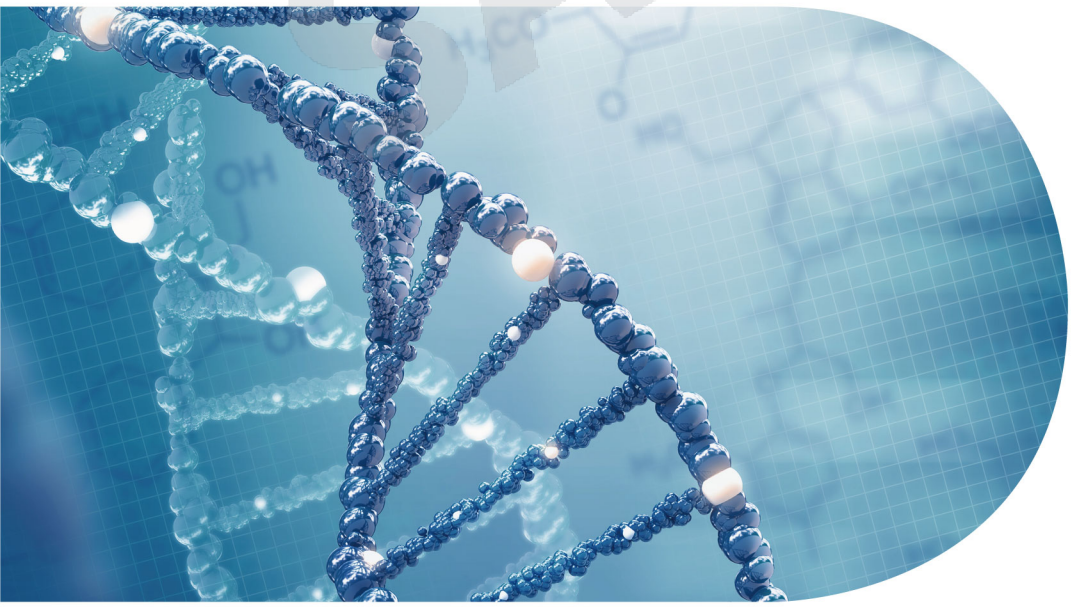
Research use only

Homo sapiens ChIP Sequencing

Report

April 2020



Humanizing Genomics macrogen

ver. MGCP_201901_hg19

Project Information

Client Name	TESTER
Company/Institution	Macrogen
Order Number	HN00000000
Species	*Homo sapiens*
Reference	hg19
Read Length	101
Number of Samples	3
Library Kit	TruSeq DNA Sample prep Kit
Library Protocol	TruSeq ChIP Sample Preparation Guide 15023092 Rev. B
Reagent	NovaSeq 6000 S4 Reagent Kit
Sequencing Protocol	NovaSeq 6000 System User Guide Document # 1000000019358 v02
Type of Sequencer	NovaSeq 6000
Sequencing Control Software	1000000019358 v02

Table of Contents

1. Analysis Methods and Workflow

Figure 1. Analysis Workflow

1. 1. Sequence quality check

1. 1. 1. FastqQC (version: 0.11.7)

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

LINK (http://www.bioinformatics.babraham.ac.uk/projects/fastqc

1. 1. 2. Trimmomatic (version: 0.38)

Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. These adapters can pose a real problem depending on the library preparation and downstream application.

LINK http://www.usadellab.org/cms/?page=trimmomatic

1. 2. Data analysis

1. 2. 1. Bowtie (version: 1.1.2)

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

LINK http://bowtie-bio.sourceforge.net/index.shtml

1. 2. 2. MACS2 (version: 2.1.1.20160309)

MACS captures the influence of genome complexity to evaluate the significance of enriched ChIP regions, and MACS improves the spatial resolution of binding sites through combining the information of both sequencing tag position and orientation. MACS can be easily used for ChIP-Seq data alone, or with control sample with the increase of specificity.

