FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD15_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`

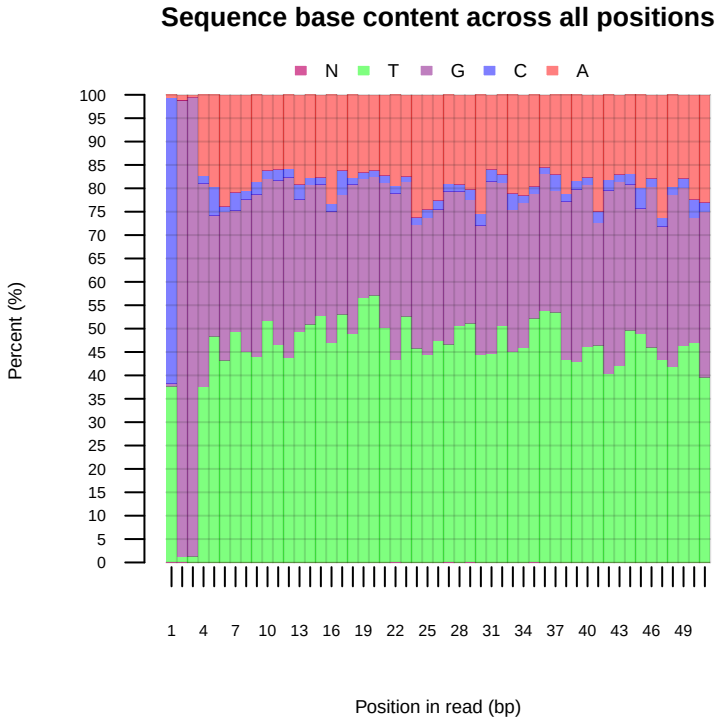
1 Sequence Duplication

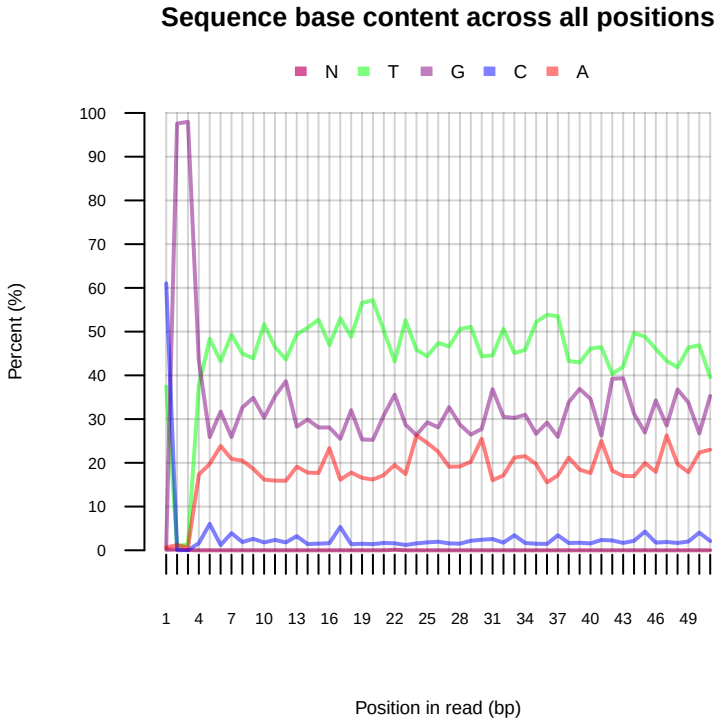
- Estimated Duplication rate 82.2141%

2 Per base sequence quality

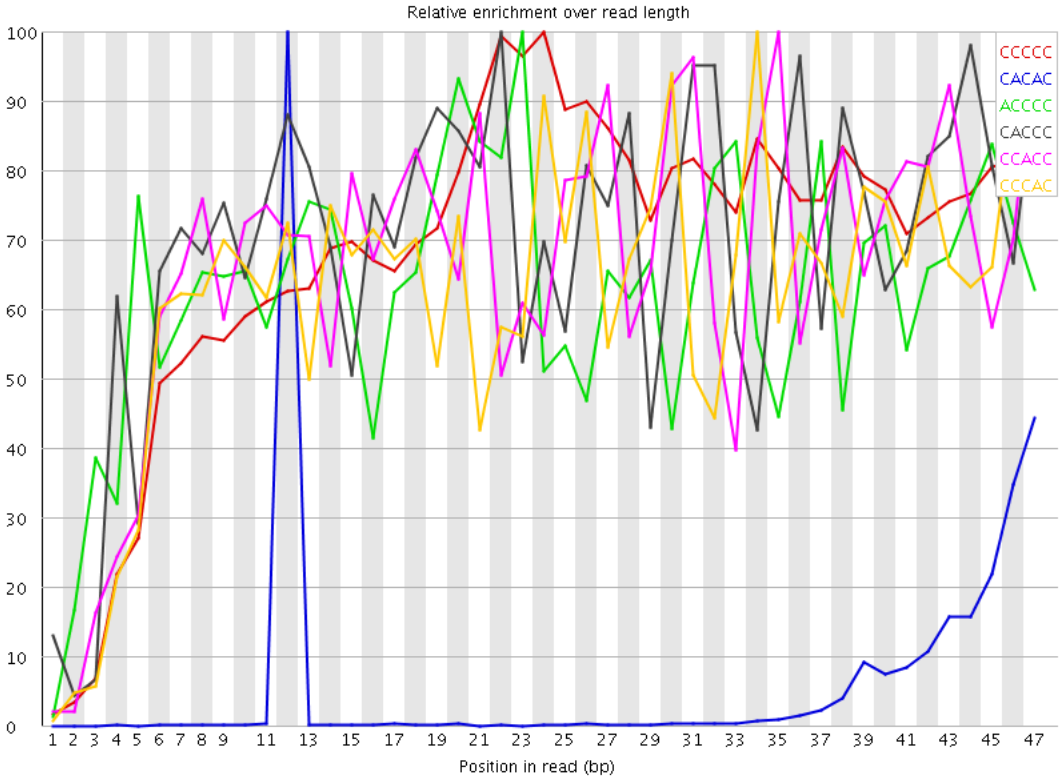
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	199290	1460.9021	2118.3499	24
CACAC	390925	90.28671	1468.1927	12
ACCCC	52120	67.81682	107.91094	23
CACCC	50095	65.18195	93.66718	22
CCACC	47830	62.23481	93.54301	35
CCCAC	47740	62.117706	99.35124	34
CCCCA	47325	61.57773	90.48624	23
CGGGC	4262645	30.020727	1134.7825	1
GCCCC	22890	16.557072	29.910006	45
CCCGC	22780	16.477505	34.667686	34
CGCCC	22485	16.264122	28.380468	44
CCGCC	22390	16.195406	28.381226	43
CCCCG	21995	15.909691	30.75973	46
CGCGG	2056575	14.483935	390.2362	5
GCGCG	1939975	13.662753	388.00745	4
ACTCC	137885	12.920561	567.9475	23
CTCCA	136205	12.763134	565.76685	24
AGCAC	513180	11.695055	150.67116	10
CGGCG	1548325	10.904461	318.71005	1
GGCGC	1545690	10.885902	389.3117	3
GCACA	448835	10.228673	150.35507	11
CGGAA	4206790	9.459878	230.22545	1
CGCGC	119750	8.547033	47.9582	13
CGGGA	6124415	7.656038	241.32825	1
ACACG	322100	7.3404603	145.73962	13
AGACG	3111205	6.9962187	85.6803	27
CGGGT	12155325	6.1650977	242.45316	1
TCGCG	1180585	6.06832	26.233	30
ACCAA	142125	5.826373	254.95006	36
CGGAG	4655185	5.8193755	174.28595	1
TCCCC	10545	5.5669007	18.604378	5
CGGTT	14706865	5.4440565	196.04097	1
ACGCG	417550	5.28988	16.757248	4
GAGAC	2343440	5.2697325	82.14377	26
CGGGG	7576595	5.265241	128.15022	1
CGCGT	985160	5.063816	24.299608	31
CTCCC	9500	5.015226	13.767239	24
AGGCG	3989230	4.986876	85.07439	47
CCTCC	9270	4.8938046	8.434001	23
AACCC	21170	4.8893514	8.534756	2
CGGTC	941590	4.839863	185.42995	1
CCCTC	8970	4.7354293	7.813839	24
TCGAG	5157935	4.705919	72.68184	44

CGGAC	370645	4.6956472	158.48198	1
CGTCG	906990	4.6620154	20.99983	41
GGGCG	6591635	4.580758	117.43437	2
CCCCT	8645	4.5638556	7.8139896	44
CGCGA	343145	4.347254	23.129765	5
ACACC	18790	4.339675	6.4656234	21
CAACC	18325	4.23228	6.8368793	31
ACCAC	17860	4.124885	6.2482915	20
CGAGG	3297595	4.122273	94.30889	45
AAAAA	3172070	4.096995	10.409064	31
ACCCA	17290	3.9932399	6.457068	8
