FASTQ QC Report

Report Date	12-21-16
Run ID	161219_D00796_0155_ACAC53ANXX
Project ID	EC-EL-4039
Sample	Sample_OD19_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.

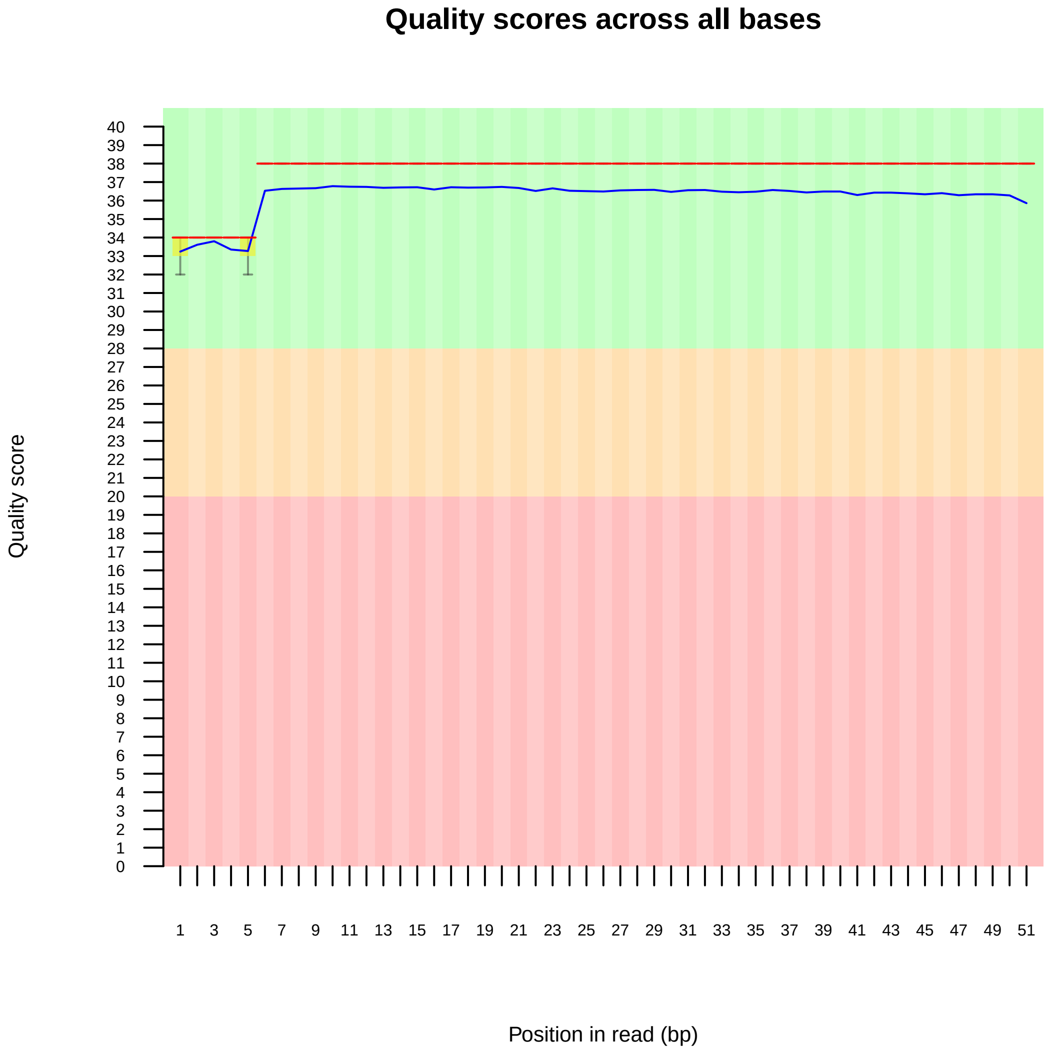
Because the duplication detection requires an exact sequence match over the whole length of the sequence. Reads over 75bp in length are truncated to 50bp for the purposes of this analysis.

