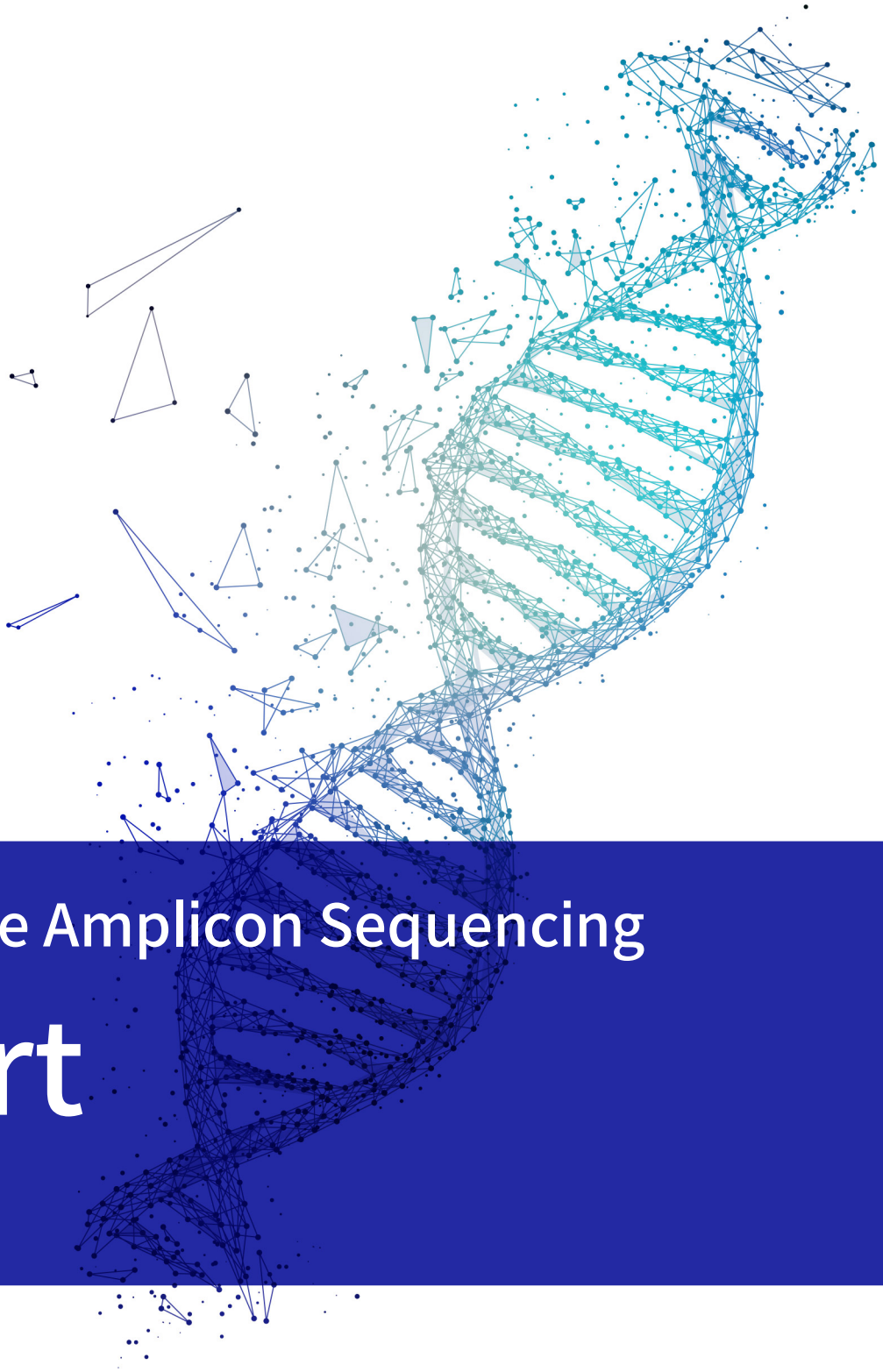
Humanizing Genomics
macrogen



Metagenome Amplicon Sequencing

Report

2024.12

RAWDATA REPORT

Research use only

Table of Contents

Order Information

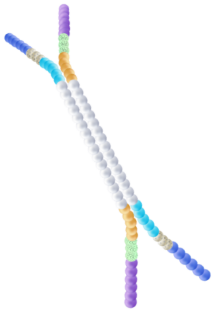
Client Name	Macrogen Europe
Client Organization	Macrogen Europe
Order Number	HN00230849
Application	Metagenome Amplicon Sequencing
Type of Read	Paired-end
Read Length	301
Library Kit	Herculase II Fusion DNA Polymerase Nextera XT Index V2 Kit
Library Protocol	16S Metagenomic Sequencing Library Preparation Part # 15044223 Rev. B
Type of Sequencer	illumina 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.

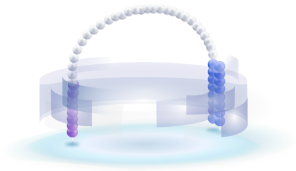


Ligate Adapters

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.

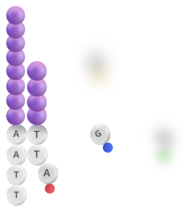
Final library Construction

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 bcl-convert 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 23 sample(s).
For example, in EF1 sample, 144,556 reads are produced, and total read bases are 43.5 Mbp.
The GC content (%) is 52.8% and Q30 is 82.9%.

