Overview:

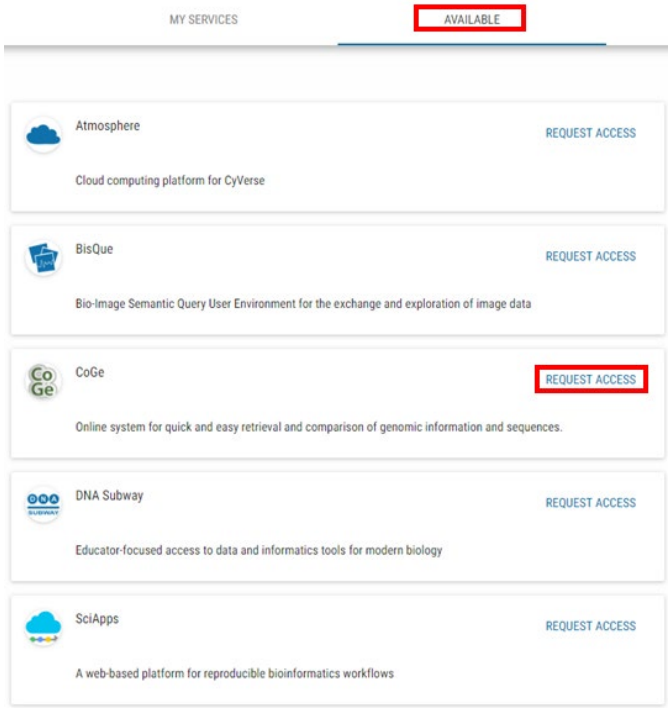
After spending a lot of time and energy assembling a genome you really want to visualize it, share it, compare it to other related species, and search for, and extract information for, your favorite gene. This tutorial is designed to walk you through some of the immediate analyses and "sanity tests" that you would perform after assembling and annotating your genome. This is not meant to be an extensive review of all the analyses CoGe can perform. For more info on CoGe, please see the (extensive) wiki at: https://genomevolution.org/wiki/index.php/New_to_CoGe

In this portion of the workshop we are going to show you how to 1) connect your Datastore to CoGe, 2) load your genome(s) into CoGe – along with pertinent genome annotation files, 3) search through your genome(s) using CoGeBLAST, 4) visualize synteny and conservation with Gevo and SynMap, 5) visualize your genome in EPIC-CoGe, and then 6) share your genomes, or your analyses with collaborators (or release it to the public).

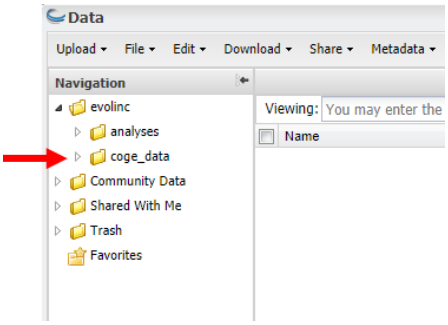
1. Connecting your CyVerse Datastore to CoGe:

If you are brand new to CyVerse's Datastore or CoGe then you likely will need to "connect" the two in order to share data between the two services. This is as simple as clicking a button!

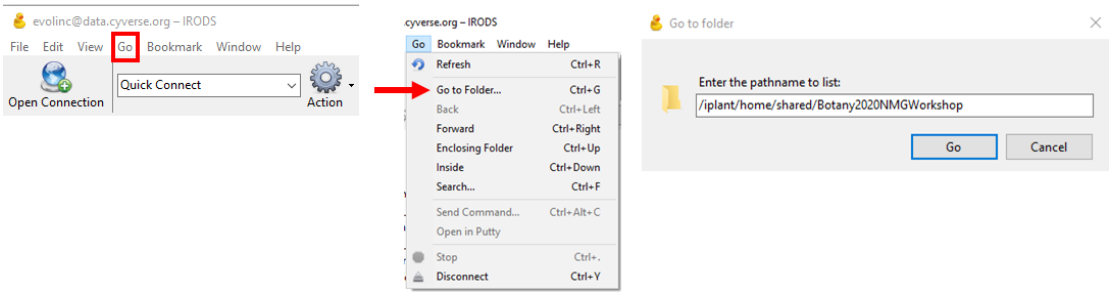
1. Go to the url: https://user.cyverse.org/services/mine
2. Click on "Available Services" and click on "REQUEST ACCESS" under CoGe.



