



**CodeLinker**

**Windows**

---

## **RNA-Seq Tutorial**

© 2017 Gene Codes Corporation

Gene Codes Corporation

**T C A G E N E**  
**A G T C O D E S**

Gene Codes Corporation

525 Avis Drive, Ann Arbor, MI 48108 USA

1.800.497.4939 (USA) +1.734.769.7249 (elsewhere)

+1.734.769.7074 (fax)

[www.genecodes.com](http://www.genecodes.com) [gcinfo@genecodes.com](mailto:gcinfo@genecodes.com)

## Introduction

In this CodeLinker RNA-Seq tutorial, you are going to learn how to import differential expression data from a Cufflinks suite run and how to perform analyses on that data. You will use data from the paper by Coffelt et al.<sup>1</sup> First you will focus on learning how to prepare your data and remove any missing values. Then you will proceed through some of the key analyses that you can do. Once you have completed this Tutorial you will be able to take your own data through the same steps and even explore some of the other analyses in the rich set of tools that CodeLinker offers.

## Importing and Preparing your RNA-Seq Data

You will begin by looking at how to take differential expression output from Cuffdiff and prepare it for analysis in CodeLinker.

You will import differential expression output from Cuffdiff into CodeLinker. The “Cuffdiff exp\_diff” import template pulls the FPKM (FPKM stands for Fragments Per Kilobase of transcript per Million mapped reads) values from any of the “exp.diff” output files from Cuffdiff, the final step in the Cufflinks suite workflow. The FPKM value gives a measure of the relative abundance of transcripts. In some cases, you may have 0 values in some loci across all your data. These 0 values may be due to technical issues with your samples and need to be removed.

### Before you start

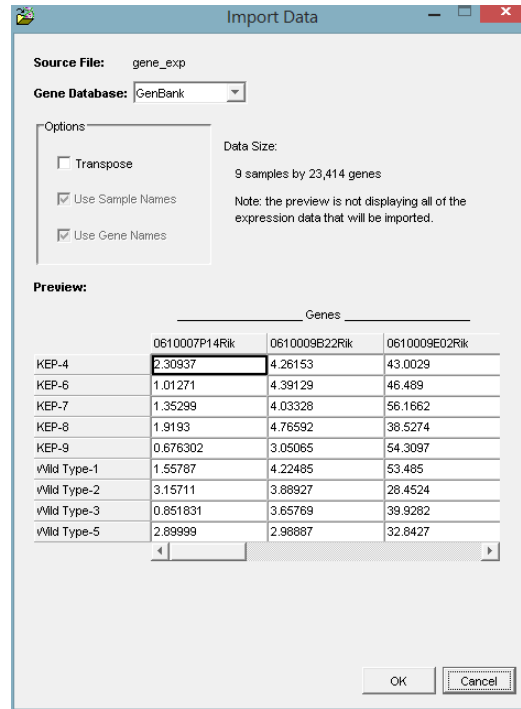
Unzip the Mouse Cancer Genome Data folder to a suitable location, you will need to use the items contained within.

### Data Import

- **Launch CodeLinker**
- **Go to File > Import > Gene Expression Data.**
- **Set the Template to Cuffdiff exp\_diff and Gene Database to GenBank.**
- **Click Select File...** and browse to the file **gene\_exp.diff** and select it. **Click Open** and then **Import**.
- **The Import Data window** (see image below) will appear. **Click OK.**

---

<sup>1</sup> S. B. Coffelt, K. Kersten, C. W. Doornebal, J. Weiden, K. Vrijland, C.-S. Hau, N. J. M. Verstegen, M. Ciampricotti, L. J. A. C. Hawinkels, J. J. K. E. de Visser, IL-17-producing  $\gamma\delta$  T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* **522**, 345–348 (2015)



## Normalizing data

Before you can normalize the data you have to remove genes that have a zero value for at least one of the samples. This is a two-step process. The first step is to remove all zero values. The second step is to remove the genes that have missing values as a result of the first step.

