FASTQ QC Report

Report Date	`12-21-16`
Run ID	`161219_D00796_0155_ACAC53ANXX`
Project ID	`EC-EL-4039`
Sample	`Sample_YD16_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.

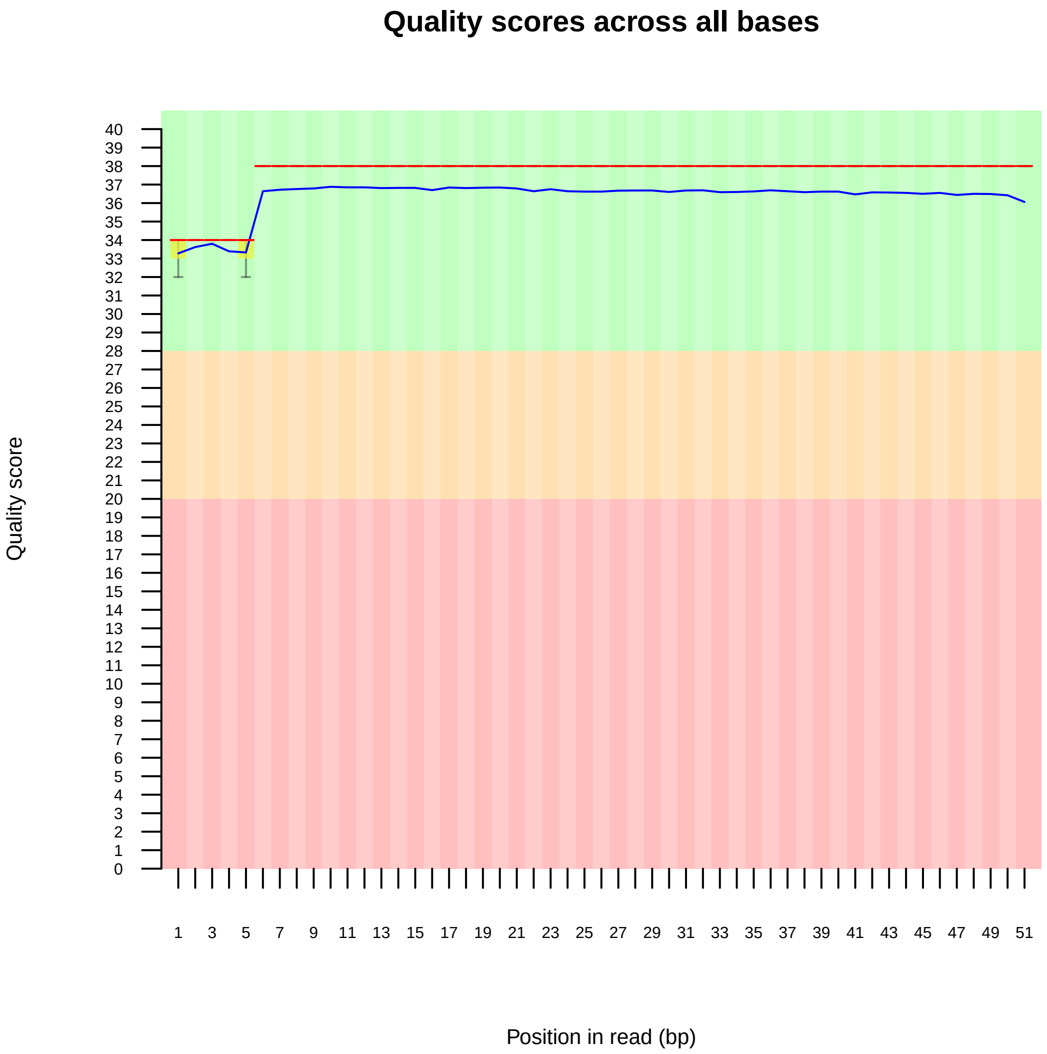
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

```
FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
FASTX-Toolkit:  http://hannonlab.cshl.edu/fastx_toolkit
```

1 Sequence Duplication

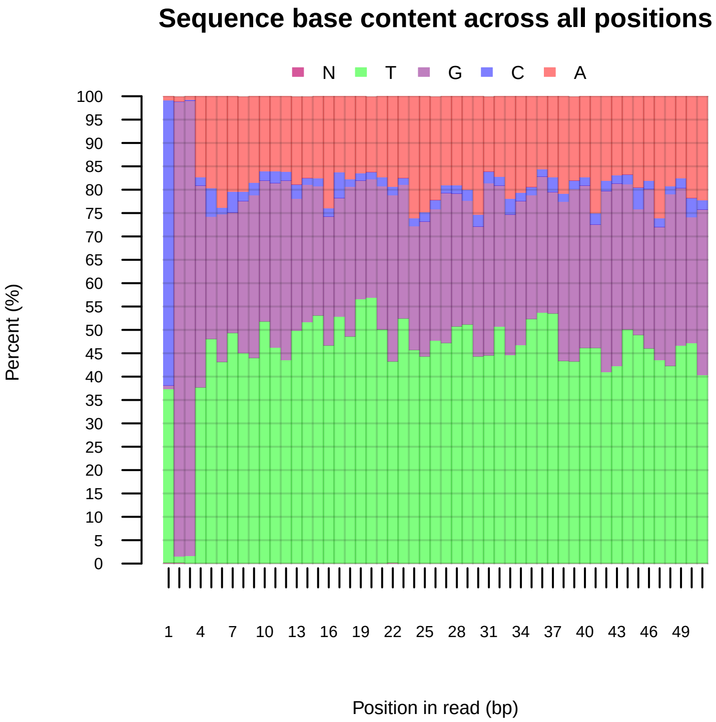
- Estimated Duplication rate 83.4168%

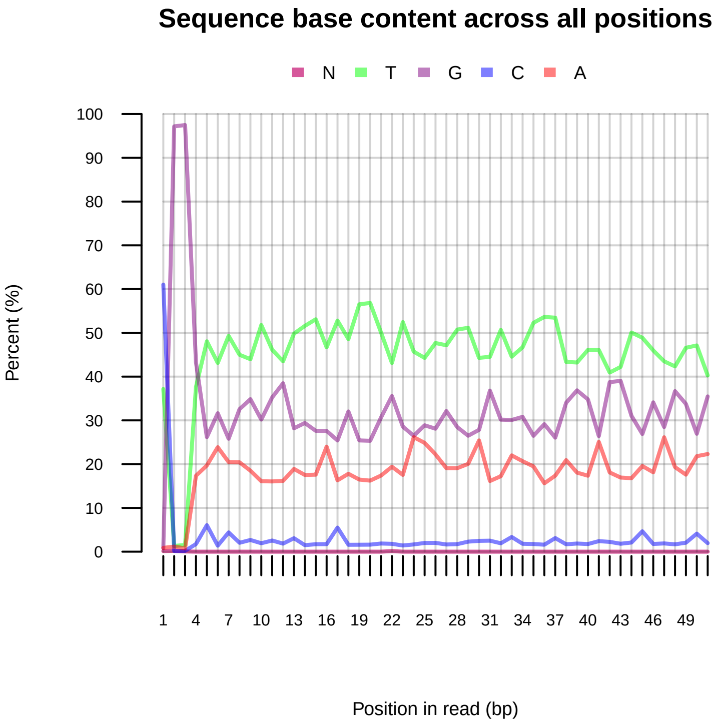
2 Per base sequence quality



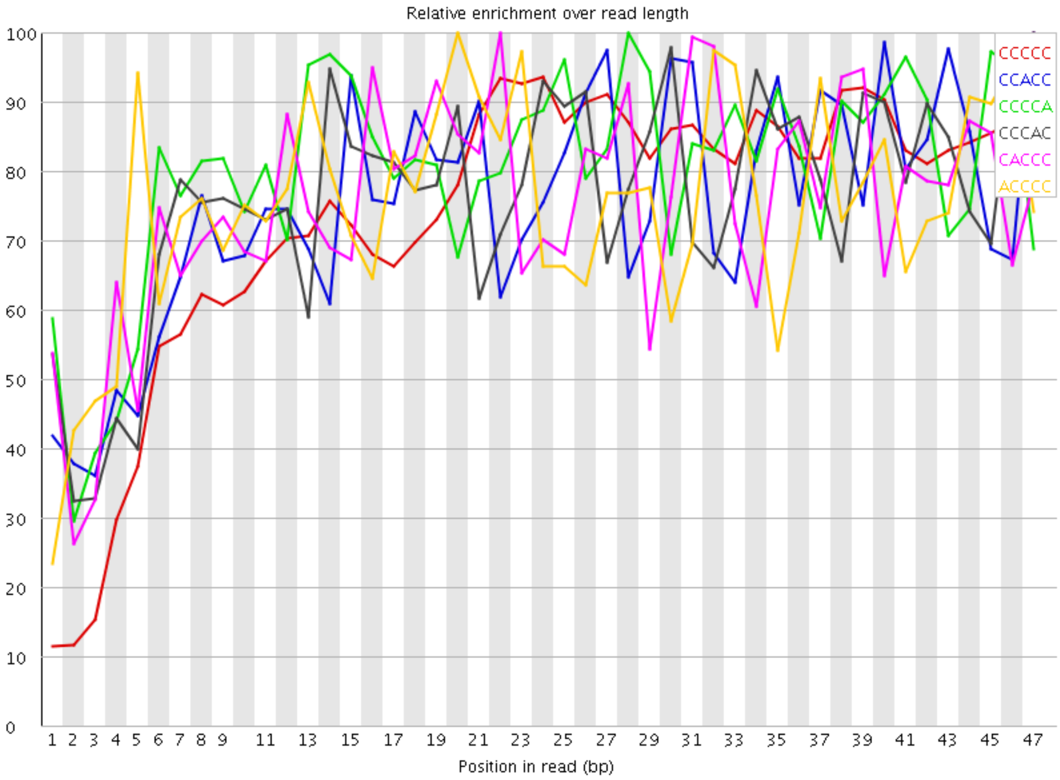
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	253450	1500.8566	2027.0826	47
CCACC	87575	94.96352	125.85446	47
CCCCA	87275	94.6382	118.20906	28
CCCAC	85020	92.192955	120.75914	47
CACCC	80625	87.42715	116.071434	22
ACCCC	79150	85.827705	114.03062	20
CACAC	300300	59.62957	544.9322	12
GCCCC	45750	27.575132	42.90671	37
CGGGC	4376425	27.32784	1060.3466	1
CCTCC	51580	22.581404	28.59496	43
CTCCC	47750	20.904654	37.439194	1
CCCCG	33995	20.48998	33.136543	46
CCCGC	32545	19.616014	35.82641	38
CCGCC	32545	19.616014	33.84391	35
CGCCC	31325	18.880678	33.13648	44
TCCCC	41655	18.2363	30.35889	3
CCCTC	39150	17.139627	24.993782	36
CCCCT	37920	16.601141	22.999836	1
CGCGG	2381965	14.873775	397.452	5
GCGCG	2269675	14.172599	395.1869	4
CGGCG	1780525	11.118185	326.5193	1
GGCGC	1773875	11.07666	396.54556	3
CCACA	55515	11.023429	14.368559	30
CGCGC	165775	10.170099	44.479336	27
CACCA	50570	10.041516	14.088654	28
ACCCA	45870	9.1082535	12.082975	39
CCCAA	44180	8.772676	11.802768	6
AACCC	43875	8.712112	12.78242	4
ACACC	42955	8.529431	10.963248	47
ACCAC	42810	8.500639	10.776708	39
CTCCA	96375	7.72616	206.63399	24
CCAAC	37725	7.490928	10.916372	34
CAACC	37070	7.3608665	11.320666	1
CGGAA	3564860	7.3334465	211.56021	1
CGGGA	6357540	7.2695026	235.11217	1
