Genomics NGS Service

Bioinformatics Analysis of de-novo genome assembly

(Illumina)

Help manual

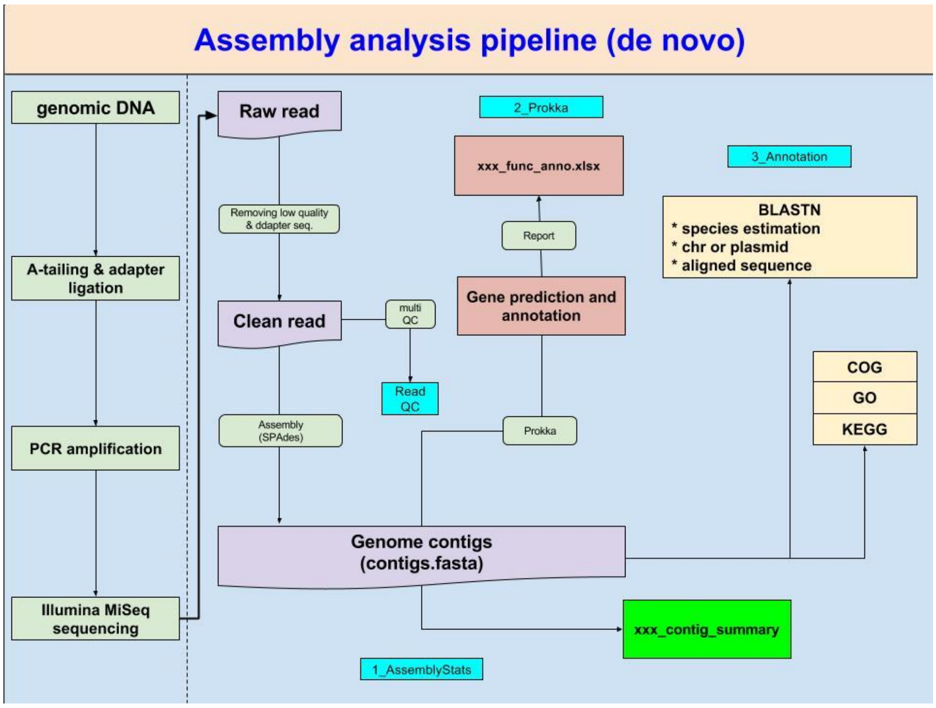
2017

Genomics NGS Analysis Team

Table of Contents

Amplicon Report Folder Instruction

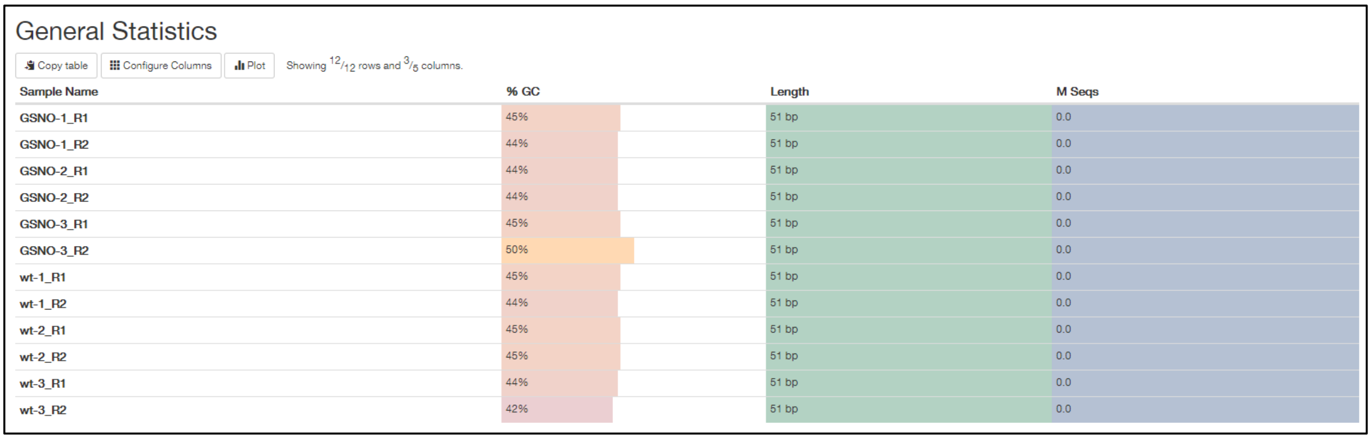
```
|--- <PB_ID>_<Sample_name>_report
|          |--- 0_ReadQC
|          |          |--- multiqc_data
|          |          |--- multiqc_report.html (Read QC report)
|          |--- 1_Assembly
|          |          |--- AssemblyStats (assembly report folder)
                                  …
|          |          |--- xxx.contig_summary.xlsx
|          |          |--- contigs.fasta
|          |--- 2_Prokka (gene prediction result)
|          |          |--- xxx.faa/fna/fsa/gbk/gff
                                  …
|          |          |--- README.txt (prokka output manual)
|          |--- 3_Annotation
|          |          |--- COG (protein group function annotation)
|          |          |--- GO (gene function ontology)
|          |          |--- KEGG (functional pathway annotation)
|--- xxx_func_anno.xlsx
|--- xxx_contig_anno.xlsx
|--- Help.pdf
```



