Genomics NGS Service

Bioinformatics Analysis of Bacteria de-novo genome assembly by Pacbio HGAP

Help manual

2017

Genomics NGS Analysis Team

Table of Contents

Pacbio Report Folder Instruction

|--- <PB_ID>_<Sample_name>_report
| |--- **1_Assembly**
| | |--- **AssemblyStats** (general assembly report folder)
…
| | |--- **xxx_polished_assembly.fasta** (raw assembly draft by HGAP)
| | |--- **xxx_polished_fixed.fasta** (fixed assembly draft by Circlator)
| | |--- **xxx_contig_summary.xlsx** (fixed assembled contig summary)
| |--- **2_Subreads**
| | |--- **xxx_filtered_subreads.fasta** (not provide in report)
| | |--- **xxx_filtered_subreads.fastq** (not provide in report)
| | |--- **filtered_subreads_dist.png** (Subread length distribution)
| |--- **3_Prokka (gene prediction result)**
| | |--- **xxx.faa/fna/fsa/gbk/gff** (gene prediction/annotation raw output)
…
| | |--- **README.txt** (prokka output manual)
| |--- **4_Annotation**
| | |--- **COG** (protein group function annotation)
| | |--- **GO** (gene function ontology)
| | |--- **KEGG** (functional pathway annotation)
|--- **Pacific-Biosciences-Glossary-of-Terms.pdf** (Pacbio terms explanation)
|--- **xxx_genomeMap.png** (circular layout for the longest fixed assembled contig)
|--- **Report.pdf** (general sequencing, assembly & annotation result)
|--- **xxx_anno_report.xls** (gene prediction/annotation final report from prokka output)
|--- **Help.pdf** (general data explanation & how to use)

Experiment Process

- Library Preparation Part

● Pipeline of Experiment

Genomic DNA is extracted and fragmented randomly and then required length DNA fragments are retained by electrophoresis. And after this, we ligate adapters to DNA fragments then conduct cluster preparation, sequencing finally. The library preparation method and sequencing pipeline is shown below.

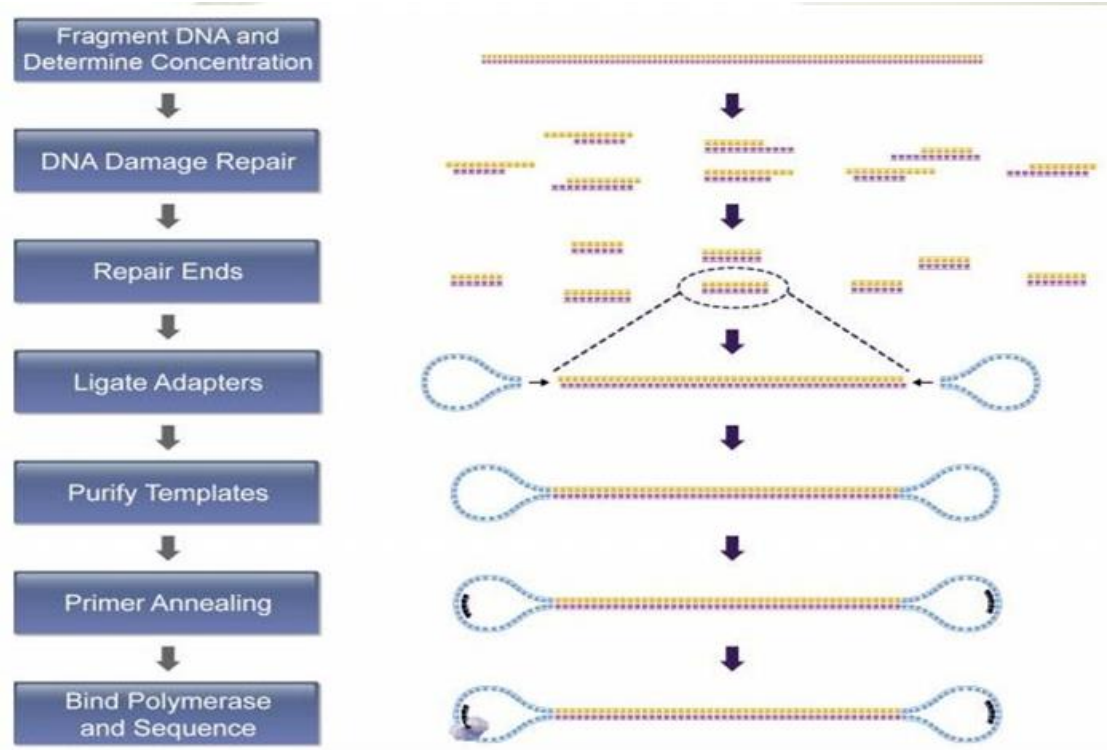
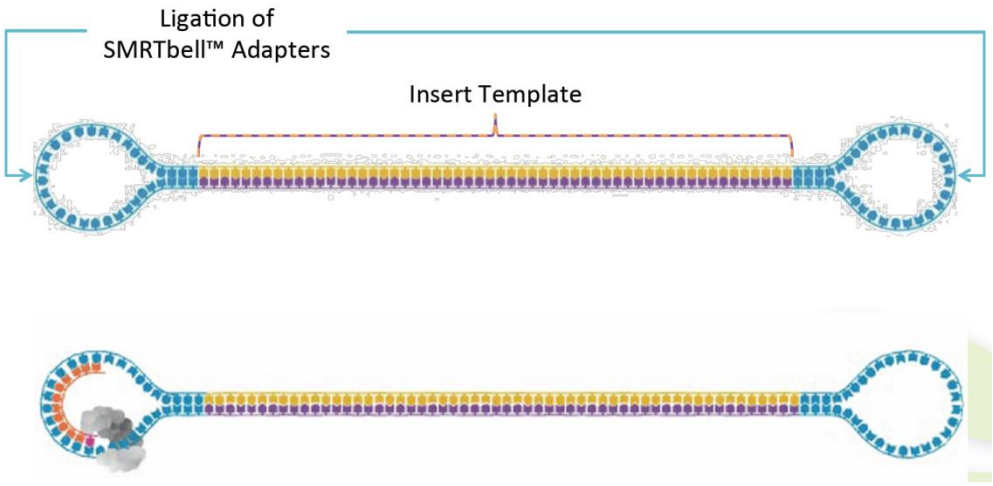