ACTCC	83605	6.702419	206.12584	23
AGACG	3219285	6.6225476	77.883255	27
TCGCG	1347630	6.1122284	24.140657	30
CCCAG	55165	6.0886345	10.087002	27
CGGGT	13023035	6.0120244	235.47139	1
CGGAG	5090090	5.820242	175.69818	1
AGCAC	271635	5.4899926	56.93751	10
ACGCG	476055	5.3480206	17.762625	4

CCAGC	47965	5.2939606	8.894193	28
CGGTT	15669850	5.254319	187.32428	1
CGGGG	8198915	5.2110095	130.34894	1
CAGCC	46460	5.127852	7.1568427	16
AGGCG	4481030	5.123815	93.44722	47
CGCGT	1125990	5.1069717	22.234339	31
CGTCG	1105320	5.013222	18.53414	41
GAGAC	2429175	4.997174	74.70987	26
GCACA	239790	4.846376	56.927975	11
TCGAG	5779955	4.8004622	80.36032	44
CGGTC	1031475	4.678295	173.57327	1
CGGAC	413695	4.647465	147.9204	1
ACACA	125980	4.5807614	96.4806	32
CGAGG	3854330	4.4072175	104.80362	45
GGGCG	6860155	4.3601294	113.05347	2
CGCGA	382315	4.294942	23.239214	5
AGCCC	37875	4.180314	5.938105	12
TTACG	6599230	3.9810252	66.38317	14
GCCAC	34680	3.8276777	5.626938	29
CGACG	340360	3.8236177	19.018944	24
GCCCA	34530	3.811122	5.2120495	25
CGGTA	4581710	3.8052769	134.62717	1
AAAAA	3114610	3.7975132	8.837901	31
GACGG	3294975	3.7676253	42.72085	28
TACGT	5968950	3.6008055	66.87838	15
ACGGG	3141320	3.5919292	42.883644	29
ACGTT	5934400	3.579963	68.12545	16
TCCCA	43120	3.4568303	6.5544276	26
GTCGA	4038275	3.3539343	79.765274	43
CGTTT	13436105	3.2724133	38.335888	17
AGAGA	8602760	3.2406583	26.729198	25
ACACG	158350	3.2003982	54.491985	13
CGGAT	3834465	3.1846626	110.48117	1
TCCAC	39090	3.133755	3.8422506	5
GAGGC	2725115	3.1160212	75.781006	46
GGCGG	4898170	3.1131449	23.332672	11
CTCAC	38385	3.0772367	6.366081	13
CCTCA	36765	2.9473648	5.741471	1
GGCGT	6089495	2.8111875	45.08847	3
CGGTG	6057235	2.796295	51.9061	1
AAGCG	1358110	2.793834	49.351883	8
CGAGA	1341150	2.7589452	28.057346	25
TTTCG	11001260	2.679397	12.224392	30
GGAGG	22741315	2.6467369	32.296505	39
AGCGA	1286495	2.6465118	50.00822	9
