PAPST User's Guide and Tutorial

Outline:

Introduction

This walkthrough will provide step-by-step instructions for utilizing the main feature of PAPST to perform both routine and novel analysis of Next Generation Sequencing (NGS) data. After working a few examples you will feel confident using PAPST and applying it to your own data. This tutorial will use NCBI GEO data from Mikkelsen et al. (2007) and Chen et al. (Chen et al., 2008) which cover chromatin histone modifications and transcription factor (TF) binding respectively. This tutorial starts with significant peaks which have been called using MACS. The datasets are provided.

A note on Peaks: PAPST was originally designed to work with ChIP-seq peaks. We quickly realized that it was useful beyond peaks derived from this one experiment type. Other types of features such as TF binding motifs, CpG islands, bed tracks form UCSC or ENCODE could be used. Though the word 'peak' is used, PAPST can be applied to any arbitrary genomic region.

Logging: PAPST creates logs in a folder in the directory where PAPST.jar is opened. Each time PAPST is opened it will write a file representing the commands that have been called. This is to facilitate debugging.

System Requirements: Java 1.7 or Higher

Download PAPST: Available at https://github.com/paulbible/papst

Open PAPST: Double click the PAPST.jar file.

Citing: If you use PAPST in your research please cite:

Paul W Bible, Yuka Kanno, Lai Wei, Stephen R Brooks, John J O'Shea, Maria Morasso, Rasiah Logananthara j, Hong-Wei Sun. **PAPST, a User Friendly and Powerful Java Platform for ChIP-Seq Peak Co-Localization Analysis and Beyond.** [Journal]

Getting Started

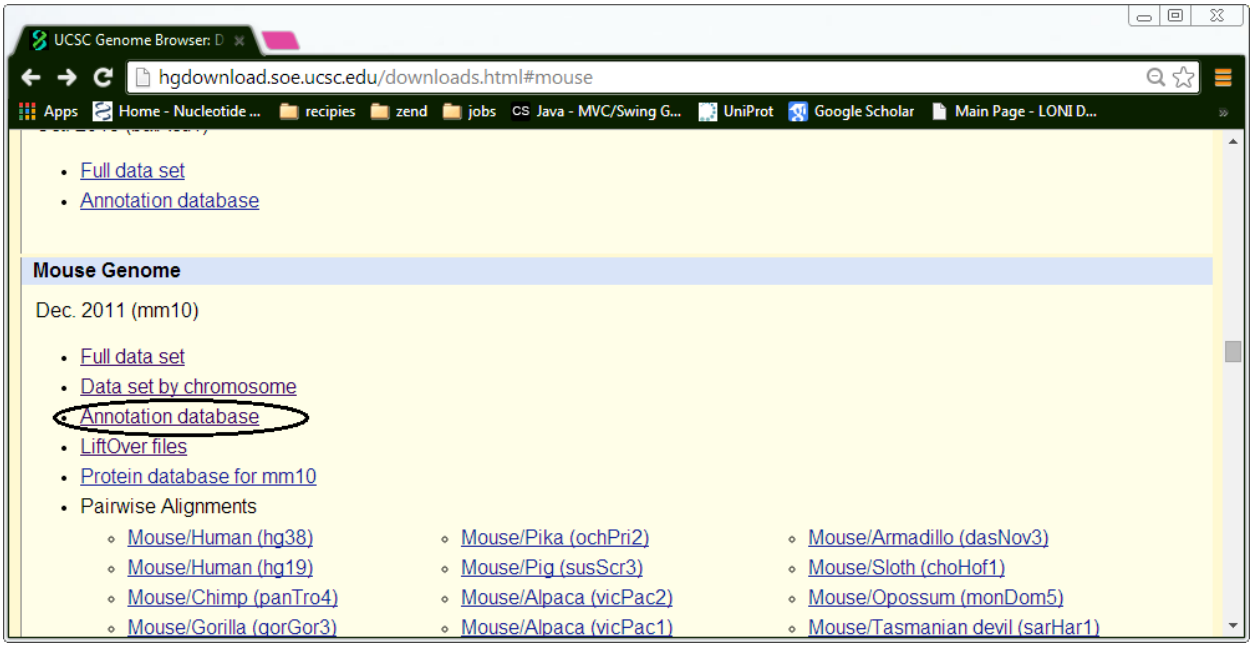
Getting a Genome File and Loading It into PAPST

Before starting with the analysis, we need to select the appropriate genome. You should use the genome that matches the genome build (hg19, mm10, etc.) to which your sequences were aligned. The provided sample data was aligned to the mm10 genome.

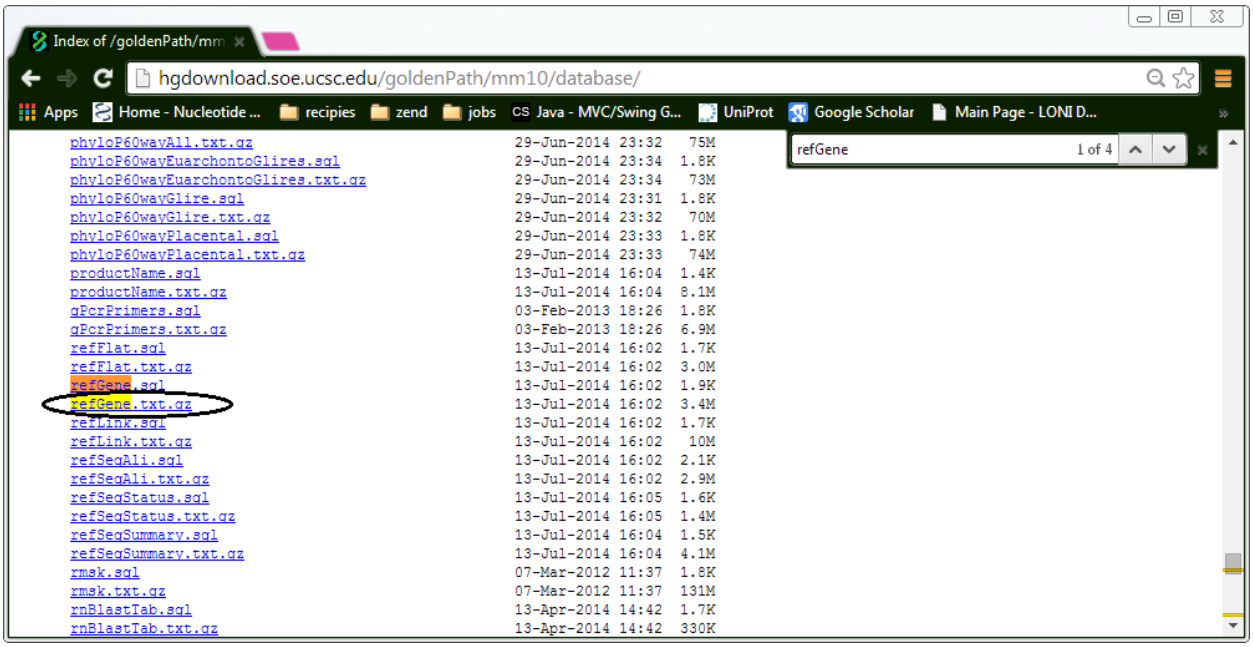
Get the refGene.txt file form UCSC Genome Browser

1. Visit the UCSB Genome Browser site: http://genome.ucsc.edu/
2. Follow the **Downloads** link

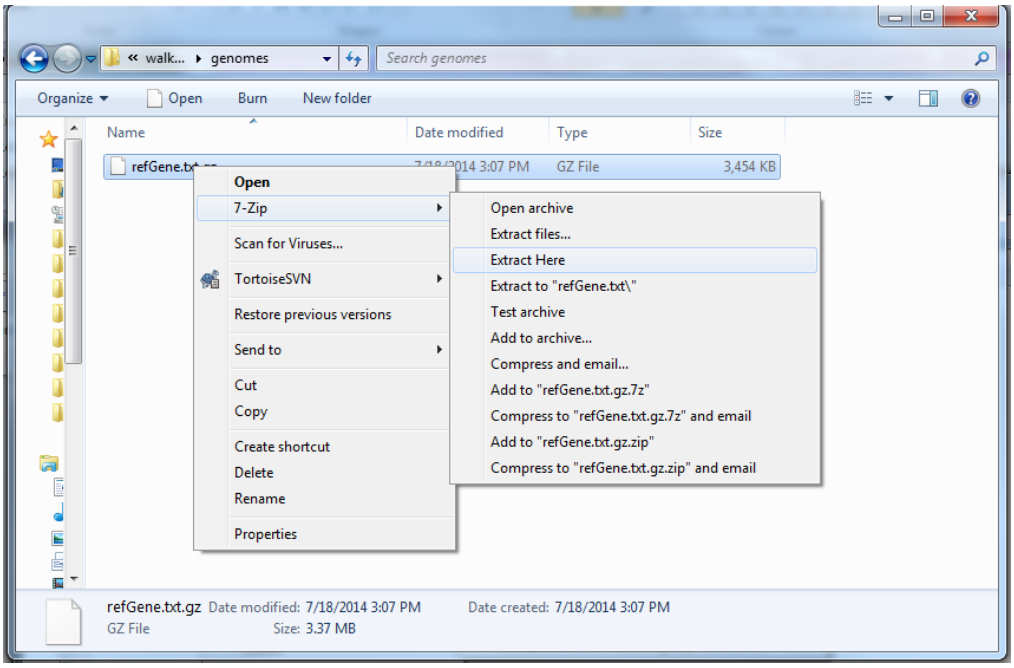
3. Click on the **Annotation Database** for the desired genome build.



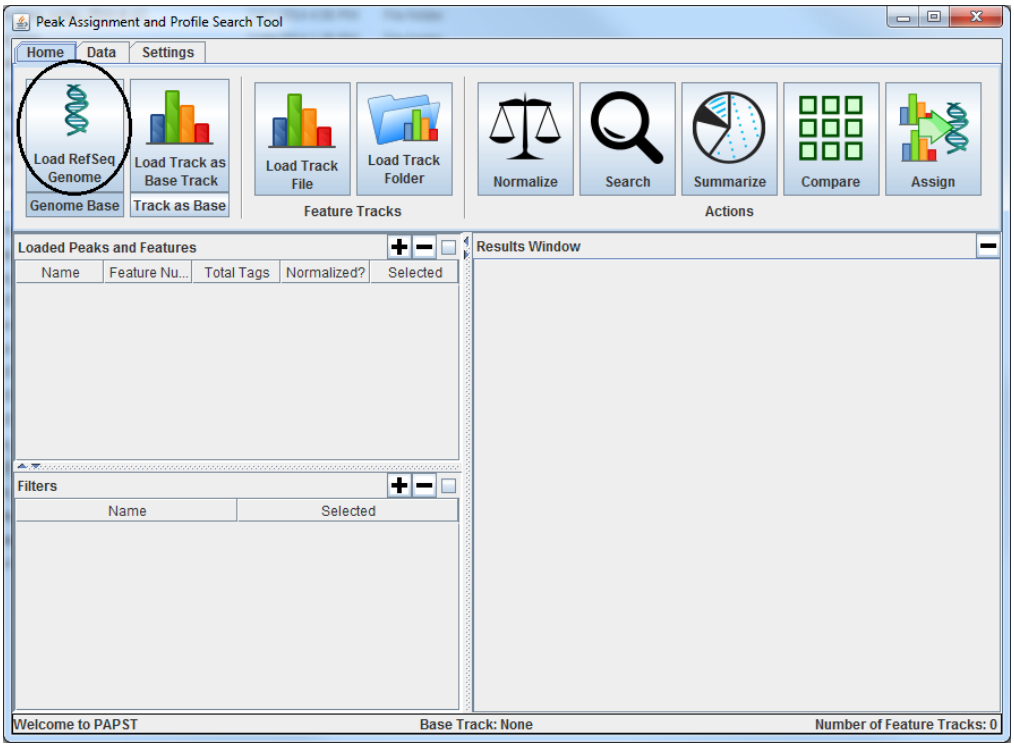
4. Download the **refGene.txt.gz** file.



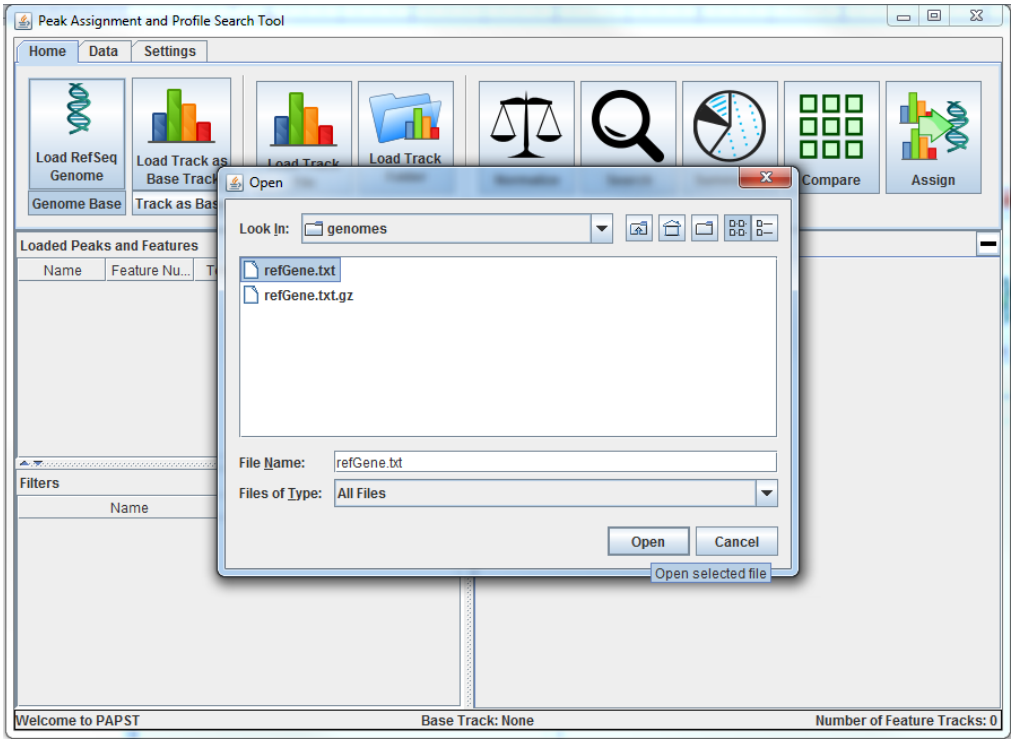
5. Extract the File using an extraction tool. 7-zip is good option. On Mac, simply double click the .gz file will extract it to the current folder.



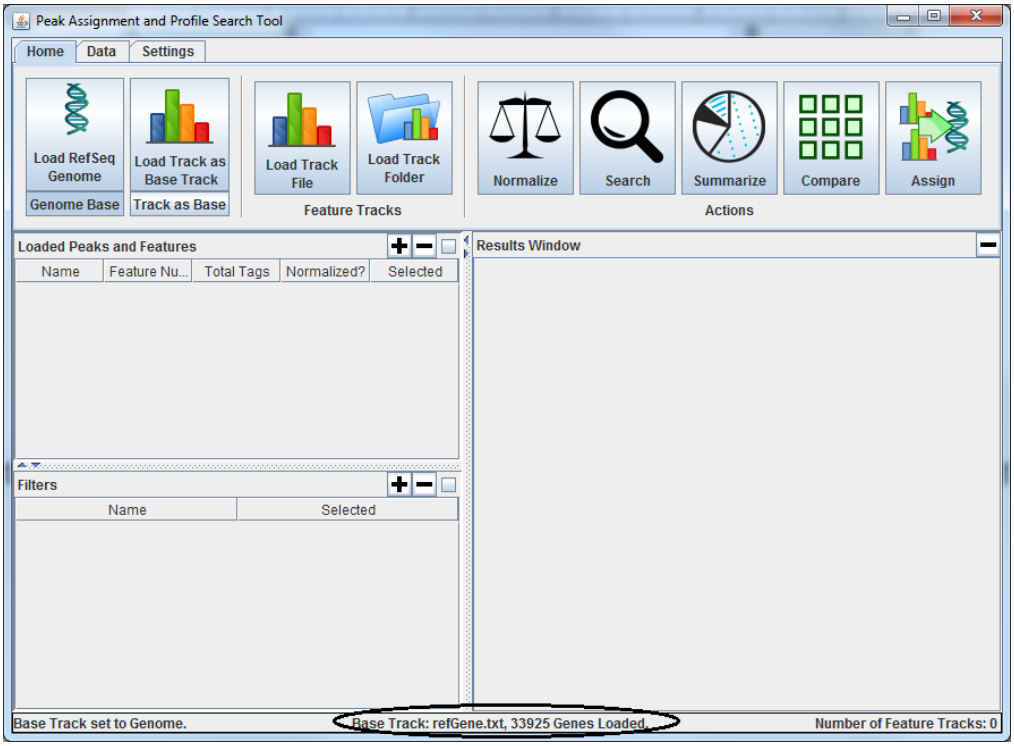
6. Load the genome into **PAPST**
    a. Open PAPST and click **Load RefSeq Genome**.



b. Select the refGene.txt



c. The genome is now loaded into PAPST.



Loading Peak Track Files into PAPST

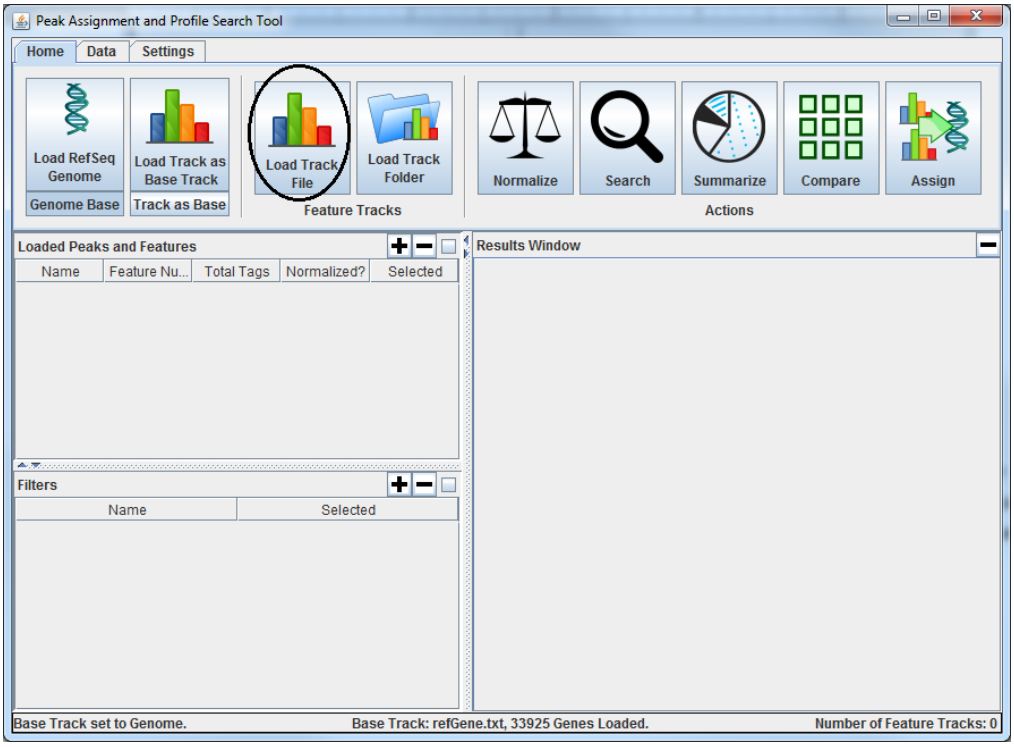
Peak files represent significant enrichment of protein binding from ChIP-seq experiments. They are typically represented as 4 column files. This section will show you how to load peak files and save parsers for later use. Tracks will refer to peaks or other genomic features.

Format of a simple 4 column ChIP-seq peak file ('bed'-like):

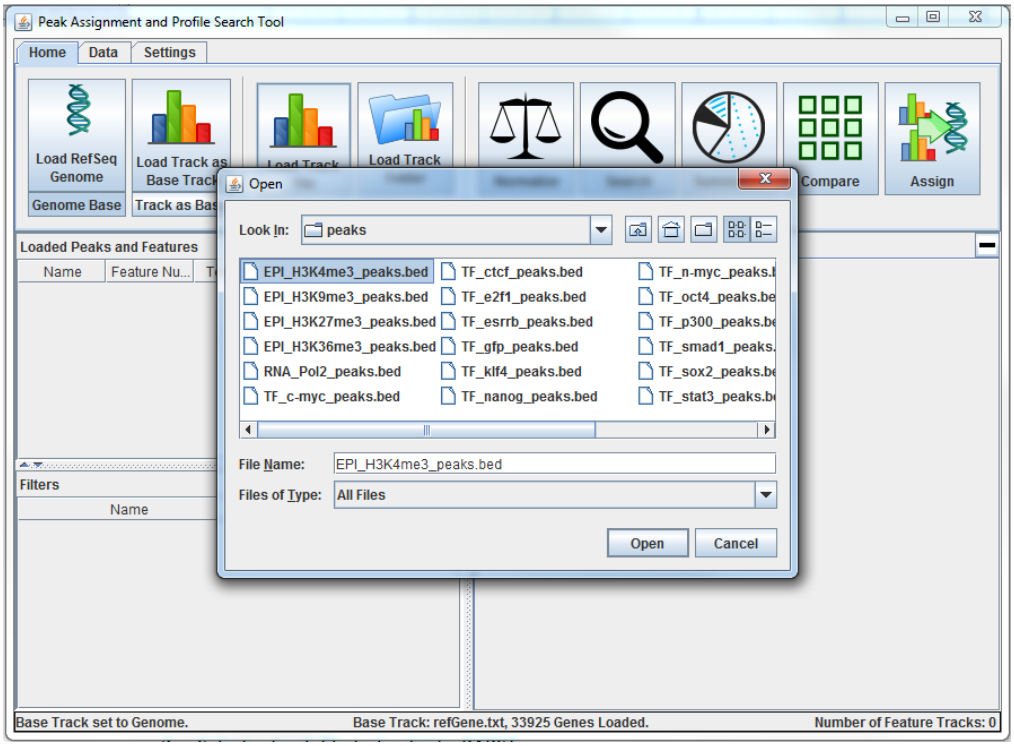
```
chr1  3670348    3672714    583.44
chr1  4491438    4493852    528.36
chr1  4571186    4572331    543.51
chr1  4784782    4786424    1435.29
…
```

Four column, tab delimited, files are the default for PAPST.

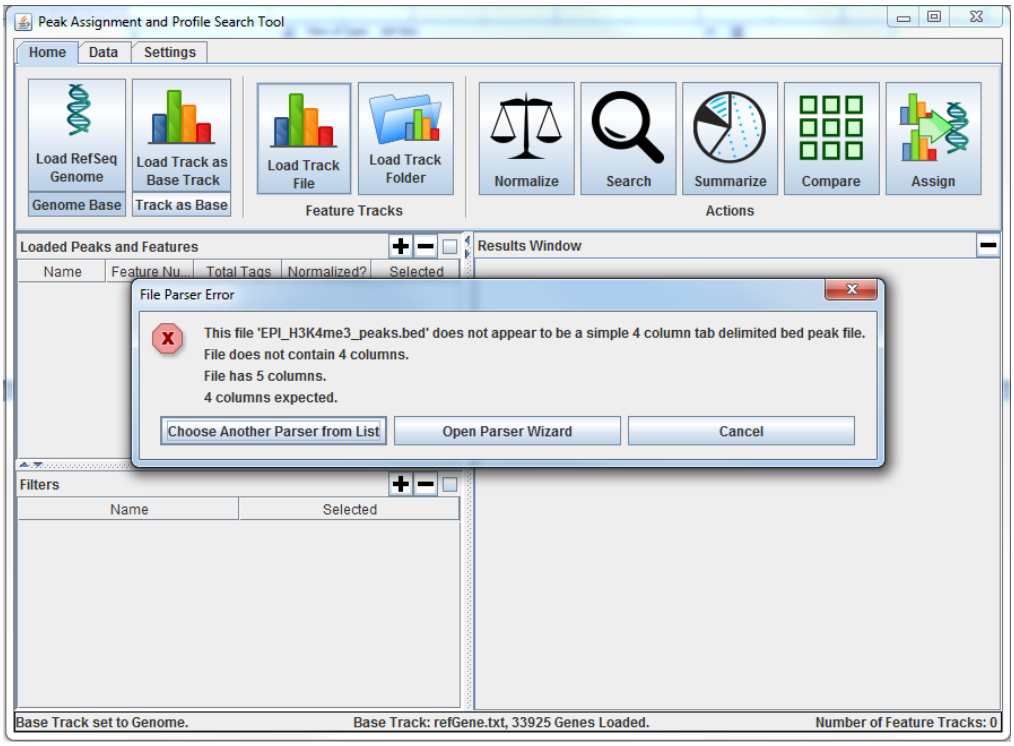
1. Click **Load Track File.**



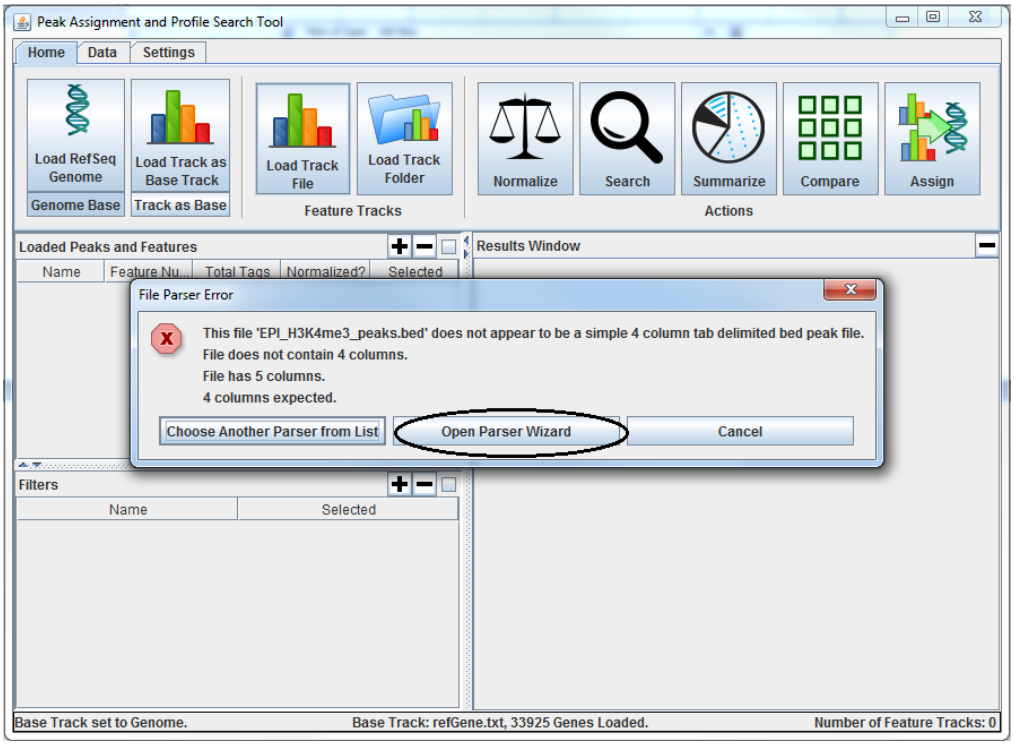
2. Select a track file to load into PAPST



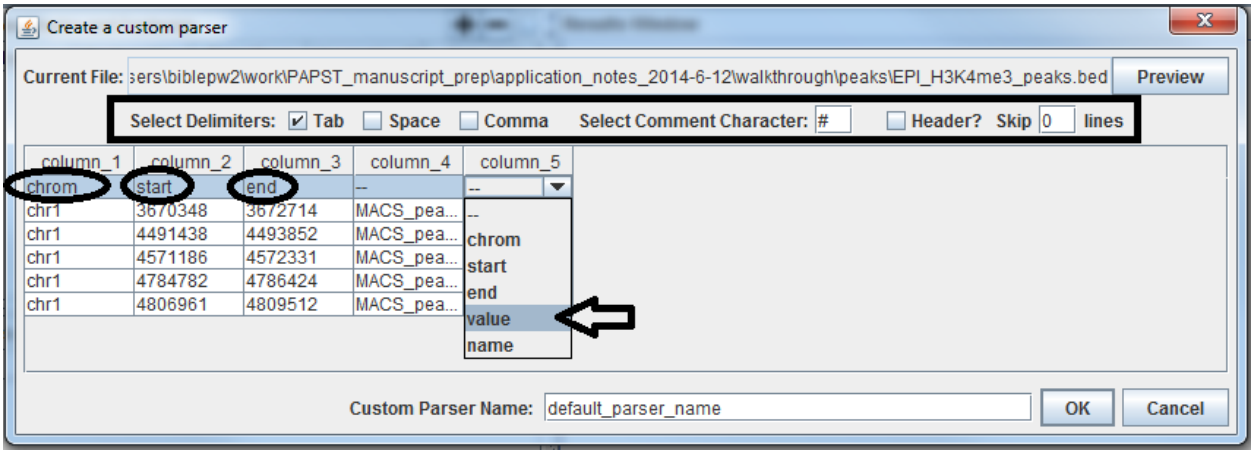
3. If a file does not match the current format, an error message will appear.



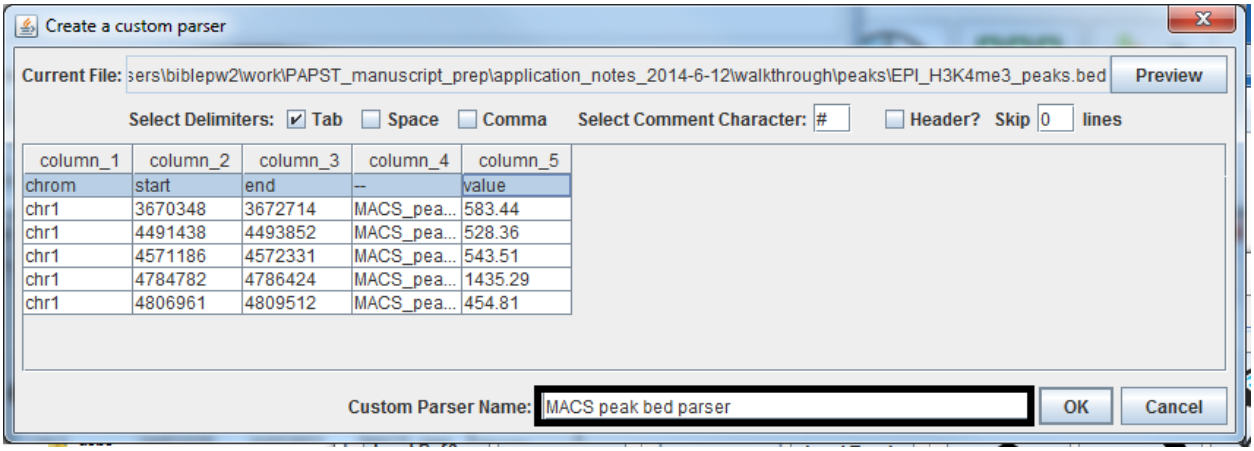
4. To create a new parser for your custom format, click **Open Parser Wizard**



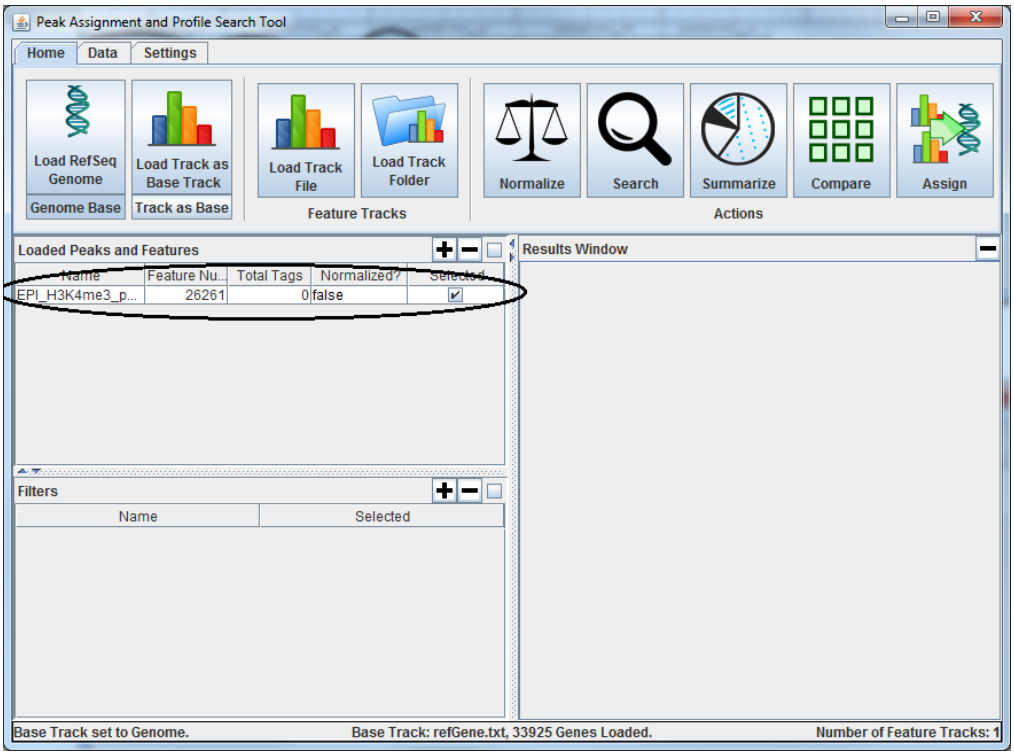
5. To create a custom parser, select the **delimiter** and set the 4 columns **chromosome**, **start**, **end**, and **value.**



6. Give the parser a descriptive name.



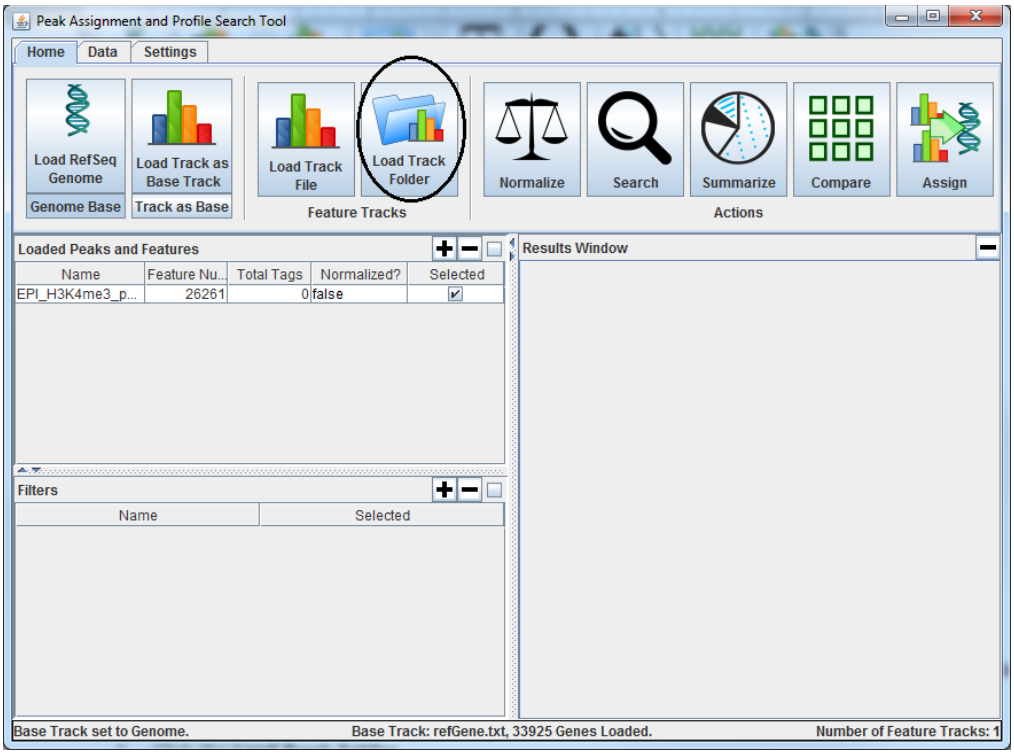
7. Your peak tracks are displayed in the **Loaded Peaks and Features** pane.



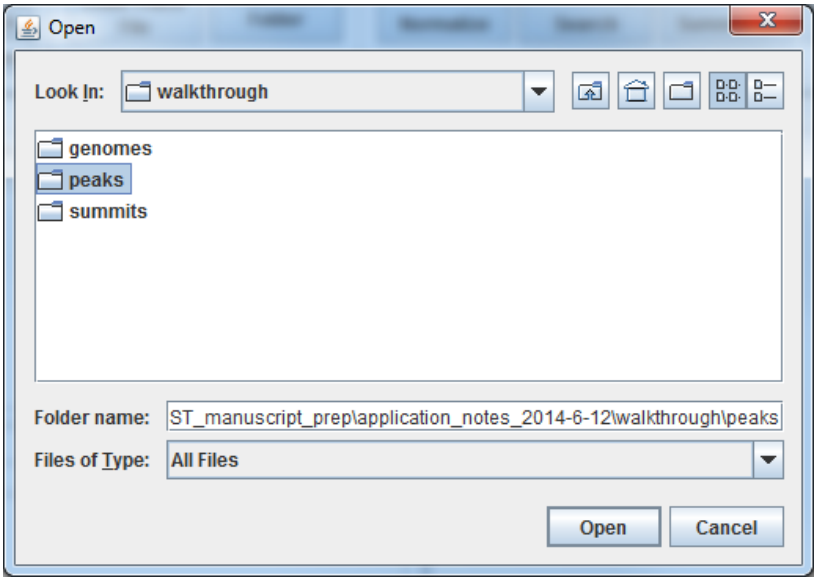
Loading a Folder of Features at Once

Often we will need to work with many files at once. PAPST allows you to load all the track files in a folder at once.

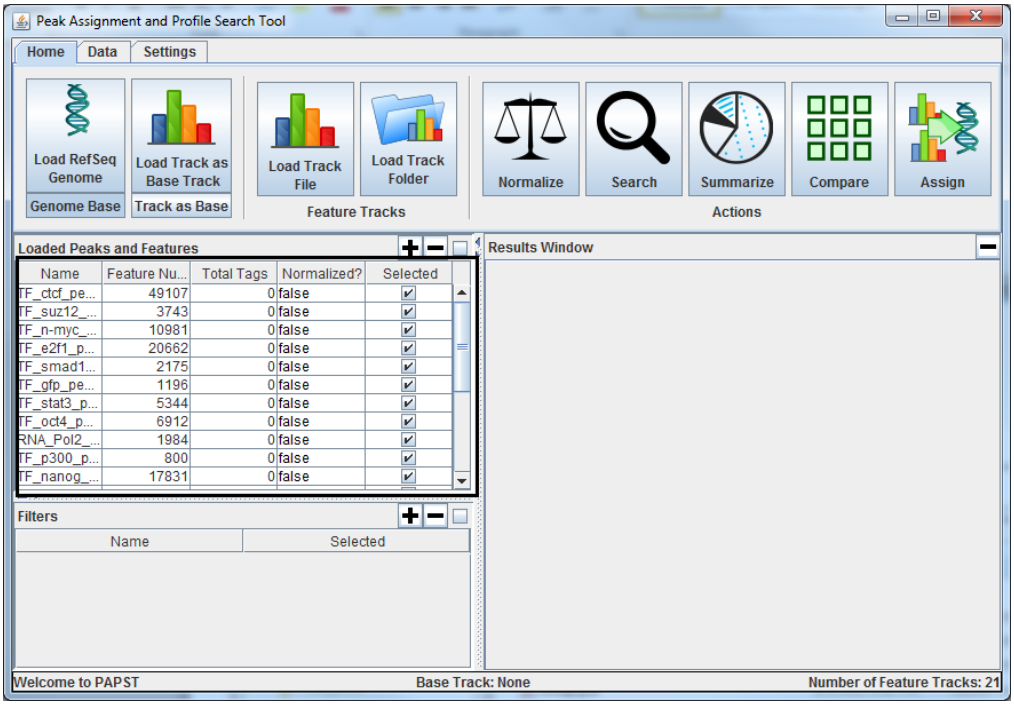
1. Click the **Load Track Folder.**



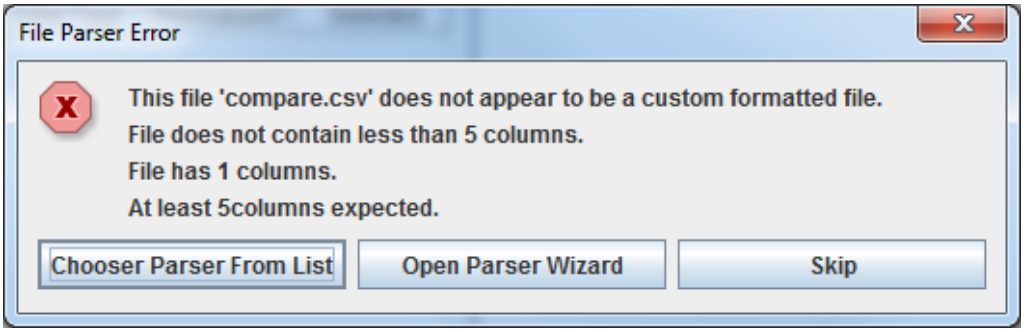
2. Select a folder to load.



3. PAPST will use the most recent parser by default. ‘MACS peak bed parser’ in our case. You will see the set of files loaded in the **Loaded Peaks and Features** pane.



