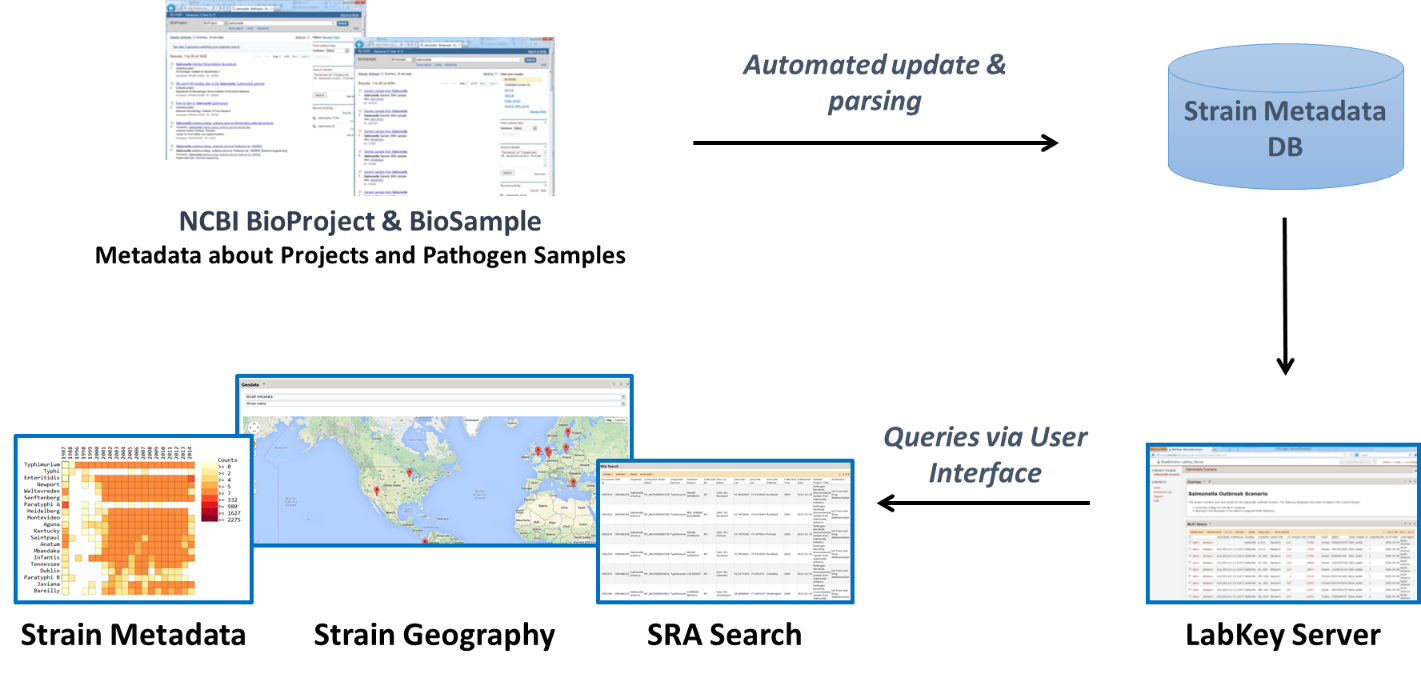
**Pathogen Metadata Toolkit: User Instructions**

# **Overview**

This toolkit is designed to provide quick access and analysis for data about pathogen samples. Pathogen metadata, such as where and when a pathogen was collected and the type of environment it came from, is a critical complement to DNA data in uncovering the origins of new pathogens, understanding patterns of disease outbreaks, and identifying DNA signatures. This toolkit enables easy query of metadata from the largest public repository for this data in US, NCBI BioSample. The query results include standardized data, are filterable, and are in an exportable format for other analyses. The toolkit also provides a heat map summary of query results, and quick mapping of the locations where samples were collected. The toolkit is built on the LabKey data platform, an open-source data management platform designed for biological data, and uses a PostgreSQL database designed for capture of metadata useful for disease outbreak investigations. This open-source toolkit can be modified and extended for other data sources, including data an institution has internally. In addition, some of the components are modular, and the database can be used with other data platforms.



# **Instructions**

***Installation*** For installation instructions, see the README document.

***Accessing the Toolkit*** Once on the LabKey site, sign in, go to the homepage, and choose the “Metadata Query Portal”. Within the Metadata Query Portal, there are three pages that appear as tabs near the top of the screen: Strain Metadata for querying of pathogen names, SRA Search for identifying metadata associated with DNA sequence read ids, and Strain Geography for mapping locations.

***Query for Metadata*** To query for pathogen metadata on the Strain Metadata page, type in a pathogen name in the query box. This could be a pathogen species name, part of a name, or a serovar name. The search will be for any string that contains the exact name (not case sensitive), and therefore results will include cases where the string is embedded in a larger text string. Click one or more of the three indicated database fields to search, and then hit submit. To query for metadata on the SRA Search page, enter NCBI sequence read identifiers (e.g. SRS#### or ERS####, not SRR## or ERR##) and hit submit. To search for more than one SRA id, enter a list of SRA identifiers using a comma-separated list without spaces.

***Query Results*** Both of these query pages return a heat map summarizing the results, and a table of metadata results. This table contains data parsed from NCBI BioSample. The variable *Collection Year* contains years extracted from the Collection Date attribute in BioSample. The variable *Strain* contains strain names extracted from the BioSample fields Strain, Isolate and Organism (in this order). The variable serovar contains serovar names extracted from the BioSample fields Serovar and Organism (in this order).

To refine the results in the table, click the built-in LabKey button “Views” and then “Customize Views” in the top left corner. Use “Columns” to select the variables that appear in the view, use “Filter” to refine the search results or eliminate samples missing data in certain fields, and use “Sort” to select variables to sort by. Use the “Export” button to export the results to EXCEL or text formats.

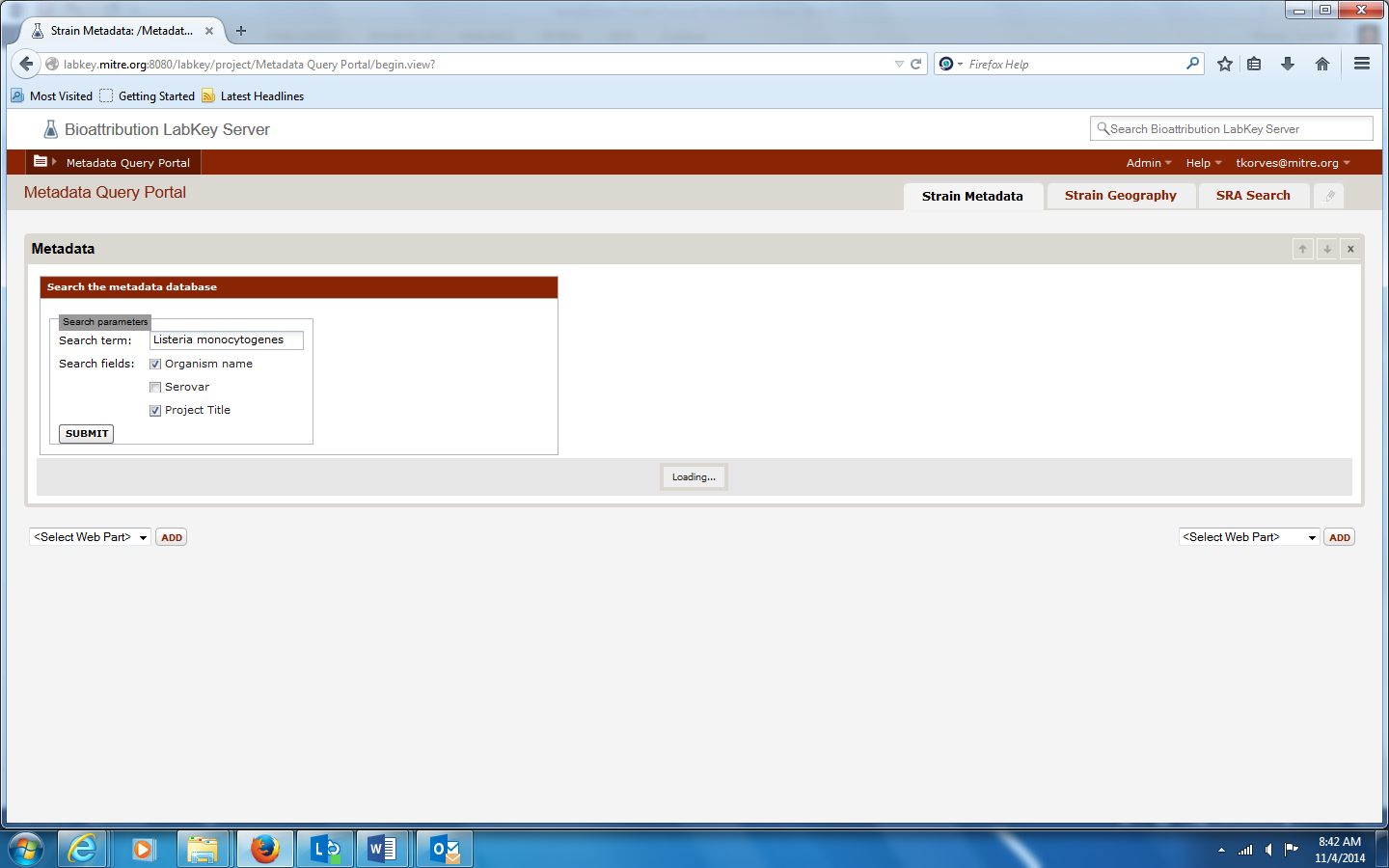
***Heat Map of Query Results*** The heat map depicts the counts for pairs of variables that appear in the results. Any changes due to filtering of the results table will result in changes on the heat map. To select the variables in the heat map, use the drop-down menus. The default setting is to show the top 20 most abundant categories for each selected variable. To adjust the number of categories shown, enter the desired numbers of row and column variables. Choose whether to include “unknown” as a value by clicking/unclicking the corresponding box. After making any of these selections, hit the “Submit” button that appears above “Plot Parameters” or “Refresh Plot” to implement the selections on the heat map. To see the exact count numbers in each box, scroll over the boxes of the heat map. Finally, the count data can be exported by clicking the “Export Data” button and saving the file.

