In this tutorial, we will be looking at some of the downstream analysis steps you can perform once you have identified one or more genes/proteins of interest from your annotated genomes.

1. **Step 1: Obtain the sequence(s) of your gene(s) or protein(s)**

There are a number of different ways to do this; in this example, I will first use a tool to extract the sequences from the Genbank file we generated using pharokka. This brings my sequences into a convenient FASTA format which makes my downstream analyses easier.

A screenshot of a computer

Description automatically generated

Figure 1. Extract sequences from Genbank tool in Galaxy.

You can find the “Extract sequences from GenBank” tool in Galaxy (Figure 1) by searching for it in the search bar on the left-hand side of your screen. You will now want to point this to the correct item in your history (for me, it is number 4 – the Genbank output from pharokka; note that I found this after clicking on the “folder” icon).

You will then want to specify what features the tool should extract. In today’s example, I will choose a protein sequence (Feature to extract: CDS, output type: amino acids) but note that you could choose genes or tRNAs (in those cases you would want to specify the output type as “nucleotide”).

Click “Run Tool”. Your output should look something like Figure 2 – with a number of FASTA formatted sequences, each with appropriate headers followed by the sequence information.

For this example I will focus on the **integrase** protein (phi-C31 type integrases are hugely useful for synthetic biology/biotechnology) – for your project, you will be focussing on your own proteins/genes of interest.

Find the sequence of your gene/protein of interest, select it, and copy it.

A screenshot of a computer

Description automatically generated

Figure 2. FASTA sequences extracted from GenBank file.

Now you have your sequence of interest – there are many potential ways you could analyse it (BLAST search; structural analysis e.g., using alphafold; etc.). What we will do first is to create a multiple sequence alignment with the other sequences from our genomes; and then use this to create a phylogenetic tree. This will allow us to analyse the evolution of our particular protein/gene of interest.

1. **Step 2: Create a FASTA file for each gene/protein of interest**

You will want to create a FASTA file that contains **just the sequences of your particular protein/gene**. (Currently we have a FASTA file on Galaxy that contains all the sequences from a particular phage.)

Copy and paste your gene/protein sequence into a **text editor** such as Notepad. You will then want to save this as a **.fasta** file (Figure 3).

In most Windows set-ups, you will be able to do this by clicking: File > Save As… and entering in the file name. Here I have chosen to name my file “integrases.fasta”.

**Important:** When saving this file as a .fasta file, you **must** change the drop-down “Save as type” from “Text documents” to “All Files”. If you do not do this, your file will be saved as a .txt file instead of a .fasta file!

Note also that I have saved this file in a specific folder with an informative name (BM432ProjectDemo). You may organise your files for your project in any way you choose, but I strongly recommend that you consider the project and data management tips outlined [here](https://sipbs-compbiol.github.io/BM432/notebooks/03-02-project_management.html), and that you use sensible file names (so that you will easily be able to identify the contents of any particular file or folder at a glance). (You may also consider including a README text file in each folder, which describes the contents of the folder: i.e., a brief description of the contents of each file, how the data were obtained and processed, etc.)