- Highlight “gene\_exp” and go to **Data > Remove Values**.
- In the **Remove Values** window, set **Removal Technique** to **by Expression Value**. For **Expression Value** make sure that **<=** is selected and that the value is **0** and click **OK**.
- Highlight “Removed: v <= 0.0” and go to **Data > Estimate Missing Values....**
- In the **Estimate Missing Values** window, set **Remove Genes That Have Missing Values:** to **1** and click **OK**.

Next you will normalize the data so the genes are on a similar scale.

- Highlight “Estimated: #mv < 1 | median” and go to **Data > Normalize**.
- In the **Normalization (Page 1 of 2)** window, select **Other Transformations** and select **Next**.
- In the **Normalization (Page 2 of 2)** window, select **Standardize** and select **Finish**.

## The IBIS Classifier Search

The next thing you will need to do is filter down the number of genes to a more reasonable number for CodeLinker's analyses to work with. You are going to use the **(IBIS)** Integrated Bayesian Inference System to create a gene list of the most important genes. This is an advanced data-mining algorithm designed to find patterns or variable sets that have predictive value. For our data set, the samples can be classified as either "Tumor bearing" or "Wild Type", **IBIS** will find genes that fit those criteria.

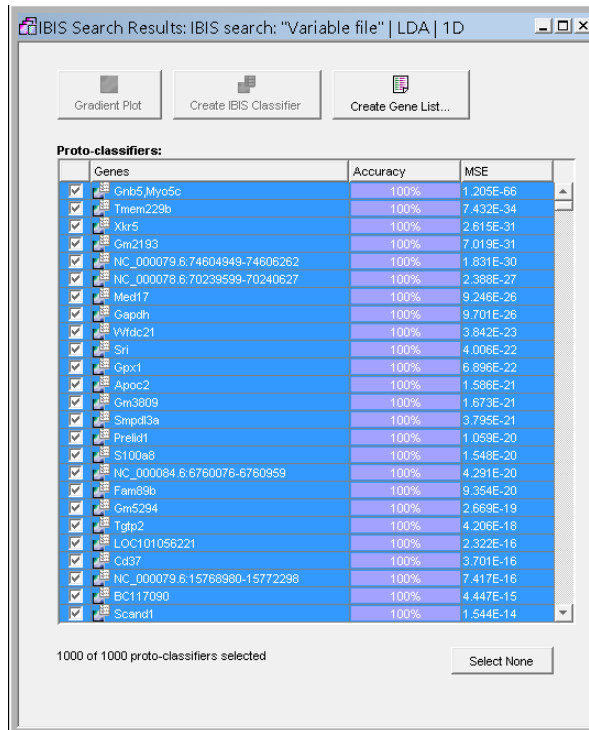
You begin by importing a variable file that classifies each of the samples into one of two categories. The variable file is just a text file with one condition for each sample *in the same order as the data file*.

- Highlight "gene\_exp" and go to **File > Import > Variable**.
- You need to import a **Source File**, click the button labeled ... and browse to the file called **Variable file.txt** and click **Open** followed by **OK**.
- Click **Preview** to confirm that the variables are in the same order as the data file. Select **Close** and then **Import**.

There will now be a small purple V on the icon for this dataset. The variables will be applied to everything downstream from that dataset.

You are now ready to use the **IBIS** Classifier Search. You have the options of using Linear Discriminate Analysis (**LDA**) or Quadratic Discriminate Analysis (**QDA**) on either a single gene or on a pair of genes. It's best to start with an LDA/single gene analysis as this will be the fastest option and will find the clearest examples. In this tutorial you are going to use LDA on a single gene.

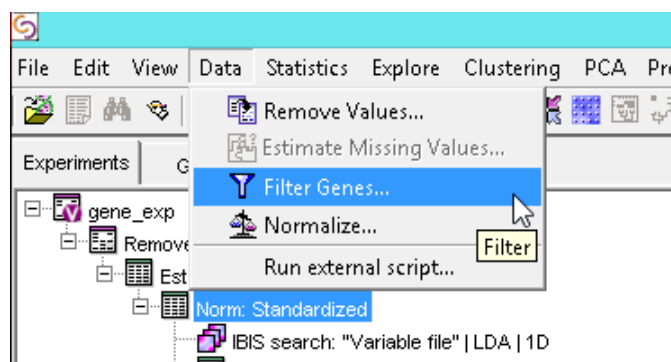
- Highlight **Norm: Standardized** and go to **Predict > IBIS Classifier Search....**
- In the **IBIS Classifier Search** window, keep the defaults and select **OK**.
- In the **IBIS Search Results** window, click anywhere in the **Proto-classifiers** pane, click **Ctrl+A** to highlight all of the genes, clicking in the top checkbox which should place a checkmark beside all genes.
- Now click **Create Gene List** button.



