FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD2_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`

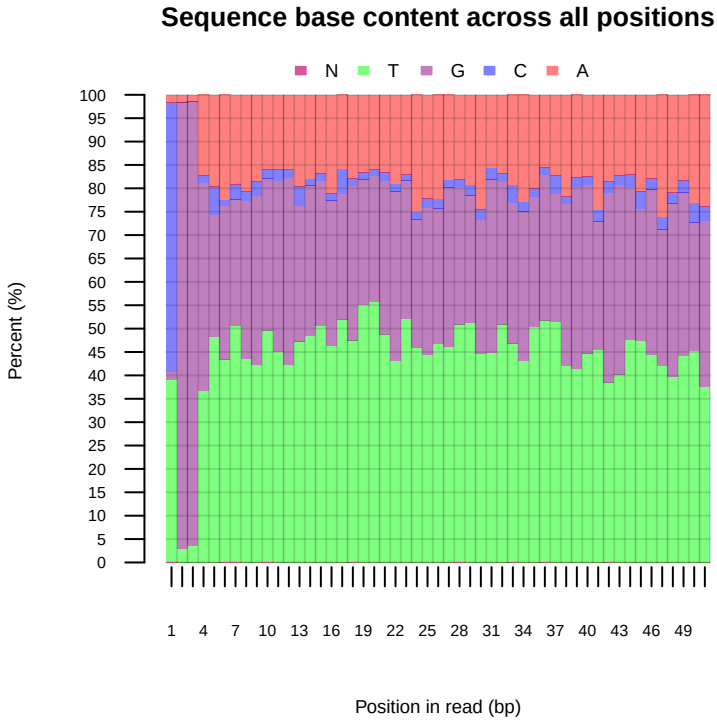
1 Sequence Duplication

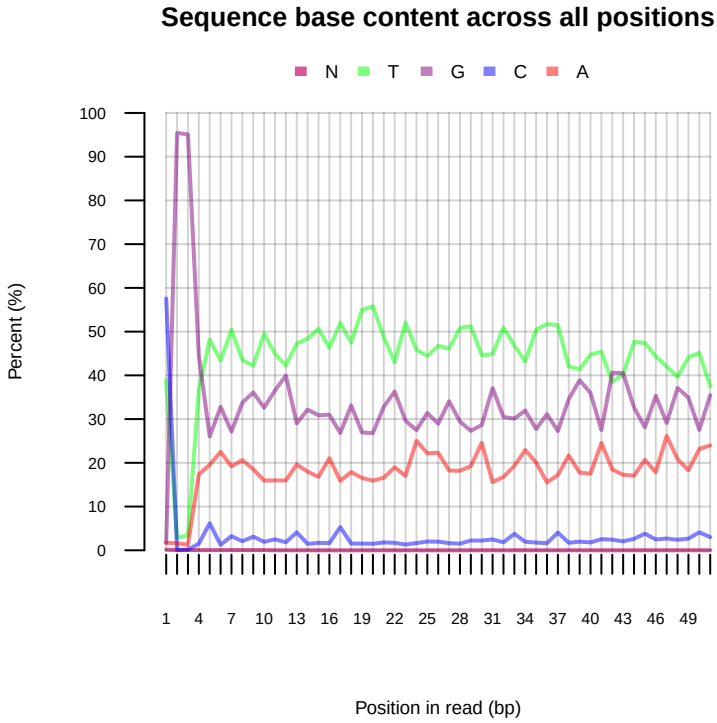
- Estimated Duplication rate 73.4601%

2 Per base sequence quality

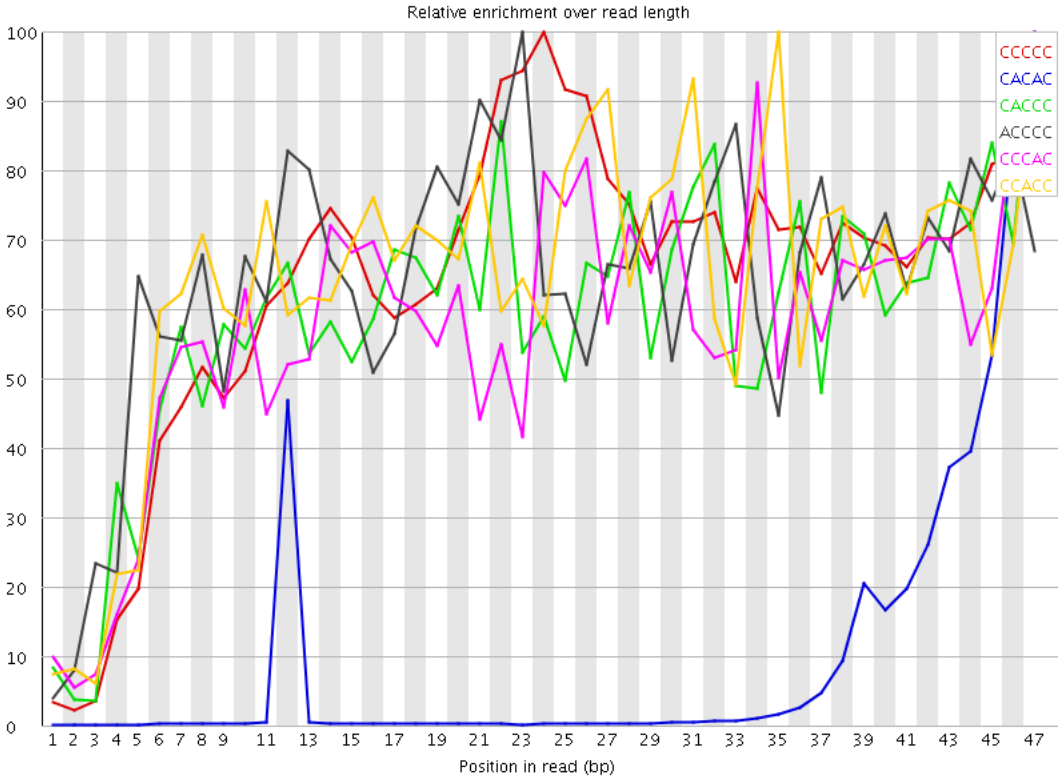
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	140170	820.1324	1270.2867	24
CACAC	1018330	205.21068	1991.4937	47
CACCC	52095	56.567394	95.42466	47
ACCCC	51570	55.997322	87.000786	23
CCCAC	50420	54.74859	94.40407	47
CCACC	50150	54.455414	84.70815	35
CCCCA	48410	52.566032	83.939606	24
AGCAC	1427705	28.469034	205.97261	45
CGGGC	4902455	27.79144	959.99164	1
GCACA	1225850	24.443962	205.3176	46
GCCCC	31880	18.45738	41.764664	47
CCCGC	30160	17.46156	31.832487	26
CCGCC	29580	17.125761	29.248741	35
CCCCG	29270	16.946283	32.104404	25
CGCCC	29155	16.879702	27.887157	23
ACACG	783880	15.630894	171.10396	47
CGGAA	6382880	12.594291	189.26236	1
CGCGG	1938025	10.986435	245.06212	5
GCGCG	1835000	10.402399	243.03522	4
CGCGC	170660	9.777016	64.08829	13
ACTCC	117255	9.680423	364.70572	23
CTCCA	115240	9.514069	363.71835	24
CCTCC	20750	9.230811	14.111171	28
CGGCG	1622540	9.1979885	248.03975	1
CACCT	106985	8.832546	362.79764	31
TCACC	106950	8.829656	364.0768	30
CTCCC	19340	8.60356	14.52966	47
TCCCC	19010	8.456758	15.478038	5
AGATC	5455205	8.270621	40.275726	43
CCCTC	18465	8.214309	13.6930895	39
CCCCT	17575	7.818385	12.856975	38
GGCGC	1355445	7.683858	243.35785	3
CGGGA	7073930	7.4421673	212.18758	1
TCGCG	1408950	6.137098	22.954617	30
CGGGT	13460410	5.80161	223.8665	1
ACACC	26995	5.439947	8.286447	47
AACCC	26840	5.408712	8.854258	22
CCACA	26825	5.4056897	8.144358	35
ACGTC	654540	5.3471465	38.31116	15
AGACG	2690770	5.309256	59.5674	27
ACCAC	25575	5.1537943	8.4281225	20
CGGAG	4891665	5.146303	149.74939	1
CACCA	25130	5.064119	7.481392	32

CGCGT	1143200	4.9795446	20.597605	31
ACGCG	466855	4.9636126	16.068132	14
CACGT	601370	4.9127836	43.6607	47
ACCCA	23835	4.8031545	8.428428	33
CAACC	23805	4.7971087	8.144289	31
GATCG	5855925	4.733738	22.510624	44
CCCAA	23050	4.6449633	8.096754	14
CGGTT	13960440	4.623378	157.06483	1
CGGAC	430545	4.5775633	149.64883	1
