FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD17_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`

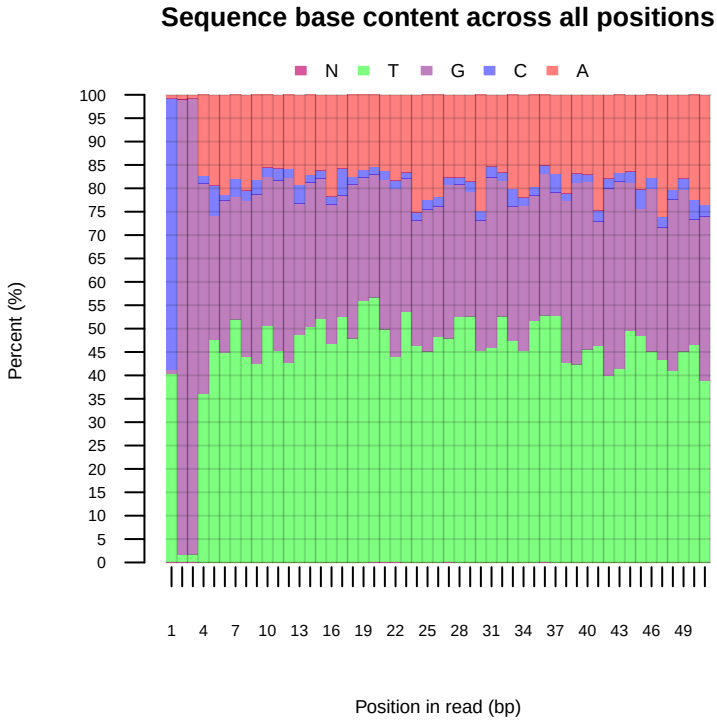
1 Sequence Duplication

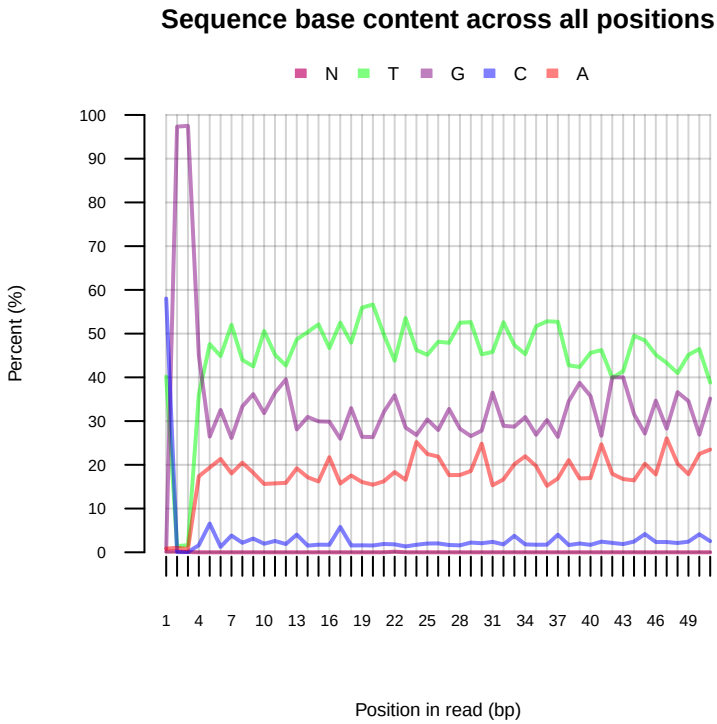
- Estimated Duplication rate 85.7028%

2 Per base sequence quality

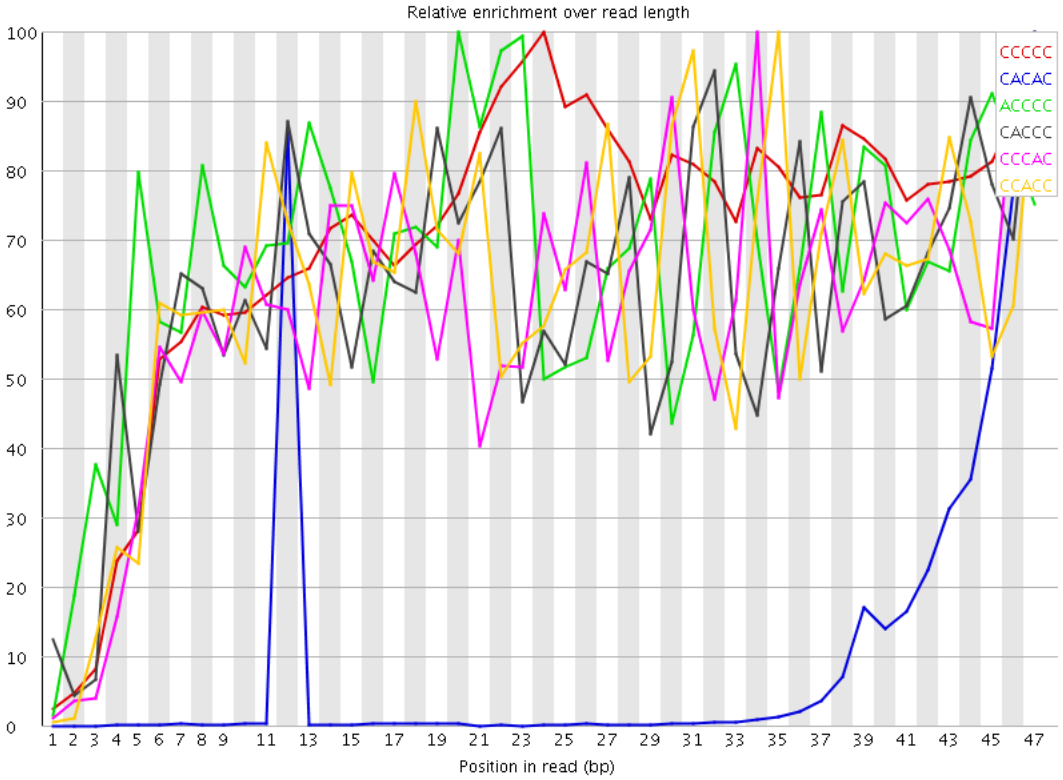
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	188545	1317.045	1870.8911	24
CACAC	472690	119.39705	1172.0034	47
ACCCC	48025	63.792374	94.05994	20
CACCC	45985	61.082607	97.36941	47
CCCAC	43960	58.39277	97.99166	34
CCACC	43735	58.0939	92.99846	35
CCCCA	42940	57.037888	85.508644	29
CGGGC	4313980	31.050133	1099.0056	1
AGCAC	687720	17.545351	134.99095	47
GCACA	655280	16.71773	120.515434	46
CCCGC	21060	14.858578	29.007812	34
GCCCC	20880	14.73158	26.521933	45
CGCCC	20705	14.60811	24.035465	44
CCGCC	20430	14.41409	26.024155	36
CCCCG	20235	14.276511	25.361616	46
CGCGG	1879275	13.5261965	318.39487	5
GCGCG	1778230	12.798919	316.1915	4
CGCGC	171205	12.200244	64.07205	13
CGCAC	90775	12.178698	508.92953	37
CCGCA	88410	11.861403	508.29916	36
CGGAA	4201945	10.827643	190.43607	1
CGGCG	1480265	10.654298	289.86502	1
GGCGC	1324035	9.529822	317.4533	3
ACACG	368190	9.393391	101.8601	13
ACTCC	91310	9.070865	386.4067	23
CTCCA	88855	8.826982	384.30545	24
CGGGA	5620285	7.692357	220.50346	1
TCGCG	1316120	7.014174	23.147577	30
AGATC	3581705	6.833909	36.441216	43
CGGGT	11078320	5.9633207	231.41544	1
CGGAG	4013360	5.4929953	161.9614	1
ACGCG	405320	5.492447	16.061535	14
CGCGT	1027980	5.478551	20.920092	31
CGTCG	1007985	5.371989	25.697588	41
TCCCC	10160	5.307723	20.262178	3
AGACG	2032390	5.2370977	51.805206	27
CGCGA	381925	5.175424	21.1522	5
CGGAC	371650	5.0361886	163.33742	1
CGGTC	923125	4.919733	178.4152	1
ACGTC	489435	4.9108725	41.08367	32
AACCC	19195	4.8484764	8.070755	4
GGGCG	6514075	4.73556	115.86762	2
TCGAG	4596080	4.6578383	60.878086	44

CTCCC	8810	4.6024647	9.941632	24
CGTCC	87165	4.599281	202.64874	33
AGGCG	3353845	4.590332	70.119965	47
TCCGC	86685	4.5739536	200.65288	35
CGACG	335210	4.542394	32.63564	24
CGGTT	11235240	4.4780855	149.5462	1
CCCTC	8425	4.4013352	7.1188846	39
CCCCT	8370	4.372603	7.118713	23
CCTCC	8310	4.341258	8.591535	28
CGGGG	5957245	4.3307595	111.650955	1
ACACC	16575	4.1866894	6.1205606	20
CAACC	16275	4.110912	8.248786	31
