC	22295	11.040192	22.104977	47
CGGAA	5514165	10.94624	183.6707	1
GCGCG	2062875	10.832178	255.68425	4
CCCCG	21785	10.787646	21.407032	41
CCCGC	21755	10.772791	21.988771	42
CGCGC	210295	10.72346	56.18715	13
CGCCC	21390	10.592048	18.847364	44
CCGCC	21215	10.50539	19.31236	27
CGGCG	1798390	9.443365	248.88281	1
CACGA	476500	9.1856785	420.961	36
CACAT	524920	8.016244	343.5415	31
GGCGC	1506120	7.9086504	256.42184	3
AACTC	495830	7.5719986	340.175	22
TCACA	491270	7.5023623	343.43747	30
ACATC	488815	7.46487	343.1217	32
CATCA	486915	7.4358554	342.68042	33
ATCAC	478310	7.3044443	333.05774	34
ACGTC	895415	7.136733	181.3751	15
CGGGA	6589305	6.8268766	186.35728	1
TCGCG	1546975	6.435104	19.947786	30
CACGT	771945	6.152639	181.70842	14
AGATC	3750290	5.8976564	31.05292	43
CGTCG	1237980	5.1497474	25.156506	41
ACGCG	510105	5.1322193	26.597113	6
CGGGT	11890955	5.0936127	192.46194	1
CGCGT	1207305	5.0221457	18.150524	31
CGATC	613815	4.8922944	181.19348	38
CGGAG	4654155	4.8219566	139.6795	1
CGGTC	1117910	4.6502805	165.35602	1

AGACG	2332300	4.629878	46.248463	27
CGCGA	451115	4.5387144	17.073454	5
CGGAC	443335	4.460439	134.9111	1
CGACG	418455	4.2101192	31.317484	24
GGGCG	7771140	4.2020807	99.62177	2
ATGCC	515610	4.109569	190.62762	47
TCCCC	10415	4.0856194	14.384903	3
CGGTT	11966340	4.0606914	132.3992	1
TCGAG	4851640	3.9819925	47.7698	44
GTCAC	492420	3.924738	179.91449	29
CAGTC	488560	3.8939734	178.9333	27
TCCAG	486770	3.8797064	177.62437	25
CGGGG	7169745	3.87689	99.38668	1
AACCC	20660	3.8676019	9.908963	22
CCAGT	484080	3.8582664	177.48395	26
TCACG	483110	3.850535	174.20511	35
AGGCG	3706075	3.8396943	53.242214	47
GATCG	4660550	3.825155	19.083382	1
AGAGC	1873945	3.7199922	48.021313	8
CTCCC	9315	3.6541088	19.722887	24
AAAAA	2468120	3.548922	9.519495	31
CCTCC	8810	3.4560063	7.7416945	24
CCCCT	8800	3.4520833	6.9123697	46
CGTTT	12811805	3.4441204	41.0979	17
TTACG	5199590	3.3807292	42.044212	14
ACACC	17700	3.3134828	5.4976816	13
ATCTC	523390	3.3046837	150.95728	40
CCCTC	8325	3.2657495	6.1749234	38
CGGTA	3968160	3.2568748	112.01755	1
CAACC	17255	3.2301776	5.4537	31
GAGAC	1603905	3.1839323	44.131554	26
GGCGT	7396850	3.1685164	50.588318	3
ACCAC	16715	3.1290884	4.7060275	10
TACGT	4765725	3.0986342	43.858353	15
GGCGG	5730070	3.0984151	39.128147	11
CCACA	16515	3.0916479	5.5856447	30
ACGTT	4745685	3.0856044	45.58435	16
CGAGG	2951415	3.0578258	57.280727	45
CCCAA	16300	3.0513992	5.409719	15
TCGGA	3674325	3.015709	19.706839	3
CCAAC	16085	3.0111508	5.7615705	30
ACCCA	16040	3.0027268	4.9259357	23
CACCA	15975	2.9905586	5.233793	16
ATCGG	3634060	2.9826617	19.08713	2
CGTTC	860165	2.8345459	34.141197	33
GAGCA	1417780	2.814453	47.600628	9
CGGAT	3426040	2.8119287	91.71087	1
TTTCG	10406130	2.7974176	13.783279	30
