

# Metagenome Amplicon Sequencing Report

2024.01

RAWDATA REPORT



# Table of Contents

---

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

---

## 01 Workflow

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

---

## 02 Raw Data Result

Raw Data Statistics	5
Total Bases	6
GC/AT Content	7
Q20/Q30 (%)	8

---

## 03 Deliverables

Download List	9
---------------	---

---

## 04 Appendix

FAQ	11
Result File Description	14

# Order Information

Client Name	META사업부
Client Organization	(주)마크로젠
Order Number	HN00sample
Application	Metagenome Amplicon Sequencing
Type of Read	Paired-end
Read Length	301
Library Kit	Herculase II Fusion DNA Polymerase Nextera XT Index V2 Kit
Library Protocol	16S Metagenomic Sequencing Library Preparation Part # 15044223 Rev. B
Type of Sequencer	Illumina platform

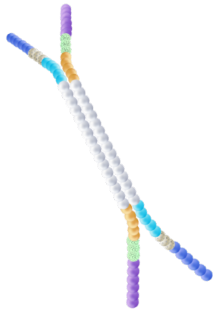
# Experimental Workflow

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.



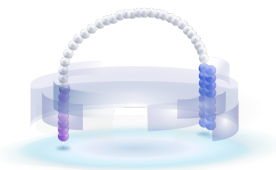
## Ligate Adapters

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.

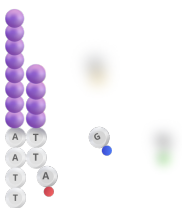


## Cluster generation using bridge amplification

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 15 samples.  
 For example, in A1 sample, 118,288 reads are produced, and total read bases are 35.6 Mbp.  
 The GC content (%) is 53.3% and Q30 is 84.1%.

## \* Raw Data

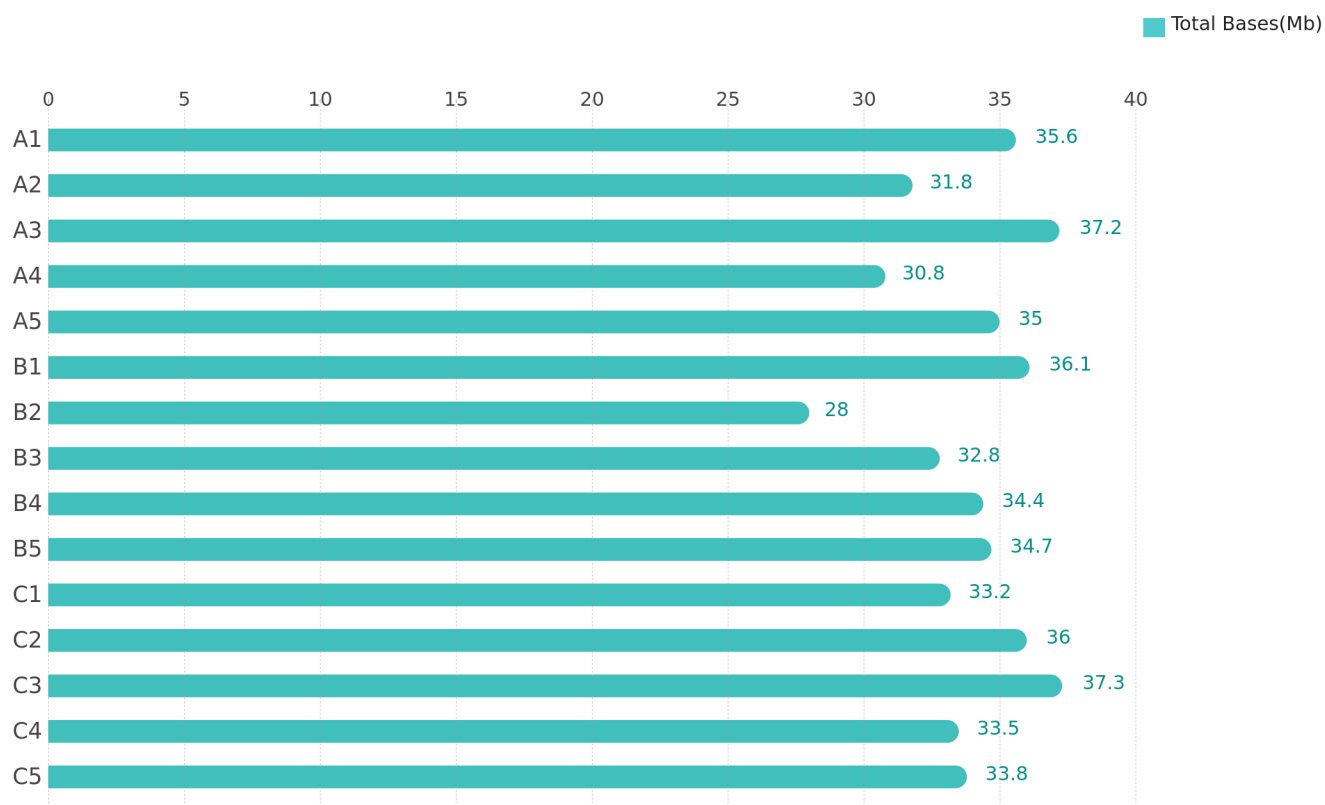
Sample ID	Total bases (bp)	Total reads	GC (%)	AT (%)	Q20 (%)	Q30 (%)
A1	35,604,688	118,288	53.3	46.7	91.9	84.1
A2	31,772,356	105,556	51.6	48.4	92.2	85.0
A3	37,246,342	123,742	50.4	49.6	94.0	87.3
A4	30,835,042	102,442	51.1	48.9	93.9	87.0
A5	35,012,922	116,322	50.7	49.3	93.2	86.1
B1	36,147,090	120,090	54.5	45.5	92.4	84.8
B2	28,007,448	93,048	53.5	46.5	92.3	84.7
B3	32,836,090	109,090	51.4	48.6	93.7	86.7
B4	34,438,614	114,414	52.2	47.8	93.6	86.6
B5	34,738,410	115,410	51.9	48.1	93.3	86.3
C1	33,187,658	110,258	53.8	46.2	91.9	84.2
C2	36,027,894	119,694	52.7	47.3	93.4	86.4
C3	37,268,014	123,814	50.6	49.4	93.6	86.8
C4	33,471,802	111,202	52.1	47.9	92.7	85.2
C5	33,809,524	112,324	51.2	48.8	92.6	85.0