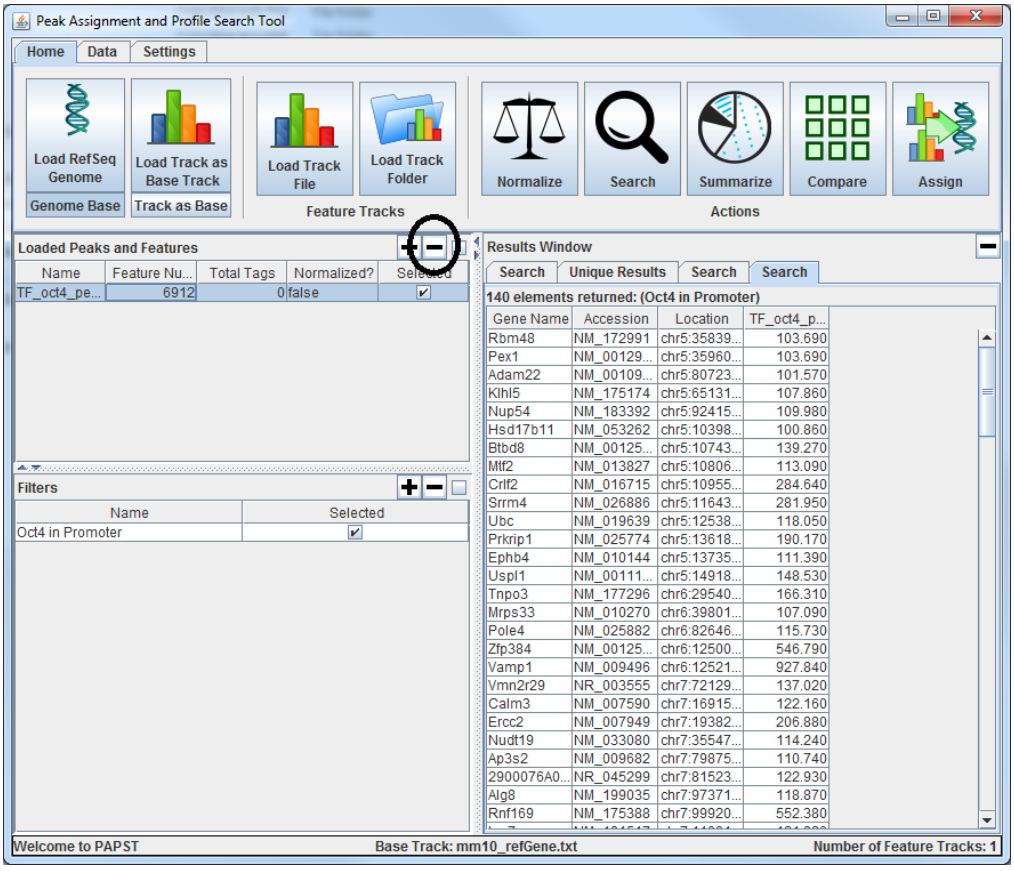
4. If you are using different formats or there are non-text files in your folder, you may encounter an error. You may choose to **use another parser,** (by clicking **Choose Parser From List**) or simply **skip** the file.



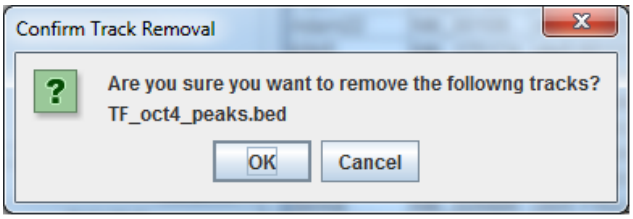
Remove a Peak Set or Feature

Removing peaks or other features in PAPST is simple. Just click the **minus**, **'-'** button in the **Loaded Peaks and Features** pane.

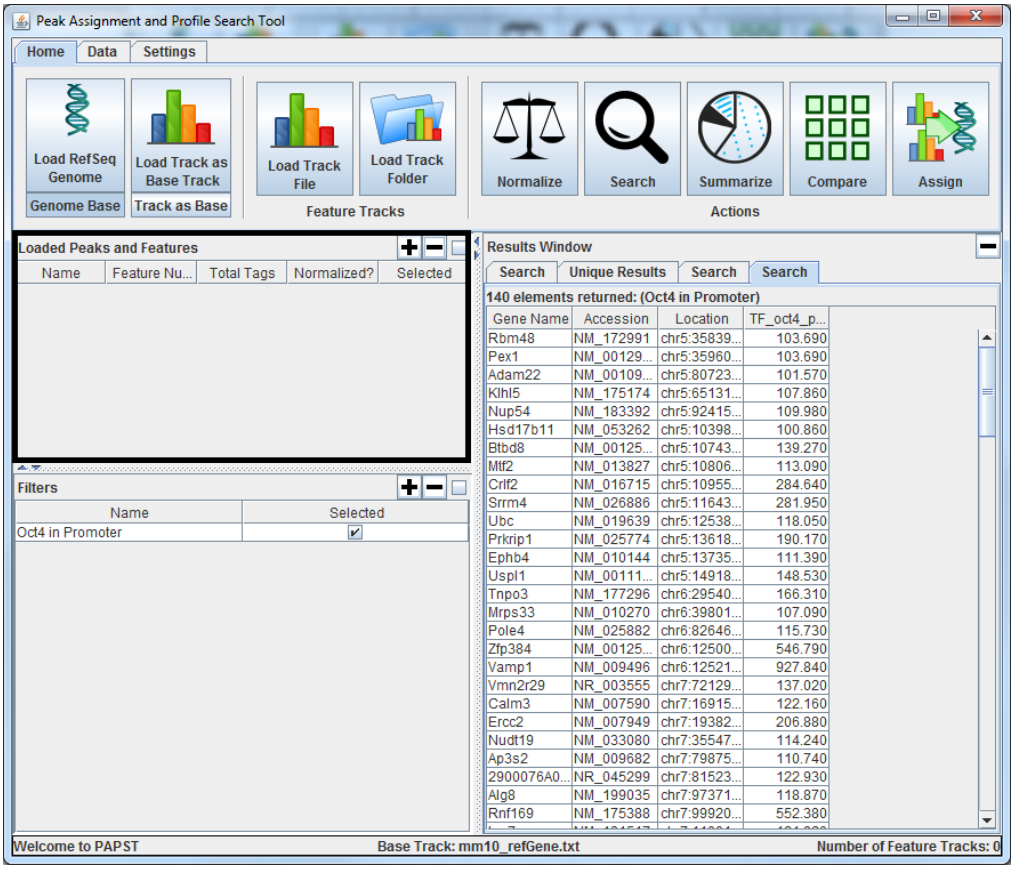
1. Select the peak set you wish to remove. For this example we will select **TF_oct4_peaks.bed**.
2. Click the **minus**, **'-'** button in the upper right corner of the **Loaded Peaks and Features** pane.



3. You will be prompted to confirm removing the peak set.



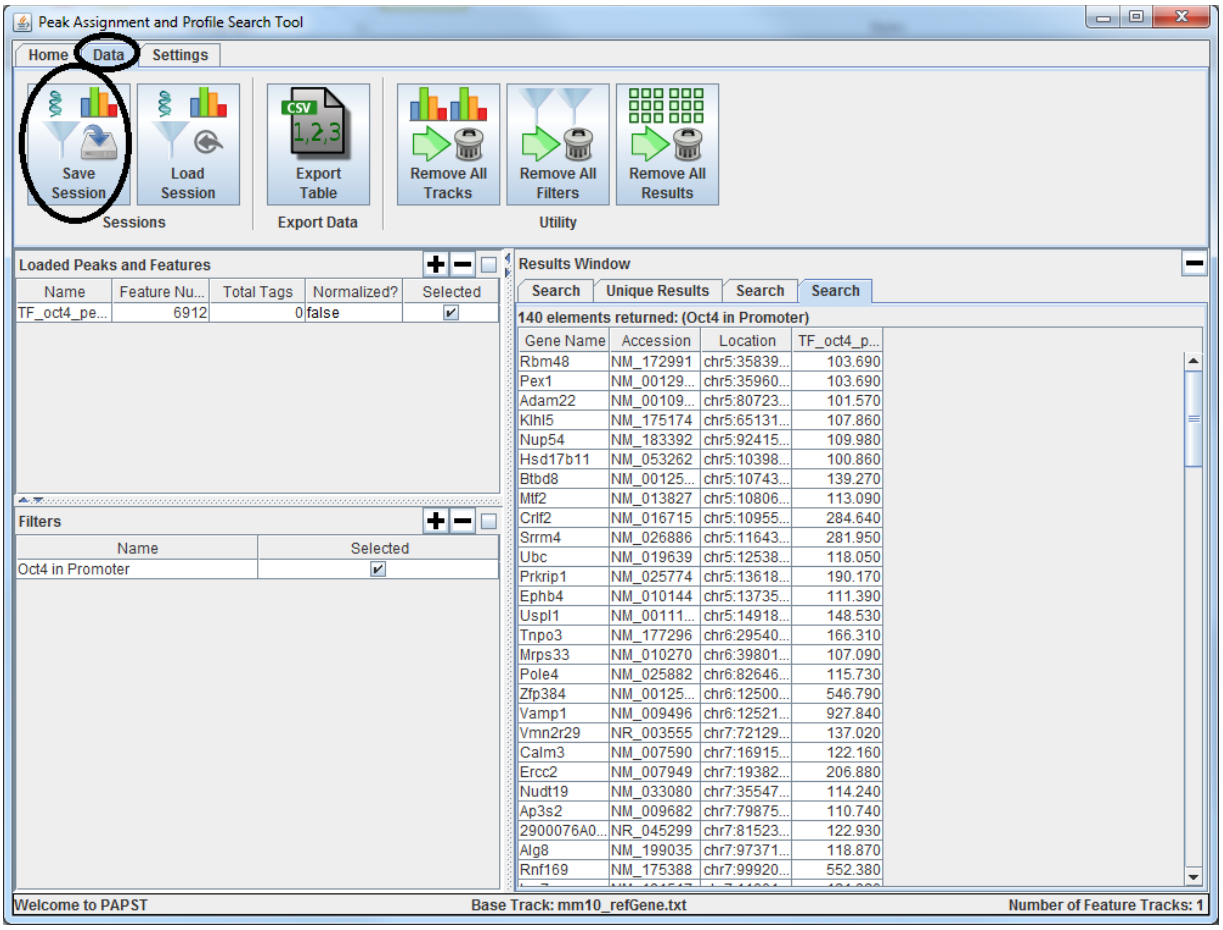
4. Click ok to remove the track. This track will no longer appear in the **Loaded Peaks and Features** pane.



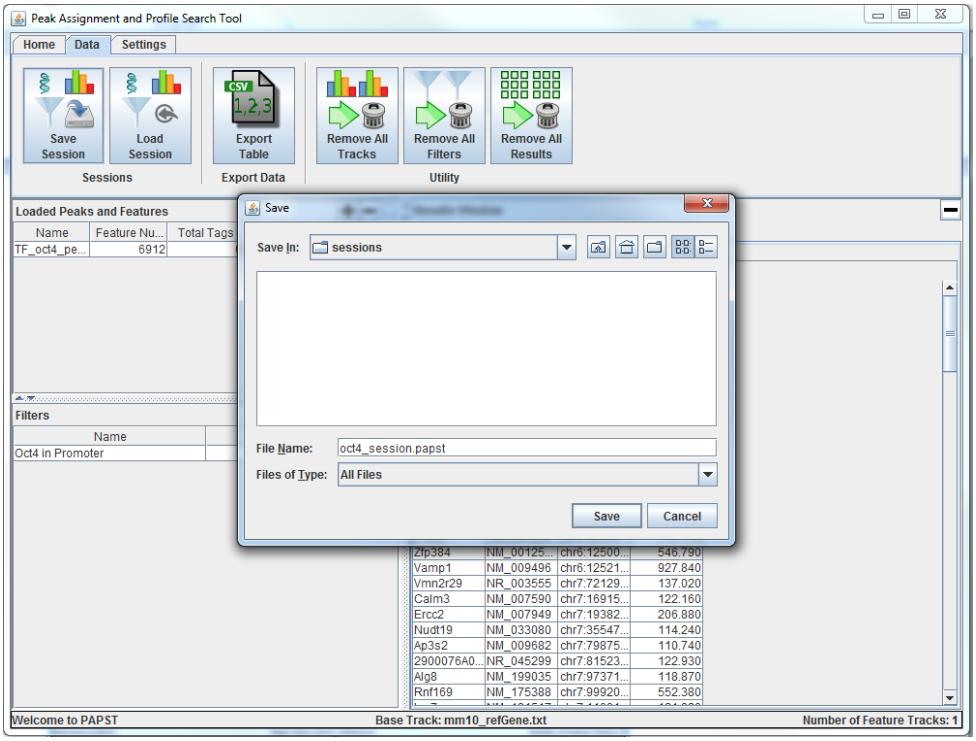
Saving a PAPST Session

You can save your PAPST session file at any time. A PAPST session holds the genome, peaks, filters, result tables, and custom parsers.

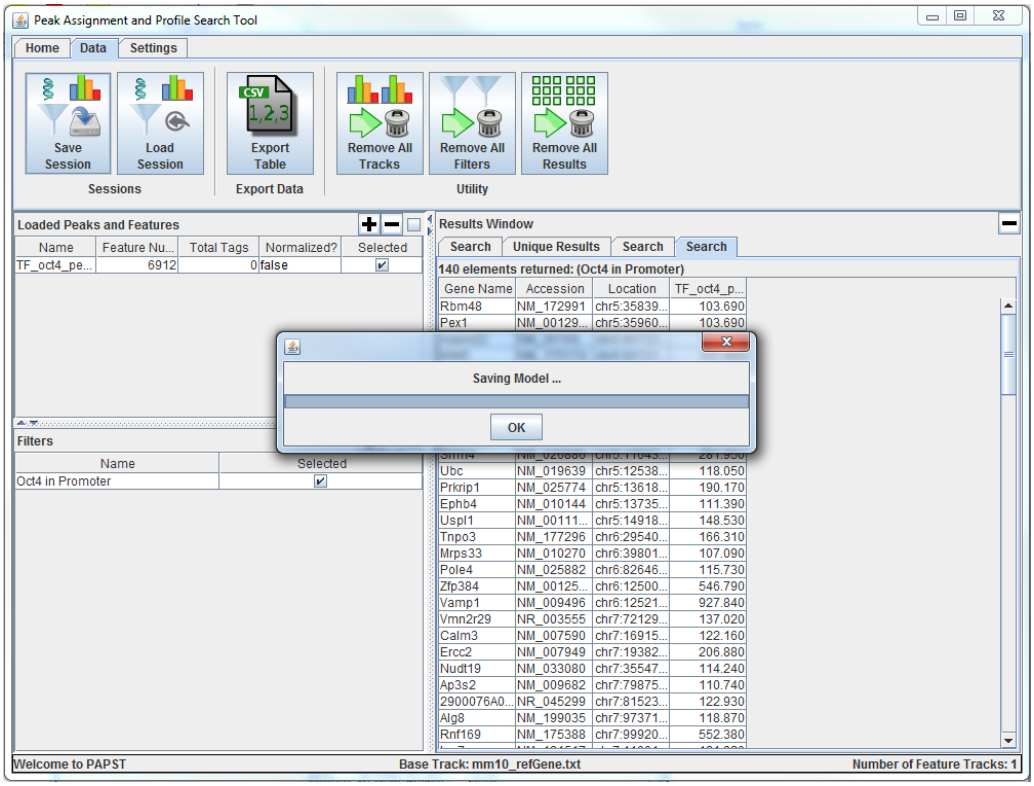
1. To save your PAPST session, navigate to the **Data** tab and click **Save Session**.



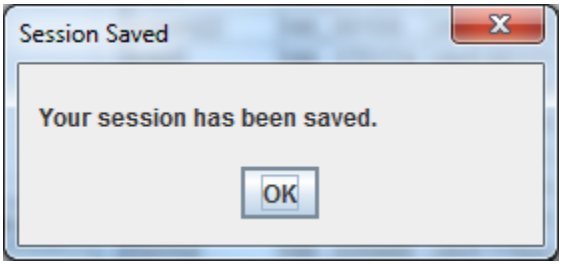
2. You will be prompted to give the session a name. Let's name the file 'oct4_session.papst'.



3. A dialog box with a progress bar will appear. Click OK when the process is complete.



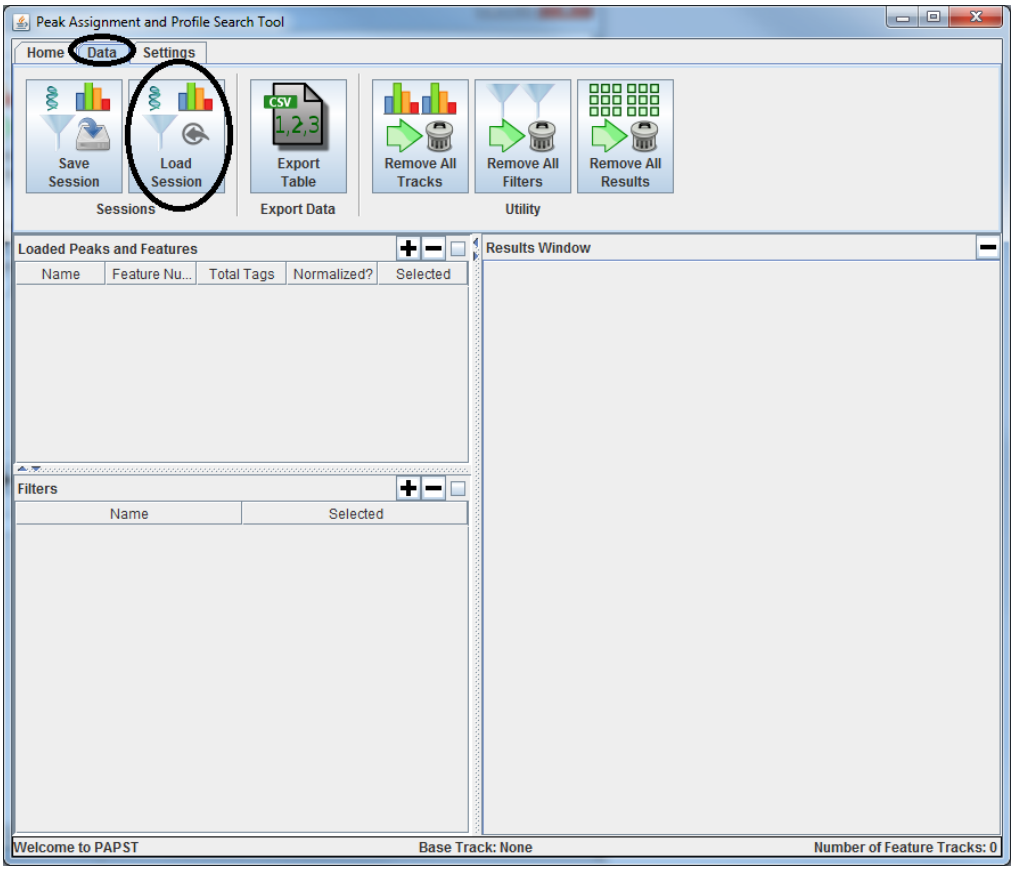
4. A dialog will appear to confirm that the session is saved.



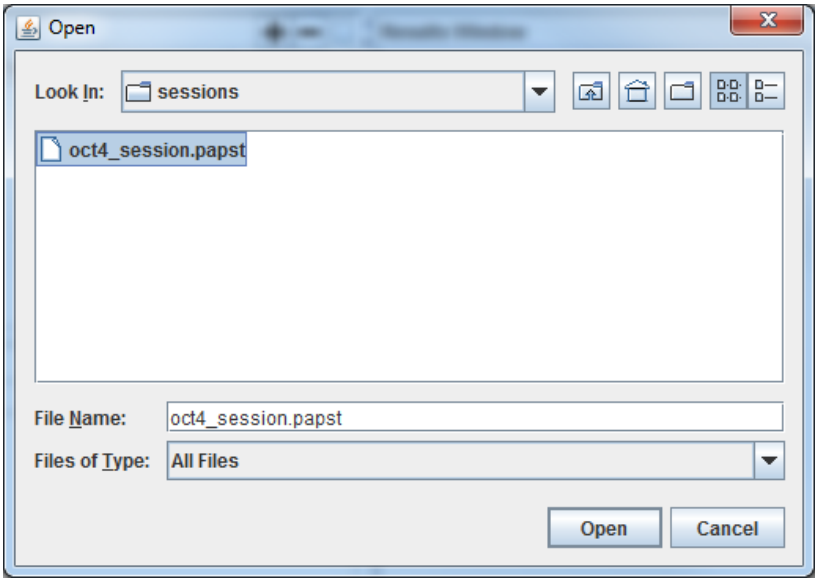
Load a PAPST Session

Loading a PAPST session is simple too. Use the **Load Session** button on the **Data** tab.

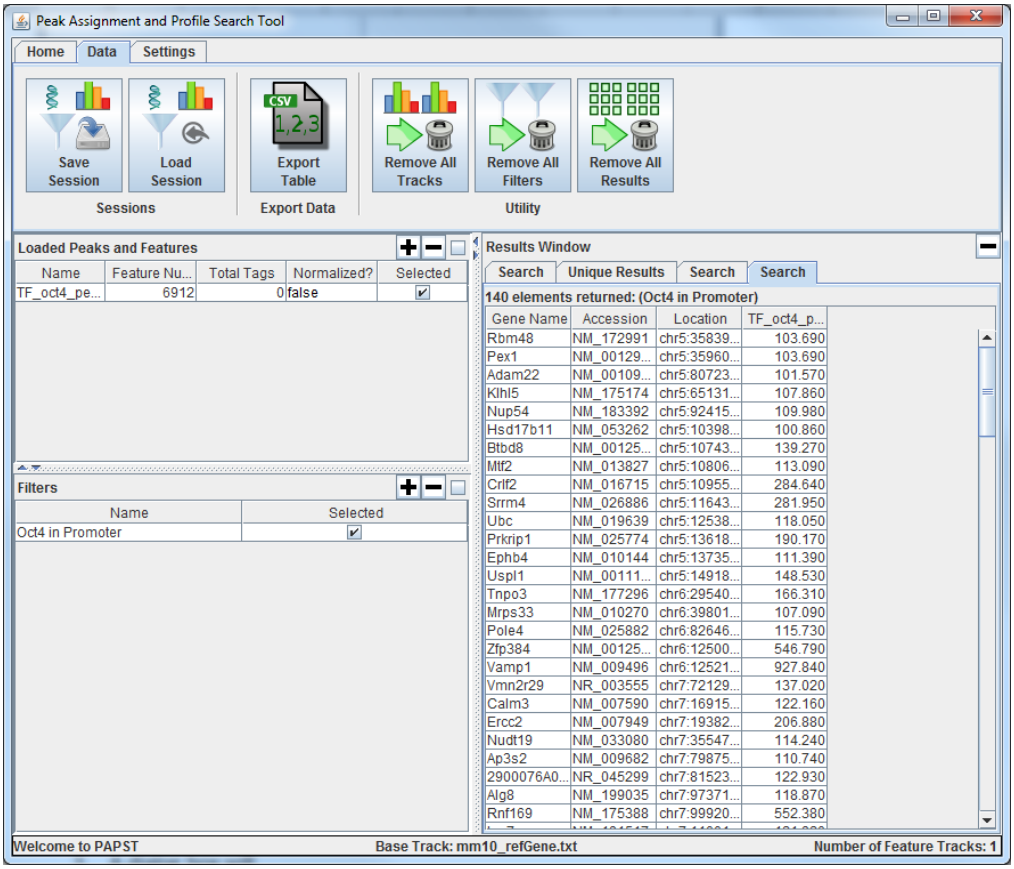
1. To Load a session, navigate to the **Data** tab and click **Load Session**.



2. Select the desired file.



3. Your data will appear once it has loaded.



Working with Filters

Searching with Filters

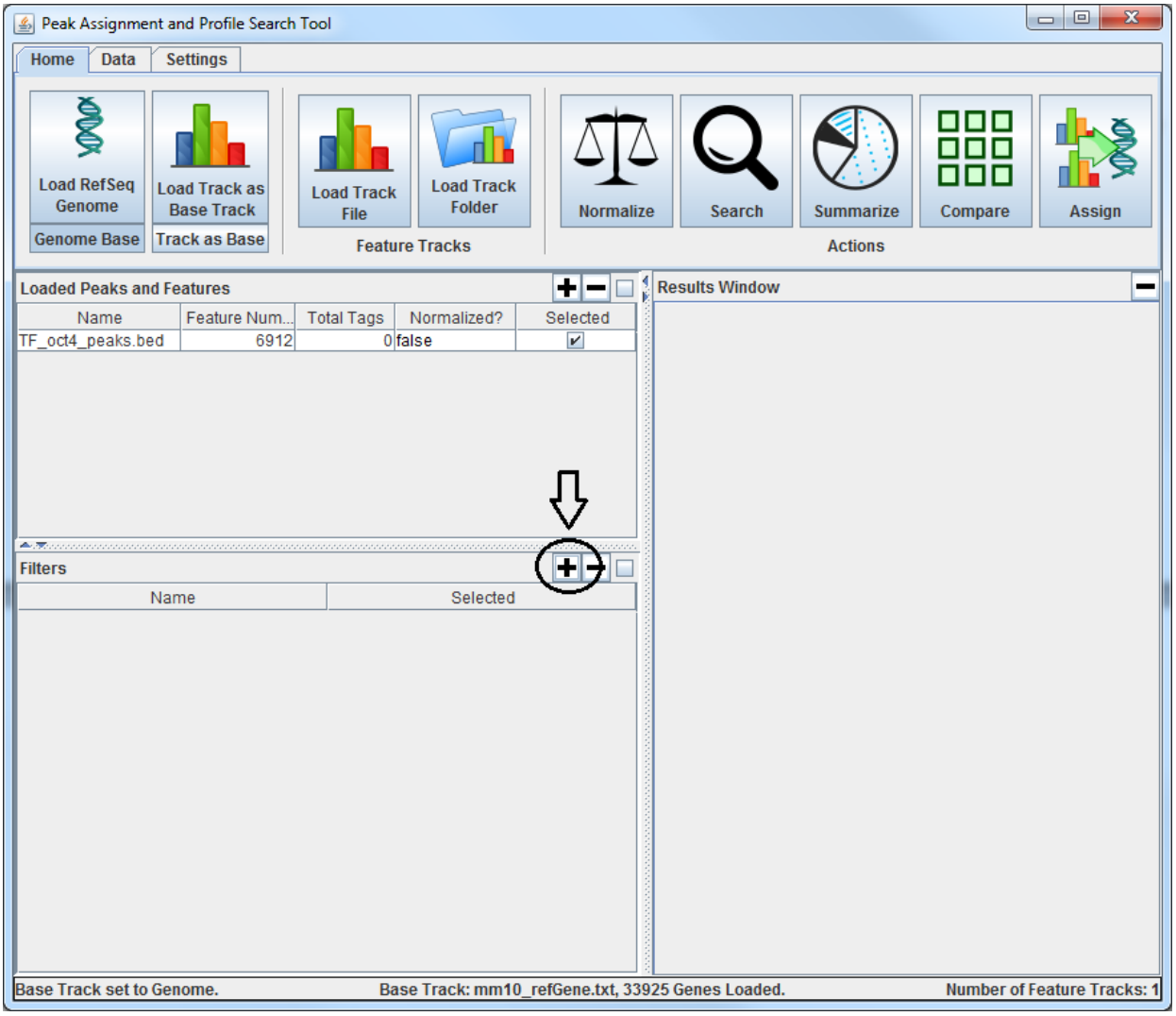
The most powerful feature of PAPST is the ability to perform gene searches using track filters. Each filter imposes a constraint on the set of genes that are returned.

Let's start with a simple example.

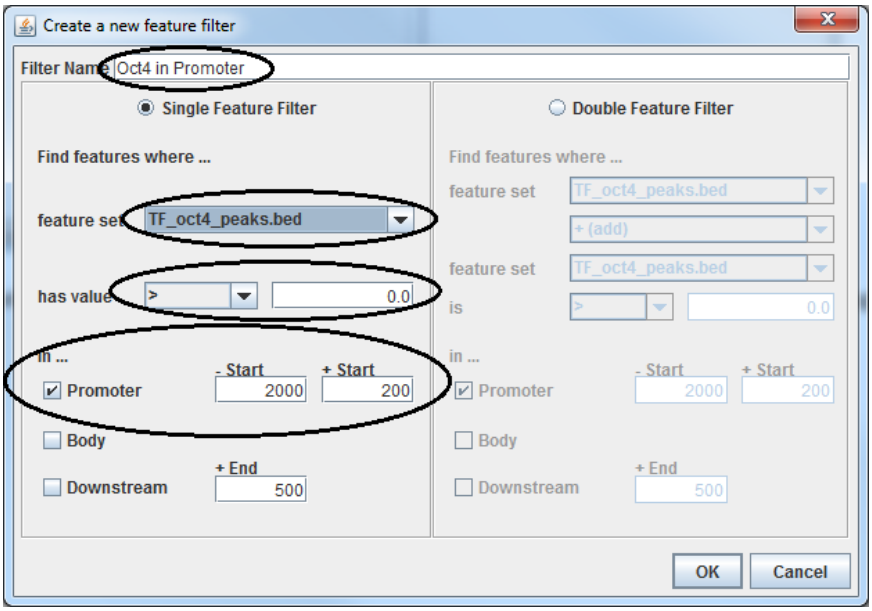
1. Using the provided data, load the mm10 genome.
2. Load the file named TF_oct4_peaks.bed. (You may need to create a new parser)

We will create a filter to select genes with Oct4 in their promoters. We will define the promoter as the region from -2000bp to +200bp relative to the TSS.

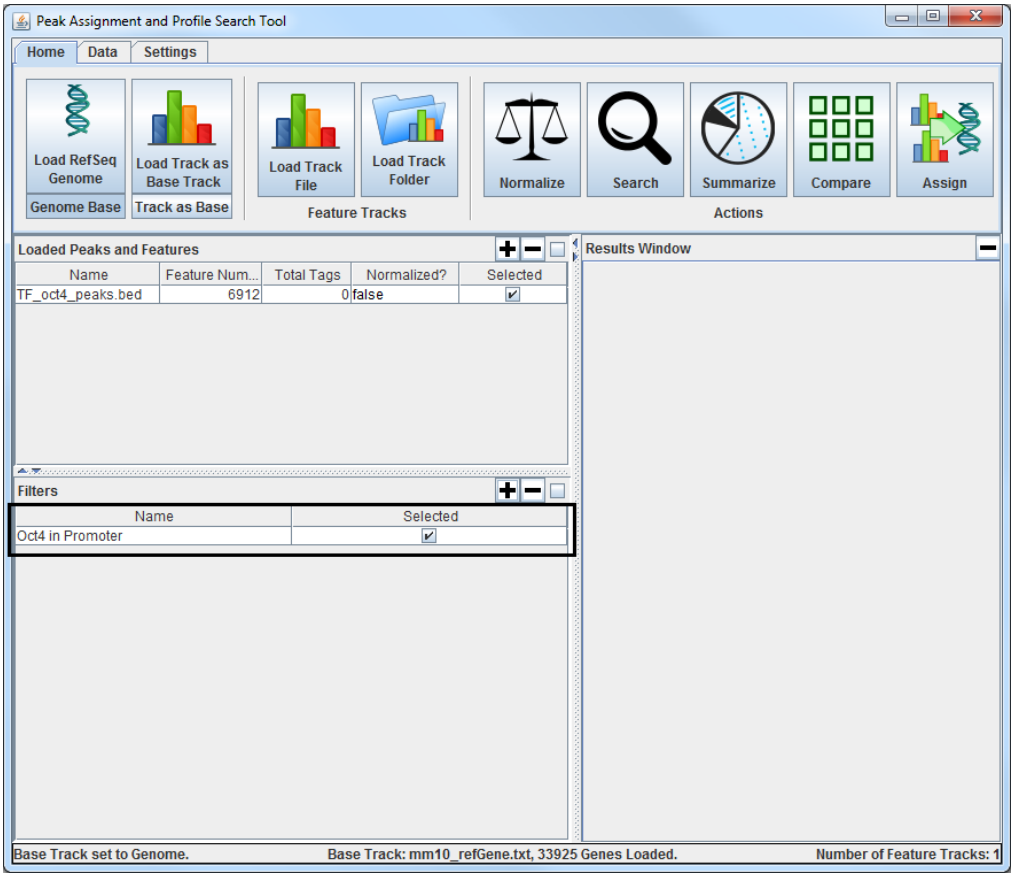
3. Click new filter button denoted by the **plus sign**, '**+**'.



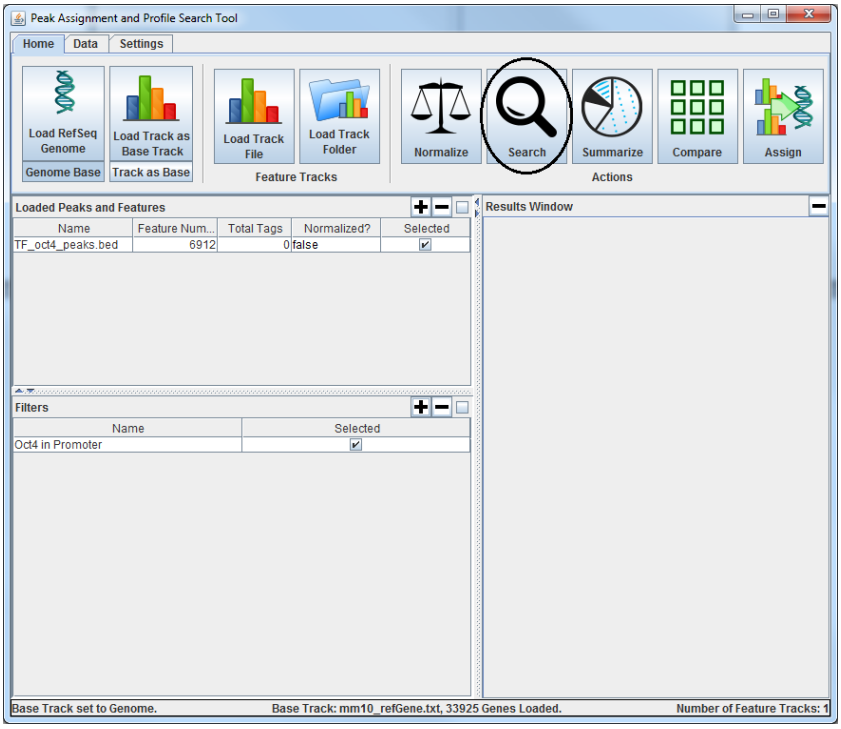
4. Select the **TF_oct4_peaks.bed** feature set and adjust the promoter settings for this single feature filter. Give the filter a name. The threshold determines the stringency of the search. Click **OK** to create the filter.



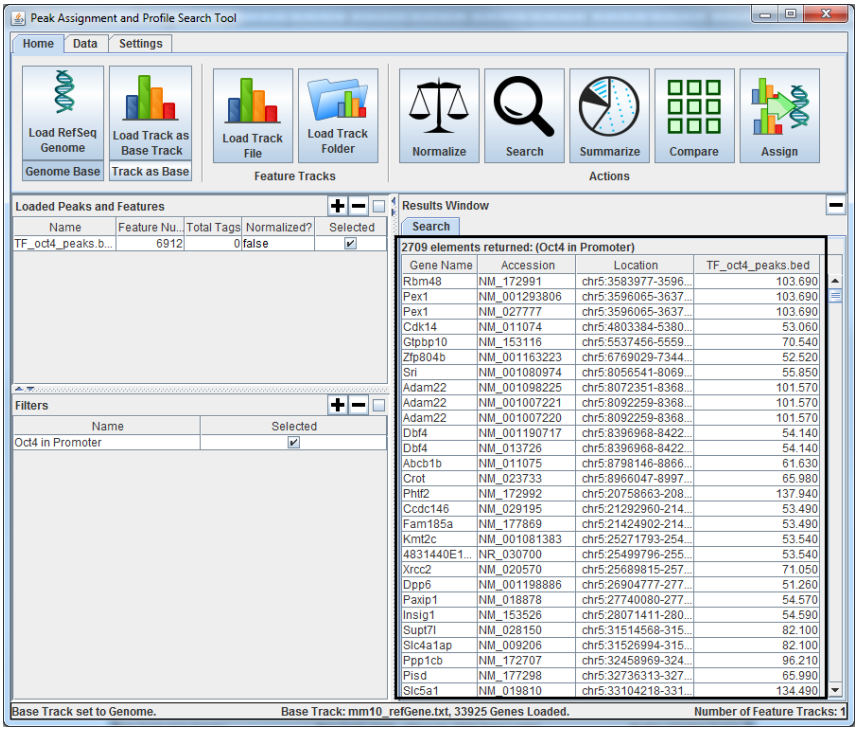
5. The created filter will appear in the **Filters** pane.



6. Click the search button to apply the selected filters to the loaded genome.



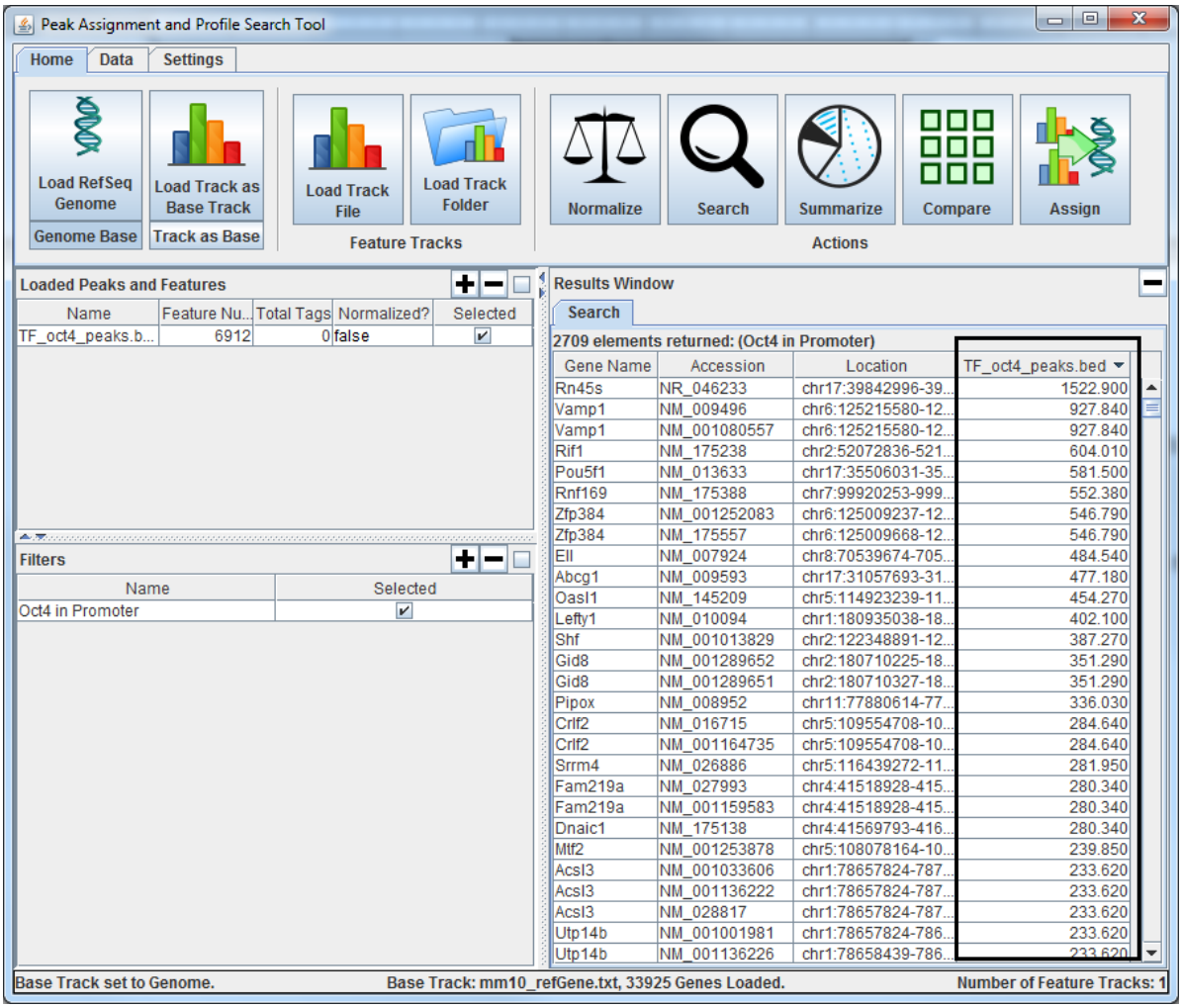
7. The results appear as a table in the **Results Window**.



This table shows a list of genes that have a significant Oct4 peak in their promoter (-2000 +200 to TSS).

We may want to prioritize this list by finding the genes with the strongest signal or p-value.