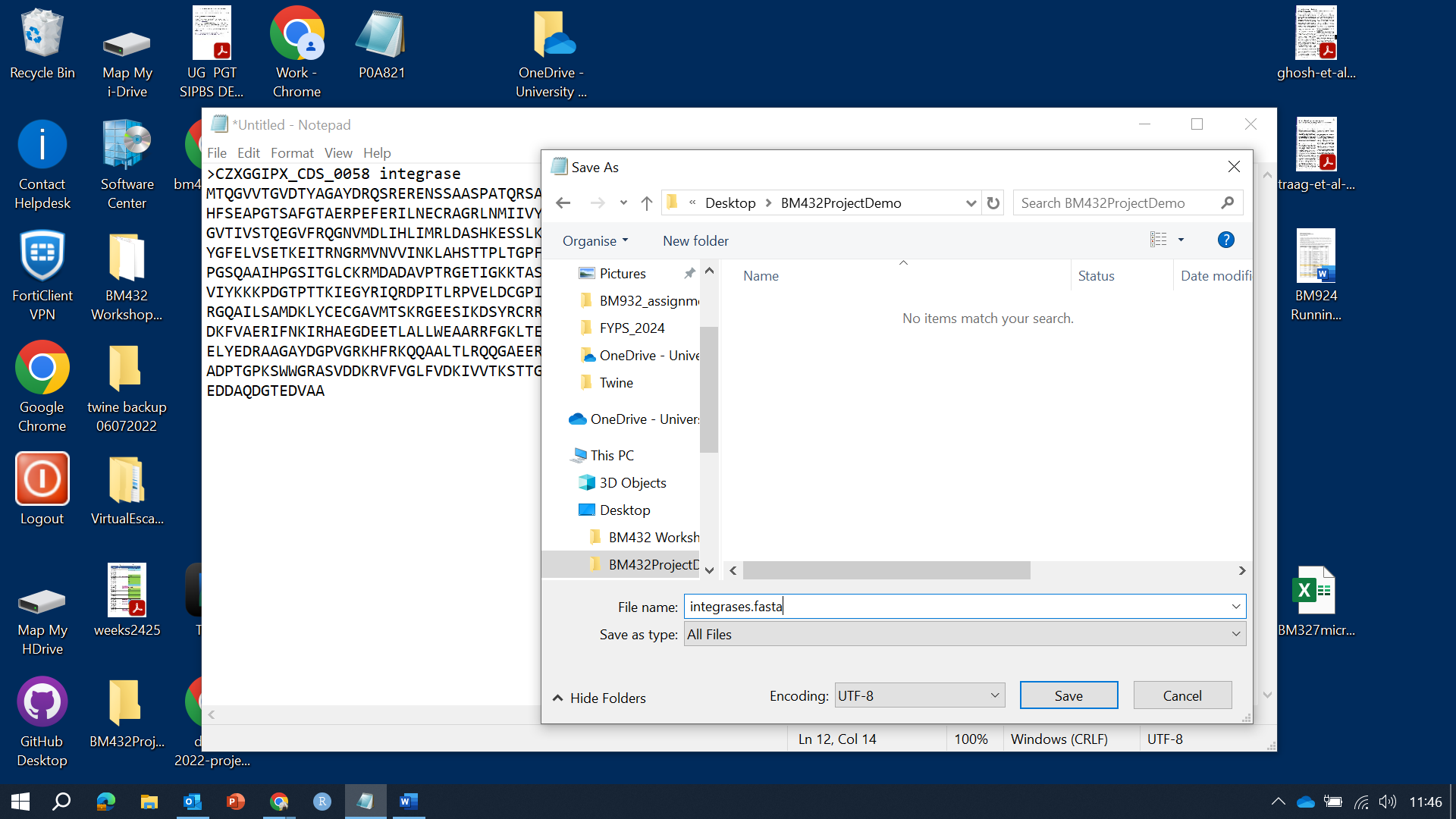


Figure 3. Saving a FASTA file

Now, you have a .fasta file for your gene/protein of interest, and the first sequence has been added to it.

Continue to add sequences according to your project needs (e.g., all the integrase protein sequences from my phages of interest…) until you have added all of the sequences that you want to align (Figure 4).

You may at this step want to consider whether you want to give your FASTA sequence headers more informative names (e.g., the name of the phage instead of the arbitrary string of characters that pharokka has assigned to it). (You can also change the sequence names at later steps, so don’t worry if you forget to do this now.)

Depending on your project/the type of tree you are wanting to build, at this step you should consider whether or not you need to add an **outgroup** (a more distantly related sequence that would serve as a reference group and allow you to root your tree. Check out [this introduction](https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/what-is-a-phylogeny/aspects-of-phylogenies/root/) to phylogenetics for more explanation about outgroups and rooting a tree….)

