FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD14_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`

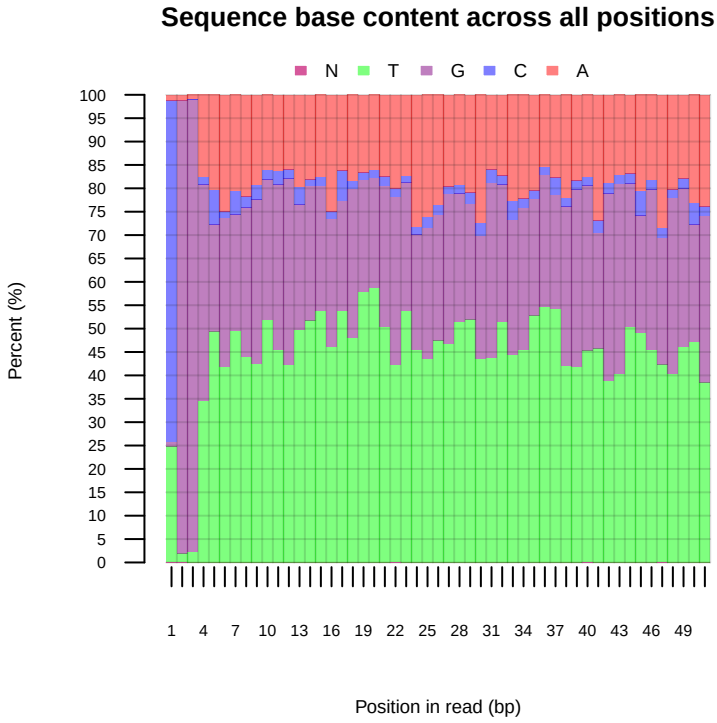
1 Sequence Duplication

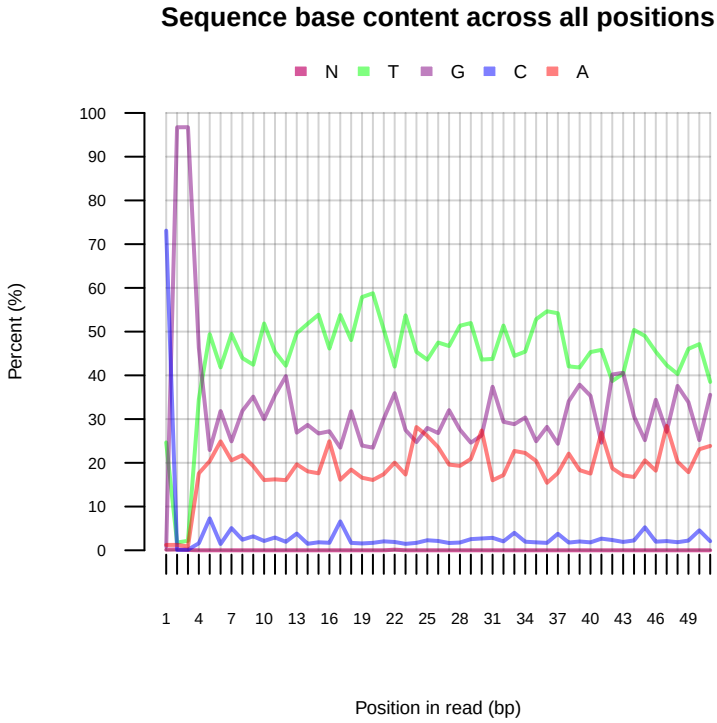
- Estimated Duplication rate 82.0519%

2 Per base sequence quality

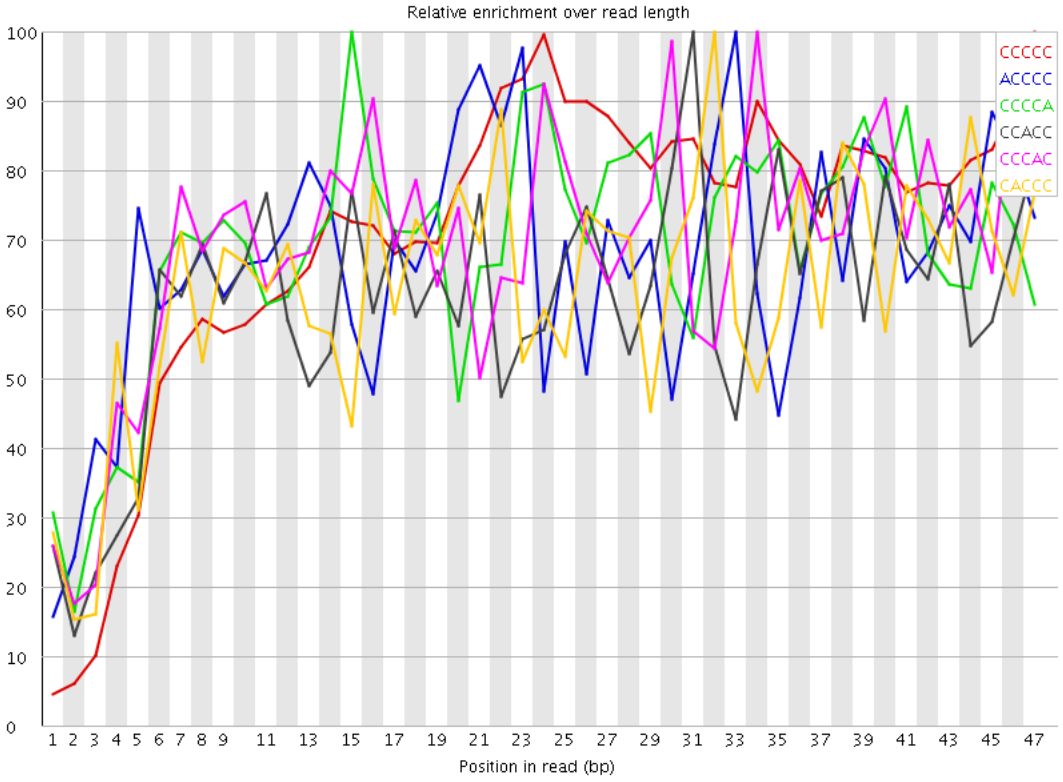
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	189500	618.4145	870.21857	47
ACCCC	55160	36.36098	53.894314	33
CCCCA	54775	36.10719	52.190754	15
CCACC	54305	35.79737	57.920902	31
CCCAC	53730	35.418335	50.796936	34
CACCC	53190	35.062374	55.443	32
CACAC	217735	28.992146	304.29376	47
CGGGC	5277835	28.195036	1076.2169	1
CGCGG	2612240	13.955003	381.65543	5
GCGCG	2479180	13.244175	380.2058	4
GGCGC	1963675	10.490263	381.55026	3
GCCCC	26535	10.205596	16.806784	37
CGGCG	1821150	9.728874	275.87842	1
CGGGA	7914525	8.540484	279.5909	1
CCCCG	21585	8.301782	14.457759	46
CCCGC	21525	8.278706	17.801126	47
CCGCC	21390	8.226783	14.457448	35
CGCCC	21155	8.1364	13.28278	32
CGGAA	4396735	8.131675	231.98402	1
CGCGC	174190	7.8957114	37.027565	13
CGGGT	15552220	7.0836086	282.4488	1
AGACG	3505315	6.4830112	74.82912	27
CGGAG	5783145	6.2405343	187.40366	1
CGGGG	9214160	5.801247	148.76678	1
TCGCG	1497300	5.7865834	28.962048	30
CCTCC	20795	5.7859507	8.759389	7
CTCCC	20405	5.6774387	9.936226	47
CGGTT	16920145	5.5752306	198.00305	1
TCCCC	19915	5.5411024	12.82058	3
AGGCG	4994180	5.3891697	95.73865	47
TCGAG	6579585	5.1363273	82.36517	44
GGGCG	7980635	5.0246177	131.25058	2
CCCCT	18055	5.0235796	7.200243	22
GAGAC	2605715	4.819218	71.675415	26
CCCTC	17260	4.802381	6.7985396	39
CGCGT	1235265	4.7739024	26.805647	31
ACGCG	512030	4.68818	17.40891	6
CGAGG	4168980	4.4987044	107.71455	45
CGTCG	1159990	4.4829884	17.91283	41
AGCAC	282790	4.4377713	36.72531	45
CGGTC	1107235	4.2791076	159.40492	1
CGGAC	452150	4.139915	135.3466	1
AAAAA	3750760	4.116175	10.62218	31

TTACG	7234230	4.0854797	67.49011	14
CGCGA	420955	3.8542912	20.179502	5
GACGG	3569195	3.8514824	43.142048	28
CCACA	28525	3.7981994	5.0366254	45
CGGTA	4822765	3.7648726	132.29068	1
