FASTQ QC Report

Report Date	`12-21-16`
Run ID	`161219_D00796_0155_ACAC53ANXX`
Project ID	`EC-EL-4039`
Sample	`Sample_OD13_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.

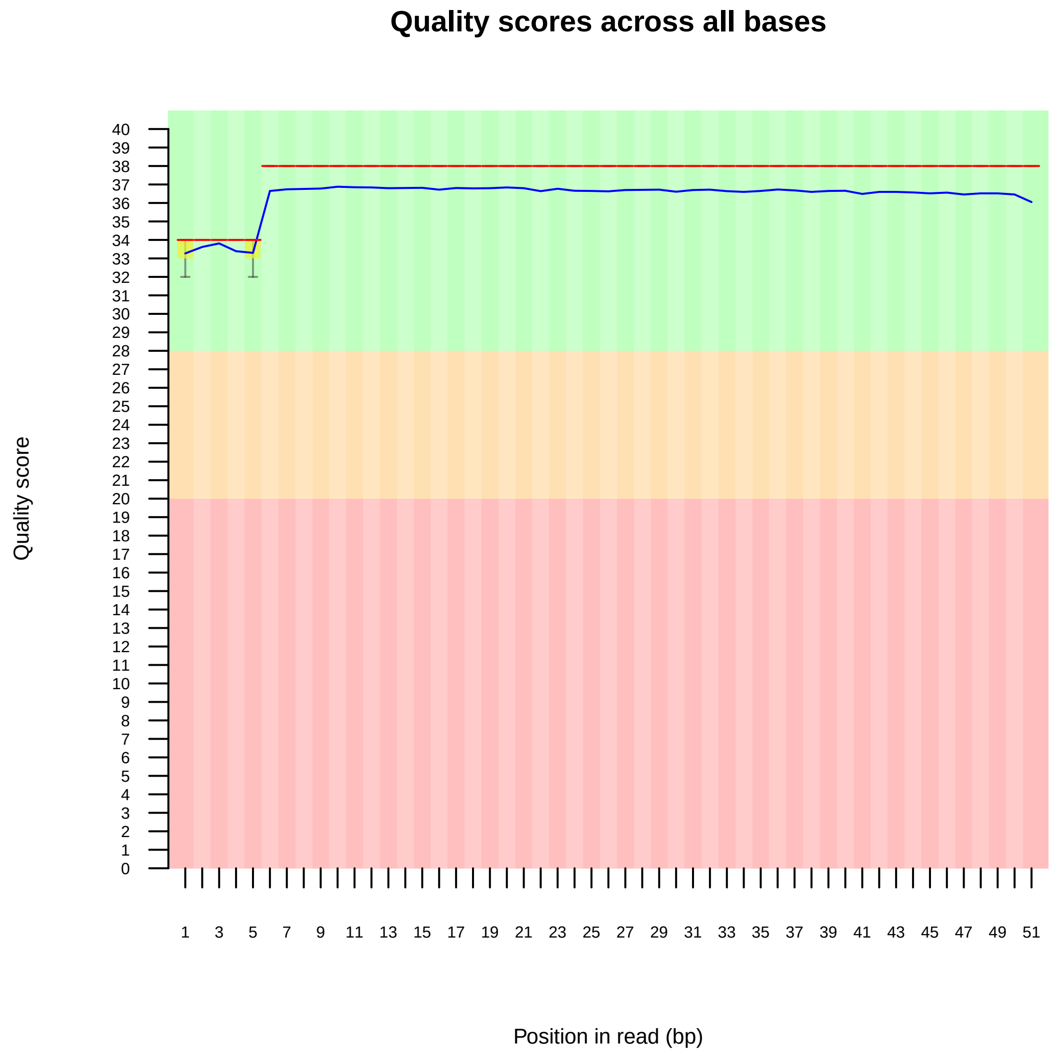
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

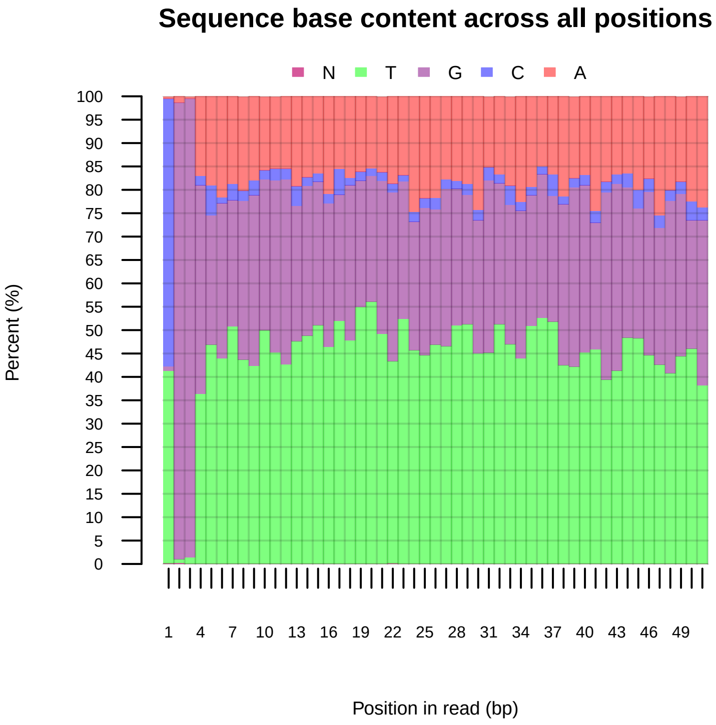
- Estimated Duplication rate 83.0447%

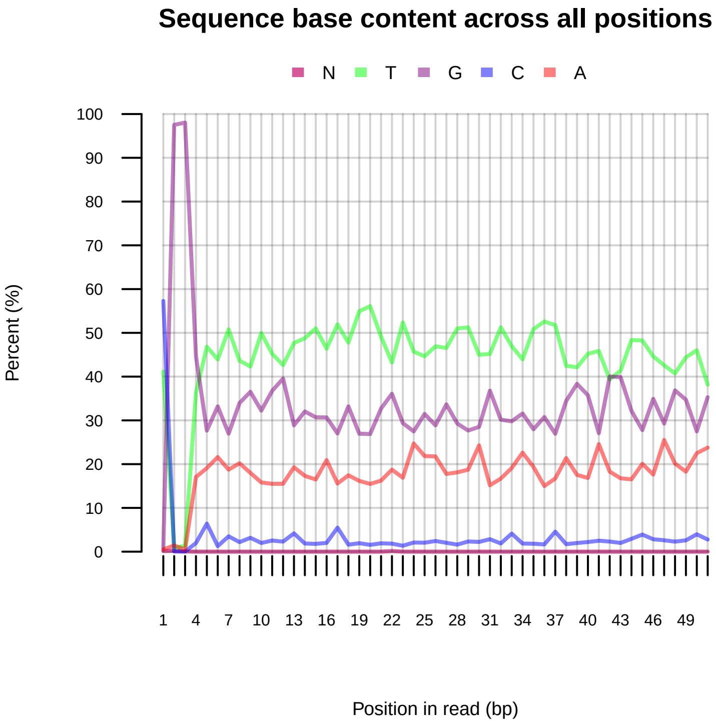
2 Per base sequence quality



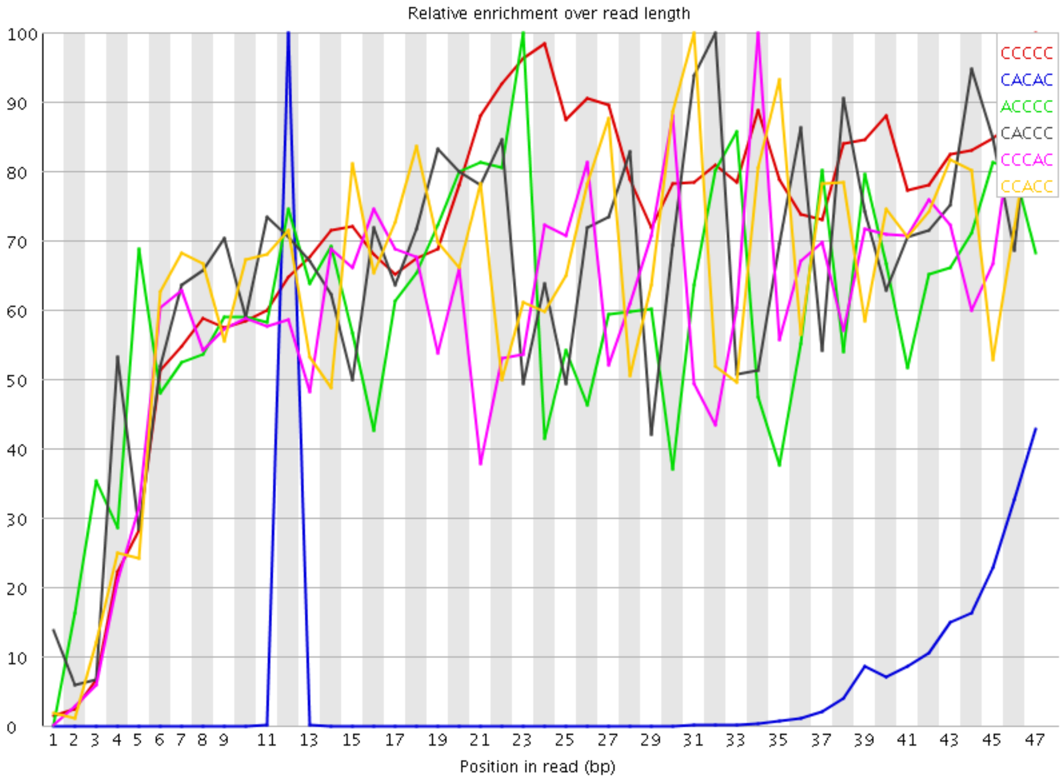
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	177285	931.7145	1324.8829	47
CACAC	916725	186.6095	3147.2915	12
ACCCC	50150	51.870895	86.26589	23
CACCC	47875	49.51783	75.573616	32
CCCAC	46150	47.73363	79.94764	34
CCACC	45560	47.12339	73.8726	31
CCCCA	44705	46.239048	76.78863	24
ACTCC	334510	27.09557	1213.1097	23
CGGGC	4737180	26.933598	934.37195	1
CTCCA	329690	26.705147	1207.9501	24
AGCAC	1222680	25.550257	331.7128	10
GCACA	1068105	22.320116	331.02515	11
ACACG	771305	16.117908	324.80438	13
CAACA	334785	13.412278	606.2526	37
CGGAA	5565640	11.939482	177.36334	1
GCCCC	21610	11.658792	24.336992	47
CGCGG	2040395	11.6008215	252.34557	5
CCCGC	21415	11.553587	24.463745	47
CCCCG	21120	11.394431	21.675117	46
CGCCC	20885	11.267648	20.407516	44
CGCGC	203025	11.244395	62.06718	13
CCGCC	20620	11.124677	19.9265	22
GCGCG	1933185	10.99127	249.67937	4
CGGCG	1752950	9.966531	265.44562	1
GGCGC	1384665	7.872618	250.57791	3
AGATC	4756370	7.7838516	42.029102	43
CGGGA	6446150	7.213014	201.37263	1
CACAG	334975	6.9999495	319.88504	31
TCGCG	1536355	6.6636868	20.990662	30
ACGTC	777085	6.461682	129.51344	15
CACGT	668645	5.5599723	129.49782	14
AACTC	348585	5.5569925	244.2295	22
CAGTC	667180	5.5477905	127.501236	33
