FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD18_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`

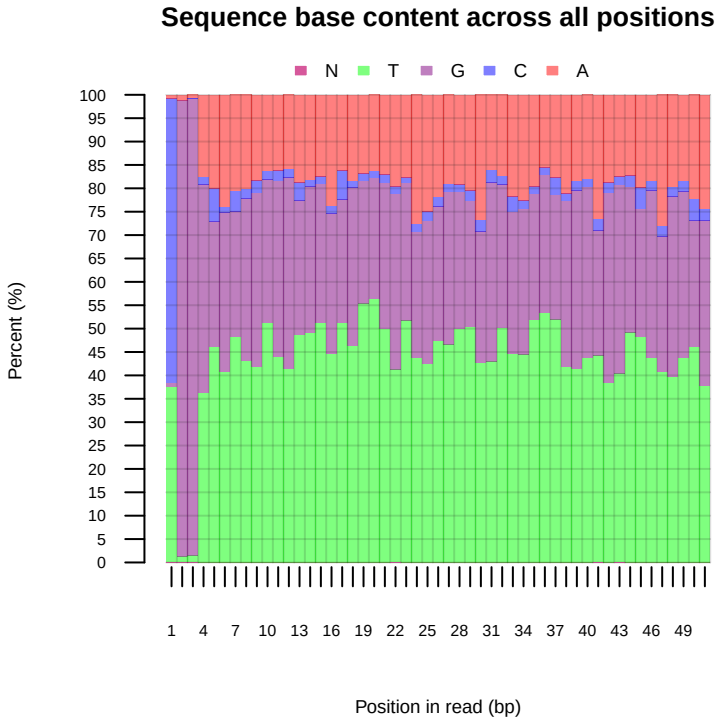
1 Sequence Duplication

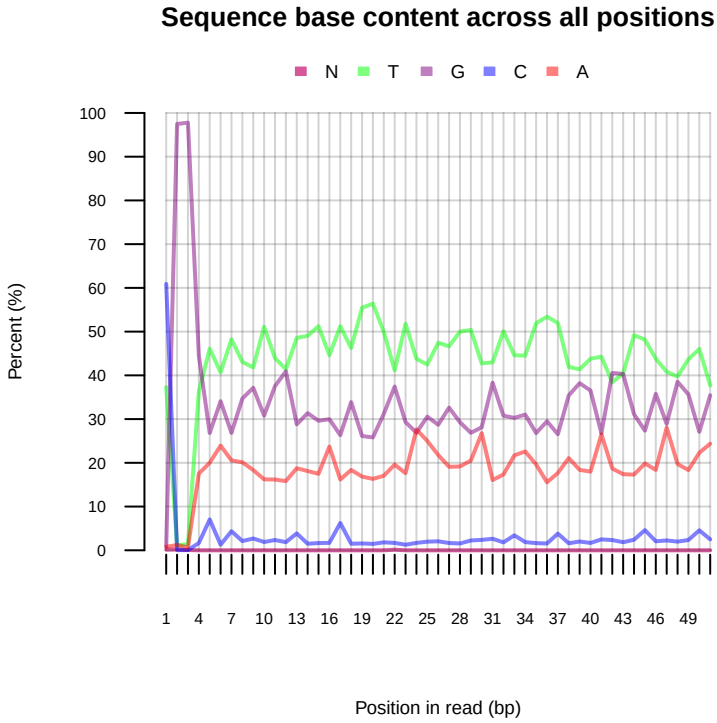
- Estimated Duplication rate 83.6507%

2 Per base sequence quality

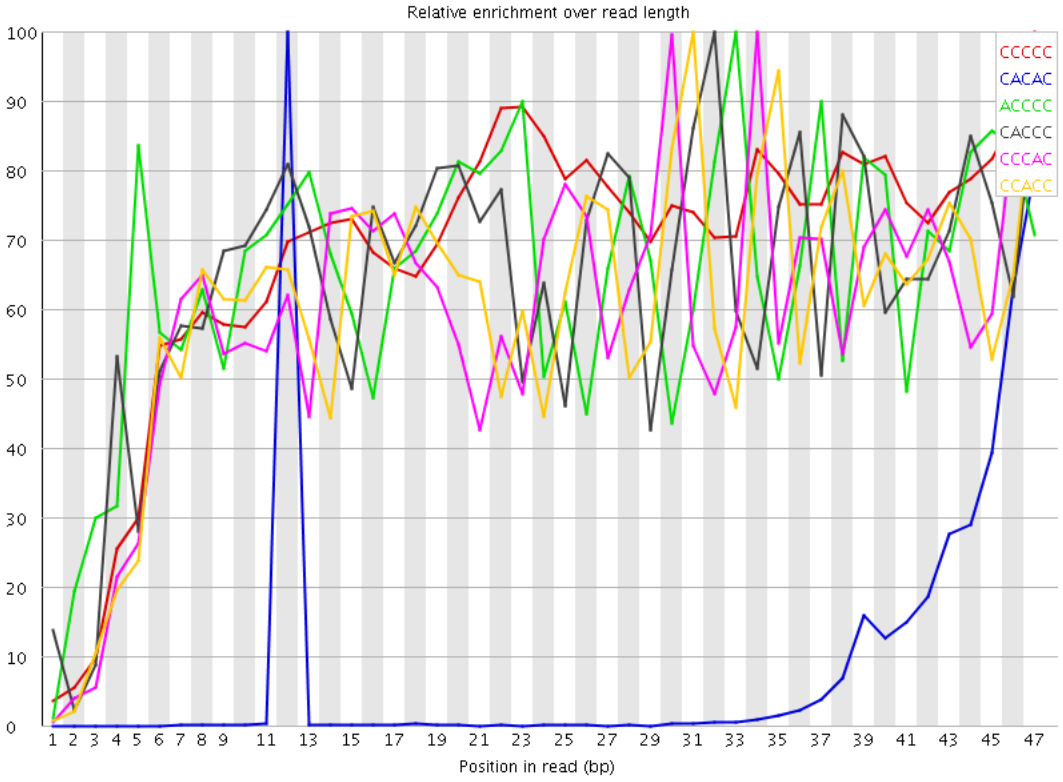
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	189420	1041.1271	1526.3779	47
CACAC	584210	109.47668	1202.1321	12
ACCCC	50890	51.64731	79.399345	33
CACCC	48770	49.49576	76.77654	32
CCCAC	46450	47.141235	79.399345	34
CCACC	46065	46.750507	77.730286	31
CCCCA	46025	46.70991	75.58436	24
CGGGC	5367220	30.646088	1125.0046	1
CACCA	147570	27.653536	1144.0616	31
AGCAC	819030	15.544199	126.77077	10
CGCGG	2454635	14.015628	370.8647	5
GCGCG	2328465	13.295215	368.56137	4
GCACA	700055	13.286198	126.3783	11
CCCGC	21625	12.037878	22.233109	26
GCCCC	21425	11.926545	23.410624	45
CCGCC	21110	11.751196	23.280348	43
CCCCG	20900	11.634295	21.448412	25
CGCCC	20890	11.628729	21.056448	44
ACTCC	141245	11.228573	486.05328	23
CTCCA	138135	10.981336	484.37128	24
TCACC	136390	10.842614	486.89267	30
GGCGC	1826070	10.426609	369.60684	3
CGGAA	5218410	10.0305	199.6407	1
CGGCG	1748650	9.984551	275.31332	1
CGCGC	171815	9.686588	53.03484	13
ACACG	468685	8.895073	122.552345	13
CGGGA	6866045	7.238851	212.44104	1
TCGCG	1406175	6.2099104	25.80457	30
AGACG	3024555	5.8136096	69.798256	27
AGATC	3794955	5.6417174	27.912786	43
CGGGT	12443335	5.565429	213.44223	1
CGGAG	5002325	5.2739363	154.38432	1
ACGCG	486480	5.0642138	20.499962	6
CGCGT	1145350	5.0580626	24.060472	31
CGGTT	14047980	4.85955	169.29831	1
CGTCG	1067810	4.7156324	21.874979	41
AGGCG	4421980	4.66208	79.34962	47
GGGCG	8028910	4.642994	119.42461	2
CGGGG	7906510	4.5722117	112.7228	1
CGGAC	437725	4.5566792	147.55391	1
CGGTC	1031685	4.5560985	168.00133	1
TCGAG	5534605	4.5130477	71.90375	44
GAGAC	2252670	4.329941	67.05125	26

CGCGA	404735	4.213256	20.898573	5
TCCCC	9785	4.2128396	18.115976	3
ACGTC	498145	4.010727	52.385426	15
CGAGG	3660820	3.8595917	87.90019	45
AAAAA	3255735	3.8407288	10.687357	31
AACCC	20360	3.8153148	6.956121	32
TTACG	6022995	3.7985349	63.145573	14
CGACG	361730	3.7655778	26.584843	24
CGGTA	4389400	3.5792208	125.51755	1
CTCCC	8285	3.5670288	10.01398	24
