FASTQ QC Report

Report Date	`10-02-16`
Run ID	`160930_D00796_0121_AC9MR4ANXX`
Project ID	`EC-EL-3883`
Sample	`Sample_YD10_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.

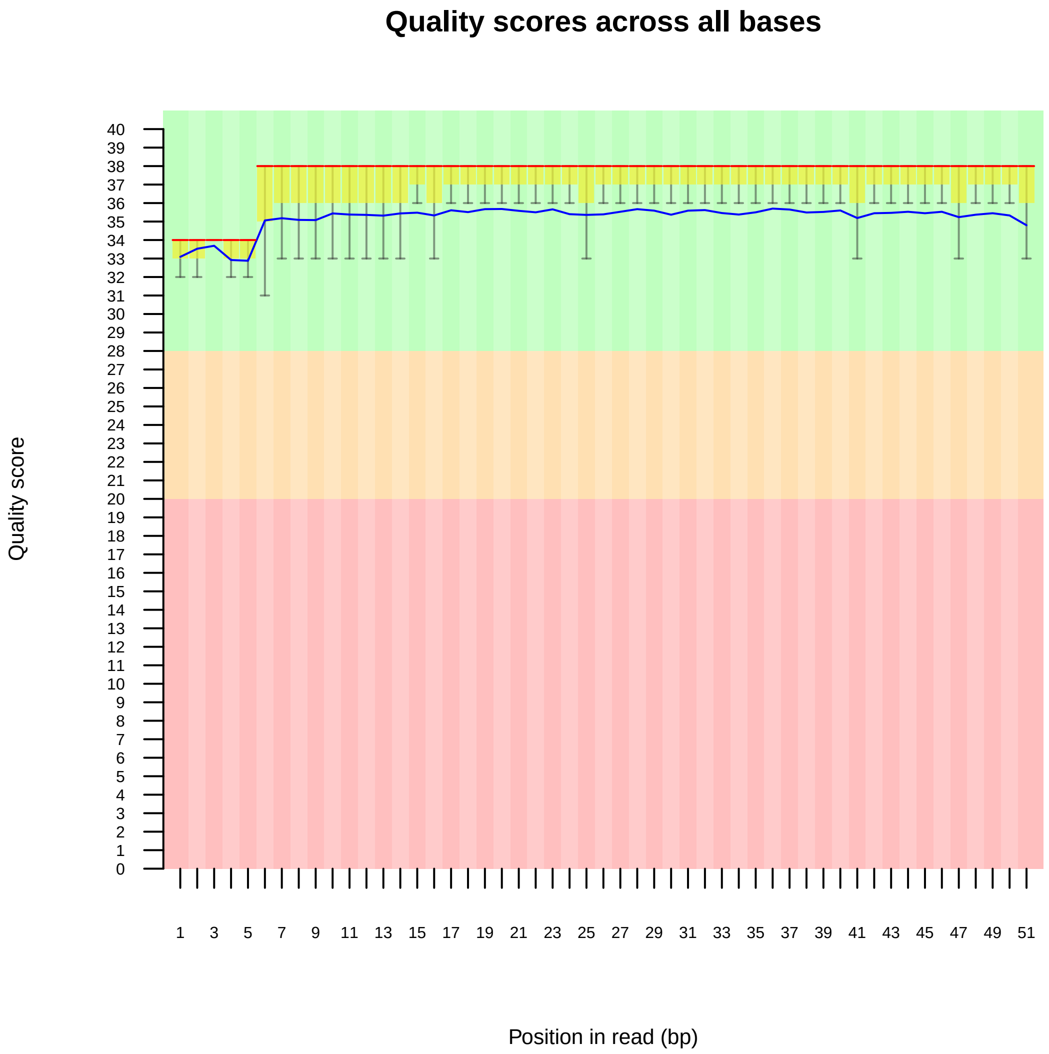
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

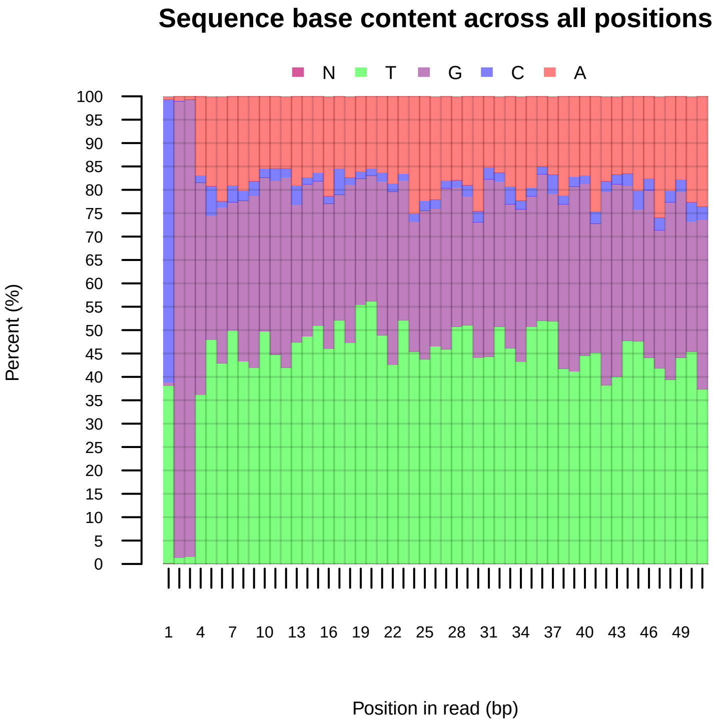
- Estimated Duplication rate 73.4677%

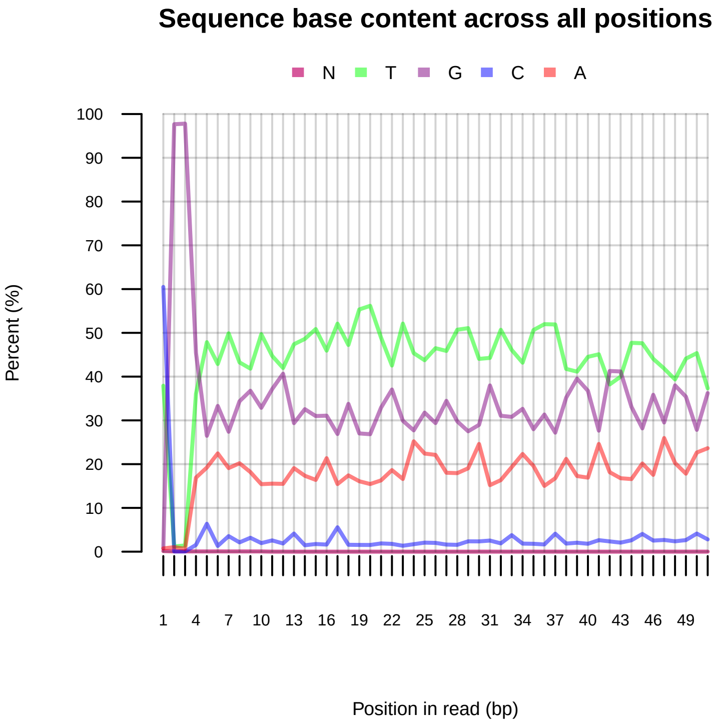
2 Per base sequence quality



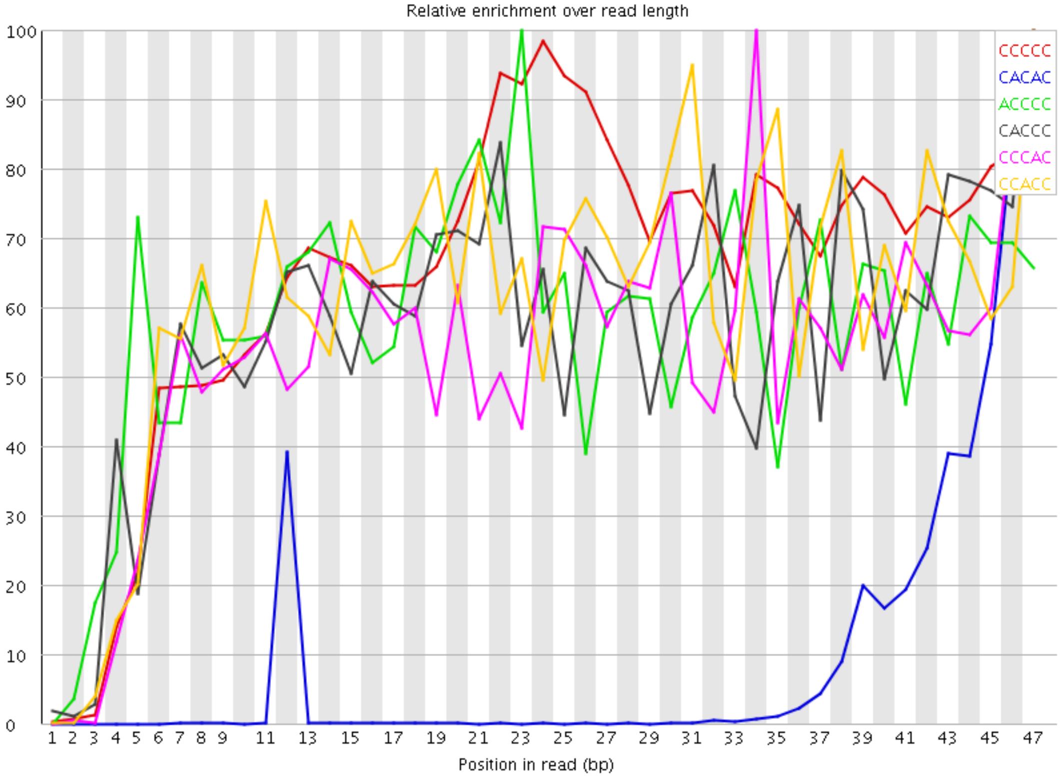
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	116095	616.5409	935.8866	47
CACAC	863980	174.1141	1764.074	47
ACCCC	41160	42.580883	73.16421	23
CACCC	40675	42.07914	72.92462	47
CCCAC	38935	40.27907	74.86881	34
CCACC	38390	39.71526	64.659836	47
CCCCA	37255	38.541077	61.74025	24
CGGGC	4885025	26.253693	917.1923	1
AGCAC	1229895	24.884178	185.39977	45
GCACA	1060105	21.448858	185.01091	46
ACACG	680120	13.760712	152.88258	47
GCCCC	25305	13.492106	26.434402	47
CGCCC	24545	13.08689	22.550123	40
CCCGC	24450	13.036238	23.051802	47
CCGCC	24180	12.89228	24.052935	22
CCCCG	24140	12.870953	24.179346	46
CGGAA	5849310	11.881875	184.8879	1
CGCGG	2052190	11.029126	246.72946	5
GCGCG	1950920	10.484869	244.65637	4
CGCGC	194690	10.421785	59.617897	13
CGGCG	1730525	9.300396	248.30101	1
AGATC	4821560	7.6739655	37.317	43
GGCGC	1427385	7.671225	245.25851	3
CGGGA	6911815	7.236134	207.96045	1
ACTCC	83535	6.798057	262.72168	23
CTCCA	79595	6.4774203	261.93924	24
TCGCG	1453835	6.121961	22.02454	30
CGGGT	13505935	5.709857	219.72653	1
AGACG	2764610	5.6158333	65.205215	27
TCCCC	12670	5.293001	12.375042	5
CCTCC	12205	5.0987434	8.736093	28
CGGAG	4852740	5.0804424	147.46086	1
ACGCG	479070	4.995609	18.116894	6
CTCCC	11940	4.988037	9.227134	47
CCCTC	11905	4.973416	8.637951	39
CGCGT	1170190	4.92756	20.242662	31
CCCCT	11585	4.839733	8.441698	38
CACGT	592285	4.8391833	39.970142	47
ACGTC	588005	4.8042145	28.437944	46