Figure 1 Pipeline of experiment. For the PacBio library construction and sequencing, genomic DNA was sheared using a Covaris g-TUBE followed by purification via binding to pre-washed AMPure PB beads (Part Number: PB100-265-900).After end-repair, the blunt adapters were ligated, followed by exonuclease incubation to remove all un-ligated adapters and DNA. The final “SMRT bells” were annealed with primers and bound to the proprietary polymerase using the PacBio DNA/Polymerase Binding Kit P6 v2 (Part Number PB100-372-700) to form the “Binding Complex”. After dilution, the library was loaded onto the instrument with DNA Sequencing Kit 4.0 v2 (Part Number PB100-612-400) and a SMRT Cell 8Pac for sequencing. A primary filtering analysis was performed with the RS instrument, and the secondary analysis was performed using the SMRT analysis pipeline version 2.3.0.

- **SMRTbell™ template:**
  A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) at both ends.

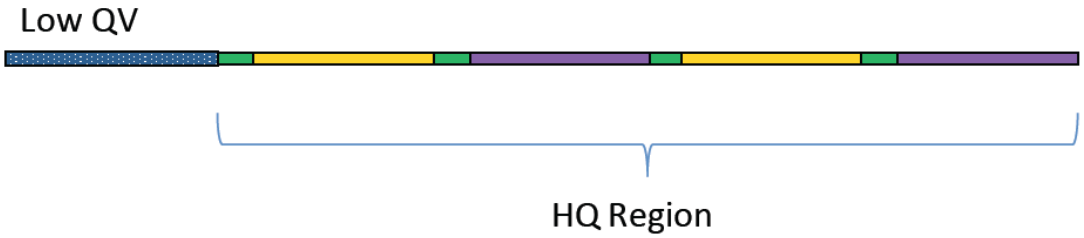


- Sequencing Part

- **Primary analysis**:

(Polymerase read -> Pre-filtered subreads):

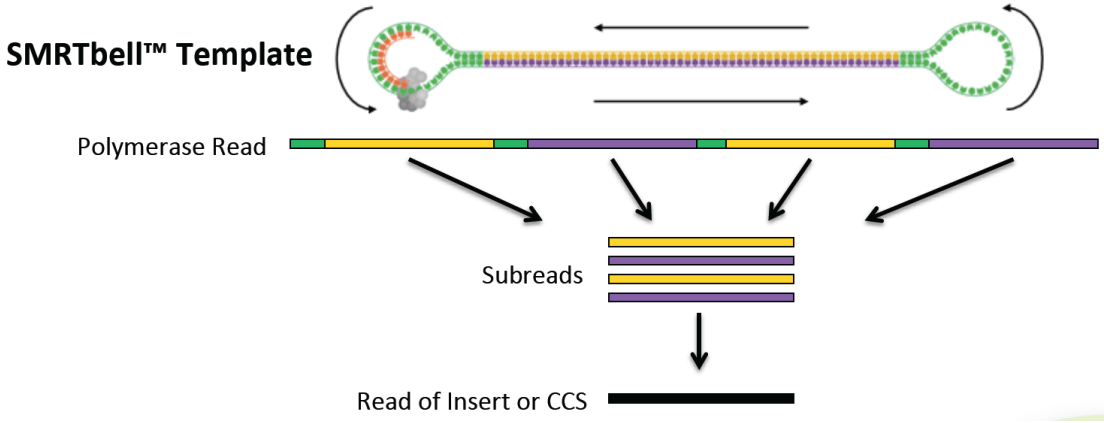
On-instrument analysis which includes signal processing of the movie, base calling of the traces and pulses, and quality assessment of the base calls. Subsequently, it trims the sequences to the high-quality (HQ) regions, identifies adapter, barcode (optional), and control sequence reads, assigns read scores, and **outputs the subread data in a BAM file**.