CGTTT	13238595	3.5419664	47.66	17
TACGT	5524915	3.4844098	64.43717	15
ACGTT	5500455	3.4689832	66.17313	16
ACACC	18250	3.4199164	5.239323	43
GATCG	4193740	3.4196749	16.472952	44
CCTCC	7925	3.4120347	6.5748367	24
GACGG	3202830	3.3767343	38.02344	28
CAACC	17895	3.353392	6.559886	31
CCCTC	7620	3.2807198	5.7656384	23
ACCAC	17385	3.2578216	5.635451	45
GGCGG	5630190	3.2558513	36.693893	11
GGCGT	7249980	3.2426395	55.303856	3
ACGGG	3028995	3.19346	38.107407	29
CCCAA	16995	3.1847384	5.8554688	14
GTCGA	3900180	3.180299	71.45868	43
CCCCT	7385	3.1795425	5.4621716	34
ACCCA	16890	3.1650622	5.8994946	33
CCAAC	16805	3.149134	6.3397555	34
CCACA	16685	3.1266468	5.0629992	35
CACGT	383925	3.0911047	52.113037	14
CGGAT	3663070	2.986954	100.76193	1
AGAGA	8277660	2.9378479	23.747734	25
GAGGC	2680125	2.8256478	65.64597	46
AGAGC	1432460	2.7533846	16.267427	47
CGAGA	1421025	2.731405	34.548695	25
TTTCG	10046770	2.6879983	12.204178	30
GCGGC	470165	2.6845775	8.644494	9
ATCGC	330190	2.6584673	38.330986	29
CCAGA	139620	2.6498182	119.32411	33
ACCAG	139590	2.6492488	119.33303	32
CGGTG	5854200	2.6183603	47.588966	1
GGAGG	24141080	2.577719	32.99802	39
TCGGA	3155890	2.5733874	14.458339	46
ATCGG	3138025	2.5588198	14.183792	45
TTTTT	154000630	2.4962127	5.2998934	16
TTCGA	3839300	2.4213395	26.075298	31
CGTTC	703105	2.4015214	26.590334	33
GCGGG	4128230	2.3872912	37.410225	12
TCGTT	8885230	2.37723	6.0886703	4
AAGCG	1232840	2.3696876	38.290188	8
GGTCG	5257325	2.3514009	40.50961	42
TTTTA	61531535	2.3510287	14.760544	26
GGGAG	21932335	2.3418748	29.829567	38
TTCGC	674060	2.3023155	7.7100415	33
CGTTA	3603290	2.2724946	22.096237	9
AGCGA	1181750	2.2714856	38.866413	9
TTTAG	45755625	2.260395	18.24755	27
ATTCG	3579330	2.2573836	40.242004	34
GAGGT	27096590	2.2377617	25.873024	40
AGGTC	2740610	2.2347581	68.20209	41
CGTAG	2720630	2.218466	28.066439	5
AACTC	150315	2.206435	92.500084	22
GCGGA	2090910	2.20444	22.942722	7
TTCGT	8220915	2.1994936	5.7476015	35
TTTAC	4493910	2.192039	47.829094	13
ACGGA	1138105	2.187594	13.859668	30
ATTTT	56675695	2.1654942	10.267965	25
GCGGT	4834415	2.1622493	37.64944	6
GGAAG	11098330	2.1605182	10.981864	2
AGTAG	13861455	2.0870337	15.152416	35
GACGC	197935	2.060486	18.36482	5
AATTT	22541125	2.030191	21.983892	24