CAATC	339675	5.4149528	248.03929	40
CGGGT	12138475	5.4047422	206.77977	1
TCACA	336705	5.367606	244.22435	30
AGACG	2499515	5.3619914	58.459553	27
TCAAC	330655	5.2711596	241.8386	36
CGTCG	1214890	5.2693853	25.499456	41
CGCGT	1205295	5.227769	19.081034	31
ACGCG	477155	5.2010136	15.498313	14
CGGAG	4592450	5.138789	149.24216	1
CGCGA	450960	4.915487	19.383137	5

CGGTC	1132555	4.912271	178.02374	1
CGGAC	434105	4.731766	147.58913	1
GATCG	5358500	4.574126	23.111591	44
CGGTT	12973325	4.406672	146.55435	1
CGACG	399340	4.352826	31.987684	24
AACCC	21020	4.278853	8.332523	22
GGGCG	7284165	4.251499	100.52824	2
TCGAG	4887375	4.1719637	45.930176	44
AGAGC	1940690	4.163193	35.036892	8
CGGGG	7085525	4.13556	103.33967	1
TCCCC	10030	4.1280813	13.537311	5
AGGCG	3617285	4.047614	52.619335	47
GAGAC	1765505	3.787384	55.987953	26
CTCCC	9190	3.7823596	17.598505	24
ACACC	18180	3.7007396	6.0260415	47
CCTCC	8935	3.6774085	7.542232	23
ATCGG	4292070	3.6637995	21.227634	45
CAACC	17725	3.6081195	6.934572	31
TCGGA	4210925	3.594532	21.52226	46
AAAAA	2290600	3.5544264	10.067516	31
ACCCA	17280	3.5175347	6.695449	33
ACCAC	17195	3.500232	5.547777	45
CCACA	16940	3.448324	5.643307	35
TTACG	5285685	3.4420302	39.13	14
CCCCT	8345	3.4345798	6.4787164	47
CGTTT	13156095	3.4090626	37.20349	17
CCAAC	16715	3.4025228	6.1693783	30
CCCAA	16585	3.3760595	6.121554	28
CGGTA	3934260	3.3583655	115.476166	1
CCCTC	8065	3.3193395	6.382019	47
CACCA	15970	3.2508698	4.973773	8
GGCGT	7189990	3.201394	49.527477	3
GGCGG	5446950	3.1791842	39.37491	11
TACGT	4816855	3.1367292	40.96821	15
ACGTT	4811280	3.1330988	42.53914	16
CGAGG	2797545	3.130354	56.870285	45
GAGCA	1447645	3.1055067	34.62815	9
GACGG	2753900	3.0815163	30.345345	28
