FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD9_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`

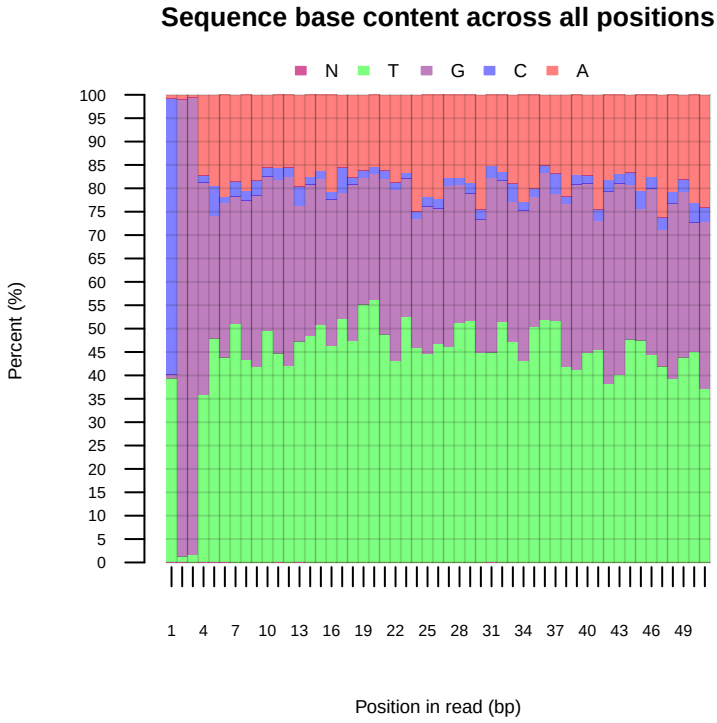
1 Sequence Duplication

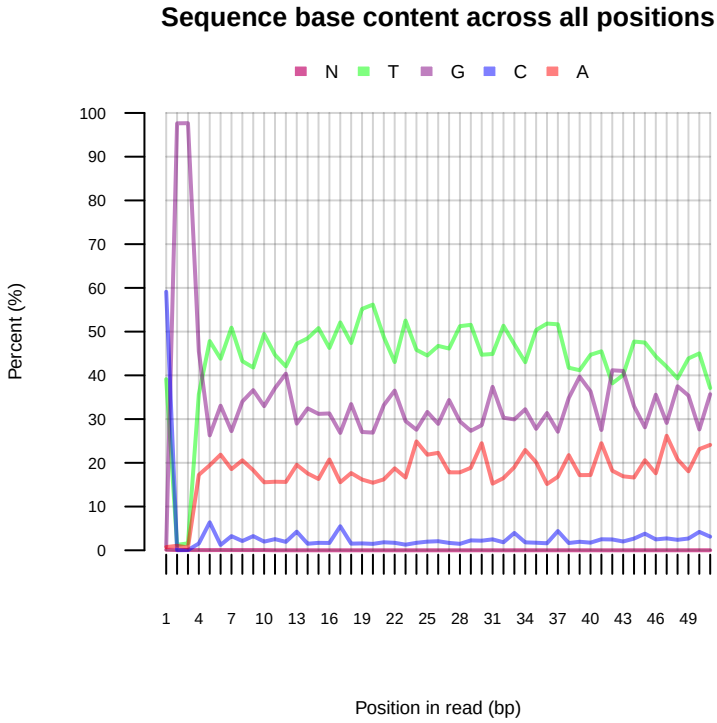
- Estimated Duplication rate 75.7612%

2 Per base sequence quality

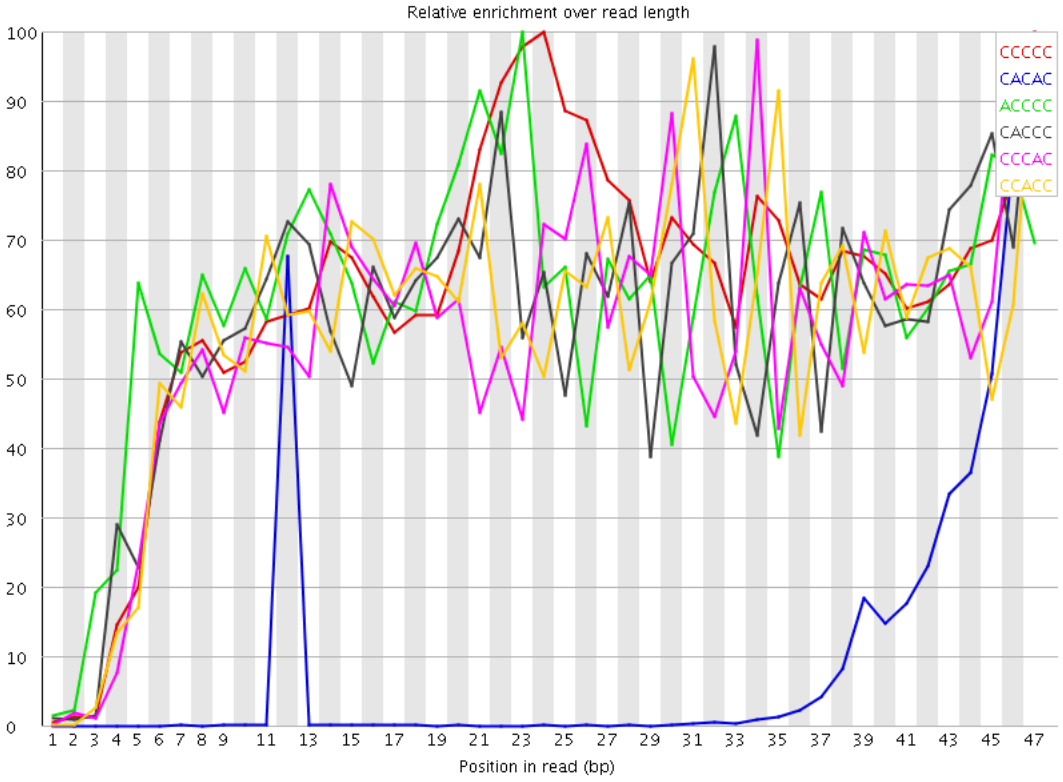
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	117885	664.8941	1066.8187	24
CACAC	907700	188.58104	1885.2867	47
ACCCC	43340	46.91521	76.30358	23
CACCC	42085	45.556686	77.57921	47
CCCAC	40645	43.997894	78.08792	47
CCACC	40065	43.370052	75.54434	47
CCCCA	39455	42.709732	68.41918	25
CGGGC	4924030	27.697271	955.51483	1
AGCAC	1317720	27.351822	196.72212	45
GCACA	1120000	23.247763	196.18666	46
GCCCC	26495	14.93019	30.586803	47
ACACG	714070	14.821901	157.05551	47
CCCGC	25325	14.270883	24.495735	34
CCCCG	25025	14.1018305	27.010485	25
CGCCC	24830	13.991946	24.362291	23
CCGCC	24545	13.8313465	24.627974	31
CGGAA	6137640	12.728348	187.62985	1
ACTCC	141655	11.892395	491.3124	23
CTCCA	137400	11.535175	491.0385	24
CGCGG	1893860	10.652811	237.98972	5
GCGCG	1785230	10.041775	236.3017	4
AGATC	5609645	9.030487	45.22338	43
CGGCG	1591160	8.950147	241.70049	1
CGCGC	158100	8.901039	64.30448	13
CGGGA	6950915	7.5039363	211.87898	1
GGCGC	1321160	7.4314184	236.7437	3
TCGCG	1357730	5.9283648	22.818586	30
TCCCC	13470	5.892155	13.265645	5
CCTCC	13415	5.8680964	10.175406	40
CGGGT	13252140	5.7811594	223.30544	1
CTCCC	13040	5.704061	10.484011	47
CCCCT	12610	5.515967	9.55895	47
CCCTC	12355	5.4044223	9.353141	39
ACGTC	625820	5.2492146	52.13405	15
AGACG	2528020	5.242653	58.45359	27
CGGAG	4722605	5.09834	148.99352	1
GATCG	5830115	4.8857303	24.768839	44
ACGCG	446695	4.826706	16.094027	14
CGCGT	1097305	4.79125	20.864182	31
CGGAC	432425	4.672513	155.33621	1
CACGT	554560	4.6515045	51.649532	14
CGGTT	13447170	4.5537057	154.24533	1
AGAGC	2178755	4.518341	28.825407	47

CGTCG	1023950	4.470954	23.0381	41
CGCGA	404275	4.368342	17.817795	24
CGGTC	998495	4.3598075	162.12663	1
GGGCG	7664945	4.3075747	100.04493	2
AAAAA	2911715	4.276586	13.224426	31
AACCC	20210	4.1987696	6.590279	32
CGGGG	7355350	4.1335874	100.445145	1
TCGAG	4867715	4.079224	45.66825	32
ATCGG	4782715	4.0079923	23.121157	45
ACACC	18830	3.912064	6.4927397	47
TCGGA	4648705	3.8956897	23.31307	46
AGGCG	3600675	3.8871481	48.56194	47