ACCAC	16210	4.0944934	6.8840103	45
GATCG	3914205	3.9668007	20.187216	44
CCCAA	15425	3.89621	7.0619106	29
ACCCA	15425	3.8962097	6.5871596	5
CACGT	382325	3.836156	40.183163	14
CACCA	15185	3.835588	6.765367	39
CCAAC	15110	3.816644	8.248786	30
AAAAA	2181050	3.7881753	10.837628	31
CCACA	14980	3.7838073	6.2905655	46
TTACG	5012915	3.7616894	49.974865	14
CGAGG	2698775	3.6937525	77.48696	45
GAGAC	1398425	3.6034853	49.470173	26
CGGTA	3524475	3.5718338	124.25296	1
TACGT	4588470	3.4431858	51.556087	15
ACGTT	4587380	3.4423683	53.37286	16
CGTTT	11632570	3.4330602	40.362617	17
GGCGT	6301400	3.3919647	54.171577	3
GGCGG	4620405	3.358912	41.34675	11
AGAGC	1251045	3.2237144	22.036957	47
CGGAT	3105445	3.1471734	106.80917	1
ATCGG	3046115	3.0870461	18.661013	45
TCGGA	3003385	3.043742	19.01532	46
AAGCG	1174880	3.027451	56.983894	8
GACGG	2185655	2.9914567	27.440786	28
GTCGA	2936585	2.9760444	60.006683	43
CGAGA	1115355	2.8740659	34.80681	25
GCGGC	395955	2.8499103	10.196744	9
AGCGA	1104600	2.846352	57.701897	9
ACGGG	2068990	2.83178	27.674559	29
ATCGC	280765	2.8171275	33.21959	29
TTTCG	9482795	2.7986085	15.4965	30
TTCGA	3572090	2.680495	35.602592	31
TTCGC	678015	2.6755686	9.184955	33
AGAGA	5456720	2.673813	18.2407	25
CGTTC	677200	2.6723523	26.87065	33
GAGGC	1938475	2.653147	57.70727	46
GGAGG	18689975	2.5837057	32.89855	39
GCGGG	3520480	2.5592957	42.327248	12
CGTTA	3357535	2.5194924	31.388739	9
TTTTT	150898675	2.4661298	5.152774	16
CGGTG	4519600	2.432844	46.069733	1
TCGTT	8162980	2.409098	6.189558	4
CGTAG	2354125	2.3857577	28.073366	5
GGGAG	17153430	2.3712935	29.251003	38
GGAAG	8994545	2.3409715	12.111497	2
ATTCG	3092150	2.3203475	41.608402	34
GCGGA	1663910	2.2773561	24.698908	7
AGAAA	2435245	2.246594	5.394919	22
CGAGT	2216085	2.245863	44.271618	33
TTCGT	7602530	2.2436957	5.3290887	35
TTTTA	53984220	2.2432833	12.687236	26
GGTCG	4142615	2.2299178	33.213634	42
TTTAG	39702915	2.2281475	16.276836	27
GAGGT	21720170	2.2232757	24.739658	40
GAAGA	4517455	2.2135699	8.607709	46
AGTAG	11362750	2.1897578	22.155266	35
GAGCA	849490	2.1889803	17.087116	47
GTCGC	409080	2.1801643	10.962922	3
TCGTC	548635	2.165012	10.65564	40
TACGC	211360	2.1207347	10.843749	13
GCGTT	5276805	2.1032026	25.956846	16
TTTAC	3785205	2.1031861	36.186966	13
AGGAG	7987975	2.0789952	9.579895	38
