

PAPST User's Guide and Tutorial

Outline:

1. [Introduction](#)
2. [Getting Started](#)
 - a. [Loading a Genome](#)
 - b. [Loading Peaks and Other Features](#)
 - c. [Loading Folders of Peaks](#)
 - d. [Remove Peak Sets](#)
 - e. [Save a PAPST Session](#)
 - f. [Load a PAPST Session](#)
3. [Working with Filters](#)
 - a. [Searching with Filters](#)
 - b. [Editing Filters](#)
 - c. [Using Multiple Filters](#)
 - d. [The ALL Filter](#)
 - e. [Remove Filters](#)
4. [Working with Results Tables](#)
 - a. [Exporting Results](#)
 - b. [Delete Result Windows](#)
 - c. [Clear All Results](#)
5. [General Analysis](#)
 - a. [Peak Distributions](#)
 - b. [Assign Peaks to Nearest Gene](#)
 - c. [Batch Assign Multiple Features](#)
 - d. [Comparing Peaks by Overlap](#)
6. [Normalization](#)
 - a. [Normalize Values by Sequencing Depth](#)
 - b. [Normalize Multiple Peak Sets at Once](#)
 - c. [Change Normalization Factor](#)
 - d. [Undo Normalization](#)
7. [Alternative Base Tracks](#)
 - a. [Search Example Using K36me3 as the Base Track](#)
 - b. [Switching Between Genome and Custom Base Tracks](#)
8. [References](#)

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 from 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 Loganantharaj, Hong-Wei Sun. **PAPST, a User Friendly and Powerful Java Platform for ChIP-Seq Peak Co-Localization Analysis and Beyond.** [Journal]

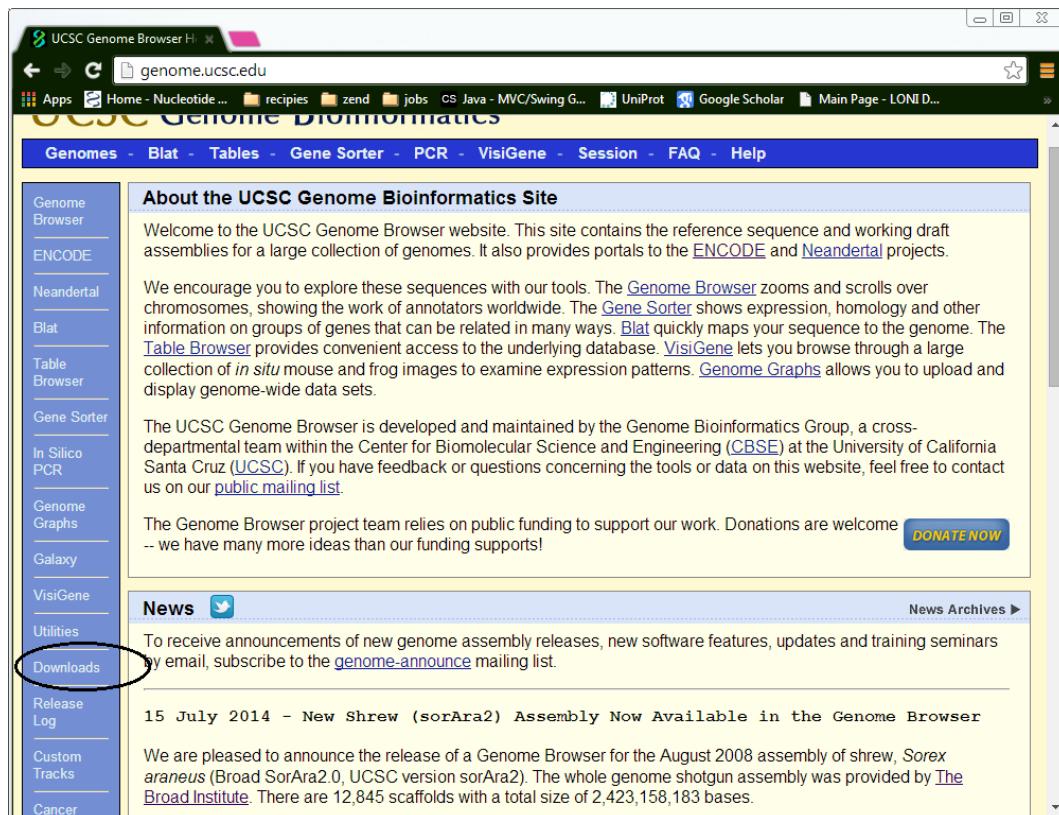
Getting Started

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 from UCSC Genome Browser

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



The screenshot shows the UCSC Genome Bioinformatics website. The URL in the address bar is genome.ucsc.edu. The main content area displays the "About the UCSC Genome Bioinformatics Site" page. On the left, a vertical sidebar lists various tools and databases: Genome Browser, ENCODE, Neandertal, Blat, Table Browser, Gene Sorter, In Silico PCR, Genome Graphs, Galaxy, VisiGene, Utilities, and Downloads. The "Downloads" link is circled in red. The main content area includes sections on the site's purpose, tools like Genome Browser and Gene Sorter, and a news section. A "DONATE NOW" button is visible in the news section.

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

The screenshot shows a web browser window for the UCSC Genome Browser. The URL in the address bar is hgdownload.soe.ucsc.edu/downloads.html#mouse. The page displays a list of links for the "Mouse Genome" (Dec. 2011, mm10). One of the links, "Annotation database", is circled in red.

- Full data set
- [Annotation database](#)

Mouse Genome

Dec. 2011 (mm10)

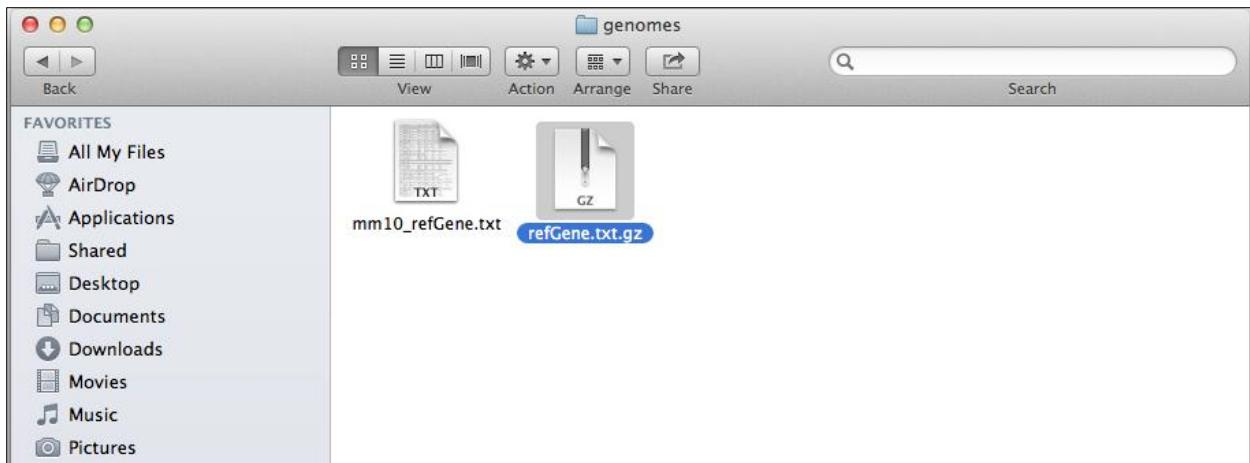
- Full data set
- Data set by chromosome
- [Annotation database](#)
- LiftOver files
- Protein database for mm10
- Pairwise Alignments
 - [Mouse/Human \(hg38\)](#)
 - [Mouse/Human \(hg19\)](#)
 - [Mouse/Chimp \(panTro4\)](#)
 - [Mouse/Gorilla \(gorGor3\)](#)
 - [Mouse/Pika \(ochPri2\)](#)
 - [Mouse/Pig \(susScr3\)](#)
 - [Mouse/Alpaca \(vicPac2\)](#)
 - [Mouse/Alpaca \(vicPac1\)](#)
 - [Mouse/Armadillo \(dasNov3\)](#)
 - [Mouse/Sloth \(choHof1\)](#)
 - [Mouse/Opossum \(monDom5\)](#)
 - [Mouse/Tasmanian devil \(sarHar1\)](#)

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

The screenshot shows a file browser window titled "Index of /goldenPath/mm...". The URL in the address bar is hgdownload.soe.ucsc.edu/goldenPath/mm10/database/. The page lists various files in the "database" directory, including "refGene.txt.gz", which is circled in red.

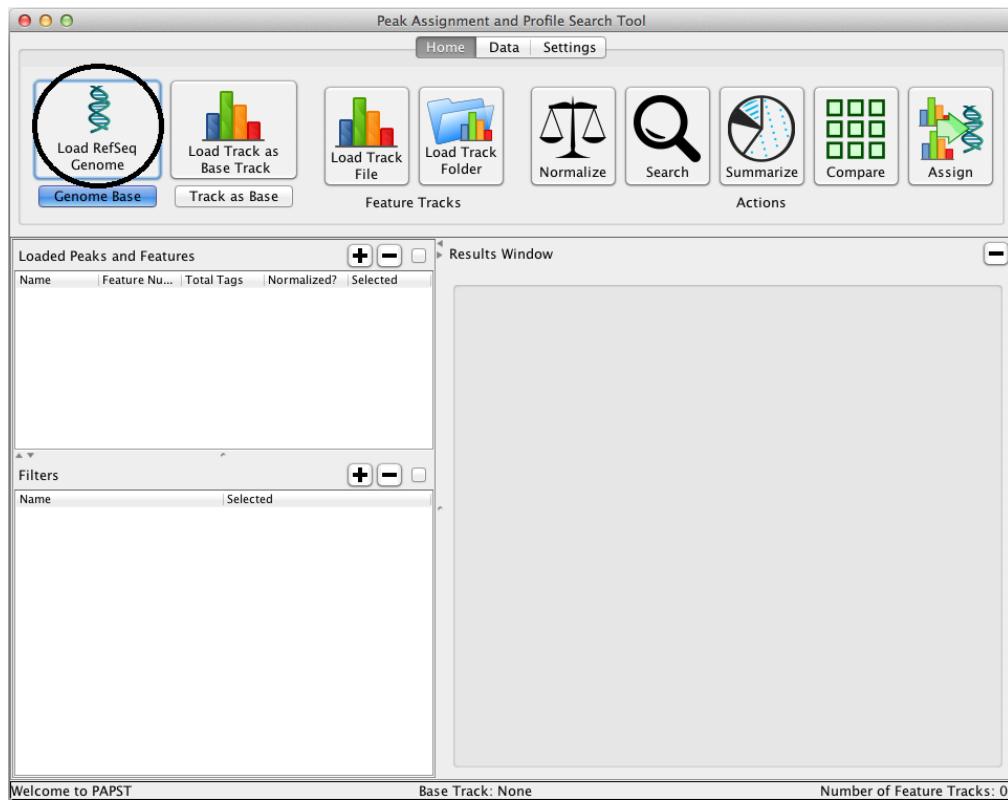
File	Last Modified	Size
phylcP60wayAll.txt.gz	29-Jun-2014 23:32	75M
phylcP60wayEuarchontaGlires.sql	29-Jun-2014 23:34	1.8K
phylcP60wayEuarchontaGlires.txt.gz	29-Jun-2014 23:34	73M
phylcP60wayGlires.sql	29-Jun-2014 23:31	1.8K
phylcP60wayGlires.txt.gz	29-Jun-2014 23:32	70M
phylcP60wayPlacental.sql	29-Jun-2014 23:33	1.8K
phylcP60wayPlacental.txt.gz	29-Jun-2014 23:33	74M
productName.sql	13-Jul-2014 16:04	1.4K
productName.txt.gz	13-Jul-2014 16:04	8.1M
gPcrPrimers.sql	03-Feb-2013 18:26	1.8K
gPcrPrimers.txt.gz	03-Feb-2013 18:26	6.9M
refFlat.sql	13-Jul-2014 16:02	1.7K
refFlat.txt.gz	13-Jul-2014 16:02	3.0M
refGene.sql	13-Jul-2014 16:02	1.9K
refGene.txt.gz	13-Jul-2014 16:02	3.4M
refLink.sql	13-Jul-2014 16:02	1.7K
refLink.txt.gz	13-Jul-2014 16:02	10M
refSeqAli.sql	13-Jul-2014 16:02	2.1K
refSeqAli.txt.gz	13-Jul-2014 16:02	2.9M
refSeqStatus.sql	13-Jul-2014 16:05	1.6K
refSeqStatus.txt.gz	13-Jul-2014 16:05	1.4M
refSeqSummary.sql	13-Jul-2014 16:04	1.5K
refSeqSummary.txt.gz	13-Jul-2014 16:04	4.1M
rmask.sql	07-Mar-2012 11:37	1.8K
rmask.txt.gz	07-Mar-2012 11:37	131M
rnBlastTab.sql	13-Apr-2014 14:42	1.7K
rnBlastTab.txt.gz	13-Apr-2014 14:42	330K

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.

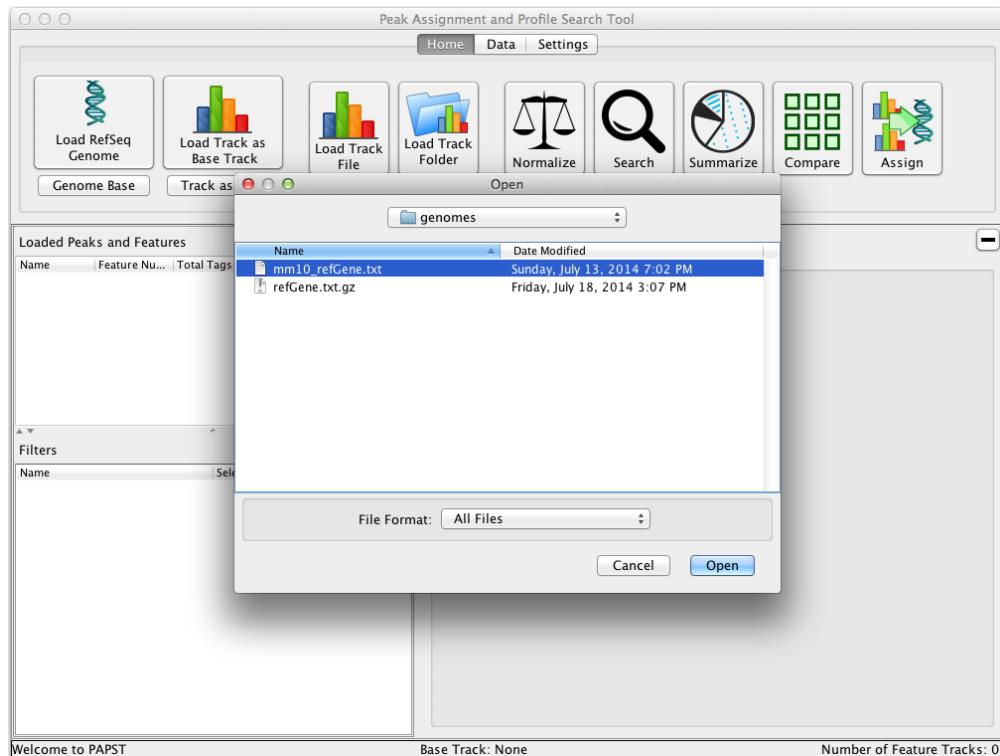


6. Load the genome into PAPST

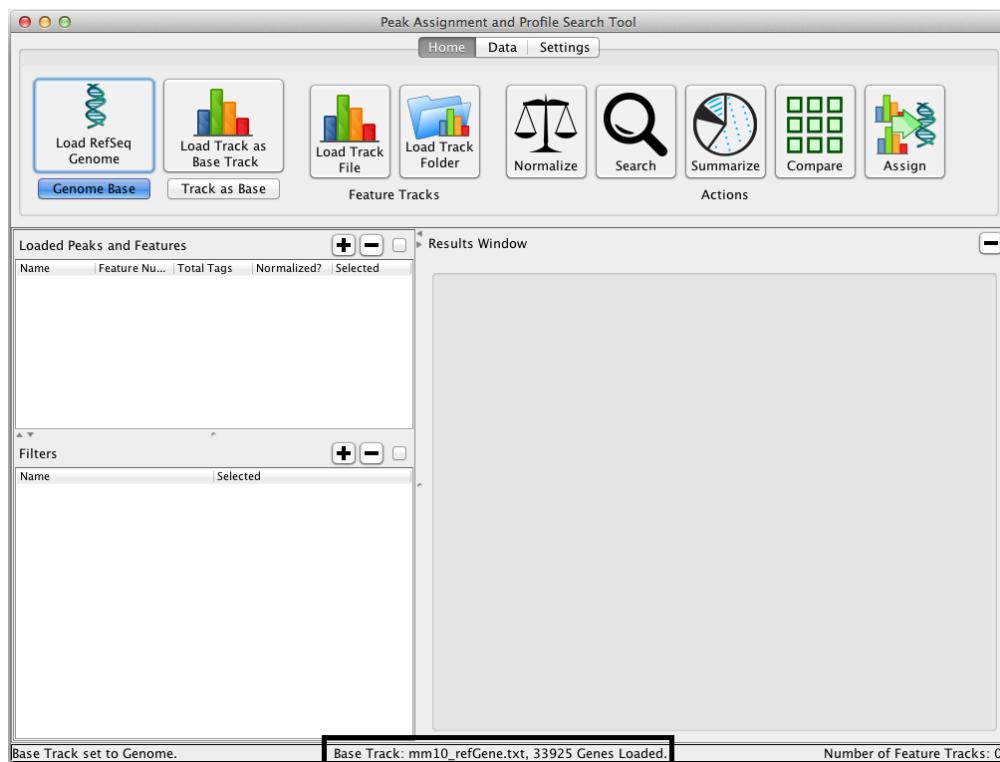
- Open PAPST and click **Load RefSeq Genome**.



b. Select the refGene.txt



c. The genome is now loaded into PAPST.



Loading Peak Track Files into PAPST

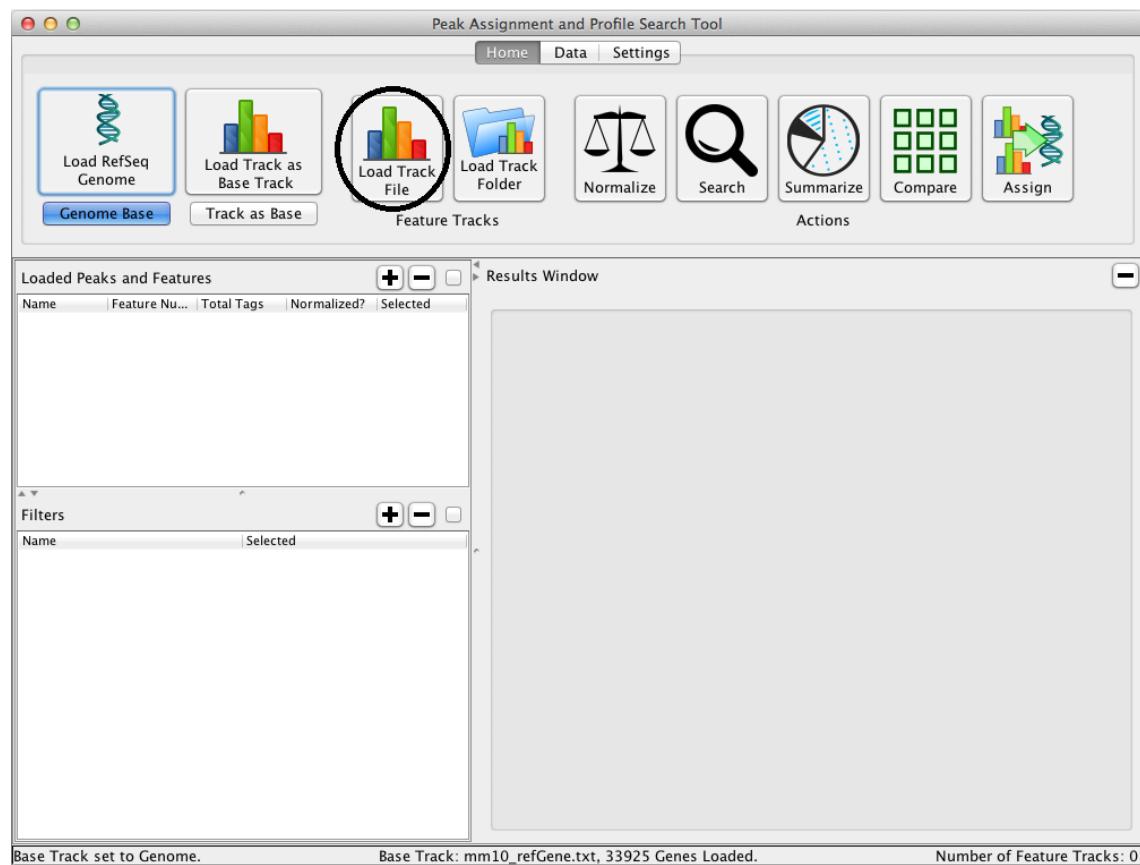
Peak files represent significant enrichment of protein binding from ChIP-seq experiments. They are typically represented as 4 column files. This section will show you how to load peak files and save parsers for later use. Tracks will refer to peaks or other genomic features.