CGTCG	1137765	4.7910204	22.050226	41
CGGTT	14205000	4.70537	161.34805	1
CGGAC	437735	4.5645785	142.164	1
CGGTC	1054055	4.438526	158.6638	1

CGCGA	422715	4.407954	17.2736	5
GATCG	5289825	4.3391805	20.48137	44
CGGGG	8012605	4.3233695	103.19724	1
GGGCG	7761565	4.187916	96.564644	2
AAAAA	2802925	4.1755576	11.920072	31
TCGAG	5013930	4.1128674	47.4915	44
GAGAC	2021025	4.105367	62.697617	26
AGAGC	2019935	4.103153	24.27896	47
AGGCG	3817920	3.9970658	53.371296	47
CGACG	376425	3.9252553	29.570328	24
AACCC	19285	3.8864217	6.723982	32
ACACC	17480	3.522668	5.587568	44
TCGGA	4268560	3.5014489	19.147545	46
ATCGG	4248105	3.4846702	18.889559	45
TTACG	5364120	3.4476	41.0177	14
ACCAC	16810	3.3876457	5.6349626	46
CGTTT	13050760	3.3871953	38.393017	17
CAACC	16725	3.370516	6.4872184	31
CGGTA	4076895	3.3442285	114.5971	1
CCACA	16185	3.2616918	6.0137386	35
ACCCA	16040	3.2324705	6.2031283	33
CACCA	15995	3.223402	5.3981256	32
ACGTT	4989450	3.206794	44.75105	16
CCAAC	15885	3.2012346	6.5819006	30
GACGG	3053835	3.197128	33.815666	28
TACGT	4931295	3.1694171	42.974976	15
GGCGG	5833040	3.1473396	37.76789	11
CCCAA	15550	3.133723	5.776782	15
GGCGT	7389750	3.1241388	48.268185	3
CGAGG	2943090	3.0811875	57.73845	45
ACGGG	2903095	3.0393157	33.83704	29
GAGCA	1479950	3.0062659	19.439587	47
AGAGA	7464470	2.9537263	23.04315	25
CGGAT	3537985	2.902167	97.5465	1
TTTCG	10830625	2.8109813	15.067196	30
AAGCG	1329905	2.7014754	51.027946	8
CGAGA	1296945	2.6345227	33.537254	25
GTCGA	3177125	2.6061578	47.049397	43
GCGGC	481960	2.5902078	9.297658	9
ATCGC	316405	2.5851436	33.570766	29
AGCGA	1270775	2.5813625	51.53394	9
TTCGA	4011195	2.5780547	33.284767	31
TTTTT	160789805	2.5721238	5.735433	16
CGTTC	762115	2.5144784	26.04887	33
CGTTA	3796180	2.4398615	30.293753	9
TCGTT	9306560	2.4154253	6.079293	4
GGAGG	22977660	2.415157	27.82542	39
GCGGG	4474515	2.4143188	38.552296	12
