FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD7_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.

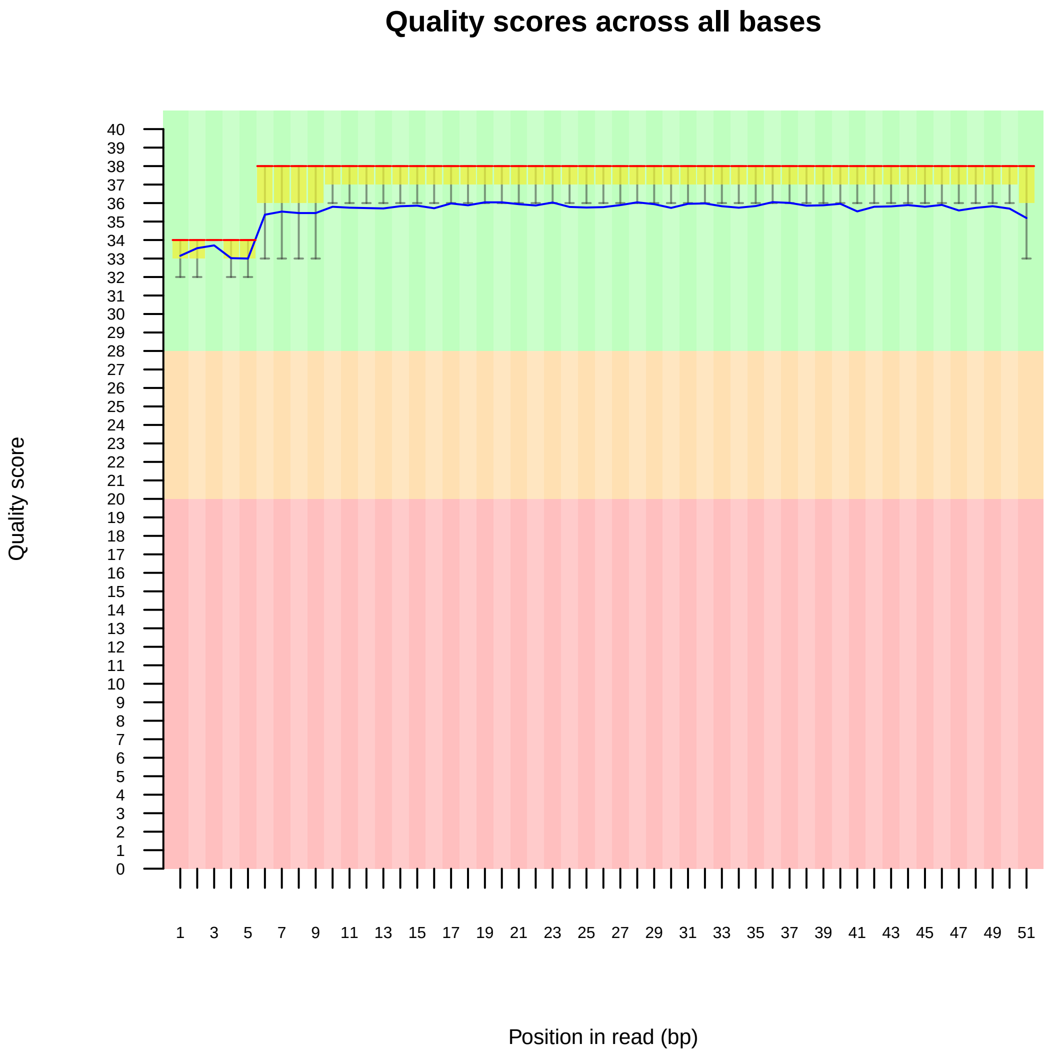
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

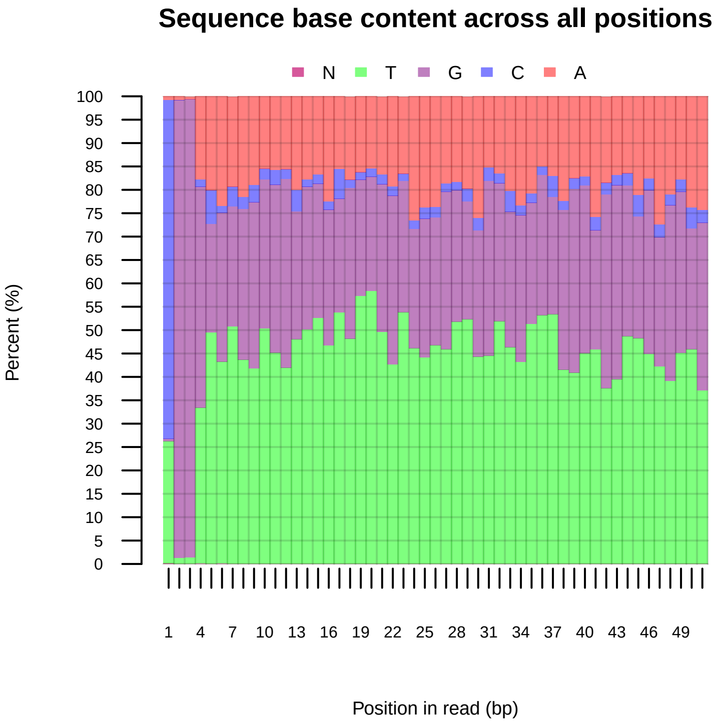
- Estimated Duplication rate 77.5289%

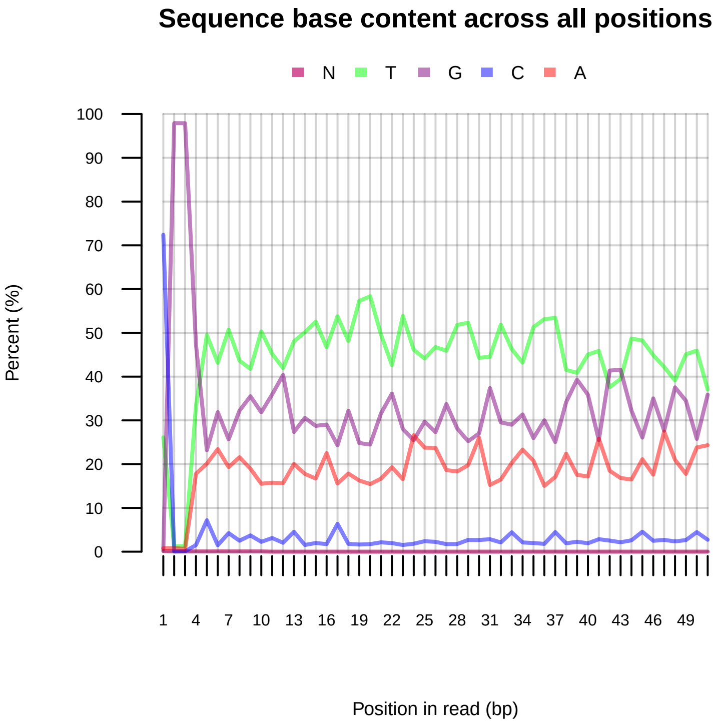
2 Per base sequence quality



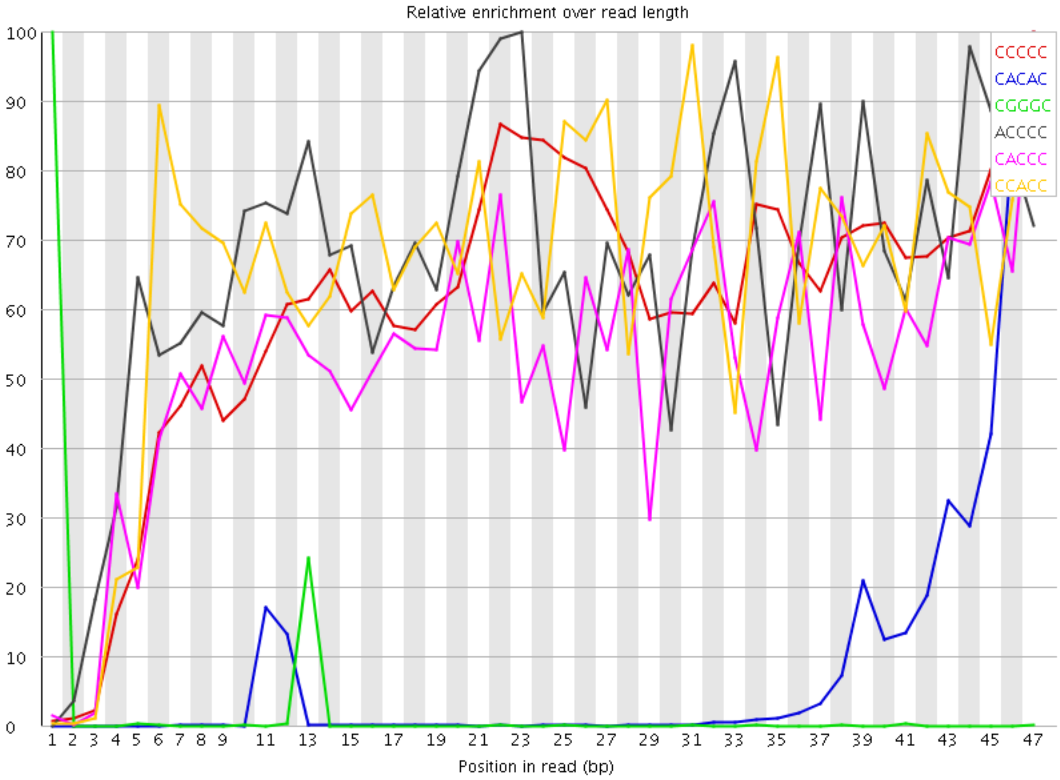
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	106685	335.40988	551.83344	47
CACAC	633010	91.03729	1051.8673	47
CGGGC	4976940	25.925076	925.4046	1
ACCCC	37560	25.256094	38.076065	23
CACCC	36945	24.842556	46.609833	47
CCACC	35080	23.588493	35.86587	47
CCCAC	35040	23.561598	41.86985	47
CCCCA	34345	23.094267	36.496143	23
AGCAC	884885	15.05881	129.73747	45
GCACA	767325	13.058195	129.59015	46
CGCGG	2167780	11.292051	265.75214	5
CGGAA	5409420	10.89305	221.93773	1
GCGCG	2050270	10.679936	264.21408	4
CACCA	59990	8.627553	177.03741	30
CGGCG	1656210	8.627263	228.30333	1
CGGGA	7599585	8.466728	260.23837	1
CGCGC	192325	8.466392	48.531376	13
GCCCC	22375	8.323972	19.668219	47
ACACG	475430	8.09078	106.10149	47
CCGCC	21580	8.028214	14.073049	22
GGCGC	1530975	7.974909	264.68866	3
CCCGC	21410	7.9649715	15.384846	34
CGCCC	20925	7.784541	14.510087	23
CCCCG	20770	7.7268777	15.210123	46
CGGGT	14753745	6.803195	268.4497	1
AGACG	2978850	5.998566	67.604546	27
TCGCG	1502430	5.85477	23.711346	30
CGGAG	5255065	5.854689	174.42511	1
AGATC	3789115	5.70814	30.169325	43
CGGTT	14991285	5.171392	178.08116	1
CGGGG	8357620	5.151519	127.31087	1
GGGCG	7734965	4.7677226	113.42871	2
CGCGT	1199075	4.672636	21.464447	31
ACGCG	491230	4.6250405	16.457367	6
TCGAG	5520950	4.6014805	55.502163	44
AGGCG	4064155	4.527891	64.40531	47
GAGAC	2165560	4.360829	65.00091	26
CGTCG	1116395	4.3504434	19.063528	41
AAAAA	3085910	4.342066	13.754245	31
CGGAC	446385	4.202815	135.79034	1
