FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD1_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`

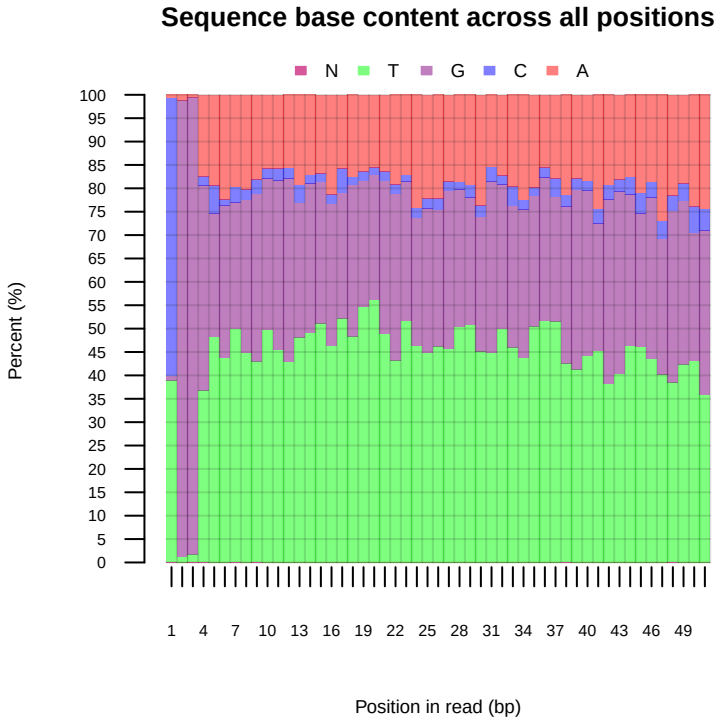
1 Sequence Duplication

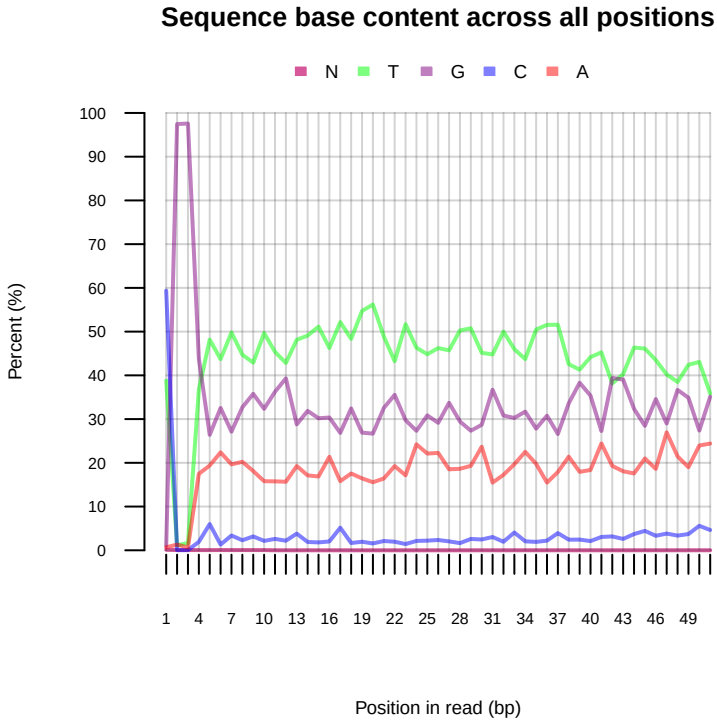
- Estimated Duplication rate 78.0103%

2 Per base sequence quality

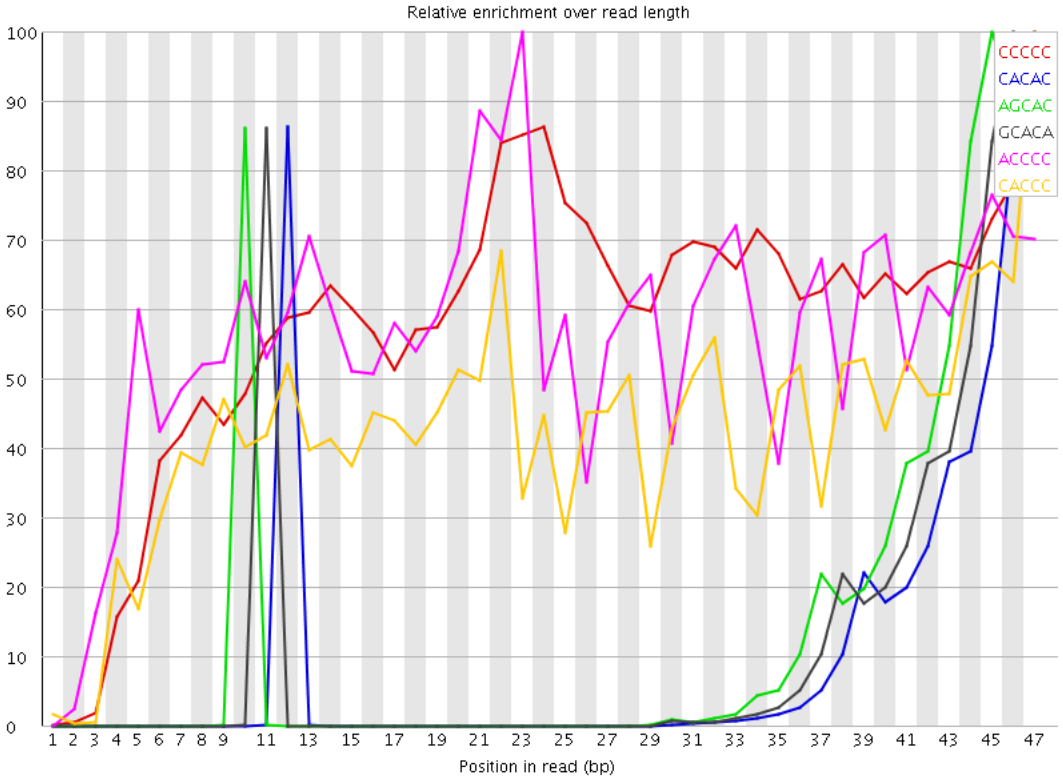
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	110540	442.97882	758.02075	47
CACAC	1742280	300.35696	2730.6082	47
AGCAC	2456145	47.250675	318.2978	45
GCACA	2125540	40.890583	317.39502	46
ACCCC	40270	33.47135	59.174618	23
CACCC	40005	33.25109	77.731384	47
CCCAC	38440	31.950308	74.41121	47
CCACC	37215	30.932121	55.46662	47
CCCCA	35775	29.735231	55.65963	24
ACACG	1413490	27.192348	265.50223	13
ACTCC	309975	22.0644	925.0943	23
CTCCA	300340	21.37857	922.10614	24
CGGGC	3813460	21.236616	744.1852	1
CGGAA	8220005	17.64657	172.53934	1
AGATC	8180980	13.478238	67.609665	43
GCCCC	25615	11.454921	31.523956	47
CCGCC	25075	11.213435	18.49387	43
CCCCG	24305	10.869094	23.01252	46
CCCGC	23975	10.721519	20.385492	47
CGCCC	23900	10.68798	19.124413	44
CGCGC	209630	10.461292	42.82255	13
CGCGG	1778775	9.905745	198.33723	5
GCGCG	1679235	9.35142	196.2618	4
CACGT	1128160	8.961286	108.71269	14
ACGTC	1100220	8.73935	108.46191	15
CGGCG	1564140	8.710473	224.86674	1
AGAGC	3651730	7.839473	46.752045	47
GATCG	8572795	7.5989976	37.248	44
CGGGA	5830990	6.734984	199.10594	1
GGCGC	1180690	6.5750947	196.6775	3
ATCGG	7409375	6.5677323	35.99281	45
TCGGA	7114355	6.3062243	36.22811	46
GAGCA	2914930	6.257723	39.52392	47
TCGCG	1439005	6.149899	20.196451	30
CACGA	293495	5.6461797	255.17433	36
CACAT	366750	5.4145665	197.20088	31
AGACG	2512245	5.393246	64.70166	27
CGGGT	10981985	5.2374606	201.64148	1
CGGAG	4305935	4.9734955	144.85068	1
AACTC	334670	4.9409494	197.76375	22
AAAAA	3098460	4.7659574	18.61345	31
CGCGT	1100875	4.704828	18.425098	31
CGGTT	12663440	4.6348023	158.92426	1

CTCCC	13490	4.6296616	14.434544	29
CGTCG	1079625	4.614011	18.190054	41
CCTCC	13155	4.5146923	15.24084	28
ACGCG	434285	4.4950533	13.716196	6
TCACA	300595	4.4378777	199.96513	30
CGCGA	425755	4.406764	15.324041	24
ACATC	294910	4.3539467	197.35732	32
CATCA	293460	4.332539	197.53148	33
TCCCC	12590	4.320789	9.842891	5
CGGAC	411720	4.2614956	137.58003	1
ATCAC	287690	4.247353	191.0348	34