3. Now you have given CoGe permission to access your Datastore. For security purposes, CoGe can only access one folder in your Datastore: your “coge_data” folder.
4. Any genomic data you want to load into CoGe will need to be placed in this folder. Once you are finished loading the genome into CoGe (the task we will do next) you can move the file back out of the coge_data directory to its original location.



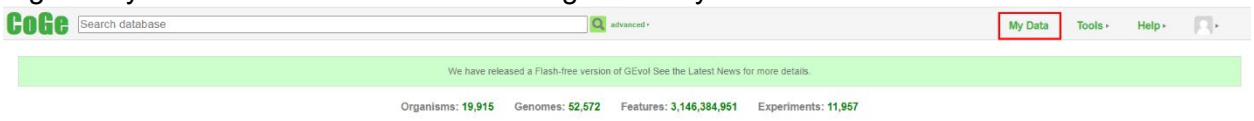
5. **IF** you are wanting to load a shared with you by another user, or if you want to keep your file in a specific location, then you will have to do a few additional steps, as the Datastore doesn’t allow you to copy a file directly through the GUI. There are multiple workarounds here such as iCommands or downloading and re-uploading the file, but the one we will use today is CyberDuck.
    a. There is extensive information on how to get CyberDuck set up with CyVerse here: https://learning.cyverse.org/projects/data_store_guide/en/latest/step1.html so for the sake of this tutorial we will simply navigate to the Community workshop folder: /iplant/home/shared/Botany2020NMGWorkshop using CyberDuck’s “Go To Folder” command and then copying a genome file in which we are interested (control/command + C). Navigate back to your coge_data folder (left arrow next to the navigation bar) and paste the file into your coge_data folder (control/command + V).



2. Loading a genome and genome annotation file into CoGe:

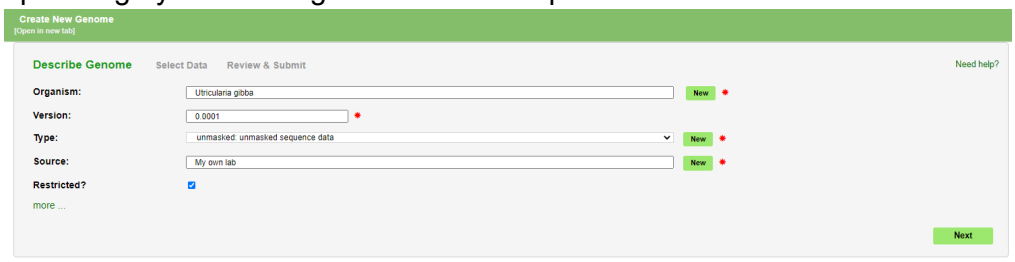
Now that you have a genome and genome annotation file loaded into your “coge_data” directory, let’s get it loaded into CoGe (https://genomevolution.org/coge/)

1. Sign into your CoGe account and then navigate to “My Data”:

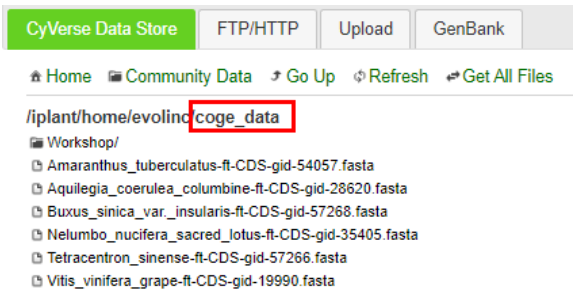


2. Click on “New” and then “New genome”. A new pane will popup asking you to provide information about the species with which your genome is associated.
    a. As you start to type “Utricularia” you will notice that it may provide you with prepopulated information based on previously added, similarly named species. If your species is completely new to CoGe you will have to add it by clicking “New” (follow the prompts).