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

1 Sequence Duplication

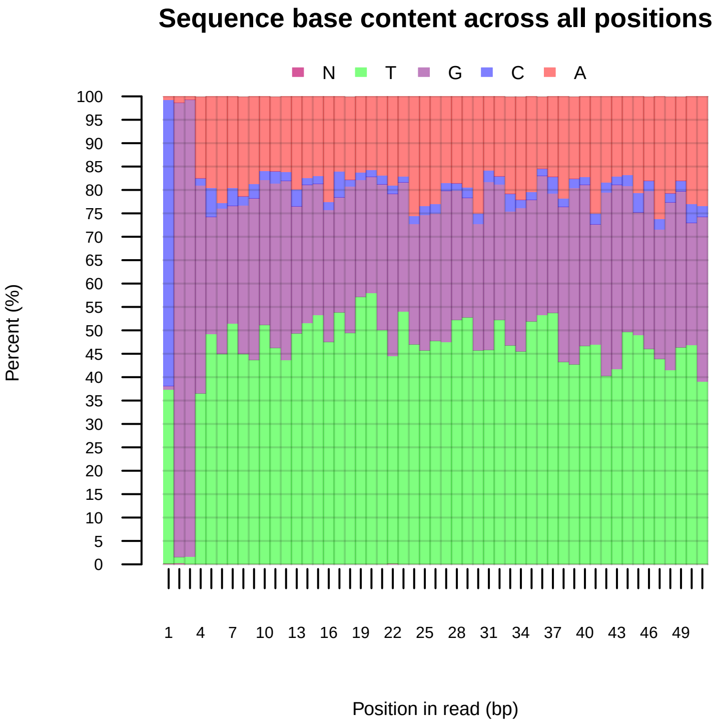
- Estimated Duplication rate 81.6730%

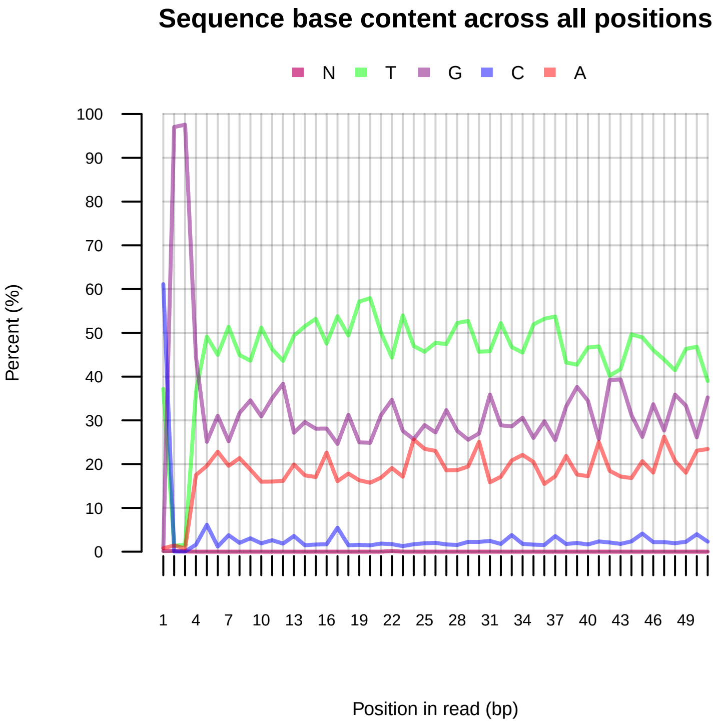
2 Per base sequence quality



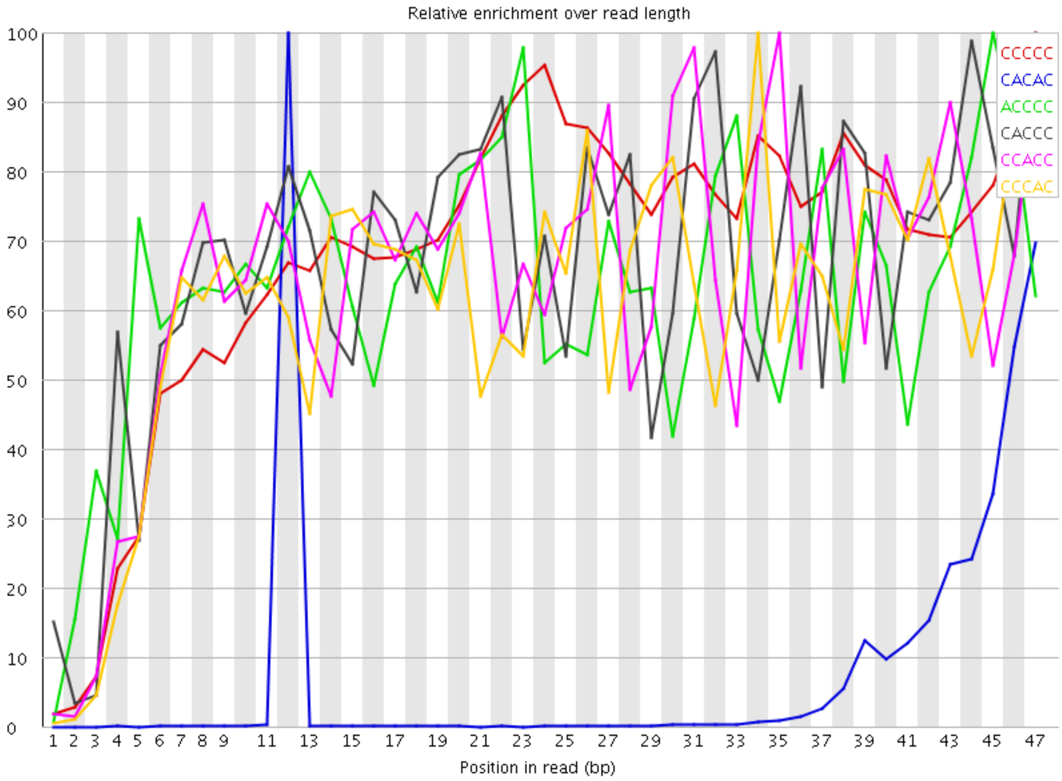
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	184205	1243.4246	1822.2307	47
CACAC	484125	109.47494	1363.1268	12
ACCCC	50390	62.256107	98.111336	45
CACCC	47425	58.59289	87.951965	47
CCACC	45535	56.257828	87.369606	35
CCCAC	45385	56.072502	91.72357	34
CCCCA	45005	55.603024	79.53247	28
CGGGC	4259060	31.30859	1143.655	1
AGCAC	670380	15.596321	144.99832	10
GCCCC	22280	15.473091	34.590965	45
CCGCC	21525	14.948756	27.900652	27
CCCGC	21500	14.931396	31.980358	47
CCCCG	21465	14.907089	31.16451	46
CGCCC	21220	14.736938	27.411665	44
CGCGG	1923715	14.141337	351.6594	5
GCGCG	1825035	13.415935	349.50403	4
GCACA	572010	13.307754	144.72482	11
ACTCC	130985	11.877478	523.0755	23
CTCCA	128810	11.680251	520.64575	24
CGCGC	156385	11.1738	61.22064	13
CGGCG	1516875	11.150634	313.14606	1
CGGAA	4368785	10.456965	213.41556	1
GGCGC	1393815	10.246013	350.96756	3
ACACG	388205	9.031549	141.10081	13
CGGGA	6003325	8.077208	246.41801	1
CGGGT	12370290	6.6741304	264.3573	1
TCGCG	1264225	6.629721	25.529238	30
AGACG	2612495	6.2531734	69.684425	27
CGGAG	4425265	5.953997	178.54903	1
ACGCG	425185	5.5603657	15.039535	4
AGATC	3244950	5.5408216	30.121181	43
CGCGT	1032495	5.414506	23.114428	31
CGGTT	13483150	5.1895213	182.17857	1
CGGGG	6810885	5.1510673	126.91212	1
CGGAC	393435	5.1451545	174.06718	1
CGTCG	961980	5.0447183	23.370195	41
TCCCC	10175	5.041008	18.168146	3
GGGCG	6552660	4.955772	123.107414	2
AGGCG	3674260	4.943553	78.36776	47
TCGAG	4989685	4.789207	64.87576	44
CGCGA	364495	4.7666907	20.82799	5
AAAAA	3437010	4.765372	14.326085	31
CGGTC	878375	4.6062856	170.54877	1

AACCC	20195	4.5666842	7.558195	2
GAGAC	1877415	4.493713	66.777115	26
CTCCC	9045	4.4811716	13.269181	24
CCCCT	8630	4.2755685	7.332971	38
CACGT	449475	4.193263	56.72842	14
CGACG	318560	4.165975	32.638405	24
CCTCC	8400	4.1616187	7.6821575	28
ACGTC	434865	4.0569625	57.008965	15
CCCTC	8130	4.0278525	7.3426957	22
CAACC	17615	3.983271	6.853354	31
CGAGG	2947920	3.966295	85.9669	45
ACCAC	17425	3.9403062	7.0660014	45
