FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD4_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`

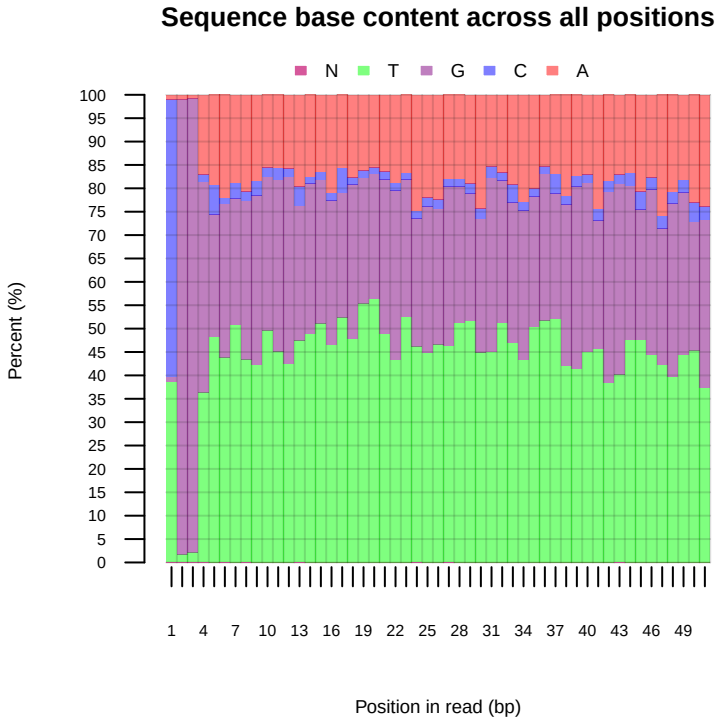
1 Sequence Duplication

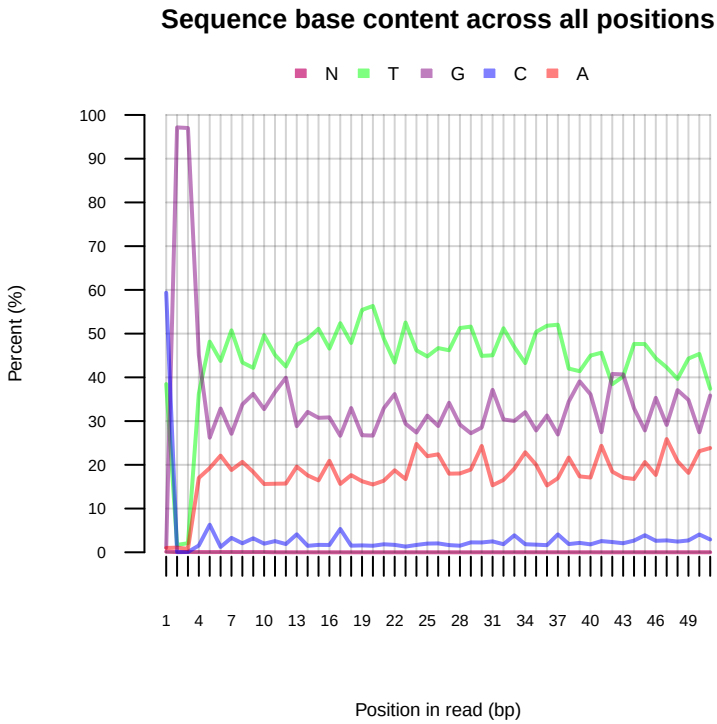
- Estimated Duplication rate 75.1598%

2 Per base sequence quality

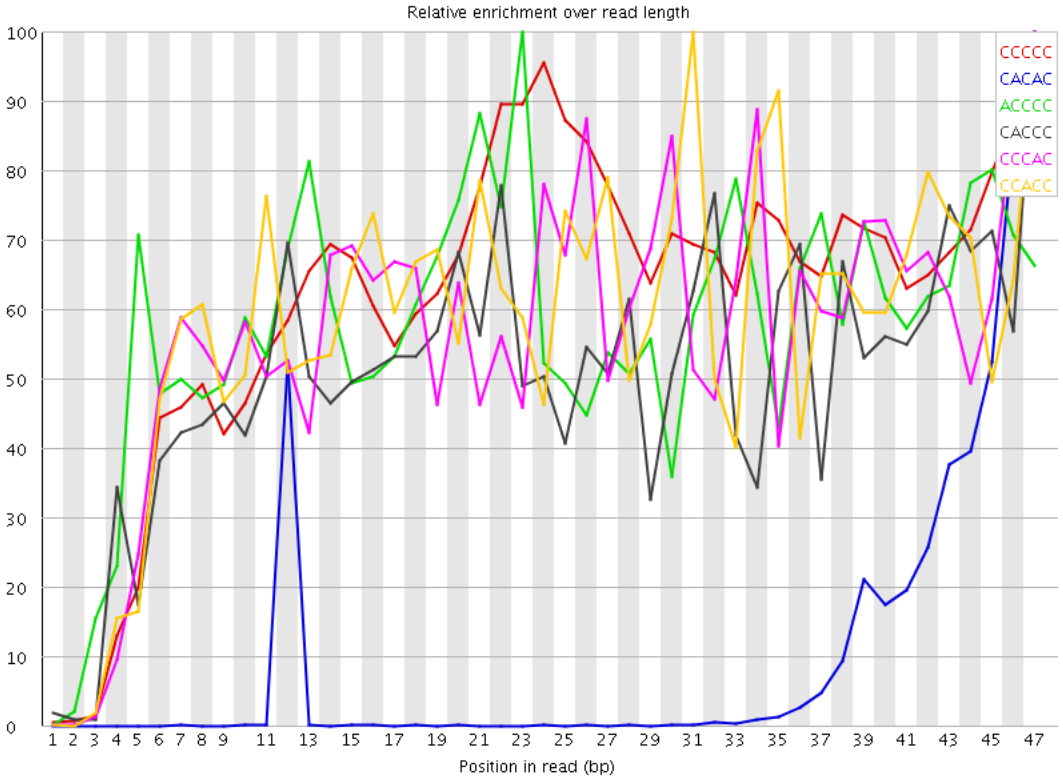
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	119975	709.0573	1138.7384	47
CACAC	972365	210.28978	2068.8586	47
ACCCC	43365	49.02629	84.738304	23
CACCC	42365	47.895744	94.03997	47
CCCAC	39545	44.7076	78.89796	47
CCACC	39350	44.48714	76.50596	31
CCCCA	38830	43.899254	67.73766	24
AGCAC	1370725	29.796728	215.62202	45
CGGGC	4559995	27.367996	953.73315	1
GCACA	1185780	25.776403	215.62303	46
ACACG	769070	16.71799	177.45166	47
GCCCC	25855	15.359053	35.175476	47
CCCCG	24890	14.7858	25.125395	46
CCCGC	24765	14.711546	26.521013	34
CGCCC	24635	14.634317	24.565872	22
CCGCC	24470	14.536301	23.030798	27
CGGAA	5830420	12.739354	186.78271	1
CGCGG	1834500	11.010229	244.06007	5
ACTCC	121500	10.596731	421.8177	23
GCGCG	1713365	10.283205	242.70512	4
CTCCA	116210	10.135361	420.54803	24
CGCGC	155570	9.289131	60.202282	13
CGGCG	1539020	9.236829	253.74101	1
AGATC	5040950	8.453499	40.99034	43
GGCGC	1269520	7.619355	243.08444	3
CGGGA	6533395	7.5009575	215.40428	1
TCCCC	14020	6.392115	15.005809	5
CTCCC	13245	6.0387707	11.355407	24
TCGCG	1309320	6.0311623	22.37752	30
CGGGT	12657560	5.8604984	227.07132	1
CCTCC	12790	5.831323	11.0342455	40
CCCTC	12775	5.8244843	10.605521	24
CCCCT	12735	5.8062463	11.141647	47
ACGTC	644940	5.65385	44.30847	15
AGACG	2511240	5.487011	61.529003	27
CACGT	592825	5.1969852	44.444237	47
CGGAG	4500255	5.16672	150.85597	1
ACGCG	435785	4.9776077	15.681508	14
CGCGT	1057350	4.8705044	19.756157	31
GATCG	5414570	4.771098	22.669918	44
AGAGC	2162005	4.7239385	27.385464	47
CGGTT	13131510	4.6663337	158.8436	1
CGGAC	407120	4.6501913	155.1404	1

CGTCG	981595	4.521552	23.605997	41
CGGTC	963355	4.437533	166.17995	1
AACCC	20170	4.3620915	7.418932	22
CGGGG	7197295	4.341875	105.32441	1
AAAAA	2866945	4.3401594	13.40788	31
GGGCG	7173330	4.327418	100.48335	2
CGCGA	376045	4.2952476	16.87405	5
TCGAG	4652215	4.099342	45.04999	32
AGGCG	3496835	4.0146985	50.482216	47
ATCGG	4454410	3.9250445	21.117323	45
GAGAC	1791835	3.9151244	59.04427	26