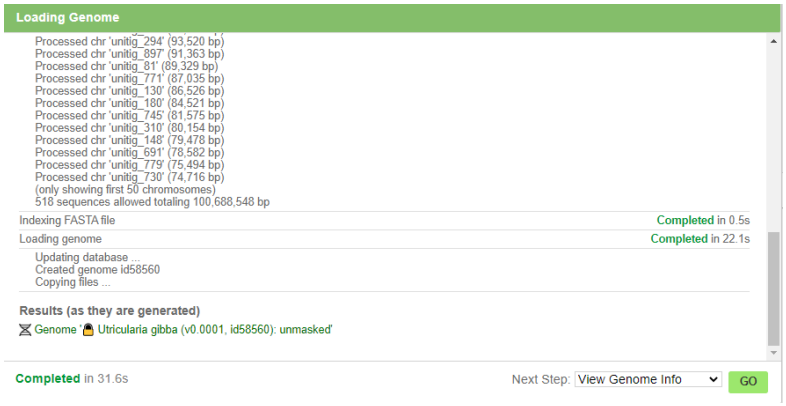
b. Add a version number associated with your genome assembly – this helps you coordinate version control.
c. Source refers to where you obtained the genome – this could be NCBI, your own lab, or a collaborator.
d. You can add more descriptive information by expanding the fields with the “more” button or you can proceed to uploading the genome by clicking “next”.
e. **NOTE** that you have the option to make your genome immediately public after uploading by deselecting the “restricted” option.



f. Now you can specify the upload source. The first option you should see is the “CyVerse” Data Store and it should point you to your coge_data folder.
g. Navigate to the genome you want to upload and select it. NOTE – these genomes can be compressed (.gz) but cannot be tar.gz.



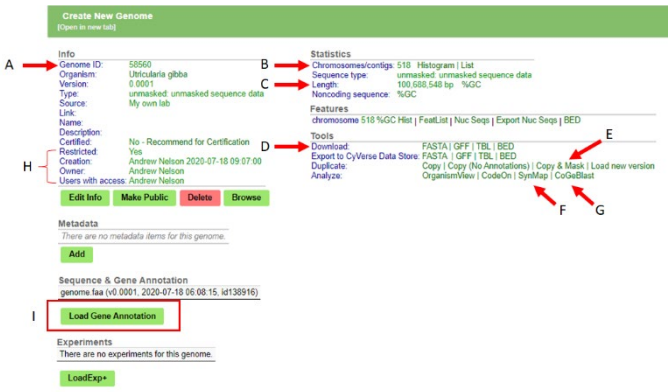
h. Click next – you will have the option to review all of the information that you entered before clicking “start” to upload the genome.
i. A popup window will appear that provides information on the uploading process. On good days when the wind is out of the east and the early birds have gotten all of the worms, this can be finished in < 1 minute. This all depends on site traffic and size of your genome. Eventually the progress report should say “completed”. You will be prompted to visit the “Genome Info” page. **Click “GO”**



j. The “Genome Info” tab has a lot of information associated with it, including genome ID (A), details about the assembly itself (N50 and average length of scaffolds can be seen by clicking on Histogram; B), Length, as well as GC content (C). You can also download specific aspects of your genome (FASTA, BED file, etc) to your computer or to your “coge_data” folder in the Data Store.

k. Copy and Mask (**E**) will duplicate your genome and then hardmask (NCBI WindowMasker) it to mask repetitive elements (useful when performing genome alignments).

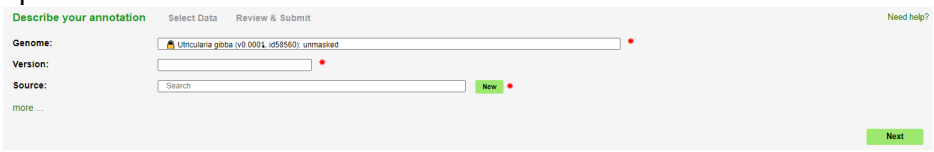
l. You can also send your genome to SynMap, or use it in a query in CoGeBLAST (F, G).



m. This page also lets you see who has access to your genome (H).

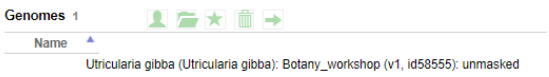
n. One thing is missing however – we don't have an annotation file uploaded! Load a gene annotation by clicking on the corresponding button (I).

o. This will start up a very similar process to the "load genome" except now some of the information is already loaded for you. Add in the additional information you want to add, navigate to the location where the annotation file is stored (Note that .gz files are also accepted here), and then click next to review and start the upload.



p. This process will take a little longer than uploading the genome, so let's let it run and look at a different genome that already is annotated. Close the progress window (the job will still be running in the background).

