FASTQ QC Report

Report Date	`12-21-16`
Run ID	`161219_D00796_0155_ACAC53ANXX`
Project ID	`EC-EL-4039`
Sample	`Sample_YD20_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.

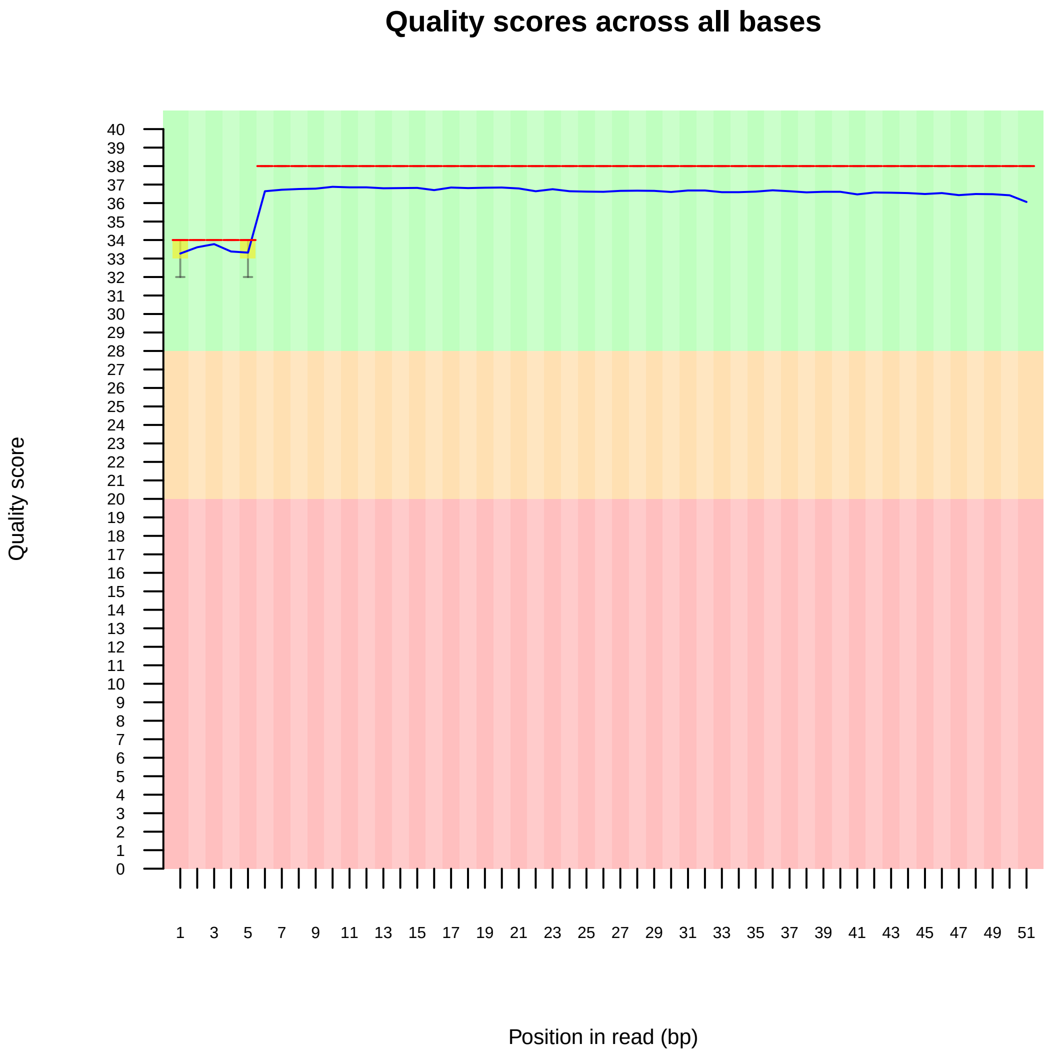
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

```
FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
FASTX-Toolkit:  http://hannonlab.cshl.edu/fastx_toolkit
```

1 Sequence Duplication

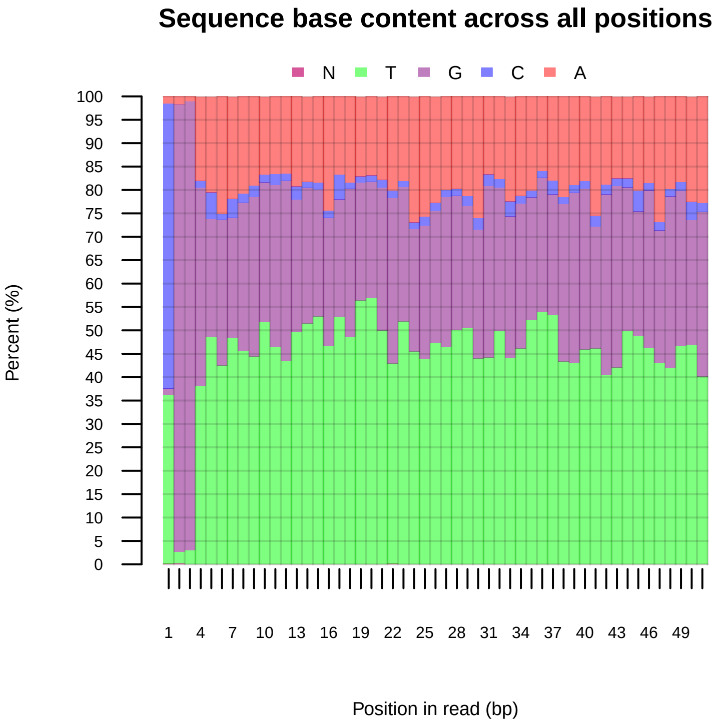
- Estimated Duplication rate 80.9245%

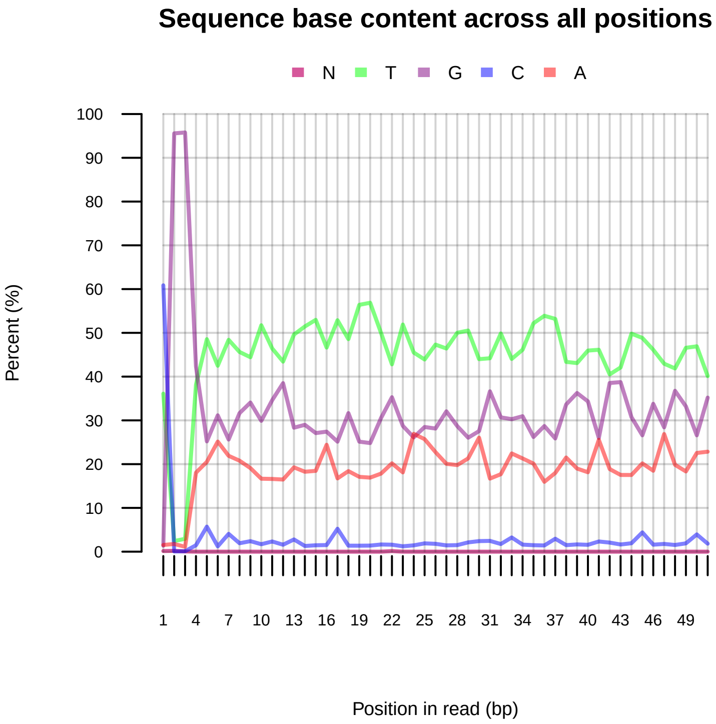
2 Per base sequence quality



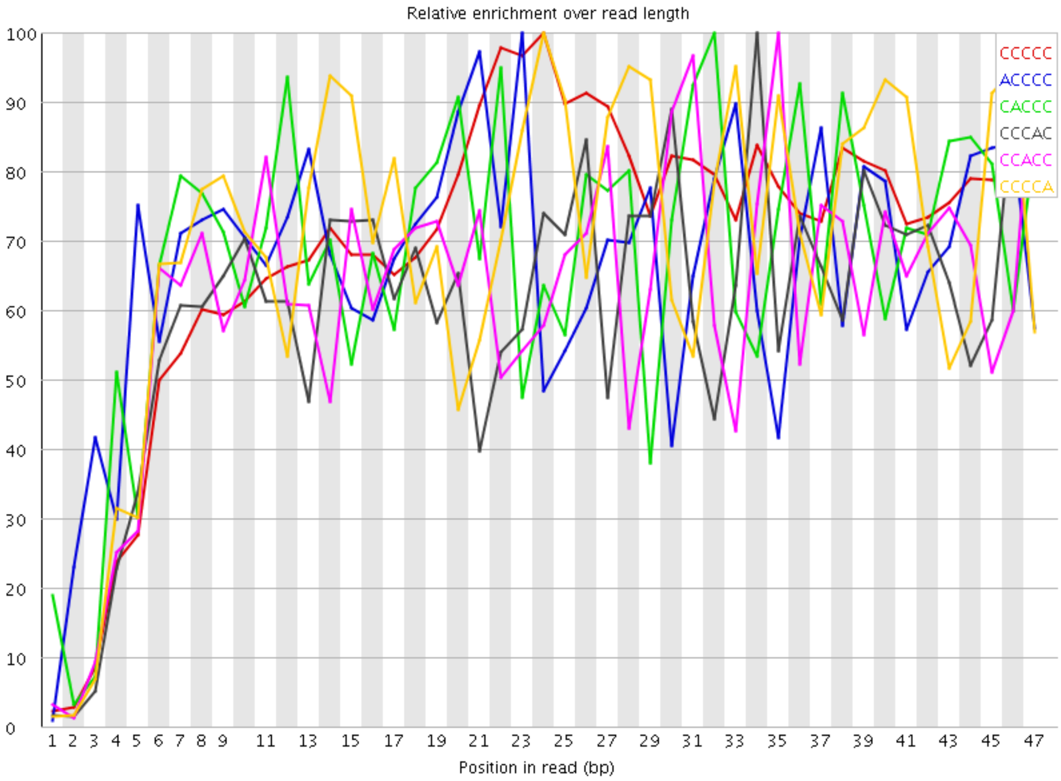
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	205645	1698.7748	2439.5408	24
ACCCC	56270	77.5955	116.30815	23
CACCC	53290	73.486115	108.85641	32
CCCAC	52280	72.09335	118.575745	34
CCACC	51565	71.107376	115.011986	35
CCCCA	51275	70.70747	102.376884	24
CACAC	245360	56.48139	634.1714	12
CGGGC	4022670	30.375044	1183.8025	1
GCCCC	24095	19.316992	32.397007	47
CCCCG	23965	19.212772	34.84599	41
CGCCC	23960	19.208763	30.701202	24
CCCGC	23910	19.168676	35.59832	34
CCGCC	23705	19.004328	33.149757	35
CGCGG	2137310	16.138758	441.49066	5
GCGCG	2010130	15.178423	438.59186	4
GGCGC	1602170	12.097933	439.99185	3
CGGCG	1530530	11.55698	340.05984	1
CGCGC	125715	9.781257	49.189957	27
CGGAA	3984965	8.640047	248.59198	1
CGGGA	6230460	7.8535247	260.64322	1
AGCAC	320280	7.155277	63.819077	10
AGACG	3237580	7.0195956	87.535385	27
TCCCC	11760	6.803609	22.983574	3
TCGCG	1188985	6.478857	35.24641	30
CGGGT	12067720	6.381785	250.82365	1
CGGAG	4939405	6.2261443	188.49005	1
GCACA	277795	6.2061324	63.761272	11
ACTCC	63175	6.101257	251.26654	23
CTCCA	61730	5.961703	249.97269	24
CTCCC	10285	5.950265	12.912514	24
CCTCC	9945	5.753562	10.330216	46
CGGTT	15029685	5.735698	209.91542	1
CGGGG	7724440	5.660626	137.41014	1