GCGGC	529200	2.7788348	9.616526	9
GACGG	2653295	2.7489572	24.248867	28
AGAGA	6796110	2.6618598	16.857904	25
TTCGA	4014855	2.6104248	30.46299	31
GTCGA	3145860	2.5819705	47.31305	43
TTTTT	146578710	2.5464473	5.4726286	16
ATTCG	3902525	2.5373886	46.966175	34
TTCGC	767575	2.5294294	8.412587	33
CGAGA	1272480	2.5260165	29.564064	25
ACGGG	2428890	2.5164616	24.045135	29
TCGTT	9232180	2.4818313	6.474104	36
ATCGC	310690	2.4762948	29.507626	29
GGAGG	23078285	2.462197	28.381214	39
CGTTA	3645890	2.3705268	28.796028	9
GCGGG	4375085	2.3657358	40.13881	12
AAGCG	1182550	2.3474948	39.373135	8
GGAAG	11481045	2.346948	10.771662	2
AGCGA	1174310	2.3311377	39.943523	9
TTTTA	55069515	2.31392	14.079785	26
TTCGT	8544140	2.2968698	6.288547	35
GAAGA	5817290	2.2784812	10.751198	6
GAGGC	2157570	2.2353592	45.004257	46
TTTAG	42080480	2.2319698	17.01878	27
CGGTG	5190835	2.2235475	44.301697	1
GGGAG	20716290	2.2101984	24.826536	38
ACGGA	1111645	2.2067409	16.42137	30
CGTCT	665445	2.1928751	74.728485	16
TCGTC	662720	2.1838953	10.232997	40
GTCGC	524420	2.1814814	11.332857	3
GAGGT	25403165	2.1470222	22.778137	40
GCGGA	2071625	2.1463156	19.688364	7
AGAAA	2844675	2.1348145	5.3410935	22
CGTAG	2583690	2.1205683	23.259684	5
AGGAG	10215010	2.0881457	9.349782	38
GCGTT	6048355	2.0524657	28.175756	16
AGTAG	12535600	2.0300026	18.241371	35
AAGAG	5173560	2.026349	10.638015	7
GGTCG	4726830	2.024786	26.181707	42
GACGC	200435	2.016597	24.701315	5
ATTTT	47688500	2.003783	8.412637	25
CGAGT	2420125	1.9863222	35.067116	33
TTTAC	3804205	1.9594523	32.47518	13
TACGC	244130	1.945791	10.312127	13
GAGAT	11872180	1.922569	7.5565166	26
TAGTT	36166455	1.9182867	10.880931	29
GCGGT	4369180	1.8715831	26.738697	6
TCGTA	2873695	1.8684522	16.977585	43
CGAGC	182710	1.8382642	7.724775	32
AATTT	17905765	1.8197144	18.11728	24
AGGTA	11002095	1.7816683	29.677502	47
AAACG	467980	1.7799802	8.227449	7
TTAGT	33551870	1.7796081	16.411026	28
TCTCG	531470	1.7513804	78.785866	41
CTCGT	530865	1.7493867	78.84713	42
GCGTA	2124210	1.7434493	22.899868	4
ATCGT	2678195	1.7413398	11.681706	39
TAGAG	10689625	1.7310673	7.799441	24
AGTTA	13489395	1.730507	23.372046	30
GCGAC	171860	1.7291013	18.345203	23

GGAGA	8400350	1.7171941	11.004435	2
GCGTC	410490	1.7075558	11.817856	40
ACGAT	1085490	1.7070272	36.250015	37
GGAAA	4327990	1.6951612	10.6636505	2
AGTTT	31797790	1.6865705	7.9541287	24
AGCGC	166850	1.678695	7.9351482	35
TAGTA	13067210	1.6763464	14.717249	29
TATTT	39464110	1.6582093	5.979165	32
AGGTC	2020030	1.6579435	44.970974	41
GAGAA	4222620	1.6538905	5.1290135	3
GGACG	1593670	1.6511284	14.589228	2
AACGG	826840	1.6413705	14.51992	29
GAGCG	1561070	1.617353	8.96167	28
GTCGT	4760175	1.6153308	10.886576	3
TGGCG	3759240	1.610309	35.432144	10
TATCG	2455650	1.5966427	12.158035	38