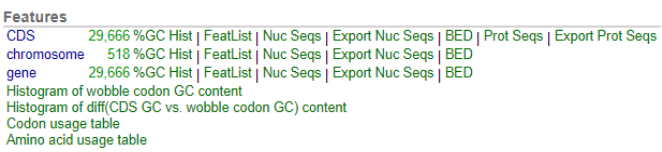
q. Search for the Utricularia genome that was already uploaded (58555) in the search bar up at the top. You can restrict your search to just genomes by clicking on the advanced button and selecting "genome" under Type. Click on the search button and you should get the following results:



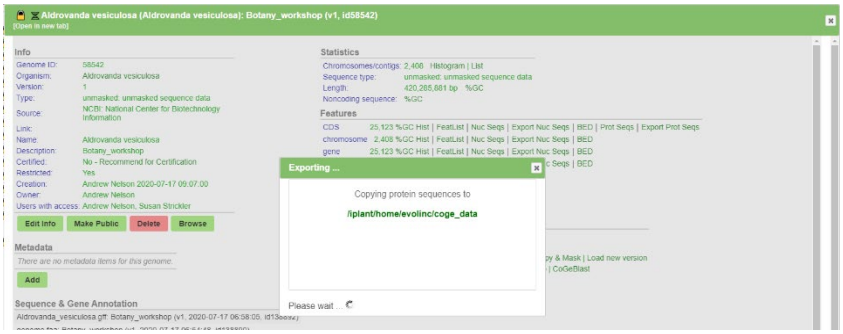
r. Double click on the search result and it should load up a new Genome info window. Under "Sequence and Gene Annotation" you should see that this genome has both .fasta and .gff files associated with it.

s. Let's verify that the annotation file makes sense by clicking on "Click for Features".

t. You should now see information regarding the genes associated with that annotation file. You can view or download the protein or nucleotide sequences associated with your genome annotation here:



u. You can download sequence data for the genes you have annotated by either directly downloading them or exporting them to your “coge_data” folder in the DataStore. Let’s export the protein sequences for Aldrovanda, Drosera, Genlisea, and Utricularia back to the DataStore so that we can work on them later. Click on the “Export Prot Seqs”. When you do this you should see a pop-up window informing you that the protein sequences are being exported to your coge_data folder.

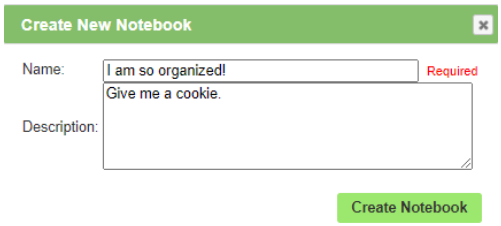


v. Let’s close out of this Genome info page and look at some of the other data management tools within CoGe.

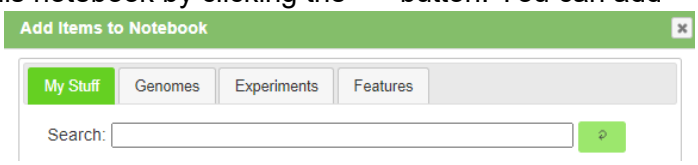
3. Notebooks, groups, and privately sharing data in CoGe.

Genome projects are often collaborative endeavors that require many the work of many pairs of hands prior to the genomes in question being made public. CoGe facilitates data sharing with collaborative groups – both public and private. Here we will discuss the ways in which you can organize and share your data in CoGe.

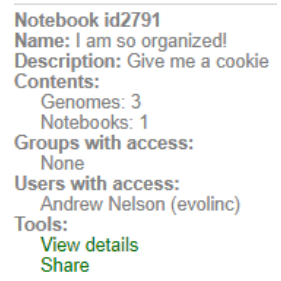
1. **Notebooks**: Notebooks are ways in which you can organize affiliated sets of genomes or experiments. These are useful when your data folder starts to get crowded with files, but are even more useful when you want to easily share a project with a collaborator.
   a. To create a notebook, click on “New” and then select “Notebook” A small popup will appear that will prompt you for a name (and an optional description).



