FASTQ QC Report

Report Date	`10-02-16`
Run ID	`160930_D00796_0121_AC9MR4ANXX`
Project ID	`EC-EL-3883`
Sample	`Sample_YD3_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.

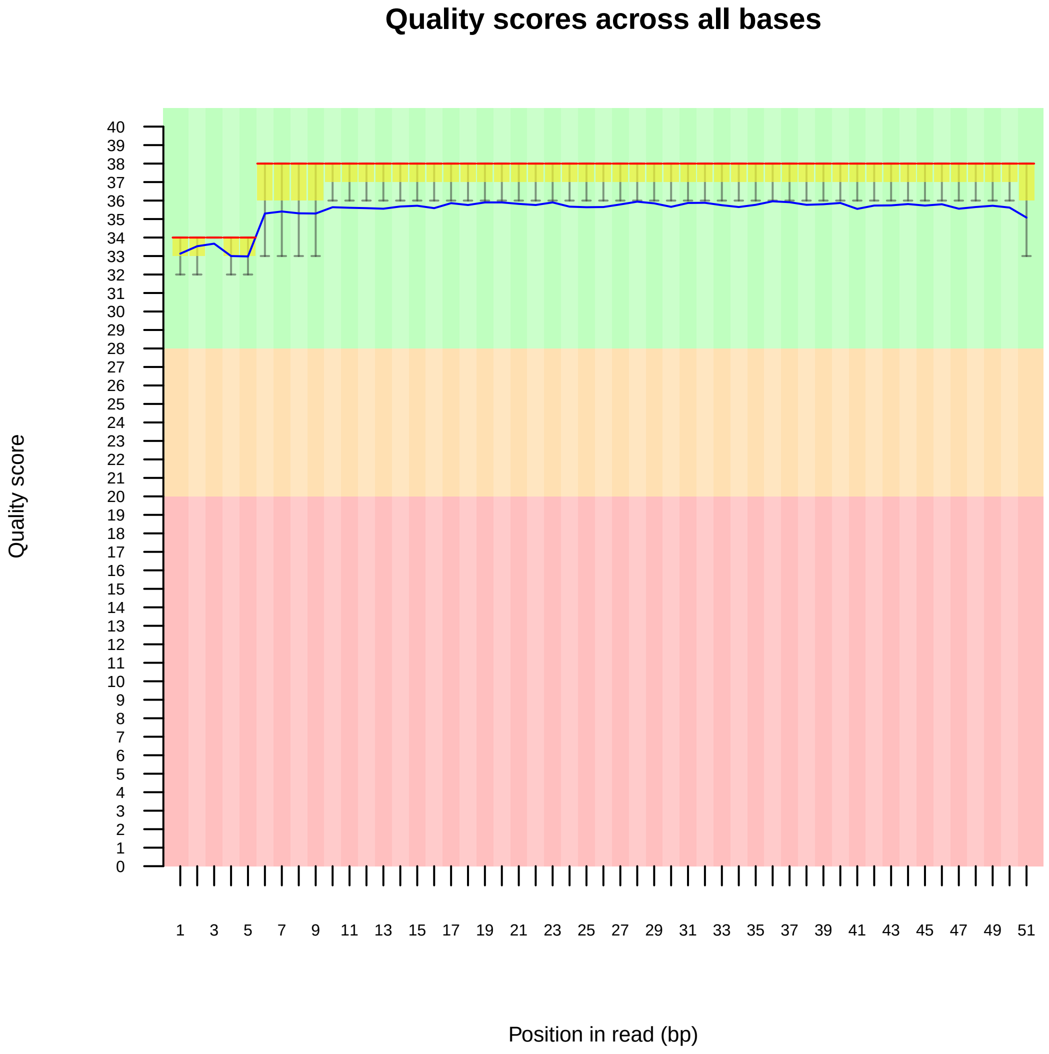
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

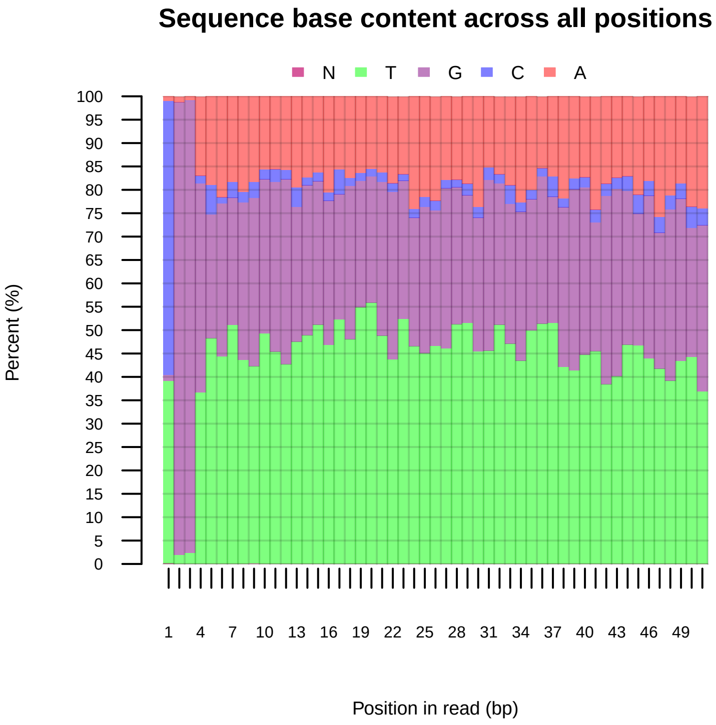
- Estimated Duplication rate 76.4465%

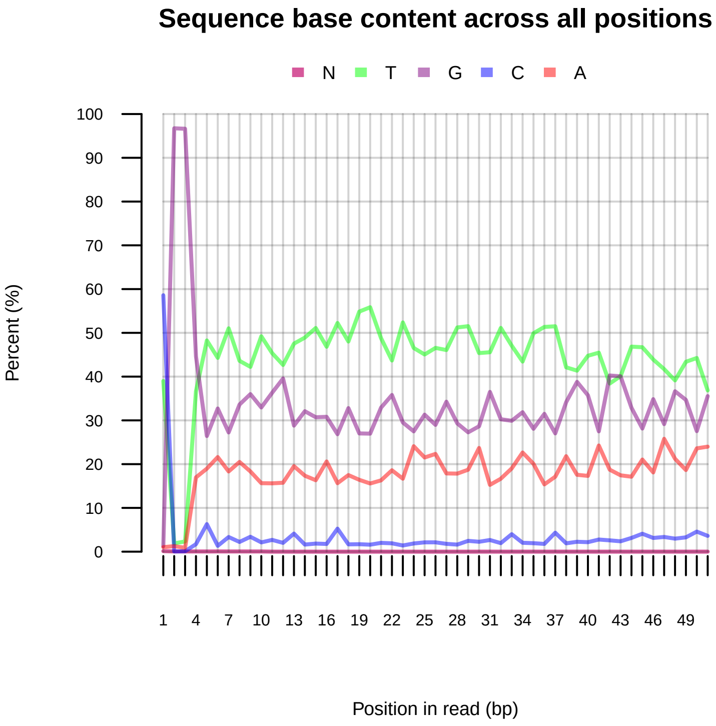
2 Per base sequence quality



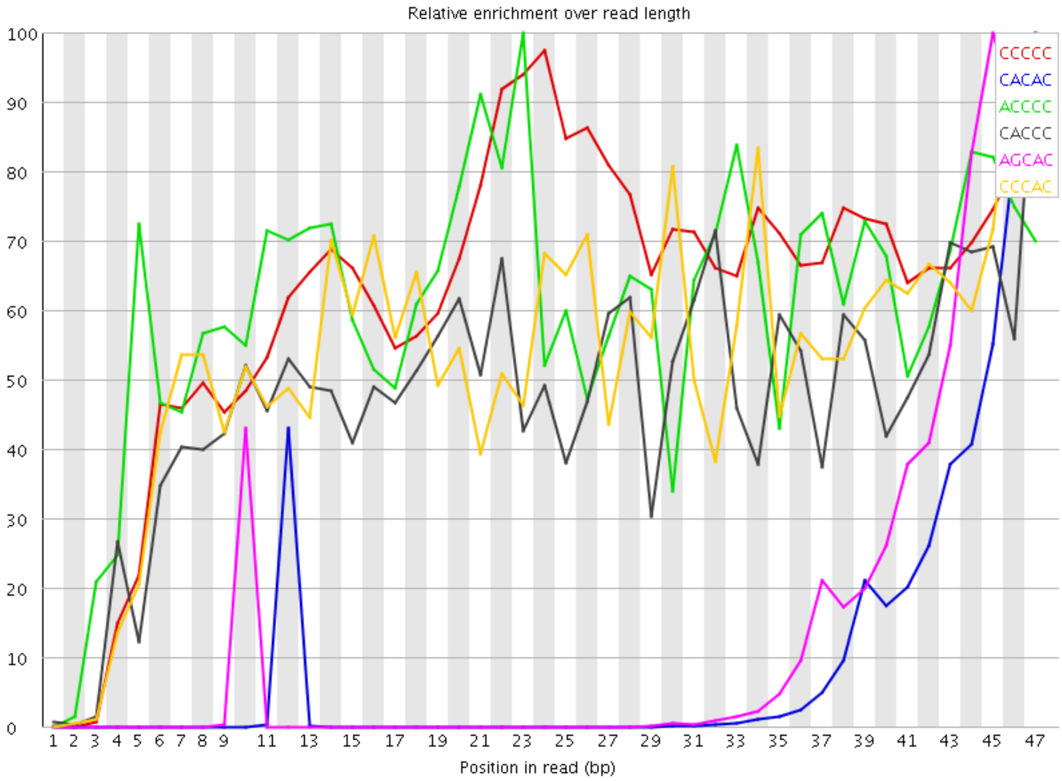
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	124520	572.9408	914.66345	47
CACAC	1378750	256.66354	2561.0964	47
ACCCC	44140	40.851406	67.4115	23
CACCC	43255	40.03234	83.72511	47
AGCAC	1945545	38.354034	279.4055	45
CCCAC	41380	38.297035	71.54691	47
CCCCA	40130	37.140163	58.061245	24
CCACC	39925	36.95044	64.370476	47
GCACA	1679355	33.106426	279.68973	46
CGGGC	4450450	24.319126	839.8084	1
ACACG	1088160	21.451744	226.67357	47
CGGAA	7249875	15.135315	173.52473	1
GCCCC	27780	13.536114	35.378597	47
CCCGC	26645	12.983072	24.615227	26
CCCCG	26420	12.873439	25.990126	46
CCGCC	26220	12.775987	23.127422	31
CGCCC	25655	12.500684	21.867182	23
AGATC	7011555	11.227334	55.919186	43
ACTCC	147735	11.105818	430.34164	23
CGCGC	208090	10.737522	56.11178	13
CTCCA	140895	10.591628	429.3906	24
CGCGG	1900295	10.384008	211.38045	5
GCGCG	1792900	9.797158	209.07066	4
CGGCG	1704925	9.316425	246.68387	1
CGGGA	6410510	7.045967	199.27115	1
GGCGC	1269690	6.9381175	209.53085	3
ACGTC	866235	6.8959475	47.728065	15
CACGT	848270	6.7529316	61.21318	47
TCGCG	1514515	6.3477383	19.384247	30
AGAGC	2981495	6.2243643	37.913452	47
GATCG	7367940	6.2114854	30.600592	44
TCCCC	14990	5.602279	11.949038	5
CGGGT	12572580	5.5803432	215.29025	1
CCTCC	14525	5.4284925	12.733237	28
CTCCC	14265	5.3313217	12.382066	29
ATCGG	6238565	5.2593737	29.219208	45
AGACG	2472295	5.161325	55.51742	27
TCGGA	6018155	5.0735593	29.415096	46
CAACA	134825	5.048393	216.66731	37
CCCTC	13460	5.0304656	9.044837	24
