FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD6_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.

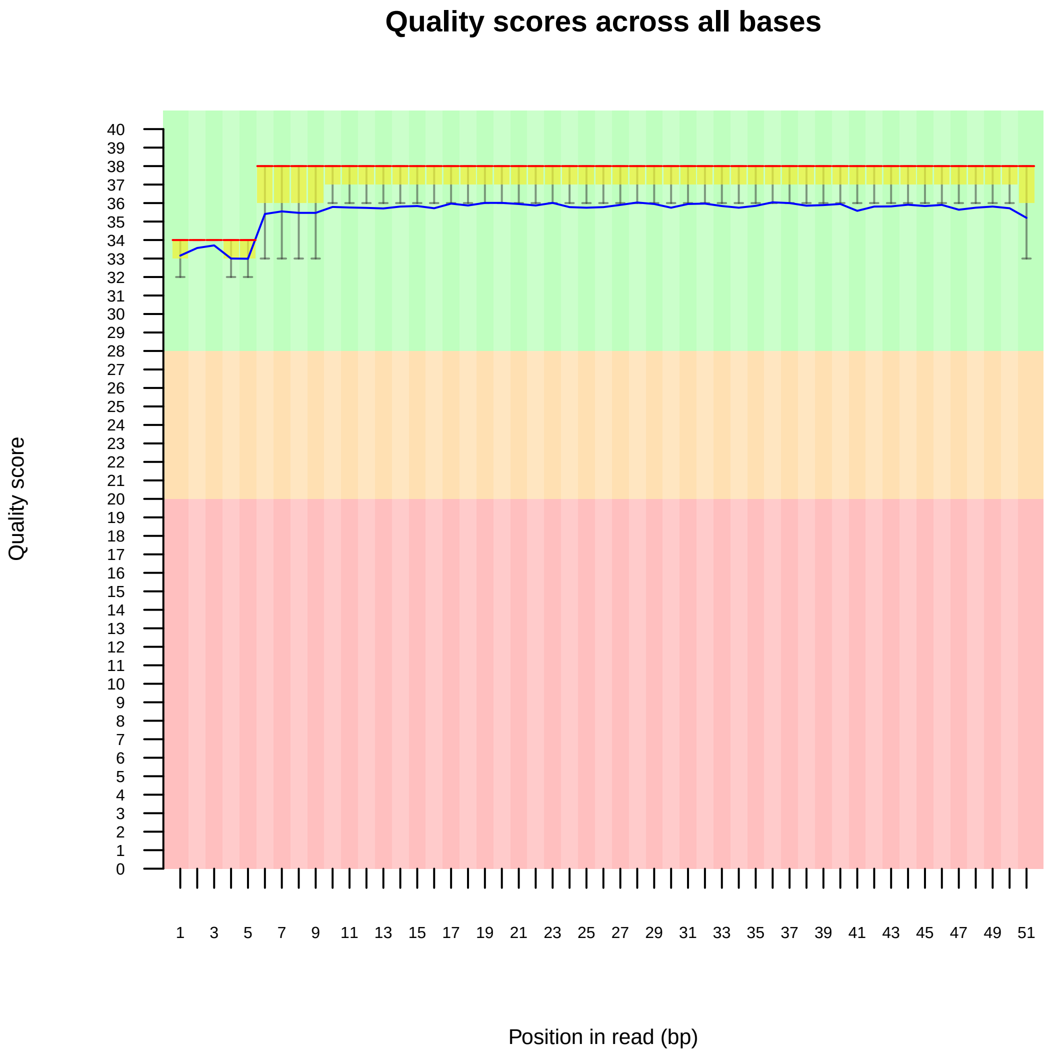
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