- In the **Create a Gene List** window, enter the name and description for your gene list and select **OK**.
- You can also highlight a single gene in the list and select **Gradient Plot**, which will show you a graphical representation of how that gene will fit the variable criteria. Finally dismiss the **IBIS Search Results** dialog.

The next step is to filter the Norm:Standardized dataset.

- Highlight **Norm:Standardized** and go to **Data > Filter**.



- In the **Filter Genes** window, set the **Filtering Operation** to **Gene List Filtering**. Select **Keep only genes that are on this list**, select the **Gene List** you just created, and select **OK**.

Using the **IBIS Classifier Search** feature, you have now condensed the dataset down to 1000 genes that best fit the experimental variables. This will help the downstream analyses be more effective.

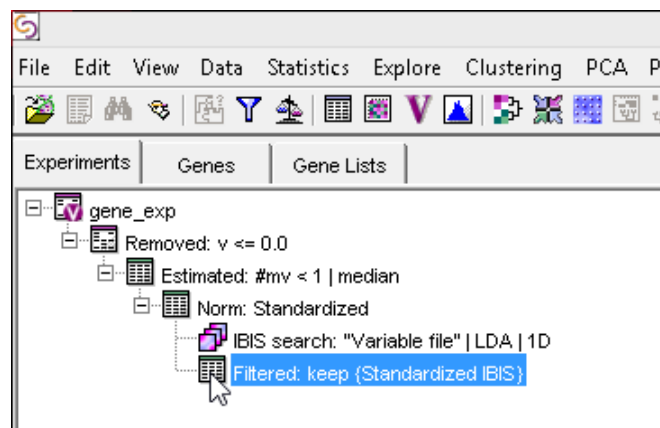
## Analysis in CodeLinker


Now you are going to briefly look at some of the analyses you can do in CodeLinker. These analyses are designed to explore and expose any relationships between the observations in your data sets. Typically, you would perform several of these analyses as part of your overall data exploration strategy.