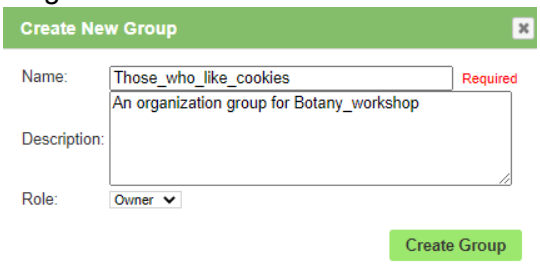
   b. Create the notebook, and then look for it under the “Notebooks” link to the left. Double click on the notebook name to open a new pane describing the notebook and its contents. You will notice that there is nothing in the notebook and that it is currently private (the lock icon tells you that).
   c. Let’s add genomes to this notebook by clicking the “+” button. You can add genomes/experiments from your own data folder, or public genomes or experiments.



d. Add the genome that you assembled and uploaded to CoGe, and then search for and add additional genomes. Here we will add a close relative of Utricularia, Sesamum indicum (id26082) as well as Arabidopsis thaliana (id1).
e. Once you have populated a notebook, you can either immediately start using it in analyses (we will discuss this in a moment) or you can share with other users/groups by clicking on the "share" link on the right of the window.
f. Click on share and then search for user by typing their name into the search bar. You can also share with a group – for example, a collaborative group working on a project – so that you don't have to give a bunch of people access every single time you update your genome.



2. **User groups:** Giving collaborators access to files can be annoying if the genomes in question are constantly being updated. Notebooks take care of part of this annoyance as anything placed in the notebook is instantly shared with whomever has access to the notebook. This annoyance can also be alleviated by creating user groups that have access to notebooks or files, and then simply adding a new user to the user group.

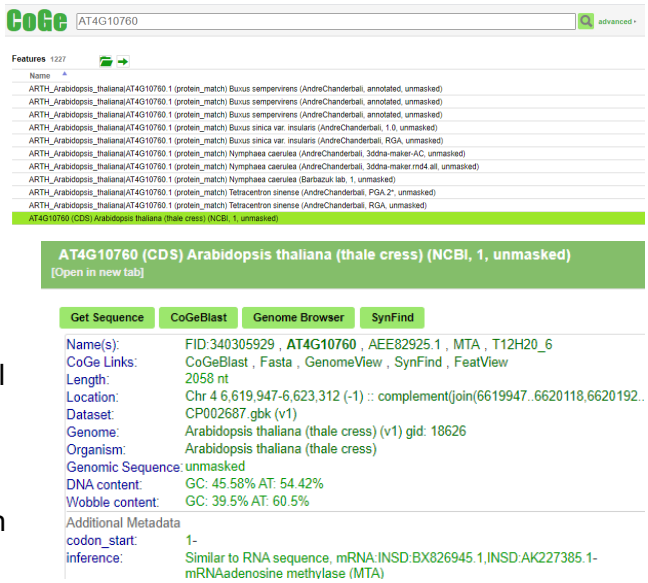


   a. To create a new user group, click on the "New" button, followed by "user group".
   b. Once it is created, then you can add users by 1) double clicking on the group (or editing membership to the right) and then entering their handle or their full name into the search bar.
   c. Then you can give that group access to one of your notebooks by clicking on the notebook and clicking on "share" to the right.
   d. Now when you double click on the notebook and look at access you see something different than before:
   e. If you add someone to the user group they will now automatically have access to the notebook you created.

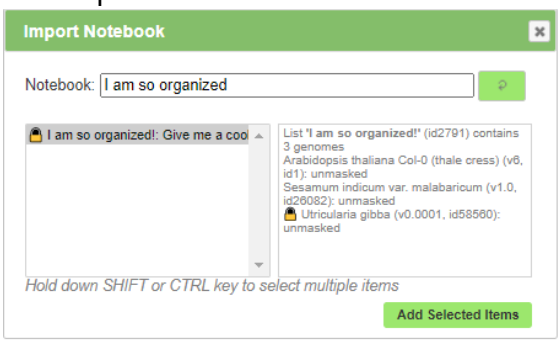
4. Searching through your genome with CoGeBLAST.

Organizing and sharing your genomes is just one of many features that CoGe offers. There a suite of analysis tools that allow you to search through your genome and compare it to others. Let's start by searching for a specific gene in your genome using CoGeBLAST.