- **Seconary analysis:**

Follows primary analysis and uses basecalled data. It is application-specific, and may include:

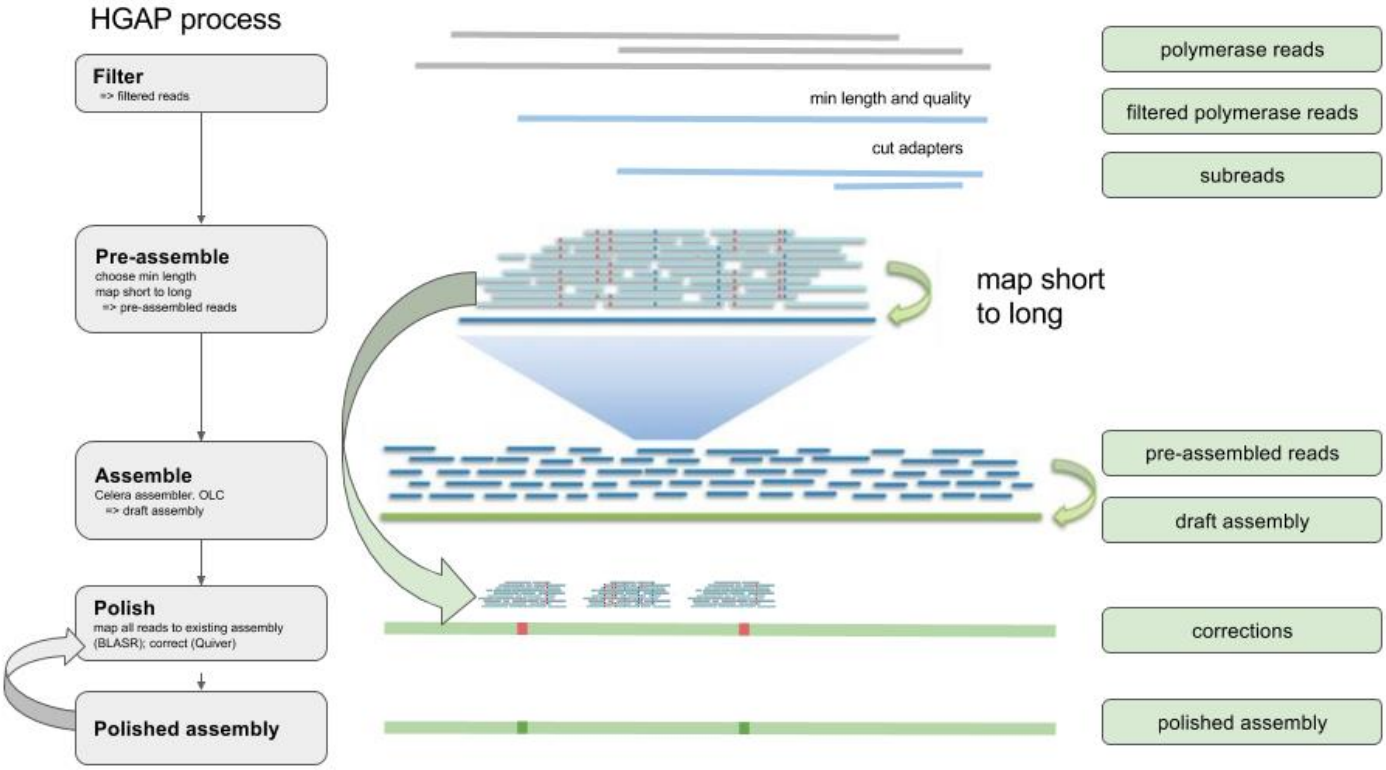
– Filtering/selection of data that meets a desired criteria (such as quality, read length, etc.).
– Comparison of reads to a reference or between each other for mapping and variant calling, consensus sequence determination, alignment and assembly (de novo or reference-based), variant identification, etc.
– Quality evaluations for a sequencing run, consensus sequence, assembly, etc.
– PacBio's SMRT Analysis contains a variety of secondary analysis applications including RNA and Epigenomics analysis tools.



- **How HGAP assembly work :**

"HGAP", The **Hierarchical Genome Assembly Process** (HGAP) for long single pass reads generated by the PacBio® Single Molecule Real Time (SMRT) sequencer was developed to allow the complete and accurate shotgun assembly of bacterial sized genomes. [1]. The three main steps involved in HGAP are **preassembly**, **assembly**, and **consensus polishing.**

If you need more detail information, please visit Pacbio official github website: https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP



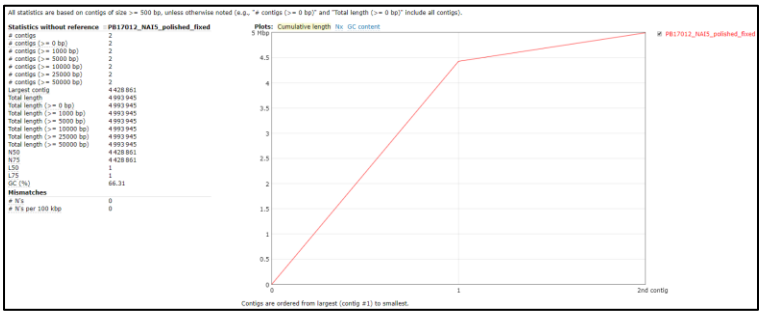
Bioinformatics analysis

1. Assembly Stats (1_Assembly)

Pacbio sequencing with official assembly tool – "**HGAP v3**" is powerful for bacterial genome assembly. [1] However, current long-read assembly software still typically assumes that the contigs they produce are linear. In contrast, the genome of almost every species contains at least one circular DNA structure, such as bacterial chromosomes and plasmids. Thus, a useful circularization of genome assembly tool – "**Circlator**" is our default pipeline to correct and linearize the genome. [2] Finally, we are using "**QUAST v4.5**" for evaluating the assembled genome quality. [3]

