



Transcriptome Resequencing Report

2023.03

RAWDATA REPORT



Table of Contents

Order Information	3
-------------------	---

01 Workflow

Experimental Workflow	4
-----------------------	---

02 Raw Data Result

Raw Data Statistics	5
Total Bases	7
GC/AT Content	9
Q20/Q30 (%)	11

03 Deliverables

Download List	15
---------------	----

04 Appendix

FAQ	19
Result File Description	22

Order Information

Client Name	Diana Martinez Alarcon
Client Organization	UMR MARBEC
Order Number	EN00001606
Application	Transcriptome Resequencing
Type of Read	Paired-end
Read Length	101
Library Kit	TruSeq Stranded mRNA LT Sample Prep Kit
Library Protocol	TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E
Type of Sequencer	illumina system

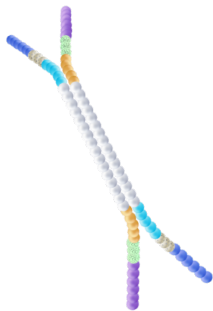
Experimental Workflow

The samples are prepared according to NGS library preparation workflow, and sequenced using Illumina platform. The workflow illustrated below shows the common ligation based method of library preparation. The process may differ based on the library preparation protocol followed.



Sample Preparation

DNA/RNA is first extracted from the sample, and samples which meet quality control standards proceed to library construction.



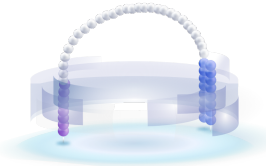
Ligate Adapters

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step which greatly increases the efficiency of the library preparation process.

Final library Construction

Adapter-ligated fragments are then PCR amplified with a PCR primer solution which anneals to the ends of each adapters.

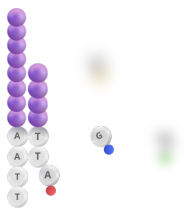
The library templates undergo quality control and quantification process.



Cluster generation using bridge amplification

The library is loaded onto a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters.

Each fragment is then amplified into distinct clonal clusters through bridge amplification. Once cluster generation is complete, the templates are ready for sequencing.



Sequencing by synthesis (SBS) technology

Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4-reversible, terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive sequence regions and homopolymers.



Generation of Raw data

The Illumina sequencer generates raw images utilizing sequencing control software for system control and base calling, through integrated primary analysis software called RTA (Real Time Analysis).

The BCL/cBCL (base call) binary files are converted into FASTQ files using bcl2fastq, which is an Illumina provided package. Adapters are not trimmed away from the reads.

Raw Data Statistics

- The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 58 samples.
For example, in C10_con_GI sample, 109,208,892 reads are produced, and total read bases are 11 Gbp.
The GC content (%) is 45.2% and Q30 is 93.9%.

* Raw Data

Sample ID	Total bases(bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
C10_con_GI	11,030,098,092	109,208,892	45.2	54.8	97.8	93.9
C1_con_MG	10,137,279,302	100,369,102	45.8	54.2	97.7	93.6
C1_poll_GI	10,789,151,886	106,823,286	43.3	56.7	98.0	94.3
C1_poll_MG	10,040,477,468	99,410,668	46.3	53.7	97.9	93.9
C2_con_GI	10,201,498,536	101,004,936	48.1	51.9	97.3	93.1
C2_con_MG	11,205,373,290	110,944,290	45.2	54.8	97.7	93.5
C2_poll_MG	11,036,823,884	109,275,484	47.0	53.0	97.7	93.7
C3_con_MG	9,833,124,670	97,357,670	47.5	52.5	97.7	93.7
C3_poll_MG	8,707,168,590	86,209,590	46.7	53.3	97.6	93.4
C4_con_MG	11,418,222,710	113,051,710	46.0	54.0	97.9	94.1
C4_poll_MG	9,476,362,774	93,825,374	47.6	52.4	97.7	93.8
C6_con_MG	10,604,541,460	104,995,460	48.6	51.4	97.9	94.2
C6_poll_GI	10,473,008,554	103,693,154	44.3	55.7	97.9	94.0
C6_poll_MG	8,115,551,596	80,351,996	45.9	54.1	97.7	93.8
C7_con_GI	10,946,755,922	108,383,722	44.1	55.9	97.8	93.7
C7_con_MG	8,920,309,294	88,319,894	47.6	52.4	97.5	93.5
C7_poll_GI	9,986,354,396	98,874,796	43.1	56.9	97.9	93.9
C7_poll_MG	9,724,776,314	96,284,914	45.6	54.4	97.0	92.6
C8_con_MG	8,166,873,736	80,860,136	47.0	53.0	97.8	93.9
C9_con_GI	10,903,246,536	107,952,936	44.8	55.2	97.6	93.4
C9_con_MG	10,212,817,606	101,117,006	48.4	51.6	97.8	93.8
C9_poll_MG	8,911,840,848	88,236,048	47.7	52.3	97.8	93.9
N1_con_MG	9,536,380,408	94,419,608	47.7	52.3	97.7	93.6
N1_poll_MG	8,299,315,238	82,171,438	47.9	52.1	97.9	94.2
N2_poll_MG	9,121,083,558	90,307,758	47.3	52.7	97.9	94.2
N3_con_MG	9,029,122,250	89,397,250	47.3	52.7	97.7	93.6
N4_con_MG	8,229,018,632	81,475,432	47.2	52.8	97.9	94.1
N4_poll_MG	10,305,598,832	102,035,632	48.1	51.9	97.9	94.1
N5_con_MG	8,162,636,584	80,818,184	47.3	52.7	98.0	94.3
N5_poll_MG	9,699,929,506	96,038,906	46.4	53.6	98.0	94.3
N6_con_GI	11,433,574,104	113,203,704	44.2	55.8	97.7	93.6
N6_con_MG	7,464,183,204	73,902,804	47.1	52.9	97.8	93.8