CGGTC	990825	4.234505	147.93365	1
CGGGG	6711145	4.170585	95.35155	1
CCCTC	12105	4.1543407	7.2575717	39
GAGAC	1842150	3.954697	62.425247	26
CCCCT	11500	3.9467094	8.386724	47
GGGCG	6202160	3.85428	87.060715	2
TCGAG	4235840	3.7546842	40.342663	44
AGGCG	3243515	3.7463655	45.291874	47
CGACG	357065	3.6957908	28.269587	24
GAAGA	7696515	3.4269712	17.225798	46
CGATC	414930	3.2959037	109.25401	38
CGTTT	11460660	3.2190576	31.971405	17
GACGG	2774490	3.2046266	34.758274	28
AACCC	18450	3.1806521	7.331628	22
ACACC	18285	3.1522071	5.8331614	47
TTACG	4544200	3.0912268	34.89107	14
ACCAC	17835	3.0746303	6.2382507	46
CGGTA	3467410	3.0735414	105.296234	1
AAGAG	6868750	3.0583982	17.192083	47
GGCGG	4904525	3.0478761	31.939766	11
ACGGG	2616410	3.022039	34.80546	29
GGAAG	12560565	3.009065	11.564985	2
GAGAT	15941565	2.9308455	9.911703	26
GGCGT	6036440	2.8788617	41.383144	3
ACGTT	4193190	2.8524497	37.823387	16
AGAGA	6383715	2.84243	23.77924	25
TACGT	4156705	2.8276308	36.42833	15
CGAGG	2429060	2.8056436	49.650486	45
CGGAT	3128705	2.7733104	92.90593	1
TTTTT	148792475	2.746714	5.75373	16
TTTCG	9446690	2.6533759	13.061935	30
ACCCA	15235	2.6264083	5.1847878	23
CGTCT	795725	2.6098065	43.895283	16
GCGGC	464765	2.58821	9.268483	9
CGAGA	1204235	2.5852318	35.02988	25
ATGCC	317525	2.522189	115.72871	47
ATCGC	314490	2.498081	29.360857	29
AAGCG	1136115	2.438993	43.700005	8
TGAGA	13229000	2.4321427	8.974639	41
TCCAG	305400	2.4258766	106.778145	25
GTCGA	2724310	2.4148514	39.905743	43
CAGTC	301430	2.3943417	107.505585	27
GTCAC	301150	2.3921177	108.243774	29
GGAGG	18547800	2.3906767	25.830725	39
CCAGT	299930	2.382427	106.67427	26
TCGTT	8449145	2.3731866	6.0879126	4
TCACG	296830	2.3578026	103.28063	35
TTCGC	705490	2.313855	7.666528	33
TTTTA	51715000	2.312083	11.734414	26
AGCGA	1069640	2.2962854	44.406914	9
GCGGG	3638735	2.2612615	32.70116	12
TTCGA	3316665	2.2561867	28.243145	31
CGTTC	685555	2.2484727	17.053171	33
CGTTA	3303790	2.2474287	26.441074	9
CACGC	24115	2.2367275	11.071622	47
TTTAG	37582575	2.1894436	14.505152	27
GGGAG	16954105	2.185261	22.685461	38
GCGGA	1841645	2.12716	24.972252	7
AGTAG	11544000	2.1223567	19.375137	35
GAGGC	1821425	2.1038053	37.17424	46
CGGTG	4368615	2.083453	40.54229	1
ATTTT	45791780	2.0472665	7.610852	25
GAGGT	20672005	2.0447993	20.036715	40
