Blast2GO
Command Line Report

October 15, 2015

Version: 1.1.0

Build: 1444836920764

1. Introduction

This report summarises the functional annotation process performed with the Blast2GO Command Line. The command line is based on the Blast2GO methodology, first published in 2005 (Conesa et al., 2005), for the automatic and high-throughput functional annotation and analysis of gene or protein sequences. The method uses sequence alignments (BLAST) to obtain a list of potential homologous for each input sequence. Blast2GO then maps Gene Ontology (GO) terms associated to the obtained BLAST hits and returns an evaluated functional annotation for the query sequences (Götz et al., 2008). Additional steps to improve the quality of the functional annotation are available. The following sections provide more detailed information about the different analysis steps as well as information about the input datasets, used parameters and the overall results.

The analysis started at: *13:56 on October, 15, 2015*
The analysis finished at: *21:53 on October, 15, 2015*

2. Command Line Parameters

The following parameters have been used for this analysis. Please be aware that, additionally to the parameters provided here and in each section, additional parameters can be adjusted in the properties file (.prop) provided with the -properties parameter.

```
-properties Tomato_analysis.prop
-useobo go_latest.obo
-loadfasta /home/mariana/Documents/Analysis/Tomato/ITAG2.3_cds.fasta
-cloudblast B2G-MONTMARI-********************************
-loadips50 /home/mariana/Documents/Analysis/Tomato/IPS/
-mapping
-annotation
-annex
-goslim goslim_plant.obo
-tempfolder /home/mariana/Documents/Analysis/Tomato/
-gograph
-statistics all
-workspace /home/mariana/Documents/Analysis/Tomato/
-nameprefix Tomato_Analysis
-savedat
-savereport
-saveb2g
-saveannot
-savelog /home/mariana/Documents/Analysis/Tomato/
-saveseqtable
```

3. Input Files

This section lists the following input file types: Fasta sequences, Blast and InterProScan XML results and annotation files.

- Import of *34727* sequences from .fasta: `/home/mariana/Documents/Analysis/Tomato/ITAG2.3_cds.fasta`
- Import of *34722* sequences from InterPro .xml (v5.0): `/home/mariana/Documents/Analysis/Tomato/IPS`

4. Sequences

This analysis has been performed with 34727 sequences provided in fasta format. The following figure shows the sequence length distribution.

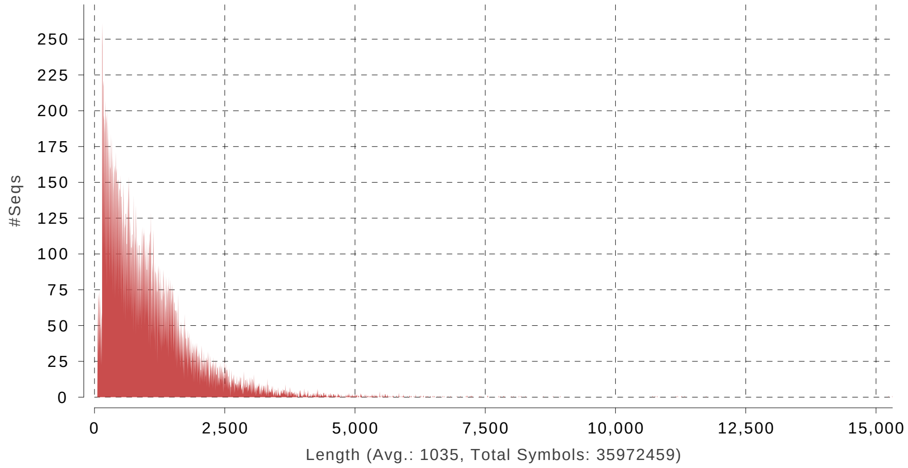

Figure: Distribution of the sequence length

5. CloudBlast

Blast finds regions of local similarity between sequences. The Blast program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches (Altschul et al.,1990). In this scenario Blast is used to infer functional relationships between sequences. The blast step has been performed with CloudBlast, a system which allows to execute NCBI Blast+ (Altschul et al.,1990) against public sequence databases on a high performance computation cluster. The following parameters have been used.

CloudBlast summary:

Blasted sequences: *34727*

Sequences with hits: *31374*

Parameters:

Name	**Value**
Blast Program	*blastx-fast*
Blast DB	*Viridiplantae (nr subset) [viridiplantae, taxa:33090] from 30.09.2015*
Blast Expectation Value (e-Value)	*1.0E-3*
Word Size	*6*
Low Complexity Filter	*true*
Filter by Description	
Number of Blast Hits	*20*