Bioinformatics analysis

Read QC (0_ReadQC)

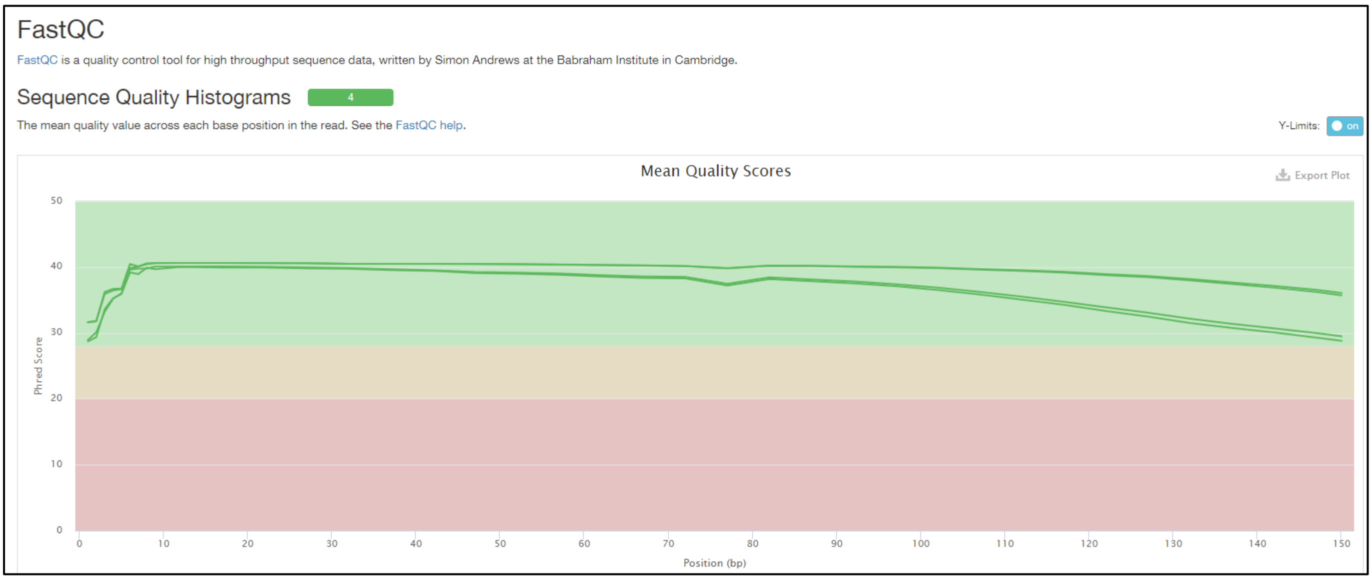
We are using “**MultiQC v1.2**” for evaluating read quality. MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples [1].



General Statistics

Showing 12/12 rows and 3/5 columns.

