FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD20_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`

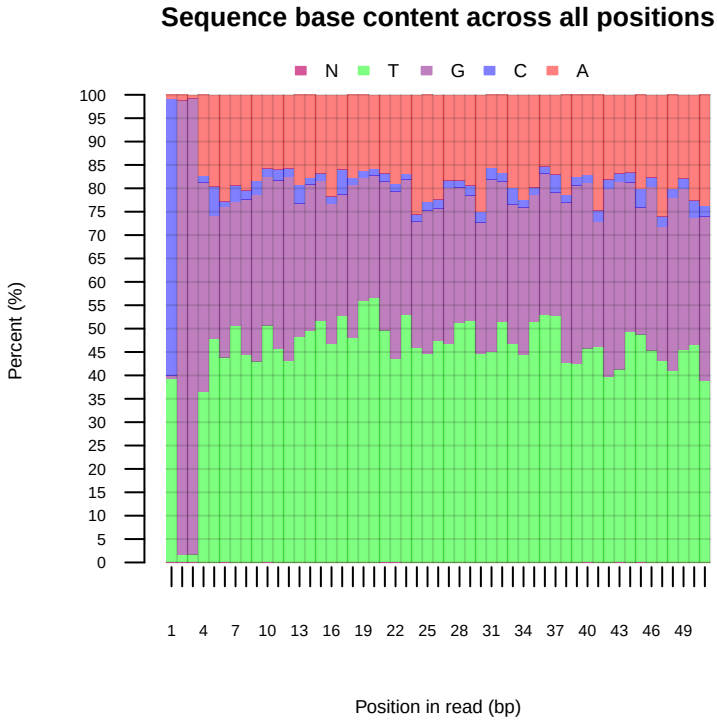
1 Sequence Duplication

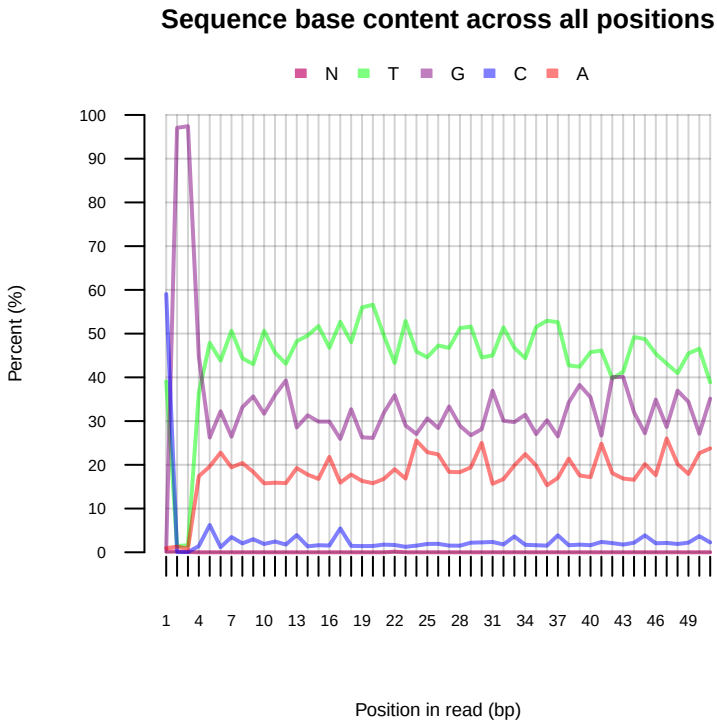
- Estimated Duplication rate 79.4679%

2 Per base sequence quality

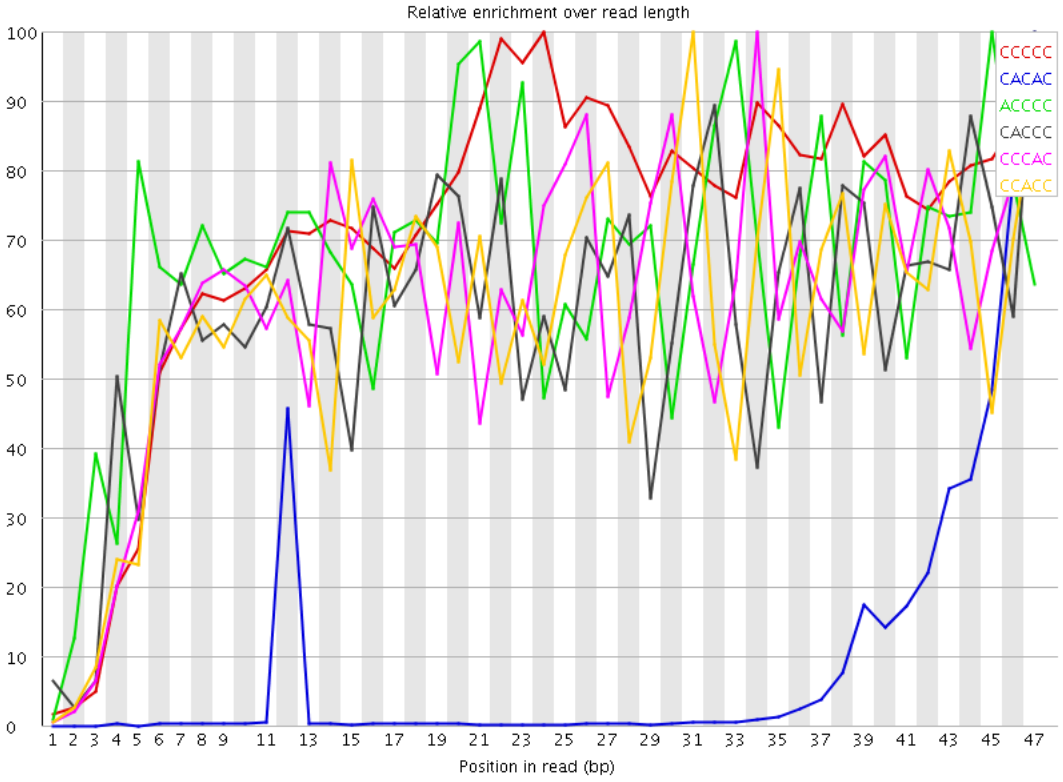
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	173450	1312.9288	1833.4965	24
CACAC	376235	92.579185	978.6403	47
ACCCC	48100	65.64559	98.11715	45
CACCC	46270	63.148056	106.13332	47
CCCAC	43985	60.029556	97.7946	34
CCACC	43665	59.592827	101.32162	31
CCCCA	42995	58.678436	87.53419	28
CGGGC	4549795	31.436293	1115.9927	1
GCCCC	21590	15.852961	29.327154	47
CCCGC	21460	15.757505	33.639973	47
CCCCG	20930	15.3683405	29.844687	46
CCGCC	20825	15.291242	33.466785	35
CGCCC	20805	15.276556	27.08394	36
AGCAC	557585	13.309331	104.477135	47
CGCGG	1900035	13.128077	315.00113	5
GCGCG	1783870	12.325448	313.68463	4
GCACA	460765	10.998277	97.192314	46
CGGCG	1500650	10.368571	288.7028	1
CGGAA	4323175	10.010108	203.53119	1
CGCGC	133355	9.498563	64.44313	13
GGCGC	1360280	9.398701	315.10544	3
CGGGA	6379745	7.94761	232.5565	1
ACACG	284990	6.8025956	75.20895	47
TCGCG	1261330	6.5067673	24.43327	30
CGGGT	12232040	6.121056	237.53598	1
AGACG	2596070	6.0110784	69.445694	27
CGGAG	4502115	5.608539	166.39566	1
AGATC	3216045	5.559749	29.577225	43
ACGCG	413000	5.3038635	16.597338	14
CGCGT	1016315	5.2428193	22.519571	31
TCCCC	9385	5.145044	17.13014	5
CGGAC	389040	4.9961624	166.19073	1
CGGTT	13246025	4.9489155	170.98347	1
CGTCG	951955	4.910808	25.363625	41
AACCC	19825	4.8782873	10.232504	32
CGCGA	373480	4.7963367	20.724798	5
CGGTC	928305	4.788806	181.11253	1
CTCCC	8595	4.7119503	10.947835	24
CGGGG	7022445	4.706724	117.66689	1
GGGCG	6955050	4.6615534	113.35381	2
CCTCC	8490	4.654387	8.500691	23
CCCCT	8315	4.5584483	8.758443	46
CCCTC	8225	4.5091085	8.372043	47

AGGCG	3598120	4.4823813	64.14044	47
TCGAG	4775920	4.4420896	54.972153	44
ACTCC	44815	4.429675	170.89241	23
GAGAC	1881185	4.3557954	66.49129	26
CTCCA	42640	4.2146907	169.8238	24
CAACC	16890	4.1560793	9.018478	31
ACACC	16485	4.0564218	6.3591833	37
ACCCA	16385	4.031815	7.746641	33
CGACG	311090	3.9951067	30.421919	24