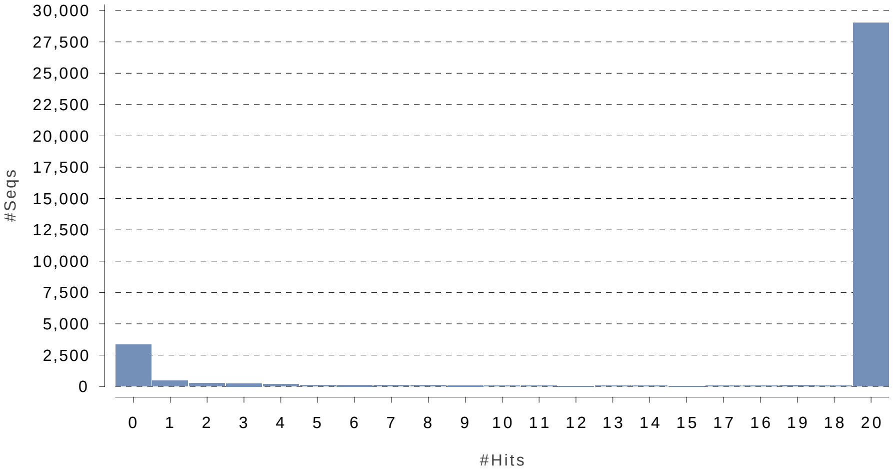

Figure: Distribution chart showing the number of Blast hits

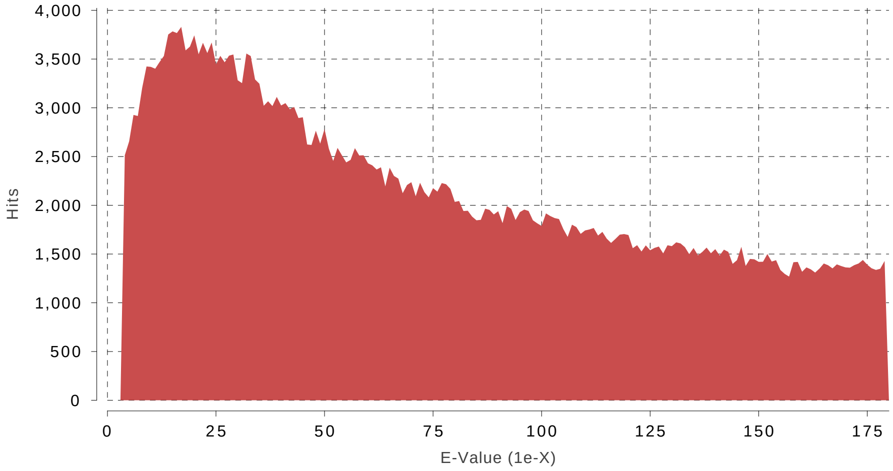

Figure: Evalue distribution without exact matches (eValue < 1e-180)

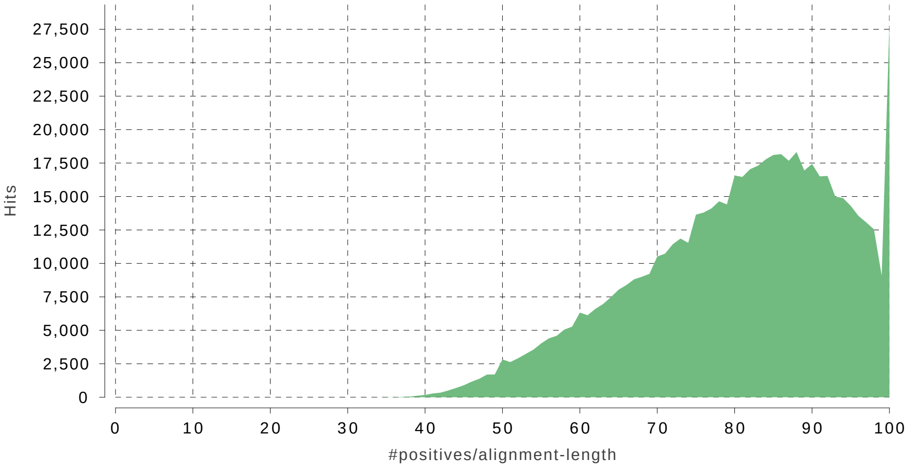

Figure: Similarity distribution

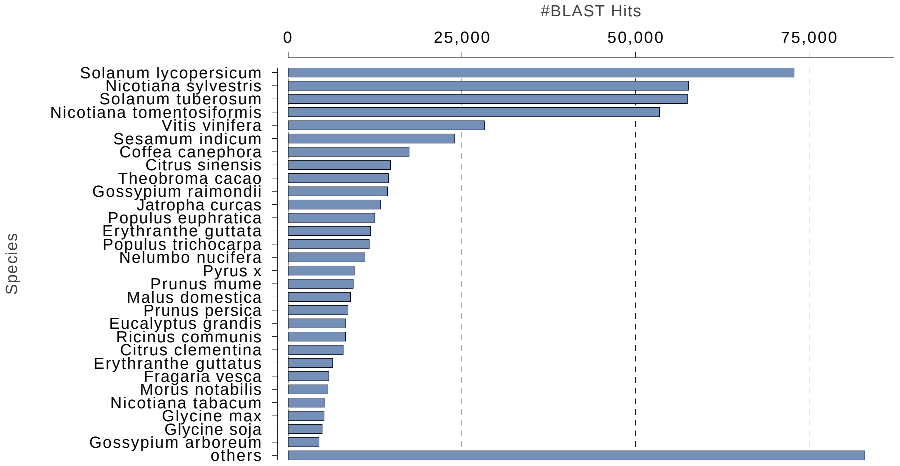

Figure: Species distribution

6. Gene Ontology Mapping

Mapping is the process of retrieving Gene Ontology (Ashburner et al., 2000) terms associated to the hits obtained after a BLAST search. To run mapping, select one or various data-sets, which contain blasted sequences and execute the mapping function.

