FASTQ QC Report

Report Date	`10-02-16`
Run ID	`160930_D00796_0121_AC9MR4ANXX`
Project ID	`EC-EL-3883`
Sample	`Sample_OD5_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.

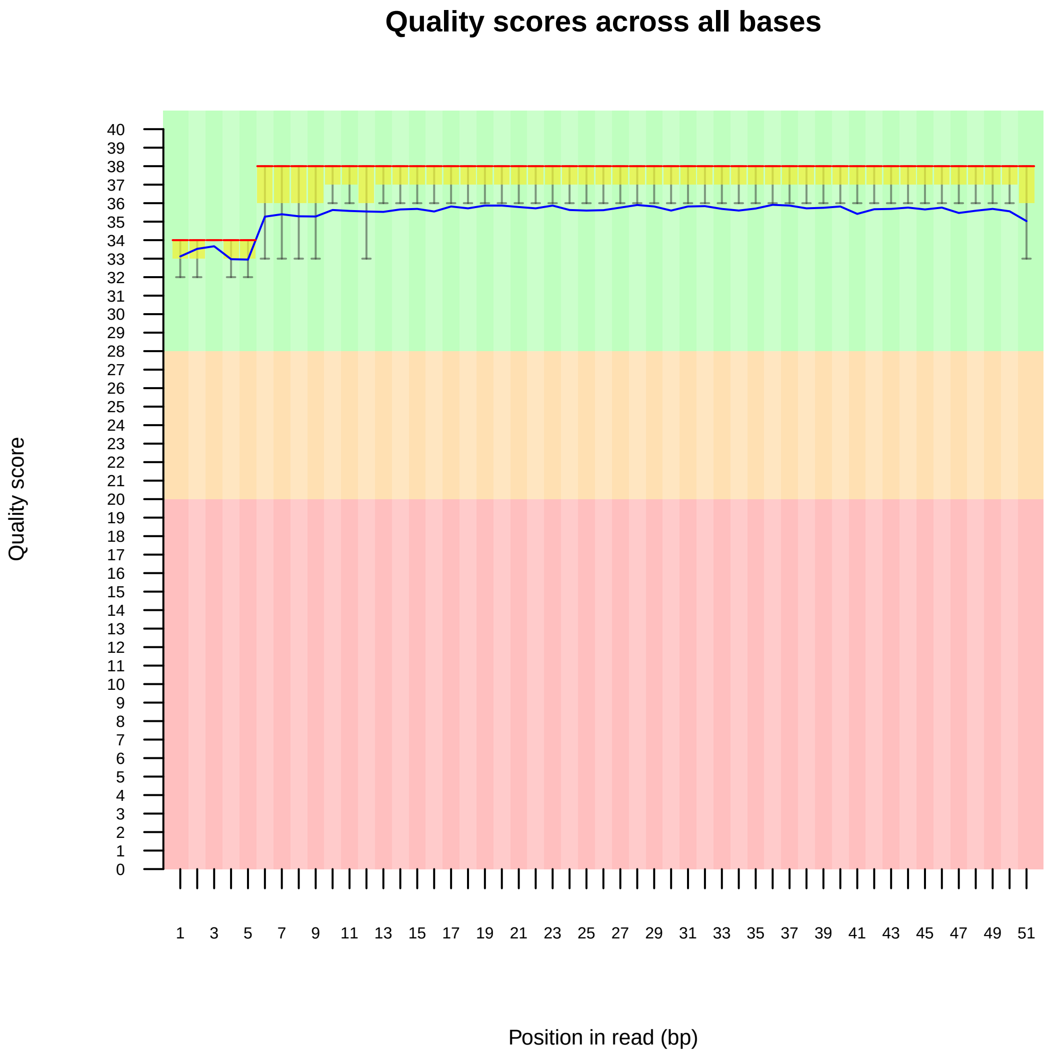
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

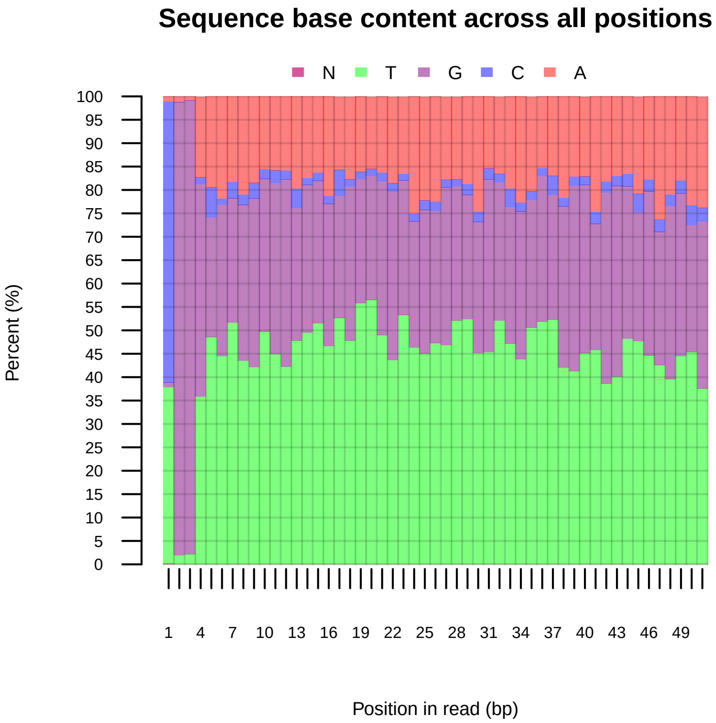
- Estimated Duplication rate 75.9013%

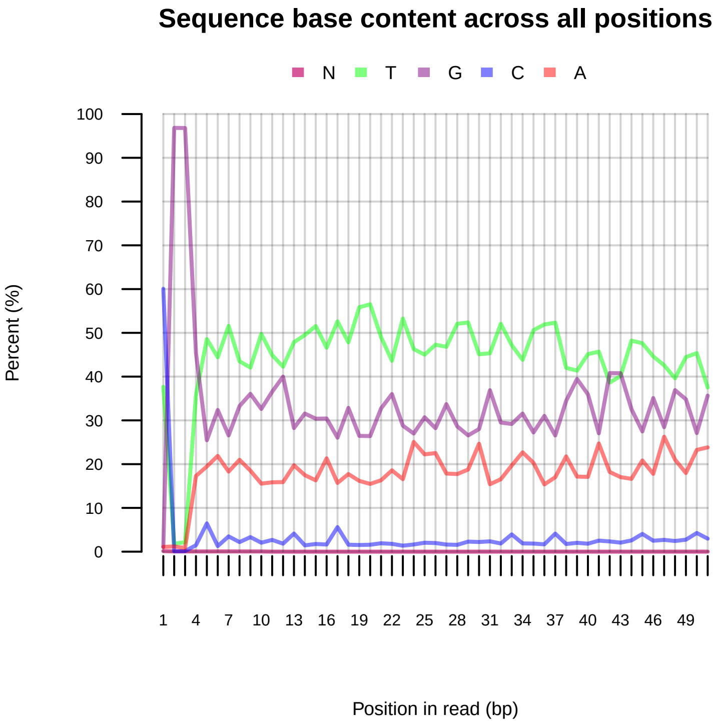
2 Per base sequence quality



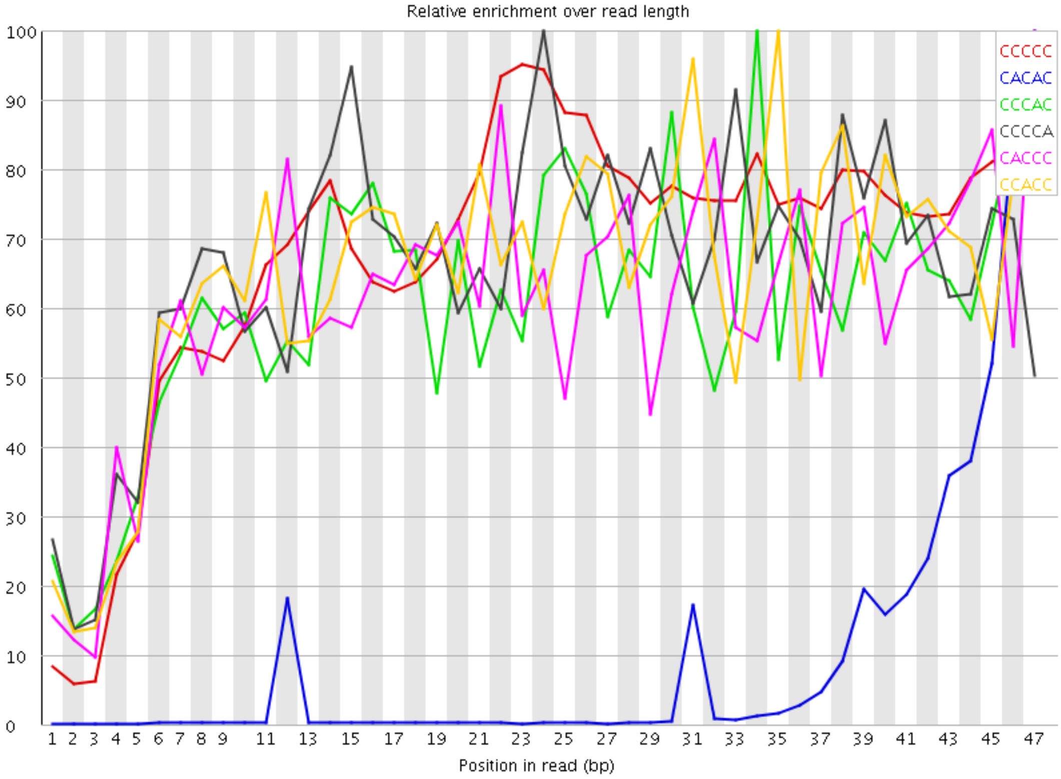
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	142945	764.7488	1117.56	47
CACAC	857085	171.05833	1738.5438	47
CCCAC	56215	58.088116	94.20701	34
CCCCA	55900	57.762623	86.91967	24
CACCC	55675	57.53012	93.96483	47
CCACC	55460	57.307957	87.65123	35
ACCCC	54080	55.881977	82.791695	23
CGGGC	4838715	28.135195	988.82904	1
AGCAC	1195885	24.539543	185.24257	45
GCACA	1013355	20.79403	184.92513	46
GCCCC	35200	19.361952	34.897118	47
CCCCG	29855	16.421906	32.05227	25
CCCGC	29330	16.133127	30.11384	26
CCGCC	28800	15.841598	27.400475	35
CGCCC	28490	15.67108	27.270071	23
ACACG	621675	12.756763	154.59682	47
CCTCC	29230	12.160207	18.670078	25
CGGAA	5683395	11.99062	187.86787	1
CTCCC	27815	11.571541	16.61778	30
CGCGG	1933410	11.242007	256.86362	5
TCCCC	26230	10.912152	20.931843	3
GCGCG	1812305	10.537829	255.22849	4
CCCTC	25125	10.452453	15.053339	24
CCCCT	23750	9.880428	15.35434	8
CGGCG	1555985	9.047431	246.5313	1
CGCGC	158610	8.970027	62.335358	13
AGATC	5076005	8.099561	39.609123	43
GGCGC	1361185	7.914747	255.75505	3
CGGGA	6799870	7.636701	220.45856	1
CCACA	32685	6.523322	9.425788	15
CGGGT	13437695	6.07585	237.08278	1
TCGCG	1365860	6.0066414	21.912718	30
CACCA	30095	6.0064063	9.660402	39
AACCC	29700	5.927572	8.206453	21
ACCCA	28770	5.7419605	9.566571	13
ACACC	28665	5.721005	8.488195	44
ACCAC	28135	5.6152263	8.816531	46
CCCAA	27605	5.509448	8.15968	14
CGGAG	4694775	5.2725415	155.69339	1
AGACG	2456795	5.1832557	54.829094	27