CGGAG	4525705	4.974326	143.97533	1
CGCGT	1178250	4.93836	17.496414	31
ACGCG	473855	4.9181566	14.578929	6

CGTCG	1172460	4.9140925	22.453915	41
CCCCT	13145	4.9127393	9.308547	38
GAGCA	2316640	4.8363695	31.451956	47
CGGTC	1122335	4.7040057	168.27455	1
CGCGA	451475	4.6858735	17.312288	24
CGGAC	442905	4.5969257	146.97035	1
CGGTT	13144900	4.4750214	149.25731	1
AAAAA	2851625	4.319994	14.675997	31
CGGGG	7176915	4.1531053	99.76766	1
GGGCG	7101660	4.1095567	92.82356	2
CGACG	385925	4.0055285	33.373528	24
TCGAG	4750290	4.0046954	43.090435	32
AGGCG	3466100	3.8096855	44.432716	47
AACCC	20325	3.7836351	6.4734926	22
GAGAC	1733920	3.619845	53.42889	26
ACACC	18895	3.517431	6.954984	47
ACCAC	18100	3.369436	6.7362413	45
CCACA	18040	3.3582666	6.7362804	46
CGTTT	12749625	3.3291745	33.68265	17
CAACC	17610	3.2782197	5.9050903	31
CCCAA	17560	3.2689114	5.6424785	16
TTACG	5052420	3.2670078	33.54188	6
CGGTA	3832530	3.230985	109.98184	1
ACCCA	17155	3.193518	5.8176284	33
GGCGG	5483970	3.1734388	37.308456	11
CACCA	16985	3.1618714	4.8553324	35
CCAAC	16730	3.1144013	5.9050646	30
GGCGT	6980020	3.0980833	45.48144	3
GACGG	2761450	3.0351853	29.365875	28
ACGTT	4632960	2.9957752	36.434	16
TACGT	4625175	2.9907415	34.837757	15
GAAGA	7114605	2.9875536	14.027735	46
ACGGG	2608460	2.86703	29.323732	29
CGGAT	3375785	2.8459296	94.73714	1
CGAGG	2583990	2.8401341	47.83773	45
TTTCG	10794300	2.8186014	15.406169	30
AGAGA	6501860	2.7302506	19.885763	25
CGAGA	1305495	2.725437	32.738182	25
GGAAG	12323945	2.724594	10.957035	2
AAGCG	1287380	2.6876187	51.380417	8
CACAG	135600	2.673188	115.465935	31
AAGAG	6349795	2.6663961	14.00335	47
TTTTT	162441085	2.6425843	5.6924653	16
GCGGC	483035	2.6395059	9.752454	9
AGCGA	1238705	2.5860016	51.912262	9
TTCGA	3996615	2.5843005	34.45926	31
ATCGC	321660	2.56068	28.084423	29
GAGAT	15040795	2.550493	8.405216	26
CGTTC	787040	2.5301337	23.877392	33
AACTC	164095	2.481227	88.6563	22
TCGTT	9434395	2.4635038	6.695668	4