Blast2GO performs different mapping steps to link all BLAST hits to the funtional information stored in the Gene Ontology database. Therefore Blast2GO uses different public resources provided by the NCBI, PIR and GO to link the different protein IDs (names, symbols, GIs, UniProts, etc.) to the information stored in the Gene Ontology database - the GO database contains several million functionally annotated gene products for hundreds of different species. All annotations are associated to an Evidence Code which provides information about the quality of this functional assignment.

1. BLAST result accessions are used to retrieve gene names or Symbols making use of two mapping files provided by NCBI. Identified gene names are than searched in the species specific entries of the GO database.
2. BLAST result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.
3. BLAST result accessions are searched directly in the GO database.

Mapping summary:

GO Mapping Database Name: *b2g_jul15*

Mapped sequences: *24839*

Sequences that could not be mapped: *9888*

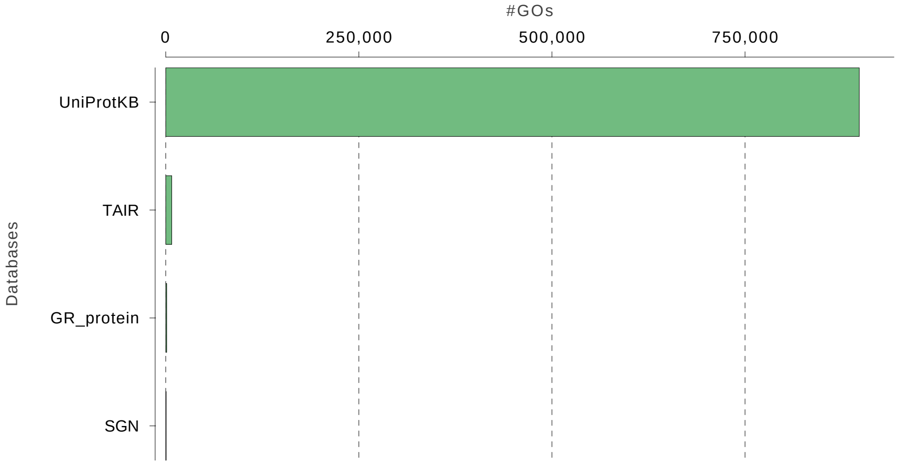

Figure: GO Mapping Source Database Distribution

7. Functional Annotation

This is the process of selecting GO terms from the GO pool obtained during the Mapping step and assigning them to the query sequences. The annotation is carried out by applying an annotation rule (AR) on the found ontology terms. The rule seeks to find the most specific annotation with a certain level of reliability. This process is adjustable in specificity and stringency. For each candidate GO an annotation score (AS) is computed. The AS is composed of two additive terms. The first, direct term (DT), represents the highest hit similarity of this GO weighted by a factor corresponding to its evidence code. The second term (AT) of the AS provides the possibility of abstraction. This is defined as annotation to a parent node when several child nodes are present in the GO candidate collection. This term multiplies the number of total GOs unified at the node by a user defined GO weight factor that controls the possibility and strength of abstraction. When GO weight is set to 0, no abstraction is done. Finally, the AR selects the lowest function (GO term) per branch that lies over a user defined threshold, the Annotation Cut-Off.

1. E-Value Hit Filter. This value can be understood as a pre-filter: only GO terms obtained from hits with a greater e-value than given will be used for annotation and/or shown in a generated graph (default: 1.0E-6).
2. Annotation Cut-Off (threshold).The annotation rule selects the lowest term per branch that lies over this threshold (default: 55).
3. GO-Weight. This is the weight given to the contribution of mapped children terms to the annotation of a parent term (default: 5).
4. Hsp-HitCoverage CutOff. Sets the minimum needed coverage between a Hit and his HSP. For example a value of 80 would mean that the aligned HSP must cover at least 80% of the longitude of its Hit. Only annotations from Hit fulfilling this criterion will be considered for annotation transference.
5. EC-Weight. EC code weights can be modified in the command line properties file (cli.prop), be default located in the blast2go_cli folder in the home directory. Note that in case influence by evidence codes is not wanted, you can set them all at 1. Alternatively, when you want to exclude GO annotations of a certain EC (for example IEAs), you can set this EC weight at 0.

