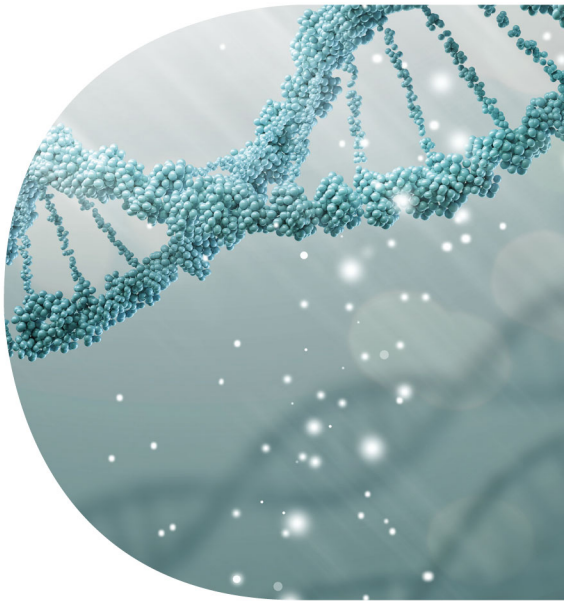
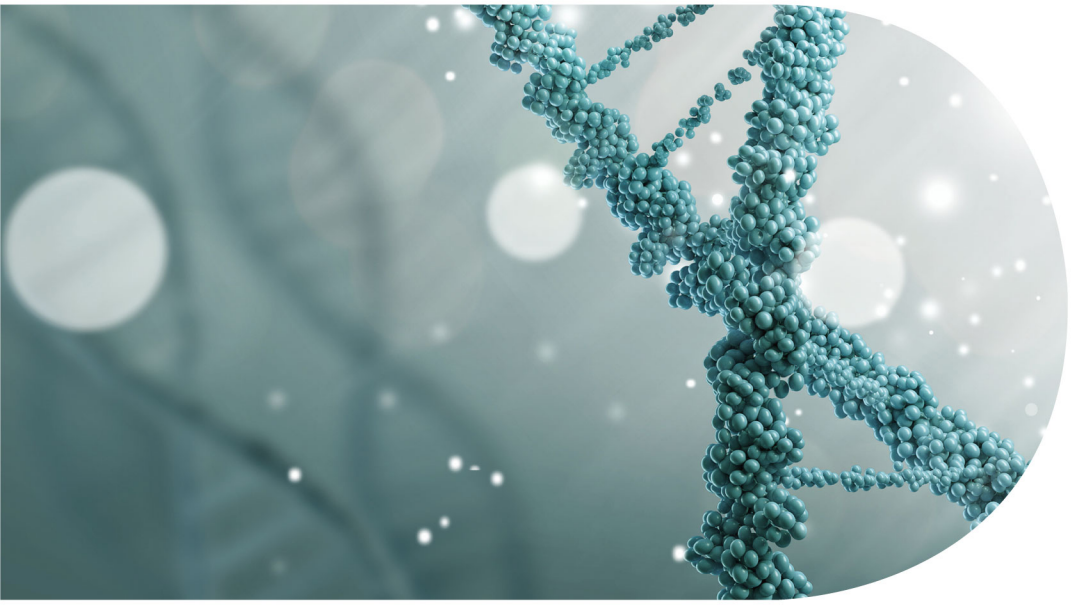
Research use only

Whole Genome De novo Sequencing

Report

September 2020





Humanizing Genomics
macrogen

Project Information

Client Name	Carlos Lopez-Otin
Company/Institute	Universidad de Oviedo ESQ3318001I
Order Number	HN00132087, HN00132604
Sample	HMforPacbio
Type of Analysis	De novo assembly, Error Correction, Annotation
Type of Sequencer	Sequel II System, Illumina platform