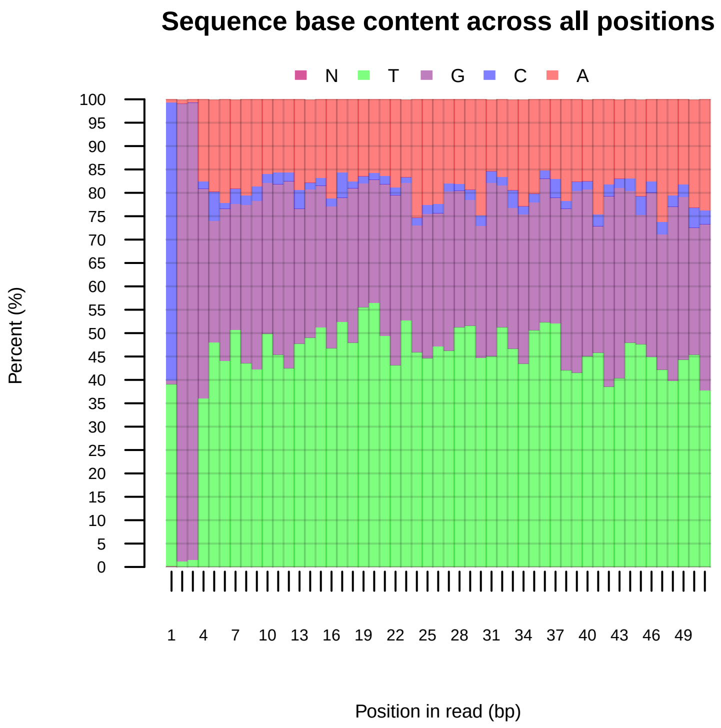
- Estimated Duplication rate 76.8984%

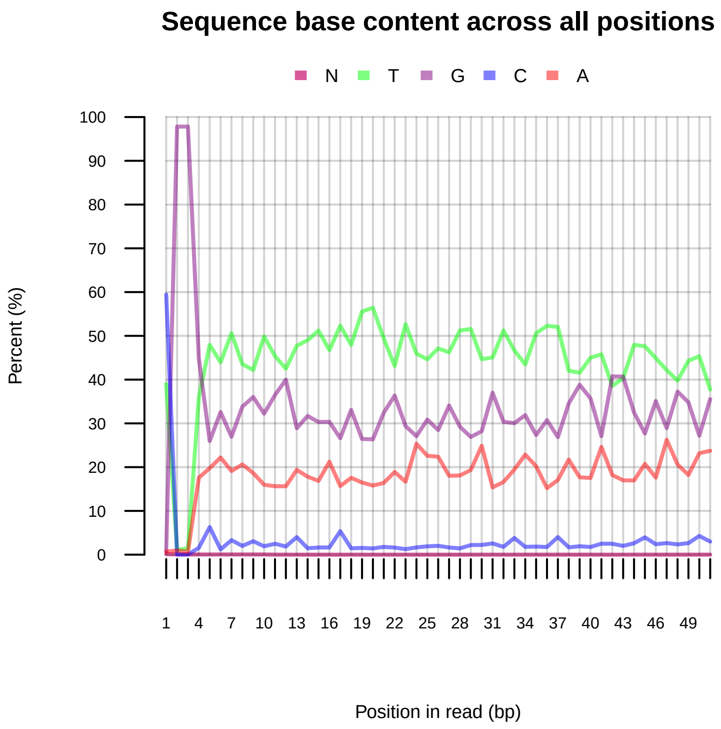
2 Per base sequence quality



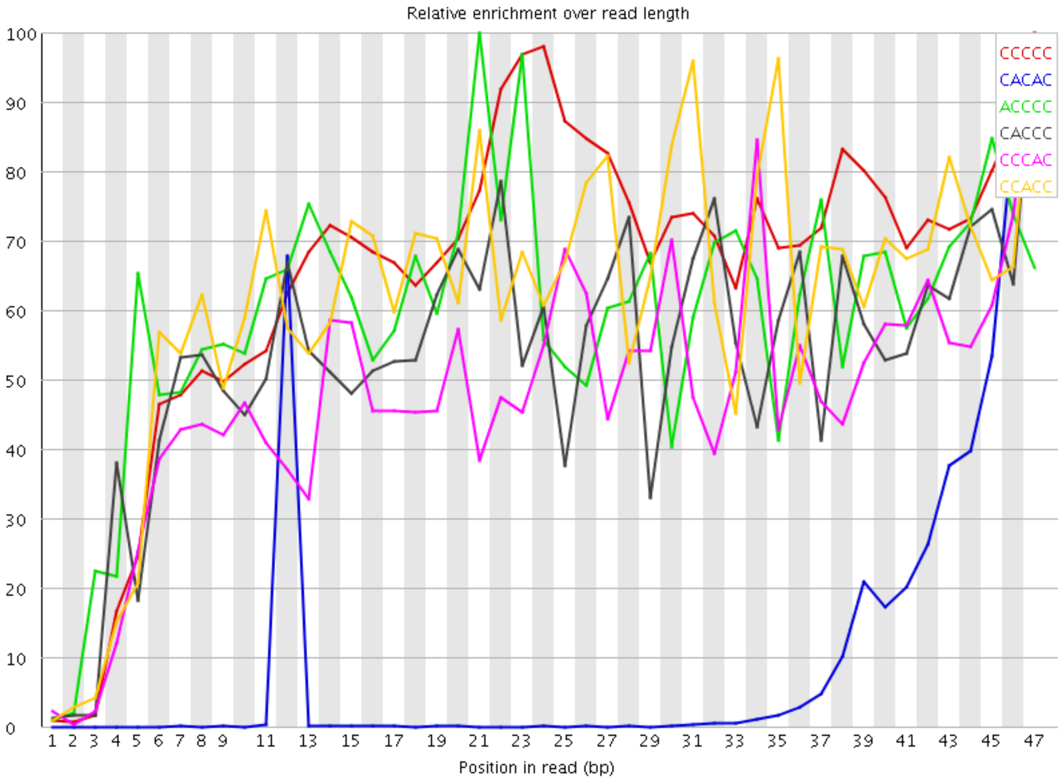
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