TTCGC	761840	2.4491222	8.693606	33
CGTTA	3771525	2.4387522	32.416615	9
GTCGA	2863635	2.4141655	38.250587	43
GCGGG	4154635	2.4041858	38.146095	12
GGAGG	20462675	2.3817794	26.635666	39
TTTTA	56538865	2.2776742	11.857197	26
TTCGT	8627810	2.2528887	5.375331	35
ATTCG	3442295	2.225865	39.618774	34
AGAAA	2767505	2.207325	5.543563	22
TTTAG	41933020	2.202416	14.726046	27
GGGAG	18665105	2.172549	23.026806	38
GCGGA	1967530	2.1625662	22.86177	7
AGTAG	12722995	2.15746	22.43401	35
CGAGT	2536295	2.138204	42.050392	33
GAGGC	1921185	2.111627	37.119156	46
CGGTG	4728825	2.0988903	42.18905	1
TGAGA	12366865	2.0970702	7.5286045	41
CAGTC	262055	2.0861747	46.656624	27
CGTAG	2448520	2.064206	20.806707	5
GAGGT	23095695	2.0619187	20.633558	40
GCGTT	5967535	2.0315747	24.872862	16
GTCGC	483600	2.0268967	10.174478	3
TCACA	133820	2.0234485	87.980545	30
TCGTC	629320	2.0231042	8.688201	40
AGGAG	9109360	2.0139093	9.462192	38
CAATC	133050	2.0118055	89.29817	40
CACGC	20275	1.9871317	8.981539	47
ATTTT	48820855	1.9667531	7.501939	13
TCAAC	129245	1.9542714	86.76642	36
GGTCG	4368635	1.9390198	21.718819	42
TAAAA	3141985	1.9221323	5.9975977	30
TAGTT	36535040	1.9189019	9.695135	29
AAACG	479185	1.9001025	15.860704	7
CGCAC	19275	1.8891227	9.257894	47
AAAAT	3074085	1.880594	5.8331823	32
CGAGC	180435	1.8727409	8.835671	32
TACGC	234670	1.8681674	9.760456	13
ACGGA	890520	1.8591079	9.142167	30
TTTAC	3689260	1.8297482	24.90673	5
GCGGT	4120805	1.8290204	21.95501	6
CGTCT	565085	1.8166047	18.605024	16
GACGC	173725	1.8030976	12.7709	5
TAGAG	10557290	1.7902178	8.903951	24
AATTT	17923210	1.788018	15.776065	24
TTAGT	33678050	1.7688464	14.15749	28
AGCGC	168525	1.7491266	9.177163	35
ACGGC	168365	1.747466	7.7225623	6
AGGTA	10210465	1.7314059	26.962385	47
ATCGT	2677435	1.7312893	12.961649	39
TTGAG	25249605	1.7290056	13.777719	44
GAAAA	2126815	1.6963191	5.1396303	3
GCGTA	2010480	1.6949197	20.426462	4
GGAGA	7642310	1.689572	10.171126	2
AGTTT	32078980	1.6848596	8.977004	26