* Raw Data

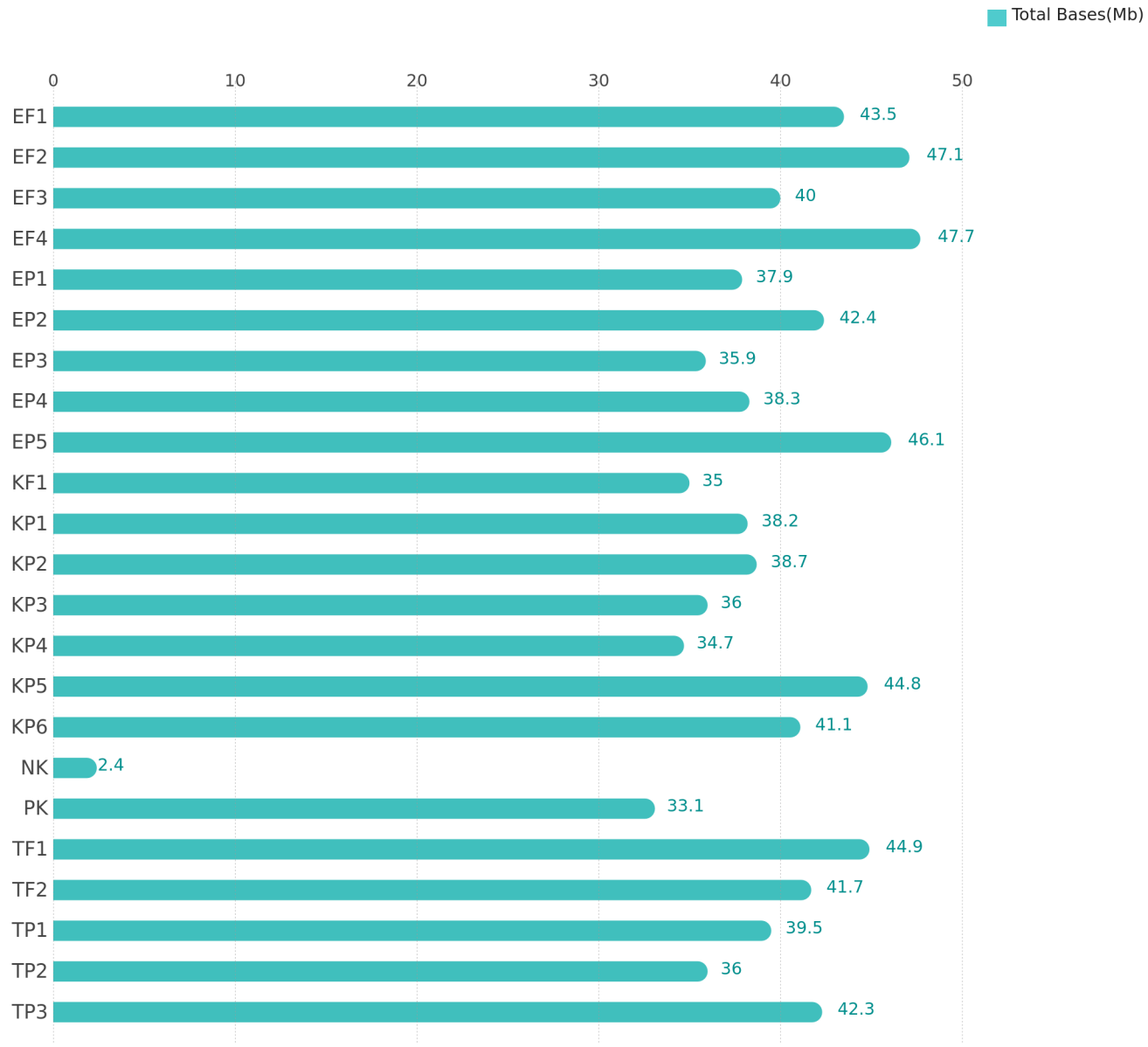
Sample ID	Total bases (bp)	Total reads	GC (%)	AT (%)	Q20 (%)	Q30 (%)
EF1	43,511,356	144,556	52.8	47.2	91.3	82.9
EF2	47,083,022	156,422	52.9	47.1	90.6	81.8
EF3	40,037,214	133,014	54.7	45.3	89.1	79.8
EF4	47,668,768	158,368	53.0	47.0	91.0	82.5
EP1	37,895,900	125,900	53.2	46.8	90.6	81.9
EP2	42,385,616	140,816	54.0	46.0	90.6	81.9
EP3	35,867,762	119,162	55.3	44.7	89.6	80.5
EP4	38,280,578	127,178	53.4	46.6	90.6	82.0
EP5	46,084,304	153,104	53.0	47.0	91.4	83.0
KF1	34,979,812	116,212	54.3	45.7	89.5	80.3
KP1	38,178,840	126,840	54.6	45.4	88.9	79.8
KP2	38,709,202	128,602	54.6	45.4	91.0	82.5
KP3	36,009,232	119,632	54.6	45.4	90.5	81.5
KP4	34,736,002	115,402	52.3	47.7	91.0	82.5
KP5	44,756,292	148,692	52.3	47.7	91.6	83.5
KP6	41,142,486	136,686	52.4	47.6	91.1	82.5
NK	2,415,224	8,024	46.5	53.5	40.9	30.2
PK	33,137,692	110,092	55.3	44.7	91.0	82.9
TF1	44,921,240	149,240	52.2	47.8	91.4	83.2
TF2	41,654,788	138,388	52.2	47.8	90.9	82.2
TP1	39,536,350	131,350	53.5	46.5	90.8	82.2
TP2	35,979,734	119,534	52.8	47.2	90.8	82.2
TP3	42,300,132	140,532	55.0	45.0	90.4	81.8