GCGTT	5855060	2.0254128	25.419365	16
AGGAG	10210200	1.9876254	8.419073	38
GTCGC	449525	1.9851798	11.256154	3
GAAGA	5536415	1.9649448	6.39555	46
GAGAT	12946715	1.9493068	9.622725	26
ATCGT	3070375	1.9363998	17.32954	39
TCGTC	565405	1.9311941	9.269456	40
TAGTA	16546235	1.9268156	20.993183	29
GAGCA	985545	1.8943526	13.163351	9
TAGAG	12580050	1.8941005	10.934977	24
CGAGT	2288245	1.8658892	30.276646	33
TTAGT	37448460	1.8500091	17.606424	28
GCGTA	2245505	1.831038	27.737122	4
TACGC	226765	1.8257589	9.319817	13
AGCGC	173365	1.8047146	11.9301605	35
TAGTT	36232360	1.7899318	9.145838	29
TACGG	2194480	1.7894309	17.262709	5
AAACG	510435	1.7887361	9.393163	7
TATCG	2834280	1.7875013	17.782654	38
AAGAG	4948460	1.7562721	6.391134	47
ACGGC	166080	1.7288781	8.726283	12
AGGTA	11451315	1.7241538	27.122677	47
TATTT	45098275	1.7231382	8.105249	32
CGAGC	165270	1.7204462	5.568876	32
GTAGA	11268420	1.6966165	10.643486	23
TAATT	18417350	1.6587785	21.78916	23
GGAAA	4673370	1.6586391	11.077883	2
GGAGA	8496620	1.6540418	10.35838	2
AACGG	858580	1.6503086	12.678317	29
GGACG	1561430	1.6462108	15.6388	2
AGTTA	14048400	1.6359417	18.869751	30
CGTGG	3642215	1.6290238	35.664204	5
GAGCG	1539270	1.6228476	9.092701	28
AGCGG	1537455	1.6209342	5.49789	6
TCGTA	2562750	1.6162549	5.7116103	43

GGAAT	10671800	1.606787	10.3534975	2
AGTTT	32463450	1.603742	6.7143917	24
GCGTG	3577135	1.599916	35.7924	4
TGGGA	19268025	1.5912428	18.03949	37
AGTCG	1923275	1.568284	14.616352	22
GCGTC	352855	1.5582684	9.427295	40
TGGCG	3471240	1.5525534	32.492542	10
GTACG	1888895	1.5402497	16.969408	4
GTCGT	4449835	1.53931	10.528571	3
GAACG	795025	1.5281471	12.375301	28
GCGAC	146395	1.5239592	16.44008	23
TGGAA	10064740	1.5153859	8.453515	1
ACGGT	1847550	1.5065359	16.637054	6
AACGC	79265	1.5043539	7.6068654	11
AGGTT	23441510	1.4972873	15.341581	41
GGTTT	54892330	1.4874079	8.664146	2
TAGGA	9849030	1.4829077	6.7037387	37
CGTAC	183570	1.4779819	8.462931	13
TTATT	38368795	1.4660147	7.070907	32
ACGAG	759960	1.4607472	5.429887	32
TTCGG	4219915	1.4597749	22.398169	35
AGCGT	1785940	1.4562978	7.5544105	29
GGGAA	7466820	1.45357	12.723785	2
CGATT	2299990	1.4505396	14.821795	11
TTGAG	22388440	1.4300241	13.676678	44
GTTTA	28181710	1.392218	10.143396	12
TTGTA	27997535	1.3831193	17.060587	20