ACACC	17340	3.9210851	5.8440666	47
TTACG	5595775	3.8315265	51.93683	14
CCAAC	16635	3.761664	6.6939735	30
CCCAA	16450	3.7198303	5.843946	15
CGGTA	3864875	3.7095902	131.53853	1
GACGG	2721515	3.6616771	38.727093	28
ACCCA	16065	3.6327703	6.1627064	33
CACCA	16005	3.6192026	5.4189453	9
CCACA	15855	3.585283	6.215958	46
GGCGG	4701560	3.5557861	38.321293	11
GATCG	3670400	3.5229287	17.825657	44
ACGGG	2604075	3.5036669	38.971123	29
TACGT	5081435	3.4793487	53.244198	15
ACGTT	5057505	3.4629633	54.79624	16
CGTTT	12175950	3.3431811	38.48887	17
GGCGT	6184105	3.336504	53.313118	3
CGGAT	3470575	3.3311324	116.25771	1
AAGCG	1307350	3.1292255	62.285267	8
GTCGA	3157465	3.0306027	64.275154	43
AGAGA	6839790	2.9964526	23.247366	25
AGCGA	1237550	2.9621549	62.98201	9
GAGGC	2157420	2.9027126	63.71045	46
CGAGA	1209730	2.895566	34.770233	25
AGAGC	1202710	2.8787632	17.479107	47
GCGGC	380780	2.7991352	10.329514	33
TTTCG	10107380	2.7752085	15.903147	30
ATCGC	292820	2.7317898	34.92868	29
GGAGG	19471865	2.695385	34.410538	39
TCGGA	2783060	2.6712408	16.521824	46
ATCGG	2779015	2.6673584	16.048935	45
TTCGA	3865360	2.6466806	35.86457	31
CGGTG	4845580	2.614331	49.130096	1
GCGGG	3439530	2.6013138	39.12919	12
TTTTT	178315725	2.563497	5.5083194	16
CGTTA	3624665	2.4818726	31.82527	9
GGGAG	17869055	2.4735167	30.642344	38
CGTTC	638290	2.3878665	25.81821	33
TTCGC	636835	2.3824232	7.013764	33
TCGTT	8553725	2.3486176	5.5465236	4
GGTCG	4302410	2.321275	37.48204	42
GCGGA	1718900	2.3127036	23.835644	7
AGTAG	13117065	2.304341	23.54969	35
CGTAG	2383095	2.287346	26.631336	5
AACTC	137475	2.2816339	98.51581	22
GGAAG	9257845	2.279811	12.417083	2
CGAGT	2374245	2.2788515	46.196175	33
TTTTA	62971615	2.257577	11.615717	26
GAGGT	22530540	2.2248769	24.939156	40
TTTAG	44224370	2.222482	15.312642	27
TTCGT	7980470	2.1912177	5.3206816	35
GCGGT	4036030	2.1775553	33.774914	6
AAACG	498515	2.1227505	14.798067	7
ATTCG	3090345	2.116014	34.66305	34
ACGGA	876875	2.0988562	8.162409	30
GACGC	159880	2.090834	15.100344	3
AGGTC	2160780	2.0739632	61.330067	41
AGGAG	8404110	2.0695727	9.4510565	38
GAAGA	4695230	2.0569394	6.8375487	46
TTTAC	4206630	2.0547879	36.207684	13
GCGTT	5290495	2.0362554	22.95214	16
TAAAA	3625100	2.0154936	5.8734374	30
ATTTT	56036700	2.0089552	8.168175	25
ACGGC	153165	2.0030184	12.209828	12