TAAAA	3143800	1.9966631	7.519537	30
GTCGC	465980	1.9914664	9.100321	3
ATCTC	321765	1.9614536	88.35921	40
AGGAG	8092465	1.938667	8.3643	38
ATTCG	2848690	1.9378432	32.571648	34
AAAAT	3041680	1.9318057	7.275128	32
CGTAG	2175870	1.9287095	19.148394	5
GCGTT	5244805	1.9195917	21.50452	16
GGTCG	4019830	1.9171124	22.708828	42
AAACG	474020	1.8913802	21.303457	7
TCGTC	569255	1.8670337	6.5574613	40
AATTT	17224080	1.8649931	16.398405	24
CGAGT	2101105	1.8624371	34.353603	33
TTGAG	24456230	1.8565096	13.243782	44
TAGTT	31854370	1.8557364	8.787854	29
TAGAG	10037610	1.8454077	10.687368	24
CGCAC	19685	1.8258338	9.415237	47
GCGGT	3788165	1.8066282	22.2575	6
TTAGT	30668300	1.7866395	13.922896	28
ATCGT	2593140	1.7640035	15.464823	39
ACGGA	819805	1.7599437	8.664117	30
GACGC	165460	1.7125889	12.098087	5
TTTAC	3277255	1.7108946	26.006462	13
AAATA	2675165	1.6990279	6.879834	33
TAGTA	11924930	1.6825093	13.424847	29
CGAGC	162395	1.6808647	7.740955	18
GGAGA	7007540	1.6787577	10.301242	2
AGCGC	161730	1.6739815	9.97886	35
AGTTT	28348415	1.6514904	8.154639	26
TCGTA	2421765	1.6474243	11.173914	43
TATCG	2414060	1.6421828	15.787654	38
GAAAA	1983940	1.6418701	5.341932	3
GGAAA	3685290	1.6409222	11.962475	2

TATTT	36701240	1.6408452	5.456783	32
AGGTA	8921265	1.6401683	24.175755	47
GTAGA	8850290	1.6271198	10.311136	23
TACGC	204325	1.6230099	7.5851336	13
GAGCG	1399580	1.6165606	9.452192	28
ACGGC	154755	1.6017867	7.3994946	6
GCGTA	1787020	1.5840296	18.878004	4
AGTTA	11222135	1.5833508	18.467047	30
GGACG	1368150	1.5802579	16.038414	2
AGCGG	1347335	1.5562159	7.2346797	6
GCGAC	149695	1.5494137	16.96311	23
ATAAA	2430325	1.543527	6.6276484	37
AAAAC	207945	1.5421429	37.377155	6
TGGGA	15582200	1.5413344	12.655411	37
GCGTC	358575	1.5324479	9.286786	40
AGCCC	16455	1.5262432	5.6883583	45
GGTTT	48367620	1.5160308	10.125742	2
AACGC	78510	1.5103549	10.789658	23
TACGG	1702530	1.5091369	10.614333	5
AGTCG	1698955	1.505968	14.4030075	22
GCACC	16180	1.5007362	9.480623	47
GTCGT	4073695	1.490967	10.1346855	3
TTATT	33251745	1.4866246	6.6987767	32
AGGTC	1675525	1.4851992	37.81075	41
GGGAA	6172500	1.4787117	13.6900015	2
TGGCG	3064160	1.4613404	27.975727	10
TAATT	13457305	1.4571334	16.067364	23
ACGCC	15580	1.4450846	8.978923	23
AACGG	671020	1.4405346	11.78532	8
GCGTG	3012970	1.4369271	27.662281	4