CGCGA	438725	4.1306953	15.926156	5
CGGTC	1047545	4.0821433	149.17787	1
TTACG	5893590	3.6746979	46.062057	14

GACGG	3177940	3.5405557	37.319828	28
GGCGG	5739845	3.5379593	41.245586	11
CGTTT	13703240	3.5363064	40.023495	17
GATCG	4222180	3.51901	17.810654	44
CGAGG	3143415	3.5020912	70.37692	45
CGGTA	4168525	3.4742908	119.94201	1
CGACG	366825	3.4537396	23.993696	24
ACTCC	57345	3.413416	72.124954	22
ACGGG	3024395	3.3694906	37.192295	29
TACGT	5390455	3.3609893	47.81514	15
GGCGT	7251260	3.3436756	51.895718	3
ACGTT	5361725	3.343076	49.51438	16
CGGAT	3968865	3.307882	115.28248	1
CTCCA	54790	3.2613316	71.285736	23
ACGTC	462355	3.2565987	17.10634	46
AGAGA	7205490	3.1033483	26.20516	25
TCCCC	10985	3.057212	8.244966	3
CCTCC	10945	3.0460796	6.3430953	28
TCACC	51030	3.0375204	72.210815	29
AGAGC	1505350	3.0313516	16.948814	47
CTCCC	10850	3.0196404	6.539311	29
TTTCG	11538410	2.9776425	16.956888	30
AAGCG	1451900	2.9237182	58.66966	8
CGAGA	1415880	2.8511841	38.44556	25
GTCGA	3405035	2.8379536	54.98909	43
AGCGA	1373090	2.765017	59.542427	9
TCGGA	3300360	2.7507117	16.347092	46
TTCGA	4408375	2.7486553	37.12733	31
TTTTT	158774460	2.7134275	6.4790244	16
ATCGG	3241955	2.7020335	16.04699	45
GGAGG	20324510	2.6794205	36.221024	39
GAGGC	2379245	2.6507263	53.084908	46
GCGGG	4289090	2.6437345	41.95774	12
CGTTA	4083570	2.5461369	33.447533	9
TCGTT	9803605	2.5299528	6.498759	4
ATCGC	351175	2.4735022	33.45892	29
TTTTA	59444490	2.454507	14.279488	26
GGGAG	18610860	2.4535067	31.82609	38
GCGGC	467245	2.4338975	8.472281	33
CACGT	344930	2.4295154	22.001854	47
AGTAG	13469540	2.401069	26.449728	35
TTTAG	43355400	2.3929734	18.255644	27
TTCGT	9121280	2.3538697	5.9982924	35
GCGGA	2100955	2.3406823	26.769964	7
GAGGT	23690600	2.336441	27.407251	40
ACCAC	16135	2.3204794	5.105463	5
ATTCG	3720175	2.3195572	40.824116	34
CGAGT	2767920	2.306945	46.098087	33
GGAAG	9651070	2.2996902	12.636573	2