AAAAA	2716795	3.9182498	10.443608	31
ACCAC	15705	3.8644893	6.706179	45
CCCAA	15675	3.8571079	6.9372916	29
CCAAC	15520	3.8189673	8.555993	30
CCACA	15225	3.7463775	6.7061825	46
CACCA	15020	3.695933	6.6482368	36
TTACG	5313715	3.6899934	45.951294	14
CGGTA	3919290	3.6453366	127.23524	1
CGAGG	2820110	3.5131707	69.68206	45
CGTTT	12458590	3.4752855	40.11954	17
GACGG	2716200	3.3837242	36.9182	28
TACGT	4848150	3.3666916	47.739883	15
ACGTC	351010	3.3655717	18.25548	15
ACGTT	4844730	3.3643167	49.479206	16
GGCGG	5009465	3.3575442	41.79324	11
GGCGT	6702010	3.3537643	53.693832	3
GATCG	3559965	3.3111281	17.31973	44
ACGGG	2632235	3.279124	37.183655	29
CGGAT	3416925	3.1780863	108.46104	1
AGAGA	7083540	2.9571953	23.072437	25
AAGCG	1246750	2.886791	54.019905	8
TTTCG	10246370	2.8581934	15.54533	30
ATCGC	296105	2.8391287	34.99943	29
AGCGA	1200220	2.7790532	54.69718	9
CGAGA	1198510	2.7750938	34.86537	25
GTCGA	2960170	2.753258	54.2199	43
TTCGA	3873460	2.6898394	34.9636	31
GCGGC	385855	2.6660213	9.431291	9
GAGGC	2069810	2.5784793	53.313236	46
GGAGG	21212135	2.563353	30.69518	39
GCGGG	3813115	2.5557024	42.58098	12
CGTTA	3670165	2.548666	31.840206	9
AGAGC	1095875	2.5374472	16.728558	47
TCGGA	2709115	2.5197513	15.46911	46
TTTTT	164723950	2.4846394	5.396405	16
ATCGG	2670995	2.4842958	15.163176	45
CACGT	258665	2.4801447	18.668093	47
CGTTC	641660	2.471375	31.576487	33
TCGTT	8831955	2.4636467	6.4504547	4
TTCGC	638885	2.4606867	9.128369	33
ATTCG	3443795	2.391468	43.49738	34
CGGTG	4740035	2.371969	46.399483	1
GGGAG	19221470	2.3227937	26.900766	38
CGTAG	2479470	2.3061583	25.612131	5
AGAAA	2953095	2.2914474	5.2484574	22
TTCGT	8156165	2.275137	6.018117	35
GCGGA	1822960	2.2709646	23.601185	7
TTTTA	60170125	2.2594006	12.230346	26
AGTAG	13449240	2.2553902	20.782688	35
TTTAG	44488525	2.2375011	15.599894	27
CGAGT	2394795	2.227402	42.876053	33
GGAAG	9876705	2.2183952	11.6178055	2
GAGGT	24062510	2.1710114	23.203632	40
ACGGA	932805	2.159866	11.755068	30
GGTCG	4255505	2.1295044	31.167864	42
GCGTT	5653880	2.1123753	26.849297	16
AGGAG	9377170	2.1061954	9.618683	38
TTTAC	3950335	2.0481336	33.443596	13
GCGGT	4032515	2.0179176	29.430492	6
GACGC	157065	2.0170736	13.494252	3
GAAGA	4811180	2.0085435	6.625787	46