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

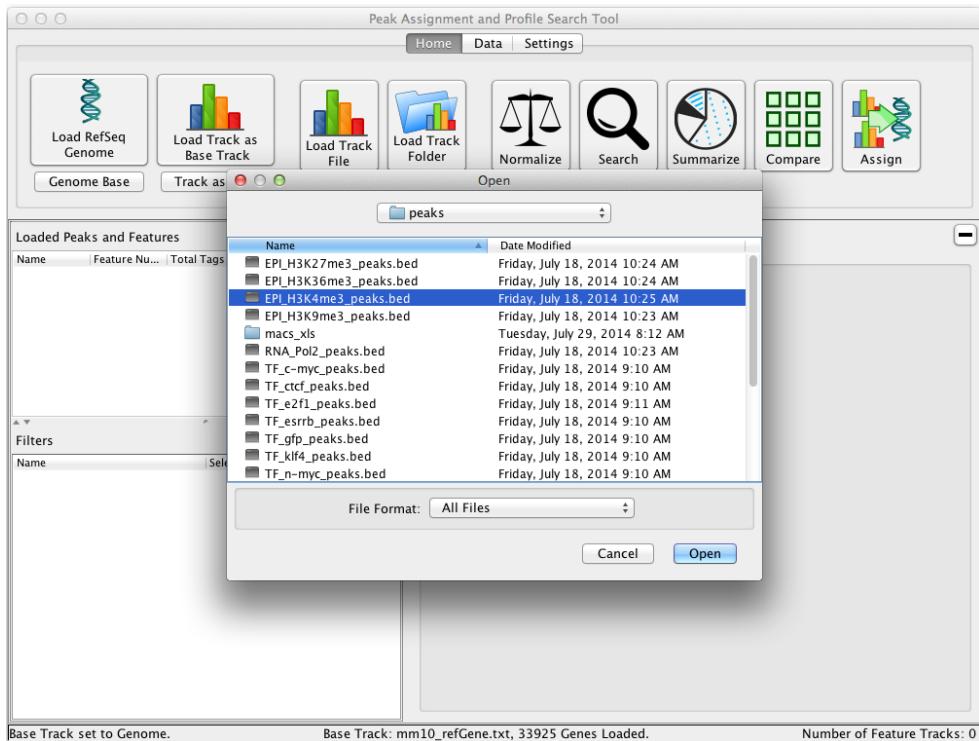
```
chr1 3670348    3672714    583.44  
chr1 4491438    4493852    528.36  
chr1 4571186    4572331    543.51  
chr1 4784782    4786424    1435.29  
...
```

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

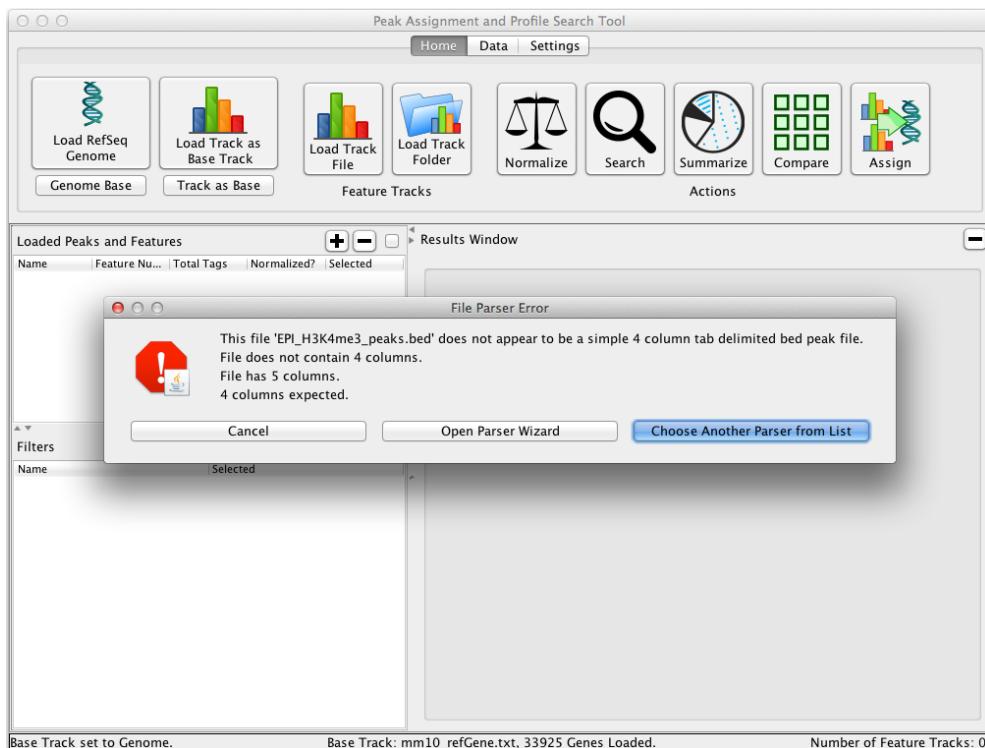
1. Click Load Track File.



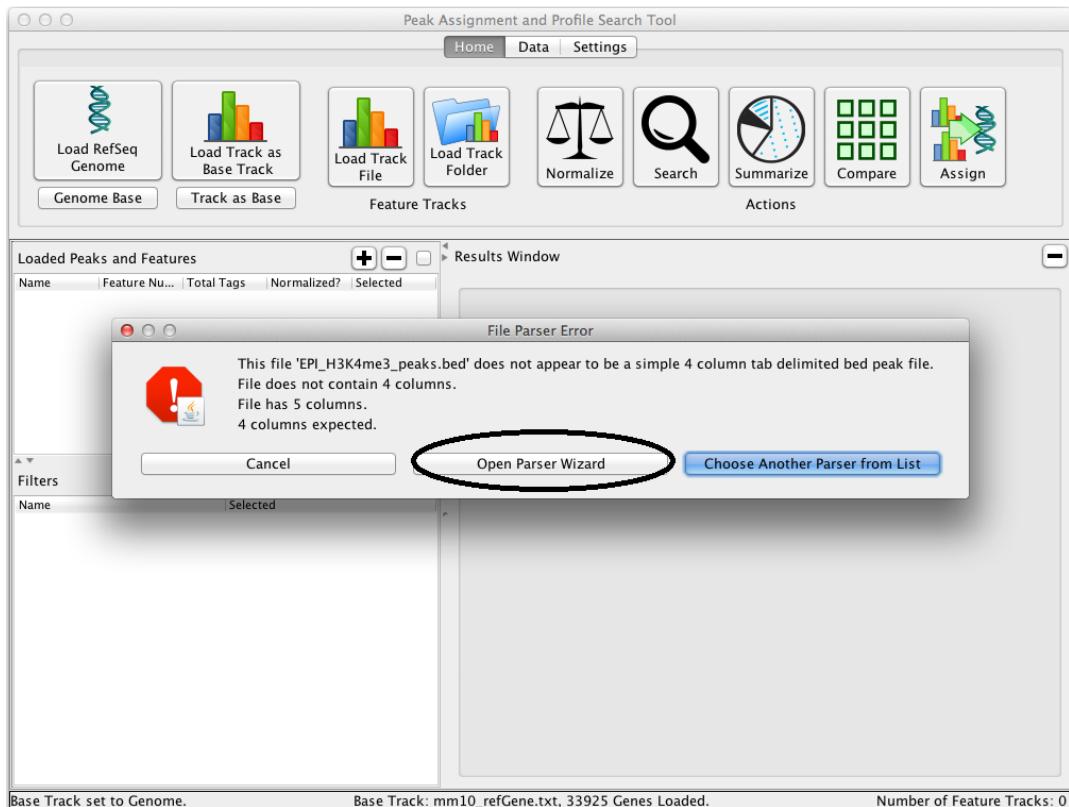
2. Select a track file to load into PAPST



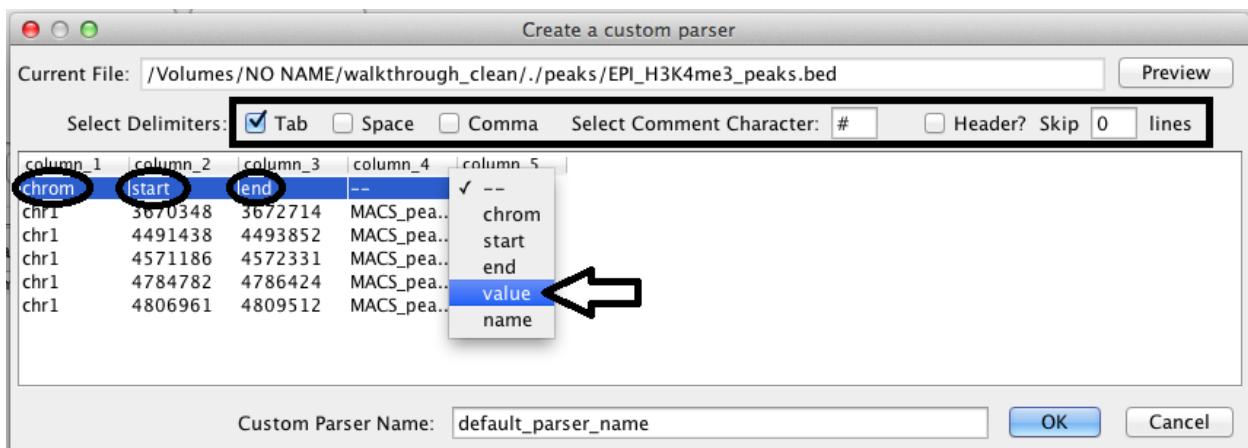
3. If a file does not match the current format, an error message will appear.



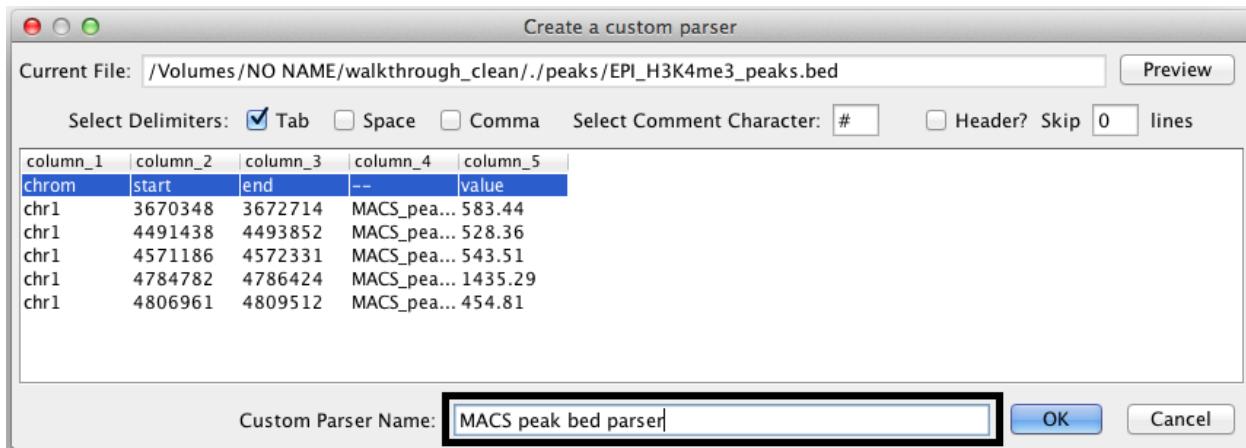
4. To create a new parser for your custom format, click **Open Parser Wizard**



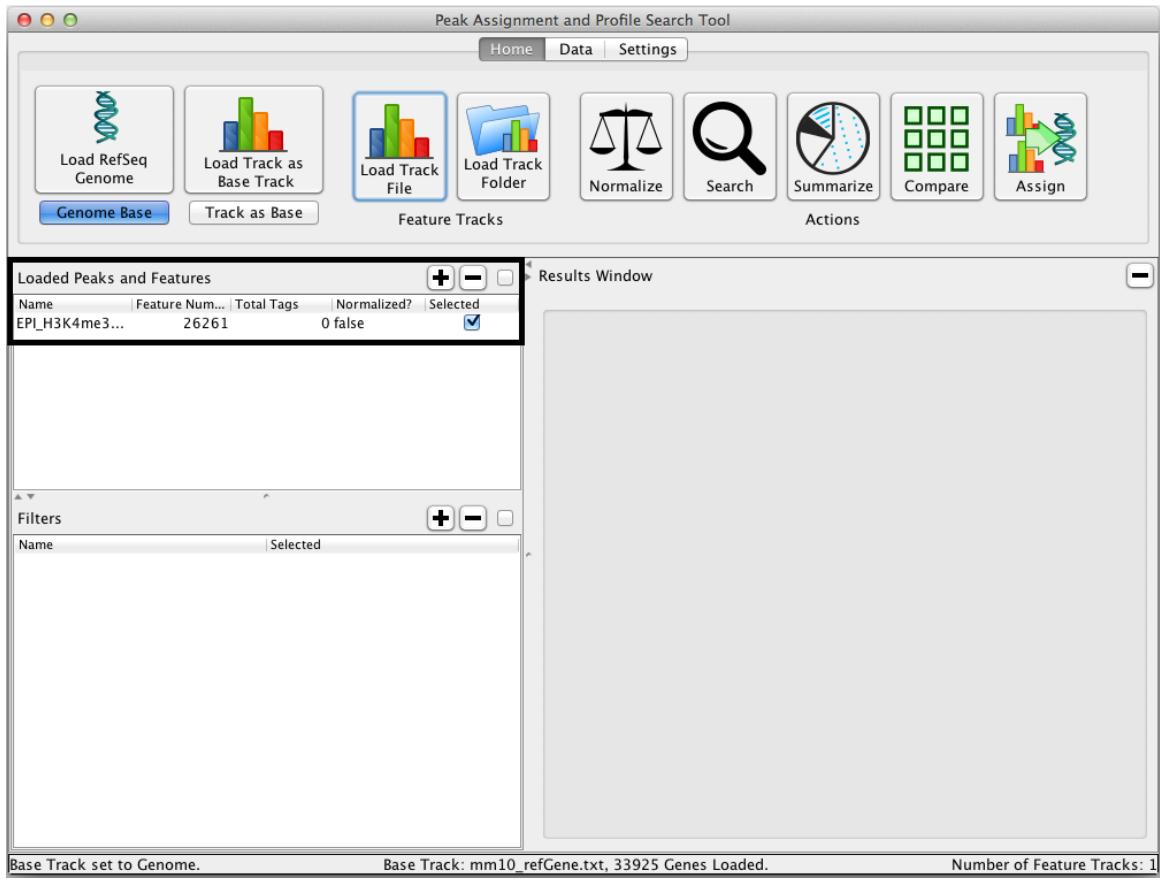
5. To create a custom parser, select the **delimiter** and set the 4 columns **chromosome**, **start**, **end**, and **value**.



6. Give the parser a descriptive name.



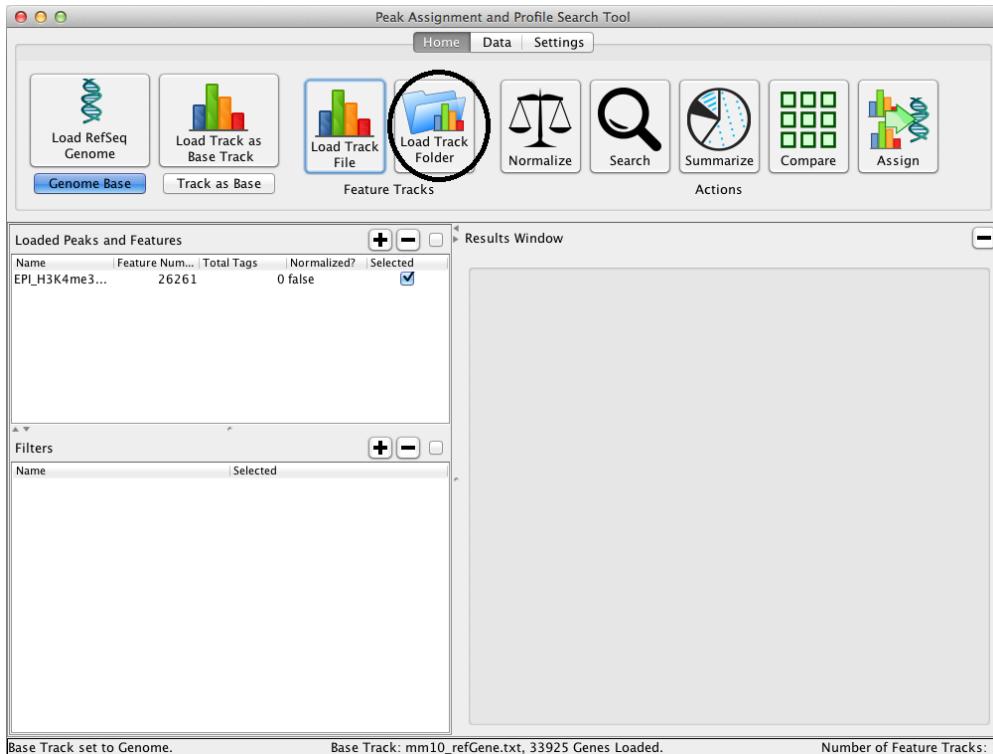
7. Your peak tracks are displayed in the **Loaded Peaks and Features** pane.



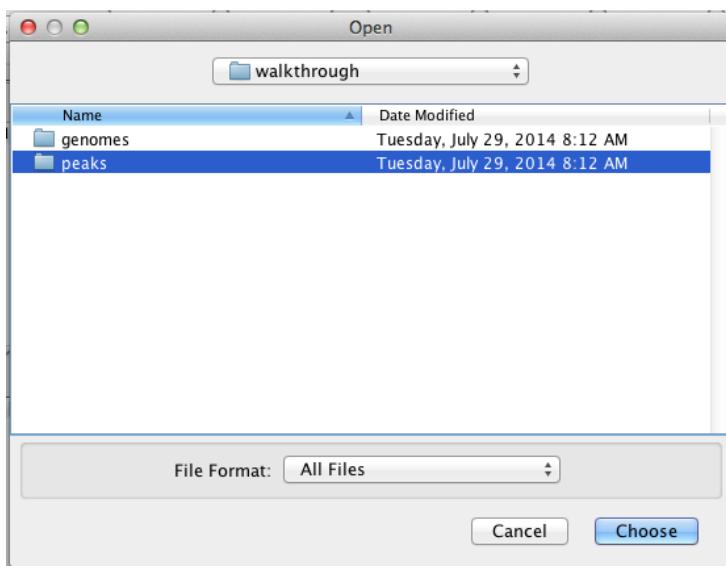
Loading a Folder of Features at Once

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.

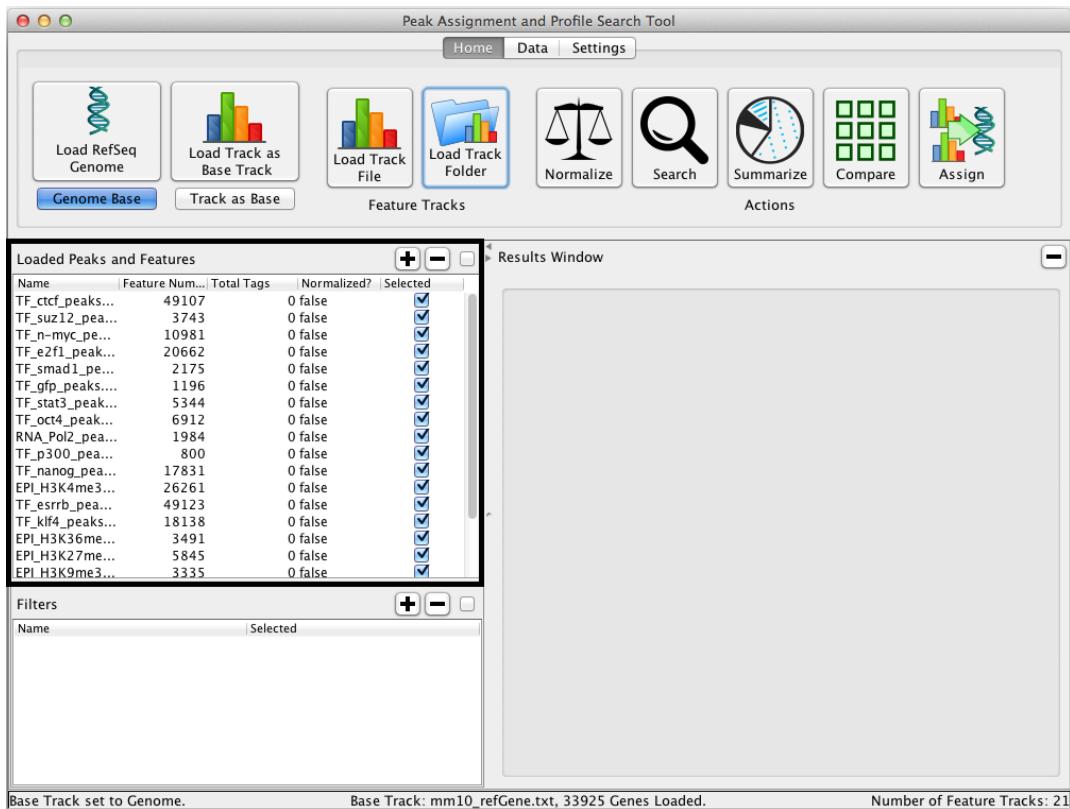
1. Click the Load Track Folder.



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.