CCCAA	17285	3.9920852	6.4570527	14
CCACA	17170	3.9655252	6.29427	35
GACGG	3170520	3.9634187	46.884575	28
CACCA	16915	3.9066312	6.619836	28
TTACG	5830440	3.8823893	60.095814	14
CCAAC	16805	3.8812258	7.325228	30
CGGTA	4177415	3.811327	135.06256	1
ACGGG	3037095	3.796626	47.128044	29
AGATC	2298365	3.772097	19.063461	43
ACGTC	398625	3.6857934	59.406384	15
CGACG	287490	3.6421688	22.629725	24
TACGT	5301065	3.5298877	61.018982	15
ACGTT	5267580	3.5075905	62.46764	16
CGTTT	12405520	3.3515573	39.62235	17
AGAGA	8185960	3.2673957	28.148197	25
GGCGG	4697310	3.2643254	30.569038	11
CGGAT	3533105	3.2234812	112.37029	1
GACCA	137870	3.141972	141.5634	35
GTCGA	3440320	3.1388273	72.07757	43
GGCGT	6017380	3.0519738	49.322536	3
GAGGC	2404165	3.0054102	69.732574	46
AAGCG	1226780	2.7586806	50.90207	8
CGAGA	1226385	2.7577922	32.47543	25
TTTCG	10058820	2.7175567	13.551217	30
CGGTG	5283140	2.6795723	50.86487	1
AGCGA	1186955	2.6691253	51.780125	9
ATCGC	287445	2.6577933	37.58545	29
GGAGG	21279050	2.6247857	31.979725	39
GATCG	2844915	2.595601	11.660066	44
CACGT	277920	2.5697227	59.29125	14
TTCGA	3849350	2.5632157	30.247526	31
GCGGC	360315	2.537607	9.062368	33
TTTTT	176281175	2.5032144	5.312041	16
ACGGA	1070210	2.4065988	14.41556	30
GGTCG	4738570	2.40337	42.15342	42
AACTC	144310	2.4002593	103.95306	22
GGGAG	19456895	2.4000216	29.016613	38
CGTTA	3598660	2.3962855	25.714514	9
CAATC	141610	2.3553512	107.22275	38
TCGTT	8698310	2.349993	6.1687455	4
GCGGG	3369205	2.3413787	31.10958	12
AGTAG	14312870	2.31789	18.7429	35
CGTTC	617380	2.3160768	28.691175	33
CCAAT	136630	2.2725205	103.69412	37
TTTTA	64658315	2.2629895	11.714473	26
CGTAG	2454245	2.239167	25.870026	5
AGGTC	2448740	2.2341444	68.84154	41
TTTAG	46361745	2.2232525	15.163647	27
GCGGA	1774105	2.2177815	21.041824	7
TTCGT	8103445	2.1892798	5.7477913	35
GCGGT	4300460	2.1811638	36.78547	6
AGAGC	959570	2.157801	15.66189	8
GAGGT	23895255	2.1512084	23.751842	40
TTCGC	571455	2.1437905	5.635467	33
GGAAG	9648235	2.1408427	11.616772	2
GACGC	168710	2.1373622	16.819994	3
TTTAC	4368340	2.1229675	42.94934	13
CGAGT	2288820	2.0882392	37.60681	33
ATTTT	59085490	2.0679452	8.703331	25
TAGAG	12760955	2.0665658	12.389775	24
AGGAG	9266280	2.0560908	8.056544	38
ATTCG	3062695	2.0393958	30.943695	34
ATCGT	2979665	1.9841075	18.733728	39