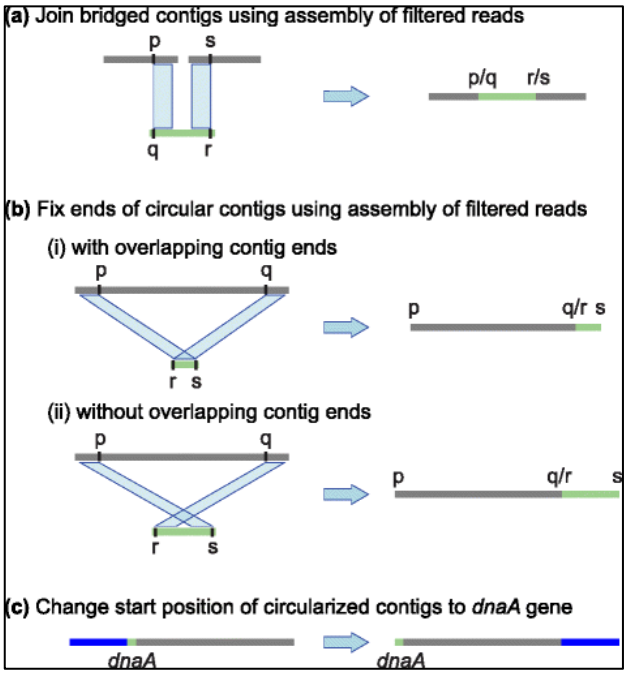
Generally, all of the following pipelines will be using Circlator output – "**xxx.polished.fixed.fasta**"

Report.html (Assembly Summary)



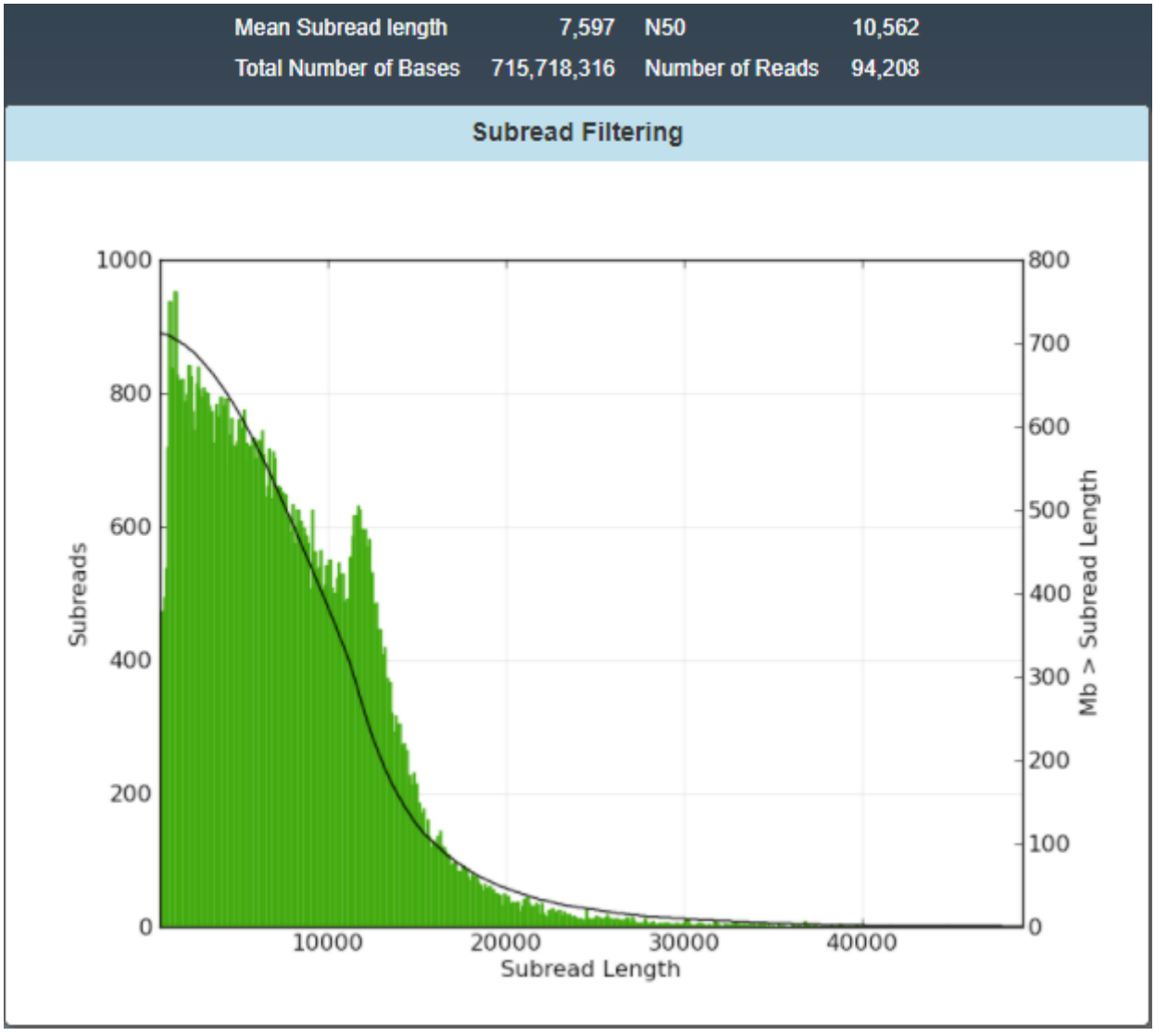
Note:

- **N50: the shortest sequence length at 50% of the genome.**
  - ✓ Commonly we said that if N50 is much larger, the assembly result is much better.
- **L50: the smallest number of contigs whose length sum produces N50**
  - ✓ Commonly we said that if L50 is much lower, the assembly result is much better.



2. Subreads (2_Subreads)

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™ template and **no adapter** sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for applications like de novo assembly, resequencing, base modification analysis, and so on.



*** Due to fasta/fastq are too large, we're not provided in this folder. If user required, we suggested these files could be using hard-copy and transferred by our salesperson to you.**

3. Gene prediction / annotation (3_Prokka)

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. "**Prokka v1.12**" is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files. [4]

As de-novo assembled genome finished, we would like to know what kind of and how much of genes, RNAs, and other elements existed in genome. Prokka is a powerful tool which could help for finding bacteria origin of replication (ori), following with downstream gene/CDS/RNA sequence prediction and using blast method against the uniprot/swissprot bacterial database.

[**files explanation**]:

gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV (genome browser).
gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
fna	Nucleotide FASTA file of the input contig sequences.
faa	**Protein FASTA file of the translated CDS sequences.**
ffn	**Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)**
fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
err	Unacceptable annotations - the NCBI discrepancy report.
txt	Statistics relating to the annotated features found.
tsv	Tab-separated file of all features: locus_tag, ftype, gene, EC_number, product
tbl	Feature Table file, used by "tbl2asn" to create the .sqn file

*** We suggest that user could view the "**xxx_anno_report.xls**" in the root path directly. ***

4. Annotation (4_Annotation)

After gene prediction work complete. Our functional annotation works including:

- COG (protein group ortholog)
- GO (gene functional ontology)
- KEGG pathway (gene functional pathway)

[Protein group function annotation by COG]

In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG.

cog_barchart.png (COG enrichment plot)

rps_blast_cog.txt (informative COG annotation)

query id	subject id	% identity	alignment l	mismatche	gap opens	q. start	q. end	s. start	s. end	evalue	bit score	COG#	functional categories				COG protein description
HMMICGI	CDD:2236	52.012	323	142	7	1	321	97	408	4.39E-133	382	COG0593	L				ATPase involved in DNA replication initiation
HMMICGI	CDD:2244	23.894	113	77	4	2	113	91	195	2.73E-04	38.9	COG1484	L				DNA replication protein
HMMICGI	CDD:2241	27.737	137	67	6	111	243	246	354	4.39E-04	38.6	COG1223	R				Predicted ATPase (AAA+ superfamily)
HMMICGI	CDD:2243	21.25	160	110	8	85	238	126	275	6.76E-04	38.1	COG1474	L	O			Cdc6-related protein, AAA superfamily ATPase
HMMICGI	CDD:2237	39.604	404	228	8	5	398	14	411	5.45E-156	443	COG0635	H				Coproporphyrinogen III oxidase and related Fe-S oxidoreductases
HMMICGI	CDD:2239	18.145	248	176	7	31	261	208	445	7.18E-09	54.2	COG1032	C				Fe-S oxidoreductase
HMMICGI	CDD:2241	21.491	228	141	8	34	226	81	305	8.71E-09	53.9	COG1243	K	B			Histone acetyltransferase
HMMICGI	CDD:2241	28.283	99	71	0	136	234	146	244	1.51E-07	49.6	COG1242	R				Predicted Fe-S oxidoreductase
HMMICGI	CDD:2236	26.957	115	75	3	140	249	262	372	4.06E-05	42.6	COG0621	J				2-methylthioadenine synthetase
HMMICGI	CDD:2241	19.728	147	106	7	30	169	60	201	0.001	37.8	COG1244	R				Predicted Fe-S oxidoreductase

[GO annotation of predicted genes]:

Gene ontology concern with annotation of genes and gene products and to provide centralized access to resources and tools. both GO and COG provide specific information about gene or gene products.

There are three main classes in GO database:

1. **Cellular Component:** These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
2. **Biological Process:** A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions.
3. **Molecular Function:** Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity".