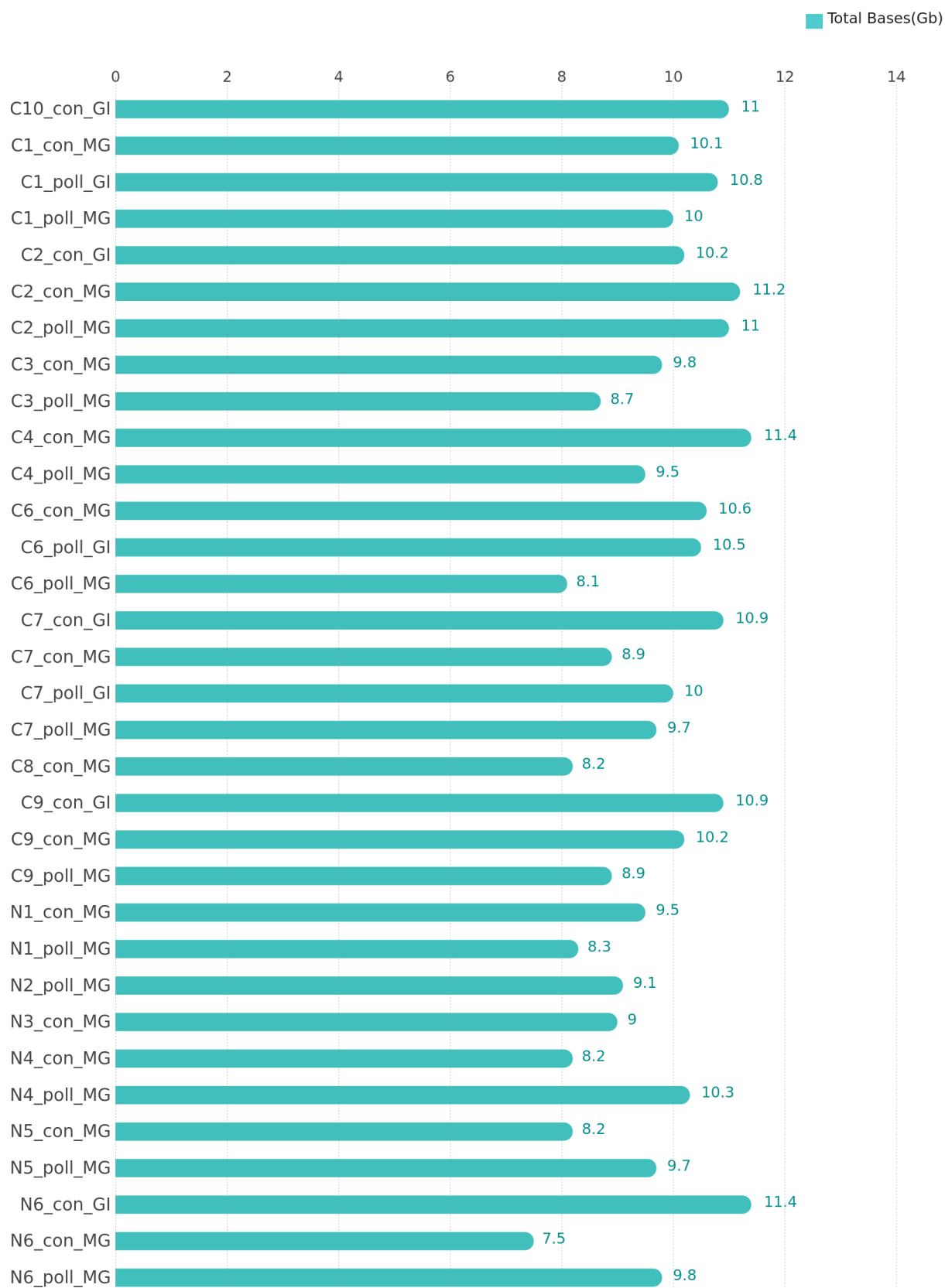
Sample ID	Total bases(bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
N6_poll_MG	9,792,120,084	96,951,684	47.8	52.2	97.7	93.6
N7_con_MG	8,080,336,936	80,003,336	48.4	51.6	97.7	93.6
N7_poll_GI	11,709,726,284	115,937,884	44.7	55.3	97.8	93.7
N7_poll_MG	10,285,066,744	101,832,344	47.7	52.3	97.8	94.0
N8_con_MG	11,269,196,806	111,576,206	48.3	51.7	97.2	92.6
N8_poll_MG	9,639,055,190	95,436,190	48.4	51.6	97.9	94.2
N9_poll_MG	9,557,640,706	94,630,106	48.8	51.2	97.7	93.9
S11_con_GI	10,718,813,264	106,126,864	44.9	55.1	97.8	93.8
S1_poll_MG	11,330,574,102	112,183,902	46.1	53.9	97.9	94.0
S2_poll_GI	8,712,565,424	86,263,024	45.8	54.2	97.7	93.6
S2_poll_MG	10,299,514,996	101,975,396	46.5	53.5	97.9	94.0
S3_con_MG	11,468,728,164	113,551,764	46.4	53.6	98.0	94.2
S3_poll_MG	9,288,132,508	91,961,708	47.6	52.4	97.7	93.6
S4_con_mg	9,137,477,878	90,470,078	46.1	53.9	97.8	93.9
S4_poll_MG	10,026,289,392	99,270,192	48.0	52.0	97.8	93.9
S5_con_MG	8,136,621,408	80,560,608	47.3	52.7	97.9	94.1
S5_poll_MG	9,467,741,010	93,740,010	46.1	53.9	97.9	94.1
S6_con_GI	9,994,143,314	98,951,914	44.8	55.2	97.8	93.8
S6_con_MG	8,392,941,228	83,098,428	46.1	53.9	97.9	94.1
S6_poll_MG	11,407,135,132	112,941,932	47.0	53.0	97.7	93.5
S7_poll_GI	10,659,417,992	105,538,792	42.4	57.6	97.4	92.8
S7_poll_MG	11,701,263,898	115,854,098	48.7	51.3	97.8	94.0
S8_con_MG	11,158,203,866	110,477,266	45.6	54.4	97.8	93.8
S8_poll_GI	10,402,262,498	102,992,698	46.5	53.5	97.9	94.0
S8_poll_MG	10,727,715,202	106,215,002	46.1	53.9	97.7	93.6
S9_con_MG	9,995,893,038	98,969,238	47.1	52.9	97.9	94.1

- **Sample ID** : Sample name.
- **Total bases(bp)** : Total number of bases sequenced.
- **Total reads** : Total number of reads. For illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- **GC(%)** : Ratio of GC content.
- **AT(%)** : Ratio of AT content.
- **Q20(%)** : Ratio of bases that have phred quality score of over 20.
- **Q30(%)** : Ratio of bases that have phred quality score of over 30.

