FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD3_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.

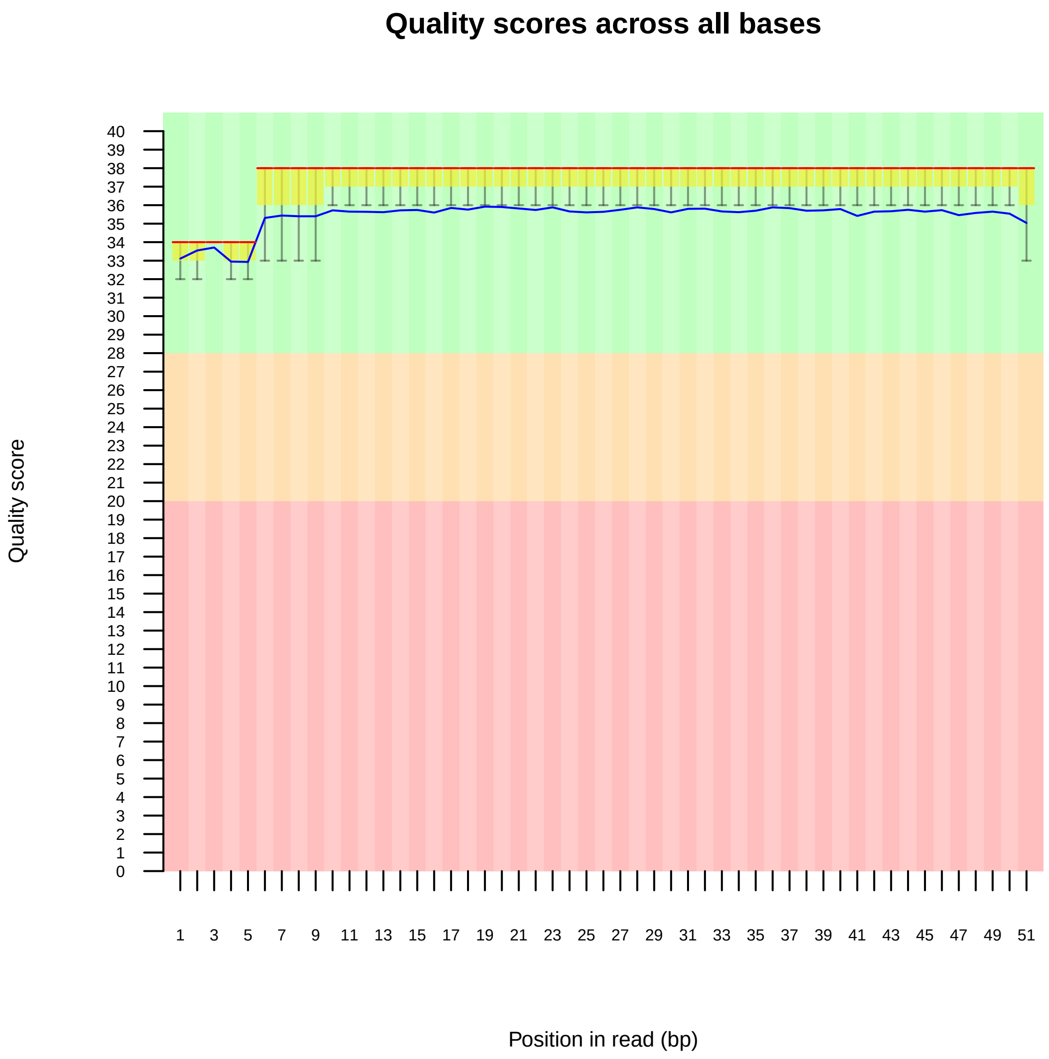
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

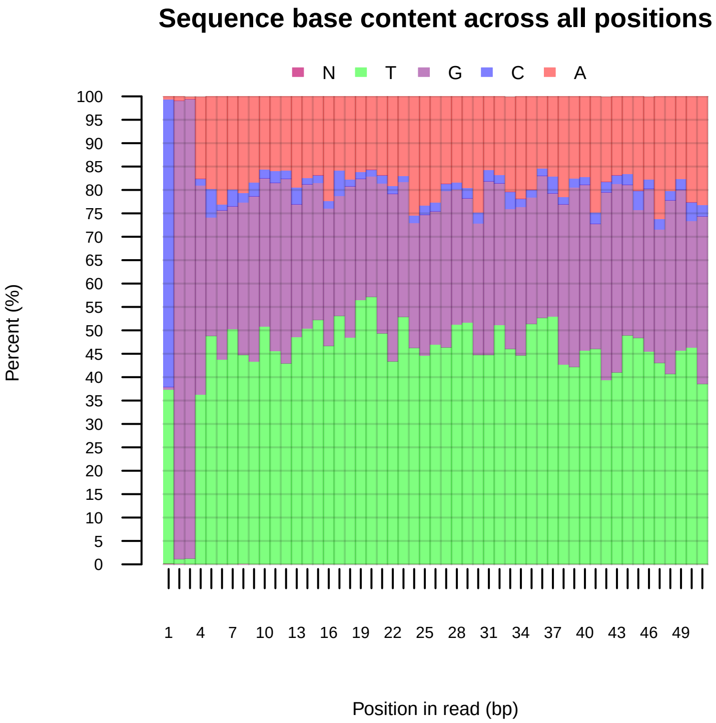
- Estimated Duplication rate 76.2011%

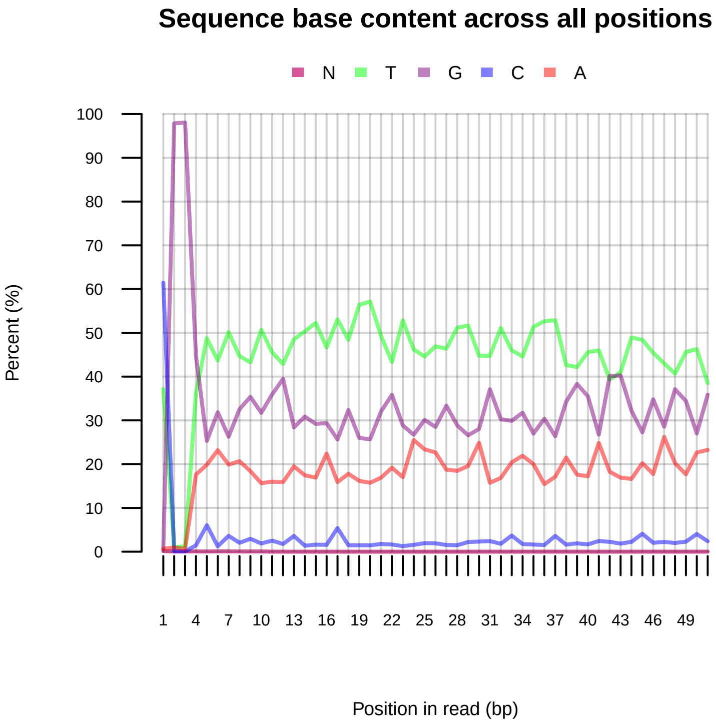
2 Per base sequence quality



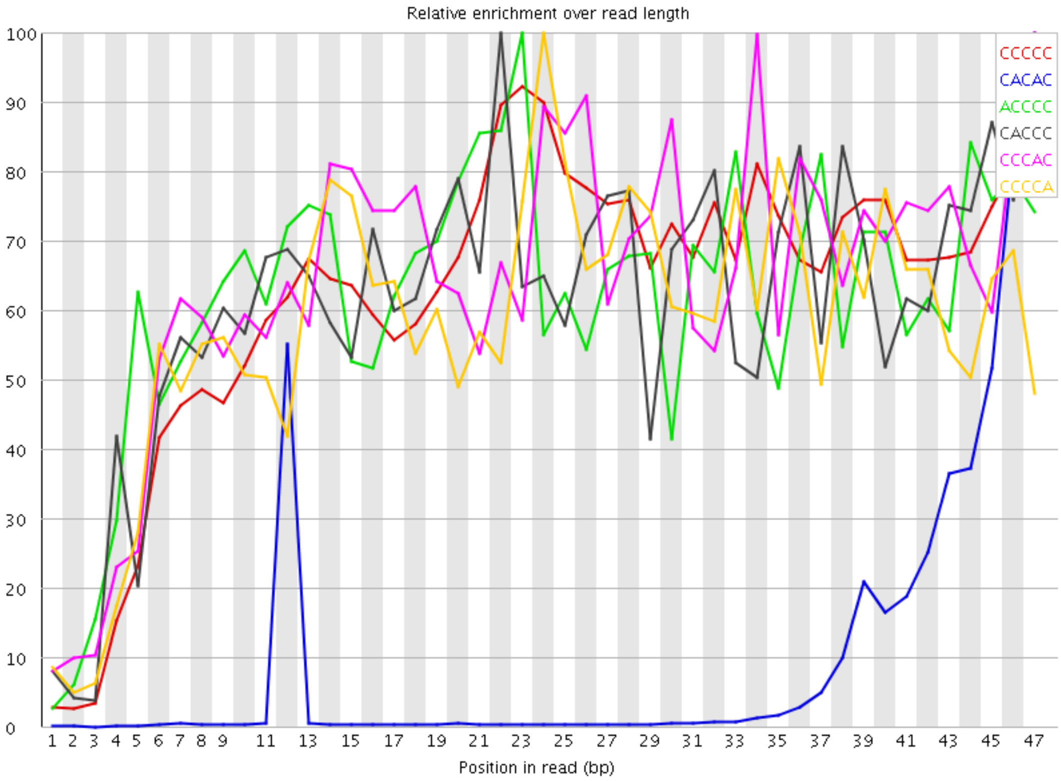
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	114430	830.5456	1321.7344	47
CACAC	513875	125.43485	1219.104	47
ACCCC	42190	56.156715	90.07055	23
CACCC	40990	54.559464	88.50682	22
CCCAC	40565	53.99377	83.50689	47
CCCCA	40445	53.834045	92.26018	24
CCACC	39980	53.21511	77.2513	35
CGGGC	4070650	28.873442	1051.5446	1
GCCCC	24805	17.866243	34.3565	47
AGCAC	722260	17.495436	126.10145	45
CCCGC	23240	16.739021	29.956036	34
CGCCC	23010	16.573362	30.463478	32
CCGCC	22885	16.483328	29.447567	27
CCCCG	22490	16.198822	30.462605	25
GCACA	621040	15.043563	125.94261	46
CGCGG	1781095	12.633447	317.5193	5
GCGCG	1681260	11.925309	315.72055	4
CGGAA	4269965	10.264216	212.09207	1
CGGCG	1372665	9.736421	272.9318	1
ACACG	396960	9.615634	101.28569	47
GGCGC	1298460	9.210079	316.35666	3
CGCGC	127845	9.137937	56.245094	13
CTCCC	16795	9.030739	14.276814	29
CCTCC	16655	8.955461	13.897709	28
TCCCC	16400	8.818346	16.94055	3
CCCTC	14945	8.035986	13.013034	22
CCCCT	14620	7.8612323	12.129037	38
CGGGA	6007210	7.8140497	240.73703	1
ACTCC	70045	6.9069924	256.85312	23
CTCCA	69235	6.8271203	256.64578	24
CGGGT	11890205	6.2480397	244.70288	1
AGACG	2577315	6.1953955	73.35956	27
TCGCG	1118780	5.9241867	26.602547	30
CGGAG	4368605	5.6825867	170.68391	1
AGATC	3065705	5.501501	26.57111	43
CGGTT	13145150	5.156669	182.91411	1
ACGCG	391915	5.1371803	14.583303	14
AACCC	20660	5.0430236	8.201811	32
CGCGT	935560	4.9539967	24.654167	31
CGGGG	6879860	4.842669	116.88244	1