Sample Name	% GC	Length	M Seqs
GSNO-1_R1	45%	51 bp	0.0
GSNO-1_R2	44%	51 bp	0.0
GSNO-2_R1	44%	51 bp	0.0
GSNO-2_R2	44%	51 bp	0.0
GSNO-3_R1	45%	51 bp	0.0
GSNO-3_R2	50%	51 bp	0.0
wt-1_R1	45%	51 bp	0.0
wt-1_R2	44%	51 bp	0.0
wt-2_R1	45%	51 bp	0.0
wt-2_R2	45%	51 bp	0.0
wt-3_R1	44%	51 bp	0.0
wt-3_R2	42%	51 bp	0.0



FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

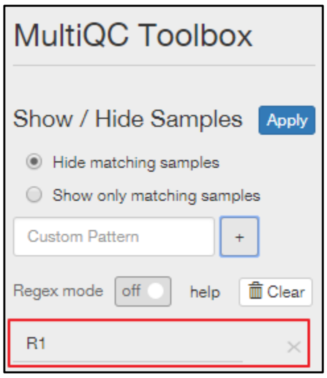
Sequence Quality Histograms

The mean quality value across each base position in the read. See the FastQC help.

[Notice]:

Using “Toolbox” in the right panel to help you show/hide samples.

Red square: mask all name containing “R1” sample.



MultiQC Toolbox

Show / Hide Samples

● Hide matching samples
○ Show only matching samples

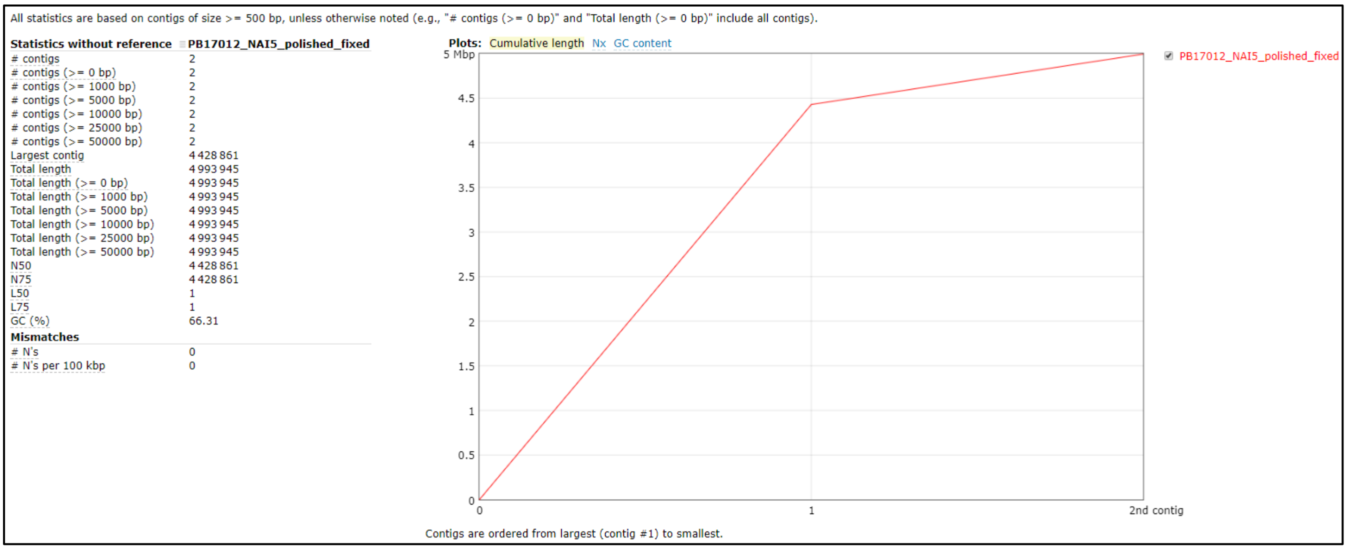
Regex mode off

R1

1. Assembly Stats (1_Assembly)

Sequencing work is completed by MiSeq, and we are using popular assembly tool “**SPAdes v3.10.1**” for genome assembly [2]. As assembly work complete, we are using “**QUAST v4.5**” for evaluating the assembled amplicon quality. [3]

Report.html (Assembly Summary)



All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	PB17012_NAI5_polished_fixed
# contigs	2
# contigs (>= 0 bp)	2
# contigs (>= 1000 bp)	2
# contigs (>= 5000 bp)	2
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	2
Largest contig	4 428 861
Total length	4 993 945
Total length (>= 0 bp)	4 993 945
Total length (>= 1000 bp)	4 993 945
Total length (>= 5000 bp)	4 993 945
Total length (>= 10000 bp)	4 993 945
Total length (>= 25000 bp)	4 993 945
Total length (>= 50000 bp)	4 993 945
N50	4 428 861
N75	4 428 861
L50	1
L75	1
GC (%)	66.31
Mismatches	
# N's	0
# N's per 100 kbp	0

Plots: Cumulative length Nx GC content
5 Mbp
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2nd contig
☑ PB17012_NAI5_polished_fixed
Contigs are ordered from largest (contig #1) to smallest.

