Blast2GO User Manual

Blast2GO File Creation for NCBI Submission of Complete Eukaryotic Genomes or Chromosomes

May, 2016

BioBam Bioinformatics S.L.
Valencia, Spain

Contents

1 NCBI Database

The most important source of new data for GenBank is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries, updates existing entries, assists authors with submission of new data.

2 NCBI Submission Submission Tool

Thi NCBi data submission tool facilitates the creation of a Genbank ready for submission. The tool combines a reference genome (fasta file), the gene coordinates (gff file) and the functional annotations of Blast2GO, creating a feature table which will be validated with the tbl2asn program. The 'tbl2asn' command-line program is used to automate the creation of sequence records (.sqn files). For more information about tbl2asn visit: `http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/`.

When fisrt executed, the App will download the NCBI program 'tbl2asn' for your operating system allowing execution on Win, Mac and Linux.

Note: This tool requires an internet connection to download and execute the tbl2asn program and to execute it, as it requieres a conection to the NBI databases in order to validate the annotations.

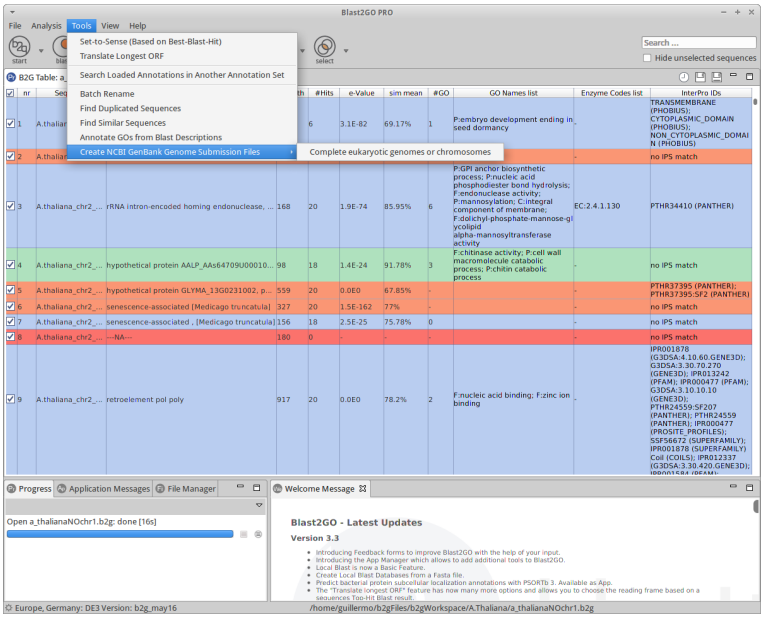


Figure 1: NCBI Submission Tool

3 General Workflow

To succesfully submit the annotated sequences, it is first necessary to prepare the source of the annotations, i.e. the reference genome to which the sequences belong, the position on the genome, and the functional annotation. These files are processed by Blast2GO and validated by 'tbl2asn' program to create the ASN1 file (.sqn) and the validation files.

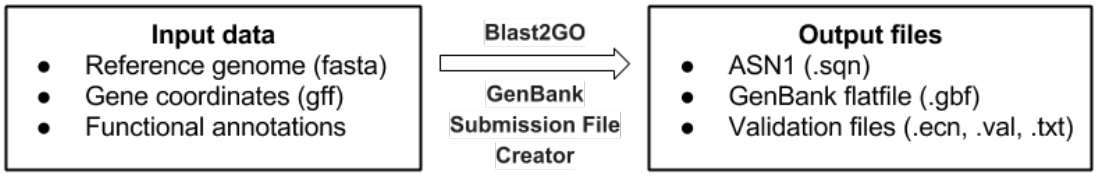


Figure 2: General Workflow

3.1 Input Files

Three elements are necessaty to create the submission files:

- **Reference genome**: This file provides the foundational nucleotide sequence and may contain one or more chromosomes The chromosome names in the fasta description line have to match the GFF file name.
- **Genomic annotation**: This data is provide by the GFF3 files, and is also used to link the Blast2GO annotations and the genome reference sequences in the Fasta file. A GFF3 is necessary for each chromosome, with the file name matching the chromosome name as it appears in the fasta file. The sequences name used in the Blast2GO project should appear in the feature column in the GFF3 file. The corresponding feature ID can be specified as parameter (default is seqName).
- **Functional annotation**: This information is provided by your Blast2GO project, and is intended to provide the functional features of your sequences, including gene names, Gene Ontology terms and enzyme numbers. The option to create the submission file is only activated when a Blast2GO Project file is loaded and selected. The sequence name of the functional annotation in your Blast2GO project has to match with a feature of your choice in the gff file.

