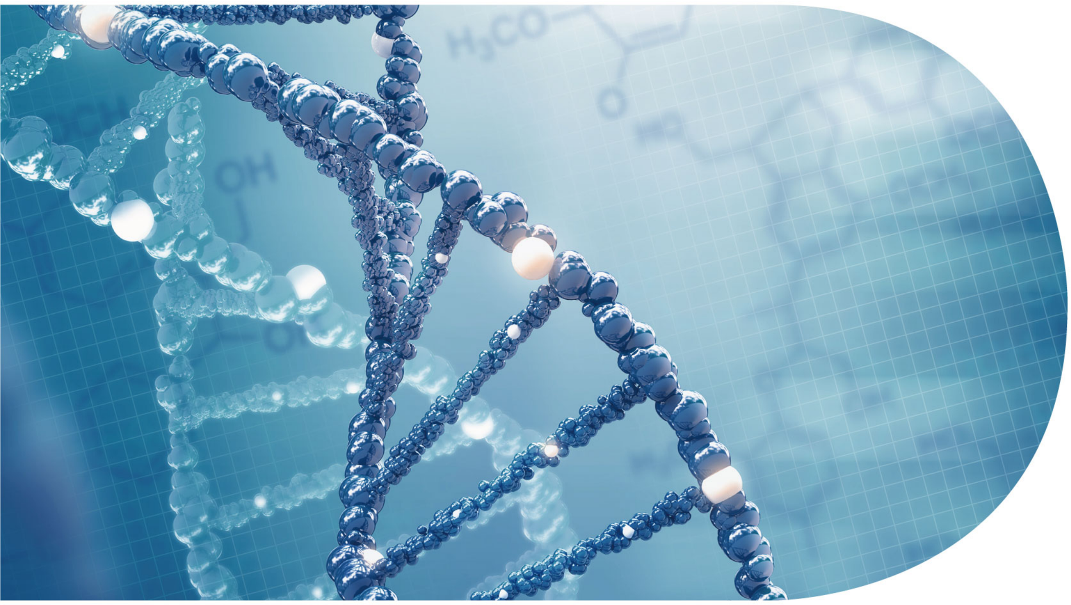
Research use only

Raw Data

Report

May 2020



Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen Oceania PL
Company / Institution	Macrogen Oceania PL
Order Number	HN00126866
Type of Read	Paired-end
Read Length	151
Number of Samples	18
Library Kit	TruSeq RNA Sample Prep Kit v2
Library Protocol	TruSeq RNA Sample Preparation v2 Guide, Part # 15026495 Rev. F
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis Results

Download link	File size	md5sum
P1A2_1.fastq.gz	1.9G	e1b119ef028b371b187a0cd0ae1685f2
P1A2_2.fastq.gz	1.9G	714ca007d6d3012811ba6c6f99848a81
P1A3_1.fastq.gz	1.9G	0c839bca5e505bb2eb6c909c0d0e2c9a
P1A3_2.fastq.gz	1.9G	093c76f39b430a7def08141f8a69521e
P1B2_1.fastq.gz	1.9G	328a97fbe7143beaede1f4d68bcc4d86
P1B2_2.fastq.gz	2.0G	22e02a9121f885666e1f4cf1e74d5112
P1B3_1.fastq.gz	1.6G	5efbb4b7ec12d811724175d8fe486496
P1B3_2.fastq.gz	1.6G	4481c63d454a33c8725a3aa3cace70da
P1C2_1.fastq.gz	1.9G	05149fb45a90f133708f3f4c8c918d69
P1C2_2.fastq.gz	1.9G	59b540ffdd73279d06748f830fe021fd
P1C3_1.fastq.gz	1.6G	d0ce85227a8dd3cc076e65395516789a
P1C3_2.fastq.gz	1.7G	2ba4e5a12d7dc9c4e94692578d83246b
P2A2_1.fastq.gz	1.9G	ba334190b2f37bda5364b23da6019d66
P2A2_2.fastq.gz	1.9G	c1d5bad7e7c01f0c0bc3d56026febf90
P2A3_1.fastq.gz	1.6G	05d912c4ab91e8f925b76f60585ea9cb
P2A3_2.fastq.gz	1.7G	9a87fc06635e396515ad7f214c643834
P2B2_1.fastq.gz	1.9G	6ab1fa1c19779033e2c076cffcbf2541
P2B2_2.fastq.gz	1.9G	20b9d9cb92cb3ab87492212469db755b
P2B3_1.fastq.gz	1.9G	6c73b72dc427dbefa22cea8d8332c384
P2B3_2.fastq.gz	2.0G	3ee03d49ed4daa294942d0e3fa68d8e1
P2C2_1.fastq.gz	1.9G	51f6fabb9ad6bcd5c55eb356aca1b421
P2C2_2.fastq.gz	1.9G	2d67871cb7626c1132fe500f72245468
P2C3_1.fastq.gz	1.8G	7ed061f4a1001fc330302c6efbc8b570
P2C3_2.fastq.gz	1.9G	3a3ef8afd308331e9861874992ee524c
P3A2_1.fastq.gz	1.9G	ea3a0ec2be04e5ce4824bae525c46d32
P3A2_2.fastq.gz	1.9G	891a3e9561ee97bc1d5a3c46d0a5e01e
P3A3_1.fastq.gz	2.1G	40eba46ba3cbe4dd02a9ccd86d282b2b
P3A3_2.fastq.gz	2.2G	1ca639b80c39d9b97a23f4b359cff6fe
P3B2_1.fastq.gz	1.8G	535672cdaed667246e6f469d7fbee66a
P3B2_2.fastq.gz	1.9G	a3340bf332b5231c7c838aca7bfa9bb7
P3B3_1.fastq.gz	1.9G	406b9a83d65ceaa3bd1e0f19452b4683
P3B3_2.fastq.gz	2.0G	8b677a384e847a4bf9f549c4610ee32e
P3C2_1.fastq.gz	1.5G	d586532625d0a0a4d1d13792c62af609