Total Bases

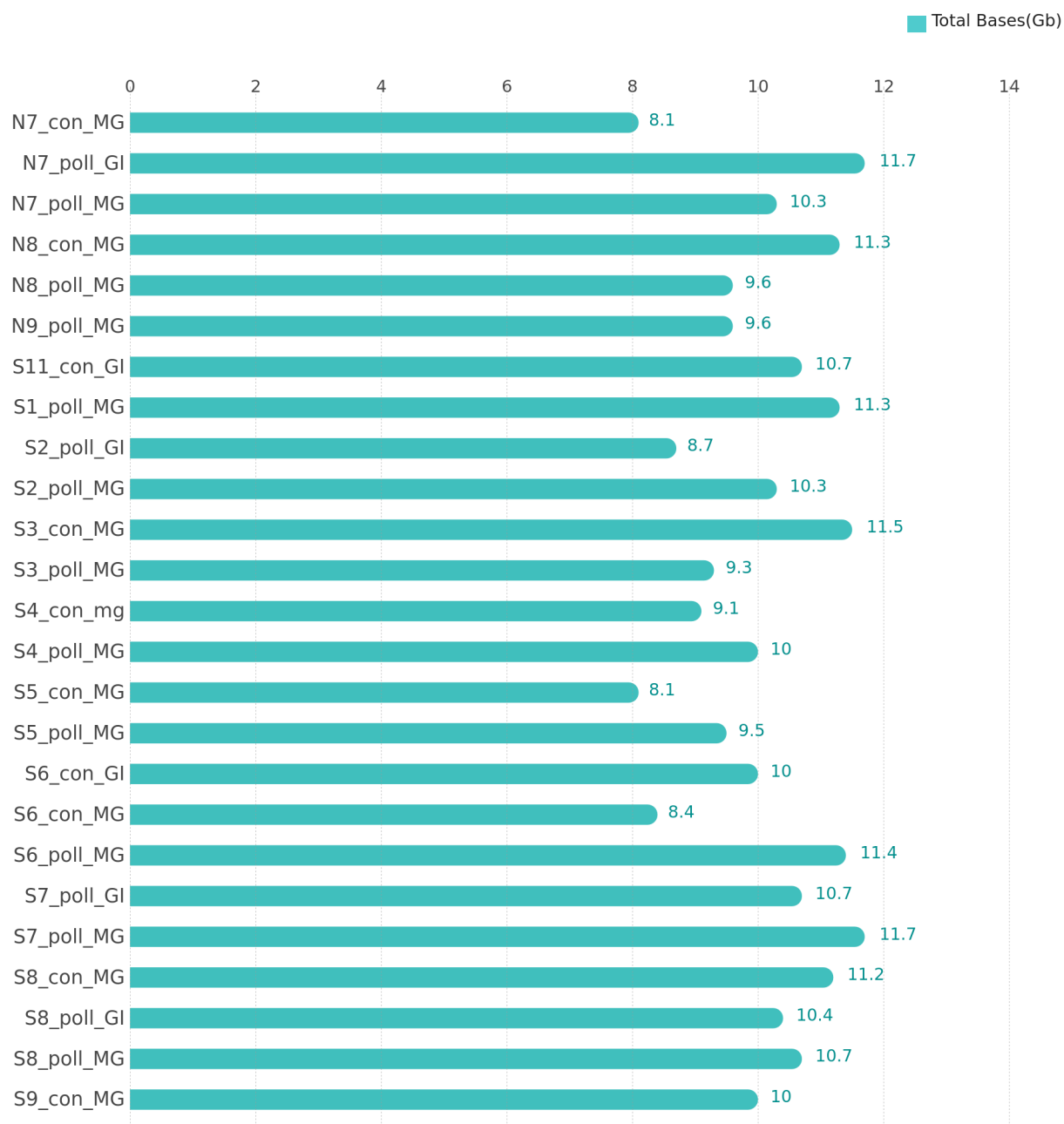
Total number of samples : 58

* Raw Data



Total Bases

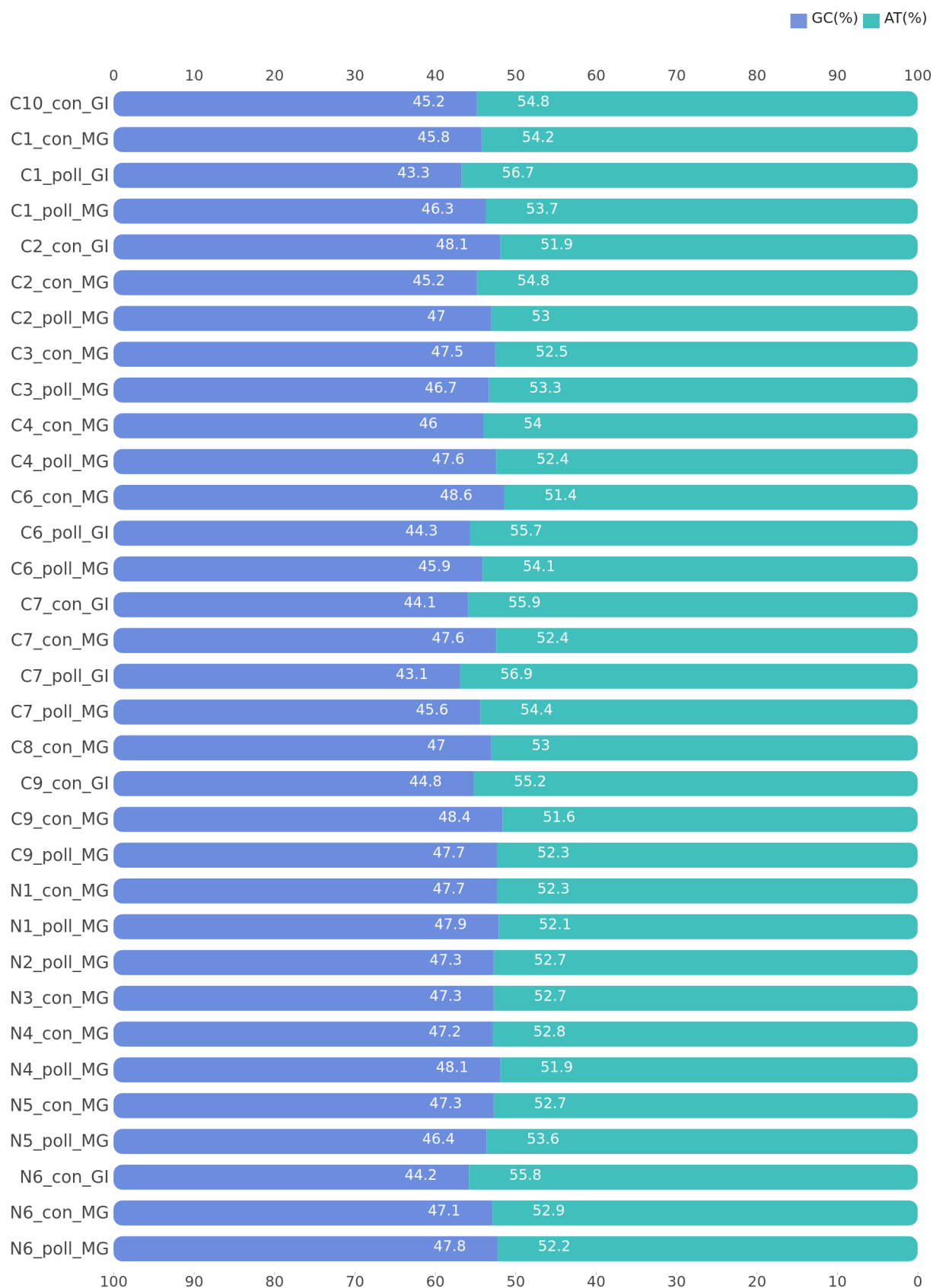
* Raw Data



GC/AT Content

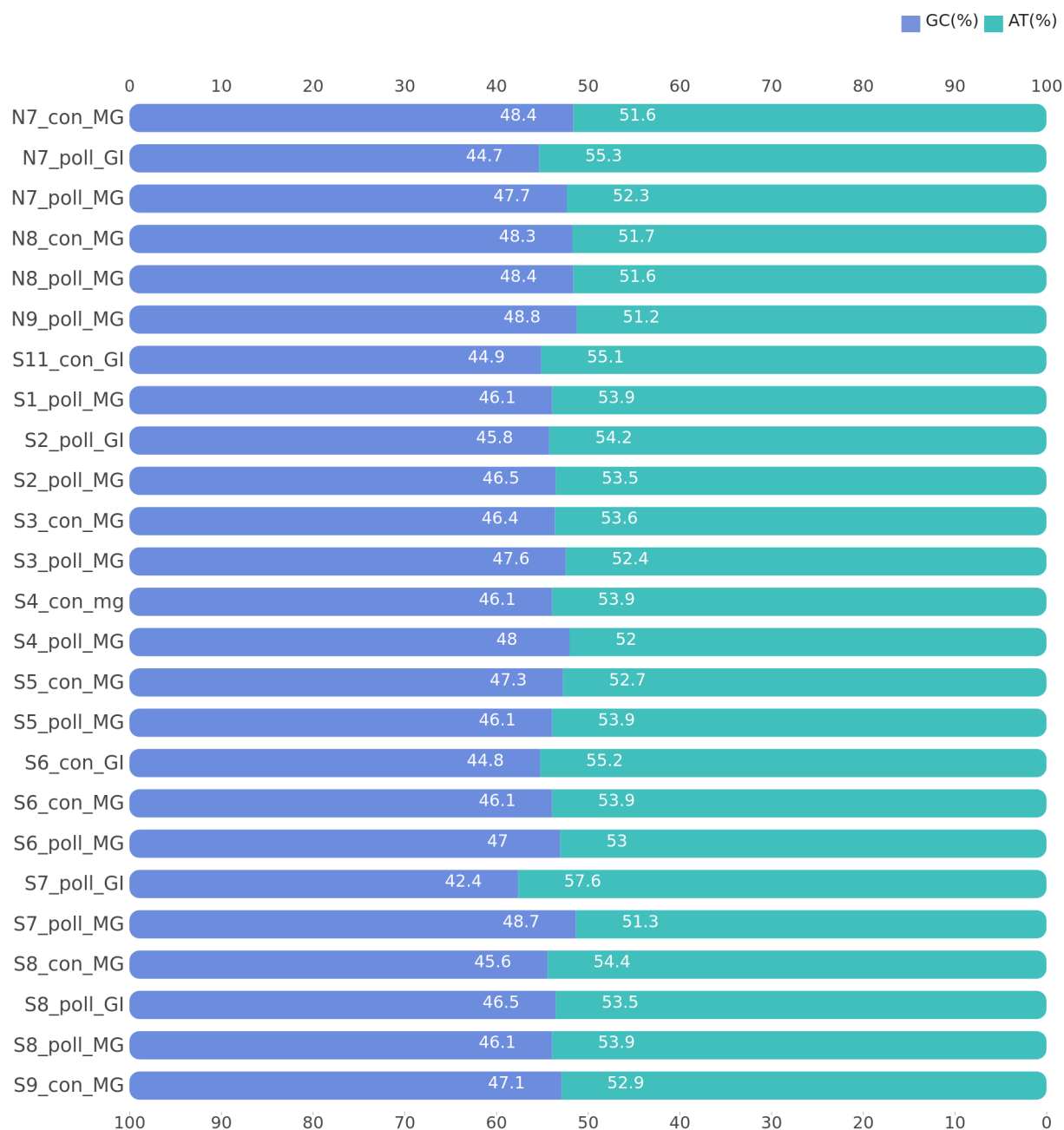