ACGCG	425665	5.5286293	20.173063	4
CCCTC	9470	5.478757	10.058369	45
CGCGT	1002595	5.4632053	32.79995	31
CCCCT	9410	5.444045	8.29156	43
GAGAC	2490605	5.400033	83.86373	26
AGGCG	4271185	5.38385	96.99253	47
AACCC	23265	5.3555574	10.620327	2
TCGAG	5500340	5.0032654	82.184525	44
CGTCG	917680	5.0004983	18.212654	41
ACACC	21060	4.8479705	7.3552203	37

CAACC	20440	4.705248	8.923612	31
CGGAC	359575	4.670238	153.82147	1
CGGTC	848500	4.6235313	173.29047	1
GGGCG	6275370	4.598718	120.53551	2
ACCAC	19680	4.5302973	7.517483	10
CGAGG	3585765	4.5198746	107.59311	45
ACCCA	19535	4.4969187	7.409302	33
CCCAA	19400	4.4658422	7.084808	29
CCACA	19290	4.4405203	7.841962	35
CCAAC	18985	4.37031	8.166458	30
CACCA	18960	4.364555	7.517467	16
ACACG	194210	4.3387856	61.47802	13
AAAAA	3954535	4.234704	9.008463	31
CGCGA	325930	4.2332497	22.647783	5
GACGG	3249085	4.095488	50.057446	28
TTACG	6165185	4.0469666	67.81083	14
ACGGG	3143405	3.9622774	50.265034	29
CGGTA	4258575	3.8737206	137.77097	1
CGACG	284745	3.69833	19.242374	24
TACGT	5606685	3.6803544	68.22182	15
ACGTT	5539580	3.636305	69.51063	16
GTCGA	3840405	3.4933412	81.76243	43
CGGAT	3675495	3.3433344	116.823784	1
CGTTT	12066560	3.3230693	40.445827	17
GAGGC	2613830	3.294746	78.45071	46
GGCGG	4475905	3.2800343	24.696678	11
AGAGA	8963305	3.2441564	27.521181	25
GGCGT	5584575	2.9532971	48.245796	3
ATCGC	315050	2.9528983	50.3319	29
CGAGA	1360880	2.9506073	36.75931	25
CGGTG	5514395	2.9161832	53.72345	1
ACGTC	307150	2.8788533	26.52791	15
AGATC	1784660	2.7923288	19.654417	27
AAGCG	1285775	2.7877674	50.863792	8
TTTCG	9805985	2.7005184	13.040294	30
GGAGG	21870450	2.6754522	31.998049	39
AGCGA	1222600	2.650794	51.637043	9
GCGGC	345180	2.6064425	7.8038983	33
TTCGA	3961865	2.6006577	28.01686	31
GGTCG	4910495	2.5968225	49.04972	42
AGGTC	2748130	2.499777	77.972176	41