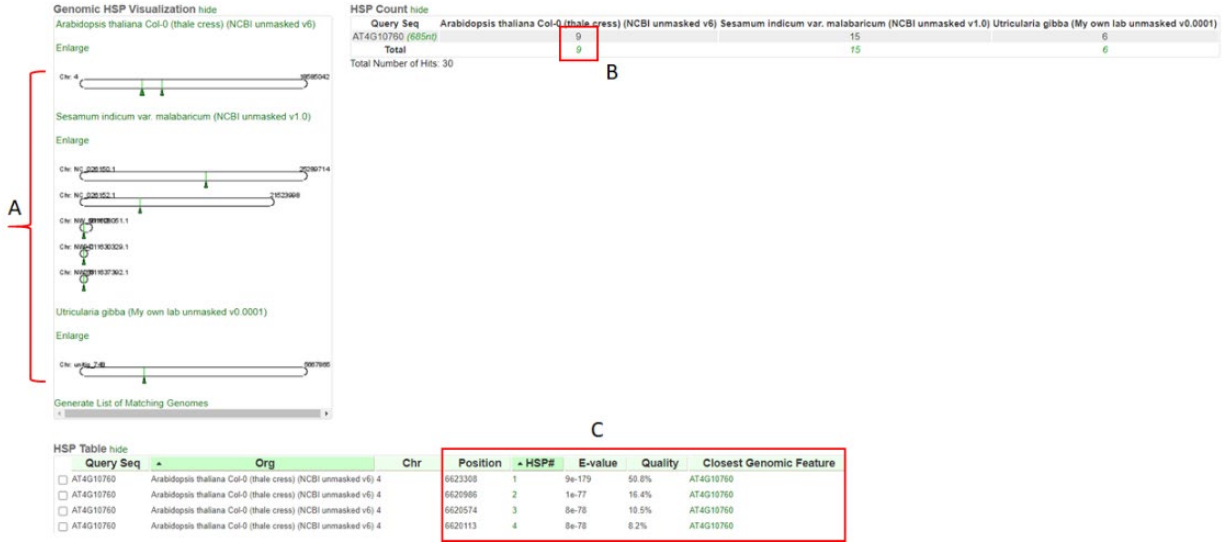
1. First we need to grab a gene to serve as query. Let's search for the Arabidopsis gene "AT4G10760" in the search bar at the top of the CoGe page.
2. Double click on the first search result that is associated with Arabidopsis thaliana (thale cress). This will load the information page for this gene. AT4G10760 is a methyltransferase that adds methyl groups to RNA molecules. This gene is highly conserved (origins likely in the ancestor to eukarya) and thus should be present in each of our newly assembled genomes (let's call it a quality test).



3. Next to the "CoGe Links" are several informative links or analyses that can be performed on this gene. Click on "Fasta".
4. The Fasta viewer window allows you to now download the nucleotide/amino acid sequence for your gene. It also allows you to send your sequences out to external analysis services (we will save that for later). For now, let's just click on "Protein Sequence". This will convert the nucleotide sequence to amino acid based on the first frame. Now click "CoGeBlast". A new tab will open.
5. CoGeBlast utilizes a number of popular BLAST algorithms to search for hits in any of the public (or your private) genomes found within CoGe. First let's make sure the query sequence is in amino acid format (sometimes cookies screw things up). If it is the nucleotide sequence, delete and replace with the amino acid sequence from the previous tab.
6. Now we need to load in our favorite genomes. This can happen one of two ways – either by searching for individual genomes and loading as many as you want – or by importing one of the notebooks that we created earlier. Let's import a notebook:
7. By clicking on the notebook you can add all of the genomes found within that notebook – a real time saver! Let's add Brassica rapa (id24668) for fun (you can combine notebooks with other genomes).
8. Under BLAST parameters, switch to tblastn and E-value of 1e-10 (but note the other options available for future reference). Then click "Run CoGe BLAST" at the bottom.