CAACC	25580	5.1052957	7.643739	20
CCAAC	24720	4.9336553	7.081319	12
ACGCG	441795	4.8257995	15.5559435	14

CGCGT	1088695	4.7877536	19.889486	31
CTCCA	58550	4.7046175	121.2272	24
CGGAC	427895	4.6739674	158.43474	1
ACTCC	57830	4.6467643	121.339745	23
CGGTT	13415940	4.587848	156.01811	1
GATCG	5356210	4.5495496	22.320753	44
ACGTC	547435	4.522585	28.215546	47
CGTCG	1023850	4.5025845	22.492357	41
GGGCG	7433975	4.444242	105.881805	2
CGCGA	405820	4.4328384	17.858002	24
TCGAG	5106150	4.337149	49.04235	44
CGGGG	7128785	4.2617908	105.859375	1
CGGTC	959600	4.220032	157.27539	1
AGAGC	1997350	4.2139354	26.712893	47
AAAAA	2929630	4.212976	12.932798	31
AGGCG	3687640	4.141462	56.55974	47
CACGT	469520	3.8788965	37.871143	47
CGACG	341780	3.733319	30.036243	24
ATCGG	4352330	3.6968565	20.66865	45
GAGAC	1724935	3.6392045	52.693176	26
TCGGA	4249975	3.6099167	20.935986	46
TTACG	5489370	3.5264604	40.968716	14
CGGTA	4045125	3.4359176	120.14069	1
CGTTT	12856160	3.3251011	38.03682	17
GGCGT	7224510	3.26656	51.71678	3
TACGT	5060920	3.251217	42.959152	15
ACGTT	5019895	3.224862	44.707943	16
CCCAG	30320	3.221224	11.2084255	27
GGCGG	5374090	3.212784	41.239697	11
CGAGG	2841785	3.1915116	61.380806	45
CGGAT	3620270	3.0750475	105.207115	1
GAGCA	1452935	3.0653489	21.22655	47
AAGCG	1408435	2.9714649	60.08793	8
GACGG	2628285	2.951737	29.280853	28
ACGGG	2539580	2.8521154	29.353844	29
AGCGA	1334045	2.8145196	60.838135	9
CCAGC	26145	2.7776682	7.189396	28
TTTCG	10656375	2.7561517	16.934805	30
AGCCC	25645	2.7245479	5.3672214	47
CGAGA	1283625	2.7081454	34.84251	25
AGAGA	6592095	2.6862257	19.112677	25
ATCGC	325135	2.686073	31.540207	29
TTCGA	4087875	2.6261175	37.905144	31
GTCGA	3042290	2.5841124	48.184986	43
TTTTT	168949965	2.5699208	5.6537366	16
GGAGG	21648435	2.4997027	30.562197	39
ACACA	64795	2.4977396	58.703037	32
GCGGG	4159860	2.4868827	42.0949	12
CGTTA	3856235	2.477308	34.250423	9
GAAGA	5940445	2.4206834	9.886757	46
