FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD15_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`

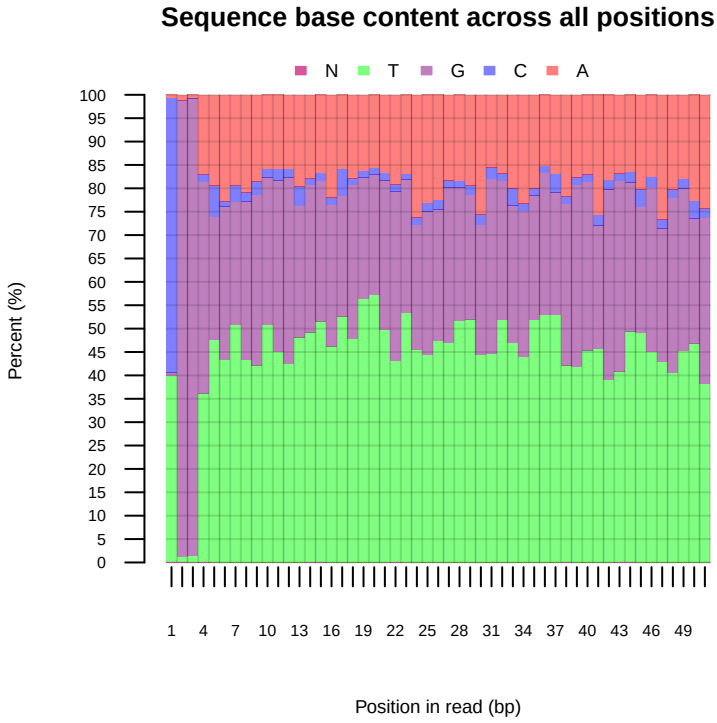
1 Sequence Duplication

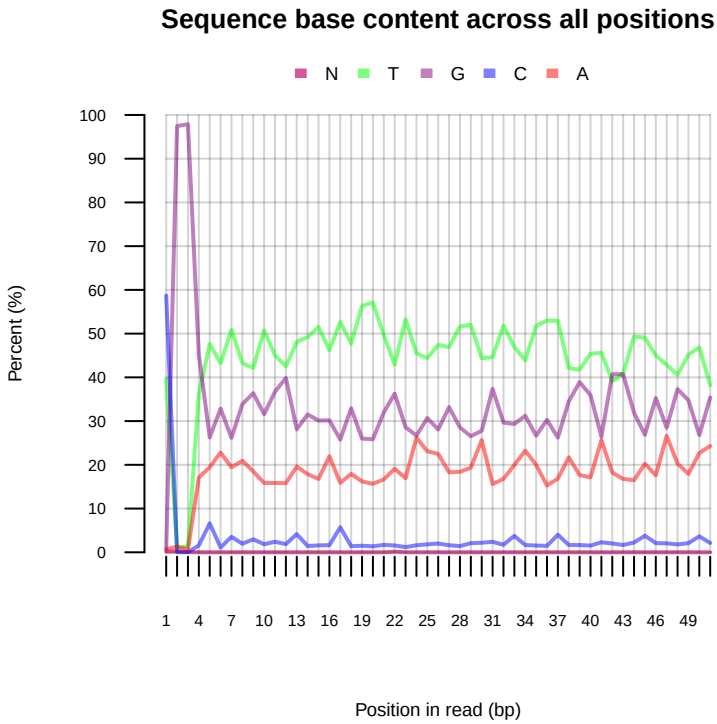
- Estimated Duplication rate 81.8470%

2 Per base sequence quality

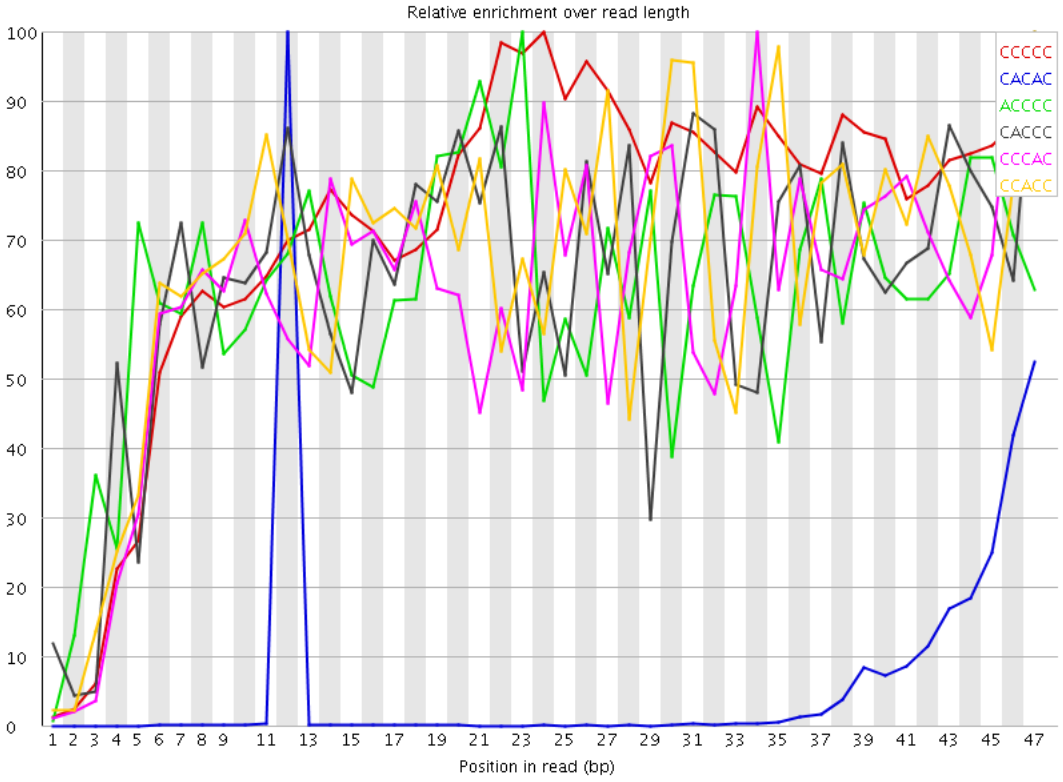
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	185955	1311.3949	1804.3018	24
CACAC	471885	106.40592	1626.0151	12
ACCCC	51555	65.01273	103.98996	23
CACCC	49335	62.213234	97.17752	47
CCCAC	46715	58.909317	94.50916	34
CCACC	46580	58.73908	88.881874	47
CCCCA	46445	58.568836	89.17637	24
CGGGC	5362805	33.976204	1203.7681	1
GCCCC	22240	15.133735	25.100077	45
CCCCG	21705	14.769683	26.379066	46
CCCGC	21600	14.698234	27.817938	47
CGCCC	21415	14.572344	23.821056	44
CCGCC	21230	14.446458	24.77985	35
ACTCC	155125	14.178423	631.1061	23
AGCAC	651405	14.173152	162.41649	10
CTCCA	152770	13.963176	628.6567	24
CGCGG	2095395	13.275433	336.1995	5
GCGCG	1980990	12.550618	334.73486	4
GCACA	557160	12.122585	162.01762	11
CGGCG	1633175	10.347027	294.81403	1
GGCGC	1542500	9.772553	335.9894	3
CGGAA	4487765	9.42174	192.75938	1
CGCGC	134510	8.831849	73.211685	13
ACACG	383615	8.346626	157.53952	13
CGGGA	6980750	7.908369	223.27956	1
TCGCG	1297850	6.176467	24.531702	30
CGGGT	13446835	6.1747794	240.17278	1
AGACG	2783205	5.8431396	67.3354	27
CGGAG	4746875	5.3776507	158.72676	1
AGATC	3305095	5.21216	27.963636	43
ACGCG	442210	5.191916	18.21361	14
CGCGT	1069805	5.0912004	22.41518	31
TCCCC	9935	5.078226	18.503628	3
CGGAC	425870	5.00007	169.04222	1
GGGCG	7986730	4.8824472	121.950325	2
CGGTT	13984810	4.8238153	166.58966	1
CGTCG	997980	4.749386	26.995575	41
CGGTC	994790	4.734205	182.02435	1
AACCC	20525	4.628207	9.124184	22
CGGGG	7447860	4.5530252	113.81266	1
CGCGA	384960	4.5197525	20.508585	5
AGGCG	3978750	4.507456	66.440895	47
TCGAG	5245185	4.4635296	56.8229	44

CTCCC	8535	4.3626227	12.008805	24
GAGAC	2004490	4.2082834	64.50718	26
CCTCC	8230	4.2067237	7.205283	24
CCCCT	8205	4.193945	7.325514	46
CCCTC	8115	4.147942	8.286241	47
ACACC	18390	4.1467834	6.1453023	37
ACGTC	462010	4.0745835	64.54661	15
AAAAA	3098290	3.9944563	11.371261	31
CAACC	17615	3.9720278	7.2048373	31
ACCAC	17400	3.923547	7.628799	45
CGACG	327420	3.8441849	31.591772	24
ACCCA	16790	3.7859974	7.0988836	13