CCAAC	22670	4.568387	7.386775	47
CGTCG	1044740	4.5506735	23.10706	41
AGAGC	2297125	4.532541	27.210722	47
CGCGA	422230	4.4891586	16.420275	24
CGGTC	1006625	4.384652	161.74388	1
GGGCG	7715665	4.328064	99.94332	2
CGGGG	7665590	4.2999744	103.88723	1
AAAAA	3266595	4.2075605	11.3898325	31
TCGAG	5005215	4.046052	42.902172	32
AGGCG	3744935	3.93988	48.868755	47
ATCGG	4800395	3.880482	21.100908	45
GAGAC	1929750	3.8076599	57.3129	26
TCGGA	4700550	3.7997708	21.316029	46
CGACG	345495	3.6733103	29.493044	24
GAGCA	1710860	3.37576	21.69625	47
CGTTT	13258510	3.3738446	37.575424	17
TTACG	5402065	3.355355	35.886856	14
CGGTA	4131210	3.3395348	114.96962	1
GGCGT	7550315	3.254283	50.012245	3
GGCGG	5763790	3.2331693	41.72596	11
TACGT	4967065	3.085166	37.93652	15
GACGG	2932490	3.0851429	31.779623	28
ACGTT	4916380	3.0536845	39.657036	16
CGGAT	3657220	2.9563768	99.97713	1
CGAGG	2801770	2.9476182	52.071144	45
ACGGG	2791860	2.9371922	31.794909	29
TTTCG	11008385	2.8012636	15.724315	30
AGAGA	7580540	2.775866	20.25815	25
AAGCG	1382850	2.7285516	51.327187	8
ATCGC	331075	2.704657	32.86245	29
AGCGA	1322950	2.6103609	51.89615	9
CGAGA	1322050	2.6085851	34.064617	25
CACGC	24240	2.6045074	10.502803	47
TTCGA	4181195	2.5970428	34.307808	31
TTTTT	171897675	2.5554252	5.594315	16
GCGGG	4473705	2.5095024	42.543495	12
GCGGC	435720	2.4700453	9.232497	9
CGTTA	3960120	2.459728	32.143383	9
GAAGA	6657635	2.437914	9.551685	46
GGAGG	23313225	2.4269671	27.702938	39
TCGTT	9536080	2.4266114	6.232021	4
GTCGA	3000590	2.4255786	41.53108	43
GGAAG	12329035	2.4071825	10.921992	2
CGCAC	22265	2.3923001	9.669647	47
CGTTC	711295	2.3806043	27.78016	33
ATTCG	3821115	2.3733883	44.14994	34
TTCGC	704465	2.3577452	8.914816	33
TTTTA	62436980	2.2656012	12.27892	26
AGAAA	3286115	2.2568333	5.3007593	22
TTCGT	8819145	2.2441754	5.6739573	35
TTTAG	47003290	2.2197282	15.286648	27
GAGGC	2100430	2.2097692	40.58201	46
GGGAG	21106445	2.1972356	23.963009	38
AAGAG	5960620	2.1826785	9.5595	47
GAGAT	14504440	2.1759598	8.562068	26
GCGGA	2054475	2.1614218	23.632513	7
CGTAG	2670895	2.159064	22.731813	5
CGGTG	4998510	2.1544225	43.15141	1
AGTAG	14331825	2.1500642	21.02496	35
CGAGT	2654875	2.146114	41.958084	33
GAGGT	26249615	2.0996866	21.49763	40
GCGTT	6318520	2.092549	27.68174	16
