FASTQ QC Report

Report Date	10-02-16
Run ID	160930_D00796_0121_AC9MR4ANXX
Project ID	EC-EL-3883
Sample	Sample_YD9_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`

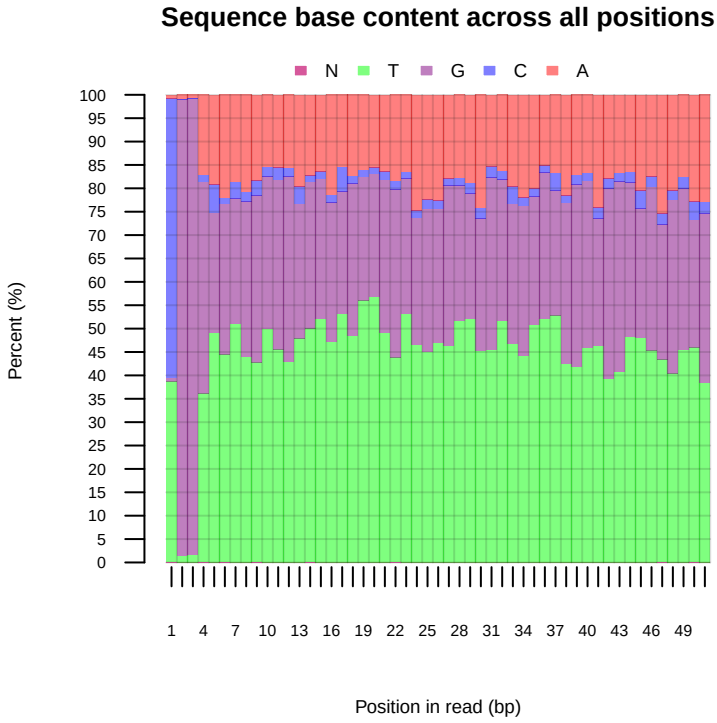
1 Sequence Duplication

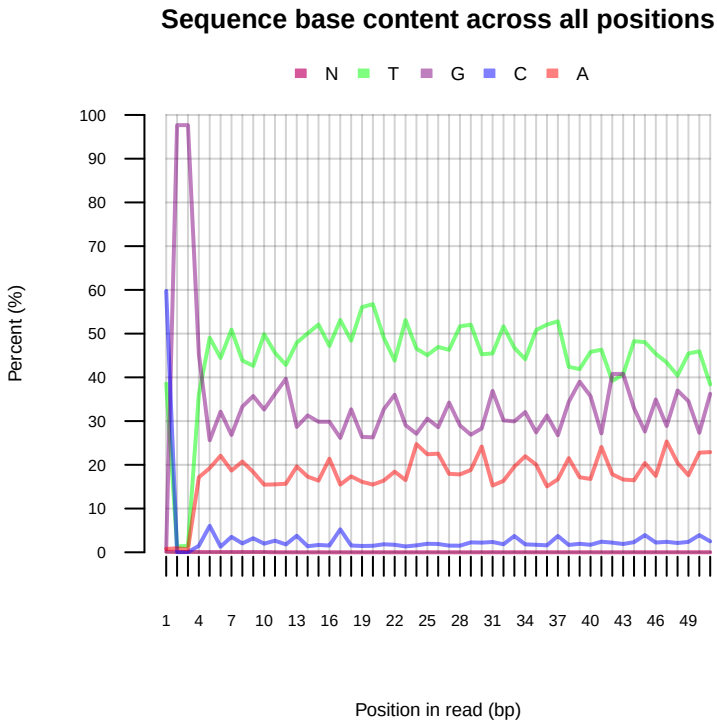
- Estimated Duplication rate 73.7639%

2 Per base sequence quality

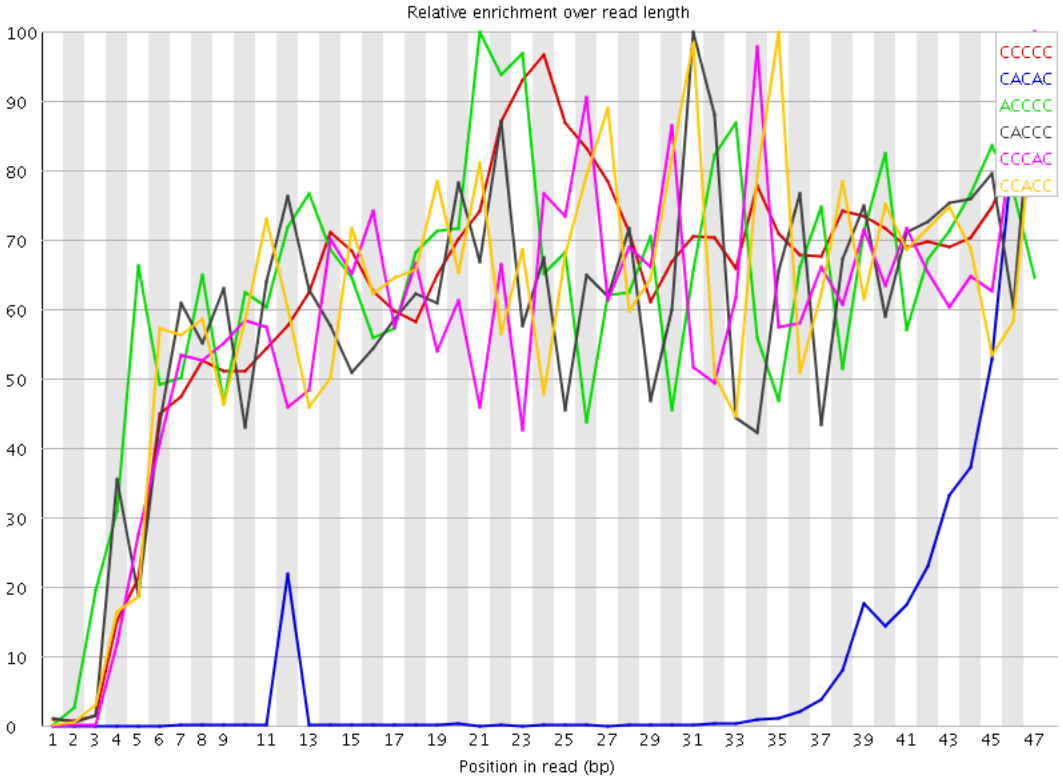
Background colors	Green - calls of very good quality Orange - calls of reasonable quality Red - calls of poor quality
Yellow boxes	Inter-quartile range
Upper and lower whiskers	Maximum and minimum quality excluding outliers
Red line	Median quality
Blue line	Mean quality

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	121450	803.7796	1273.6152	47
CACAC	568595	132.65565	1458.8103	47
ACCCC	43435	53.972298	85.837	21
CACCC	42515	52.829105	89.34323	31
CCCAC	40860	50.77261	87.3006	47
CCACC	40360	50.151314	82.92072	35
CCCCA	39005	48.467583	78.24631	24
CGGGC	4364840	27.572926	981.30853	1
AGCAC	827545	19.009588	148.53992	45
GCCCC	26125	17.023706	35.216045	47
CCCGC	25330	16.505665	29.702982	26
CCCCG	25285	16.476341	29.703985	46
CGCCC	25065	16.332985	26.793686	22
GCACA	698930	16.055164	147.9737	46
CCGCC	24010	15.645521	25.26335	31
CGCGC	203295	13.043161	62.531265	13
CGCGG	1984190	12.534234	284.61453	5
GCGCG	1907360	12.048895	282.27283	4
CGGAA	4839405	10.9453945	204.70792	1
CGGCG	1638770	10.352198	280.11584	1
ACACG	432290	9.9301605	120.79698	47
GGCGC	1390080	8.781209	282.89487	3
CGGGA	6346385	7.527185	224.21703	1
TCGCG	1406625	6.6986017	23.090815	30
TCCCC	13470	6.616932	17.207605	5
AGATC	3723975	6.34947	31.568323	43
CGGGT	13193340	6.1861286	241.01753	1
CCTCC	12530	6.155172	11.311468	28
CTCCC	12460	6.1207848	11.195893	24
CCCTC	12280	6.032363	10.041678	24
CCCCT	12255	6.020082	10.849873	38
AGACG	2549600	5.766489	63.225292	27
ACGCG	468450	5.643013	15.612499	14
CGGAG	4583855	5.436721	160.92105	1