GAAGA	6048410	2.3933847	9.278565	46
TTCGC	715395	2.3603332	8.618353	33
GGAAG	11543810	2.354261	10.860011	2
TTTTA	58600260	2.3213832	13.177547	26
ATTCG	3595705	2.3110132	42.692547	34
AGAAA	2978070	2.2865045	5.3282213	22
GAGGC	2165900	2.2675292	44.128265	46
TTTAG	44498650	2.2497911	16.147121	27
TTCGT	8651245	2.2453449	5.621982	35
CGGTG	5219870	2.2067862	42.970833	1
AGTAG	13770005	2.2003484	20.817135	35
GGGAG	20900655	2.196845	24.349047	38
GAGAT	13491795	2.1558924	9.054224	26
GCGGA	2057640	2.1541893	21.96485	7
AAGAG	5406415	2.1393442	9.27726	47
CGTAG	2591360	2.1256618	23.25835	5
GAGGT	25398050	2.0916636	21.915487	40
CGAGT	2539245	2.0829127	39.709034	33
ATTTT	51565840	2.0427225	8.248914	25
GCGTT	6147175	2.036236	26.082624	16
TCGTC	611805	2.0185542	8.777926	40
AGGAG	9824995	2.0037234	9.219838	38
ACGGA	980530	1.9917796	9.953504	30
GGTCG	4694405	1.9846371	26.054968	42
GTCGC	470905	1.9829363	9.767347	3
TTTAC	3906870	1.96743	31.22634	13
GACGC	187535	1.9555626	16.216982	5
TAGTT	37789920	1.9106069	9.95334	29
AATTT	19465410	1.9095227	18.223629	24
TACGC	232585	1.9003038	10.33969	13
CGAGC	181865	1.8964376	7.2917256	32
GCGGT	4439935	1.8770558	25.222153	6
TAGAG	11695455	1.8688502	10.233543	24
CACGC	17920	1.8612399	8.736974	47
ATCGT	2888660	1.856585	15.532535	39
AAACG	469445	1.8502523	10.943787	7
TTAGT	36019160	1.8210797	15.552312	28
AGCGC	172550	1.7993033	10.739188	35
ACGGC	170905	1.7821498	8.164044	12
CGCAC	16985	1.7641273	8.663759	47
TAGTA	14072515	1.761899	14.868383	29
AGGTA	11014635	1.7600598	28.029497	47
GAAAA	2288750	1.7572579	5.228051	3
GCGTA	2132115	1.7489487	22.925627	4
TATCG	2678475	1.7214957	15.950966	38
GGAAA	4332835	1.7145234	11.439584	2
AGTTA	13567990	1.6987317	21.473228	30
GGAGA	8255535	1.6836455	10.154019	2
GAGCG	1602725	1.6779286	9.465917	28
AGTTT	33102055	1.6735948	8.501345	26
TACGG	2039975	1.6733673	12.510495	5
GGACG	1588955	1.6635126	14.994709	2