GCACC	19205	2.0635135	7.422654	47
ACGGA	1029580	2.031502	10.872528	30
AGGAG	10305240	2.0120466	9.540329	38
AACTC	129180	1.9792447	69.624146	22
ATTTT	54258930	1.9688506	7.4016438	25
TAGTT	40448835	1.9101944	10.009993	29
GGTCG	4410145	1.9008297	23.841372	42
TTTAC	3935610	1.8782818	26.723942	13
TACGC	229405	1.8740826	11.256042	13
GACGC	174165	1.851726	11.4275875	5
GTCGC	424730	1.850037	9.096173	3
AAACG	498515	1.8448188	11.839122	7
GCGGT	4275175	1.8426555	23.968328	6
TCGTC	548490	1.8357189	9.337768	40
TAGAG	12127395	1.819355	9.177444	24
AATTT	20489945	1.814808	16.013092	24
AGCGC	169955	1.8069652	7.2898426	35
ACGGC	169180	1.7987255	8.551452	12
ATCGT	2875150	1.7858262	14.207823	39
CGAGC	167910	1.7852228	6.230548	32
TTAGT	37714915	1.7810851	14.704779	28
AGGTA	11729025	1.7595916	27.943792	47
GCGTA	2169145	1.7534657	22.315098	4
ACGCC	16270	1.7481574	5.099681	23
GAAAA	2536560	1.7420549	5.0548806	3
AACGC	86730	1.7294323	7.065713	11
GGAAA	4687455	1.7164671	11.233929	2
TGAGA	11268465	1.6904984	5.5697	41
AGTTA	14594030	1.6822681	21.401072	30
GGAGA	8609990	1.6810576	10.18091	2
AGTTT	35543045	1.6785188	8.514964	26
AGCGG	1594535	1.6775397	6.8655357	6
TGGCG	3888865	1.6761513	36.402027	10
GAGCG	1591005	1.6738257	9.737158	28
TATCG	2672615	1.6600267	14.582143	38
TACGG	2039640	1.6487781	11.503589	5

TAGTA	14299960	1.6483704	12.548658	29
CACCG	15325	1.6466203	6.9176235	31
TAGCG	2026830	1.638423	5.3759437	10
GGACG	1552760	1.6335899	15.582342	2
GCGTG	3736015	1.6102709	34.543003	4
GTAGA	10697745	1.6048788	8.8556795	23
TATTT	43948470	1.5947232	5.088722	33
CGTGG	3686485	1.5889229	34.33882	5
AGTCG	1961545	1.5856488	14.256618	22
TTGAG	25656675	1.576892	13.954586	44
CGATT	2512185	1.5603799	18.484795	11
TGGGA	19413795	1.5528947	12.690396	37
GCGAC	145685	1.5489259	19.397953	23
AGGTC	1898485	1.5346731	39.50079	41
AACGG	776195	1.5315386	9.643877	29
TTCGG	4621220	1.5304422	24.11292	35
GTCGT	4597160	1.5224743	10.171003	3
GCGTC	348150	1.51647	10.44658	40
TAGGA	10072785	1.511122	7.558782	37
GGGAA	7648675	1.4933655	13.214151	2
AGCGT	1843440	1.4901766	7.8711615	29
GGAAT	9834700	1.4754043	9.83611	2
GGTTT	58124020	1.4635543	9.023414	2
AGGTT	23714930	1.4575498	14.14423	41