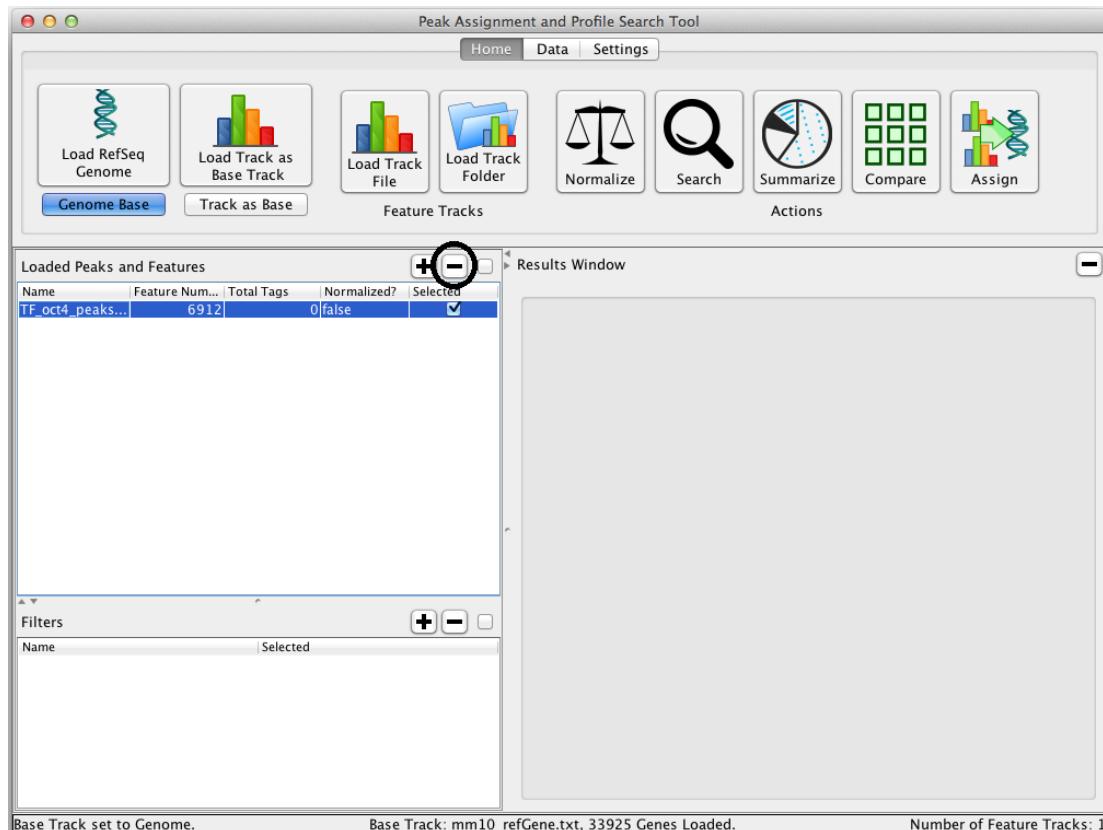
4. If you are using different formats or there are non-text files in your folder, you may encounter an error. You may choose to **use another parser**, (by clicking **Choose Parser From List**) or simply **skip** the file.



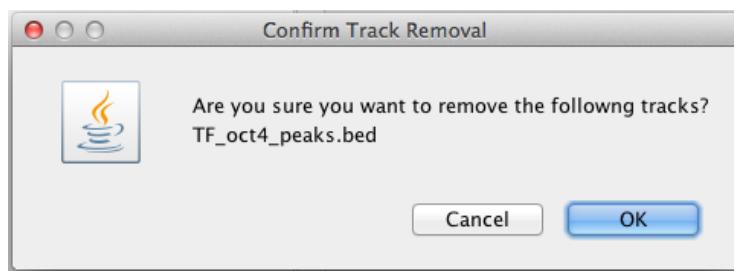
Remove a Peak Set or Feature

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

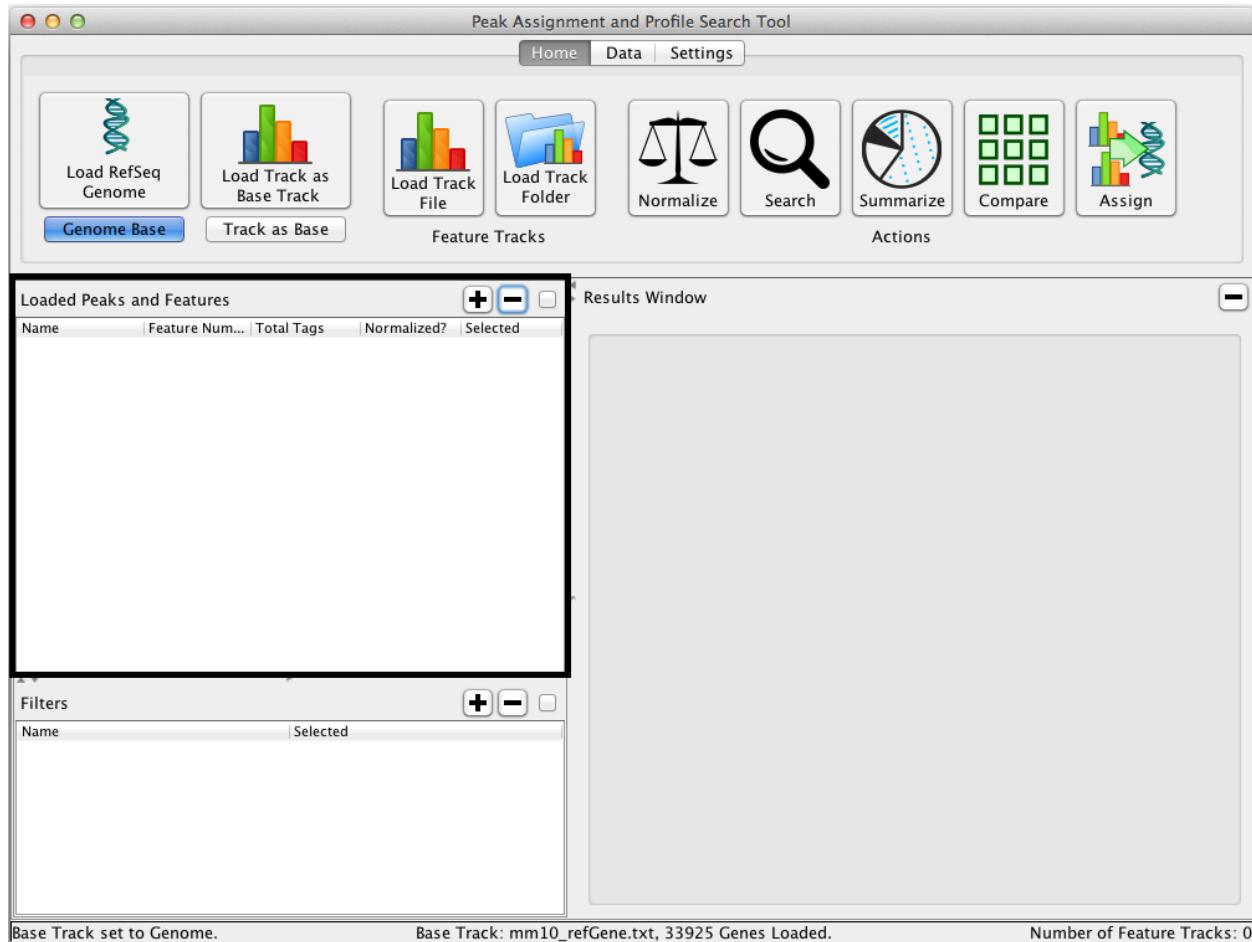
1. Select the peak set you wish to remove. For this example we will select **TF_oct4_peaks.bed**.
2. Click the **minus '-'** button in the upper right corner of the **Loaded Peaks and Features** pane.



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



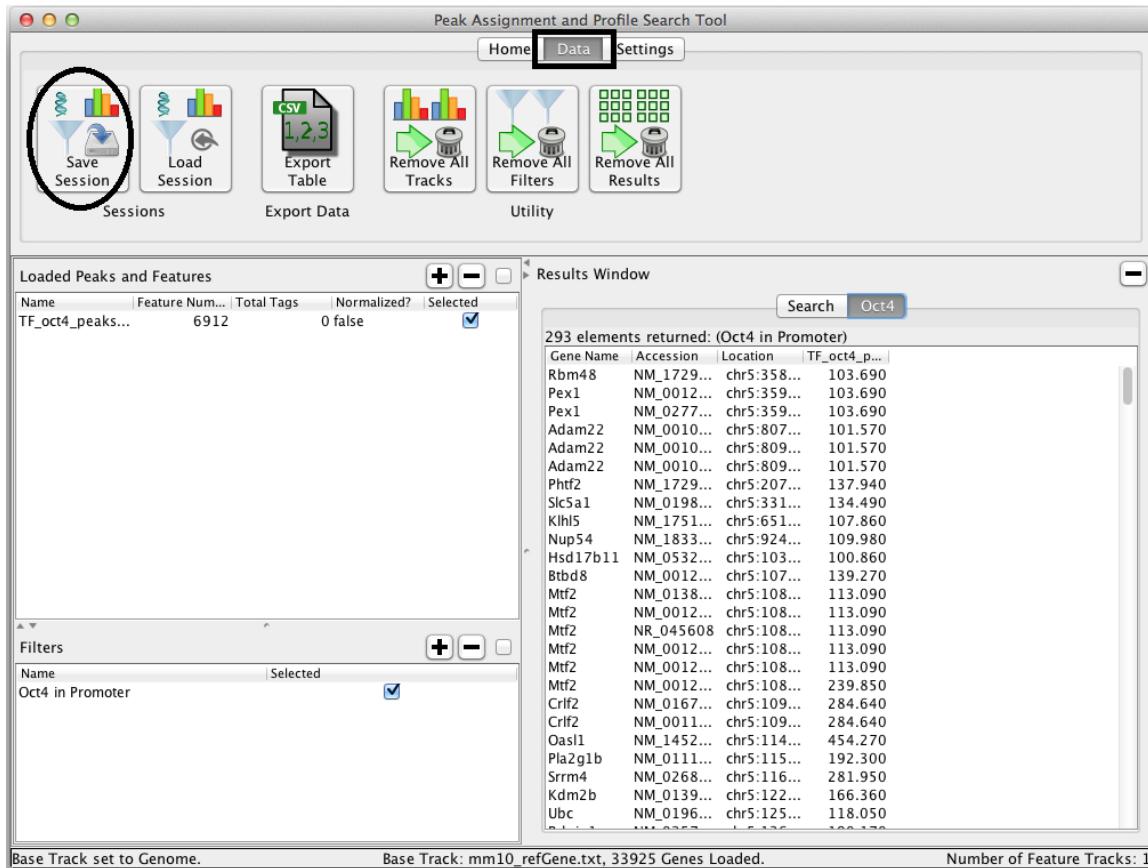
4. Click ok to remove the track. This track will no longer appear in the **Loaded Peaks and Features** pane.



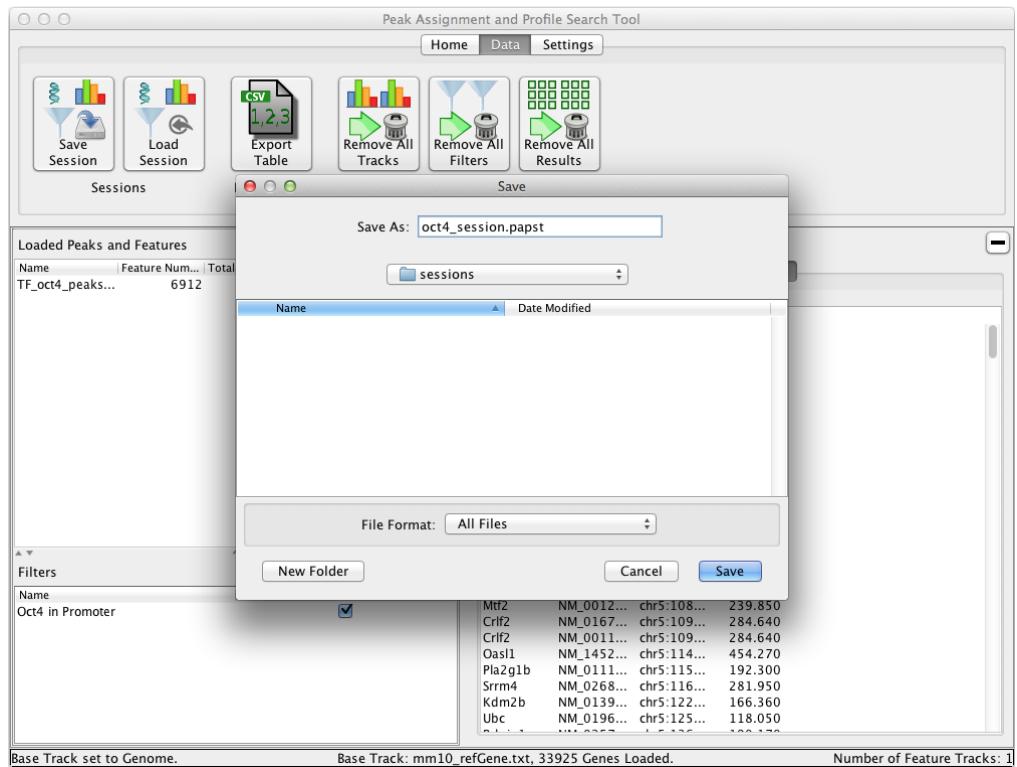
Saving a PAPST Session

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

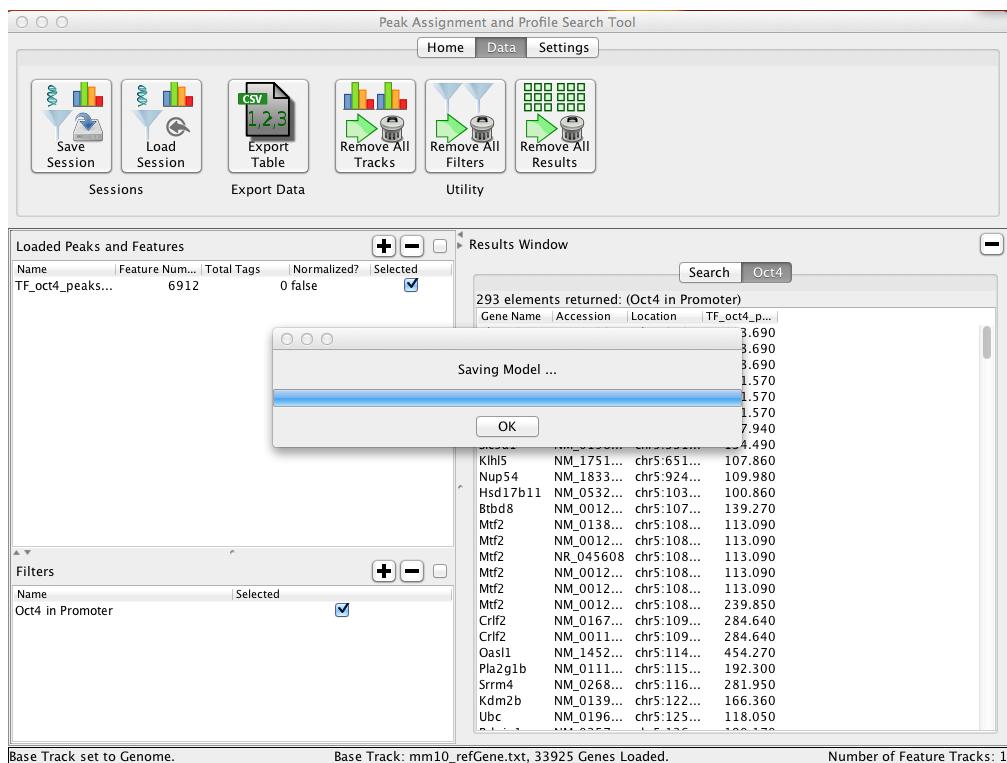
1. To save your PAPST session, navigate to the **Data** tab and click **Save Session**.



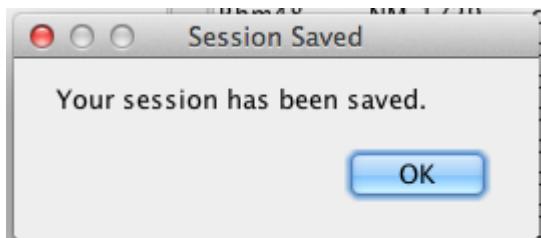
2. You will be prompted to give the session a name. Let's name the file 'oct4_session.papst'.



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



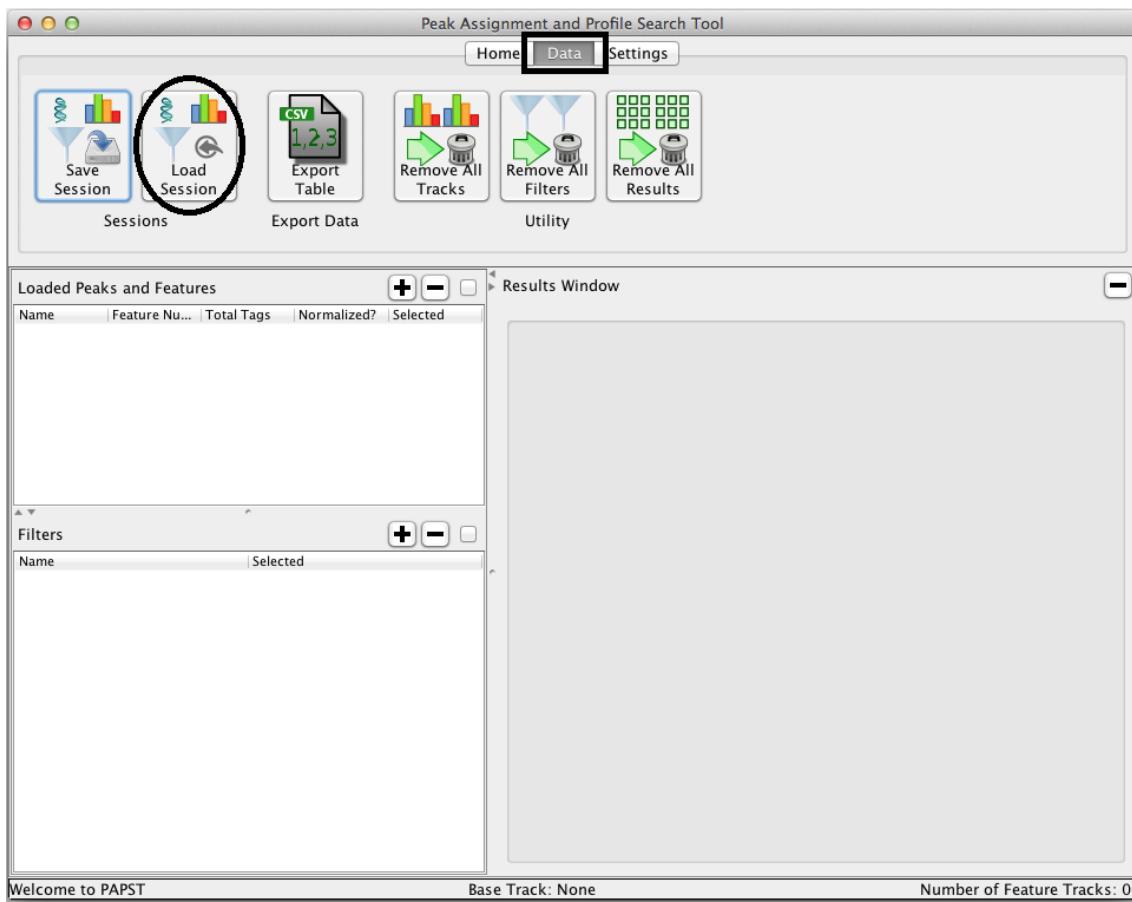
4. A dialog will appear to confirm that the session is saved.



