FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_YD21_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`

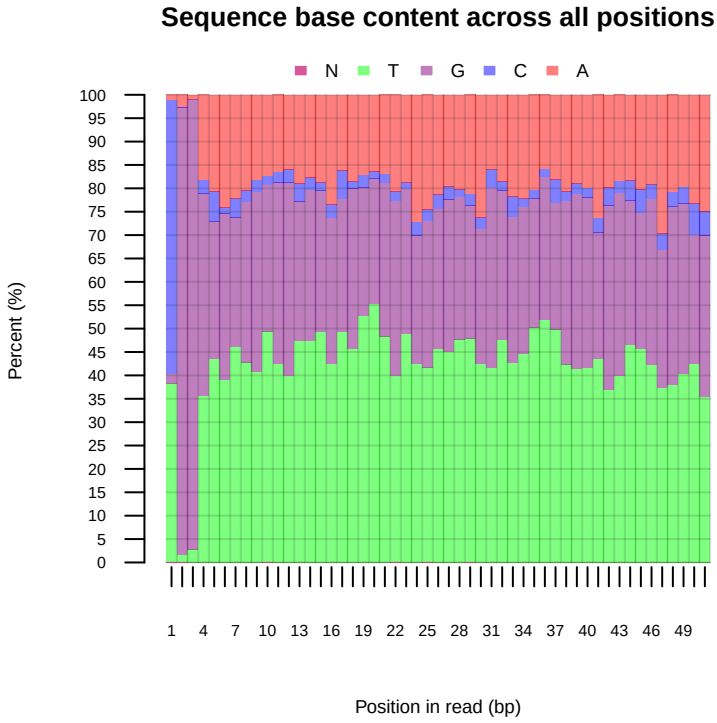
1 Sequence Duplication

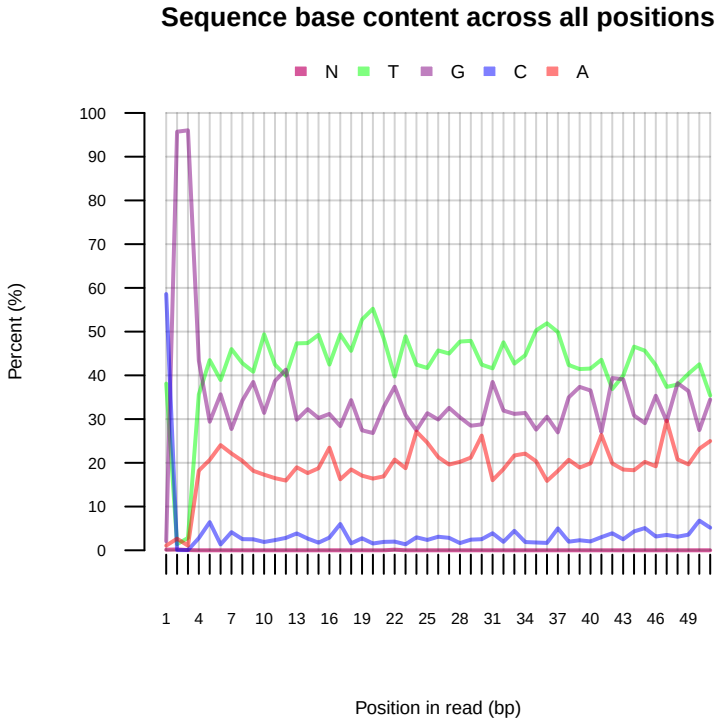
- Estimated Duplication rate 81.5438%

2 Per base sequence quality

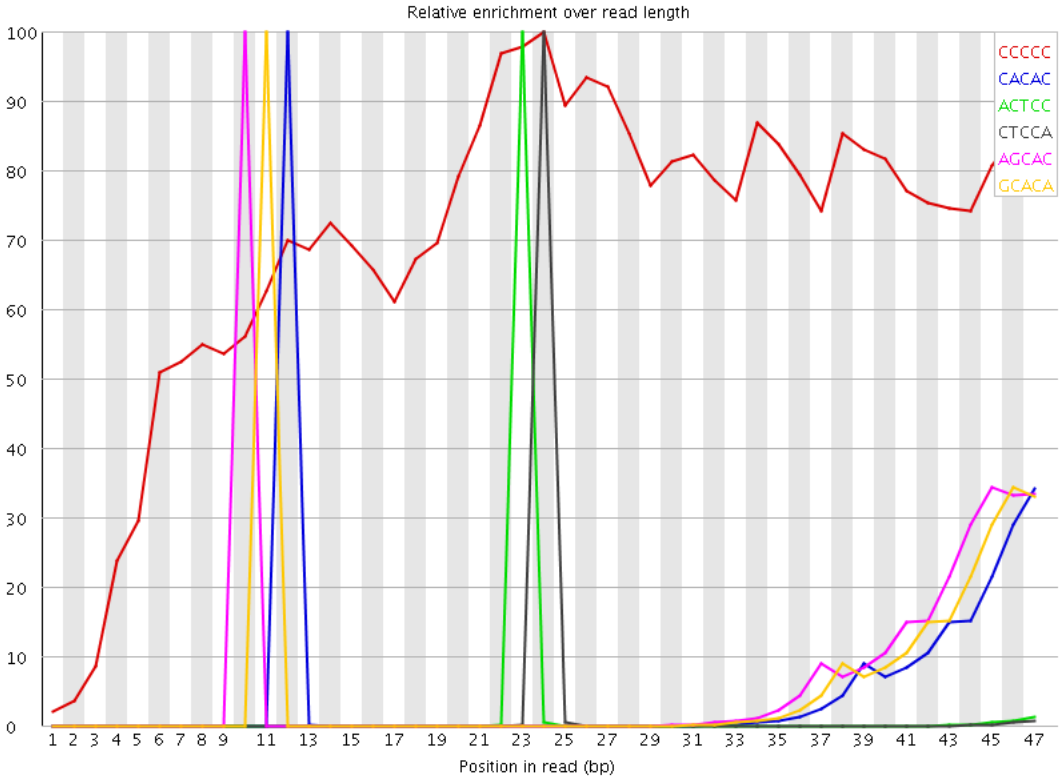
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	172490	506.28284	719.9246	24
CACAC	2329115	295.97144	5281.129	12
ACTCC	890335	50.99221	2280.012	23
CTCCA	876370	50.19239	2272.8496	24
AGCAC	3032550	44.460556	634.6889	10
GCACA	2716785	39.831093	633.59296	11
ACACG	2042700	29.948254	615.9461	13
ACCCC	46800	28.581823	51.367104	23
CACCC	44950	27.451986	46.34613	47
CCCAC	43040	26.285503	42.902447	47
CCCCA	42370	25.876318	45.4842	24
CCACC	41425	25.299185	37.879585	30
CGGGC	4525345	20.39883	733.981	1
CGGAA	8978300	15.1869	160.16045	1
CACGT	1785825	11.800418	277.8299	14
ACGTC	1743530	11.52094	276.4824	15
AACTC	937740	11.175001	488.95114	22
AGATC	7977975	10.968972	50.90485	43
CACTA	892400	10.634687	491.15253	31
CGCGG	2138535	9.639842	229.22949	5
GCGCG	2047550	9.229711	227.57951	4
CGCGC	214815	8.392862	36.18423	27
CGGCG	1736315	7.8267612	197.03969	1
AGAGC	4200865	7.105813	75.386925	8
GCCCC	20895	7.0758686	14.16199	47
CCCGC	20580	6.9691973	14.002867	47
GATCG	9090030	6.9299784	32.83137	1
CGCCC	20215	6.845594	12.172677	36
GGCGC	1517420	6.8400517	228.39792	3
CCCCG	19835	6.716911	13.684613	46
CCGCC	19625	6.645797	13.206956	35
ATGCC	952655	6.2949767	293.1699	47
CGGGA	6371125	5.9756403	179.85951	1
CAGTC	902380	5.9627686	274.16925	27
TCCAG	899365	5.942846	271.35776	25
ATCGG	7780190	5.931394	32.834827	2
GTCAC	896470	5.923716	273.39923	29
GAGCA	3499035	5.918659	75.04476	9
CCAGT	893615	5.9048505	271.26773	26
CTAGC	886535	5.8580675	270.24622	33
TCGGA	7502475	5.719672	33.844505	3
TCGCG	1444520	5.292677	24.225014	30
ATCTC	967195	5.194823	238.59645	40

TCACT	895675	4.8106875	222.19603	30
AGACG	2804565	4.7439547	59.202072	27
CGGAG	4968200	4.6598015	133.40083	1
CGGGT	10257515	4.3361278	162.06073	1
CGTCT	1454380	4.331396	124.45782	16
CGCGT	1166980	4.2757783	23.056026	31
CGGTT	12335160	4.238409	148.29936	1
CGTCG	1144305	4.192698	10.924043	41
ACGCG	465260	3.7822998	10.395751	4