- 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 : 23

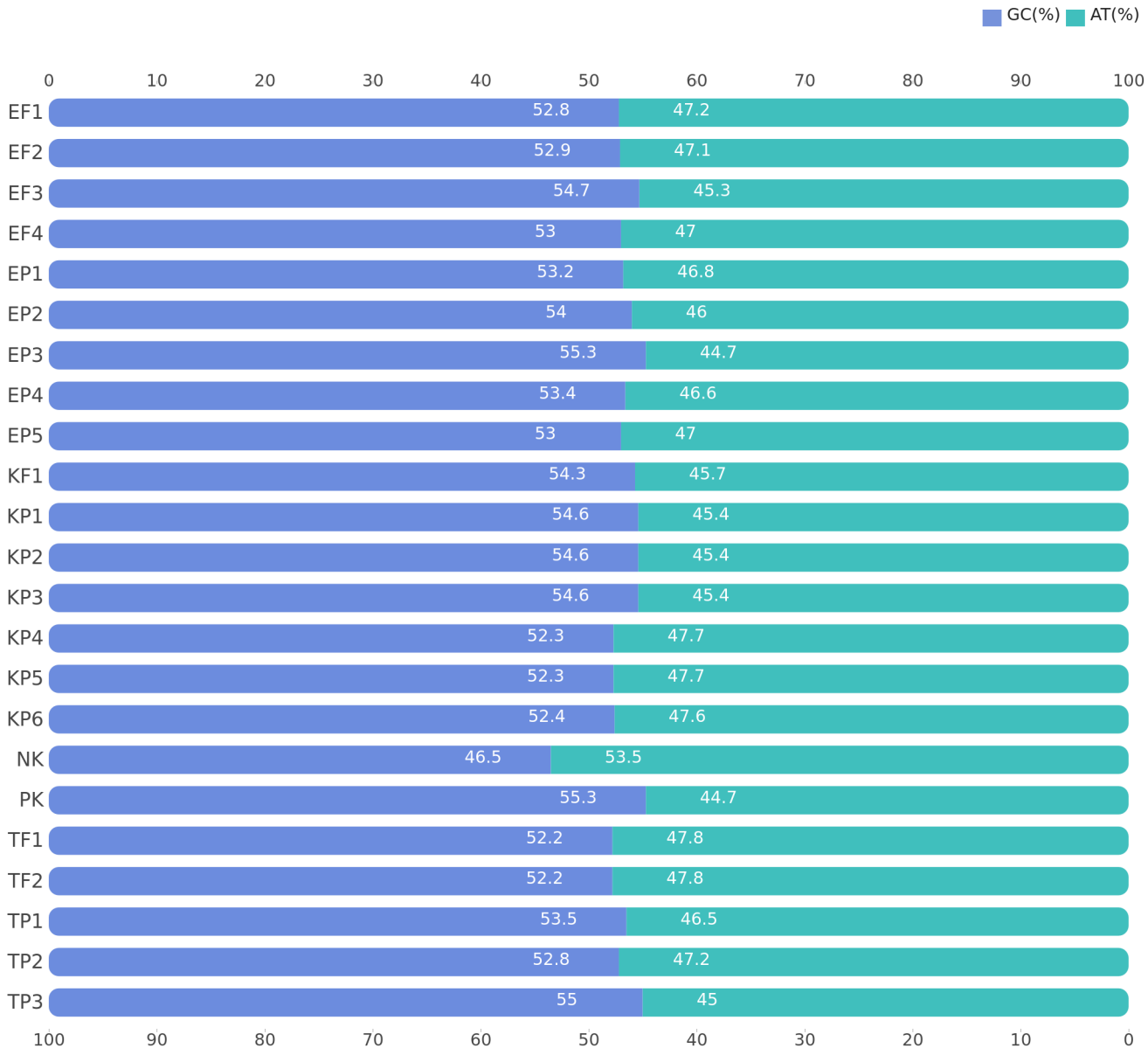
* Raw Data