3 Sequence base content


Sequence base content across all positions


Sequence base content across all positions

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	122330	756.37115	1152.1187	47
CACAC	970560	212.38983	2003.4836	47
ACCCC	42810	49.796856	83.63353	21
CACCC	41350	48.098576	89.650475	47
CCCAC	40760	47.412273	98.67017	47
CCACC	39735	46.21999	74.89094	47
CCCCA	39125	45.510433	73.79474	24
AGCAC	1352990	29.624414	209.36078	45
CGGGC	4580845	28.3713	999.7189	1
GCACA	1177395	25.779673	208.9346	46
CACGC	146105	17.004568	695.1443	31
ACACG	775340	16.976469	173.70226	47
GCCCC	27195	16.82422	40.266853	47
ACGCC	141305	16.445915	693.6677	32
CGCCA	139885	16.28065	696.45905	33
CCCCG	25120	15.540519	28.781664	25
CCCGC	25040	15.491027	27.038393	47
CCGCC	24575	15.203354	27.037632	27
CGCCC	24495	15.153862	26.455769	23
CGGAA	6009395	13.165273	212.13963	1
ACTCC	146900	13.042555	528.02356	23
CTCCA	141595	12.57155	526.6282	24
CGCGG	1833615	11.356429	267.39685	5
GCGCG	1730480	10.717666	266.05383	4
CGGCG	1502200	9.303822	260.26648	1
CGCGC	136295	8.436644	60.214394	13
GGCGC	1308390	8.103467	266.49103	3
AGATC	4774035	7.9785695	38.0315	43
CGGGA	6427450	7.4890547	217.86888	1
TCCCC	13900	6.5599575	15.644799	3
CCTCC	13860	6.5410795	13.3067045	23
CTCCC	13630	6.4325337	13.972104	24
CCCCT	12995	6.132852	11.976558	47
CCCTC	12780	6.031385	10.202256	46
ACGTC	663430	5.8935885	55.89649	15
CGGGT	12393380	5.8587856	227.7938	1
TCGCG	1222735	5.7770524	23.093704	30
AGACG	2555585	5.5987287	64.382164	27
CACGT	613885	5.453455	55.37497	14
CGGAG	4462765	5.1998677	153.07567	1
ACGCG	425750	4.957915	15.976837	14
CGCGT	1011440	4.7787476	20.623652	31
CGGTT	13146475	4.740973	163.50372	1