Parameters:

Name	**Value**
Annotation CutOff	*55*
E-Value-Hit-Filter	*1.0E-6*
GO Weight	*5*
HSP-Hit Coverage CutOff	*0*
Filter GO by Taxonomy	*No Filter*

Annotation summary:

Annotated sequences: *21329*

Sequences that could not be annotated: *13398*

Assigned Gene Ontology terms: *90452*

Assigned enzyme codes: *6819*

Sequences with enzyme codes assigned: *5967*

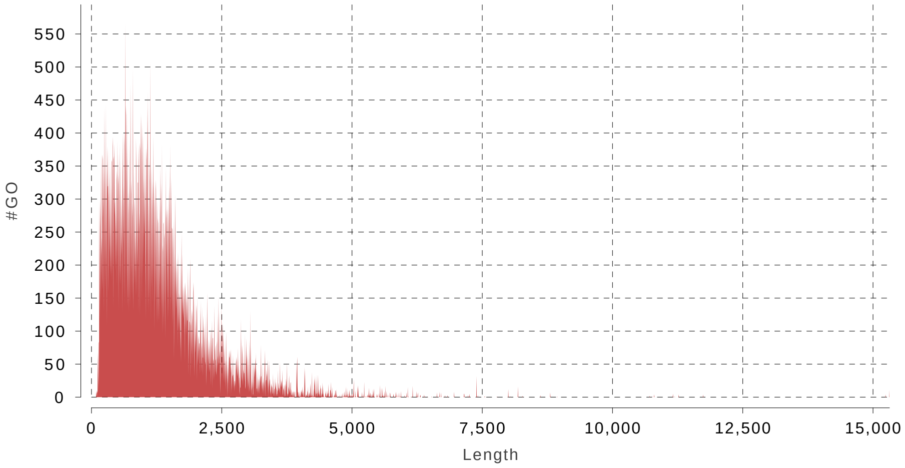

Figure: This chart shows the number of Gene Ontology terms corresponding to its sequence length.

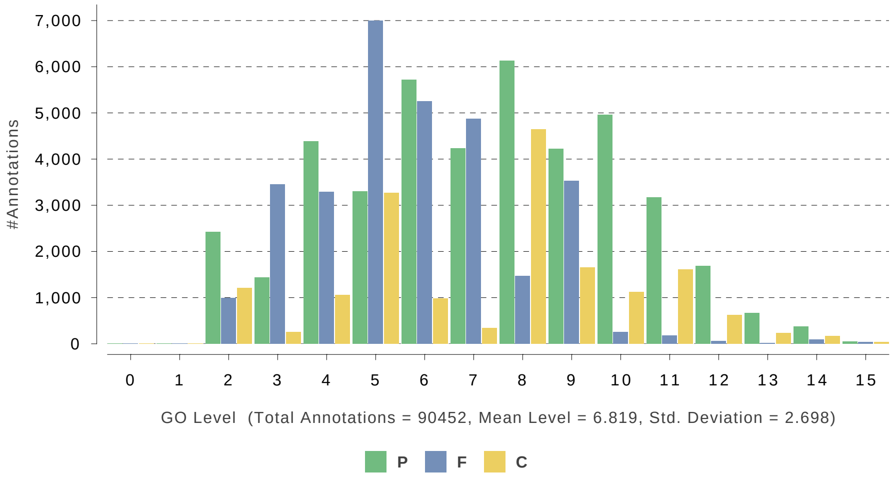

Figure: GO Annotation Level Distribution

8. Merge InterProScan

The Merge InterProScan step adds functional information obtained through domain based searches to the existing annotations. Once annotations are added, a validation is done which removes redundant, more general functions based on the true path rule.

Merge InterProScan summary:

Annotations before merge: *90452*

Afterwards: *95879*

Comfirmed: *107340*

Too general: *12260*

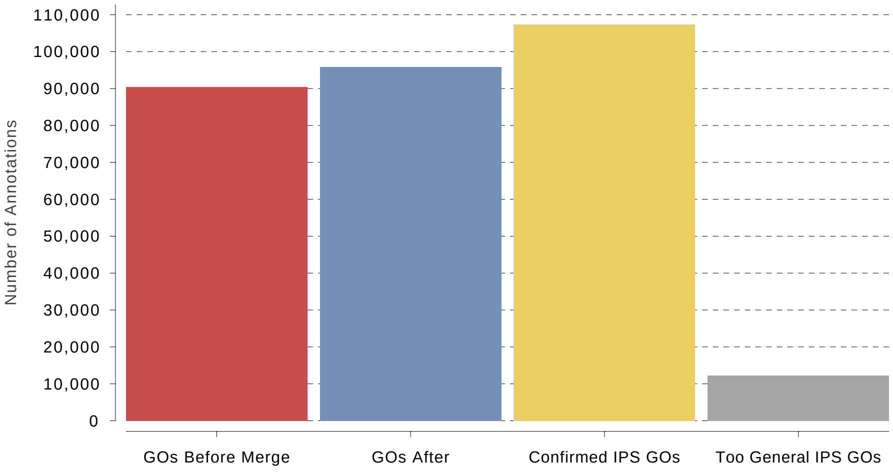

Figure: Result details of the Merge InterProScan step showing the amount of annotations which could be added as domain based functional information.

9. ANNEX

ANNEX is a set of relationships between the terms of different Gene Ontology categories. These relationships consist of over 6000 manually reviewed links between molecular functions involved in biological processes and molecular function terms acting in cellular components (Myhre et al., 2006). In this way this analysis step complements existing functional annotations by adding further implicit terms based on these relationships.

ANNEX summary:

Total original annotations: *95879*

New annotations added via ANNEX: *9574*

More general annotations replaced by more specific ANNEX annotations: *1364*

Annotations confirmed by ANNEX: *3161*

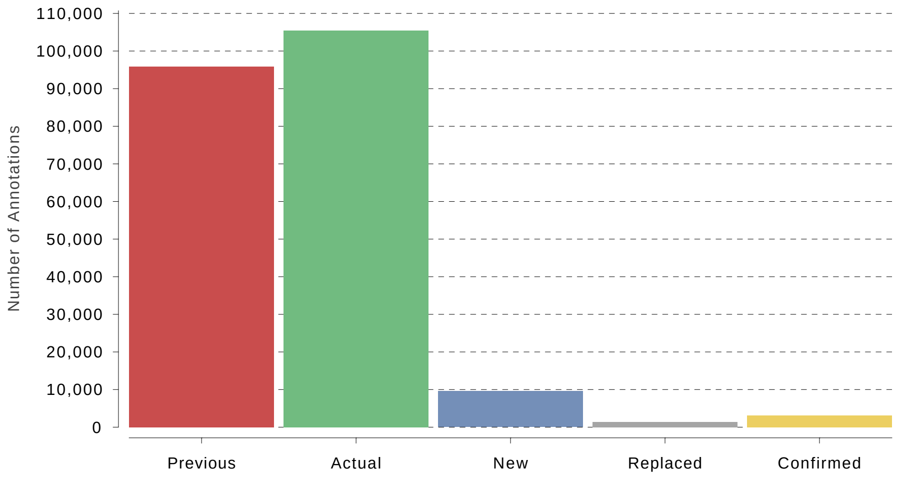

Figure: Annotations before and after applying the ANNEX augmentation.

10. Annotation Results

From a total of *34727* CDS sequences, *9.45%* were analysed with BLAST but fail to obtain significant hits. *15.37%* of the sequences return significant sequence alignments but can not be linked to any Gene Ontology entries. *6.59%* of the GO mapped dataset does not obtain an annotation assignment. Overall we can assign functional labels to *68.59%* of the input sequences. Enzyme codes could be assigned to *17.18%* of the sequences.

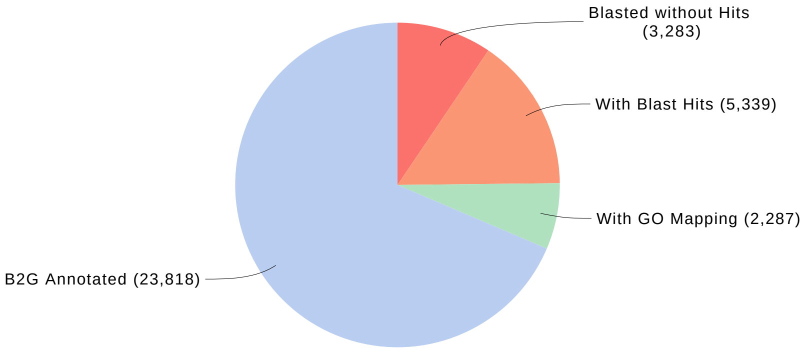

Figure: The data distribution pie chart shows the amount of sequences which could finally be annotated in comparison to the ones not annotated due to missing results in the blast, mapping or annotation step.

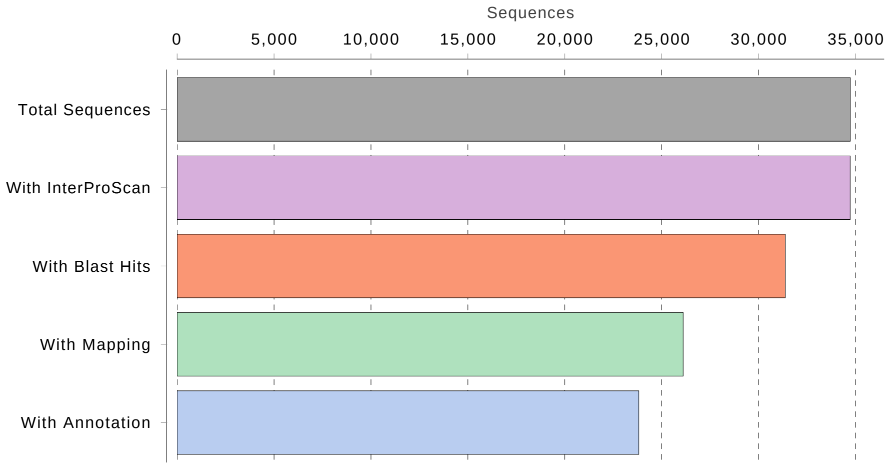

Figure: The Analysis Progress shows the total amount of sequences which obtained results during the different analysis steps. Please note that for example the total amount of mapped (green) sequences cannot be higher than the number of blasted (orange) sequences.