"**InterProscan v5**" is a database which integrates together predictive information about proteins' function from a number of partner resources, giving an overview of the families that a protein belongs to and the domains and sites it contains. We input all of the predicted protein sequences to the database and try to parse their GO terms. [5]

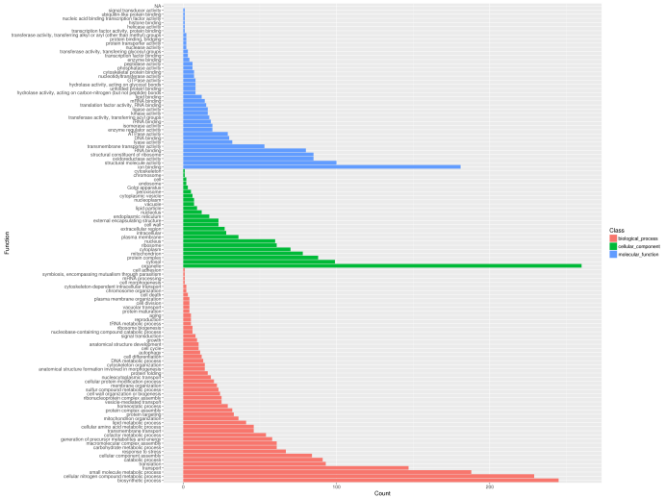
GeneID_GO.txt (GO terms extraction)

Prokka_ID	GOs
NMAHGBEO_00001	GO:0016491,GO:0055114
NMAHGBEO_00002	GO:0016491,GO:0055114
NMAHGBEO_00003	GO:0016491,GO:0055114
NMAHGBEO_00007	GO:0003677,GO:0003700,GO:0006352,GO:0006355,GO:0016987,GO:0003677,GO:0003700,GO:0006352,GO:0006355,GO:0016987
NMAHGBEO_00008	GO:0004252,GO:0006508
NMAHGBEO_00009	GO:0005215,GO:0006810,GO:0016020
NMAHGBEO_00012	GO:0016491,GO:0055114

GO_mapping.txt (informative GO annotation)

biological_process	GO:0009058	biosynthetic process	245	The chemical reactions and pathways resulting in the formation of substances; typically the energy-requiring part of metabolism in which simpler substances are transformed into more complex ones. [GOC:cu
biological_process	GO:0034641	cellular nitrogen compound metabolic p	229	The chemical reactions and pathways involving various organic and inorganic nitrogenous compounds, as carried out by individual cells. [GOC:mah]
biological_process	GO:0044281	small molecule metabolic process	188	The chemical reactions and pathways involving small molecules, any low molecular weight, monomeric, non-encoded molecule. [GOC:curators, GOC:pde, GOC:vw]
biological_process	GO:0006810	transport	147	The directed movement of substances (such as macromolecules, small molecules, ions) or cellular components (such as complexes and organelles) into, out of or within a cell, or between cells, or within a mul
biological_process	GO:0006412	translation	93	The cellular metabolic process in which a protein is formed, using the sequence of a mature mRNA molecule to specify the sequence of amino acids in a polypeptide chain. Translation is mediated by the ribos
biological_process	GO:0009056	catabolic process	91	The chemical reactions and pathways resulting in the breakdown of substances, including the breakdown of carbon compounds with the liberation of energy for use by the cell or organism. [ISBN:019854768
biological_process	GO:0022607	cellular component assembly	84	The aggregation, arrangement and bonding together of a cellular component. [GOC:isa_complete]
biological_process	GO:0006950	response to stress	67	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a disturbance in organismal or cellular
biological_process	GO:0005975	carbohydrate metabolic process	61	The chemical reactions and pathways involving carbohydrates, any of a group of organic compounds based of the general formula Cx(H2O)y. Includes the formation of carbohydrate derivatives by the additic
biological_process	GO:0065003	macromolecular complex assembly	61	The aggregation, arrangement and bonding together of a set of macromolecules to form a complex. [GOC:jl]
biological_process	GO:0006091	generation of precursor metabolites and	58	The chemical reactions and pathways resulting in the formation of precursor metabolites, substances from which energy is derived, and any process involved in the liberation of energy from these substances.
biological_process	GO:0051186	cofactor metabolic process	54	The chemical reactions and pathways involving a cofactor, a substance that is required for the activity of an enzyme or other protein. Cofactors may be inorganic, such as the metal atoms zinc, iron, and coppe

GO_barchart.png (according to GO_mapping.txt)



[KEGG pathway annotation]:

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:

1. Metabolism

   Global/overview, Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Other amino, Glycan, Cofactor/vitamin, Terpenoid/PK, Other secondary metabolite, Xenobiotics, Chemical structure
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development