CGGAC	401025	4.669989	156.61568	1
AGAGC	2113925	4.6311483	25.95106	47
GATCG	5176030	4.6007156	21.533506	44
GGGCG	7100865	4.4003654	104.99581	2
CGGTC	921410	4.353383	165.4833	1
CGGGG	6992805	4.333401	105.0441	1
AACCC	19725	4.316466	6.632907	22
CGTCG	912135	4.3095613	22.75844	41
CGCGA	365175	4.252511	17.334352	5
AAAAA	2906255	4.234567	13.000685	31
TCGAG	4666975	4.148242	48.000618	44
AGGCG	3540710	4.1255193	55.592136	47
GAGAC	1840095	4.031246	61.879505	26
ACACC	18385	4.023231	7.507343	47
CAACC	17630	3.8580124	7.198721	31
ACCAC	17520	3.833941	6.3244085	20
ATCGG	4256195	3.78312	19.72487	45
CACCA	17250	3.7748566	5.656124	38
ACCCA	17115	3.7453141	6.6331277	33
TCGGA	4193170	3.7271004	20.006495	46
CCACA	16905	3.6993592	6.3760533	35
CCCAA	16830	3.6829467	6.0673466	15
CCAAC	16710	3.6566865	6.9416003	30
CGACG	304050	3.5407023	29.318493	24
GAGCA	1604010	3.5140357	21.605143	44
TTACG	5091660	3.452458	40.44527	14
CGGTA	3833470	3.407381	119.22755	1
CGTTT	12208155	3.3585246	38.291584	17
GGCGT	6870705	3.2480233	51.021015	3
GGCGG	5172615	3.20544	40.631622	11
TACGT	4671985	3.1678922	42.40501	15
GACGG	2714985	3.1634116	34.1503	28
ACGTT	4634585	3.1425326	44.147648	16
CGAGG	2655205	3.093758	59.649406	45
ACGGG	2617945	3.050344	34.060978	29
CGGAT	3410770	3.031664	104.28949	1
GCCAA	137880	3.0189538	133.7564	34
AGAGA	6893335	2.8410778	21.91673	25
AAGCG	1258950	2.7580843	54.00856	8
TTTCG	9986945	2.7474585	15.855776	30
AACTC	158950	2.6549482	102.29933	22
AGCGA	1209405	2.6495423	54.751057	9
ATCGC	293390	2.6063337	33.034626	29
TTTTT	161792305	2.591677	5.716006	16
TTCGA	3818935	2.5894718	35.190037	31
CGAGA	1169545	2.5622177	33.63384	25
GTCGA	2800465	2.4891942	47.35492	43
GAAGA	5977390	2.4635725	9.493107	46
GGAGG	21121025	2.462335	29.223406	39
CGTTA	3620815	2.4551344	32.801357	9
GGAAG	11150430	2.4441879	11.460601	2
GCGGG	3942065	2.4428751	41.42349	12
TCGTT	8667445	2.3844576	6.328643	4
GAGGC	2030395	2.36575	46.20405	46
GCGGC	378395	2.343576	8.504741	33
ATTCG	3429145	2.3251705	43.66473	34
TTTTA	58042300	2.2915962	12.564373	26
CCAAT	134430	2.2453897	101.45536	35
TTTAG	43221725	2.2369444	15.726862	27
GGGAG	19159845	2.2336965	25.38071	38
CGTTC	616590	2.2223413	29.00205	33
TTCGC	616040	2.2203588	8.41394	33
AGAAA	2861680	2.2176123	5.36232	22
AAGAG	5371160	2.2137156	9.526128	47
AGTAG	13220035	2.210629	22.161846	35
TTCGT	7972685	2.1933255	5.807085	35
ACGGA	995110	2.1800687	15.207864	30
GAGAT	12994960	2.1729925	8.822072	26
CGGTG	4596220	2.1727943	42.42077	1
CGAGT	2438430	2.1673992	43.284462	33
CGTAG	2424215	2.1547642	23.834349	5
GCGGA	1823615	2.1248167	22.575993	7
GAGGT	23859935	2.1219823	22.583578	40
GCGTT	5709080	2.058848	26.728577	16
AGGAG	9243495	2.0261853	9.7307205	38
ATTTT	50622855	1.9986653	7.8574142	25
TTTAC	3764865	1.9474156	29.9894	13
GGTCG	4088505	1.9327798	26.9563	42
CGCAC	16465	1.9162945	7.848824	46
TAGTT	36889530	1.9092213	9.806	29
GACGC	161995	1.8864532	11.559188	3
GCGGT	3968765	1.8761744	26.171858	6
TAGAG	11129860	1.8611139	9.775139	24
AAACG	451580	1.8601353	13.390929	7
TACGC	208960	1.8562988	11.140178	13
AATTT	18939195	1.8430004	17.249725	24
TAAAA	3108300	1.8375016	5.236597	30
AGCCC	15765	1.8348243	5.059326	45
AAAAT	3060615	1.8093123	5.084237	32
ACGGC	155325	1.8087801	9.442997	12
AGCGC	155125	1.8064511	7.2214727	10
TTAGT	34827995	1.8025265	15.115225	28
ATCGT	2613800	1.7723166	14.851535	39
GCGTA	1984155	1.7636168	23.548931	4
AGGTA	10539130	1.7623333	28.335623	47
GGAAA	4220840	1.7396129	12.14506	2
GAAAA	2220060	1.7203993	5.4471755	3
TCGTC	475660	1.714395	9.427569	40
TAGTA	13333600	1.7008697	14.1281185	29
CGAGC	145020	1.688777	5.986899	13
AACGG	769110	1.6849518	14.104675	29
GTCGC	356430	1.6840237	8.547661	3
GGAGA	7653060	1.6775601	10.286853	2
TACGG	1885260	1.6757138	12.532021	5
AGTTT	32322910	1.6728754	8.784778	26
AGTTA	13112435	1.6726573	21.165699	30
GAGCG	1433975	1.6708207	10.342801	28