GC/AT Content

Total number of samples : 23

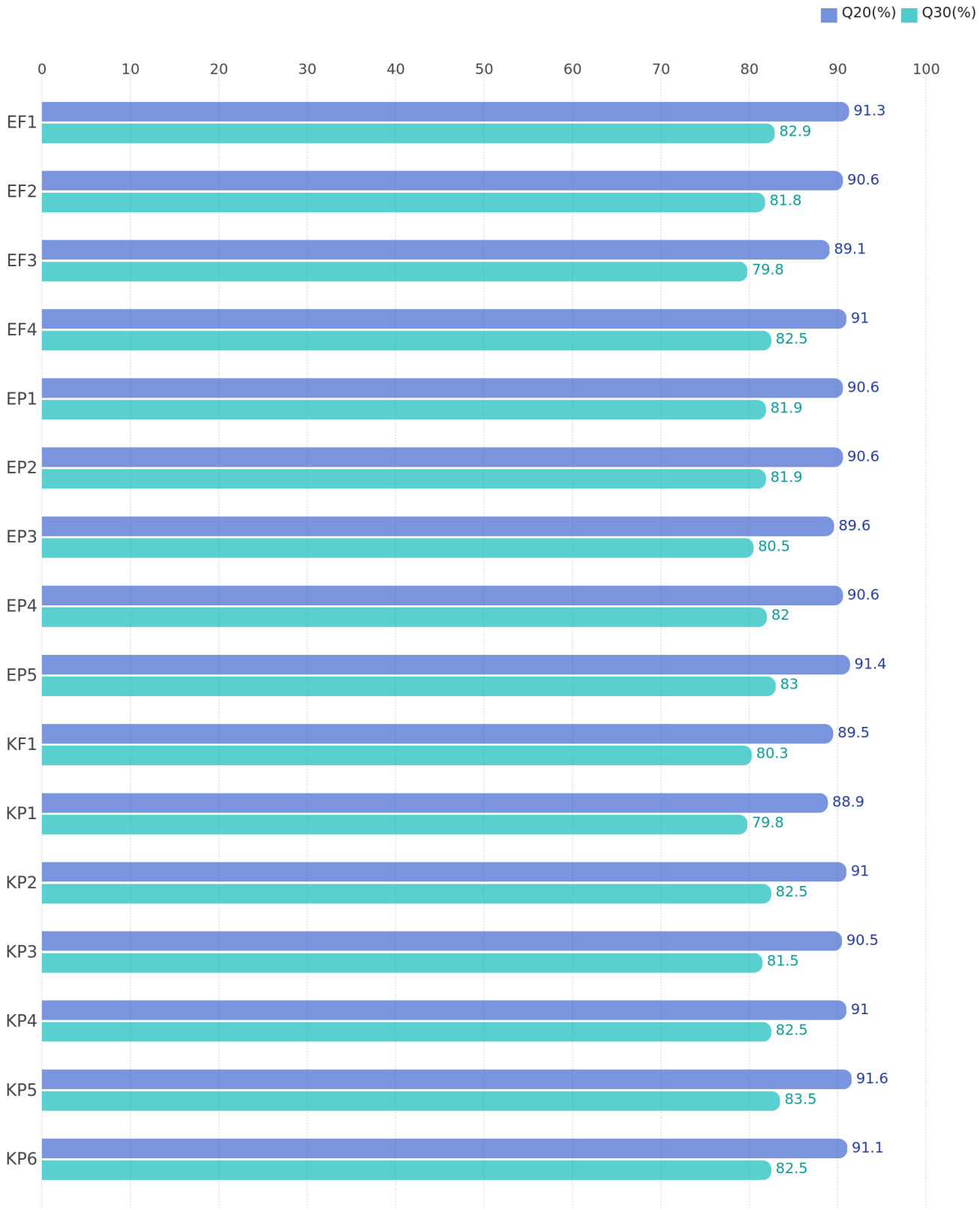
* Raw Data



Q20/Q30 (%)