TACGC	209420	2.0079713	11.28578	13
ATTTT	52825750	1.9836178	7.8254976	25
AAACG	458765	1.9743725	9.578155	7
AGCGC	153405	1.9700708	8.641135	10
GAGAT	11568880	1.9400603	8.789425	26
TCGTC	503415	1.9389197	10.774638	40
GTCGC	375110	1.9350634	9.92918	3
TAGAG	11506900	1.9296664	10.202723	24
CGAGC	148905	1.9122806	6.094642	13
ATCGT	2745935	1.9068544	15.799041	39
ACGGC	146420	1.8803674	9.295836	12
GCGTA	2015245	1.874382	25.323906	4
TAGTT	37079745	1.8648847	9.294875	29
AATTT	19851865	1.8557496	17.32049	24
AGGTC	1988995	1.8499668	51.810593	41
TACGG	1952000	1.8155577	13.876251	5
TTAGT	35694535	1.7952173	14.993345	28
AGGTA	10703320	1.7949089	28.020344	47
AAGAG	4296815	1.7938094	6.627946	47
TAGCG	1923270	1.7888359	5.511005	10
TATCG	2574130	1.7875482	16.217402	38
GAGCG	1433290	1.7855303	11.04268	28
TAGTA	14238395	1.7827153	15.26144	29
AGCGG	1427865	1.7787724	6.2597766	6
GGAAA	4241625	1.7707691	11.863249	2
GAAAA	2254260	1.749188	5.167106	3
GGAGA	7679865	1.7249656	10.848935	2
GGACG	1376885	1.7152635	16.937714	2
GTAGA	10150425	1.7021904	9.915493	23
AGTTA	13434035	1.6820058	20.210592	30
TGGCG	3354015	1.6783884	35.831043	10
AGTTT	32893675	1.6543511	8.016363	26
AGTCG	1776440	1.6522691	14.120715	22
CGTGG	3297795	1.6502552	35.77394	5
CGATT	2356070	1.6361215	19.032225	11
GCGTG	3259780	1.631232	35.97827	4
AGCGT	1741170	1.6194646	8.848837	29

AACGG	698665	1.6177261	10.749776	29
TGGGA	17893505	1.6144202	14.597186	37
GAGCA	695090	1.6094482	12.69697	47
TCGTA	2307280	1.6022402	5.1435056	45
TATTT	42593080	1.599379	5.7089334	32
CGTAC	165715	1.588917	10.05358	13
GCGAC	123125	1.5812063	21.111042	23
GTCGT	4222820	1.5777096	10.914235	3
TAGGA	9343130	1.5668099	7.4289594	37
AACGC	65380	1.5605943	5.6639786	11
GGGAA	6916080	1.5534129	14.077852	2
TTCGG	4156280	1.5528492	23.630177	35
TCGAA	895650	1.5483581	5.1703024	32
GCGTC	299925	1.5472099	10.397763	40
GTACG	1640340	1.5256823	13.65686	4
ACGAG	647960	1.5003209	5.189115	32
TAATT	15932505	1.4893683	17.061815	23
AGGTT	22011455	1.4827466	14.283642	41
ACGGT	1586570	1.4756708	13.341351	6
GGTTT	54208200	1.4668221	9.134016	2
GAACG	629765	1.4581914	10.439702	28
TCGAC	151105	1.4488326	6.0168443	23
GTAGT	21502150	1.4484385	8.967549	36