AGGCG	4020410	3.7708452	56.897938	47
CGGGG	7238400	3.7644734	90.78457	1
TCGAG	4863585	3.7078578	54.94449	44
GGGCG	6981715	3.6309793	89.314575	2
GAGAC	2120635	3.5870793	57.044174	26
CGGTC	907655	3.3256197	110.82395	1
CGCGA	403200	3.2777872	14.82297	5
CGGAC	392940	3.1943793	97.604004	1
CGTTT	11225600	3.1352043	41.21908	17
TTACG	5038555	3.1222725	51.28237	14
GAAGA	8749880	3.0795832	17.005993	6
AAAAA	2688890	3.0780537	7.542778	31
CGAGG	3194805	2.996489	63.859776	45
CGACG	366210	2.9770796	15.078898	24
CGGTA	3875600	2.9546463	101.886246	1
ACGTT	4710295	2.9188576	53.578133	16
TACGT	4689640	2.906058	52.207287	15
TCTCG	973895	2.9004285	132.14502	41
CTCGT	973410	2.898984	132.20113	42
GTCGA	3720005	2.8360252	54.705456	43
GACGG	3016505	2.8292568	32.41971	28
GGCGT	6615820	2.7966852	44.94752	3
AAGAG	7867680	2.7690866	16.84401	7
GGAAG	14084500	2.7486823	10.185102	2
GGCGG	5193335	2.7008967	28.054754	11
AGAGA	7601155	2.675281	22.540243	25
GAGAT	16804035	2.6655996	11.367775	26
ACGGG	2830775	2.655056	32.71146	29
TTTTT	123083885	2.6203666	5.509876	8
CTCCC	9375	2.5805182	23.862684	24
AACCC	20200	2.5669074	6.8757186	22
TCCCC	9240	2.5433586	9.187821	3
CGAGA	1469745	2.486091	40.65602	25
GGAGG	22404175	2.424406	27.387459	39
TTTTA	51032400	2.4105453	15.051932	26
CGGAT	3150905	2.4021597	79.491	1
ACACC	18840	2.394086	7.493588	13
GCGGC	522865	2.3569107	8.677763	9
ATTTT	48140825	2.2739596	10.535844	25
TTTCG	8122105	2.2684276	8.789472	30
CCTCC	8195	2.2557168	8.730249	24
TTTAG	38519740	2.2384863	17.858627	27
CGGTG	5287330	2.2350967	41.226933	1
GAGGC	2368165	2.2211623	46.244114	46
ATCGC	336075	2.2207246	37.60328	29
GGGAG	20261250	2.1925156	25.127047	38
GAGGT	24263300	2.1341481	22.524536	40
TGAGA	13282390	2.1069665	7.163832	41
AATTT	19687255	2.0633025	21.289244	24
TCGTA	3298315	2.0438871	28.630775	43
AAGCG	1196175	2.023344	27.718596	8
GGTCG	4720140	1.9953302	30.682375	42
CGTTA	3166645	1.9622942	14.807197	33
GCGGG	3771495	1.9614407	28.985527	12
TTCGC	654845	1.9502422	7.468519	33
GCGGA	2078545	1.9495202	28.07082	7
CGTAG	2556745	1.949189	23.306725	5
AGTAG	12210750	1.9369737	11.233237	35
TTCGT	6908645	1.9295197	5.0998883	44
TAGTA	14649695	1.8888962	19.954868	29
TTCGA	3045550	1.8872547	17.135279	31
TTAGT	32021535	1.860858	17.2106	28