3.1.1 Preparing your data

In order to integrate all the information and create the NCBI submission files, we need to create informative links between them. As discussed above, the GFF3 files act as a link between the genome sequence and the functional annotation.

The FASTA, or multi-FASTA file may contain one or more chromosomes. The chromosome names in the fasta description line have to match the GFF3 name, additionally, the sequence name used in the Blast2GO project should appear in the feature column in the GFF3 file. The corresponding feature ID can be specified as a parameter (for the GFF files created by Augustus and Glimmer included in Blast2GO, the IDs correspond to 'seqName').

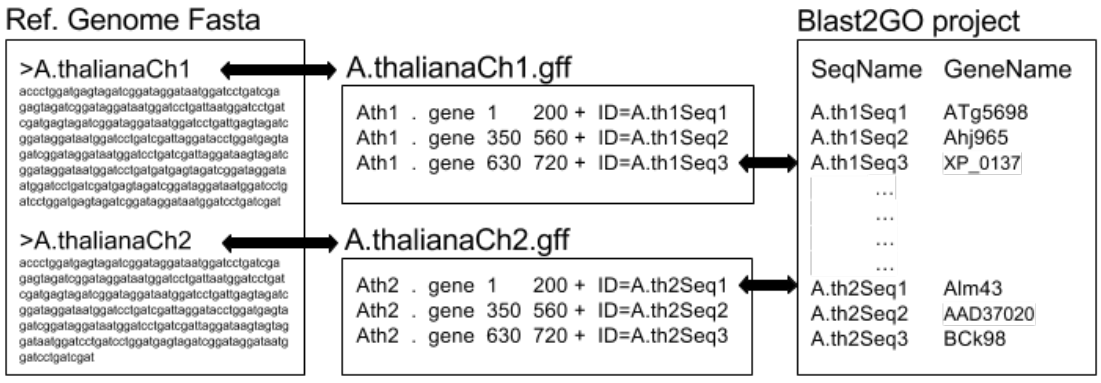


Figure 3: Information integration schema

4 Wizard pages

4.1 Page 1: Project Details

- **Locus tag**: The 'locus tag' is an alfanumeric identifier of your project provided by the NCBI or user determined at the moment of the BioProject registration at: `https://submit.ncbi.nlm.nih.gov/subs/bioproject/`.
- **Laboratory ID**: The laboratory ID is a unique tag that refers to your own laboratory and allows the sequences to be associated with it.
- **Submission type**: Here you can choose the type of submission you want to perform.
  - One or a few nucleotide sequences: use this option if you have a small dataset containing few sequences (less than a cromosome, or a chromosome on scaffold stage).
  - Complete eukaryotic genomes or chromosomes: use this option if you have a complete data set without N's, conforming a chromosome or a whole genome.
  - Incomplete genomes (WGS): use this option if your dataset consist of incomplete genomic or chromosomal assemblies derived from shotgun sequencing methods.
- **Assembly details (only for WGS submission)**: These details provide information about the more technical steps of the assembly. Here we can find:
  - Assembly method: The program or algorithms used assemble the genome.
  - Assembly name: This is a short project identifier.
  - Long assembly name: This is a larger and more explanatory name of your project.
  - Genome coverage: This is the mean genome coverage obtained by the assembler, and has a general format of one or more digits followed by an 'x' (e.g: 12x or 76x).
  - Sequencing technology: The name of the technology used to perform the sequencing of the query genome. If the technology used is not in the list shown, you can manually enter the name.
- **Optional source qualifiers**: These are addtional sequence qualifiers to your all project, specifications of optional qualifiers allows you to add useful information regarding the organism chromosome, type, etc. If you are going to submit a WGS project,add the source information as organism and the relevant strain, breed, cultivar or isolate, if exists for the sequenced organism. *Note: the 'gcode' corresponding to the genetic code is only mandatory if the submitting organism is not specified or is not in the NCBI Taxonomy Browser.*