***Geocoding and Mapping*** The Strain Geography page will geocode and map locations listed in any table. The first step is to import a table into LabKey. In the upper right, click the Admin pull down menu and select Manage Lists. Click Create New List. On the next screen, enter a name for the list, select the box for Import from File, and click Select List. On the next screen, choose Browse, select the desired file, and click Import List. Next, select the Strain Geography page, and select the table of interest from the first drop down menu. Then select the variable to be geocoded (e.g. Geo Loc Name). The sample locations will then appear on the map below.

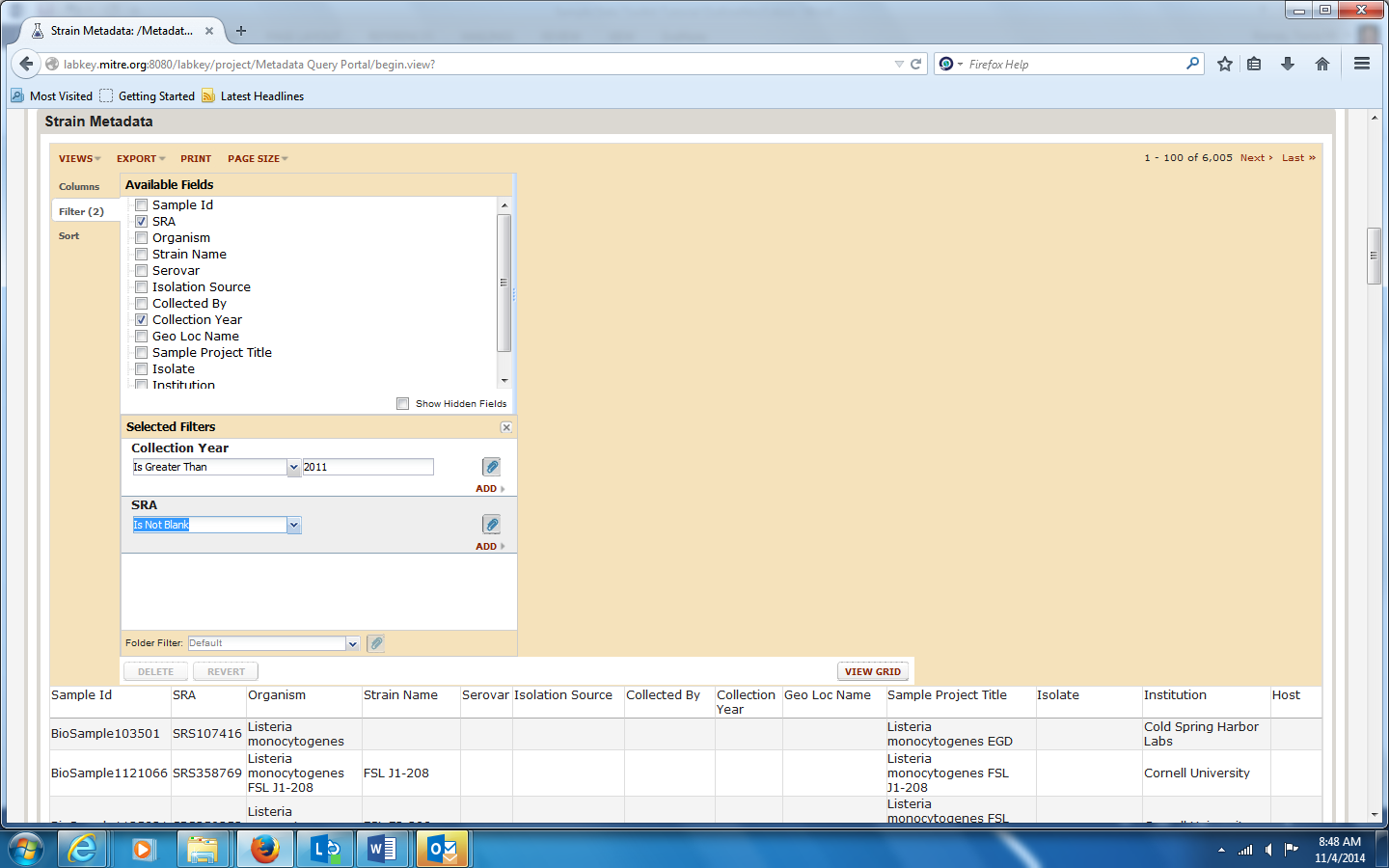
# **Examples**

**Example 1**. There is a new disease outbreak, and investigators want to determine whether it is closely related to other recent outbreaks. What genome sequences are available for this pathogen that were collected in the last three years, and where were they collected from?

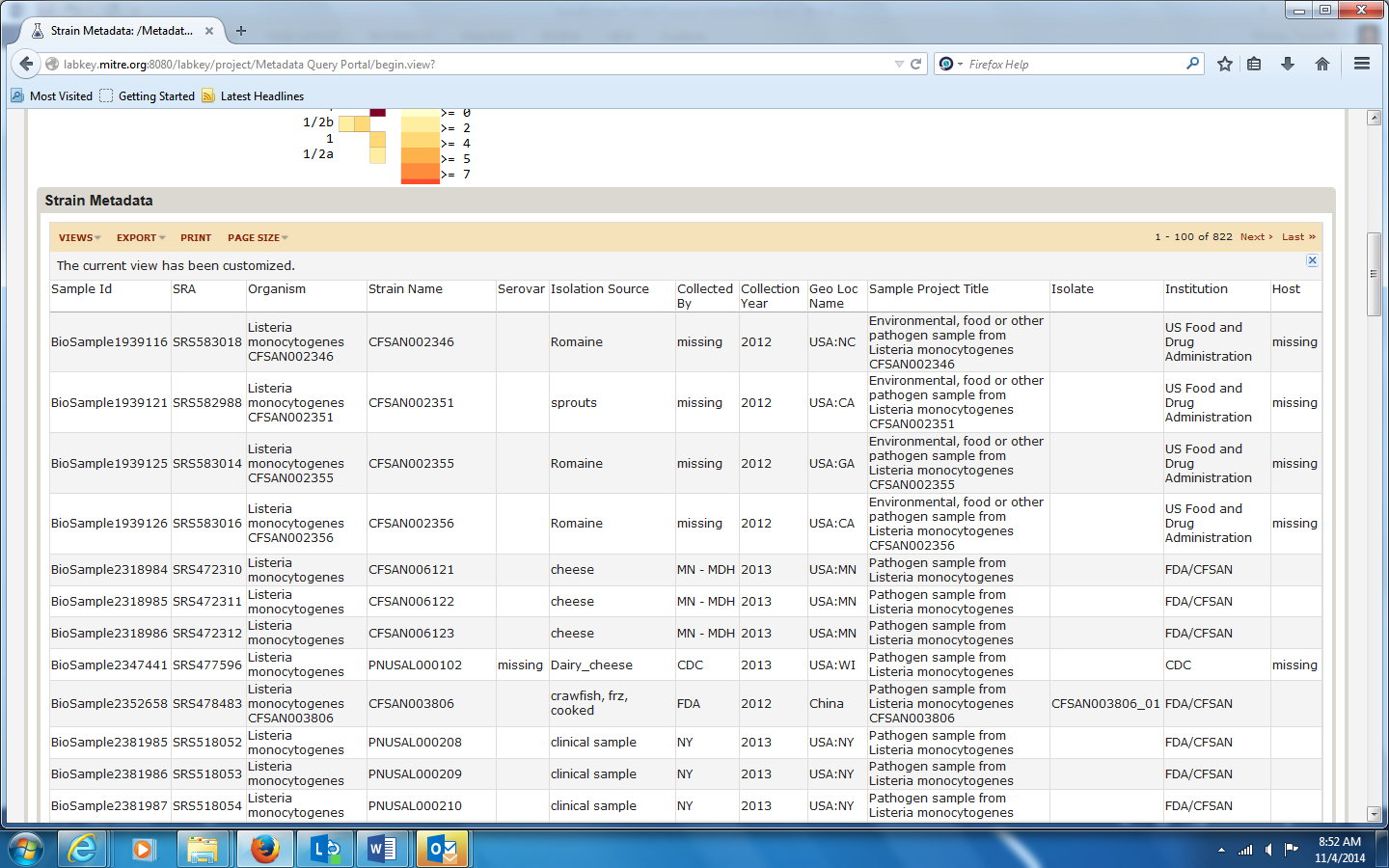
* 1. Query for pathogen name: Click on the Metadata Query Portal folder, and select the Strain Metadata tab at the top-right. On the Strain Metadata page, enter the pathogen name, select the fields Organism name and Project Title to search, and click Submit:



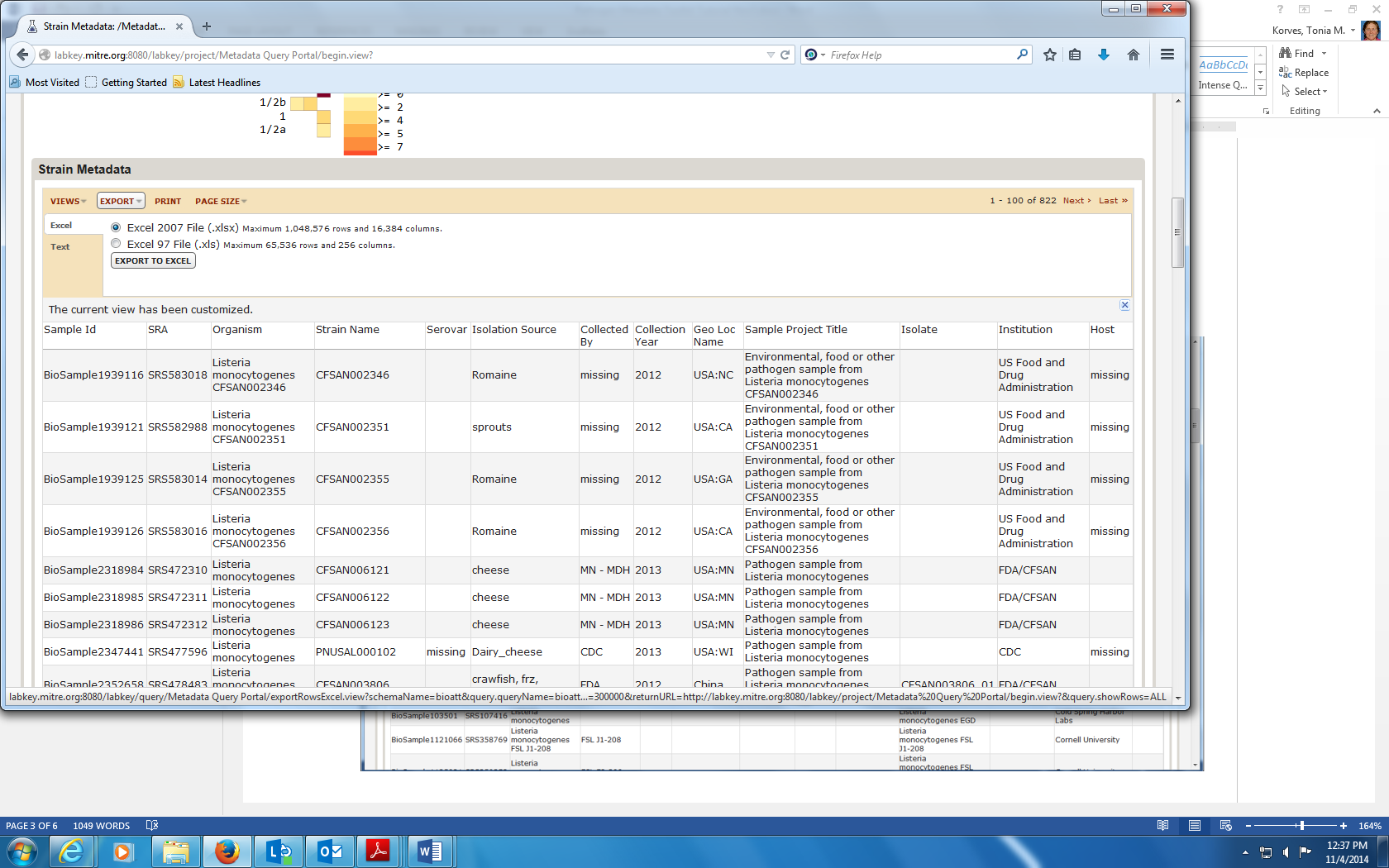
* 1. Filter the results in the table. Filter by Collection Year, and by presence of SRA id (indicates that a genome sequence is available for that strain). Scroll down to the results table, use the Filter menu under VIEWS, select the variables and types of filtering, and click VIEW GRID:



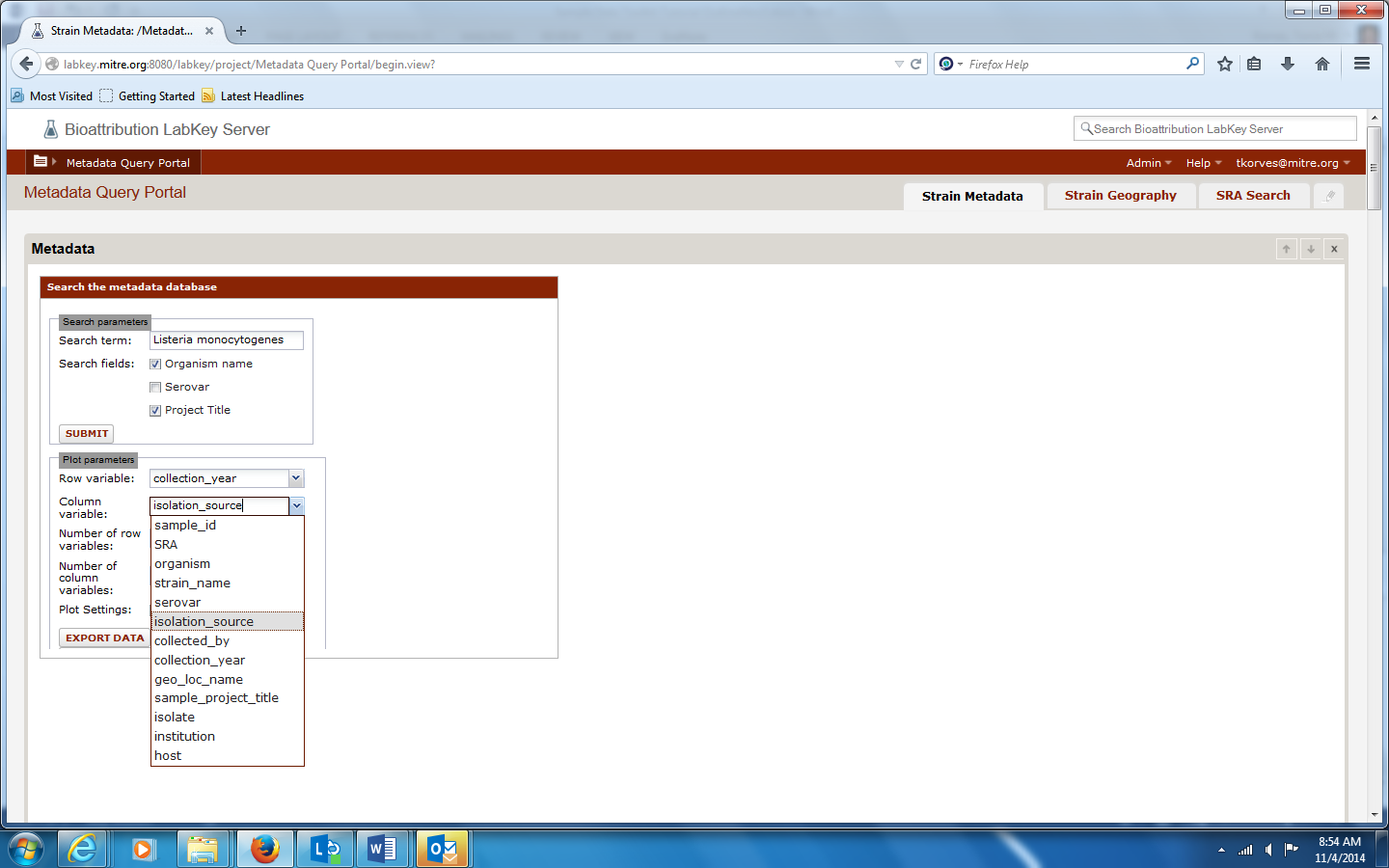
New results table appears:



* 1. Export and save table of query results

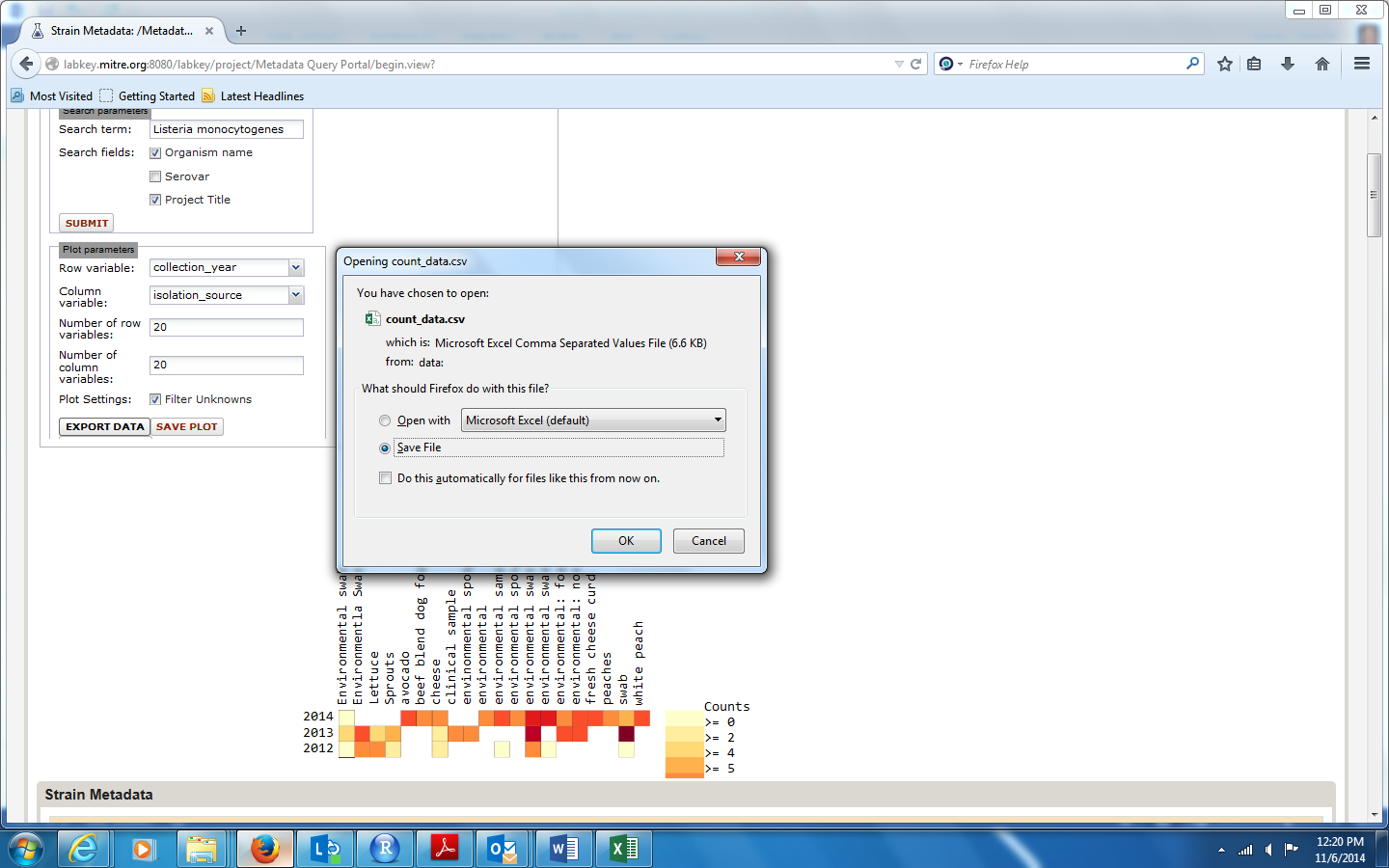


* 1. View a summary of isolation sources by year. Select Collection Year and Isolation Source from the drop down menus, adjust number of variables to display, and click the Submit button:

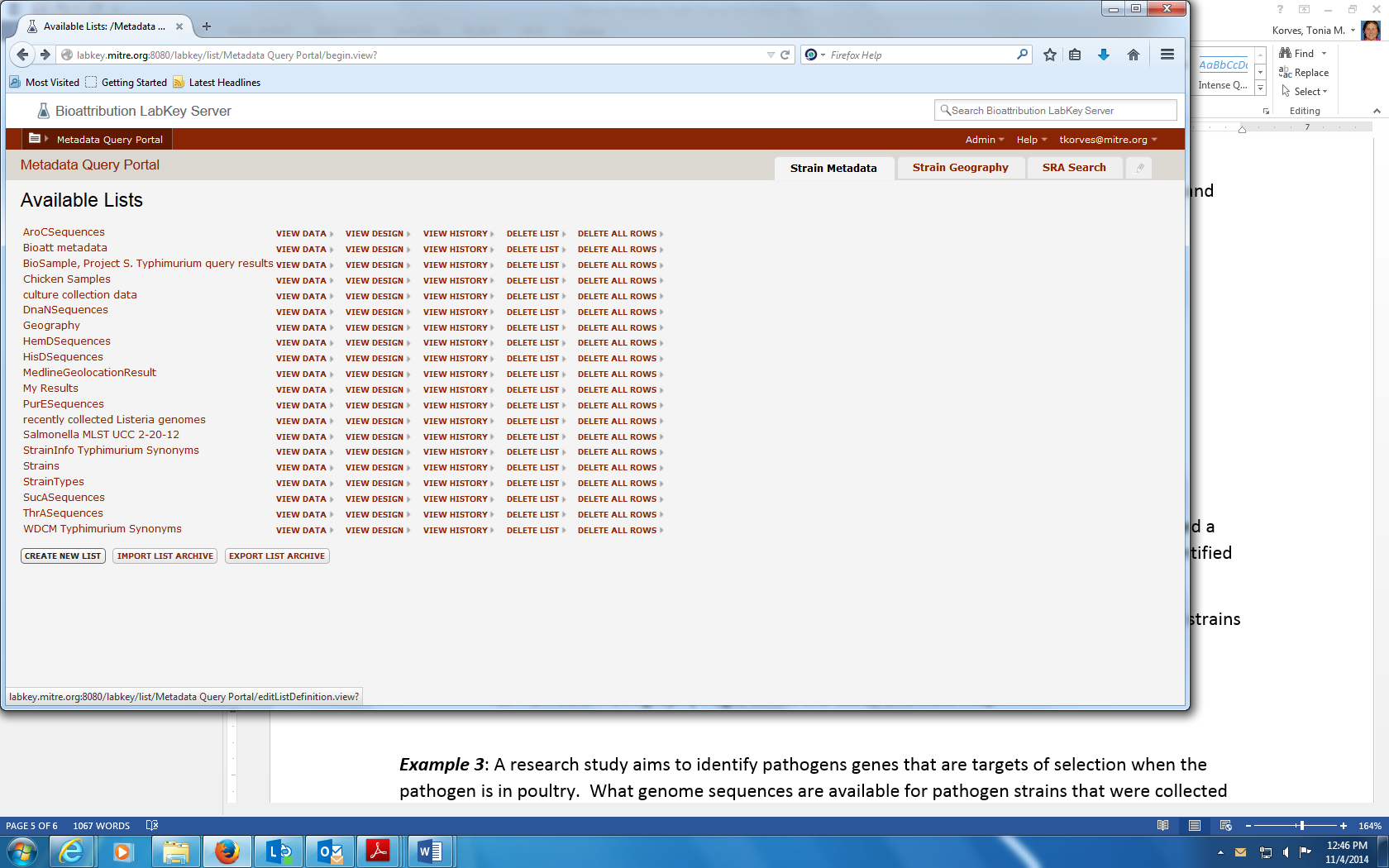
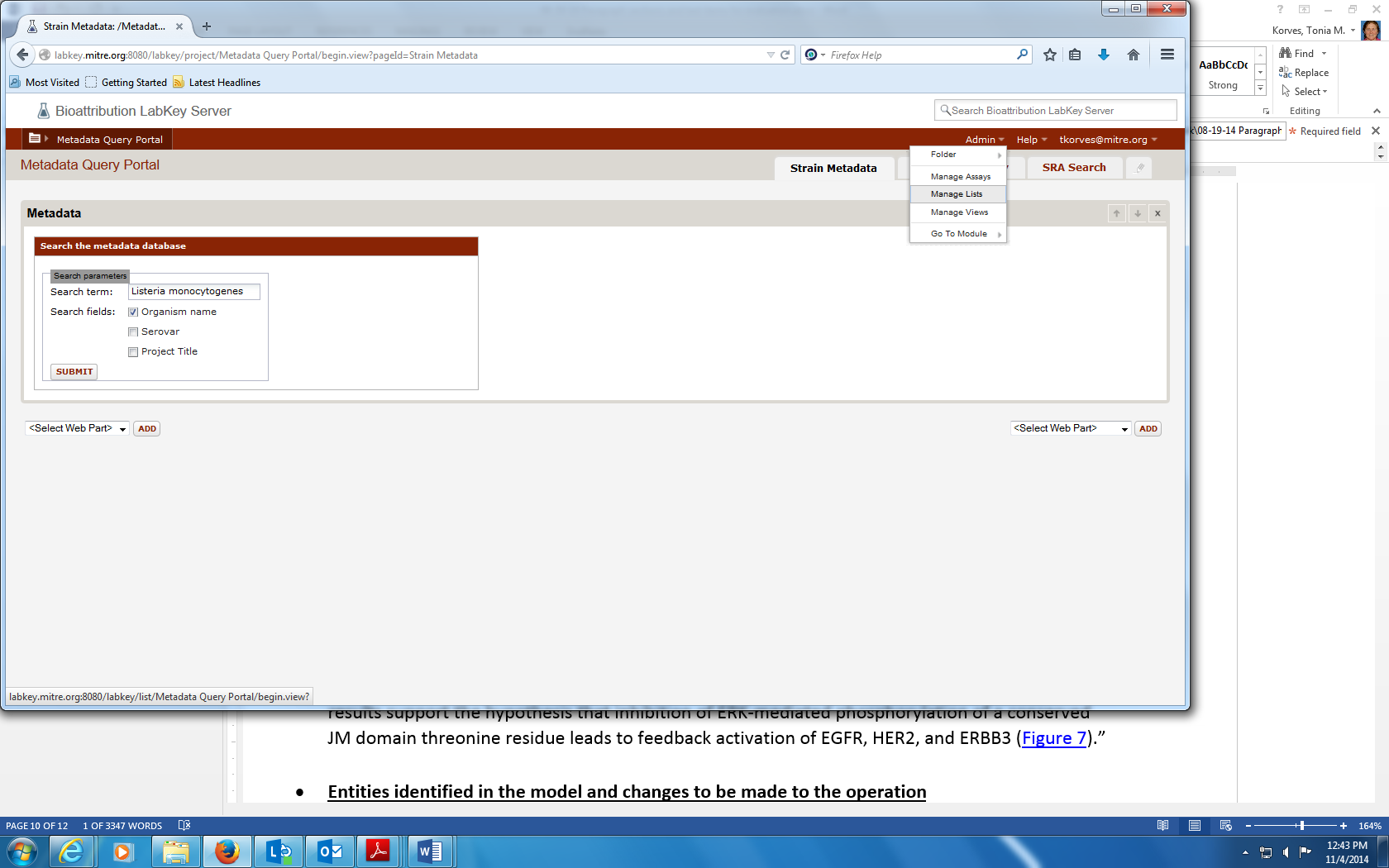