TAGTA	16532760	1.9540719	19.665665	29
AATTT	22646645	1.9535664	19.168633	24
AGCGC	153310	1.9422619	11.066331	35
GCGTT	5217100	1.9312197	19.438044	16
TCGGA	2113390	1.928183	10.02348	46
TACGG	2112810	1.9276538	17.757483	5
AAACG	474220	1.9182705	8.131278	7
ACGGC	150630	1.9083095	11.459218	12
GTCGC	365565	1.879039	11.083487	3
TATCG	2794590	1.8608692	19.218489	38
TACGC	199665	1.846156	8.737063	13
TCGTC	490885	1.8415357	9.199979	40
GTAGA	11328140	1.8345293	12.077506	23
ATCGG	1999395	1.8241779	9.637211	45
GAGAT	11222750	1.8174618	9.836882	26
TTAGT	37686590	1.8072402	14.524357	28
GCGTA	1977665	1.8043522	25.530922	4
CGAGC	141170	1.788462	5.0242095	7
GGAAA	4471825	1.7849123	12.3933115	2
GAAAA	2457245	1.7643138	5.2797785	3
GAGCG	1406840	1.7586694	11.605062	28
AACGG	777055	1.7473766	13.232139	29
TAGTT	36082635	1.7303234	7.00785	25
GGACG	1380210	1.7253795	16.343811	2
GGAGA	7743200	1.7181351	10.855609	2
AGGTA	10347805	1.6757694	23.135912	47
TGGGA	18452210	1.6611894	17.369772	37
TATTT	46177525	1.6161767	6.969432	32
TCGTA	2423785	1.6139566	5.9428983	43
GGAAT	9938340	1.6094587	11.396478	2
GAACG	715255	1.6084058	12.811074	28
TAATT	18485960	1.5946534	18.920708	23
AGTCG	1745055	1.5921272	13.271775	22

AGCGT	1732285	1.580476	9.028036	29
GTACG	1731645	1.5798922	17.482899	4
CGATT	2365555	1.5751822	17.580248	11
AGTTA	13318885	1.5742115	14.4330015	30
ACGGT	1725150	1.5739664	17.123253	6
AGTTT	32754310	1.5707155	7.1941524	26
CGTGG	3081620	1.5629766	30.746609	5
ACGAG	693870	1.560317	5.3671365	32
AACGC	68275	1.555945	7.9669666	11
GGGAA	6979395	1.5486549	13.72357	2
GTCGT	4157390	1.5389457	10.892855	3
TAGGA	9369300	1.517306	6.079642	37
GGTTT	56770810	1.513418	9.451629	2
GCGTG	2968515	1.5056105	30.843843	4
TGGAA	9214075	1.4921682	8.884719	1
GCGTC	285065	1.4652613	8.128451	4
AGGTT	22245315	1.4616336	13.077449	41
CGTAC	157835	1.4593847	8.254808	13
GGTTA	21744825	1.4287486	20.484934	2
CACGC	11100	1.425142	5.580381	15
TGGCG	2807375	1.4238814	26.037586	10
GGCGA	1135365	1.4193025	10.068866	2
AGATA	4843295	1.4109193	6.493779	26
GTTAA	11925110	1.4094757	25.643045	3
GCGAC	110660	1.4019353	15.3881235	23
CGAAG	619415	1.3928888	5.113654	45
TTAAG	11715385	1.3846872	8.969265	6
TTATT	39374840	1.3780881	5.154038	32
GAGCA	611670	1.3754725	15.22022	9
GTTTA	28668470	1.3747813	9.164987	12