TTTTT	178944935	2.4906301	5.1925073	16
ACGGA	1144910	2.4823494	13.295984	30
GGGAG	20116975	2.4609466	29.724	38
GACGC	178795	2.3222284	20.126377	3
CGTTA	3535795	2.3209758	20.936079	9
GCGGG	3154990	2.3120406	25.089878	12
GCGGA	1828050	2.3042657	23.693462	7
GCGGT	4352130	2.3015413	41.06489	6
TCGTT	8303575	2.2867625	5.3520026	36
CGTTC	581465	2.2864678	24.15379	33
AGTAG	14860915	2.2565844	16.60179	35
CGTAG	2470910	2.2476099	26.3813	5
TTTTA	67428430	2.2369716	11.149286	26
TTTAG	47650850	2.1906276	14.619501	27
TTTAC	4610465	2.1839752	47.956078	13
TTCGT	7915215	2.17981	5.3565316	35
TTCGC	554100	2.1788616	5.2594194	13
GAGGT	24198195	2.1362002	23.443588	40
ATTTT	63235585	2.0978718	8.919088	25
GATCG	2287005	2.0803246	12.546323	28
ATCGT	3160425	2.074574	19.464087	39
CGAGT	2265855	2.061086	34.904797	33
TAGAG	13537105	2.0555677	12.475289	24
TACGG	2242050	2.0394323	20.209864	5
AGCGC	156190	2.0286298	12.248476	35
ACGGC	156090	2.0273309	14.344813	12
GGAAG	9555260	2.0106156	12.030595	2
AAACG	533060	1.9879924	5.8607855	7
TAGTA	18057845	1.9787521	21.193844	29
AGGAG	9388580	1.9755429	6.6596975	38
GTCGC	362340	1.9744143	10.567561	3
AATTT	24931445	1.97148	19.01781	24
ATTCG	2965695	1.9467492	24.996782	34
TCGTC	493620	1.941039	8.023769	40
TATCG	2929485	1.9229801	20.023062	38
GCGTT	4901350	1.8704759	15.665046	16
TACGC	199035	1.8655138	7.2358856	13
CGAGC	143250	1.8605621	8.055799	33
AACGG	852705	1.848802	12.251234	29
GCGTA	2002250	1.8213034	26.20777	4
GTAGA	11991190	1.8208253	12.171973	23
GAAAA	2919920	1.8178228	5.203608	3
GGAAA	4943655	1.7892942	12.693449	2
TTAGT	38762070	1.7819884	13.986815	28
GGACG	1379775	1.7392133	15.603119	2
GAGCG	1377740	1.7366482	9.387103	28
GGAGA	8151945	1.7153306	11.097603	2
GGAAT	11261600	1.7100394	11.644441	2
GAGAT	11229450	1.7051575	10.428128	26
GAACG	780105	1.6913934	11.714852	28
TCGAA	1074665	1.6814508	5.2879095	44