AACCC	28265	3.7635796	6.139696	19
GCACA	239250	3.754506	36.599964	46
TACGT	6595650	3.724846	68.39286	15
ACGTT	6575930	3.7137094	69.91376	16
ACGGG	3416965	3.6872127	43.232548	29
GGCGG	5661355	3.5643961	32.6314	11
CGTTT	14948250	3.5632443	44.207558	17
ACCCA	26600	3.5418794	5.4433093	45
CCCAA	26345	3.507925	4.8488216	28
GTCGA	4467020	3.4871616	81.77236	43
CACCA	25710	3.4233727	4.754974	16
CGGAT	4365675	3.4080474	120.12623	1
ACACC	25170	3.35147	4.855372	19
GAGGC	3076060	3.3193455	78.17925	46
CGACG	356090	3.2603831	16.817327	24
CCAAC	24405	3.2496078	5.2242146	34
GGCGT	7084195	3.2266557	53.41578	3
CAACC	24160	3.2169852	5.113682	1
AGAGA	8594105	3.2106364	28.197134	25
ACCAC	23980	3.1930175	4.9740586	45
CGAGA	1596185	2.9521124	38.415234	25
TTTCG	12077990	2.879055	14.654016	30
GGAGG	22568755	2.870212	41.00751	39
AGATC	2112210	2.8260727	19.362043	27
AAGCG	1515855	2.8035436	53.025803	8
CGGTG	6010205	2.7374833	52.527977	7
GGTCG	5901940	2.6881714	49.63814	42
TTCGA	4706605	2.6580215	31.49249	31
GGGAG	20822240	2.6480966	37.464535	38
AGCGA	1427225	2.639624	53.753613	9
TTTTT	177706390	2.6127644	5.980803	16
GCGGG	3983630	2.5080986	33.171417	12
ATCGC	375270	2.4857032	40.765507	29
AGGTC	3181335	2.4834967	77.90669	41
TCGTT	10272985	2.448792	6.315462	4
CGTTA	4288345	2.421812	25.670704	9
TTTTA	69482635	2.4202983	14.18131	26
GCGGC	452515	2.4174073	8.50689	33
GAGGT	26267350	2.4166806	30.145828	40
GCGGT	5278735	2.4043186	43.205032	6
GCGGA	2207870	2.3824904	26.058224	7
TTTAG	49384030	2.3778398	18.688292	27
AGTAG	15034695	2.370771	21.83455	35
TTCGT	9651460	2.300638	5.772069	35
CGTAG	2924075	2.2826676	28.232706	5
ACGGA	1232435	2.2793639	11.745736	30
ATTTT	64515625	2.2472818	10.762825	25
TTTAC	5479935	2.2388382	47.85887	13
ACACG	139650	2.1915016	28.776356	47
CGTTC	778650	2.176969	24.402416	33
CGAGT	2786975	2.175641	39.312447	33
GATCG	2692085	2.1015654	12.45912	28
TAGAG	13246935	2.088865	12.808205	24
TAGTA	18198410	2.0759888	24.78093	29
AATTT	25090365	2.0705943	23.182108	24
ATTCG	3657590	2.0655978	30.456268	34
TACGG	2607530	2.035558	20.797728	5
AGGAG	9326685	2.0329456	8.995366	38
