Research use only

Whole Genome De novo Sequencing

Report

March 2020

Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen
Company/Institute	Macrogen Corp.
Order Number	1234APB-5678
Sample	Macrogen
Type of Analysis	De novo assembly,Annotation
Type of Sequencer	PacBio RSII

Table of Contents

1. Data Download

Download link	File size	md5sum
PacBio raw data (1)	12G	1a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p
Analysis Results	1.2M	1a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p
Functional Annotation	1.2M	1a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p
COG Analysis	1.2M	1a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p

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 email (ngskr@macrogen.com) or contact our sales team.

2. Sequencing and Analysis Workflow

Figure 1. Workflow overview

2. 1. Sequencing & Preprocessing

A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell (TM) template. Polymerase reads are most useful for quality control of the instrument run. Polymerase read metrics primarily reflect movie length and other run parameters rather than insert size distribution. Polymerase reads are trimmed to include only the high quality region; they include sequences from adapters; and can further include sequence from multiple passes around a circular template.

Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell (TM) template and no adapter sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for de novo assembly.

2. 2. Analysis

2. 2. 1. De novo Assembly

At first, preassembly step is performed. It is accomplished by mapping single pass reads to seed reads, which represent the longest portion of the read length distribution. Subsequently, a consensus sequence of the mapped reads is generated, resulting in long and highly accurate fragments of the target genome.

The next step is correcting and filtering reads. Some reads that are fully contained in other reads do not provide extra information for constructing the genome, so they are filtered. And reads that have too high or too low overlaps are also filtered.

After then, given the overlapping data, they contain the information of each contig. So we can construct contigs.

3. Summary of Data Production

3. 1. Subreads Filtering

Table 1. Stats of filtered subreads

Mean subread length	12,345	N50	12,345
Total subread bases	1,234,567,890	**Total subreads**	123,456

- Mean subread length : The mean length of the subreads that passed filtering
- N50 : 50% of all bases come from subreads longer than this value
- Total subread bases : The total number of bases in the subreads that passed filtering
- Total subreads : The total number of subreads that passed filtering