GGAAA	4010295	1.6839967	11.496439	2
AAATA	2749055	1.6817545	5.499998	33

GAGCG	1527040	1.6784115	10.347154	28
AACGC	85115	1.677938	9.060169	23
GCGAC	160490	1.6657311	20.438566	23
AGTTA	12753840	1.6588064	21.074898	30
AGCCC	16890	1.6553713	5.273781	46
GGACG	1499695	1.6483558	16.257189	2
TAGTA	12463185	1.6210029	12.146444	29
TATCG	2496850	1.6145191	13.26076	38
AGCGG	1466810	1.612211	6.7990646	6
GCGTC	384645	1.61215	10.708891	40
TACGG	1894640	1.5972614	10.55358	5
TCGTA	2453170	1.5862746	8.24542	45
TGGCG	3571265	1.585107	32.64389	10
GTCGT	4641035	1.579984	11.077246	3
GTAGA	9283770	1.5742646	8.566146	23
AGTCG	1865175	1.5724214	13.299504	22
CGATT	2426315	1.5689094	18.31621	11
TGGGA	17441095	1.5570918	12.14969	41
GCGTG	3469245	1.5398253	31.128492	4
ATAAA	2501800	1.5304945	5.170115	37
TAGGA	8841215	1.4992198	7.4513507	37
GGGAA	6778840	1.4986749	13.319182	2
CGTGG	3362280	1.4923489	30.888006	5
TTCGG	4359270	1.4840603	21.638504	35
CGTAC	185620	1.4776889	7.8693457	13
TTATT	36603875	1.4745908	7.400331	32
GGTTT	53321650	1.474463	9.583459	2
CGAAA	370310	1.4683827	6.7205386	32
AGCGT	1734800	1.4625098	8.351232	29
TCGAA	913230	1.4623201	5.056425	32
AGGTC	1733615	1.4615107	36.226986	41
GTAGT	21324085	1.4601996	9.698842	36
GCACC	14855	1.4559231	6.954935	47
TCGAC	182260	1.4509406	8.413475	23
AACGG	692560	1.4458336	8.722411	8
AGGTT	20996660	1.4377787	13.51186	41
CCCAG	14485	1.4196597	9.073365	27
ACGCC	14425	1.4137793	5.158404	16
AAGTA	4362520	1.4050887	11.819931	34
TTTAA	14063735	1.4029971	8.453857	5
TATAG	10784655	1.4026875	17.264725	47
CGAAC	70905	1.3978052	7.1229076	9
TAATT	13991320	1.3957729	15.452753	23
TTATA	13919970	1.3886548	13.510735	46
GTACG	1638490	1.3813164	10.351475	4
AAAAC	181500	1.3669838	27.707932	6
ACGGT	1620855	1.3664494	10.023819	6
TTAAG	10343630	1.3453263	10.103767	6
GCGAT	1586360	1.3373686	21.884333	10
GTTTA	25383125	1.3331784	8.511225	4
GGAAT	7857510	1.3324113	9.446045	2
CCAGC	13335	1.3069495	6.4941545	28