GCGTT	6149635	2.0263202	20.362198	16
GGAAG	9219955	2.0096817	12.493827	2
TTCGC	715290	1.9998255	5.4137096	33
ATCGT	3514175	1.9846051	17.068684	39
TTAGT	40341510	1.9424428	17.993198	28
GACGC	209735	1.9203472	15.135165	5
GTAGA	11782340	1.857918	12.51102	23
GCGTA	2362910	1.8445964	27.997032	4
ACGTC	276405	1.8308439	6.4490085	41
TAGTT	38010625	1.8302107	8.259177	25
AAACG	576630	1.8278431	7.2878613	7
TATCG	3228510	1.8232781	17.511547	38
GGACG	1681990	1.8150185	16.448044	2
GAGAT	11455235	1.806338	11.5283575	26
AGGTA	11416355	1.800207	26.972944	47
TCGTC	643570	1.7993089	8.399339	40
GAGCG	1657435	1.7885214	11.56654	28
GTCGC	460935	1.7813656	9.432573	3
TATTT	50130630	1.7462074	8.805108	32
TGGGA	18943570	1.7428693	22.64507	37
TAATT	20781285	1.7149858	22.960457	23
GGAGA	7737555	1.6865616	10.901473	2
AGTCG	2155390	1.6825967	16.427197	22
GTACG	2154035	1.6815389	20.581495	4
AGCGC	183280	1.6781236	12.971153	35
ACGGT	2139685	1.6703367	19.982534	6
CGTGG	3637830	1.6569315	31.917583	5
GGAAA	4388040	1.6393099	11.692439	2
AACGG	884800	1.63642	10.620349	29
TACGC	246940	1.6356745	7.324906	13
AGTTT	33833675	1.6290905	8.444991	26
CTCCA	28960	1.6276283	15.448855	23
ACGGC	177565	1.6257967	9.544445	12
CGAGC	177455	1.6247895	5.113182	7
GTCGT	4926940	1.6234392	10.976202	3
GGAAT	10254145	1.616942	11.695884	2
AGTTA	14165120	1.6158899	15.97067	30
CGATT	2840290	1.6040334	19.011236	11
GGTTT	56951630	1.599964	9.196049	2
AGCGT	2027395	1.582678	8.907336	29
GCGTG	3456320	1.5742587	32.21913	4
ACGAG	843835	1.560656	5.6495285	32
AGGTT	23384365	1.5564139	15.929324	41
TCGAA	1157580	1.5488068	5.137875	32

TAGGA	9688065	1.5276787	6.795443	37
GAACG	816135	1.5094256	10.341394	28
GGTTA	22627215	1.5060196	21.626373	2
GGGAA	6834685	1.4897622	13.894486	2
GTTTA	30730810	1.4796877	11.809432	12
TGGCG	3246975	1.4789078	27.48837	10
CGAAA	466105	1.4774932	5.714996	15
TTAAG	12831315	1.4637359	10.697331	6
GTTAA	12821245	1.462587	26.33238	3
TTATT	41787450	1.4555881	5.7600517	32
TTCGG	4392810	1.4474418	18.122112	35
GGCGA	1336970	1.4427106	9.428057	2
TATAG	12549310	1.4315661	16.264229	47