TGGGA	18908965	1.6692706	18.612267	37
GTACG	1824360	1.6594896	19.874557	4
ACGGT	1815615	1.6515349	19.535227	6
TAGTT	35561280	1.6348401	6.389712	25
AGTCG	1786365	1.6249282	15.5941725	22
TAATT	20456090	1.6175866	18.761028	23
ACGAG	743395	1.6118001	6.2435393	32
TATTT	48505520	1.6091946	7.5274606	32
CGATT	2442380	1.6032332	17.909037	11
CGTGG	3010460	1.5920249	28.408052	5
AGGTA	10454595	1.5874982	19.348421	47
TGGAA	10439720	1.5852394	8.836938	1
CGTAC	167295	1.5680213	7.169825	13
AGCGT	1720865	1.5653478	7.3438344	29
GGTTT	57594310	1.5393255	9.338165	2

AGAGC	709390	1.5380719	6.917596	47
CACGT	163915	1.5363413	25.904734	14
AACGC	68525	1.5308958	9.736348	11
AGTTT	33299310	1.5308518	6.5717907	26
GGGAA	7271655	1.5301	13.7034645	2
GCGTC	280420	1.5280267	9.057401	4
GTCGT	3977765	1.5180132	9.486442	3
CACGC	11255	1.5062605	6.508487	15
GCGTG	2832440	1.4978822	28.518751	4
GGCGA	1185960	1.4949086	10.998555	2
AGTTA	13411715	1.4696361	10.68462	30
GGTTA	22877235	1.4574108	21.006636	2
AGGTT	22705370	1.446462	12.119418	41
GTTAA	13129120	1.4386697	25.901297	3
ACGCC	10700	1.4319847	6.2883935	16
CGAAG	652745	1.4152563	5.717458	45
GCGAT	1554510	1.4140264	22.695751	10
TCGGA	1549165	1.4091644	6.2811933	46
CACGA	62795	1.4028838	59.693462	31
AGATA	5326040	1.3910972	6.3807945	26
TCGTG	3640325	1.3892378	10.779494	40
CGATC	147515	1.3826275	26.82078	33
GTTTA	29996285	1.3790035	10.117561	12
TTAAG	12573990	1.3778393	8.574018	6
TTTAA	17388860	1.3750424	6.869884	5
GCGAC	105320	1.367919	12.102035	23
AAGTA	5115150	1.3360153	8.108905	34
TTTTG	69213075	1.33493	5.051311	34
TGGCG	2501220	1.322723	20.86269	10
AAGGC	608680	1.3197163	27.186962	46
GGGTT	35624280	1.3194032	13.564687	2
GGAGT	14930210	1.3180288	10.6645975	2
GGTAG	14877145	1.3133442	7.649463	2
TATAG	11950415	1.3095089	12.330365	47
CCAGC	9765	1.3068534	5.7538815	28
