FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_OD4_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.

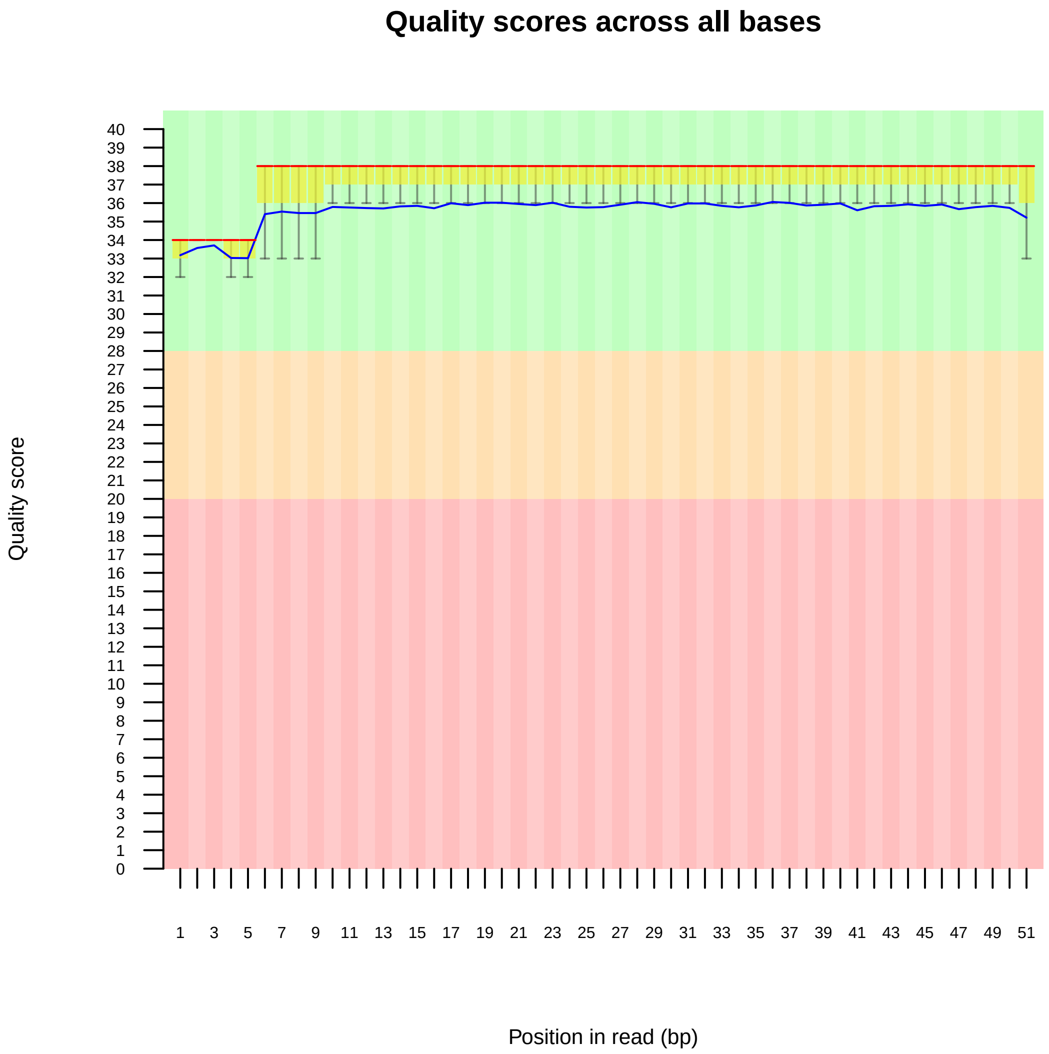
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

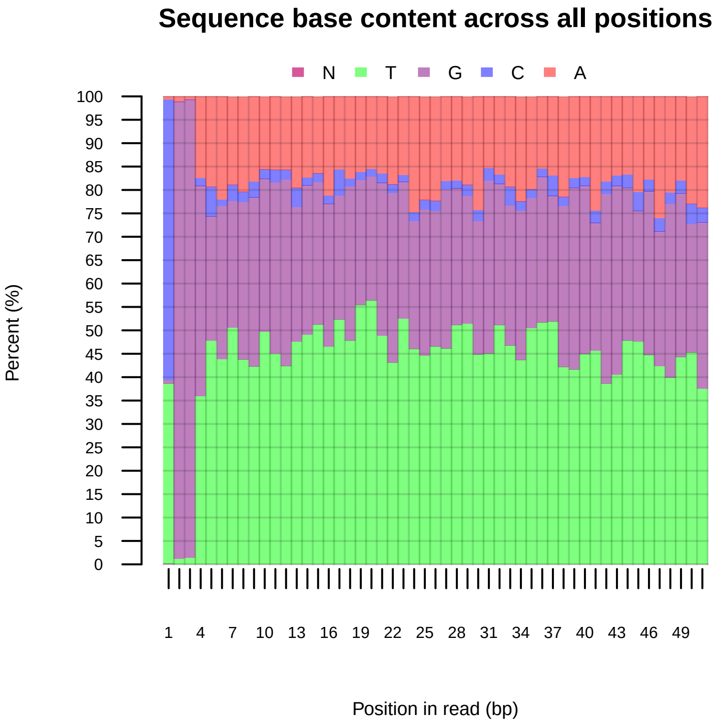
- Estimated Duplication rate 77.1047%

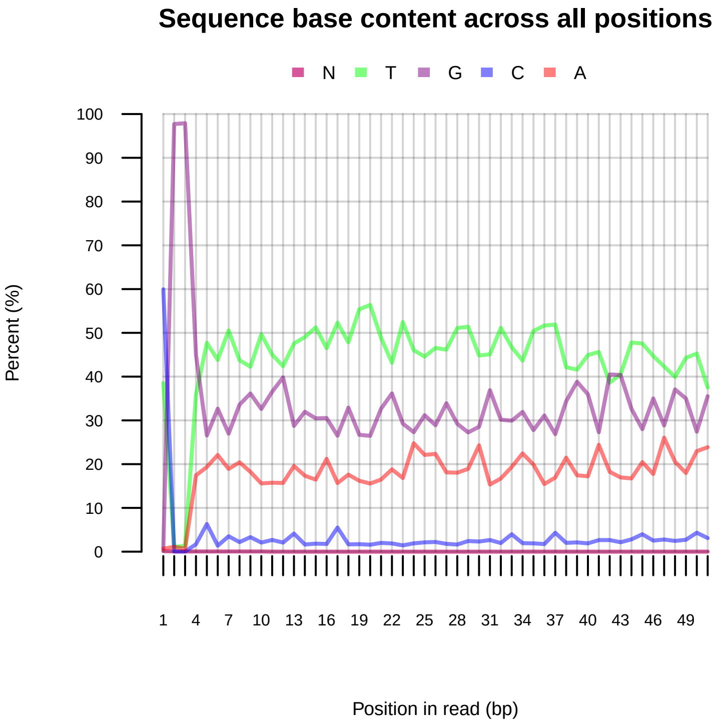
2 Per base sequence quality



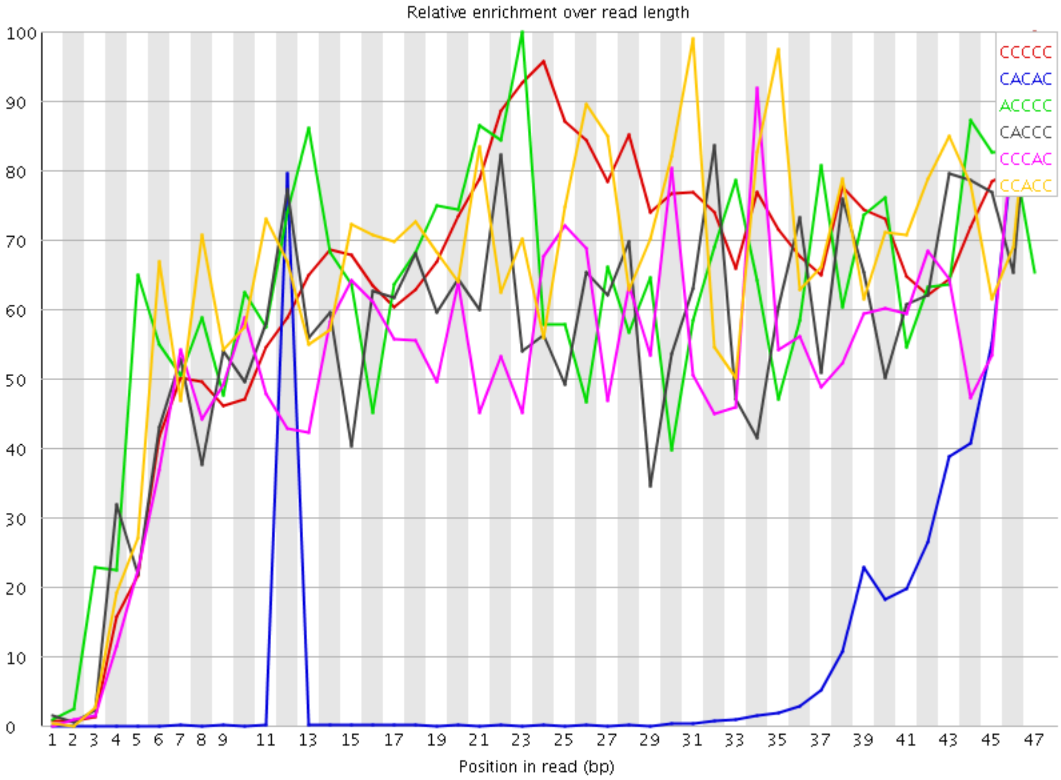
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	119600	638.71075	997.6126	47
CACAC	845880	178.09781	1617.4825	47
ACCCC	40255	42.68559	69.26416	23
CACCC	38935	41.28589	73.75219	47
CCCAC	37235	39.48324	75.24716	47
CCACC	36755	38.97426	60.546555	47
CCCCA	36135	38.316822	61.540646	24
AGCAC	1182255	26.031271	175.92784	45
CGGGC	4205265	25.684412	888.5614	1
GCACA	1024140	22.549843	175.60275	46
ACACG	681575	15.007137	143.26646	47
CGCGC	235495	13.753827	62.64572	13
GCCCC	24475	13.668792	27.42682	47
CCGCC	24245	13.540341	22.702467	35
CCCGC	23760	13.269479	25.195827	34
CGCCC	23720	13.247139	22.177526	44
CCCCG	23255	12.987446	24.53983	46
CGGAA	5260475	12.112773	182.92123	1
ACTCC	142240	12.083711	491.87735	23
CTCCA	137860	11.711616	491.32	24
CGCGG	1884780	11.511634	237.30705	5
GCGCG	1804185	11.019383	234.14793	4
CGGCG	1613645	9.855627	254.38304	1
AGATC	4499080	7.9364276	40.003716	43
GGCGC	1263330	7.716016	234.56346	3
CGGGA	6026660	7.30872	209.63492	1
TCGCG	1477470	6.913177	21.302849	30
ACCAA	137365	5.7426753	251.98737	36
CGGGT	11598300	5.675286	218.22977	1
ACGTC	626600	5.566771	54.73938	15
ACGCG	477230	5.53424	19.961098	6
AGACG	2390540	5.5044594	60.702293	27
CCTCC	12855	5.4999948	13.672232	28
CTCCC	12810	5.4807415	11.561148	29
TCCCC	12765	5.4614882	12.271988	3
CGTCG	1164505	5.448793	24.067905	41
CGCGT	1161280	5.4337034	19.542694	31
CCCTC	12665	5.418703	9.048038	47
CGGAG	4338750	5.261739	153.99309	1
CCCCT	11925	5.102095	9.148422	38
CGCGA	432115	5.0110598	16.474316	5
CACGT	543965	4.8326344	54.476627	14
CGGAC	412855	4.7877097	148.00871	1

CGGTC	996560	4.6629677	161.89265	1