CCACA	19575	4.77818	7.513342	15
ACACC	19545	4.770858	7.9147534	21
CGGAC	357610	4.6875143	157.3611	1

AAAAA	3110740	4.6830654	13.1391535	31
CCCAA	18985	4.6341634	7.283958	14
GAGAC	1915765	4.6051497	70.57224	26
CACCA	18830	4.5963287	6.4238405	36
AGGCG	3524050	4.584009	70.46573	47
ACCCA	18745	4.5755806	7.1120896	33
ACCAC	18685	4.5609345	7.513652	46
CGTCG	854985	4.527334	20.743555	41
TCGAG	4649040	4.514565	60.62756	44
GGGCG	6369200	4.48322	109.4475	2
CAACC	18105	4.419359	7.915022	31
CCAAC	18015	4.3973904	8.029711	30
CGGTC	800730	4.2400417	157.94542	1
CGCGA	318415	4.1737504	17.740398	5
ACGTC	401815	3.931954	27.696587	15
CGACG	281350	3.6879063	27.407804	24
TTACG	5075585	3.679493	50.222363	14
CGAGG	2803490	3.64672	77.29117	45
CGGTA	3718525	3.6109653	127.43839	1
GACGG	2704900	3.5184762	39.552605	28
ACGGG	2602300	3.3850162	39.434845	29
GATCG	3480100	3.379437	15.683177	44
TACGT	4639195	3.3631368	51.671703	15
ACGTT	4590840	3.3280823	53.303383	16
CGTTT	11205460	3.281577	38.42571	17
GGCGG	4627440	3.2572114	35.92715	11
CGGAT	3329685	3.2333727	112.740326	1
GGCGT	5930805	3.1165068	49.80495	3
AGAGA	6993735	3.083041	24.839325	25
CACGT	311630	3.0494497	26.983936	14
AGAGC	1264075	3.038606	17.36863	47
AAGCG	1205360	2.8974657	56.97804	8
GTCGA	2968155	2.8823004	60.03254	43
CGAGA	1157675	2.7828398	36.091904	25
AGCGA	1156370	2.7797027	57.68478	9
ATCGC	279340	2.7334766	39.33608	29
GAGGC	2096460	2.72703	57.811596	46
TTTCG	9286740	2.71967	15.362207	30
TCGGA	2656465	2.579626	14.085361	46
TTCGA	3555195	2.5773022	34.19658	31
TTTTT	158716295	2.5706527	5.6106725	16
GGAGG	19848320	2.5621085	30.628736	39
ATCGG	2632075	2.5559413	13.675265	45
GCGGG	3468920	2.4417403	36.5154	12
CGGTG	4611360	2.4231677	45.567375	1
GCGGC	340725	2.4167893	8.338349	33
CGTTA	3318465	2.4056869	29.547272	9
GGGAG	18155100	2.3435402	27.258852	38
TCGTT	7899060	2.313281	5.691021	36