GCGGC	417680	2.6081316	7.75022	3
TTCGA	4216895	2.5438676	27.378868	31
GGTCG	5480330	2.5299695	46.197147	42
ACGTC	308005	2.5132563	22.697962	15
TTTTT	188747745	2.4685519	5.0541477	16
ATCGC	301290	2.458463	33.86482	29
GGGAG	21037795	2.4484737	30.01426	38
AGGTC	2943360	2.4445672	76.05235	41
AGATC	1586795	2.370989	14.456104	27
CGTTC	718640	2.3674638	24.994041	33
TCGTT	9405615	2.2907724	5.497164	4
AGTAG	15019750	2.2842894	17.427912	35
ACGGA	1101450	2.2658467	9.365013	30
GACGC	201470	2.2633219	17.759165	3
CGTTA	3735120	2.253234	20.946274	9
TTTTA	69467885	2.2503579	11.55649	26
GCGGT	4849340	2.2386756	38.887314	6
CGTAG	2694455	2.2378426	25.95037	5
TTTAG	49775840	2.2199492	15.090243	27
GCGGG	3466660	2.2033157	23.782843	12
CACAG	108375	2.1903577	53.376125	33
TTTAC	4993025	2.1878045	47.259724	13
TTCGC	664095	2.1877725	5.038591	13
GCGGA	1912565	2.1869144	17.912725	7
TTCGT	8905455	2.1689565	5.2934012	35
GAGGT	25311210	2.1396916	23.836514	40
AGCGC	187650	2.1080675	16.14472	35
ATTTT	64445180	2.0876513	9.216037	25
TAGAG	13498710	2.0529609	11.75795	24
AGGAG	9778735	2.0475192	7.0205507	38
TCGTC	620835	2.0452583	8.336731	40
GGAAG	9736355	2.0386455	11.906444	2
CGAGT	2437800	2.0246816	33.700768	33
TAGTA	18326805	2.024502	22.51249	29
TACGG	2424480	2.0136187	19.809175	5
GTCGC	441985	2.0046403	11.538005	3
ACGGC	177895	1.9984795	12.373133	12
AATTT	24676375	1.9799576	20.37303	24
ATCGT	3205495	1.9337344	16.858799	39
CGAGC	170340	1.9136063	6.2815948	7
TACGC	233650	1.9065349	7.085444	13
ACGCC	17205	1.8989387	5.2120485	16
AAACG	510835	1.890586	5.6031	7
ATTCG	3102135	1.8713816	23.902262	34
GATCG	2193235	1.8215612	9.075108	28
GAAAA	2685490	1.8199903	5.279813	3
GTAGA	11930285	1.8144256	11.443681	23
GGAAA	4802675	1.8091669	12.926979	2
GCGTT	5394875	1.8089769	15.019517	16
TATCG	2993535	1.805868	17.255472	38
GCGTA	2169450	1.8018072	25.68949	4
TTAGT	40054565	1.7863905	14.420682	28