CGCGT	1137840	5.418599	21.019682	31
CGTCG	1077475	5.13113	23.67152	41
CGCGA	415145	5.0008936	18.929073	5
CGGAC	414665	4.9951115	163.26903	1
CGGTT	13632605	4.818749	165.32008	1
CGGTC	975235	4.6442447	167.05438	1
CGGGG	7408330	4.6077957	111.49467	1
AACCC	19470	4.5424347	7.2909517	32
GGGCG	7047345	4.3832726	102.150986	2

AAAAA	2828355	4.3674865	12.602641	31
TCGAG	4871490	4.355713	52.04323	44
AGGCG	3652080	4.3315816	61.33907	47
CGACG	353945	4.26367	31.71165	24
CAACC	17885	4.1726475	6.7975674	31
GAGAC	1825055	4.1277695	60.725124	26
ACACC	17525	4.088658	6.742546	21
ACCAC	17390	4.057162	6.6880226	46
ACCCA	16965	3.9580076	6.6331463	33
ACGTC	427865	3.8854904	22.754858	47
CCAAC	16540	3.8588533	6.907283	34
CACCA	16535	3.8576868	6.907121	28
CCACA	16380	3.8215249	6.0301847	46
CCCAA	16060	3.7468672	6.852214	15
GATCG	4143490	3.7047913	17.807634	44
TTACG	5352865	3.6080709	43.67738	14
ACTCC	38990	3.5961185	119.1249	23
CGGTA	3959095	3.5399194	122.56202	1
AGAGC	1486175	3.3613162	20.579487	47
CGAGG	2829030	3.3553956	65.64575	45
CTCCA	36230	3.3415582	118.71365	24
TACGT	4896740	3.300622	45.371456	15
CGTTT	12382580	3.2995741	36.07479	17
GACGG	2761335	3.2751055	33.27248	28
ACGTT	4853850	3.271712	47.00514	16
GGCGG	5253515	3.267555	36.237686	11
CACGT	351430	3.1913757	30.44508	47
ACGGG	2654785	3.1487308	33.324516	29
CGGAT	3511885	3.1400585	107.520424	1
GGCGT	6663275	3.1242943	47.50428	3
AAGCG	1341540	3.0341926	60.846714	8
AGAGA	6873030	2.9186316	22.073503	25
AGCGA	1288735	2.914762	61.41334	9
TCGGA	3215535	2.8750854	16.53478	46
ATCGG	3214830	2.8744545	16.200884	45
GCGGC	447545	2.8271656	10.719084	9
TTTCG	10521165	2.8035648	16.17376	30
ATCGC	304005	2.760704	33.777416	29
CGAGA	1186080	2.6825848	31.978178	25
TTCGA	3970795	2.6764936	36.528557	31
GTCGA	2991270	2.6745641	51.611298	43
CGTTC	730840	2.6237319	28.690678	33
TTTTT	171600175	2.5586176	5.670412	16
TTCGC	712465	2.5577652	8.041522	33