CGGTG	4917460	2.2675219	37.84373	1
CGTTC	771925	2.2503428	25.34329	33
GGTCG	4820040	2.2226	32.26344	42
CGTAG	2648700	2.2075803	24.445845	5
TTCGC	745590	2.1735702	7.693817	33
ATTTT	52569170	2.17062	9.334477	25
GCGTT	6225080	2.1474028	26.269579	16
GAGAT	11959320	2.1318583	10.961603	26
GAGCA	1053040	2.1205263	15.867068	44
GAAGA	4918135	2.1182022	8.355118	46
ACGGA	1049660	2.11372	11.312749	30
GCGGT	4552225	2.099106	30.714449	6
AGGAG	8590830	2.0470526	10.7460985	38
TTTAC	4358685	2.033083	33.516743	13
TAGTT	36561955	2.0180137	10.557935	29
TAGAG	11230465	2.0019333	11.760666	24
AATTT	19847765	1.9800649	20.440058	24
TTAGT	35327265	1.9498658	17.623129	28
AAGAG	4433595	1.9095147	8.338502	47
ATCGT	3032190	1.8905933	15.650884	39
AGGTA	10550125	1.8806564	31.664322	47
TAGTA	14003255	1.8674062	17.934944	29
AAACG	509870	1.855801	12.67341	7
TAAAA	3169140	1.845608	5.6706934	30
AGGTC	2212585	1.8440967	52.241768	41
TACGG	2179160	1.8162385	14.9311695	5
AAAAT	3094785	1.802306	5.3557134	32
GACGC	190770	1.7961423	14.6976595	5
GCGTA	2151290	1.79301	24.176544	4
GGACG	1595270	1.7772967	16.565384	2
GTCGC	456015	1.7770298	8.412899	3
GTAGA	9921510	1.7686	11.4765005	23
GTTTT	76671600	1.7515161	5.184611	27
GAGCG	1571640	1.7509705	10.946947	28
TCGTC	597995	1.743296	8.618414	40
AGTTA	13064230	1.7421825	22.506495	30
TATCG	2793465	1.7417465	16.066782	38
AGCGG	1559945	1.737941	6.720433	6
AGTTT	31289185	1.7269866	10.286219	26
TATTT	41764255	1.724477	6.711742	32
GGAGA	7117570	1.6959991	11.151008	2
AGTCG	2029890	1.6918283	16.146353	22
TACGC	239725	1.6885037	9.057662	13
TAGCG	2005010	1.6710918	5.231715	10
TGGCG	3623790	1.6709894	35.2576	10
TGGGA	16813230	1.6581732	17.490301	37
AAATA	2838910	1.6532924	5.0576987	33
CGAGC	174910	1.646817	5.3161106	32
CGATT	2633935	1.6422783	21.09167	11
GGAAA	3808020	1.6400844	12.093752	2
CGTGG	3548245	1.6361543	33.80867	5