TGAGA	9989135	1.6703643	5.501116	41
AACGC	75775	1.6591326	9.193969	11
AGCGG	1419860	1.6543745	5.983873	6
TATCG	2436510	1.6521031	15.296312	38
TAGCG	1857890	1.6513861	5.375777	10
AGGTC	1851830	1.6459997	45.157837	41
GTAGA	9806360	1.6398009	9.483365	23
TGGCG	3453085	1.6323944	35.45824	10
GGAAT	9725970	1.6263583	11.407974	2
GGACG	1385405	1.6142286	16.409317	2
GCACC	13855	1.6125271	6.6181703	47
TATTT	40807820	1.6111532	5.5184584	32
TGGGA	17897365	1.5917014	13.739071	37
GCGTG	3363890	1.5902288	34.73986	4
CGTGG	3315320	1.567268	34.49467	5
AGTCG	1761050	1.5653099	13.855223	22
TTGAG	23035250	1.5628084	13.947448	44
TCGTA	2301575	1.5606089	5.8109407	43
CGATT	2295435	1.5564456	19.024246	11
GAACG	700785	1.5352669	14.0824585	28
CGTCT	425090	1.5321283	22.08135	16
TGGAA	9119245	1.5249028	9.469091	1
TAGGA	9052335	1.5137143	7.645166	37
AGCGT	1693935	1.5056547	8.362919	29
GGGAA	6859085	1.5035198	13.631413	2
GTCGT	4118650	1.4852961	10.51004	3
GCCCA	12745	1.483339	5.0866947	46
GGTTT	53664155	1.4771569	9.4177685	2
TTCGG	4079375	1.4711325	23.373053	35
TAATT	15100755	1.4694762	16.91015	23
GCGAC	125460	1.4609983	20.332687	23
AGGTT	21455605	1.4556385	14.42708	41
GTAGT	21432180	1.4540493	9.616921	36
TTATT	36795210	1.4527296	7.456341	32
AAGTA	4492670	1.4125329	11.549523	34
GTACG	1588695	1.412112	12.1895685	4
GCGTC	297130	1.4038492	9.930397	40
TATAG	10951275	1.396974	17.109528	47
TTTAA	14291385	1.3907152	8.4119215	5
CGTAC	154325	1.3709481	8.944287	13
TTAAG	10722980	1.3678521	10.136233	6
ACGGT	1538615	1.3675984	11.900976	6
TTATA	14031355	1.3654115	13.3168745	46
GCGAT	1514375	1.3460526	23.013582	10
GTTTA	25951180	1.3431059	8.358813	4
ATGCC	149485	1.3279519	58.70827	47
GGCGA	1135915	1.323531	9.253272	2
AAAAC	168140	1.3022419	22.73531	6
AGATA	4137585	1.3008914	5.266372	26
TAAGC	774720	1.2947447	39.3621	7
GGTAG	14530595	1.2922778	7.479231	2
GGAGT	14458140	1.2858341	10.2427845	2
GGTTA	18874760	1.2805432	16.78456	2
TTGTA	24648430	1.275682	14.493275	20
GGGTT	35340560	1.2751915	14.244418	2
GAGTA	7600340	1.2709144	15.85246	34
TCACG	142380	1.2648346	54.269516	30
TCCAG	142365	1.2647014	54.635674	25
TCGGG	2674280	1.264226	29.351269	36
GACGT	1412030	1.2550833	5.924427	3
TCGAC	141225	1.2545741	5.938409	23
ATTAT	12857615	1.251193	13.238778	45
GTTAA	9692670	1.2364229	20.49908	3
CAGTC	138720	1.2323209	54.408775	27
TATTC	2378940	1.2305315	31.435753	33
CCAGT	138000	1.2259247	53.899124	26
GTCAC	136770	1.214998	54.440548	29
TTTGT	57633580	1.210203	6.9651747	19
CGTAT	1766240	1.197619	5.3487496	44
GGGAT	13428370	1.1942515	11.608152	42
GGGGA	10093475	1.1767193	9.798214	2
GTAAT	9215870	1.175601	20.918007	22
GATTA	9151935	1.1674453	16.73893	44
GGTGG	24547650	1.1611053	11.096255	8
TGAGG	12986305	1.1549364	16.284302	45
TCGTG	3166235	1.1418296	6.801967	40
CGTAA	675345	1.128665	9.031701	21
GGATT	16313045	1.1067456	9.13067	43
GGGGT	23026400	1.0891501	8.136709	2
TTTTC	5107620	1.0719076	11.249096	29
CGTGA	1204480	1.0706024	8.535516	26
GTGGC	2262660	1.0696387	33.393517	9
TAGGC	1202785	1.0690958	9.1560755	13
GTATT	20556305	1.0638936	5.8758645	31
TGGAG	11835070	1.0525514	9.922472	1
GGGTA	11819070	1.0511285	14.62114	2
TCGAT	1549425	1.050605	6.9263086	11
AGTAT	8232040	1.050101	13.288411	30
TGTAA	8176080	1.0429627	20.253815	21
GTTAT	20102190	1.0403908	8.890903	31
AGTAA	3287735	1.0336913	7.108481	9
TTAAT	10545910	1.0262378	13.935523	4
GTTGA	15104325	1.024741	13.08891	43
TGTAG	15067600	1.0222495	8.316443	21
GGAAC	462135	1.0124369	13.202691	27
ATCTC	148825	1.0085582	44.52994	40
AGTTG	14836915	1.0065988	9.601134	38
ATTTC	1900530	0.98306894	6.040561	22
GGTTG	27191230	0.9811397	7.1042356	42
CGTGT	2712925	0.9783538	6.4959345	41
TAAGT	7596260	0.96899927	6.5169053	7
TGGGG	19859550	0.9393579	8.668409	1
AAGGC	426290	0.9339082	14.306751	46
GAATA	2959710	0.9305576	5.434928	3
TTGGG	25749985	0.9291353	6.180571	36
TTATC	1793800	0.92786187	11.301629	37
GTTTG	33539180	0.92319787	6.8237433	18