4.2 Page 2: Sequence Data and Annotation Files

- **Output Directory**: The creation and validation of the submitting sequences will produce multiple files that may be checked. This option allows the files to be saved in an existing folder, or to create a new one.
- **Fasta File**: The reference genome is the FASTA or multi-FASTA file containing the sequences to be submitted. This tool is designed to submit complete eukaryotic genomes or chromosomes. If you are submitting a single complete chromosome it must be in a single fasta entry, however, if you are submitting a complete genome, you must have a single entry for each chromosome. **Important note**: if this is a complete genome or chromosome submission, remove all the 'Ns' present in the fasta file.
- **Genome annotation**:The genome annotation refers to the .gff file containing the gene coordinates for each annotated gene. This file must be named according to the dasta entry to which it corresponds.
- **Feature ID**: The feature ID of annotation refers to the flag on the ninth column of the gff file, which contains the name of the sequence, displayed as SeqName in Blast2GO.

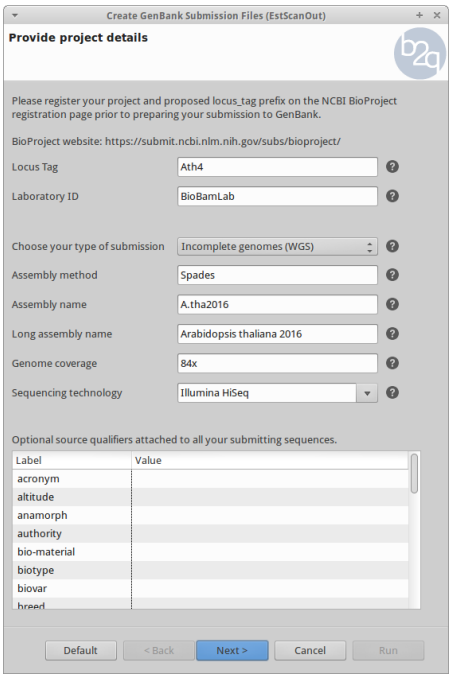


Figure 4: Project Details page

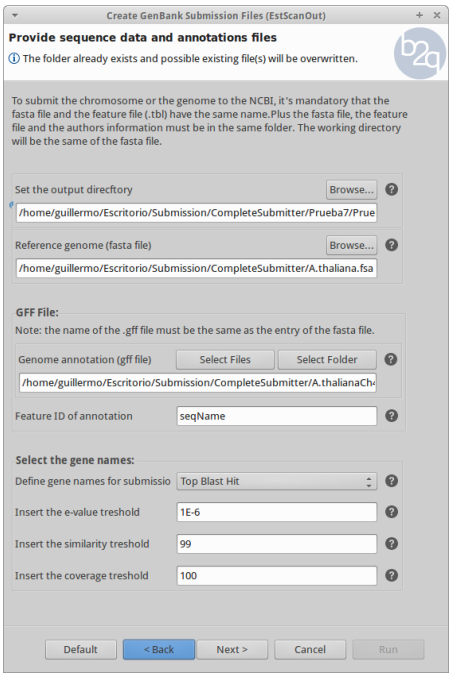


Figure 5: Sequence Data and Annotation Files page

- **Gene names**: Here you can choose how to assign the names for your annotated sequences, the options are: "hypothetical protein", the "SeqName" assigned in the Blast2GO project or assign the name of the "Top BLAST Hit". If this last option is selected, you can set the threshold for:
  - **E-value**:The minimum E-value obtained in the BLAST between the top BLAST hit and your query (default value is 1E-6).
  - **Similarity**:The minimum percent similarity between the two sequences (0-100).
  - **Coverage**:The minimum percent coverage between the two sequences (0-100).

  Important note: If the threshold is not reached, the name of the gene will be "hypothetical protein".

Note: all manual gene names annotations has higher priority.