ACGGC	158075	1.5904089	6.389249	12
TAGCG	1935950	1.5889345	5.243777	10
TACGG	1933395	1.5868376	11.727221	5
AGCGG	1516080	1.5707409	5.4840636	11
GCGTG	3665040	1.5699575	34.75405	4
CGTGG	3654210	1.5653182	34.569176	5
GGAAT	9660345	1.5643867	10.43812	2
AACGC	80855	1.5586736	6.825279	11
GAACG	779620	1.5476335	14.387934	28
TCGAA	983495	1.5466311	5.5234947	31
TCGAC	193800	1.5446452	8.168078	23
CGATT	2373400	1.5431646	15.846826	11
CGTAC	193150	1.5394645	8.600616	13
TAGGA	9484680	1.5359397	7.578432	37
TGGGA	18130865	1.5323828	13.617462	37
GTAGA	9452970	1.5308046	7.5004177	23
AGTCG	1861735	1.5280224	12.222792	22
TTCGG	4500535	1.5272238	24.682426	35
TTATT	35856250	1.5066137	8.578828	32
TTGAG	22421175	1.5011932	14.513031	44
TGGAA	9232715	1.4951367	8.969467	1
AGGTT	22248140	1.4896077	15.068252	41
GGGAA	7270710	1.4862739	12.680085	2
CGTAT	2248965	1.4622581	16.42606	44
GTAGT	21720025	1.4542482	9.065925	22
TAATT	14223430	1.4454887	17.919842	23
GGTTT	52025570	1.4401964	8.662551	2
AGCGT	1746370	1.4333364	7.453957	29
ACGAG	718815	1.4269288	5.479075	32
GTACG	1723920	1.4149106	11.489656	4
TTTAA	13757250	1.398112	7.3430386	5
TATTC	2699385	1.3903866	35.707653	33
AAGTA	4449760	1.3806729	9.676918	34
CGAAA	362610	1.3792013	6.9227653	32
GTTTA	25813360	1.3691536	7.3952756	4
TTATA	13365530	1.3583026	11.993162	46
ACGGT	1648100	1.3526812	11.158016	6
TATAG	10488855	1.3455783	14.72398	47
CGAAC	69195	1.3338993	7.1105976	29
CACGC	13510	1.3199701	8.493147	12
TTAAG	10269245	1.3174053	8.400288	6
TTGTA	24571255	1.3032718	15.775773	20
TCGGG	3035860	1.3004417	29.713833	36
GTTCG	3825500	1.2981557	5.3759637	34
GGAGT	15246995	1.2886441	9.834342	2
GGTAG	15058470	1.2727103	7.4397373	2
ACGAC	66015	1.2725971	6.4720025	28
GACGT	1535780	1.2604942	5.88836	3
GGCGA	1214460	1.2582463	7.8298707	2
TTTGT	56590575	1.241019	7.6658382	19
GAGTA	7629110	1.2354505	13.101577	34
ATTAT	12081920	1.2278528	11.9175	45
AAAAC	167790	1.2228036	13.533081	6
GGTTA	18146645	1.2149951	14.163394	2
GCGAT	1475580	1.2110851	17.499954	10
GGGTT	34282500	1.1979755	12.6658125	2
GGGGA	11215695	1.1965903	10.482477	2
GGTGG	27056635	1.1934928	11.022929	8
GTAAT	9274280	1.1897649	21.424793	22
TCGAT	1813745	1.1792817	8.412004	11
TCGTG	3463355	1.1752644	5.246611	40
TGAGG	13874740	1.1726639	16.496689	45
ATTTA	11534100	1.1721793	5.031278	34
GTTAA	9013120	1.1562616	16.443632	3
CGTAA	728575	1.1457473	7.992137	21
CGCAC	11705	1.1436158	7.3224697	12
GATTA	8844260	1.134599	14.407116	44
GGGAT	13313625	1.1252397	9.398151	42
AGTAT	8645465	1.1090964	13.787742	30
TTTTC	5149010	1.0965317	10.179492	29
CGTGA	1331965	1.0932127	9.689797	26
GTTAT	20580255	1.0915871	9.922515	31
TGTAA	8473670	1.0870574	20.896679	21
GTGGC	2531470	1.0843811	33.594566	9
CGTGC	260080	1.0818803	5.1328096	13