P3C2_2.fastq.gz	1.6G	8e20bfbcc96d3d122697e8d0708ad9fa
P3C3_1.fastq.gz	1.9G	9d1b89ac658b41b1d728b6a3e4443894
P3C3_2.fastq.gz	1.9G	f392db46588deb8ee8481735fd296bc5

- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please contact us.

2. Experimental Methods and Workflow

2. 1. Experiment Overview

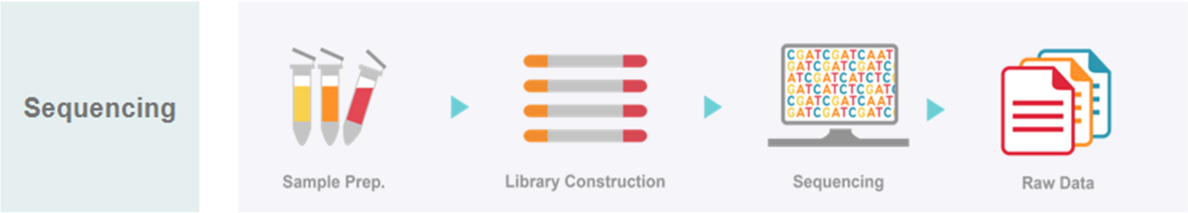


Fig1. Experiment overview

The Illumina NGS workflow includes 4 basic steps :

1) Sample Preparation

For library construction, DNA/RNA is extracted from a sample. After performing quality control (QC), qualified samples proceed to library construction.

2) Library Construction

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 that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

3) Sequencing

For cluster generation, the library is loaded into 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. When cluster generation is complete, the templates are ready for sequencing.

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 persent 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.

4) Raw data

Sequencing data is converted into raw data for the analysis.

2. 2. Generation of Raw Data

The Illumina sequencer generates raw images utilizing sequencing control software for system control and base calling through an integrated primary analysis software called RTA (Real Time Analysis). The BCL (base calls) binary is converted into FASTQ utilizing illumina package bcl2fastq. Adapters are not trimmed away from the reads.

3. Summary of Produced Data

3. 1. Raw Data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 18 samples. For example, in P1A2, 59,887,438 reads are produced, and total read bases are 9.0G bp. The GC content (%) is 50.35% and Q30 is 94.8%.