GCGAT	1505975	1.3739988	21.981323	10
AAGTA	4712195	1.3727281	9.574389	34
TTTAA	15851205	1.3673717	7.2097683	5
ATGCC	147345	1.3623914	62.322975	47
GTAGT	20728755	1.3619876	8.157724	36
TTCGG	3656310	1.3534602	17.213835	35
TATAG	11428475	1.3507762	14.047626	47
GGTAG	14824820	1.3346279	8.005111	2
ACGCC	10285	1.3205032	6.1836653	16
GGAGT	14655975	1.3194273	10.556137	2
TTATA	15271110	1.317331	10.590139	46
TCGTG	3515675	1.3014013	9.68816	40
GGGTT	35628900	1.3013889	13.823027	2
CAGTC	140150	1.2958643	58.572216	27
TTGAG	19690755	1.2937855	11.142147	44
TCCAG	139730	1.2919809	57.7728	25
TTGTA	26940165	1.2919011	13.969062	20
GACGT	1413635	1.2897509	5.816371	3
GTCAC	139105	1.2862021	58.39408	29
CCAGT	137590	1.2721938	57.72718	26
CTGAC	137150	1.2681257	57.805386	33
CACTG	137080	1.2674785	57.924866	31
GTAAT	10691480	1.2636681	24.174574	22
TGACC	136560	1.2626703	57.568604	34
CCAGC	9730	1.249246	5.2485743	28
CGTAT	1867500	1.243536	5.4518137	44
TAAGC	753695	1.2369709	35.93744	7
GGGGA	9943255	1.2265075	10.866858	2
GAGTA	7569060	1.2257671	13.360853	34
ATTAT	14098855	1.2162088	10.429319	45
AAGGC	538040	1.2098994	23.217226	46
CCCAG	9390	1.2055931	5.3692317	27
AGTAT	10148800	1.1995267	18.765303	30
GGGAT	13231100	1.1911508	10.186859	42
GGTGG	23524345	1.177316	11.620127	8
TTTGT	60268370	1.1726067	6.664709	19
GTATT	24035505	1.1526097	8.107779	31
GGGGT	22861490	1.1441423	8.911191	2
TGTAA	9607885	1.1355938	23.81853	21
CGTAA	689620	1.1318105	6.7451224	21
TCGGG	2230245	1.131165	22.315205	36
AAAAC	154590	1.124881	12.416499	6
TTAAT	13007670	1.12208	17.198353	4
TCGAT	1683605	1.1210835	6.6499095	11
GATTA	9458630	1.1179527	13.680131	44
CGTGT	2987170	1.105764	9.249411	41
TGAGG	12257060	1.1034613	13.759042	45
GGGTA	12164395	1.0951189	15.574681	2
GGATT	16489330	1.0834352	7.753992	43
TAGGC	1176580	1.0734704	6.7713485	13
TGGAG	11805775	1.0628335	9.674549	1
AGTAA	3624805	1.0559564	6.378882	9
TTATC	2121090	1.0308275	13.749132	37
CGTGA	1126340	1.0276331	8.504377	26
GGAAC	456605	1.0267752	11.893921	27
ATCTC	151830	1.0245973	45.522724	40
TTTTC	5142175	1.0139292	9.181524	29
CGAAC	42570	0.97014403	5.26313	9
TATTC	1994545	0.969328	20.928326	33
TAGAC	587855	0.96479285	10.396545	25
TAAGG	5941525	0.9621969	6.3915358	45
GTTAT	20039395	0.96097857	6.068016	31