<u>Note</u>:

- **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. Gene prediction / annotation (2_Prokka)

Assembled genome annotation is the process of identifying features of interest in a set of 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 amplicon finished, we would like to know what kind of and how much of genes, RNAs, and other elements existed in. Prokka is a powerful tool which could help for downstream gene/CDS/RNA sequence prediction and using blast method against the uniprot/swissprot bacterial database (reviewed).

[**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.
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_func_anno.xlsx**" in the root path directly. ***

3. Annotation (3_Annotation)

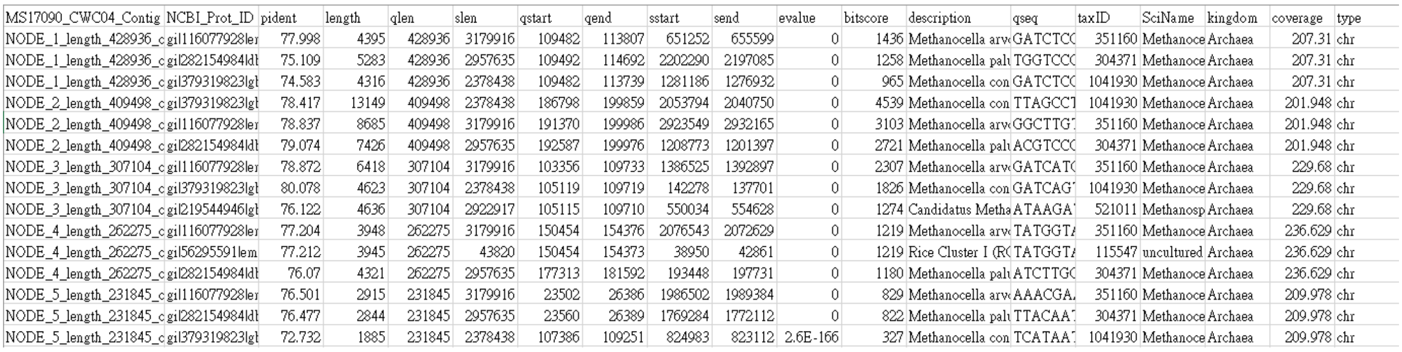
After gene prediction work complete. Our functional annotation works including:

- Targeted BLAST work (advanced annotation for contigs & proteins)
- COG (protein group ortholog)
- GO (gene functional ontology)
- KEGG pathway (gene functional pathway)

[Assembled Contigs BLAST work (blastn)]

In addition to prokka focusing on protein functional annotation, we also used BLAST method against to whole nt database for all of the assembled contig annotation.

[**xxx_contig_anno.xlsx**]