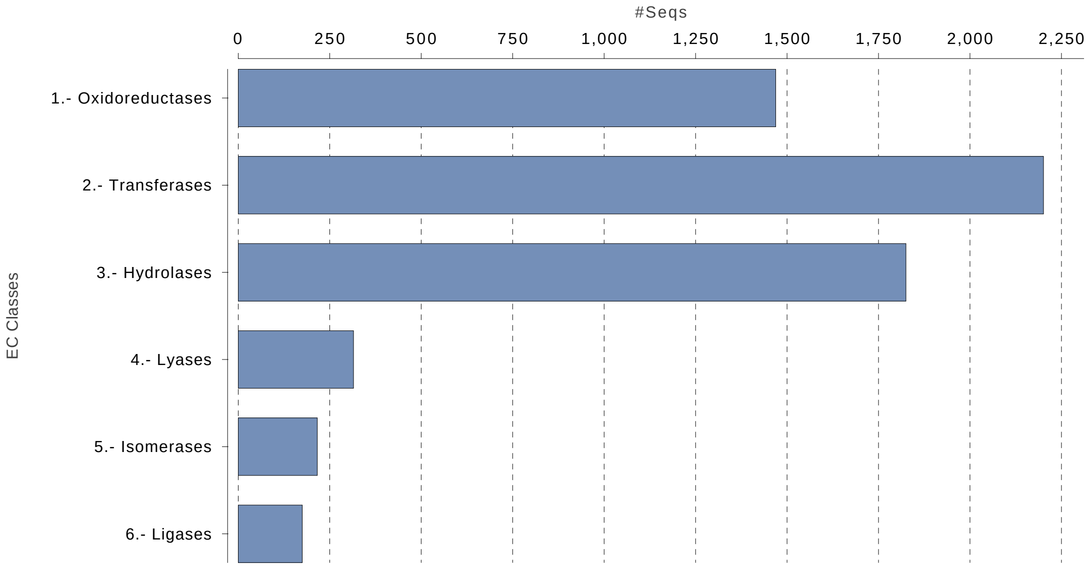

Figure: Main Enzyme Code Level Distribution

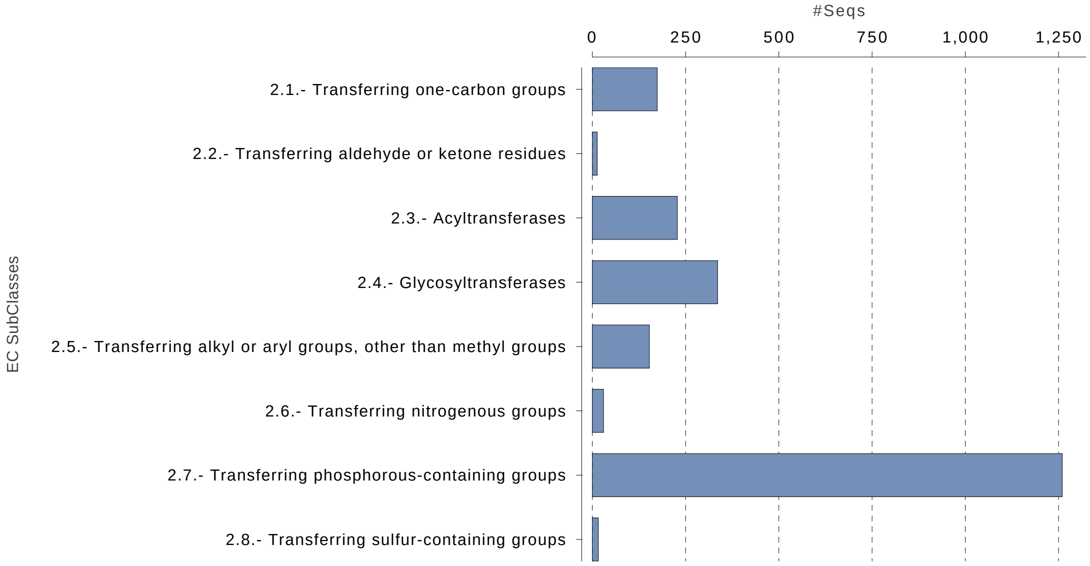