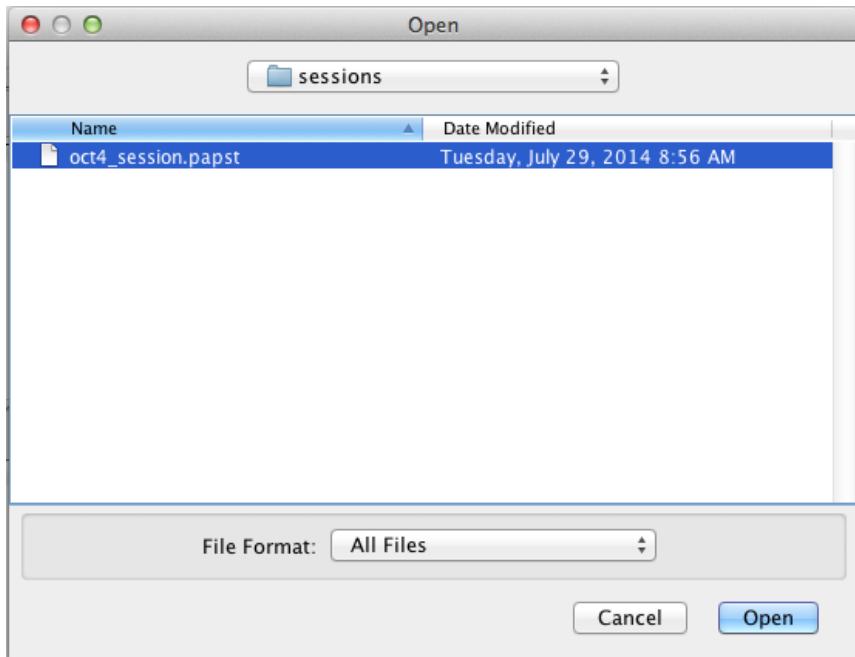
Load a PAPST Session

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

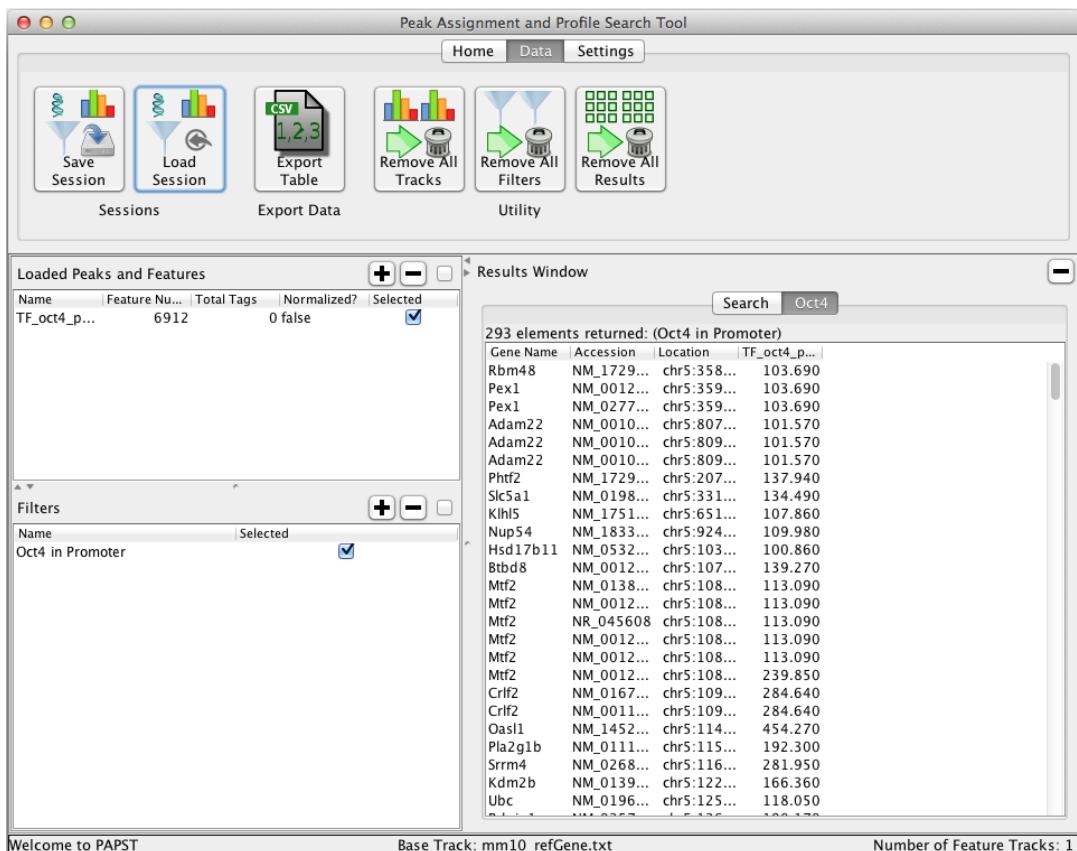
1. To Load a session, navigate to the **Data** tab and click **Load Session**.



2. Select the desired file.



3. Your data will appear once it has loaded.



Working with Filters

Searching with Filters

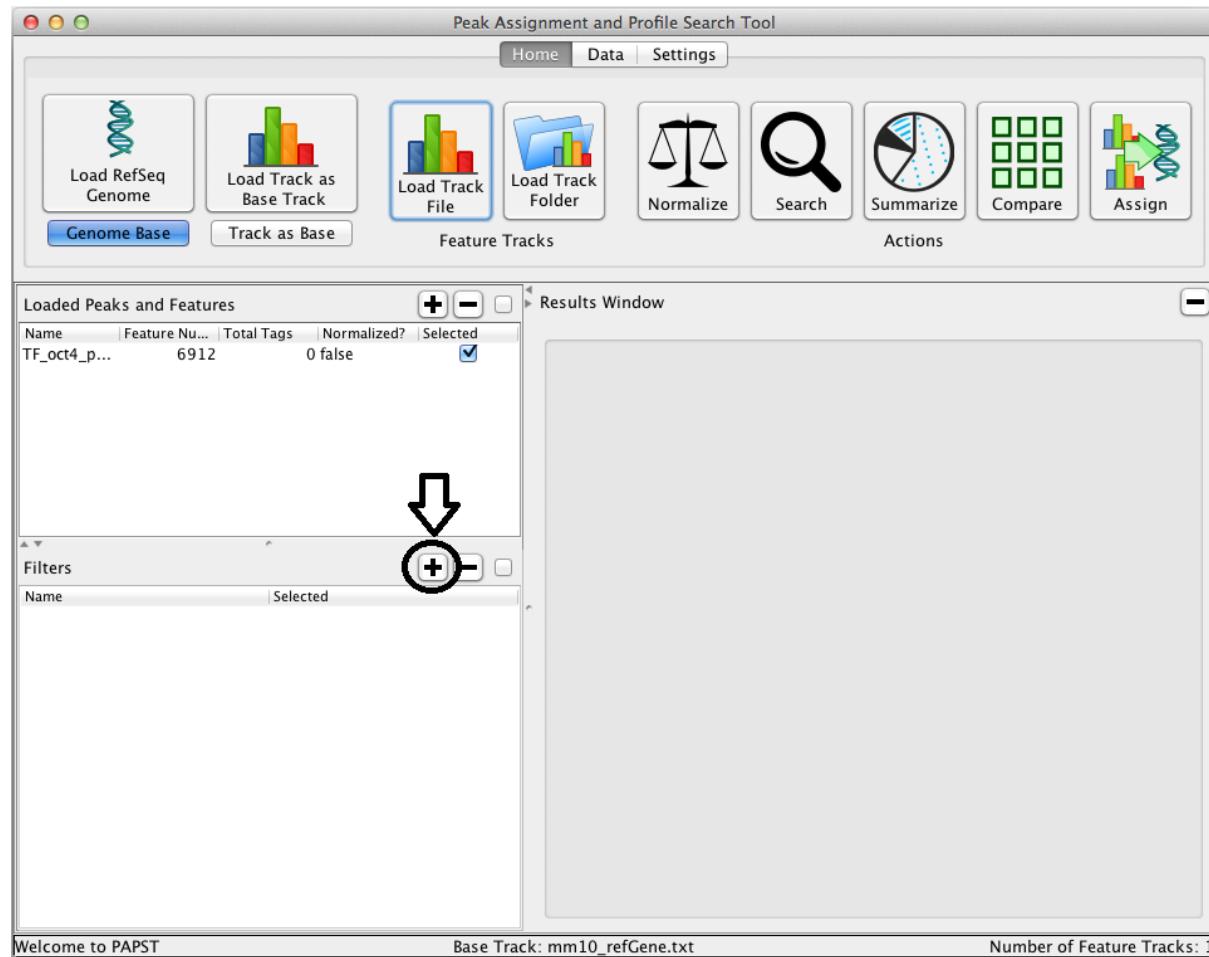
The most powerful feature of PAPST is the ability to perform gene searches using track filters. Each filter imposes a constraint on the set of genes that are returned.

Let's start with a simple example.

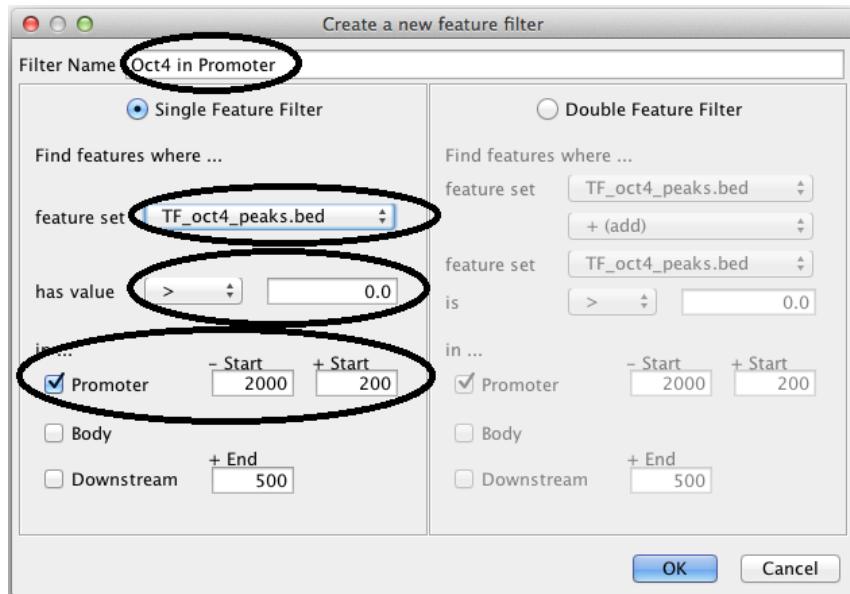
1. Using the provided data, load the mm10 genome.
2. Load the file named TF_oct4_peaks.bed. (You may need to create a new parser)

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.

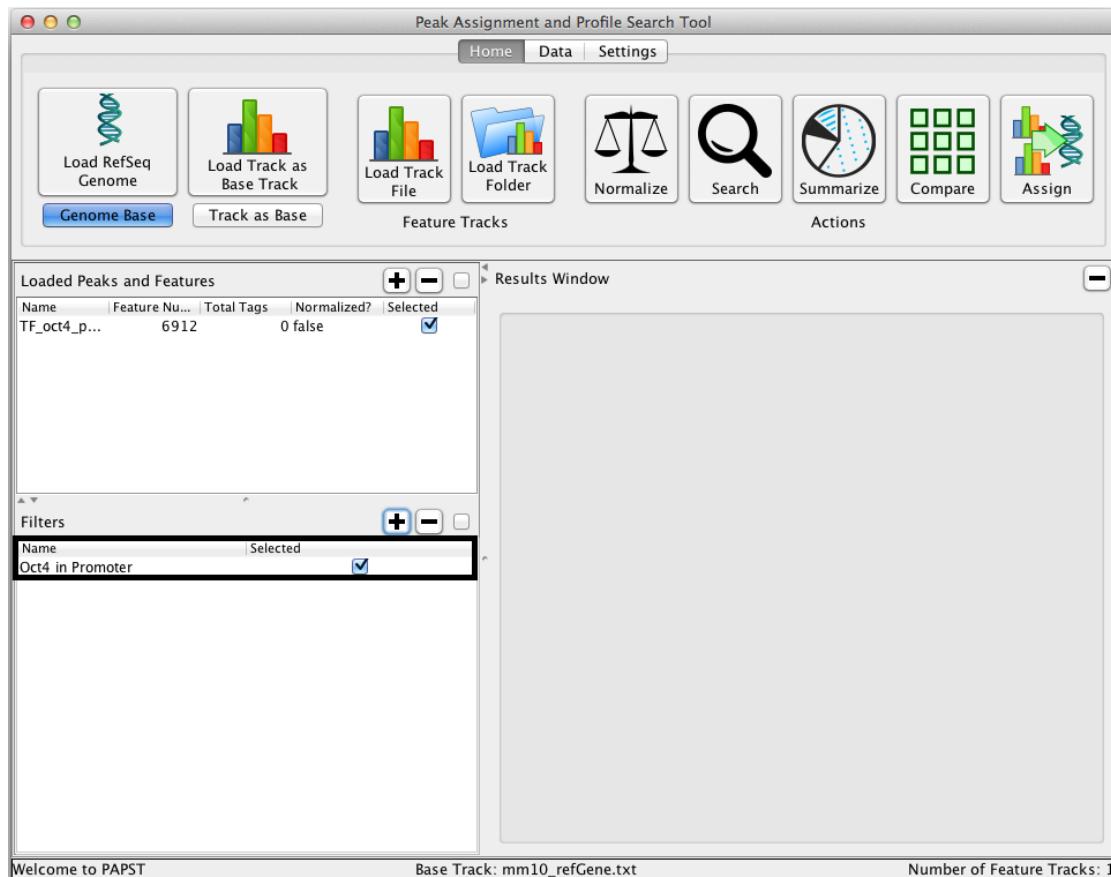
3. Click the new filter button denoted by the plus sign, '+'.



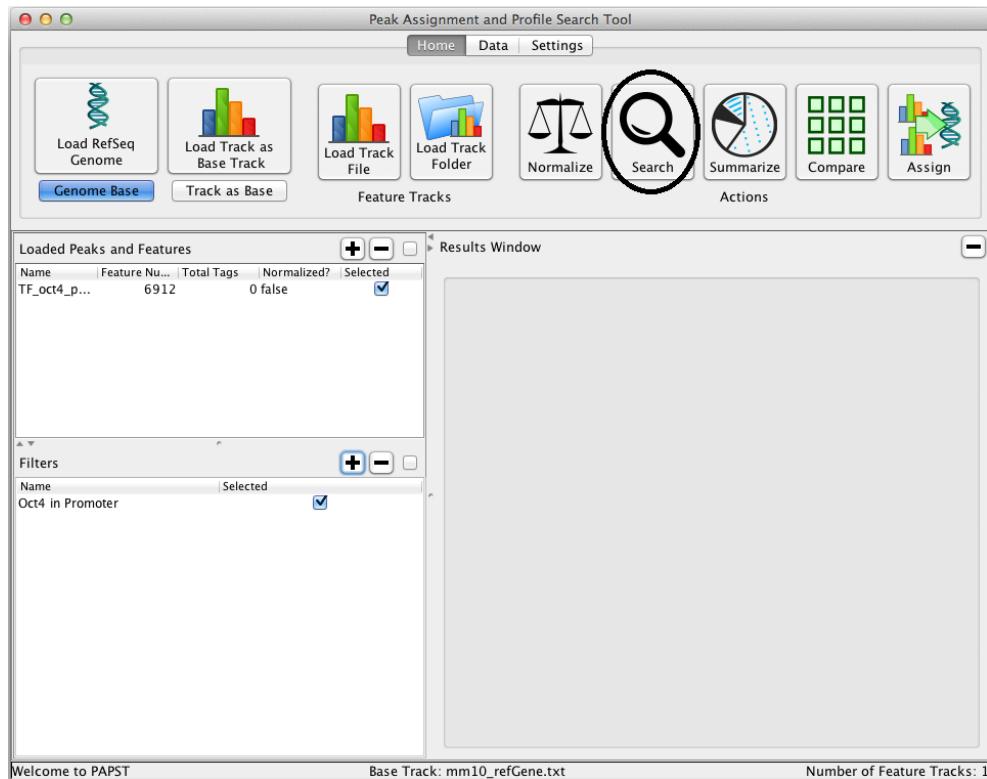
4. Select the **TF_oct4_peaks.bed** feature set and adjust the promoter settings for this single feature filter. Give the filter a name. The threshold determines the stringency of the search. Click **OK** to create the filter.



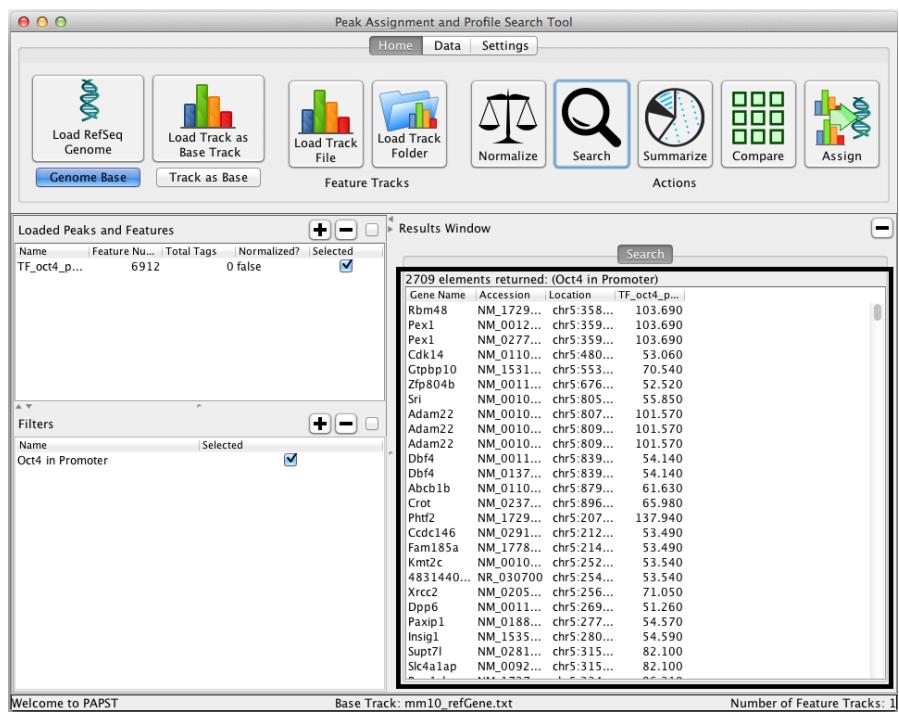
5. The created filter will appear in the **Filters** pane.



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



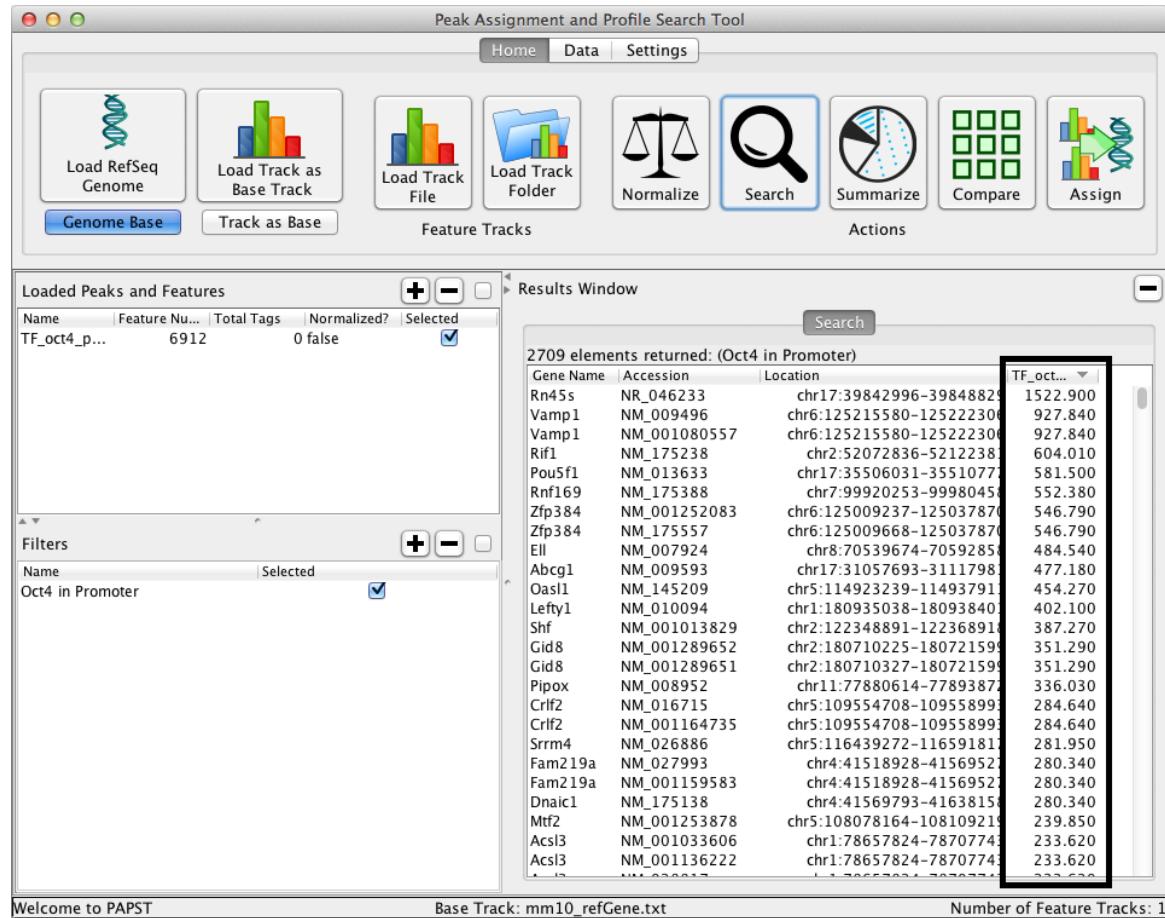
7. The results appear as a table in the **Results Window**.



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 values. 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.