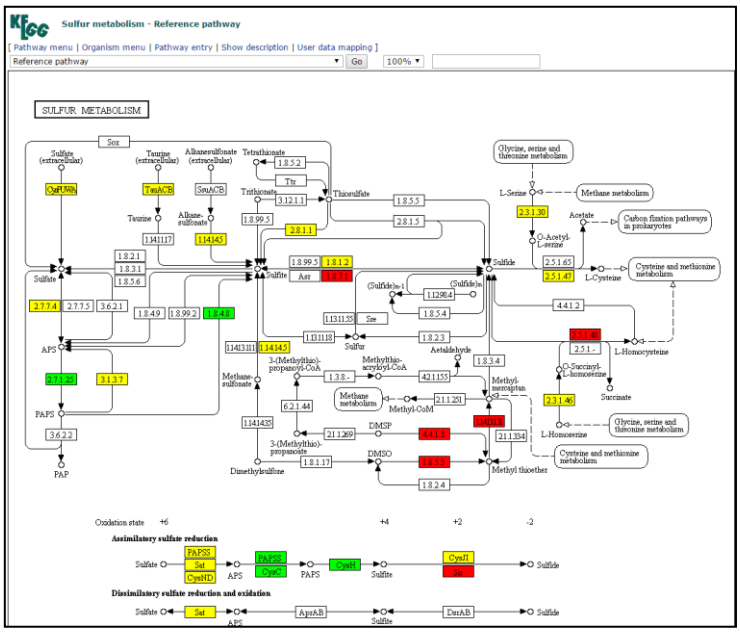
ec2kegg.xls

PathwayID	PathwayNa	Category	Total(EC_All)	Total(EC_Ref(sce))	Total(EC_Given)	Total(EC_Shared)	Total(EC_Unique_Ref)	Total(EC_Unique_Given)	EC_All	EC_Ref(sce)	EC_Given	EC_Shared	EC_Unique_Ref	EC_Unique_Given	P-value	FDR	URL
10	Glycolysis	Carbohydra	47	25	19	17	8	2	1.1.1.1,1.1	1.1.1.1,1.1.1	1.1.1.1,1.1	1.1.1.1,1.1.1	2.3.1.12,2.7.1.11	1.2.1.59,2.7.1.2	0	0	http://www
20	Citrate cyc	Carbohydra	25	16	8	8	8	0	1.1.1.286,	1.1.1.37,1.1.1	1.1.1.41,1.1	1.1.1.41,1.2.	1.1.1.37,1.1.1.42	2.3.1.12,2.3.1.61,4	0	0	http://www
30	Pentose pl	Carbohydra	55	17	6	6	11	0	1.1.1.215,	1.1.1.343,1.1.	1.1.1.44,2.2	1.1.1.44,2.2.	1.1.1.343,1.1.1.363	1.1.1.49,2.7.1.11	0.0001	0.00035	http://www
40	Pentose ai	Carbohydra	68	8	2	1	7	1	1.1.1.10,1.1.1	1.1.1.14,1.1.1	1.1.1.2,1.1.	1.1.1.2	1.1.1.14,1.1.1.30	1.1.1.21	0.04523	0.07563	http://www
51	Fructose a	Carbohydra	75	14	9	6	8	3	1.1.1.11,1.1.1	1.1.1.14,1.1.1	1.1.1.21,2.7.	2.7.1.1,2.7.7	1.1.1.14,1.1.1.67	1.1.1.21,2.7.1.4,2.	0	0	http://www
52	Galactose	Carbohydra	48	10	3	1	9	2	1.1.1.120,	2.7.1.1,2.7.1.	1.1.1.21,2.	2.7.1.1	2.7.1.11,2.7.1.6,2	1.1.1.21,2.7.1.2	0.00921	0.02088	http://www
53	Ascorbate	Carbohydra	46	0	1	0	0	1	1.1.1.122,	1.1.1.129,1.1.	1.1.1.2.1.3			1.2.1.3	0.03484	0.05923	http://www
61	Fatty acid	Lipid meta	17	6	1	1	5	0	1.1.1.100,	1.1.1.100,2.3.	6.2.1.3	6.2.1.3	1.1.1.100,2.3.1.179	2.3.1.39,2.3.1.86	0.21991	0.30727	http://www
62	Fatty acid	Lipid meta	13	7	0	0	7	0	1.1.1.211,	1.1.1.330,1.3.	1.3.1.38,1.3.	1.93,2.3.1.16,2.	1.1.1.330,1.3.1.38	1.3.1.93,2.3.1.16,	1	1	http://www
71	Fatty acid	Lipid meta	29	8	3	3	5	0	1.1.1.1,1.1	1.1.1.1.1.14	1.1.1.1,1.2	1.1.1.1,1.2.1	1.14.14.1,1.3.3.6	2.3.1.16,2.3.1.9,5.3	0.0056	0.01298	http://www
72	Synthesis	Lipid meta	6	2	1	1	1	0	1.1.1.30,2.	2.3.1.9,2.3.3.	2.3.3.10	2.3.3.10	2.3.1.9		0.10094	0.156	http://www
100	Steroid bio	Lipid meta	25	14	1	1	13	0	1.1.1.170,	1.1.1.170,1.1.	1.1.14.13.70	1.14.13.70	1.1.1.170,1.1.1.270	1.14.13.72,1.14.	0.41286	0.53306	http://www
130	Ubiquinone	Metabolism	40	5	2	1	4	1	1.1.1.237,	2.1.1.114,2.1.	1.6.5.2,2.6.	2.6.1.5	2.1.1.114,2.1.1.2	1.6.5.2	0.02261	0.04435	http://www
190	Oxidative p	Energy me	11	6	4	4	2	0	1.10.2.2,1.	1.10.2.2,1.3.5	1.10.2.2,1.3.	1.10.2.2,1.3.	3.6.1.1,3.6.3.14		0.00026	0.00086	http://www
220	Arginine bi	Amino acid	28	16	4	4	12	0	1.14.13.16	1.2.1.38,1.4.	1.1.4.1,2.1.4.	1.4.1.2,1.4.1	1.2.1.38,2.1.3.3	2.3.1.1,2.3.1.35,2.6.	0.00448	0.01088	http://www
230	Purine met	Nucleotide	109	42	13	11	31	2	1.1.1.154,	1.1.1.205,1.1.	1.1.1.205,1.	1.1.1.205,1.	2.1.2.2,2.4.2.1,2.	3.6.1.15,3.6.1.3	0	0	http://www
240	Pyrimidine	Nucleotide	64	23	4	4	19	0	1.1.98.6,1.	1.1.17.4.1,1.3.9	1.17.4.1,2.	1.17.4.1,2.7.	1.3.98.1,2.1.1.45	2.1.3.2,2.4.2.1,2.4.	0.01341	0.0285	http://www

