FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD10_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.

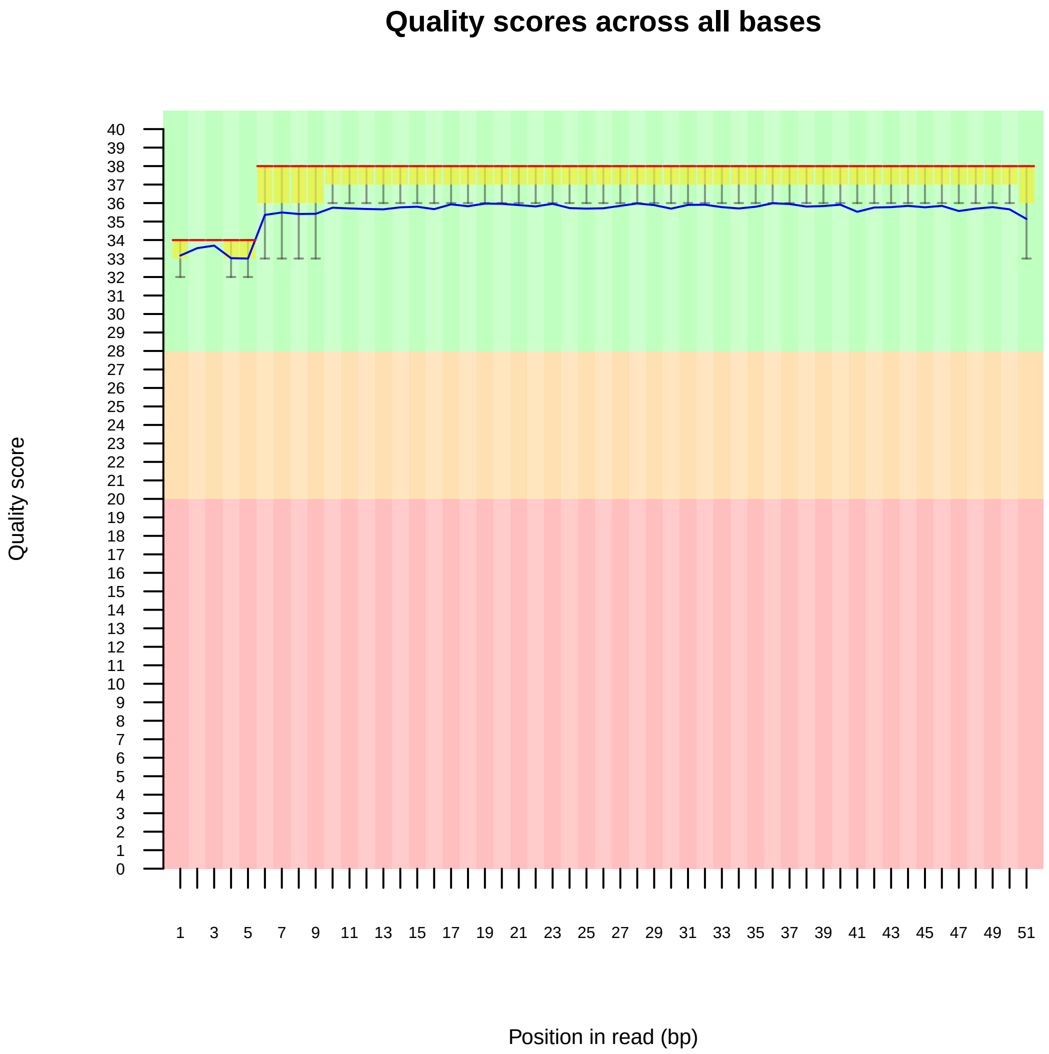
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

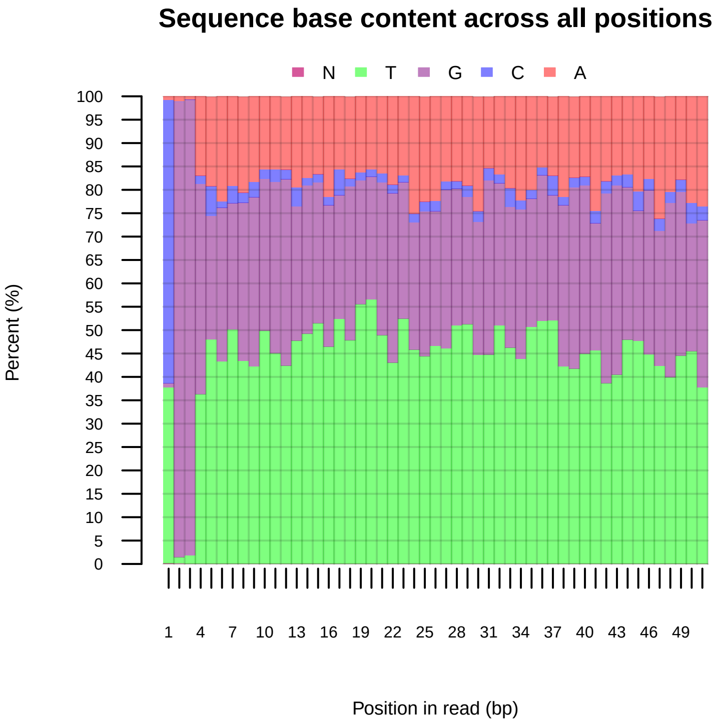
- Estimated Duplication rate 76.0512%

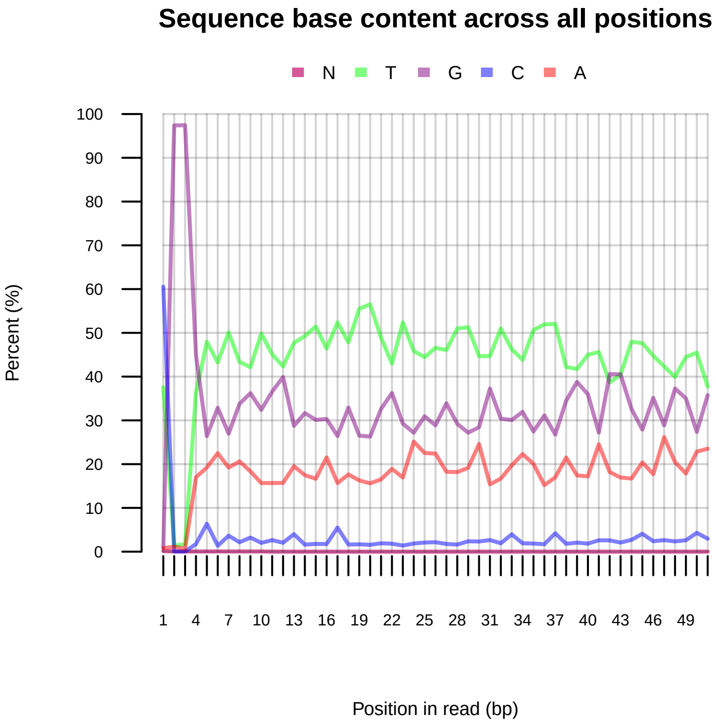
2 Per base sequence quality



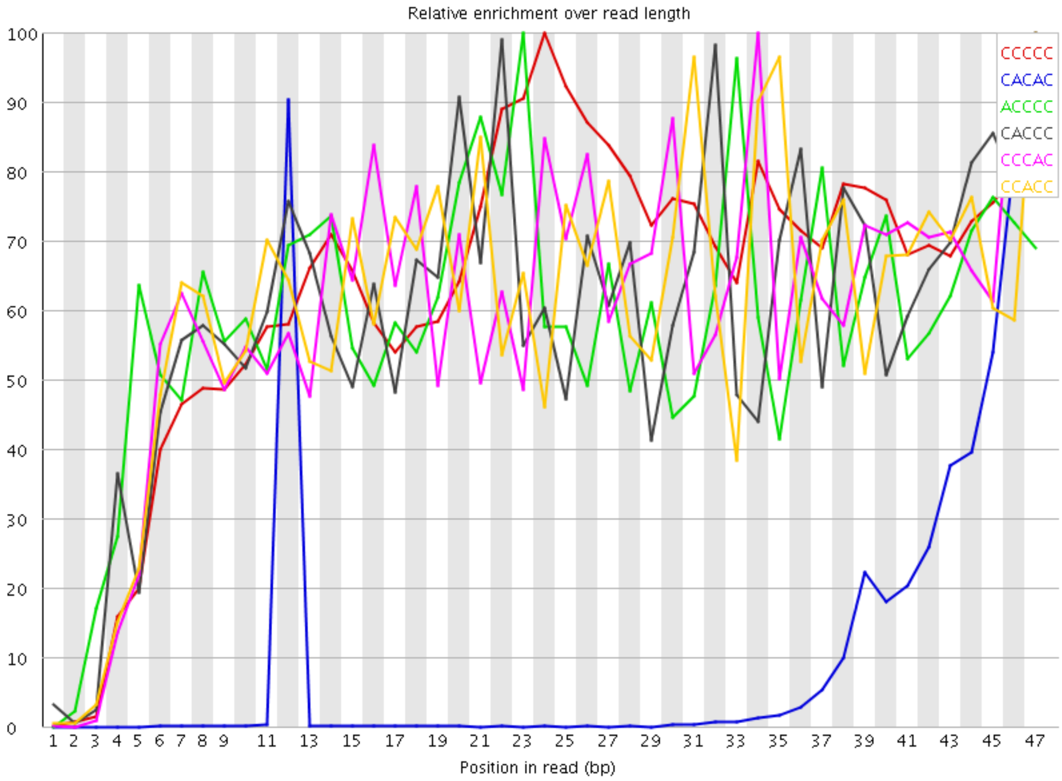
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	133200	662.8908	1037.2025	24
CACAC	867160	165.69048	1490.5872	47
ACCCC	45785	44.646896	75.839516	23
CACCC	44785	43.671753	73.09366	47
CCCAC	42980	41.911617	69.65614	34
CCACC	41945	40.902348	67.823586	47
CCCCA	41795	40.756077	66.90415	15
CGGGC	4718210	26.128048	918.5018	1
AGCAC	1207490	23.90814	161.93764	45
GCACA	1048040	20.751053	161.41751	46
GCCCC	27430	14.145802	30.779137	47
ACACG	700565	13.871095	140.93703	13
CCCGC	25990	13.403184	23.993183	47
CGCCC	25820	13.315516	22.7803	22
CCGCC	25600	13.202061	24.477713	35
CCCCG	25490	13.145333	23.143942	24
CGCGC	243835	13.030521	58.350502	13
ACTCC	161105	12.463521	517.2191	23
CTCCA	156375	12.097596	516.07715	24
CGCGG	2138505	11.842406	256.8166	5
CGGAA	5564520	11.417051	184.94374	1
GCGCG	2049615	11.35016	253.7473	4
CGGCG	1820650	10.08222	269.22165	1
GGCGC	1452160	8.041632	254.16676	3
CGGGA	6704860	7.2752795	211.21855	1
AGATC	4494635	7.0602546	34.653927	43
TCGCG	1578865	6.693825	22.47901	30
CGGGT	13205955	5.8018055	223.28995	1
AGACG	2747135	5.6364574	62.74146	27
TCCCC	14005	5.529484	12.159204	5
CGTCG	1247665	5.2896547	23.3676	41
ACGTC	657970	5.2747483	56.77479	15
CGCGT	1241785	5.2647266	20.369135	31
CGGAG	4836870	5.248369	153.30206	1
CTCCC	13260	5.235341	11.967222	24
CCTCC	13225	5.2215223	9.648128	28
ACGCG	495510	5.1885786	16.37986	6
CCCCT	12850	5.073464	10.7614765	38
CCCTC	12790	5.0497746	9.926463	39
CGGAC	456535	4.780464	145.06483	1
CGGTT	14032970	4.720007	161.38528	1
CGCGA	450550	4.7177944	17.864277	5