GGAAG	11140770	2.416605	11.706585	2
GCGGC	408510	2.3753226	9.169188	9
GAGGC	2113400	2.373487	46.63483	46
TCGTT	9072785	2.3465734	5.9178276	4
TTCGC	695200	2.3122828	8.963367	33
AGAAA	2997505	2.2946029	5.5860515	22
ATTCG	3568845	2.2926838	43.18723	34
GGGAG	19794825	2.28567	26.629866	38
TTTTA	60337785	2.2796803	12.685612	26
CACGC	21415	2.2751489	8.03835	47
CGAGT	2661130	2.2603567	47.4269	33
TTTAG	44990835	2.2475166	16.004148	27
CGTTC	667895	2.2214646	25.292656	33
AGTAG	13482505	2.2119064	24.081299	35
CGGTG	4870575	2.202229	42.7346	1
CGTAG	2587165	2.197531	24.699068	5
GAGAT	13311025	2.1837735	8.256635	26
AAGAG	5324145	2.1695464	9.941522	47
TTCGT	8381450	2.1677678	5.3563333	35
GAGGT	24822365	2.1677575	23.350859	40
GCGGA	1926190	2.163238	24.069864	7
GCACC	19850	2.108882	7.2395077	47
GCGTT	5996675	2.0506825	26.540327	16
AGGAG	9357620	2.0298119	9.997157	38
CGCAC	18785	1.9957355	8.063319	46
TTTAC	4099565	1.9918654	30.073685	13
ATTTT	52490255	1.9831848	7.9947877	25
GGTCG	4328985	1.9573492	27.131157	42
TAGTT	38717920	1.9341533	9.879614	29
TAAAA	3304025	1.9129193	5.3222766	30
TCCCA	23780	1.9107739	8.288305	26
AAACG	480290	1.9035571	13.101791	7
AAAAT	3256140	1.8851955	5.159584	32
GCGGT	4147505	1.8752931	25.849894	6
TACGC	225370	1.8618736	10.765637	13
AATTT	19581335	1.8375949	17.774324	24
ACGGA	867040	1.8292494	6.8143134	30
GACGC	166190	1.8153206	11.472181	3
ACGGC	165945	1.8126445	8.890849	12
TAGAG	11035460	1.8104502	8.56745	24
AGGTA	11004135	1.805311	29.153334	47
TTAGT	35959965	1.7963797	15.40226	28
GCGTA	2111715	1.7936847	24.379393	4
TCGTC	538660	1.79162	8.959325	40
GAAAA	2318380	1.7747298	5.4228587	3
CGAGC	162200	1.7717372	5.9698577	13
GTCGC	401370	1.7651044	8.247963	3
AGCGC	161340	1.7623433	5.9699945	35
ACGCC	16545	1.7577558	5.766376	16
GGAAA	4297375	1.7511458	12.240479	2
TACGG	2018880	1.7148309	13.21916	5
TGAGA	10446345	1.7138013	5.7531643	41
AGGTC	2013465	1.7102314	46.00229	41