GATCG	4988415	4.634576	22.177305	44
CGGTT	12279765	4.603267	155.61739	1
CGACG	392960	4.5569954	35.47684	24
AGAGC	1906435	4.3897586	26.112875	47
AAAAA	2643720	4.357429	13.418652	31
CGGGG	6817680	4.3545904	104.44905	1
GGGCG	6699130	4.27887	97.606964	2
TCGAG	4445585	4.13025	44.798428	44
AGGCG	3329065	4.037262	51.717686	47
GAGAC	1702575	3.92035	58.39539	26
AACCC	18260	3.8445942	6.9259667	22
ATCGG	4006830	3.7226167	20.618883	45
TCGGA	3938120	3.6587803	20.861525	46
ACACC	17210	3.6235201	6.480974	37
ACCAC	16155	3.4013927	6.233626	45
TTACG	4774700	3.3984132	37.981915	14
CGTTT	11802700	3.3895357	36.179314	17
CGGTA	3579035	3.3251657	114.05165	1
CAACC	15725	3.3108573	5.9861965	31
CCCAA	15615	3.2876966	6.0355124	15
ACCCA	15600	3.284539	6.382003	33
CACCA	15580	3.280328	5.5409827	44
GGCGG	5117550	3.2686832	38.29533	11
CCACA	15495	3.262431	5.986247	35
GAGCA	1414625	3.2573164	20.913345	47
CCAAC	15445	3.2519033	5.738816	30
GACGG	2656865	3.2220638	31.934238	28
GGCGT	6443440	3.1529076	47.915318	3
ACGTT	4381515	3.1185622	41.503826	16
TACGT	4365085	3.106868	39.86533	15
CGAGG	2542565	3.0834486	55.53372	45
ACGGG	2486765	3.0157783	31.958652	29
CGGAT	3208190	2.9806256	100.29421	1
GACCA	135255	2.978088	132.75272	35
GCGGC	484155	2.9570637	11.008057	9
AGAGA	6400685	2.9263985	21.780579	25
TTTCG	9946265	2.8563995	15.191782	30
CGTTC	789690	2.8307292	27.473658	33
AAGCG	1189885	2.7398298	51.549625	8
CGAGA	1171490	2.6974735	32.64636	25
AGCGA	1157575	2.665433	52.209282	9
TTCGC	742585	2.6618767	8.369789	33
TTCGA	3689440	2.625975	34.090656	31
ATCGC	294070	2.6125445	31.479454	29
TTTTT	147848575	2.606009	5.720424	16
GTCGA	2786910	2.5892282	44.20297	43
AACTC	152445	2.5714617	100.39329	22
GAAGA	5483390	2.5070105	9.869253	46
TCGTT	8666955	2.4890032	6.0761285	4
GCGGG	3882870	2.4800677	39.172157	12
GGAAG	10248585	2.4678395	11.557359	2
GGAGG	19439595	2.4653935	28.501497	39
CGTTA	3439875	2.4483457	31.32986	9
CAATC	136770	2.307054	104.537964	38
TTTTA	52711965	2.3027096	12.531498	26
