FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD18_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.

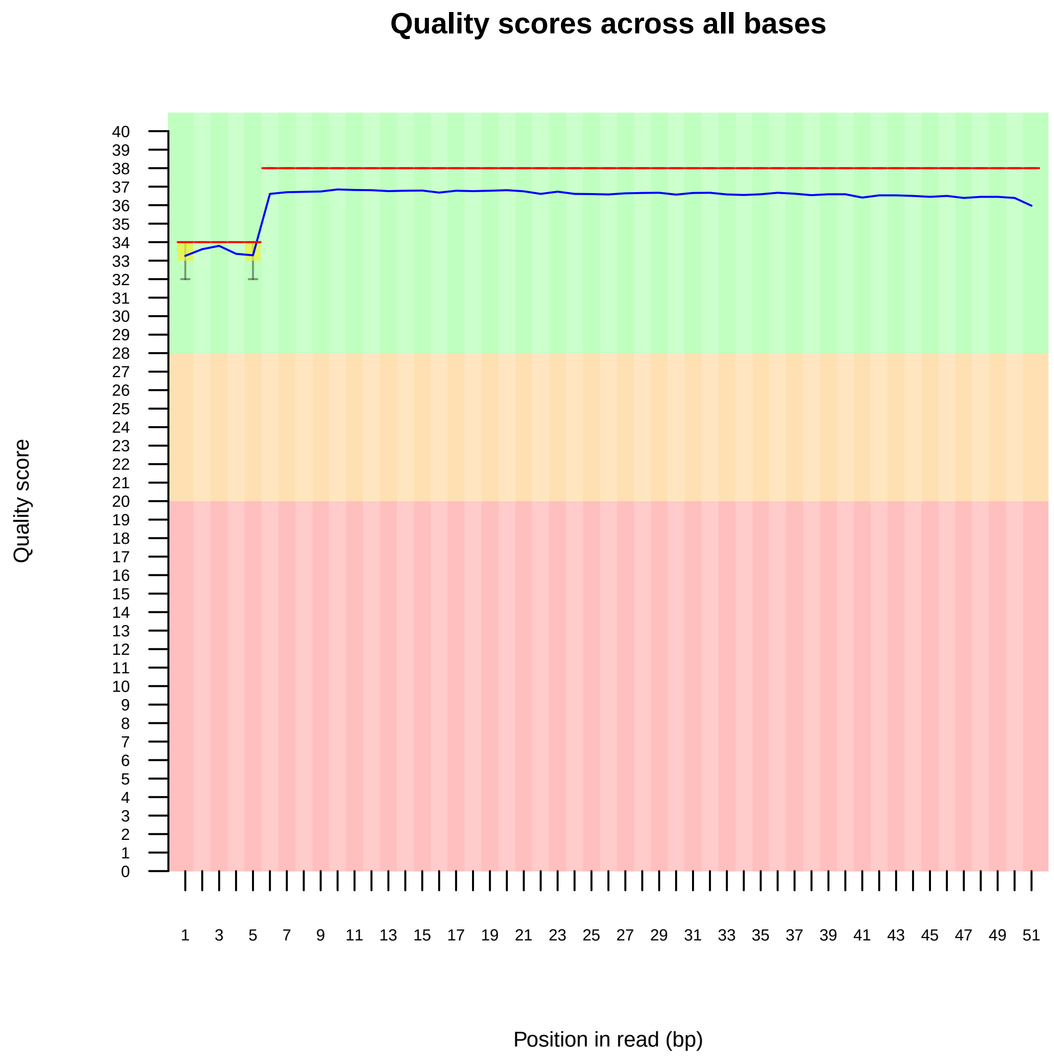
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

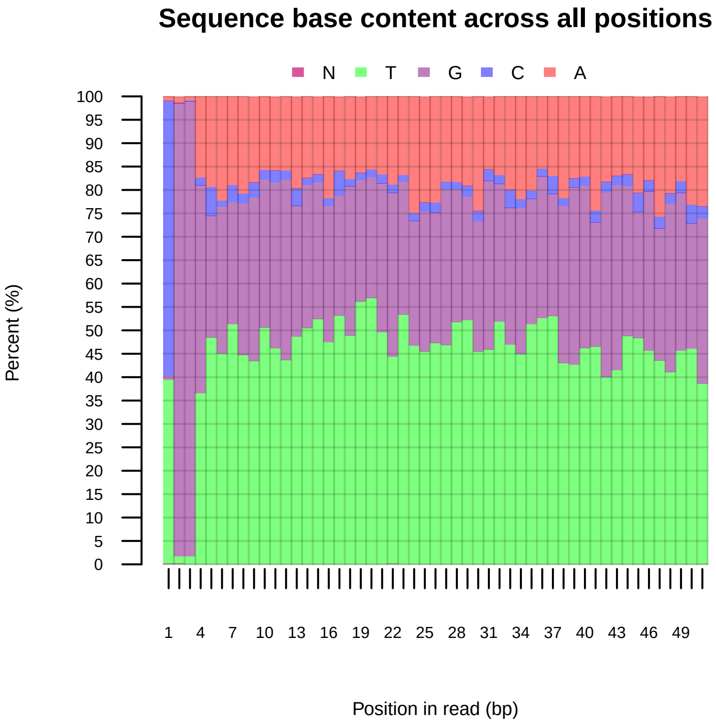
- Estimated Duplication rate 82.0264%

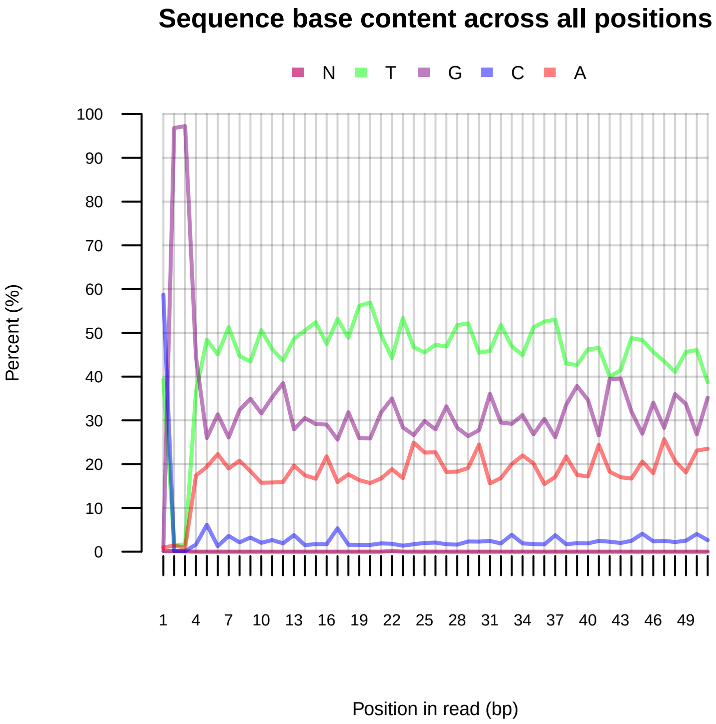
2 Per base sequence quality



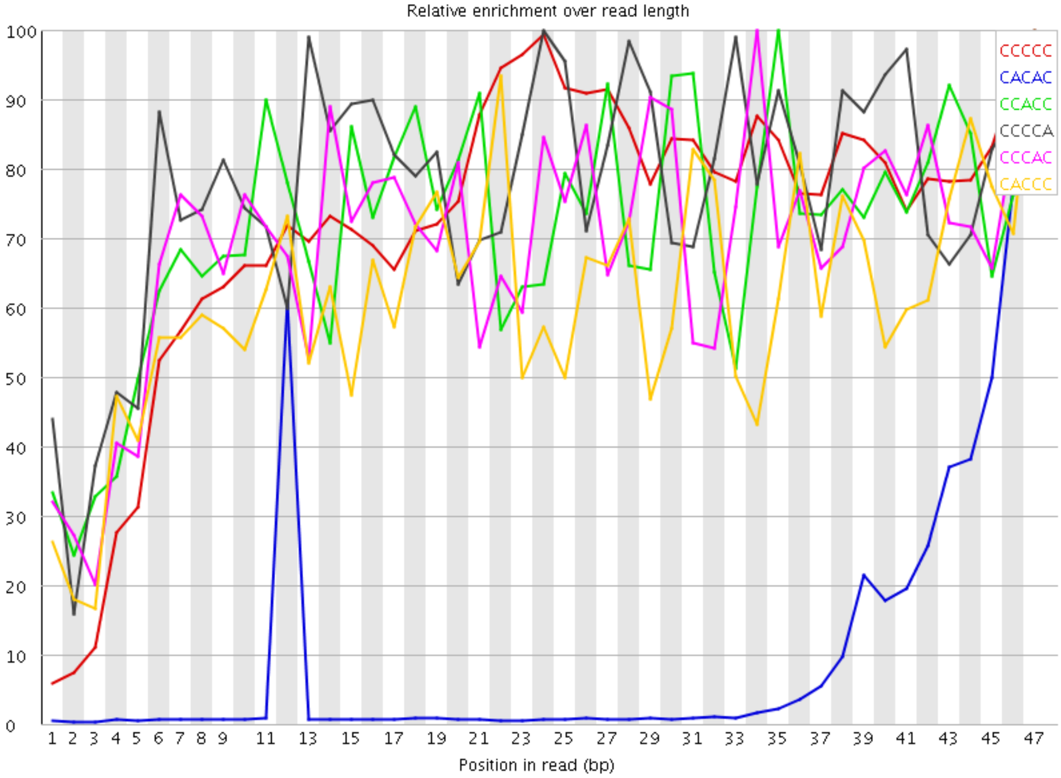
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	203450	1317.6798	1822.9467	47
CACAC	606920	139.06062	1305.579	47
CCACC	64650	78.755516	110.18646	35
CCCCA	64055	78.0307	101.314316	24
CCCAC	63790	77.70788	111.617455	34
CACCC	63475	77.32416	125.64406	47
ACCCC	61710	75.174065	125.64142	23
CGGGC	4084355	28.514711	1025.7124	1
GCCCC	33415	22.19004	38.84907	45
AGCAC	836465	19.65101	137.09952	45
CCCCG	26975	17.913403	32.296234	46
GCACA	709740	16.673868	136.54211	46
CCCGC	24960	16.57529	31.20338	34
CCGCC	24910	16.54209	30.151756	22
CGCCC	24520	16.2831	25.742838	23
CCTCC	30215	14.661824	19.495132	45
CTCCC	28900	14.023721	19.72271	24
CGCGG	1840920	12.852286	302.5211	5
TCCCC	25985	12.609218	28.060307	3
GCGCG	1734180	12.107087	300.89236	4
CCCTC	23530	11.417927	15.390956	39
CCCCT	22860	11.09281	15.2765465	35
CGGAA	4604670	11.091749	194.44983	1
ACACG	453310	10.649579	104.05971	47
CGGCG	1494225	10.431853	291.15292	1
CGCGC	145445	9.903298	58.180763	13
GGCGC	1297760	9.060245	301.90662	3
CTCCA	98855	9.022446	301.8069	24
ACTCC	93960	8.575682	302.72946	23
CCACA	35605	8.157999	10.604596	38
CACCA	33290	7.6275754	11.304686	40
CGGGA	5772220	7.579641	225.01677	1
AACCC	30685	7.0307035	10.894336	2
ACCCA	30440	6.974569	10.01268	45
ACCAC	30185	6.916142	9.635865	46
ACACC	29610	6.784395	10.12035	47
AGATC	3841535	6.761688	35.09156	43
CCCAA	29425	6.742006	10.38927	14
TCGCG	1251560	6.3847804	22.255526	30
CGGGT	11667475	6.1028905	238.44202	1
CAACC	25835	5.9194474	9.258832	31
AGACG	2449400	5.9001265	65.39174	27
CCAAC	25640	5.8747673	9.635644	15

CGGAG	4248615	5.5789585	165.15024	1
ACGCG	413045	5.289794	15.19154	14
CGCGT	1010820	5.156655	20.004683	31
CGGAC	385985	4.94324	163.0373	1
CGTCG	960465	4.8997717	23.74599	41
CGGTT	12626105	4.825884	164.928	1
CGGGG	6569300	4.7025023	116.99563	1
CGCGA	365945	4.6865916	20.348656	5
CGGTC	913045	4.65786	174.9399	1
GGGCG	6345040	4.5419707	109.985855	2
AGGCG	3449775	4.5299826	65.89241	47
TCGAG	4656790	4.468288	55.185913	44
ACGTC	465525	4.3564534	33.0031	15
GAGAC	1758935	4.236931	62.542885	26
CACGT	439575	4.1136093	32.418278	14
GATCG	4166040	3.997403	20.47275	44
CGACG	308875	3.9557064	29.014065	24
AAAAA	2584985	3.941046	11.106987	31
CCCAG	31355	3.9163697	5.3994536	25
TTACG	5226395	3.664422	45.556713	14
CCAGC	28275	3.5316648	4.460418	26
CGAGG	2683555	3.5238407	71.24657	45
CGGTA	3664830	3.5164814	123.15333	1
CAGCC	28030	3.5010636	5.1353498	31
AGAGC	1441875	3.473195	22.251997	47
GACGG	2584655	3.3939726	35.36505	28
TACGT	4759805	3.3372781	46.95275	15
GGCGG	4646465	3.3260796	37.330173	11
CGTTT	11905130	3.3249903	36.340565	17
ACGTT	4711720	3.3035643	48.450436	16
AGCCC	26155	3.266868	5.7223763	46
ACGGG	2486690	3.2653325	35.38541	29
GGCGT	6102425	3.1919873	49.401855	3
ATCGG	3277450	3.144782	18.27705	45
CGGAT	3265545	3.1333592	106.881935	1
TCGGA	3238140	3.1070635	18.618822	46
AAGCG	1236870	2.9793785	57.780228	8
AGAGA	6483390	2.937406	22.568037	25
AGCGA	1176715	2.8344762	58.514225	9
TTTCG	10043710	2.8051133	15.938445	30
GTCGA	2851120	2.7357094	54.60159	43
CGAGA	1133530	2.730452	31.993841	25
GCGGC	382955	2.6735804	8.996574	33
TTCGA	3802955	2.666395	36.34631	31
ATCGC	284485	2.6622534	30.437338	29
GAGGC	1987000	2.6091776	54.26524	46
GGAGG	19111625	2.573166	30.946424	39
TTTTT	165603205	2.5321352	5.4696217	16
GCACC	20245	2.5286846	7.8352547	47
CGTTA	3584380	2.5131438	33.496494	9
GCGGG	3459175	2.4761813	38.002308	12
GAGCA	1012215	2.438228	17.23838	47
CGTTC	647780	2.4147413	29.105879	33
TCGTT	8601340	2.402273	6.269828	4
TTCGC	641265	2.3904552	7.8558025	33
GGAAG	9636430	2.3800287	11.939846	2
CGGTG	4540790	2.375145	46.42765	1
GGGAG	17446495	2.348975	27.329338	38
AAACG	520415	2.299573	16.276865	37
AGTAG	12730455	2.2975166	23.186396	35
CGAGT	2363085	2.2674298	46.090965	33
GCGGA	1722275	2.2615607	22.385723	7
GAAGA	4966350	2.2500863	8.550544	46
CGTAG	2339380	2.2446842	23.924541	5
TTTTA	58477595	2.2446816	11.467926	26
TTCGT	7961385	2.2235396	5.6905675	35
TTTAG	42295405	2.221824	14.851454	27
ATTCG	3103025	2.1756475	36.21683	34
GAGGT	21857330	2.1503842	22.951822	40
GGTCG	4087725	2.138161	31.454576	42
AGGAG	8483505	2.0952766	9.288831	38
GAGAT	11565480	2.087269	8.552244	26
ACGGA	849045	2.0451837	7.676167	30
GCGGT	3881200	2.0301342	29.277678	6
AAGAG	4456590	2.0191312	8.552248	47
TTTAC	3938760	2.0179553	32.49437	13
GCGTT	5258770	2.0099795	23.05407	16
CACGC	15930	1.9897233	5.1941576	46
GACGC	155275	1.9885789	13.131637	3
ATTTT	51463975	1.9754614	7.6954823	25
TACGC	208865	1.9545901	10.098144	13
AGCGC	151440	1.9394648	8.827903	10
ACGGC	150750	1.9306282	10.6030855	12
TCGTC	517790	1.9301752	9.92467	40
TAGAG	10646320	1.9213843	9.934581	24
GTCGC	374135	1.9086335	10.002401	3
CGAGC	147695	1.8915032	5.6987066	32
AGGTC	1928005	1.8499614	52.11234	41
TAGTT	35184895	1.8483009	8.381824	25
AATTT	18992795	1.830211	17.135315	24
GCGTA	1907075	1.8298786	23.872469	4
TACGG	1881340	1.8051852	14.226364	5
ATCGT	2570685	1.8024039	14.206053	39
TTAGT	33851315	1.7782466	14.232966	28
GGAAA	3914805	1.7736666	12.382921	2
AACTC	103260	1.7726296	58.29635	22
TAGTA	13428335	1.7708658	15.847222	29
GAGCG	1342535	1.762915	11.038891	28
AGGTA	9735810	1.7570611	26.268446	47
GAAAA	2105045	1.7495205	5.4378242	3
AGCGG	1322860	1.7370793	6.176263	6
GGAGA	7001680	1.7292917	11.057928	2