CCCAA	16580	3.7386444	6.9929304	14
TTACG	5812980	3.715773	47.724583	14
CACCA	16250	3.6642323	6.1453023	36
CCAAC	16170	3.646193	6.410186	30
CGGTA	4281450	3.6434138	128.39096	1
CCACA	16085	3.627026	5.774465	15
GGCGT	7687435	3.530066	59.06583	3
CGTTT	13491405	3.4956188	43.95282	17
CGAGG	3078595	3.487686	71.60762	45
GGCGG	5592055	3.4185352	46.81687	11
TACGT	5339990	3.4134285	49.844643	15
ACGTT	5265805	3.366008	51.81708	16
GACGG	2902225	3.2878797	36.024033	28
ACGGG	2821640	3.1965864	36.158173	29
GATCG	3748805	3.1901453	16.409369	44
CGGAT	3677535	3.1294959	107.67919	1
AGAGA	7740140	2.9057074	22.558401	25
CACGT	322640	2.8454444	64.01826	14
AAGCG	1354035	2.8426998	54.977924	8
CACAT	173435	2.8345551	116.49823	31
TTTCG	10843245	2.8094816	16.479942	30
AGCGA	1313390	2.7573683	55.68473	9
ATCGC	310805	2.7410684	35.19688	29
GTCGA	3200675	2.7236996	56.215088	43
CGAGA	1280295	2.6878872	33.849457	25
TTCGA	4162420	2.6607022	37.07547	31
AACTC	161305	2.6363072	116.23876	22
GCGGG	4279700	2.6162658	47.713017	12
GAGGC	2296440	2.6015966	55.50872	46
GGAGG	23692135	2.589851	33.244015	39
TCACA	158205	2.5856414	116.7862	30
CGTTA	3995160	2.5537863	34.10787	9
GCGGC	401710	2.5450451	9.727135	9
AGAGC	1205265	2.5303676	16.535843	8
ATTCG	3865700	2.4710329	48.65719	34
TTTTT	174413775	2.4603617	5.5339985	16
TCGTT	9345315	2.4213684	6.4729595	4
CGTTC	671565	2.400692	34.08129	33
TCGGA	2817815	2.3978946	14.439671	46
CGGTG	5203795	2.3895798	46.835552	1
ATCGG	2802880	2.3851852	14.181154	45
TTCGC	653715	2.3368824	9.278712	33
GGGAG	21359485	2.3348627	28.904503	38
CGTAG	2727735	2.3212383	27.905525	5
TTTTA	65606605	2.2832277	13.418011	26
TTTAG	49020065	2.271134	17.065329	27
AGAAA	3250225	2.2611675	5.6961894	22
AGTAG	14829590	2.2565699	21.977272	35
GAGGT	27370095	2.2473962	25.311768	40
TTCGT	8615805	2.2323525	6.2200327	35
CGAGT	2611130	2.2220101	45.52016	33
GCGGA	1946015	2.2046063	23.147326	7
GGAAG	10649635	2.1573546	11.196021	2
GCGTT	6207975	2.1413321	30.210064	16
AGGAG	10375050	2.1017303	10.622076	38
TTTAC	4365755	2.0962472	34.9828	13
GGTCG	4556760	2.0924616	32.07927	42
ACGGA	959435	2.014265	9.5875635	30
GCGGT	4373775	2.008435	31.212519	6
ATTTT	57467780	1.9999819	8.377602	25
GAAGA	5218380	1.9590195	6.084722	46
TACGC	221165	1.9505104	12.724086	13
GACGC	164885	1.9358879	13.396487	3
TAGTT	41502550	1.9228423	10.488016	29