CGGTC	1108275	4.698692	165.49425	1

CACGT	563415	4.5167294	56.643158	14
CGGGG	7814960	4.4845657	108.02302	1
CGACG	421095	4.409365	31.317884	24
AAAAA	3016235	4.335651	12.04133	31
GGGCG	7488090	4.2969933	99.89388	2
TCGAG	5096615	4.233908	49.565296	44
GATCG	5083100	4.2226815	19.638964	44
AGGCG	3823160	4.1484165	56.762646	47
GAGAC	1986870	4.076577	60.364815	26
AGAGC	1956435	4.014132	22.579197	47
AACCC	20500	3.9169874	7.0038567	32
ACACC	19190	3.6666825	6.016175	37
CAACC	18245	3.4861188	7.273224	31
ACCAC	18195	3.4765651	5.5222945	33
TTACG	5455670	3.469824	41.823578	14
CGGTA	4079895	3.3892894	116.321594	1
CGTTT	13048325	3.360064	36.951324	17
CCACA	17470	3.3380375	6.1957674	35
ATCGG	4017820	3.3377213	17.801016	45
ACCCA	17420	3.328484	5.881469	33
TCGGA	3990325	3.314881	18.072647	46
GACGG	3043535	3.3024652	33.24394	28
CCAAC	17240	3.294091	6.5997596	30
CCCAA	17145	3.275939	5.656913	29
GGCGG	5684945	3.26227	37.031414	11
CGAGG	2988370	3.242607	61.594463	45
CACCA	16965	3.241546	5.6569366	38
ACGTT	5013180	3.188399	45.285152	16
TACGT	4997925	3.1786969	43.65119	15
GGCGT	7113260	3.1250863	47.957516	3
ACGGG	2857025	3.100088	33.058254	29
AGAGA	7397400	2.9739635	22.677086	25
GAGCA	1442205	2.9590561	18.388683	47
CGGAT	3548570	2.947902	99.18553	1
GCGGC	518425	2.870884	9.423699	9
TTTCG	10956490	2.8213973	15.164736	30
AAGCG	1342925	2.7553582	52.07101	8
GTCGA	3237395	2.6893997	48.997517	43
CGAGA	1309360	2.6864905	32.878056	25
CGTTC	826310	2.6820822	26.257723	33
AGCGA	1305395	2.6783552	52.793804	9
TTCGA	4137980	2.631769	33.8322	31
AACTC	173350	2.627761	104.46218	22
TTTTT	166361580	2.6020112	5.7427077	16
ATCGC	323085	2.5900757	32.85885	29
TTCGC	778810	2.5279043	8.099716	33
GGAGG	22061885	2.4806557	28.870773	39
GCGGG	4292690	2.4633334	37.81389	12
TCGTT	9475445	2.4400146	6.091644	4
CGTTA	3829815	2.435775	30.253508	9
GAGGC	2191310	2.3777363	46.779793	46
GAAGA	5908355	2.3753254	8.776603	46