GGACG	1316715	1.7290102	17.494701	2
TATCG	2412585	1.6915541	14.631663	38
AACGC	71950	1.690316	10.277146	11
GTAGA	9318605	1.6817663	9.648649	23
TGGGA	16985795	1.6711091	15.033583	37

AGTTT	31296275	1.6440276	8.617154	26
AGTTA	12318245	1.6244724	18.138048	30
GCGAC	125815	1.61129	20.222324	23
CGATT	2292430	1.6073089	19.44049	11
AGTCG	1667880	1.600366	12.868532	22
TATTT	41316145	1.5859337	5.768425	32
TCGTA	2254260	1.5805466	6.186203	45
AGCGT	1642830	1.57633	8.658273	29
TGAGA	8731995	1.5758984	5.2479634	41
CGTAC	168090	1.5730116	8.398631	13
TGGCG	3004015	1.5713061	31.571793	10
GCGTC	306050	1.561301	9.792256	40
CGTGG	2983655	1.5606564	32.179646	5
GGGAA	6316880	1.5601583	14.284284	2
GCGTG	2977390	1.5573795	32.375652	4
TAGGA	8615750	1.5549192	7.012684	37
GTCGT	4065120	1.5537487	10.818372	3
TCGAA	866390	1.5249786	5.1633162	32
GTACG	1562180	1.4989444	13.936237	4
CGATC	159725	1.4947306	32.733532	40
GGTTT	52181490	1.4942975	9.705243	2
TTGAG	20770225	1.4931688	12.710735	44
AACGG	618980	1.4910018	6.3411565	29
ACGGT	1537325	1.4750959	13.536584	6
AGGTT	20396415	1.4662957	13.578762	41
ACGAG	607800	1.4640716	5.374562	32
TAATT	15147260	1.459642	16.89136	23
AAGTA	4380000	1.4500558	11.797219	34
GCGAT	1506460	1.4454801	24.353144	10
GTAGT	20097880	1.4448339	9.85509	36
TTCGG	3769150	1.4406247	20.231297	35
GGCGA	1094985	1.4378512	10.700943	2
TATAG	10871620	1.4336985	17.29866	47
AACGA	320770	1.4173957	15.963357	38
TTATT	36800170	1.4125865	6.17061	32
TTTAA	14585310	1.4054906	8.196483	5
TTAAG	10632685	1.4021889	10.247906	6
AAAAC	172745	1.4002277	17.144958	6
TCGAC	148880	1.3932416	5.4810963	33
TTATA	14402700	1.3878938	13.055471	46
TAAGC	777965	1.3693372	40.88897	7
GTTTA	25953015	1.3633403	8.349968	4
AGATA	4060165	1.3441702	5.436483	26
GAACG	553705	1.3337673	6.160629	28
GGAGT	13502425	1.3284056	10.975967	2
GGTAG	13461615	1.3243908	7.7513266	2
GGAAT	7264805	1.3111088	9.904796	2