CGTTA	3734670	2.517335	33.924656	9
GAGGC	2117800	2.5118353	50.61233	46
GCGGG	4017245	2.4986258	37.102314	12
GGAGG	21122370	2.4666483	28.895853	39
TCGTT	8976405	2.391934	5.90462	4
AGAAA	2890740	2.3408487	5.236307	22
CGGTG	4918350	2.3061292	43.995872	1
GAGCA	1015330	2.2963955	16.67234	47
GGAAG	10245860	2.2816355	11.7319355	2
TTTTA	60229550	2.2716389	11.857782	26
AGTAG	13504775	2.2671301	23.869076	35
GGGAG	19399970	2.2655082	25.422108	38
CGAGT	2516545	2.2501016	45.48906	33
GACGC	185295	2.232089	13.030022	5
TTCGT	8374280	2.2314863	5.574632	35
CGTAG	2493655	2.2296352	23.640888	5
GCGGA	1875830	2.2248445	21.475595	7
TTTAG	44393230	2.2210336	14.957802	27
ATTCG	3280085	2.2109241	37.122257	34
GAAGA	5167485	2.1943722	7.794962	46
GTCGC	439235	2.091716	10.370594	3
GAGGT	23717395	2.087963	21.94835	40
TACGC	228870	2.0783944	10.685767	13
GGTCG	4431565	2.0778842	28.957617	42
ACGGA	907795	2.0531812	8.798381	30
TCGTC	569810	2.0456307	9.407069	40
CGAGC	168520	2.0300152	6.9346375	32
AGGAG	9109350	2.0285478	9.245173	38
TTTAC	3987645	2.0262663	32.012505	13
GCGTT	5729125	2.0250874	23.337662	16
AGCGC	167325	2.01562	9.702915	35
GAGAT	11892555	1.9964769	8.537983	26
ATTTT	52899300	1.9951686	7.875066	25
GCGGT	4224025	1.9805721	27.390388	6
AAGAG	4647145	1.9734101	7.8010454	47
AAACG	453410	1.9555271	12.074743	7
CACGC	15895	1.9446883	8.020632	47
TAAAA	3160235	1.9291899	5.268756	30
CGCAC	15740	1.9257247	7.5319195	47
ACGGC	159035	1.9157575	10.087854	12
TAGAG	11242735	1.8873874	9.672159	24
AATTT	19576970	1.867743	17.630941	24
TAGTT	37314140	1.8668603	8.837528	25
GCGTA	2037880	1.8221161	23.304937	4
GAAAA	2241365	1.8150011	5.6138206	3
ATCGT	2684785	1.8096653	14.014174	39
GGAAA	4226350	1.7947196	12.606001	2
GAGCG	1509280	1.7900949	10.0367565	28
TACGG	1996690	1.7852871	13.789223	5
TTAGT	35532925	1.7777444	14.390977	28
AGGTC	1969625	1.7610878	49.12233	41
AGCGG	1473390	1.747527	5.70891	6
TAGTA	13801190	1.7466139	15.462394	29