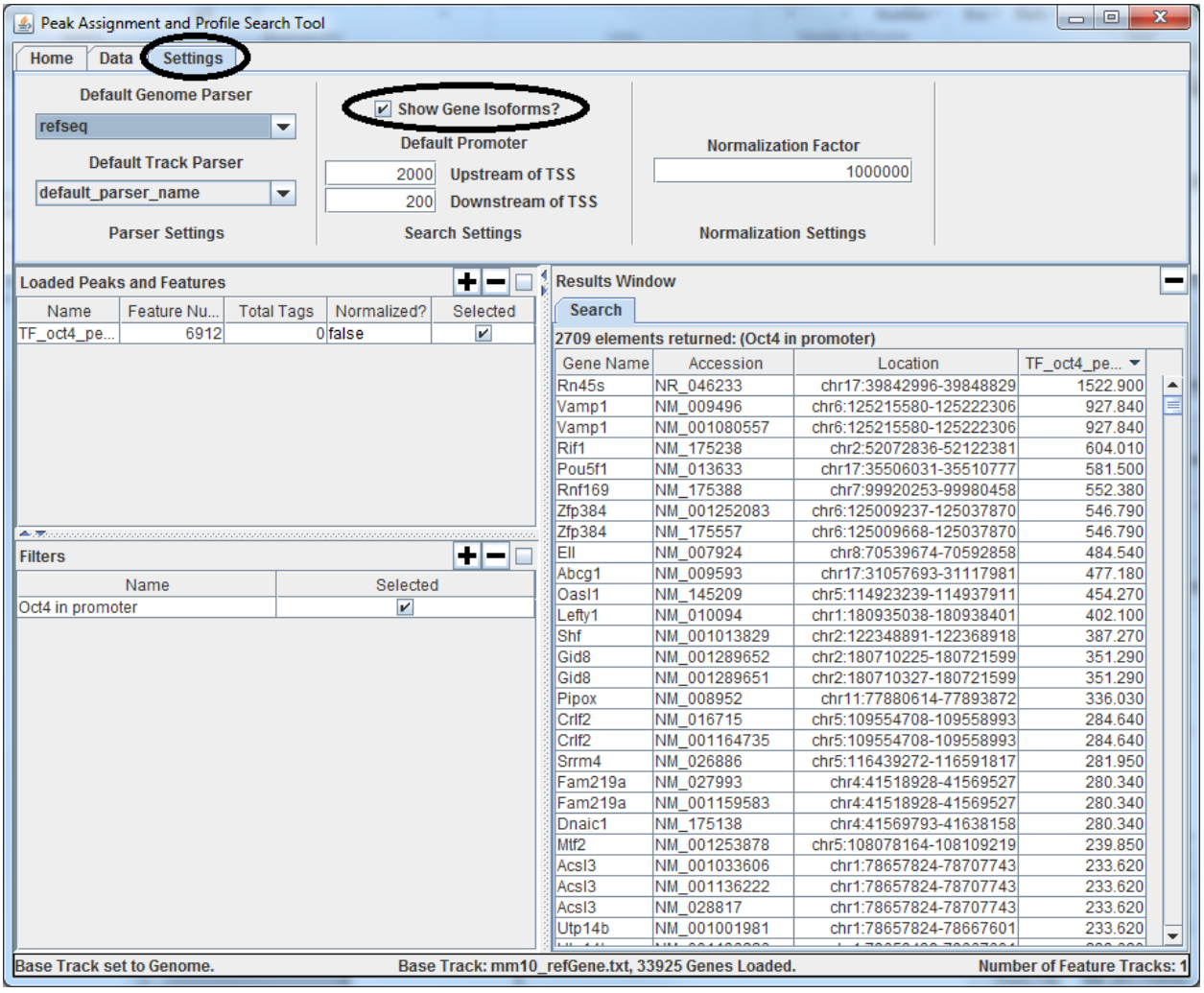
8. Click on the features column to sort it by the Oct4 peak's value. You will see the genes with the strongest value at the top.



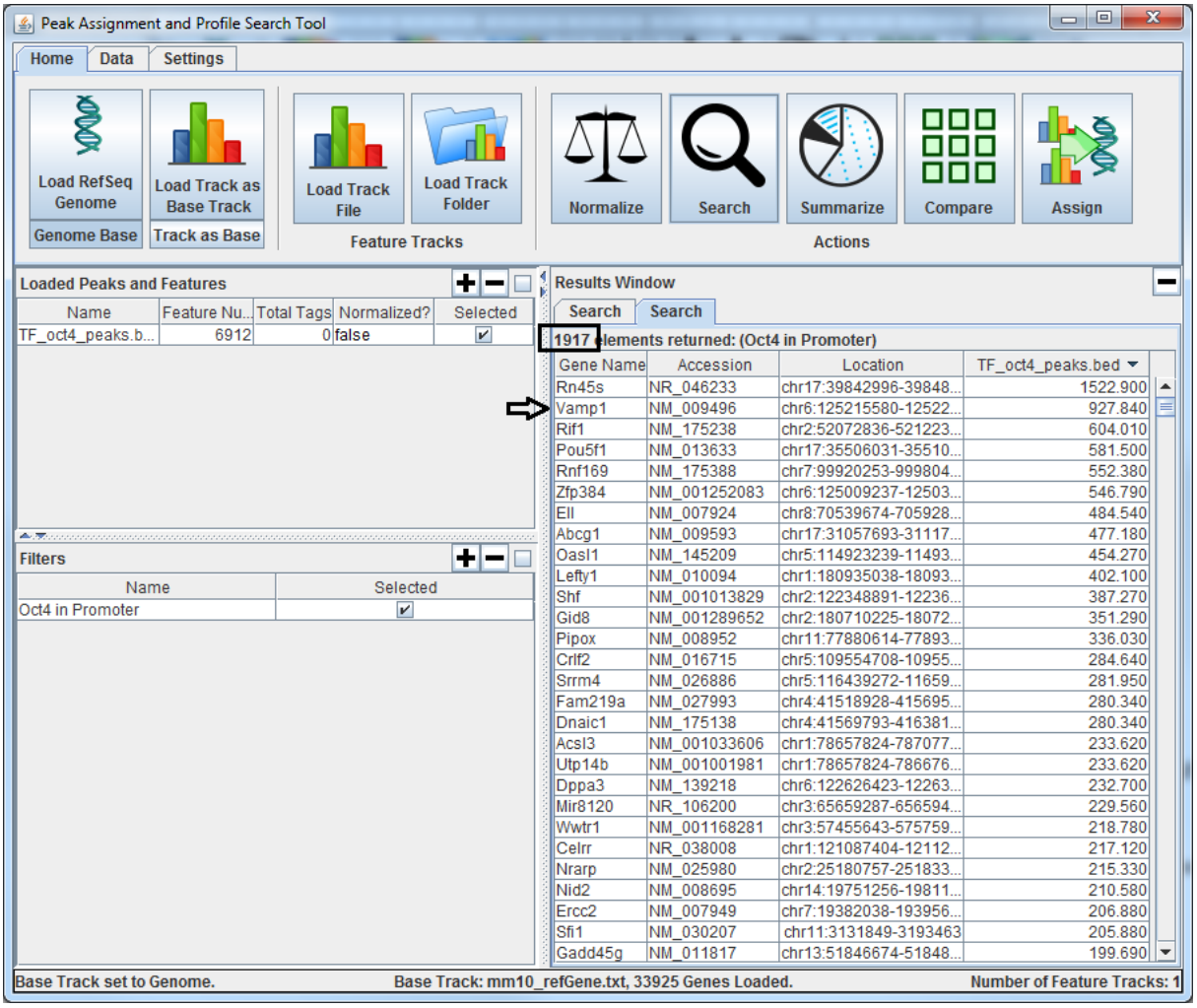
The top 5 genes with the strongest Oct4 signal in their promoters are Rn45s, Vamp1, Rif1, Pou5f1, and Rnf169.

Some gene isoforms are in this list (e.g. Vamp1). We can use settings to hide the isoforms and show only one row per unique gene.

9. To hide isoforms, navigate to the **Settings** Tab and uncheck the **Show Gene Isoforms?** option box**.**

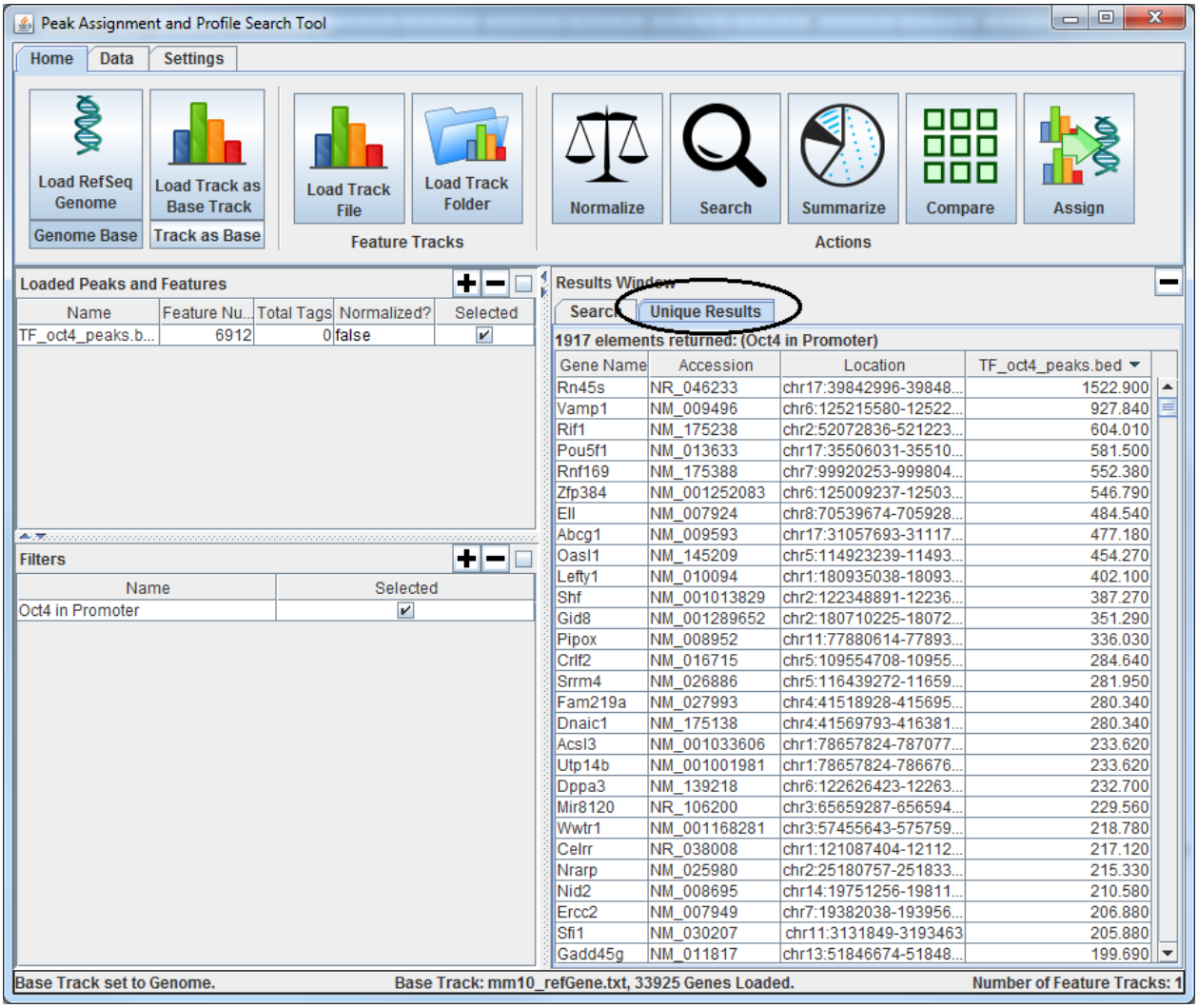


10. Repeat the search.



Notice that there are now fewer rows returned and Vamp1 only shows a single isoform (the first one encountered).

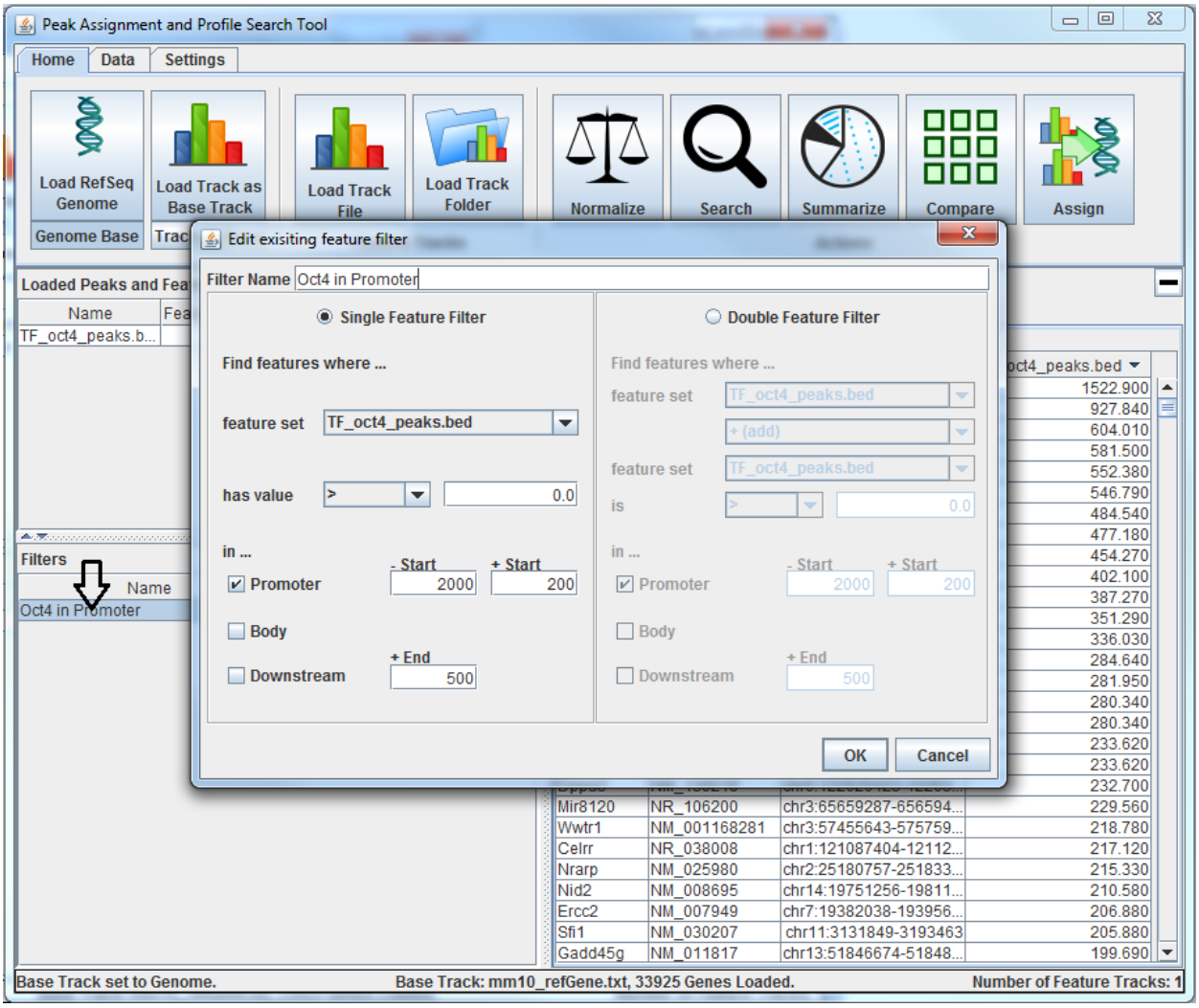
11. Edit a **Results** tab by double clicking on the tab label and typing the new label. Press enter to apply.



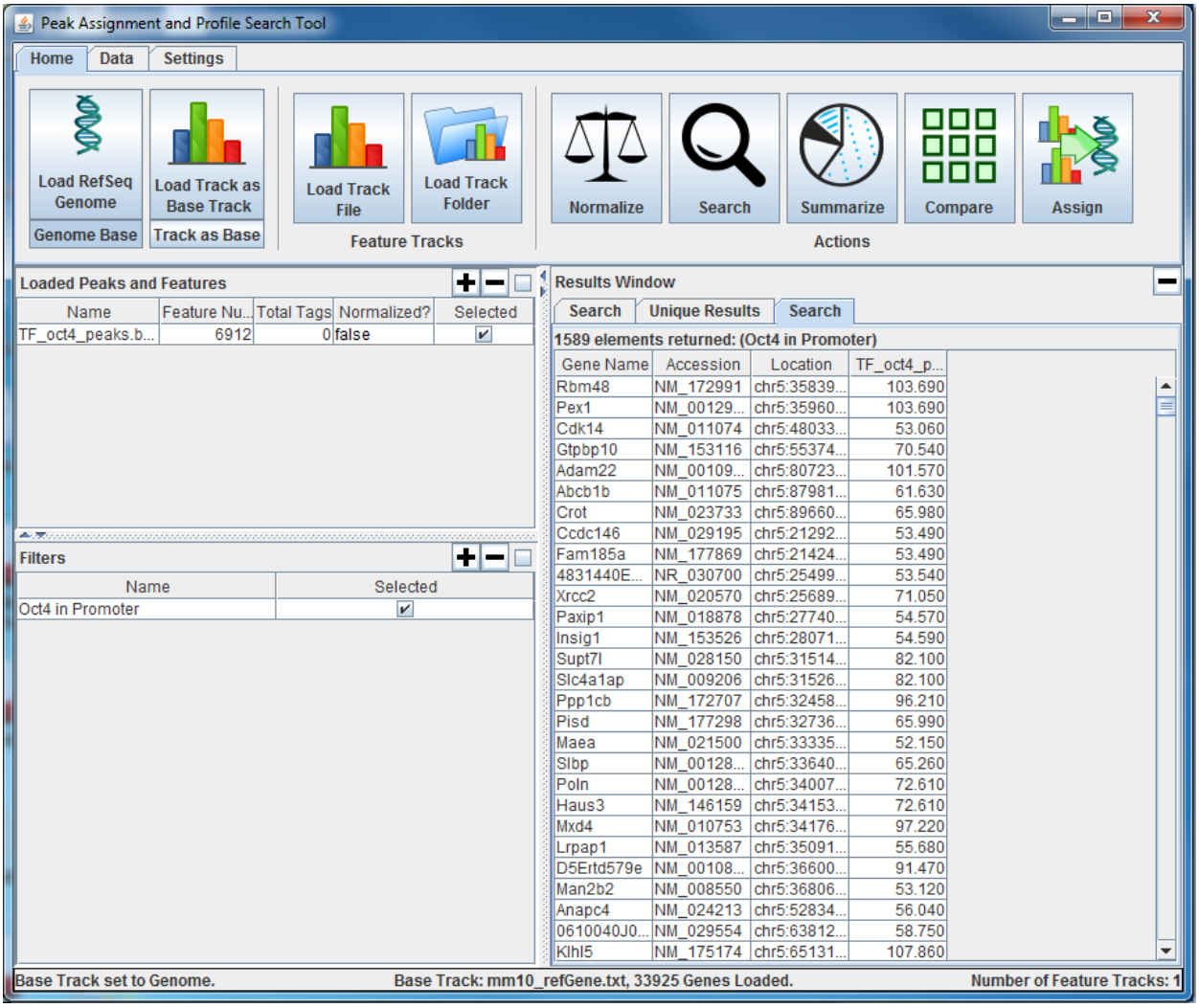
Editing Filters

Editing filter options is done easily by double clicking the filter in the window.

1. Double click the filter in the **Filters** pane to open the **Filter Editor**.

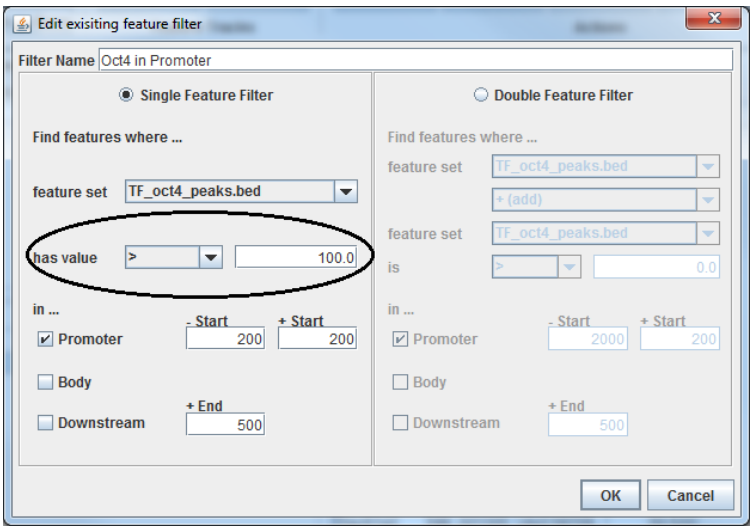


2. Let's change the promoter region to be -200bp +200bp relative to the promoter region and **Search** again.

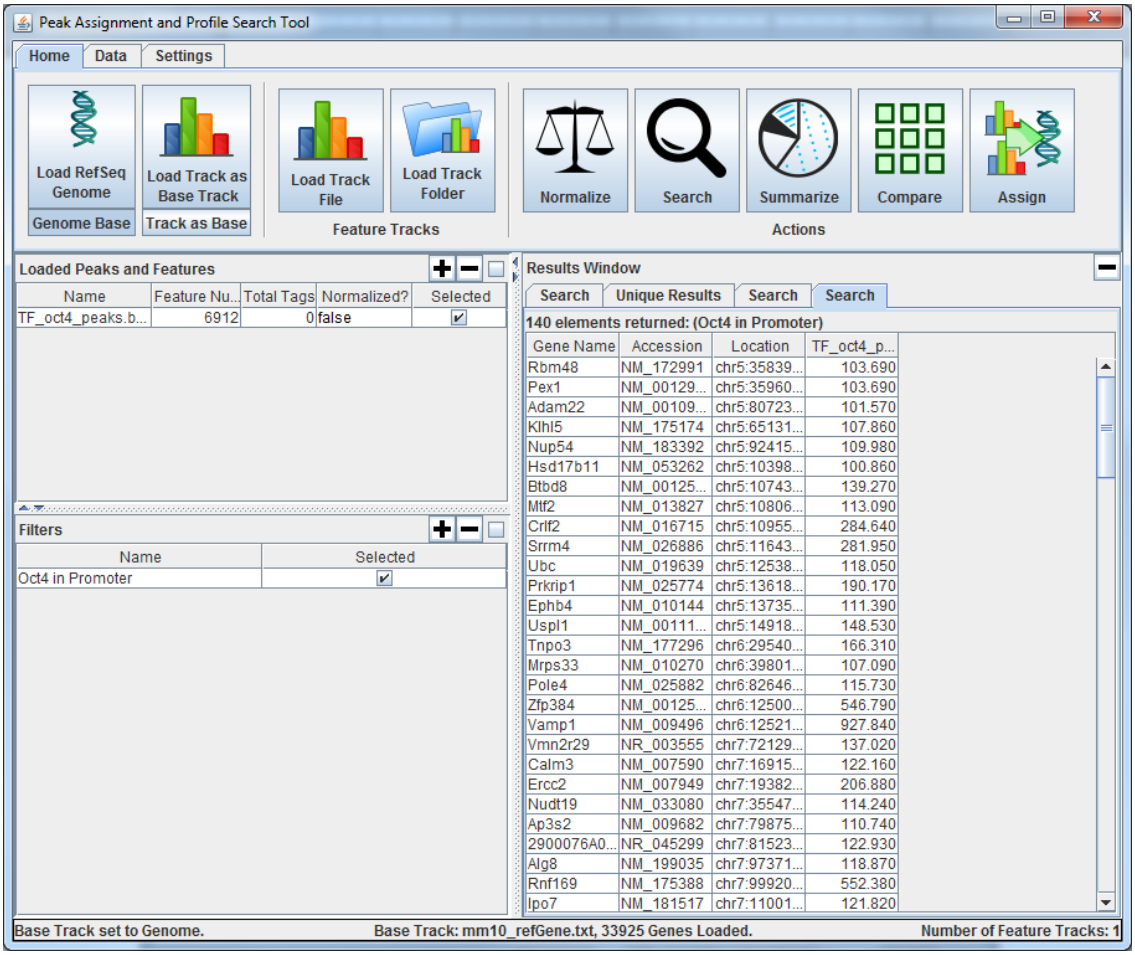


This reduces the number of unique genes to 1589 from 1917.

3. Double click to edit the filter again and adjust the threshold to '> 100'. Click OK and **Search** again.



4. Increasing the threshold filters more of the genes. Now we have only 140 unique genes that pass the threshold.

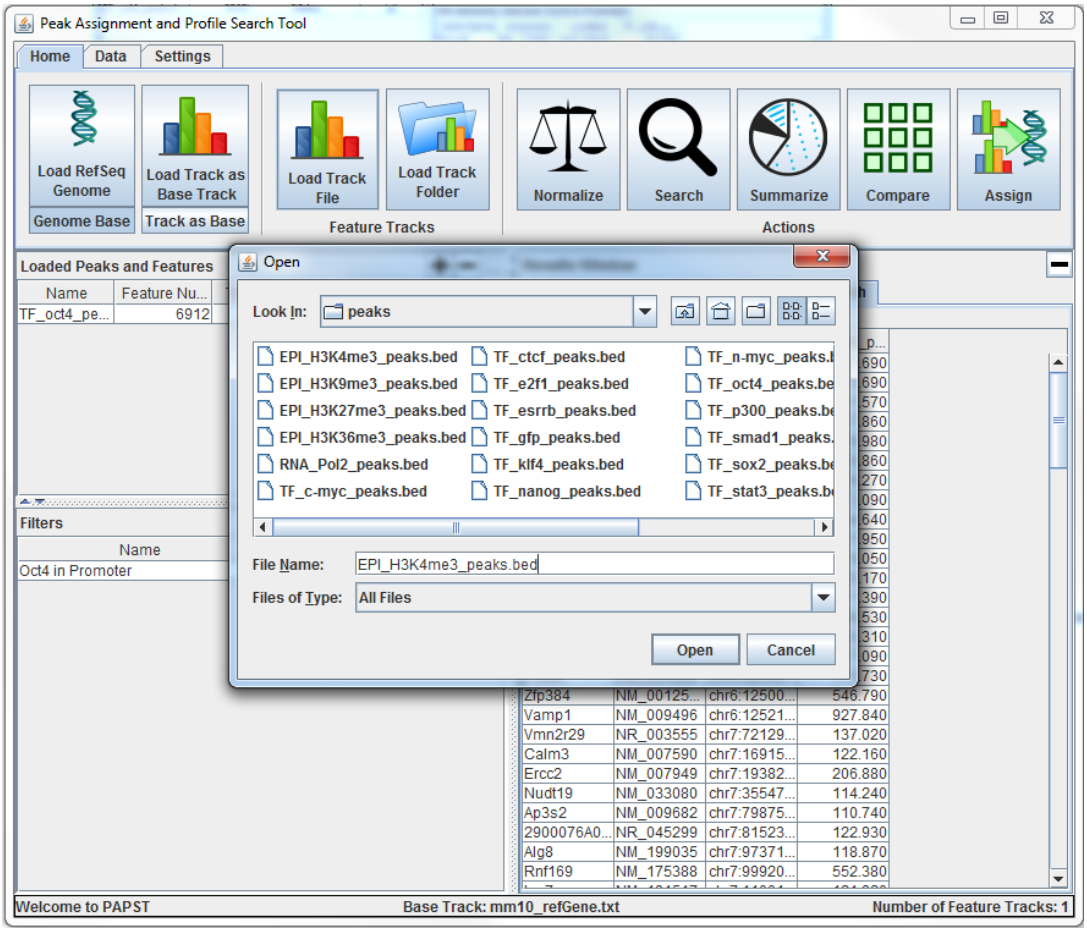


Using Multiple Filters

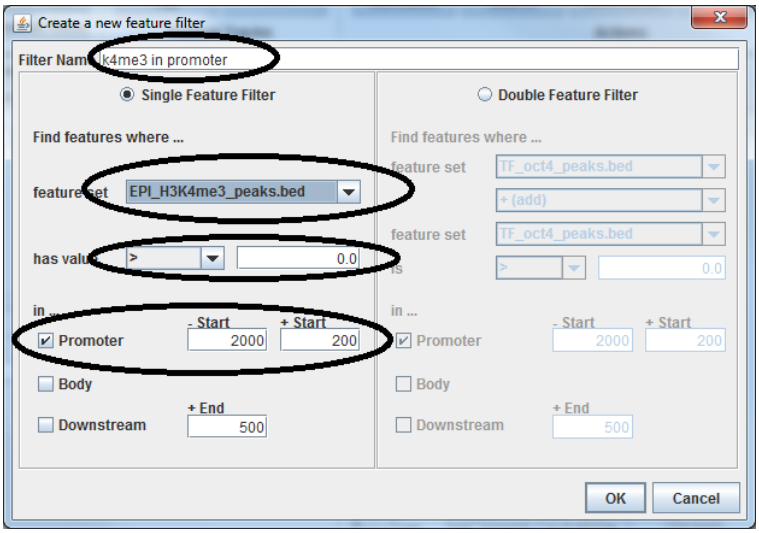
Using multiple filters allows the user to create complex patterns. Filters have '**and**' semantics, meaning that for a gene to appear in the results it **must pass all filters (Filter1 and Filter2 and …)**. In this way the user can iteratively define a set of interesting genes.

In this example we will find activated genes using k4me3 histone modifications in the promoter and refine this list to genes that have Oct4 binding in their promoters.

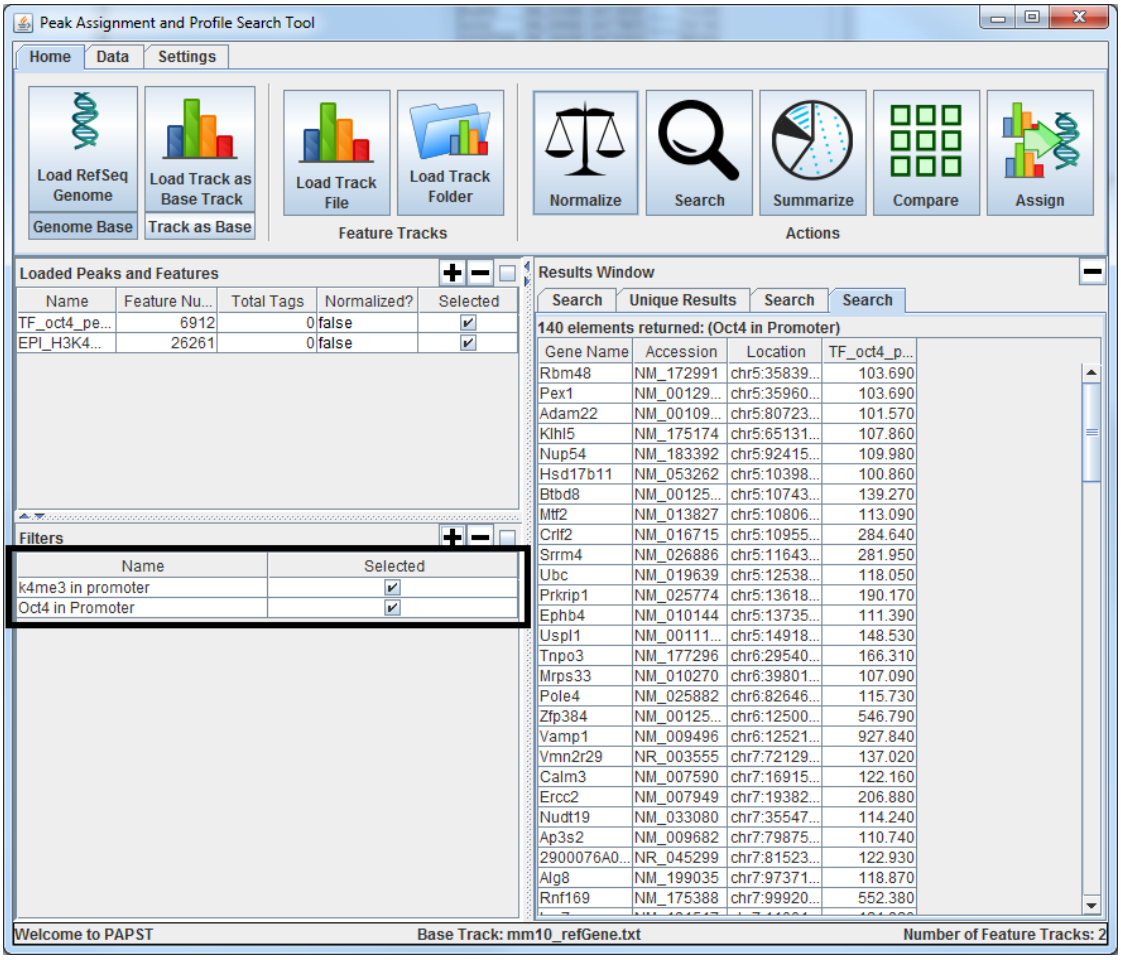
1. Load the mm10 RefGene genome and Oct4 TF binding as in the previous example.
2. Load the peak file for K4me3 modifications, **EPI_H3K4me3_peaks.bed.**



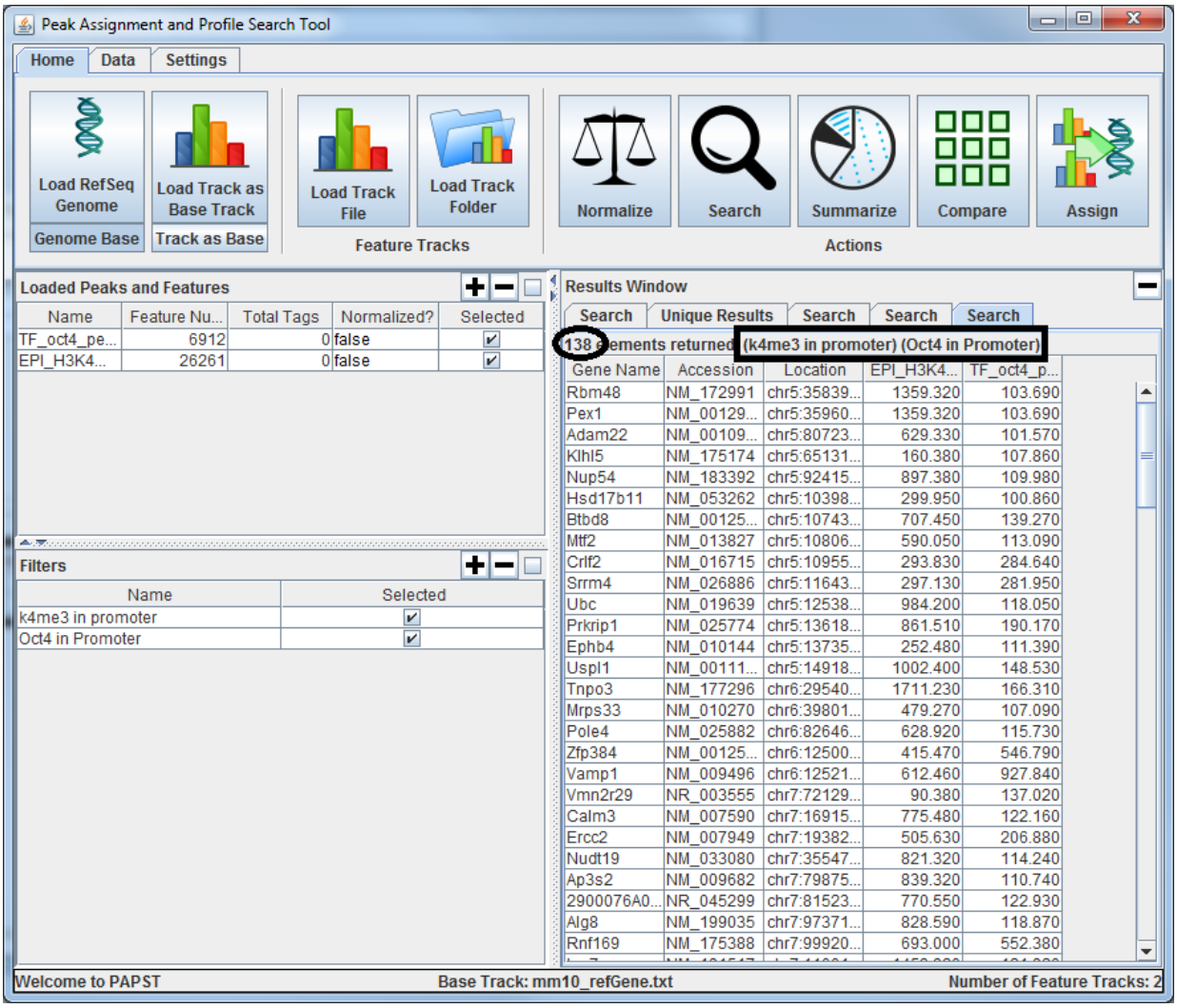
3. Create a new **Filter** for K4me3 in the promoter. Set the Filter **name**, **threshold**, and **promoter description**. Name the filter 'k4me in promoter'. Select the k4me peaks. Set the threshold to '> 0.0'. Keep the default settings for the promoter. Click OK.



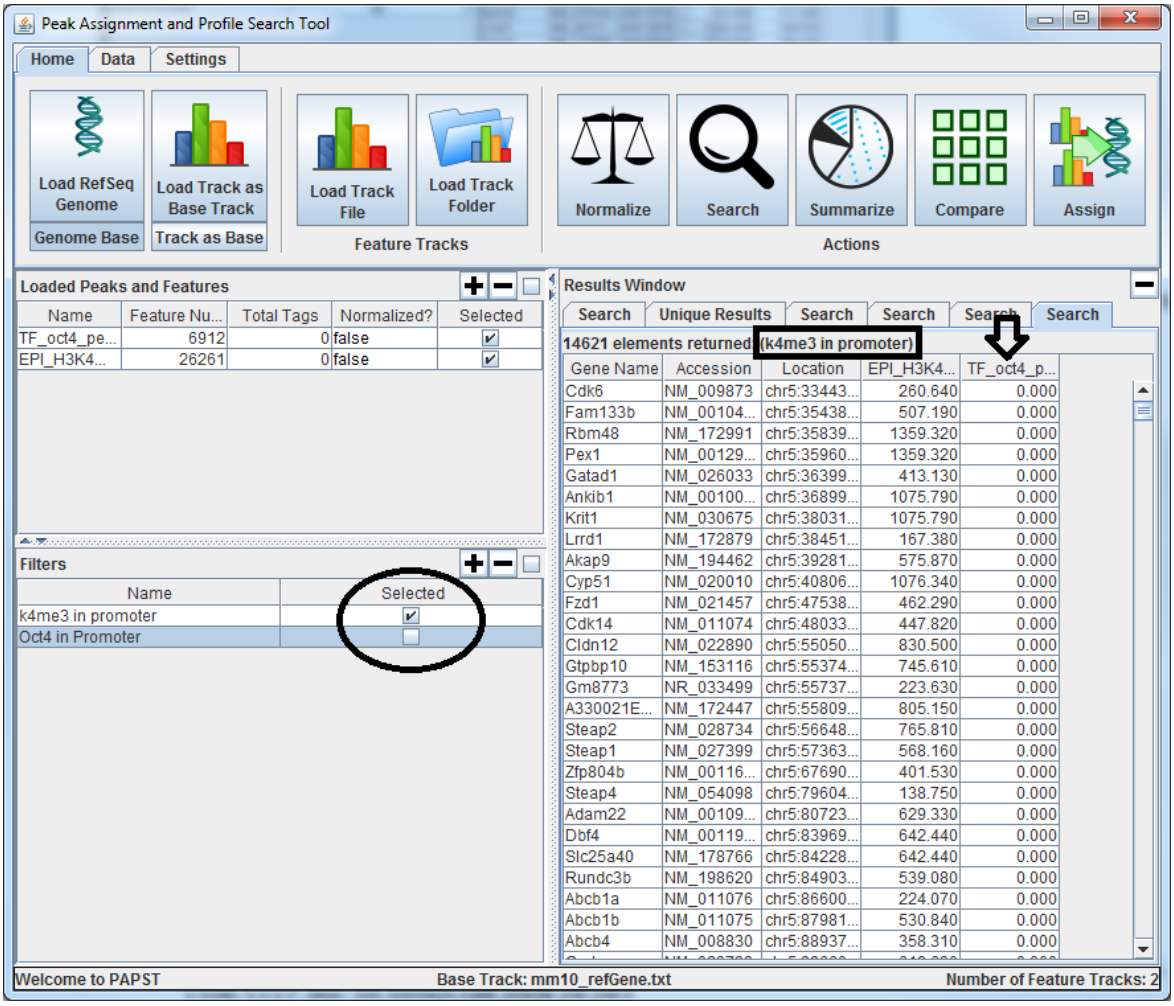
4. Now we have two filters loaded into PAPST.



5. Search again using both features. You will see that adding the second filter on K4me3 removed two of Oct4's strong proximal binding targets. PAPST also displays the name of each Filter used in the search.