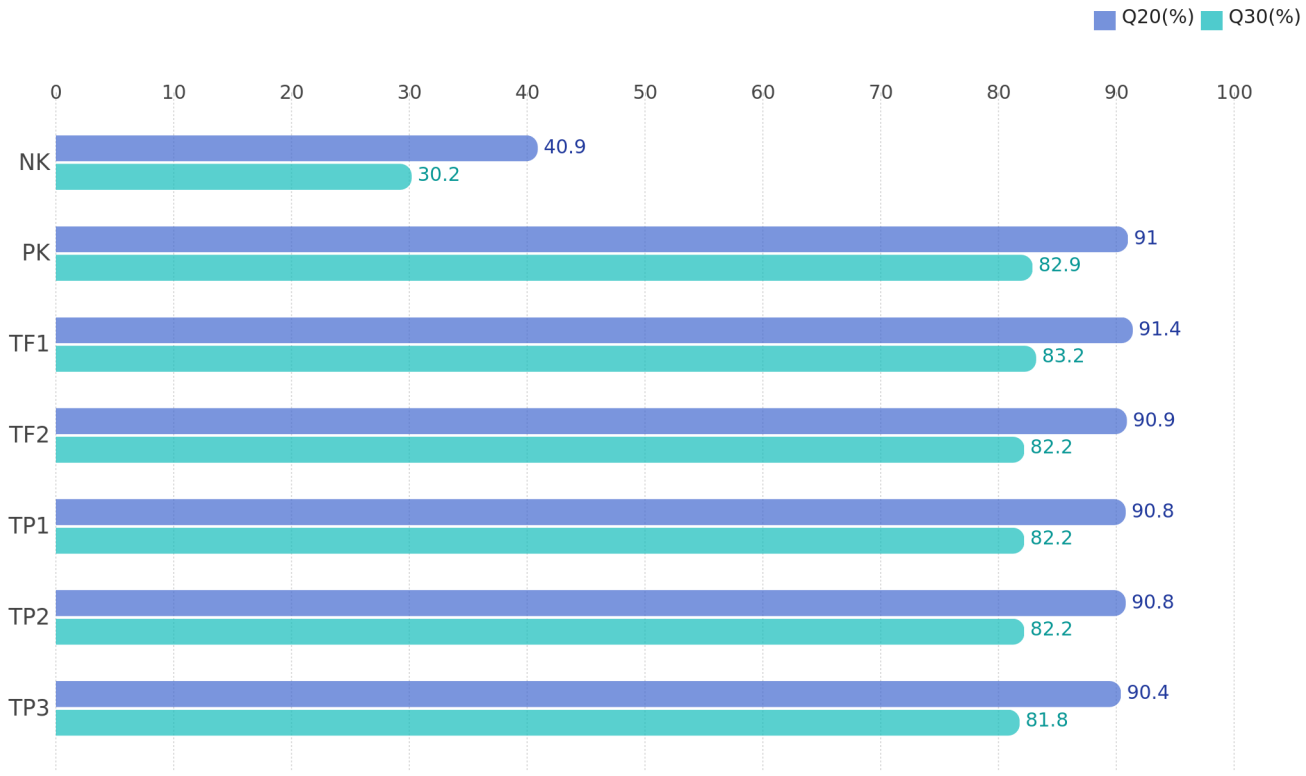
Total number of samples : 23

* 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
HN00230849.zip	416,656,586	421ab0bf3e141a1e9295f49807142867

File Name	File Size(byte)	md5sum
TF1_1.fastq.gz	9,046,328	8475d04e38cef8f4abb9d5ac68cdf430
TF1_2.fastq.gz	12,593,125	4ff7155e7925fb5c6f5fad4a5729eafd
TF2_1.fastq.gz	8,693,842	77368ff8cce1f3a9907553312cd0c967
TF2_2.fastq.gz	11,956,069	cb1a0f9612323edd2b1e0e3665ba1322
KP1_1.fastq.gz	7,042,079	4bfc5285352d17c851e8845a5f2002de
KP1_2.fastq.gz	10,948,777	0891a98e1a01be602f1f109e8c2f2936
KP2_1.fastq.gz	7,274,079	f223dd08a6ca9ba7668e49d923d12349
KP2_2.fastq.gz	10,066,451	3f4fe2d4ff6c529e54ace0bfb54efbab
KP3_1.fastq.gz	7,058,792	807c414b454046ae9b2eb7b3dd42d831
KP3_2.fastq.gz	9,576,217	d766a63f2a0b6dc95f59b9a4db030450
KP4_1.fastq.gz	7,413,414	024d7b46e95f8742c18247a6f78c3cda
KP4_2.fastq.gz	9,759,523	cf2da007422bec370b178e7f178ea03e