View the heat map. Scroll over boxes for counts for pairs of variables:

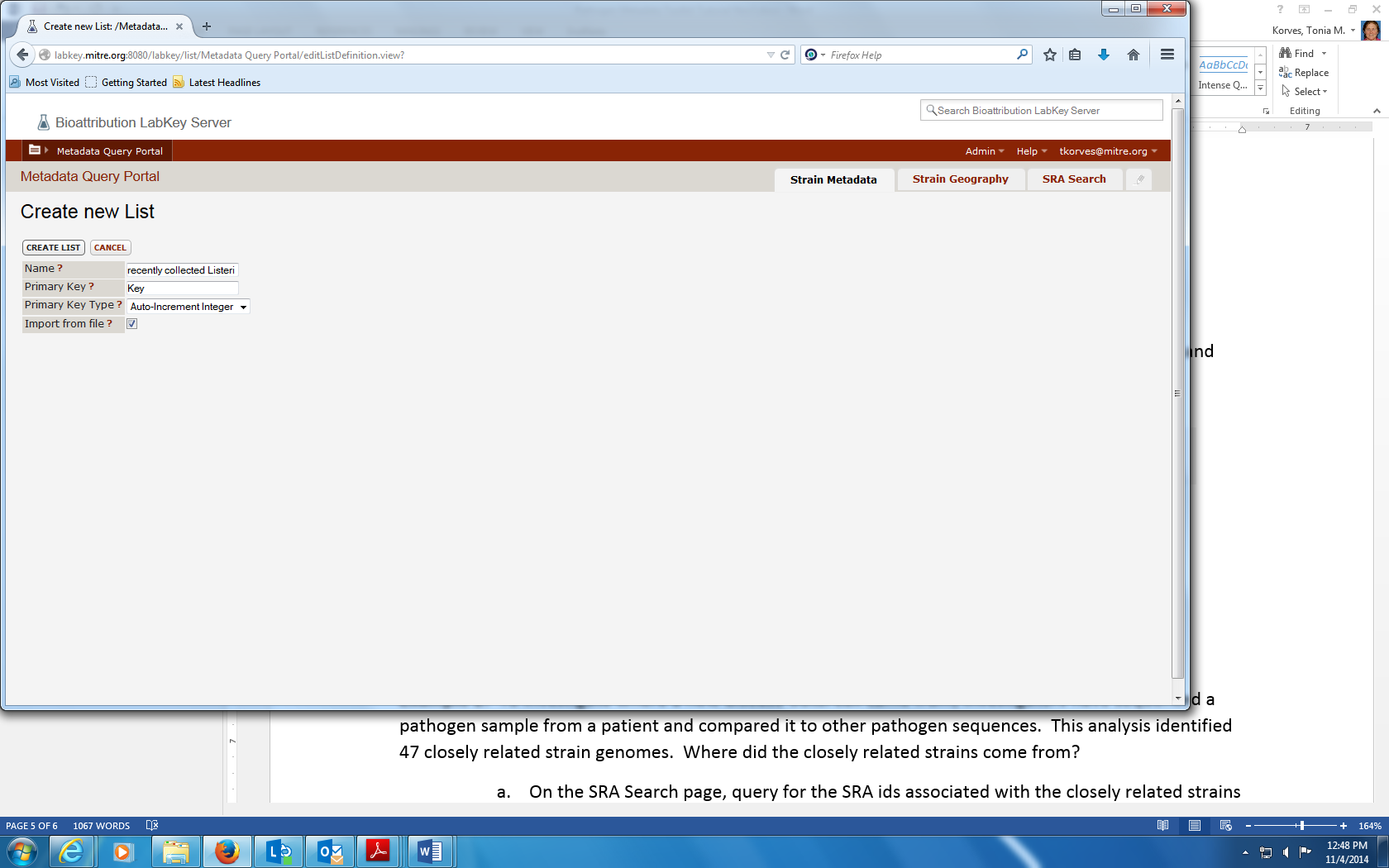


Export and save count data:

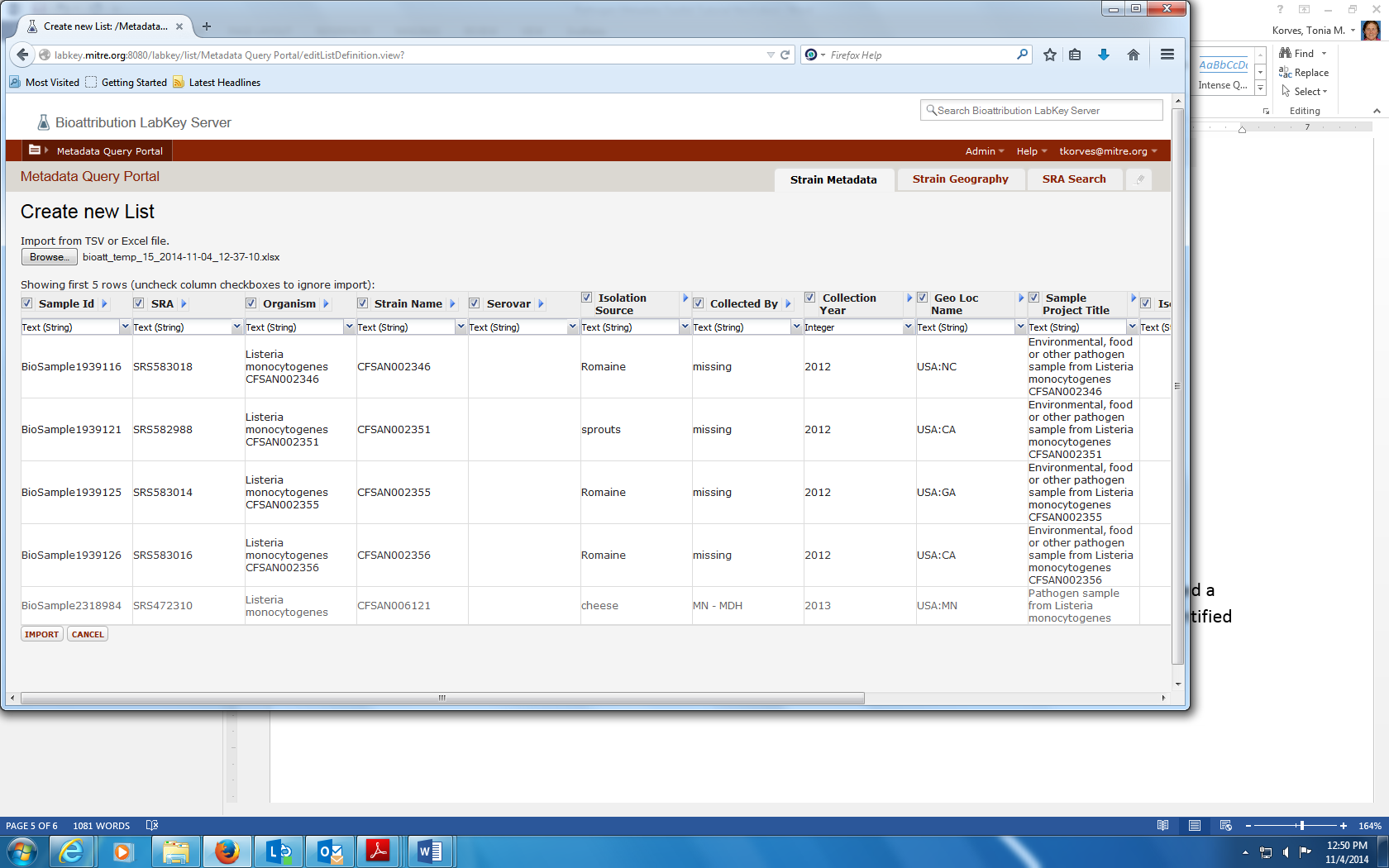
* 1. To geocode and map locations, first import the results table into LabKey as a list. At the top-right, choose Admin, then Manage Lists, and then click the button Create New List



