FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD19_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`

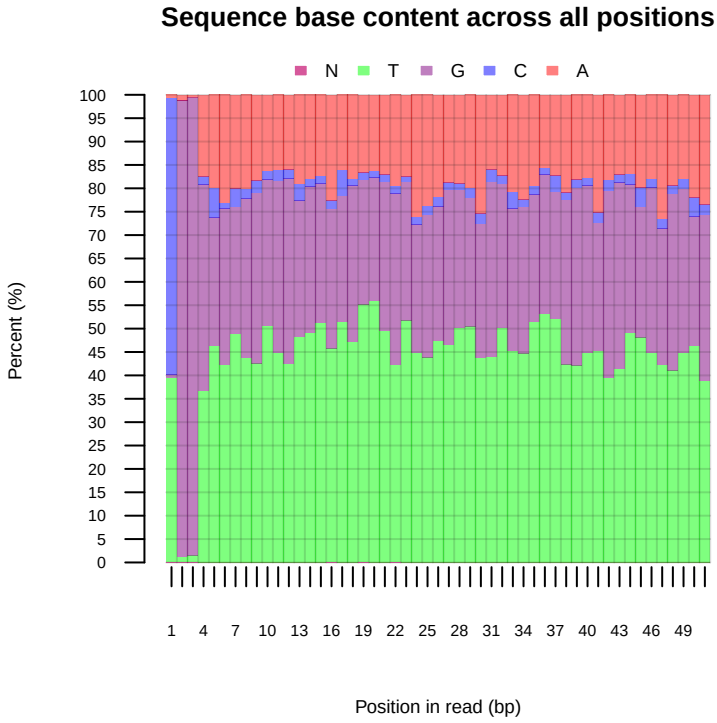
1 Sequence Duplication

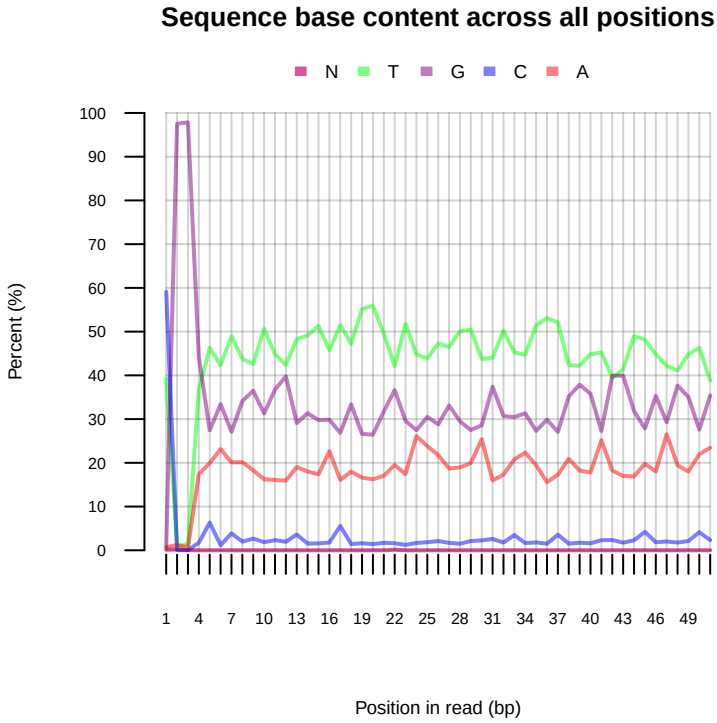
- Estimated Duplication rate 83.1077%

2 Per base sequence quality

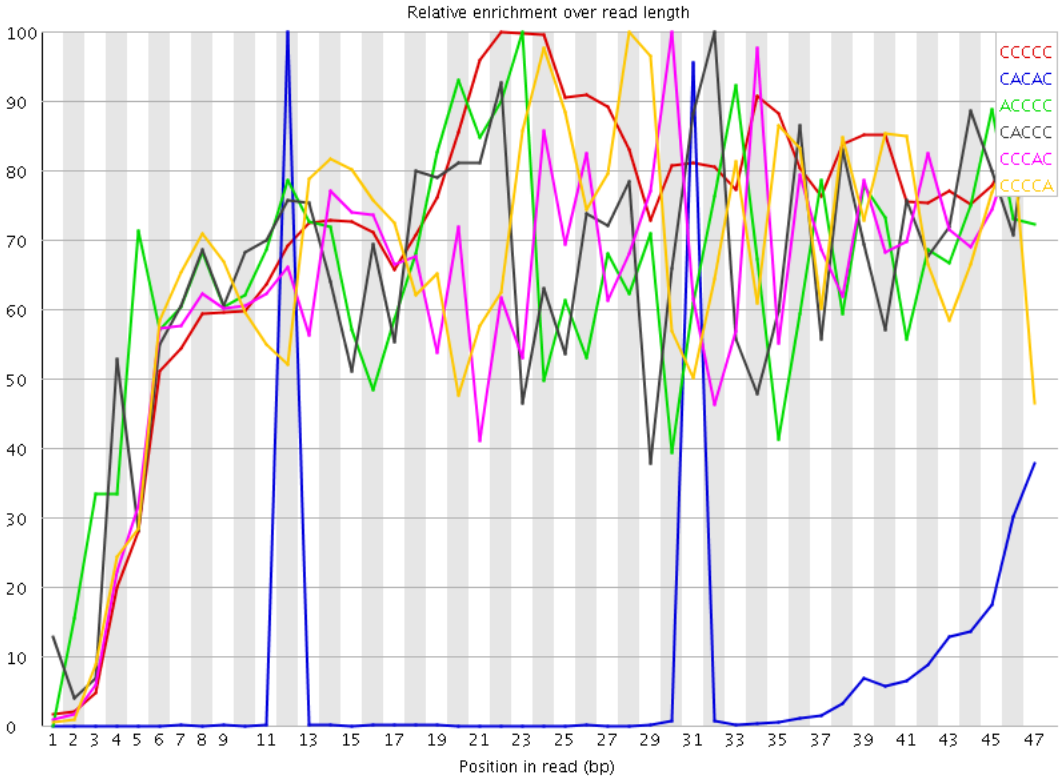
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	166520	1310.7797	1833.3102	22
CACAC	627080	159.41338	2134.4617	12
ACCCC	45645	64.56926	100.03606	23
CACCC	43340	61.308613	94.71827	32
CCCAC	41670	58.946236	93.72123	30
CCCCA	41495	58.698677	89.733086	28
CCACC	41165	58.231865	97.04468	31
CGGGC	4151535	29.55527	1088.8938	1
ACTCC	179545	18.78938	841.4236	23
CTCCA	177450	18.570135	838.7658	24
GCCCC	20310	15.460642	28.079033	45
CCCGC	19700	14.996289	28.078478	34
CGCCC	19675	14.977259	24.680445	36
CCGCC	19475	14.825012	24.85929	27
AGCAC	599875	14.74745	215.09709	10
CCCCG	19230	14.63851	25.396545	41
CGCGG	1932315	13.75638	358.75812	5
GCACA	527705	12.973208	214.53082	11
GCGCG	1802420	12.831642	357.45688	4
CGGCG	1456890	10.371773	296.35794	1
GGCGC	1424935	10.144281	358.56577	3
CGGAA	3960935	9.416904	208.01933	1
ACACG	381070	9.368303	207.72102	13
CGCGC	106220	7.8194857	51.78214	13
CGGGA	5746630	7.3520546	221.01573	1
TCGCG	1109795	6.043955	23.286602	30
AGACG	2461715	5.8525915	68.26448	27
CGGGT	10675415	5.622345	216.50778	1
CGGAG	4204835	5.3795314	159.51305	1
TCCCC	9085	5.290476	18.616503	3
ACGCG	387840	5.1309032	15.260861	4
CGCGT	912755	4.9708724	21.275251	31
CGGTT	12098560	4.8743715	169.14119	1
AACCC	18640	4.7385745	8.779614	32
CGGTC	856835	4.6663322	177.64162	1
CGGAC	352315	4.660927	154.33499	1
CTCCC	7960	4.6353536	13.818073	24
AGGCG	3599520	4.6051106	74.24576	47
ACGTC	454765	4.6023583	85.64331	15
CGGGG	6651660	4.5794206	115.93663	1
CGTCG	835070	4.5477996	21.718576	41
CCCCT	7595	4.422803	7.798542	41
TCGAG	4508520	4.4124694	66.21376	44

GGGCG	6397930	4.404737	110.34994	2
CCTCC	7490	4.361658	7.3880305	46
AGATC	2373465	4.3166366	21.933125	43
GAGAC	1799220	4.2775464	65.31307	26
CCCTC	7280	4.239369	6.977586	47
CGCGA	317095	4.1949863	20.787094	5
CAACC	16315	4.147524	7.943462	31
ACACC	15865	4.0331273	6.5698957	44
AAAAA	2625775	3.8740668	10.71502	31
ACCCA	15010	3.815773	7.4059343	33
CCCAA	14975	3.8068755	6.868407	29