Table of Contents

1. Data Download

Download link	File size	md5sum
PacBio raw data (1)	189G	4e1d6d815ce2dce3451f6a3a18c2660b
Illumina raw data (1)	4.9G	74013c2e70eb8a0fdcd002770c4dd14f
Illumina raw data (2)	5.3G	cc7ca784f4534a9fcbe5cb1f1db8d3c2
Filtered Illumina raw data (1)	2.5G	bea941ec81aada32345bc57f40681730
Filtered Illumina raw data (2)	2.7G	d9deab73835803b2079605c4e2ac8990
Assembly data	175M	434bde89b23ce6b3eb131a527d149f66
Annotation Data	915M	9b562a8c89b206673ff98896325cff2d

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. Sequencing and Analysis Workflow

Figure 1. Workflow overview

2. 1. 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.

In the case of Illumina data, the sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. This library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. 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. Then, sequencing data is converted into raw data for the analysis.

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 reads were corrected and filtered. 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.

2. 2. 2. Error Correction

In the next step after de novo assembly step, Illumina reads are applied for sequence compensation to construct contigs more accurately. By mapping the Illumina reads to first assembled genome sequence, we can see the mapping result that shows a slight difference from the assembly result. We use this information to correct the consensus sequence. Also, we can get a consensus sequence with higher quality through the self-mapping step.

2. 2. 3. Annotation

After whole genome or draft genome is assembled, the location of protein-coding sequences, tRNA genes, and rRNA genes are identified. Then their functions are annotated.

3. Summary of Data Production

3. 1. Subreads Filtering

Table 1. Stats of filtered subreads

Mean subread length	9,223	**N50**	12,978
Total subread bases	88,646,436,290	**Total subreads**	9,610,778