LINK https://github.com/taoliu/MACS

1. 2. 3. Picard (MarkDuplicates) (version: 0.118)

Identifies duplicate reads. This tool locates and tags duplicate reads (both PCR and optical/sequencing-driven) in a BAM or SAM file, where duplicate reads are defined as originating from the same original fragment of DNA. Duplicates are identified as read pairs having identical 5' positions (coordinate and strand) for both reads in a mate pair (and optinally, matching unique molecular identifier reads; see BARCODE_TAG option). Optical, or more broadly Sequencing, duplicates are duplicates that appear clustered together spatially during sequencing and can arise from optical/imagine-processing artifacts or from bio-chemical processes during clonal amplification and sequencing.

LINK http://broadinstitute.github.io/picard/

1. 2. 4. ChIPseeker (version: ChIPseeker v1.16.1)

ChIPseeker is an R package for annotating ChIP-seq data analysis. It supports annotating ChIP peaks and provides functions to visualize ChIP peaks coverage over chromosomes and profiles of peaks binding to TSS regions.

LINK http://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html

2. Summary of Data Production

Analyses were successfully performed on all 3 paired-ends samples as requested. Figure 2 below shows the amount throughput between raw data and trimmed data. Figure 3 shows the % Q30 score (% of bases with quality over phred score 30) per sample between raw and trimmed data.

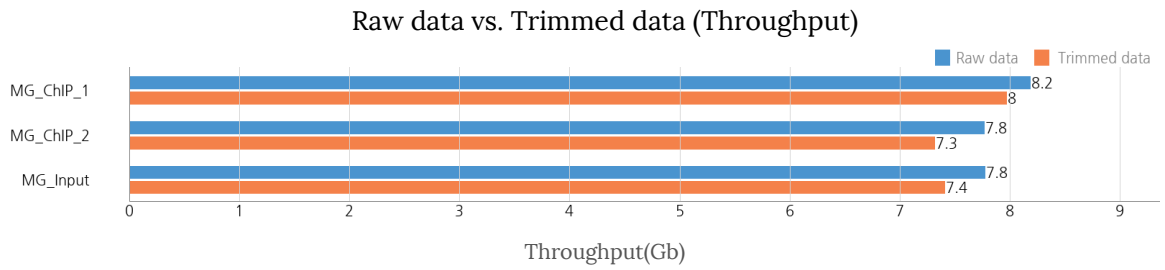


Figure 2. Throughput output of Raw and Trimmed data

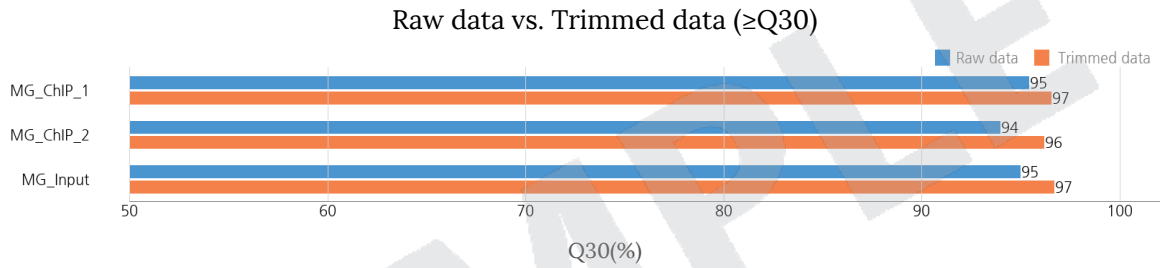


Figure 3. Q30 score of Raw and Trimmed data

After the read trimming, Bowtie (read mapping), Picard (remove duplicates), MACS2 (peak calling) and ChIPseeker (peak annotation) were used for downstream analysis.

2. 1. Raw Data Statistics

(Refer to Path: result_ChIPseq/0.Analysis_statistics/rawData/raw_throughput.stats)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 3 samples. For example, in MG_ChIP_1, 81,023,690 reads are produced, and total read bases are 8.2Gbp. The GC content (%) is 47.94% and Q30 is 95.41%.

Table 1. Raw data stats

Index	Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
1	MG_ChIP_1	8,183,392,690	81,023,690	47.94	97.28	95.41
2	MG_ChIP_2	7,766,432,168	76,895,368	44.34	96.31	93.95
3	MG_Input	7,771,731,032	76,947,832	40.95	96.94	94.95

(* Total read bases = Total reads x Read length)

- Total read bases: Total number of bases sequenced
- Total reads: Total number of reads
- GC (%): GC content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20
- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

2. 2. Average Base Quality at Each Cycle

(Refer to Path: 0.Analysis_statistics/rawData/A_fastqc/)

The quality of produced data is determined by the phred quality score at each cycle. Box plot containing the average quality at each cycle is created with FastQC.