[**Column definition**]

- Total(EC_All) = number of ECs associated with the KEGG pathway;
- Total(EC_Ref(ead)) = number of ECs in reference genome ead (*E. adhaerens OV14*) associated with the KEGG pathway;
- Total(EC_Given) = number of tested ECs found to be associated with the KEGG pathway;
- Total(EC_Shared) = number of tested ECs that are shared with reference genome;
- Total(EC_Unique_Ref) = number of ECs that are unique to the reference genome;
- Total(EC_Unique_Given) = number of ECs that are unique to the tested genome.

Click URL and get the pathway information

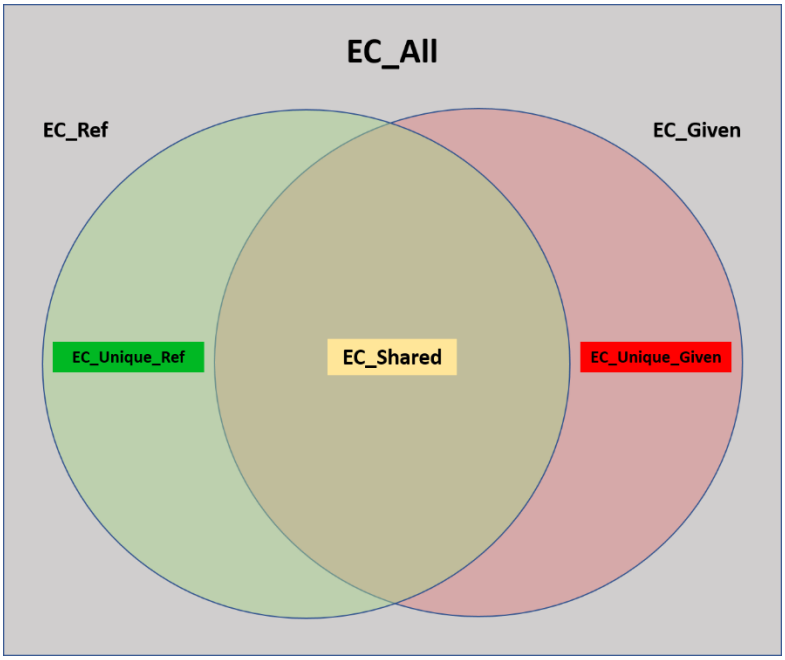


[**Pathway map color definition**]

green – an enzyme unique to a reference organism, (EC_Unique_Ref)

red – an enzyme unique to a given list, (EC_Unique_Given) ← **most important!**

yellow – a shared enzyme. (EC_Shared) ← **most important!**



***** All of the data including 'Prokka ID', 'Genome unitig', 'Region in genome' and functional annotation report is merged in "xxx_anno_report.xls" *****

5. Reference & Useful tools

1. Alexander, D.H., Chin, C., Clum, A., Copeland, A., Drake, J., Eichler, E.E., Heiner, C., Huddleston, J., Klammer, A.A., Korlach, J., Marks, P., & Turner, S.W. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods, 10 6*, 563-9.
2. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biology. 2015;16:294. doi:10.1186/s13059-015-0849-0.
3. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 2013.
4. Seemann T., Prokka: rapid prokaryotic genome annotation, Bioinformatics 2014 Jul 15;30(14):2068-9. PMID:24642063
5. Jones P, Binns D, Chang H-Y, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031.

- **Useful tools:**

- Comma separator: https://delim.co/ (分隔符號轉換行)
- Venny diagram: http://bioinfogp.cnb.csic.es/tools/venny/ (例如：若有多個 samples，可查看彼此之間交集的基因)
- REVIGO: http://revigo.irb.hr/ (視覺化 GO data)
- Uniprot database: http://www.uniprot.org/ (全球三大基因/蛋白質資料庫)
- Uniprot ID mapping: http://www.uniprot.org/mapping/ (Transform Uniprot gene ID to what you want)
- KEGG mapping: http://www.genome.jp/kegg/tool/map_pathway1.html (透過 KEGG 網站搜索 enzyme 或基因的代謝路徑)