GGACG	1545605	1.7673155	15.671567	2
GAGCG	1535110	1.7553151	10.522655	28
GGAGA	8358430	1.7501289	11.29916	2
TGGGA	20017265	1.6921664	18.848198	37
TAGTT	37360000	1.666216	6.534702	25
GAGAT	10911780	1.6595258	9.118968	26
AGGTA	10862400	1.6520156	20.678457	47
ACGGT	1983230	1.6471447	19.12315	6
GTACG	1981755	1.6459194	19.415802	4
CGAAA	444505	1.6451005	7.304709	15

TATTT	50009225	1.6200098	7.7765503	32
TCGAA	1081000	1.6152302	5.0786843	44
CGTAC	197735	1.613476	6.7231207	5
TAATT	20095215	1.6123792	20.13368	23
ACGAG	780495	1.6055944	6.088684	32
AGCGT	1916425	1.5916607	8.223578	29
GGGAA	7514400	1.5734018	13.834996	2
CGATT	2604400	1.5711201	16.996414	11
AACGG	761765	1.5670638	8.086185	29
AGTTT	34920015	1.5573953	6.6913066	26
GTCGT	4628870	1.5521247	10.102028	3
AGTCG	1857710	1.5428956	11.4717655	22
GGTTT	61960870	1.5360023	9.485687	2
TAGGA	10097195	1.5356388	5.306367	37
GCGTC	337565	1.5310392	8.188964	4
AGTTA	13778190	1.5220314	11.508469	30
CGTGG	3286925	1.517394	26.534117	5
AACGC	73985	1.4953046	7.687579	11
GGCGA	1290860	1.4760284	11.0498705	2
AGGTT	23971905	1.4719191	12.376669	41
GAACG	707470	1.4553709	7.826651	28
CGAAG	700265	1.4405494	5.7132683	45
GGAAT	9451705	1.4374692	10.465701	2
GCGAC	127850	1.4362718	11.760834	23
GGTTA	23390080	1.436194	19.70537	2
GCGTG	3098885	1.4305862	26.550495	4
AGATA	5193400	1.4209876	6.3968444	26
TTAAG	12833840	1.4177122	8.83338	6
GTTTA	31770190	1.4169166	10.407389	12
GTTAA	12791495	1.4130346	24.10705	3
TTTAA	17310020	1.3889036	7.091893	5
GCGAT	1669350	1.3864559	21.159685	10
AAGTA	5058220	1.3840004	8.882851	34
AAGGC	671050	1.3804495	27.156033	46
AGAGC	666235	1.3705443	6.394146	8
TCACA	93110	1.3668643	39.038513	30
TATAG	12368590	1.3663176	13.048606	47
GGTAG	15978265	1.350728	8.203451	2
TGGAA	8835015	1.3436793	9.094594	1
TTATA	16650980	1.3360242	9.823162	46
GTAGT	21739435	1.3348413	7.5506377	36
GGAGT	15770195	1.3331387	10.726024	2