GCGTT	5362065	1.8424264	21.247257	16
GCGGT	4343290	1.8360256	28.715155	6
CGTTC	612360	1.8237145	7.4930077	33
GTCGC	496655	1.8197285	8.525732	3
ATTCG	2934480	1.8184273	31.625242	34
AGGTC	2383805	1.8173447	51.971657	41
TTTAC	3603910	1.8152522	40.816296	13
TAGAG	11392625	1.8071955	11.112451	24
TATTT	38114965	1.8003825	8.621619	32
TAGTT	30612930	1.779	8.641015	29
ATCGT	2859920	1.7722243	16.512302	39
TTGAG	24714690	1.7669654	13.200776	44
AGGAG	9049275	1.7660253	6.2628503	38
TAATT	16244925	1.7025324	21.056274	23
TGAAC	1226005	1.6856426	59.34692	20
ACGGA	995135	1.6832825	8.024383	30
AGCGA	994875	1.6828427	28.02579	9
CTGAA	1190865	1.6373284	59.313923	19
CGTAT	2636355	1.6336861	28.24047	44
GTAGA	10267010	1.628641	10.765091	23
TATCG	2619285	1.6231085	16.913694	38
AGTTT	27708275	1.6102028	6.7206798	24
AGGTA	10135850	1.6078349	22.688496	47
GCGTA	2079520	1.5853665	23.184752	4
GGAAT	9910090	1.572023	10.133301	2
AGTTA	12160245	1.5679127	16.563124	30
GGAGA	7891405	1.5400591	9.276279	2
TTATT	32255635	1.5236136	6.834428	32
GGTTT	46836275	1.5091997	8.455147	2
TACGG	1963070	1.4965883	14.07746	5
TGGGA	17006045	1.4958154	16.172527	37
ACGGC	183900	1.4950026	9.032519	18
CGAGC	183295	1.4900844	6.591169	26
AGGTT	20688305	1.4791008	14.041945	41
TGGAA	9213680	1.4615525	8.500745	1
TTGTA	25103095	1.4588088	16.631514	20

TTATC	2895125	1.4582442	26.373512	38
GGAAA	4139290	1.4568526	10.132464	2
AGCGG	1544605	1.4487244	9.286716	6
GCGTC	394880	1.446828	6.92008	4
AGTCG	1896775	1.446047	18.319025	22
TTTTG	55016870	1.4409817	5.548181	34
AAACG	466340	1.422605	5.8189216	7
AGCGC	174925	1.422041	8.869716	35
GCGTG	3351255	1.4166657	28.37425	4
GAACT	1027345	1.4125035	59.048367	21
CGAGT	1847440	1.4084353	18.606476	33
GACGC	172715	1.4040749	10.387911	3
TTTAT	29653855	1.4007171	5.062119	13
CGTGG	3310130	1.399281	28.189722	5
TACGC	211755	1.3992399	5.796823	13
GTAGT	19515745	1.3952692	7.3519745	22
GTTTA	23936355	1.3910064	10.308636	12
GTTAA	10774205	1.3892001	24.474836	3
GGGAA	7088740	1.383414	12.706516	2
GGTTA	19316545	1.3810273	19.20168	2
GTCGT	3993770	1.372275	7.919193	3
TAGGA	8518855	1.3513335	5.1711135	37
GGACG	1440280	1.3508753	11.947743	2
GAGCG	1431625	1.3427576	6.197743	28
AACGC	91055	1.3349675	8.263328	11