TTTAA	17339325	1.4309363	8.516948	5
TGGAA	9055260	1.4278939	6.442443	1
GCGAT	1815050	1.4169118	23.783756	10
GTAGT	21270240	1.4157021	9.754249	36
TTATA	17150165	1.4153259	12.18746	46
TCGGA	1806595	1.4103113	7.996913	46
GTAAT	12342660	1.4079925	29.85223	22
AGATA	5176995	1.3991516	6.4704566	26
AAGTA	5146130	1.3908098	10.557246	34
TTGTA	28712010	1.3824825	17.305693	20
CGTAC	208320	1.3798643	6.7740207	13
GGTAG	14896605	1.3705355	7.655091	2
GCGTC	354065	1.3683474	8.267867	4
GGGTT	35028475	1.3602812	15.902039	2
TCGTG	4104340	1.3523903	9.303892	40
CGAAG	731035	1.3520347	5.3750315	45
TTTTG	65828780	1.3378776	5.7618227	34
AGAGC	721240	1.3339192	7.0462084	47
GACGT	1705265	1.3312085	5.856945	3
ACTCC	23540	1.3230101	15.099462	22
AGTAT	11569910	1.3198408	23.787512	30
GGAGT	14258265	1.3118061	10.558937	2
ATTAT	15746240	1.2994663	11.996807	45
ATCGG	1639450	1.2798302	7.6377997	45
TTGAG	19201280	1.2779965	13.090177	44
GAGTA	7992505	1.2603115	15.639484	34
AAGGC	680855	1.2592278	25.539444	46
TAAGC	933855	1.2494696	37.116623	7
GTATT	25946290	1.2493132	10.558492	31
TGTAA	10939310	1.2479048	29.37295	21
AACGC	79260	1.2438126	6.5330715	11
TCGGG	2724215	1.240805	23.738457	36
GGGAT	13389005	1.2318314	12.433363	42
GCGAC	131800	1.2067695	9.503605	23
GATTA	10500225	1.197816	15.891918	44
GGGGA	9297940	1.1824781	11.273835	2
TTTGT	57416880	1.1669176	8.342804	19
GGGTA	12682315	1.1668136	18.570986	2
TCGAT	2040695	1.1524678	6.600934	11
TTAAT	13929865	1.149569	17.587044	4
GGTGG	21294045	1.1430602	15.021782	8
CGTGT	3466515	1.1422254	8.849491	41
GGGGT	21262940	1.1413903	9.11001	2
GGATT	16895185	1.1245078	9.366924	43
CGTAA	840070	1.123988	7.169692	21
TGAGG	12136995	1.1166425	16.691162	45
TAGGC	1425165	1.1125494	6.7171373	13
AAAAC	199850	1.0857702	10.553279	6
TTTTC	6210655	1.0709994	9.828514	29
AGTAA	3940475	1.0649656	7.3298745	9
AAGGT	6622120	1.0442201	5.7946076	46
TGCGG	2283990	1.0402946	9.966045	5
TTGGG	26727370	1.0379196	10.111684	36
CGTGA	1308765	1.0216823	8.230763	26
GTTAT	20873230	1.0050454	6.9363236	31
TAAGG	6299635	0.99336857	8.103544	45
TTATC	2423935	0.9903034	12.422659	37
ATTTC	2418215	0.9879665	5.3923717	22
GGAGC	914495	0.98682237	10.589731	27
GGTAC	1261130	0.9844962	20.721554	3
TATTC	2391600	0.97709286	20.588226	33
TAAGT	8459655	0.96503764	6.8254204	7
TTTGG	34019385	0.95571965	7.7938547	35
TAGAC	711705	0.95223963	10.106637	25
TGGAG	10304410	0.94803876	6.5337605	1
ATTAC	975540	0.9442519	6.29018	29
AGTTG	14092670	0.93797827	9.742225	38
TGTAG	14092460	0.9379643	6.752636	21
GGTTG	24059005	0.93429744	7.500019	42
GGATA	5918425	0.93325675	8.069228	2
GGAAC	497065	0.9193118	9.407192	27
GTGGC	2005160	0.91329527	25.96287	9
GTTTG	31908995	0.8964315	8.746786	18
AGGTG	9697765	0.8922255	6.259044	47
GTTGA	13352745	0.88873047	12.713243	43
TGGTT	31587690	0.8874051	6.9582033	10
GTGGT	22802445	0.8855006	11.655322	9
AGTGA	5550180	0.8751894	6.928318	18
GGTAT	12771455	0.85004103	6.7475834	3
TGGGG	15817445	0.84907734	6.399043	1
GTGCG	1857840	0.84619504	9.53855	4
GGGGG	10951605	0.8126286	6.647967	2
GGGTG	14959045	0.8029986	9.522013	2
GAAGC	431695	0.7984114	6.620665	4
CGATC	118145	0.78256565	7.0029144	44
GGGGC	1214650	0.7647452	5.0046573	2
GTTGG	19656110	0.7633172	5.8018007	39
GGTAA	4753085	0.7494982	5.858781	2
TAGTG	11222465	0.74694353	5.339448	7
AGTGG	8117875	0.7468705	7.316715	8
GAGTC	938830	0.73289394	14.988661	21
CACGT	105560	0.69920546	5.9991794	47
GGGAC	639835	0.6904395	5.29598	2
TGGGT	17414390	0.67626315	6.7956624	1

TGGTG	15873800	0.6164366	6.459709	7
GATTC	905335	0.51128143	5.6063633	29
GGTGC	1047420	0.47707102	9.626688	3

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	303714	0.3798054127842461	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	230264	0.2879535140604373	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	140587	0.17580916114205736	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	135362	0.16927510844182728	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	98025	0.12258383080192461	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATTTTGTTAGTTA	91148	0.11398389196566003	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82714	0.10343686795154697	No Hit