- 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 : 15

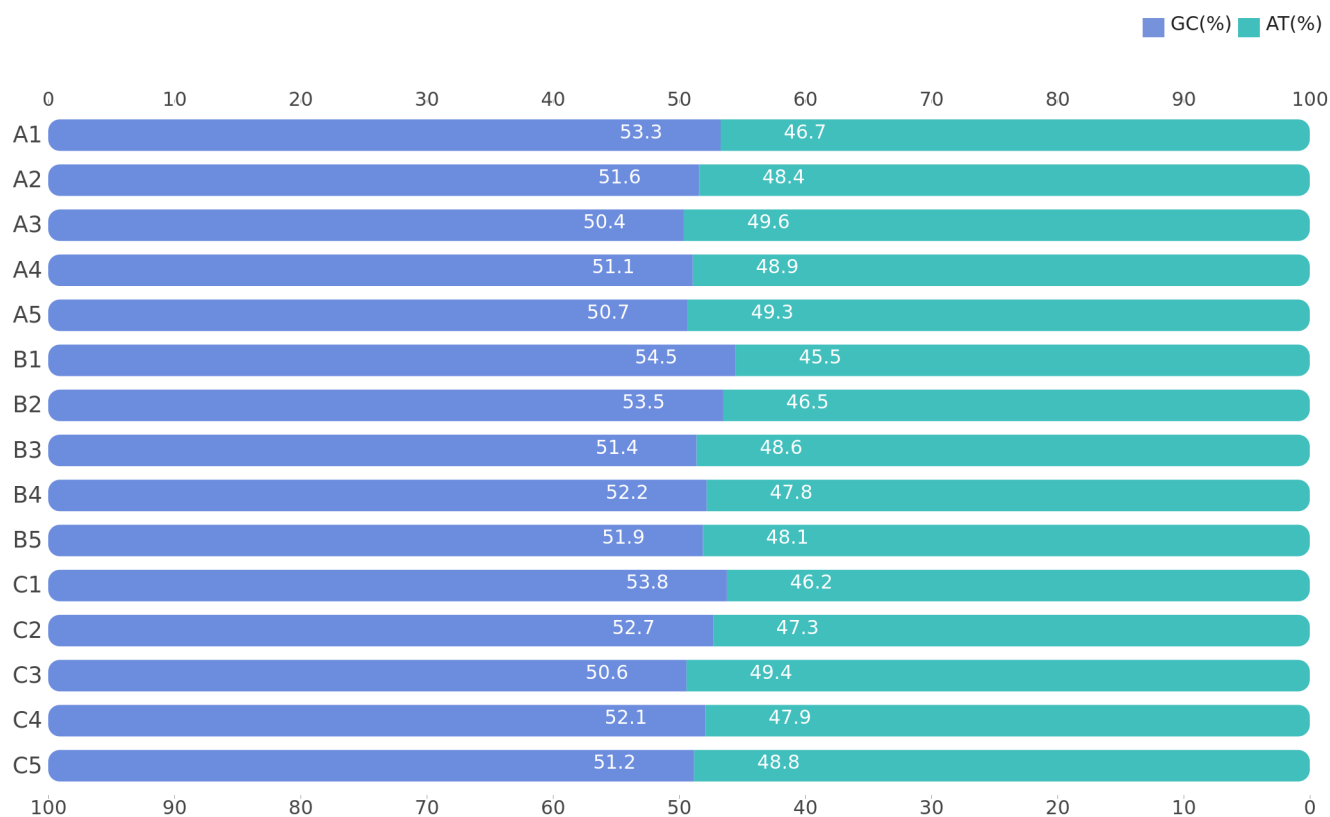
\* Raw Data



# GC/AT Content

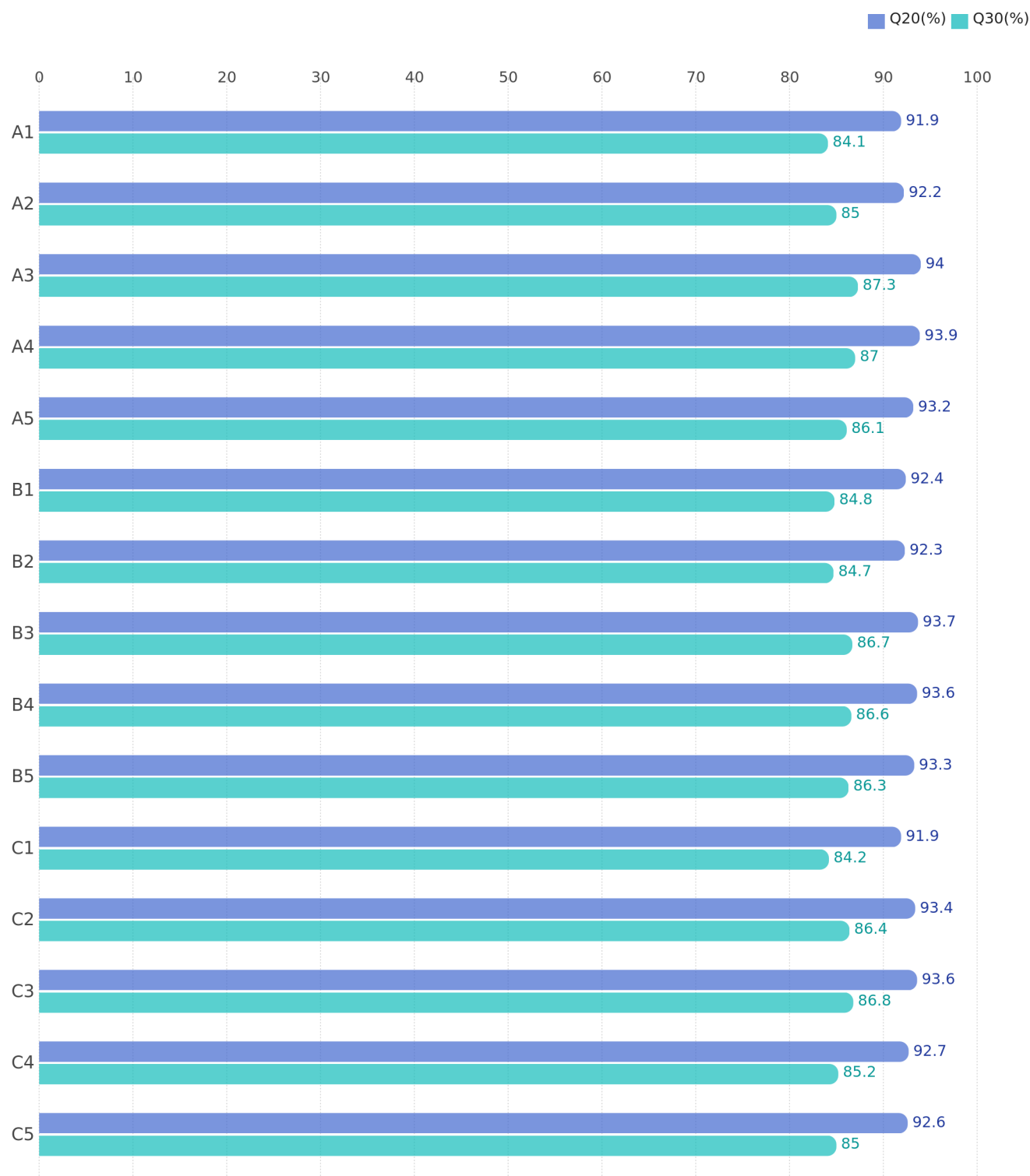
Total number of samples : 15

\* Raw Data



# Q20/Q30 (%) Total number of samples : 15

\* 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
<a href="#">HN00sample.zip</a>	259,728,480	b73c0df1d0734f9024b8628375285533

File Name	File Size(byte)	md5sum
<a href="#">A1_1.fastq.gz</a>	8,581,042	bb0a6a84c4868271f5f979ccae3b18eb
<a href="#">A1_2.fastq.gz</a>	11,515,498	ec9497144b6ab5480b26da3359adc315
<a href="#">A2_1.fastq.gz</a>	6,965,763	76b2fc22eae8351dbc9beddc8b45bf6c
<a href="#">A2_2.fastq.gz</a>	9,183,979	34542dc75a577b0e237abd5577770b0a
<a href="#">A3_1.fastq.gz</a>	7,644,734	eaafc4cfb93e7949ac8fc26150d2b9b3