4.3 Page 3: Author's and Affiliation data

- **Contact data**: This page allows provison of the contact details for the submitting person. This information will not be publically visible, and only can be used by the NCBI staff for validation.
- **Institution data**: Information about the institution where the sequencing was performed is provided here.
- **Title of the manuscript**: This title is provisional and can be modified at any time via email request to the NCBI.
- **Release date**: This is the date when your submited and validated data will be accesible in the NCBI database. If the release date is The same day or before the submission, it will be automaticaly available once the data is validated.
- **Names and Initials**: Insert the names and initials of the individuals who must receive scientific credit for the generation of the sequences and annotations in this submission. If the authors are part of a consortium, it is not necessary that they appear as individual authors, as they are represented in the 'Consortium' option.

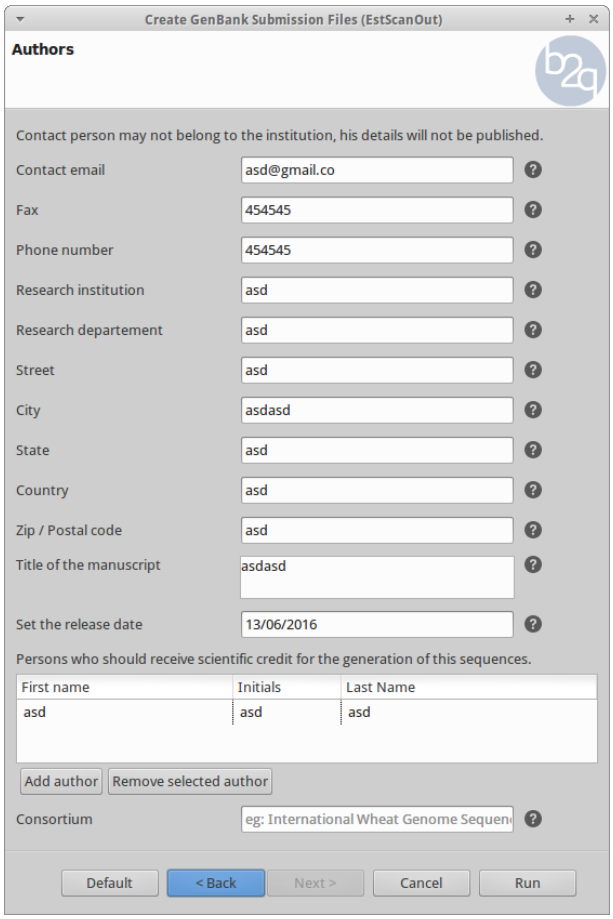


Figure 6: Author's and Affiliation page

5 Results Files

Once the input data has been analysed and processed via the tbl2asn tool, several result files are created.