TTACG	5051160	3.781734	57.044155	14
ACCAC	14865	3.7789114	5.55456	45
CCAAC	14830	3.7700138	7.8240104	30
CGAGG	2944030	3.766498	81.947495	45
CGGTA	3812050	3.7308369	131.46527	1
CCACA	14450	3.6734123	6.868407	35
CGACG	273465	3.6177866	23.27048	24
CACCA	13965	3.550118	6.1517043	38
AACTC	186665	3.5105178	155.65173	22
TACGT	4605220	3.447865	58.26382	15
CGTTT	11178150	3.4451418	42.294334	17
TCACA	182590	3.433882	156.03566	30
ACACT	182025	3.4232564	156.16821	32
ACGTT	4556910	3.4116962	59.996048	16
CACGT	329300	3.332615	85.686104	14
GACGG	2560185	3.275419	36.2706	28
ACGGG	2464565	3.153086	36.357765	29
GGCGG	4576230	3.1505642	33.48056	11
CGGAT	3151035	3.0839045	104.740585	1
GGCGT	5849130	3.0805197	49.75464	3
GTCGA	3032785	2.9681737	65.60542	43
AGAGA	6830800	2.9184434	22.604593	25
GATCG	2927340	2.8649752	13.544295	44
GAGGC	2170450	2.7768042	61.96945	46
TTTCG	8858835	2.7303212	12.8337965	30
CGAGA	1127775	2.6812232	31.242834	25
GCGGC	360080	2.5634522	9.704211	12
ATCGC	252810	2.5585132	34.06235	29
GGAGG	20502465	2.5366225	29.59544	39
TTCGA	3385590	2.534745	28.024647	31
CGGTG	4811850	2.534223	48.77101	1
AGAGC	1054345	2.506647	21.70664	8
TTTTT	142552680	2.4864013	5.2098856	16
AAGCG	1022560	2.4310799	40.06194	8
TCGTT	7781620	2.3983202	6.317192	36
AGCGA	1004125	2.387252	40.556824	9
CGTTA	3171660	2.3745785	25.815258	9
CGTTC	566200	2.358851	33.115204	33
GCGGG	3405385	2.344481	34.133675	12
GGGAG	18566835	2.2971408	26.519318	38
TTTTA	54065190	2.2907429	12.804488	26
TTCGC	545480	2.2725291	6.366012	33
ACGGA	953370	2.2665849	15.482615	30
GGTCG	4269515	2.2485952	37.161137	42
ATTCG	2991760	2.2398896	37.67149	34
CGTAG	2281935	2.23332	25.268185	5
TTCGT	7230680	2.2285187	6.2034373	35
TTTAG	40050060	2.2182453	15.941203	27
TTTAC	3797480	2.1749415	42.67523	13
GCGGA	1694050	2.1673133	20.257835	7
TCGGA	2203025	2.1560912	10.959356	46
GGAAG	9366805	2.153556	11.221107	2
AGTAG	12215890	2.1485333	16.425644	35
GAGGT	22625835	2.1414435	22.956772	40
AGGTC	2167205	2.1210341	62.612473	41
ATCGG	2137655	2.0921137	10.703659	45
GCGGT	3955615	2.0832756	34.396866	6
AGGAG	9032595	2.0767167	8.284786	38
ATTTT	48577295	2.0582206	8.848572	25
GACGC	155110	2.0520172	15.29082	3
GCGTT	4908600	1.9776188	22.92106	16
CGAGT	2001800	1.9591532	33.24967	33
ATGCC	193400	1.9572661	90.24591	47
