FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD5_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`

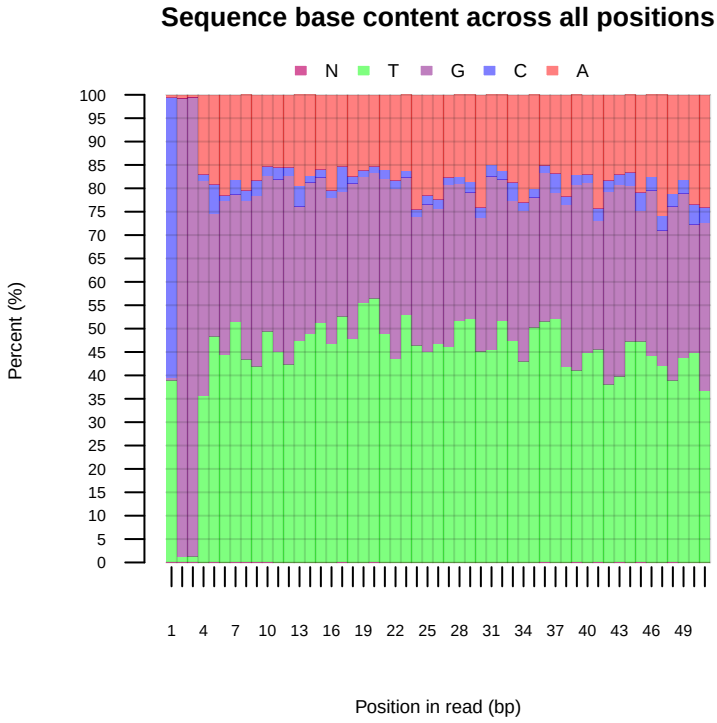
1 Sequence Duplication

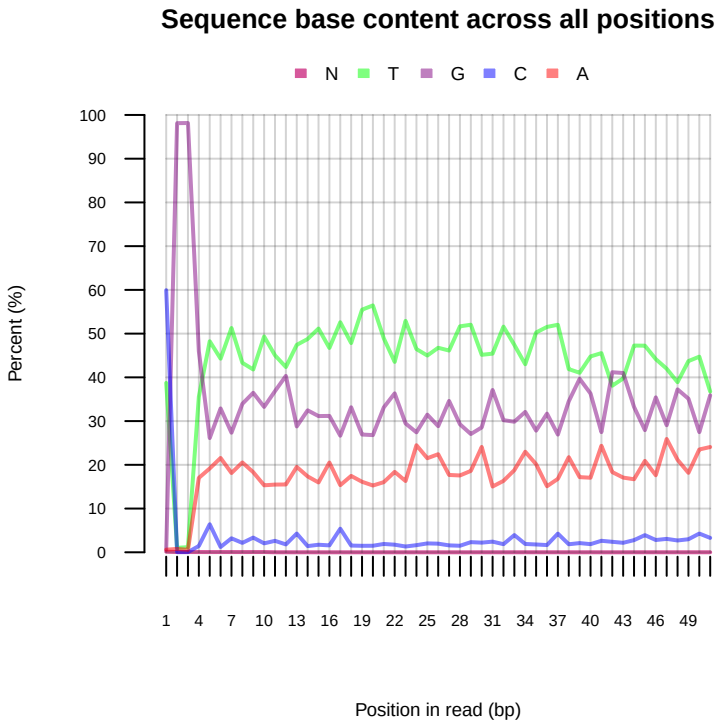
- Estimated Duplication rate 75.8545%

2 Per base sequence quality

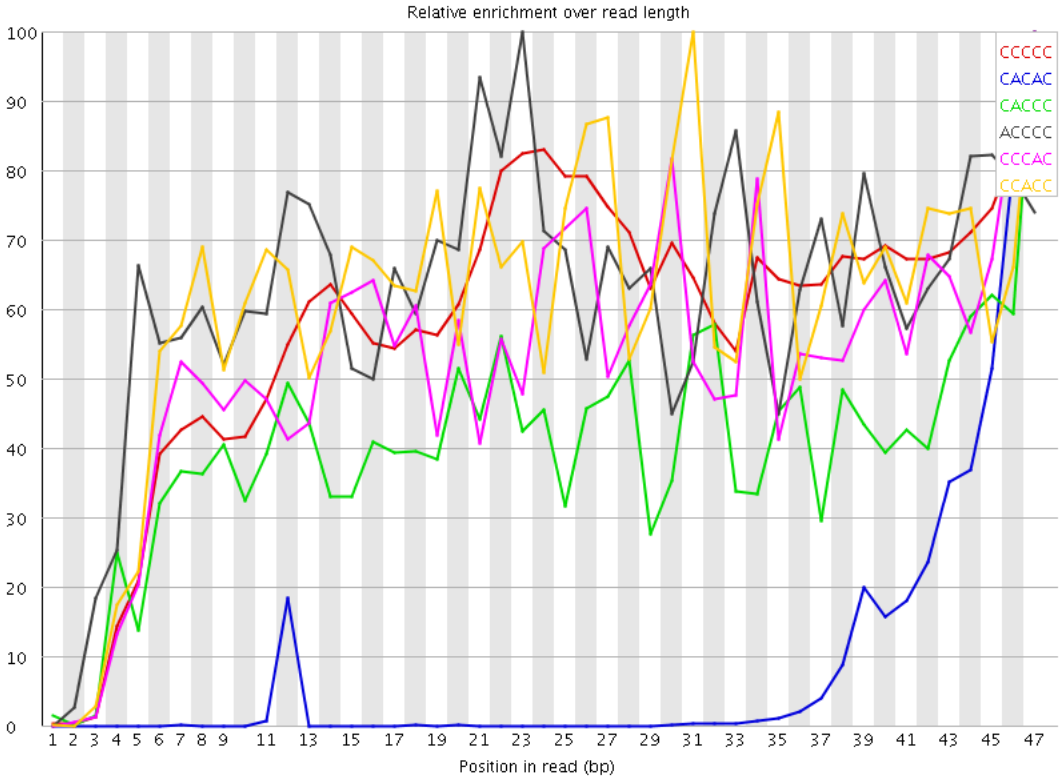
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	120470	646.2584	1106.7231	47
CACAC	1120500	231.62604	2561.5325	47
CACCC	43390	45.692097	112.08997	47
ACCCC	42470	44.723286	72.00135	23
CCCAC	40985	43.1595	81.90239	47
CCACC	39125	41.200813	67.30253	31
CCCCA	38330	40.363636	65.81598	24
AGCAC	1625655	34.154766	271.46057	45
GCACA	1387395	29.14896	270.425	46
CGGGC	4759840	26.807844	926.6812	1
ACACG	858130	18.029182	213.93355	47
GCCCC	26905	14.669246	42.53339	47
CGGAA	6618225	14.132267	182.28244	1
CGCCC	25020	13.641499	23.571589	23
CCCGC	25020	13.641497	23.95616	26
CCCCG	24635	13.431585	25.622555	46
CCGCC	24130	13.156249	21.77905	35
AGATC	6241415	10.304586	51.89585	43
CGCGG	1818085	10.239617	221.9771	5
GCGCG	1708625	9.623129	220.4749	4
CGGCG	1555150	8.758743	237.98833	1
CGCGC	157190	8.710579	60.297626	13
CGGGA	6798645	7.516506	213.72418	1
GGCGC	1251605	7.049151	220.94893	3
TCGCG	1350990	5.8830004	20.834225	30
CGGGT	13283910	5.8792243	228.58746	1
ACGTC	667320	5.612516	39.364807	47
GATCG	6494495	5.5515757	28.061224	44
AGAGC	2569775	5.487385	33.75574	47
TCCCC	12990	5.475971	13.379074	5
ACTCC	65970	5.4591184	179.92924	23
CACGT	644540	5.420924	54.668842	47
CTCCC	12455	5.25044	11.093975	47
AGACG	2458590	5.249965	57.688927	27
CGGAG	4631875	5.120949	148.6367	1
CCTCC	12075	5.0902495	10.202506	46
CCCTC	12055	5.081819	9.509093	45
CCCCT	11950	5.0375557	11.391133	47
CTCCA	60655	5.0192943	179.79402	24
ACGCG	435245	4.7345643	15.522131	14
CGCGT	1084225	4.7213492	18.986378	31
CGGAC	432315	4.7026916	158.97502	1
ATCGG	5457105	4.664801	26.549263	45

CGGTT	13313540	4.5558023	153.50546	1
TCGGA	5294540	4.525839	26.774134	46
AAAAA	2825465	4.4181213	15.6868515	31
CGGTC	1004235	4.3730264	163.69594	1
CGCGA	397705	4.3262067	18.601868	24
CGTCG	992805	4.3232536	22.814045	41
GGGCG	7499470	4.2928677	98.64809	2
CGGGG	7309565	4.1841617	101.6732	1
GAGCA	1951960	4.1681294	28.430994	44