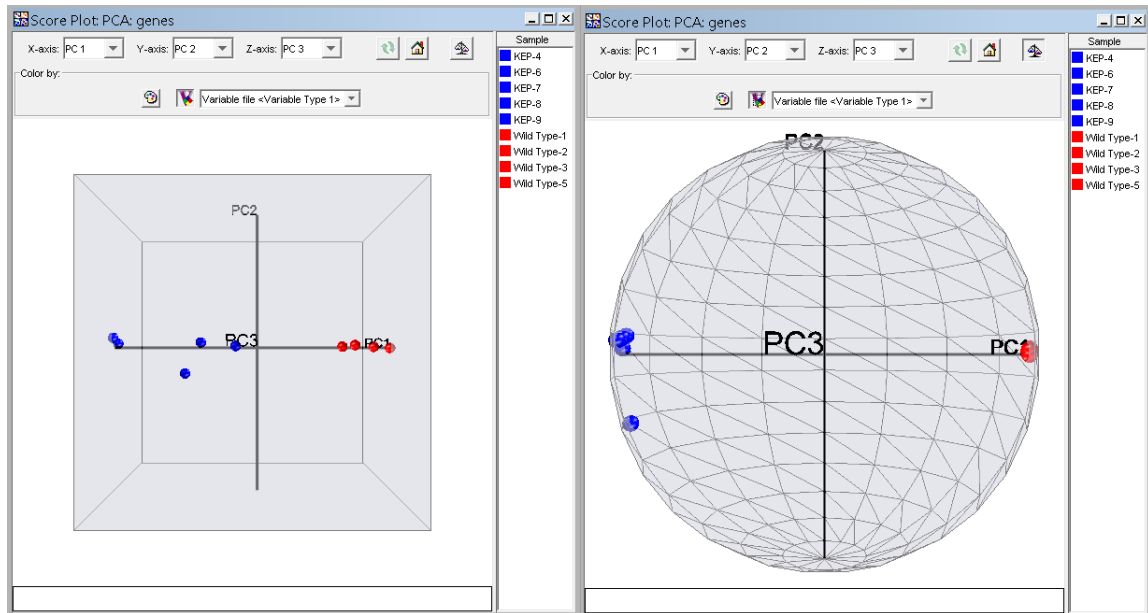
### Principle Components Analysis

This is a statistical method that converts a large set of observations into a smaller set of variables or principal components. It is typically used as one of a series of analytical steps and is useful for exposing patterns in a set of data.

- Highlight the filtered dataset.



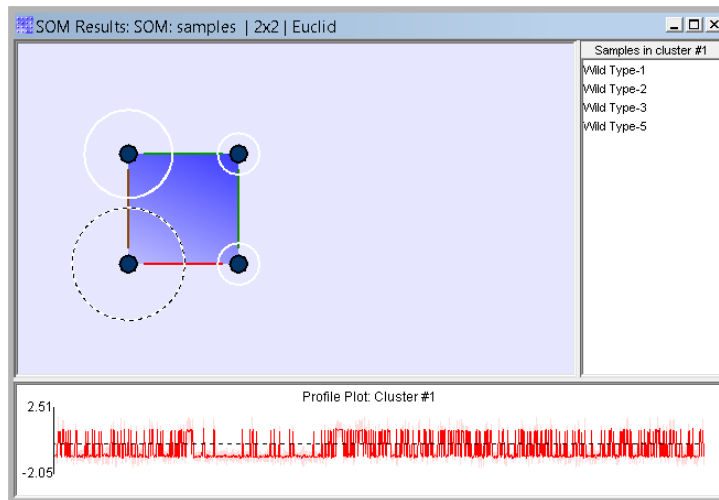
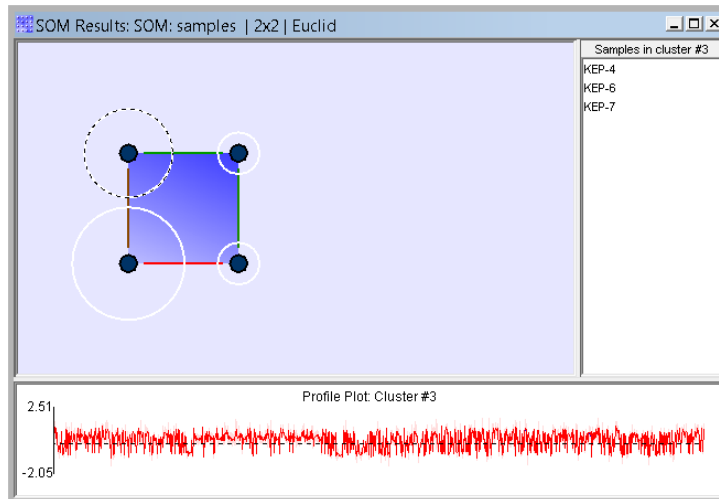
- Go to **PCA** menu and select **Principle Components Analysis....**
- In the **PCA** window, make sure that **Genes** is selected for **PCA Orientation** and select **OK**.
- Click on the  button to display the two conditions as different colors.



## Self-Organizing Maps (SOMs)

The **SOM** is a type of Clustering algorithm that is used to produce a low dimension representation of the input samples and automatically classify them. The **SOM** algorithm tries to assign the data observations to nodes, but unlike other clustering algorithms won't force data into each node. You determine the size and shape of the network. Not all of the nodes will have data clustered on them.

- Highlight the filtered dataset.
- Go to **Clustering > Self-Organizing Map....**
- In the **Self Organizing Map** window, keep all of the defaults except for **Orientation** and **Map Dimensions**. Set **Orientation** to **Samples**. For the **Map Dimensions**, you will need to decide how many nodes you want. (For this tutorial 2 x 2 is a good starting point and these are shown in the images below. For other data sets you should explore other combinations.) Once you've set those, select **OK**.
- If you click on a cluster, the pane on the right will show the samples that are in that cluster.