GCGGT	3860985	2.0783198	30.859303	6
CGAGC	153175	2.0756576	7.7394633	32
GAGAT	10593970	2.041603	7.8234677	26
GACGC	148265	2.0091228	14.021855	3
AGGTC	1974995	2.0015333	57.18671	41
ACGGC	146520	1.9854765	10.97724	12
ACGGA	766730	1.9757229	7.5238986	30
AAGAG	4031980	1.9756852	8.60437	47
ATTTT	47467915	1.9725015	8.343648	25
GCGTA	1932455	1.9584217	27.66503	4
AGCGC	143975	1.9509895	9.942553	35
TAGTT	33308770	1.8693048	9.134663	29
CACAT	98815	1.8666762	73.66227	39
AAACG	382850	1.8573545	9.19583	7
AATTT	17365935	1.8348571	18.990734	24
TACGG	1805650	1.8299128	15.70277	5
AACTC	96450	1.8219999	75.047356	22
AGGTA	9437855	1.8188039	29.081972	47
TAGAG	9394210	1.8103931	8.0505295	24
GCGAC	133470	1.8086374	20.894363	23
AGCGG	1317060	1.8026303	7.1591773	6
GAGCG	1295275	1.7728137	10.381175	28
TTAGT	31540320	1.7700586	15.656919	28
TAGTA	12386560	1.7674984	17.688799	29
GGAAA	3599355	1.7636975	12.628936	2
TAGCG	1732235	1.7555113	5.521016	10
GCGTC	327335	1.74451	10.828446	40
GGACG	1270095	1.7383504	17.227636	2
ATCGT	2305205	1.7298248	11.39223	39
GAAAA	1868450	1.7237068	5.5441165	3
GGAGA	6600125	1.7177861	10.991713	2
ACATC	89585	1.6923158	75.162415	40

TGGGA	16443085	1.6831135	16.164686	37
CGTAC	166840	1.6740319	9.28555	13
AGTTT	29784750	1.671535	8.325263	26
AGTTA	11656135	1.6632705	20.207216	30
TGGCG	3085360	1.6608106	35.104507	10
TCGTA	2210670	1.6588855	7.3193817	45
AACGC	64700	1.6506488	6.731147	23
AGTCG	1625880	1.6477271	13.700995	22
TCGAC	163460	1.6401178	7.484559	23
GCGTG	3025360	1.6285133	35.13201	4
CGATT	2167735	1.6266675	19.350811	11
CGTGG	3021500	1.6264355	34.8309	5
TATCG	2165790	1.6252077	11.790016	38
TATTT	38664410	1.606677	6.3643827	32
AGCGT	1566665	1.5877165	8.226508	29
TGAGA	8237895	1.5875552	5.3867626	41
GTCGT	3978265	1.5856367	10.770899	3
GTAGA	8202570	1.5807476	7.7301683	23
GGGAA	6023850	1.5678012	14.462921	2
GTACG	1543475	1.564215	15.463244	4
CGAAA	321880	1.5615653	5.6396728	32
TAGGA	8098780	1.5607458	7.3352537	37
TCGAA	802830	1.5318031	5.0990453	32
TTGAG	20142955	1.5266833	13.877108	44
ACGGT	1499995	1.5201507	15.082077	6
TTCGG	3755540	1.4968643	22.60776	35
AGGTT	19713710	1.4941498	14.7546215	41
GTAGT	19645045	1.4889455	9.356291	36
ACGAG	571665	1.4730763	5.200983	32
GGTTT	49081900	1.4630541	9.469917	2
TAATT	13834455	1.4617265	18.732557	23
