RAW DATA

↓

CLC Genomics Workbench:
Converts raw transcriptome
reads to contigs.

.xlsx
Excel file, contains all the
expression values for each
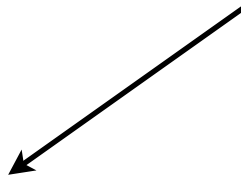contig, within every
transcriptome you sequenced.

.fasta
Fasta file, contains the
DNA sequence for each
contig in your data set.

↓

Blast2go:
Uses the BLAST algorithm for gene
homology search against a non-redundant
database hosted by NCBI.

Load the .fasta file into blast2go (in red on the
next page), and run the BLAST, GO-Mapping,
Annotations, InterProScan, and KEGG steps.

Congratulations! You got this far.
Now it is time to create a Database!

DATABASE CREATION:
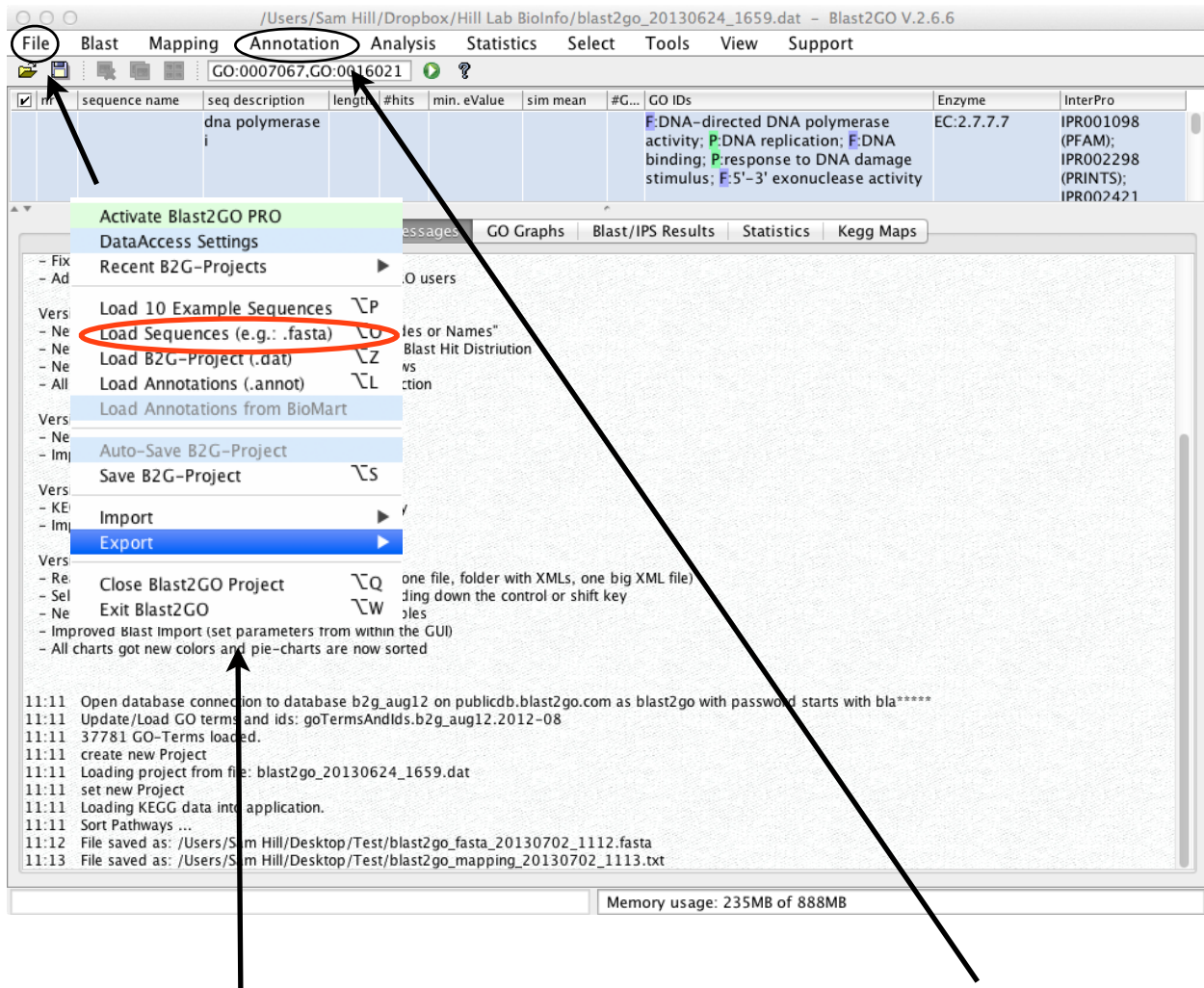
There are six necessary files for database creation after you have finished BLASTing, Mapping, and Annotating. Four of the six are under **File > Export**, one is online at the gene ontology website, and one is under the **Annotation**. (Suggested that you keep all of these files in one folder for easy access)
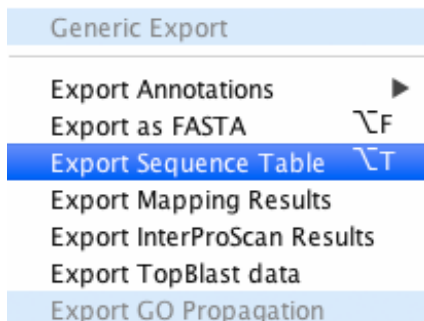
For the initial creation of your database, you will need a Sequence Table to fill your database with contig IDs, most significant blasts, base pair counts, and evalues for the contigs in your data set. To get a Sequence Table go to the **Export Sequence Table** option under **File > Export.** Name your file and save it to where you want to keep it. (Suggested naming: *transcriptome_info_*table.txt)