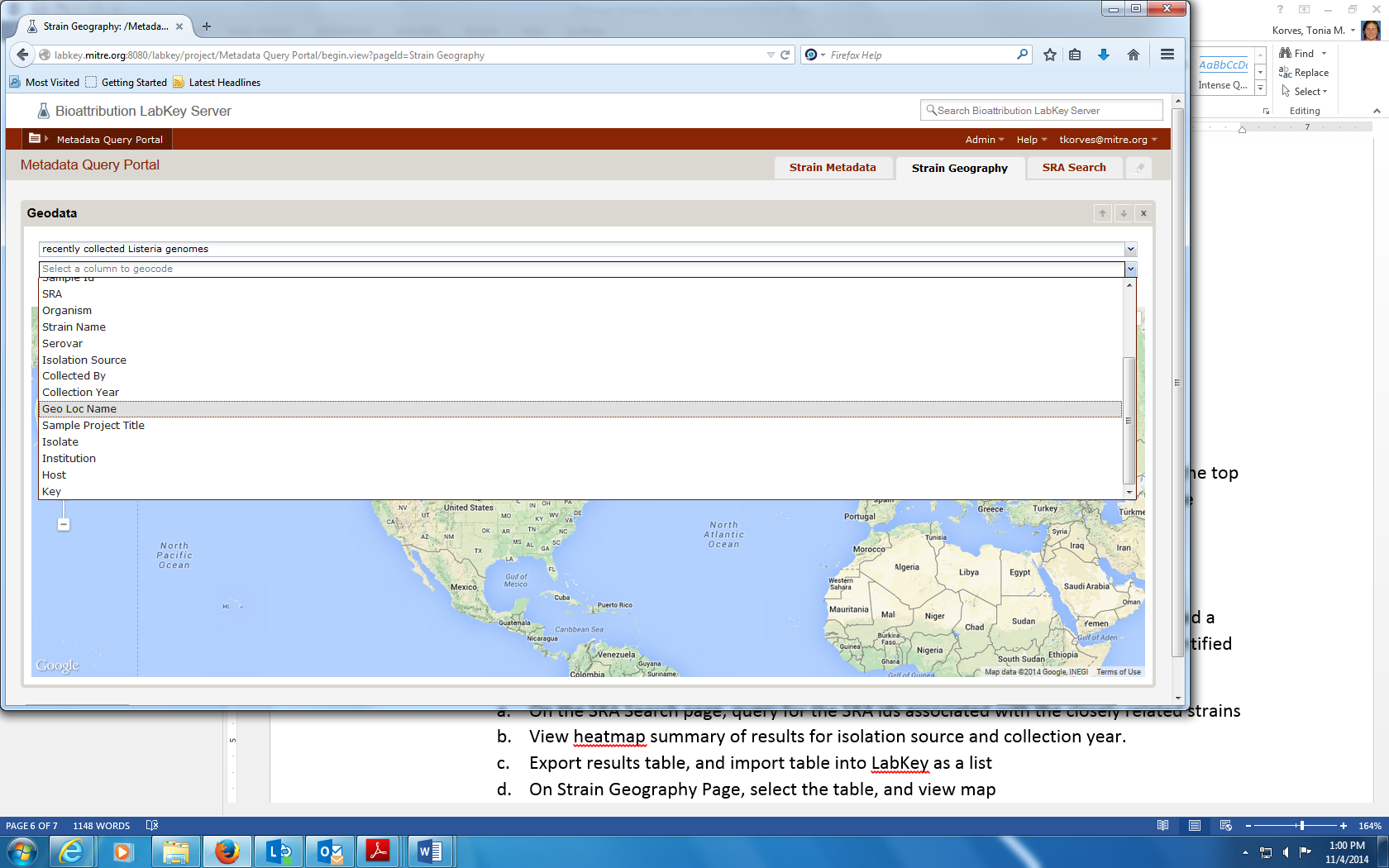
`Enter a name for the list, click Import from file, and click Create List:



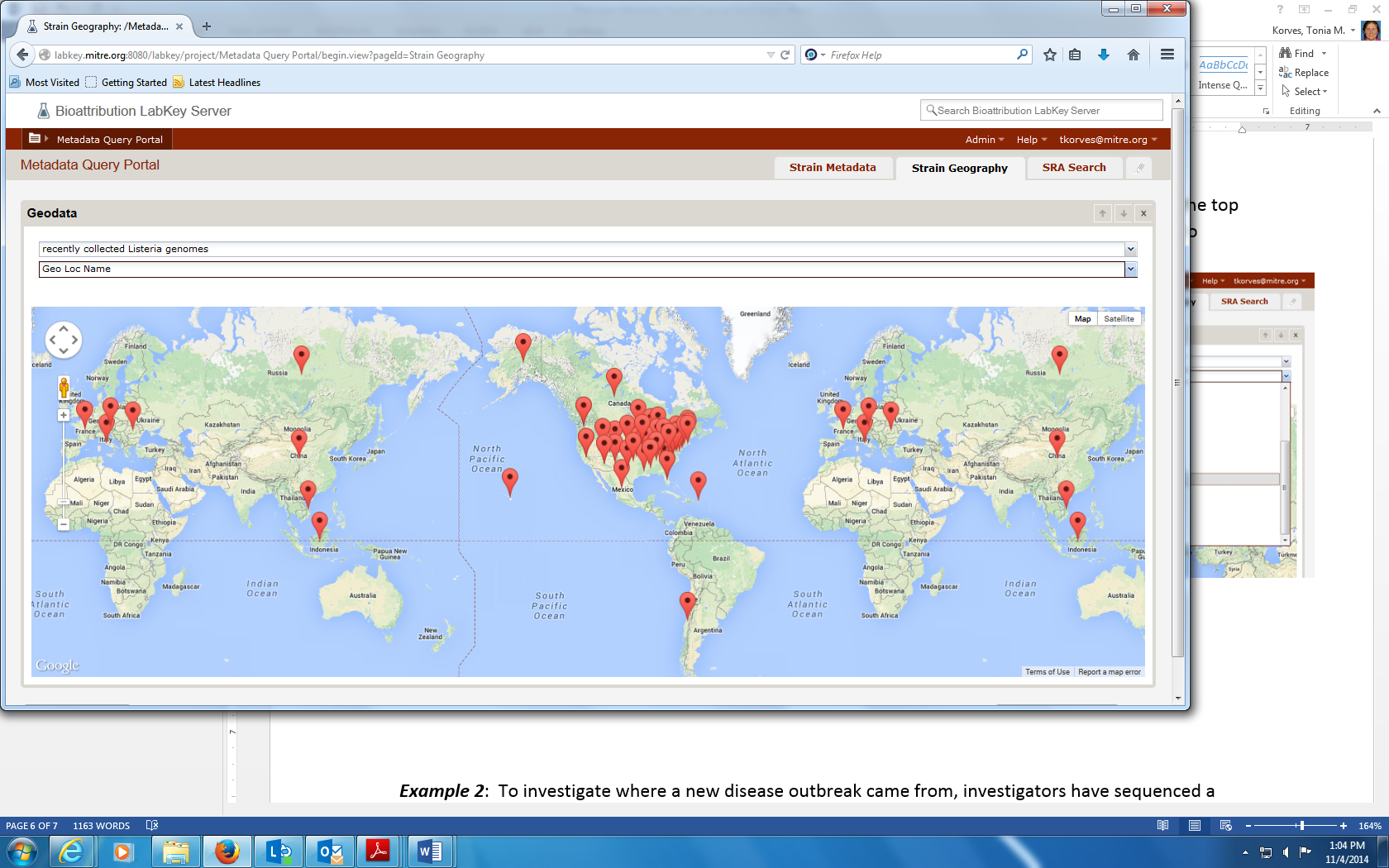
Browse, select list, and click Import at the bottom:



* 1. To geocode and map, at the top-right of the screen, select the Strain Geography tab. Use the top pull down menu to select the list with the data, and then use the second pull-down menu to selection the location variable to geocode.

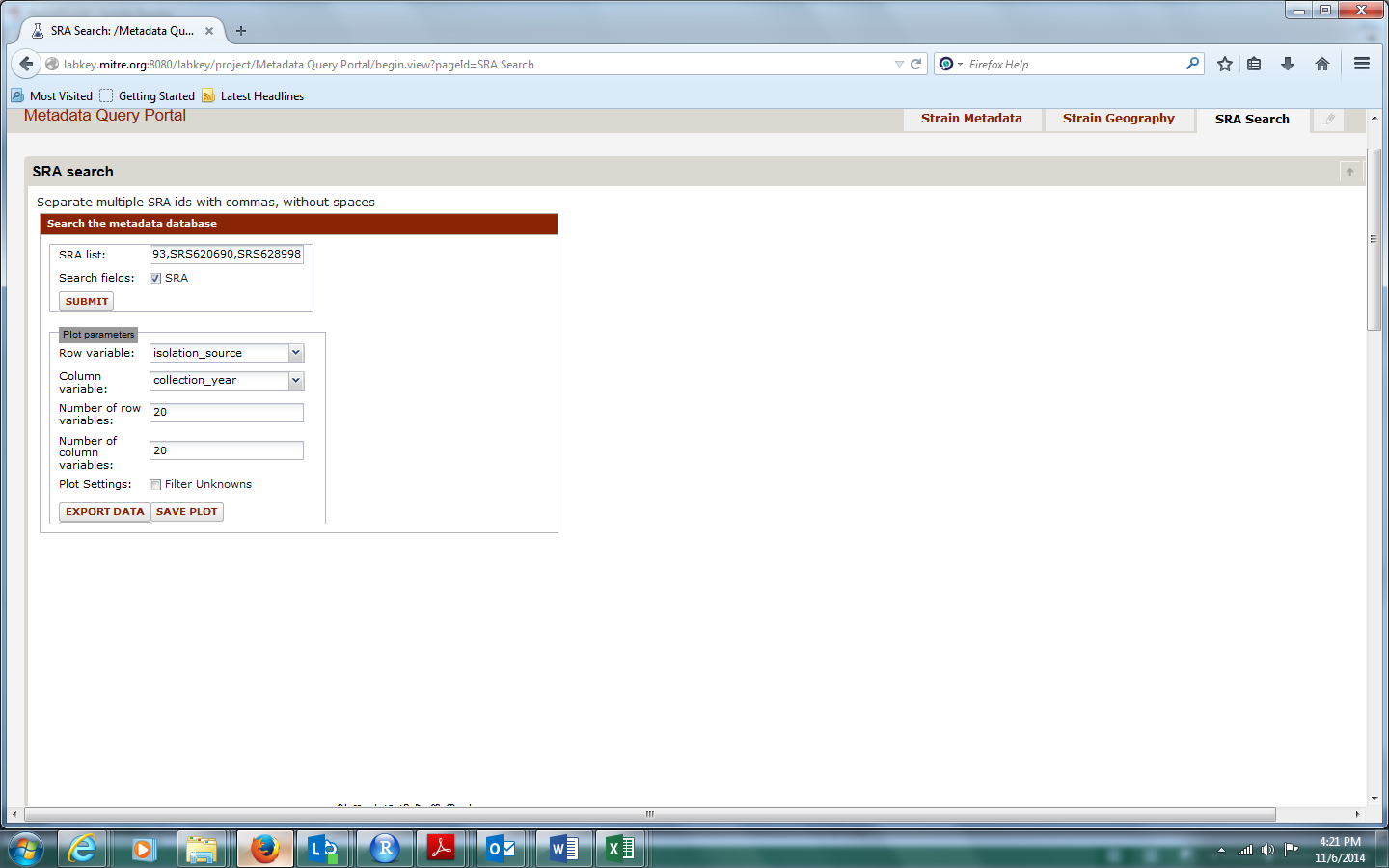


The location names will then be geocoded and the places where the strains were collected will appear on the map:

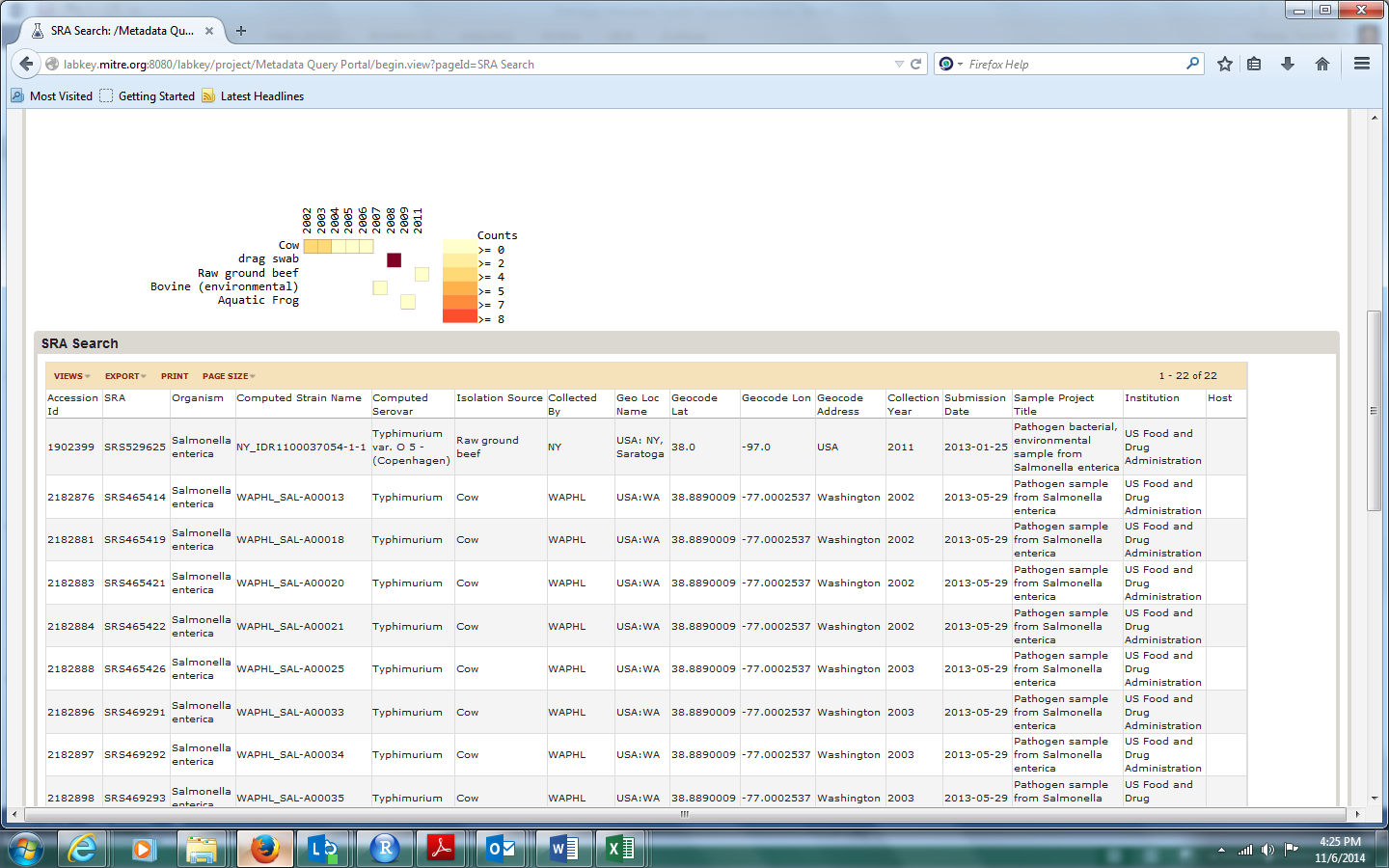
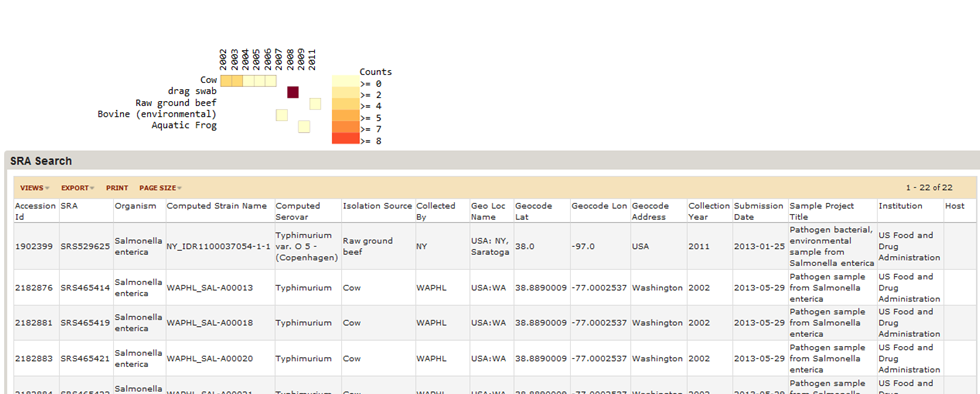


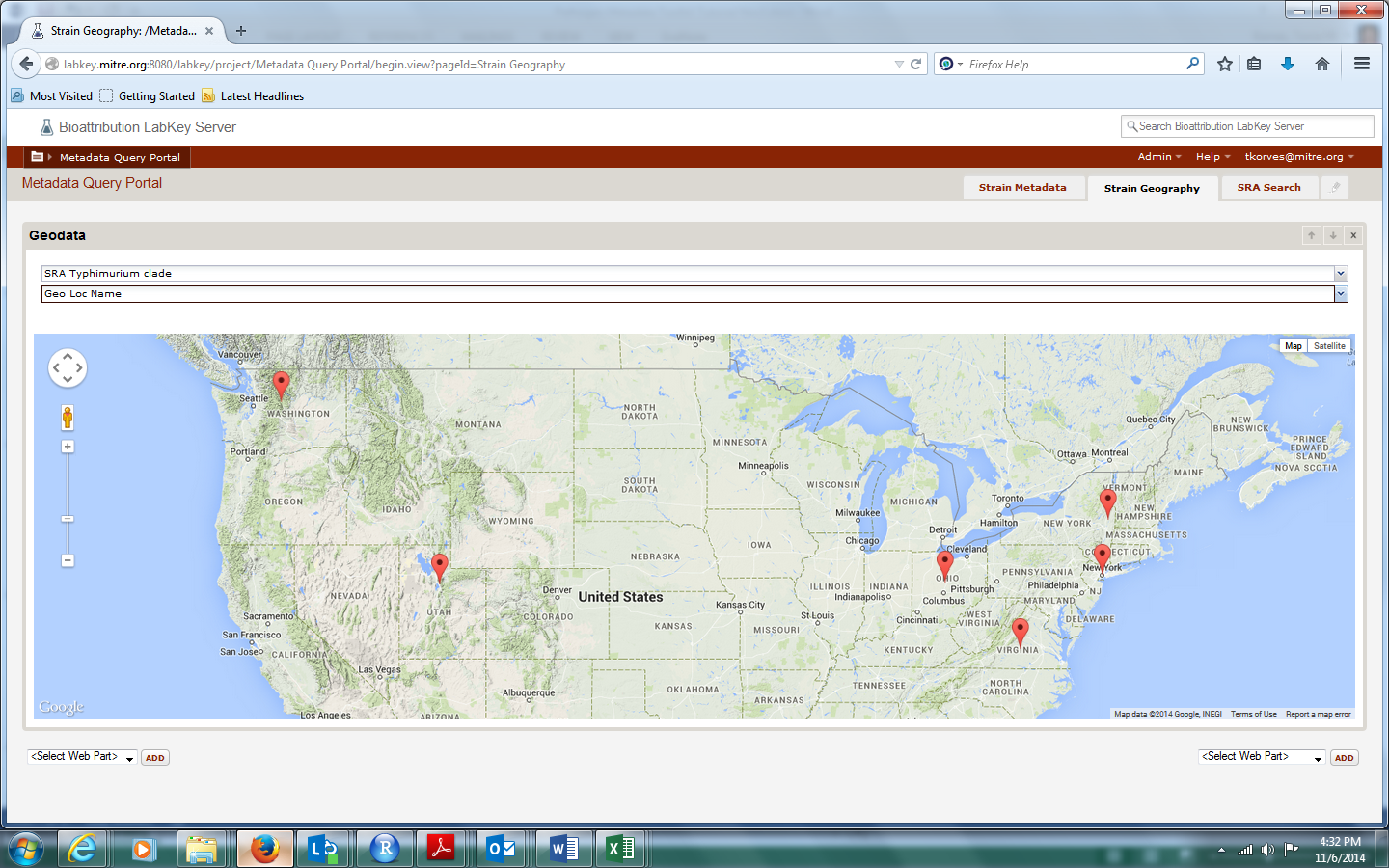
**Example 2**. To investigate where a new disease outbreak came from, investigators have sequenced a pathogen sample from a patient and compared it to other pathogen sequences. This DNA analyses identified 22 closely related strain genomes. Where did the closely related strains come from?