TTTTG	73655945	1.3262495	5.0231028	34
AAAAC	197115	1.3124583	8.017088	6
GTAAT	11870165	1.3112583	25.974298	22
GACGT	1572460	1.3059852	5.901695	3
GGGTT	38243035	1.3052198	13.6118965	2
AACTC	88315	1.2964734	38.97338	22
AGTAT	11710515	1.2936223	21.603228	30
TCGTG	3855015	1.2926403	8.948928	40
CGAAC	63730	1.2880418	5.622051	9
TTGTA	28767800	1.283013	13.850926	20
TGGCG	2775885	1.2814747	19.812532	10
TTCGG	3800500	1.2743607	13.495861	35
GGGGA	10791355	1.2559468	11.269686	2
TAAGC	831870	1.2429801	34.623104	7
TCGGA	1485075	1.2334088	5.707155	46
GAGTA	8095855	1.2312636	12.494254	34
ATTAT	15001145	1.2036464	9.649124	45
TTGAG	19557185	1.2008471	9.783556	44
GGGAT	14102960	1.1921986	9.956998	2
CACGT	145345	1.1859846	22.134344	14
TGTAA	10665660	1.1782005	25.674034	21
GGTGG	25050910	1.1770955	11.467094	8
GTATT	26349440	1.1751568	9.254111	31
CGTAA	784460	1.1721401	5.8647065	21
TTTGT	64736240	1.1656412	6.5854115	19
TCGAT	1915015	1.1552445	6.1181536	11
GGGGT	24457635	1.1492187	9.083427	2
GGGTA	13235990	1.118909	15.770097	2
TTAAT	13824995	1.1092757	16.03935	4
GATTA	10008780	1.1056371	12.688575	44
CGTGT	3255075	1.091472	8.540711	41
ATCGG	1310085	1.0880731	5.2757206	45
GGATT	17682905	1.0857629	7.186717	43
AGTAA	3960540	1.0836596	6.3505087	9
AAGGT	7119675	1.0828007	5.342384	46
TAGGC	1299790	1.0795227	5.220396	13
TGAGG	12688035	1.0725874	12.317885	45
TGGAG	12625280	1.0672823	9.688553	1
TCGGG	2303730	1.0635065	17.506224	36
TAAGG	6711590	1.0207369	7.447834	45
TTTTC	5766990	1.0202038	8.170709	29
TGCGG	2139445	0.98766506	8.950019	5
CGTGA	1182615	0.9822048	7.640934	26
TTATC	2232470	0.9782062	12.26598	37
TTGGG	28440440	0.970661	8.147418	36
ATTAC	894210	0.9704906	5.477503	29
GGAGC	844785	0.9659658	9.564414	27
GGATA	6322630	0.9615815	8.0525875	2
GGTAC	1155180	0.9594189	19.54964	3
TGGGG	20408310	0.95894843	9.021639	1
TAAGT	8601355	0.95016354	5.6778283	7
CGGCC	15445	0.9475324	6.214768	1