TGGCG	3149890	1.3315432	25.579367	10
CACGC	18765	1.322211	10.44573	12
TTTAA	12599245	1.320451	6.218578	5
GTACG	1728050	1.3174158	13.959426	4
TTATA	12413515	1.3009857	7.722081	46
AACGG	765705	1.295199	6.687515	29
AGTCA	932350	1.2818943	58.710323	28
TTTGT	48928565	1.281519	8.566409	19
ACTAG	930120	1.2788281	58.64701	32
GGAGT	14445355	1.2705824	10.239213	2
GTAAT	9826235	1.2669711	24.380966	22
ACGGT	1660970	1.266276	13.808102	6
GTATT	21722050	1.2623271	9.1866455	31
AGATA	4366555	1.2491856	5.3939457	26
CGATT	2013905	1.2479689	11.647396	11
TTAAG	9676270	1.2476349	6.992246	6
GCGAC	153320	1.2464046	7.8326364	23
GGGTT	31315020	1.2414203	10.891591	2
GGTAG	14071745	1.2377205	7.2224474	2
CGTAC	187280	1.2375131	5.6229486	13
GAACG	730270	1.2352602	6.647737	3
ATTAT	11768070	1.2333405	7.751062	45
TTCGG	3572385	1.2274854	17.903008	35
CGAAC	83705	1.2272084	7.925775	9
AGCGT	1602330	1.2215705	5.2916746	29
CACAT	101830	1.2135031	19.962362	12
TATAG	9405005	1.2126586	9.156025	47
GGTGG	24516075	1.1956908	11.130907	8
TTAAT	11291760	1.1834213	18.250196	4
AGTAT	9116575	1.1754693	19.037395	30
TGTAA	8926685	1.1509852	23.746435	21
AAGTA	3979610	1.1384882	5.610166	34
GGGGA	10331390	1.117983	9.905347	2
GTTGA	15624615	1.1170746	12.049584	43
GGCGA	1185620	1.1120232	6.79688	2
TCGTG	3234335	1.1113302	7.526557	40
GGGGT	22116425	1.0786557	8.110665	2
GACGT	1406640	1.0723821	5.3564987	3
CGCAC	15125	1.0657309	6.025746	12
TCGGG	2509895	1.0610001	20.990536	36
TGAGG	12042105	1.0591978	13.492363	45
GAGTA	6639400	1.0531982	8.242909	34
GTTAT	17900245	1.0402316	7.665113	31
TGGAG	11825855	1.0401769	10.0727005	1
TGTAG	14408440	1.0301248	7.3851924	21
GCGAT	1338150	1.0201672	13.267188	10
TATTC	2020975	1.017944	23.922989	33
GATTA	7876100	1.0155253	8.902875	44
GGTTG	25499800	1.0108877	6.5878067	42
GGGAT	11471015	1.008966	7.8209114	2
GTTTG	31246710	1.0068592	8.114223	18
GGATT	14005370	1.0013074	6.0411363	2
CGTGT	2908815	0.9994802	7.297553	41
GGGTA	11279885	0.9921545	13.613418	2
GTGGT	24729785	0.9803622	9.263415	9
TGGTT	29997345	0.9666009	7.265638	1
CGTAA	698905	0.96092904	7.988845	21
TTTTC	4170825	0.94683975	6.363338	29
AGTTG	13195885	0.9434337	5.6324506	38
TTTGG	29065450	0.9365726	6.7833805	35
AAGGC	548540	0.9278619	16.32174	46