1. On the SRA Search page, query for the SRA ids associated with the closely related strains, and then select variables for the heat map:



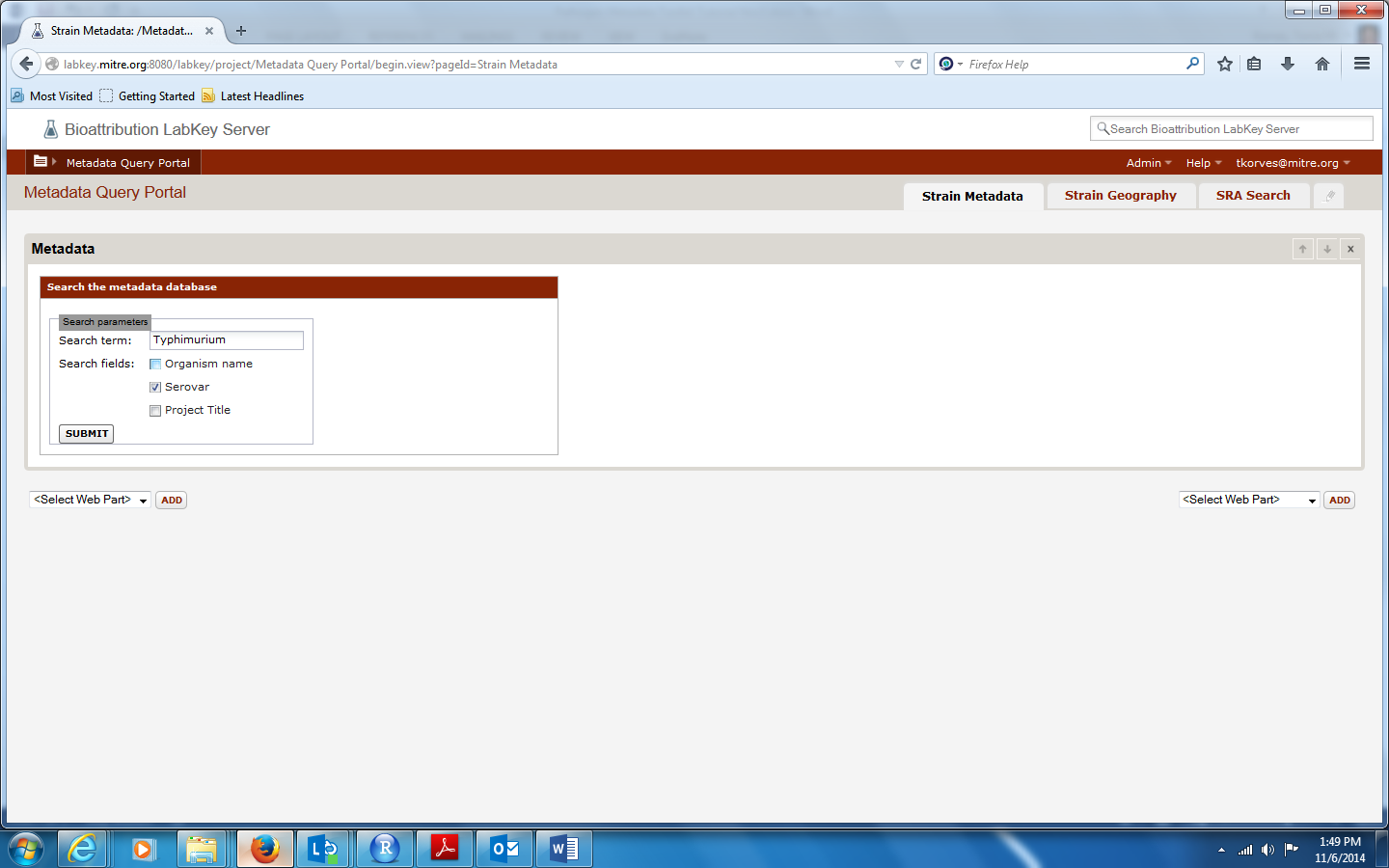
1. View results on Isolation Sources, Collection Years, and Collection Locations for insight into the origin of the new strain. (See instructions in Example 1, steps c-f.)



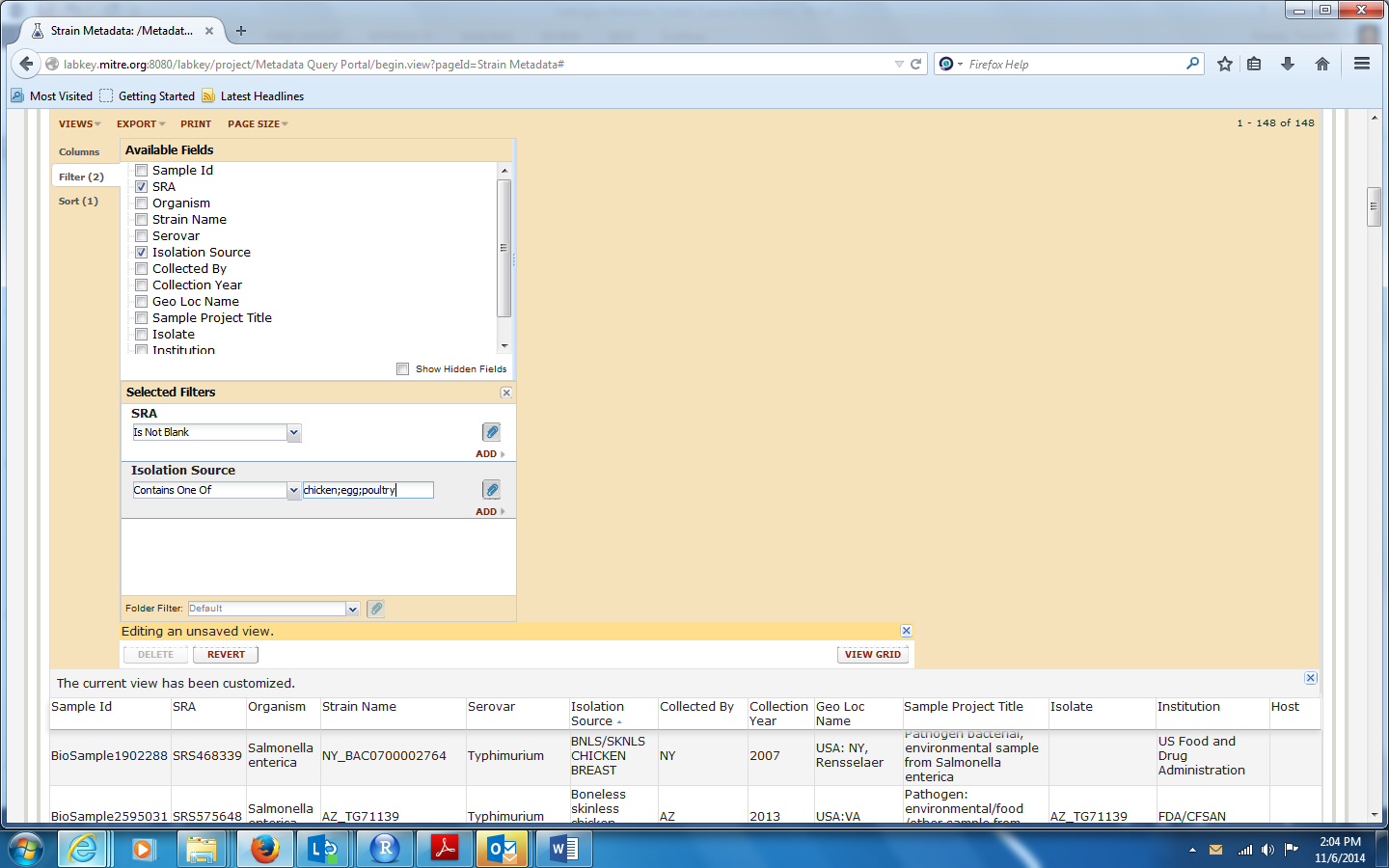


**Example 3**. A research study aims to identify pathogens genes that are targets of selection when the pathogen is in poultry. What genome sequences are available for pathogen strains that were collected from poultry for *Salmonella enterica* serovar Typhimurium?

1. Use the Strain Metadata query page to search the serovar field:



1. Filter results by Isolation Source, using “Contains one of” and multiple synonyms for chicken, separated by semicolons. Select View Grid.



1. Export the table with SRA ids and other metadata for these samples for bioinformatic analyses:

