|  |
| --- |
| **GENOMICS Tech Solutions Co.,Ltd** |
| De novo Genome Assembly Report |
| **[PB17025]: NAI4** |



**Summary:**

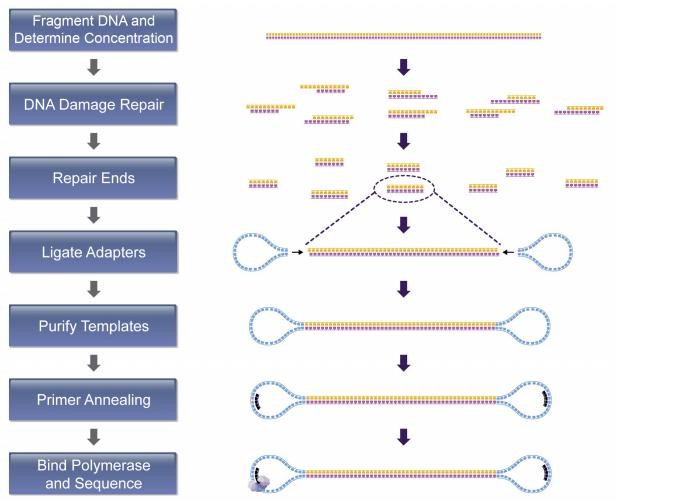
Using PacBio RSII sequencing platform, nearly **1,043 Mb** data was produced from 20K bp library.

Based on the assembly result, Genome size was assembled to nearly **7.36 Mb**, GC content was **62.2%**, the number of contig was **3**.

Based on the gene prediction / annotation result, the genome contained **7,096 genes**, **6,972 CDS**. The RNA part, number of **rRNA was 9**, and number of **tRNA was 62**, number of **tmRNA was 1** and number of **miscRNA was 52**.

INTRODUCTION OF WORKFLOW

1. **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.

1. Pipeline of Bioinformatics Analysis

Bioinformatics analysis will be proceeding after data filtering. The content of bioinformatics analysis pipeline is shown below.



**Figure 2 Pipeline of Bioinformatics Analysis.** (1) These longest reads were used as alignment seeds for the

multi-molecule consensus error correction step of the hierarchical genome assembly process (HGAP). (2) After genome assembly completed, we are using Prokka as gene prediction tool which is packaged with multiple functions including: (a) various RNA prediction, like rRNA, tRNA, tmRNA and miscRNA; (b) Gene / CDS prediction. While we got multiple predicted proteins, followed with protein group function annotation which is blasted against with COG database and using other scripts to generate KEGG functional pathway annotation according to their EC (enzyme commission) number.

**RESULTS**

1. **Raw Data and QC Statistics**

* **[Sequencing / QC statistics]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sample name** | **Pre-filtered** | | | **Filtered** | | |
| **Total output (MB)** | **# of read** | **Read length** | **Total output (MB)** | **# of subread** | **Subread length** |
| **NAI5** | **926.03** | **67319** | **13,755** | **924.35** | **91,191** | **10,136** |

**Subread filtered parameters (cutoff):**

* Minimum Subread Length: 500
* Minimum Polymerase Read Quality: 0.8
* Minimum Polymerase Read Length: 100

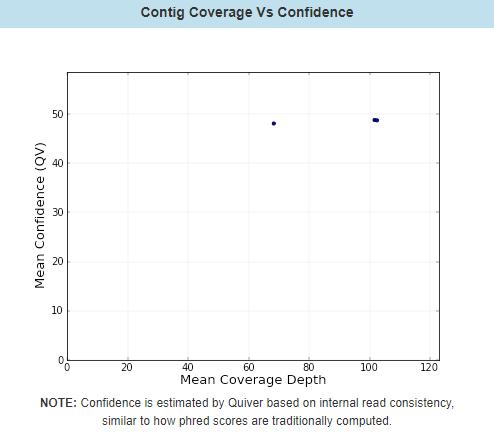
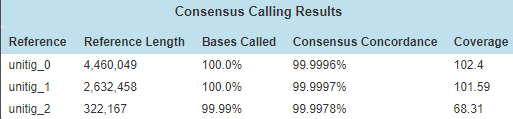
1. Summary of Assembly

* **[Genome assembly statistics]**

|  |  |
| --- | --- |
| **Genome Size** | 7,356,018 |
| **Total scaffold** | 3 |
| **Longest length** | 4,436,225 |
| **N50 / L50** | 4,436,225 / 1 |
| **N75 / L75** | 2,614,809 / 2 |
| **GC content** | 62.2% |
| **Scaffold > 1kb** | 3 |

* **[Genome contig coverage]:**

**[NOTICE]: genome contig coverage data is generated from raw polished assembled contigs instead of fixed polished contigs.**



1. Genome Prediction / Annotation

For the de-novo assembled genome, we are using following tools/database for digging out more information.