TAGAC	633850	0.9470987	8.400197	25
TGGTT	38034015	0.9428585	7.371326	1
CGATC	115200	0.94000787	7.961684	44
AGGTG	11038535	0.9331463	5.494733	47
TTTGG	37567355	0.93129003	6.3659573	35
AGTTG	15078070	0.9258212	7.5265775	38
GGTTG	27095785	0.92476845	5.4417124	42
GTGGT	26577320	0.9070735	9.288223	9
GTTTG	35576635	0.88194036	6.768072	18
GGAAC	425630	0.8755841	6.92867	27
TATTC	1977585	0.86652267	15.956109	33
GGGTG	18374490	0.86338305	8.685114	2
GTTGA	13892360	0.85301644	9.509172	43

GTGCG	1794220	0.8282935	8.375188	4
GGGGG	12801085	0.82811743	6.471169	2
GGTAT	13348015	0.81959265	6.4475846	2
TCCAG	99480	0.8117359	21.839125	25
GGTAA	5235695	0.7962743	6.7539244	2
AGTGG	9316650	0.78758615	6.01681	8
TGGGT	22560120	0.76996803	9.070514	1
GTGGC	1609855	0.74318224	18.642586	9
TGGTG	20618665	0.7037069	5.886663	1
CCAGT	86050	0.70215	21.670424	26
CAGTC	85040	0.6939086	21.908138	27
ATGCC	83890	0.6845248	23.21604	47
GAGCA	329840	0.67852986	5.96207	9
GTCAC	81975	0.66889876	21.975227	29
GAGTC	773955	0.6427977	10.212496	21
TGGTA	10368020	0.63661546	5.4621935	1
AAGTC	407735	0.6092376	5.396099	41
ATCTC	92555	0.54855716	16.906134	40
TGGAT	8467905	0.5199449	5.092254	1
GATTC	838695	0.5059479	5.152689	29
GGTGC	997170	0.46033898	8.463015	3
CGTCT	118470	0.39028364	8.862038	16
TCTCG	90200	0.2971519	9.380881	41
CTCGT	85600	0.2819978	9.377018	42
CGGCT	43375	0.196729	5.377818	1

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	273825	0.3235597511497316	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	173101	0.20454128178131903	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	129227	0.15269846055629094	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	101588	0.12003939742462862	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATTTTGTTAGTTA	87979	0.10395859890953067	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	85041	0.10048697086651813	No Hit