GTCGC	380355	1.9946193	10.426573	3
GAGAT	11272125	1.980231	8.915098	26
TACGC	211855	1.9764475	9.571633	13
TCGTC	522335	1.9540745	8.882599	40
AAAAT	3508205	1.950502	5.425794	32
TAGAG	11075120	1.9456222	10.250941	24
GAGCA	809580	1.9377813	15.271025	9
CGAGC	146930	1.9214801	6.166363	32
TACGG	1986355	1.9065462	16.737518	5
AGCGC	145105	1.8976136	11.60456	35
GCGTA	1954825	1.876283	26.2773	4
AATTT	20980295	1.8757023	18.124187	24
AAGAG	4258095	1.8654343	6.874194	47
TAGTA	14483820	1.8151573	17.856224	29
ATCGT	2646570	1.8121535	14.651092	39
TAGTT	36032415	1.8107979	8.388958	25
GGAAA	4129315	1.809017	13.583408	2
GAAAA	2311995	1.8018855	5.542944	3
GGACG	1334825	1.7959476	18.65812	2
AGCGG	1330910	1.7906802	6.041332	6
AAATA	3210150	1.7847886	5.138686	33
TTAGT	35416185	1.7798296	14.697046	28
GAGCG	1319360	1.7751402	10.114594	28
AGGTA	9994735	1.7558255	26.280571	47
GGAGA	7070720	1.7412157	11.467973	2
TGGGA	17446085	1.7227905	17.12011	37
CGTGG	3166705	1.7085292	33.924416	5
GCGAC	130460	1.7060933	21.215157	23
GTAGA	9681875	1.7008638	9.930147	23
TATCG	2476855	1.6959466	15.026139	38
AACGC	71675	1.6675112	11.412698	11
CGAAA	383640	1.6335957	5.5822744	32

GCGTG	3009275	1.6235912	34.101494	4
CGATT	2370515	1.6231337	20.525566	11
AGTCG	1690025	1.6221223	13.843526	22
AGTTT	32191200	1.6177588	8.578634	26
TCGTA	2350930	1.6097234	7.6460156	45
TATTT	44634800	1.6001891	6.296993	32
TGGCG	2958605	1.5962534	31.758732	10
ACGGT	1651915	1.5855435	16.120586	6
GCGTC	302200	1.5847667	10.481321	40
AGTTA	12596270	1.5786037	17.000895	30
GGGAA	6387700	1.5730172	14.782901	2
AGCGT	1636670	1.570911	7.8550706	29
GTACG	1633755	1.5681132	16.461056	4
AACGG	653620	1.5644811	8.729272	8
TAGGA	8741630	1.5356861	7.0181484	37
GTCGT	3987420	1.5347157	10.072371	3
ACGAG	638350	1.5279315	5.537954	32
TCGAA	891155	1.5216663	5.0988007	32
GGTTT	53455740	1.5100638	9.966727	2
GCGAT	1564370	1.501516	26.350655	10
TAATT	16783275	1.5004759	17.877588	23
CGTAC	160190	1.494452	8.5634	13
AGGTT	20934450	1.4747455	14.013242	41
GGCGA	1095670	1.4741753	10.152944	2
AAGTA	4656685	1.4553337	12.10391	34
TTCGG	3748685	1.442829	19.98898	35
TATAG	11487815	1.4396887	17.254965	47
TTAAG	11480560	1.4387795	10.768581	6
TTTAA	16057360	1.4355767	8.439387	5
GTAGT	20238960	1.4257512	10.051052	36
TTGAG	20213560	1.4239619	12.641126	44
