FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD8_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.

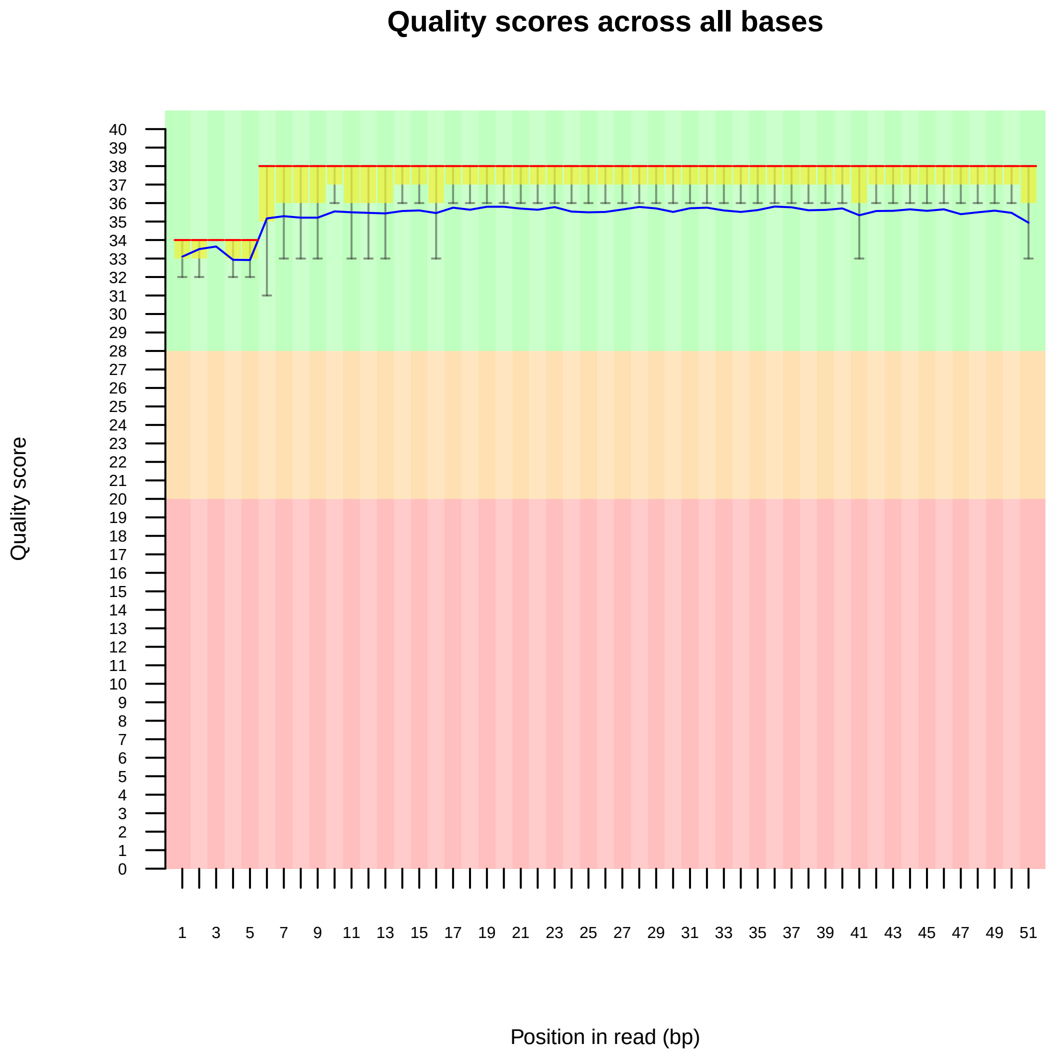
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

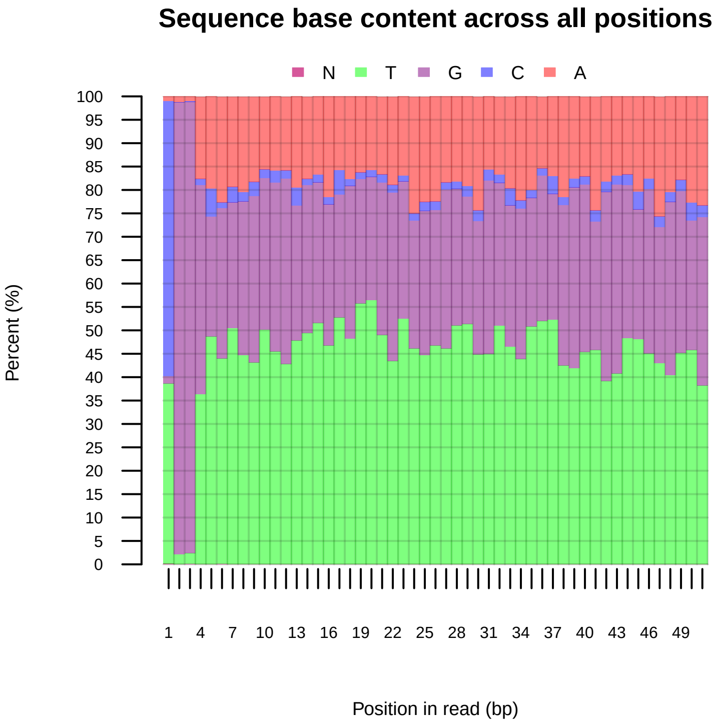
- Estimated Duplication rate 74.3788%

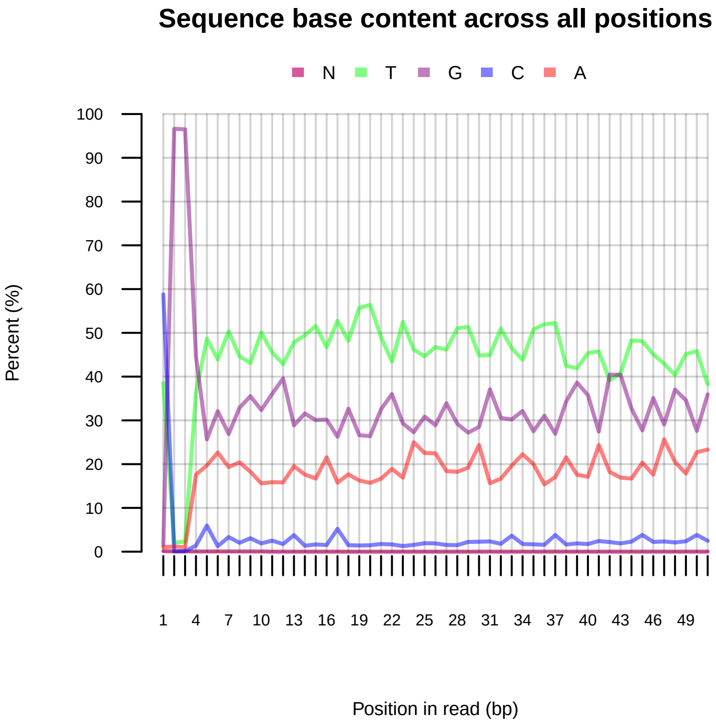
2 Per base sequence quality



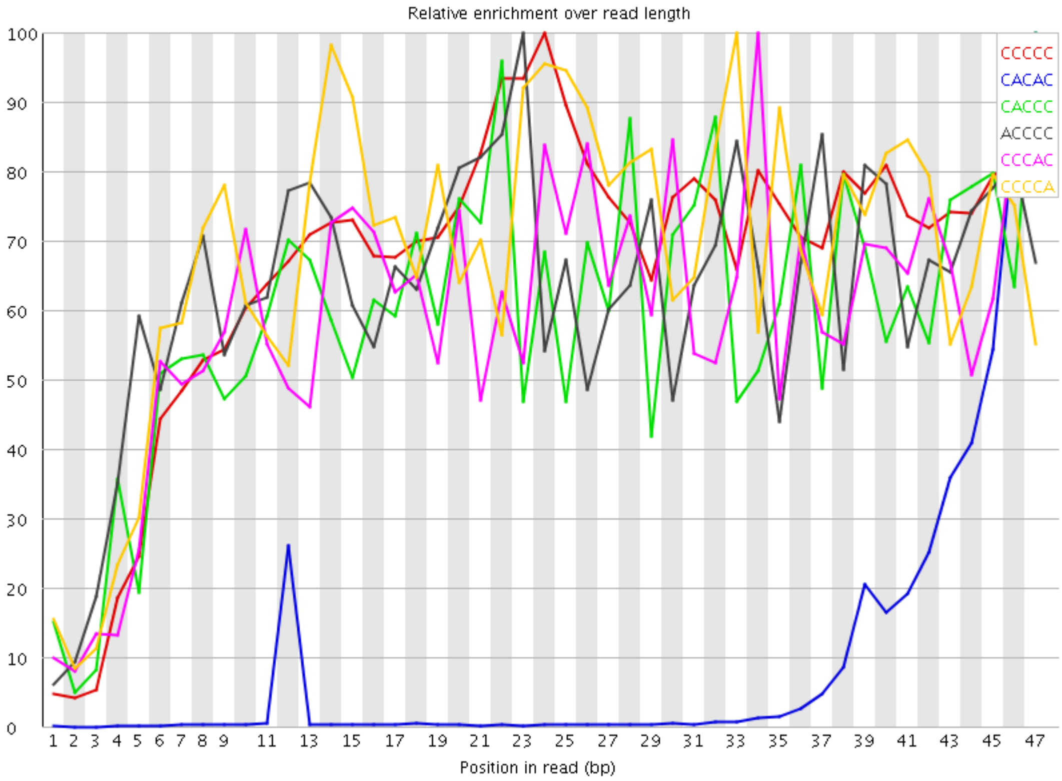
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	137800	944.3281	1405.68	24
CACAC	671915	152.68082	1560.2084	47
CACCC	50425	62.924217	105.26493	47
ACCCC	50220	62.668404	98.80937	23
CCCAC	49660	61.969597	103.79868	34
CCCCA	49205	61.401814	90.01698	33
CCACC	48690	60.759155	101.74634	47
CGGGC	4603510	28.625425	1011.22925	1
AGCAC	951085	20.922771	157.45113	45
GCCCC	31030	20.586628	36.165123	24
CCCGC	28125	18.659327	33.82706	26
CCCCG	27840	18.470243	34.450413	25
CGCCC	27655	18.34751	32.8914	22
CCGCC	27400	18.178331	31.644331	22
GCACA	811780	17.858223	156.64551	46
CGCGG	1894665	11.781356	279.78458	5
ACACG	513735	11.301577	131.49226	47
GCGCG	1796845	11.173094	278.0598	4
CGGAA	5105540	10.8735485	204.8994	1
CTCCC	20745	10.419906	17.585114	29
CCTCC	20615	10.35461	15.342627	28
TCCCC	20550	10.32196	18.775204	5
CCCTC	19060	9.573557	15.342364	24
CGGCG	1518285	9.44096	259.01025	1
CGCGC	146740	9.425007	62.132477	13
CCCCT	18460	9.272184	13.808638	44
GGCGC	1363955	8.481308	278.6107	3
CGGGA	6823770	7.726547	228.83813	1
AGATC	3920705	6.3218	29.830673	43
CGGGT	13278105	6.051675	235.66617	1
AGACG	2824760	6.016046	69.27658	27
TCGCG	1269950	5.9785576	24.04471	30
CCACA	25445	5.781928	8.703099	35
AACCC	25250	5.737617	9.877667	32
ACACC	24290	5.5194736	8.32896	21
CGGAG	4830105	5.4691224	162.49852	1
ACCCA	23905	5.431989	9.343774	33
CACCA	23580	5.3581386	8.062312	32
CCCAA	23550	5.3513227	8.062117	14
ACCAC	23250	5.283152	7.8488417	46
ACGCG	445160	5.2065196	16.23855	14
ACTCC	54905	5.0218134	156.34282	23
CGCGT	1061350	4.9965296	22.24066	31

CAACC	21775	4.947984	8.062306	31
CCAAC	21675	4.925262	8.489536	34
CTCCA	53700	4.9115987	156.70894	24
CGGTT	14200515	4.8999376	170.3337	1
CGGAC	402505	4.7076335	155.76198	1
CGGGG	7631320	4.5940127	109.63269	1
CGTCG	972755	4.5794497	22.137638	41
GAGAC	2072520	4.4139595	66.54424	26
AAAAA	3216880	4.4136696	12.353214	31
CGCGA	374825	4.3838925	17.792622	5
GGGCG	7272295	4.3778815	102.914	2
CGGTC	929070	4.3737936	161.26309	1
TCGAG	5025255	4.307905	50.32247	44
AGGCG	3767385	4.2658057	58.885574	47
ACGTC	473465	4.192434	25.570992	47
CACGT	424870	3.762135	33.51069	47
CGACG	317695	3.7157094	31.679905	24
GATCG	4332400	3.713955	16.998417	44
TTACG	5480230	3.5567536	41.912918	14
AGAGC	1626480	3.4640036	20.165997	47
CGGTA	4037700	3.461323	119.7271	1
GACGG	3003580	3.4009502	36.628216	28
CGTTT	12660600	3.3074076	37.289658	17
CGAGG	2899570	3.2831798	63.217888	45
ACGGG	2887890	3.2699544	36.78514	29
TACGT	5015315	3.2550168	43.751953	15
GGCGG	5361955	3.2278671	38.633015	11
ACGTT	4960870	3.219681	45.431854	16
GGCGT	6985260	3.183626	49.263668	3
CGGAT	3646280	3.1257777	107.362656	1
AGAGA	7725850	2.9962232	23.0658	25
AAGCG	1365115	2.9073603	56.840733	8
ATCGG	3382130	2.899335	15.462369	45
TCGGA	3371650	2.8903503	15.761565	46
AGCGA	1310905	2.791906	57.43985	9
TTTCG	10650935	2.7824101	15.803992	30
CGAGA	1275420	2.716332	34.388863	25
ATCGC	304425	2.69562	35.848595	29
TTCGA	4061955	2.6362715	35.315887	31
GTCGA	3066950	2.6291463	49.69143	43
TTTTT	176071855	2.5523899	5.5634727	16
CGTTA	3838660	2.4913495	32.225056	9
GGAGG	22618550	2.479451	28.271725	39
GCGGG	4118420	2.4792662	39.361057	12
GAGGC	2170155	2.4572637	48.50934	46
GAGCA	1146455	2.441668	16.059935	47
GCGGC	391820	2.4364047	8.767734	9
TCGTT	9057015	2.3660204	5.8223066	36