GGAAG	11171185	2.3751411	11.1335745	2
CACTA	152560	2.3126116	104.82168	31
TTTTA	59835135	2.311421	12.607067	26
AGAAA	3033680	2.3061767	5.1923947	22
CGGTG	5195010	2.2823367	44.31035	1
TTCGT	8802080	2.2666168	5.6588597	35
GGGAG	20070240	2.2567136	25.364304	38
ATTCG	3543885	2.2539225	39.32234	34
TTTAG	44553260	2.2480323	15.751049	27
GTCGC	528390	2.2401855	11.091324	3
AGTAG	13692035	2.2287323	21.705263	35
TCGTC	680945	2.2102487	9.303854	40
GCGGA	2034990	2.2081177	21.899921	7
CGAGT	2573510	2.1378908	40.984596	33
AAGAG	5312920	2.135944	8.773192	47
CGTAG	2564855	2.1307006	22.847216	5
GAGAT	13076280	2.1285024	9.105941	26
GAGGT	24564220	2.1145918	22.32214	40
GGTCG	4764520	2.0932086	27.690613	42
ATTTT	52698780	2.0357447	8.101037	25
AGGAG	9563720	2.0333729	9.232736	38
GCGTT	6024340	2.0262942	24.580252	16
ACGGA	973300	1.9969765	8.758344	30
CGAGC	190140	1.990992	7.794629	32
TACGC	247490	1.9840531	9.612302	13
GACGC	189200	1.981149	14.314561	5
TTTAC	3993470	1.944507	31.145355	13
GCGGT	4425705	1.9443561	26.303041	6
TAGTT	37561955	1.8952705	9.392165	29
AATTT	19840055	1.89292	17.767145	24
TAGAG	11577065	1.8844664	10.09391	24
AAACG	479900	1.8618443	10.4646435	7
ATCGT	2919115	1.8565668	14.612006	39
AGCGC	176660	1.8498403	10.304344	35
CACGC	17995	1.8183768	7.265649	47
TTAGT	35937740	1.8133173	15.15194	28
GCGAC	172090	1.8019869	18.585424	23
GAAAA	2346990	1.7841611	5.3749633	3
ACGGC	169230	1.7720393	9.1307125	12
TAGTA	14096595	1.7567252	15.222522	29
GCGTA	2108365	1.7514812	22.56743	4
AGGTA	10759930	1.7514564	27.283045	47
GGAAA	4347215	1.7477031	11.98987	2
GGACG	1600335	1.736484	15.779265	2
GGAGA	8034835	1.7083119	10.64994	2
TATCG	2685740	1.7081395	15.018742	38
GAGCG	1568535	1.7019787	9.52114	28
TACGG	2048310	1.7015915	13.047685	5
AGGTC	2046345	1.699959	46.53854	41
GCGTC	399720	1.6946706	10.7777	40
CGCAC	16665	1.6839817	6.933234	47
AGTTA	13414870	1.6717683	20.208687	30
AGTTT	33074040	1.6688231	8.615904	26
GTAGA	10232070	1.6655337	9.789063	23