ACCAC	17990	3.7375488	6.492746	46
GAGAC	1791740	3.7157423	56.16632	26
CGACG	343115	3.7074854	31.301737	24
ACCCA	17625	3.661717	6.785565	33
CAACC	17565	3.6492515	6.980813	31
CCACA	17425	3.6201656	6.9808593	35
CACCA	16910	3.5131707	5.6627984	36
CCCAA	16700	3.4695423	6.102069	29
CCAAC	16605	3.4498053	5.80919	30
CGTTT	12920970	3.3965216	38.463287	17
TTACG	5217260	3.3939075	36.16675	14
CGGTA	4009625	3.3601303	116.3969	1
GAGCA	1599100	3.3162422	22.808784	47
GGCGT	7473350	3.2602	50.98982	3
GGCGG	5651005	3.1757731	42.534	11
TACGT	4795460	3.1195202	38.378094	15
ACGTT	4752710	3.0917106	40.196663	16
CGGAT	3562930	2.9857926	101.04985	1
GACGG	2757975	2.977402	30.402933	28
CGAGG	2706050	2.9213464	51.576073	45
CACGA	138640	2.8777409	126.79945	31
ACGGG	2663600	2.875519	30.573092	29
TTTCG	10699325	2.8125203	16.527079	30
AAGCG	1350240	2.8001518	54.360996	8
AGAGA	6897900	2.7454808	20.247353	25
ATCGC	325065	2.7265604	34.277237	29
AGCGA	1281470	2.6575353	55.05357	9
TTCGA	4055600	2.63823	36.808506	31
CGAGA	1270625	2.6350448	36.07567	25
TTTTT	162761615	2.5757818	5.736219	16
GAAGA	6317165	2.5143385	10.585408	46
CGTTA	3864060	2.51363	34.334873	9
GCGGG	4414550	2.4809055	43.471405	12
GGAAG	11805960	2.4461293	11.252637	2
TCGTT	9282185	2.4399981	6.364613	4
AACTC	151355	2.4387379	97.00953	22
GGAGG	22585760	2.4360678	28.326654	39
ATTCG	3744510	2.4358609	47.185516	34
GTCGA	2893950	2.4251766	41.65615	43
GCGGC	425550	2.3936844	9.622367	9
CGTTC	692160	2.3460276	29.415775	33
TTCGC	679240	2.3022363	9.380237	33
TTTTA	58762655	2.3013184	12.976826	26
AGAAA	2956485	2.2604797	5.769597	22
GAGAT	14016965	2.2544324	8.636055	26
TTTAG	44588060	2.249514	16.026222	27
TTCGT	8546975	2.2467341	5.8725586	35
AAGAG	5626490	2.239438	10.58886	47
GAGGC	2041240	2.203643	40.52573	46
GGGAG	20420350	2.2025096	24.400116	38
CGAGT	2605560	2.1835012	44.71769	33
AGTAG	13575040	2.183355	22.628025	35
CGTAG	2585295	2.1665187	23.431177	5
GCGGA	2001050	2.1602554	24.248325	7
GAGGT	25386505	2.125508	22.154789	40
CGGTG	4813545	2.0998776	41.852715	1
GCGTT	6187250	2.09523	28.683886	16
AGGAG	9840120	2.0388181	10.048137	38
ATTTT	50502330	1.9778199	7.6728144	25
TAGTT	38923905	1.9637517	10.701111	29
CACGC	17855	1.9310439	9.17403	47
TTTAC	3822390	1.9301795	27.216991	13
AAACG	481095	1.916571	13.70205	7
ACGGA	916960	1.901608	8.789657	30
TAAAA	3153055	1.8713787	5.3062325	30