KP5_1.fastq.gz	9,187,007	15d01d7d121ce3154779acf23655f3c6
KP5_2.fastq.gz	12,258,802	b2c106b7cc845b3301b78dd8b462f243
KP6_1.fastq.gz	8,486,080	dbb4b71760f559ac91787e86ef018ec0
KP6_2.fastq.gz	11,409,545	c1293af219659d872963824778cd3d22
EF1_1.fastq.gz	8,467,783	247c172d6d21cc71bebae8582eb46bf2
EF1_2.fastq.gz	11,432,096	1f76df8ba869161131b1e3c7b385eb3d
EF2_1.fastq.gz	9,429,920	a4bdeb43737dc778beb3cc1635f4f892
EF2_2.fastq.gz	13,361,490	f54743dc8972796120f39d15441426b4
EF3_1.fastq.gz	8,293,667	15b4e540501b00562ced36e54bbb0265
EF3_2.fastq.gz	11,870,334	032c8aa3e9b16bf92bee9f5a3b2975fb
EF4_1.fastq.gz	9,606,476	be6fc2d3825b60745646d33dfbbcdec0
EF4_2.fastq.gz	12,857,631	dc4702e1f5142bff129ca5ccb22234c2
TP3_1.fastq.gz	8,105,739	2914a8549a5d6b849200ae9cba93f82f
TP3_2.fastq.gz	11,156,799	971d4ce98661a772a149f85972f61899
TP2_1.fastq.gz	7,580,244	47ec7acdf7d99f16def61f053e7de1ed
TP2_2.fastq.gz	10,537,110	2c884ad2a9dd84492f9348a488938d9a