TTTTA	56795000	2.2770936	12.075825	26
AGTAG	12759590	2.2722595	21.50972	35
CGTTC	568065	2.2455945	26.931232	33
GGAAG	9392140	2.2404566	12.317713	2
TTTAG	41433585	2.2252254	15.41085	27
GCGGA	1697915	2.2086112	23.467957	7
CGTAG	2254070	2.18887	25.011217	5
TTCGC	550475	2.1760602	6.7146955	33
TTCGT	7402490	2.1678581	5.603913	35
CGAGT	2222535	2.1582475	42.701977	33
GGTCG	4083230	2.1456473	34.14728	42
GAGGT	22237420	2.1429272	23.097816	40
ATTCG	2914860	2.1130977	36.147533	34
GAGCA	870780	2.0931964	13.454899	47
GAAGA	4706250	2.0746512	6.4734297	46
GACGC	157225	2.0608888	14.355288	3
TTTAC	3773030	2.0419326	36.49133	13
ATTTT	50878990	2.0399017	8.348359	25
AGGAG	8497990	2.0271606	8.853993	38
GCGGT	3822975	2.0088887	30.776314	6
ACGGA	828180	1.9907937	7.5765414	30
AAACG	447650	1.9885582	12.399912	7
CACGC	14985	1.9793352	6.859204	47
GAGAT	11066290	1.9707123	9.66543	26
GCGTT	5015295	1.967434	22.779364	16
TAAAA	3225680	1.9617281	5.3584204	30
AGGTC	2016665	1.9583324	57.248356	41
TAGAG	10992975	1.9576563	10.958853	24
AAAAT	3173555	1.930028	5.0733595	32
AATTT	19365190	1.9219475	18.40408	24
ACGGC	143475	1.8806553	11.128022	12
AAGAG	4256770	1.8765075	6.4963193	47
ATCGT	2578575	1.8693115	16.428387	39
GAAAA	2269875	1.849142	5.8161464	3
TACGC	187470	1.8344843	9.838517	13
AGCGC	139935	1.8342533	8.842658	35
TAGTA	13720930	1.8241225	17.20736	29
TAGTT	33843630	1.8176007	8.229334	29
TACGG	1871580	1.8174438	15.076041	5
GGAAA	4108625	1.811201	13.446777	2
CGCAC	13670	1.8056397	6.8281693	46
TTAGT	33457590	1.796868	14.7979	28
CGAGC	136645	1.7911284	5.2882066	13
GCGTA	1838575	1.7853936	24.735943	4
TCGTC	450730	1.7817624	8.497985	40
GTCGC	330160	1.74827	8.660103	3
GGAGA	7243305	1.7278607	11.160762	2
TATCG	2381710	1.726596	16.842424	38
AGGTA	9676140	1.7231512	25.624376	47
GTAGA	9665445	1.7212466	10.632306	23
GGACG	1303645	1.6957536	16.307241	2