- 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

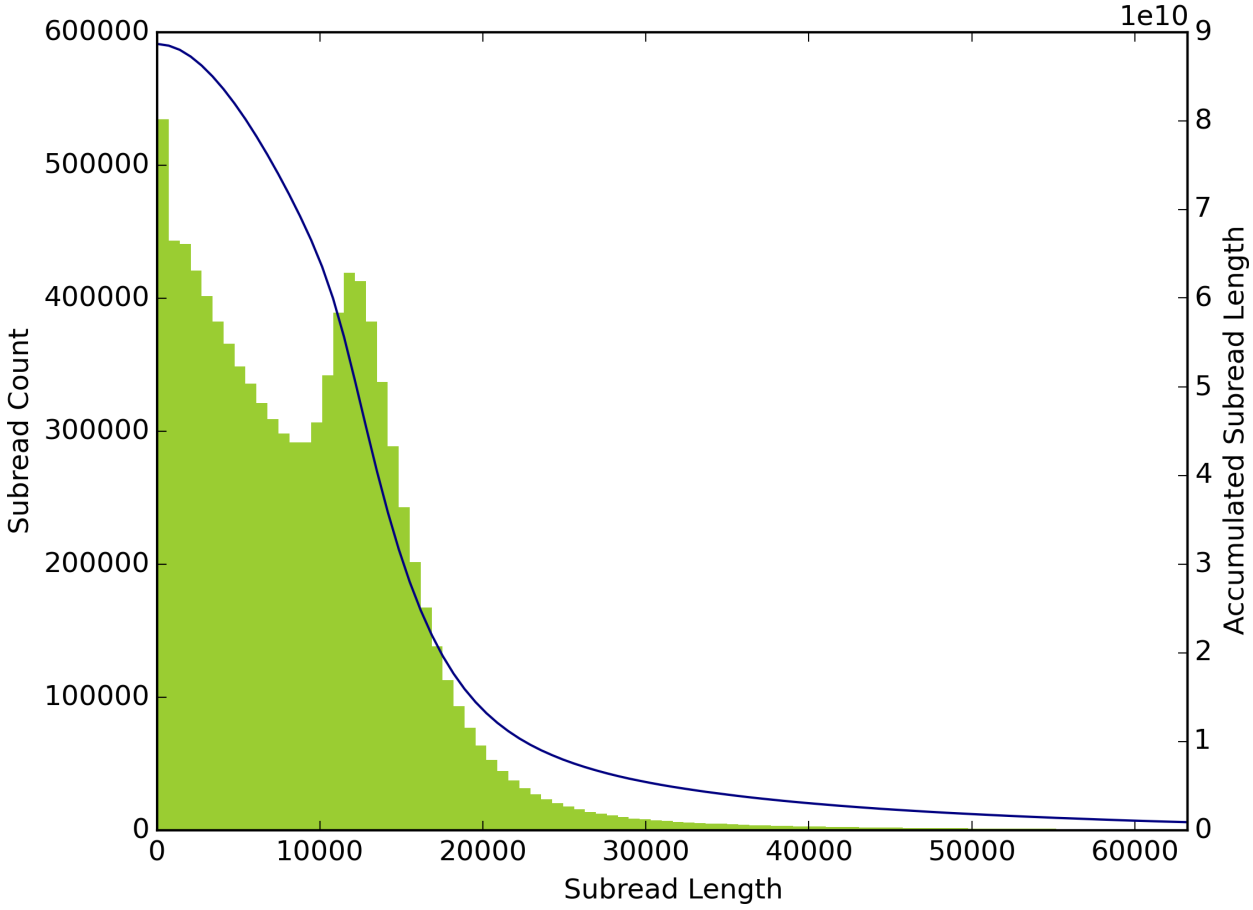


Figure 2. Filtered subread length distribution

3. 2. Illumina Raw Data Filtering

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) were calculated. Assembled contigs were corrected using this Illumina data. By revising contigs, more accurate nucleotide genomic sequences could be obtained and applied to other analysis protocols.

Table 2. Stats of Illumina raw data

	Total read bases	Total reads	GC (%)	Q20 (%)	Q30 (%)
Raw dataset	20,506,036,466	135,801,566	44.11	95.94	90.48
Filtered dataset	12,301,853,730	81,469,230	43.29	99.17	96.1

- Total read bases : The total number of bases sequenced
- Total reads : The total number of reads. For Illumina paired-end sequencing, this value refers to the sum of read1 and read2
- GC (%) : GC content
- Q20 (%) : Ratio of bases that have phred quality score over 20
- Q30 (%) : Ratio of bases that have phred quality score over 30

Figure 3. Base quality of filtered dataset read1 (up) and read2 (down)

4. Analysis Results

4. 1. De novo Assembly

Bioinformatics software such as HGAP, FALCON and CANU can assemble the PacBio long-reads. In this analysis, wtdbg2 was used and the detailed information is attached in the appendix. As the result of mapping reads against assembled contigs and error correction using Arrow, the consensus sequence with higher quality is generated. The assembly results are summarized in the table below.

Table 3. Stats before assembly correction

Contigs	Total contig bases	N50	Max length	Min length	Mean length
776	603,162,511	3,284,247	23,002,843	1,133	777,271

After assembly, Illumina reads were applied for accurate genome sequence using Pilon. And then, by mapping the subreads against assembled contigs, the consensus sequence with depth of coverage data was generated.

Table 4. Stats after assembly correction

Contigs	Total contig bases	N50	Max length	Min length	Mean length
657	602,437,578	3,289,651	23,036,873	1,091	916,952

- 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 5. Result of assembly: 657 contigs were formed