AGCGG	1521225	1.7084358	7.01532	6
ATCGT	2657830	1.707433	12.652822	39
AGTTT	33997395	1.6983396	9.3049	26
GGAGA	7808450	1.6937733	10.470919	2

TAGTA	13607055	1.6883634	14.793747	29
GAGCG	1496540	1.6807126	9.479511	28
AGTTA	13488560	1.6736606	21.463102	30
TAGCG	1961260	1.6658887	5.489118	10
TGGCG	3626675	1.6398	35.5248	10
TGGGA	18682370	1.6315465	14.295568	37
GGACG	1452000	1.6306915	16.919014	2
AGTCG	1889000	1.6045113	14.1725	22
TATTT	42452150	1.6039255	5.6781454	32
GCGTG	3532050	1.5970154	34.936317	4
TTGAG	24140725	1.5944963	14.270042	44
CGTGG	3517235	1.5903169	34.703022	5
TATCG	2472465	1.5883516	12.885228	38
GTAGA	9676505	1.5875032	8.247224	23
CGATT	2459735	1.5801737	20.7037	11
AACGC	75820	1.5558255	5.299008	11
GCGAC	140675	1.5366161	20.97109	23
TAGGA	9347320	1.5334982	7.79092	37
GGGAA	6992445	1.5167692	14.093499	2
AGCGT	1776620	1.5090562	7.6481814	29
TTCGG	4403570	1.5058887	23.654163	35
GCGTC	339760	1.4941623	10.135804	40
GTAGT	22565275	1.4904376	10.344092	36
GTCGT	4358330	1.4904176	9.894571	3
GGTTT	55707875	1.4813814	9.489997	2
AGGTT	22310770	1.4736276	14.743089	41
GTACG	1727260	1.4671297	12.943738	4
AAGTA	4750450	1.4640621	12.309388	34
TAATT	15469180	1.4516929	17.441706	23
TATAG	11699070	1.4516206	18.43047	47
TTATT	38318250	1.4477388	7.4183316	32
TTTAA	15341340	1.4396958	9.010271	5
TTATA	15142855	1.4210693	14.230681	46
CGTAC	171590	1.4175752	9.061306	13
GCGAT	1667365	1.4162554	25.32143	10
TTAAG	11412005	1.4160017	10.974827	6
ACGGT	1666120	1.4151978	12.593728	6
TAAGC	884485	1.4113343	43.32691	7
AAAAC	187950	1.3993723	22.096441	6
AACGG	653105	1.3778971	7.3651147	8
GTTTA	27377085	1.367622	8.879158	4
ATCCC	16775	1.3479071	5.51291	25
TCGAC	161185	1.3316151	5.916521	23
GAGTA	8046905	1.3201551	17.373318	34
GGCGA	1174580	1.3191305	8.903423	2
TCGGG	2903400	1.3127713	29.943579	36
GGTAG	14915040	1.3025426	7.434352	2
GACGT	1527460	1.2974204	6.2292485	3
GGAGT	14820510	1.2942873	10.378035	2