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

Peak Assignment and Profile Search Tool

Settings (selected)

Show Gene Isoforms? (checked and circled)

Parser Settings

Search Settings

Normalization Settings

Loaded Peaks and Features

Name	Feature Nu...	Total Tags	Normalized?	Selected
TF_oct4_p...	6912	0	false	<input checked="" type="checkbox"/>

Filters

Name	Selected
Oct4 in Promoter	<input checked="" type="checkbox"/>

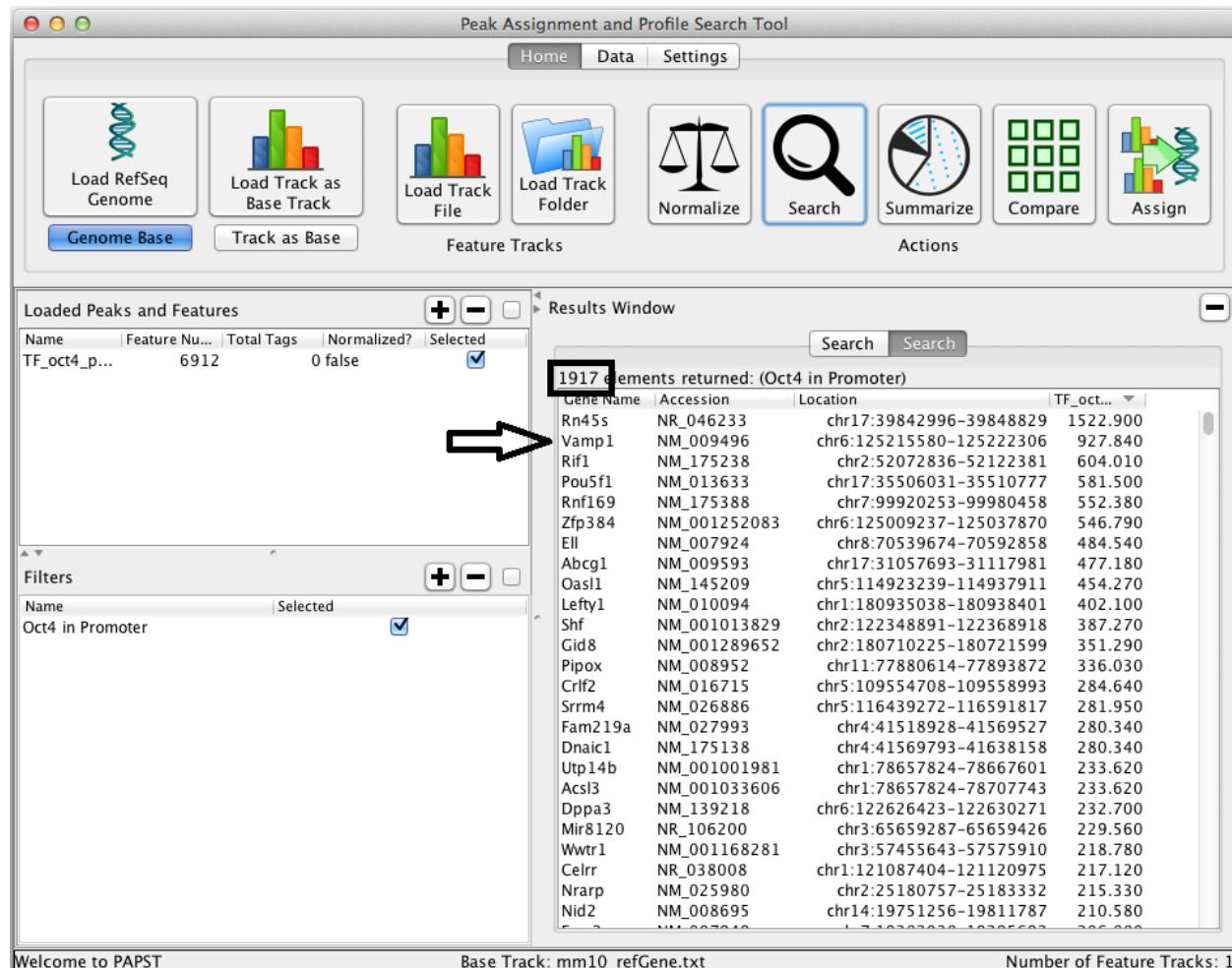
Results Window

2709 elements returned: (Oct4 in Promoter)

Gene Name	Accession	Location	TF_oct...
Rn45s	NR_046233	chr17:39842996-39848829	1522.900
Vamp1	NM_009496	chr6:125215580-125222306	927.840
Vamp1	NM_001080557	chr6:125215580-125222306	927.840
Rif1	NM_175238	chr2:52072836-52122381	604.010
Pou5f1	NM_013633	chr17:35506031-35510777	581.500
Rnf169	NM_175388	chr7:99920253-99980458	552.380
Zfp384	NM_001252083	chr6:125009237-125037870	546.790
Zfp384	NM_175557	chr6:125009668-125037870	546.790
Ell	NM_007924	chr8:70539674-70592858	484.540
Abcg1	NM_009593	chr17:31057693-31117981	477.180
Oasl1	NM_145209	chr5:114923239-114937911	454.270
Lefty1	NM_010094	chr1:180935038-180938401	402.100
Shf	NM_001013829	chr2:122348891-122368918	387.270
Gid8	NM_001289652	chr2:180710225-180721599	351.290
Gid8	NM_001289651	chr2:180710327-180721599	351.290
Pipox	NM_008952	chr11:77880614-77893872	336.030
Crlf2	NM_016715	chr5:109554708-109558993	284.640
Crlf2	NM_001164735	chr5:109554708-109558993	284.640
Srrm4	NM_026886	chr5:116439272-116591817	281.950
Fam219a	NM_027993	chr4:41518928-41569527	280.340
Fam219a	NM_001159583	chr4:41518928-41569527	280.340
Dnaic1	NM_175138	chr4:41569793-41638158	280.340
Mtf2	NM_001253878	chr5:108078164-108109219	239.850
Acsl3	NM_001033606	chr1:78657824-78707743	233.620
Acsl3	NM_001136222	chr1:78657824-78707743	233.620

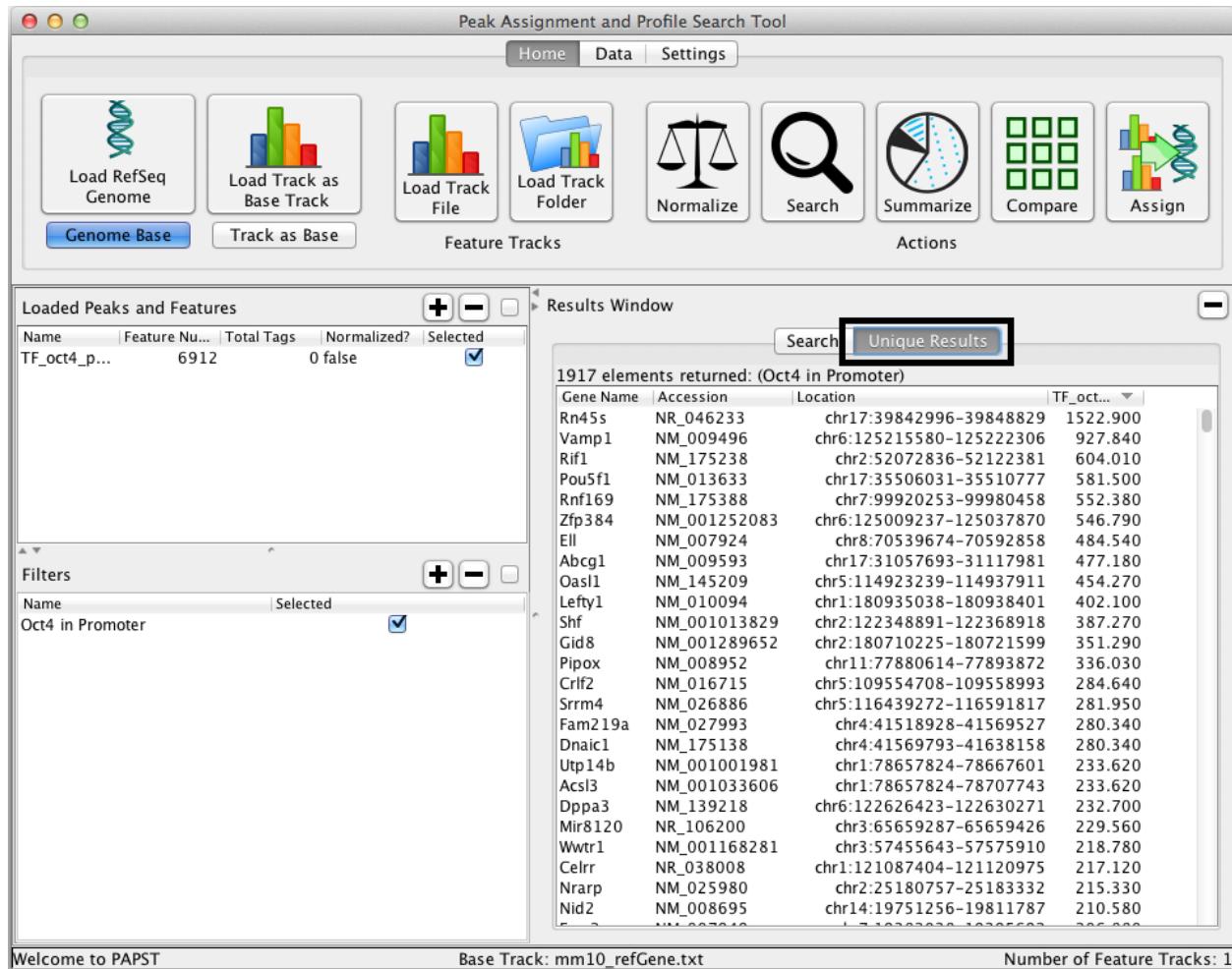
Welcome to PAPST Base Track: mm10_refGene.txt Number of Feature Tracks: 1

10. Repeat the search.



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

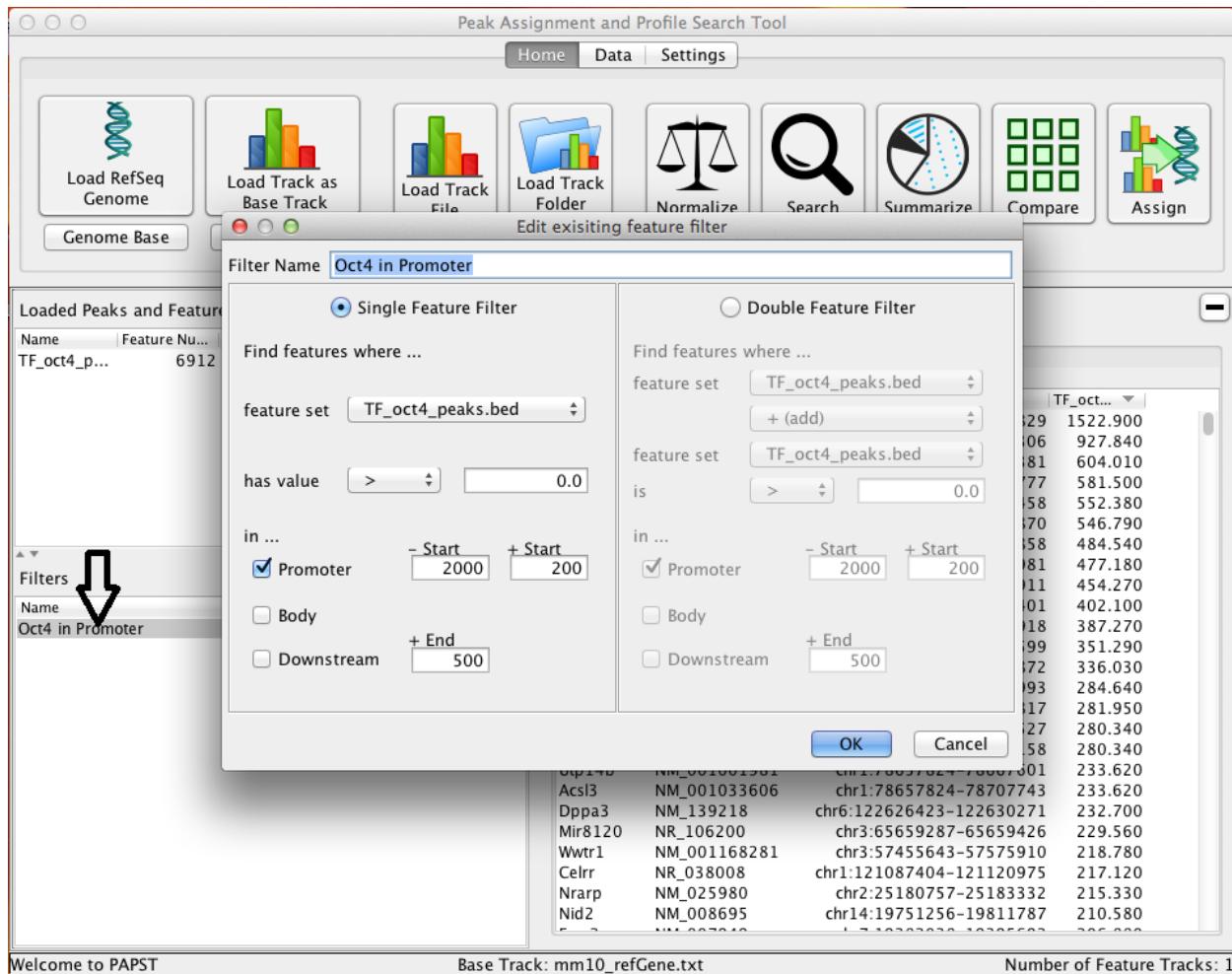
11. Edit a **Results** tab by double clicking on the tab label and typing the new label. Press enter to apply.



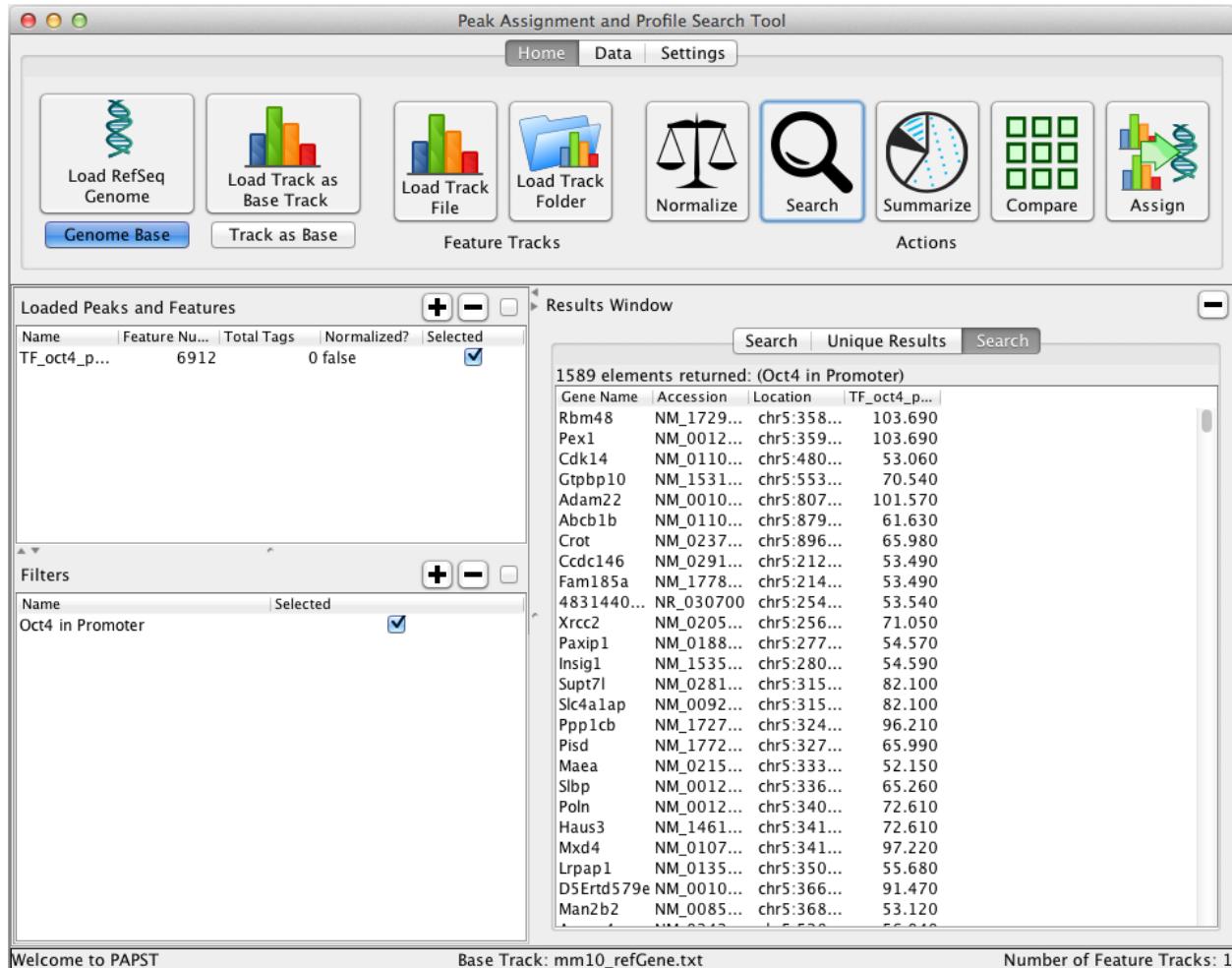
Editing Filters

Filters can be edited easily by double clicking the filter in the window.

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

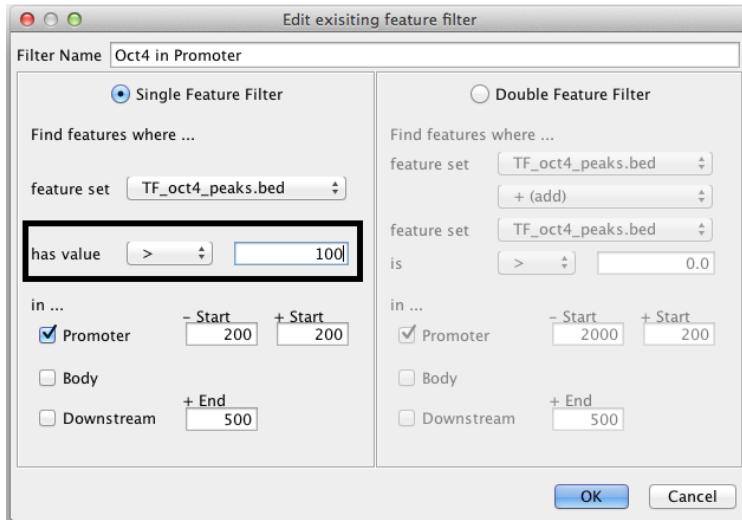


2. Let's change the promoter region to be -200bp +200bp relative to the promoter region and **Search** again.

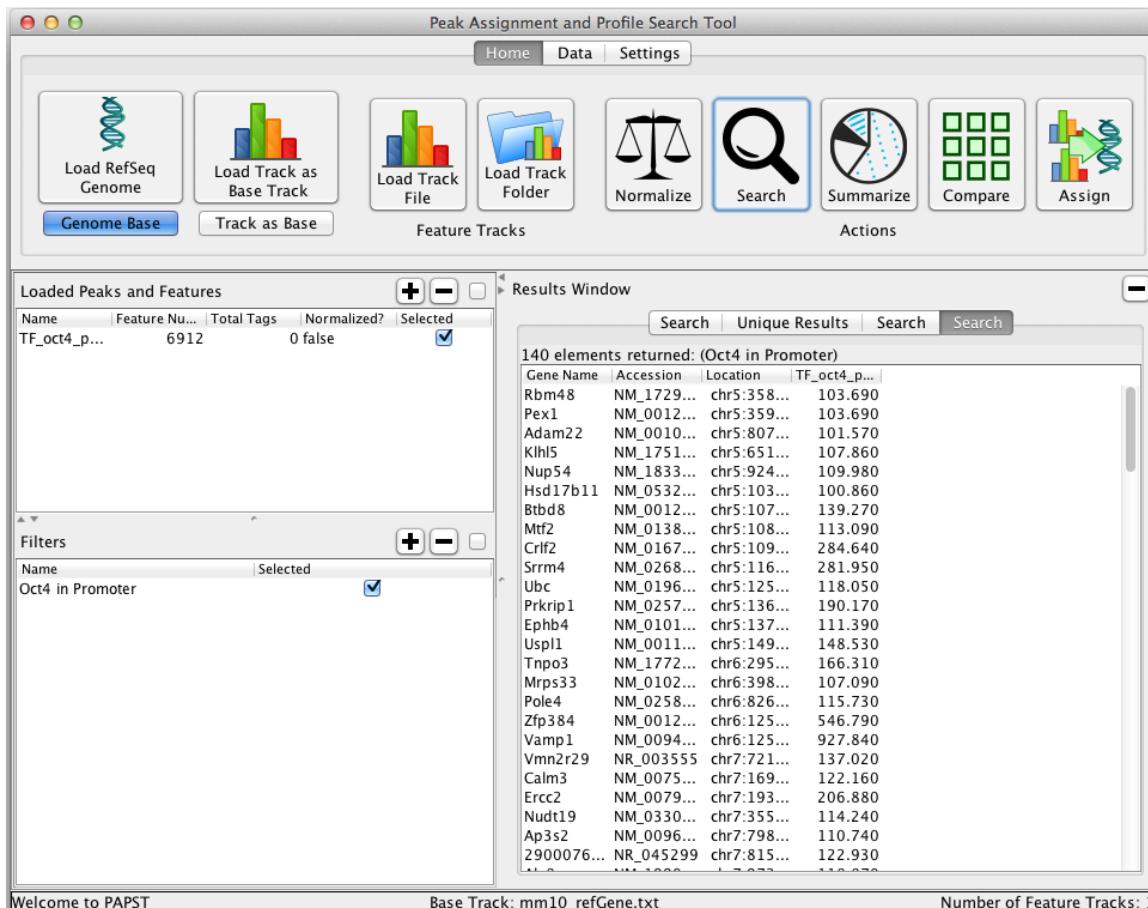


This reduces the number of unique genes to 1589 from 1917.

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.