Figure: Enzyme Code Level 2 Distribution

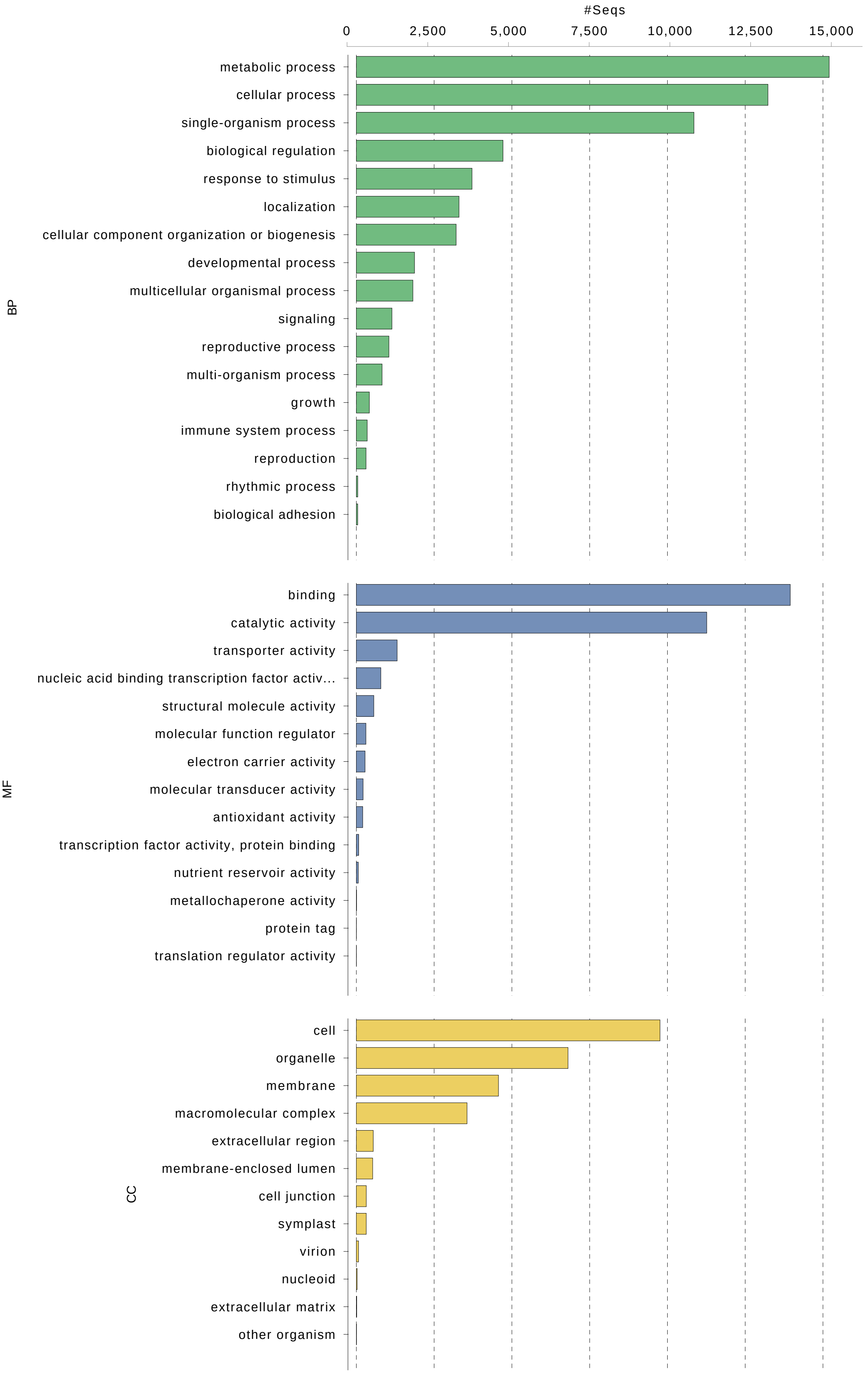

Figure: GO Distribution in Level 2 (Top 50)