AGGTC	2021860	1.6585077	44.72455	41
GTAGA	10368765	1.6568546	9.910785	23
AGCGG	1573290	1.6471124	6.2553906	6
TATTT	41382850	1.6393349	5.835207	32
TGAGA	10258710	1.6392686	5.4542217	41
TAGCG	1953630	1.6025395	5.120283	10
TGGCG	3770220	1.5939227	33.49163	10

GCGAC	152805	1.5934079	17.623507	23
TCGTA	2455585	1.578241	6.303207	45
AGTCG	1922835	1.5772787	13.691293	22
GCGTG	3716490	1.5712075	32.754704	4
CGATT	2444555	1.5711521	18.526886	11
CGTGG	3698320	1.5635256	32.570747	5
GCGTC	369620	1.556435	10.381932	40
TGGGA	18831175	1.5508469	13.568657	37
GTCGT	4676150	1.5489626	10.162421	3
AACGC	76240	1.5425464	6.2988906	23
TTGAG	23856750	1.5394114	13.750276	44
TTCGG	4627025	1.53269	22.749695	35
TAATT	15549390	1.525368	17.868036	23
CGAAA	382395	1.5071568	6.0121894	32
TAGGA	9428450	1.506599	7.4140506	37
GGGAA	7307205	1.4902416	13.173412	2
CGTAC	182280	1.4892937	8.861495	13
AACGG	731820	1.4865676	8.682913	29
AACTC	93005	1.4743941	53.10047	22
GGTTT	56496185	1.4721384	9.003194	2
TCGAC	179770	1.4687861	7.8784633	23
AGCGT	1789825	1.4681721	7.7381096	29
TTATT	37056955	1.4679694	7.668208	32
GTACG	1788130	1.4667819	12.284871	4
AGGTT	22601025	1.458383	14.249802	41
GTAGT	22398500	1.4453146	9.034654	36
GCACC	13665	1.4192994	6.8821974	47
AAGTA	4563275	1.4148157	11.069822	34
TTTAA	14349150	1.4076265	8.476525	5
ACGGT	1699475	1.3940593	11.954295	6
GGAAT	8715800	1.3927226	9.608141	2
TATAG	11114690	1.3915751	16.458683	47
GAACG	683700	1.3888198	8.50626	28
TTATA	14104285	1.3836057	13.132655	46
TTAAG	10959325	1.3721231	9.953746	6
GTTTA	27014225	1.3658026	8.113242	4
AGATA	4355320	1.3503404	5.5174026	26
CGAAC	66730	1.3501326	5.1438365	29
ACGCC	12900	1.3398435	5.002776	23
GCGAT	1612085	1.3223742	21.703157	10
TCGGG	3093625	1.3078809	27.855398	36
TTGTA	25822215	1.3055359	14.871359	20
GGTTA	20092510	1.2965152	17.193298	2
TGGAA	8103230	1.2948382	8.642582	1