TTCGT	8013140	2.3012385	5.755953	35
GTCGC	489285	2.2893958	10.715869	3
ATTCG	3205765	2.281717	40.390385	34
AGAAA	2626370	2.2799053	5.513682	22
GAGGC	1875595	2.2745929	42.68333	46
GGGAG	17692475	2.243818	24.845453	38
TCGTC	625900	2.2436066	9.459577	40
TTTAG	39327640	2.2425656	15.637276	27
GCGGA	1848110	2.2412612	22.414648	7
CCAAT	132530	2.235533	101.570335	37
AAGAG	4884840	2.2333527	9.908984	47
GAGAT	12079650	2.2283843	8.821668	26
CGGTG	4535555	2.21934	43.579098	1
AGTAG	11973940	2.2088838	22.11367	35
CGAGT	2312545	2.148511	41.41824	33
CGTAG	2291665	2.129112	22.254198	5
GAGGT	21838520	2.1218026	22.116673	40
GACGC	182330	2.114406	18.11275	5
TACGC	234410	2.0825198	10.00319	13
GCGTT	5509420	2.0652947	25.504757	16
AGGAG	8496260	2.045883	9.562123	38
GGTCG	4163030	2.0370562	24.918018	42
ATTTT	45962580	2.0078642	7.791676	25
CGAGC	172035	1.9950191	9.106534	32
TAGTT	33622345	1.9172347	9.837966	29
ACGGA	829095	1.9090747	7.570838	30
TAAAA	2870035	1.9086698	5.5139217	30
AAACG	436035	1.9063115	13.349358	7
GCGGT	3873960	1.8956081	24.500158	6
TTTAC	3472120	1.8932483	28.256273	13
GCGAC	162210	1.8810827	20.329634	23
AAAAT	2820335	1.8756179	5.287362	32
AATTT	17136320	1.8553196	17.096777	24
TAGAG	10037950	1.8517433	9.690608	24
CACGC	16635	1.8446699	7.2958455	47
ATCGT	2556910	1.8198917	14.067253	39
AGCGC	156285	1.8123728	10.095735	35
TTAGT	31567420	1.8000574	15.03985	28
ACGGC	154520	1.7919049	8.055864	6
GCGTC	380485	1.7803136	11.275829	40
AGGTA	9614930	1.773707	28.237564	47
GAAAA	2027075	1.759668	5.577628	3
GGAAA	3826505	1.749481	12.389661	2
GCGTA	1876300	1.7432097	21.881937	4
GGACG	1434740	1.7399544	16.248808	2
GGAGA	7193840	1.7322628	11.094044	2
TGAGA	9330805	1.7212931	6.013049	41
GAGCG	1412925	1.7134986	9.517182	28
TAGTA	12070385	1.7058452	13.950617	29
AACGC	77405	1.704328	8.257328	11
AGTTA	11964335	1.6908578	21.403107	30

TATCG	2371915	1.6882207	14.399348	38
AGTTT	29547620	1.6848831	8.743623	26
TACGG	1799765	1.6721036	11.785824	5
CGCAC	14945	1.657264	6.4620333	47