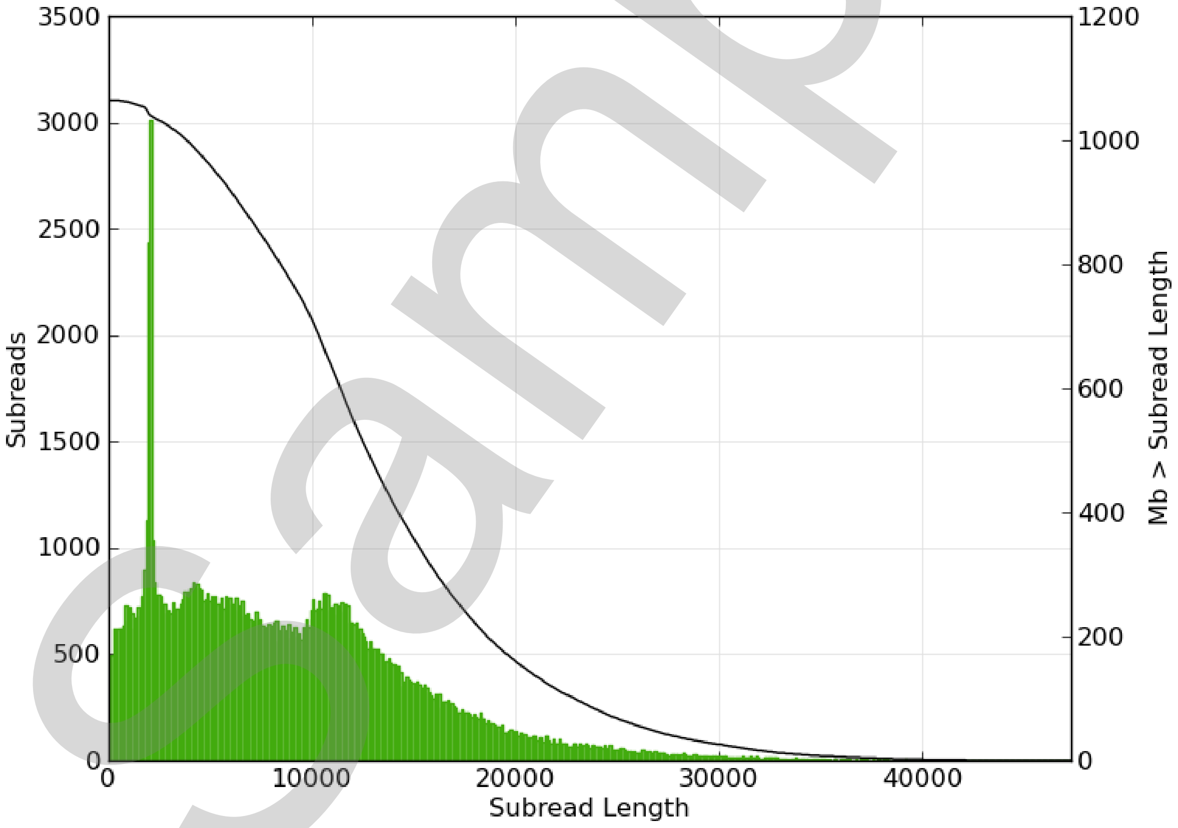


Figure 2. Filtered subread length distribution

4. Analysis Results

4. 1. De novo Assembly

Bioinformatics software such as HGAP, FALCON, CANU, and Unicycler can assemble the PacBio long-reads. In this analysis, HGAP3 was used and the detailed are attached in the appendix.

When both ends of the contig overlap each other, the contig is regarded as a circular form. On the other hand, if there are no overlaps, the contig might have been originally linear or there might be gaps at the end of the contig.

The assembly results are summarized in the table below.

Table 2. Summary of assembly

Contigs	Total contig bases	N50	Max length	Min length	Mean length
2	3,456,789	2,345,678	2,345,678	23,456	2,345,678

- Contigs : The number of contigs assembled
- Total contig bases : The total length of contigs
- N50 : 50% of all bases come from contigs longer than this value.
- Max length : The length of maximum contig
- Min length : The length of minimum contig
- Mean length : The average length of contigs assembled

Table 3. Result of assembly: 2 contigs were formed

Contig name	Length	GC (%)	Depth	Circular	Alias
contig1	2,345,678	12.3	456	Yes	Chromosome1
contig2	23,456	12.3	456	Yes	Plasmid2
Total	3,456,789	12.3	456		

- Length : The number of bases in each contig
- GC (%) : GC content
- Depth : The number of reads that aligned to each contig
- Circular : 5' end and 3' end are connected
- Alias : The alias is named based on the BLASTN (v2.7.1+) result

The following two conditions are used to create an alias:

a. Query cover 80% or more

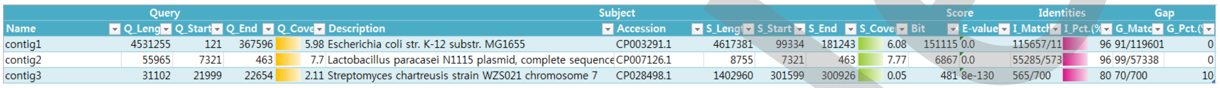
b. Similarity between genome size

If both conditions are met, it is named Chromosome or Plasmid. If not, named it Contig.

4. 2. Assembly Validation

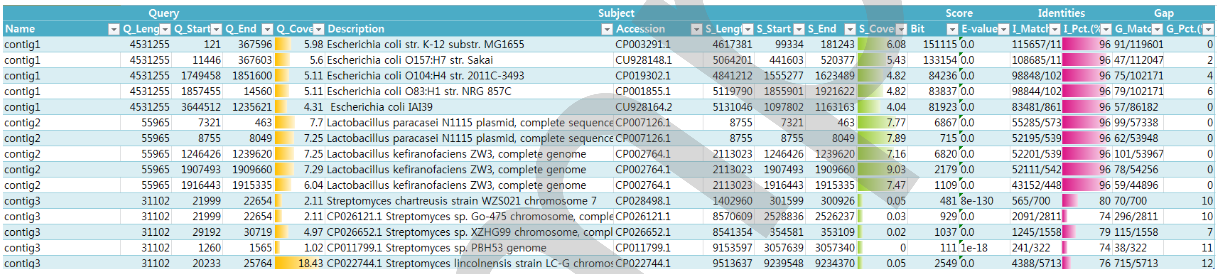
4. 2. 1. BLAST Results

After complete genome or draft genome was assembled, BLAST analysis was carried out to identify to which species each scaffold show similarity. Best hit and top 5 hit results were identified using NCBI NT database. Each result was prepared separately by the sheet of excel. Following is the example.