TTATT	39272415	1.4079438	5.7636976	32
TAAGC	824550	1.4079367	43.029408	7
TCGAC	150485	1.4039116	6.9962707	23
TTATA	15655115	1.3996152	12.734818	46
AAAAC	183895	1.3930488	24.227434	6
GTTTA	27187860	1.3663176	8.40975	4
AGATA	4329880	1.3531988	5.6703677	26
GGTTA	19153085	1.3492558	18.001602	2
GAACG	558520	1.3368533	6.680073	28
GGTAG	13499445	1.3330618	7.838295	2
GACGT	1381785	1.326267	6.29548	3
GGGTT	33458855	1.3249239	15.770338	2
GGAGT	13373785	1.320653	11.117645	2
GTTAA	10335465	1.2952728	21.26065	3
TCACG	138710	1.2940598	55.976627	30
GAGTA	7342305	1.2898598	16.6074	34
GGAAT	7325065	1.2868311	10.177437	2
GGGAT	12818655	1.2658342	12.918814	42
ATTAT	14156915	1.2656714	12.588579	45
TCGGG	2336045	1.260364	26.575819	36
CGTAT	1813975	1.2420609	5.1727934	46
GGGGA	8966495	1.2411835	10.882267	2
CAGTC	132555	1.2366382	55.94813	27
GTCAC	132525	1.2363584	56.11033	29
TCCAG	132480	1.2359384	55.457165	25
TTGTA	24536885	1.2330937	14.058562	20
GTAAT	9797370	1.2278371	23.25954	22
CCAGT	130600	1.2183995	55.433064	26
GATTA	9594135	1.2023668	16.86176	44
CGAAC	51585	1.2001197	8.882022	9
TCGTG	3088790	1.1888425	7.1357026	40
GGTGG	21361235	1.185728	12.3332815	8
TGGAA	6737980	1.1836951	9.191905	1
CGTAA	690635	1.1792741	8.660579	21
AAGGC	483255	1.1567017	21.589907	46
AGTAT	9202355	1.1532677	16.97195	30
TGAGG	11603060	1.145795	15.7091055	45
TTTGT	56839775	1.1454455	6.659823	19
GGATT	16165360	1.1387829	9.628536	43
GGGGT	20475140	1.1365423	9.048963	2
TAGGC	1165275	1.1184559	7.9094157	13
GGGTA	11235430	1.1094918	16.287415	2
TCGAT	1603065	1.0976471	6.051031	11
AGTAA	3487340	1.0898833	7.7738566	9
GTATT	21603735	1.0856892	7.2960005	31
GTGGC	2011485	1.0852547	30.004414	9
TGTAA	8580825	1.0753758	22.731922	21
TTTTC	5415135	1.0606874	10.667597	29
CGTGA	1099850	1.0556597	7.7357817	26
TTAAT	11734670	1.0491152	13.597292	4
TGGAG	10563300	1.0431195	10.024198	1
TATTC	2130260	1.0405557	23.570818	33
ACGTG	1072020	1.0289478	6.894891	32
CGTGT	2631775	1.0129423	6.8155136	41
TAAGT	7916680	0.992143	6.9313393	7
AGTTG	14040180	0.9890727	10.082351	38
GTTAT	19629465	0.98647285	7.0400996	31
TGCGG	1809115	0.97607005	7.5889263	5
TTGGG	24492265	0.9698594	7.4227996	36
TGTAG	13745280	0.9682982	6.9070287	21
TAGAC	565410	0.96544963	11.411925	25
TTATC	1967505	0.96105576	10.501765	37