GGTCG	4276505	1.8655974	23.329973	42
CGATC	221945	1.8616167	51.568108	33
TACGC	220135	1.846435	11.513654	13
AAAAT	3085610	1.8313493	5.176833	32
AATTT	18871280	1.8289208	16.953444	24
TAGAG	11282810	1.814682	9.051308	24
TCGTC	534470	1.8115486	9.314811	40
CGCAC	16705	1.8066697	8.36082	47
GCGGT	4139815	1.8059672	23.53052	6
AGGTA	11227980	1.8058633	29.852861	47
TTAGT	35789875	1.8056368	15.446371	28
CGAGC	165425	1.7874788	6.095911	13
GACGC	164940	1.782238	10.233089	3
ACGGC	164845	1.7812116	7.8060956	6
TGAGA	11073230	1.7809741	6.1502447	41
GCGTA	2120240	1.7767951	23.030048	4
ATCGT	2709875	1.7628152	13.95159	39
GTCGC	400320	1.7479489	8.623396	3
AGCGC	160940	1.7390165	6.156993	35
GGAAA	4357835	1.734492	11.872543	2
GAAAA	2267865	1.7339722	5.323888	3
AGTTA	13757760	1.7176559	23.320265	30
AGTTT	33868370	1.7086947	9.052822	26
GGAGA	8144545	1.6875043	10.362969	2
TGGCG	3861690	1.6846371	37.414	10
AGCGG	1558265	1.682242	7.1348677	6
GAGCG	1544575	1.6674628	9.770813	28
TACGG	1977610	1.6572691	11.439877	5
TAGTA	13264955	1.6561291	12.880526	29
AACGC	79105	1.6419772	10.74967	11

TAGCG	1957050	1.6400394	5.607585	10
TATCG	2517140	1.6374379	14.25876	38
TTGAG	25052400	1.628225	14.746633	44
GGACG	1501220	1.6206585	16.285538	2
TATTT	41144325	1.6113329	5.394909	33
AGTCG	1915560	1.6052701	14.728514	22
TCGTA	2461440	1.6012046	5.615048	43
GTAGA	9941930	1.5990204	8.742105	23
GCGTG	3659950	1.5966293	35.383507	4
CGTGG	3615155	1.5770878	35.1566	5
GCACC	14575	1.5763072	8.742012	47
CGATT	2417215	1.5724354	19.354727	11
TGGGA	18725345	1.5677965	12.800656	37
GCGAC	144435	1.5606738	21.044603	23
TTCGG	4554010	1.5421554	25.217758	35
TAGGA	9515465	1.5304294	8.035908	37
AGGTC	1819730	1.524963	39.678837	41
GTCGT	4485505	1.5189568	10.454491	3
GGGAA	7270590	1.5064257	13.730312	2
AGCGT	1776300	1.488568	8.00156	29
GTAGT	22903495	1.4885614	9.806031	36
TTATT	37677695	1.4755696	8.20308	32
ACGCC	13610	1.471941	6.734061	23
AGGTT	22591295	1.4682705	14.771278	41
GGTTT	55687120	1.4625189	9.269157	2
GCGTC	333235	1.4550303	10.556128	40
AACGG	696830	1.4450988	7.6917768	29
TAATT	14881030	1.4422035	16.611544	23
AAGTA	4649095	1.436399	11.879339	34
CGTAC	170660	1.4314516	9.460037	13
GTACG	1695675	1.421003	11.182264	4
TATAG	11349465	1.4169803	17.732481	47
TTTAA	14554665	1.4105737	8.683535	5
CGAAA	352745	1.4052542	5.095966	32