Query					Subject						Score		Identities		Gap	
Name	Q_Leng	Q_Start	Q_End	Q_Cove	Description	Accession	S_Lengt	S_Start	S_End	S_Cove	Bit	E-value	I_Match	I_Pct.(%	G_Matc	G_Pct.(
contig1	4531255	121	367596	5.98	Escherichia coli str. K-12 substr. MG1655	CP003291.1	4617381	99334	181243	6.08	151115	0.0	115657/11	96	91/119601	0
contig2	55965	7321	463	7.7	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	7321	463	7.77	6867	0.0	55285/573	96	99/57338	0
contig3	31102	21999	22654	2.11	Streptomyces chartreusis strain WZS021 chromosome 7	CP028498.1	1402960	301599	300926	0.05	481	8e-130	565/700	80	70/700	10

Figure 3. Best hit result example



Query					Subject						Score		Identities		Gap	
Name	Q_Leng	Q_Start	Q_End	Q_Cove	Description	Accession	S_Lengt	S_Start	S_End	S_Cove	Bit	E-value	I_Match	I_Pct.(%	G_Matc	G_Pct.(
contig1	4531255	121	367596	5.98	Escherichia coli str. K-12 substr. MG1655	CP003291.1	4617381	99334	181243	6.08	151115	0.0	115657/11	96	91/119601	0
contig1	4531255	11446	367603	5.6	Escherichia coli O157:H7 str. Sakai	CU928148.1	5064201	441603	520377	5.43	133154	0.0	108685/11	96	47/112047	2
contig1	4531255	1749458	1851600	5.11	Escherichia coli O104:H4 str. 2011C-3493	CP019302.1	4841212	1555277	1623489	4.82	84236	0.0	98848/102	96	75/102171	4
contig1	4531255	1857455	14560	5.11	Escherichia coli O83:H1 str. NRG 857C	CP001855.1	5119790	1855901	1921622	4.82	83837	0.0	98844/102	96	79/102171	6
contig1	4531255	3644512	1235621	4.31	Escherichia coli IAI39	CU928164.2	5131046	1097802	1163163	4.04	81923	0.0	83481/861	96	57/86182	0
contig2	55965	7321	463	7.7	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	7321	463	7.77	6867	0.0	55285/573	96	99/57338	0
contig2	55965	8755	8049	7.25	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	8755	8049	7.89	715	0.0	52195/539	96	62/53948	0
contig2	55965	1246426	1239620	7.25	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1246426	1239620	7.16	6820	0.0	52201/539	96	101/53967	0
contig2	55965	1907493	1909660	7.29	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1907493	1909660	9.03	2179	0.0	52111/542	96	78/54256	0
contig2	55965	1916443	1915335	6.04	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1916443	1915335	7.47	1109	0.0	43152/448	96	59/44896	0
contig3	31102	21999	22654	2.11	Streptomyces chartreusis strain WZS021 chromosome 7	CP028498.1	1402960	301599	300926	0.05	481	8e-130	565/700	80	70/700	10
contig3	31102	21999	22654	2.11	CP026121.1 Streptomyces sp. Go-475 chromosome, comple	CP026121.1	8570609	2528836	2526237	0.03	929	0.0	2091/2811	74	296/2811	10
contig3	31102	29192	30719	4.97	CP026652.1 Streptomyces sp. XZHG99 chromosome, compl	CP026652.1	8541354	354581	353109	0.02	1037	0.0	1245/1558	79	115/1558	7
contig3	31102	1260	1565	1.02	CP011799.1 Streptomyces sp. PBH53 genome	CP011799.1	9153597	3057639	3057340	0	111	1e-18	241/322	74	38/322	11
contig3	31102	20233	25764	18.43	CP022744.1 Streptomyces lincolnensis strain LC-G chromos	CP022744.1	9513637	9239548	9234370	0.05	2549	0.0	4388/5713	76	715/5713	12

Figure 4. Top 5 hit result example

Because the BLAST analysis is based on registered information, it is difficult to determine the information of the species that is not registered. In particular, the assembly results could be matched with a relative species or an evolutionarily distant species due to sequence differences or error that may occur during the assembly process. Therefore, it would be more appropriate to use the analysis results to identify patterns rather than to use it as an absolute criterion for species determination.