AGCGG	1303020	1.6949406	5.2683263	6
GAGCG	1302610	1.6944072	9.235839	28
GCACC	12790	1.6894025	5.586682	47
TGGGA	17026200	1.6407437	15.667979	37
AGTTT	30346880	1.629805	8.29074	26
AGTCG	1657530	1.6095854	14.679053	22
TATTT	40124495	1.6087196	6.3276763	32

ACGCC	12150	1.6048664	5.4622974	16
AGTTA	12071445	1.6048326	17.476692	30
AACGC	65620	1.5895247	11.298256	11
CGTGG	3013525	1.5835406	32.200527	5
CGATT	2157900	1.5643475	19.723557	11
GCGTG	2942590	1.546266	32.35595	4
TAATT	15499315	1.5382689	18.0612	23
GGGAA	6437310	1.5355939	14.269148	2
TGGCG	2919655	1.5342143	30.74059	10
TAGGA	8518295	1.5169591	6.8118033	37
GTACG	1559335	1.5142307	14.807687	4
GGTTT	51814170	1.505816	9.554944	2
AGCGT	1548735	1.5039374	7.3656187	29
ACGGT	1547815	1.503044	14.546416	6
GCGAC	113405	1.4865007	17.949429	23
GTCGT	3787055	1.4856119	9.516993	3
ACGAG	616140	1.4810883	5.370929	32
AACGG	609800	1.4658478	7.0896907	8
AGGTT	20256370	1.4572496	13.64887	41
GCGTC	275175	1.4571123	8.863739	40
AAGTA	4369420	1.4379479	11.123935	34
CGTAC	145230	1.4211454	8.530239	13
TTCGG	3622470	1.4210472	20.156425	35
TTTAA	14282125	1.4174656	8.3662615	5
TTAAG	10621525	1.4120736	10.235323	6
TTGAG	19581780	1.4087192	12.384278	44
GTAGT	19554640	1.406767	9.308804	36
TTATT	35059280	1.4056389	6.1473637	32
GCGAT	1442910	1.4011734	24.262226	10
GGCGA	1076085	1.3997482	9.652516	2
TATAG	10520625	1.3986595	16.376284	47
AACTC	76315	1.3800355	48.484417	22
TTATA	13827780	1.372373	12.504081	46
AGATA	4169830	1.3722643	5.933491	26
AAAAC	166530	1.3670689	20.470331	6
GTTTA	25298095	1.3586556	8.073966	12
GGTTA	18797535	1.3523004	18.471188	2
TAAGC	750845	1.3474141	40.647457	7
GTTAA	10036290	1.3342699	23.154327	3
TCGAC	133725	1.3085636	6.033151	23
GAACG	543425	1.3062947	6.4106884	28
GGGTT	33538035	1.3056047	14.44104	2
GGAAT	7312295	1.3021916	9.987094	2
GGTAG	13505945	1.3015115	7.429831	2