Welcome to PAPST

Base Track: mm10_refGene.txt

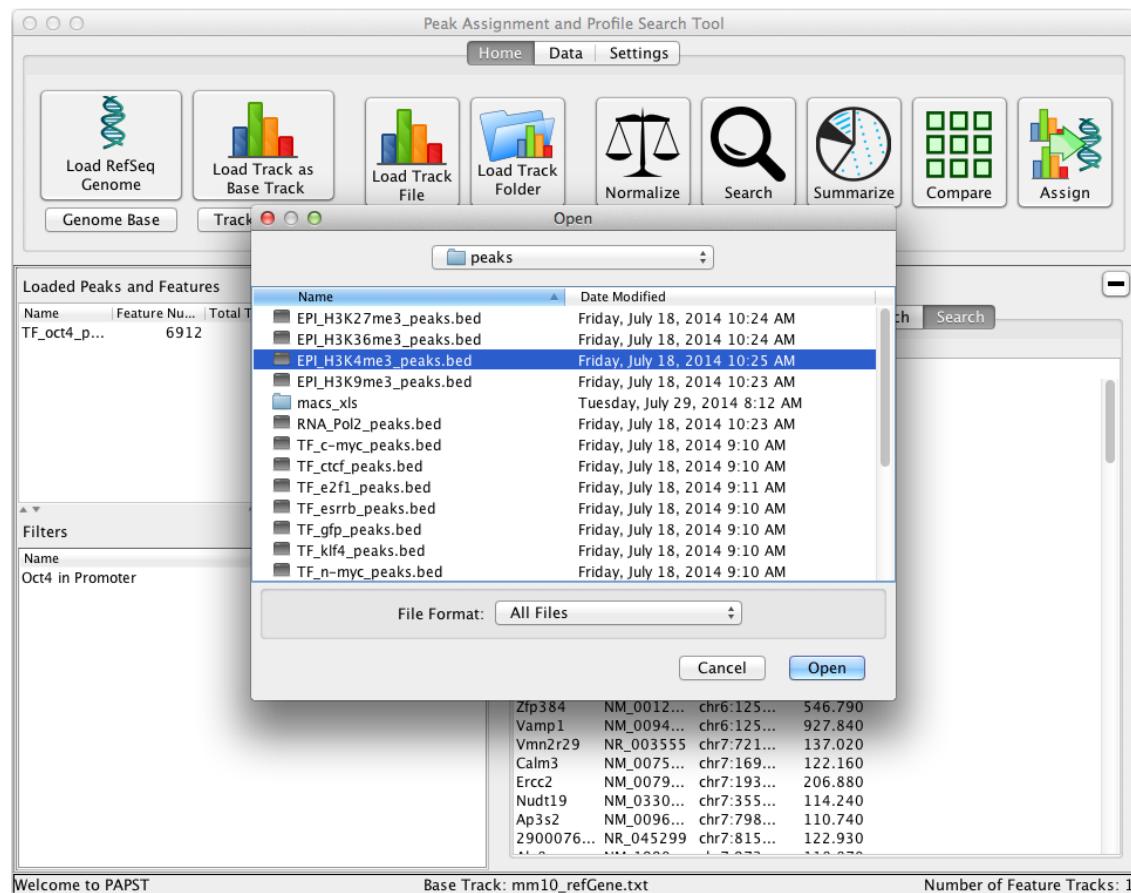
Number of Feature Tracks: 1

Using Multiple Filters

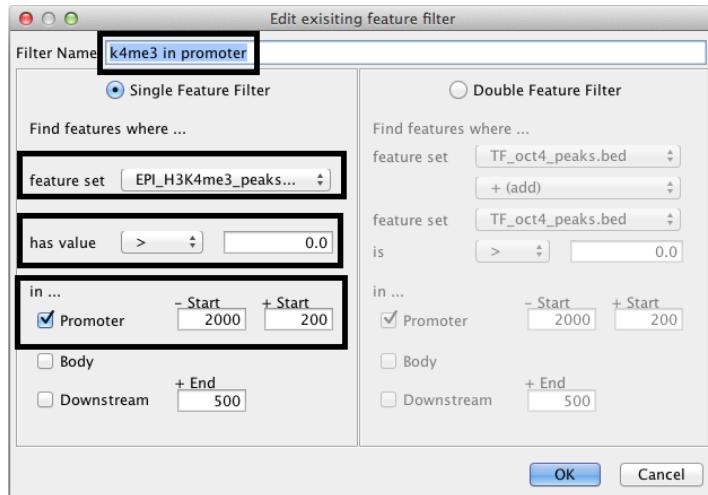
Using multiple filters allows the user to create complex patterns. Filters have ‘**and**’ semantics, meaning that for a gene to appear in the results it **must pass all filters (Filter1 and Filter2 and ...)**. In this way the user can iteratively define a set of interesting genes.

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 examples.
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.

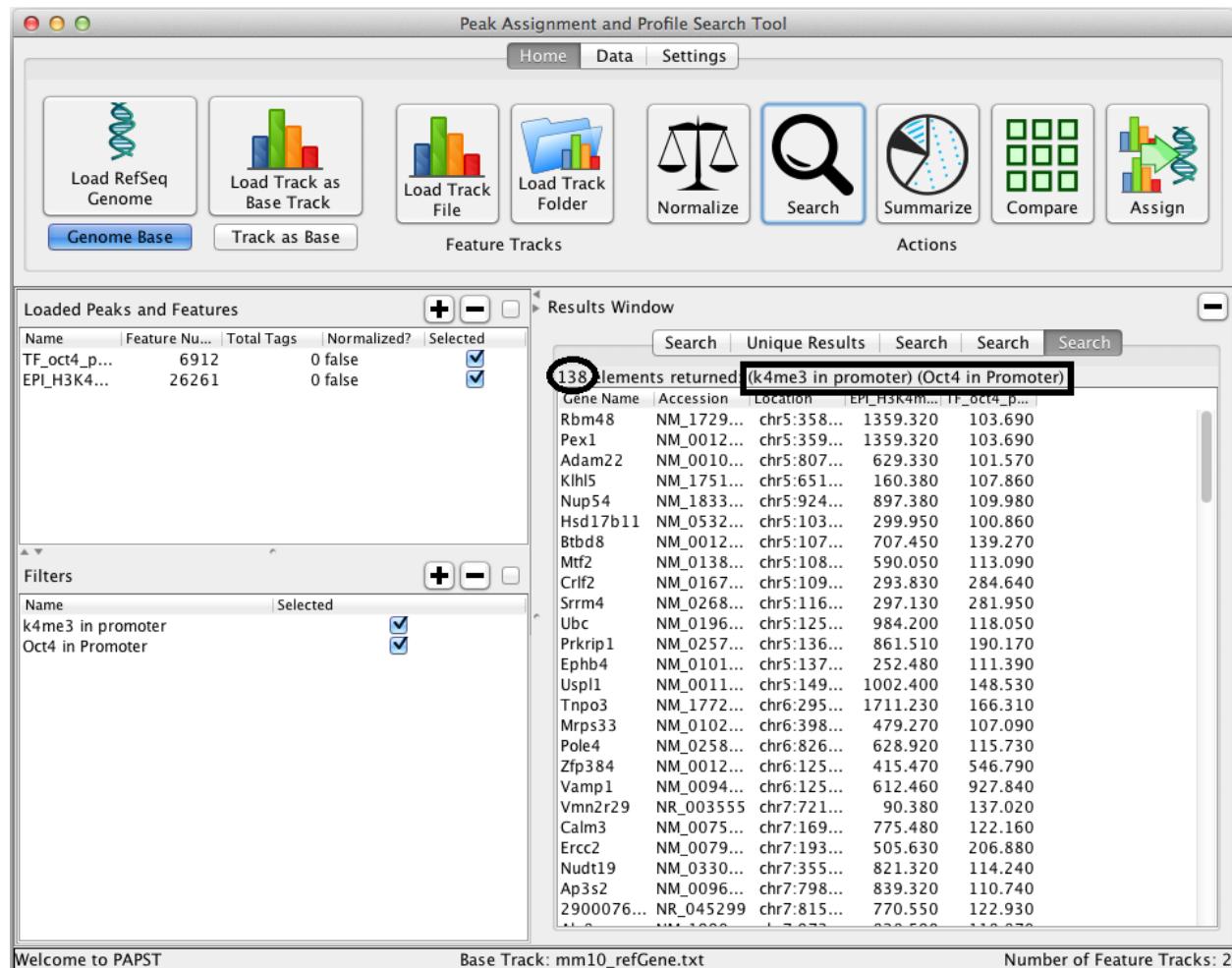


4. Now we have two filters loaded into PAPST.

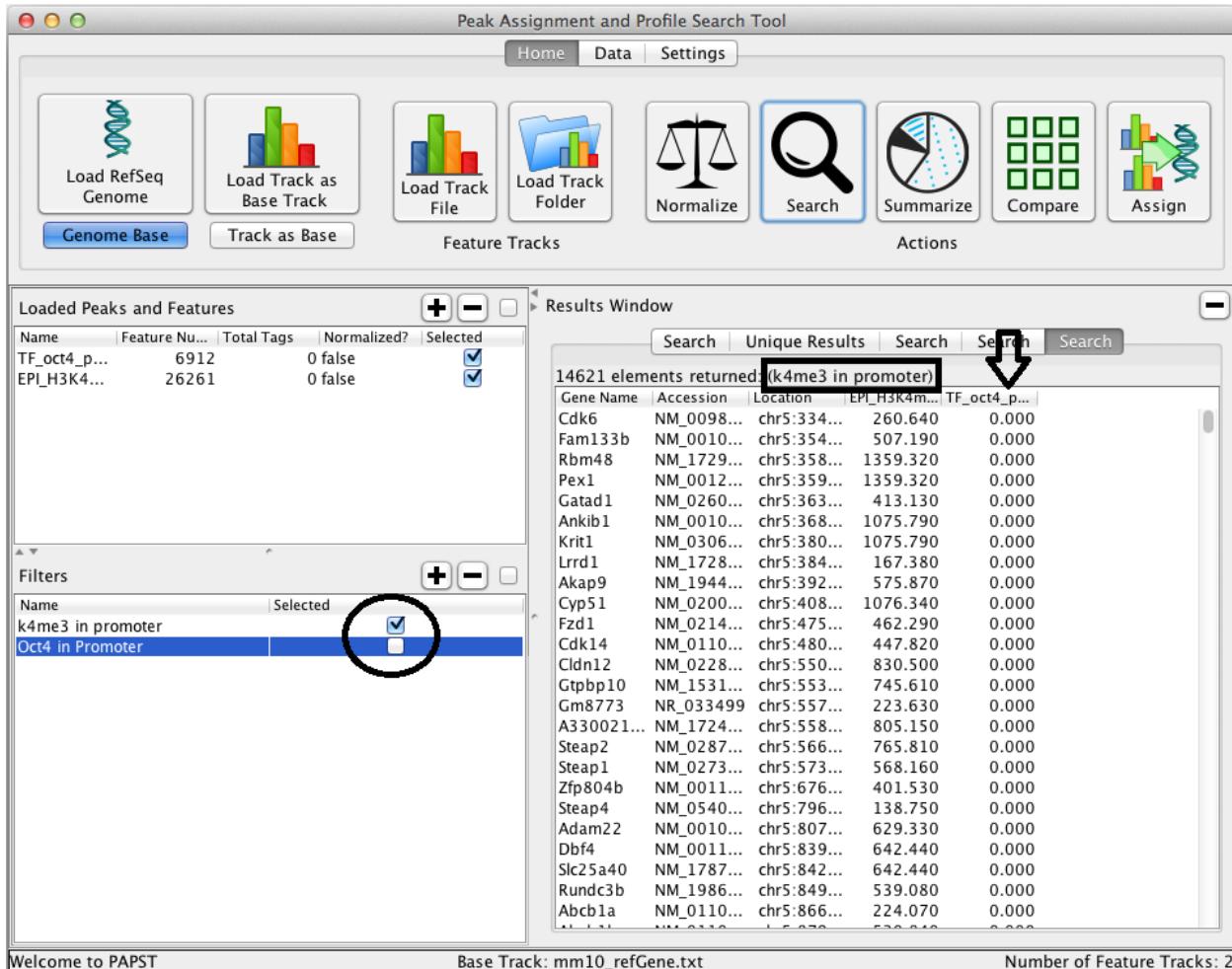
Name	Feature Nu...	Total Tags	Normalized?	Selected
TF_oct4_p...	6912	0	false	<input checked="" type="checkbox"/>
EPI_H3K4...	26261	0	false	<input checked="" type="checkbox"/>

140 elements returned: (Oct4 in Promoter)			
Gene Name	Accession	Location	TF_oct4_p...
Rbm48	NM_1729...	chr5:358...	103.690
Pexl	NM_0012...	chr5:359...	103.690
Adam22	NM_0010...	chr5:807...	101.570
Klf15	NM_1751...	chr5:651...	107.860
Nup54	NM_1833...	chr5:924...	109.980
Hsd17b11	NM_0532...	chr5:103...	100.860
Btbd8	NM_0012...	chr5:107...	139.270
Mtf2	NM_0138...	chr5:108...	113.090
Crlf2	NM_0167...	chr5:109...	284.640
Srrm4	NM_0268...	chr5:116...	281.950
Ubc	NM_0196...	chr5:125...	118.050
Prkrp1	NM_0257...	chr5:136...	190.170
Ephb4	NM_0101...	chr5:137...	111.390
Usp11	NM_0011...	chr5:149...	148.530
Tnpo3	NM_1772...	chr6:295...	166.310
Mrps33	NM_0102...	chr6:398...	107.090
Pole4	NM_0258...	chr6:826...	115.730
Zfp384	NM_0012...	chr6:125...	546.790
Vamp1	NM_0094...	chr6:125...	927.840
Vmn2r29	NR_003555	chr7:721...	137.020
Calm3	NM_0075...	chr7:169...	122.160
Erc2	NM_0079...	chr7:193...	206.880
Nudt19	NM_0330...	chr7:355...	114.240
Ap3s2	NM_0096...	chr7:798...	110.740
2900076...	NR_045299	chr7:815...	122.930
...

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.

Name	Feature Nu...	Total Tags	Normalized?	Selected
TF_oct4_p...	6912	0	false	<input checked="" type="checkbox"/>
EPI_H3K4...	26261	0	false	<input checked="" type="checkbox"/>

Filters

Name	Selected
k4me3 in promoter	<input checked="" type="checkbox"/>
Oct4 in Promoter	<input checked="" type="checkbox"/>

Results Window

14621 elements returned: (k4me3 in promoter)			
Gene Name	Accession	Location	EPI_H3K4m...
Cdk6	NM_0098...	chr5:334...	260.640
Fam133b	NM_0010...	chr5:354...	507.190
Rbm48	NM_1729...	chr5:358...	1359.320
Pex1	NM_0012...	chr5:359...	1359.320
Gata1	NM_0260...	chr5:363...	413.130
Ankib1	NM_0010...	chr5:368...	1075.790
Krit1	NM_0306...	chr5:380...	1075.790
Lrrd1	NM_1728...	chr5:384...	167.380
Akap9	NM_1944...	chr5:392...	575.870
Cyp51	NM_0200...	chr5:408...	1076.340
Fzd1	NM_0214...	chr5:475...	462.290
Cdk14	NM_0110...	chr5:480...	447.820
Cldn12	NM_0228...	chr5:550...	830.500
Gtpbp10	NM_1531...	chr5:553...	745.610
Gm8773	NR_033499	chr5:557...	223.630
A330021...	NM_1724...	chr5:558...	805.150
Steap2	NM_0287...	chr5:566...	765.810
Steap1	NM_0273...	chr5:573...	568.160
Zfp804b	NM_0011...	chr5:676...	401.530
Steap4	NM_0540...	chr5:796...	138.750
Adam22	NM_0010...	chr5:807...	629.330
Dbf4	NM_0011...	chr5:839...	642.440
Slc25a40	NM_1787...	chr5:842...	642.440
Rundc3b	NM_1986...	chr5:849...	539.080
Abcb1a	NM_0110...	chr5:866...	224.070

Welcome to PAPST Base Track: mm10_refGene.txt Number of Feature Tracks: 2

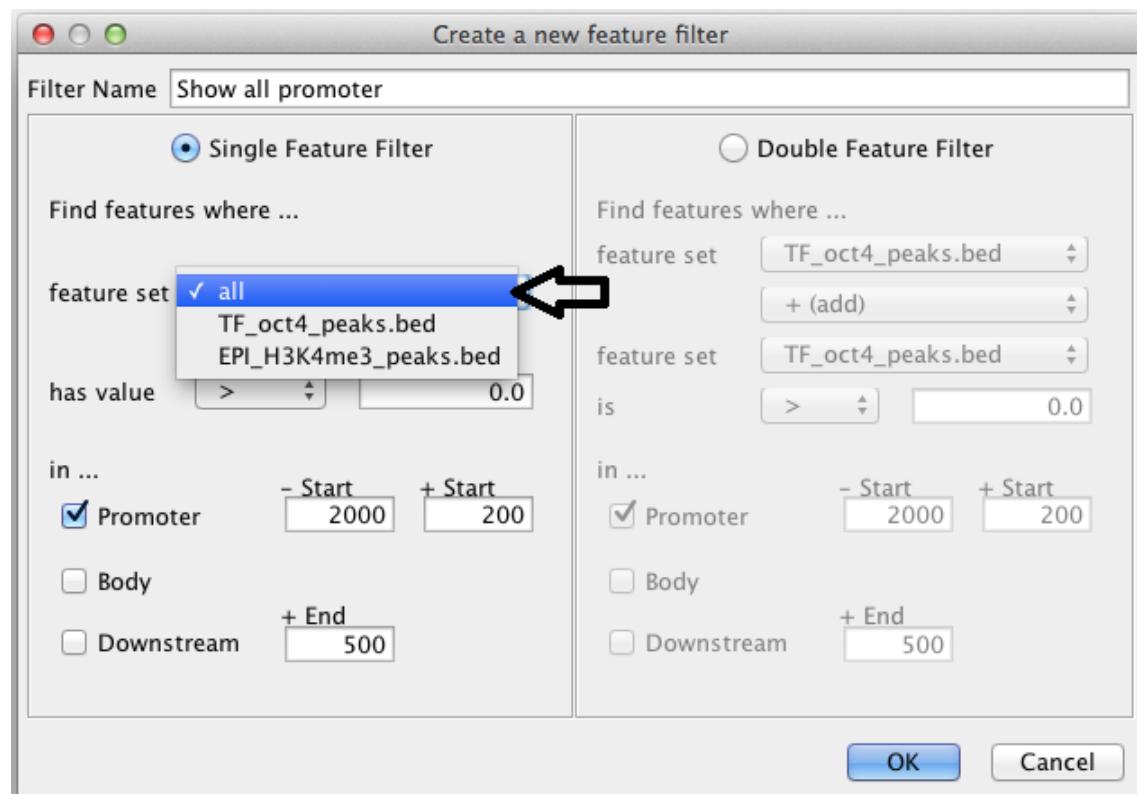
Now we see that the Oct4 column is hidden.

The ALL filter, displaying a value for all columns

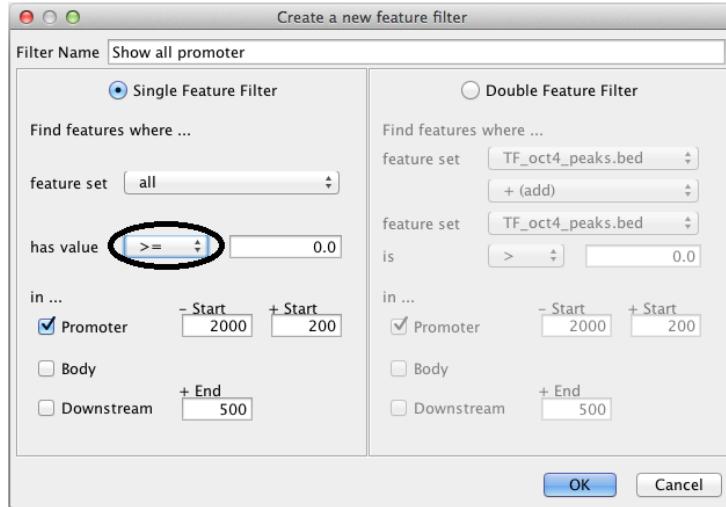
As we saw in the previous example, if a feature is not part of a filter its value is not reported. This is an optimization that saves on computation time. There are times when we want to see all values even if we are not filtering on them. PAPST has an **all** filter for exactly these situations.

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.

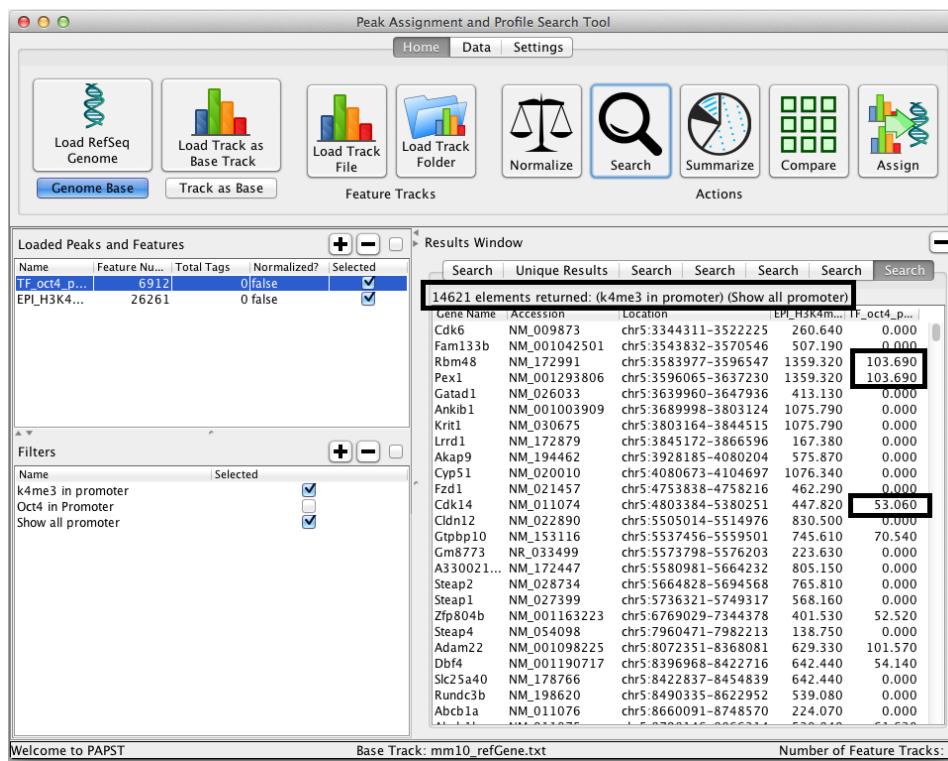
1. Click the selected checkbox to enable the **TF_oct4_peaks.bed** feature track.
2. Click the **plus '+'** button to create a new filter and select **all** from the feature set dropdown.
Keep the default promoter settings. Name the filter 'Show all promoter'.