TTGAG	21413240	1.4424493	13.436403	44
GGCGA	1152525	1.4357656	10.099756	2
AAGTA	4580190	1.4276121	10.720825	34
TTATT	37987205	1.4264275	6.9625573	32
GCGAT	1525880	1.419223	23.041552	10
GGAAT	8410870	1.4104731	9.93014	2
TATAG	11144820	1.3953849	15.906608	47
TTTAA	14879360	1.3909205	7.7938423	5
TTAAG	11000850	1.3773593	9.529685	6
AGATA	4359265	1.3587514	5.533086	26
GTTTA	26977165	1.3567868	7.7805276	4
TTATA	14465430	1.3522265	12.230595	46
GACGT	1430260	1.3302867	6.211339	3
GGTTA	19640310	1.3230205	17.348808	2
TCGGG	2639810	1.3209919	30.139587	36
GGTAG	14637030	1.3206085	7.628239	2
GGAGT	14583230	1.3157545	10.496261	2
TAAGC	760980	1.3155468	38.95769	7
TGGAA	7841020	1.3149114	9.049178	1
TTGTA	25481030	1.2815402	14.377395	20
GTTAA	10235195	1.2814956	21.357548	3
GAGTA	7582740	1.2715987	14.925566	34
GGGTT	35057690	1.2705704	14.016723	2
TATTC	2405870	1.2473736	31.09719	33
ATTAT	13223450	1.2361263	12.0720625	45
GCACC	9280	1.2285687	5.5053887	47
TCGTG	3275245	1.2236813	7.6126447	40
GGGGA	10044455	1.2138093	10.5791	2
CGTAA	698760	1.2079837	9.306709	21
CGTAT	1735330	1.205062	5.333146	13
GGGAT	13309690	1.200851	11.003954	42
AAAAC	149470	1.1956267	15.453515	6
TTTGT	58745910	1.1868263	6.8198533	19
GTAAT	9469835	1.1856688	21.419836	22
TGAGG	12932945	1.1668594	16.14035	45
GGTGG	23996235	1.164827	11.28615	8
GATTA	9266005	1.1601484	15.5397835	44
TCGAT	1624930	1.1283971	7.2852225	11
TAGGC	1212155	1.127427	9.286088	13
AGTAT	8824435	1.1048617	14.356947	30
GGATT	16367270	1.1025403	8.544199	43
GTGGC	2192530	1.0971677	33.928112	9
TTTTC	5262755	1.0960536	10.88428	29
CGTGA	1175285	1.093134	9.225341	26
GGGGT	22424455	1.0885297	8.54266	2
GTATT	21547430	1.0837041	6.2492943	31
GGGTA	11928365	1.0762225	14.8781595	2
TGGAG	11859305	1.0699916	10.14802	1
AGTAA	3428415	1.0686121	6.728999	9
TGTAA	8444130	1.0572457	20.864002	21
TTAAT	11192405	1.0462644	14.497908	4
AAGGC	447290	1.035679	16.737257	46
GTTAT	20550785	1.0335791	8.370417	31
CGTGT	2753500	1.0287493	7.256191	41
TGTAG	15232300	1.0260859	8.032887	21
ATTTC	1974430	1.0236845	6.132199	22
CGTGC	194700	1.0043902	5.3229623	13
TTATC	1909530	0.9900358	11.870883	37
AGTTG	14693145	0.9897671	8.992761	38
GTTGA	14539705	0.9794309	12.665171	43