GGAGT	13482790	1.2992802	10.556029	2
GACGT	1329845	1.2913789	5.948715	3
TTGTA	23810500	1.278763	14.176058	20
GAGTA	7136105	1.2708154	15.429783	34
ATTAT	12691185	1.2595688	12.348317	45
TCGGG	2359040	1.2396234	25.758226	36
TCGTG	3141775	1.2324768	7.9862185	40
TGGAA	6854730	1.2207074	9.319529	1
GTAAT	9161380	1.2179554	22.762674	22
GGGGA	9400120	1.2134087	10.26368	2
GGGAT	12558460	1.2102063	11.460084	42
TTTGT	54628885	1.185209	6.786941	19
GATTA	8855360	1.1772717	15.930472	44
GGTGG	22351680	1.165561	11.363496	8
CGTAA	632540	1.1351122	7.8544273	21
AGTAT	8529055	1.1338912	16.32948	30
GGGGT	21726500	1.1329601	8.255275	2
TGAGG	11660985	1.1237205	15.020495	45
TAGGC	1151090	1.1177943	7.8805213	13
GGATT	15444240	1.1110635	8.852561	43
GTATT	20610565	1.1069078	7.105425	31
AAGGC	459045	1.1034604	19.300365	46
TATTC	1994525	1.0794204	25.59885	33
AGTAA	3278795	1.0790303	7.293328	9
TGTAA	8106785	1.0777528	22.196821	21
TTAAT	10843190	1.0761598	15.528183	4
GGGTA	11152105	1.0746818	15.035432	2
TCGAT	1469850	1.0655527	6.43041	11
CGAAC	43945	1.0644877	9.965251	9
TTTTC	4836870	1.0574665	10.615973	29
CGTGT	2688720	1.0547494	7.629636	41
TGGAG	10932420	1.0535117	9.791349	1
CGTGA	1070860	1.0398849	8.022404	26
GTTAT	18567040	0.99715865	7.348762	31
TGTAG	13660760	0.98275924	7.0311036	21
AGTTG	13656380	0.98244417	9.28985	38
TAAGT	7351560	0.97734964	6.5864844	7
GTGGC	1855305	0.9749217	28.875206	9
TTGGG	24715550	0.96215355	6.9910154	36
GGTTG	24663220	0.9601164	6.5579715	42
TGGGG	18385165	0.9587213	8.395774	1
TTATC	1766860	0.95620996	12.18595	37
GTTGA	13228410	0.95165586	11.761707	43
TGCGG	1804375	0.9481592	7.066427	5
ATTTC	1750650	0.9474373	5.3883886	22
TAAGG	5248495	0.93466496	5.6631927	45
ATTAC	696630	0.93325907	5.8946047	29
GGGGG	13264120	0.92651826	5.838435	2
GGATA	5202035	0.92639136	7.8575706	2
AAGAC	207405	0.9213379	7.6113095	32
TTTGG	31301115	0.9096685	5.5755286	35