GGGTT	36448365	1.281512	14.859394	2
ATTAT	13644970	1.2805012	14.102635	45
TTGTA	25232150	1.2604718	14.727126	20
GGAAT	7660560	1.2567725	9.345855	2
GGTTA	18899310	1.2483004	15.216412	2
ACGAC	60710	1.2457683	5.968863	18
TATTC	2562475	1.2450355	31.384104	33
GAACG	586965	1.2383572	5.7629743	3
GGGAT	14015920	1.2240218	12.577213	42
GATTA	9817670	1.2181765	17.983046	44
CACAG	58520	1.2008295	31.3356	33
TTTGT	59538600	1.1974422	7.011499	19
GTTAA	9623910	1.1941347	17.827274	3
GTAAT	9594640	1.1905029	21.712511	22
GGGGA	10222065	1.1803219	9.752337	2
TGGAA	7170785	1.176421	8.920395	1
TGAGG	13390705	1.1694211	16.750408	45
GGTGG	25095240	1.1666224	11.423139	8
CGTAA	726525	1.1592845	10.313624	21
GGATT	17165455	1.1337793	9.819023	43
CGTAT	1762065	1.1319792	5.280854	13
TTTTC	5669300	1.1089928	11.976986	29
TCGTG	3240610	1.108191	5.628566	40
GTGGC	2406195	1.0879602	33.54306	9
GGGGT	23340130	1.0850312	8.128764	2
AGTAT	8724495	1.0825354	13.9662895	30
TAGGC	1270195	1.0789002	9.265095	13
AGTAA	3484525	1.073911	7.7317944	9
CGTGA	1261505	1.0715189	7.738472	26
GGGTA	12177365	1.0634594	14.891253	2
GTATT	21049685	1.0515368	6.11034	31
TGTAA	8472635	1.0512846	20.936747	21
GTTAT	20970495	1.0475808	8.941903	31
TGGAG	11939640	1.0426984	9.70644	1
AACTC	67170	1.042457	24.06714	22
TCGAT	1611325	1.0351412	6.1908064	11
TGTAG	15663980	1.0346068	8.433571	21
GTTGA	15644590	1.033326	13.302336	43
AGTTG	15583165	1.029269	10.334078	38
CGTCT	303425	1.0092124	10.365548	47
TAAGT	8088895	1.0036701	6.969957	7
CGAAC	48280	0.9907049	6.7837377	20
TTAAT	10540840	0.9891968	11.901958	4
AAGGC	468120	0.9876225	15.653959	46
GGTTG	28060320	0.98659116	7.20632	42
ATTTC	2004875	0.97411317	5.435917	22
TTGGG	27067305	0.9516771	6.393712	36
CGTGT	2765620	0.945759	5.381762	41
AAGAC	237625	0.94179094	9.267992	32
TGGGG	20094635	0.9341553	8.332005	1
GTTTG	34822575	0.9260003	6.916874	18
GTGGT	25913135	0.9110968	7.926997	9