MS17090_CWC04_Contig	NCBI_Prot_ID	pident	length	qlen	slen	qstart	qend	sstart	send	evalue	bitscore	description	qseq	taxID	SciName	kingdom	coverage	type
NODE_1_length_428936_c	gi|116077928|er	77.998	4395	428936	3179916	109482	113807	651252	655599	0	1436	Methanocella arv	GATCTCC	351160	Methanoce	Archaea	207.31	chr
NODE_1_length_428936_c	gi|282154984|db	75.109	5283	428936	2957635	109492	114692	2202290	2197085	0	1258	Methanocella palu	TGGTCCC	304371	Methanoce	Archaea	207.31	chr
NODE_1_length_428936_c	gi|379319823|gb	74.583	4316	428936	2378438	109482	113739	1281186	1276932	0	965	Methanocella con	GATCTCC	1041930	Methanoce	Archaea	207.31	chr
NODE_2_length_409498_c	gi|379319823|gb	78.417	13149	409498	2378438	186798	199859	2053794	2040750	0	4539	Methanocella con	TTAGCCT	1041930	Methanoce	Archaea	201.948	chr
NODE_2_length_409498_c	gi|116077928|er	78.837	8685	409498	3179916	191370	199986	2923549	2932165	0	3103	Methanocella arv	GGCTTGT	351160	Methanoce	Archaea	201.948	chr
NODE_2_length_409498_c	gi|282154984|db	79.074	7426	409498	2957635	192587	199976	1208773	1201397	0	2721	Methanocella palu	ACGTCCC	304371	Methanoce	Archaea	201.948	chr
NODE_3_length_307104_c	gi|116077928|er	78.872	6418	307104	3179916	103356	109733	1386525	1392897	0	2307	Methanocella arv	GATCATC	351160	Methanoce	Archaea	229.68	chr
NODE_3_length_307104_c	gi|379319823|gb	80.078	4623	307104	2378438	105119	109719	142278	137701	0	1826	Methanocella con	GATCAGT	1041930	Methanoce	Archaea	229.68	chr
NODE_3_length_307104_c	gi|219544946|gb	76.122	4636	307104	2922917	105115	109710	550034	554628	0	1274	Candidatus Metha	ATAAGAT	521011	Methanosp	Archaea	229.68	chr
NODE_4_length_262275_c	gi|116077928|er	77.204	3948	262275	3179916	150454	154376	2076543	2072629	0	1219	Methanocella arv	TATGGTA	351160	Methanoce	Archaea	236.629	chr
NODE_4_length_262275_c	gi|56295591|em	77.212	3945	262275	43820	150454	154373	38950	42861	0	1219	Rice Cluster I (R	TATGGTA	115547	uncultured	Archaea	236.629	chr
NODE_4_length_262275_c	gi|282154984|db	76.07	4321	262275	2957635	177313	181592	193448	197731	0	1180	Methanocella palu	ATCTTGC	304371	Methanoce	Archaea	236.629	chr
NODE_5_length_231845_c	gi|116077928|er	76.501	2915	231845	3179916	23502	26386	1986502	1989384	0	829	Methanocella arv	AAACGAA	351160	Methanoce	Archaea	209.978	chr
NODE_5_length_231845_c	gi|282154984|db	76.477	2844	231845	2957635	23560	26389	1769284	1772112	0	822	Methanocella palu	TTACAAT	304371	Methanoce	Archaea	209.978	chr
NODE_5_length_231845_c	gi|379319823|gb	72.732	1885	231845	2378438	107386	109251	824983	823112	2.6E-166	327	Methanocella con	TCATAAT	1041930	Methanoce	Archaea	209.978	chr

* pident   Percentage of identical matches

* length   Alignment length

* qlen   Query sequence length

* slen   Subject sequence length

* qstart   Start of alignment in query

* qend   End of alignment in query

* sstart   Start of alignment in subject

* send   End of alignment in subject

* evalue   Expect value

* bitscore   Bit score

* Alnseq   Aligned part of query sequence

* taxID   Subject Taxonomy ID(s), separated by a ';'

* SciName Subject Scientific Name(s), separated by a ';'

* kingdom   Subject Super Kingdom(s), separated by a ';'   (in alphabetical order)