Total number of samples : 58

* Raw Data



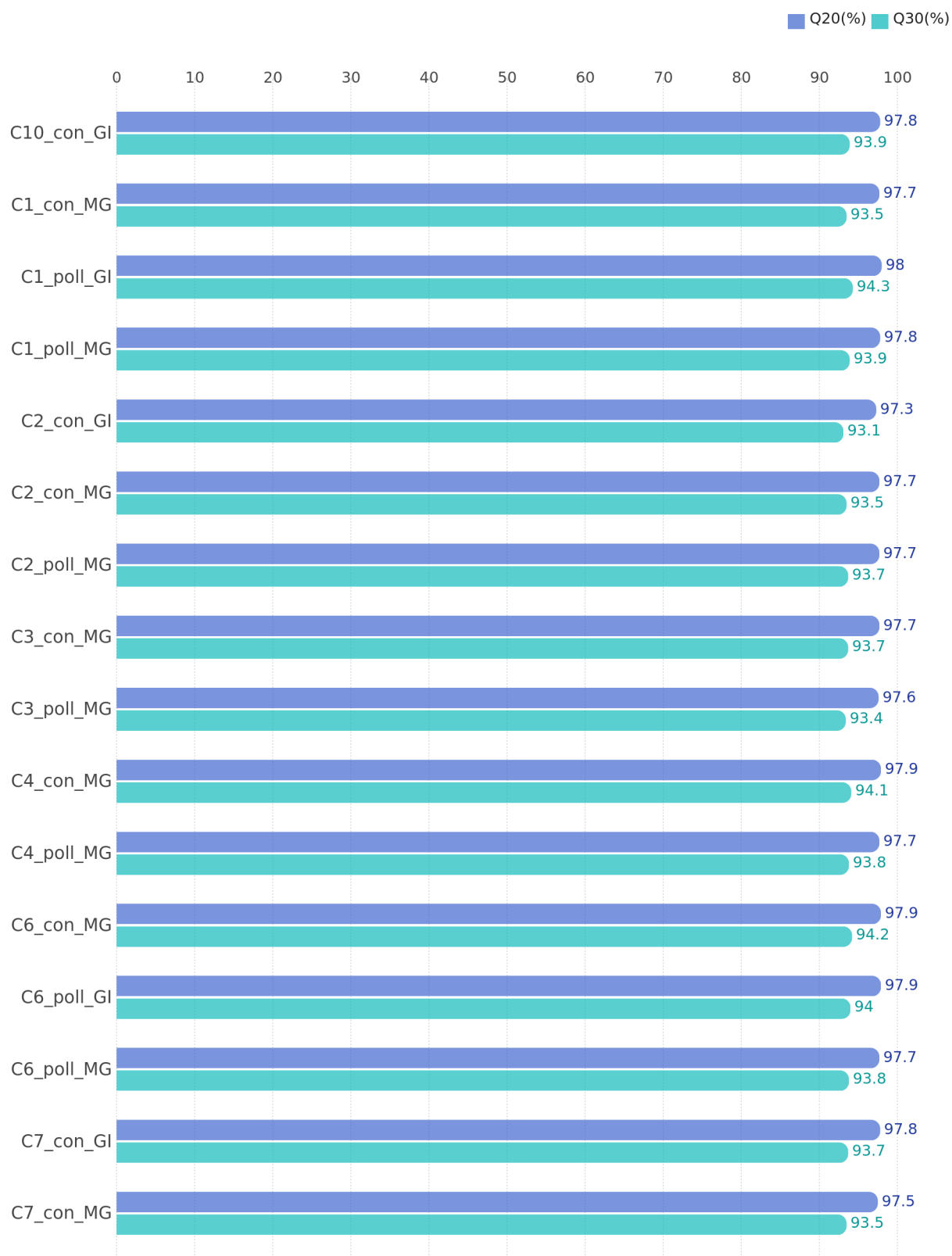
GC/AT Content

* Raw Data



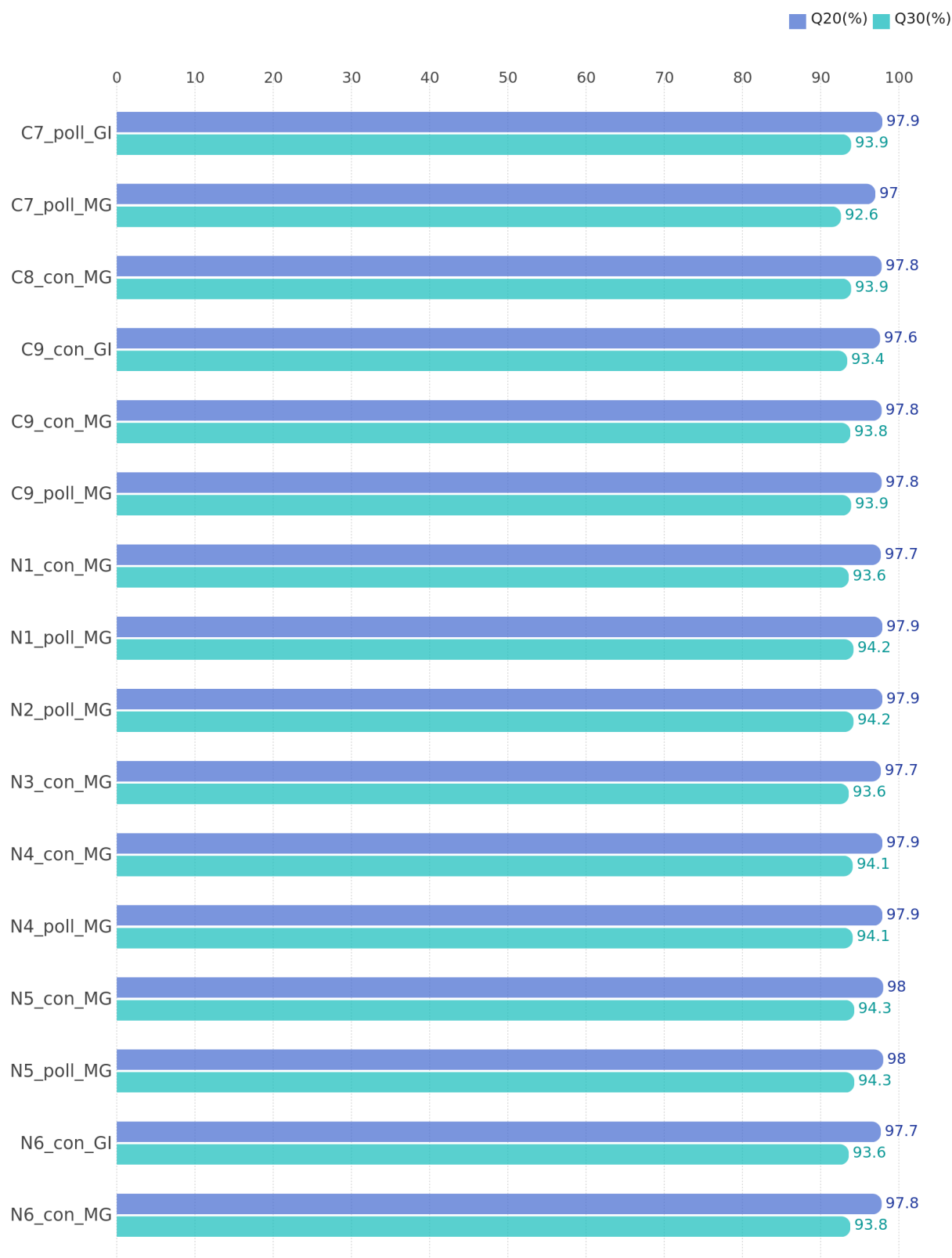
Q20/Q30 (%) Total number of samples : 58