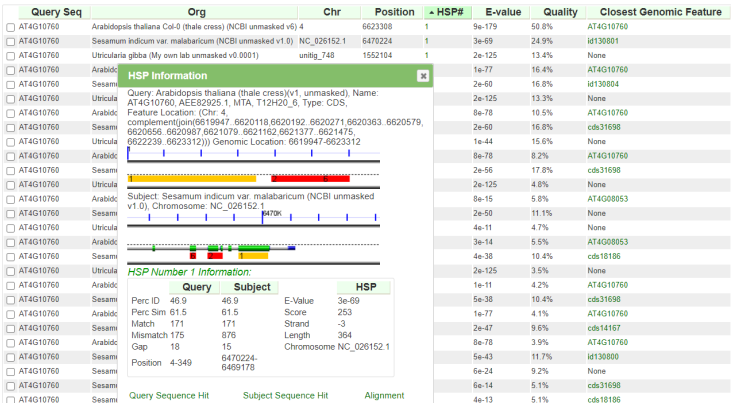


9. If the winds are favorable and out of the southwest, this simple BLAST will be finished in seconds. But if this is the first time that your genomes have been indexed, you are using large, repetitive genomes, many query sequences, or if there is a high user load, it might take a bit longer. You will see the progress of your analysis in the pop-up progress window. Once it is finished, click “OK”.
10. You will now see your BLAST results. These include the locations of the hits along the chromosomes/scaffolds of your subject genomes (A), the number of hits per species (B),

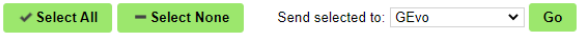


and a table of hits (C). The table of hits holds even more information so let’s explore it further.

11. The HSP table is sortable, so let’s sort by HSP ascending (so that the top HSPs are at the top of the table). If you click on the link for the 1st HSP for Arabidopsis thaliana, a pop up window will appear that tell you where the query matched to the genome. You can click anywhere within the gene model that appears to open a new Genome Browser tab.

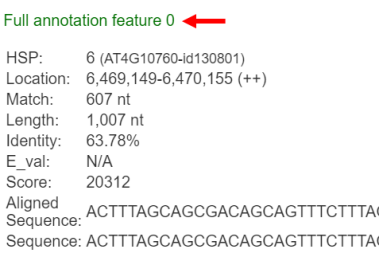


12. You can also click on the closest genome feature (the gene corresponding to what you see in the HSP information tab) to open the “Feature Information” pane, which provides additional information about what gene your BLAST return corresponds to. Notice that you can immediately send this gene to CoGeBLAST (which is nice for a quick reciprocal BLAST search).
13. Let’s click on the square buttons next to the first three #1 HSPs corresponding to Arabidopsis, Sesamum, and Utricularia. Scroll down to the bottom of the table and select Send selected to: “Gevo” and then GO.



14. Gevo allows us to compare multiple genomic regions (brought in from CoGeBLAST or elsewhere) to get a close up view of collinearity within gene order (synteny) or look for duplications between close relatives.
15. The initial view for Gevo asks you to order your sequences and add a specific amount of sequence on either end (if you want to look at synteny outside of the immediate area around your gene of interest). If you want to add, say 50kb to all sequences, scroll to the bottom and add “50000” to the “Pad CoGe Sequences with additional sequence”.

16. Click Run Gevo. Gevo uses LASTz by default to line up the 100kb regions (50 + 50kb) around each of your BLAST results. In the visualized version of your results it highlights regions that are significantly similar to one another with bars above the gene model.
17. You can quickly tell from Gevo that there are several similar regions between Arabidopsis and Brassica, but really only our gene of interest is similar between all four species. Interestingly, in Sesamum you see two regions of similarity to our gene in Arabidopsis. If you click on the middle colored bar in Arabidopsis, connector lines will appear that show there looks to have been a local duplication of our gene in Sesamum – it just isn’t annotated as a gene. Sequencing error or real duplication event? We can get a better sense by clicking on each of the colored boxes and then selecting the feature annotation name.



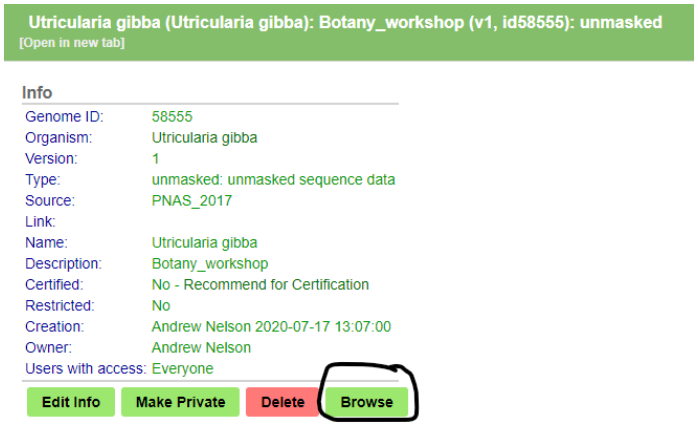
18. Alignment results can be found in the “Alignment reports underneath the Gevo viewer. Check out the “Results Visualization Options” for some really useful additional features. Also check out the Gevo wiki page for more info: https://genomevolution.org/wiki/index.php/GEvo