TAGGA	7810460	1.4359478	6.6455007	37
AGGTT	18735775	1.422261	12.721644	41
CGATT	2078340	1.4138068	15.937892	11
GTAGT	18559165	1.4088544	8.593625	36
CGTGG	2879500	1.3732734	27.466042	5
TTCGG	3744705	1.370557	18.433601	35
CGAAA	342080	1.3649284	6.260076	32
AGCGT	1537290	1.3626668	7.6261544	29
TTTAA	12530540	1.3567848	7.694119	5
ACGAT	817245	1.3464185	23.243912	37
ACGGT	1508875	1.3374794	10.247168	6
TTATA	12311250	1.3330405	11.566083	46
TCGAC	166230	1.3204108	7.6894994	23
TATAG	9337115	1.31739	14.744483	47
GTTTA	22564145	1.3145169	7.645603	4
CGTAC	164080	1.3033328	6.6967173	13
GGGTT	31899210	1.3028477	14.254588	2
GTACG	1466895	1.300268	10.50158	4
AAGTA	3796845	1.2974148	10.066284	34
GGTTA	17042330	1.2937092	18.053959	2
TTAAG	9150490	1.2910584	9.113272	6
CGCCA	13895	1.2887967	9.327675	24
GGAGT	12991005	1.2850227	10.79454	2
AGATA	3710615	1.2679492	5.2278466	26
TTGTA	21739345	1.2664665	13.305486	20
GGAAT	6865790	1.2622707	9.8150835	2
CGTAT	1853565	1.2609019	10.656066	44
GGTAG	12675580	1.253822	7.15896	2
GTTAA	8854760	1.2493336	22.464499	3
ATTAT	11488520	1.2439568	11.636843	45
CCCAG	13400	1.2428842	7.9111824	27
CCAGC	13280	1.231754	7.5624957	28
TTTGT	51104315	1.229279	6.6191845	19
GCGAT	1384220	1.2269843	19.127628	10
GAACG	561115	1.2045923	7.4539123	28
GGCGA	1035535	1.1960768	7.3096924	2
GAGTA	6475315	1.1904821	13.845014	34
GACGT	1339875	1.1876764	5.687272	3
TGGAA	6421960	1.180673	9.238941	1
GGGGA	8977025	1.1570733	9.627141	2
TCGGG	2417485	1.152932	22.836811	36
GGGAT	11653465	1.1527182	10.088617	42
GGTGG	21525405	1.1455799	10.202644	8
GATTA	7914390	1.1166552	14.429572	44
GTAAT	7902585	1.1149895	19.586092	22
GGGGT	20927990	1.1137856	8.194812	2
TGAGG	11200735	1.1079358	14.135603	45
GTTGA	14576220	1.1065031	11.746681	43
TCGTG	2995305	1.0962777	6.521102	40
CGTAA	661780	1.0902886	10.350426	21
GGATT	14304450	1.0858724	8.020751	43
CTCGT	330050	1.082493	47.67835	42
TCTCG	329970	1.0822306	47.651833	41
GTATT	18550800	1.0807118	5.7320666	31
TAAGC	655165	1.0793903	31.902538	7
TTTTC	4895920	1.0553415	9.230944	29
TTAAT	9603105	1.0398072	15.201338	4
TGGAG	10495055	1.0381325	10.053932	1
TATTC	1986420	1.0370129	23.980341	33
GTTAT	17540140	1.0218339	7.914705	31
AGTAT	7219335	1.0185885	12.613348	30
GGGTA	10218595	1.010786	14.298709	2
AGTTG	13274370	1.0076776	8.671968	38