TAGAG	10951335	1.926123	10.198741	24
GTCGC	353560	1.9254911	10.255445	3
GAAGA	4496540	1.9211363	5.1125264	46
AATTT	18620180	1.916484	19.25536	24
ATCGT	2556865	1.9142897	16.323069	39
AGCGC	144595	1.9129097	9.513965	10
AAACG	427040	1.8866613	9.205705	7
TAGTA	13890555	1.8689115	18.306086	29
CAGTC	183480	1.8568727	84.58287	27
TCCAG	182695	1.8489282	83.91712	25
TACGG	1888465	1.8482326	15.712799	5
GTCAC	182490	1.8468536	84.54482	29
TACGC	182265	1.8445765	9.841133	13
CCAGT	181225	1.8340514	83.81489	26
GCGTA	1851955	1.8125004	24.833838	4
ACGGC	136585	1.806942	8.883015	6
TCGTC	432235	1.8007381	9.770486	40
TTAGT	32361650	1.7924087	15.330237	28
GAGAT	10169805	1.7886674	8.646178	26
TAGTT	32132825	1.7797348	8.2573395	29
TATCG	2367895	1.7728105	16.841173	38
CGAGC	131010	1.7331879	5.0165024	13
GGAAA	4028595	1.7212076	11.444235	2
AAGAG	4025640	1.7199453	5.06314	47
AGGTA	9759465	1.7164967	24.976301	47
GTAGA	9733110	1.7118614	9.959972	23
GAGCG	1338020	1.71182	10.441349	28
AACGG	718755	1.7088002	14.503464	29
GGAGA	7427945	1.7077858	10.659702	2
GAAAA	2122900	1.6854823	5.0515437	3
GAGCA	706385	1.6793914	21.297218	9
GGACG	1302100	1.665865	15.58822	2
TCGTA	2217770	1.6604141	8.177587	43

TATTT	38968100	1.6510789	6.9280505	32
AACGC	66950	1.6459123	11.366789	11
AGTTA	12088520	1.6264559	17.336481	30
AGTTT	29061765	1.6096387	6.7796993	26
TGGGA	16676960	1.5784066	15.737542	37
CGTGG	2990770	1.5751276	32.18544	5
CGTAC	155480	1.5735042	8.806851	13
AGTCG	1603865	1.569696	12.836471	22
AGCGT	1602490	1.5683502	8.517691	29
GAACG	656785	1.56147	14.295124	28
TAATT	15059565	1.5500075	19.061333	23
TAGGA	8810045	1.5495125	6.529002	37
GTACG	1569775	1.536332	15.411815	4
ATCTC	198295	1.5351733	69.10805	40
GCGTG	2907845	1.5314542	32.301754	4
GTCGT	3778625	1.5223647	10.608035	3
ACGGT	1550440	1.5174091	15.166039	6
GGAAT	8620130	1.5161101	10.430864	2
GGGAA	6532560	1.5019246	12.991961	2
CGATT	1999025	1.4966427	14.864636	11
TGGCG	2824125	1.4873619	29.616896	10
CACGC	10850	1.4842827	6.7493587	31
ACGAG	622260	1.479389	5.2688594	32
AGGTT	20383365	1.4758087	13.590161	41
GGTTT	49457085	1.4740763	8.868393	2
GCGTC	263730	1.4362762	8.453776	40
TGGAA	8155730	1.4344314	8.774419	1
TTATT	33809830	1.4325229	6.390662	32
GCGAC	108275	1.4324166	16.780773	23
GGCGA	1110245	1.420412	9.961525	2
CACTT	183145	1.417884	64.57356	33