GCGTG	3509045	1.6180786	34.20824	4
GTCGT	4675855	1.6129824	10.809692	3
TAATT	16075725	1.6037563	20.115273	23
AGCGC	169895	1.5995996	9.433229	35
ACGGC	168615	1.587548	8.256286	12
TTCGG	4591530	1.5838938	23.112986	35
AACGG	784210	1.5791785	10.116097	29

GGTTT	51182135	1.5629317	9.361356	2
TAGGA	8643315	1.54075	8.312682	37
AGGTT	20867435	1.5395925	16.602417	41
AGCGT	1843580	1.5365465	8.622808	29
GTACG	1842050	1.5352713	14.715398	4
TTATT	36914030	1.5242076	7.8592873	32
ACGGT	1820525	1.5173311	14.2714	6
GTAGT	20429605	1.5072895	11.583376	36
TTGAG	20397815	1.5049442	15.3726425	44
TATAG	11242055	1.499186	20.14523	47
TGAGA	8389195	1.4954509	5.537033	41
GGAAT	8366745	1.4914489	11.579416	2
GGGAA	6220345	1.4822053	14.406777	2
TCGAA	981760	1.4789797	5.533258	32
AAGTA	4559300	1.4690046	13.356286	34
TTATA	14707185	1.4672271	15.399084	46
TTAAG	10969420	1.4628288	12.150311	6
CGAAA	400890	1.4591407	5.820676	32
TTTAA	14570860	1.453627	9.867386	5
GAACG	714425	1.4386511	9.952803	28
GCGAT	1724370	1.4371899	25.652075	10
GTTTA	25933315	1.4313728	10.048795	4
GGTTA	18930735	1.3967035	19.486168	2
GCGTC	355215	1.3842256	7.07232	40
TCGGG	2984495	1.3761998	29.50914	36
AGATA	4231240	1.3633038	6.0428576	26
GGCGA	1218585	1.3576305	8.285312	2
AACGC	79775	1.3575962	6.961741	11
ATTAT	13593340	1.3561071	15.298166	45
TTTTG	59277285	1.3541535	5.07462	17
GGGTT	33167965	1.3538872	16.61446	2
GCGAC	143430	1.3504256	12.77105	23
TAAGC	895110	1.3484452	41.99718	7
TTGTA	24417420	1.3477039	16.81318	20
GTTAA	10103845	1.3473998	24.447855	3
GGTAG	13565325	1.3378547	7.1846795	2
GAGTA	7467145	1.331087	18.882742	34
GACGT	1590585	1.3256857	5.7509584	3
CGTAC	187615	1.3214668	7.70886	13
TGGAA	7357845	1.3116032	7.010705	1
GGAGT	13029955	1.2850548	10.535834	2
GTAAT	9569820	1.2761848	25.286003	22
GATTA	9539025	1.2720782	19.692648	44
TCGTG	3672520	1.2668722	7.651791	40
GGGAT	12815730	1.2639273	14.390635	42