GCGAT	1420300	1.4393848	23.734797	10
AAGTA	3965400	1.4387394	11.206415	34
TATAG	10077355	1.4379866	16.8325	47
AACGG	556535	1.4340891	6.2301617	29
TTATT	34472155	1.4324707	6.774002	32
TTTAA	13468830	1.4230952	8.242155	5
TTATA	13399750	1.4157963	12.887869	46
TTAAG	9886575	1.4107636	10.211444	6
GGCGA	1017675	1.3928688	9.105562	2
CGAAC	54465	1.38953	9.122855	20
GTTTA	24750665	1.3890196	8.054707	12
TAAGC	713050	1.3605025	40.92781	7
GACGT	1335585	1.3535315	6.434176	3
GAACG	524295	1.3510125	6.179308	28
GGAGT	13010835	1.3317885	11.00078	2
GGTAG	12897270	1.3201641	7.8700833	2
ACGAC	51345	1.3099314	5.1255946	18
GAGTA	6762445	1.3032157	16.090145	34
TCGGG	2414920	1.2999212	29.067987	36
GGAAT	6545905	1.2614857	9.500144	2
TTGTA	22429165	1.2587359	15.003645	20
ATTAT	11843320	1.2513462	12.738917	45
GGGTT	31078395	1.2511281	14.436373	2
GGTTA	16465910	1.2479912	14.639139	2
CGTAA	652640	1.24524	10.696911	21
CGTAT	1646160	1.2352777	5.081651	13
GTAAT	8630865	1.23158	23.703068	22
GGGAT	11926260	1.2207714	11.558683	42
TATTC	2189325	1.2164618	29.606657	33
GGGGA	8687365	1.200943	10.306865	2
GATTA	8344180	1.1906716	16.428688	44
GTTAA	8324400	1.187849	16.565218	3
TGAGG	11555495	1.1828197	16.48066	45
TGGAA	6121780	1.1797509	9.592442	1
GGTGG	21662255	1.1777449	11.630701	8
TTTGT	53305350	1.1765369	6.942215	19
AAAAC	128195	1.1708999	15.057712	6
AGTAT	8170380	1.1658711	16.777225	30
TCGAT	1515255	1.1370467	6.223372	11
TAGGC	1112960	1.127915	8.9993725	13
TCGTG	2825015	1.1259803	5.2997756	40
GGATT	14839025	1.1246858	8.938459	43
AAGGC	432685	1.1149502	19.61713	46
TTTTC	5055300	1.1047103	10.792838	29
GTGGC	2032660	1.0941556	33.2313	9
TGTAA	7654735	1.0922912	23.02154	21
GGGTA	10526255	1.0774668	15.34655	2
CGTGA	1061770	1.0760372	7.945553	26
AGTAA	2964380	1.075546	7.328176	9
TGGAG	10450425	1.0697051	10.60906	1
GTATT	19049760	1.069082	7.0963564	31
GGGGT	19402365	1.054878	8.658462	2
TGTAG	13735730	1.0410643	8.067285	21
GTTAT	18398695	1.0325439	8.212175	31
CGTGC	192065	1.0235975	5.492964	13
AGTTG	13287085	1.0070603	9.381166	38
GTTGA	13276925	1.0062903	12.925623	43
ATTTC	1783890	0.9911886	5.3207493	22
TAAGT	6928790	0.98870265	6.489243	7
TCACG	98185	0.98516434	39.61976	30
TTAAT	9295080	0.9821034	10.91723	4
TGCGG	1795890	0.9667051	6.7347984	5
GGTTG	23873115	0.96106404	6.8961887	42