GTAGT	21613005	1.3804946	7.7216415	22
GGTTA	21505485	1.3736268	18.751251	2
GTTAA	11741435	1.3672948	23.537552	3
TTTAA	15126535	1.3623878	6.515507	5
TTTTG	64789945	1.357833	5.151323	34
TCGAC	167630	1.3496438	6.490027	23
AGATA	4881565	1.3399886	5.577251	26
GTAAT	11446965	1.3330036	26.66658	22
GGCGA	1264140	1.332779	8.827163	2
TTATA	14681700	1.3223232	9.570512	46
TTAAG	11288170	1.3145119	7.610355	6
TATAG	11048825	1.28664	11.9583025	47
GGAGT	15573720	1.2861499	9.914962	2
GGTAG	15543625	1.2836645	7.5426373	2
AAGTA	4669005	1.2816409	7.6303234	34
TCGGG	2808170	1.2559873	27.617916	36
GACGT	1535500	1.2520831	6.1176744	3
TCGTG	3616545	1.251054	8.401922	40
CGTAT	1959660	1.2359029	5.2260613	44
GGGTT	35152960	1.2315725	12.124827	2
AGTAT	10569475	1.2308196	20.035841	30
ATTAT	13620975	1.2267878	9.484993	45
TTTGT	58465920	1.2252976	8.308884	19
GCGAT	1489555	1.2146184	17.323095	10
TGTAA	10407950	1.2120099	26.167582	21
GGTGG	26508175	1.2007638	12.644521	8
GTATT	24287170	1.1998218	9.002092	31
TATTC	2454860	1.1974314	29.498964	33
ATGCC	148360	1.1944947	54.758636	47
CGAAC	62065	1.1779187	5.578074	9
GGGGA	10944785	1.1686544	10.328327	2
GAGTA	7620505	1.1473724	11.009542	34
TCCAG	141805	1.1417183	50.953503	25
CAGTC	141485	1.1391417	51.409367	27
TGAGG	13745115	1.1351352	15.928491	45
GTCAC	140320	1.129762	51.293987	29
CCAGT	140035	1.1274673	50.913773	26
TTAAT	12512480	1.12695	16.615587	4
AAAAC	175080	1.1185715	15.019175	6
CGTAA	750415	1.1155941	7.9032173	21
AAGGC	577970	1.1109376	21.050283	46
TCGAT	1734625	1.0939796	6.4266005	11
GATTA	9330400	1.0865288	11.673292	44
GGGAT	13155525	1.086444	8.778902	2
GGGGT	23725290	1.074705	8.311488	2
CGTGT	3065795	1.0605357	8.03377	41
GGGTA	12791500	1.0563812	15.198834	2
TAGGC	1294015	1.0551705	8.211899	13
TGGAG	12678770	1.0470715	9.521393	1
GGATT	16211195	1.035463	6.566737	43
TTATC	2113365	1.0308571	13.476285	37
TAAGC	692015	1.0287746	28.362244	7
TTTTC	4968145	1.0280569	8.6753025	29
GTTAT	20733950	1.0242877	8.233971	31
GTGGC	2255495	1.008797	30.733818	9
CGTCT	293080	1.0010424	21.895138	16
TGTAG	15621305	0.9977848	7.757765	21
AGTAA	3633165	0.9973032	5.102093	9
CGTGA	1219735	0.99460083	7.549621	26
ATTTC	2036375	0.9933029	5.510989	22
GTTGA	15541155	0.99266535	13.029522	43
GGTTG	27711720	0.9708711	7.214713	42