TCGAG	4734015	4.046695	46.832363	32
AACCC	19325	3.9947999	6.4113026	20
AGGCG	3500480	3.8700917	45.156788	47
ACACC	18390	3.8015196	8.305941	47
ACCAC	17810	3.681624	7.917335	45
GAGAC	1723070	3.679368	55.462883	26
CGACG	332790	3.6200662	32.053593	24
CCACA	16950	3.5038476	6.945912	46
CAACC	16585	3.428396	5.2943554	31
ACCCA	16470	3.4046237	6.022968	33
CGTTT	12798135	3.3860643	36.73806	17
CACCA	16215	3.3519108	5.342937	32
TTACG	5068445	3.34983	37.05922	6
CGGTA	3889360	3.3246732	114.42413	1
CCCAA	15760	3.257855	5.974237	14
GGCGT	7347205	3.2517433	49.80671	3
CCAAC	15705	3.2464855	5.5856376	15
GGCGG	5615440	3.2144058	41.79667	11
TACGT	4671555	3.087518	36.561523	7
ACGTT	4644845	3.069865	37.5919	16
GACGG	2709450	2.9955378	30.02619	28
CGGAT	3453700	2.9522655	99.81039	1
ACGGG	2599750	2.8742545	30.081753	29
AAGCG	1323570	2.8262932	56.106155	8
CGAGG	2555255	2.8250613	48.650208	45
TTTCG	10652800	2.8184626	16.712368	30
GAAGA	6617520	2.7738926	12.540224	46
AGAGA	6568350	2.7532818	20.484371	25
AGCGA	1269720	2.7113042	56.87651	9
ATCGC	312860	2.631319	30.716145	29
CGAGA	1230915	2.6284418	35.245914	25
TTTTT	163483360	2.6279912	5.896362	16
TTCGA	3940175	2.6041353	37.592087	31
GGAAG	11925690	2.5882204	11.258743	2
CGTTA	3846645	2.5423195	36.027058	9
GCGGG	4325445	2.4759834	42.62492	12
AAGAG	5898645	2.472559	12.563552	47
TCGTT	9254300	2.4484549	6.602989	4
GAGAT	14472475	2.4284978	8.725355	26
GGAGG	21544920	2.4209507	27.960325	39
GTCGA	2755915	2.3557906	39.100594	43
ATTCG	3561990	2.3541856	45.667633	34
GCGGC	414510	2.3345573	9.264087	9
TTTTA	57449815	2.3069515	12.736507	26
TTCGC	684190	2.3035657	9.339853	33
CGTTC	670615	2.2578611	26.253803	33
TTTAG	43258000	2.2466688	15.762012	27
TTCGT	8453515	2.236587	5.589484	35
AGAAA	2745255	2.2225614	5.9009686	22
CGAGT	2589870	2.2138534	45.798923	33
AGTAG	13187635	2.2129002	24.060253	35
GGGAG	19546530	2.1963964	23.938528	38
GCGGA	1968095	2.1759038	24.375664	7
GAGGC	1942680	2.1478052	37.742107	46
CGTAG	2507720	2.14363	22.570698	5
GAGGT	24300390	2.1112103	21.788404	40
GCGTT	6121985	2.0949016	27.960281	16
CACGC	19320	2.067788	11.040292	47
CGGTG	4643760	2.0552464	40.86777	1
AGGAG	9390210	2.0379477	10.325357	38
TAGTT	38093430	1.9784391	10.71038	29
ATTTT	49221065	1.9765182	7.9454	13