Contig name	Length	GC %	Depth
contig1	23,036,873	43.9	93
contig2	19,131,563	43.3	95
contig3	18,214,800	43.6	97
contig4	17,938,902	42.5	97
contig5	17,300,051	42.4	100
contig6	16,493,625	43.3	94
contig7	14,369,762	44.7	93
contig8	13,111,320	43.5	95
contig9	12,376,668	44.6	88
contig10	9,969,525	43.7	91
contig11	8,532,378	43.6	94
contig12	7,707,451	43.8	99
contig13	7,578,755	43.1	93
contig14	7,480,624	43.9	101
contig15	6,848,578	44.4	93
contig16	6,011,400	43.8	89
contig17	5,601,369	43.1	93
contig18	5,434,093	42.2	101
contig19	5,301,733	44.2	94
contig20	5,235,463	42.6	99
contig21	5,054,663	43.5	91
contig22	5,020,441	43.1	93
contig23	4,956,152	42.4	102
contig24	4,849,865	44.3	91
contig25	4,824,257	43.6	97
...	...	...	...
Total	602,437,578	43.65	95

If more than 25 contigs are created, the omitted contig information is provided as an xlsx file.

Please refer to the attached excel file (HN00132604_Pac_Denovo_len_gc.xlsx HN00132087, HN00132604_Pac_Denovo_len_gc.xlsx)

- Length : The number of bases in each contig
- GC % : GC content
- Depth : The average number of reads aligned to each bases on contig

4. 2. Assembly Validation

4. 2. 1. K-mer Analysis

K-mer analysis was performed to estimate the genome size of sample. The graph was plotted with the coverage and frequency of k-mers. The sharp left-side peak represents random sequencing error while the right represents appropriate data. The genome size can be estimated using total k-mer number and volume peak.

For the accurate analysis, Illumina sequencing data were randomly sampled into 40-fold of total contig bases. K-mer Analysis results can estimate genome size, not perfectly identify that. It means that estimated genome size can be different with total contig bases as well as real genome size.

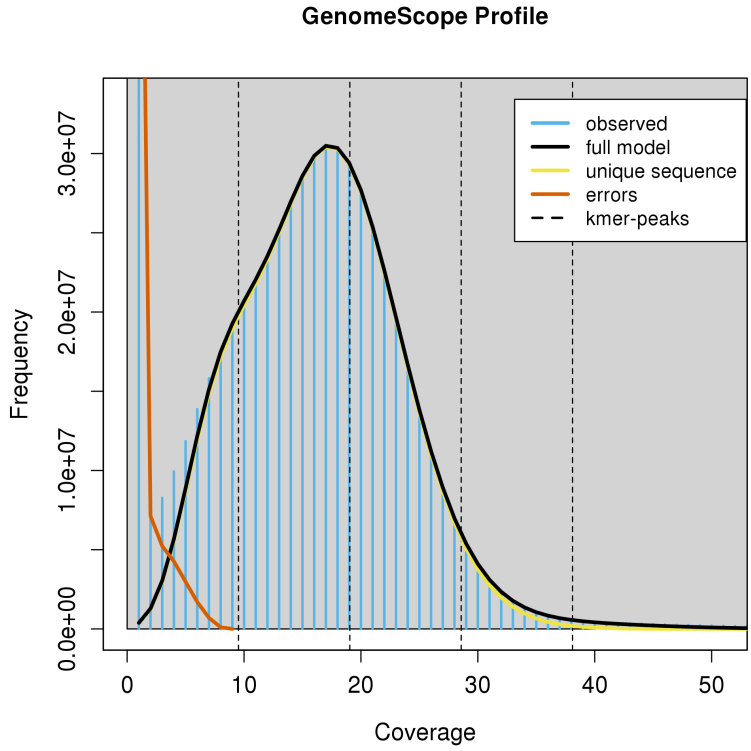


Figure 4. K-mer graph

Table 6. K-mer analysis result

	K-mer coverage	Heterozygosity	Genome length	Genome repeat length
21mer	19.05	0.789	552,071,163	103,672,848

- K-mer coverage : The mean k-mer coverage for heterozygous bases
- Heterozygosity : The overall rate of heterozygosity
- Genome length : The inferred genome length
- Genome repeat length : The length of the genome that is repetitive

4. 2. 2. Mapping Results

In order to validate accuracy of the assembly, Illumina reads were mapped to the assembly result. After mapping, the necessary stats were calculated.