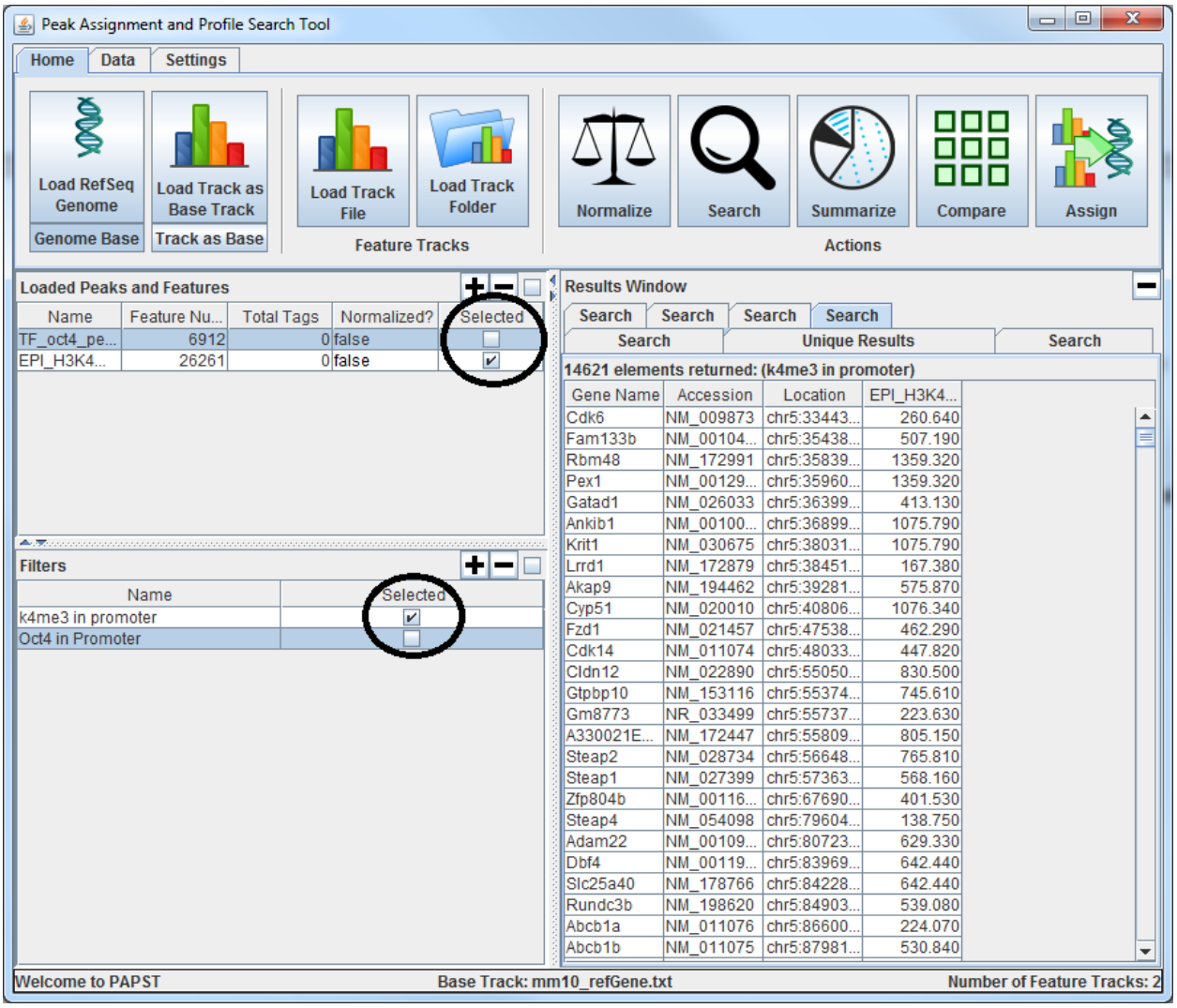


6. **To disable a feature**, uncheck the box in the **Filters** pane under the selected column. Click on the 'Oct4 in Promoter' Selected check box to disable it and search again.



Notice that only 'K4me3 in promoter' is shown in the results table. Also the 'Oct4 in promoter' column now contains only 0's. This is due to an optimization. **If a column is not used in at least one filter, its value is not reported**.

7. **To hide a peak or feature from output**, click on the selected checkbox in the **Loaded Peaks and Features** pane. Let's hide the Oct4 column from this search since it does not display any information. Click the selected checkbox to the right of the TF_oct4_peaks.bed feature and search again.



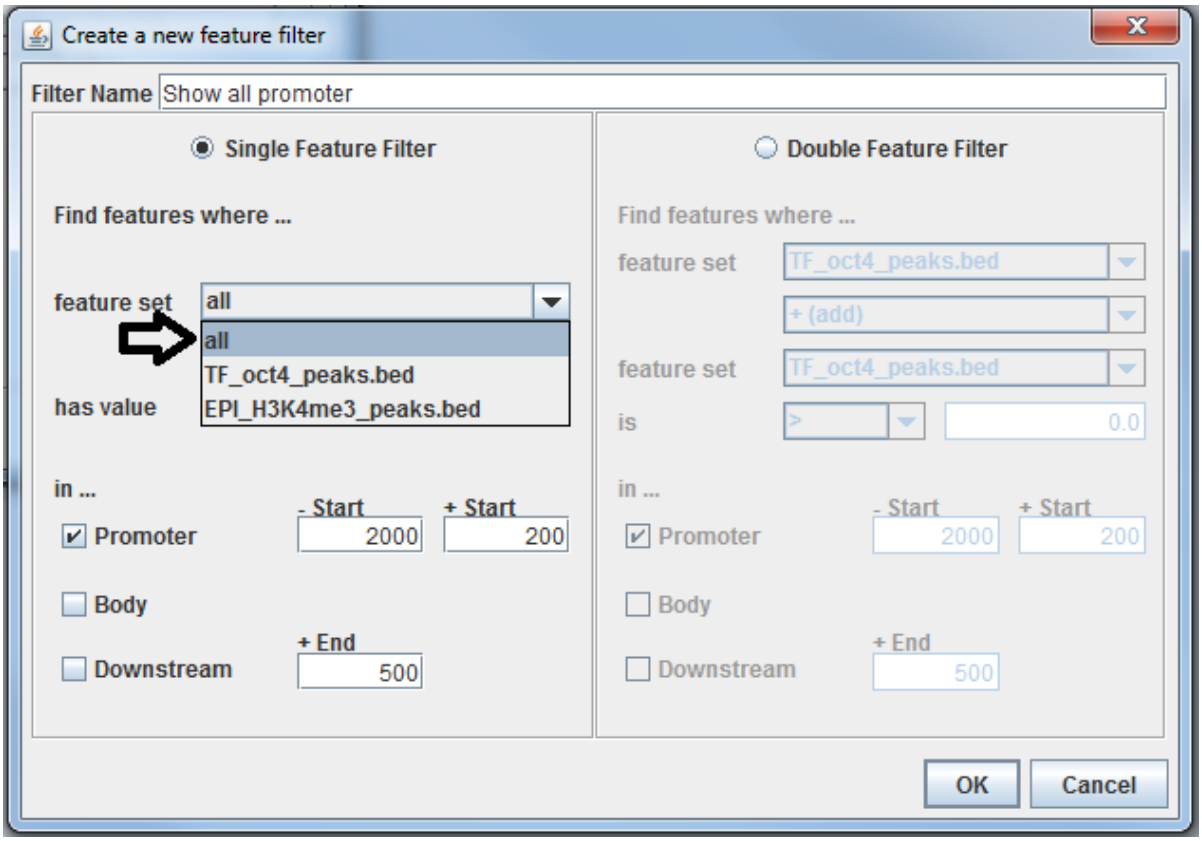
Now we see that the Oct4 column is hidden.

The ALL filter, displaying a value for all columns

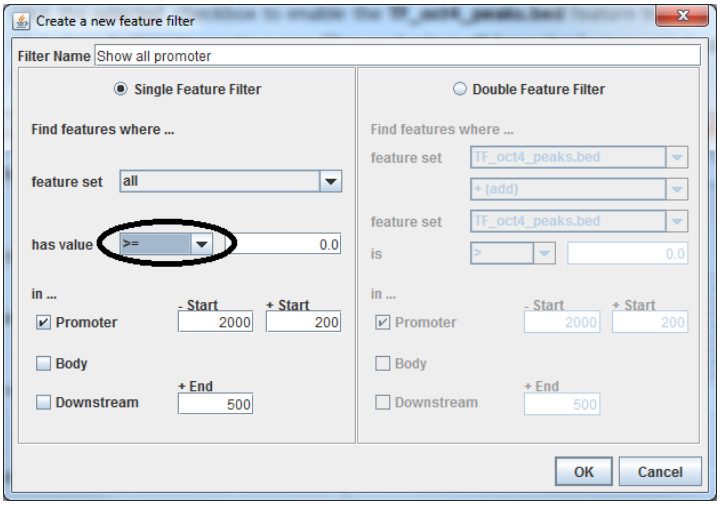
As we saw in the previous example, if a feature is not part of a filter its value is not reported. This is an optimization that saves on computation time. There are times when we want to see all values even if we are not filtering on them. PAPST has an **all** filter for exactly these situations.

Let's create a search that will only filter on K4me3 but will also display Oct4 values. We will do this with an **all** filter.

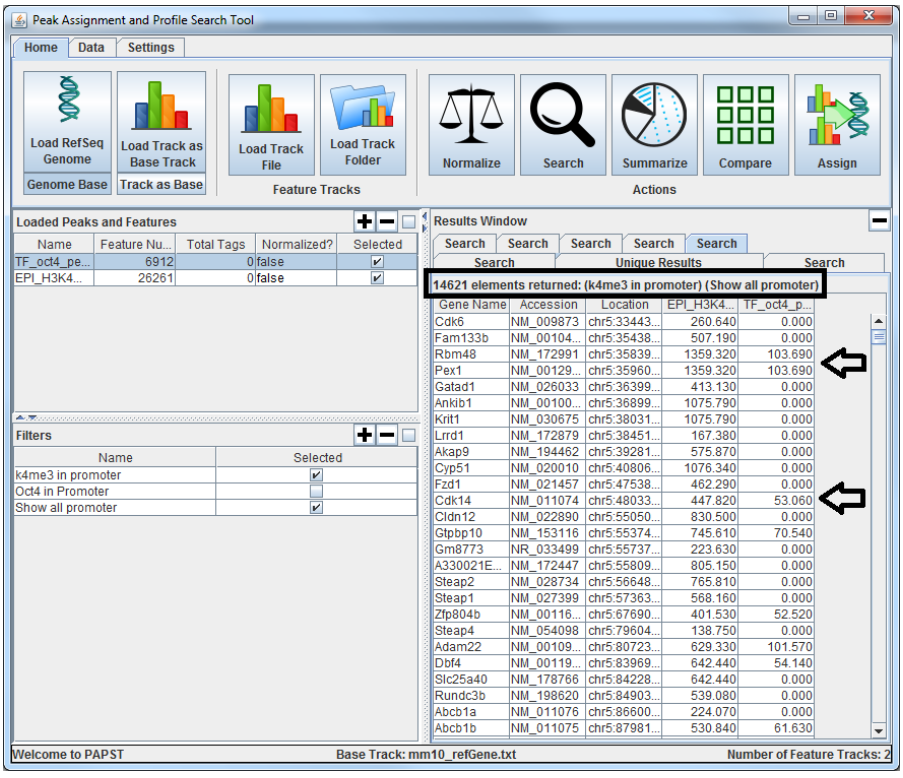
1. Click the selected checkbox to enable the **TF_oct4_peaks.bed** feature track.
2. Click the **plus** '**+**' button to create a new filter and select **all** from the feature set dropdown. Keep the default promoter settings. Name the filter 'Show all promoter'.



3. Change the condition from '>' to '>='. This will ensure that even genes with no Oct4 peaks will still be reported.



4. Click ok to save the filter. Now our **all** filter will be visible in the **Filters** pane. Search again to view the results. Now we see the Oct4 values for the returned genes.

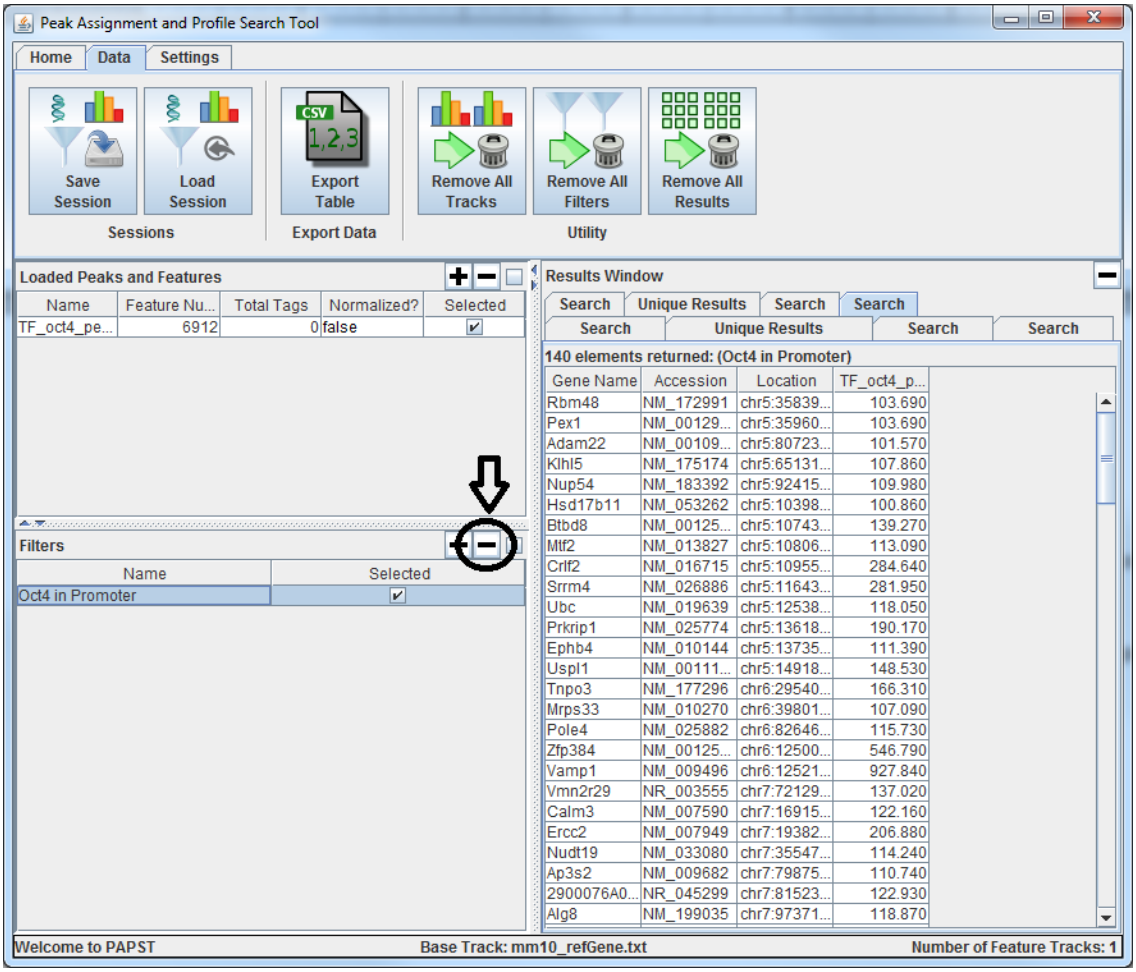


The **all** filter is just like a normal filter but it applies its constraints to all the selected features. Using '>= 0' will accept all features. Changing it to '> 0' will filter out all genes where any of the features are 0. It is easy create a set of filters that do not return any results, so be careful when working with many feature files.

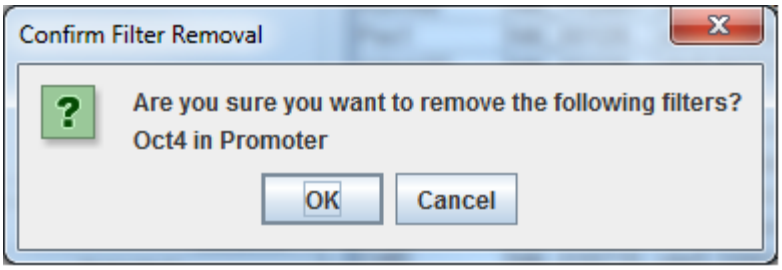
Removing Filters

Removing filters in PAPST is simple. Simply click on the **minus '-'** button in the upper right corner of the **Filters** pane.

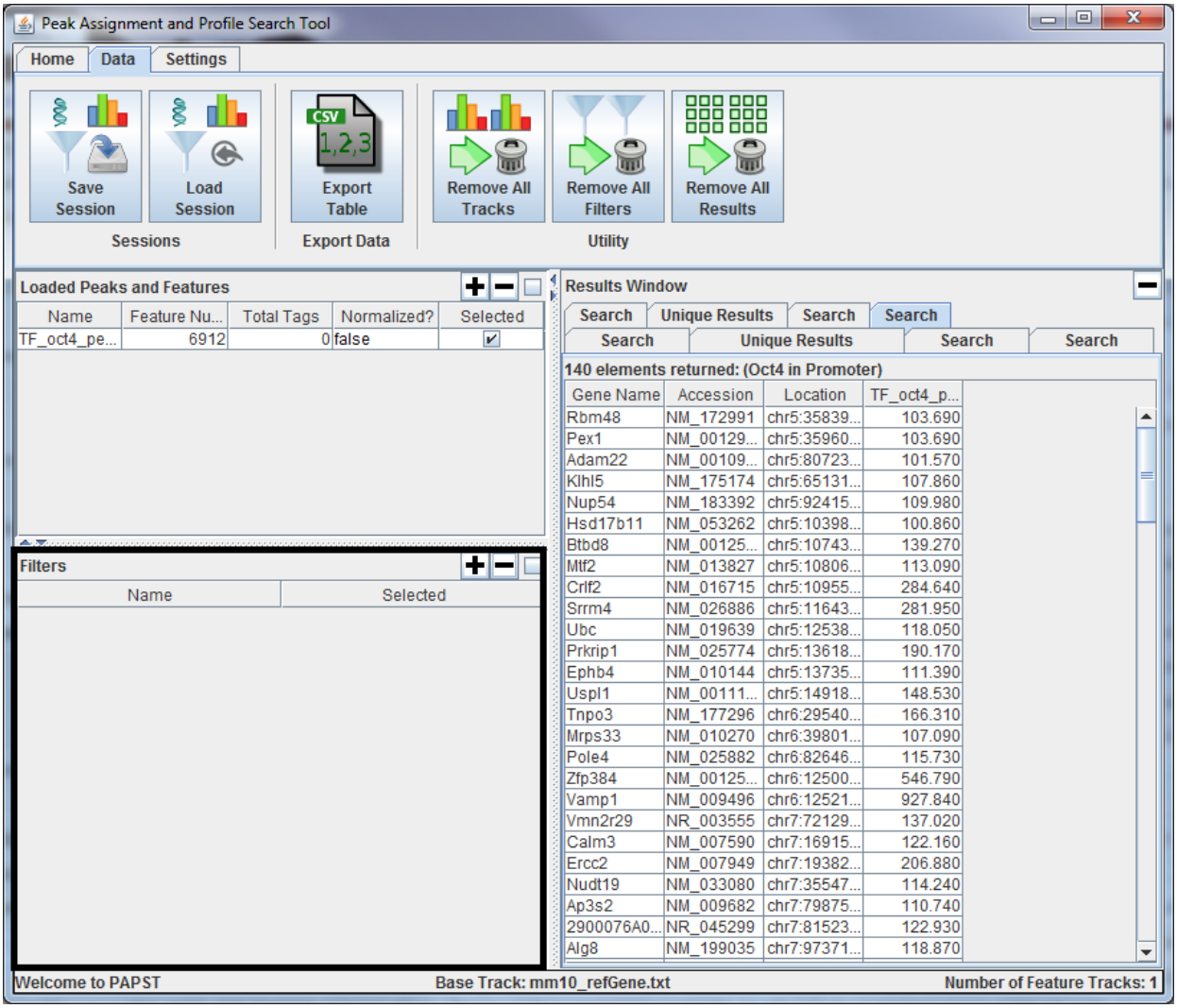
1. Click the filter you wish to remove to select it. Click on the **minus '-'** button in the upper right corner of the **Filters** pane.



2. A dialog will appear to confirm you wish to delete the selected filter.



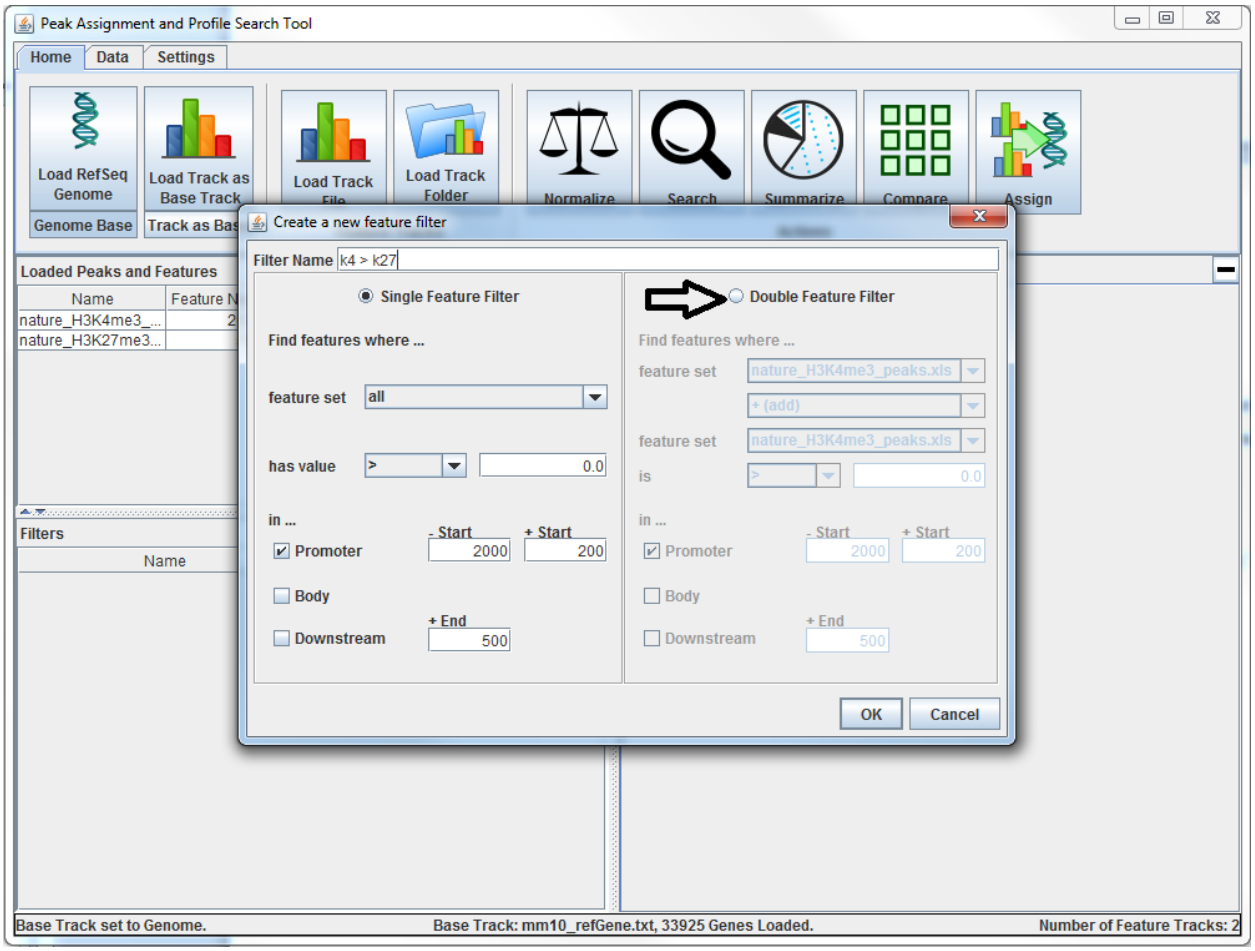
3. Click OK to confirm removing the filter. The filter is no longer available in the **Filters** pane.



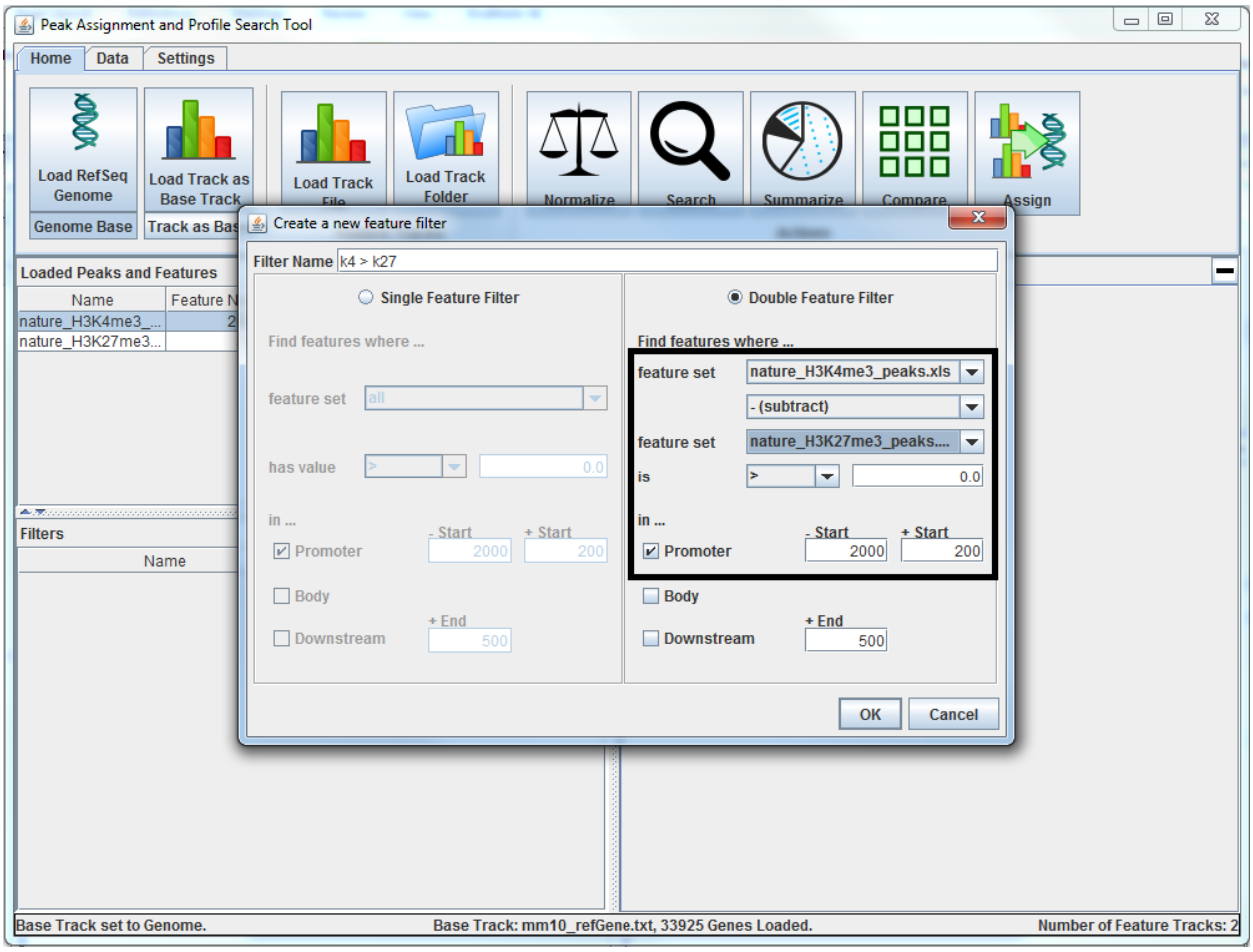
Advanced: Two Peak Filter

Filtering genes based on a comparison of two peaks can be beneficial. PAPST has a double filter that will consider an operation on two peaks and return regions that pass the filter. In this example we will find genes that have more K4me3 than K27me3 in their promoter regions.

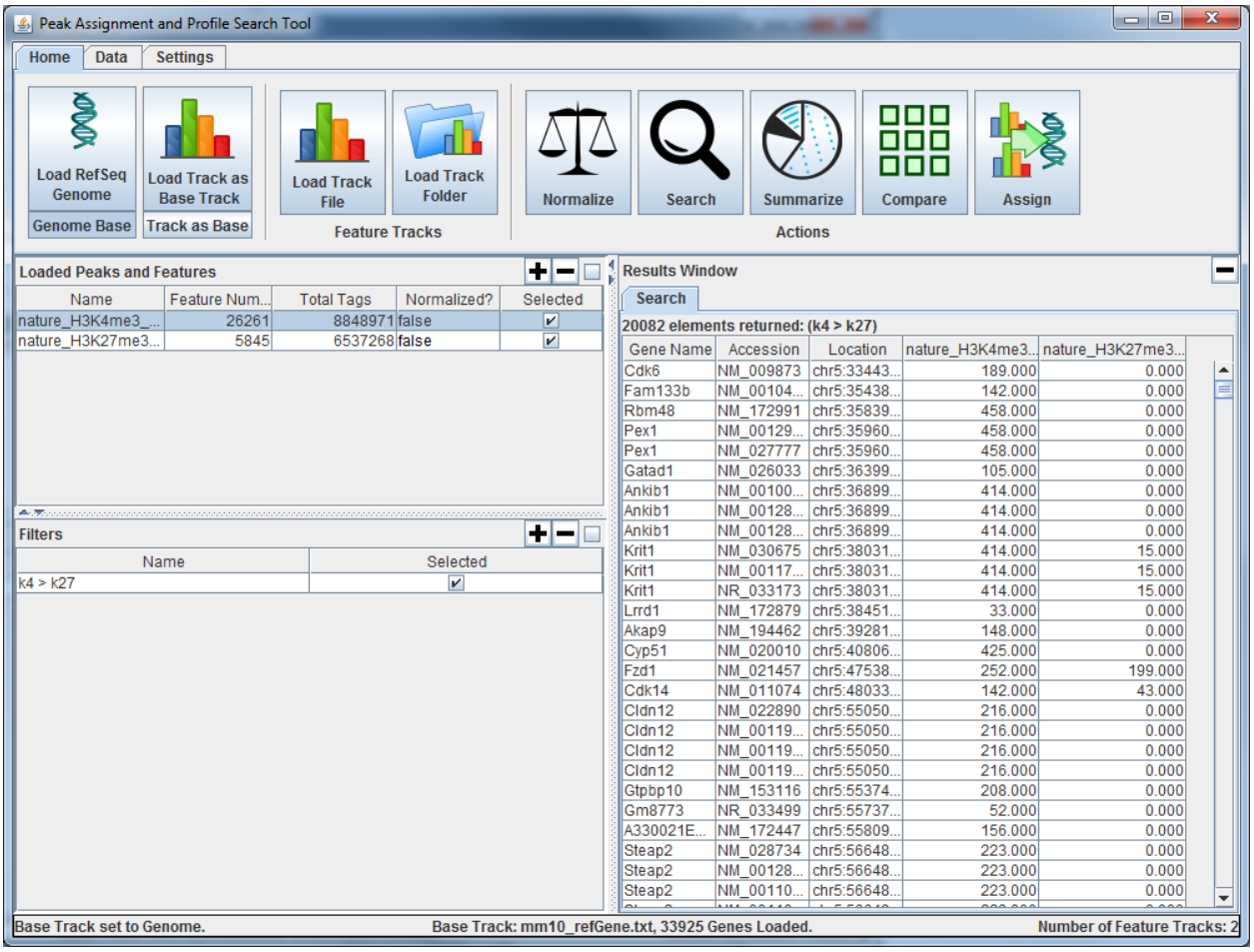
1. Load the mm10 RefGene genome into PAPST.
2. Load the files '**nature_H3K4me3_peaks.xls**' and '**nature_H3K27me3_peaks.xls**' under peaks/macs_xls/.
3. Create a new filter:
    a. Click on the **Double Feature Filter** button to change modes.



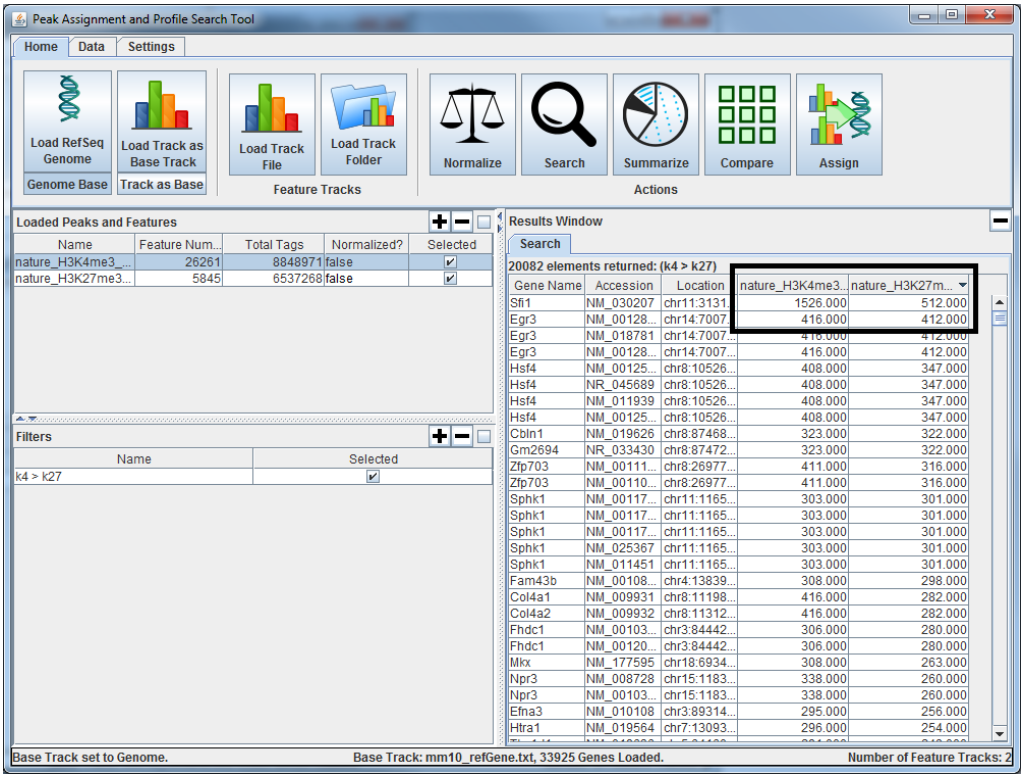
b. Set the first set to **nature_H3K4me3_peaks.xls** and the operation to '**- (subtract)**'. Set the second set to **nature_H3K27me3_peaks.xls** and set the relationship to > and the threshold to 0. Keep the default promoter region. Click OK.



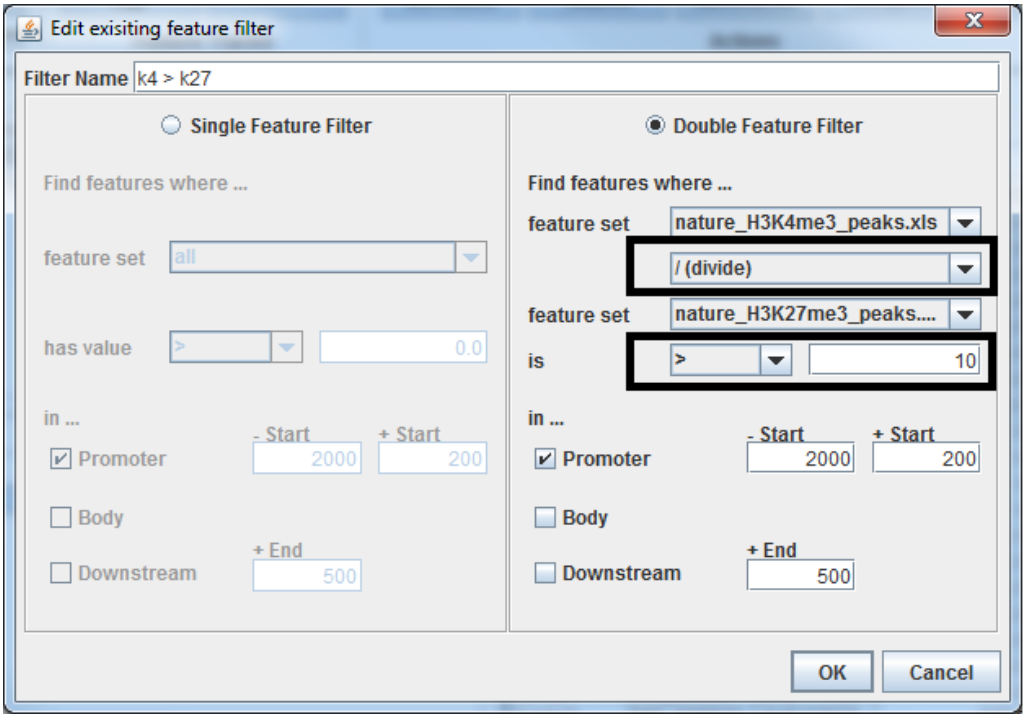
4. Click search to see the results. With this filter, PAPST takes the K4 value and subtracts the K27 value. Any regions where K4 – K27 > 0 will be returned. Genes which have more K27 are not returned.



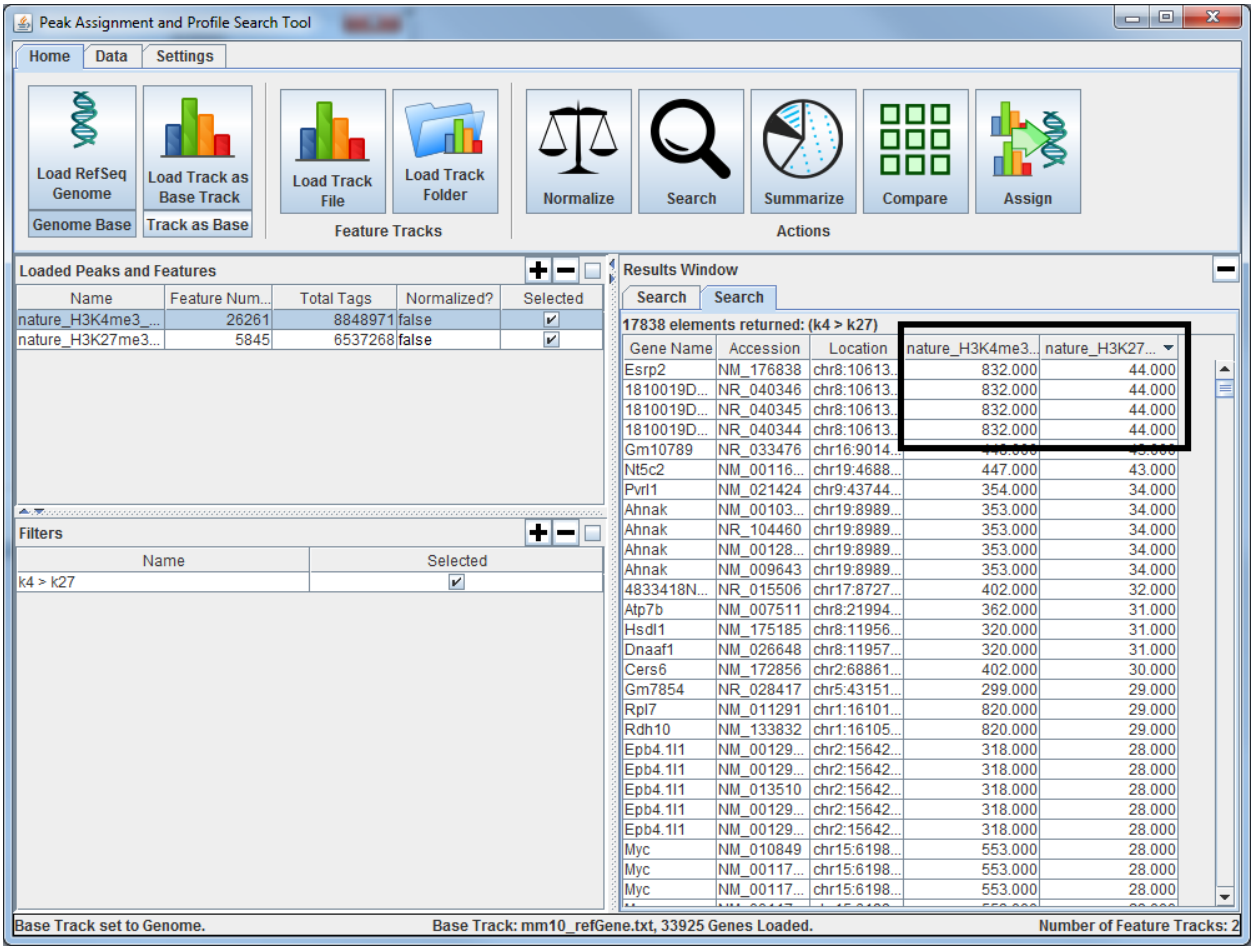
5. Sort the results by K27 to confirm that even the highest K27 region has a higher K4 region.



6. Let's edit the search to find gene that have 10 times as much K4 as K27. Double click to the filter to start editing. Change the operation to '**/ (divide)**' and set the threshold to 10.