TGAGA	11706390	1.9643456	6.983831	41
AAACG	470695	1.9412729	17.367613	7
CGCAC	18125	1.939889	10.28583	47
TAAAA	3047105	1.9073732	6.230672	30
TTTAC	3727870	1.9049604	27.828772	5
ACGGA	873250	1.8646997	8.836179	30
AAAAT	2969695	1.8589175	6.054051	32
GGTCG	4191570	1.8551149	21.966263	42
TACGC	216230	1.8186094	11.031044	13
AATTT	18105205	1.8161545	16.633928	24
TAGAG	10779270	1.8087739	9.090002	24
GCGGT	4084190	1.8075905	22.367523	6
TTAGT	34754965	1.8050508	15.16962	28
AGGTA	10740430	1.8022565	29.858343	47
AGCGC	161460	1.7563505	7.0724626	35
GACGC	161450	1.7562418	9.531163	5
GCGTA	2052615	1.7546008	22.250158	4
GCACC	16370	1.7520542	10.86425	47
TCGTC	520295	1.751756	9.270088	40
CGAGC	160685	1.7479202	6.530539	32
ATCGT	2630575	1.7385961	13.69567	39
GTCGC	399115	1.73798	8.714987	3
AGCCC	16080	1.7210159	5.8596535	47
ACGGC	157440	1.7126212	8.14342	12
AGTTA	13198075	1.7123115	23.569479	30
GGAAA	4083715	1.7117875	11.775898	2
AGTTT	32941910	1.7108874	9.548585	26
GAAAA	2102260	1.7019919	5.4676957	3
TTGAG	25249240	1.6960696	14.799023	44
GGAGA	7791950	1.6910789	10.28181	2
GAGCG	1517255	1.6774602	10.410257	28
TGGCG	3786570	1.6758691	36.74167	10
AAATA	2654365	1.6615328	5.693866	33
AGCGG	1502145	1.6607547	7.1747293	6
GGACG	1486160	1.6430819	16.958138	2
TAGTA	12595675	1.6341561	12.214723	29

TAGCG	1910840	1.6334097	5.489994	10
TATCG	2459135	1.6252882	14.0958185	38
TACGG	1891400	1.6167922	10.989737	5
TATTT	40197870	1.6141833	5.4442806	33
CACAT	98535	1.600629	35.752937	31
GCGTG	3609155	1.5973483	34.613068	4
GTAGA	9515375	1.5966908	8.773205	23
AGTCG	1862995	1.5925114	14.398584	22
TGGGA	18095130	1.5720991	12.897216	41
CGATT	2368350	1.5652869	19.793507	11
CGTGG	3508135	1.5526386	34.384373	5
TCGTA	2317245	1.5315105	6.147444	45
TTCGG	4466130	1.5282791	24.445488	35
AACGC	72690	1.527206	6.2051806	23
ATAAA	2439335	1.526932	5.3836784	37
TAGGA	9073370	1.5225219	8.220583	37
GTCGT	4437720	1.5185577	10.770596	3
GCGAC	138990	1.5119233	21.252241	23
GGGAA	6938360	1.5058252	13.7137575	2
AGCGT	1749170	1.4952124	8.47963	29
GTAGT	22183975	1.4901665	10.315913	36
TTATT	37105780	1.4900175	8.201964	32
AGGTC	1733780	1.4820569	37.21961	41
GGTTT	54441430	1.4639473	9.4262705	2
AAAAC	182795	1.456089	30.820559	6
AGGTT	21668365	1.4555314	14.559642	41