* Raw Data



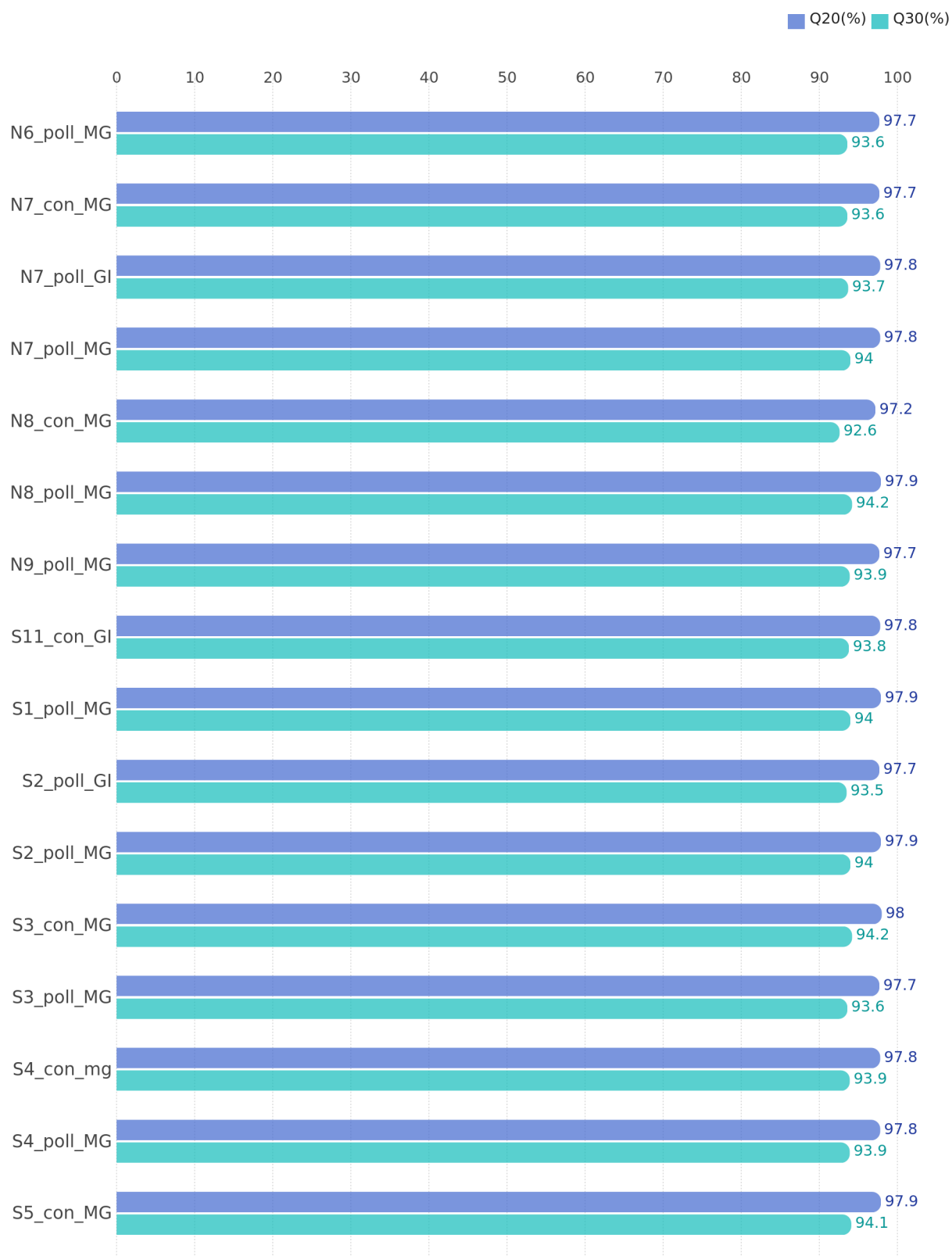
Q20/Q30 (%)

* Raw Data



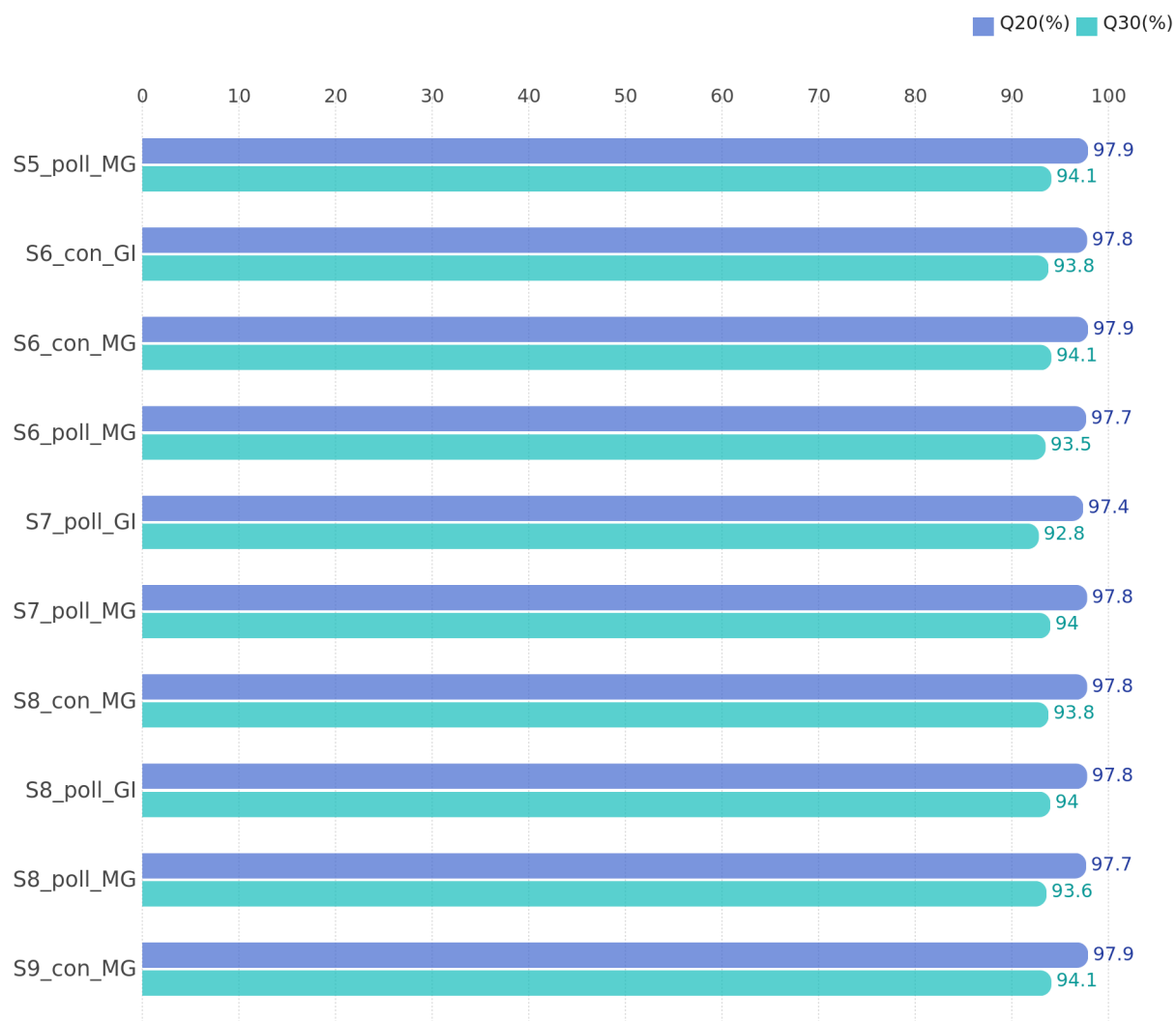
Q20/Q30 (%)

* Raw Data



Q20/Q30 (%)

* Raw Data



Download List

- The data can be downloaded from the links below. The download links are active for 2 weeks only, so please download your data within this period.
- Once you receive/download the data, please make sure to check the integrity of the files.
Please note that the sequencing files will be deleted from our server 3 months after the analysis report is released; please contact us within 3 months if you encounter a problem with the data.