The BLAST results are in the "Analysis Result" file

4. 2. 2. BUSCO Results

In order to assess the completeness of the genome assembly, BUSCO analysis was performed based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs.

The recovered matches are classified as 'Complete' if their lengths are within the expectation of the BUSCO profile match lengths. If these are found more than once, they are classified as 'duplicated'. The matches that are only partially recovered are classified as 'Fragmented', and BUSCO groups for which there are no matches that pass the tests of orthology are classified as 'Missing'.

Higher complete BUSCOs may indicate good assembly, however, for species other than model organisms, relatively low BUSCOs can appear due to characteristics of the sample as well as the incompleteness of the assembly.

By default, bacteria or eukaryota DB was used for analysis.

Table 4. BUSCO analysis result

Used Lineage : bacteria_odb9 (number of species: 3663, number of BUSCOs: 148)

Status	# of BUSCOs	Percentage
Complete BUSCOs (C)		
Complete and single-copy BUSCOs (S)	137	92.57 %
Complete and duplicated BUSCOs (D)	0	0.00 %
Fragmented BUSCOs (F)	1	0.68 %
Missing BUSCOs (M)	10	6.76 %
Total BUSCO groups searched	148	100.00 %

- Status : A quantitative assessment list of the completeness in terms of expected gene content

  The following two conditions are used to create a status:
  a. Expected range of scores
  b. Expected range of length alignments
  If both conditions are met, it is classified as Complete (These complete busco matches are either single-copy or duplicated). If length alignments is not met, it is classified as Fragmented. If both conditions are not met, it is classified as Missing.

- # of BUSCOs : Identified count in sample
- Percentage : Identified percentage in sample

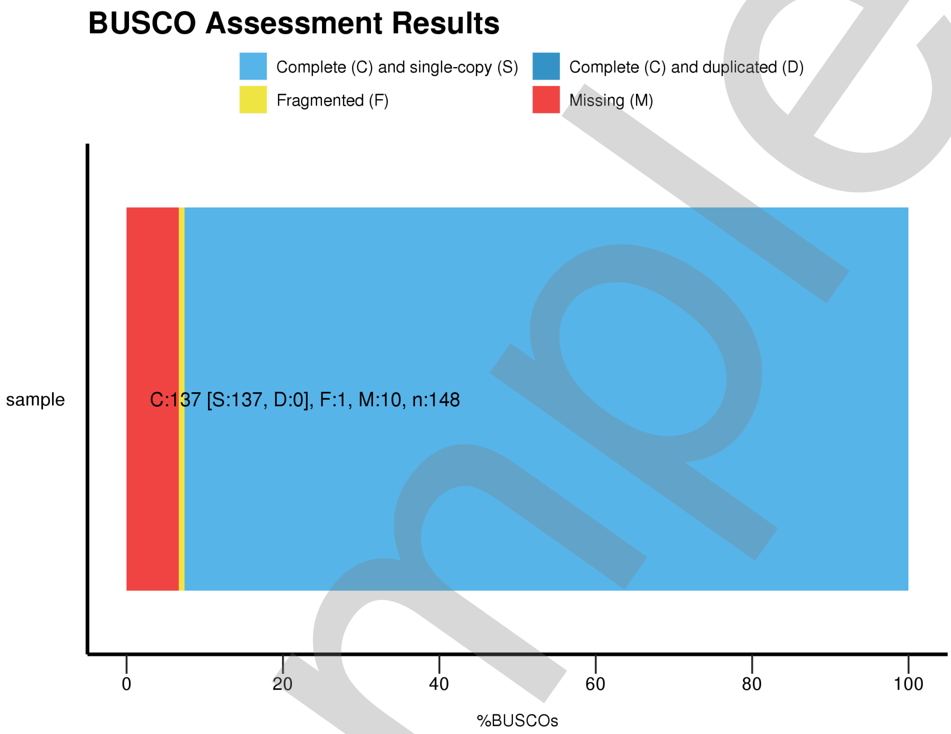


Figure 5. BUSCO result plot

5. Details of File Extensions

Raw Data

File extensions	Description
Bax.h5	The files contain base call information from the sequencing run.
Bas.h5	The file is essentially a pointer to the three bax.h5 files.
mcd.h5	Gridding identifies the location of the ZMWs with respect to the camera.
Matadata.xml	The *.metadata.xml contains top level information about the data, including what sequencing enzyme and chemistry were used, sample name and other metadata.
Fastq/fasta	The files contain subreads sequence in FASTQ/FASTA format.