AACGG	679095	1.4501096	9.345504	8
AAGTA	4467695	1.4479551	12.87154	34
TATAG	11049350	1.4335366	18.757551	47
GCGTC	328075	1.4286304	10.373025	40
TAATT	14184230	1.4228368	16.298695	23
TTTAA	14044625	1.4088328	9.06822	5
TTATA	14035075	1.4078748	14.755614	46
CGAAA	341185	1.4071388	5.482076	32
TCGAA	851480	1.405795	5.3275166	32
CGTAC	166265	1.3983771	9.037077	13
GTACG	1630515	1.3937846	10.7354	4
CGTCT	407960	1.3735408	14.819217	47
TTAAG	10543245	1.3678749	10.803401	6
ACGGT	1587055	1.3566345	10.391065	6
GCGAT	1585725	1.3554974	23.583864	10
GTTTA	25812750	1.3406239	9.07945	4
TAAGC	807850	1.3337617	41.368366	7
ACGTA	804920	1.3289243	5.162125	26
GGAAT	7890220	1.323988	9.574495	2
TCGGG	2965750	1.3125888	30.306917	36
GAGTA	7778560	1.3052512	17.15693	34
TCGAC	155050	1.304053	7.2682757	23
TATTC	2546355	1.3012003	34.038483	33
ATTAT	12846125	1.2886099	14.715814	45
GGTAG	14772405	1.2834219	7.274047	2
GGCGA	1159660	1.2821071	8.215714	2
GGAGT	14742910	1.2808595	10.234715	2
GGGTT	36599910	1.2729131	14.639977	2
GACGT	1481195	1.266144	6.025067	3
GAACG	592815	1.265871	7.4849005	28
TTGTA	24340165	1.2641428	14.482322	20
AACTC	76790	1.2473973	36.305164	22
TGGAA	7424625	1.2458606	9.06147	1
GGTTA	18371370	1.234062	15.707994	2
TTTGT	58495205	1.2161674	6.9344296	19
GGGAT	13902650	1.207858	12.30872	42
GATTA	9291645	1.2054932	18.357111	44
GTTAA	9088340	1.1791166	19.066658	3
GGGGA	10466280	1.1760706	9.683993	2
TGAGG	13471565	1.1704053	16.7845	45
CGTAA	706210	1.1659539	10.683342	21
GGTGG	25624565	1.1526515	10.716434	8
CGTAT	1730325	1.143604	5.5500097	13
ATTTA	11325145	1.1360385	5.1148715	34
GTAAT	8748395	1.1350123	19.851301	22
TTTTC	5521435	1.1294779	12.087579	29
GGATT	16685365	1.1208078	9.790568	43
GTGGC	2527980	1.1188394	34.695	9
TCGTG	3246545	1.1109456	5.803964	40
TAGGC	1282120	1.0959721	9.789883	13
CGTGA	1282030	1.095895	7.8821545	26
GTTGA	16090510	1.0808495	13.649194	43
GGGGT	23924100	1.0761608	7.946309	2
GTTAT	20654855	1.0727408	9.77079	31
TCGAT	1582035	1.0455966	6.4193983	11
TGGAG	12023400	1.0445892	9.820357	1
TGTAG	15543425	1.0441002	8.95619	21
GGGTA	11974675	1.0403559	14.642784	2
AGTTG	15455085	1.038166	10.354758	38
AGTAA	3193455	1.0349811	7.6416903	9
GTATT	19914930	1.0343118	5.034912	31
ATTTC	1988875	1.0163251	5.7449393	22
CGAAC	48275	1.0142505	5.277126	20
AGTAT	7758075	1.0065286	11.368943	30
TGTAA	7741160	1.0043341	19.02403	21
GGTTG	28532625	0.9923399	7.187233	42
TAAGT	7619385	0.9885349	6.904269	7
TTAAT	9817295	0.98478436	12.928003	4
CGTGT	2780020	0.9513038	5.5590076	41
GTTTG	34733060	0.933983	6.7161565	18
TGGGG	20707930	0.93149006	8.440454	1
TAGAC	562455	0.9286141	12.599359	25
TCACA	56585	0.9191819	36.424637	30
TTGGG	26307135	0.9149393	5.682734	40
TTATC	1788000	0.91367704	10.542745	37
GGGGG	15550940	0.90473413	5.6058793	2
GTGGT	25854660	0.8992026	7.17897	9
AGTGA	5339930	0.89604634	5.470016	18
AAGAC	215465	0.88863564	9.679153	32
GGATA	5293050	0.88817984	7.1296163	2
GGAGC	802415	0.88714105	9.548287	27

TGGTT	32714760	0.87971026	7.5264034	1
CGTGC	201935	0.879343	5.018675	13
GGGTG	19291250	0.86776453	8.134491	2
AAGGC	395580	0.84470415	11.222659	46
GAAGC	390050	0.83289564	12.183275	4
GGTAT	12167815	0.8173499	5.939018	2
TGGTG	22052590	0.7669699	6.172198	7
TGGGT	22030825	0.76621294	9.015159	1
GGTAA	4435780	0.744329	5.8958917	2
GGTAC	865565	0.73989576	10.766254	3
GTTGG	21241085	0.7387465	5.5079074	39
GGAAC	344570	0.73577964	6.5852647	27
GAGTC	780405	0.66709995	13.26337	21
TGGTA	9864550	0.6626324	5.15017	1
CATAC	33680	0.5471069	5.8704805	47
TGGGC	1157160	0.5121386	5.865011	13
GATTC	760880	0.50287986	5.4932375	29
TGGAT	7471260	0.5018678	5.070733	1
TCCAG	58930	0.49563265	18.965193	25
CAGTC	56500	0.47519505	19.028662	27
CCAGT	55620	0.4677938	18.915901	26
GTCAC	55475	0.46657425	19.153374	29
ATCTC	60055	0.39052588	15.136015	42
CAGAA	61765	0.2547355	9.241082	38
CTCGT	69340	0.23345749	7.928534	44
TCTCG	68310	0.22998962	7.8683896	43

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	268560	0.3501822190179848	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	171860	0.2240926279432189	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	160876	0.20977031079363018	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	99538	0.12979013150361993	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	91324	0.11907968785224324	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	81321	0.10603652156971083	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	80170	0.10453570337604946	No Hit