7. Search again. Sort by the K27 value. Notice that even the highest K27 region has over 10x more K4.



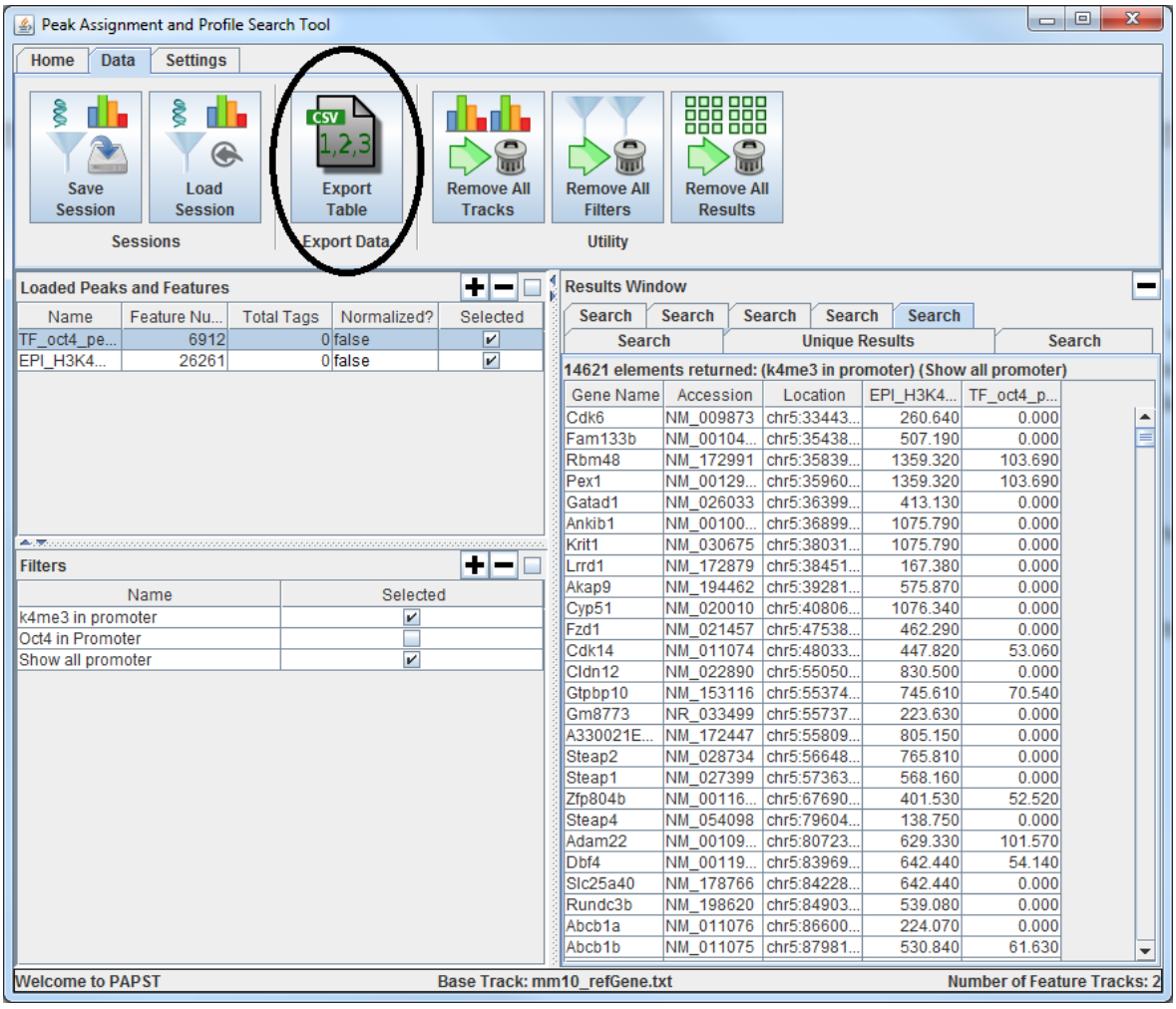
Note: Usually when you compare two peak values directly in a Two Peak filter, you should normalize the values. See the section on Normalization.

Working with Results

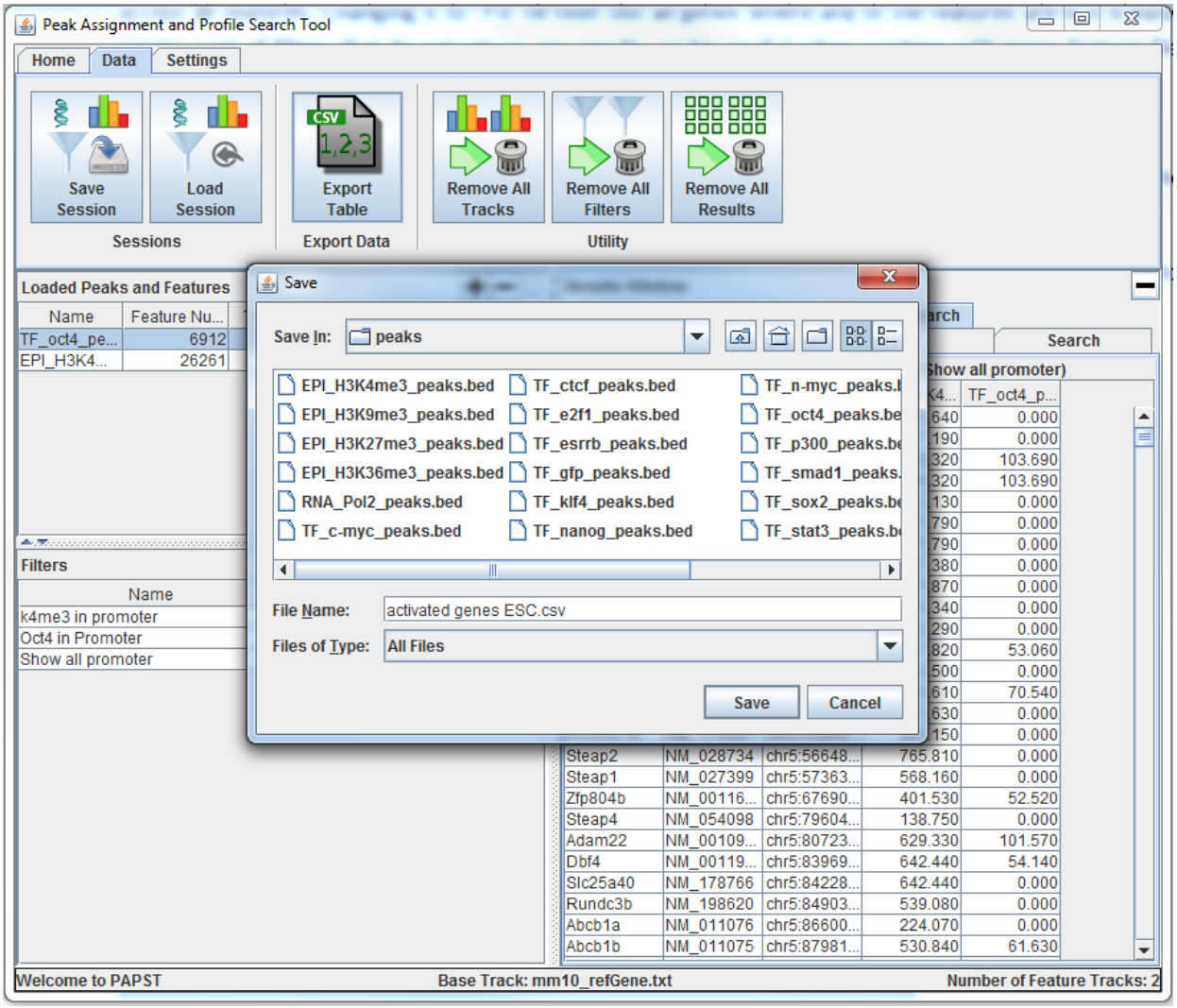
Exporting Results

PAPST enables the user to export interesting gene sets that match a particular pattern. Tables are exported in the comma separate values format, a common spreadsheet file type readable with Microsoft Excel.

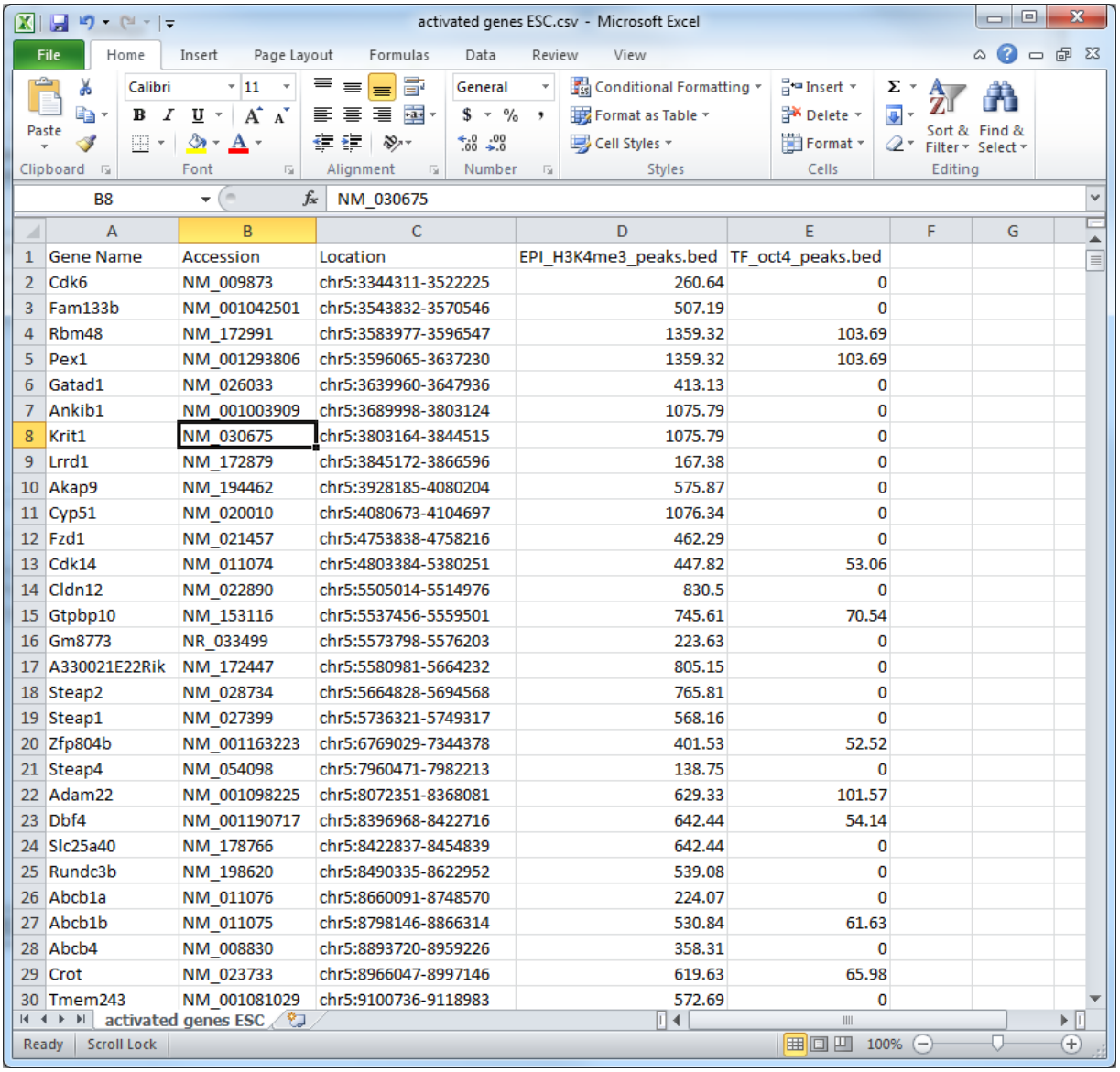
1. To export the current table, navigate to the **Data** tab, then click the **Export Table** button.



2. Next name the file ‘activated genes ESC.csv’. PAPST saves tables as portable comma separated values with a ‘.csv’ extension. This file type is opened by most spreadsheet programs.



3. Next open this file to view the contents. The values from PAPST appear in the spreadsheet ready to share with collaborators.

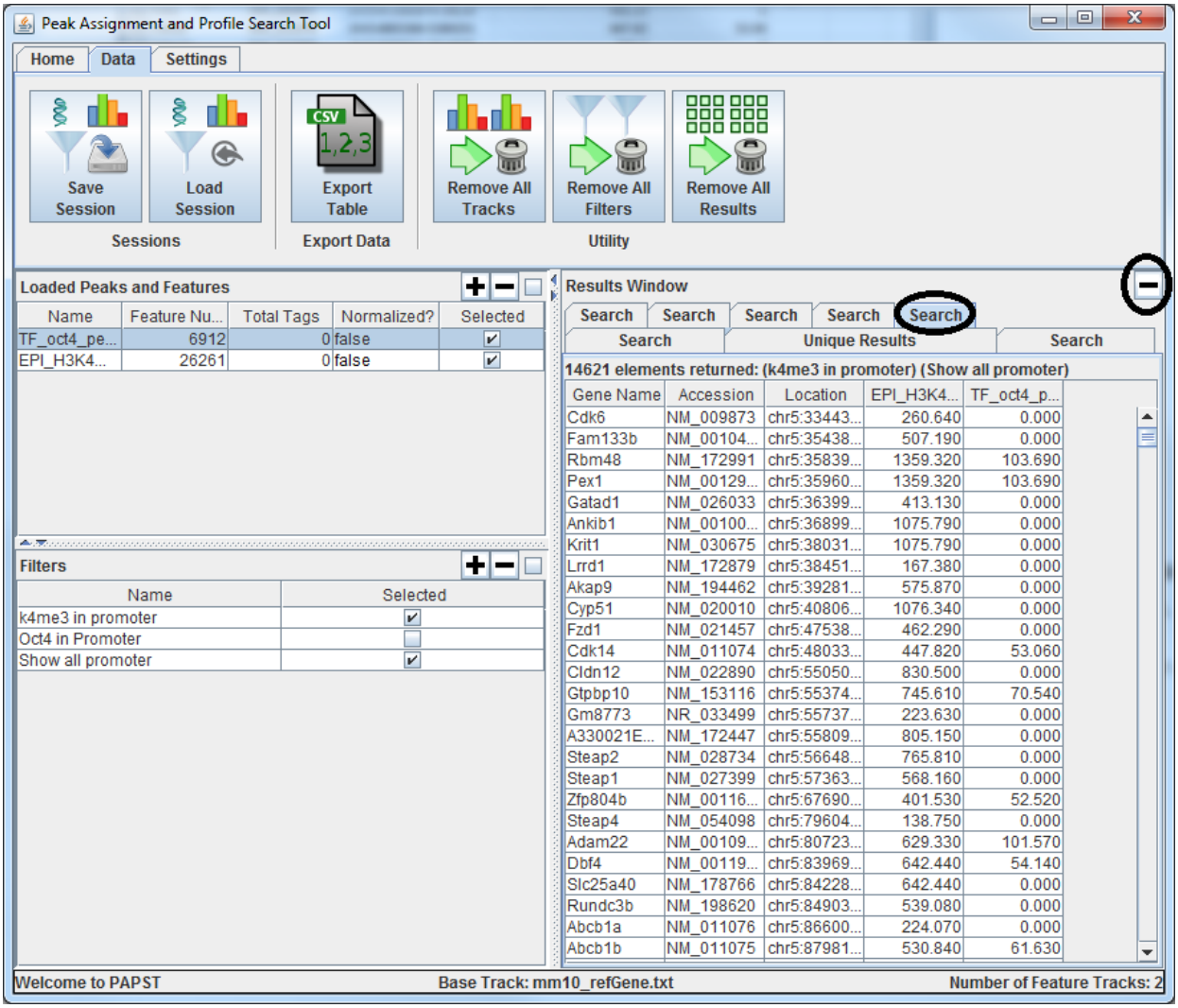


	A	B	C	D	E	F	G
1	Gene Name	Accession	Location	EPI_H3K4me3_peaks.bed	TF_oct4_peaks.bed		
2	Cdk6	NM_009873	chr5:3344311-3522225	260.64	0		
3	Fam133b	NM_001042501	chr5:3543832-3570546	507.19	0		
4	Rbm48	NM_172991	chr5:3583977-3596547	1359.32	103.69		
5	Pex1	NM_001293806	chr5:3596065-3637230	1359.32	103.69		
6	Gatad1	NM_026033	chr5:3639960-3647936	413.13	0		
7	Ankib1	NM_001003909	chr5:3689998-3803124	1075.79	0		
8	Krit1	NM_030675	chr5:3803164-3844515	1075.79	0		
9	Lrrd1	NM_172879	chr5:3845172-3866596	167.38	0		
10	Akap9	NM_194462	chr5:3928185-4080204	575.87	0		
11	Cyp51	NM_020010	chr5:4080673-4104697	1076.34	0		
12	Fzd1	NM_021457	chr5:4753838-4758216	462.29	0		
13	Cdk14	NM_011074	chr5:4803384-5380251	447.82	53.06		
14	Cldn12	NM_022890	chr5:5505014-5514976	830.5	0		
15	Gtpbp10	NM_153116	chr5:5537456-5559501	745.61	70.54		
16	Gm8773	NR_033499	chr5:5573798-5576203	223.63	0		
17	A330021E22Rik	NM_172447	chr5:5580981-5664232	805.15	0		
18	Steap2	NM_028734	chr5:5664828-5694568	765.81	0		
19	Steap1	NM_027399	chr5:5736321-5749317	568.16	0		
20	Zfp804b	NM_001163223	chr5:6769029-7344378	401.53	52.52		
21	Steap4	NM_054098	chr5:7960471-7982213	138.75	0		
22	Adam22	NM_001098225	chr5:8072351-8368081	629.33	101.57		
23	Dbf4	NM_001190717	chr5:8396968-8422716	642.44	54.14		
24	Slc25a40	NM_178766	chr5:8422837-8454839	642.44	0		
25	Rundc3b	NM_198620	chr5:8490335-8622952	539.08	0		
26	Abcb1a	NM_011076	chr5:8660091-8748570	224.07	0		
27	Abcb1b	NM_011075	chr5:8798146-8866314	530.84	61.63		
28	Abcb4	NM_008830	chr5:8893720-8959226	358.31	0		
29	Crot	NM_023733	chr5:8966047-8997146	619.63	65.98		
30	Tmem243	NM_001081029	chr5:9100736-9118983	572.69	0		

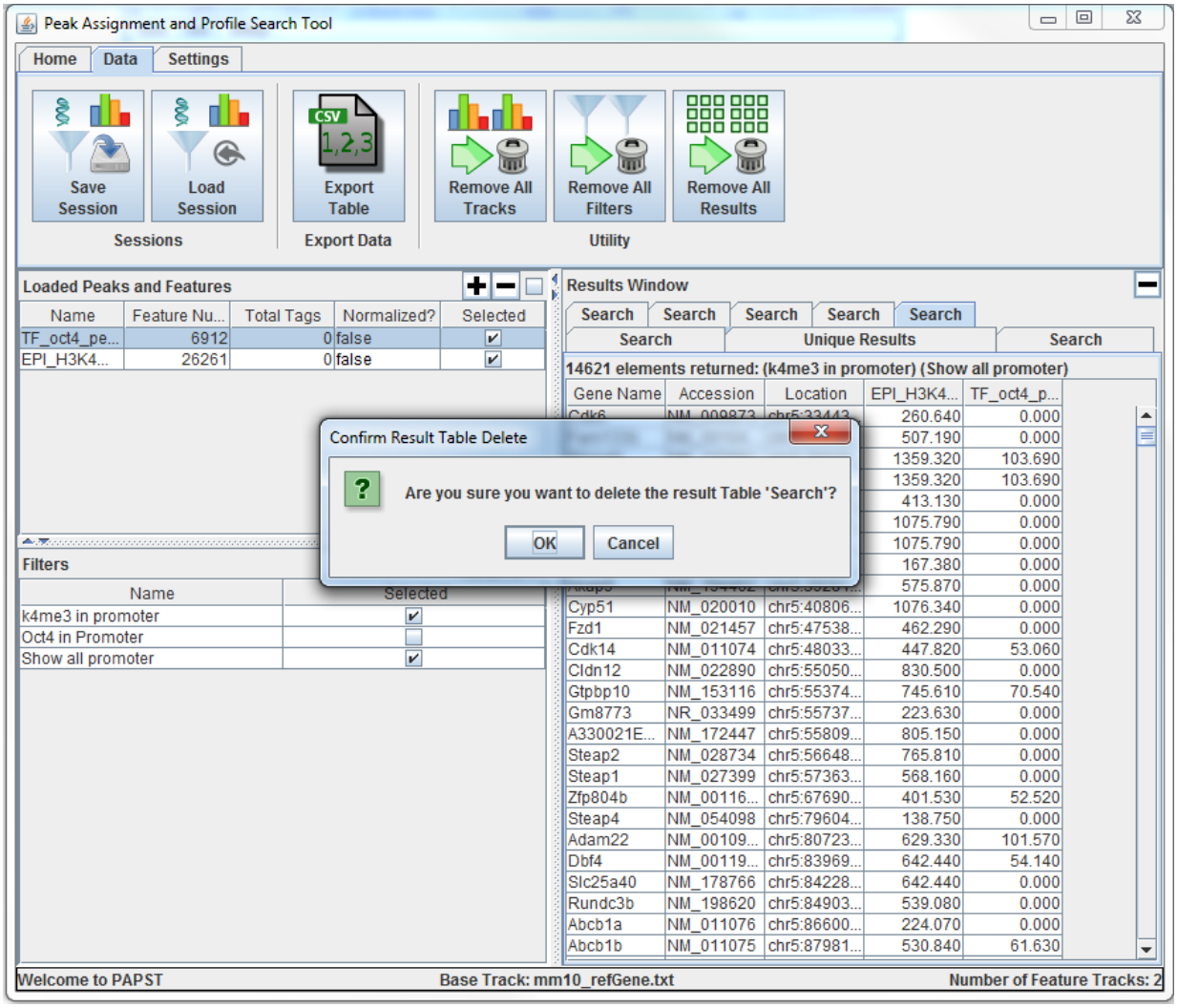
Deleting Result Tables

We have a number of tables in our results window. Our tabs are getting crowded and hard to see. We can delete these results without worrying because they can be easily generated with another search.

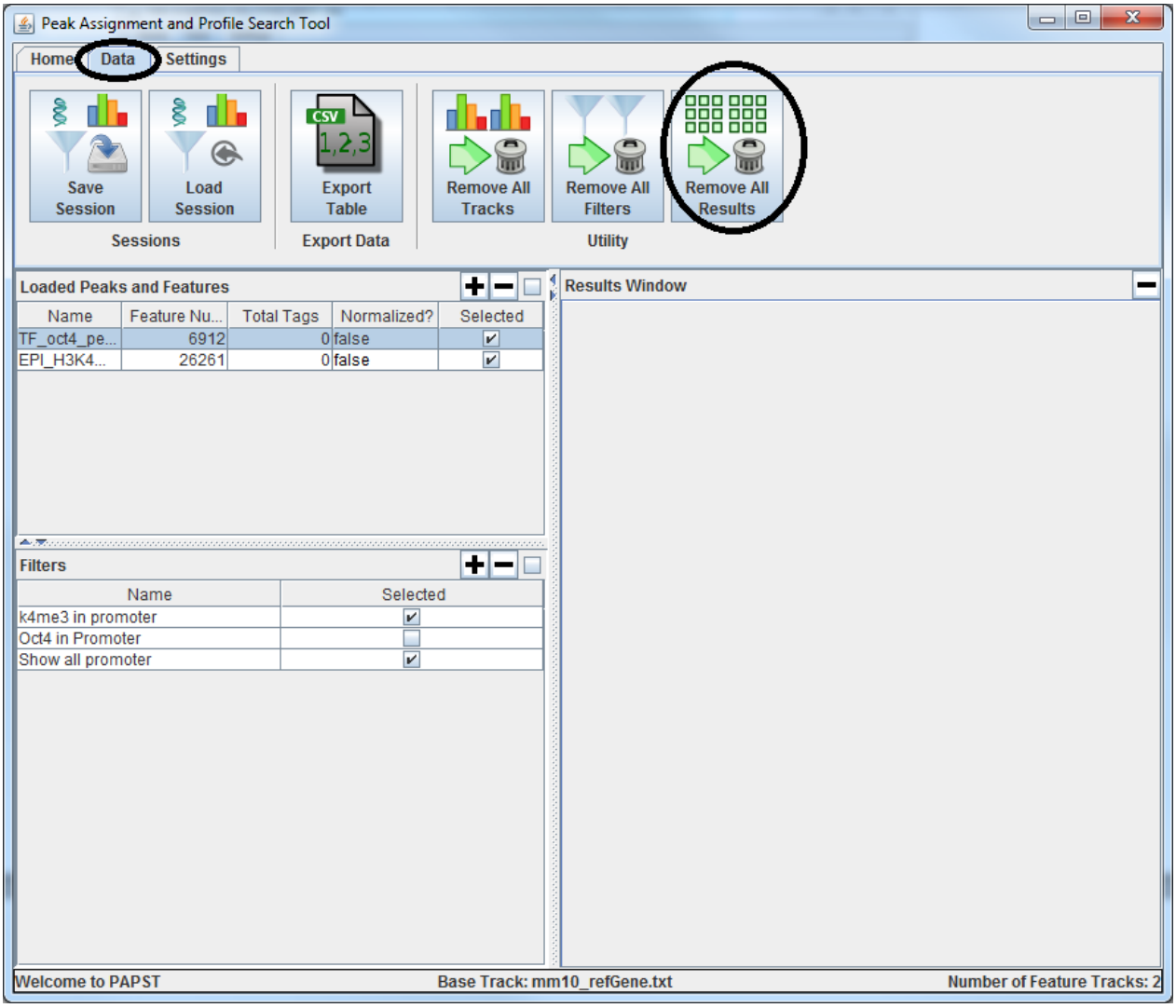
1. To delete the selected results tab, click on the **minus ('-')** button in the top right-hand corner of the **Results** window. It will delete the current selected tab.



2. After clicking the remove button, you will be asked to confirm removing the results table. Click OK to remove the table.



3. To remove all results at once, navigate to the **Data** tab and click **Remove All Results**.



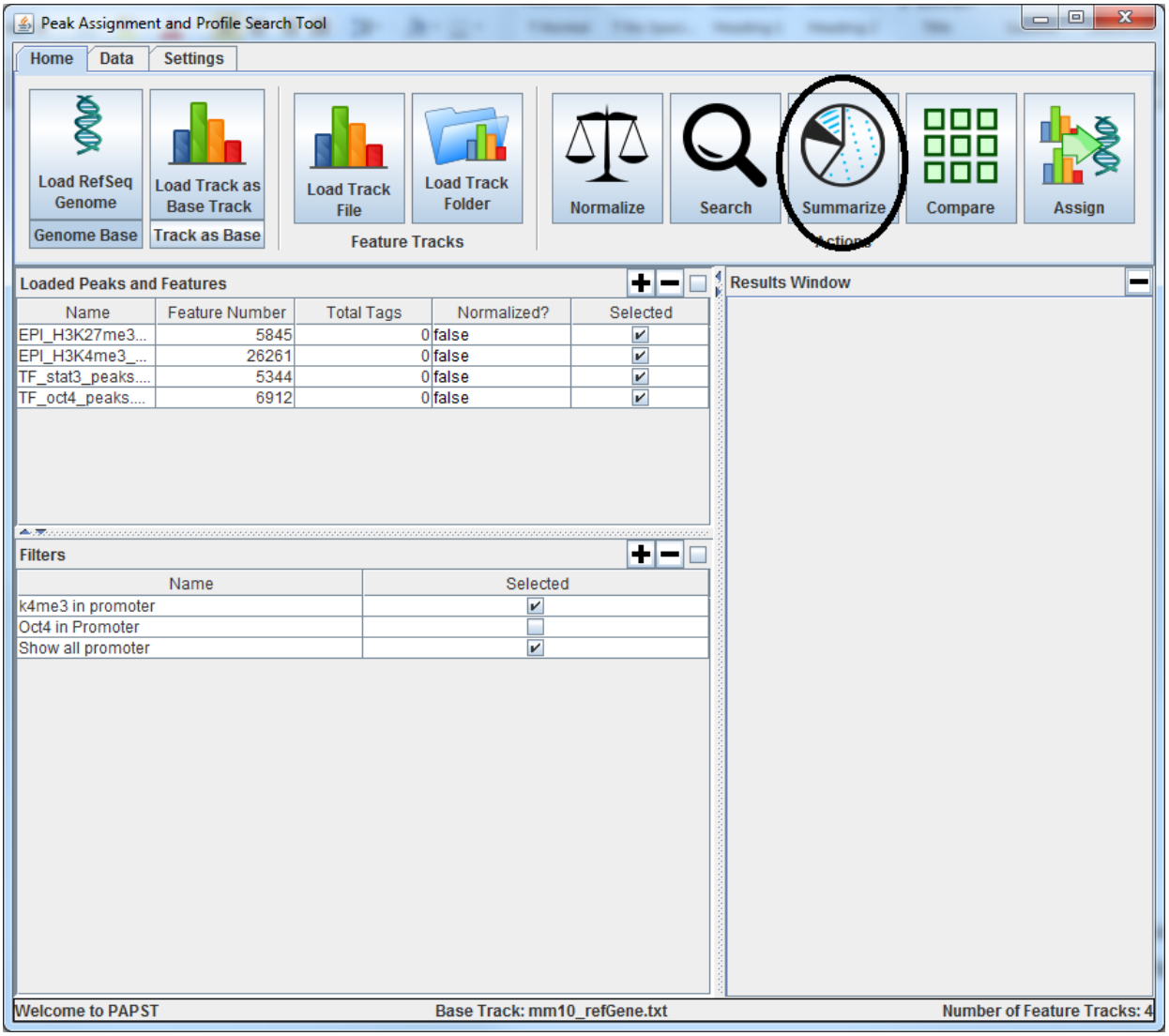
All results are now removed.

General Feature Analysis in PAPST

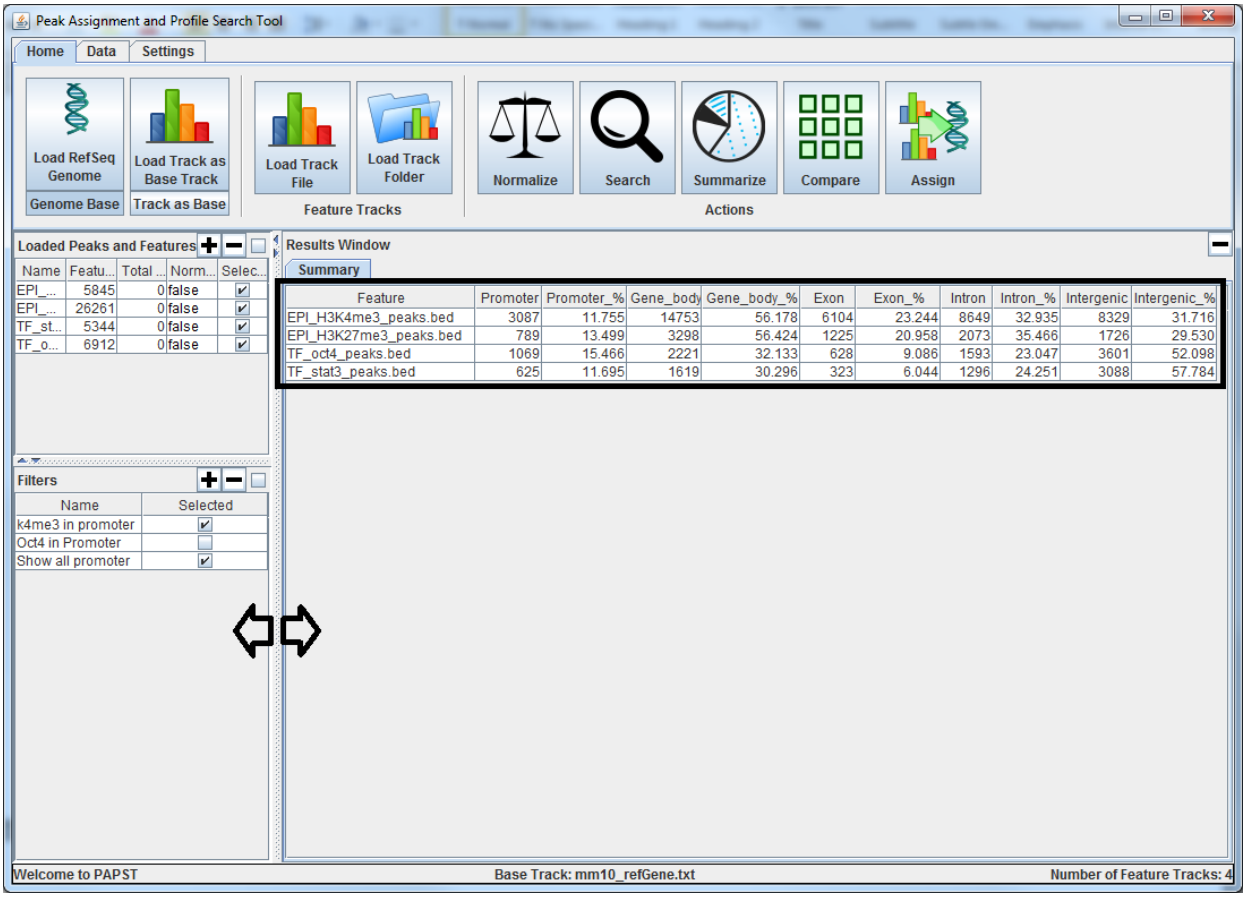
Peak Distributions

The distribution of peaks gives clues to its behavior. Often we will want to know if a protein binds primarily to promoters or to distal regions. PAPST can calculate this information quickly for multiple peak sets.

1. Load **EPI_H3K27me3_peaks.bed** and **TF_stat3_peaks.bed** into PAPST.
2. Click on the **Summarize** button to display a summary of the peaks' distributions.

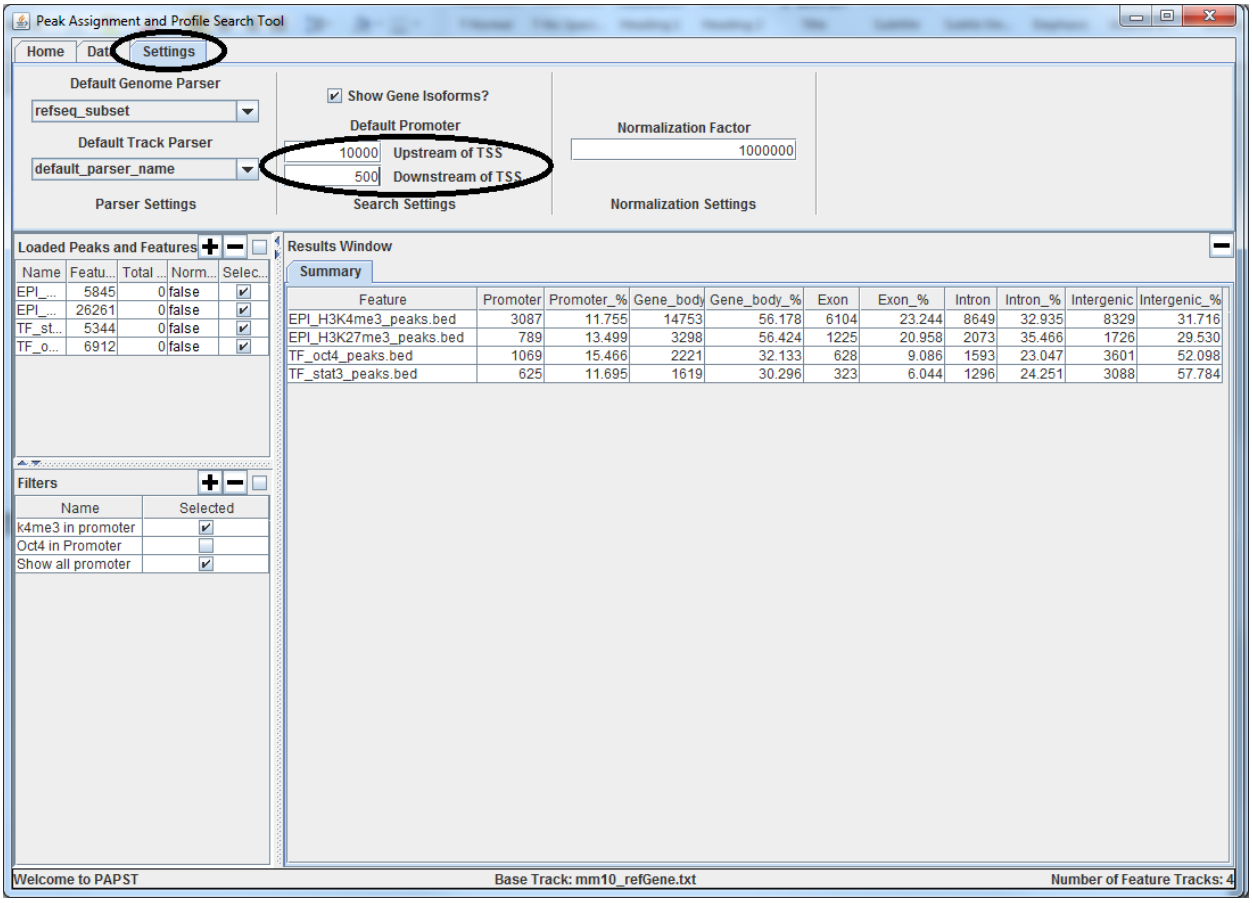


3. The summary results will be displayed in the **Results Window**. You can **move the pane separator** to create more viewing room in the results window.

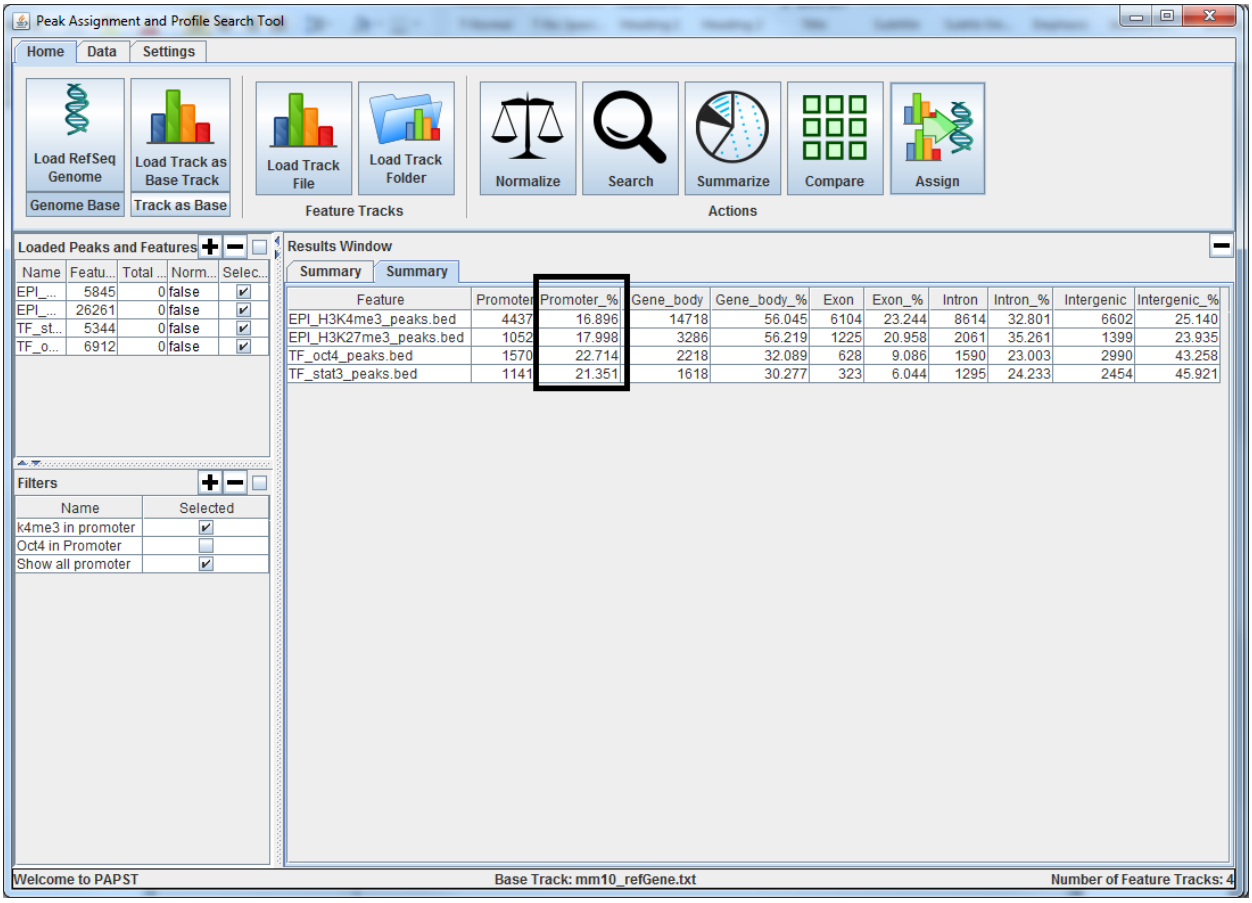


Notice that K4 and K27 more closely associated which gene bodies and promoters. The two transcription factors have more intergenic binding. **The gene body's percent includes the exon and intron percentages.**

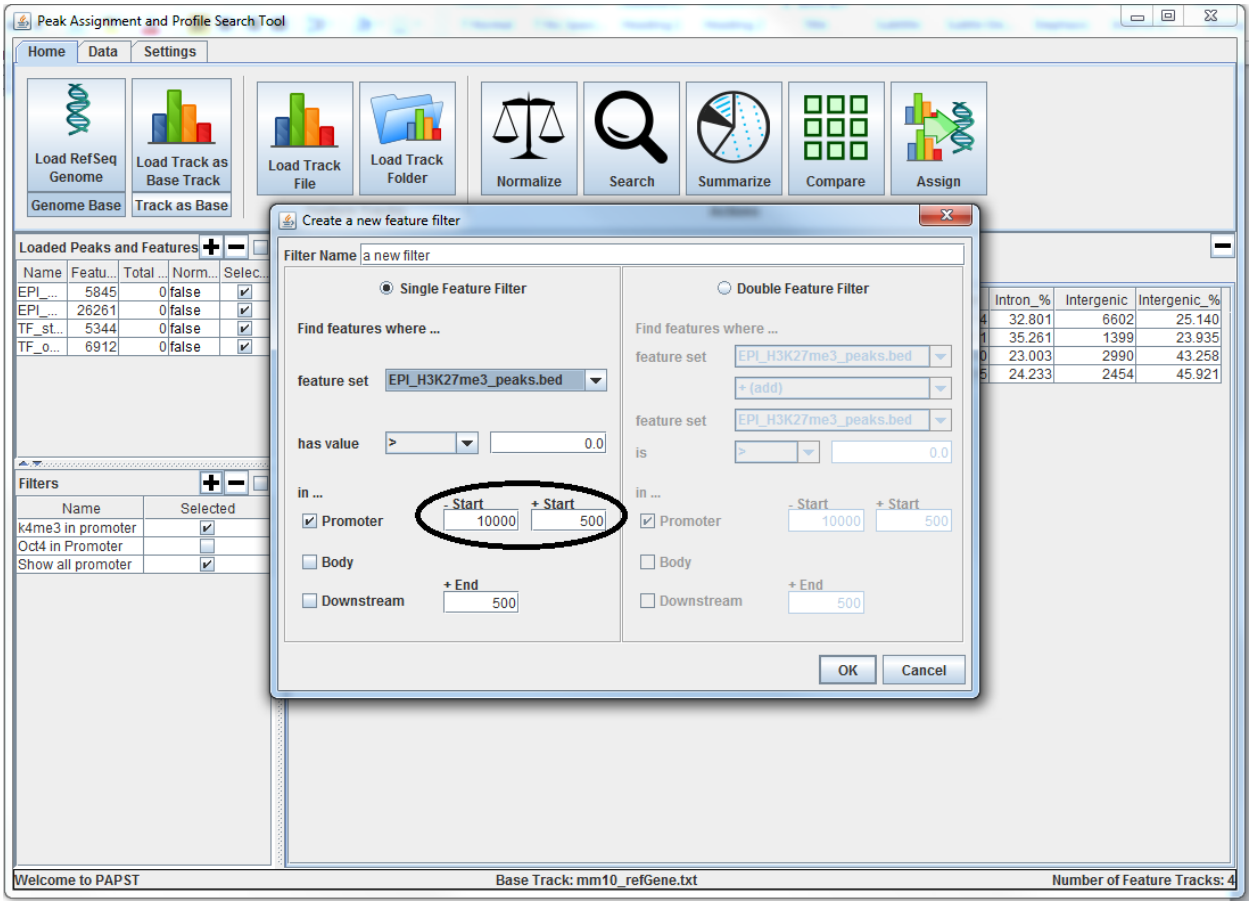
4. To adjust the promoter setting for **Summary**, Navigate to the **Settings** tab. Let's change the promoter settings to be from -10,000 to +500 relative to the TSS. Press **Summarize** again.