TAAGG	5450655	0.9575441	6.1752276	45
ATCTC	143695	0.95633334	42.191814	42
GGTTG	24039035	0.95191216	6.743031	42
TGGGG	17071280	0.9475996	8.961697	1
GTTGA	13405025	0.94432867	11.915259	43
CGTCT	250955	0.9388319	22.527542	16
AAGAC	219265	0.93366265	8.449442	32
GGATA	5289500	0.92923313	7.958642	2
GGAGC	689225	0.92732155	9.177356	27
ATTAC	758030	0.92336553	5.25632	29
GTGGT	22774955	0.9018564	8.848215	9
TGGTT	31590560	0.89239746	7.4009285	1
GGTAC	929260	0.8919237	16.621132	3
TTTGG	31265510	0.88321507	5.6984296	35

GTTTG	31116620	0.8790092	6.8562994	18
AGTGA	4988295	0.8763189	5.2991195	18
GGGGG	11136040	0.86649966	6.380646	2
GGGTG	15592410	0.86550975	8.828897	2
CCTTA	129710	0.863259	39.117996	38
CGATC	91690	0.85539854	5.1333113	44
GAAGC	346770	0.83001614	8.888955	4
GGTAT	11668410	0.8219913	6.424019	2
GTGCG	1479410	0.7981847	7.1672077	4
GGTAA	4441195	0.78020716	6.6689973	2
AGTGG	7771050	0.7673863	5.7745066	8
TGGGT	19292225	0.7639452	9.384225	1
GTTGG	18665890	0.73914313	5.7361393	39
GGAAC	304395	0.72858894	5.9264154	2
TGGTG	18128380	0.71785855	5.8399267	7
TGGCC	131130	0.6876587	30.80856	35
GGCCT	130270	0.6831488	30.734634	36
GAGTC	697400	0.66937953	12.628726	21
TGGTA	9004045	0.63429785	5.2710576	1
TCTCG	154095	0.5764751	23.714268	43
CTCGT	152520	0.5705829	23.720669	44
TGGGC	980705	0.52911884	5.050623	13
TGGAT	7150305	0.5037095	5.1725097	1
GCCTT	130720	0.48902836	21.96064	37
GATTC	712390	0.48778608	5.131829	29
TGAAC	265670	0.4536372	10.877141	20
GGTGC	793410	0.4280677	7.25968	3
CTGAA	188455	0.32179096	10.582295	19
GAACT	156660	0.26750028	10.576671	21
AGTCA	144705	0.24708688	10.525352	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	231837	0.3124624520304156	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	198732	0.26784459778598135	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	151617	0.20434451614494464	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	87267	0.11761565583292694	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	83325	0.11230275501940752	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGGCCTTATCTCGTATG	81930	0.11042261888676937	TruSeq Adapter, Index 1 (97CGGGTTTAC
81648	0.1100425483567307	No Hit	
CGGGATGGTTTCGATTTTTTGATTTCGTGATTCGTTCGTTTCGGTTTTTTA	74302	0.10014184582600681	No Hit