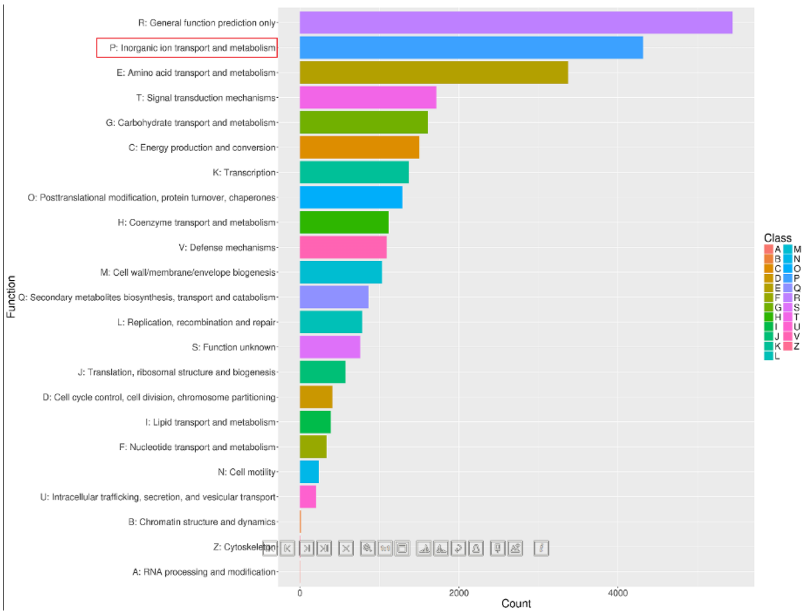
* coverage read coverage on genome

* type   Chromosome / Chromid / Plasmid

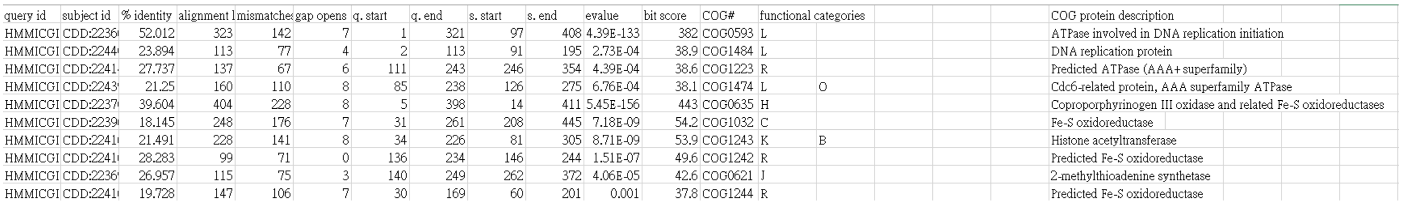
[Protein group function annotation by COG]

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	mismatches	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]:

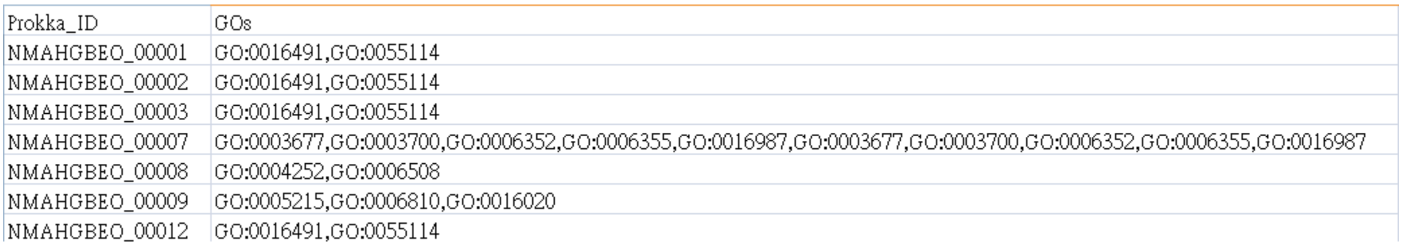
Gene ontology concern with annotation of genes and gene products and to provide centralized access to resources and tools. both GO and COG provide specific information about gene or gene products.

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".

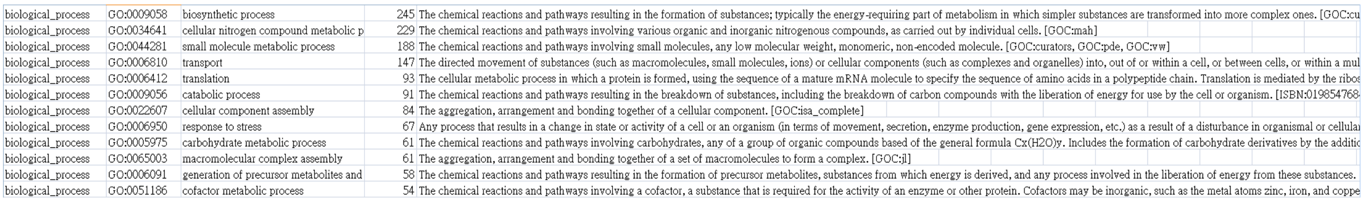
"**InterProscan v5**" is a database which integrates together predictive information about proteins' function from a number of partner resources, giving an overview of the families that a protein belongs to and the domains and sites it contains. We input all of the predicted protein sequences to the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