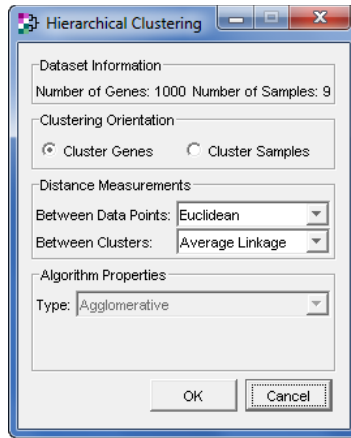


## Hierarchical Clustering

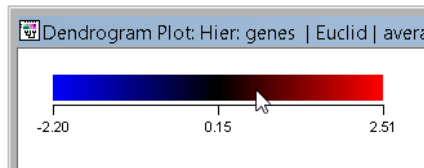
Hierarchical clustering is a method of dividing observations into groups or clusters such that the observations in any particular group are more closely related to each other than they are to observations in another cluster. The method goes through several steps to produce a dendrogram of the clusters.

- Highlight the filtered dataset.
- Go to **Clustering > Hierarchical Clustering**.
- Keep all of the settings as shown below and select **OK**.

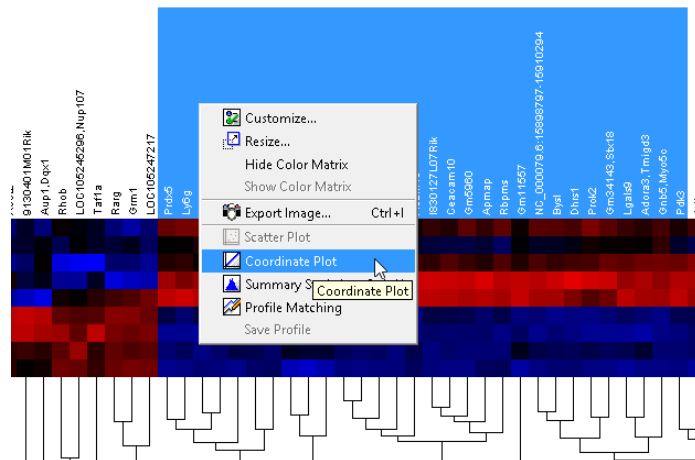




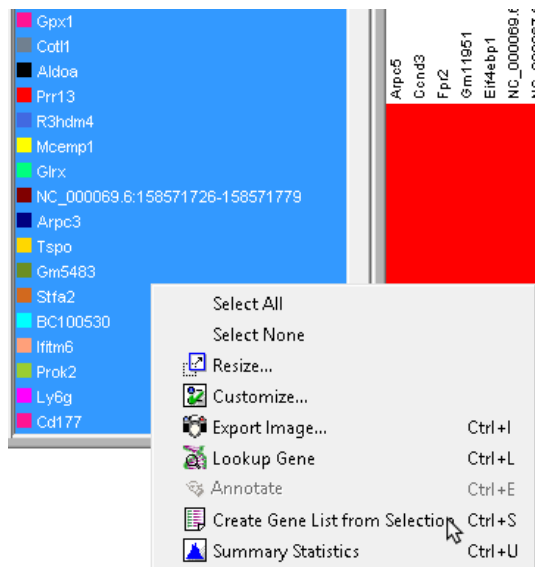
- You will need to change the **Data Range**. Double-click on the color scale as shown below.



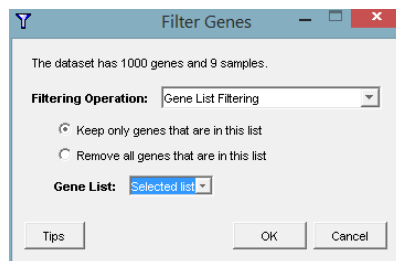
- Set the **Minimum** to **-2** and the **Maximum** to **2** and close the window. (You can also change the color palette in this menu if you so desire.)
- You are going to look at some of the genes in closer detail. Go to the **Edit** menu and select **Find**. Type **Prdx5** into the **Find what** text entry box. Select **Find**.
- Highlight 30 genes to the right-hand of **Prdx5**, right click, and select **Coordinate Plot** as shown below.