CGGAT	3422045	2.9211283	97.17435	1
ACGGG	2595125	2.9038532	30.471006	29
TTTCG	10973565	2.843516	14.803678	30
AGAGA	6698125	2.8279092	21.094706	25
GCGGC	497355	2.8277497	10.246709	9
TCCAG	336380	2.7970948	126.80577	25
GTCAC	336090	2.7946832	127.9232	29
CCAGT	333890	2.7763898	126.74522	26
AAGCG	1282480	2.7511926	49.69314	8
CGAGA	1268370	2.7209237	34.121567	25
CGTTC	822305	2.7208495	29.389307	33
AACAA	344675	2.7176166	122.24007	38
ATCGC	318180	2.6457567	30.690945	29
TTCGA	4061200	2.6446474	33.812283	31
AGCGA	1223650	2.6249895	50.447617	9
GTCGA	3043610	2.598088	45.324314	43
TTCGC	778445	2.5757248	10.048303	33
CGTTA	3838580	2.499678	30.681665	9
TCGTT	9633360	2.4962366	6.863702	4
TTTTT	161119495	2.494256	5.350258	16
GGAGG	21399140	2.4581056	28.794174	39
GAAGA	5819120	2.4567986	9.680494	46
GGAAG	11141545	2.4535995	11.284567	2
GCGGG	4122390	2.4060874	40.396862	12
ATTCG	3641880	2.3715868	43.47652	34
TTCGT	8866460	2.2975142	5.8716235	35
GAGGC	2018825	2.2589939	43.951378	46
TTTTA	57962445	2.2549925	12.51524	26
ATCTC	355415	2.2545578	101.37656	42
GGGAG	19441640	2.2332487	25.161417	38
TTTAG	43741060	2.2306898	15.673674	27
CGTAG	2601570	2.220754	23.66867	5
TCGTC	671145	2.2206898	10.581833	40
GCGGA	1982685	2.2185545	23.586765	7
AGTAG	13130350	2.2058861	20.802702	35
AGAAA	2718165	2.2000983	5.4968715	22
AAGAG	5180630	2.1872318	9.6616535	47
CGGTG	4910375	2.1863794	43.549572	1
GAGAT	12988225	2.1820092	8.5043	26
CGAGT	2514425	2.1463652	40.65683	33
GTCGC	493755	2.141581	11.41992	3
GAGGT	24367315	2.1353076	22.304482	40
AGGAG	9484855	2.0887618	9.493684	38
GCGTT	6114070	2.0767767	26.483095	16
GGTCG	4544315	2.0233886	25.395506	42
TACGC	242460	2.0161235	10.78774	13
ACGGA	939150	2.0146766	10.118734	30
ATTTT	50325225	1.9578714	7.6368256	25
CGAGC	178680	1.947621	8.453393	32
TTTAC	3874150	1.9245911	29.127584	13
TAGTT	37493695	1.9120888	9.941023	29
GACGC	175235	1.9100702	13.050133	5
GCGGT	4260770	1.8971381	25.115347	6
AGCGC	172865	1.884237	10.302331	35
TAGAG	11013060	1.8501835	9.309363	24
GCGTA	2125070	1.8140037	23.42681	4
AATTT	18538130	1.8124657	17.090803	24
ATCGT	2777925	1.8089808	13.406604	39
CGTCT	545700	1.8056165	51.2109	16
TCGTA	2772215	1.8052623	14.904096	45
ACGGC	165000	1.7985083	8.847765	12
AGGTA	10697030	1.7970908	28.848663	47
TTAGT	35112285	1.7906429	15.0738735	28
GCGAC	163560	1.7828122	20.07717	23
AAACG	431935	1.7764082	8.787899	7
TAGTA	13477465	1.727283	14.057717	29
AGTTA	13467785	1.7260425	21.976706	30
GAGCG	1540405	1.7236588	10.898615	28
GGAAA	4080155	1.7226177	11.481969	2

GGAGA	7767425	1.710548	10.750667	2
TGAGA	10173140	1.7090776	5.9657264	41
TATCG	2596535	1.6908598	13.759238	38
AGTTT	33053450	1.6856469	8.146776	26
TACGG	1967750	1.679712	12.028769	5
GGACG	1500195	1.6786652	15.927561	2
AGCGG	1500160	1.6786257	7.0362525	6
GCGTC	386070	1.6745151	11.236934	40
TAGCG	1958780	1.6720549	5.349673	10
GAAAA	2061940	1.6689461	5.058467	3
AACGC	78125	1.6325728	6.9077177	23
GTAGA	9714320	1.6319963	9.018134	23
AGGTC	1906200	1.6271718	43.08583	41
GTCGT	4784485	1.6251543	11.47998	3
TGGCG	3639080	1.6203262	34.430252	10
AGTCG	1894405	1.617103	13.654458	22
CGTAC	193870	1.6120839	8.730606	13
CGATT	2459210	1.6014341	17.875065	11
TCGAC	192195	1.5981557	7.531116	23
TATTT	41046550	1.5968903	5.393748	32
TTGAG	23693055	1.5838803	14.069898	44
TGGGA	18065210	1.583054	13.405925	37
GCGTG	3509180	1.5624874	33.4047	4
TAGGA	9241845	1.5526211	7.4612727	37