TGCGG	2139620	0.9569704	8.087891	5
ATCTC	152860	0.95187867	42.401814	40
GGAAC	494670	0.9508236	11.450902	27
GTGGT	27081510	0.948792	9.698438	9
GTTTG	34783575	0.94252455	8.260541	18
TGGTT	34490510	0.9345834	7.258932	1
TCATC	149845	0.93310386	40.80022	38
ATTAC	809175	0.93039536	6.2884965	29
TTGGG	26544970	0.9299944	8.170005	36
AGTTG	14534420	0.9283618	6.950911	38
TTTGG	34129845	0.9248106	6.7397003	35
TAAGG	6089480	0.91685545	6.3003945	45
TGGGG	20134790	0.91206324	8.754969	1
GGATA	5957320	0.89695686	7.4594703	2
AAGAC	254135	0.8905746	6.7741923	32
CATCT	142815	0.88932705	40.653477	39
GGTAC	1086420	0.88589257	17.067421	3
CGATC	109625	0.8826266	7.6534753	44

GGAGC	822985	0.86767066	8.263661	27
GGGTG	18833680	0.8531255	8.575705	2
TAGAC	568195	0.8446993	8.697539	25
AGTGA	5583455	0.8406664	5.710452	18
GGTAT	12922700	0.82541585	6.357501	2
GTGCG	1750310	0.782847	7.475065	4
AGTGG	9356930	0.7727386	6.2549386	8
GGGGG	13043460	0.7639241	5.915427	2
TGGTG	21712425	0.76068777	6.311045	7
GGTAA	5030270	0.7573767	6.4349613	2
GAAGC	393415	0.75619763	7.9926476	4
TGGGT	21386480	0.7492685	8.6771	1
GAGTC	836255	0.6819021	13.445239	21
TGGTA	10207475	0.6519854	5.3650866	1
TCTCG	161310	0.5509696	23.2744	41
CTCGT	160215	0.5472294	23.280834	42
TGGGC	1157030	0.51749545	5.155111	13
GATTC	780170	0.49203146	5.401843	29
TGAAC	301235	0.4478269	10.049866	20
GGTGC	948660	0.42429942	7.531084	3
CTGAA	215095	0.319768	9.830934	19
GAACT	172060	0.25579062	9.794212	21
GATCA	153290	0.22788648	9.705884	36
AGTCA	153275	0.22786419	9.782374	28
CAGAT	152165	0.22621399	9.66432	34
ATCAT	151985	0.17475344	7.510051	37

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	289670	0.3612437024442844	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	182610	0.22773056410173909	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	116745	0.14559117631048427	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	105448	0.13150283403647217	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	102251	0.12751589677436573	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCC	87330	0.10890811107280476	TruSeq Adapter, Index 7 (100CGGGATG
86469	0.10783436913265035	No Hit	
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	81203	0.10126720878787317	No Hit