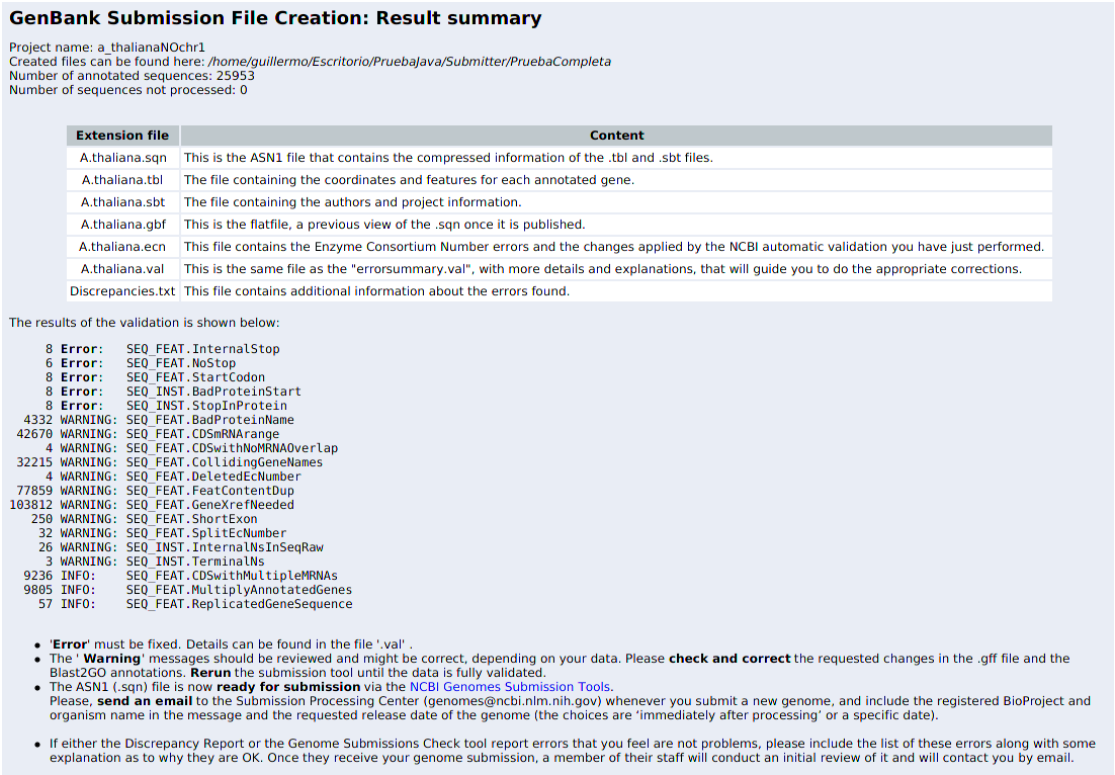
- **.sqn**:This is the ASN1 file containing the compressed information of the .tbl and .sbt files.
- **.tbl**: The file containing the coordinates and the features for each annotated gene.
- **.sbt**: The file containing the authors and project information.
- **.gbf**: This is the GeneBank flatfile, a previous view of the .sqn once it is published.
- **.ecn**: This file contains the Enzyme Consortium Number errors and the changes applied by the previous NCBI automatic validation you have just performed.
- **.val**: This is the same file as the "errorsummary.val", with more details and explanations, that will guide you to make the appropriate corrections.
- **.txt**: This file contains additional information about the errors found.

A results page provides a a summary of the different types of errors and warnings. Errors must be corrected and the warnings should be reviewed (they may be correct, depending on your data).

Modifications can be made by editing the gff or the annotation in the Blast2GO project. Once errors are corrected the tool can be reruned until an error-free validattion is acheived.

Once the submission files have no errors, the ASN1 (.sqn) file is ready for submission via the NCBI Genomes Submission Tools (`www.ncbi.nlm.nih.gov/projects/GenomeSubmit/genome_submit.cgi`).

Whenever you submit a new genome, it is necessary to send an email to the Submission Processing Center (genomes@ncbi.nlm.nih.gov) specifying the registered BioProject and organism name in the message as well as the requested release date of the genome.



GenBank Submission File Creation: Result summary

Project name: a_thalianaNOchr1
Created files can be found here: */home/guillermo/Escritorio/PruebaJava/Submitter/PruebaCompleta*
Number of annotated sequences: 25953
Number of sequences not processed: 0

Extension file	Content
A.thaliana.sqn	This is the ASN1 file that contains the compressed information of the .tbl and .sbt files.
A.thaliana.tbl	The file containing the coordinates and features for each annotated gene.
A.thaliana.sbt	The file containing the authors and project information.
A.thaliana.gbf	This is the flatfile, a previous view of the .sqn once it is published.
A.thaliana.ecn	This file contains the Enzyme Consortium Number errors and the changes applied by the NCBI automatic validation you have just performed.
A.thaliana.val	This is the same file as the "errorsummary.val", with more details and explanations, that will guide you to do the appropriate corrections.
Discrepancies.txt	This file contains additional information about the errors found.

The results of the validation is shown below:

```
     8 Error:   SEQ_FEAT.InternalStop
     6 Error:   SEQ_FEAT.NoStop
     8 Error:   SEQ_FEAT.StartCodon
     8 Error:   SEQ_INST.BadProteinStart
     8 Error:   SEQ_INST.StopInProtein
  4332 WARNING: SEQ_FEAT.BadProteinName
 42670 WARNING: SEQ_FEAT.CDSmRNArange
     4 WARNING: SEQ_FEAT.CDSwithNoMRNAOverlap
 32215 WARNING: SEQ_FEAT.CollidingGeneNames
     4 WARNING: SEQ_FEAT.DeletedEcNumber
 77859 WARNING: SEQ_FEAT.FeatContentDup
103812 WARNING: SEQ_FEAT.GeneXrefNeeded
   250 WARNING: SEQ_FEAT.ShortExon
    32 WARNING: SEQ_FEAT.SplitEcNumber
    26 WARNING: SEQ_INST.InternalNsInSeqRaw
     3 WARNING: SEQ_INST.TerminalNs
  9236 INFO:    SEQ_FEAT.CDSwithMultipleMRNAs
  9805 INFO:    SEQ_FEAT.MultiplyAnnotatedGenes
    57 INFO:    SEQ_FEAT.ReplicatedGeneSequence
```