AGCGT	1811670	1.5464787	8.867102	29
CGTGG	3470180	1.5451223	33.17875	5
GGGAA	6951965	1.5309669	13.6892395	2
TTCGG	4447425	1.5106645	23.044563	35
TCGAA	921775	1.5084947	5.0651646	32
GTAGT	22202595	1.4842429	8.896231	36
CGAAA	360235	1.4815296	5.879522	32
AGGTT	22152530	1.4808962	14.34922	41
GGTTT	54596720	1.4523202	9.156781	2
TTATT	37308455	1.4514621	7.5352407	32
GTACG	1697975	1.4494263	11.814981	4
AACGG	673090	1.4439213	8.6853695	29
TAATT	14734880	1.4406234	16.829117	23
ACGAG	667500	1.4319296	5.0092196	32
CGAAC	68165	1.4244393	5.8865213	29
AAGTA	4341120	1.3981783	10.667357	34
TATAG	10877710	1.3940964	16.127554	47
ACGGT	1627770	1.3894981	11.585777	6
GAACG	646410	1.386687	8.542739	28
TTTAA	14093425	1.3779087	7.9287605	5
CGTAT	2105815	1.3713038	11.29493	46
GTTTA	26702530	1.3617651	7.7944393	4
TTATA	13918340	1.3607907	12.644573	46
TTAAG	10593250	1.3576398	9.52454	6
GGAAT	8073945	1.356415	9.592964	2
GCGAT	1576335	1.3455921	21.094059	10
AGATA	4125885	1.3288559	5.1065207	26
CACGC	12500	1.3272437	5.787429	12
GGAGT	15094305	1.3227137	10.675694	2
GGCGA	1180735	1.3212007	9.030464	2
GGTAG	14821080	1.2987709	7.7011223	2
GACGT	1512700	1.2912718	6.1327934	3
TGGAA	7643415	1.2840862	9.330512	1
GAGTA	7636175	1.28287	15.1134205	34
TTGTA	25066670	1.2783402	14.397697	20
TATTC	2568755	1.2761	31.949938	33
TCGGG	2862345	1.2744794	28.758507	36
GGTTA	18927045	1.2652726	15.8947115	2
TAAGC	761310	1.2458922	36.58315	7
GGGTT	35585620	1.2408571	13.68426	2
ATTAT	12667080	1.2384554	12.60177	45
GTTAA	9510920	1.2189274	19.154284	3
CGTAA	744750	1.2187916	10.478325	21
TCTCG	367375	1.2155733	52.89224	43
TTTGT	59814450	1.2138094	6.809663	19
CTCGT	364275	1.205316	52.916706	44
