Using the Genome2-ID Web Portal

For more information on how Genome2-ID works, visit: www.dna4tech.com

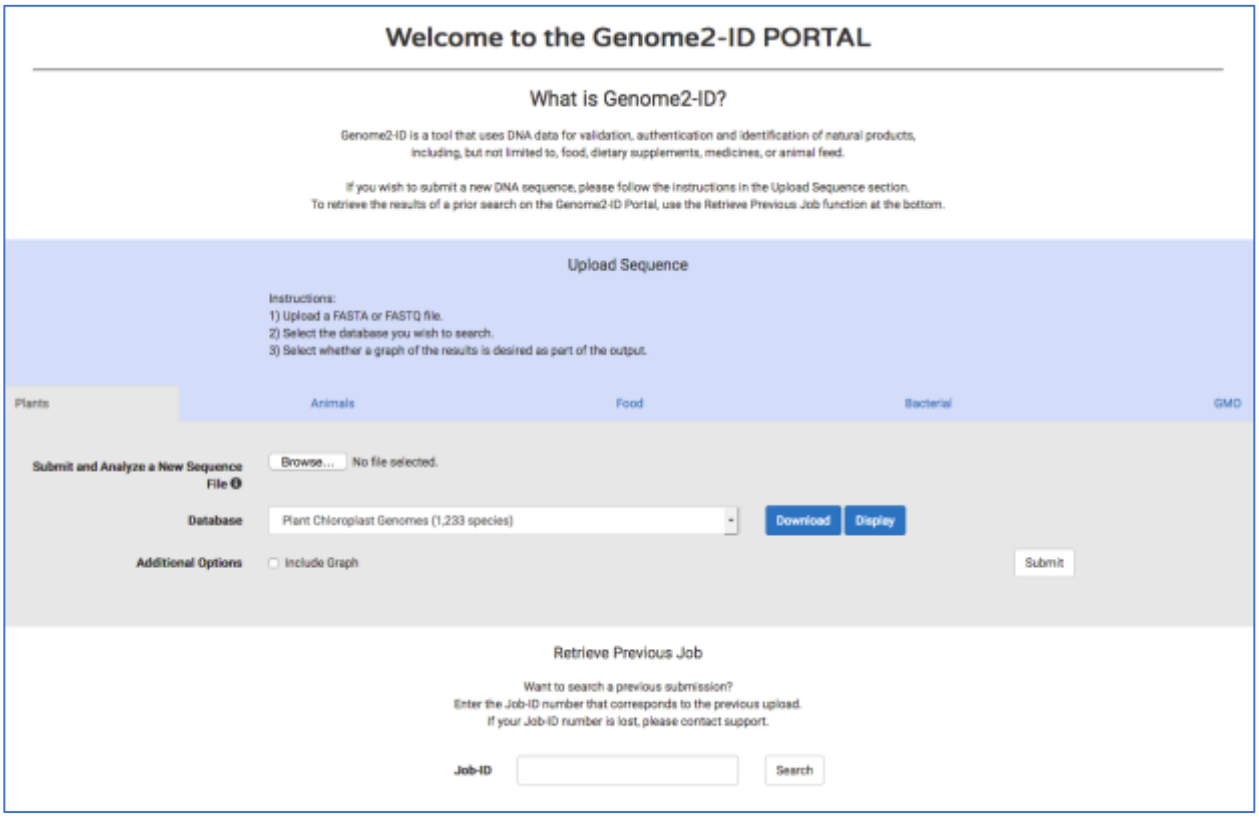

Figure 1

Overview

The Genome2-ID portal encompasses two functions. The first function is an analysis of new DNA sequence data. The Upload Sequence section allows for selection of user provided DNA sequence data, selection of the and additional options such as generation of a graph that plots the results. Separate tabs allow for the selection of taxon specific databases, while the drop down database menu allows for genome versus DNA barcode databases. Submissions of DNA data through the Upload Sequence section will result and analysis of DNA data using the database specified by the user and will display the results on the page when finished with an option to download the results.

The second function is archival and retrieval of prior analysis jobs. Each new submission is given a unique Job-ID. Each Job-ID may be used in the Retrieve Previous Job section to display and download the results of that analysis. Users may navigate away from the page once their data is uploaded and they receive a Job-ID, the results will be displayed on the page when they return to the page. Users may also close the web page and return to retrieve their results using their Job-ID.

Using your own DATA:

Data Format: The portal accepts both FASTA and FASTQ format data. Data in any format will not be analyzed by Genome2-ID.