CGTTC	661295	2.356959	28.98746	33
AGAAA	3219100	2.3481772	5.1381884	22
GGAAG	11062315	2.2808952	12.045783	2
ATTCG	3496450	2.26925	40.26489	34
TTTTA	62932700	2.2664962	11.78218	26
GGGAG	20614335	2.2597485	24.7587	38
CGGTG	4957215	2.2593174	43.720543	1
TTCGC	632515	2.2543826	7.785988	33
AGTAG	14400165	2.2478836	21.80111	35
CACGC	18370	2.2192733	8.430929	47
TTTAG	46650905	2.219176	14.85041	27
TTCGT	8494005	2.218942	5.728219	35
CGAGT	2576505	2.2087119	43.82835	33
CGTAG	2563745	2.1977732	23.433197	5
GCGGA	1936300	2.1924703	23.082441	7
GAAGA	5646340	2.1897519	7.274103	46
AAACG	531720	2.1300075	12.397681	7
GAGGT	25249170	2.0954864	21.596573	40
GCACC	17070	2.0622208	7.068355	47
GACGC	175810	2.0562453	12.862599	5
CGCAC	16920	2.0440996	6.9548078	47
GCGTT	5883155	2.0300033	25.089437	16
AGGAG	9838305	2.0285213	9.141141	38
ACGGA	945545	2.013779	8.29656	30
GAGAT	12853180	2.006397	8.924233	26
GGTCG	4390135	2.000863	28.155869	42
ATTTT	55211535	1.9884214	7.584609	25
TTTAC	4038715	1.984473	30.807003	13
AAGAG	5067780	1.9653759	7.2806516	47
TACGC	216900	1.9206043	11.058248	13
TAAAA	3472005	1.9174485	5.072207	30
TAGAG	12227985	1.9088035	10.169543	24
GCGGT	4181900	1.9059572	26.533	6
ACGGC	160645	1.878878	9.40706	12
AATTT	20845085	1.8651078	16.70396	24
TAGTT	39100090	1.8599851	8.926941	29
ATCGT	2853885	1.8522153	15.448434	39
AGCGC	158020	1.8481764	9.091028	35
CGAGC	157685	1.8442583	6.2988095	32
GAAAA	2509340	1.8304417	5.564568	3
TCGTC	513055	1.828608	8.985945	40
GGAAA	4631425	1.7961496	12.6602335	2
GTCGC	379840	1.7881771	8.761992	3
TTAGT	37508545	1.7842753	14.2846	28
GCGTA	2080340	1.7833738	23.070824	4
TACGG	2051280	1.7584621	12.890627	5
AGGTA	11127575	1.7370281	26.465963	47
TAGTA	14678440	1.7347351	14.328995	29
AGGTC	2018840	1.7306529	47.343765	41
GGAGA	8356025	1.7228959	10.928815	2
AGCGG	1521175	1.7224247	6.274184	6