GeneID_GO.txt (GO terms extraction)



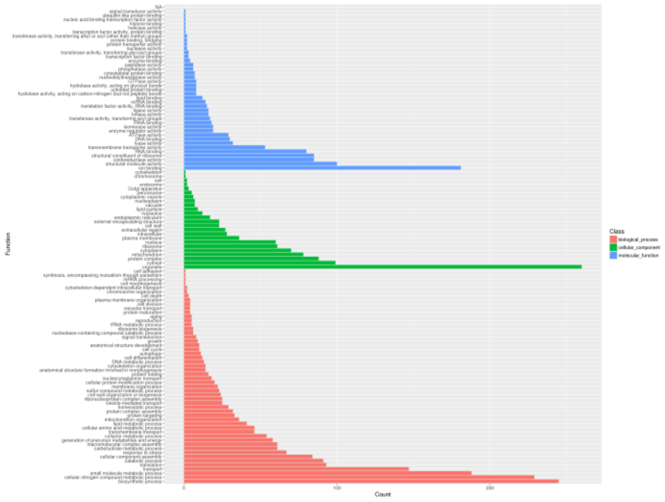
Prokka_ID	GOs
NMAHGBEO_00001	GO:0016491,GO:0055114
NMAHGBEO_00002	GO:0016491,GO:0055114
NMAHGBEO_00003	GO:0016491,GO:0055114
NMAHGBEO_00007	GO:0003677,GO:0003700,GO:0006352,GO:0006355,GO:0016987,GO:0003677,GO:0003700,GO:0006352,GO:0006355,GO:0016987
NMAHGBEO_00008	GO:0004252,GO:0006508
NMAHGBEO_00009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



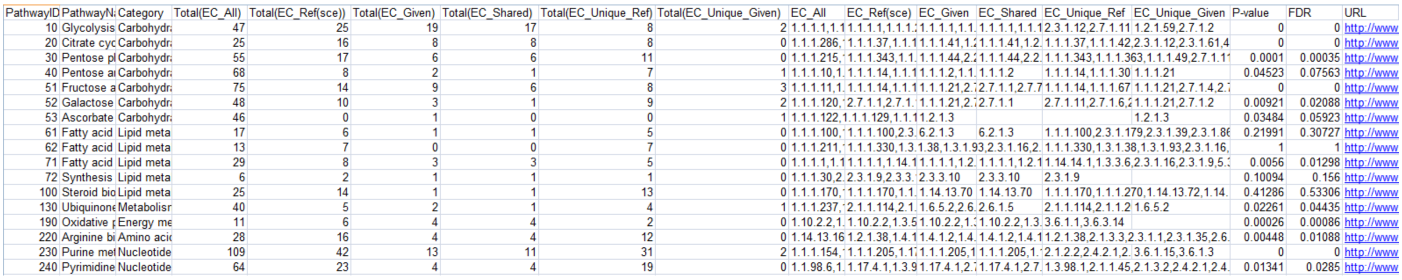
[KEGG pathway annotation]:

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

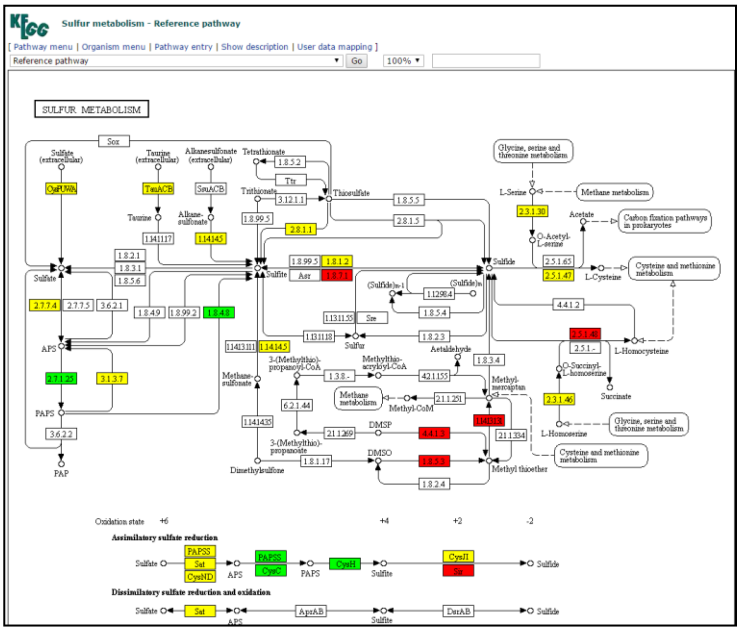


PathwayID	PathwayN	Category	Total(EC_All)	Total(EC_Ref(sce))	Total(EC_Given)	Total(EC_Shared)	Total(EC_Unique_Ref)	Total(EC_Unique_Given)	EC_All	EC_Ref(sce)	EC_Given	EC_Shared	EC_Unique_Ref	EC_Unique_Given	P-value	FDR	URL
10	Glycolysis	Carbohydr	47	25	19	17	8	2	1.1.1.1,1.1	1.1.1.1,1.1.1	1.1.1.1,1.1.	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	Carbohydr	25	16	8	8	8	0	1.1.1.286,	1.1.1.37,1.1.1	1.1.1.41,1.2	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	Carbohydr	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	Carbohydr	68	8	2	1	7	1	1.1.1.10,1.	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	Carbohydr	75	14	9	6	8	3	1.1.1.11,1.	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	1.1.1.21,2.7.1.4,2.7	0	0	http://www
52	Galactose	Carbohydr	48	10	3	1	9	2	1.1.1.120,	2.7.1.1,2.7.1.	1.1.1.21,2.7	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	Carbohydr	46	0	1	0	0	1	1.1.1.122,1.1.1.129,1.1.1		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,	1.1.1.330,1.3.1.38,1.3.1.93,2.3.1.16,2.			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	1.1.1.1,1.2.	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.	1.14.13.70	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	Metabolisr	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.	1.10.2.2,1.3.	3.6.1.1,3.6.3.14		0.00026	0.00086	http://www
220	Arginine bi	Amino acic	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.	1.17.4.1,1.3.9	1.17.4.1,2.7	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

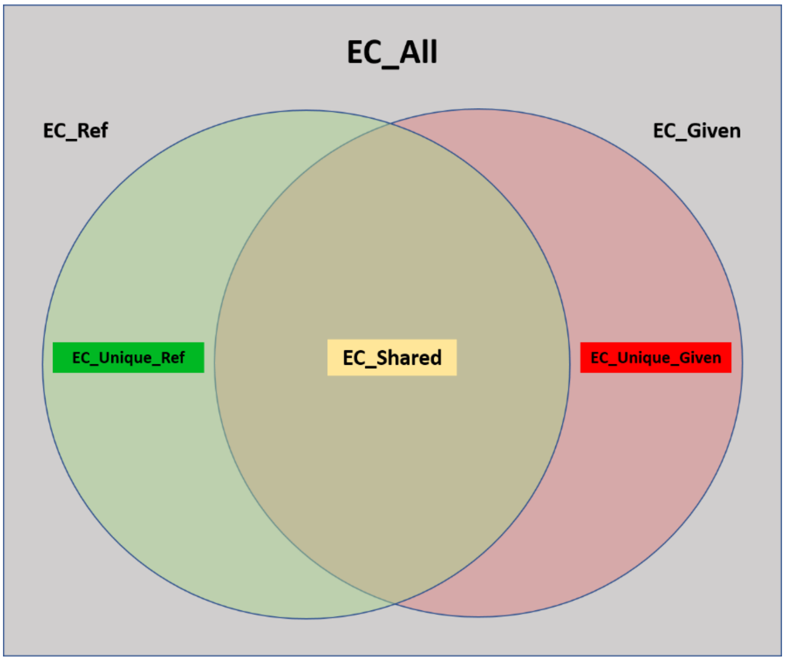


[**Pathway map color definition**]

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 integrated in "xxx_func_anno.xlsx"** ***

4. Reference

1. MultiQC: Summarize analysis results for multiple tools and samples in a single report; Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller; Bioinformatics (2016); doi: 10.1093/bioinformatics/btw354; PMID: 27312411
2. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021.
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:**

- **Notepad++:** https://notepad-plus-plus.org/download (適合觀看所有文字 or 序列檔)
- **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)