ACACC	18020	3.8971188	6.9111013	47
TCGGA	4394380	3.8721483	21.416403	46
ACCAC	17480	3.780335	6.707848	46
CGACG	329895	3.7681148	30.608635	24
CAACC	16925	3.6603072	6.7585497	31
ACCCA	16840	3.6419246	6.55531	33
CCCAA	16505	3.5694761	6.3010335	15
GAGCA	1629440	3.5602949	22.28561	44
CCACA	16380	3.5424423	6.2504954	46
CACCA	16135	3.489457	5.8946795	32
TTACG	5053590	3.417677	36.743187	14
CGGTA	3858630	3.400067	117.48619	1
CCAAC	15705	3.3964624	6.5553384	34
CGTTT	12357275	3.3702366	37.0076	17
GGCGT	6965425	3.2250185	49.61033	3
GGCGG	5334685	3.2182279	40.490444	11
GACGG	2735140	3.1402004	32.393276	28
TACGT	4638800	3.1371593	38.81937	15
ACGTT	4598850	3.1101415	40.499115	16
ACGGG	2620670	3.0087779	32.369766	29
CGAGG	2593025	2.9770389	53.965916	45
CGGAT	3373140	2.972273	101.23426	1
AGAGA	6823845	2.852173	21.29101	25
TTTCG	10293285	2.8073184	16.15914	30
AAGCG	1268220	2.7710361	54.07181	8
AGCGA	1230105	2.6877556	54.821552	9
ATCGC	306210	2.6843824	32.057365	29
CGAGA	1201405	2.6250465	33.694332	25
TTCGA	3878690	2.623107	35.99077	31
TTTTT	160054340	2.584566	5.7379975	16
GAAGA	6053420	2.5301573	10.155772	46
CGTTA	3699040	2.501612	33.72216	9
CACAG	114560	2.4902973	108.00888	31
GCGGG	4108725	2.4786494	41.272625	12
GGAAG	11198675	2.4594817	11.088617	2
GGAGG	21173640	2.4434457	28.031654	39
GTCGA	2760820	2.432722	42.912445	43
TCGTT	8905680	2.4288728	6.4695177	4
GCGGC	401990	2.4126475	8.404723	9
CGTTC	661605	2.3390017	28.253914	33
TTCGC	660080	2.3336105	8.354393	33
ATTCG	3448050	2.3318708	43.721325	34
TTTTA	57225880	2.2914264	12.432733	26
AGAAA	2846020	2.2638857	5.553261	22
AAGAG	5413920	2.2628646	10.16724	47
GAGGC	1969810	2.2615287	42.047337	46
TTTAG	42898565	2.2381008	15.45768	27
TTCGT	8176780	2.230078	5.8465023	35
GAGAT	13223540	2.2289546	8.713094	26
AACTC	133005	2.219028	82.97231	22
GGGAG	19190425	2.214582	24.223354	38
AGTAG	13116900	2.2109792	22.502506	35
CGAGT	2476385	2.182089	43.95712	33
CGTAG	2444605	2.154086	22.816343	5
GCGGA	1863160	2.1390846	22.588236	7
CGGTG	4611615	2.1351953	42.66025	1
GAGGT	23719080	2.1007848	21.697008	40
GCGTT	5838260	2.0746486	26.939493	16
AGGAG	9292400	2.0408206	9.828453	38
CACGC	17620	2.0022807	9.719412	47
ATTTT	49565310	1.9846834	7.593276	25
ACGGA	891320	1.9475168	8.9383545	30
TAGTT	36928480	1.9266298	10.038915	29