TATCG	2652870	1.7217535	15.845055	38
GAGCG	1509695	1.7094259	9.285491	28
GGACG	1492075	1.6894748	15.85389	2
GTAGA	10768060	1.6809074	9.866167	23
TAGCG	1953805	1.6749016	5.104785	10
AGTTT	34683240	1.6498762	8.469933	26
AGTTA	13910730	1.6440053	19.25472	30
TGGCG	3537460	1.612245	33.377808	10

AACGC	73185	1.6099856	6.4872794	11
TGGGA	19245295	1.5972112	13.615338	37
AGTCG	1861625	1.5958802	13.940138	22
CGTGG	3489360	1.5903227	32.86248	5
TATTT	44086635	1.5877626	5.4656644	32
CGATT	2444690	1.5866414	19.775778	11
GCGTG	3465045	1.5792409	33.0798	4
GGGAA	7422315	1.5303779	13.964139	2
GCGAC	130350	1.5245525	20.264265	23
TAGGA	9739930	1.5204151	7.14011	37
CGAAA	378500	1.5162263	5.264479	32
AGCGT	1760900	1.5095335	7.4481473	29
TCGAA	933630	1.5053982	5.0082693	32
GTCGT	4342245	1.4983069	9.689995	3
AACGG	699990	1.490807	7.194596	29
GCGTC	316190	1.4885312	9.802771	40
TTCGG	4308800	1.4867666	22.02887	35
GTACG	1731100	1.4839875	12.677143	4
GGTTT	58543300	1.4806087	9.288351	2
TAATT	16511665	1.4773761	16.368578	23
CGTAC	166280	1.4723748	9.266873	13
TTGAG	23155925	1.4549457	12.795671	44
ACGGT	1691475	1.4500189	12.44079	6
AGGTT	23071190	1.4496218	13.489717	41
AAGTA	4900940	1.4389782	11.347407	34
GTAGT	22691125	1.4257414	9.400372	36
TTATT	39429390	1.4200338	6.7461114	32
TTTAA	15776305	1.41158	8.277574	5
TATAG	11883640	1.4044385	16.77246	47
GCGAT	1622915	1.3912457	23.96212	10
TTAAG	11734515	1.3868146	10.026854	6
AAAAC	183235	1.3806232	20.953249	6
TTATA	15333340	1.3719459	12.962541	46
GGAAT	8758335	1.3671868	9.846777	2
AGATA	4612560	1.3543062	5.4978223	26
TAAGC	838325	1.351727	40.832928	7
GAACG	630270	1.3423206	7.0554357	28
GTTTA	28211070	1.341996	8.1096945	4
GGCGA	1184710	1.3414457	8.330853	2
TCGAC	149895	1.3272889	7.5627384	23
GACGT	1516320	1.299867	5.785055	3
GGTTA	20661320	1.2982033	16.954342	2
GGAGT	15591710	1.2939917	10.270748	2