GTAGT	23639870	1.4529366	9.246161	36
TTATT	40036730	1.452781	7.634717	32
CGTAC	177615	1.4509935	8.749154	13
TAATT	16141895	1.4296983	15.671857	23
GTACG	1757370	1.4206004	11.205727	4
GAACG	714820	1.4104373	9.539035	28
AAGTA	4968850	1.3980548	10.814869	34
TTTAA	15771615	1.3969027	8.078811	5
TGGAA	9273430	1.3912026	8.809676	1
TATAG	12019670	1.385519	16.335312	47
ACGGT	1690720	1.3667227	10.923577	6
TTATA	15322010	1.3570806	12.805204	46
TTAAG	11707390	1.3495222	9.614801	6
CGTCT	402690	1.3477468	15.108863	16
GTTTA	28277970	1.3354259	8.073595	4
GCGAT	1648145	1.3323064	22.056189	10
TCGGG	3078695	1.3269575	30.063995	36
TCGAC	160265	1.3092558	6.783456	23
AAAAC	187465	1.3011101	19.534058	6
GGCGA	1219600	1.2830869	7.8660035	2
GGTAG	16038965	1.2829447	7.1423564	2
GGAGT	15976030	1.2779104	9.870063	2
TATTC	2675815	1.2770407	32.430904	33
TTGTA	26969765	1.273646	14.053707	20
GGGTT	38630220	1.2659318	13.80868	2
TAAGC	834665	1.2654333	37.548786	7
GACGT	1561535	1.2622938	5.7869267	3
GAGTA	8385775	1.2580363	15.122926	34
GGTTA	20400355	1.2538319	15.812704	2
ATTAT	14086770	1.2476746	12.701314	45
TTTGT	62842335	1.2158375	6.857234	19
GTTAA	10436255	1.2029973	19.095083	3
CGTAT	1906710	1.1843045	5.257467	13
GGGGA	11317675	1.1781993	9.601973	2
GGGAT	14660130	1.1726526	11.042184	42
GGTGG	27386275	1.1680089	10.765228	8
TCGTG	3525675	1.1676229	6.4307146	40
GATTA	10074225	1.1612657	15.905321	44
CGTAA	763960	1.1582378	9.641259	21
GTAAT	10029315	1.156089	19.24268	22
TGAGG	14441710	1.1551815	16.14202	45
CGAAC	56340	1.1234431	5.034455	9
GTGGC	2581605	1.1127052	34.348198	9
TTTTC	5629185	1.1006414	11.217632	29
GGATT	17835850	1.096214	8.746812	43
GGGGT	25498560	1.0874989	7.9044714	2
CGTGA	1343795	1.0862799	7.874637	26
TAGGC	1334535	1.0787944	9.562489	13
TCGAT	1732135	1.0758716	6.556437	11
TGGAG	13200085	1.0558648	9.653116	1
GTTAT	22312400	1.0537021	9.096489	31
GTATT	22163945	1.0466914	5.2640786	31
AGTAA	3711590	1.0443074	6.6644106	9
TGTAG	16858745	1.03616	8.495221	21
GGGTA	12945220	1.0354782	14.020316	2
GTTGA	16829335	1.0343524	13.029322	43
TGTAA	8963770	1.0332625	18.54331	21
AGTAT	8855535	1.020786	11.72342	30
TTAAT	11384440	1.0083274	13.159544	4
AGTTG	16365390	1.0058378	9.227712	38
ATTTC	2093655	0.99920315	5.6729975	22