TTTAC	3706135	1.9236618	27.351797	13
GGTCG	4141020	1.9173081	24.332434	42
AAACG	460495	1.9148834	13.600621	7
TAAAA	3096110	1.8902073	5.3307915	30
TACGC	215480	1.8890002	11.083916	13
GACGC	164400	1.877804	11.420328	5
TCACA	112395	1.8751751	82.35908	30
GCGGT	3999610	1.8518345	24.18191	6
AAAAT	3032220	1.851202	5.192244	32
TAGAG	10980985	1.8509507	9.494154	24
CGCAC	16195	1.8403484	7.1293635	46
AATTT	18474560	1.8343439	16.671434	24
TCGTC	514680	1.8195713	9.881095	40
ACGGC	159140	1.8177233	9.37483	12
GTCGC	390875	1.8005002	8.765919	3
TTAGT	34460145	1.7978522	14.854625	28
ATCGT	2642650	1.7871895	14.414492	39
AGGTA	10524530	1.7740107	28.472	47
AGCGC	155035	1.7708354	7.3753242	35
GCGTA	1988235	1.7519515	22.344078	4
CGAGC	152945	1.7469629	6.478867	32
GAAAA	2195255	1.7462302	5.333902	3
GGAAA	4176155	1.7455139	11.78889	2
TGAGA	10292105	1.7348334	5.8417015	41
GGAGA	7705285	1.6922544	10.194023	2
AGTTA	13078080	1.6918975	21.811686	30
AGTTT	32257635	1.6829429	8.945123	26
GAGCG	1465065	1.6820337	10.222172	28
TAGTA	12965690	1.6773577	13.017354	29
TACGG	1901620	1.67563	11.724805	5
AGCGG	1454720	1.6701566	6.1168733	6
TATCG	2462765	1.6655356	14.721594	38
GGACG	1447745	1.6621487	16.310064	2

TAGCG	1875665	1.6527596	5.34968	10
TGGCG	3534160	1.6363297	35.35561	10
GTAGA	9673270	1.6305227	9.170084	23
TATTT	40076690	1.6047422	5.1602435	32
AACGC	73620	1.6003466	7.8252172	11
GCACC	14045	1.5960292	7.556576	47
AGTCG	1806380	1.5917083	13.866703	22
GCGTG	3428845	1.5875683	34.22806	4
TTGAG	23342455	1.586744	14.000395	44
AGGTC	1784230	1.5721906	40.86188	41
TCGTA	2323880	1.5716093	7.474139	45
TGGGA	17732815	1.5705849	12.83871	37
CGTGG	3387725	1.5685296	34.01034	5
CGATT	2310125	1.562307	19.081078	11
GCGAC	133830	1.5286282	20.238426	23
GTCGT	4283815	1.5222708	10.634758	3
TAGGA	9029775	1.5220555	7.7353997	37
ACGCC	13310	1.5125061	5.9007974	23
TTCGG	4251495	1.5107856	23.598343	35
GGGAA	6853180	1.505113	13.416587	2
AGCGT	1700165	1.498116	8.257352	29
GGTTT	53539205	1.4677057	9.242313	2
TCGAA	874540	1.4665736	5.0180674	32
AACGG	669335	1.462484	7.6794586	29
TTATT	36467485	1.4602234	7.659546	32
GTAGT	21477125	1.459945	9.739063	36
GCGTC	315795	1.4546566	10.127463	40
AGGTT	21369925	1.4526578	14.137814	41
CGTAC	165400	1.449975	9.15386	13
TAATT	14553290	1.4449999	16.324327	23