* Raw Data Download

File Name	File Size(byte)	md5sum
C10_con_GI_1.fastq.gz	2,953,928,113	114df1052514a7c1b710ba04ad9f2ef3
C10_con_GI_2.fastq.gz	2,998,422,897	b76fce0278ef25379f00626d892770ed
C1_con_MG_1.fastq.gz	2,744,558,056	4d7d0068f6b91e98ed1e9ca00a1f4c7b
C1_con_MG_2.fastq.gz	2,805,508,871	90d7faf1791f57c1f36db544d13a6179
C1_poll_GI_1.fastq.gz	2,880,818,841	e6dbb3f95b3e69a4c9409a845d067cc4
C1_poll_GI_2.fastq.gz	2,913,377,755	f6cb0805a870c928361253bad030cd3f
C1_poll_MG_1.fastq.gz	2,701,401,439	35cdc393bd321d4aa0168133aab8cc7e
C1_poll_MG_2.fastq.gz	2,770,253,504	976aaaf2184d57f2bace461ca611796f
C2_con_GI_1.fastq.gz	2,717,146,167	50ed215bcc2e6daaa6a0f29a784f319f
C2_con_GI_2.fastq.gz	2,803,059,260	540039742ad016ac0fe5292c883a22b1
C2_con_MG_1.fastq.gz	3,009,656,630	ec76e8a7034a03301621af44648882ff
C2_con_MG_2.fastq.gz	3,107,972,843	33ee213050938beec6fc15452862293f
C2_poll_MG_1.fastq.gz	2,973,612,981	c285287623c33d78e8c3932693a0dd8d
C2_poll_MG_2.fastq.gz	3,037,269,474	1b3d950e32659a94f9791bccc130bf06
C3_con_MG_1.fastq.gz	2,654,146,890	a4b851da271c3bcc658e10e5c1458b2f
C3_con_MG_2.fastq.gz	2,716,569,568	778887a56ce3791f68767438a7133e27
C3_poll_MG_1.fastq.gz	2,357,683,094	ee557e9b68fb64eb6788fe98f64dfa41
C3_poll_MG_2.fastq.gz	2,444,222,476	6c36b803f003edb4ff04d37363bde070
C4_con_MG_1.fastq.gz	2,665,162,937	f4fd384330510dc10755ddb4c0d4228e
C4_con_MG_2.fastq.gz	2,720,761,728	cee4b77e2eb8d3d4dbe39a8ba28050c1
C4_poll_MG_1.fastq.gz	2,564,620,251	96ccf974d44515625810b01c3b99591a
C4_poll_MG_2.fastq.gz	2,612,507,613	806beec21049a6321f0eb762aaa906fd
C6_con_MG_1.fastq.gz	2,852,678,192	2713e278c6fe12d47d87adef8f6a17f
C6_con_MG_2.fastq.gz	2,893,284,194	046c286aa8ece4d30be2773cd5cb5b45
C6_poll_GI_1.fastq.gz	2,800,677,703	75efe7b933084b93e887a42be69660a0
C6_poll_GI_2.fastq.gz	2,845,891,936	65b419661606da83da92b211917d74dc
C6_poll_MG_1.fastq.gz	2,194,990,687	051d0c8d87079584f533b893ea5f4b5e
C6_poll_MG_2.fastq.gz	2,222,469,532	b21f106085cf4911748438409102258a
C7_con_GI_1.fastq.gz	2,946,829,660	9acb0747d2ef54198e29f514c8c167af
C7_con_GI_2.fastq.gz	3,002,160,538	7fb5c4f7049550a117dccb99b00a3e9f
C7_con_MG_1.fastq.gz	2,414,504,544	41a9acdb831583f35ef5355fcd0c6959

File Name	File Size(byte)	md5sum
C7_con_MG_2.fastq.gz	2,464,144,048	20742e5b2da36a1ff50ec6b95f2960b5
C7_poll_GI_1.fastq.gz	2,655,407,186	ebe3399e8407cd0bc2774bcad76eae05
C7_poll_GI_2.fastq.gz	2,696,766,410	b981bf380b3fb87904cc85f775f90522
C7_poll_MG_1.fastq.gz	2,623,532,635	3b24f59da5fa1986558a6e3a86f14a0d
C7_poll_MG_2.fastq.gz	2,735,346,738	444a7661e1605f81c73b4bcab651726d
C8_con_MG_1.fastq.gz	2,205,951,981	a9aaab4af67b6ef29d636f060816846d
C8_con_MG_2.fastq.gz	2,250,362,019	3c55e4414d88693f354975112c54494b
C9_con_GI_1.fastq.gz	2,928,630,550	4c33441e0a7823f57060e833b00f6b23
C9_con_GI_2.fastq.gz	3,002,221,934	415d652005d1edf1d940f31532d0d734
C9_con_MG_1.fastq.gz	2,772,281,101	4b9cc86766be3a74c38cfd032d3b47de
C9_con_MG_2.fastq.gz	2,801,523,015	937f5cd6787c3af3aa91f885ffd215b9
C9_poll_MG_1.fastq.gz	2,399,491,431	239fddad69a5b40fcee7c919d1af25fc
C9_poll_MG_2.fastq.gz	2,440,457,401	637c342b20988347b2a9b91252dfa1b
N1_con_MG_1.fastq.gz	2,564,467,170	e8312f448e272ffda850679f6525c682
N1_con_MG_2.fastq.gz	2,653,551,552	0e69070213d81fbb0e569805886fdeb1
N1_poll_MG_1.fastq.gz	2,235,146,418	c7127457bc46b7256a9adb0beb7ddf8c
N1_poll_MG_2.fastq.gz	2,251,741,981	37647067af8bb1fd4fba78d199f20bee
N2_poll_MG_1.fastq.gz	2,450,796,483	1bd979084dec84ca1d87b90e3cd704a4
N2_poll_MG_2.fastq.gz	2,493,722,942	9ed10e76746e9ca69442b585ca2e466f
N3_con_MG_1.fastq.gz	2,422,742,966	24dc37577143653586297f032fa19147
N3_con_MG_2.fastq.gz	2,501,269,917	ad1673d713842282c427f95c51af7549
N4_con_MG_1.fastq.gz	2,215,367,747	be6ef213f7fa4b16bfb5eff824288b6c
N4_con_MG_2.fastq.gz	2,261,630,548	bfe3c33f3bab6c330ce327d72eb26891
N4_poll_MG_1.fastq.gz	2,772,716,760	f314a4f31f1f2621b83eecd829083d62
N4_poll_MG_2.fastq.gz	2,825,538,068	81dd25d61e2c777cf4fab27785f23392
N5_con_MG_1.fastq.gz	2,203,609,375	8305290bd77dcbc627e00e70ce0fbdf1
N5_con_MG_2.fastq.gz	2,215,856,190	b9948f245d98a8346565c03f7452e381
N5_poll_MG_1.fastq.gz	2,596,312,356	441b66fb70c416f6ce40505fe258ca82
N5_poll_MG_2.fastq.gz	2,633,205,131	493ff0f471ad255a5d0b5b360f43a280
N6_con_GI_1.fastq.gz	2,729,073,833	105f026df2c562ff97467532381faea0
N6_con_GI_2.fastq.gz	2,801,138,437	7899f22f58082fd2427168d536658b2f
N6_con_MG_1.fastq.gz	2,009,729,434	7d258db049ea42f457a64fbf4edc6ea1
N6_con_MG_2.fastq.gz	2,067,162,187	7a7151de7daa628f1936d9c88e7749f9
N6_poll_MG_1.fastq.gz	2,642,705,774	255008f29678f5c3164ab21bf00a67f8
N6_poll_MG_2.fastq.gz	2,713,449,461	ee72061a17cc10347c9594bdf7da28b
N7_con_MG_1.fastq.gz	2,184,385,778	f88cf224aa6b70852262b15f9b7648f6
N7_con_MG_2.fastq.gz	2,249,940,300	bb7748c33c2bdf7c35865e17ede3807d
N7_poll_GI_1.fastq.gz	2,711,811,417	751c56b2bdb32812add500697d08e02f