TCGTA	2586115	1.6447777	6.2585273	43

AGCGG	1509585	1.6380137	5.650727	6
TATTT	42163685	1.6287757	5.868154	32
TCGAC	200405	1.6065868	7.651297	23
CGATT	2519835	1.6026233	18.807598	11
TGAGA	9831065	1.6002598	5.331538	41
GTCGT	4749405	1.5974683	10.491915	3
TGGGA	18452435	1.5884635	14.132021	37
AGTCG	1907375	1.5845128	13.520871	22
CGTAC	197070	1.5798512	8.235349	13
TGGCG	3587585	1.5761427	32.26977	10
GCGTG	3564025	1.5657921	32.179214	4
CGTGG	3555225	1.5619259	32.043938	5
TCGAA	972965	1.5283512	5.097595	32
AACGC	77000	1.5245898	5.587336	23
CGAAC	76950	1.5235997	6.299289	29
CGAAA	391110	1.5173699	6.3028393	32
TTGAG	23005295	1.5161812	13.360137	44
TAGGA	9309660	1.5153875	7.2730603	37
GGGAA	7106610	1.510959	13.336684	2
TAATT	15730490	1.5008303	17.411245	23
AGCGT	1794160	1.4904617	7.70984	29
GGTTT	55663910	1.4853555	9.261052	2
TTCGG	4409120	1.4830132	21.461761	35
GTACG	1783345	1.4814774	12.716359	4
AGGTT	22153510	1.4600435	14.021787	41
TTATT	37652695	1.454517	7.1500287	32
AACGG	708450	1.4535683	7.094579	29
ACGAG	708005	1.4526553	5.179723	32
GTAGT	21904900	1.443659	9.416577	36
AAGTA	4652230	1.4319156	11.383941	34
ACGGT	1723610	1.4318538	12.404552	6
GCACC	14110	1.4258014	6.00722	47
TTTAA	14869800	1.4187129	8.52841	5
ACGCC	13900	1.4045811	5.5795527	23
TATAG	11265375	1.4038972	16.88459	47
TTATA	14557790	1.3889441	13.194875	46
TTAAG	11126650	1.3866091	10.205121	6
GCGAT	1660225	1.379198	22.45383	10
GGAAT	8441225	1.3740274	9.572893	2
GTTTA	27072785	1.3660165	8.257259	4
GGCGA	1250430	1.3568109	9.149398	2
GAACG	659200	1.3525192	6.9151998	28
AGATA	4371150	1.3454015	5.4557076	26
ATGCC	165410	1.3260424	60.169827	47
CGTCT	405480	1.3161292	22.425962	16
GACGT	1574905	1.3083202	6.105546	3
GGTTA	19739540	1.3009492	17.13951	2
GGTAG	15025800	1.2934843	7.3969254	2
TCGGG	2942285	1.2926414	26.758146	36
TGGAA	7925380	1.2900602	8.713642	1
GGAGT	14911765	1.2836678	10.14143	2
TTGTA	25408670	1.2820499	14.461764	20