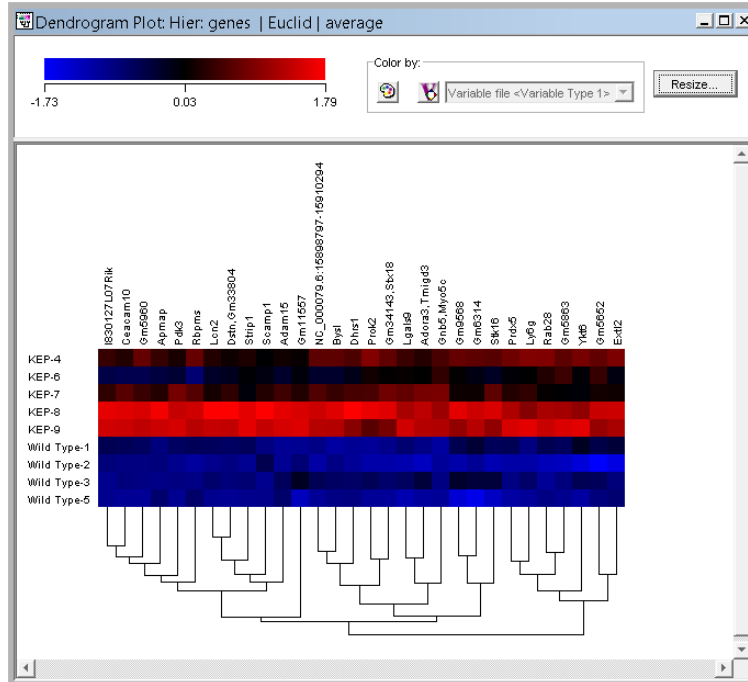
- Highlight all the genes in the **Gene** pane, right-click, and select **Create Gene List from Selection**.



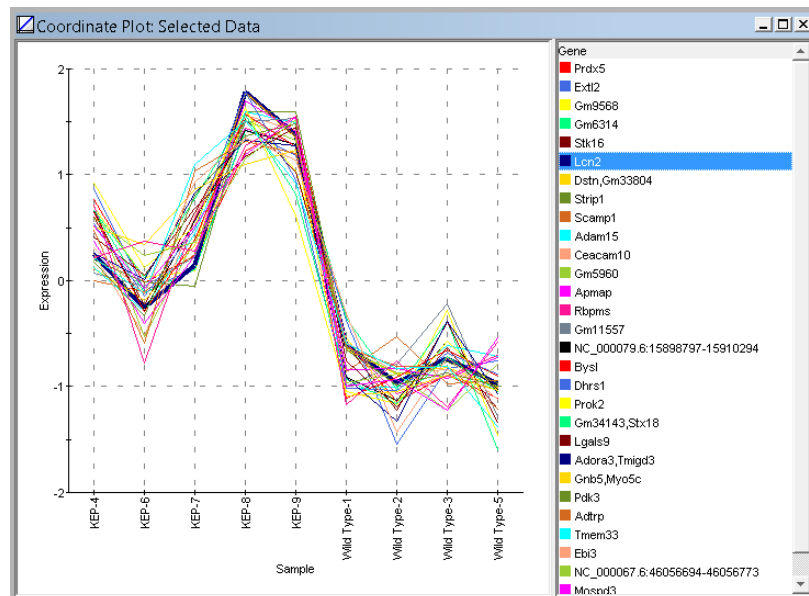
- Choose a name for the list and select **OK**. Close the **Coordinate Plot**.
- Highlight the filtered dataset and go to **Data > Filter Genes....**
- Select **Gene List Filtering** from the **Filtering Operation** drop-down list. Make sure that the gene list you just created is selected in the **Gene List** drop-down list. Click the radio button **Keep only genes that are in this list**, and select **OK**.



- Repeat the steps to create a **Hierarchical Clustering** on the new filtered list you just created.
- Now you can see the dendrogram more clearly.
- You will need to change the **Data Range**. Double-click on the color scale and click **Use actual range**. Close the window.
- Click on the  button, this will show you which rows correspond to which samples.
- Highlight all of the genes in the **Hierarchical Plot**.
- Right-click and select **Coordinate Plot**.