TTATA	14339985	1.3897681	14.021817	46
TTAAG	10962525	1.3686708	10.29582	6
ACGGT	1624110	1.3610303	10.883709	6
GCGAT	1611345	1.350333	23.017986	10
ACGTA	838085	1.3491613	5.039436	26
GGAAT	8384250	1.3484892	9.669357	2
GTTTA	26682140	1.3461418	8.596504	4
AAAAC	175535	1.3433254	23.56787	6
TATTC	2655765	1.3410727	35.09989	33
TCGAC	159140	1.3348249	6.907675	23
TCGGG	3044720	1.3282393	31.168285	36
TAAGC	816685	1.3147112	40.16046	7
GGCGA	1215225	1.3119094	8.60515	2
GAACG	630870	1.3083094	7.6434836	28
CGCCA	11985	1.2961949	5.9971523	24
CGTCT	381715	1.2937961	20.322311	16
GAGTA	8006610	1.2877511	16.292841	34
GGTAG	15334020	1.2838547	7.287359	2
TTGTA	25420910	1.2825116	14.775137	20
GGAGT	15276280	1.2790203	10.1138525	2
GACGT	1522900	1.2762147	6.045832	3
TGGAA	7890450	1.2690684	9.092251	1
ATTAT	13087285	1.2683617	13.930651	45
GGGTT	37244285	1.2600889	14.110629	2
GGTTA	19164785	1.2455722	15.667147	2
TTTGT	59835125	1.2198547	7.074794	19
ATGCC	144960	1.2158866	55.011887	47
CGTAT	1868110	1.215234	5.549662	13
TCACG	143095	1.2002435	51.327347	30
GTTAA	9576355	1.1956075	19.061493	3
GATTA	9527350	1.1894894	17.298643	44
GGGAT	14177455	1.1870202	11.590932	42
GGGGA	10884870	1.1740265	9.646718	2
TCCAG	139945	1.1738223	50.921906	25
TGAGG	13994885	1.1717343	16.835974	45
CGTAA	727315	1.1708422	10.45207	21
CAGTC	138660	1.1630439	51.350384	27
GGTGG	26596250	1.1591958	10.84744	8
GTAAT	9238265	1.1533971	20.211178	22
CCAGT	136700	1.146604	50.798126	26
GTCAC	136685	1.146478	51.47104	29
ATTTA	11757890	1.1395227	5.0268955	34
TCGTG	3358275	1.1372355	5.9762726	40
ACGAT	705650	1.1359655	11.324719	32
GTGGC	2588455	1.1291965	35.32747	9
TTTTC	5474825	1.1171582	11.953512	29
GGATT	17098035	1.1112484	9.262772	43
CGTGA	1309505	1.0973862	8.638414	26
CGAAC	52355	1.086729	9.178362	9
GTTAT	21299050	1.0745593	9.748077	31
TCGAT	1647710	1.0718607	7.0248823	11
GGGGT	24586110	1.071584	7.8637543	2
GTTGA	16329595	1.0613054	13.712665	43
TAGGC	1264295	1.0594996	9.938834	13
TGGAG	12566335	1.0521277	9.91457	1
TGTAG	16176255	1.0513395	9.021322	21
GTATT	20799855	1.0493745	5.3351326	31
AGTAA	3374910	1.0427227	7.1431513	9
GGGTA	12409875	1.0390279	14.26297	2
AGTAT	8251545	1.0302051	12.030998	30
AGTTG	15772350	1.0250885	9.800328	38
TGTAA	8210045	1.0250238	19.411564	21
ATTTC	2018140	1.0190934	6.1669865	22
TTAAT	10306070	0.9988187	13.0885515	4
GGTTG	29231550	0.9889934	7.2814736	42