File Name	File Size(byte)	md5sum
TP1_1.fastq.gz	7,819,025	c7f5865e9e696381f0225bc98a260e47
TP1_2.fastq.gz	11,050,590	e64e98a483d7d2e2eb99068ac3a2bf1a
PK_1.fastq.gz	4,810,133	fa9dace6e697156deeeddeb831ce4bcd
PK_2.fastq.gz	7,990,957	9edcc522c4be0bb49f55c321b920db5d
KF1_1.fastq.gz	7,077,448	f6e65d74728ee33a078153613810cd57
KF1_2.fastq.gz	9,996,496	92f0b389c29fe5036b7aa7b6166cae4a
NK_1.fastq.gz	938,321	c175ec1c6fb0aedc7f688eb11d7dd09e
NK_2.fastq.gz	983,713	d8f6862558587a1f76be9b33b46b4865
EP4_1.fastq.gz	6,960,579	53d0f6d3bc6156406956c339aa9fba97
EP4_2.fastq.gz	10,001,509	ae0def4dc2706d6c9ebca7e2d35caf3c
EP5_1.fastq.gz	8,616,772	91d9839d4572162c9eefba26d4c22ef4
EP5_2.fastq.gz	11,670,350	71e8c4a7c0e6622b8155f165e297b24d
EP2_1.fastq.gz	8,135,095	a5b138081bd50530a58c52d894c72317
EP2_2.fastq.gz	11,287,533	39ac325911bc6bd606f384d95d33c557
EP3_1.fastq.gz	6,856,589	8ff5bf13f3093bbf5518d5bd21668f21
EP3_2.fastq.gz	9,942,378	4f3f92290a631b230b9b9923e1ca37eb
EP1_1.fastq.gz	7,127,956	1ffa999fe357e0a67c8a518bec6ae207
EP1_2.fastq.gz	9,914,533	e725523b677b25751ef634131eec6ce0