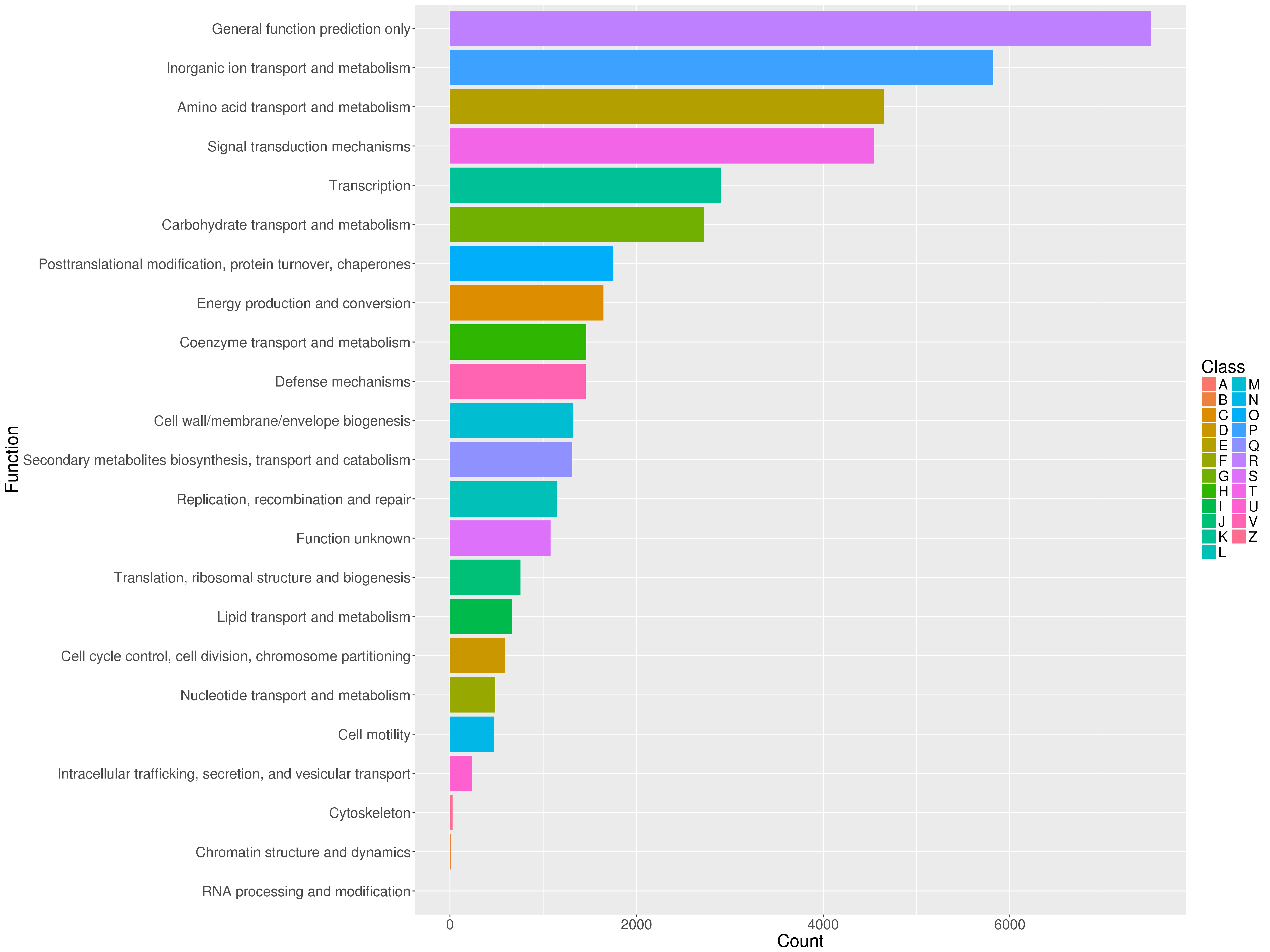
* **[Prokka general gene prediction stats]**

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 is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

|  |  |
| --- | --- |
| **Genome Size** | 7,356,018 |
| **Genes** | 7,096 |
| **CDS** | 6,972 |
| **rRNA** | 9 |
| **tRNA** | 62 |
| **tmRNA** | 1 |
| **misc\_RNA** | 52 |
| **# of Hypothetical protein** | 2,527 |

* **[Protein group function annotation by COG]**
  + **For more detail of explanation, please visit Help.pdf**

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.

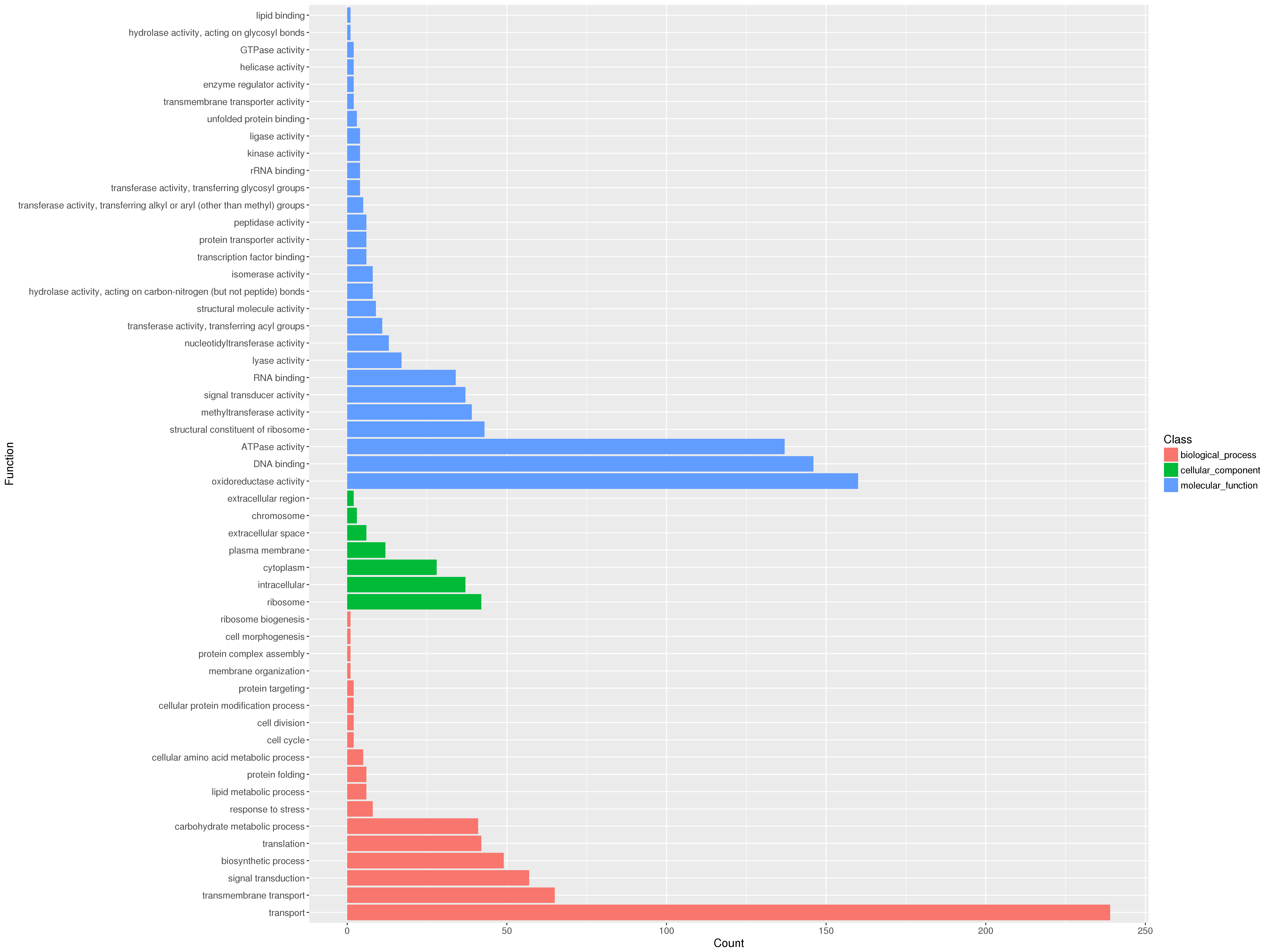


* **[GO functional annotation of predicted genes]**
  + **For more detail of explanation, please visit Help.pdf**

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:

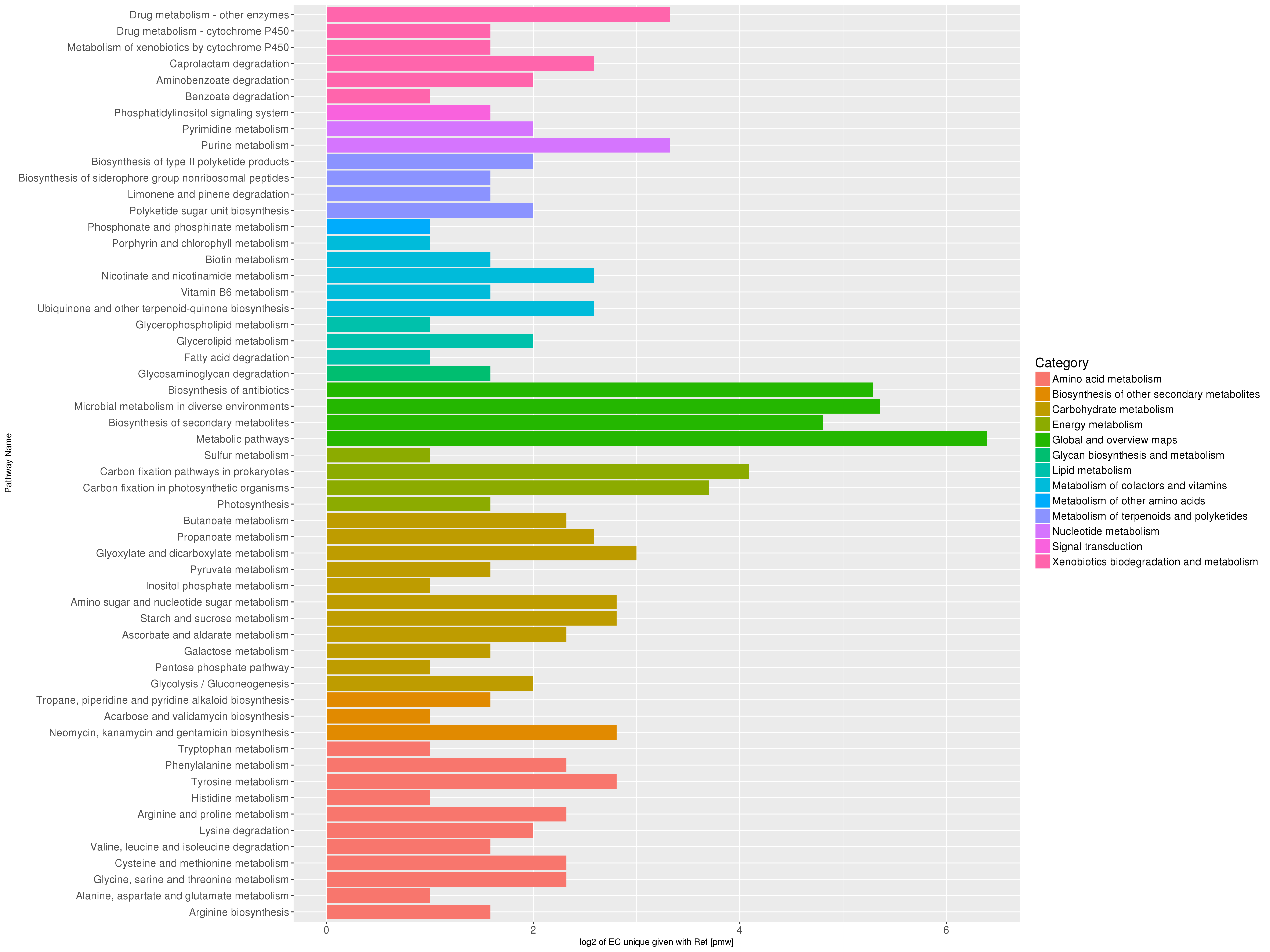
1. **Cellular Component:** These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
2. **Biological Process:** A biological process 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".



* **[Functional pathway annotation by KEGG]:**
  + **For more detail of explanation, please visit Help.pdf**
* Notice: we’re using ***Ensifer adhaerens OV14*** as reference for KEGG pathway search

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

In this step, we are using Enzyme Commission number (EC number) to parse functional pathway result. A set of perl scripts has been developed to perform automated data retrieval from the KEGG database using its new REST program application interface. Enrichment or depletion in metabolic pathways is evaluated using the two-tailed Fisher exact test followed by Benjamini and Hochberg correction (<http://dx.doi.org/10.1186/1751-0473-9-19>).



**[CGview genome plot]**

(only plot the longest scaffold)