TGGTT	31160345	0.90557754	7.24824	1
TAGAC	502655	0.90202963	10.092598	25
GTTTG	31015375	0.90136445	6.805911	18
GTGGT	22992390	0.8950725	8.316222	9
GGAGC	685115	0.89118284	8.35588	27
AGTGA	4975075	0.88597375	5.702439	18
GGGTG	16813825	0.87678146	8.544173	2
CGTCT	215780	0.85299104	10.720486	16
GGTAC	862965	0.8380035	14.855463	3
CGGCC	11685	0.8352051	6.2740545	1
GGTAT	11303040	0.81314427	6.19875	2
GAAGC	333245	0.80106026	8.069568	4

CGATC	80695	0.78963953	5.0148044	44
GGTAA	4413635	0.78599113	6.5336685	2
GTGCG	1467730	0.7712597	6.65729	4
TGGGT	19736735	0.7683329	8.953083	1
AGTGG	7881830	0.75953907	5.293362	8
GTTGG	19281560	0.75061333	5.20816	39
GGAAC	310015	0.7452195	6.430682	2
TGGTG	18656105	0.72626495	5.688867	7
GAGTC	710330	0.6897835	13.45698	21
TCCAG	69505	0.68014	26.535437	25
ATGCC	68535	0.6706479	28.649637	47
CAGTC	66840	0.6540617	26.705927	27
GCATC	66645	0.6521535	27.554714	38
TGGTA	9016850	0.64867496	5.175776	1
GTCAC	66220	0.64799464	27.271757	29
CCAGT	65980	0.6456461	26.452816	26
ATCTC	71835	0.5247683	21.284435	40
TGGAT	7100760	0.51083094	5.098984	1
CATCT	68615	0.50124556	20.53947	39
GATTC	687245	0.4982112	5.346739	29
CACTT	67410	0.49244285	19.871918	31
TCACT	66095	0.48283648	20.028067	30
GGTGC	787000	0.4135511	6.718141	3
TGAAC	205100	0.36805817	5.309803	20
TCTCG	76870	0.30387166	11.563231	41
CTCGT	76250	0.30142078	11.579086	42
CTGAA	137920	0.24750163	5.065655	19
AGGCA	83535	0.2008029	6.8378277	36
GAACT	95055	0.17057909	5.065238	21
AGTCA	75310	0.13514608	5.053123	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	198718	0.2776040174086516	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGAGGTCGAGGCG	186185	0.26009573355825744	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	120927	0.1689319589225738	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	87553	0.12230932545707823	No Hit