ACGAC	64575	1.2785764	5.4663777	23
GGGTT	36664895	1.277931	14.084991	2
GAGTA	7836505	1.2755935	15.569003	34
ATTAT	13323735	1.2712044	13.098622	45
GTTAA	10195125	1.270522	21.162	3
TCCAG	158305	1.269084	55.580524	25
TAAGC	807710	1.2687656	38.082718	7
CAGTC	156975	1.2584215	56.05216	27
CCAGT	155220	1.2443522	55.46221	26
GTCAC	154995	1.2425485	55.97909	29
CTAGC	154830	1.2412257	55.930832	33
CGTAT	1948120	1.2390107	5.6214676	44
TCGTG	3619310	1.2173595	6.77446	40
TTTGT	59177555	1.2089642	6.960252	19
AAAAC	164490	1.2066975	17.40129	6
GGGGA	10696705	1.2027459	10.093563	2
GTAAT	9583975	1.1943601	21.440369	22
GGGAT	13833625	1.1908569	11.307945	42
CGCCA	11700	1.1822734	5.413387	24
TATTC	2420050	1.1783746	28.730238	33
GATTA	9430840	1.1752764	16.434452	44
CGTAA	741115	1.164157	8.165982	21
GGTGG	25462290	1.1591902	10.97368	8
TCGAT	1797885	1.1434608	6.637431	11
TGAGG	13184180	1.13495	15.602778	45
TTTTC	5674035	1.1186249	10.751496	29
GGATT	16692345	1.1001214	8.927132	43
GGGGT	24084615	1.0964706	8.119438	2
AGTAT	8761990	1.091924	14.338768	30
GTATT	21575530	1.0886406	6.3065276	31
TAGGC	1291970	1.0732776	8.467307	13
TGTAA	8501925	1.0595145	20.825754	21
CGTGA	1272205	1.0568583	7.728084	26
GGGTA	12273885	1.0565879	14.651026	2
AGTAA	3414330	1.0509009	7.1372566	9
TTAAT	10948445	1.0445802	14.518473	4
TGGAG	12107290	1.0422467	9.494137	1
GTTAT	20635855	1.0412272	8.478658	31
ATCTC	167460	1.027795	45.75019	40
CGATC	128185	1.0276208	6.3801165	44
GTGGC	2336285	1.0264059	30.449877	9
CGTGT	3044825	1.0241307	6.4943314	41
TTATC	2099860	1.0224673	11.133668	37
TGTAG	15319885	1.0096686	7.928557	21
GTTGA	15281395	1.0071318	12.592689	43
AGTTG	15125185	0.99683666	9.402605	38
ATTTC	2017800	0.98251045	5.5651803	22
TAAGT	7830065	0.97578686	6.5183926	7
GGTTG	27780475	0.9682704	6.8048625	42
AAGGC	469505	0.963311	14.752623	46
TGCGG	2138350	0.9394467	5.97746	5
AAGAC	241640	0.9374787	8.003896	32
TGGGG	20568230	0.9363843	8.400895	1
TCACT	152275	0.9345961	42.42924	30
TTGGG	26674100	0.9297084	6.3619213	36

GTTTG	34426675	0.9186536	6.838436	18
GGATA	5549410	0.90330976	7.224853	2
GGGGG	15094105	0.897563	5.8257	2
GTGGT	25743450	0.89727116	7.773259	9
TTTGG	33523760	0.89456004	5.210434	35
ATTAC	743695	0.8943768	5.0892057	29
TGGTT	33465225	0.8929979	7.2690887	1
GGAGC	821245	0.8911127	8.647131	27
AGTGA	5452325	0.8875067	5.4396076	18
TAGAC	561795	0.8824778	10.342165	25
GGGTG	18929775	0.86179245	8.211575	2
GGTAT	12376955	0.8157124	5.9952335	2
GGTAC	966515	0.8029125	12.733168	3
GAAGC	385600	0.7911581	8.73465	4
GTGCG	1786950	0.78506523	5.544194	4
GGTAA	4725985	0.7692761	6.2147064	2
TGGGT	21806700	0.7600583	8.822078	1
TGGTG	21336985	0.7436868	5.88864	7
GGAAC	358720	0.7360068	6.250568	2
GTTGG	21067655	0.7342994	5.134308	39
GAGTC	820785	0.68185025	12.371763	21
TGGTA	10006265	0.65947044	5.197132	1
CACAT	42330	0.64166784	5.674123	47
TCTCG	176220	0.57198447	24.27986	41
CTCGT	176000	0.57127047	24.305222	42
TGAAC	344870	0.54172814	11.549397	20
GATTC	818820	0.5207722	5.5329065	29
TGGGC	1153140	0.5066118	5.207257	13
CTGAA	271165	0.42595094	11.26849	19
GGTGC	913290	0.40123796	5.5814157	3
GAACT	202775	0.3185227	11.360035	21
AGTCA	166395	0.26137632	11.299002	28
ACTAG	162930	0.25593343	11.258558	32

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	233820	0.2981383490212138	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	191362	0.24400115792232277	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	144622	0.18440408995015817	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCC	93004	0.11858719960811295	TruSeq Adapter, Index 10 (100CGGGCGT
87861	0.11202948200903629	No Hit	
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	84783	0.1081047970450157	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82082	0.10466081585988911	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	80214	0.10227897326314107	No Hit