TCGAC	186530	1.6571496	9.535455	23
AGCGG	1366070	1.656676	6.032603	6
TCGTA	2289995	1.6299138	5.9530125	43
GTAGA	8831475	1.6291796	9.370778	23
TATTT	37062890	1.6190835	5.4444904	32
TAGCG	1742200	1.618622	5.013843	10
GTCGT	4302405	1.6128256	10.415882	3
CGATT	2259035	1.6078779	18.741955	11
TGGCG	3277910	1.6039486	33.45839	10
AGGTC	1716255	1.5945172	41.977837	41
AGTCG	1712490	1.5910193	13.171453	22
TGGGA	16364135	1.5899183	13.502739	37
TTGAG	21188775	1.5771402	13.896162	44
GCGTG	3206590	1.5690502	32.25733	4
CGTAC	175345	1.5577809	8.301896	13
CGTGG	3165810	1.5490957	32.03705	5
TAGGA	8275900	1.5266905	7.5716524	37
GGGAA	6329055	1.5240242	13.664617	2
CGAAA	348400	1.5231781	7.859721	32
TCGAA	856950	1.5116694	5.2893515	32
TTCGG	4001840	1.5001543	21.988869	35
AGCGT	1598820	1.485412	7.7037044	29
CGAAC	67375	1.4834844	6.767135	29
GGTTT	49388325	1.4832611	9.400595	2
AGGTT	19719195	1.4677554	14.212538	41
ACGCC	13230	1.4670864	10.891168	23
TTATT	33551435	1.4656864	7.501948	32
GTACG	1572150	1.4606338	11.542272	4
TAATT	13457745	1.4570467	16.745712	23
GTAGT	19540475	1.4544526	9.55226	36
ACGGT	1537195	1.4281583	11.354025	6
AAGTA	4068370	1.4249843	11.597566	34
TTTAA	13092145	1.4174639	8.48567	5
TATAG	10000465	1.413314	17.160334	47
AACGG	610800	1.4064285	7.3804708	8
TTATA	12867380	1.393129	13.413342	46
GCACC	12465	1.382255	6.0711856	47
CGTCT	385375	1.3814185	21.497795	16
CGCCA	12445	1.380037	10.734868	24
TTAAG	9742985	1.3769257	10.181748	6
GCGAT	1471630	1.3672439	22.078205	10
GTTTA	23807175	1.3575478	8.362057	4
AGATA	3803250	1.3321238	5.148883	26
GGCGA	1093140	1.3256853	8.499806	2
GACGT	1422870	1.3219426	6.238142	3
AAAAC	157950	1.3111311	22.849186	6
TCGGG	2658875	1.3010422	27.472952	36
ACGAC	58875	1.2963287	5.458068	23
GGAGT	13302485	1.2924525	10.4842615	2
CCAGC	11635	1.2902153	5.263292	28
GAGTA	6978560	1.2873646	15.851009	34
GAACG	558865	1.286843	6.171637	28
GGTAG	13193100	1.2818248	7.277197	2
CCCAG	11540	1.2796807	6.5400133	27
GGAAT	6918625	1.2763081	9.576113	2