TAAGT	7855840	0.98080134	6.559369	7
CGTGT	2836235	0.9604534	5.7064805	41
ATCTC	147440	0.95998716	42.32287	40
GTTTG	35605185	0.93510413	6.8575206	18
TGGGG	21418030	0.93350345	8.466296	1
TTATC	1827870	0.9230134	10.699795	37
TTGGG	27074625	0.9160181	5.802057	36
CGTGC	208450	0.91017175	5.3924174	13

GTGGT	26764155	0.9055138	7.360786	9
TGCGG	2058990	0.8982209	5.1049294	5
GGATA	5581045	0.8976329	7.222746	2
TAGAC	554985	0.8934228	11.314153	25
AGTGA	5548015	0.8923206	5.5609236	18
AAGAC	223925	0.8920652	9.3944	32
TGGTT	33804535	0.88781345	7.516759	1
GGGGG	15640670	0.87818617	5.4416676	2
GGAGC	808015	0.87230146	8.906316	27
ATTAC	697845	0.872047	5.107305	29
GGGTG	19956715	0.86981213	8.051425	2
AAGGC	418260	0.86739504	12.278365	46
GAAGC	406215	0.8424159	11.856288	4
GGTAT	12616845	0.8200035	6.0249944	2
TGGTG	22769500	0.7703624	6.2145977	7
TGGGT	22717730	0.76861084	8.994141	1
GGAAC	368050	0.76326865	6.81948	27
GGTAC	902630	0.75641847	11.199714	3
GGTAA	4686050	0.75368553	5.9982867	2
GTTGG	21716715	0.7347434	5.24881	39
CACAT	43705	0.7042055	6.4552546	47
CGGCC	12180	0.6857347	5.220054	1
GAGTC	807080	0.676346	13.643362	21
TGGTA	10296565	0.66920215	5.293468	1
GATTC	813770	0.52936983	6.1206217	29
TCTCG	155700	0.5277342	22.148268	41
CTCGT	155195	0.5260226	22.161953	42
TGGGC	1182640	0.51591897	6.120851	13
TGGAT	7859140	0.5107872	5.140171	1
TGAAC	312160	0.50251967	10.272425	20
CTGAA	245185	0.39470235	10.025016	19
GAACT	179510	0.28897777	10.158551	21
TCAGA	149960	0.24140775	10.009959	36
CAGAT	148005	0.23826057	10.088625	37
AGTCA	147865	0.2380352	10.113396	28
GATCA	147635	0.23766494	9.961162	34
ATCAG	144125	0.23201449	9.958888	35

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	264942	0.33573370433144023	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	177263	0.2246271396415219	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	161317	0.20442041647468104	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	100913	0.1278766496259507	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	91149	0.11550373823744989	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	88543	0.11220142288734408	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	86415	0.10950482769738815	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCC	82210	0.1041762643638521	TruSeq Adapter, Index 9 (100