Now you will want to add a list of associated GO Terms and a list of their associated definitions to your current database. There are two files you will need to run the GOStat/Defs function, a GOStat Mapping result and an ontology from:
http://www.geneontology.org/GO.downloads.ontology.shtml
The ontology you will need from gene ontology is the RDF-XML filtered ontology under the XML format section.

Download the gzip file and unzip it (double click the downloaded file). You should get a file called "go_daily-termdb.rdf-xml".

For the GOStat Mapping result go to the **Export Mapping Results** under **File > Export**

When you are prompted for where to save and for the file type, choose where you want to save and name your file. Then under the File Format drop down menu select ***.txt - TSV - GOstat Format**. (Suggested naming: *transcriptome_info_*GOStat.txt)

Generic Export

Export Annotations ▶
Export as FASTA ⌥F
Export Sequence Table ⌥T
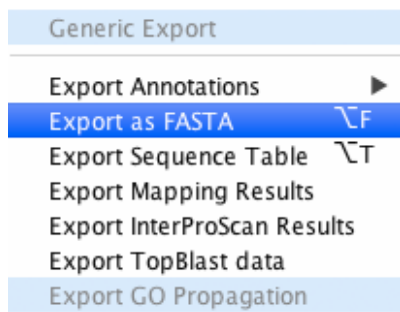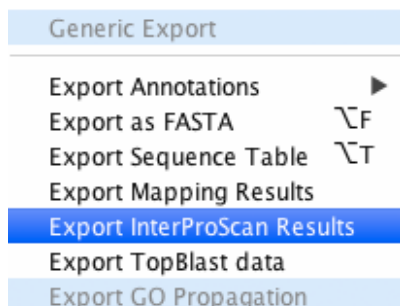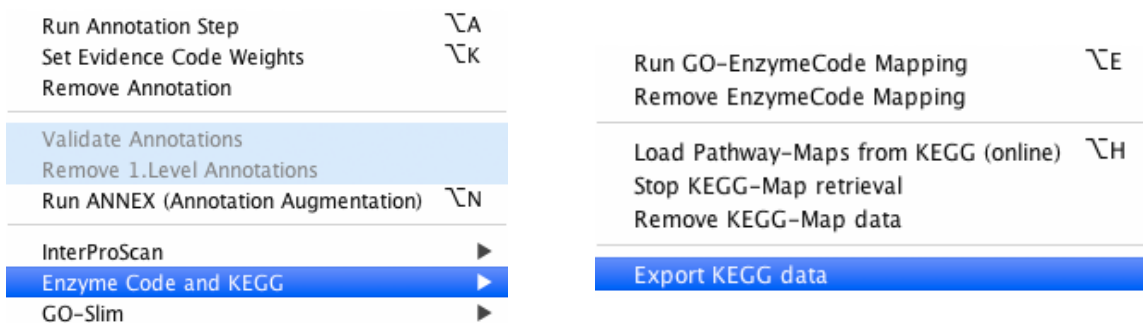Export Mapping Results
Export InterProScan Results
Export TopBlast data
Export GO Propagation