GGGTT	32456670	1.2723732	14.251341	2
ATTAT	11740510	1.2711246	13.372851	45
ATGCC	143055	1.2709136	57.712116	47
GGTTA	17064280	1.2701426	16.410435	2
TTGTA	22248300	1.2686566	14.348466	20
TAAGC	712665	1.257149	37.74545	7
TCCAG	139115	1.2359103	53.323975	25
GTTAA	8697505	1.2291734	20.106428	3
CGTAT	1724925	1.2277228	5.437753	44
CAGTC	137935	1.225427	54.286972	27
CCAGT	136825	1.2155657	53.601864	26
TATTC	2222995	1.2121359	29.784718	33
TTTGT	52600785	1.2102311	6.901282	19
GTCAC	135435	1.2032168	54.124622	29
GGGAT	12370270	1.2018796	11.478873	42
TGGAA	6512420	1.2013737	9.165269	1
GGGGA	9452025	1.198737	10.141344	2
TCGTG	3186630	1.1945596	6.2618446	40
GATTA	8403385	1.187607	16.74666	44
CTGAC	131820	1.171101	53.19834	33
CACTG	131565	1.1688355	53.26076	31
TGACC	130710	1.1612395	52.718506	34
GTAAT	8180220	1.1560682	20.597164	22
CGTAA	655255	1.1558771	8.869039	21
TGAGG	11847235	1.1510623	16.058094	45
GGTGG	22477680	1.1502157	10.809535	8
TCGAT	1615075	1.1495366	6.894333	11
TTTTC	5115735	1.1255108	10.795306	29
GGATT	14977315	1.1148039	9.077308	43
CGTGA	1171215	1.088138	8.151747	26
GGGGT	21109865	1.0802226	8.092006	2
TAGGC	1159280	1.0770497	8.80256	13
CGATC	120640	1.0717766	6.878399	44
GTATT	18741595	1.0686951	5.7625046	31
AGTAT	7525515	1.0635422	13.07041	30
GTGGC	2159645	1.0567585	31.522413	9
GTTAT	18503520	1.0551194	8.929518	31
TGGAG	10802075	1.0495158	9.949588	1
GGGTA	10758510	1.0452831	14.500092	2
AGTAA	2977890	1.0430337	7.0454707	9
GTTGA	13902640	1.0348128	13.006804	43
TGTAA	7253080	1.0250403	19.941944	21
TGTAG	13736670	1.022459	8.2625675	21
TTAAT	9387340	1.016351	13.695184	4
CGTGC	216095	1.0111222	5.0980587	13
AGTTG	13546360	1.0082939	9.563508	38
CGTGT	2676690	1.0034003	5.993048	41
ATTTC	1833650	0.99983716	5.8393054	22

ATCTC	146110	0.99443316	44.05201	40
TAAGT	6892615	0.9740976	6.450322	7
GGTTG	24847730	0.974086	6.9552097	42
AAGAC	217075	0.9490352	8.941447	32
AAGGC	405175	0.93295616	13.029136	46
TTATC	1710435	0.9326516	10.687885	37
TGGGG	18134060	0.9279463	8.465294	1