TTCGG	3503900	1.4116812	20.483896	35
GTAGT	19143720	1.3860552	7.321072	36
GTTTA	24822630	1.3748465	8.678309	12
CGCAC	10025	1.3714224	6.5243793	31
TTTAA	13244085	1.3631489	6.510036	5
AAGTA	4147445	1.3555406	8.717792	34
AGATA	4145040	1.3547546	5.520531	26
GGTTA	18602615	1.3468778	17.53154	2
TTGAG	18520745	1.3409503	12.135235	44
TTAAG	9869465	1.327892	7.6654854	6
GGTAG	14008805	1.3258765	7.901413	2
TCGAC	131005	1.32581	6.1462502	33
TATAG	9845895	1.3247207	12.8958435	47
TTATA	12806235	1.3180832	10.157201	46
TTGTA	23718120	1.3136712	14.835079	20
GGAGT	13876470	1.3133515	10.025285	2
GTTAA	9725965	1.3085848	21.605854	3
GACGT	1321695	1.2935374	5.7521367	3
CGTAT	1721545	1.288897	7.644402	44
TCGTG	3101280	1.2494702	8.052381	40
GTAAT	9281045	1.2487227	23.573027	22
GGGTT	31839965	1.240543	12.512476	2
GCGAT	1266405	1.2394252	17.544254	10
AAAAC	150690	1.2371567	14.72673	6
ATTAT	11718725	1.206151	10.02509	45
GAGTA	6855080	1.2056729	11.661388	34
GGGGA	9713950	1.2018372	10.661811	2
TTTGT	52557070	1.1983255	7.1537595	19
AGTAT	8853110	1.191146	17.425365	30
GGTGG	23338310	1.1886566	11.158706	8
TCGGG	2230915	1.1749401	25.457277	36
TATTC	2023180	1.1587417	27.58526	33
CGTCT	277005	1.1540331	35.00915	16
GTATT	20549685	1.1381816	7.644344	31
TGTAA	8453870	1.1374302	23.159966	21
GGGAT	11931340	1.1292529	9.21572	2
CGTAA	620055	1.1276981	7.3257914	21
TGAGG	11874340	1.1238581	14.403796	45
TCGAT	1480620	1.1085198	7.661511	11
AAGGC	461825	1.0979635	19.297943	46
GATTA	8155310	1.0972602	12.58359	44
GGGGT	21319045	1.0858122	8.547748	2
TAAGC	594280	1.080821	29.381107	7
TGGAG	11416175	1.0804946	9.966805	1
GGGTA	11362440	1.0754088	14.657762	2
TTAAT	10422760	1.0727637	15.007814	4
CGTGA	1090835	1.0675955	9.667822	26
TAGGC	1081850	1.058802	7.8343253	13
GGATT	14600230	1.0570946	6.915642	43
CGTGT	2565440	1.0335864	7.6279497	41
ATTTC	1803305	1.032812	6.6497164	22
TTTTC	4370090	1.0303379	9.077334	29
AGTAA	3151860	1.030146	5.396452	9
GTTAT	18197995	1.0079291	7.393931	31
TTATC	1757780	1.0067384	12.585177	37
TGTAG	13867185	1.0040203	6.978752	21
CGTGC	183730	1.0005953	6.8748055	47
GGAAC	416280	0.9896826	13.46455	27
CGAAC	39570	0.972797	8.836995	9
GAATA	2958295	0.96688175	5.2158737	3
GGTTG	24676430	0.9614386	6.3752294	42
AGTTG	13112315	0.9493658	7.3044205	38