Next you will need a Fasta file, so go to **Export as FASTA** under **File > Export**. Name your file and save it to where you want to keep it. This .fasta file needs to be changed to a .txt file. You can do this by navigating to where you are keeping the file and renaming the file with a .txt extension. (Suggested naming: *transcriptome_info_*FASTA.txt)

To add Kegg information (pathway, pathway ID, enzyme, enzyme ID) go to **Enzyme Code and KEGG** under **Annotation**, then select **Export KEGG data**. Name your file and save it to where you want to keep it. (Suggested naming: *transcriptome_info_*KEGG.txt)

Run Annotation Step ⌥A
Set Evidence Code Weights ⌥K
Remove Annotation

Validate Annotations
Remove 1.Level Annotations
Run ANNEX (Annotation Augmentation) ⌥N

InterProScan ▶
Enzyme Code and KEGG ▶
GO-Slim ▶

Run GO-EnzymeCode Mapping ⌥E
Remove EnzymeCode Mapping

Load Pathway-Maps from KEGG (online) ⌥H
Stop KEGG-Map retrieval
Remove KEGG-Map data

Export KEGG data

Generic Export

Export Annotations ▶
Export as FASTA ⌥F
Export Sequence Table ⌥T
Export Mapping Results
Export InterProScan Results
Export TopBlast data
Export GO Propagation

Finally, for InterProScan information (we only take PFAM hits from this) go to **Export InterProScan Results** under **File > Export**. Name your file and save it to where you want to keep it. (Suggested naming: *transcriptome_info_*IPS.txt)

Now you have the files you need to make a database!
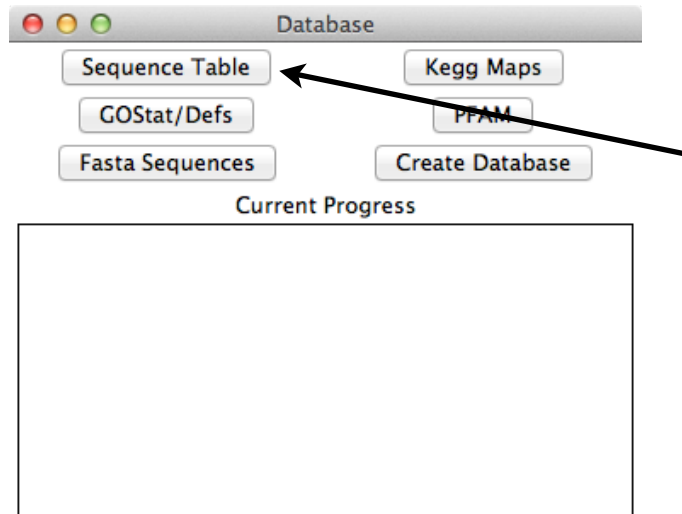
Here is the Database creation program. The first thing you will need to do is load in your sequence table. (This step presets the database with all of the basic information you need)

After you have loaded in your sequence table, you can load the rest of the files in any order you want. Once you are done, hit Create Database and save your new database.
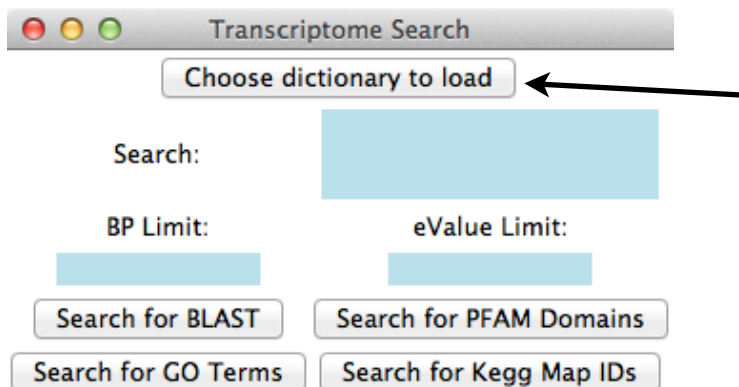
Now that you have a database, you can search through it for specific information using the Transcriptome Search program. First you will have to load the dictionary you would like to use. After that you can search the database in 4 ways; by the blast output, the GO Terms (accession number and definition), the PFAM Domains, and KEGG Maps ID's. You can also refine the search by having a minimum base pair count and e-Value. Once you have decided what to search, added limits on the bp and eValue, and decided how to search, click the appropriate button, and save your output.