5. The results will show new percentages. The summary naturally reports more promoter peaks based on the new definition provided.



6. Changing the promoter setting also changes the default filter setting. This lets the user define the promoter once for the whole program.

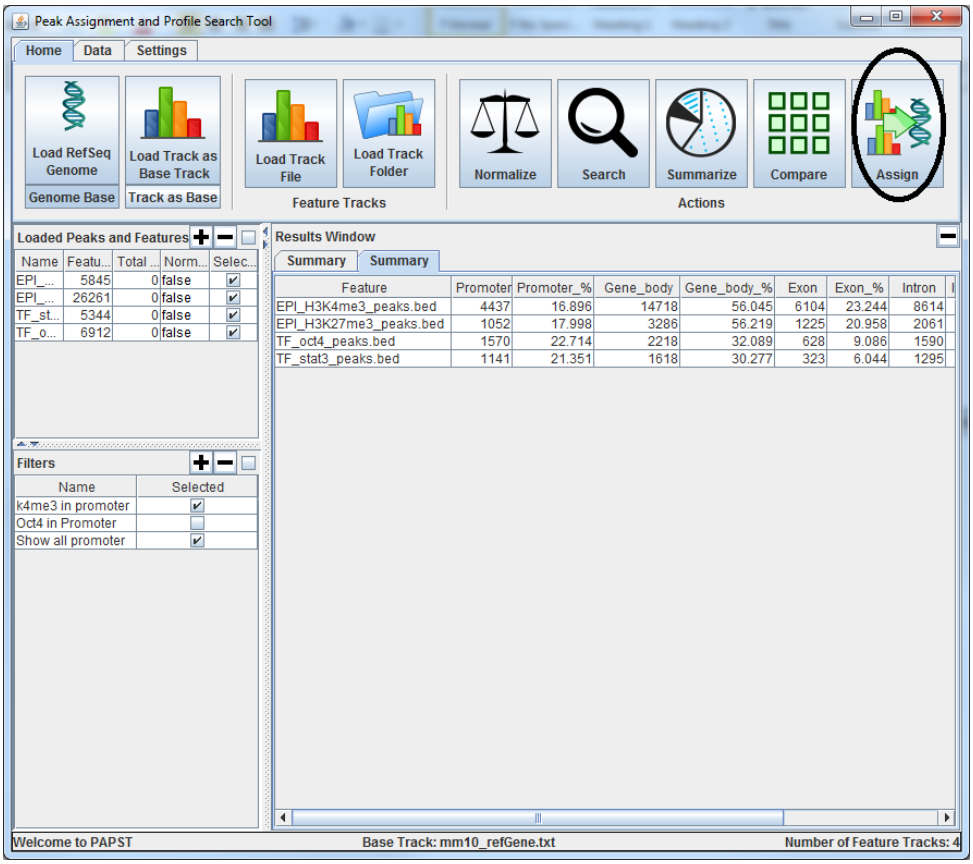


Now a new filter will use the newly defined promoter region.

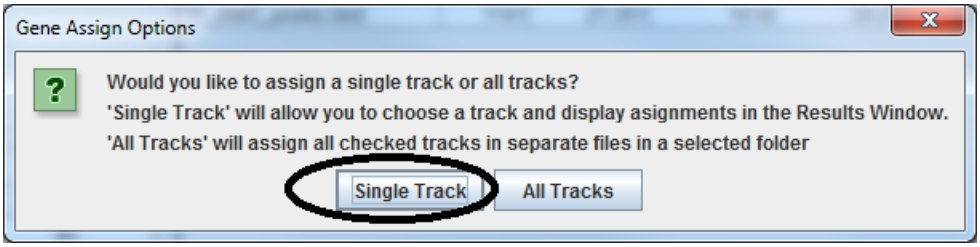
Assign Peaks to Their Nearest Genes

PAPST can quickly assign peaks to their nearest gene. The program will choose the nearest TSS.

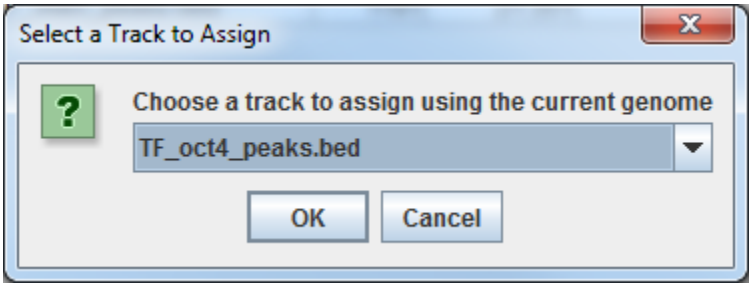
1. To assign a **single feature set** to genes, click on the **Assign** button on the **Home** tab.



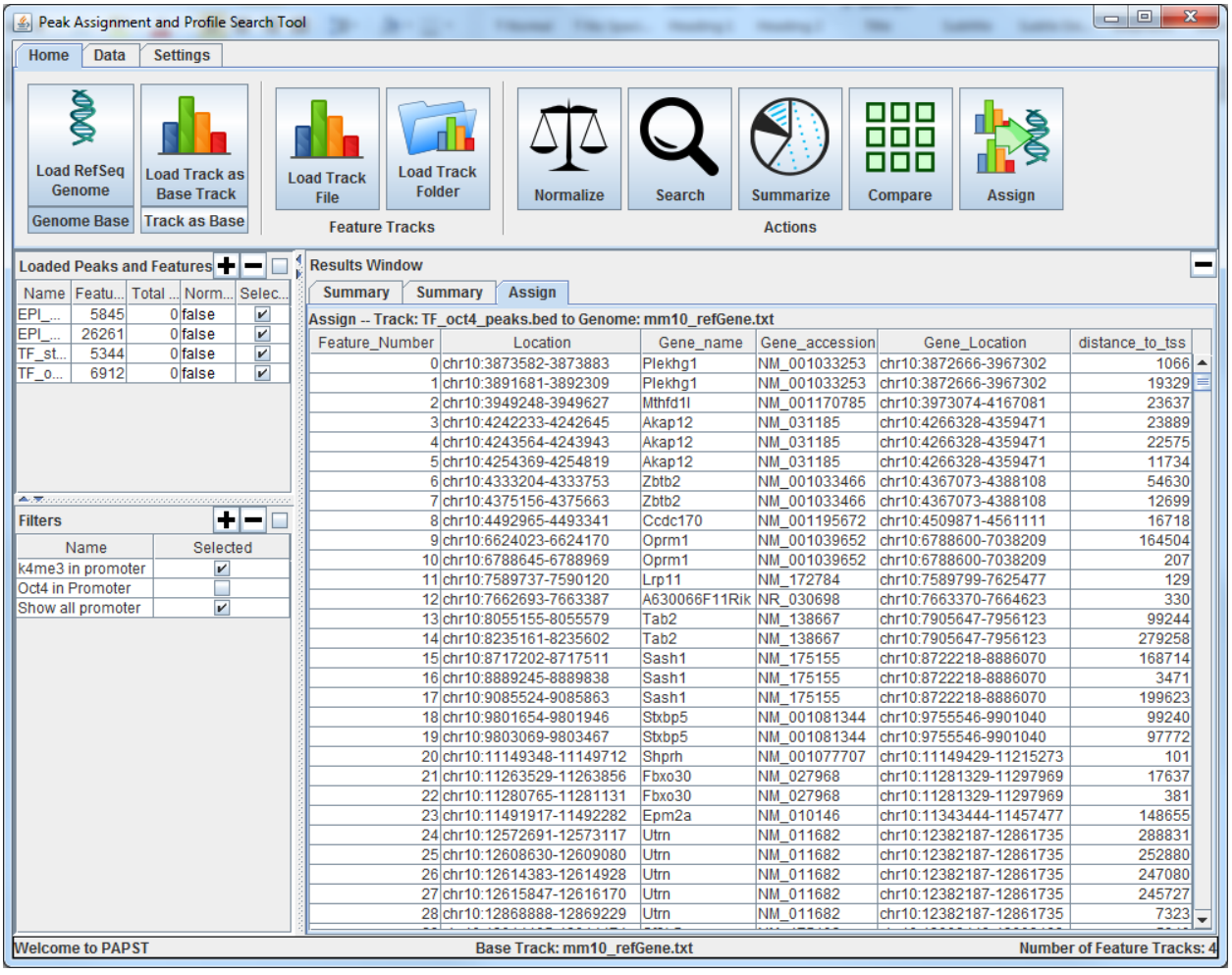
2. You may choose to assign all feature sets or a single feature set. Choose **Single Track**.



3. Next select TF_oct4_peaks.bed. Click OK.

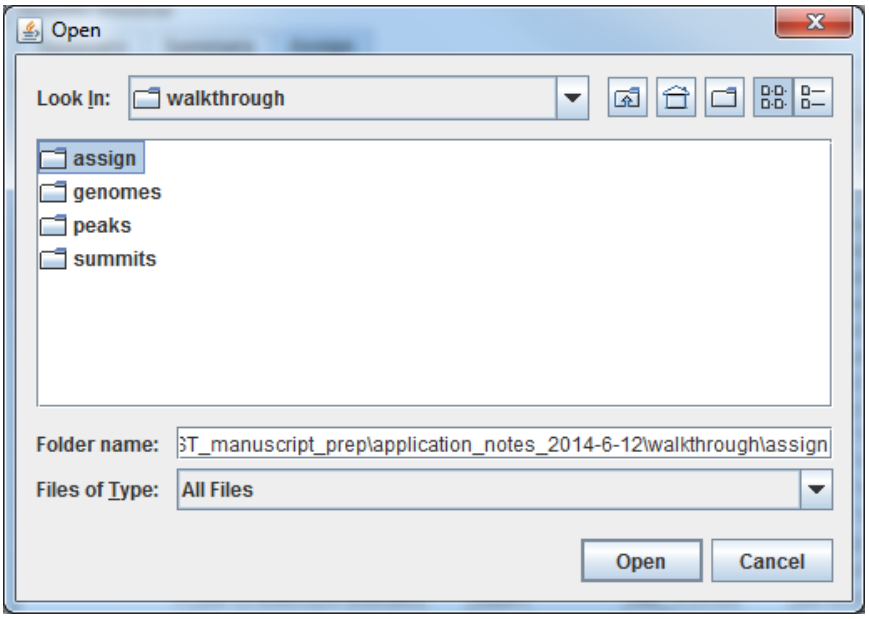


4. The results will be displayed in the **Results Window.** Each peak is given a feature number. **Assign** reports peak location, gene location, and distance to TSS as well as Refseq identifier and gene name.

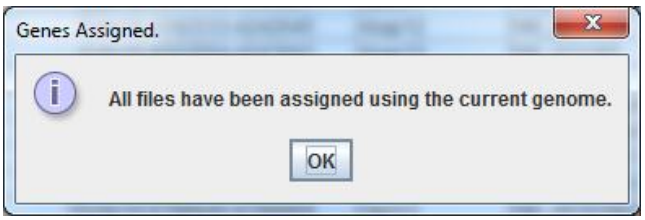


5. To assign all loaded peaks at once, click the **Assign** button, and then select **All Tracks**.

6. You will be prompted for a file location to save the results. PAPST will save each feature into a separate file.



7. A dialog confirms that all features have been assigned.



8. The files will be tab separated files with same information as the single track version. For example:

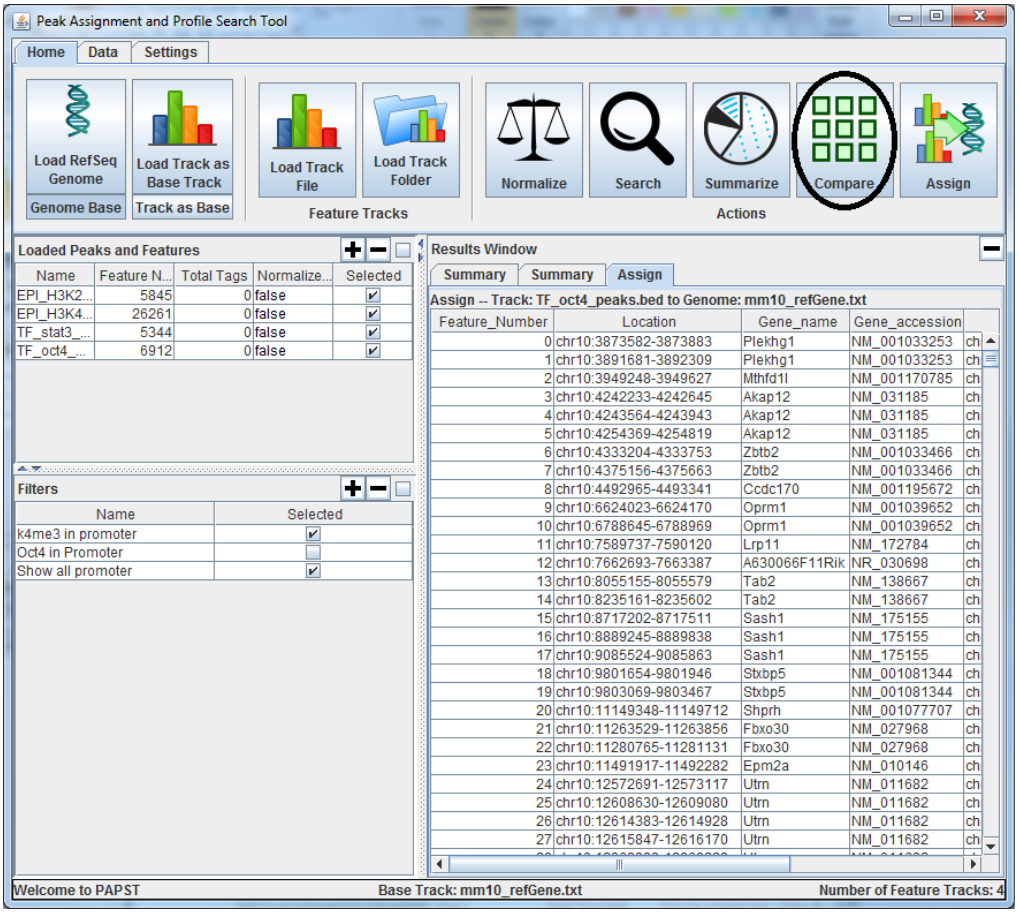
Feature_Number	Location	Gene_name	Gene_accession	Gene_Location	distance_to_tss
0	chr10:3116879-3117258	B020014A21Rik	NR_045946	chr10:3125090-3133193	16125
1	chr10:3174041-3175260	B020014A21Rik	NR_045946	chr10:3125090-3133193	41457
2	chr10:3365613-3367828	Ppp1r14c	NM_133485	chr10:3366149-3464975	571
3	chr10:3739835-3741818	Plekhg1	NM_001159942	chr10:3740376-3967302	450
4	chr10:3872410-3874366	Plekhg1	NM_001033253	chr10:3872666-3967302	722
5	chr10:3972282-3976100	Mthfd1l	NM_001170785	chr10:3973074-4167081	1117

These files can be easily viewed in a spreadsheet or regular text editor.

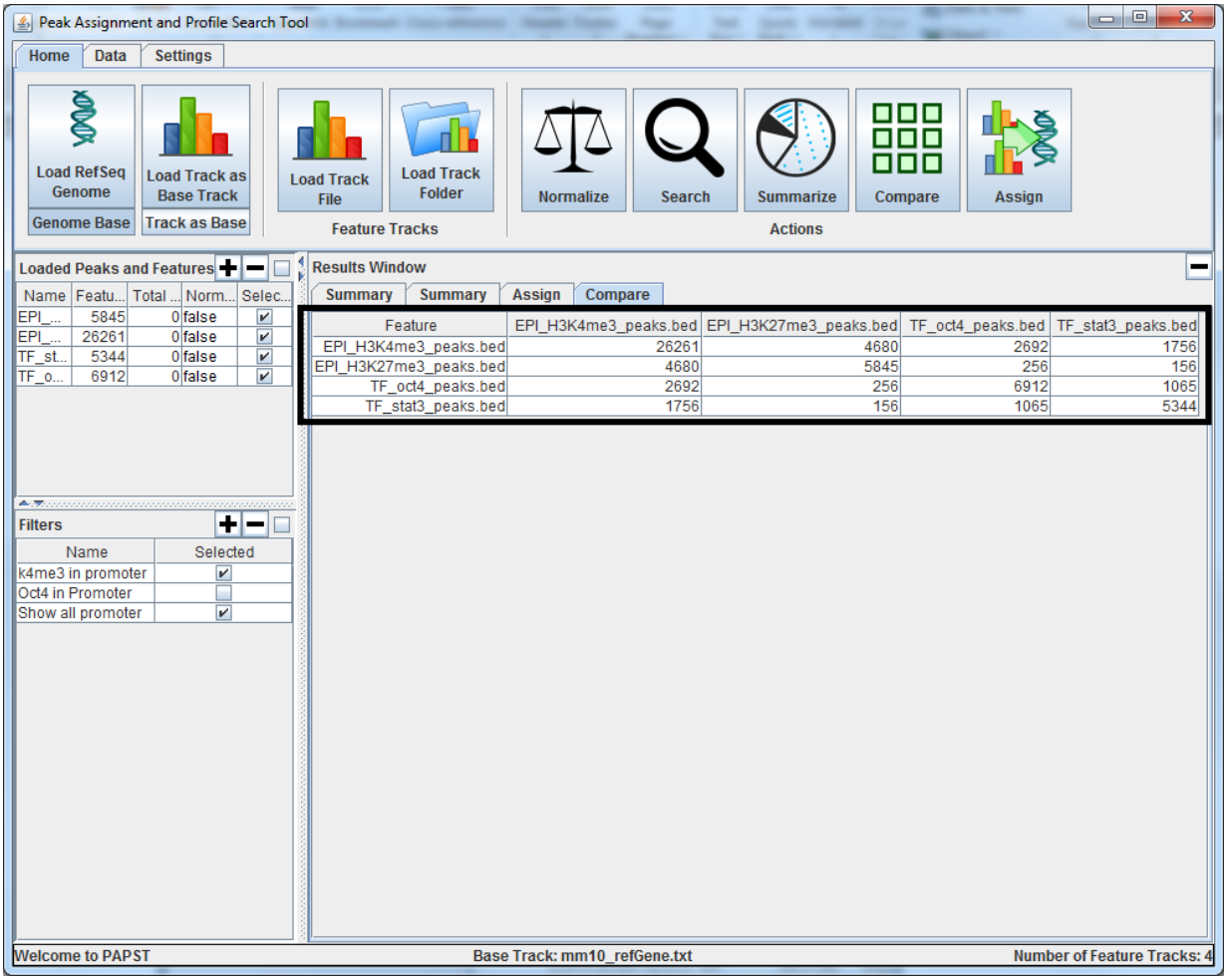
Comparing Peaks

When dealing with multiple peak sets from many conditions, it is helpful to quantify the overlap between each set. PAPST provides this functionality with the **Compare** button on the **Home** tab.

1. To compare different peaks or features by overlap, click the **Compare** button on the **Home** tab.



2. A symmetric matrix will be created showing the number of shared overlaps between all the enabled peaks in PAPST.



3. PAPST uses a symmetric overlap measure which counts the number of overlaps to ensure that the number from ‘A overlapping B’ is equal to ‘B overlapping A’. This facilitates further downstream analysis such as clustering.

Normalization

Because sequencing depth varies between experiments, directly comparing peak values can be misleading without normalization. PAPST normalizes peak tags using the following formula:

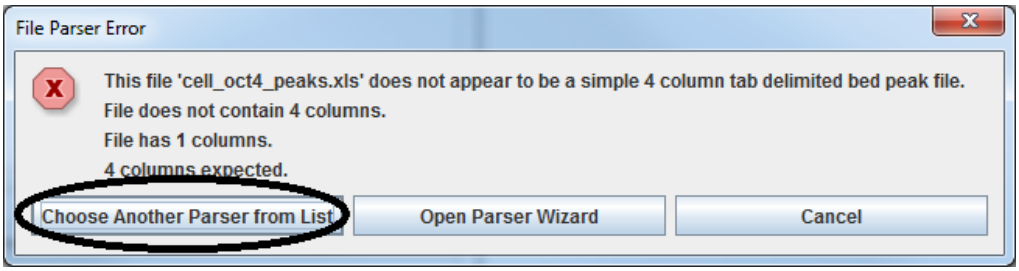
$$Normalized\ Tag\ Score_p = \frac{\#\ of tags\ for\ peak\ \ p}{Total\ \#\ of\ tags\ in\ all\ peaks} * 1{,}000{,}000$$

In this formula, 1,000,000 is the normalization factor. You can change the normalization factor on the **Settings** tab.

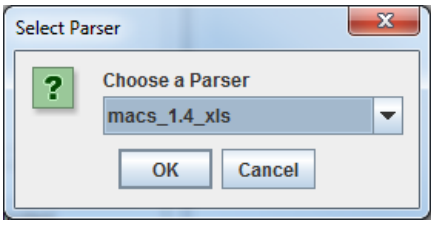
Advanced: Normalize Values by Sequencing Depth

Peaks called with MACS 1.4.2 in the xls format report tag counts within peaks. This section will show you how to normalize MACS 1.4.2 peaks using total tag number.

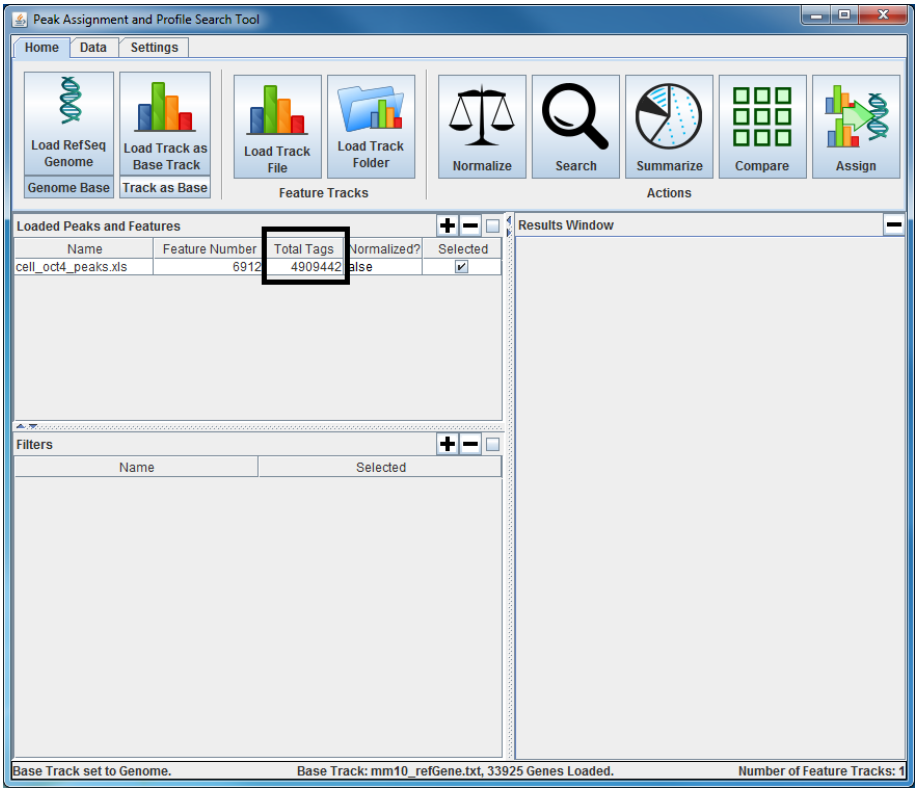
1. Load the **mm10_refGene.txt** into papst.
2. Load the peak file **cell_oct4_peaks.xls** under peaks/macs_xls/ in the walkthrough folder. The MACS 1.4 parser is not enabled by default. Click **Choose Another Parser from List**.



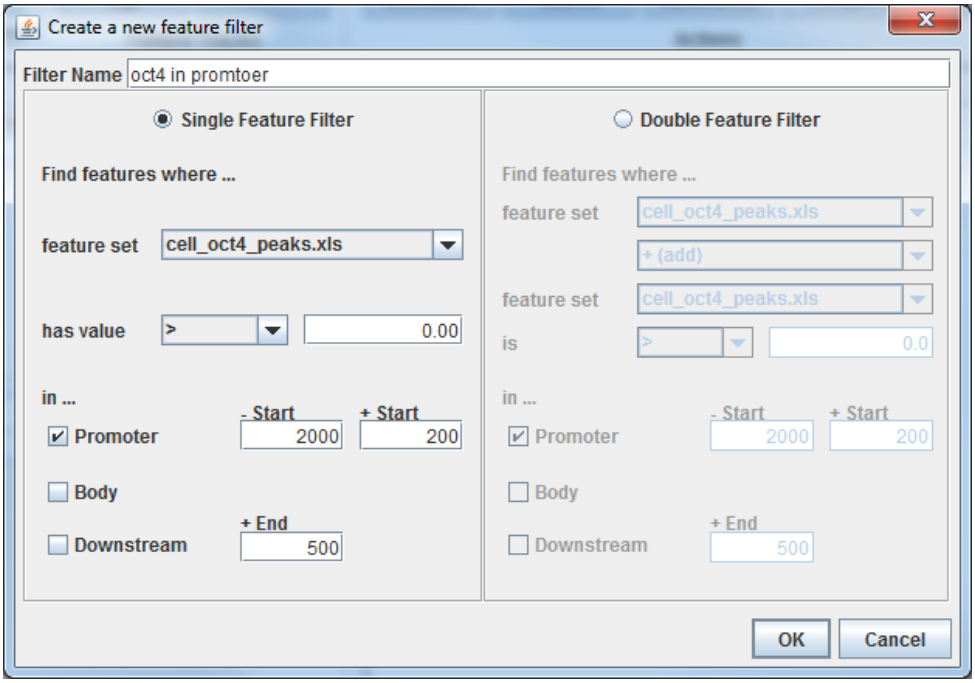
3. Choose the **macs_1.4_xls** parser from the dropdown box and click Ok.



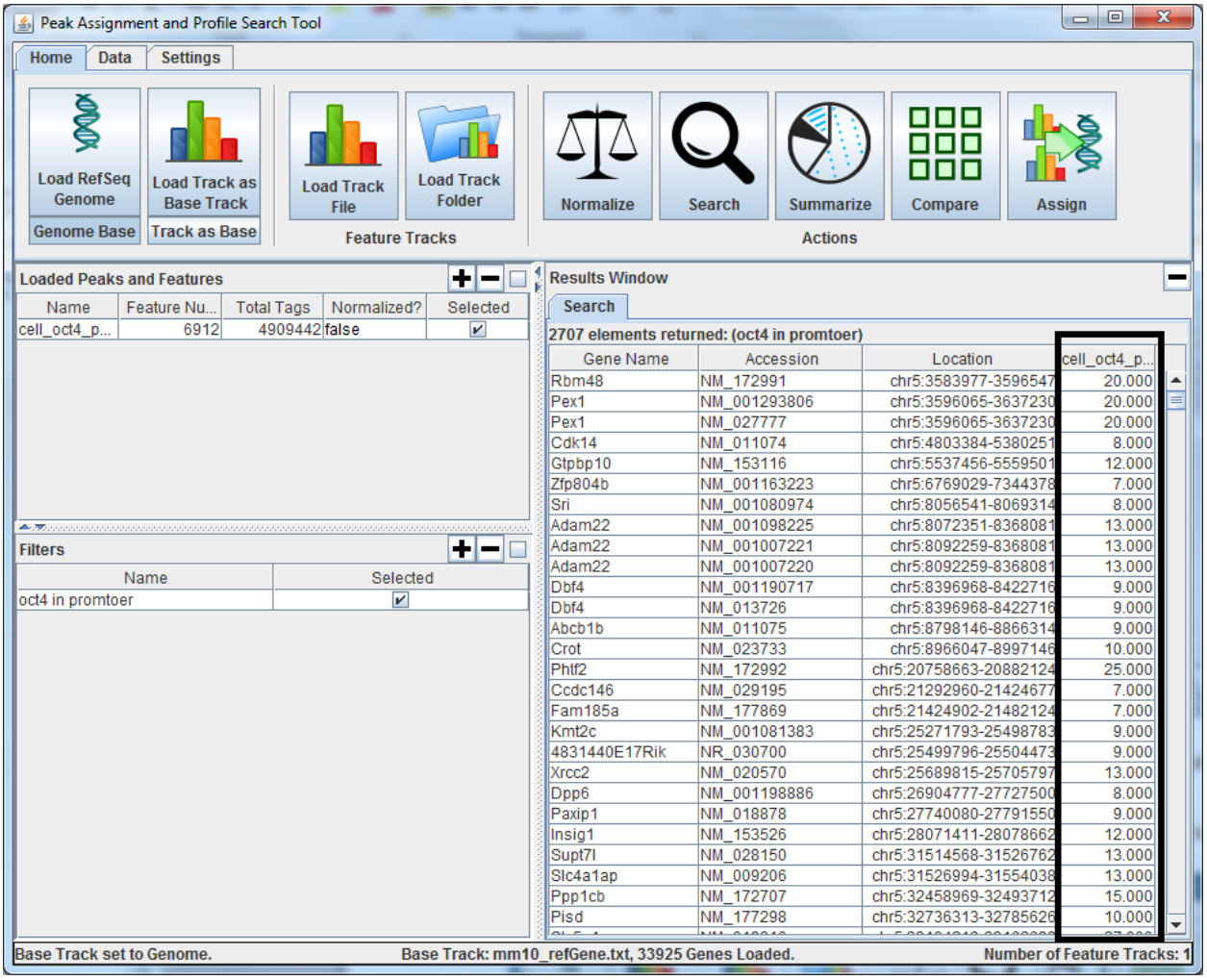
4. The total tags will be displayed for the peak set when it is loaded. Only Macs 1.4 xls files provide the total tag number.



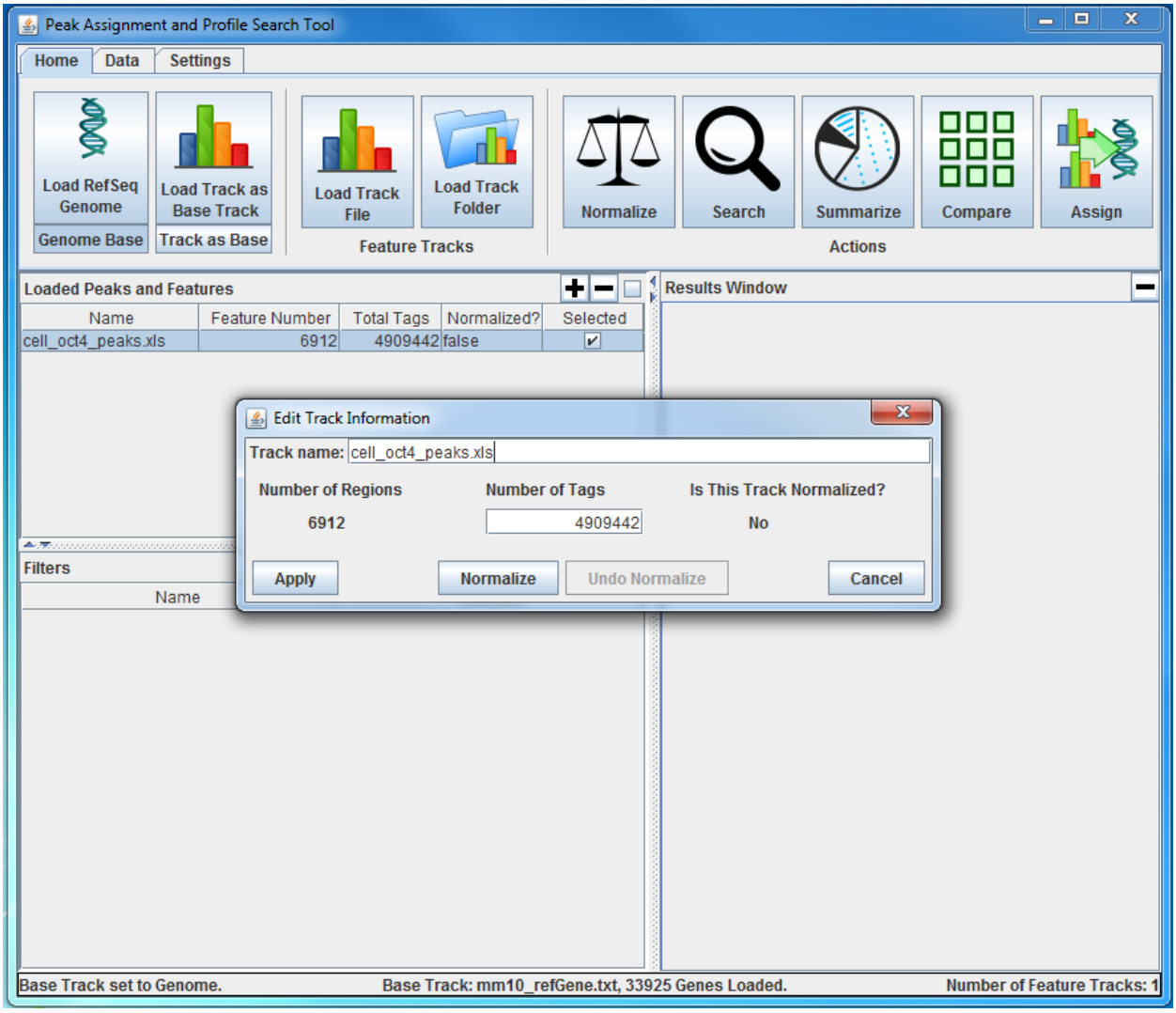
5. Before normalizing, let's take a look at results from an un-normalized search. Create a new filter for Oct4 to find genes in the promoter region -2000bp +200bp.



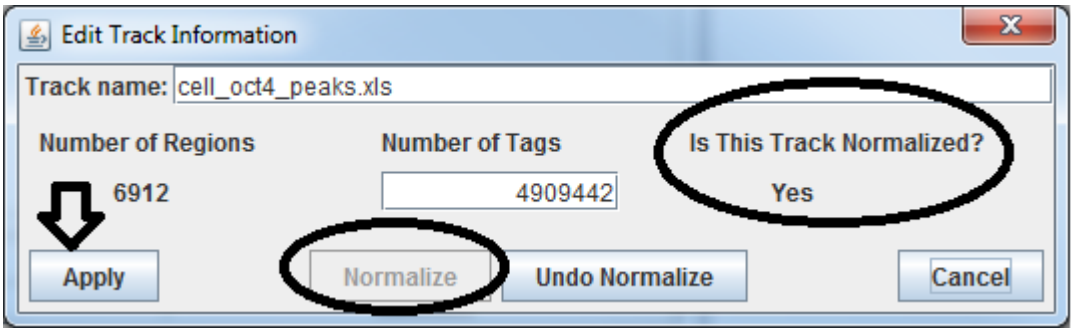
6. Click the **Search** button to show genes with Oct4 binding their promoters. Notice the range of values.



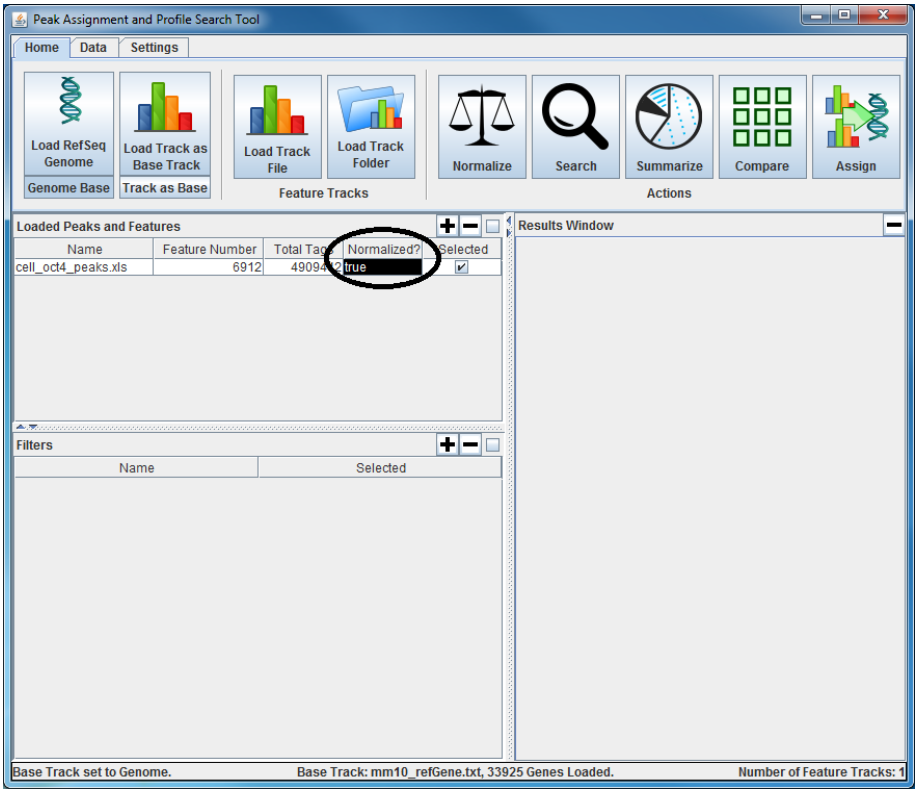
7. Now double-click the peak set that was just loaded. A track editor will appear.



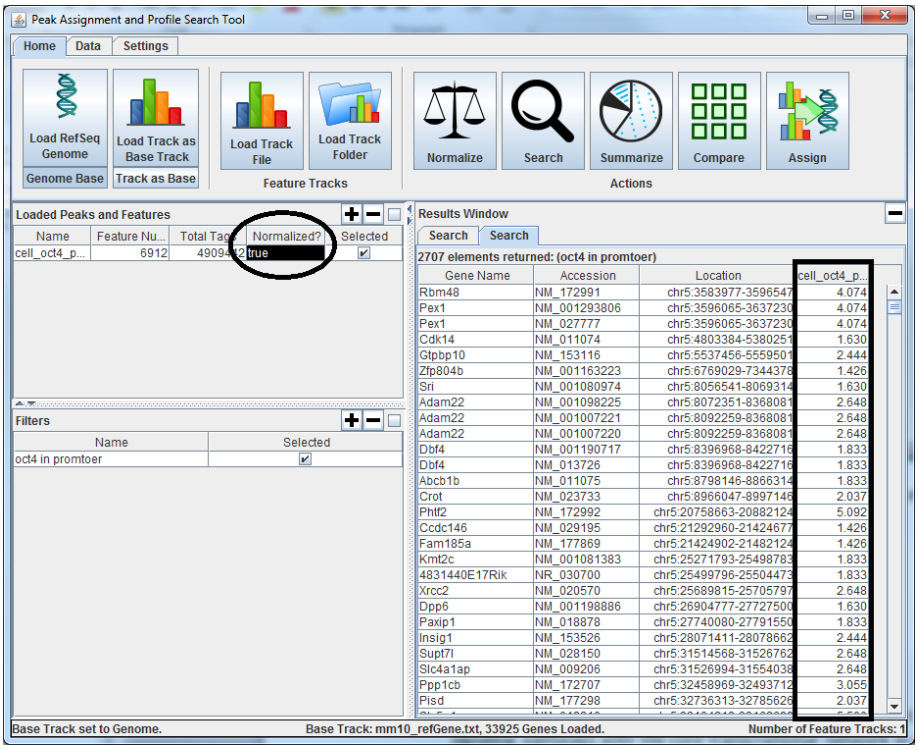
8. This editor enables the user to change the label of the peak set, edit the tag number, and normalize the values. Click the **Normalize** button. When the peaks are normalized, the normalize option will be unavailable. Click **Apply** to commit these changes to the current peak.