CGAAC	41805	0.9153421	7.030933	9
TAGAC	545215	0.9111863	11.677272	25
GTGGT	25069315	0.90457475	7.691711	9
TGGTT	32661935	0.89905083	7.6226964	1
GGATA	5356335	0.8956762	7.352172	2
GGAGC	765370	0.8917842	9.4718895	27
TTTGG	32296280	0.8889858	5.048011	35
TGCGG	1870075	0.8840501	5.624181	5
AAGAC	214115	0.8819764	9.0381565	32
AGTGA	5257080	0.8790789	5.2852554	18
GGGGG	13961995	0.86570287	5.6369367	2
GGGTG	18254080	0.86341906	8.174479	2
GGTAT	12006930	0.8146007	6.05058	2
GAAGC	358045	0.78439844	9.5817175	4
TGGGT	21370475	0.7711097	9.193142	1
GGTAC	859020	0.7635402	12.219366	3
TGGTG	21085190	0.7608158	6.0660367	7
GGTAA	4513365	0.75471634	6.159883	2
CACAT	44365	0.7410304	6.923348	47
GTTGG	20483295	0.73909765	5.249419	39
GTGCG	1502785	0.7104192	5.248854	4
CGGCC	11430	0.70751554	5.331171	1
TGGTA	9746920	0.6612721	5.365911	1
GAGTC	729495	0.64841187	12.688161	21
TCTCG	155890	0.5618657	23.76384	41
CTCGT	155625	0.5609106	23.76992	42
TGAAC	321840	0.53787255	10.88205	20
TGGGC	1086795	0.5137661	5.684194	13
ATATC	402265	0.5128519	8.531081	38
TGGAT	7515985	0.50991607	5.1886373	1
GATTC	736510	0.49939892	5.8354173	29
CTGAA	268160	0.4481603	10.635801	19
CAATA	141770	0.44548646	19.808754	36
GGTGC	773220	0.36552823	5.2842984	3
GAACT	190600	0.31853878	10.707671	21
AGTCA	147345	0.24624917	10.497007	28

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	238632	0.31505238165682525	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	184219	0.24321396416423063	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	149246	0.19704108314372984	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	96994	0.12805571216945802	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	83274	0.10994196935067578	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82718	0.10920791388367558	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	81812	0.10801177314068602	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCC	80656	0.10648557148627551	TruSeq Adapter, Index 6 (100