TGGGG	18978275	0.9256029	9.222781	1
TAGGC	1209805	0.92232066	6.3471794	13
TTGGG	22972620	0.91070294	7.840006	36
ATTAC	786520	0.87898445	5.87662	29
TAAGG	5455195	0.86534953	5.8795605	45
TGCGG	2034315	0.85995966	6.6276236	5
GTGGC	2015660	0.8520738	24.247248	9
GGGTG	17454155	0.8512689	7.91666	2
GGATA	5364300	0.85093105	6.930073	2
AAAAC	153440	0.8441643	8.972865	6
AGTGA	5230450	0.82969874	7.2124205	18
TAAGC	596655	0.8203449	21.646963	7
GGTAT	11386075	0.81404215	6.159202	2
TCTGA	1280635	0.7935789	26.832842	18
TGGTG	19967420	0.79156786	6.3737764	1
GAAGC	457665	0.7741458	14.524723	4
TATGC	1241480	0.7693156	27.620604	46
CATAC	64215	0.765247	12.621365	12
AGTGG	8650835	0.76090896	5.772032	8
TGGGT	19178900	0.76030856	8.448733	1
GGGGG	12570950	0.7542891	6.0153747	2

GGTAC	988875	0.75389	14.05503	3
GAGTC	972600	0.7414824	17.35108	21
GGTAA	4566705	0.7244097	6.1495776	2
GTGCG	1686370	0.712874	6.1975875	4
GGAAC	411255	0.69564265	6.6847553	2
GGAGC	737390	0.69161683	5.39851	27
TAGAC	499530	0.6868072	5.357946	25
TGGTA	9098985	0.6505277	5.516318	1
GCTTA	946685	0.5866381	26.323841	36
AGCTT	944690	0.5854019	26.238092	35
TAGCT	944115	0.5850456	26.243626	34
AGTTC	878875	0.5446179	5.2187176	42
TACAC	43290	0.5158848	6.8958344	12
TATCT	996500	0.50192666	22.386374	39
CTTAT	951920	0.47947216	21.423832	37
GTCTG	1383485	0.47537085	14.926546	17
ATTGC	654030	0.40528682	6.8090954	29
GGTGC	914585	0.38661963	6.258919	3
ACTTC	49165	0.26406613	10.246385	23
CTTCA	43635	0.2343644	10.184537	24
ATTCC	40525	0.21766055	7.8803787	23
CTCCG	6695	0.21261519	6.08077	24
TTCCA	35605	0.19123513	7.857652	24
GCTCC	5835	0.1853039	5.11084	23
ACTCT	32705	0.17565916	7.30749	23
CTCTA	31915	0.17141607	7.2532177	24

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
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCC	577308	0.7933626820778541	TruSeq Adapter, Index 10 (100CGGGCGC
188180	0.2586054402734946	No Hit	
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	103647	0.14243638042314216	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	96011	0.13194264494685135	No Hit
CGGGAAGCGGAGTTTGTAGTGAGTCGAGATTGCGTTATTGTAGTTCGTAGT	93568	0.128585364201883	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	88374	0.1214475352254746	No Hit