9. The peaks will now be normalized. This will be indicated by a black highlighted cell.



10. Click the **Search** button again. The value returned will reflect their normalized tag counts.

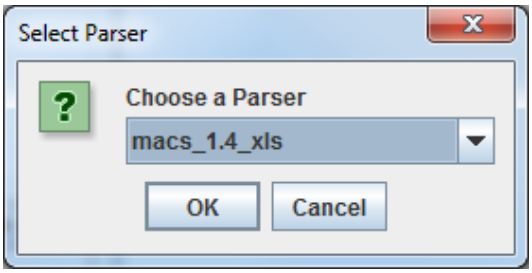


Advanced: Normalize Multiple Peaks Sets at Once

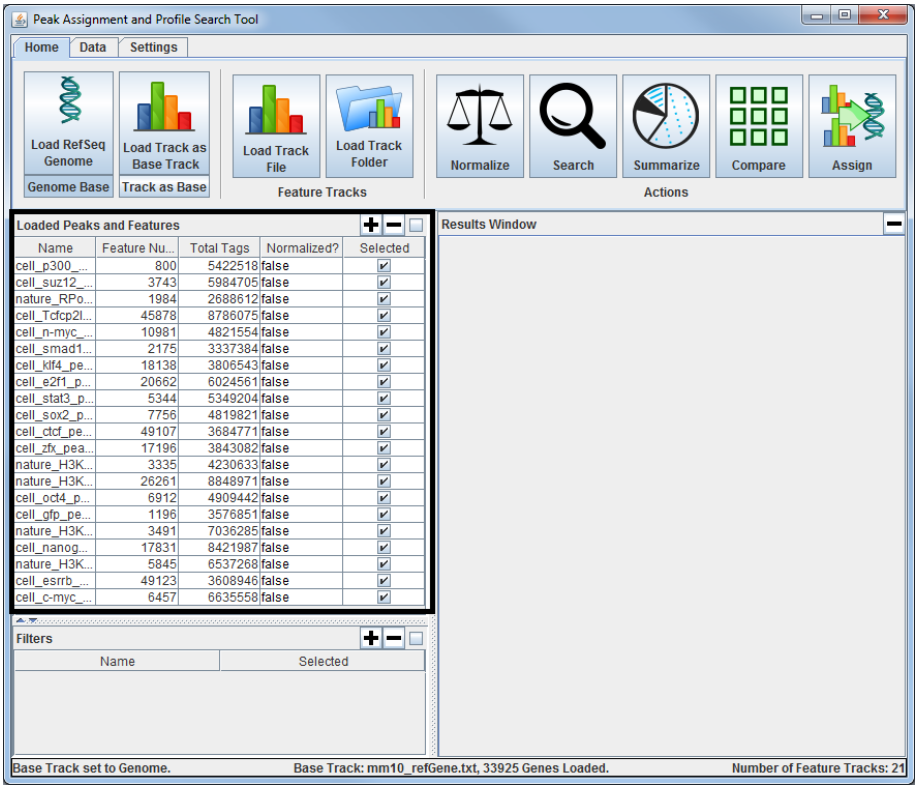
PAPST provides a simple method to normalize many peak sets at once. This will require the tag counts for each sample used to generate the peak set. This example demonstrates how to normalize many tracks at once.

1. Load the **mm10_refGene.txt** gene models.
2. Click '**Load Track Folder**' and select the folder **/peaks/macs_xls**
3. If the prompted, click '**Choose Parser from List**'

4. Select the **macs_1.4_xls** parser and click OK.



5. All peak sets in the folder will be loaded.

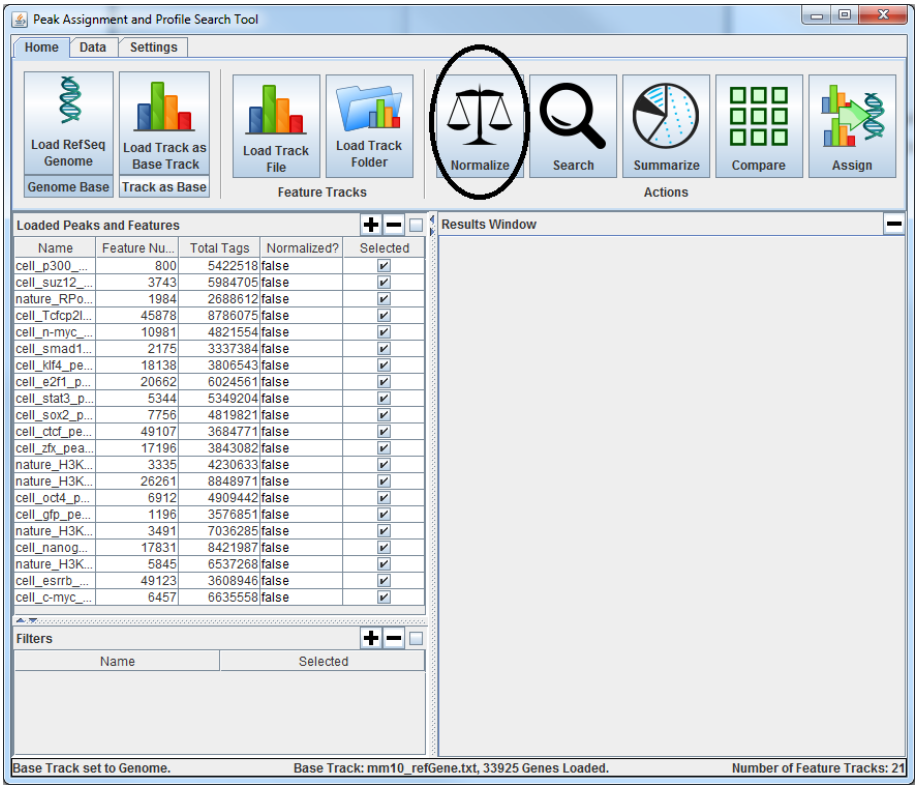


6. **Create a tab delimited file with the peak set names followed by the tag count**. Use one line per peak set like below. The example file, **tag_counts.txt**, is provided in the walkthrough folder.

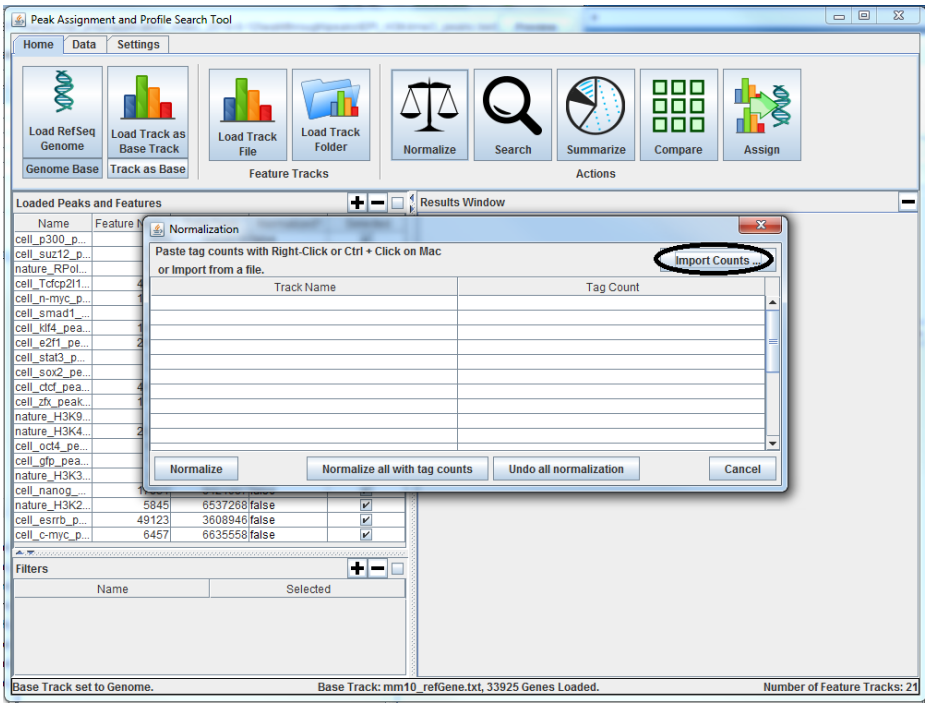
tag_counts.txt

cell_Tcfcp2l1_peaks.xls 8786075
cell_c-myc_peaks.xls 6635558
cell_ctcf_peaks.xls 3684771
cell_e2f1_peaks.xls 6024561
cell_esrrb_peaks.xls 3608946
cell_gfp_peaks.xls 3576851
…

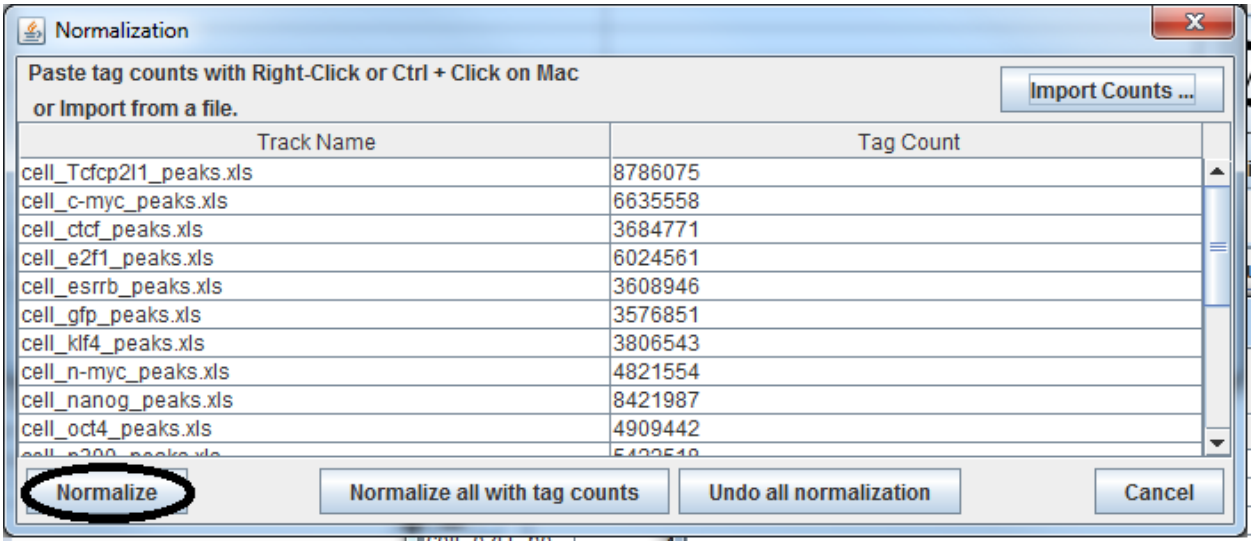
7. Click the **Normalization** button on the **Home** tab.



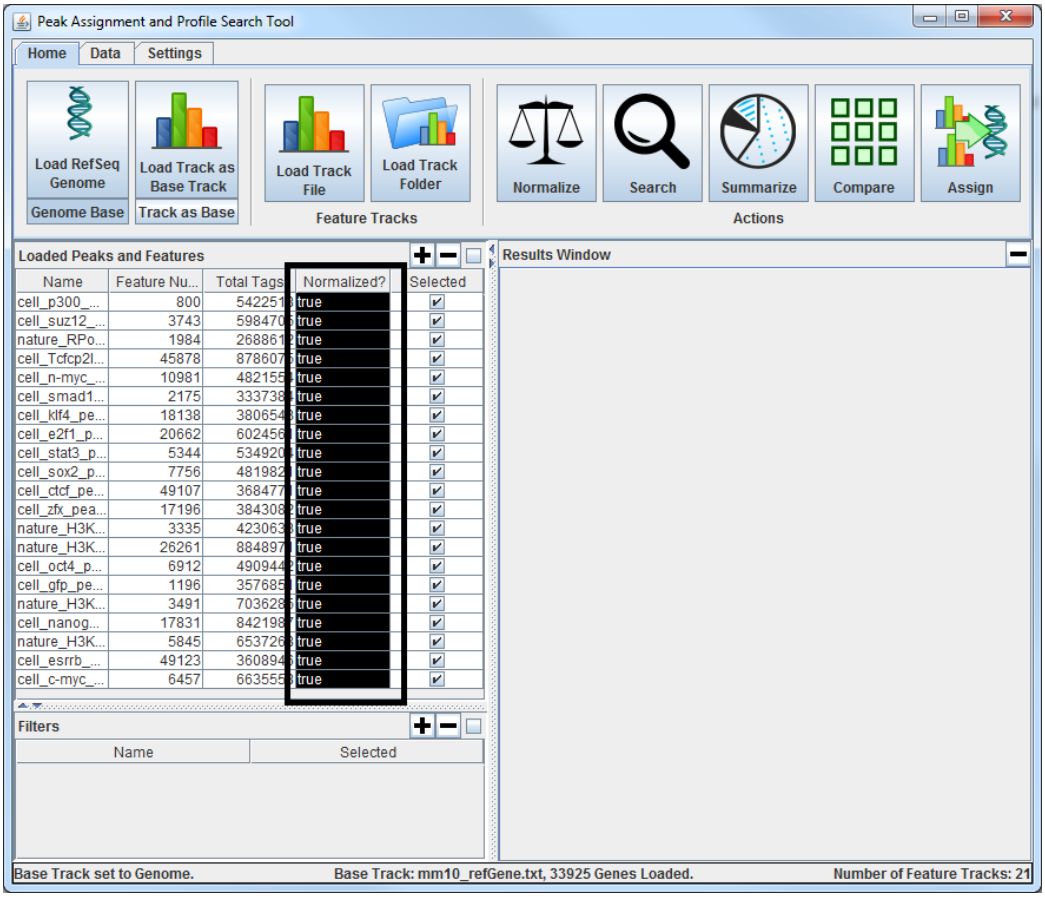
8. The Normalization dialog will appear. Click the **Import Counts …** button.



9. Select the **tag_counts.txt** file included in the walkthrough folder. The tag counts will appear in the table. Click the **Normalize** button.



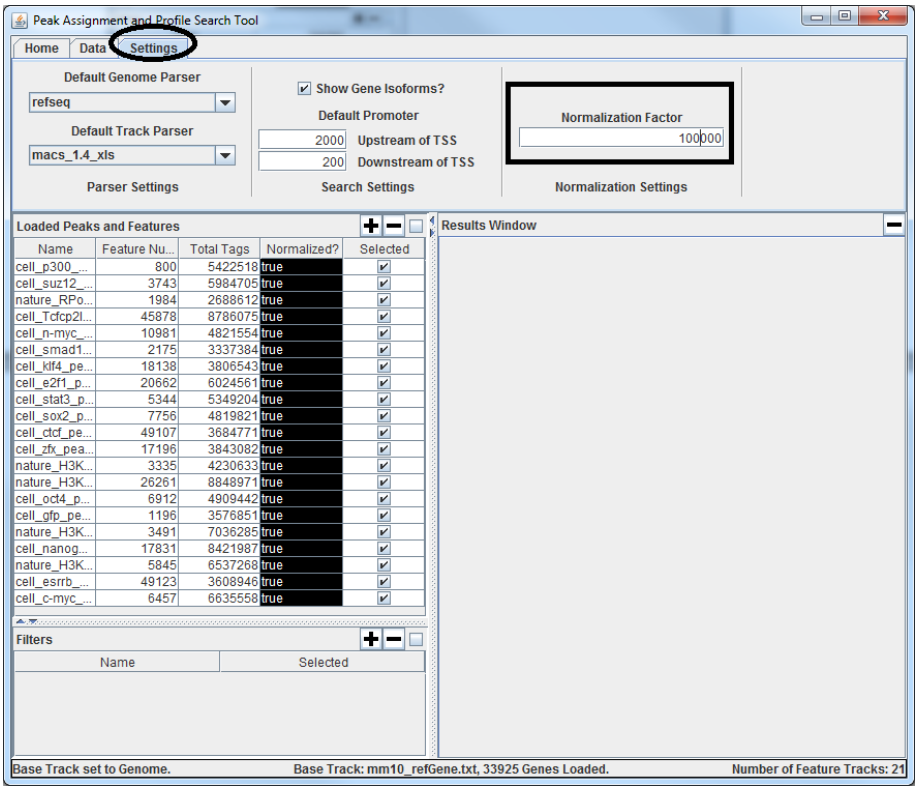
10. After normalization, highlighting will denote the normalized peak sets.



Change the Normalization Factor

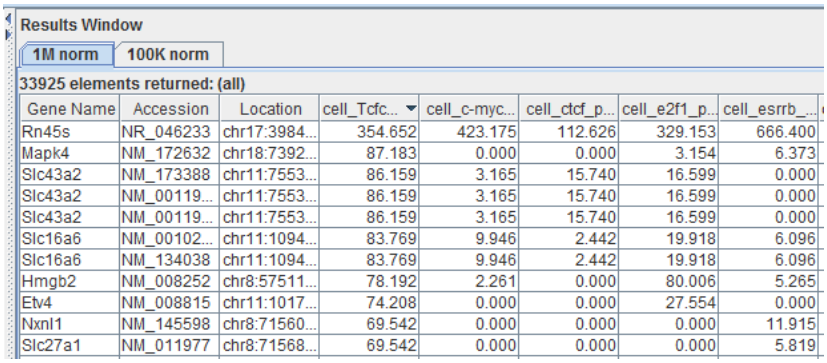
You can change the normalization factor on the settings tab.

1. Navigate to the settings tab and adjust the normalization factor to 100,000.



2. Repeat the normalization process to apply the new normalization factor. The new values will reflect the new scaling factor.

Results using 1 million as the normalization factor.



Gene Name	Accession	Location	cell_Tcfc...	cell_c-myc...	cell_ctcf_p...	cell_e2f1_p...	cell_esrrb_...
Rn45s	NR_046233	chr17:3984...	354.652	423.175	112.626	329.153	666.400
Mapk4	NM_172632	chr18:7392...	87.183	0.000	0.000	3.154	6.373
Slc43a2	NM_173388	chr11:7553...	86.159	3.165	15.740	16.599	0.000
Slc43a2	NM_00119...	chr11:7553...	86.159	3.165	15.740	16.599	0.000
Slc43a2	NM_00119...	chr11:7553...	86.159	3.165	15.740	16.599	0.000
Slc16a6	NM_00102...	chr11:1094...	83.769	9.946	2.442	19.918	6.096
Slc16a6	NM_134038	chr11:1094...	83.769	9.946	2.442	19.918	6.096
Hmgb2	NM_008252	chr8:57511...	78.192	2.261	0.000	80.006	5.265
Etv4	NM_008815	chr11:1017...	74.208	0.000	0.000	27.554	0.000
Nxnl1	NM_145598	chr8:71560...	69.542	0.000	0.000	0.000	11.915
Slc27a1	NM_011977	chr8:71568...	69.542	0.000	0.000	0.000	5.819

Results using 100k as the normalization factor.

Results Window

1M norm | 100K norm

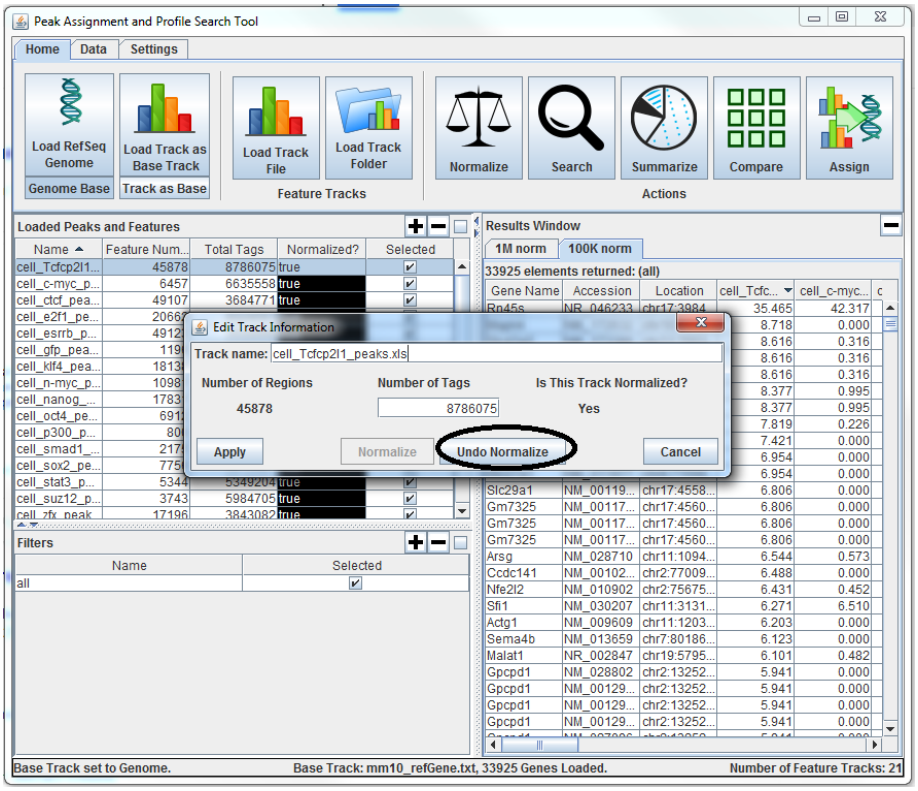
33925 elements returned: (all)

Gene Name	Accession	Location	cell_Tcfc...	cell_c-myc...	cell_ctcf_p...	cell_e2f1_p...	cell_esrrb_...
Rn45s	NR_046233	chr17:3984...	35.465	42.317	11.263	32.915	66.640
Mapk4	NM_172632	chr18:7392...	8.718	0.000	0.000	0.315	0.637
Slc43a2	NM_173388	chr11:7553...	8.616	0.316	1.574	1.660	0.000
Slc43a2	NM_00119...	chr11:7553...	8.616	0.316	1.574	1.660	0.000
Slc43a2	NM_00119...	chr11:7553...	8.616	0.316	1.574	1.660	0.000
Slc16a6	NM_00102...	chr11:1094...	8.377	0.995	0.244	1.992	0.610
Slc16a6	NM_134038	chr11:1094...	8.377	0.995	0.244	1.992	0.610
Hmgb2	NM_008252	chr8:57511...	7.819	0.226	0.000	8.001	0.526
Etv4	NM_008815	chr11:1017...	7.421	0.000	0.000	2.755	0.000
Nxnl1	NM_145598	chr8:71560...	6.954	0.000	0.000	0.000	1.191
Slc27a1	NM_011977	chr8:71568...	6.954	0.000	0.000	0.000	0.582

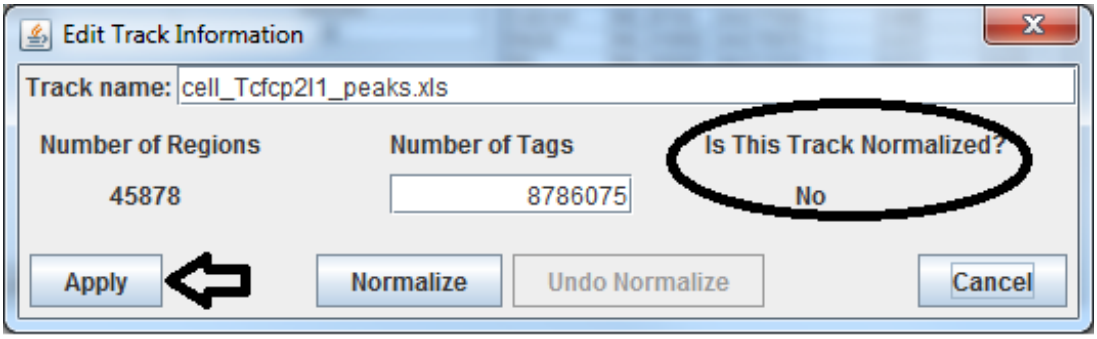
Undo Normalization for a Peak Set

PAPST can reverse the normalization on a peak set and return its original value.

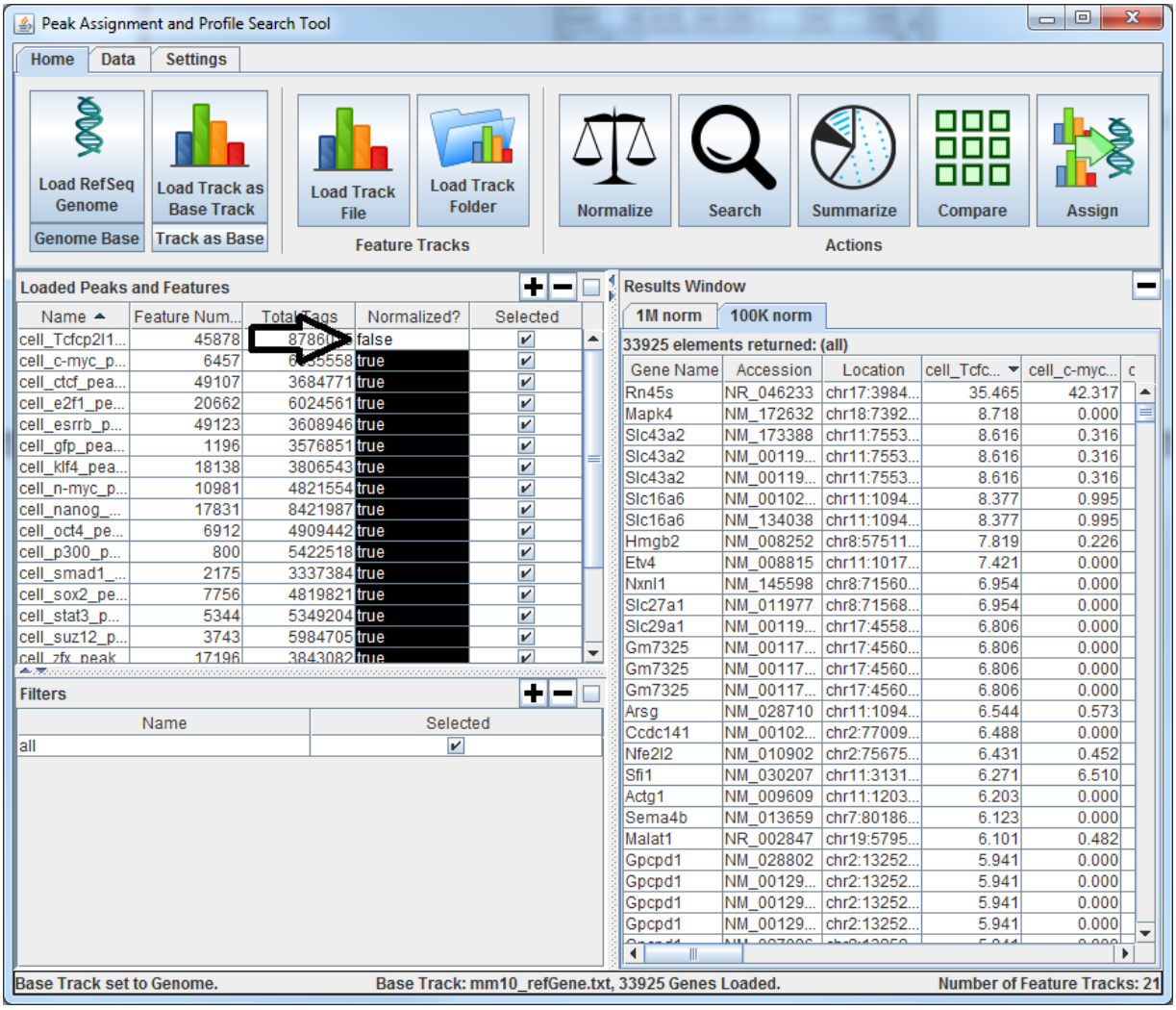
1. Double click a feature to open the track information editor. Click the '**Undo Normalization**' button.



2. The track will have its scores returned to their original values. Click **Apply** to commit the changes and save the track's information.



3. The track shows that it has been returned to its original values.



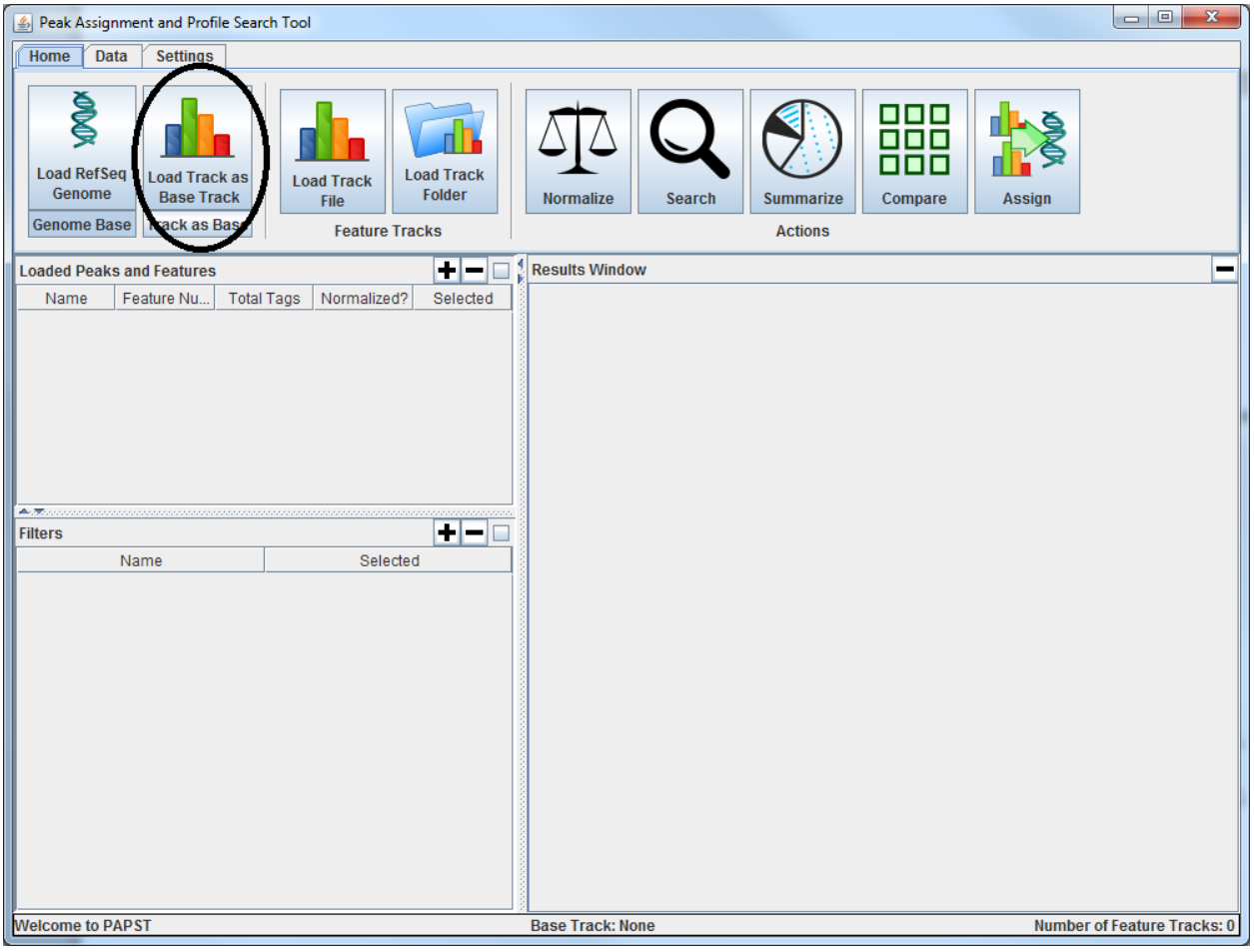
Alternative Base Tracks

For most analysis, PAPST uses the genes of an organism as the 'Base Track'. This means that the coordinates are relative to genes and genes are the rows returned from a search. Other situations may call for another base track to be used. One example would be searching for un-annotated transcripts that have certain histone modifications. For this use case, PAPST has a feature for using any arbitrary feature as the basis for searches.

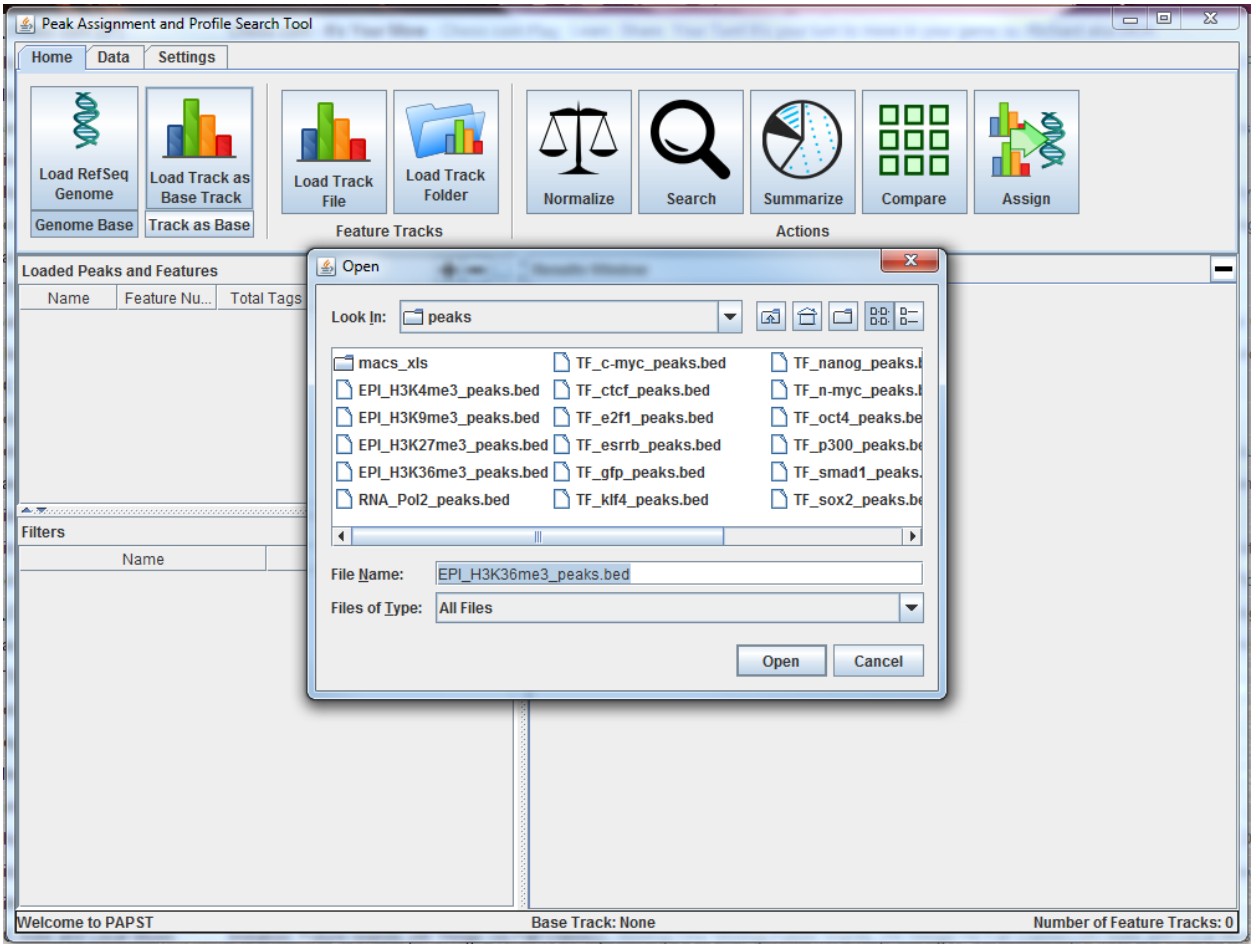
Search Example using K36me3 as Base

This example demonstrates a search for regions of K36me3 that are flanked by K4me3 and a transcription factor. Actively transcribed genes are enriched with K36me3 along the body of the gene (Barski et al., 2007); therefore, this method could potentially find novel transcripts.

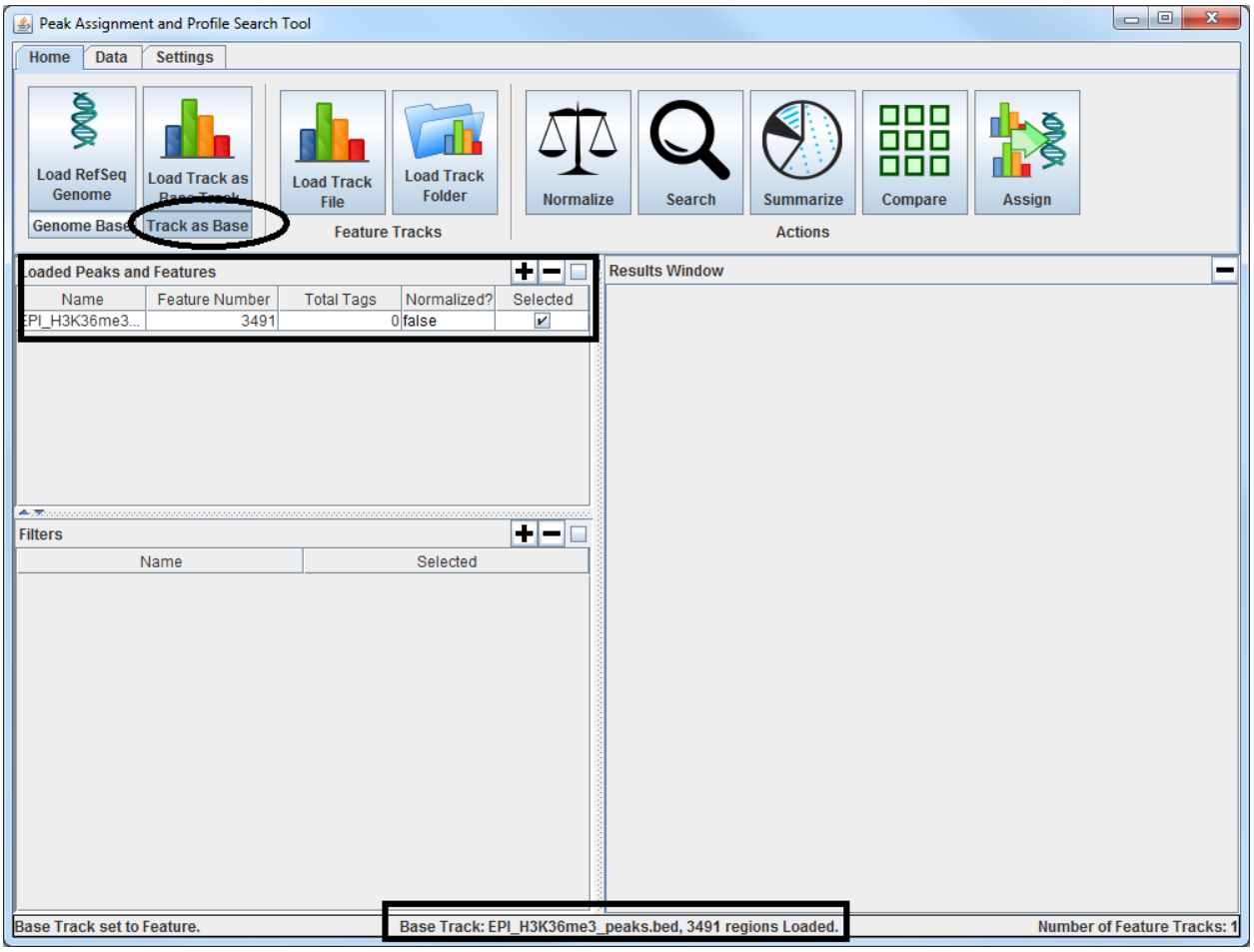
1. To load an alternative base track, click the **Load Track as Base Track** button on the **Home** tab.