Data Types: The program is designed to utilize Whole Genomic Shotgun sequence data (WGS) which captures a large, random sample of the DNA in a sample. Data from other genomic library formats, such as RAD-Seq or RNA-Seq may work but are not supported. Data from the sequencing of PCR on Next-Generation platforms is supported, but the PCR must target chloroplast DNA sequences in order to match the database.

Selecting your Data: Use the 'Browse' button to navigate to the file on your computer you wish to submit to the portal.

Number of Sequences to Use: We have observed that between 300,000 to 1 Million sequences produce high-quality results. More sequences will not improve results significantly but will increase the time required to obtain results.

Selecting the Database: Currently, there are two databases available. The Plant Chloroplast Genomes database contains the entire chloroplast genome (approx. 150,000 bp sequence) for each of the 1,233 species for which we have the chloroplast genomes. The DNA barcode database contains DNA barcode sequences (rbcL and matK) for 43,223 species. The Plant Chloroplast Genomes database is the default database that will be searched; the DNA barcode sequence can be selected in the drop-down menu next to 'Database'.

Species in the Database: The species that are included in the database may be seen with the 'Display' button which will allow users to view and search for the names of species. Only the formal taxonomic names are included in the database at this time. You may search for different taxonomic levels, Family, Genus, and species.

Results Output: The default results from Genome2-ID are a table describing the species in the database that most closely match the DNA sequences in your input file. You may download the table in tab-delimited text format. To learn more about what the table says about your data, visit the Genome2-ID page at www.dna4tech.com.

Job-ID: Each job that is submitted and run to completion will have a job-ID assigned to it. The job-ID can be used to retrieve outputs from longer searches that complete after the user closes the web page to the portal. The job-ID can be entered into the Job-ID window, after which the user clicks the 'Select' button. The retrieved results will display on the page and may be downloaded.

Graph Option: Selection of the Graph option will result in the production of a graph that plots the 5 best matches when using 20%, 40%, 60%, 80% and 100% of the input sequence data. The graph may be downloaded with the tabular results which are produced by default.

Using GenBank DATA with Genome2-ID

Users who do not have their own DNA sequence data may use data from publically available databases, such as the Short Read Archive from GenBank. While there is not a very wide range of medicinally important species in the Short Read Archive section of GenBank, there are data from many species that can be used to evaluate the performance of the Genome2-ID system. In the pages below, we outline the steps you can take to obtain data from GenBank which may be used as described above, in testing and evaluation of the program.

PART 1 – find data on GenBank

1- Navigate to the NIH web page which hosts genetic data: https://www.ncbi.nlm.nih.gov/. A screenshot of the NIH web page is below in Figure 2.