Table 7. Overall mapping stats

Library name	Total reads	Mapped reads	Coverage (%)	Depth	Ins.size (Std.)
HMforPacbio	81,469,230	81,030,860 (99.46%)	98.63	18.63	424.31 (91.36)

- Library name : Sample's library name
- Total reads : Total number of reads
- Mapped reads : Total number of mapped reads
- Coverage (%) : The percentage of mapped sited (>= 1x)
- Depth : Average mapping depth
- Ins.size (Std.) : The length between adapters and standard deviation of predicted length

This is insert size plot based on mapping status of HMforPacbio. Please refer to the insert_size_plot file in Analysis Result if the sample has 2 or more libraries.

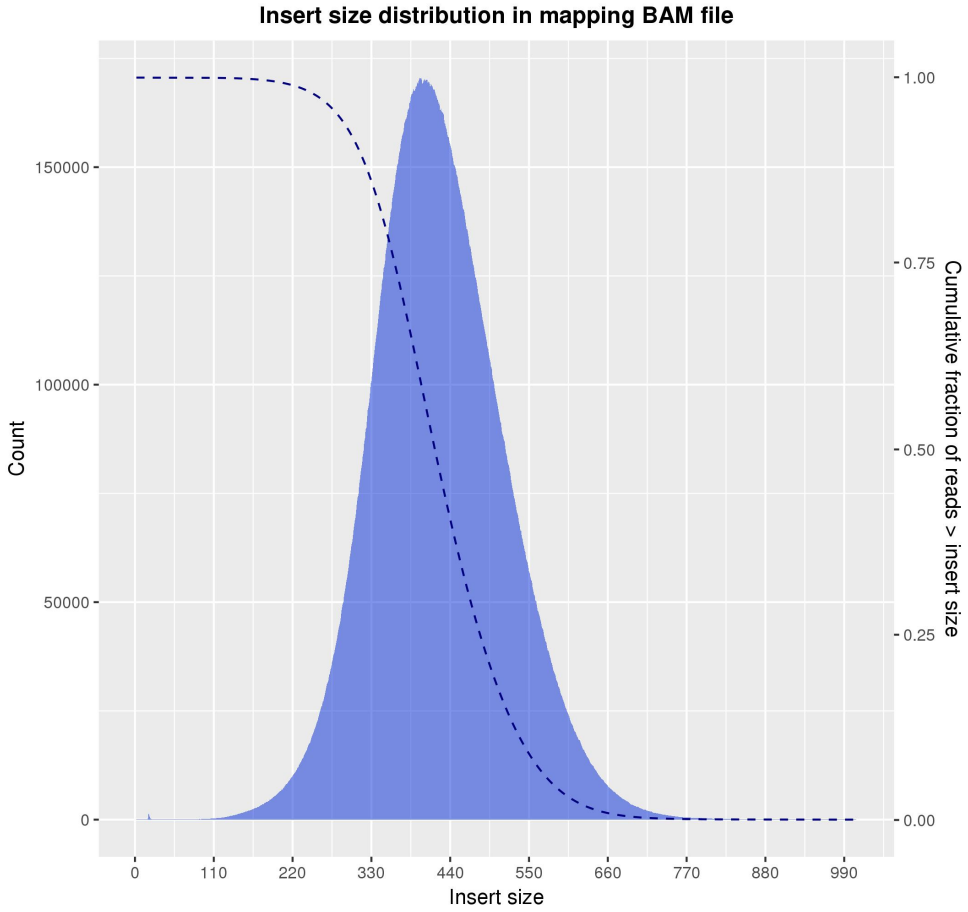


Figure 5. Insert size plot

4. 2. 3. 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 8. BUSCO analysis result

Used Lineage : eukaryota_odb9 (number of species: 100, number of BUSCOs: 303)