CGCAC	11275	1.1971738	5.1637836	12
GGGGA	10352260	1.1891575	10.251393	2
GGGAT	13516485	1.1844494	10.642511	42
TGAGG	13407445	1.1748941	16.09551	45
TCGTG	3427170	1.1641127	6.1741467	40
GGTGG	25370765	1.1596642	10.615422	8
GATTA	9047565	1.1595434	15.799444	44
GTAAT	9013440	1.1551698	20.616144	22
TCGAT	1771830	1.1538131	7.1666403	11
AAAAC	145125	1.1442493	14.693288	6
AGTCA	689745	1.1287752	25.61073	28
TTTTC	5648960	1.1166704	10.581966	29
GGATT	16479385	1.1016465	8.451455	43
TAGGC	1288730	1.1000866	9.250353	13
CGTGA	1265355	1.0801332	8.641686	26
AGTAT	8417685	1.0788175	13.126411	30
GTGGC	2417440	1.0763823	32.59813	9
TGGAG	12278275	1.0759449	10.571145	1
ACAAT	342540	1.0746925	48.761463	39
GTATT	20904810	1.0660952	5.7113676	31
GTTAT	20836045	1.0625885	8.998592	31
CGTGC	243360	1.055534	5.467009	13
TGTAG	15774135	1.0545006	8.50109	21
GGGGT	22896005	1.0465462	8.463263	2
GTTGA	15645905	1.0459285	13.040189	43
GGGTA	11915535	1.0441581	14.448267	2
TGTAA	8072120	1.0345296	19.92565	21
AGTAA	3203905	1.0319066	6.6188545	9
ATTTC	2071360	1.0290053	6.039959	22
AGTTG	15124990	1.0111053	8.94536	38
TTAAT	10246735	1.0018193	13.172488	4
CGATC	119860	0.9966698	6.3473473	44
GGTTG	27984210	0.97579885	6.8355303	42
CGTGT	2844090	0.966057	5.9274077	41
TAAGT	7521385	0.96394694	6.015049	7
AAGGC	446795	0.95847034	13.384424	46
TTATC	1890995	0.939404	10.239927	37

AAGAC	227255	0.9346259	8.83136	32
GGAGC	832825	0.93190175	10.046063	27
GTTTG	34962260	0.9300265	6.6812487	18
GTGGT	26437285	0.92185813	7.337185	9
TGCGG	2063990	0.9190062	5.202192	5
GGATA	5431335	0.9124591	7.383967	2
TGGGG	19954665	0.9121014	9.136963	1
TAGAC	555110	0.9084436	10.34641	25
TGGTT	33828690	0.89987254	7.9213157	1