AAACG	493355	1.9194504	11.206354	7
GCGTA	2245810	1.9111317	27.585644	4
TAGAG	12521065	1.9052892	10.046841	24
GAGAT	12468060	1.8972234	8.748153	26
AGGTC	2204965	1.8763735	53.766094	41
AATTT	21827060	1.8740468	18.528906	24
ACGGC	159485	1.8724874	9.461257	12
AGGTA	12221270	1.8596702	31.250927	47
AGCGC	158380	1.8595138	8.2917185	10
ATCGT	2906605	1.8579601	15.777642	39
TCGTC	518880	1.854878	11.056128	40
TTAGT	39540325	1.831931	16.436258	28
GTCGC	381440	1.8152726	10.072637	3
CGAGC	154420	1.8130201	6.918026	13
TACGG	2103465	1.7899995	14.520759	5
TAGTA	15606645	1.7838646	15.823118	29
GGAAA	4695005	1.7625406	11.751439	2
AAGAG	4682955	1.758017	6.057736	47
TGGCG	3804265	1.7469164	40.566513	10
TAGCG	2052275	1.7464379	5.983252	10
TATCG	2711650	1.733341	16.230927	38
AGTTA	15161225	1.7329525	22.848269	30
GAAAA	2480545	1.7257044	5.2462106	3
AGCGG	1522085	1.7243433	6.289601	11
GAGCG	1516475	1.7179879	10.634633	28
GCGTG	3703025	1.700427	40.120872	4
CGTGG	3700555	1.6992928	39.8161	5
GGAGA	8355930	1.6927065	10.428513	2
GTAGA	11100365	1.6891057	9.746205	23
ATACG	1059460	1.6707768	5.097354	14
AGTTT	35929615	1.6646442	8.470262	26

GGACG	1463810	1.6583246	17.116598	2
GAGCA	788860	1.6561551	16.050003	9
TGGGA	20004760	1.6426184	15.585755	37
TCGTA	2545085	1.6268693	8.68045	45
TATTT	46622815	1.6225576	5.98052	32
AACGC	74205	1.614539	9.13985	11
AGTCG	1889915	1.6082734	13.84049	22
GCGAC	136115	1.598104	23.71388	23
CGATT	2473480	1.5810978	19.419825	11
AGCGT	1843605	1.5688647	8.553312	29
TAGGA	10306090	1.5682436	8.2459	37
TTCGG	4522510	1.5599607	26.662457	35
CGTAC	175245	1.5455301	10.695607	13
GCGTC	320950	1.5274007	11.810536	40
GTCGT	4417180	1.523629	10.97952	3
GGGAA	7493750	1.5180497	13.620046	2
TAATT	17601375	1.511234	18.287573	23
GTACG	1769340	1.5056669	14.216269	4
AGGTT	24329605	1.5006216	15.856067	41
AACGG	713915	1.4988136	8.697269	29
GTAGT	23967685	1.4782988	9.516543	36
TTGAG	23695625	1.4615185	14.96384	44
GGTTT	58243690	1.4561346	8.944149	2
ACGGT	1704740	1.4506938	13.786851	6
TTATT	41578810	1.4470172	7.8579555	32
ACGTA	905220	1.4275391	5.273713	26
AAGTA	5051485	1.4244753	11.132329	34
TATAG	12360675	1.412845	16.89367	47
GGCGA	1241860	1.4068813	9.9758835	2
GTCAC	157910	1.3926485	63.45674	29
TTTAA	16215190	1.3922178	8.1299305	5
CAGTC	157605	1.3899586	63.098297	27