Status	# of BUSCOs	Percentage
Complete BUSCOs (C)		
Complete and single-copy BUSCOs (S)	280	92.41 %
Complete and duplicated BUSCOs (D)	12	3.96 %
Fragmented BUSCOs (F)	1	0.33 %
Missing BUSCOs (M)	10	3.30 %
Total BUSCO groups searched	303	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

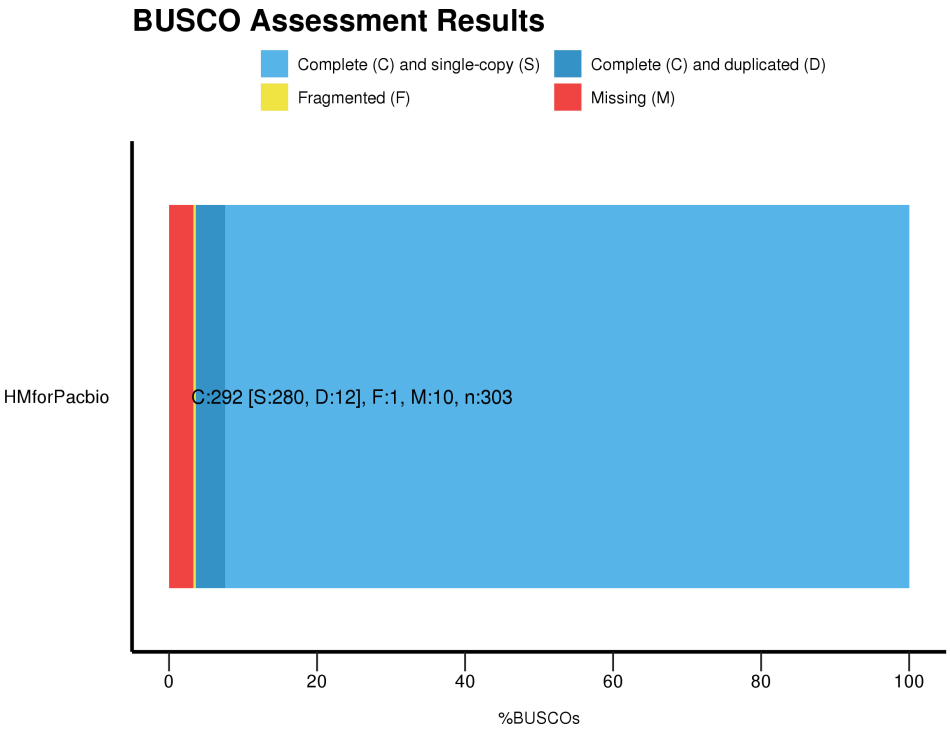


Figure 6. BUSCO result plot

4. 3. Genome Annotation

After complete genome or draft genome is analyzed, the locations of protein genes were predicted and their functions were annotated. **Maker (v2.31.8)** was used to predict the location while **Protein BLAST+ (v2.7.1+)** was performed with UniProt Swiss-Prot (201806). Three files (GFF, faa, ffn) are provided as a result of gene prediction.

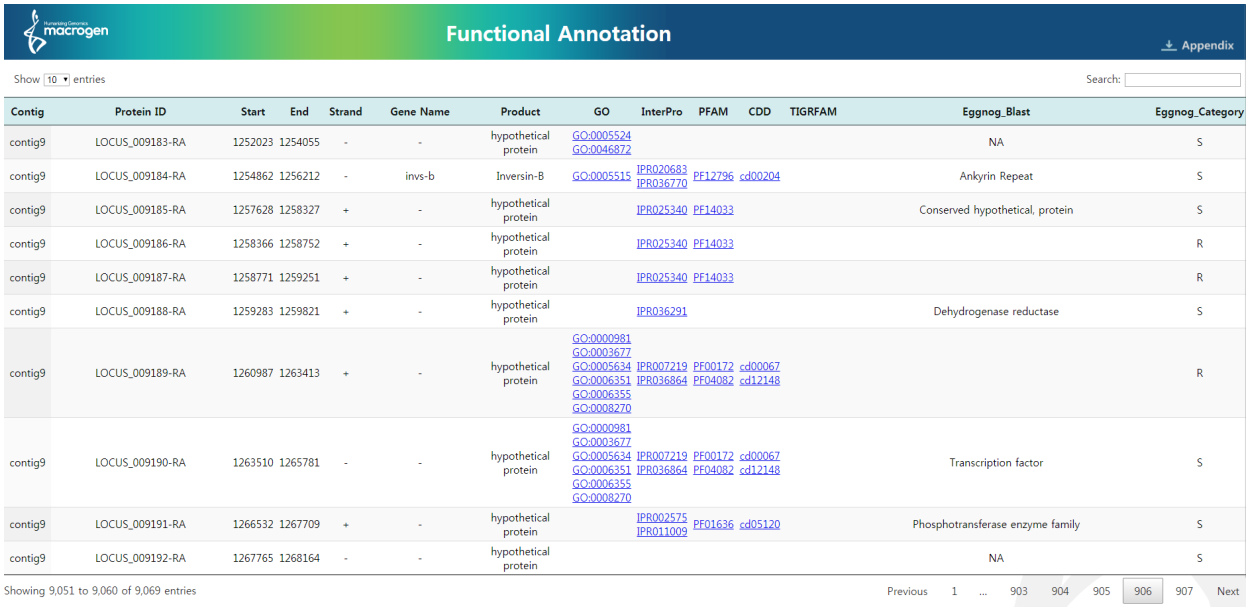
Table 9. Result of annotation

	Number of genes	Total length
Predicted transcripts	74,284	61,130,637 bp
Predicted proteins	74,284	20,276,701 aa

- Number of genes : The number of genes for MAKER produced gene annotations
- Total length : The length of sequences for MAKER produced gene annotations

We provide BLAST results in Excel and HTML format for various databases.
[UniProt(v201806), InterPro(v69.0), Pfam(v31.0), CDD(v3.16), TIGRFAM(v15.0) and EggNOG(v4.5.1)]

Following is the HTML example.



macrogen — Functional Annotation — Appendix

Show 10 entries — Search:

Contig	Protein ID	Start	End	Strand	Gene Name	Product	GO	InterPro	PFAM	CDD	TIGRFAM	Eggnog_Blast	Eggnog_Category
contig9	LOCUS_009183-RA	1252023	1254055	-	-	hypothetical protein	GO:0005524 GO:0046872					NA	S
contig9	LOCUS_009184-RA	1254862	1256212	-	invs-b	Inversin-B	GO:0005515	IPR020683 IPR036770	PF12796	cd00204		Ankyrin Repeat	S
contig9	LOCUS_009185-RA	1257628	1258327	+	-	hypothetical protein		IPR025340	PF14033			Conserved hypothetical, protein	S
contig9	LOCUS_009186-RA	1258366	1258752	+	-	hypothetical protein		IPR025340	PF14033				R
contig9	LOCUS_009187-RA	1258771	1259251	+	-	hypothetical protein		IPR025340	PF14033				R
contig9	LOCUS_009188-RA	1259283	1259821	+	-	hypothetical protein		IPR036291				Dehydrogenase reductase	S
contig9	LOCUS_009189-RA	1260987	1263413	+	-	hypothetical protein	GO:0000981 GO:0003677 GO:0005634 GO:0006351 GO:0006355 GO:0008270	IPR007219 IPR036864	PF00172 PF04082	cd00067 cd12148			R
contig9	LOCUS_009190-RA	1263510	1265781	-	-	hypothetical protein	GO:0000981 GO:0003677 GO:0005634 GO:0006351 GO:0006355 GO:0008270	IPR007219 IPR036864	PF00172 PF04082	cd00067 cd12148		Transcription factor	S
contig9	LOCUS_009191-RA	1266532	1267709	+	-	hypothetical protein		IPR002575 IPR011009	PF01636	cd05120		Phosphotransferase enzyme family	S
contig9	LOCUS_009192-RA	1267765	1268164	-	-	hypothetical protein						NA	S

Showing 9,051 to 9,060 of 9,069 entries — Previous 1 ... 903 904 905 906 907 Next

5. Details of File Extensions

Raw Data

File Extensions	Description
*.subreads.bam/pbi	These files contain Analysis-ready subreads.
*.scraps.bam/pbi	These files contain Excised adapters, barcodes, and rejected subreads.
*.xml	The *.xml contains top level information about the data, including what sequencing enzyme and chemistry were used, sample name and other metadata.
*.fasta	The files contain subreads sequence in FASTA format.

Assembly Data

File Extensions	Details
consensus.fasta	Whole nucleotide sequence.

Annotation Data

File Extensions	Details
annotation_viewer/*.html	Generate match results and sequence information for various types of Functional annotation DB in HTML format.
*_annotation.xlsx	Excel file that saves contents written in HTML.
*_EggNOG.xlsx	Excel file containing the results of the analysis of EggNOG database.
*.fsa	Whole nucleotide sequence for running tbl2asn.
*.statistics	Summary of Genome Annotation.
*.tbl	Tab Separation Formality of NCBI. (http://www.ncbi.nlm.nih.gov/projects/Sequin/table.html)
*.gff	GFF3 Format. (http://www.sequenceontology.org/gff3.shtml)
*.ffn	Fasta format of CDS nucleotide.
*.faa	Fasta format of Amino Acid.
*.sqn	NCBI's Sequin Format (Edited with Sequin). (http://www.ncbi.nlm.nih.gov/Sequin/)
*.gbf	GenBank Format. (http://www.ncbi.nlm.nih.gov/genbank/)

[ver.2020-01]

6. Appendix

6. 1. FAQ

Q: I would like to see the result. How can I open the 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 .gbf 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.

6. 2. Programs used in Analysis

6. 2. 1. Denovo Assembly

wtdbg2 (v2.3)

LINK (https://github.com/ruanjue/wtdbg2)

Wtdbg2 is an assembler which performs assembly without error corrections. Long noisy reads such as PacBio and Oxford Nanopore Technologies (ONT) can be assembled using wtdbg2. Wtdbg2 broadly follows the overlap-layout-consensus algorithm (OLC) for assembly, however, it also used a distinguishable method, fuzzy Bruijn graph (FBG). Wtdbg2 produced 1024 bp segments and similar segments were merged into a vertex. The vertex were connected each other based on the segment adjacency on reads. After complete of FBG step, wtdbg2 can construct final consensus from FBG.

- preset : SQ (Sequel)
- Genome size : expected genome size

6. 2. 2. Error Correction

Pilon (v1.21)

LINK (https://github.com/broadinstitute/pilon/wiki)

Pilon is a software tool which can be used to automatically improve draft assemblies. It significantly improves draft genome assemblies by correcting bases, fixing mis-assemblies and filling gaps.

6. 2. 3. Gene Prediction

MAKER (v2.31.8)

LINK (http://www.yandell-lab.org/software/maker.html)

MAKER is a portable and easy to configure genome annotation pipeline. MAKER allows smaller eukaryotic genome projects and prokaryotic genome projects to annotate their genomes and to create genome databases. MAKER identifies repeats, aligns ESTs and proteins to a genome, produces ab initio gene predictions and automatically synthesizes these data into gene annotations with evidence-based quality values.

6. 2. 4. Annotation

BLAST+ (v2.7.1+)

LINK (http://blast.ncbi.nlm.nih.gov/))

The Basic Local Alignment Search Tool (BLAST) 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.

- E-vale : 1e-3
- Databases : NT, NR

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

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

Humanizing Genomics
macrogen