- '**Error**' must be fixed. Details can be found in the file '.val' .
- The ' **Warning**' messages should be reviewed and might be correct, depending on your data. Please **check and correct** the requested changes in the .gff file and the Blast2GO annotations. **Rerun** the submission tool until the data is fully validated.
- The ASN1 (.sqn) file is now **ready for submission** via the NCBI Genomes Submission Tools.
  Please, **send an email** to the Submission Processing Center (genomes@ncbi.nlm.nih.gov) whenever you submit a new genome, and include the registered BioProject and organism name in the message and the requested release date of the genome (the choices are 'immediately after processing' or a specific date).
- If either the Discrepancy Report or the Genome Submissions Check tool report errors that you feel are not problems, please include the list of these errors along with some explanation as to why they are OK. Once they receive your genome submission, a member of their staff will conduct an initial review of it and will contact you by email.

Figure 7: Results summary

5.1 Most common 'ERRORS' and 'WARNINGS'

- **ERROR(s): InternalStop + StartCodon + BadProteinStart + StopInProtein**: These error codes usually appear grouped, and they refer to the same sequence. This may be due to an error in the gff that has shifted its reading frame, you can correct that by changing the frame on the .gff.
- **StartCodon**: An illegal start codon was used. Some possible explanations are: (1) the wrong genetic code may have been selected; (2) the wrong reading frame may be in use; or (3) the coding region may be incomplete at the 5' end, in which case a partial location should be indicated. This can be fixed in the .gff file, or by selecting the correct code in the 'source qualifiers' on the first wizard page.
- **InternalStop**: Internal stop codons are found in the protein sequence. Some possible explanations are: (1) the wrong genetic code may have been selected; (2) the wrong reading frame may be in use; (3) the coding region may be incomplete at the 5' end, in which case a partial location should be indicated; or (4) the CdRegion feature location is incorrect. This can be fixed in the .gff file by modifying the start of the sequence or selecting the correct code on the 'source qualifiers' on the first wizard page.
- **WARNING CDSmRNArange**: This error alerts you that two or more 'CDS' features are under the same 'mRNA' feature, but there are not colliding. If you are working with prokaryotes, this is a feature you must fix it in the .gff file, but if working with eukaryotes, it's a normal feature, as eukaryotic genes contain introns.
- **WARNING CDSwithNoMRNAOverlap**: This warning alerts you that a 'CDS' feature out of the 'mRNA' bounds, and should be fixed in the .gff file by extending the mRNA range.
- **WARNING BadProteinName**: The name assigned to this protein is not adequate. Remember that the protein name should not contain the names 'hypothetical' or 'partial', and must follow the Uni-Prot protein product names. Modify it in the Blast2GO project or directly in the .sqn file.
- **WARNING CollidingGeneNames**: Two gene features should not have the same name, this can be fixed in the Blast2GO project.
- **WARNING MissingMRNAproduct**: The mRNA feature indicates to a cDNA product that is not contained in the record. This must be fixed on the .gff file.
- **WARNING DuplicateInterval**: The location has identical adjacent intervals, e.g., a duplicate exon reference. This can be fixed eliminating the duplicated 'exon' or 'CDS' from the .gff file.
- **WARNING mRNAgeneRange**: An mRNA is overlapped by a gene feature, but is not completely contained by it. This can be corrected in the .gff by extending the range of the 'mRNA'.
- **NoOrgFound**: This entry does not specify the organism that was the source of the sequence. Please enter a name for the organism on the first page of the wizard, in the 'Optional source qualifiers'.

For more information about these errors, please refer to `http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/errmsg/valid.msg`.