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

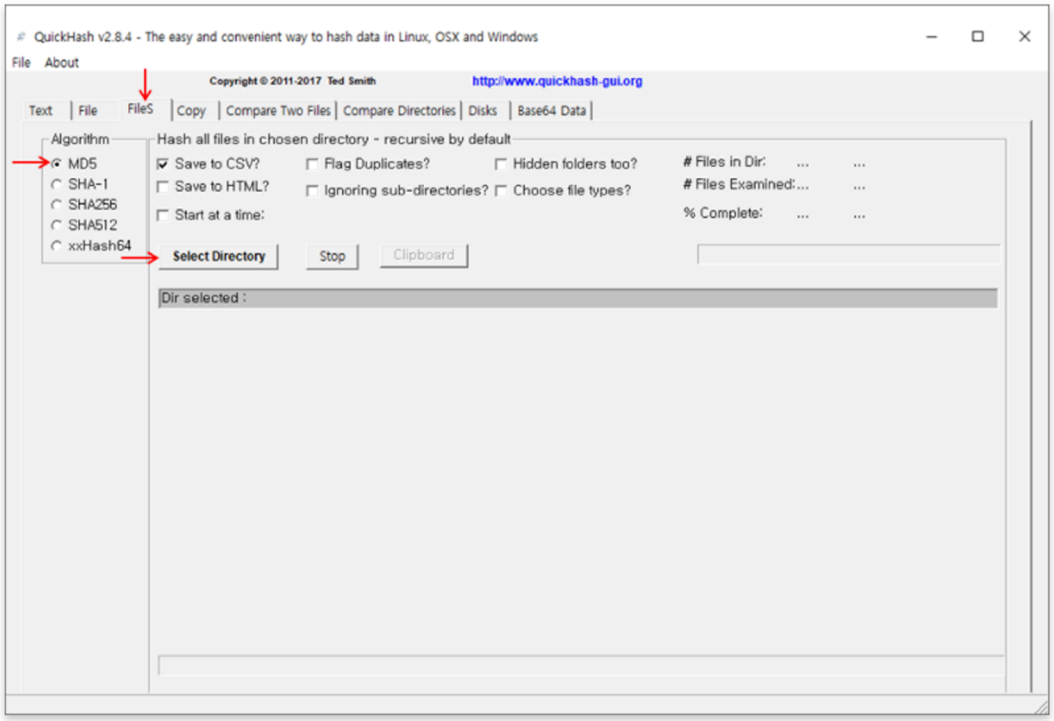
Windows does not provide a program for checking md5sum by default. An external program such as QuickHash-Windows can be used instead.

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.

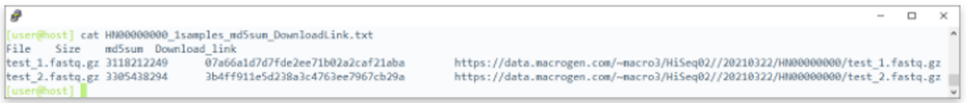
Checking md5 hash in a Linux system

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:

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



- Via HDD : **<OrderNumber>_#samples_md5sum.txt**

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

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 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)

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?**

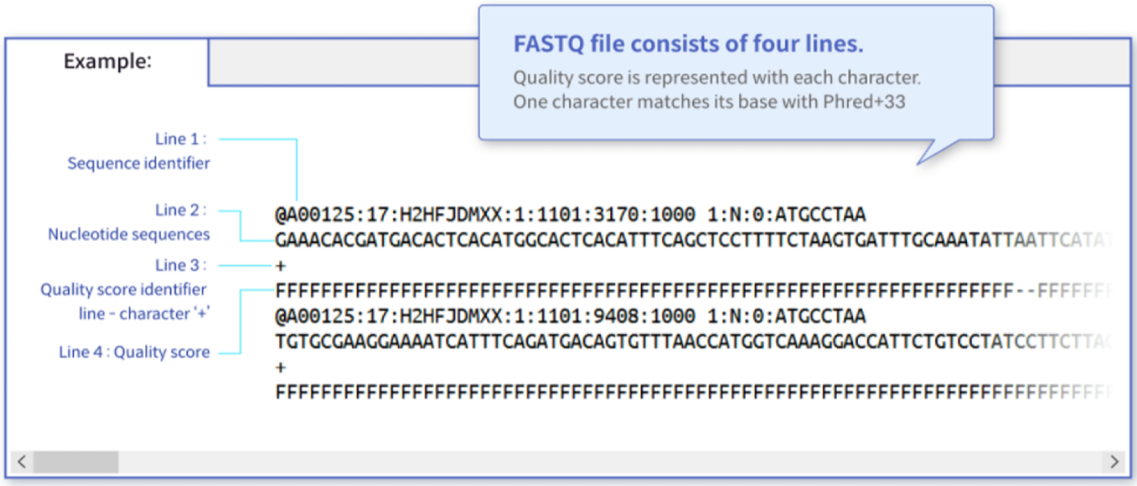
A Information on Illumina adapters can be found in this support document:
Adapter Sequences Intro

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	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. 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.

FASTQ Format



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%

Research use only

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

Macrogen Singapore

**Laboratory,
Business & Support Center**

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

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 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

Humanizing Genomics
macrogen