GGCGA	1230160	1.2878821	8.289226	2
GACGT	1566245	1.2847722	5.846932	3
GTTAA	10224500	1.280122	21.902246	3
GGTAG	15508520	1.2772086	7.171767	2
ATTAT	12996975	1.2749805	13.068923	45
GGGTT	38006430	1.2639638	13.507119	2
GAGTA	7906410	1.2633878	14.977072	34
GGAGT	15333340	1.2627817	9.821162	2
TATTC	2502675	1.2603025	31.891655	33
TAAGC	789660	1.2568182	37.923454	7
TTTGT	60009500	1.2251843	7.131126	19
TCGTG	3660010	1.2123688	7.025929	40
AAAAC	158130	1.2092786	18.798687	6
CGTAT	1845420	1.1860789	5.2880692	13
GTAAT	9447165	1.1827985	21.479979	22
GGGGA	11234025	1.1807961	9.6356535	2
GATTA	9299410	1.1642994	16.056385	44
GGGAT	14099470	1.161166	10.642961	42
CGTAA	729285	1.1607256	8.998223	21
GGTGG	27318170	1.1595173	10.636736	8
TGAGG	13782410	1.1350545	15.751238	45
CGTCT	342120	1.1287711	11.25041	47
TTTTC	5526120	1.1237681	10.907399	29
TCGAT	1736490	1.116068	6.4782705	11
GTATT	21619065	1.0930305	6.1418424	31
GGGGT	25723155	1.091817	7.763463	2
GGATT	16837695	1.086491	8.574187	43
AGTAT	8650555	1.0830618	13.982188	30
TAGGC	1320160	1.0829116	9.039391	13
CGTGA	1293890	1.0613625	7.4448338	26
TGTAA	8453770	1.0584241	20.782993	21
TTAAT	10788850	1.0583673	15.336826	4
GTTAT	20882195	1.0557753	9.007947	31
GTGGC	2491740	1.0534242	31.650146	9
AGTAA	3378870	1.0475981	6.862088	9
GGGTA	12660485	1.042658	14.321302	2
TGGAG	12603150	1.0379361	9.363395	1
CGTGT	3115175	1.0318936	6.7087255	41
GTTGA	15939365	1.0285242	12.91638	43
TGTAG	15918850	1.0272002	8.345548	21
CGTGC	240575	1.0130385	5.1746025	13
ATTTC	1989810	1.0020329	5.5632415	22
AGTTG	15486625	0.99930996	9.037087	38
TTATC	1956635	0.98532647	12.10973	37
GGTTG	29439685	0.97906315	6.8635526	42
TAAGT	7741665	0.9692675	6.323494	7
CGATC	117295	0.95834273	5.9973984	44
TGGGG	22159305	0.94054973	8.177875	1
AAGGC	461540	0.9375397	14.255609	46
GTTTG	35771400	0.9321064	6.904386	18
TGCGG	2196235	0.92849445	5.8109517	5