TGCGG	1893230	0.9602333	8.145776	5
ATTTC	1974265	0.9594721	5.532654	22
GGAGC	764495	0.9556834	10.68139	27
TGTAG	14520855	0.95409614	5.9210024	21
ATTAC	795620	0.9530127	5.755817	29
TGGGG	19018670	0.9518218	9.039521	1
TTGGG	26044770	0.9513168	7.576535	36
AGTTG	14366040	0.94392395	8.129521	38
GAATA	3234300	0.94219667	5.270253	3
GGTTG	25746330	0.94041586	6.0570993	42
TAAGT	7941260	0.93860877	5.7694993	7
GGATA	5754860	0.93196744	7.763934	2
TCACT	138100	0.9319429	42.415546	30
TGGTT	34748165	0.92633	7.42457	1
AAGAC	228695	0.92509556	6.452925	32

TTTGG	33971080	0.90561396	5.944239	35
GGTAC	990990	0.9041445	17.596666	3
GTGGT	24737405	0.90356374	8.865314	9
GTTGA	13687155	0.8993177	10.660719	43
GTTTG	33166990	0.8841784	6.8056836	18
CGTCT	231355	0.86791915	23.786362	16
AGTGA	5345615	0.8656926	5.2289786	18
GTGGC	1701895	0.86318946	24.48612	9
GGGTG	17154140	0.8585082	8.575279	2
CGATC	88560	0.81884944	6.5149083	44
GGGGG	11913010	0.81689805	6.3458853	2
GGTAT	12341745	0.8109172	6.3382134	2
GTGCG	1536595	0.77935046	7.6167054	4
GGTAA	4794165	0.77638835	6.5753093	2
AGTGG	8592180	0.77352464	5.682125	8
GAAGC	342560	0.7703203	6.089341	4
TGGGT	20877500	0.76257604	9.114671	1
TGGTG	19688725	0.7191546	5.951725	1
GAGTC	714040	0.65146506	11.971941	21
TGGTA	9707330	0.63782233	5.372553	1
CGGCC	8765	0.6255929	6.2717485	1
TCTCG	160170	0.60087144	25.31541	41
CTCGT	159145	0.59702617	25.31632	42
TGGAT	7754430	0.5095066	5.0218163	1
GATTC	742975	0.49473426	5.6732306	29
TGAAC	272980	0.4480172	11.040456	20
GGTGC	835410	0.42371416	7.6953745	3
CTGAA	185710	0.30478892	10.746636	19
GAACT	161495	0.26504704	10.74432	21
AGTCA	152255	0.2498823	10.698461	28
ACTGA	144785	0.23762242	10.55502	32
AATCT	160675	0.19246037	7.9975243	39

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	244558	0.31122381966962076	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	193805	0.24663569529956436	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	128091	0.16300824460987332	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	121522	0.15464855377412173	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATTTTGTTAGTTA	87454	0.11129371325160 91	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC	87447	0.11128480507139135	TruSeq Adapter, Index 4 (100CGGGATG
85679	0.10903485326782783	No Hit	