TCGAC	178405	1.2565961	7.218862	23
CACGC	15615	1.2424464	5.5712094	15
AAAAC	181425	1.1935543	20.643219	6
TATTC	2547960	1.188481	28.960064	33
TTTGT	52005475	1.1880335	8.0630865	19
CGTAT	1878515	1.1712682	5.199504	13
TGAGG	11844900	1.1681811	18.805456	45
GGATT	15766115	1.1632189	11.085549	43
GGGGA	8799705	1.1600826	10.347751	2
GTATT	20880885	1.1525072	7.508686	31
CGTAA	762020	1.1479508	9.61065	21
AGTAT	8594220	1.1460836	16.987833	30
TTTTC	5922055	1.1432915	11.78801	29
TCGAT	1816455	1.1325734	6.9307857	11
GGGTA	11369440	1.1212897	17.33063	2
TGTAA	8359195	1.1147418	24.436872	21
GGTGG	20399385	1.1130706	13.612133	8
GGGGT	20378865	1.1119511	8.339073	2
TAGGC	1326815	1.1058445	8.959161	13
GTGGC	2372450	1.0939758	33.271687	9
CGTGA	1310570	1.0923052	8.455249	26
TTAAT	10920230	1.0894307	16.50059	4
CGAAC	63360	1.0782487	5.1209097	9
GTTAT	19532210	1.0780679	9.593726	31
CGTGT	3097200	1.0684099	7.2956085	41
AGTAA	3266400	1.0524328	8.3557415	9
ATTTC	2188025	1.0205914	5.7908287	22
AGTTG	13651805	1.0072258	11.565219	38
TAAGT	7545935	1.0062895	7.6788177	7
TGTAG	13531215	0.9983287	8.997877	21
TTGGG	24396660	0.9958502	7.8736305	36
GTTGA	13490770	0.9953447	14.674598	43
ACGCC	12410	0.9874326	5.5151043	16
AAGGC	490140	0.9870041	16.438307	46
TGCGG	2139400	0.98651254	6.972131	5
GGTTG	23701545	0.96747625	8.242312	42
TTATC	2063915	0.962701	11.683826	37
GGAGC	857260	0.9550768	10.008709	27
TAGAC	627510	0.94531703	12.334795	25
TGGAG	9540225	0.94088686	7.1070714	1
TTTGG	30061160	0.91796756	6.263076	35
GAATA	2843545	0.916189	5.1756587	3
TAAGG	5132455	0.9149072	5.89537	45
GTTTG	29820875	0.91063017	8.141778	18
GGATA	5105710	0.91013956	7.7707496	2
AGTGA	5029090	0.89648134	6.747005	18
ATTAC	788795	0.8889532	5.4077864	29
CCAGA	52230	0.8888405	21.448757	32
GGGGG	12172940	0.8878573	5.952866	2
ACCAG	51875	0.8827992	21.400764	31
CCAGC	11005	0.8756403	5.440423	28