GCGAC	144955	1.7461479	20.888163	23
AGGTA	10400900	1.7460634	26.509354	47
GGACG	1469965	1.743465	16.82018	2
AACGC	74820	1.7186949	7.912824	11
GGAGA	7700320	1.7147729	10.805076	2
GCGTC	356215	1.69636	10.659108	40
TAGCG	1896155	1.6953964	5.01211	10
TATCG	2508545	1.6908717	14.397563	38
GTAGA	9865400	1.6561657	9.360546	23
AGTTT	33059270	1.6539853	9.088248	26

CGATT	2435835	1.6418617	20.725363	11
AGTTA	12899565	1.6325084	18.8934	30
TGGGA	18510245	1.629551	14.336199	37
AGTCG	1796980	1.6067218	12.874819	22
CGTAC	176190	1.6000015	9.068393	13
GCACC	13060	1.5978377	6.755729	47
TATTT	42340285	1.596921	5.749033	32
TGGCG	3344290	1.5680797	31.235723	10
CGTGG	3329305	1.5610535	31.196875	5
AGCGT	1736560	1.5526989	7.9638534	29
GTCGT	4379900	1.5481735	10.099833	3
CGAAA	358950	1.5481274	5.2119365	32
GCGTG	3296980	1.5458969	31.366552	4
TCGAA	904390	1.5420076	5.2005434	32
TCGAC	169585	1.5400207	6.7041135	23
GGGAA	6879790	1.5320503	13.87481	2
TAGGA	9091310	1.5262144	7.192638	37
GTACG	1703245	1.5229112	13.554589	4
AACGG	669665	1.5145969	7.694571	29
AAGTA	4672375	1.4957536	12.743484	34
GGTTT	56587120	1.4846451	9.425738	2
TAATT	15557005	1.4842176	17.280249	23
GCGAT	1636525	1.4632553	25.34011	10
ACGGT	1635295	1.4621555	13.146393	6
TTCGG	4126990	1.458777	20.191656	35
TTGAG	21942195	1.4562207	12.708148	44
TATAG	11449640	1.4490128	18.339354	47
AGGTT	21771355	1.4448826	13.508174	41
TTTAA	15122550	1.4427686	8.964228	5
GTAGT	21647415	1.4366573	10.129835	36
TAAGC	838455	1.4295866	43.749832	7
TTAAG	11294100	1.4293284	10.930624	6
TTATT	37731640	1.4230998	6.536276	32
TTATA	14887855	1.4203775	14.116422	46
GGAAT	8289895	1.3916761	10.145136	2
GAACG	611730	1.3835639	7.495771	28
GGCGA	1164130	1.3807265	9.212904	2
GTTTA	27279275	1.3648069	8.749835	4
AGATA	4219400	1.350744	5.429023	26
GACGT	1477040	1.320656	6.163254	3
GAGTA	7835125	1.3153309	17.03176	34
CGAAC	56915	1.307398	6.708853	21
GGTAG	14797830	1.3027284	7.417182	2