<a href="#">A3_2.fastq.gz</a>	9,828,889	5e8558e5f87a2afa6ba486bc83e62f23
<a href="#">A4_1.fastq.gz</a>	6,396,795	2f230933df1197a32f692b43aaa9b15e
<a href="#">A4_2.fastq.gz</a>	8,152,317	d93c9bea03299aa8432b4d368c5238ec
<a href="#">A5_1.fastq.gz</a>	7,183,289	a90d3504e58f0aac8a95fd6d10b1e074
<a href="#">A5_2.fastq.gz</a>	9,843,660	a2db06efcbaa418f75ce0f8fb9ad04b7
<a href="#">B1_1.fastq.gz</a>	8,864,176	3ed8bde65e9fd4cd0c4d0597b802e240
<a href="#">B1_2.fastq.gz</a>	11,366,756	d186916d34b3bb07784bd27f9505f28c
<a href="#">B2_1.fastq.gz</a>	6,581,470	4904bb20297c753448d9430ade2436d0
<a href="#">B2_2.fastq.gz</a>	8,645,841	f2bb19ca289041c379fd9d23995d0f36
<a href="#">B3_1.fastq.gz</a>	7,303,108	12425b744570394a534e9fa8d8ae32c2
<a href="#">B3_2.fastq.gz</a>	8,872,323	01e9c4dadd276abe9f3c6a517351fc6b
<a href="#">B4_1.fastq.gz</a>	6,973,303	b034f8785cfd1118bc18eeaf494b76bf