TGGAA	8262525	1.2897903	9.417655	1
TCGGG	2829130	1.2894139	27.8231	36
GGGTT	38522660	1.2868632	14.173451	2
GGTAG	15505115	1.2868049	7.0912232	2
GAGTA	8242900	1.2867268	15.636968	34
GTTAA	10770060	1.2728329	21.238993	3
TTGTA	26540845	1.2625437	13.626875	20
ATTAT	14066860	1.2586279	12.822248	45
AACGA	310620	1.244307	8.034581	38
TCGTG	3529395	1.2178302	7.2629743	40
GGGGA	11013830	1.2073387	10.00737	2
GGGAT	14486305	1.2022516	11.468186	42
TTTGT	62470265	1.1961439	6.5503364	19
TATTC	2420585	1.1893846	29.09015	33
GATTA	9953065	1.1762784	16.348925	44
CGATC	131905	1.1679912	16.51361	40
GTAAT	9876605	1.1672422	20.39356	22
CGTAT	1797975	1.1669135	5.1731367	13
GGTGG	26401685	1.1649327	10.625425	8
CGTAA	719995	1.1609302	8.972036	21
TGAGG	13689665	1.1361368	15.243706	45
CGTGA	1314860	1.1271652	8.396521	26
GGGGT	25246760	1.1139735	7.9349084	2
CGAAC	50540	1.1118215	5.735212	9
GGATT	17636130	1.1081232	8.93946	43
TAGGC	1284040	1.1007447	8.699533	13
TCGAT	1673610	1.0861987	6.643878	11
TTTTC	5486920	1.0851982	11.11116	29
AGTAT	9046610	1.0691513	13.492306	30
AGTAA	3635815	1.0675215	7.045576	9
GTATT	22416310	1.0663403	5.909231	31
TGGAG	12840360	1.0656509	9.901831	1
GTGGC	2308100	1.0519476	31.453625	9
GGGTA	12657540	1.0504782	14.204031	2
TTAAT	11683655	1.0453914	14.433655	4
CGTGT	3014970	1.040326	6.91981	41
TGTAA	8774250	1.036963	19.736153	21
AACTC	61930	1.0314494	29.130615	22
GTTAT	21438020	1.019803	8.052624	31
TGTAG	16025135	1.0069001	7.606034	21
AGTTG	15868300	0.9970459	9.358905	38
ATTTC	2022185	0.9936258	5.708715	22
AAGGC	462775	0.9855973	15.279282	46
TAAGT	8277100	0.97820854	6.3804674	7
GTTGA	15550965	0.97710675	12.061055	43
GGTTG	28989530	0.9684056	6.59873	42
TTATC	1964540	0.9653012	11.592147	37
TGGGG	21788525	0.9613843	8.434091	1
AAGAC	238460	0.95524263	8.679379	32
TTGGG	28251690	0.94375783	6.1273813	36
GAAAC	234835	0.94072133	7.372012	36