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 to create a set of filters that do not return any results, so be careful when working with many feature files.

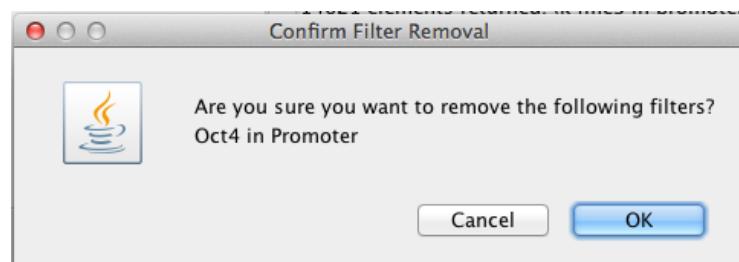
Removing Filters

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

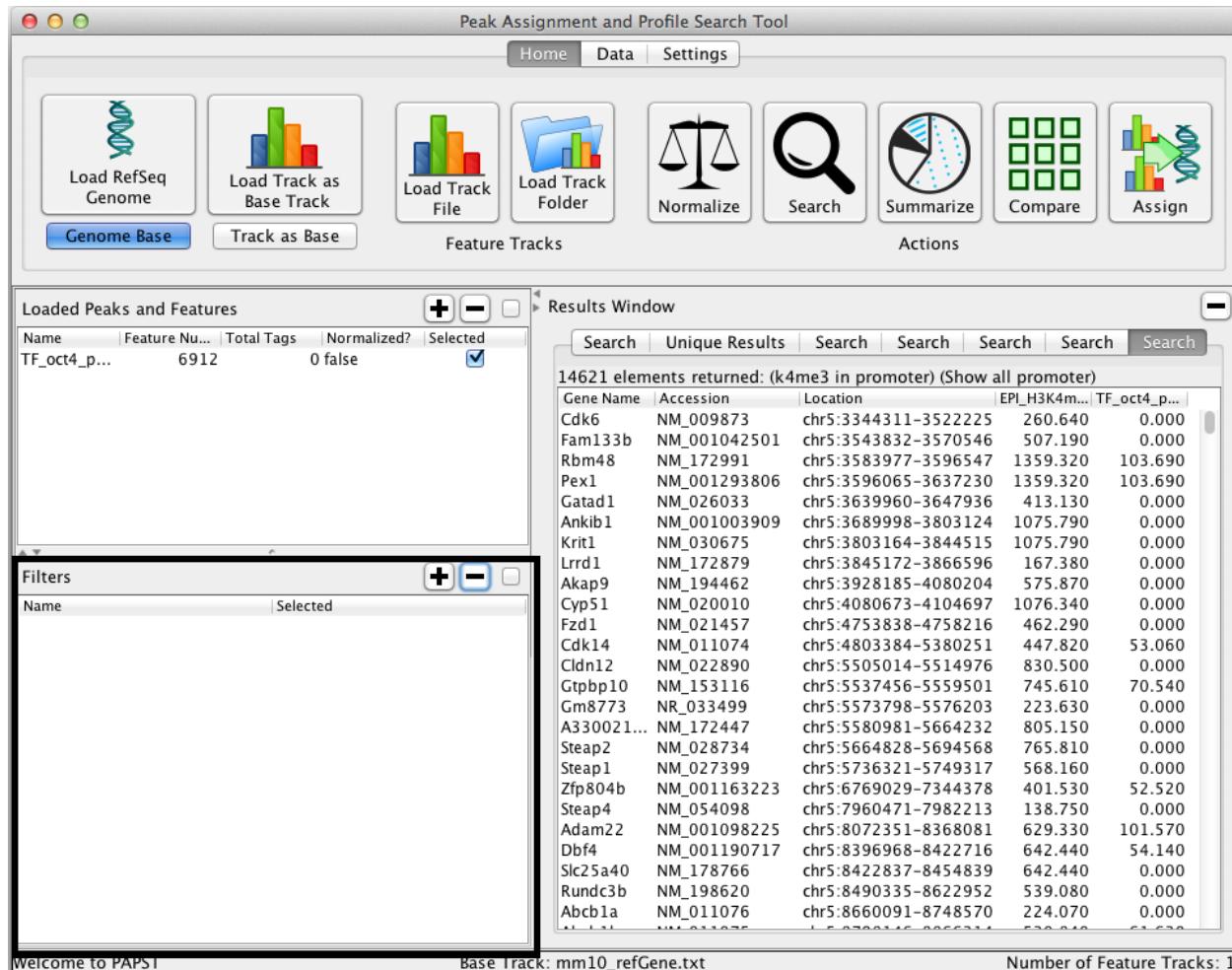
1. Click the filter you wish to remove to select it. Click on the **minus '-'** button in the upper right corner of the **Filters** pane.

The screenshot shows the PAPST application window. At the top, there's a toolbar with various icons: Load RefSeq Genome, Load Track as Base Track, Load Track File, Load Track Folder, Normalize, Search, Summarize, Compare, and Assign. Below the toolbar are three main sections: 'Genome Base' (with Load RefSeq Genome and Load Track as Base Track buttons), 'Feature Tracks' (with Load Track File and Load Track Folder buttons), and 'Actions' (with Normalize, Search, Summarize, Compare, and Assign buttons). The 'Results Window' displays a table of 14621 elements returned, filtered by 'k4me3 in promoter'. In the 'Filters' pane, the 'Oct4 in Promoter' filter is selected, indicated by a blue border and a checked checkbox. A large red arrow points to the minus '-' button in the 'Filters' pane, which is highlighted with a black circle. At the bottom of the window, there are status messages: 'Welcome to PAPST', 'Base Track: mm10_refGene.txt', and 'Number of Feature Tracks: 1'.

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



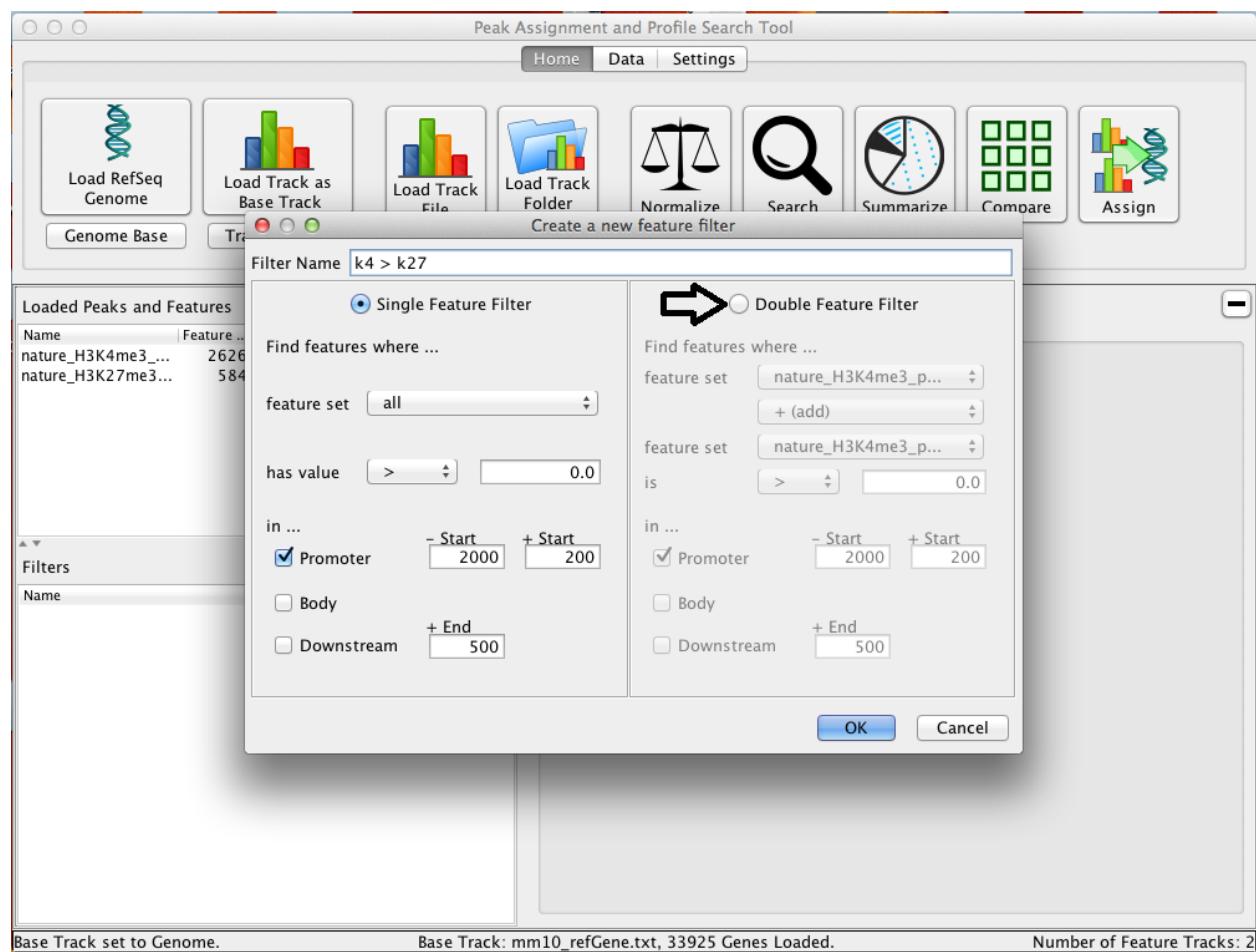
3. Click OK to confirm removing the filter. The filter is no longer available in the **Filters** pane.



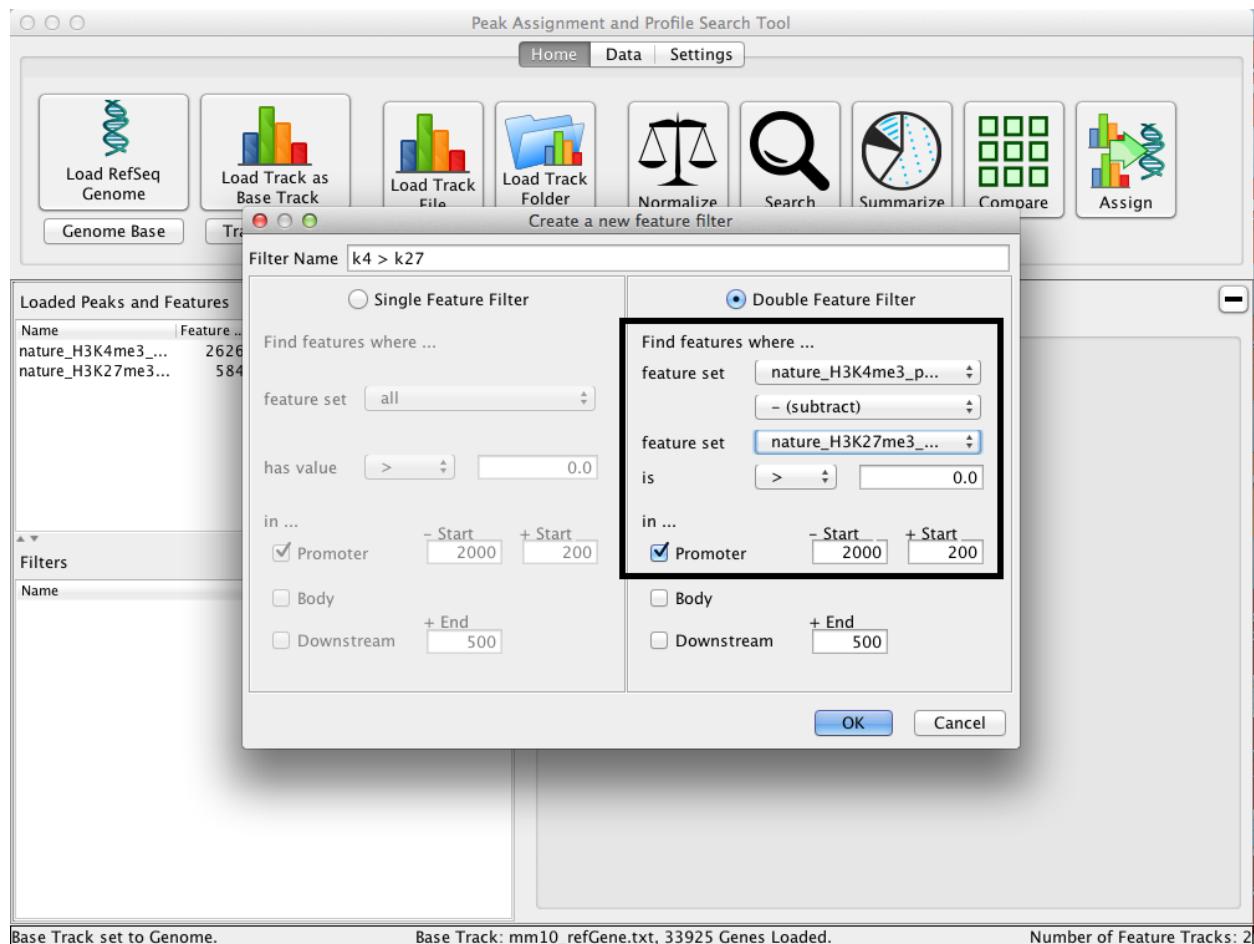
Advanced: Two Peak Filter

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.

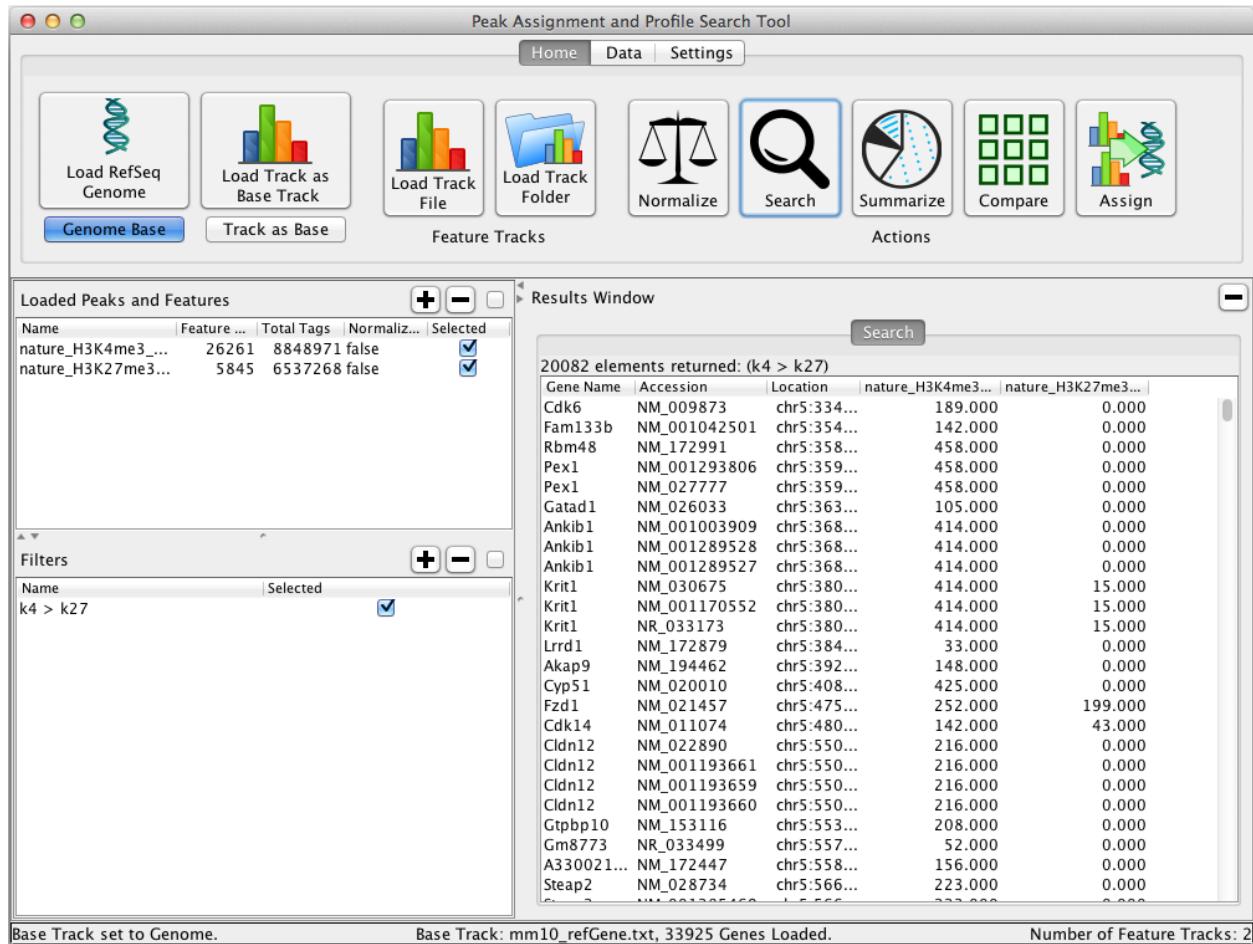
1. Load the mm10 RefGene genome into PAPST.
2. Load the files '**nature_H3K4me3_peaks.xls**' and '**nature_H3K27me3_peaks.xls**' under **peaks/macs_xls/**.
3. Create a new filter:
 - a. Click on the **Double Feature Filter** button to change modes.



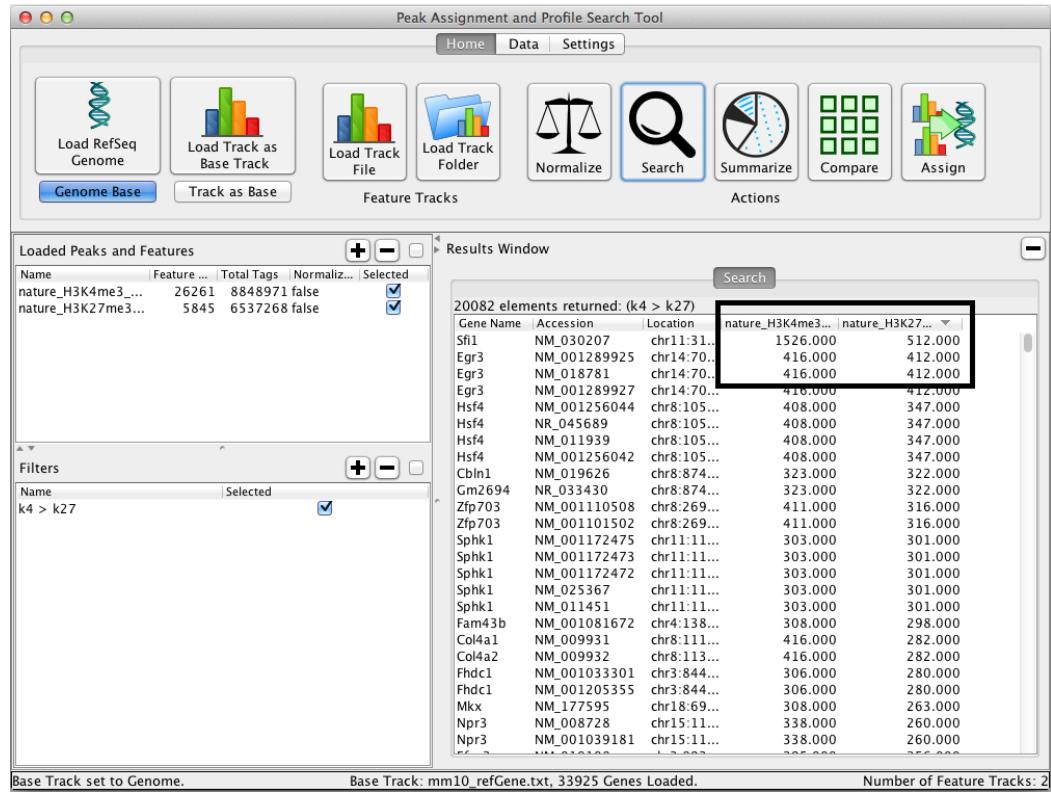
- b. Set the first set to **nature_H3K4me3_peaks.xls** and the operation to '**- (subtract)**'. Set the second set to **nature_H3K27me3_peaks.xls** and set the relationship to **>** and the threshold to **0**. Keep the default promoter region. Click OK.