AAGTA	4468190	1.4333589	11.834744	34
CGAAA	343010	1.4263433	5.0427165	32
GTACG	1617880	1.42561	11.332899	4
TATAG	10941735	1.4155208	17.539223	47
TTTAA	14175035	1.4074429	8.552272	5
CGTCT	395260	1.3973804	16.512987	16
ACGGT	1575725	1.3884645	10.9840145	6
TTATA	13966980	1.3867853	13.723939	46
TTAAG	10576705	1.3682975	10.227983	6
GGAAT	8056865	1.3580619	9.473847	2
TCGAC	154020	1.3502126	7.0383263	23
GCGAT	1527840	1.3462701	22.919146	10
AAAAC	169720	1.3431368	23.428835	6
GTTTA	25728235	1.3422915	8.500901	4
GAACG	602655	1.3167895	7.534113	28
AGATA	4101755	1.3158094	5.09578	26
GGCGA	1142770	1.3120085	8.501395	2
TAAGC	779530	1.307245	39.57433	7
TCGGG	2815595	1.3036313	29.442932	36
CGCCA	11375	1.2926189	5.9275103	24
GGTAG	14579385	1.2912874	7.383818	2
GAGTA	7624005	1.2850993	16.09582	34
GGAGT	14478960	1.2823927	10.047945	2
GACGT	1455000	1.2820866	5.8746	3
TGGAA	7602085	1.2814045	8.739371	1
GGGTT	35635365	1.2728338	14.1676235	2
ATTAT	12764570	1.2673974	13.610344	45
CGTAT	1872075	1.2660598	5.3341055	13
TTGTA	24247540	1.2650408	14.20472	20
GGTTA	18590515	1.2637225	16.28352	2
TATTC	2423400	1.2578609	32.268406	33
ACGAC	56995	1.2389534	5.3578315	18
GTTAA	9400005	1.2160687	19.898314	3
TTTGT	57558260	1.21102	6.838913	19
GGGAT	13516730	1.1971688	11.685376	42
GGGGA	10278610	1.1861553	9.744334	2
GATTA	9138320	1.1822149	17.070208	44
GGTGG	25008805	1.1638765	10.759749	8
TGAGG	13056285	1.1563876	16.203102	45
GTAAT	8925060	1.1546259	19.970419	22
TCGTG	3237940	1.1506146	6.4905424	40
CGTAA	682430	1.1444117	9.220176	21
TTTTC	5302030	1.10983	11.548943	29
GGATT	16315520	1.1090758	9.238299	43
CGTGA	1241505	1.0939635	8.529884	26
GGGGT	23456475	1.0916331	8.053171	2
TAGGC	1238620	1.0914215	9.457497	13
GTGGC	2328880	1.0782803	33.365482	9
CGAAC	48855	1.0620066	7.339645	20
TCGAT	1568875	1.0610095	6.7878265	11
GTTAT	20216870	1.054753	9.127882	31
GTATT	20140540	1.0507708	5.3855968	31
TGGAG	11861365	1.0505539	9.658542	1
GGGTA	11817535	1.0466721	14.418237	2
AGTAA	3261730	1.0463363	7.190384	9
GTTGA	15259770	1.0373094	13.036026	43
AGTAT	7994665	1.0342615	12.177973	30
TGTAG	15116655	1.0275809	8.443473	21
TGTAA	7930545	1.0259663	19.287348	21
TCCAG	116685	1.0229162	43.49888	25
AGTTG	14921930	1.0143441	9.736855	38
TTAAT	10197720	1.0125344	13.595622	4