GTTTG	30804980	0.9251545	6.77618	18
TGCGG	1889705	0.9246714	5.504847	5
TTGGG	23488450	0.9207992	6.0817127	36
GGATA	4928825	0.9092413	7.4160433	2
TAGAC	508120	0.8963294	10.580052	25
TCACT	131665	0.89611983	40.602787	30
GTGGT	22832880	0.8950993	7.5047994	9
GGAGC	738025	0.8950261	8.685946	27
AGTGA	4824425	0.8899822	5.3404875	18
TGGTT	29625780	0.8897399	7.4438334	1
GGGGG	13003600	0.86857784	5.6754503	2
GGGTG	16689990	0.8540512	8.090996	2
GGTAT	11008620	0.8194027	6.0567904	2
GAAGC	346685	0.79827714	9.456419	4
GGTAC	841785	0.78207535	11.576572	3
GGTAA	4148540	0.76529884	6.2079844	2
TGGGT	19504590	0.764623	9.058732	1
TGGTG	19169220	0.75147575	5.950509	7
CGGCC	12710	0.74231356	7.421453	1
GTTGG	18582075	0.7284584	5.1711564	39
GAGTC	734525	0.68242353	12.107889	21
GGAAC	288590	0.6645075	5.676734	2
TGGTA	8894555	0.6620468	5.2485504	1
CACAT	37280	0.62884384	5.3310275	47
TGAAC	314655	0.5550549	11.260193	20
TCTCG	153800	0.55131286	23.268774	41
CTCGT	153320	0.5495922	23.284954	42
GATTC	734945	0.52310026	5.849425	29
TGGAT	6853820	0.5101491	5.183481	1
TGGGC	1037880	0.50785595	5.4207106	13
CTGAA	251090	0.4429256	10.977901	19
GGTGC	802090	0.39247912	5.029888	3
GAACT	181445	0.320071	10.971281	21
AGTCA	146195	0.2578896	11.037392	28
ACTGA	137345	0.24227813	10.855585	32
AATCT	150315	0.20313573	8.681473	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
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	217417	0.31250089833389966	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	161502	0.23213235433623616	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	139007	0.19979952062028444	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	87230	0.12537866570537753	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC	79607	0.1144218667982115	TruSeq Adapter, Index 4 (100CGGGTTT
76258	0.10960823443036433	No Hit	
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	74950	0.1077282012451914	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	70497	0.10132775187701479	No Hit