TGCGG	2045610	0.9323143	5.9980426	5
GGGGG	15974975	0.9310273	5.6034555	2
TAGAC	570095	0.91922927	10.329916	25
ATTAC	749750	0.9152506	5.166471	29
GGATA	5832340	0.91043544	7.317479	2
GTTTG	35946730	0.90912277	6.47881	18
GGAGC	793840	0.89886403	8.41042	27
TGGTT	35466870	0.8969867	7.3846645	1
GTGGT	26816430	0.8958126	7.463185	9
CGTCT	249175	0.8880988	8.729048	47
AGTGA	5662615	0.8839412	5.3257394	18
GGGTG	20016695	0.88320506	8.224912	2
GAAGC	397925	0.8474827	9.759819	4

GGTAT	12820525	0.8055462	5.8435493	2
GGTAC	938280	0.8043415	12.697814	3
GGAAC	367910	0.7835581	6.350174	2
TGGGT	23258090	0.77694494	9.010233	1
GGTAA	4932695	0.7699996	6.0666676	2
GTGCG	1660675	0.756875	5.640952	4
GTTGG	22493230	0.7513944	5.1689086	39
TGGTG	22391235	0.74798733	5.6681333	7
GAGTC	792310	0.67920864	12.828283	21
TGGTA	10447675	0.65645415	5.085503	1
TGGGC	1169960	0.533225	5.4509287	13
GATTC	804250	0.5219706	5.833193	29
TGGAT	8137475	0.51129836	5.0444837	1
TCCAG	53390	0.47275728	15.706033	25
TCACG	51810	0.45876673	15.604495	30
CCAGT	49000	0.43388477	15.685313	26
CAGTC	48900	0.43299928	15.7332945	27
GTCAC	46435	0.41117227	15.727193	29
GGTGC	871770	0.39732087	5.678437	3
ATCTC	52215	0.3500423	12.376405	42
TCTCG	58370	0.2080398	6.6109967	43
CTCGT	56875	0.20271139	6.6327944	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	240567	0.2952317254320777	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	188039	0.23076763819859938	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	149923	0.18399043082365155	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	90892	0.1115456483556448	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	86086	0.10564756727043127	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	86077	0.10563652217476609	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	85565	0.10500817895470171	No Hit