AGTGA	5351780	0.8990939	5.277644	18
ATTAC	716805	0.8948876	5.05957	29
TTGGG	25275555	0.881349	5.8430877	36
GGGTG	18442075	0.8429629	7.8501415	2
TGAAC	513690	0.8406592	26.172129	20
GAAGC	389850	0.8363112	11.521861	4
GGTAT	12281860	0.82104206	6.194426	2
TGGTG	22406315	0.7812997	6.482976	1
GGTAC	910400	0.77713615	11.905613	3
TGGGT	22147355	0.7722699	9.437477	1
GGAAC	359535	0.7712791	7.468219	27
GGGGG	12586545	0.7541481	5.76966	2
GGTAA	4483755	0.75326645	6.0962443	2
CTGAA	436830	0.714877	25.877995	19
GAGTC	797205	0.68051064	12.584514	21
TGGTA	10008600	0.6690747	5.69448	1
GAACT	379400	0.6208923	25.839113	21
ACAGT	350525	0.573638	25.651865	32
GTCAA	347545	0.5687611	25.500378	35
GATTC	814675	0.53051513	6.282182	29
TGGAT	7804010	0.5216979	5.4682593	1
CGGCC	9400	0.5206123	5.62365	1
TGGGC	1097340	0.48859844	5.289749	13
CACAT	28970	0.46182725	6.550552	12
AATCT	366320	0.4573283	19.885368	41
TCTGA	480830	0.31311578	10.333169	18
GTCTG	527610	0.1792142	5.3980794	17

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	256594	0.33163478758114223	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATG	232152	0.300044736847071	TruSeq Adapter, Index 8 (97CGGGCGC
197949	0.25583908652150683	No Hit	
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	145897	0.1885645050302264	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	118764	0.15349647268559194	No Hit
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	97972	0.12662386263474465	No Hit
CGGGCGCGGTGGCGGGCGTTTGTAGTTTTAGTTATTCGGGAGGTTGAGGTA	84587	0.10932442604708635	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	84333	0.10899614387351406	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	82443	0.10655341431425563	No Hit