The x-axis shows number of cycles and y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads over score of 20 are accepted as good quality.

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc

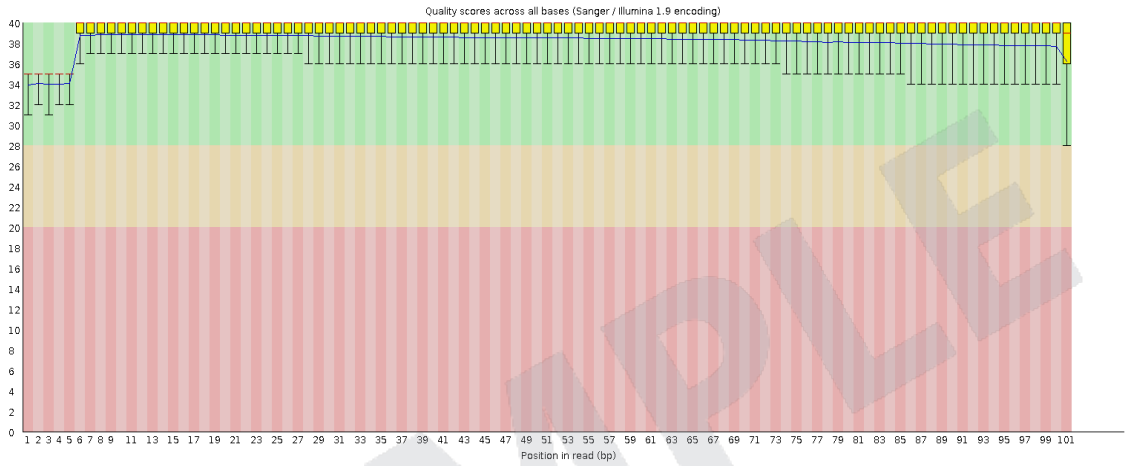


Figure 4. Read quality at each cycle of MG_ChIP_1 (read1)

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

2. 3. Trimming Data Statistics

(Refer to Path: result_ChIPseq/0.Analysis_statistics/trimmedData/trim_throughput.stats)

Trimmomatic program is used to remove adapter sequences and bases with base quality lower than three from the ends. Also using sliding window method, bases of reads that does not qualify for window size 4, and mean quality 15 are trimmed. Afterwards, reads with length shorter than 36bp are dropped to produce trimmed data.

Table 2. Trimming Data Stats

Index	Sample id	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
1	MG_ChIP_1	7,965,491,783	79,159,406	47.91	98.04	96.51
2	MG_ChIP_2	7,313,507,591	72,731,902	44.22	97.82	96.17
3	MG_Input	7,408,120,444	73,954,220	40.68	98.10	96.67

- Total read bases: Total number of read bases after trimming
- Total reads: Total number of reads after trimming
- GC (%): GC Content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20
- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

2. 4. Average Base Quality at Each Cycle after Trimming

(Refer to Path: result_ChIPseq/0.Analysis_statistics/trimmedData/A_fastqc/)

Figure 5 and 6 show average base quality at each cycle after trimming.

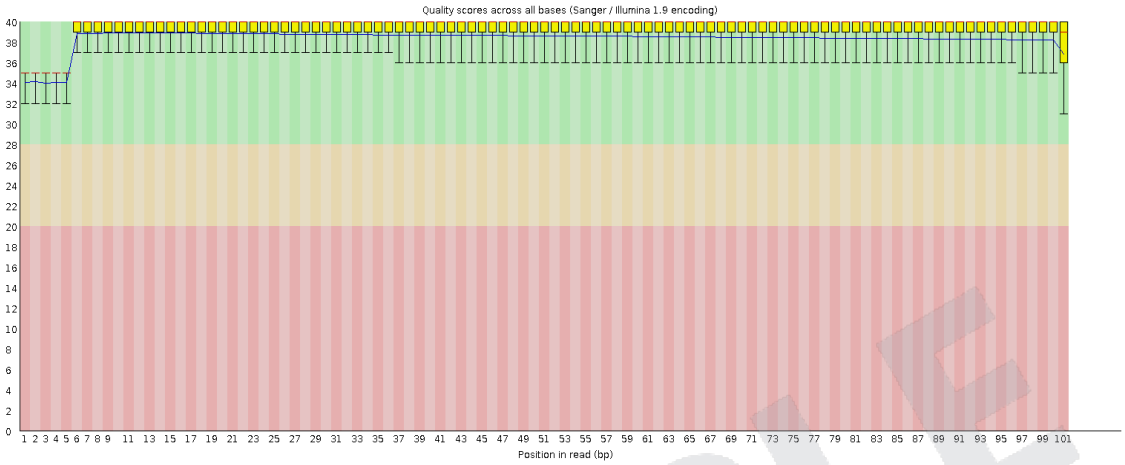


Figure 5. Average base quality of MG_ChIP_1 (read1) at each cycle after trimming

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

3. Reference Mapping Results

3. 1. Mapping Data Statistics

(Refer to Path: result_ChIPseq/0.Analysis_statistics/mapping.bowtie.stats)

In order to map reads obtained from ChIP sequencing, hg19 was used as a reference genome. Figure 6 and Table 3 shows the statistic obtained from Bowtie aligner. You can check number of processed reads, mapped reads, unmapped reads and reads removed by multiple mapping.

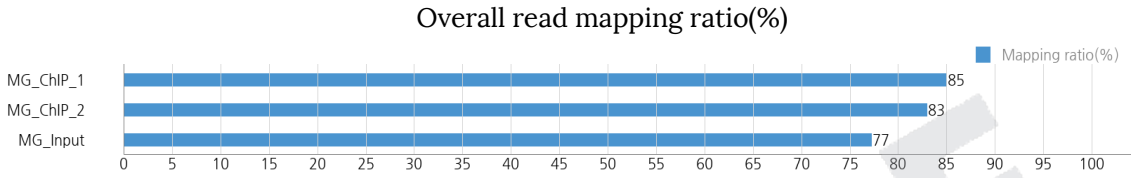


Figure 6. Overall read mapping ratio(%)

Table 3. Mapped Data Stats

Sample ID	# of processed reads	# of mapped reads	# of failed to align reads	# of suppressed reads by multiple mapping
MG_ChIP_1	79,159,406	67,247,538 (84.95%)	9,641,134 (12.18%)	2,270,734 (2.87%)
MG_ChIP_2	72,731,902	60,355,476 (82.98%)	10,049,532 (13.82%)	2,326,894 (3.20%)
MG_Input	73,954,220	57,092,338 (77.20%)	13,738,938 (18.58%)	3,122,944 (4.22%)

- Processed reads: Number of cleaned reads after trimming
- Mapped reads: Number of reads mapped to reference
- Read that failed to align: Number of reads that failed to align
- Suppressed multiple mapped reads: Number of reads removed due to multiple mapping

4. Data Download Information

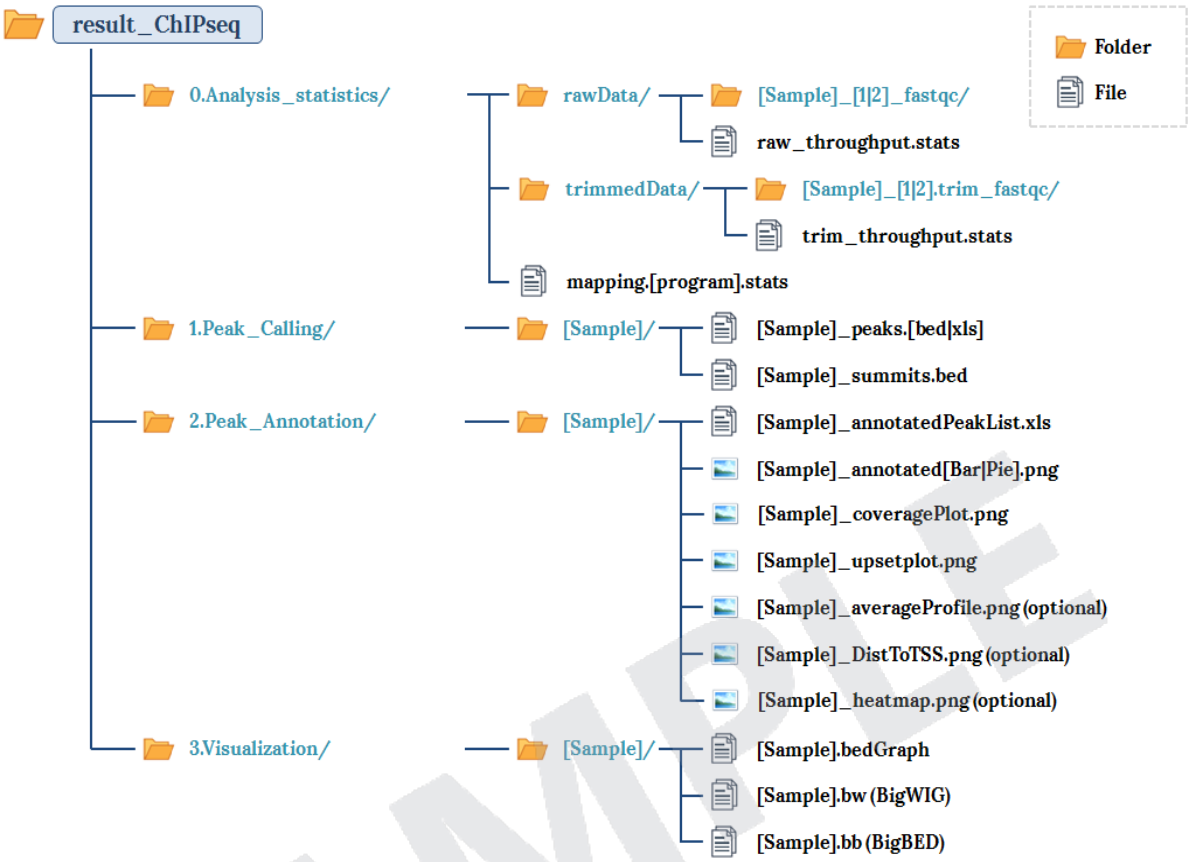
4. 1. Raw Data

Raw data is the FASTQ file that isn't trimmed adapter sequence.

Download link	File size	md5sum
MG_ChIP_1_1.fastq.gz	2.65G	dcd88d1bdee3919d7837b1fce3c590a4
MG_ChIP_1_2.fastq.gz	2.74G	0a6fa8b1b132955059019118cf902da0
MG_ChIP_2_1.fastq.gz	2.46G	7b32fb21351031c8024581bae07f9046
MG_ChIP_2_2.fastq.gz	2.61G	1df2f10ba801b60517d01c2fd26b35d3
MG_Input_1.fastq.gz	2.47G	b7b2b22febd0b7d384015a7746b733d0
MG_Input_2.fastq.gz	2.56G	4b11290d44741c23bf0aae7083a1cdd8

- 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.

4. 2. Analysis Results



Download link	File size
TEST_result_ChIPseq.zip (md5sum: a083973798046c6ee3b5df058259ec53)	2.24G

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

5. Appendix

5. 1. Phred Quality Score Chart

Phred quality score numerically express 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.

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

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

5. 2. File Description

5. 2. 1. [Sample]_peaks.xls

This file is a tabular file which contains information about called peaks. You can open it in excel and sort/filter using excel functions. Information include:

- chromosome name
- start position of peak
- end position of peak
- length of peak region
- absolute peak summit position
- pileup height at peak summit, -log10(pvalue) for the peak summit (e.g. pvalue = 1e-10, then this value should be 10)
- fold enrichment for this peak summit against random Poisson distribution with local lambda, -log10(qvalue) at peak summit

Coordinates in XLS is 1-based which is different with BED format.

5. 2. 2. [Sample]_summit.bed

This file is in BED format, which contains the peak summits locations for every peaks. The 5th column in this file is -log10pvalue the same as NAME_peaks.bed. If you want to find the motifs at the binding sites, this file is recommended. The file can be loaded directly to UCSC genome browser. Remove the beginning track line if you want to analyze it by other tools.

5. 2. 3. [Sample]_annotatedBar.png & [Sample]_annotatedPie.png

To annotate the location of a given peak in terms of genomic features which includes whether a peak is in the TSS, Exon, 5' UTR, 3' UTR, Intronic or Intergenic. Pie and Bar plot are supported to visualize the genomic annotation. The result of ChIPseeker.

5. 2. 4. [Sample]_averageProfile.png

Average profile of ChIP peaks binding to TSS region. The result of ChIPseeker/

5. 2. 5. [Sample]_heatmap.png

Heatmap of ChIP peaks binding to TSS regions. The result of ChIPseeker.

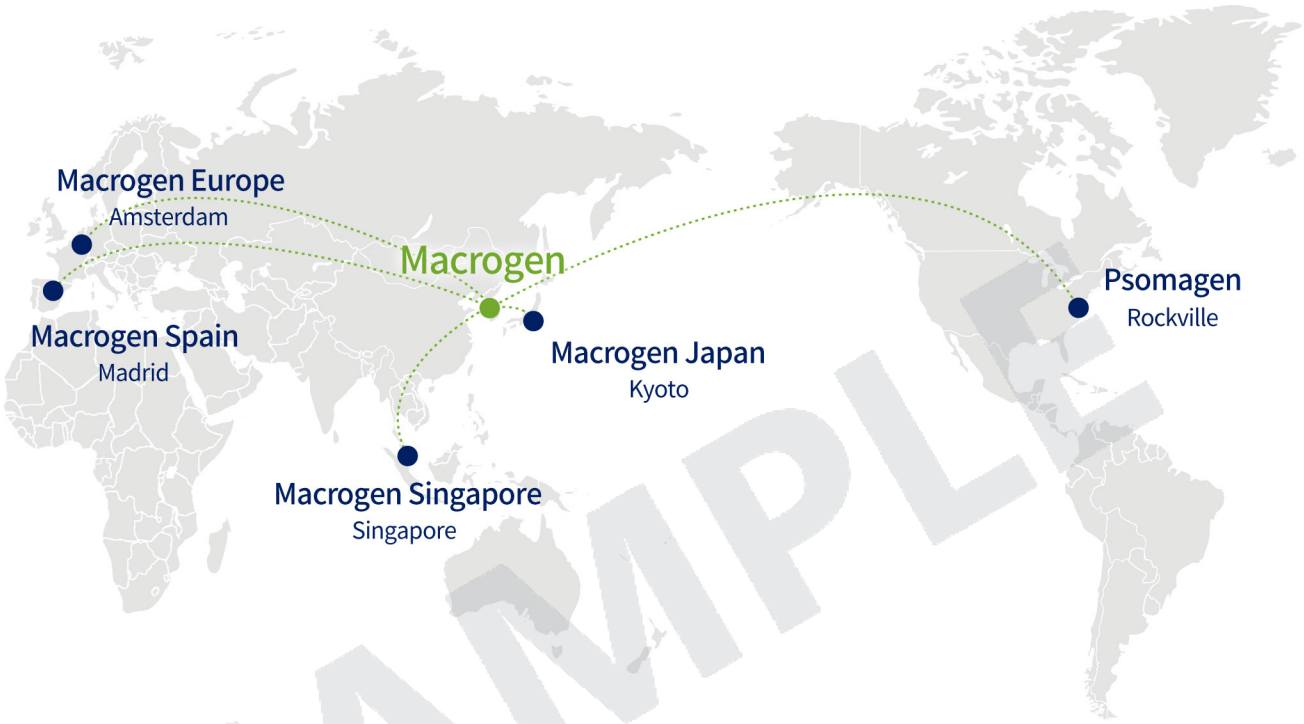
5. 2. 6. [Sample]_heatmap.png

Upset is an effective way to visualize sets and intersections. The result of ChIPseeker.

5. 3. References

1. BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, btu170.
2. LANGMEAD, Ben, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10.3: R25.
3. LI, Heng, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16: 2078-2079.
4. ZHANG, Yong, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology, 2008, 9.9: R137.
5. YU, Guangchuang; WANG, Li-Gen; HE, Qing-Yu. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 2015, 31.14: 2382-2383.

Research use only



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

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