GGTTA	18234575	1.3108811	16.93669	2
GAGTA	7252215	1.3088367	16.617857	34
GGGTT	33223800	1.3020306	15.116803	2
GACGT	1348965	1.2943603	6.105463	3
ATTAT	13050990	1.2576382	12.938956	45
GTTAA	9482465	1.250503	20.19568	3
GGGAT	12695970	1.2490644	12.379836	42
TTGTA	23559050	1.2375827	13.622121	20
TGGAA	6799860	1.2271984	9.3308	1
GGGGA	9086190	1.2233536	10.731097	2
TCGGG	2331430	1.219498	26.281332	36
CGTAA	687340	1.2098234	9.403446	21
CGAAC	51435	1.2083586	8.008676	9
GATTA	9031695	1.1910576	16.990372	44
GTAAT	8977215	1.1838729	21.559912	22
TCGTG	3072715	1.174437	6.9797564	40
TTTGT	55929340	1.170332	6.4508944	19
GGTGG	21742820	1.1661085	11.24594	8
CGTGA	1201835	1.153186	8.935326	26
TGAGG	11708985	1.151962	15.289239	45
GAAAC	259120	1.1449811	15.386148	36
GGATT	15791210	1.135228	9.421203	43
TAGGC	1172155	1.1247073	8.199749	13
AGTAT	8438495	1.1128291	14.979491	30
GGGGT	20684510	1.1093495	8.894877	2
TCGAT	1573105	1.1029631	6.896204	11
TATTC	2137315	1.0950162	25.46085	33
ACGAT	619020	1.0895697	6.750616	39
TTTTC	5330870	1.0879353	10.995266	29
AAGGC	451510	1.0875994	17.733572	46
GGGTA	11014305	1.0836177	15.213668	2
AGTAA	3235185	1.0710499	7.3760905	9
GTATT	20322675	1.0675724	6.4434876	31
TGGAG	10802950	1.062824	10.27661	1
TGTAA	7953050	1.048811	20.958788	21
TTAAT	10578765	1.0194062	13.376361	4
GTGGC	1931970	1.0105529	29.82934	9
GTTAT	19230775	1.0102136	7.4563503	31
AGTTG	13992985	1.0059538	9.88754	38
TGTAG	13864515	0.9967181	7.2220874	21
AAGAC	223085	0.98575217	8.694352	32
CGTGT	2576430	0.98474956	6.694572	41
TAAGT	7463825	0.98429424	6.528371	7
ATTTC	1906080	0.976547	5.6678658	22
GTTGA	13567985	0.9754006	11.874463	43
CGTCT	260665	0.9716855	12.807496	16
GGTTG	24477870	0.95928025	6.5282416	42
TAGAC	544930	0.9591599	11.438611	25
GGAGC	728505	0.9566175	10.168601	27
TGCGG	1815130	0.9494375	6.331251	5
TGGGG	17573340	0.9424914	9.134404	1
TCCAG	100260	0.9382483	31.800476	25