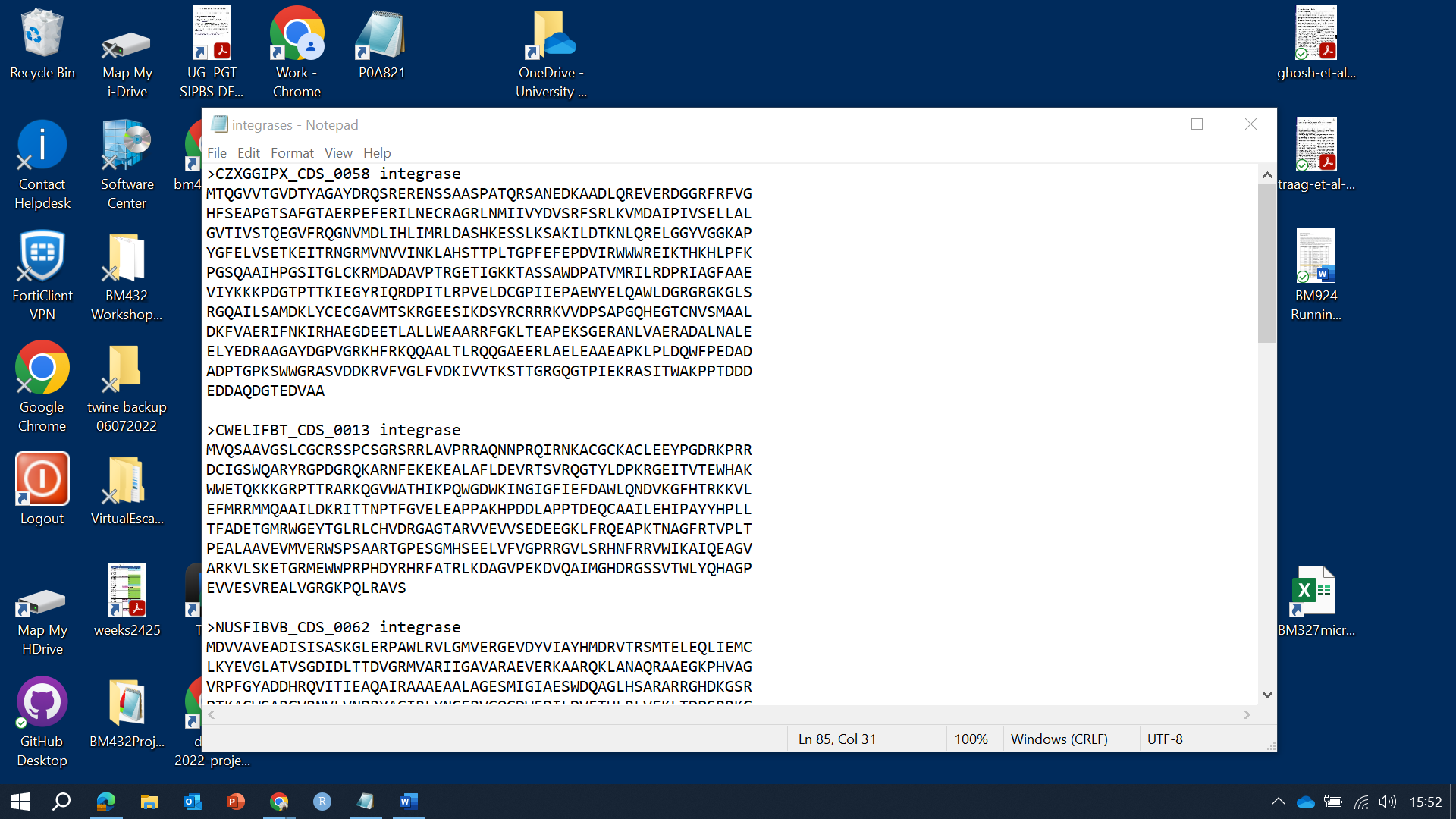


Figure 4. Screenshot of the integrases.fasta file

Upload this file to Galaxy (Click “Upload” in the top-left hand menu), then select your file in the pop-up box (Figure 5), and click “Start”.