Alignment and Assembly results

File Extensions	Details
consensus.fasta	Whole nucleotide sequence.

6. Appendix

6. 1. FAQ

Q: How can I open the sequence files?
A: After unzipping the file, the data can be opened with any kind of text editor. However, if you are dealing with big sized data, we recommend using Vim (http://www.vim.org/) or Notepad++ (http://notepad-plus-plus.org/)

Q: How can I see the annotation results?
A: Since all the annotation result files are text files, they can be viewed with Vim, Notepad++, Microsoft word, Excel, and any program that can open text files.

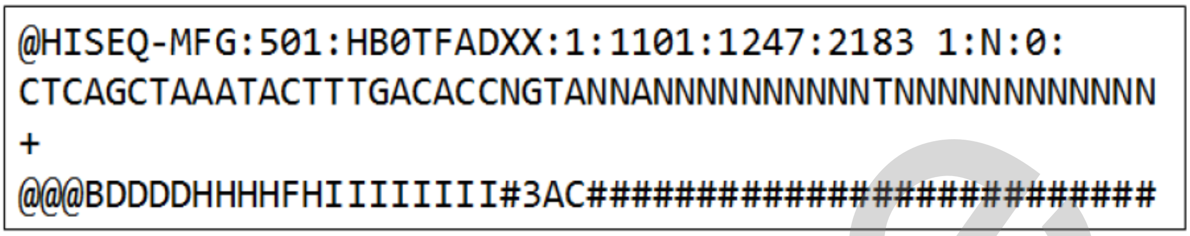
Q: How can I view annotation gene with sequence at the same time?
A: You can view the result by opening .gbk file with Genome browser such as Artemis.
(https://www.sanger.ac.uk/resources/software/artemis/)

Q: How can I register the analyzed genome to NCBI?
A: First you have to sign up for NCBI. Then you can register the genome through Genome (WGC) submission portal (https://submit.ncbi.nlm.nih.gov/subs/wgs/). In case of microorganism, you can use specific genome annotation pipeline provided by NCBI.

Q: Is there any other gene annotation pipeline that can be used?
A: You can use Prokaryotic Genome Annotation Pipeline (PGAP) (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/) of NCBI. When registering the genome, you can decide whether you are going to use it or not. Additionally you can request through NCBI.

6. 2. FASTQ File

6. 2. 1. Example of FASTQ file format



```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNTNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIII#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

6. 2. 2. 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 -10log10P, 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

6. 3. Programs used in Analysis

6. 3. 1. De novo Assembly

6. 3. 1. 1. RS HGAP Assembly

LINK http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/

SMRT Portal (v2.3) allows the execution of all HGAP steps in the web-based GUI. HGAP (v3.0) performs high quality de novo assembly using a single PacBio library preparation. It includes preassembly, de novo assembly with PacBio's assembleUnitig, assembly polishing with Quiver, and a significant speed improvement for microbial assembly.

- Option Details (Default)

1. Filtering : PreAssembler Filter v1

Minimum Subread Length: 500bp
Minimum Polymerase Read Quality: 0.80
Minimum Polymerase Read Length: 100bp

2. Assembly : PreAssembler v2

Minimum Seed Read Length: 6000 bp
Number of Seed Read Chunks: 6
Alignment Candidates Per Chunk: 10
Total Alignment Candidates: 24
Minimum Coverage for Correction: 6

3. BLASR v1

Maximum Divergence (%): 30 %
Minimum Anchor Size: 12bp

6. 3. 2. Validation Check

6. 3. 2. 1. BLAST

LINK https://blast.ncbi.nlm.nih.gov/Blast.cgi

The Basic Local Alignment Search Tool (BLAST, v2.7.1+) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

6. 3. 2. 2. BUSCO

LINK http://busco.ezlab.org/

The Benchmarking Universal Single-Copy Orthologous (BUSCO, v3.0) can assess assembly quality by comparison between predicted genes from genome assembly and near-universal single-copy orthologs DB. If assembly display higher complete BUSCOs, the assembly could be regarded as assembly with good quality.

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
Email: ngs@macrogen.com
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

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**

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