TTGTA	28411020	1.3061249	13.48411	20
GTAGT	20488095	1.3052089	7.446841	36
TTCGG	3413640	1.3027291	14.242615	35
GACGT	1421310	1.2928641	5.6999907	3
CCCAG	9640	1.2901245	6.068301	27
GTAAT	11768995	1.2896292	24.433016	22
TTATA	16301535	1.2890611	9.245643	46
ATCGG	1401050	1.2744348	5.822782	45
AGTAT	11276855	1.2357012	20.301445	30
GGGGA	10096035	1.2350665	11.181943	2
TAAGC	786150	1.2300322	35.41206	7
ATTAT	15170970	1.1996604	9.082023	45
CGTGT	3130705	1.1947542	10.301183	41
TTGAG	18717945	1.1924403	9.551909	44
GAGTA	7828845	1.188786	11.94616	34
GTATT	25850590	1.1884156	9.019224	31
GGTGG	22977925	1.1792959	11.700959	8
TTTGT	60677815	1.1703082	6.6228576	19
GGGGT	22769720	1.1686101	8.880595	2
GGGAT	13236330	1.1684942	9.655036	2
TCGAT	1776690	1.1662594	5.440861	11
TGTAA	10583015	1.1596713	24.144112	21
TTAAT	14501435	1.1467162	17.341755	4
CGTAA	731670	1.1447912	5.8032355	21
AAAAC	172735	1.1080669	7.9303517	6
TCGGG	2085090	1.1026605	18.598936	36
CGAAC	49345	1.1024013	7.4426713	9
GATTA	9934565	1.0886151	11.999342	44
GGGTA	12300980	1.0859221	15.126164	2
GGAAC	498665	1.0811862	10.838709	27
GGATT	16870715	1.0747612	7.092602	43
TAGGC	1180350	1.0736798	5.3362684	13
AGTAA	4101015	1.0711355	5.9628	9
AACTC	66390	1.0703326	43.339706	22
CGTGC	196235	1.0692973	5.632984	47
ACGAT	680040	1.0640093	5.6248937	32
TGGAG	12002330	1.0595576	9.484343	1
TTATC	2230015	1.056357	14.182638	37
TGAGG	11917475	1.0520666	12.0909395	45
AAGGT	6868335	1.0429356	5.003308	46
TGCGG	1936385	1.0240207	9.650952	5
TTTTC	5109415	1.015422	8.63428	29
ATTAC	883840	0.9979394	6.472551	29
TTGGG	26612905	0.9856522	8.341652	36
TAAGG	6490770	0.9856035	7.014834	45
CGTGA	1074225	0.9771455	6.9397144	26
TGGGG	18819395	0.96586764	8.869977	1
GGTAC	1057990	0.9623777	19.971027	3
TGGTT	35119965	0.93865275	7.11305	1
TTTGG	35049565	0.93677115	6.4717393	35