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
P1A2	9,043,003,138	59,887,438	50.35	49.65	98.28	94.8
P1A3	8,976,955,436	59,450,036	50.4	49.6	98.35	94.95
P1B2	8,134,225,946	53,869,046	50.02	49.98	98.19	94.55
P1B3	6,844,404,784	45,327,184	50.29	49.71	98.3	94.85
P1C2	9,008,294,580	59,657,580	50.44	49.56	98.18	94.59
P1C3	7,098,403,696	47,009,296	50.8	49.2	98.31	94.91
P2A2	9,018,764,014	59,726,914	49.75	50.25	98.26	94.8
P2A3	7,072,656,686	46,838,786	50.27	49.73	98.3	94.84
P2B2	8,965,561,882	59,374,582	50.55	49.45	98.23	94.7
P2B3	8,242,593,512	54,586,712	49.85	50.15	98.25	94.72
P2C2	8,952,539,944	59,288,344	50.13	49.87	98.26	94.75
P2C3	7,860,717,230	52,057,730	50.37	49.63	98.28	94.79
P3A2	8,972,362,016	59,419,616	49.78	50.22	98.23	94.69
P3A3	9,032,326,532	59,816,732	49.84	50.16	98.25	94.76
P3B2	7,796,249,290	51,630,790	50.33	49.67	98.31	94.88
P3B3	8,174,337,888	54,134,688	50.58	49.42	98.27	94.75
P3C2	6,540,168,172	43,312,372	50.79	49.21	98.32	94.9
P3C3	8,978,070,118	59,457,418	49.47	50.53	98.3	94.85

- Sample ID : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. For Illumina paired-end sequencing, this value refers to the sum of read 1 and read 2.
- GC(%) : GC content.
- AT(%) : 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.