5. Loading experiments into CoGe

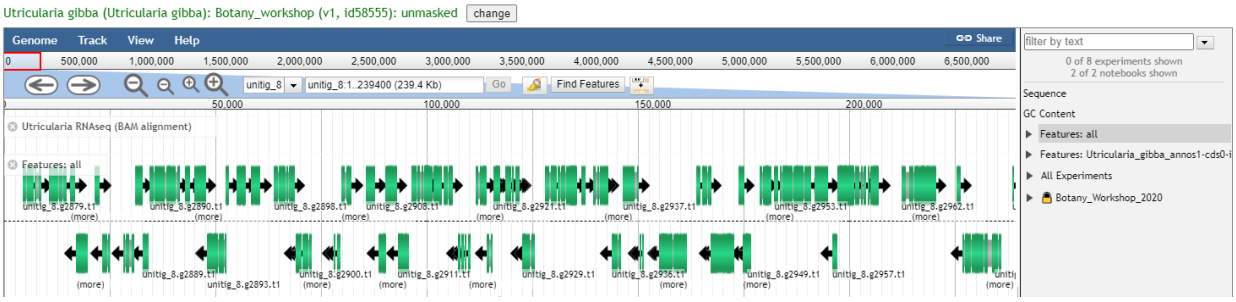
Often times you will generate data (RNA-seq, SNP, etc) that you need to sanity check by visualizing in a genome browser, or you simply want to get that good screenshot of your favorite gene being expressed under certain conditions and not in others. You can easily do that in CoGe using the “Load-Exp+” tool.

1. As with the genome data, it helps if you have your genome-associated files already stored in your “coge_data” folder in the DataStore. We have some pre-mapped RNA-seq data for you to play around with in the BotanyWorkshop community folder. Let’s use CyberDuck to copy some of those data into your coge_data folder.

2. With CyberDuck open and logged into your CyVerse account, navigate to /iplant/home/shared/Botany2020NMGWorkshop/Genome_assembly/sorted_bam using CyberDuck's "Go To Folder" command and then copy "Ugibba_leafR1.bam" and "Ugibba_stemR1.bam" (control/command + C) and then paste into your coge_data folder (control/command + C).
3. Go to CoGe (genomevolution.org) and (after logging in) navigate to "My Data", then click on "New" (sound familiar?) and select new experiment.
4. Click on the two files that you copied into your coge_data folder. Then give your experimental data a name and provide the appropriate metadata. Assign the genome (it helps if you have the gene ID to search for Utricularia since there are so many of them now).
5. Select the specific expression parameters you prefer (it is running cufflinks for those familiar).
6. Review your metadata and click "Start".
7. LoadExp + will go through and start the process of assessing how many reads mapped to each transcript and quantifying those read depths (again, using cufflinks). This is not definitive, but is useful in giving you a rough idea of how your genes are expressed.
8. Once you start the job you can close the window and proceed with your life.
9. In the search bar at the top, search for genome 58555 (a public version of Utricularia). Double click on the search result, then click "Browse":



10. This will load up EPIC-CoGe, CoGe's genome browser. You can now load in any experiments associated with this genome (either public or your own private experiments) and visualize the data alongside the gene models that you generated. An example is below:



11. You can also resize this window to make it larger or smaller, depending on your needs.
12. Let's get a qualitative perspective on how well our genome annotation is: we can test for regions where RNA-seq reads mapped but for which there are no annotated genes by holding the control/command button down while dragging one track onto another (you should see the "+" symbol appear when you try and do this).
13. This will create a new track that you can quickly search to identify novel transcribed regions that were not picked up by Maker.

14. There are many other things you can do with EPIC-CoGe – see the Wiki for more details!

Last but not least:

Synopsis: CoGe has a number of useful tools that we don't have time to talk about today.

For example, you can also perform large scale synteny analyses between your two (or three) favorite genomes. See https://genomevolution.org/wiki/index.php/SynMap for more details (SynMap is an amazing tool, I have given it way too short of a mention here!)

Be sure to check out the wiki https://genomevolution.org/wiki/index.php/New_to_CoGe and let the CoGe team (coge.genome@gmail.com) know if you have any questions or issues.