TAGGC	1135950	1.0069157	7.37138	13
TGTAG	13127225	0.9965076	7.340782	21
GGTTG	24396305	0.9964093	6.304956	42
TGTAA	7055470	0.9954687	18.71644	21
CGTGA	1114975	0.9883232	5.355823	26
AGTAA	2833630	0.9682759	6.2778497	9
GTTTG	30393145	0.95264035	6.421424	18
CGTGT	2593120	0.9490785	6.2586646	41
TGGGG	17702750	0.9421388	8.710428	1
GTGGC	1961135	0.9352925	26.309824	9
TTATC	1779290	0.92888033	11.79185	37
TAAGT	6545145	0.9234659	5.790799	7
TGGTT	29114315	0.9125567	8.007624	1
TTGGG	22231475	0.9079919	5.848192	36

TGAAC	549740	0.9057016	23.032825	20
GGGGG	13012030	0.90235937	5.762149	2
GTGGT	21887955	0.8939617	7.2836304	9
TGCGG	1858095	0.8861512	5.164598	5
AGTGA	4753065	0.8738478	5.8658915	18
CTGAA	528595	0.8708651	22.917406	19
GGATA	4704395	0.8648997	7.09578	2
GGGTG	16121590	0.8579895	8.107579	2
AAGAC	212785	0.84903014	7.657975	32
AAGGC	391090	0.8395855	11.79888	46
GGAGC	722150	0.8341068	8.633372	27
TAGAC	495675	0.81662905	10.857189	25
GGTAT	10463500	0.7943001	5.988005	2
GAAGC	364440	0.7823737	11.528778	4
TGGGT	18854930	0.770085	9.326478	1
CATAC	51115	0.7546437	6.934706	47
TGGTG	18390620	0.75112134	5.9260993	1
GGTAA	3982065	0.73209983	6.1445317	2
CGGCC	14650	0.7310878	8.654737	1
GGTAC	790400	0.70061713	10.529893	3
GGAAC	312940	0.6718144	6.302798	27
TCCCA	9335	0.66447675	5.870545	26
GAGTC	749470	0.66433644	13.375363	21
GAACT	389980	0.64249563	23.078066	21
TGGTA	8446195	0.64116347	5.247922	1
AGTCA	312315	0.5145418	22.93489	28
TGGAT	6630450	0.50332755	5.291544	1
TCTGA	606415	0.41251847	9.437285	18
TATGC	589670	0.40112758	10.005263	46
CACAG	15390	0.29606876	5.7272477	31
CAGCA	14495	0.278851	5.682075	33
GTCTG	710235	0.25994506	5.1026626	17
ACAGC	13030	0.25066772	5.5780864	32
GATCT	339120	0.230689	9.864835	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	185479	0.2693223609290196	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCC	169626	0.2463032192051169	TruSeq Adapter, Index 1 (100CGGGCGC
133825	0.19431884445854275	No Hit	
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	121900	0.17700330386322707	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	76376	0.11090077387906341	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTTTTAGTTG	71128	0.1032804839801773	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	70012	0.10166001074710625	No Hit