CGTGT	3006495	0.99568236	6.1212735	41
GGTTG	30224220	0.990463	7.0798645	42
TAAGT	8388430	0.9669424	6.1785135	7
TGGGG	22176345	0.94580835	8.457591	1
CGTGC	216065	0.9411349	5.2135353	13
TCCAG	114690	0.9369392	37.23626	25
TTATC	1960635	0.93571895	10.829246	37
GTTTG	37070640	0.93343323	6.697923	18
TTGGG	28295950	0.9272727	5.8308744	36
ATGCC	112920	0.92247945	40.055737	47
TGCGG	2129620	0.9178938	5.3747597	5
CAGTC	111335	0.9095312	37.561092	27
GTGGT	27649155	0.90607685	7.3192625	9
TGGTT	35485855	0.89352864	7.3802943	1
CCAGT	109075	0.89106846	37.17502	26
GGGGG	16051045	0.89093596	5.537226	2
AAGAC	239825	0.8875032	8.400622	32
GGATA	5915090	0.88738346	6.9241486	2
GTCAC	108590	0.8871063	37.434784	29
GGAGC	840225	0.88396347	8.84449	27
AGTGA	5850740	0.87772965	5.3429403	18
GGGTG	20544830	0.8762251	8.127147	2

AAGGC	441820	0.8717711	12.284485	46
TAGAC	574995	0.8717483	10.233716	25
CACAT	56220	0.86138064	6.9375296	47
CGGCC	15035	0.8613468	5.1097245	1
GGAAC	434860	0.85803807	8.716536	27
GAAGC	412960	0.81482637	10.997746	4
GGTAT	13221920	0.8126361	5.7937303	2
TGGGT	23619210	0.7740135	8.945494	1
TGGTG	23445430	0.76831865	6.07949	7
GGTAC	939410	0.75938827	11.17796	3
GGTAA	5006500	0.7510765	5.799973	2
GTTGG	22729140	0.74484545	5.071338	39
ATCTC	117810	0.7394997	30.616102	40
GAGTC	834185	0.67432785	13.102928	21
TGGTA	10829245	0.6655792	5.1059427	1
ACCTT	105145	0.6600008	27.951292	32
TGGGC	1229395	0.5298852	5.9296703	13
GATTC	827180	0.51378185	5.5227027	29
TGAAC	308815	0.46819365	7.3867173	20
TAATC	379200	0.44173864	5.726529	38
TCTCG	123470	0.4132367	16.369886	41
CTCGT	123115	0.41204858	16.395958	42
CTGAA	242225	0.3672367	7.1109705	19
CCTTG	104220	0.3488097	15.004826	33
GAACT	158220	0.23987694	7.2000422	21
AGTCA	119895	0.1817725	7.1610036	28
AATCT	123515	0.14388542	5.622743	39

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	254937	0.30604618679538864	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	174832	0.20988191957154664	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	156491	0.1878639578319238	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	101874	0.12229746656465487	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	89729	0.10771766473663463	No Hit