TTGGG	23911130	0.93706995	6.50845	36
TAAGG	5145730	0.9286707	5.3283625	45
GGATA	5107160	0.9217099	7.5816126	2
TTATC	1789970	0.9170601	10.442792	37
TGGTT	31416185	0.899651	7.760305	1
GTGGT	22955140	0.8996049	7.9244704	9
GTTTG	31303440	0.8964223	6.517779	18
TTTGG	30618150	0.87679803	5.140131	35

GGGTG	16030550	0.8597488	8.332766	2
CCAGT	91100	0.8525276	31.752106	26
TCACG	90620	0.8480357	32.062103	30
GAAGC	351620	0.8469839	9.533523	4
CAGTC	90040	0.842608	32.022533	27
GGTAC	859965	0.8251545	14.055715	3
GTCAC	88140	0.8248274	32.178623	29
GGGGG	11159635	0.81907636	6.064591	2
GGTAT	11346350	0.81568766	6.2181554	2
TGGGT	19767575	0.7746852	9.581609	1
GGTAA	4272855	0.7711395	6.3723574	2
AGTGG	7805080	0.76788527	5.108518	8
GGAAC	314840	0.75838804	5.537261	2
GTGCG	1445725	0.7562135	5.9382524	4
TGGTG	18796735	0.7366383	5.9457407	1
GTTGG	18700550	0.73286885	5.4741383	39
ATCTC	96595	0.66053134	24.330313	42
GAGTC	680890	0.6533283	11.6915655	21
TGGTA	8989105	0.6462256	5.4173374	1
GATTC	751355	0.52680326	6.039131	29
TGGAT	7168430	0.5153375	5.2826376	1
TGAAC	237345	0.4177634	6.484564	20
GGTGC	744990	0.38968086	6.0142813	3
TCTCG	99395	0.37051648	13.291538	43
CTCGT	95990	0.3578236	13.32624	44
CTGAA	180870	0.3183588	6.2108545	19
GAACT	123720	0.21776609	6.19926	21
AGTCA	103280	0.18178856	6.1810102	28

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	249488	0.34468039862057104	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	190824	0.26363309011404096	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	154084	0.21287490597163822	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	89179	0.1232053376057522	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	87142	0.12039111819644154	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	87133	0.12037868423734295	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	83820	0.1158016057380566	No Hit