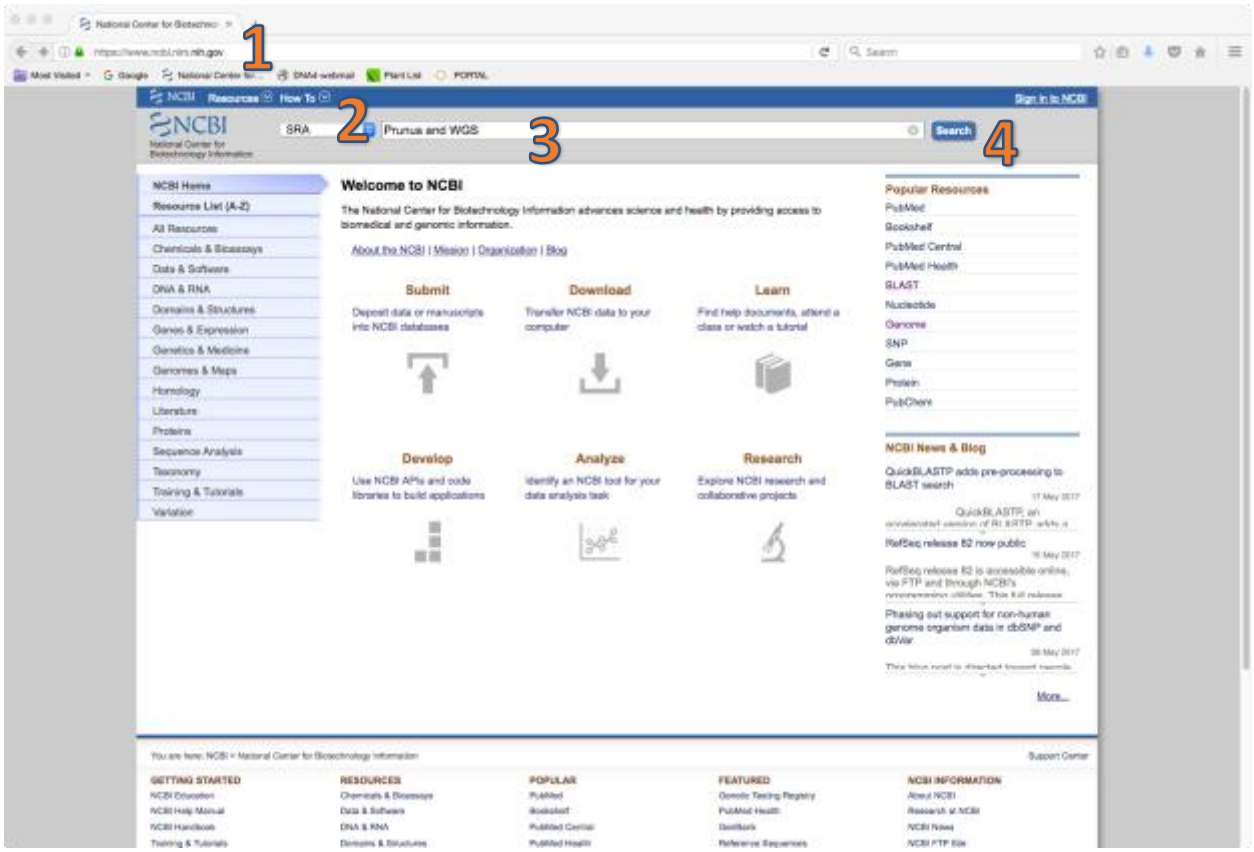

Figure 2.

2- Select the SRA tab from the drop-down menu. SRA is the 'Short Read Archive' where whole genome shotgun data is stored (along with other types of Next Generation Sequencing data).

3- Type in the name of the organism you want to use in your test in the search bar (use the scientific Latin name of the species, I use the genus name to broaden my results). To get the right type of data add the words “and WGS” to the search. The Genome2-ID algorithm works best with WGS data. Not every species will have SRA or WGS data, so you may have to be flexible.

4- Click the search button and off you go.

The next page (Figure 3, below) will be the search results matching what you typed into the search bar. Below is my example having use the search term: Prunus and WGS.

5 – Select a record that looks like it has the species you want to use in your test. I chose the 7th one on the list.