CGAAC	40800	0.9738797	5.154885	20
TAAGT	7713170	0.96572596	6.0903654	7
GGTTG	26531625	0.9615663	6.909906	42
GGAGC	768945	0.9579182	10.138602	27
TGCGG	1904940	0.9532543	6.386423	5
AAGAC	219500	0.944655	8.757119	32
TAGAC	546015	0.94392526	10.468136	25
TGGGG	19197170	0.93187046	8.975317	1
GGATA	5542490	0.929456	7.537342	2
ATTAC	719605	0.9288044	5.4816456	29
TAAGG	5500425	0.9224019	5.0687413	45
TTGGG	25336710	0.9182598	6.404203	36
GTGGT	25002130	0.90613407	7.8641033	9
GTTTG	33371215	0.9029932	6.801281	18
TGGTT	32978830	0.89237565	7.4419527	1
AGTGA	5269605	0.8836943	5.2270136	18
TTTGG	32579515	0.8815705	5.162002	35
GGAAC	375755	0.87004304	9.545927	27
AACTC	48695	0.86781377	31.791744	22
GGGTG	17604910	0.85457885	8.242457	2
GAAGC	363955	0.84272075	9.897345	4
GGTAC	899170	0.83631915	13.821534	3
CGATC	86380	0.8282331	5.0347652	44
GGTAT	12218635	0.8230778	6.123501	2

GGGGG	12370400	0.80427593	5.918413	2
GGTAA	4609110	0.77293146	6.2043786	2
AGTGG	8492025	0.76618284	5.048685	8
TGGGT	21101465	0.7647651	9.116573	1
GTGCG	1508630	0.7549361	5.95378	4
TGGTG	20708670	0.7505292	5.9302287	7
GAGTC	737525	0.6859729	12.977011	21
TGGTA	9780210	0.65881944	5.311302	1
CGTCT	156905	0.6043248	7.0200133	16
GATTC	774185	0.53761584	6.2743464	29
TGGGC	1056650	0.52876	5.7633233	13
TGGAT	7628135	0.5138502	5.1330795	1
TCACG	47665	0.45702395	16.933174	30
TCCAG	43450	0.4166095	17.061577	25
CAGTC	42875	0.41109622	17.255304	27
GTCAC	42280	0.40539125	17.219263	29
CCAGT	41565	0.39853567	17.025536	26
GGTGC	782830	0.39173728	6.04571	3
ATCTC	47830	0.3424024	13.235007	42
TCTCG	59910	0.23074536	7.1660137	43
CTCGT	58225	0.22425553	7.1730657	44

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	238248	0.31263541770811004	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	204668	0.2685708407687933	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	143540	0.18835703912654927	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	103675	0.1360451165629441	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	93572	0.12278768890308953	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	92161	0.12093613684646723	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	80343	0.10542824017377977	No Hit