<a href="#">B4_2.fastq.gz</a>	9,850,986	c3054362499d1e349b13473f9cfe8111
<a href="#">B5_1.fastq.gz</a>	7,072,595	c9fdbb623320ec14161d6fb24ddb134c
<a href="#">B5_2.fastq.gz</a>	9,847,072	2864703d87d0f8ffa5dec6d7ba837638
<a href="#">C1_1.fastq.gz</a>	8,166,696	060b7099e231a9965dbc07f3318e54de
<a href="#">C1_2.fastq.gz</a>	10,495,013	8491361d892e8f1afb6ab797d749a7d3
<a href="#">C2_1.fastq.gz</a>	8,061,474	569288666ab1c5328dc698a90d997fbe
<a href="#">C2_2.fastq.gz</a>	10,531,840	a8ccb8a1e3324789d631101421c5016c
<a href="#">C3_1.fastq.gz</a>	7,552,100	4ea32d54615e4f3bf6162b786c0a97fb
<a href="#">C3_2.fastq.gz</a>	10,148,316	0557745c2cd749df4e60a64bdb6c8872
<a href="#">C4_1.fastq.gz</a>	7,307,994	f7310f59e1ec807ad387594cd9a83e49
<a href="#">C4_2.fastq.gz</a>	9,731,464	38e262878c61e7e75c85df9d0999a0b6