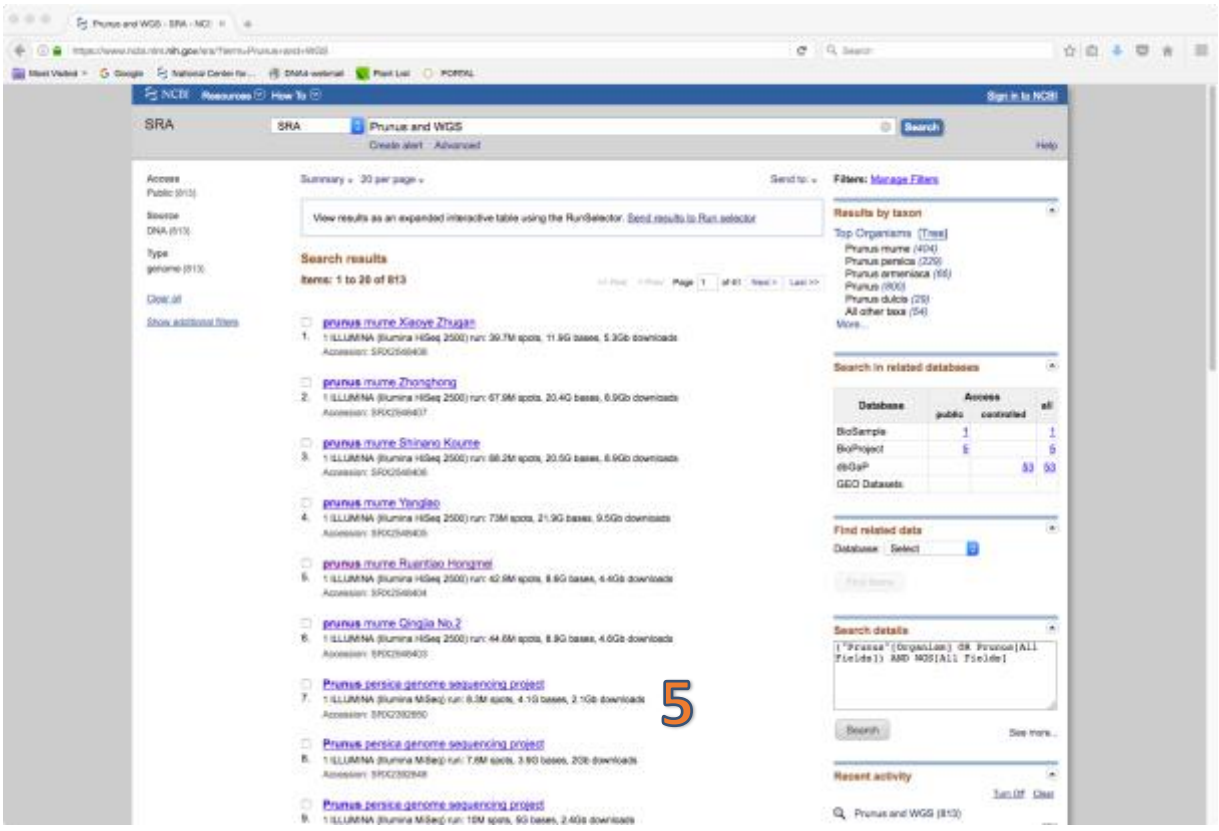

Figure 3

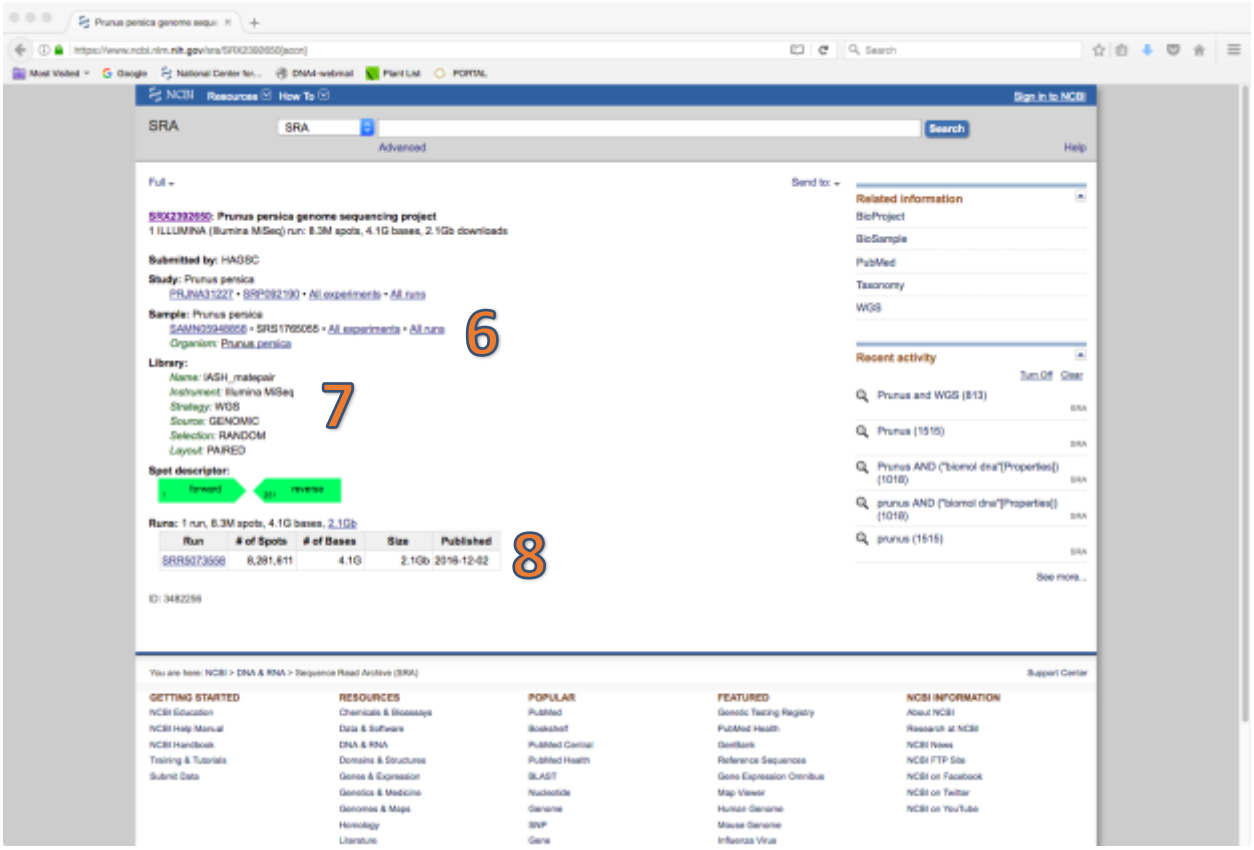

Figure 4

6 – When you select that 7th record, it will take you to the records page (Figure 4) that describes the sample, what was sequenced, and how the sample was prepared for sequencing.