GAACG	624320	1.3033713	7.8588367	28
GGAGT	14402065	1.2857758	10.269482	2
GGTAG	14324950	1.2788912	7.3292356	2
GGCGA	1160230	1.2752404	8.418961	2
GGGTT	35350030	1.2744381	14.484577	2
GAGTA	7498260	1.2714926	16.014957	34
TCGGG	2853400	1.2664824	26.91177	36
ATTAT	12678710	1.264827	13.5239935	45
GACGT	1499670	1.2642853	5.67575	3
TGGAA	7366250	1.2491076	8.789659	1
TAAGC	776630	1.2435876	37.526604	7
TTGTA	23543970	1.2365819	13.483855	20
GGTTA	17897945	1.2255894	15.509261	2
TTTGT	57354965	1.2164735	6.5611987	19
TATTC	2406655	1.1936194	29.208862	33
GGGAT	13352345	1.1920598	11.560026	42
CGTAT	1839285	1.1893228	5.173092	13
GATTA	9024320	1.1737328	16.903162	44
GGGGA	10083290	1.1736575	9.62962	2
GTTAA	8941165	1.1629174	18.27151	3
CGTAA	721975	1.1560708	10.169655	21
GGTGG	24455200	1.1494731	10.242166	8
TGAGG	12838985	1.1462283	15.388913	45
TCCAG	142965	1.1381198	46.6841	25
TCGTG	3320835	1.1305379	5.64385	40
TTTTC	5627225	1.1270282	10.988873	29
GTAAT	8552600	1.1123793	18.959225	22
GGATT	16195560	1.109016	9.156233	43
CCAGT	138270	1.1007437	46.52913	26
TCGAT	1698115	1.0980392	5.959542	11
CGTGA	1290790	1.0881906	7.12831	26
GTTGA	15813520	1.0828551	12.585907	43
GGGGT	22953890	1.0789067	8.127305	2
TAGGC	1271910	1.0722739	8.70773	13
GTCAC	133770	1.0649201	46.69826	29
GTTAT	20040060	1.0525486	8.813194	31
AACAA	139380	1.0497532	44.601746	38
TGGAG	11679400	1.042704	9.750375	1
GTGGC	2346150	1.0413393	30.765331	9
GTATT	19656235	1.0323894	5.052646	31
GGGTA	11518055	1.0282996	14.309847	2
AGTTG	14986165	1.0262008	9.722048	38
TGTAG	14923220	1.0218905	8.107189	21
AGTAA	3172475	1.0217967	7.0734906	9
AGTAT	7760415	1.0093452	11.317609	30
ATTTC	2022350	1.0030171	5.2370095	22
GGTTG	27550825	0.99326146	6.6916695	42
TGTAA	7603170	0.98889333	18.16031	21
CGTGC	234385	0.98237014	5.1594286	13
TTAAT	9817665	0.97940946	12.347362	4
TAAGT	7492055	0.9744412	6.4211593	7
CGATC	121250	0.9652504	5.9465547	44
CGTGT	2800355	0.95334685	5.4022193	41