What you see in this plot is the expression profile for each of these genes, in your cell lines. If you click on a specific gene in the **Gene** list, its plot line is emphasized and the variation in expression of that gene across all the cell lines is clearly displayed. In this case we can see the elevation of the expression curve in the cancer lines and a concomitant depression in the Wild Type samples.



## Searching GenBank

Once you have a list of interesting genes, you will want to find out more about them. CodeLinker has a preference which allows you to link to GenBank and search the Nucleotide database using just a gene name.

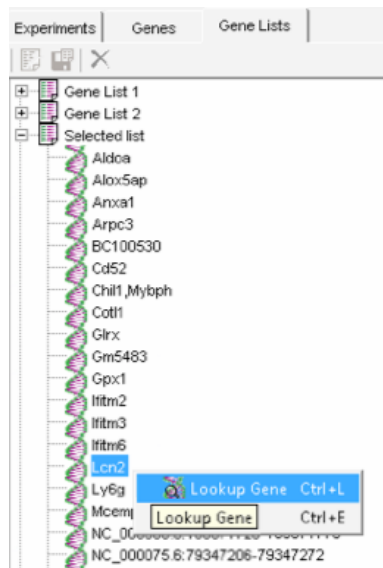
- Go to the **Tools** menu and select **Preferences**.
- Click on the **Gene Database** tab.
- Ensure that the **Gene Display Name** is set to **Gene Name**.
- Copy and paste the following text into the **GenBank** text box :  
[https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&term="MMC\\_ID"&db=Nucleotide&doptcmdl=GenBank](https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&term=)
- If you cannot copy and paste the URL, you may need to enter it manually.
- Click **OK**.

You can now test this as follows:

- In the **Coordinate Plot**, right-click on the gene called **Lcn2**. Select **Lookup Gene**. Your default browser will open at the GenBank entry for **Lcn2**.

You can also perform a **Lookup Gene** search from any **Gene List**.

- Go to the **Gene Lists** tab.
- Click on the plus symbol (+) next to your chosen **Gene List**.



- Pick a gene name and right click on it. Select **Lookup Gene**. Your default browser will open showing a display similar to the one shown below.

The screenshot shows the NCBI Nucleotide search interface. At the top, there's a search bar with 'Lcn2' entered and a 'Search' button. Below the search bar, there are options for 'Create alert' and 'Advanced'. A notification banner states: 'NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! Read more...'. On the left side, there are navigation menus for 'Species' (Animals 361, Bacteria 29), 'Molecule types' (genomic DNA/RNA 251, mRNA 205), 'Source databases' (INSDC (GenBank) 174, RefSeq 323), and 'Genetic compartments' (Plasmid 2). The main content area shows search results for 'Lcn2'. It includes a summary box: 'See LCN2 lipocalin 2 in the Gene database' and 'Lcn2 reference sequences Transcript (1) Protein (1)'. Below this, it says 'Items: 1 to 20 of 499'. A summary line indicates 'Found 505 nucleotide sequences. Nucleotide (499) EST (3) GSS (3)'. The first result is a checkbox followed by a link: 'Homo sapiens full open reading frame cDNA clone RZPD0834D0537D for gene LCN2, lipocalin 2 (oncogene 24p3); complete cds, without stopcodon'. Below the link, it specifies '594 bp linear mRNA' and provides 'Accession: CR542092.1 GI: 49457136'. At the bottom of the result, there are links for 'GenBank', 'FASTA', and 'Graphics'.

## Conclusion

This has been a sampling of just some of the vast array of tools available in CodeLinker Miner and how they relate to RNA-Seq data. You could repeat this tutorial using the isoform\_exp.diff data file and see if you get similar results or even experiment with some of the settings.

Please continue exploring CodeLinker's features with both RNA-Seq data and Microarray data. CodeLinker has an in-depth **Help** utility with information on all CodeLinker's features, and even more tutorials.

If you have any questions about features, installation, or want help using the program, please let us know at [support@genecodes.com](mailto:support@genecodes.com) or call 734-769-7249.