7 – Every SRA record will have 6 descriptors of how the data was prepared. The Name, Instrument, and Layout are not important. Data that works best will Genome2-ID have:

Strategy = WGS
Source = Genomic
Selection = RANDOM
These three parameters conform how we prepare data in our laboratory for use in the forensic identification of products. Genome2-ID is meant to work with our laboratory data, and thus when we seek to test and validate Genome2-ID with GenBank data, we use data that most closely approximates what we would produce ourselves.

8 – Near the bottom of the record, you will find the Run which is the identifier for this sequence record in GenBank (in this case: SRR5073558), as well as the number of sequences that are contained in this record (in this case 8,281,611). They call this the # of spots, which equals the number of sequences. You will need between 500,000 and 1,000,000 sequences to correctly analyze an unknown sample.

Part II – Download GenBank data using fastq-dump.

These data files are BIG, so you will need to download and utilize custom software made by NIH to download SRA data onto your computer for testing. This is command line software, so you will have to be comfortable to open and utilize a terminal window to use the software.

This software is called the SRA Toolkit. You can find the software here: https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/.

Each user will have to identify, download and install the right version of the software. In general, most windows users will download and install "MS Windows 64 bit architecture ", while MAC users will select "MacOS 64 bit architecture ".

SRA Toolkit is actually a set of programs, not just one. However, to download data from SRA, you only need to use the program called 'fastq-dump'. There may be several versions of this in your download, but any will work.

9 - Open a terminal window and navigate to the SRA toolkit folder; then find the subfolder in which fastq-dump program is located. See Figure 5 below.

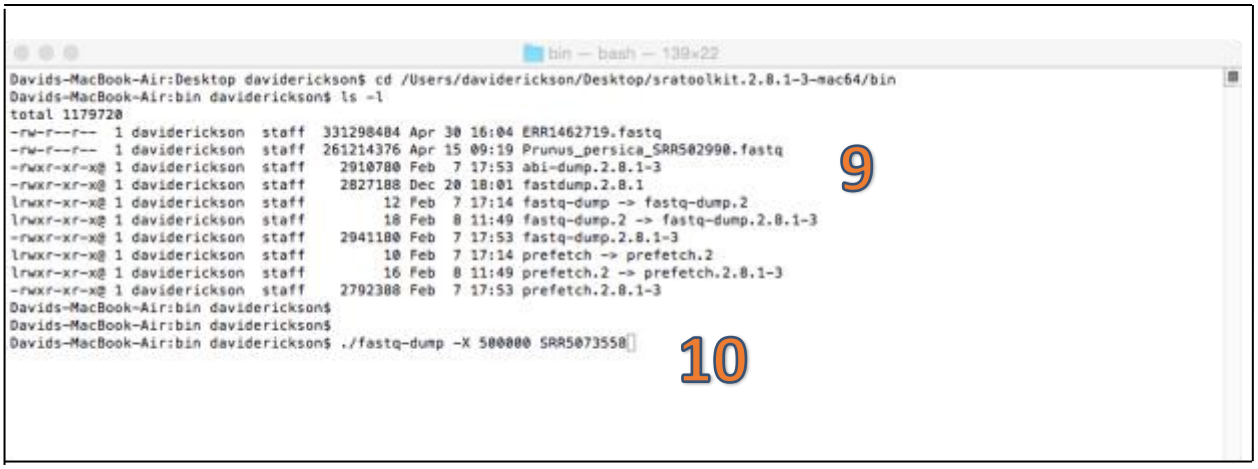

Figure 5

10 – Call the fastq-dump program (you may need './" in front as above). We specify two parameters when we use fastq-dump, the number of sequences we want to download, and the SRA record from which we want to download sequences. You don't need all 8,281,611 sequences in the record, so we will use the –X command (upper case X) to download only as many sequences as we want, in this case 500,000. Then we must list the record ID, in this case it is SRR5073558.

The complete command will look like:

./fastq-dump -X 500000 SRR5073558

Because the files are large, it can take several minutes to download them. This also depends on your connection speed. They will be downloaded into the same location/folder as the fastq-dump program and will be named same as the record ID (e.g. SRR5073558.fastq).

Once you have the data file downloaded onto your desktop, you can go to Genome2-ID portal and submit it as described in the first part of this tutorial.