GCGAT	1627000	1.3845388	23.53335	10
TCCAG	156735	1.3822857	62.60309	25
TTAAG	12056305	1.3780551	9.913877	6
TTATA	15988065	1.372717	13.074107	46
CCAGT	155310	1.3697186	62.619667	26
GTTTA	29414160	1.3627788	8.193655	4
TCGGG	2961450	1.3598962	33.843853	36
AGATA	4772030	1.3456713	5.459932	26
GGAAT	8841735	1.3454175	9.409933	2
GAACG	636910	1.337147	8.459036	28
TATTC	2767035	1.3286108	35.39477	33
TCGAC	149820	1.3213006	5.2110653	23
GGTAG	16077490	1.3201448	7.727558	2
TAAGC	835490	1.3175744	39.88897	7
GGTTA	21358420	1.3173623	17.330402	2
TTGTA	28215825	1.3072591	15.798197	20
GGAGT	15796245	1.2970514	10.23331	2
GACGT	1516455	1.2904677	6.3661675	3
GAGTA	8405900	1.2790979	15.844548	34
GTTAA	11155875	1.2751347	21.1217	3
TGGAA	8366200	1.273057	8.994443	1
ATTAT	14691580	1.2614024	12.920516	45
CGTAT	1971895	1.2604747	5.793562	13
GGGTT	37780830	1.2574517	14.0264015	2
AAAAC	171470	1.2362939	18.414486	6
GTAAT	10727005	1.2261139	22.898396	22
GTTCG	3543295	1.2221977	5.2503023	34
TGAGG	14628260	1.2011465	18.00817	45
GGGAT	14603025	1.1990744	11.647356	42
TTTGT	63534375	1.1931477	7.4495835	19
GGGGA	10900715	1.1915864	10.2314	2
GGTGG	26815940	1.1881742	12.335637	8
TCGTG	3428280	1.1825253	7.422397	40
GATTA	10340080	1.1818877	16.537113	44
GTGGC	2558740	1.1749719	38.40106	9
CGTAA	743095	1.1718665	9.258216	21
ATCTC	170305	1.1282156	50.083195	42
TAGGC	1321920	1.124923	10.496611	13
AGTAT	9696265	1.1082987	14.923404	30
GGATT	17862255	1.101723	9.086189	43
CGTGA	1288220	1.096245	9.856242	26
GTATT	23573340	1.0921694	6.5198817	31
TGTAA	9527265	1.0889819	22.36114	21
GGGTA	13186625	1.0827719	15.486755	2
TTTTC	5545680	1.0793306	11.67368	29
GGGGT	24265095	1.07515	8.467757	2
TCGAT	1667390	1.065829	7.7479877	11
TGGAG	12942285	1.0627087	10.146221	1
AGTAA	3764210	1.0614748	6.959203	9
GTTAT	22888730	1.060451	9.516953	31
TGTAG	16905090	1.0426862	8.983934	21
TTAAT	12115640	1.040235	14.349547	4
AAGGC	494880	1.0389652	16.988443	46
ATTTC	2107505	1.011933	6.5974565	22
GTTGA	16328695	1.0071348	14.195812	43
AGTTG	16179100	0.99790794	9.531803	38
CGTGT	2886430	0.99562347	7.072725	41
CGTGC	206040	0.98054415	5.7558537	13
GGTTG	29288580	0.9748059	7.733751	42
TTATC	2022365	0.9710524	11.92039	37
TAAGT	8442505	0.9649919	6.402314	7
CGAAC	43600	0.9486408	7.5040874	9
TAGAC	590765	0.9316411	11.064275	25
CGTCT	260235	0.93028086	25.847359	16