2. Let's select the **EPI_H3K36me3_peaks.bed** track.

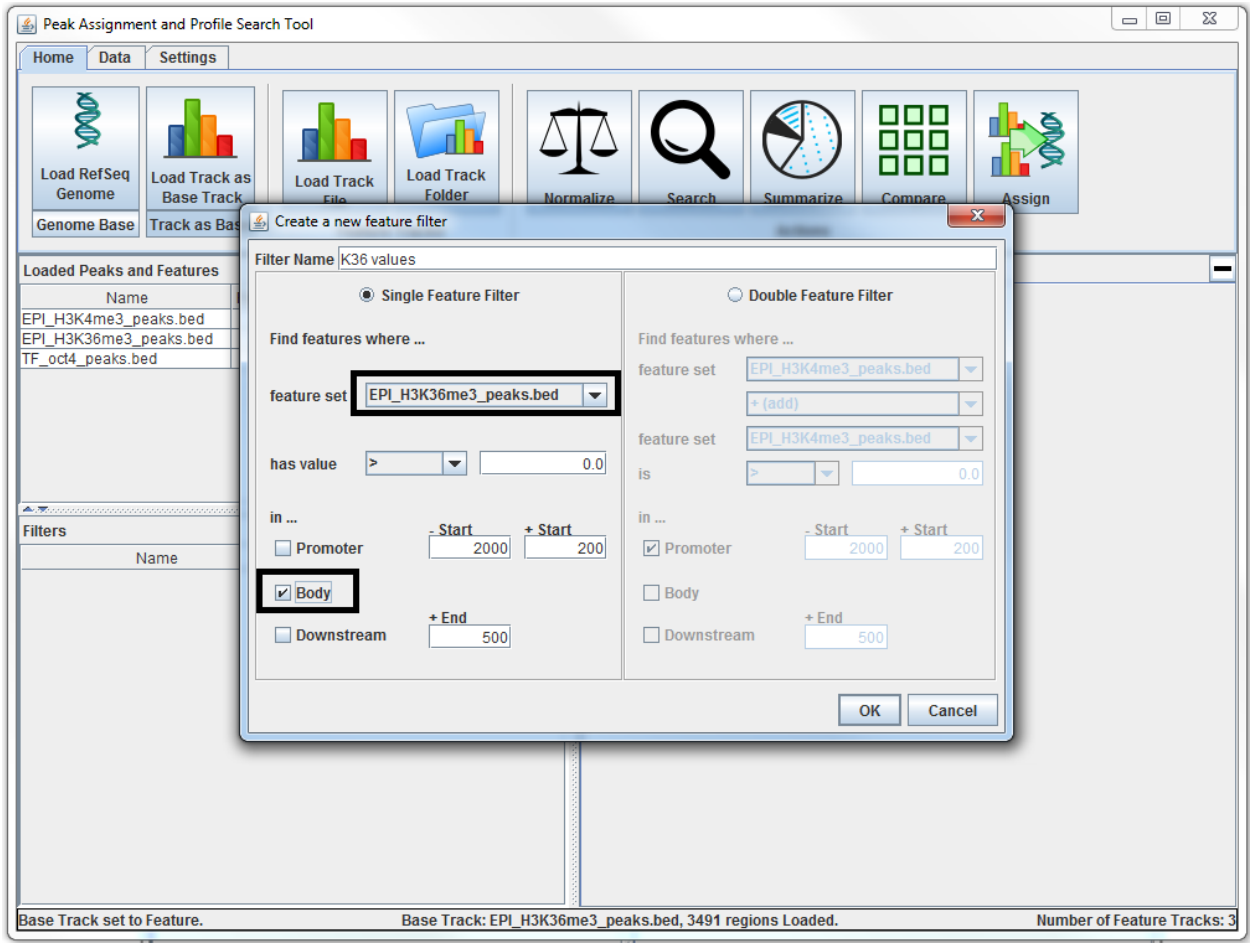


3. The peak track will be loaded and selected as the current base track. The **Track as Base** toggle button will be selected. The status bar at the bottom of the application window will show that the current base is **EPI_H3K36me3_peaks.bed**.



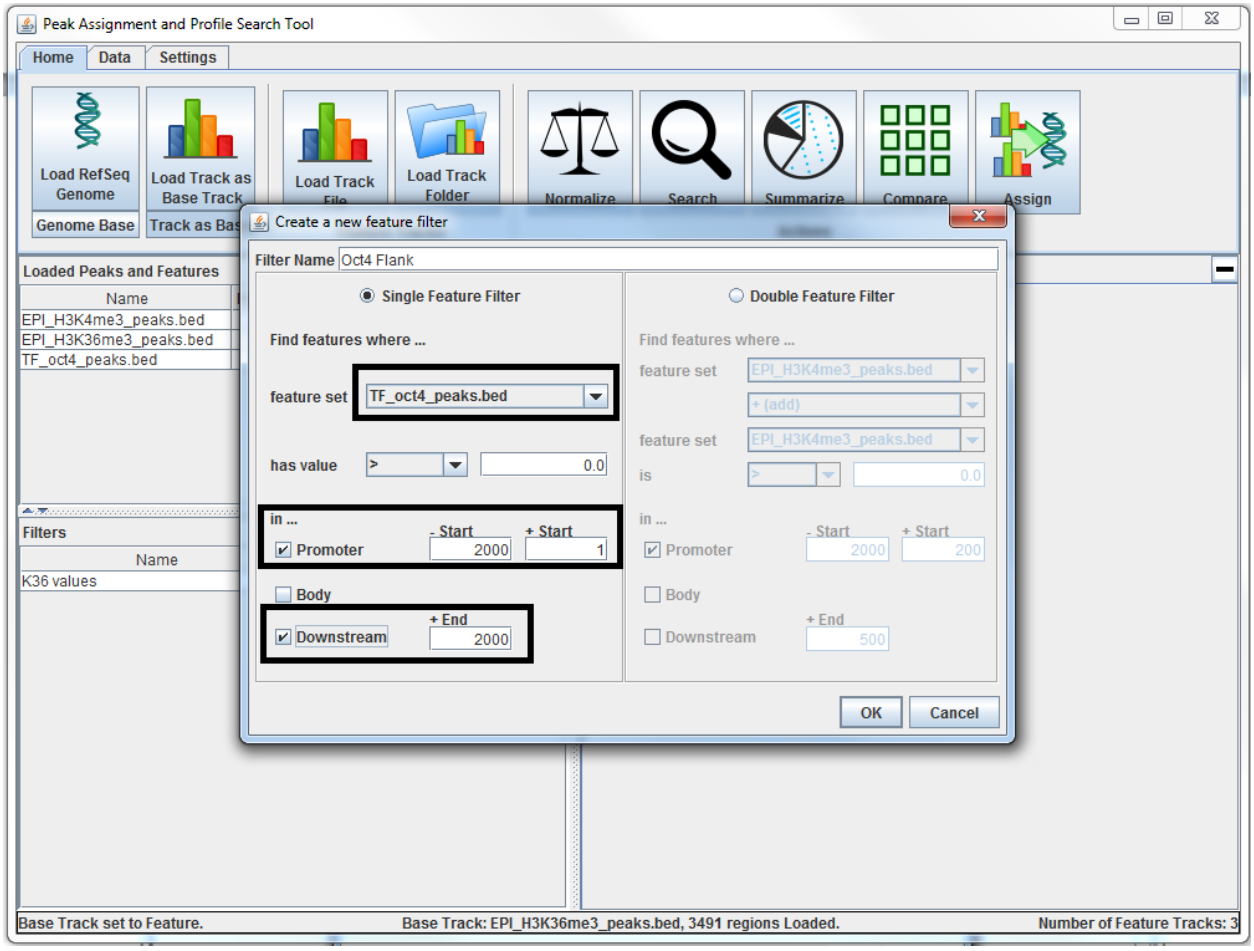
4. Load the histone modification track, **EPI_K4me3_peaks.bed.**
5. Load the transcription factor track, **TF_oct4_peaks.bed**. We will search for regions with K4 and Oct4 binding in the flanking regions of K36 sites.

6. Create a filter for K36 in the gene body. **When <u>Track as Base</u> is selected, the gene body refers to the peak region**. Uncheck the promoter checkbox. Name the filter 'K36 values'. Click OK.

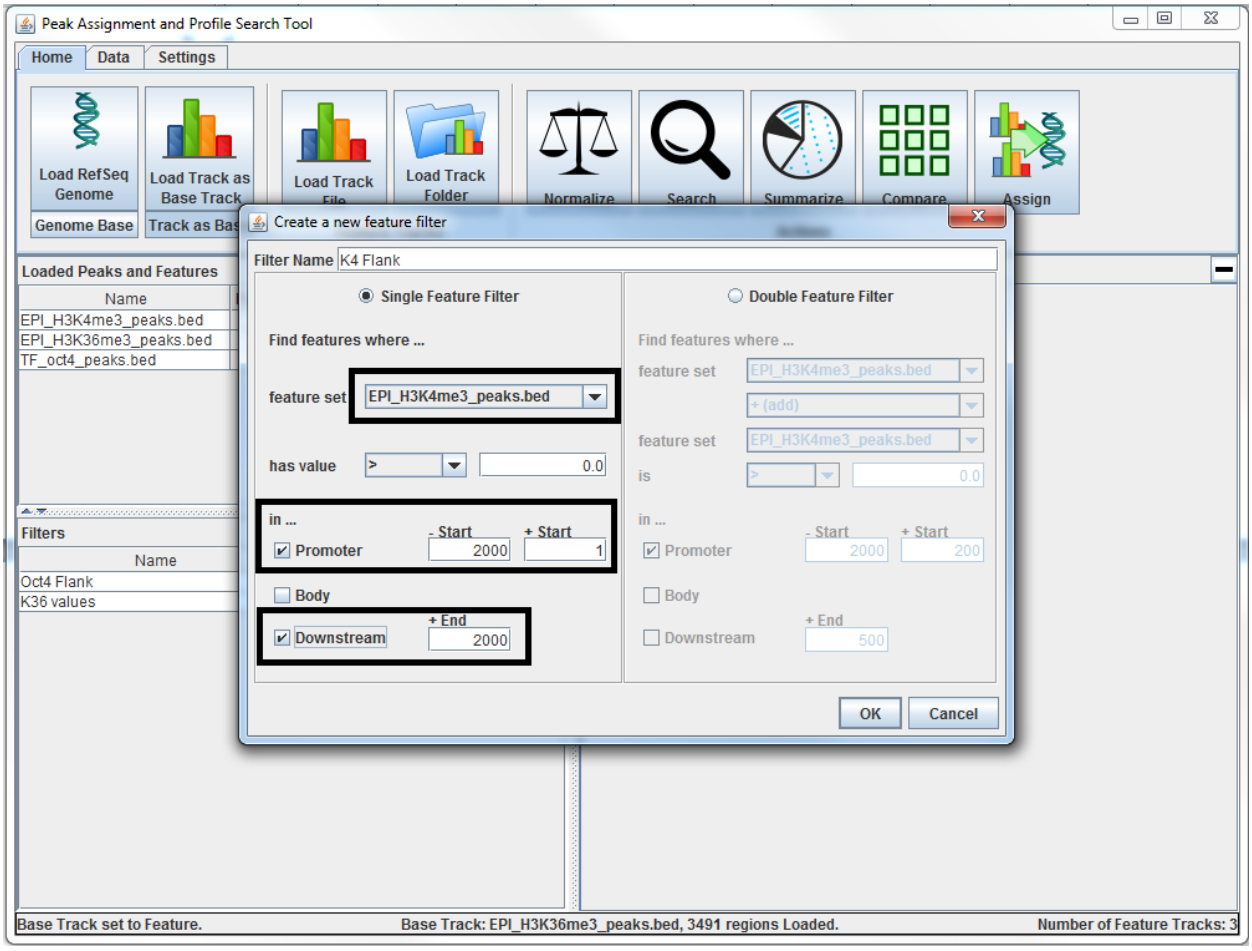


This filter will always accept because K36me3 is the current Base Track. This filter is necessary to report the K36me3 values to our results table.

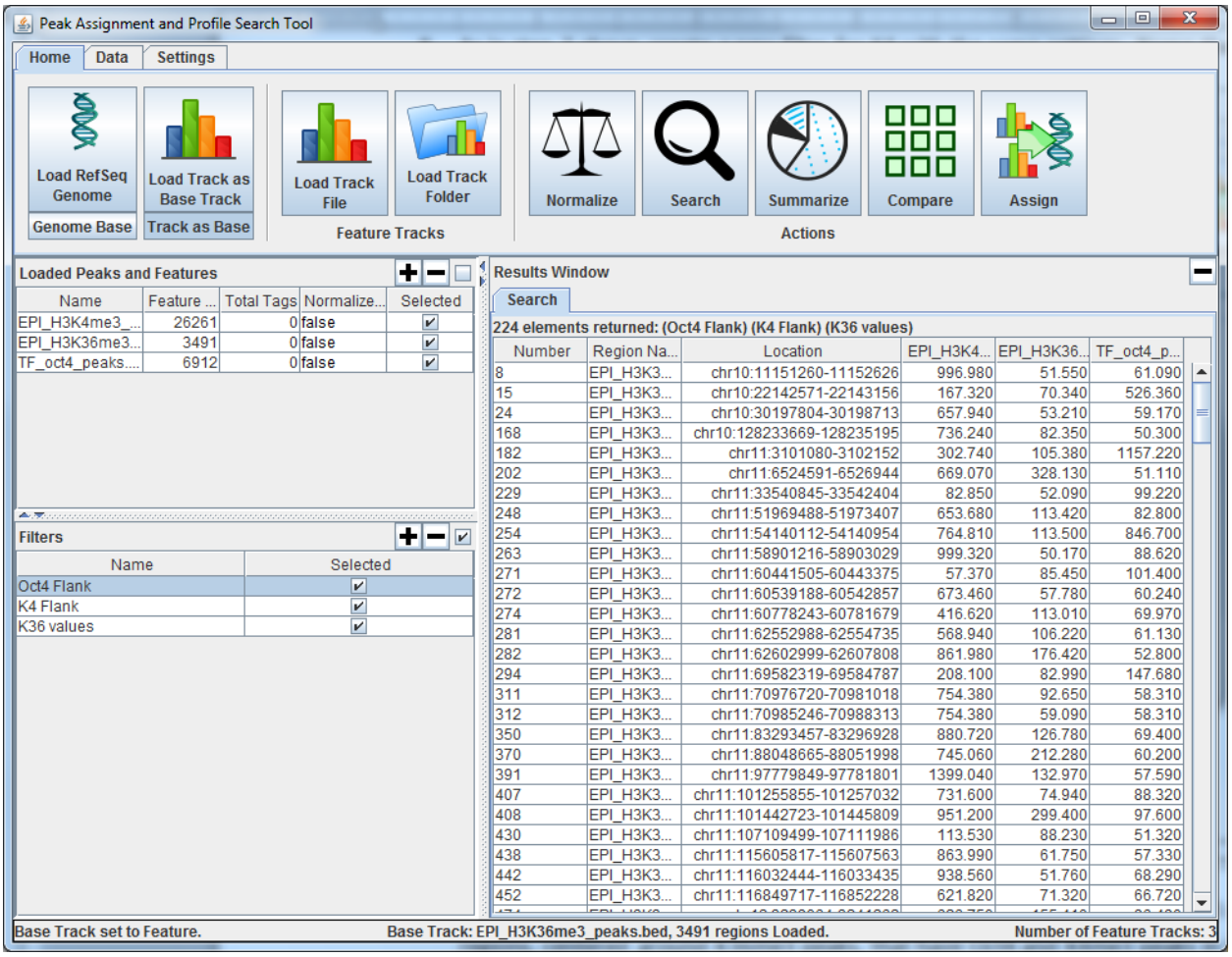
7. Create a new filter for Oct4. Select the Oct4 peak set. Select the promoter and downstream regions and set them to 2000. **When using <u>Track as Base</u>, the promoter represents the 5' flank and downstream represents the 3' flank**. Peaks are strand-less. Set the + Start to 1. Name the file 'Oct4 Flank'.



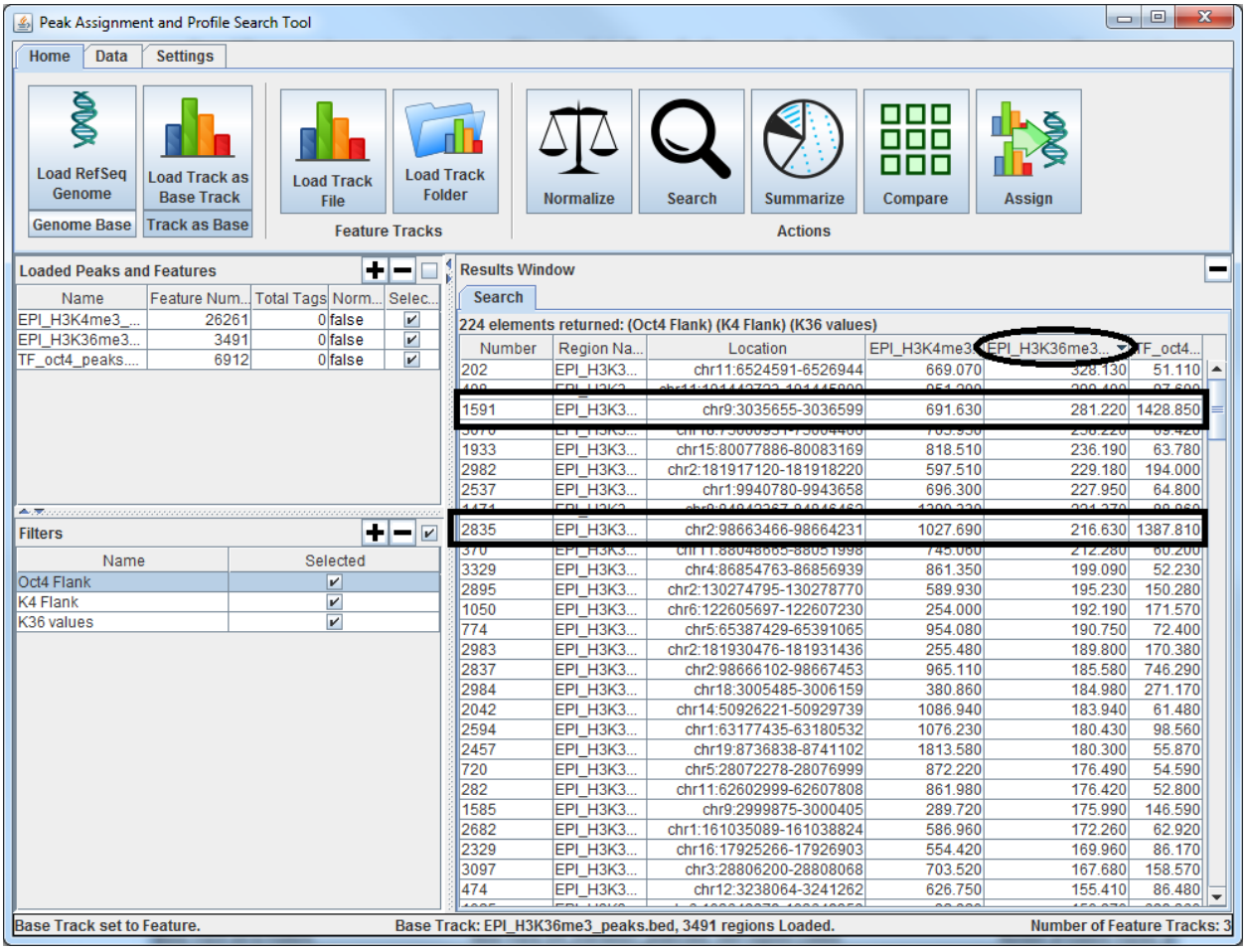
8. As in step 7 above, create a new filter for K4 with the same settings. Name the filter K4 Flank.



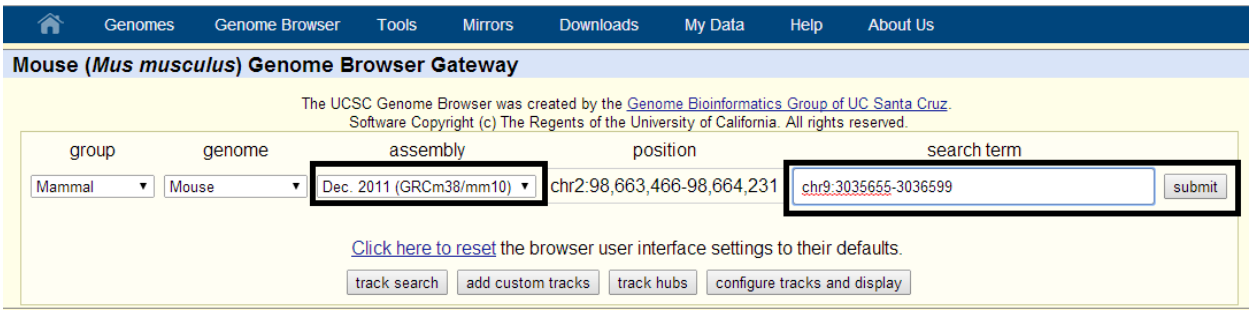
9. After creating the filters, click the **Search** button. PAPST will return all genomic regions that have Oct4 and K4me3 peaks in their flanking regions centered on K36me3 peaks. These regions may also include genes.



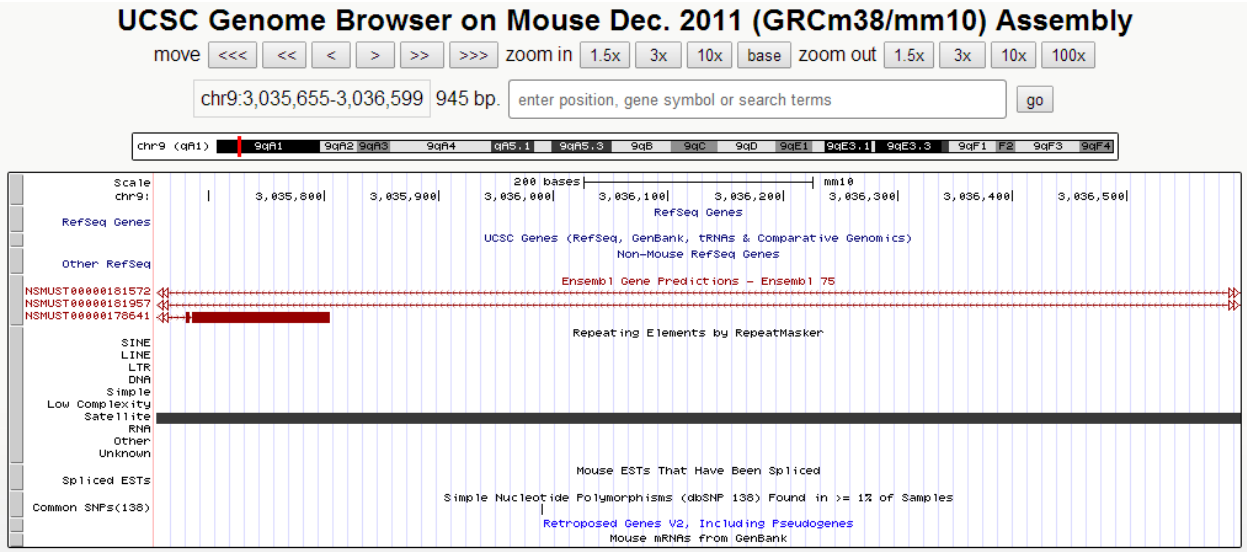
10. Sort the results table by the K36me3 value in descending order by clicking its column twice. Notice there are two regions with extremely high Oct4 signal among the highest K36me3 peaks.



11. Let's view these regions in the UCSC Genome Browser to learn more. Click on the **Location** cell of feature number 1591. Press **Ctrl + c** to copy the location value (chr9:3035655-3036599). In Genome Browser, select the MM10 genome and paste the location value in the search field. Press '**submit**'.

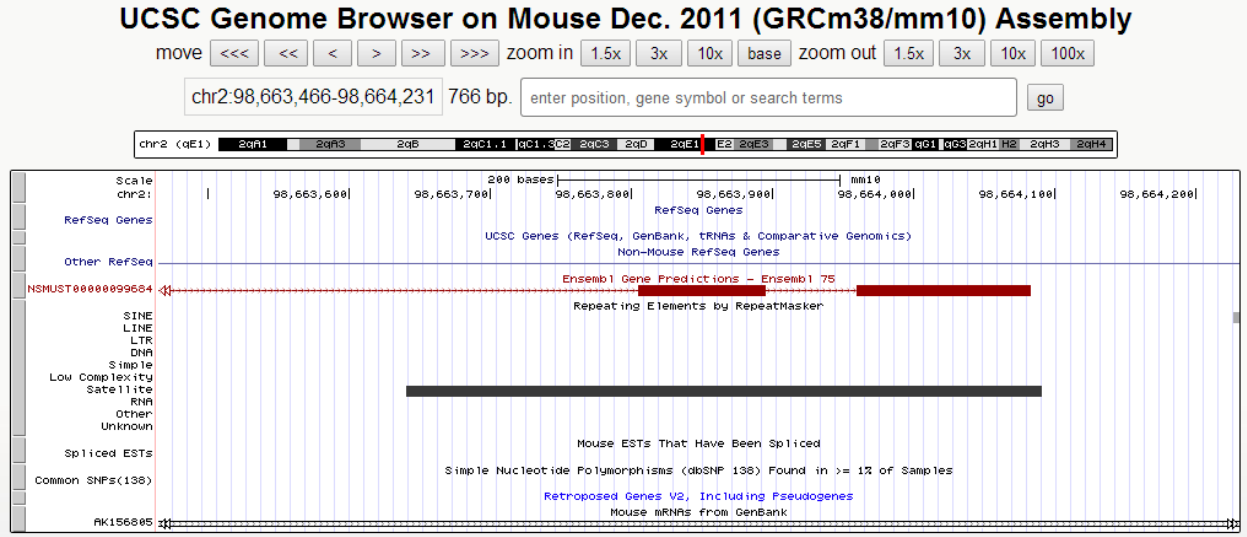


12. A region of Chr9 is return. A satellite region and two predicted genes occupy this region. There are no recognized RefSeq genes in this region though.



This region could represent a novel area of transcribed RNA relevant to mouse stem cells.

13. Check the other interesting region in Genome Browser. Copy the location (chr2:98663466-98664231) from PAPST and paste it into the search field for mm10. A predicted gene and satellite region occupy this region too.



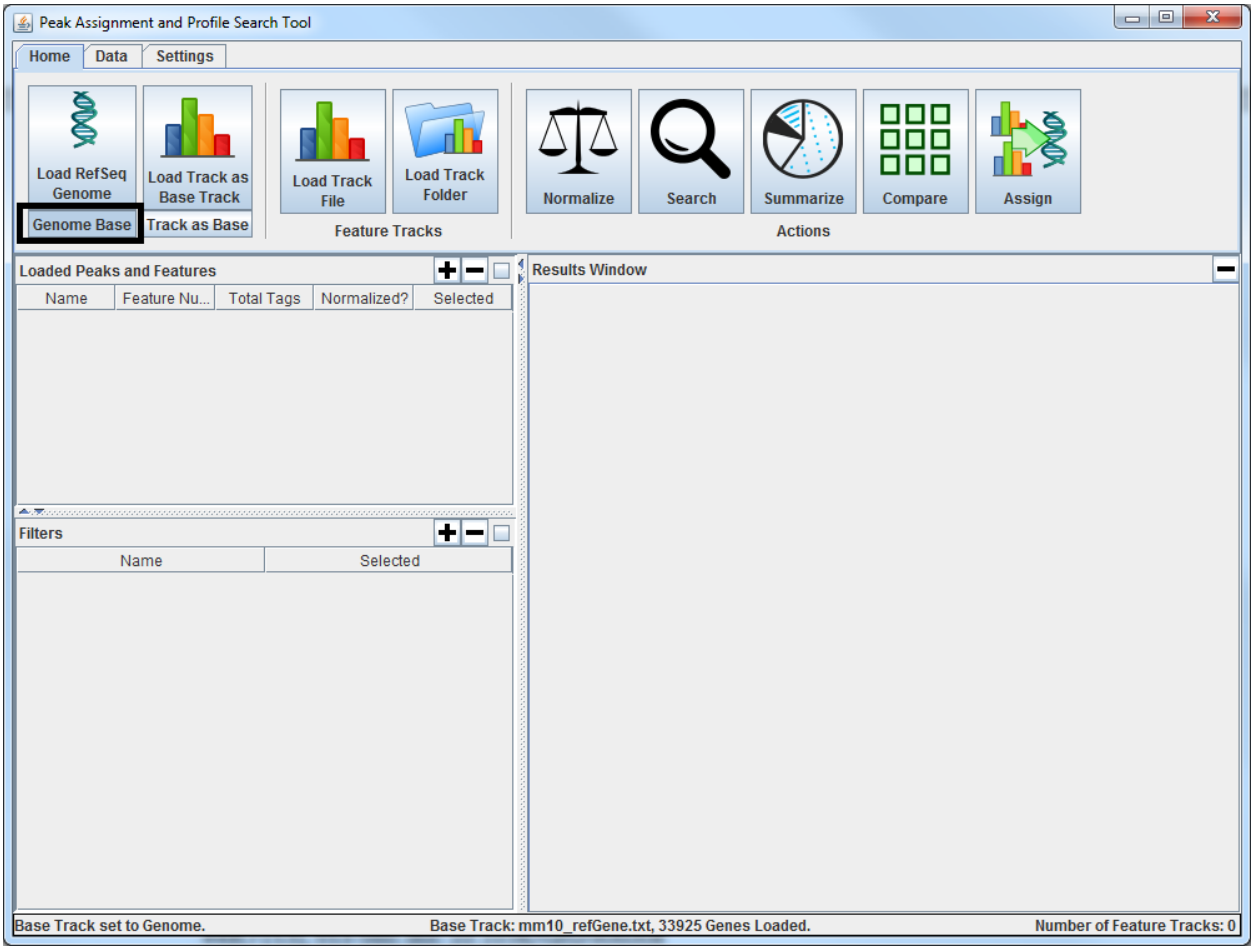
These regions have a particular epigenetic profile, specifically K36me3 flanked by K4me3 and Oct4 within 2000bp.

In this way PAPST can be used to search for arbitrary genomic regions with specific patterns defined in terms of the provide features.

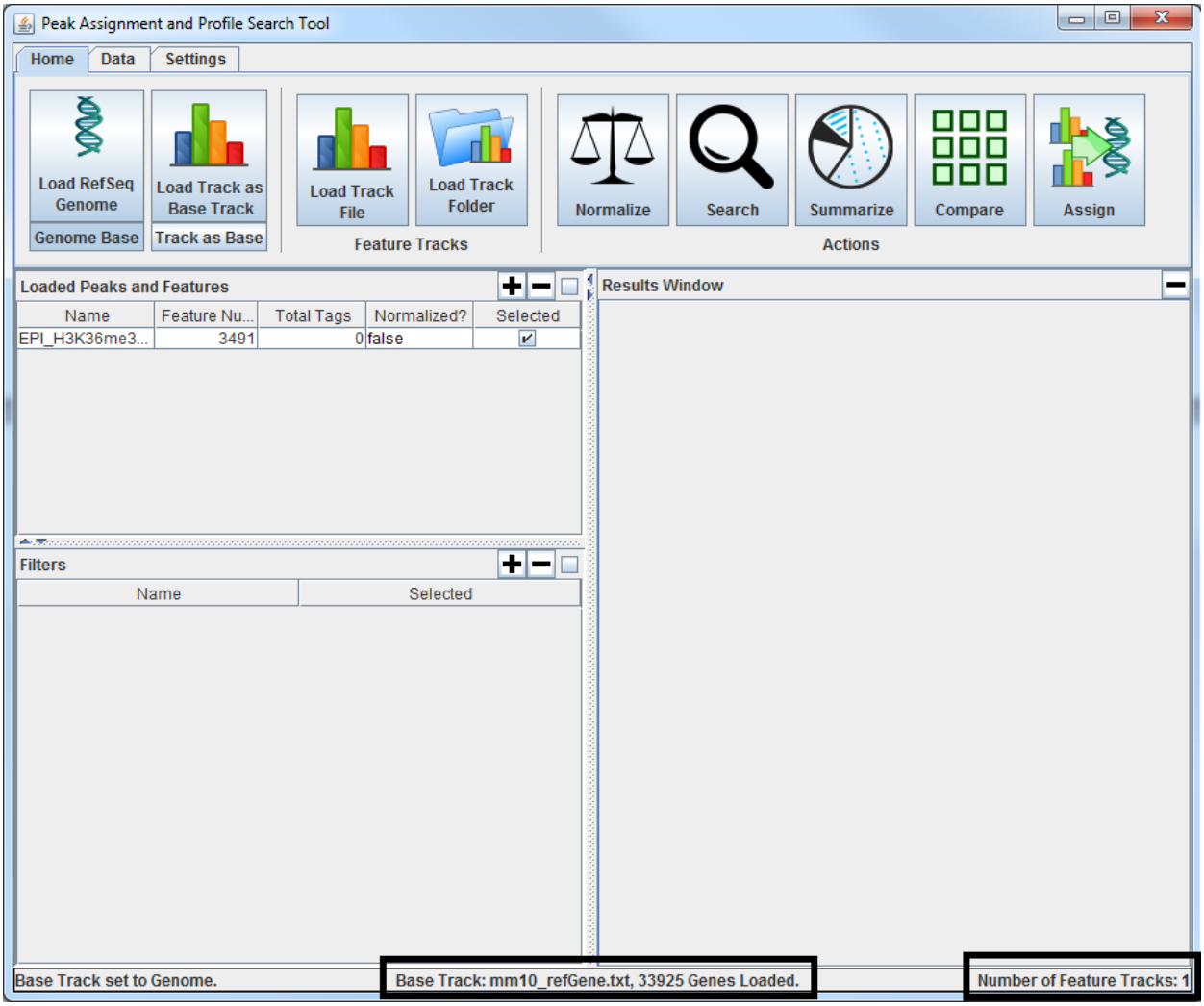
Switching between Genome and Custom Base Tracks

PAPST makes switching between base tracks easy. Simply click the toggle buttons for **Genome Base** or **Track as Base** on the **Home** tab.

1. Add a genome to PAPST by clicking **Load RefSeq Genome** and selecting mm10. Notice that the **Genome Base** button is selected.

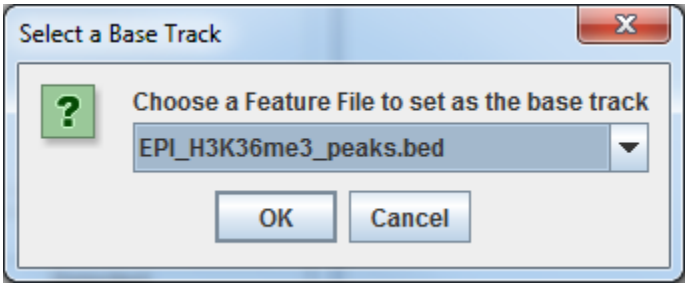


2. Add a peak track to PAPST by clicking **Load Track File**. Let's select **EPI_H3K36me3_peaks.bed**.

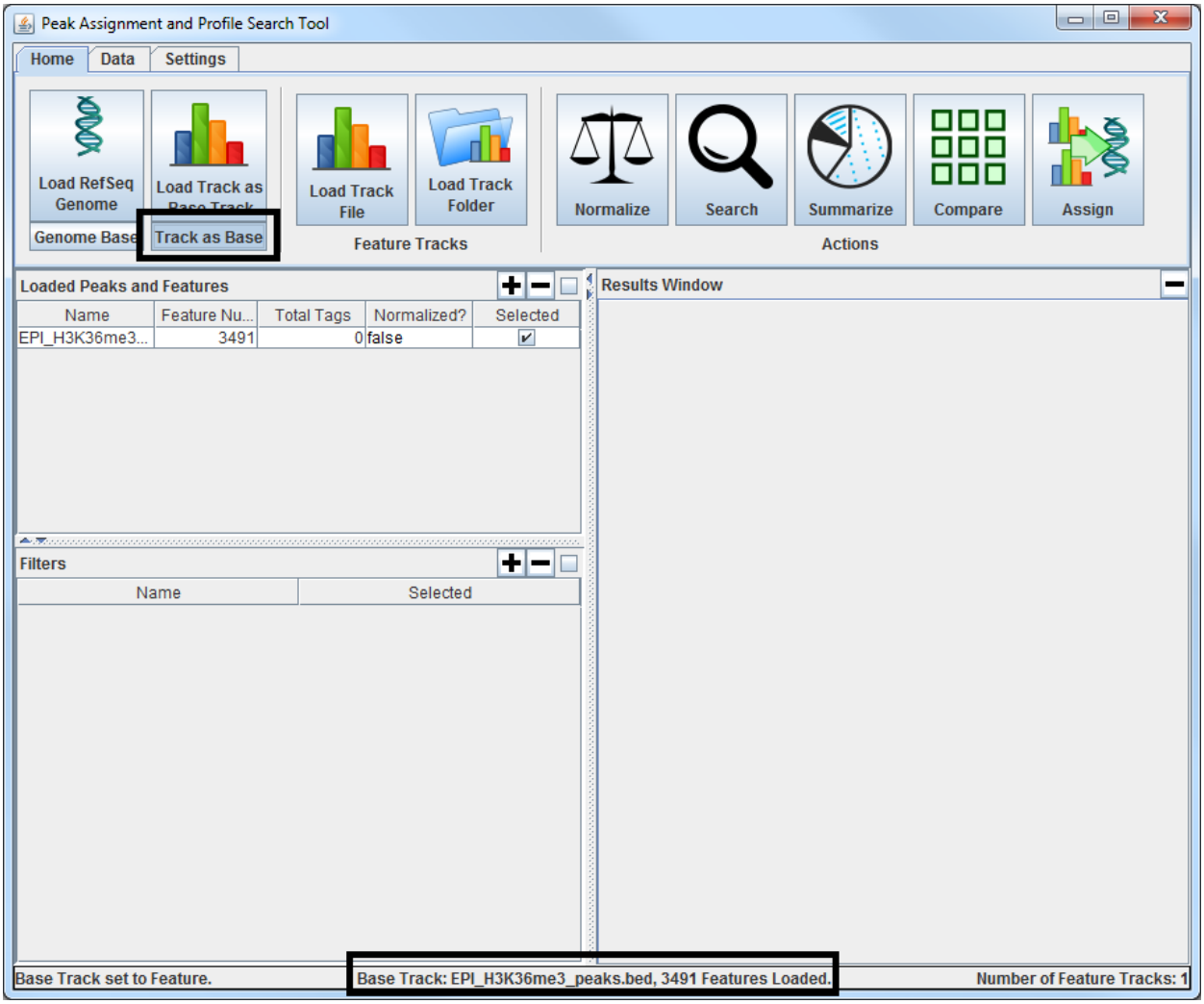


Notice that we have one feature track loaded and **mm10_refGene.txt** is the current base track.

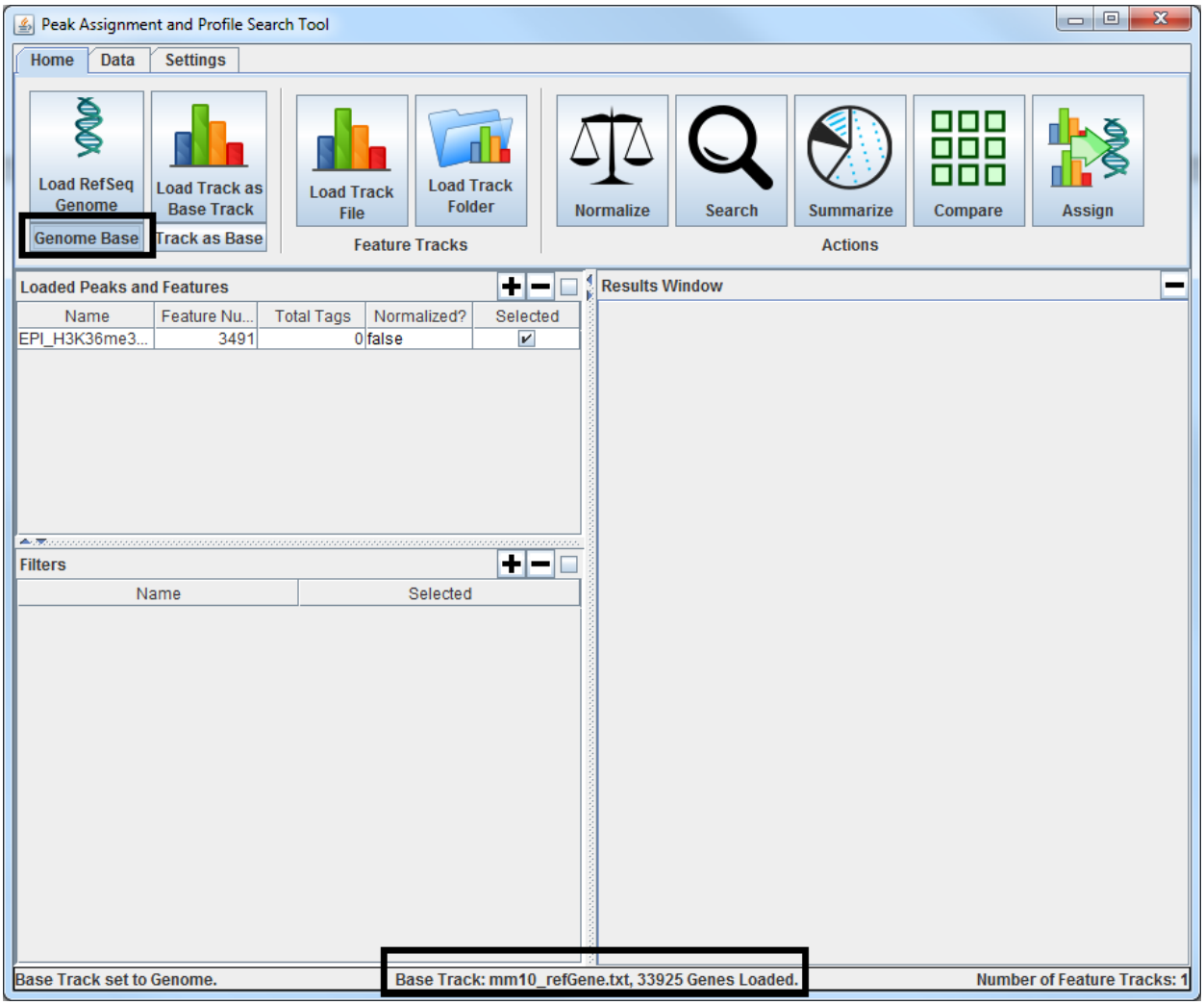
3. Let's change the current base track to K36me3. Click the **Track as Base** button. A track selection dialog will appear. Select **EPI_H3K36me3_peaks.bed** from the dropdown menu. Click OK.



4. The selected track will now be used as the base track. The **Track as Base** button is highlighted to indicate that searches now use a non-gene base track. The bottom status bar will indicate that the current base is **EPI_H3K36me3_peaks.bed**.



5. Click the **Genome Base** button to switch back to mm10. The **Genome Base** button is how highlighted to show that PAPST uses genes for searching again. Text in the status bar indicates that PAPST is now using mm10 as the base track.



References

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi: 10.1016/j.cell.2007.05.009

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., . . . Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell, 133*(6), 1106-1117. doi: 10.1016/j.cell.2008.04.043

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., . . . Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature, 448*(7153), 553-560. doi: 10.1038/nature06008