GGATA	6124785	0.9300298	7.7885284	2
GGTTG	24980075	0.9251777	5.523363	42
TAAGT	8391390	0.9195162	5.5170865	7
AGGTG	10394965	0.9176604	5.4061604	47
AGTTG	14324055	0.91252434	7.43433	38
GGAGC	720470	0.90815604	8.42938	27
GTGGT	24320375	0.9007446	9.462605	9
GTTTG	32981775	0.88150525	6.8482122	18
TAGAC	562945	0.8807994	7.520568	25
GGGTG	16787640	0.8615919	8.731001	2
TATTC	1818305	0.8613302	15.952491	33
AGTGA	5609950	0.85185367	5.879006	18
GTTGA	13228185	0.84271115	9.26641	43
GTGCG	1593225	0.842547	9.046137	4
GGGGG	11752580	0.8358449	6.37908	2
GAGCA	384320	0.8332678	6.377522	9
GGTAT	12549100	0.7994495	6.222995	2
GGTAA	5174870	0.7857881	6.584396	2
AGTGG	8797910	0.7766735	5.9694734	8
TGGGT	20580985	0.76225024	8.8106365	1
GTGGC	1426930	0.7546049	19.526077	9
GAGTC	782410	0.71170235	14.145646	21

TGGTG	18658630	0.6910527	5.503775	1
CGGCC	8750	0.68079376	5.994079	1
TCACG	67085	0.6287738	25.307985	30
TGGTA	9762455	0.6219243	5.1126842	1
ATGCC	65400	0.61298066	27.29697	47
TCCAG	63175	0.5921262	25.18027	25
CAGTC	62595	0.58668995	25.46433	27
GTCAC	62140	0.58242536	25.534794	29
CCAGT	61320	0.5747397	25.151646	26
AAGTC	363445	0.5686562	5.425826	41
CGTCT	131600	0.5174845	10.795066	16
ATCTC	69755	0.47180644	19.760658	40
GGTGC	889155	0.47021276	9.118964	3
TCTCG	77990	0.30667645	11.511396	41
CTCGT	76500	0.30081734	11.51048	42

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	248564	0.30751131433770573	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTCGAGTAGTTGGGATTATAG	156662	0.19381462129179472	No Hit
CGGTTAATTTTTTTGTATTTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	139045	0.17201972410359623	No Hit
CGGTTAATTTTTTTGTATTTTTTAGTAGAGACGGGGTTTTATTTTGTTAGTTA	93683	0.1159000597878184	No Hit
CGGGTTTACGTTATTTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	91954	0.11376102492158717	No Hit