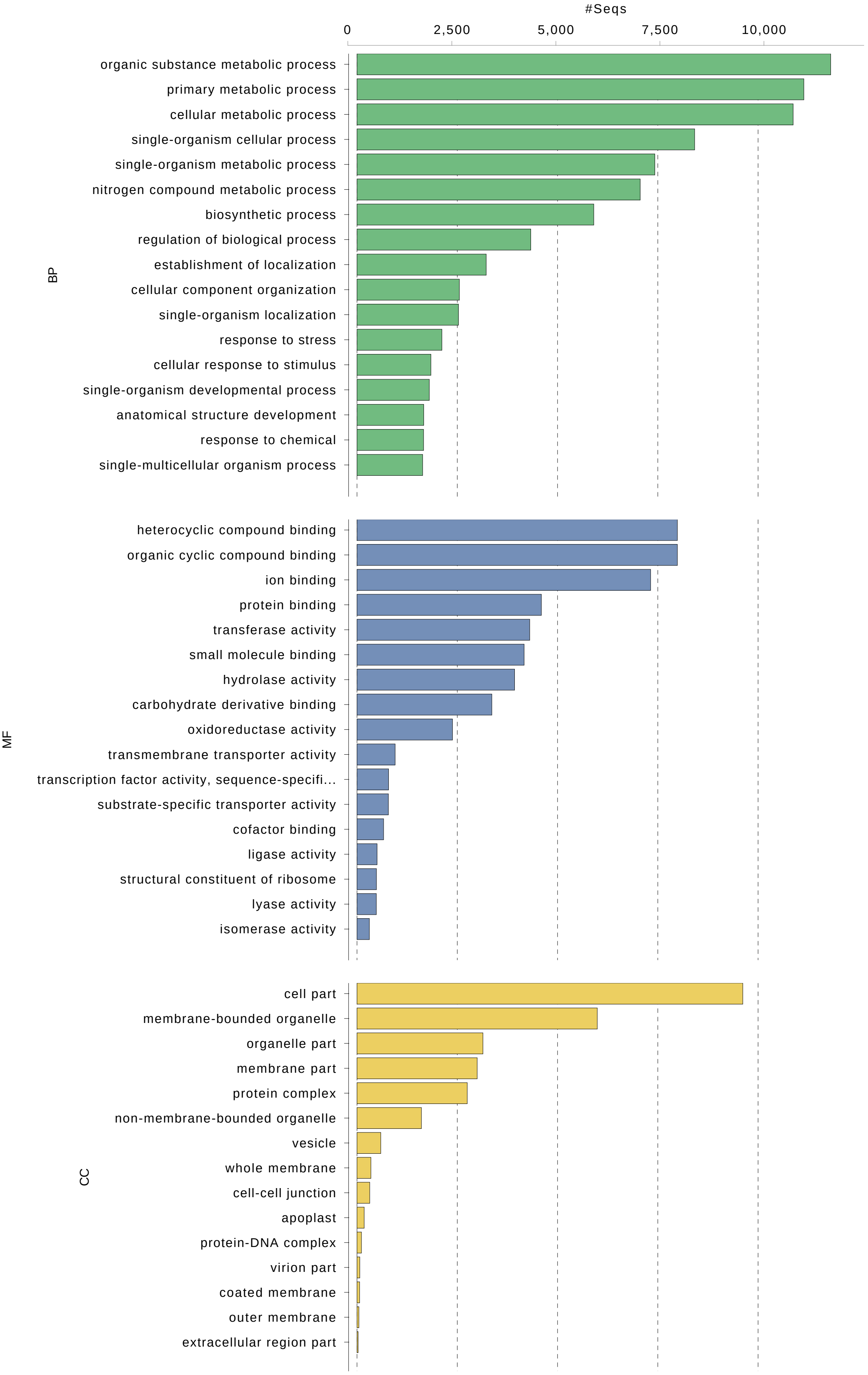

Figure: GO Distribution in Level 3 (Top 50)

11. GO-Slim

GO slims are cut-down versions of the Gene Ontology containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. GO slims are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required.

GO-Slim Summary:

The main Gene Ontology .obo file data version is from *July, 2015*, the filepath to the GO-Slim file is */home/mariana/Documents/Testing/CommandLine/blast2go_cli_v1.1.0/goslim_plant.obo*.

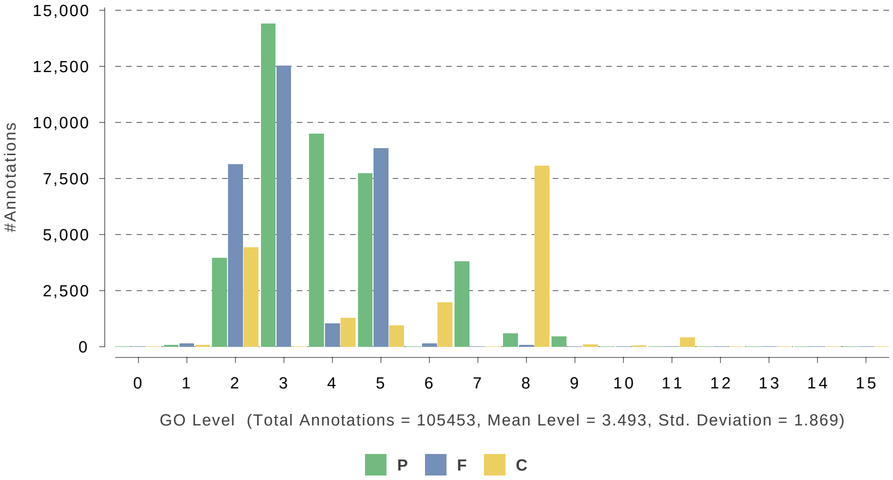

Figure: GO Annotation Level Distribution after GO-Slim

12. References

- Conesa et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics , 21(18):3674 - 3676.
- Götz et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucl. Acids Res. 36 (10):3420-3435.
- Ashburner et al. (2000) Gene ontology: tool for the unification of biology Nat Genet 25(1):25-9.
- The Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. Nucl Acids Res 43 Database issue D1049–D1056.
- Altschul et al. (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.
- Zdobnov E.M. and Apweiler R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9):847-8.
- Myhre et al. (2007) Additional gene ontology structure for improved biological reasoning. Bioinformatics, 22(16):2020-2027.