GTTGA	13073685	0.94656885	11.594291	43
TGCGG	1792605	0.94409853	7.3606877	5
GTGGC	1779920	0.9374178	27.958235	9
TGGGG	18357585	0.9349805	9.147248	1
GGATA	5307700	0.9335192	7.5673475	2
GTGGT	23850370	0.9292538	8.408058	9
TAAGG	5281725	0.9289509	5.5451713	45
TGGTT	31126790	0.9277389	7.4732184	1
GGAGC	719590	0.92062044	9.589821	27
TAGAC	505615	0.9195654	10.359983	25
GTTTG	30705115	0.9151708	7.105406	18
TTGGG	23482645	0.9149266	6.987376	36
ATTAC	653880	0.90973073	5.998991	29
TTTGG	30506135	0.9092402	5.738891	35

AAGAC	201545	0.89042515	7.126651	32
GGGTG	16966570	0.86413395	8.1475725	2
TCTCG	205965	0.85807264	37.16452	41
AGTGA	4878525	0.858036	5.0441613	18
CTCGT	204605	0.85240674	37.16651	42
GGTAC	868000	0.84950787	15.518322	3
GGTAT	11322645	0.819789	6.317874	2
CGATC	77765	0.7870051	6.3817506	44
GGTAA	4431210	0.77936214	6.502255	2
AGTGG	8200335	0.7761284	5.2694707	8
GGGGG	11582355	0.7711372	5.8649426	2
TGGGT	19765145	0.770086	8.918962	1
TGGTG	19696310	0.7674041	6.295238	1
GAATC	421845	0.7672123	16.314648	38
GAAGC	316725	0.7529962	6.7160783	4
GTGCG	1420835	0.74830115	6.6634474	4
TGGTA	9102245	0.65902627	5.551511	1
GAGTC	656255	0.64227396	11.654175	21
TGAAC	306665	0.55773365	16.06095	20
GATTC	719580	0.53873974	6.6509833	29
TGGAT	7294070	0.52810967	5.106474	1
CACAT	26725	0.50260407	6.5922465	12
CTGAA	231820	0.42161253	15.80295	19
GGTGC	750925	0.39548433	6.7299137	3
GAACT	206520	0.3755993	15.763588	21
AGTCA	195740	0.35599363	15.61475	28
TATGC	462935	0.34659308	6.85393	46
AATCT	206230	0.28692386	12.336947	39
TACAC	15170	0.2852948	5.580434	31
TCTGA	250845	0.18780422	6.532705	18
CTTGA	200280	0.1499469	6.419856	35
ACTTG	198455	0.14858057	6.4165144	34

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
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	246523	0.349886839553871	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	176772	0.250890166035692	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCC	118989	0.1688795169281388	TruSeq Adapter, Index 8 (100CGGGTTT
113469	0.16104505379756937	No Hit	
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	95049	0.13490179095969093	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	90241	0.12807786003002103	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	76546	0.10864072731749416	No Hit