AAGAC	196255	0.9521095	8.701144	32
GGAGC	692500	0.9478091	9.484993	27
CGTGT	2375355	0.94675696	5.080755	41
TAAGG	4885505	0.9415038	5.831152	45
TAGAC	489140	0.93328124	9.969474	25
GTGGT	23131920	0.9312258	8.299137	9
GGATA	4819125	0.9287115	7.594653	2
TTGGG	23009255	0.92628753	6.8984227	36
GTTTG	30707600	0.91534513	7.003305	18
TGGGG	16823095	0.91464686	9.015537	1
TCCAG	91040	0.9134732	39.577328	25
CAGTC	89790	0.9009311	39.716415	27
TGGTT	30068850	0.89630497	7.734166	1

CCAGT	89165	0.89466006	39.570263	26
GTCAC	89145	0.8944592	39.7612	29
AGTGA	4639840	0.89416075	5.1646047	18
GGTAC	868435	0.88010436	15.614117	3
TTTGG	29471245	0.8784914	5.496161	35
GAAGC	340500	0.8774063	11.409956	4
CGTCT	221815	0.8753217	15.781398	16
TTATC	1557090	0.86517113	8.487373	37
GGGTG	15730495	0.8552437	8.267031	2
GGTAT	10891595	0.8255004	6.389773	2
AGTGG	7670180	0.7851192	5.563488	8
GTGCG	1457425	0.7845136	6.240444	4
GGGGG	10633870	0.780806	5.8914394	2
TGGGT	19215030	0.7735428	9.482596	1
GGTAA	3988455	0.76862997	6.378877	2
TGGTG	18894095	0.7606228	6.153922	1
GGAAC	282445	0.72780925	6.1461883	2
ATCTC	97935	0.7276076	30.370686	42
GTTGG	17477840	0.7036084	5.0608473	39
GAGTC	689945	0.6992159	12.590697	21
TGGTA	8723995	0.6612127	5.6674104	1
CATCT	88690	0.65892196	29.534536	41
CGGCC	8530	0.6078566	6.244619	1
TGGAT	6868160	0.52055454	5.5264173	1
GATTC	688445	0.5166088	5.3850627	29
TGGGC	952390	0.51265967	5.28525	13
GTCCG	93705	0.49939457	20.85373	34
TCTCG	110750	0.43703932	16.166834	43
CTCGT	107075	0.42253712	16.188686	44
TGAAC	216700	0.41346455	8.105612	20
GGTGC	756695	0.4073195	6.3283944	3
CTGAA	147150	0.28076285	7.7954454	19
GAACT	113385	0.2163391	7.796793	21
AGTCA	104500	0.19938646	7.7254457	28
CGGCT	35105	0.18708976	6.32987	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
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	224739	0.3284601011934625	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	186250	0.2722077336255941	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	125509	0.18343366678987752	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	100104	0.14630380116433003	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	74005	0.10815964202395753	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	73054	0.10676973837468	No Hit