GGAAC	431625	0.8691714	8.97145	27
CCCAG	10885	0.8660922	5.4217095	27
GTGGT	21079575	0.86044973	9.56596	9
TGGGG	15665470	0.85476965	6.1907334	1
GGTAT	11505290	0.8488565	6.6373706	3
TGGTT	27790845	0.84863967	5.1656165	1
GGTAC	1018100	0.8485436	14.754529	3
AAGAC	229155	0.83406764	5.8762655	32
GAAGC	413900	0.8334782	10.368799	4
GGGTG	14810025	0.8080932	8.866497	2
AACTC	63270	0.805489	15.868696	21

GTGCG	1722025	0.79405415	6.621681	4
CGATC	110615	0.77911705	6.1765056	44
GTTGG	19015215	0.7761844	6.5553994	39
GGTAA	4146735	0.7391935	5.626567	2
GAGTC	873745	0.72822976	14.86482	21
AGTGG	7238060	0.7138401	5.8444457	8
CGTCT	236390	0.68913245	6.888327	47
TGGGT	16713950	0.6822488	6.9266405	1
GGGAC	603635	0.6725122	5.317364	2
TGGTG	16041870	0.654815	6.9306297	7
CGGCC	12695	0.55885005	6.6301985	1
GATTC	840635	0.52414227	5.800076	29
TGGGC	1111285	0.5124318	5.6002927	13
GGTGC	909685	0.41947073	6.6618834	3
TCCAG	55300	0.3895057	8.753044	24
ATGCC	54155	0.3814409	9.418749	46
CAGTC	53120	0.37415087	8.782901	26
CCAGT	52825	0.37207302	8.718298	25
GTCAC	51405	0.36207125	8.778012	28
ATCTC	57195	0.30137327	7.0323253	39
TCATC	54880	0.289175	6.9532156	37
CATCT	51600	0.27189195	6.8219028	38
TGCCG	36460	0.14207977	5.235699	47

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	264839	0.36840518290905094	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	208682	0.29028779892624035	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	166509	0.2316229052405543	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	96046	0.1336051117761459	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	93279	0.1297560671070853	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	89689	0.1247621855162188	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	87109	0.12117326782696099	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	79921	0.11117437621828453	No Hit