TAGAC	570175	0.90980345	10.745243	25
TGCGG	2000770	0.90464747	5.858828	5
CGGCC	15975	0.9034499	7.998567	1
GGATA	5506730	0.90342045	7.402401	2
TAAGG	5494430	0.90140253	5.0132284	45
GGGGG	14585405	0.8965031	5.6014414	2

GGAGC	797545	0.89569545	8.597841	27
TTTGG	33650150	0.8948234	5.1765475	35
AGTGA	5448425	0.8938551	5.3451514	18
TGGTT	33457430	0.88969857	7.2748384	1
GGGTG	18821380	0.8749645	8.366845	2
TTATC	1798225	0.8737079	9.419401	37
GAAGC	408430	0.8616906	11.492462	4
GGTAT	12444335	0.821949	6.1053014	2
GGTAC	943120	0.80108345	12.977533	3
TCACA	51045	0.79220206	23.520927	30
TGGGT	21876150	0.7691578	8.926768	1
GGTAA	4656845	0.76399034	6.133845	2
AGTGG	8682475	0.75824773	5.0717273	8
TGGTG	21419330	0.7530962	6.1847873	7
CACAT	48325	0.74998844	5.513835	47
GTTGG	21262340	0.74757653	5.6434603	39
GTGCG	1603330	0.7249451	5.471183	4
GGAAC	338280	0.7136908	6.090029	2
GAGTC	810000	0.6880117	13.100755	21
TGGTA	9952790	0.65738237	5.076883	1
TGGGC	1129315	0.5106194	5.691807	13
TCCAG	61245	0.50596994	12.832814	25
GATTC	786065	0.5049809	5.1433983	29
TGGAT	7645090	0.5049586	5.0232735	1
CCAGT	53695	0.44359633	12.735848	26
CAGTC	51470	0.42521477	12.869871	27
ATGCC	48645	0.4018762	13.736031	47
GTCAC	47480	0.39225167	12.753518	29
GGTGC	847605	0.38324434	5.5193777	3
ATCTC	56710	0.3543394	10.347456	40
CGGCT	48755	0.21440983	6.875902	1
TCTCG	63880	0.21246926	5.5847673	41
CTCGT	59190	0.19686998	5.5574603	42

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	267159	0.34113142567676585	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	177355	0.22646200951831236	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	158202	0.20200582351676608	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	94364	0.12049201356706056	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	89867	0.11474985993844085	No Hit
CGGGATGGTTTCGATTTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	78593	0.10035425397689787	No Hit