File Name	File Size(byte)	md5sum
<a href="#">C5_1.fastq.gz</a>	7,150,139	a854a2f3815260086c2d63c18980a739
<a href="#">C5_2.fastq.gz</a>	9,906,766	e1e87ec7157b2a7cde8c9c71c575f9be

# 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

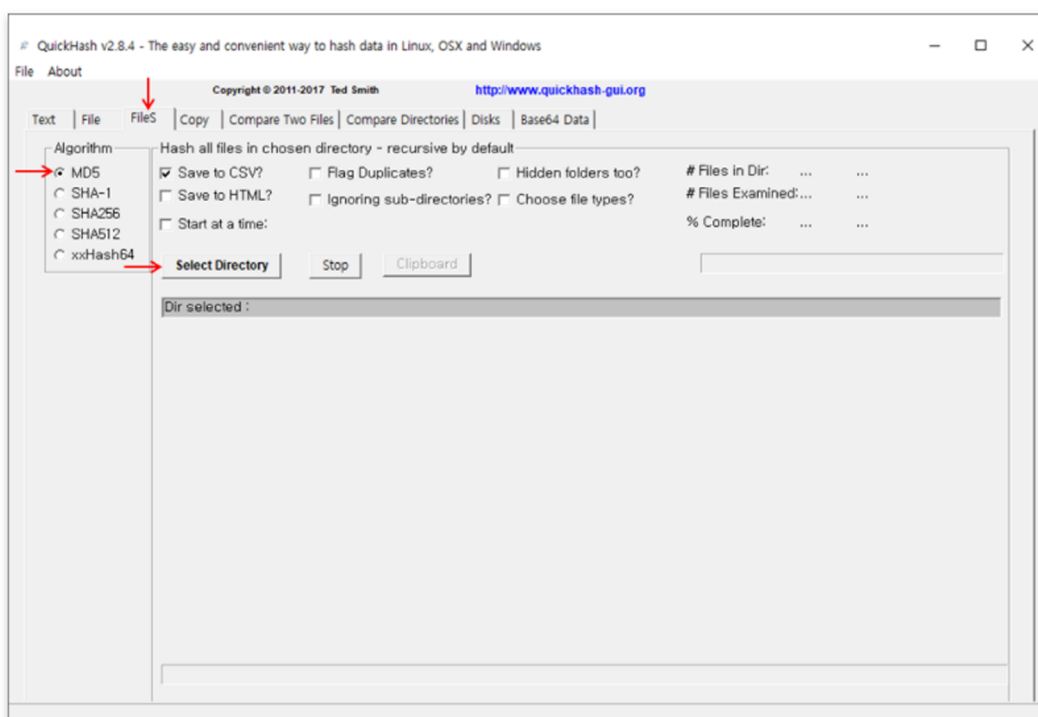
Windows does not provide a program for checking md5sum by default. An external program such as [QuickHash-Windows](#) can be used instead.

**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
<b>FASTQ</b>	 [Sample name]_[read1].fastq.gz	Raw read1 sequence data
	 [Sample name]_[read2].fastq.gz	Raw read2 sequence data
<b>md5sum</b>	[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
GAAACACGATGACACTCACATGGCACTCACATTTTCAGCTCCTTTCTAAGTGATTGCAAATATTAATTCATAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00125:17:H2HFJDMXX:1:1101:9408:1000 1:N:0:ATGCCTAA
TGTGCGAAGGAAATCATTTCAGATGACAGTGTTAACCATGGTCAAAGGACCATTCTGTCTATCCTTCTTA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

```

**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](http://www.macrogen.com)  
LIMS: [dna.macrogen.com](http://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](mailto:ngs@macrogen.com)(Overseas)  
Email2: [ngskr@macrogen.com](mailto:ngskr@macrogen.com)  
(Republic of Korea)  
Web: [www.macrogen.com](http://www.macrogen.com)  
LIMS: [dna.macrogen.com](http://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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto: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](mailto:italy@macrogen-europe.com)