3. 2. Total Read Bases

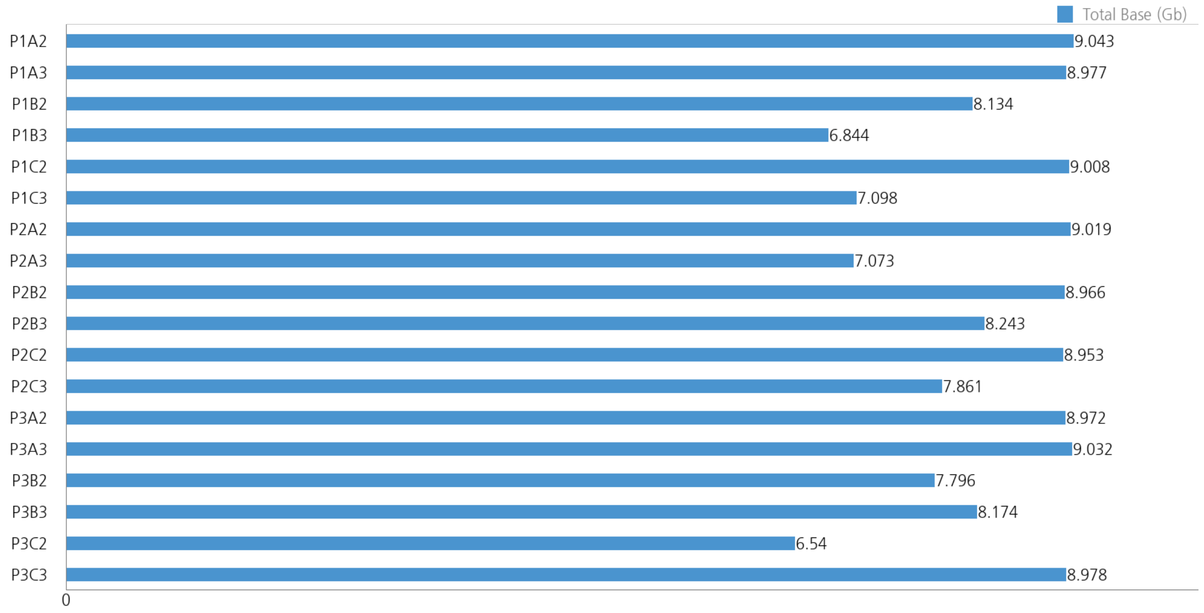


Figure 2.Throughput of Raw data

3. 3. Total Reads

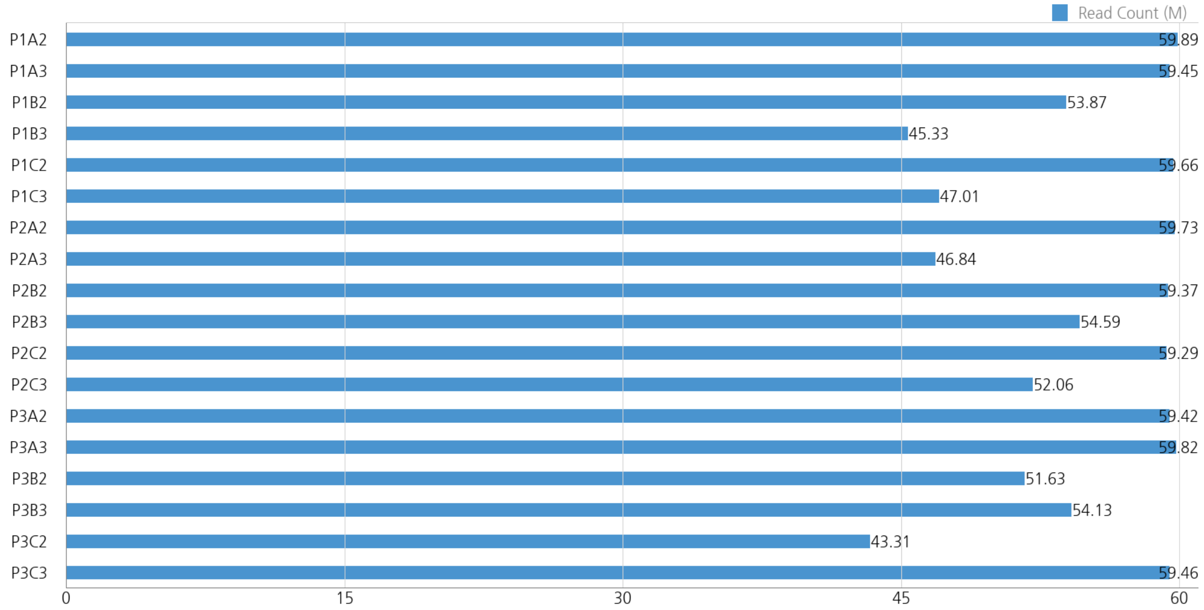


Figure 3. Total read count of Raw data

3. 4. GC/AT Content

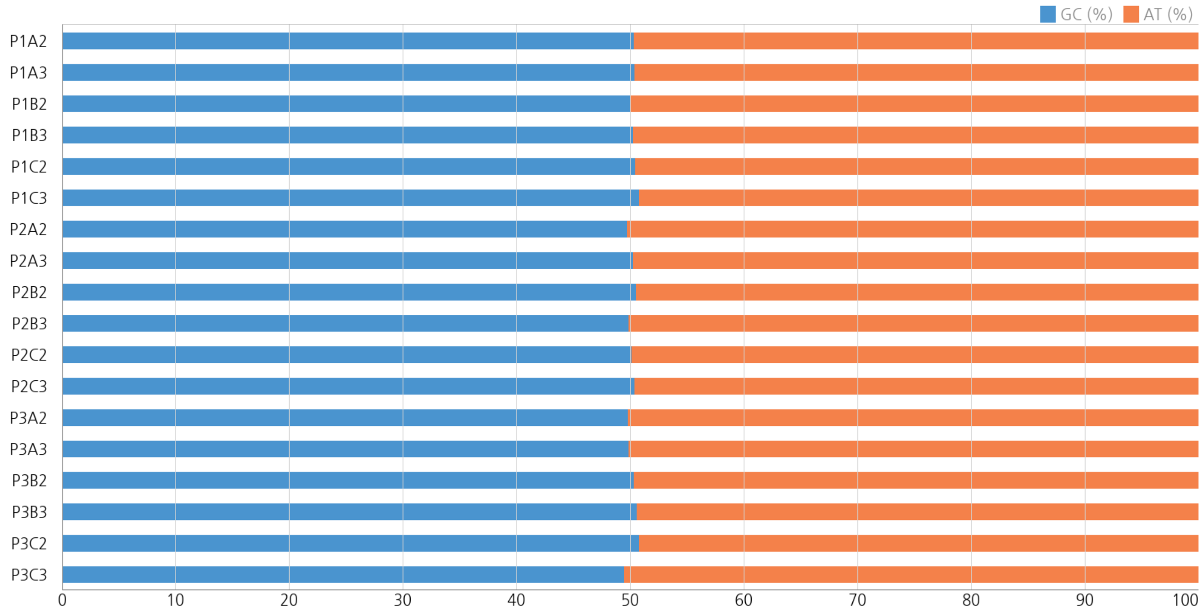


Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

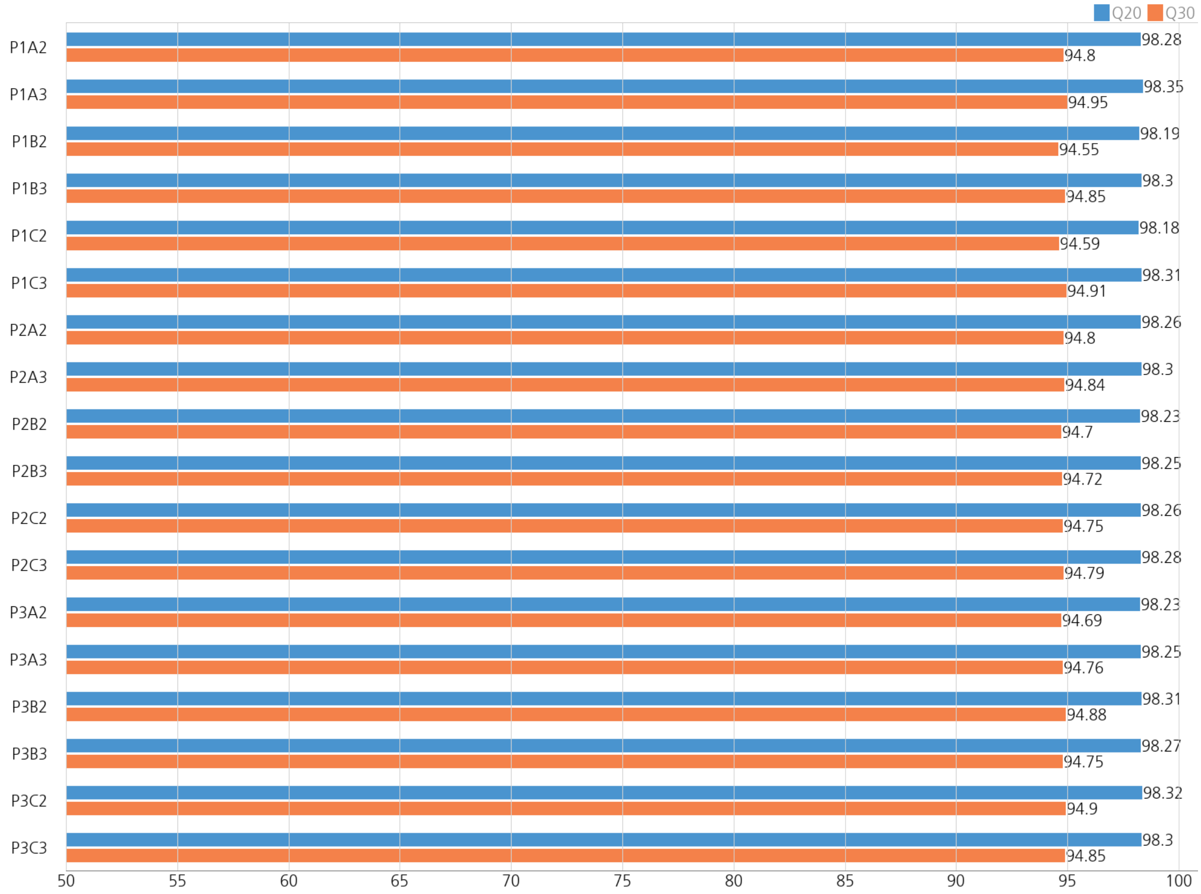


Figure 5. Q20/Q30 scores of Raw data

4. Appendix

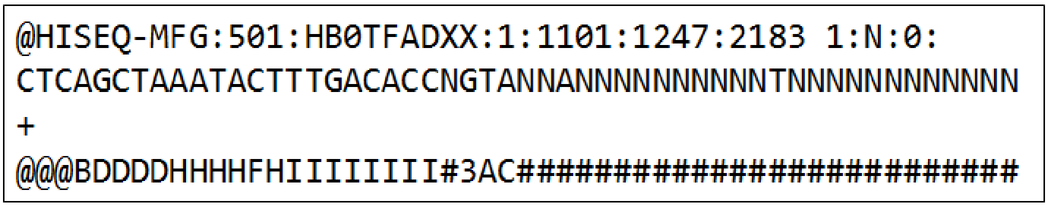
4. 1. FAQ

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

A: As the large size zip files provided by our company are hard to process in the Windows environment, we highly recommend using Linux environment for a smoother operation.

4. 2. FASTQ File

Example of FASTQ



```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNNTNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIIII#3AC#########################
```

FASTQ file is composed of four lines.
Line 1 : ID line includes information such as flow cell lane information.
Line 2 : Sequences line.
Line 3 : Separator line (+ mark).
Line 4 : Quality values line about sequences.

4. 3. Phred Quality Score Chart

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	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

- Encoding : Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

Research use only

Macrogen Europe
Amsterdam
Macrogen
Macrogen Spain
Madrid
Macrogen Japan
Kyoto
Macrogen Singapore
Singapore
Psomagen
Rockville

HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support 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 31, 1105 AZ, 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

3F Kyoto University International Science
Innovation Bldg.
36-1 Yoshida-honmachi, Sakyo-ku,
Kyoto 606-8501 JAPAN
Tel: +81-75-746-2773
Email: customer@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

Humanizing Genomics
macrogen