TGGAA	7738670	1.2991383	9.195995	1
GGGTT	37239000	1.2960204	14.76396	2
AAAAC	156950	1.2908249	20.410316	6
GGAGT	14639190	1.2887625	10.2233715	2
GGTTA	19386145	1.2865851	16.53758	2
ATTAT	13469085	1.2850195	13.95378	45
TCGGG	2694230	1.2632779	25.57874	36
GTTAA	9871050	1.2492337	20.253399	3
TTGTA	24778775	1.2397046	13.745287	20
GGGAT	13965650	1.2294673	12.418197	42
GGGGA	10395215	1.2139423	10.023294	2
GATTA	9539875	1.2073218	17.837214	44
CGTAA	703605	1.1996641	8.611211	21
GTAAT	9476115	1.1992526	21.539173	22
TCGTG	3351760	1.1847546	6.6885786	40
TTTGT	59763305	1.1820368	6.5201283	19
CGTAT	1737245	1.1709809	5.2212014	13
GGTGG	25166795	1.1618509	10.709384	8
TGAGG	12881560	1.1340294	15.216458	45
TATTC	2225845	1.1310322	26.543745	33
GGATT	16912965	1.1224496	9.641022	43
TCGAT	1660270	1.1190964	6.902436	11
GGGGT	24240375	1.119082	8.149615	2
TTTTC	5548790	1.1146443	11.32627	29
AGTAT	8694405	1.1003231	14.614512	30
AGTAA	3417055	1.0938916	7.908564	9
TAGGC	1217180	1.088309	8.317775	13
GGGTA	12160765	1.070574	14.819212	2
CGTGA	1193280	1.0669397	7.9872756	26
GTATT	21205555	1.0609331	6.271863	31
TGTAA	8367515	1.0589534	20.904531	21
TGGAG	11902720	1.047857	9.629042	1
TTAAT	10794700	1.0298696	13.56947	4
AAGGC	450275	1.0183976	16.449678	46
GTTAT	20335015	1.0173793	7.784336	31
GTGGC	2168245	1.0166525	29.405975	9
CGTGT	2853830	1.0087501	6.382234	41
TAAGT	7951425	1.006295	7.00534	7
AGTTG	15151570	1.0055525	10.103835	38
CGTGC	209370	0.9970577	5.0744085	13
TGTAG	14965870	0.9932283	7.345031	21
ATTTC	1920470	0.9758601	5.534792	22
GTTGA	14632065	0.97107494	11.961366	43
GGTTG	27767470	0.9663849	6.4999423	42
GGGGG	15726585	0.96308744	5.786227	2
TGGGG	20849065	0.9625185	8.489378	1
TGCGG	2045505	0.95910186	6.3310876	5
AAGAC	221150	0.9538052	8.797359	32
TTGGG	27359200	0.9521761	6.3489747	36
CGATC	104485	0.9488401	5.7719035	44
TTATC	1849360	0.9397266	10.478376	37
GGAGC	790215	0.93724144	9.088663	27