GTATT	20335900	1.0786264	6.161437	31
TGGAG	12727405	1.0756936	10.260086	1
GGATT	15981380	1.0700214	7.711203	43
TAAGC	678300	1.0666856	30.008722	7
TGTAG	15725275	1.0528741	9.043826	21
TGAAC	666275	1.0477753	36.80201	20
GTTGA	15589370	1.0437747	13.740613	43
ATTTC	2010340	1.0354766	7.161068	22
GGGTA	12153975	1.0272285	13.816427	2
GGGGT	23099540	1.0189416	8.091135	2
TAGGC	1239880	1.0176338	9.348107	13
AGTAA	3231215	1.0025824	5.675081	9
TTAAT	9740420	0.9898925	11.634675	4
AGTTG	14723810	0.98582166	8.153867	38
GGTTG	28119635	0.982619	7.2503557	42
GAATA	3111170	0.96533483	5.4848824	3

GTTTG	34318950	0.9500334	7.4400654	18
GTGGT	27106995	0.9472331	7.7500067	9
GGAAC	472695	0.9383529	13.320384	27
TAAGT	7282505	0.934247	5.3098164	7
AAGAC	241965	0.9203234	7.9298587	32
CTGAA	584370	0.9189725	36.663647	19
GGATA	5609550	0.90840495	7.1167336	2
TGGGG	20549275	0.9064473	9.037709	1
TGGTT	32718225	0.9057215	7.687707	1
TGCGG	2095695	0.8977124	5.517546	5
TTATC	1738550	0.8954842	9.357626	37
TTTGG	32014745	0.88624734	5.2287645	35
TTGGG	25139700	0.87848747	6.1613464	36
AGTGA	5380340	0.8712869	5.022523	18
AAGGC	435990	0.8654893	12.303455	46
ATTAC	690045	0.85964954	5.1448374	29
GGAGC	828780	0.85866094	8.208312	27
GGGTG	19412385	0.8562979	7.8534436	2
GAACT	533085	0.83832234	36.47941	21
TAGAC	530010	0.83348674	9.519992	25
GGTAT	12331180	0.825625	6.1508393	2
TGGTG	23261670	0.81286114	6.6230984	1
AGTCA	507910	0.7987325	36.23708	28
TGGGT	22108040	0.77254844	9.129457	1
GGTAC	930755	0.76391894	11.557554	3
GGTAA	4614500	0.74726754	6.073867	2
GGGGG	13397265	0.74598944	5.57942	2
GAAGC	366445	0.72743464	8.561406	4
TGGTA	10149015	0.67951983	5.682937	1
GAGTC	776240	0.63710046	11.182842	21
GATTC	885845	0.57596886	7.357512	29
TATGC	832000	0.54095936	15.692382	46
TGGAT	7978105	0.5341681	5.2571898	1
TGGGC	1171890	0.5019911	5.7085595	13
TCTGA	616790	0.40103164	15.212671	18
CATAC	26135	0.399117	8.033225	12
GATCT	540930	0.3517081	15.515523	39
GTCTG	658195	0.2233537	7.9683146	17
ACTTC	24955	0.15756583	5.693361	23
CTTCA	21810	0.1377083	5.6488433	24

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
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCC	312467	0.4074891823996055	TruSeq Adapter, Index 1 (100CGGGCGC
246845	0.3219113289705173	No Hit	
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	236711	0.3086955481858661	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	147918	0.19290032189698383	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	119722	0.1561298309749368	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	114290	0.14904594295221868	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	92173	0.12020309475662659	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	77966	0.10167570205803378	No Hit