GTTTG	33957230	0.93899333	6.3759885	18
TGGGG	19790640	0.9302237	8.528713	1

TTGGG	25181395	0.9078388	5.468817	36
TTATC	1814180	0.8997718	9.852035	37
GTGGT	24924345	0.89857167	6.966785	9
AAGAC	226580	0.8984531	9.140231	32
TAGAC	556585	0.8912382	11.835335	25
TGGTT	32181305	0.8898851	7.706222	1
AGTGA	5235610	0.88781124	5.184573	18
GGAGC	806300	0.8862264	9.528442	27
ATCTC	144975	0.88522303	37.37364	42
GGGGG	14425880	0.88403314	5.6491294	2
GGATA	5178895	0.878194	6.9067693	2
GGGTG	18318310	0.8610195	7.9800105	2
AAGGC	404370	0.8441893	10.838173	46
GAAGC	387395	0.80875117	11.058362	4
GGTAT	11802165	0.8081715	5.8901753	2
TCCCA	10370	0.7795534	6.8886704	26
TGGGT	21350675	0.76973385	9.108249	1
CGGCC	14875	0.7675556	9.677927	1
TGGTG	21235895	0.7655958	5.9953465	1
GGTAA	4371715	0.7413192	5.891841	2
GGTAC	871305	0.73454696	10.368575	3
GTTGG	20322290	0.7326585	5.210644	39
GGAAC	348915	0.7284178	6.7771893	27
CACAT	47870	0.7238266	6.7755127	47
GAGTC	798590	0.6732451	12.250048	21
TGGTA	9598700	0.6572858	5.18481	1
TGAAC	365275	0.58490086	9.82881	20
CTGAA	319930	0.5122916	9.556017	19
TGGAT	7387285	0.5058558	5.0756035	1
TCTCG	157220	0.5054224	19.764366	43
GATTC	776055	0.5018145	5.064189	29
CTCGT	154090	0.4953603	19.802946	44
TGGGC	1085030	0.48159087	5.0829053	13
AGTCA	276780	0.4431972	9.598746	28
ACAAT	135480	0.4120504	17.973808	39
GAACT	205490	0.32904324	9.678291	21
ACAGT	142025	0.22741918	9.527405	32
GTCAA	141075	0.22589798	9.487965	35
AATCT	154355	0.18957642	7.4885225	41

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	249541	0.32930222781816465	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	159024	0.2098531202349747	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	157599	0.20797264498385007	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	95513	0.1260419878320451	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATG	82612	0.10901741855853034	TruSeq Adapter, Index 8 (97CGGGTTTA
82490	0.10885642348439897	No Hit	