A screenshot of a computer

Description automatically generated

Figure 5. Upload the integrases.fasta file to Galaxy

1. **Step 3:** Create a multiple sequence alignment (MSA) using your .fasta file

There are many different tools which can be used to create multiple sequence alignments, such as T-Coffee, ClustalW, MAFFT, etc. You can use any of these you like, but if you are creating multiple MSAs through the course of your project, it is likely best to be consistent and use the same program for each … unless of course you have a very good reason for choosing different programs.

For this example, I will use T-Coffee to create a MSA of our integrase proteins.

A screenshot of a computer

Description automatically generated

Figure 6. Using T-Coffee to create a multiple sequence alignment.

You can find T-Coffee in the tools menu by searching for it (or by browsing to it – it is in the category Multiple sequence alignments). You can then point it towards your uploaded .fasta file (here: 73: integrases.fasta - Figure 6).

Depending on your downstream analyses, you may wish to tick the box for “phylip” under “Output formats” (e.g., if you plan on using PhyML to create your tree).

Click “Run Tool”.

A screenshot of a computer

Description automatically generated

Figure 7. phylip output from running T-Coffee.

Your output should look something like Figure 7. You should manually examine the MSA (either here in Galaxy or using specialised software such as Jalview for viewing MSAs - Figure 8). At this stage, it is important to ensure that everything looks OK (if you have inadvertently included some sequences in your .fasta file that don’t correspond to your gene/protein of interest, they won’t align properly and you will see huge gaps/poor alignments in your MSA. You should always manually examine your MSA before proceeding to the next step.

A screenshot of a computer

Description automatically generated

Figure 8. Manual examination of the integrase MSA using JalView.

1. **Step 4:** Use the multiple sequence alignment to create a phylogenetic tree for your sequences.

Again, for this step, there will be multiple options (different programs you can use to build your tree – for this example I will use IQ-tree, but you might also use PhyML, RAxML, or other tree-building programs … again, I would suggest (unless you have very good reasons) to use the same tools consistently throughout your thesis.

A screenshot of a computer

Description automatically generated

Figure 9. Using IQ-Tree to build your phylogenetic tree.

In IQ-tree, you will want to point the tool to the correct dataset (here I am pointing it at the phylip output from T-coffee, a.k.a. my MSA) - Figure 9.

Importantly, specify the correct type of sequence you want to build a tree for (here I have chosen AA as I have protein sequences (amino acids). If you have nucleotide sequences you will need to specify DNA. **If you do not specify the correct sequence type, you will get an error message and IQ-tree will not be able to build a tree for you.**

You will also want to specify the number of bootstraps for the tree to run (Figure 10). Click on “Bootstrap Parameters”, then “Ultrafast bootstrap parameters”, and then Specify the number of bootstrap replicates (1000 is usually sufficient).

Click “Run Tool”.

A screenshot of a computer

Description automatically generated

Figure 10. Specifying the number of bootstrap replicates in IQ-Tree.

Your results should be added to your history in an item called “Report and Final Tree” (Figure 11). Note down the best-fit model (in this example, Blosum62+F+I+64).

A screenshot of a computer

Description automatically generated

Figure 11. IQ-TREE Report and Final Tree.

If you scroll down to the bottom of the report, you will find the consensus tree (Figure 12). If you want to view/edit the tree in a tree viewing program (such as iTOL) you can copy the newick formatted tree (highlighted text).

A screenshot of a computer

Description automatically generated

Figure 12. IQ-Tree consensus tree in newick format (highlighted)

Next time we’ll look at tree visualisation and further downstream analyses….