File Name	File Size(byte)	md5sum
N7_poll_GI_2.fastq.gz	2,806,457,869	faf7df02a96563288fdef0fbd75938ca
N7_poll_MG_1.fastq.gz	2,785,468,908	22fec319c7b6a028741145c9ee126d5d
N7_poll_MG_2.fastq.gz	2,788,207,073	97356a0a953bc62c6ea0b0b61fe63025
N8_con_MG_1.fastq.gz	3,031,972,182	9460fec38bdec0eef25b77d580e9a0f1
N8_con_MG_2.fastq.gz	3,253,050,976	c227507d384378ede460feae4a8d09a2
N8_poll_MG_1.fastq.gz	2,596,289,236	e6a4e16d4199097fe9391ab703045b2e
N8_poll_MG_2.fastq.gz	2,612,391,804	a621e93d6d9f0ed8bfe75575086d58ac
N9_poll_MG_1.fastq.gz	2,576,621,897	e37c672bbef577e0b384fa121f8e4489
N9_poll_MG_2.fastq.gz	2,623,517,703	fa8d0be952b07642a146f7efe3be65e
S11_con_GI_1.fastq.gz	2,864,811,317	0874c81ab87409226486ae4b475e725a
S11_con_GI_2.fastq.gz	2,927,174,923	85ff2e8801fc157805eca3ecb235afd1
S1_poll_MG_1.fastq.gz	3,018,253,196	6ef1cc2f073f87c0b77fed0d8d044e5
S1_poll_MG_2.fastq.gz	3,110,776,221	e5b2cb389501d0745c4c8c7d559855c0
S2_poll_GI_1.fastq.gz	2,343,858,679	db66cfc5dea95d85c32a994e0f293c5
S2_poll_GI_2.fastq.gz	2,416,190,883	03ac1371c42eca58ca01395d898165a6
S2_poll_MG_1.fastq.gz	2,758,163,729	65253c6a24ea211542aa8c6bab3da915
S2_poll_MG_2.fastq.gz	2,832,550,665	516311bcb3e444c22d7c24dfc41e95cc
S3_con_MG_1.fastq.gz	2,662,547,815	03159ec54f4d6d29b2fa0b20b4655f9b
S3_con_MG_2.fastq.gz	2,697,378,141	dad3f718457e0731f2c99fa553cf92d8
S3_poll_MG_1.fastq.gz	2,518,002,334	2e4e25fa85018eea68a6ddcce0b02826
S3_poll_MG_2.fastq.gz	2,579,487,285	7582a33463ef8d932c6ddc7f6617bc31
S4_con_mg_1.fastq.gz	2,464,543,827	2dce6b1d36eeb03bf8be34655fff2e65
S4_con_mg_2.fastq.gz	2,513,674,771	5aad7b0667393acaef8fafe85fe01cd2
S4_poll_MG_1.fastq.gz	2,713,009,623	a3d601c31c36b9fe7ecb9c328deb72a
S4_poll_MG_2.fastq.gz	2,749,798,296	91c4e88c95b7ed42c46079959fee6a2b
S5_con_MG_1.fastq.gz	2,196,527,255	39d1418ec8be6fa8d27f37d1f68e5bcf
S5_con_MG_2.fastq.gz	2,237,818,251	fea12ddb1da906f2bbb6042e00309073
S5_poll_MG_1.fastq.gz	2,544,752,921	179b7ea4c3a364d8b9e9efb6b116d7e7
S5_poll_MG_2.fastq.gz	2,586,206,103	2566a2c339e50d03ffcb27e27883605e
S6_con_GI_1.fastq.gz	2,684,434,258	b90661f3dc2fb294fc734e6c263b8969
S6_con_GI_2.fastq.gz	2,730,820,081	fc17aad3d1a04344cda9905e8fc80584
S6_con_MG_1.fastq.gz	2,264,766,766	2ee7591854d2f14a2ab78395ba0284f1
S6_con_MG_2.fastq.gz	2,300,442,843	e499cf5f02e08f67d3c8800341349a51
S6_poll_MG_1.fastq.gz	2,682,453,449	c2c31a45d1e9f204563a51fa888ec0e0
S6_poll_MG_2.fastq.gz	2,795,456,944	61a929f6dd68a4c74d74ead78dfefbaa
S7_poll_GI_1.fastq.gz	2,835,758,366	330f15ff5b6ec65739bb282336515f4b
S7_poll_GI_2.fastq.gz	2,986,599,420	0e8c03fda364105aee787556970927ca
S7_poll_MG_1.fastq.gz	2,741,758,438	c855fb61841c5ddba06842d5bae45501

File Name	File Size(byte)	md5sum
S7_poll_MG_2.fastq.gz	2,772,509,152	cedc96eacde73be0e9da56cd23a4f2cc
S8_con_MG_1.fastq.gz	3,005,780,861	c1fcb7bdf9cedc38b871caa9790c1b03
S8_con_MG_2.fastq.gz	3,067,841,083	2541e914e4d0c2c8fae1454440441021
S8_poll_GI_1.fastq.gz	2,828,904,707	ffe6041e856f65d74145fb3af9992d96
S8_poll_GI_2.fastq.gz	2,813,308,370	2526bde2c62e9fd043a10908129c8cc9
S8_poll_MG_1.fastq.gz	2,883,528,057	22efc1febbba7c7021d5466d7c29b532
S8_poll_MG_2.fastq.gz	2,962,480,286	843027c280cf8eea5ae0588efe902bb2
S9_con_MG_1.fastq.gz	2,699,811,320	29d50d1cbf4210e497ae9795633cd30a
S9_con_MG_2.fastq.gz	2,723,469,019	8ca4d049fa5ff5fbb4aadeacf3a91b5d

FAQ

Q Why do I need to check the md5sum values, and how can I check it? (Windows system)

A NGS data tend to have a large files size which makes them more likely to be corrupted during file transfer. So it's important that you check the md5sum of the files after receiving them to make sure what you received are what we gave.

Checking md5 hash in a Windows system

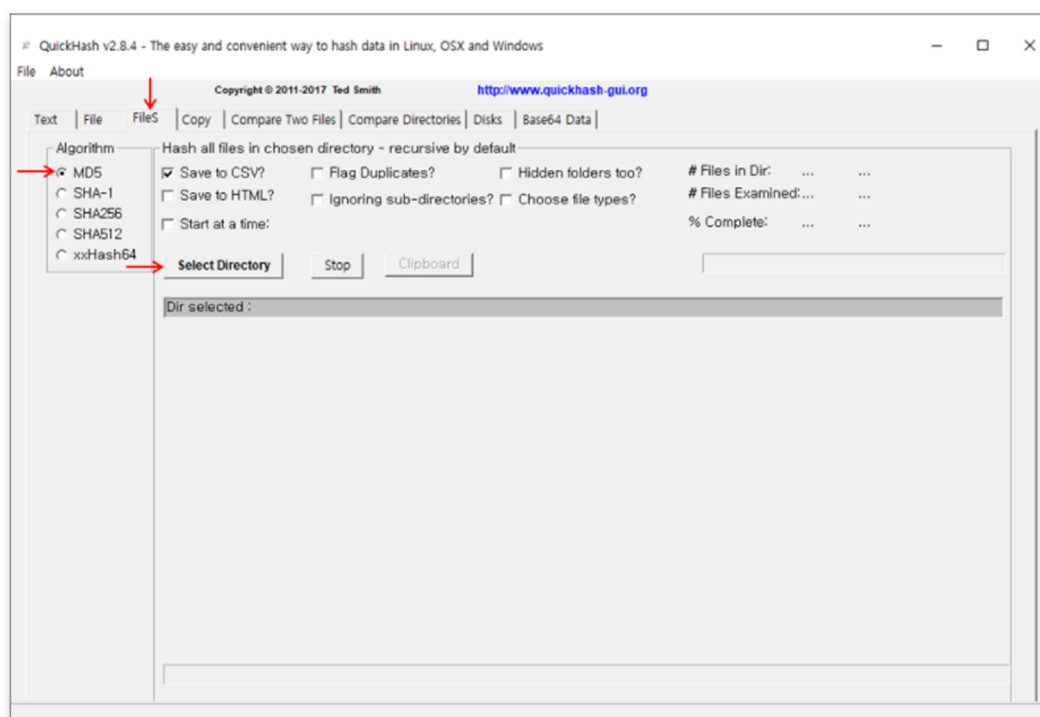
Windows does not provide a program for checking md5sum by default. An external program such as [QuickHash-Windows](#) can be used instead.