TGGGG	20911540	0.9265591	9.07441	1
GTGGT	27825360	0.9261059	8.494538	9
AAGAC	237900	0.92557544	9.8517	32
TTGGG	27800495	0.9252784	6.834751	36
GGATA	6033210	0.91805357	7.6458373	2
GGAGC	809315	0.91685873	9.772545	27
TGCGG	1996090	0.91660345	6.4962354	5
GTTTG	36347450	0.90871274	7.442841	18
TAAGG	5948325	0.905137	5.154704	45
ATTAC	758405	0.898396	5.3390183	29
TGGTT	35865540	0.89666456	7.6194434	1

TTTGG	35160315	0.8790335	5.517563	35
AGTGA	5747570	0.87458867	5.154754	18
GGGTG	19358585	0.8577499	8.415516	2
GGTAT	13602010	0.83895606	6.3396063	2
GGTAC	977305	0.83166367	14.34323	3
GAAGC	381565	0.8010684	9.77253	4
GGGGG	13298405	0.7844294	5.8892746	2
GGAAC	372030	0.7810504	7.497221	27
TGGTG	23165375	0.7710085	6.6400447	7
AGTGG	9366020	0.76905674	5.468985	8
GGTAA	5052820	0.7688709	6.282005	2
TGGGT	23001365	0.7655499	9.366919	1
GTGCG	1589190	0.72975516	6.0929646	4
GTTGG	21777980	0.72483224	5.250378	39
TGGTA	10838705	0.6685186	5.6049566	1
GAATC	419890	0.66216975	12.132738	40
GAGTC	768460	0.65394145	12.629109	21
CAGAA	166865	0.64920616	28.197725	38
TCTCG	179855	0.64294064	27.030174	43
CTCGT	178610	0.6384901	27.060593	44
TGGGC	1180655	0.5421561	6.459985	13
GATTC	828590	0.52965134	6.7449393	29
TGGAT	8325960	0.51353544	5.282361	1
TGAAC	291950	0.46040747	11.962411	20
GGTGC	830125	0.38119295	6.1696825	3
CTGAA	207730	0.32759187	11.725344	19
GAACT	179565	0.28317538	11.674888	21
ACATG	177630	0.28012392	11.553335	32
TCAGA	170605	0.2690454	11.4525585	37
AGTCA	170255	0.26849347	11.6037245	28
AATCT	178340	0.21125908	8.917446	41
GTCAG	171465	0.1459127	6.181768	36

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	280524	0.33794429412530674	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	238541	0.2873678183148137	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	166504	0.20058560675393222	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	127631	0.1537557150315375	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	109288	0.13165809704826156	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAGAATCTCGTATG	102695	0.12371557971937652	TruSeq Adapter, Index 1 (97CGGTTAAT
98606	0.11878960469164847	No Hit	
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	96634	0.11641395716054559	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	94251	0.11354318227889339	No Hit