TTGGG	27740550	0.92255574	6.1650114	36
GGGGG	17002155	0.92103827	5.5717797	2
AAGAC	230965	0.91031647	8.017179	32
ATTAC	729260	0.90942454	5.224738	29
GTGGT	27225415	0.9054242	7.5640335	9
GGATA	5625685	0.8989443	7.1254387	2
TTTGG	34437045	0.8973366	5.1475463	35
TGGTT	34424100	0.8969993	7.189197	1
TAGAC	555870	0.88471943	9.994894	25
GGAGC	843425	0.88300055	8.562123	27
AGTGA	5524865	0.88283396	5.4143333	18
GGGTG	20616150	0.87505066	8.101247	2

GAAGC	404295	0.8212563	10.065132	4
GGTAT	12633765	0.8152226	5.8615146	2
GGTAC	975525	0.80021155	12.339794	3
GGAAC	391830	0.79593575	7.5860195	27
GTGCG	1824525	0.77134806	5.375276	4
TGGGT	22956360	0.76345	8.562336	1
TGGTG	22857355	0.7601574	5.9220796	7
GGTAA	4752995	0.7594946	6.0206957	2
TCACG	84365	0.68929267	27.345104	30
GAGTC	837075	0.68664265	12.634509	21
CACAT	42900	0.68008715	6.42927	47
TGGTA	10331275	0.66664916	5.146064	1
TCCAG	79615	0.6504834	27.365555	25
CAGTC	78290	0.63965774	27.544388	27
GTCAC	77710	0.6349189	27.792301	29
CCAGT	77120	0.6300984	27.325365	26
GAATC	357570	0.5691062	5.78567	40
ATCTC	84310	0.5397249	22.400341	42
TGGGC	1231080	0.5204594	5.47105	13
GATTC	796115	0.51167494	5.4195113	29
TGAAC	265465	0.42251253	5.8334126	20
GGTGC	950930	0.40202135	5.4285617	3
CTCGT	94820	0.31284365	11.688341	44
TCTCG	94760	0.3126457	11.613883	43
CTGAA	189520	0.30163887	5.4972067	19
GAACT	119920	0.19086394	5.6064067	21
AGTCA	87525	0.13930425	5.566136	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
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	228072	0.28695226279508756	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	184006	0.2315099532948932	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	141944	0.17858900693722118	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	92656	0.11657655854967568	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	88531	0.11138662693145977	No Hit