ATTTC	1937305	1.0055542	5.9889483	22
CAGTC	113905	0.9985453	43.680626	27
CCAGT	113585	0.9957401	43.328106	26
GGTTG	27540915	0.983714	6.985206	42
GTCAC	111740	0.979566	43.679012	29
CGTGT	2754090	0.97867674	6.174294	41
CCGTA	111630	0.9786016	42.9191	38
TAAGT	7552645	0.9770779	6.540856	7
TGGGG	20235335	0.94172555	8.493867	1
TTATC	1806680	0.93775374	10.942036	37
GTTTG	33757935	0.9254287	6.676559	18
TTGGG	25839160	0.92293024	5.8191814	36
TAGAC	545450	0.9147009	11.594991	25
AAGAC	219040	0.9108373	9.020415	32
GTGGT	25255295	0.9020756	7.3472476	9

GGATA	5326830	0.89788836	7.173033	2
GGGGG	14804920	0.8977252	5.6794934	2
GGAGC	780140	0.89567477	9.352155	27
AAGGC	408965	0.8935806	12.864111	46
TGCGG	1929045	0.8931553	5.350519	5
TGGTT	32413025	0.88855976	7.4194875	1
AGTGA	5260545	0.8867154	5.32072	18
GGGTG	18740135	0.87214094	8.153957	2
GGTAT	12006395	0.81615555	5.944812	2
ATCTC	121205	0.8154965	35.09693	42
GAAGC	364660	0.796775	9.904414	4
GGAAC	364195	0.79575896	6.740379	27
TGGGT	21488770	0.7675418	8.901174	1
TGGTG	21426965	0.76533425	6.0294266	7
GGTAC	863275	0.7606827	11.359301	3
GGTAA	4479535	0.75506866	5.9961495	2
GTTGG	20738730	0.7407517	5.281996	39
GAGTC	756330	0.6664471	12.689668	21
TGGTA	9748900	0.6626984	5.1829143	1
CACAT	39460	0.6583424	6.221444	47
GATTC	767360	0.51895547	5.74938	29
TGGGC	1105135	0.51168174	5.736834	13
TGAAC	289920	0.4861859	8.42387	20
TCTCG	131000	0.4631302	18.505634	43
CTCGT	129545	0.4579863	18.52066	44
TTCCG	120730	0.42682222	17.081816	36
TCCGT	117620	0.4158273	17.113386	37
GTTCC	113370	0.40080208	17.05524	35
CTGAA	234640	0.39348322	8.155914	19
GAACT	163040	0.27341247	8.677603	21
AGTCA	123665	0.20738196	8.553234	28
ACAGT	118685	0.19903065	8.478504	32

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	245954	0.32515010565983726	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	172583	0.22815396653476544	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	152953	0.2022031929181436	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	89919	0.11887252230428011	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	83936	0.11096302263294805	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	82001	0.10840496114806963	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	76727	0.10143275635672659	No Hit