TAGAC	547625	0.9337143	11.335847	25
TAAGG	5511850	0.92530835	5.153815	45
GGATA	5479160	0.9198204	7.4836335	2
GTTTG	34317330	0.90036505	6.4458466	18
TGGTT	34231245	0.89810634	7.505487	1
TTTGG	34119145	0.8951654	5.1115656	35
GTGGT	25624405	0.8918003	7.6403875	9
GGGTG	19269220	0.8895835	8.458086	2
AGTGA	5277280	0.8859296	5.0718765	18
GAAGC	370720	0.838466	8.853551	4
GGTAC	925415	0.8274352	13.615625	3
GGTAT	12259505	0.8136171	5.968751	2

GGAAC	349625	0.7907551	6.555667	27
GTGCG	1678245	0.78690004	5.977831	4
TGGGT	22462310	0.7817506	9.310959	1
GGTAA	4615395	0.7748149	6.208761	2
CGGCC	12005	0.7702263	5.178979	1
AACTC	43600	0.7550207	22.899345	22
GTTGG	21679030	0.75449044	5.529692	39
TGGTG	21182630	0.7372142	5.577483	7
CGTCT	205265	0.73690593	7.435914	47
GAGTC	750225	0.6707937	11.739089	21
TGGTA	9896855	0.65681684	5.152734	1
TGGGC	1143640	0.536233	5.0537543	13
GATTC	784740	0.5289499	5.7668633	29
TGGAT	7660170	0.5083766	5.0672693	1
GGTGC	891500	0.4180089	6.043057	3
TCACG	41960	0.38104352	12.004573	30
TCCAG	35615	0.32342386	12.055514	25
CAGTC	34605	0.3142519	12.286092	27
GTCAC	33865	0.3075319	12.352334	29
CCAGT	33605	0.30517086	12.123859	26
ATCTC	38165	0.26127368	9.464846	42
CCTTA	31780	0.21756262	8.496429	38
TCTCG	49450	0.17752658	5.017405	43
CTCGT	48485	0.1740622	5.055383	44
GGCCT	33815	0.16103312	5.978656	36
TGGCC	33140	0.15781866	6.0077634	35

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	234444	0.30229939874598694	No Hit
CGGGCGCGGTGGTTTACGTTTGTAATTTTAGTATTTTGGGAGGTCGAGGCG	180842	0.23318330973717294	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAG	152225	0.19628365824720556	No Hit
CGGGTTTACGTTATTTTTTTGTTTTAGTTTTTTGAGTAGTTGGGATTATAG	90541	0.1167463866077204	No Hit
CGGTTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTTATCGTGTTAGTTA	78007	0.10058465645517993	No Hit
CGGGCGTAGTGGCGGGCGTTTGTAGTTTTAGTTATTTGGGAGGTTGAGGTA	77771	0.10028035070154984	No Hit