STEP 1 Download QuickHash-Windows from the website, and unzip the file.

STEP 2 Take a look at the UserManual.pdf file inside the zip file, and execute the .exe file.

Quickhash-GUI.exe	2,090,414	6,505,472
sqlite3-win32.dll	429,646	852,754
sqlite3-win64.dll	717,149	1,742,848
UserManual.pdf	512,697	576,987

STEP 3 Click on the "FileS" tab, and select [MD5] as the Algorithm.



STEP 4 Click "Select Directory" and choose the directory where the files to be checked are located in. The output can be saved as a csv or txt file. The process may take some time depending on the performance of the system being used.

STEP 5 Compare the newly calculated md5 value with the md5 value provided to you through the Analysis Report.

FAQ

Q Why do I need to check the md5sum values, and how can I check it? (Linux system)

A NGS data tend to have a large files size which makes them more likely to be corrupted during file transfer. So it's important that you check the md5sum of the files after receiving them to make sure what you received are what we gave.

Checking md5 hash in a Linux system

Linux systems have an internal md5sum program under /user/bin/md5sum.

md5sum has a "-c" option, which reads the MD5 sums from the input file and checks them simultaneously.

Usage: \$ md5sum -c [input file name]

STEP 1 Macrogen provides a text file containing the md5sum of deliverables you'll be receiving, which you can use to validate the integrity of the files. You can download this file by clicking on the "md5sum List" button in the "Download List" page. The text file will have the following name and format depending on how you're receiving your data:

o Via download link : <OrderNumber>_#samples_md5sum_DownloadLink.txt

```
[user@host] cat HN00000000_1samples_md5sum_DownloadLink.txt
File      Size      md5sum      Download_link
test_1.fastq.gz 3118212249  07a66a1d7d7fde2ee71b02a2caf21aba  https://data.macrogen.com/-macro3/HiSeq02/20210322/HN00000000/test_1.fastq.gz
test_2.fastq.gz 3305438294  3b4ff911e5d238a3c4763ee7967cb29a  https://data.macrogen.com/-macro3/HiSeq02/20210322/HN00000000/test_2.fastq.gz
```

o Via HDD : <OrderNumber>_#samples_md5sum.txt

```
[user@host] cat HN00000000_1samples_md5sum.txt
File      Size      md5sum
test_1.fastq.gz 3118212249  07a66a1d7d7fde2ee71b02a2caf21aba
test_2.fastq.gz 3305438294  3b4ff911e5d238a3c4763ee7967cb29a
```

o You can also find "md5sum.txt" located inside the HDD delivered to you.

```
[user@host]$ cat md5sum.txt
07a66a1d7d7fde2ee71b02a2caf21aba  RawData/test_1.fastq.gz
3b4ff911e5d238a3c4763ee7967cb29a  RawData/test_2.fastq.gz
```

STEP 2 Use "md5sum -c" to validate the integrity of the file you've received. The input file for md5sum -c has to be delimited by two spaces with the md5sum column appearing before the file name, just like the sample image of "md5sum.txt" file shown above. As you can see, the two other files above are not formatted this way and need to be altered to be used as input for md5sum -c. You can manually exclude the header and cut out "File" and "md5sum" column from the files, or simply run the following command:

\$ awk '{print \$3 " " " \$1}' <md5sum_file> | grep -v File

STEP 3 "md5sum -c" reads the input containing the md5 value of a file, and checks whether the md5 value of that file matches what's written inside the input file. This action outputs "OK" if the md5 value matches, and "FAILED" if otherwise. Check if the command outputs "OK" for all the files. (Refer to image below)

```
user@host
[user@host] awk '{print $3 " " " $1}' HN00000000_1samples_md5sum_DownloadLink.txt | grep -v File > md5sum.txt
[user@host] cat md5sum.txt
07a66a1d7d7fde2ee71b02a2caf21aba  test_1.fastq.gz
3b4ff911e5d238a3c4763ee7967cb29a  test_2.fastq.gz
[user@host]
[user@host] md5sum -c md5sum.txt
test_1.fastq.gz: OK
test_2.fastq.gz: OK
[user@host]
```


FAQ

Q I want to see the data produced by MacroGen. How can I open the files?



A NGS data tend to have large file sizes, and are not user-friendly to work with in a Windows environment. We recommend that you use Linux system for smoother operation.

Q Where can I find information for Illumina adapter sequences?

A Information on Illumina adapters can be found in this support document:
[Adapter Sequences Intro](#)

Result File Description

Deliverables List

File Type	File Name	Description
FASTQ	 [Sample name]_[read1].fastq.gz	Raw read1 sequence data
	 [Sample name]_[read2].fastq.gz	Raw read2 sequence data
md5sum	[Order#]_[#samples]_md5sum[DownloadLink].txt	<p>You can download this file by clicking on the "md5sum List" button found on the "Download List" page. The file is slightly different in terms content, depending on how you're receiving your data. If you're receiving via download link, the file contains the following information : File name, File size, md5sum, FTP link. Otherwise, if your receiving your data via HDD the file only contains : File name, File size, and md5sum.</p> <p>MD5 is a string of 32 hexadecimal values, which represents a 'fingerprint' of a file. By comparing the supplied MD5 value to the actual value computed by the MD5sums utility, you can make sure that the file that you downloaded off of the internet has not been tampered with or modified from the original file stored in our server.</p>

FASTQ Format

Example:

Line 1 : Sequence identifier

Line 2 : Nucleotide sequences

Line 3 : Quality score identifier line - character '+'

Line 4 : Quality score

```

@A00125:17:H2HFJDMXX:1:1101:3170:1000 1:N:0:ATGCCTAA
GAAACACGATGACACTCACATGGCACTCACATTTCTAGTCTCTTTCTAAGTGATTGCAATATTAATTCATAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00125:17:H2HFJDMXX:1:1101:9408:1000 1:N:0:ATGCCTAA
TGTGCGAAGGAAATCATTTCAGATGACAGTGTTAACCATGGTCAAAGGACCATTCTGTCTATCCTTCTTA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

```

FASTQ file consists of four lines.

Quality score is represented with each character.
One character matches its base with Phred+33

Phred Quality Score

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000. Phred Quality Score Q is calculated with $-10\log_{10}(P)$, where p is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%



HEADQUARTER

MacroGen Gangnam HQ

Business & Support Center

MacroGen Bldg, 238, Teheran-ro,
Gangnam-gu, Seoul, Republic of Korea
Tel: +82-2-2180-7000
Web: www.macrogen.com
LIMS: dna.macrogen.com

MacroGen Genome Center

Laboratory & IT Center

[08511] 1001, 10F, 254, Beotkkot-ro,
Geumcheon-gu, Seoul, Republic of Korea
(Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com
(Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

MacroGen Europe

Laboratory, Business & Support Center

Meibergdreef 57, 1105 BA, Amsterdam,
the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Psomagen (MacroGen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville,
MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

MacroGen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse,
Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

MacroGen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi,
Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

MacroGen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas,
28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

MacroGen Belgium

Laboratory, Business & Support Center

Oxfordlaan 70, 6229 EV Maastricht,
Netherlands
Tel: +31-20-333-7563
Email: info.be@macrogen.eu

MacroGen Italy

Laboratory, Business & Support Center

Viale Ortles, 22/4, 20139 Milano,
MI, Italy
Tel: +39-02-5666-0274
Email: italy@macrogen-europe.com