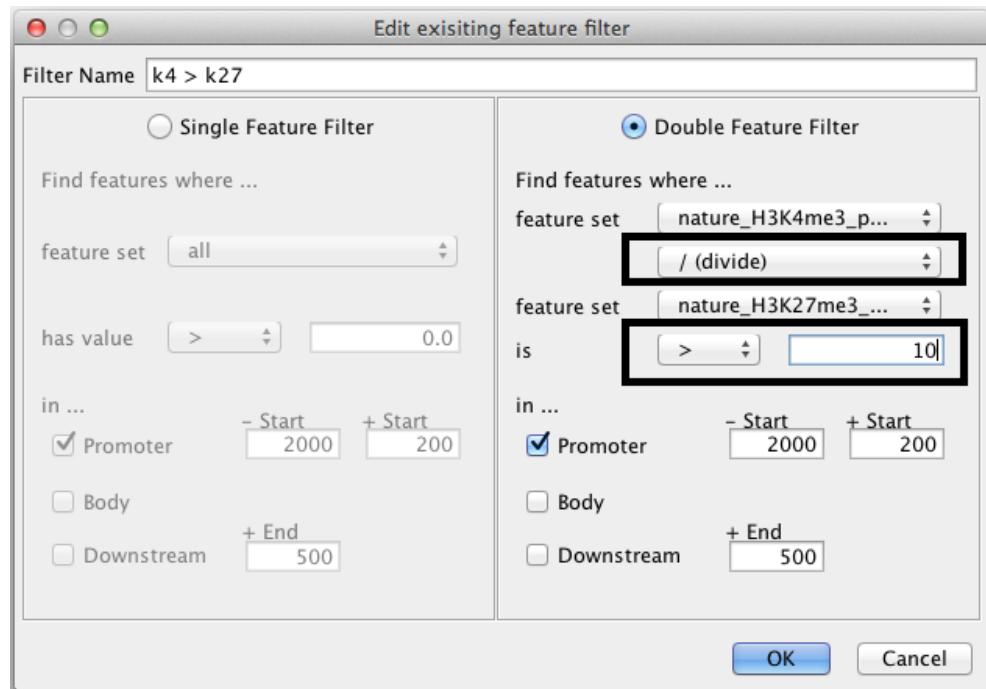
4. Click search to see the results. With this filter, PAPST takes the K4 value and subtracts the K27 value. Any regions where $K4 - K27 > 0$ will be returned. Genes which have more K27 are not returned.



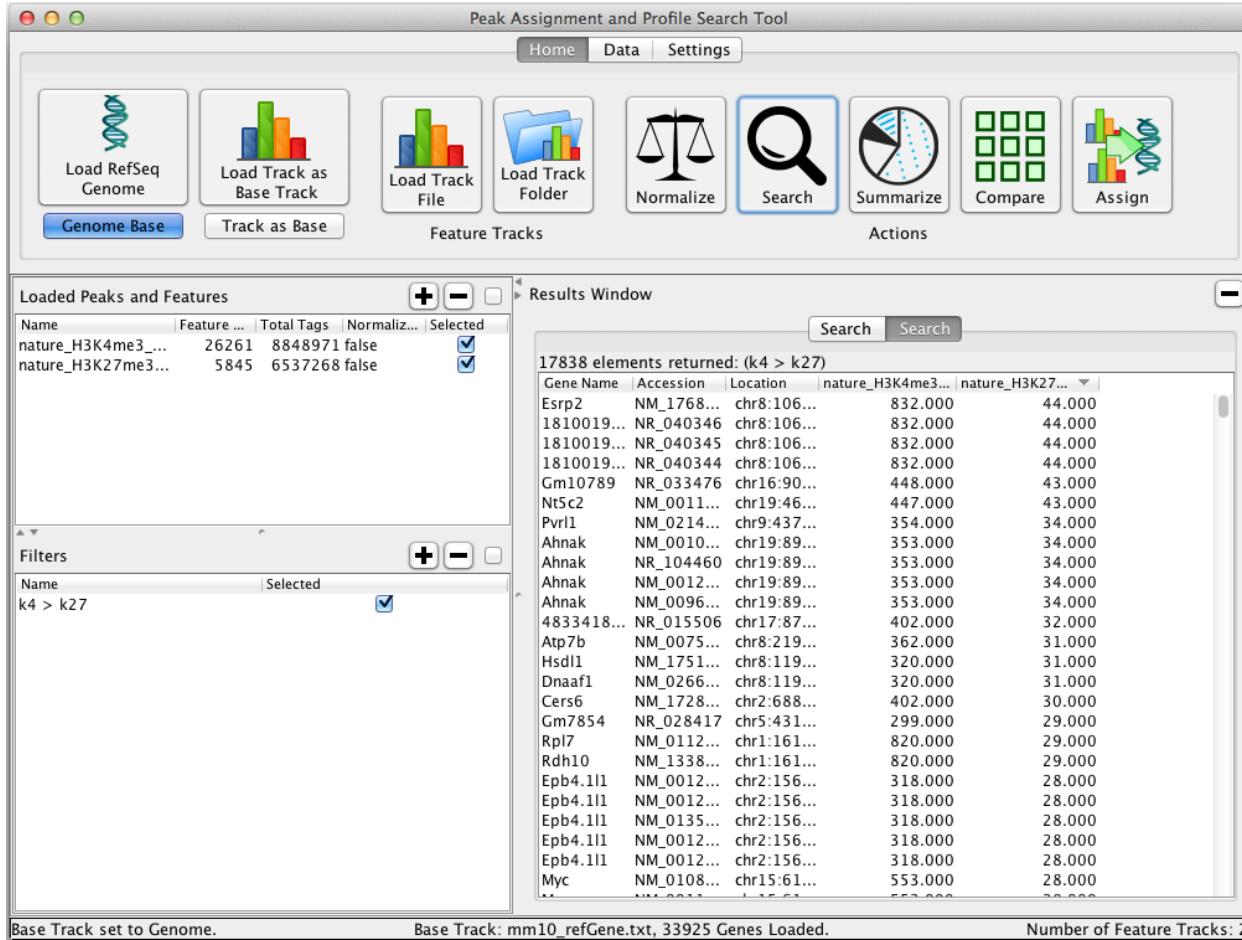
5. Sort the results by K27 to confirm that even the highest K27 region has a higher K4 region.



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

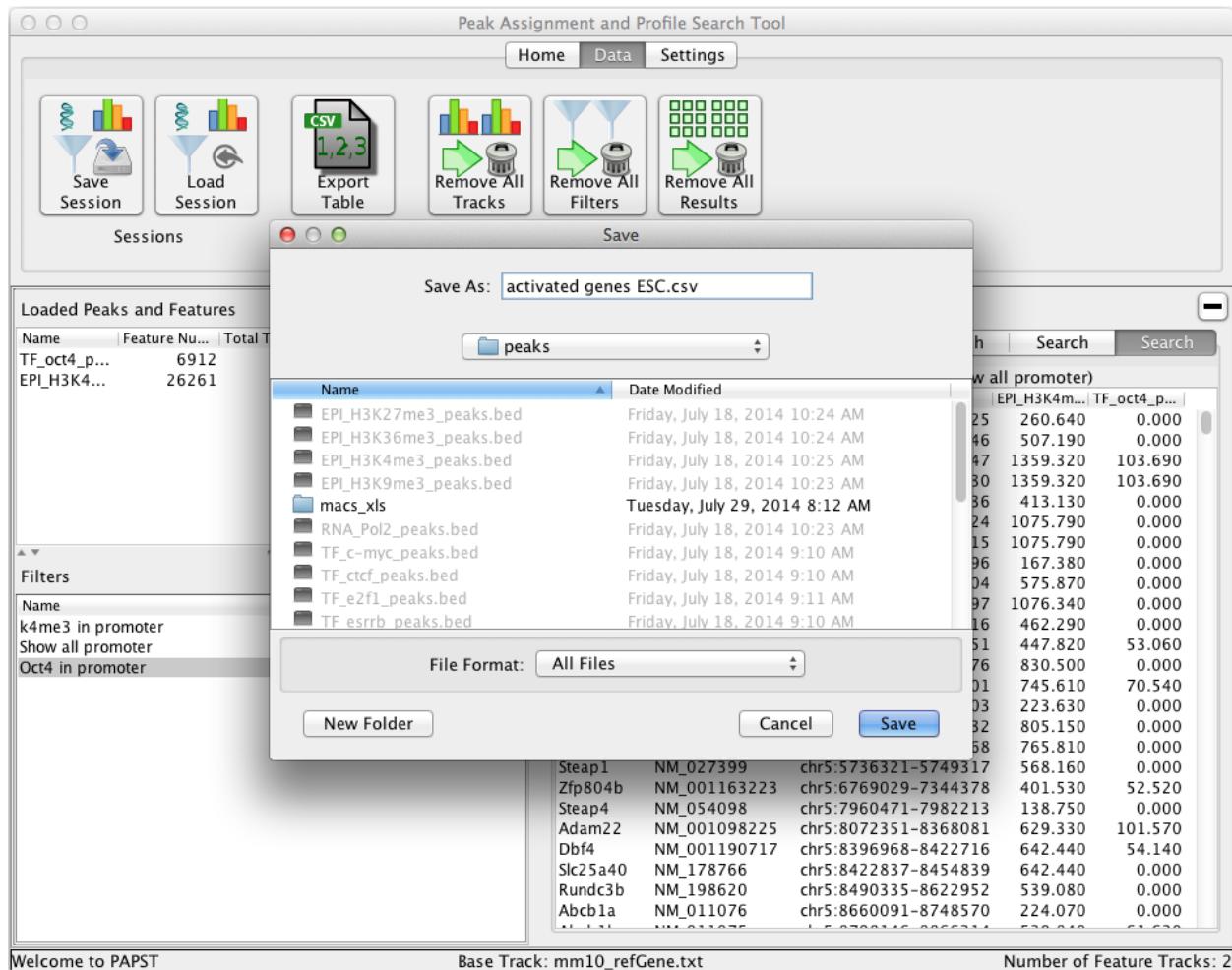
Exporting Results

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.

The screenshot shows the PAPST application window. At the top, there are three tabs: Home, Data (which is selected), and Settings. Below the tabs is a toolbar with several icons: Save Session, Load Session, Export Table (circled in red), Remove All Tracks, Remove All Filters, and Remove All Results. The main area is divided into two panes: 'Loaded Peaks and Features' on the left and 'Results Window' on the right. The 'Loaded Peaks and Features' pane contains a table with columns: Name, Feature Nu..., Total Tags, Normalized?, and Selected. Two rows are shown: TF_oct4_p... (selected) and EPI_H3K4... (selected). The 'Filters' section below the table lists 'k4me3 in promoter' and 'Show all promoter' as selected filters. The 'Results Window' pane displays a table titled '14621 elements returned: (k4me3 in promoter) (Show all promoter)'. The columns are Gene Name, Accession, Location, EPI_H3K4m..., and TF_oct4_p... . Numerous rows of data are listed, such as Cdk6, Fam133b, Rbm48, Pexl, Gatad1, Ankib1, Krt1, Lrrd1, Akap9, Cyp51, Fzd1, Cdk14, Cldn12, Gtpbp10, Gm8773, A330021..., Steap2, Steap1, Zfp804b, Steap4, Adam22, Dbf4, Slc25a40, Rndc3b, and Abcb1a. At the bottom of the window, there are status messages: 'Welcome to PAPST', 'Base Track: mm10_refGene.txt', and 'Number of Feature Tracks: 2'.

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.

The screenshot shows a Microsoft Excel spreadsheet titled "activated genes ESC.csv". The spreadsheet contains data from two BED files: "EPI_H3K4me3_peaks.bed" and "TF_oct4_peaks.bed". The columns represent gene names, accession numbers, chromosomal locations, and peak counts from the two files. Row 9 highlights the data for Lrrd1, with a value of 167.38 in cell D9.

	A	B	C	D	E	F	G	H	I
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	Gata1	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	Stear2	NM_028734	chr5:5664828-5694568		765.81	0			
19	Stear1	NM_027399	chr5:5736321-5749317		568.16	0			
20	Zfp804b	NM_001163223	chr5:6769029-7344378		401.53	52.52			
21	Stear4	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			
31	Dmtf1	NM_001110327	chr5:9118867-9161776		693.85	0			
32	9330182L06Rik	NM_172706	chr5:9266192-9480717		408.5	0			
33	Grm3	NM_181850	chr5:9485235-9725352		263	0			

Deleting Result Tables

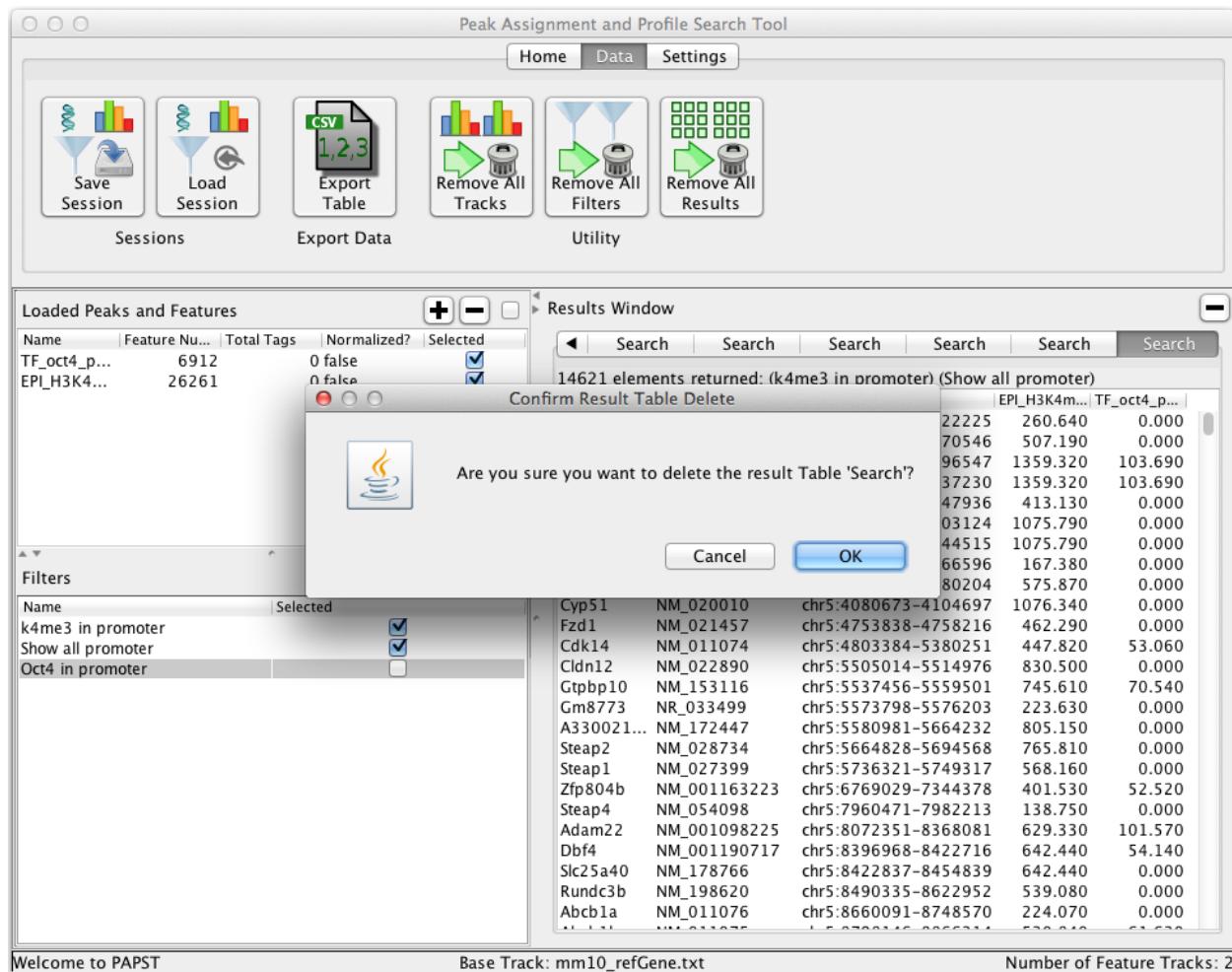
We have a number of tables in our results window. Our tabs are getting crowded and hard to see. We can delete these results without worrying because they can be easily generated with another search.

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.

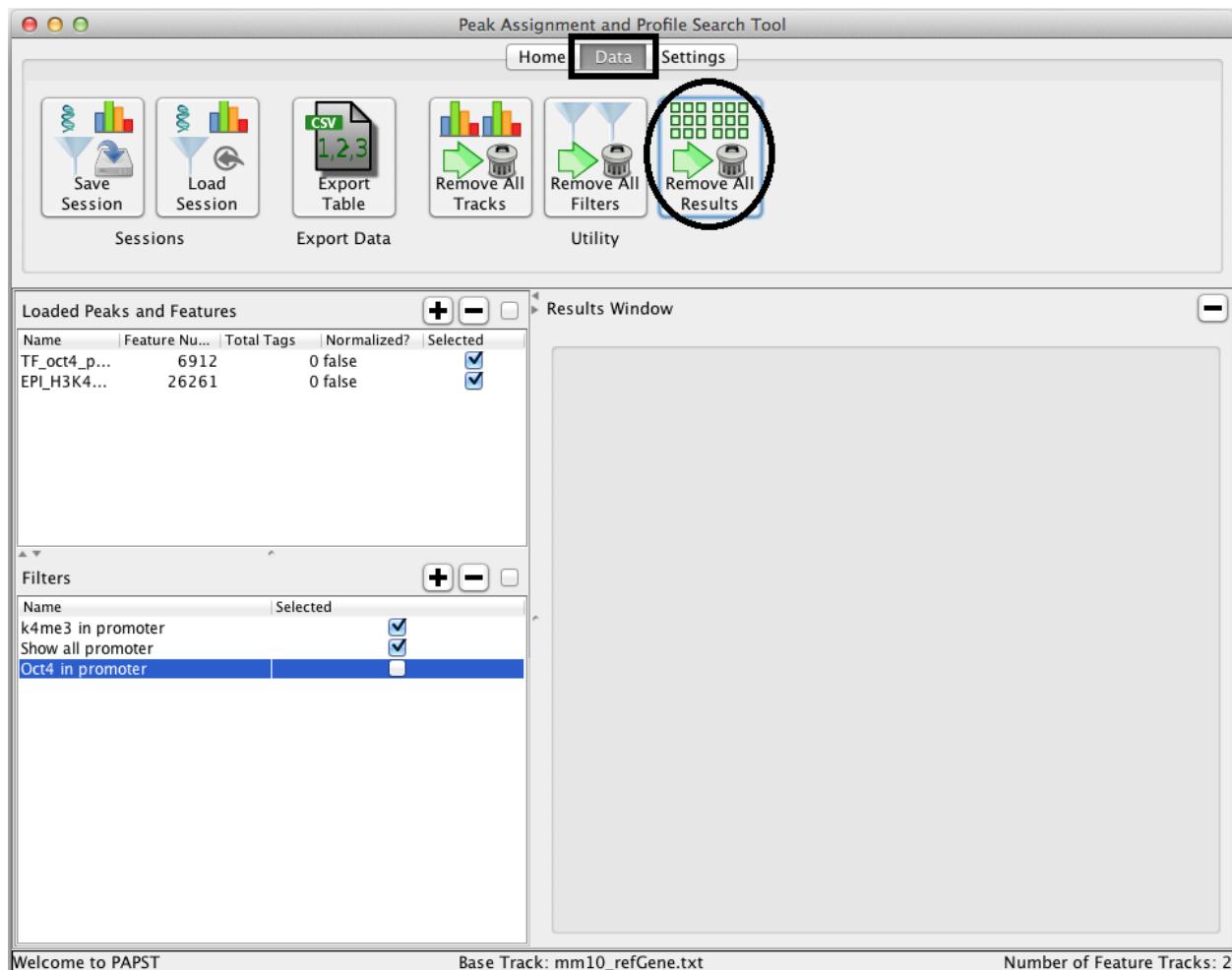
The screenshot shows the PAPST software interface. At the top, there's a toolbar with icons for Save Session, Load Session, Export Table (highlighted), Remove All Tracks, Remove All Filters, and Remove All Results. Below the toolbar are three main sections: Sessions, Export Data, and Utility. The Utility section contains the circled minus button. The main area is divided into two panes: 'Loaded Peaks and Features' and 'Results Window'. The 'Results Window' pane displays a table of 14621 elements returned, with a header row and many data rows. A large red arrow points from the circled minus button in the Utility section towards the 'Results Window' pane. The bottom of the screen shows a welcome message, the base track information (mm10_refGene.txt), and the number of feature tracks (2).

Gene Name	Accession	Location	EPI_H3K4m...	TF_oct4_p...
Cdk6	NM_009873	chr5:3344311-3522225	260.640	0.000
Fam133b	NM_001042501	chr5:3543832-3570546	507.190	0.000
Rbm48	NM_172991	chr5:3583977-3596547	1359.320	103.690
Pex1	NM_001293806	chr5:3596065-3637230	1359.320	103.690
Gata1	NM_026033	chr5:3639960-3647936	413.130	0.000
Ankib1	NM_001003909	chr5:3689998-3803124	1075.790	0.000
Krt1	NM_030675	chr5:3803164-3844515	1075.790	0.000
Lrrd1	NM_172879	chr5:3845172-3866596	167.380	0.000
Akap9	NM_194462	chr5:3928185-4080204	575.870	0.000
Cyp51	NM_020010	chr5:4080673-4104697	1076.340	0.000
Fzd1	NM_021457	chr5:4753838-4758216	462.290	0.000
Cdk14	NM_011074	chr5:4803384-5380251	447.820	53.060
Cldn12	NM_022890	chr5:5505014-5514976	830.500	0.000
Gtpbp10	NM_153116	chr5:5537456-5559501	745.610	70.540
Gm8773	NR_033499	chr5:5573798-5576203	223.630	0.000
A330021...	NM_172447	chr5:5580981-5664232	805.150	0.000
Steap2	NM_028734	chr5:5664828-5694568	765.810	0.000
Steap1	NM_027399	chr5:5736321-5749317	568.160	0.000
Zfp804b	NM_001163223	chr5:6769029-7344378	401.530	52.520
Steap4	NM_054098	chr5:7960471-7982213	138.750	0.000
Adam22	NM_001098225	chr5:8072351-8368081	629.330	101.570
Dbf4	NM_001190717	chr5:8396968-8422716	642.440	54.140
Slc25a40	NM_178766	chr5:8422837-8454839	642.440	0.000
Rundc3b	NM_198620	chr5:8490335-8622952	539.080	0.000
Abcb1a	NM_011076	chr5:8660091-8748570	224.070	0.000

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**.



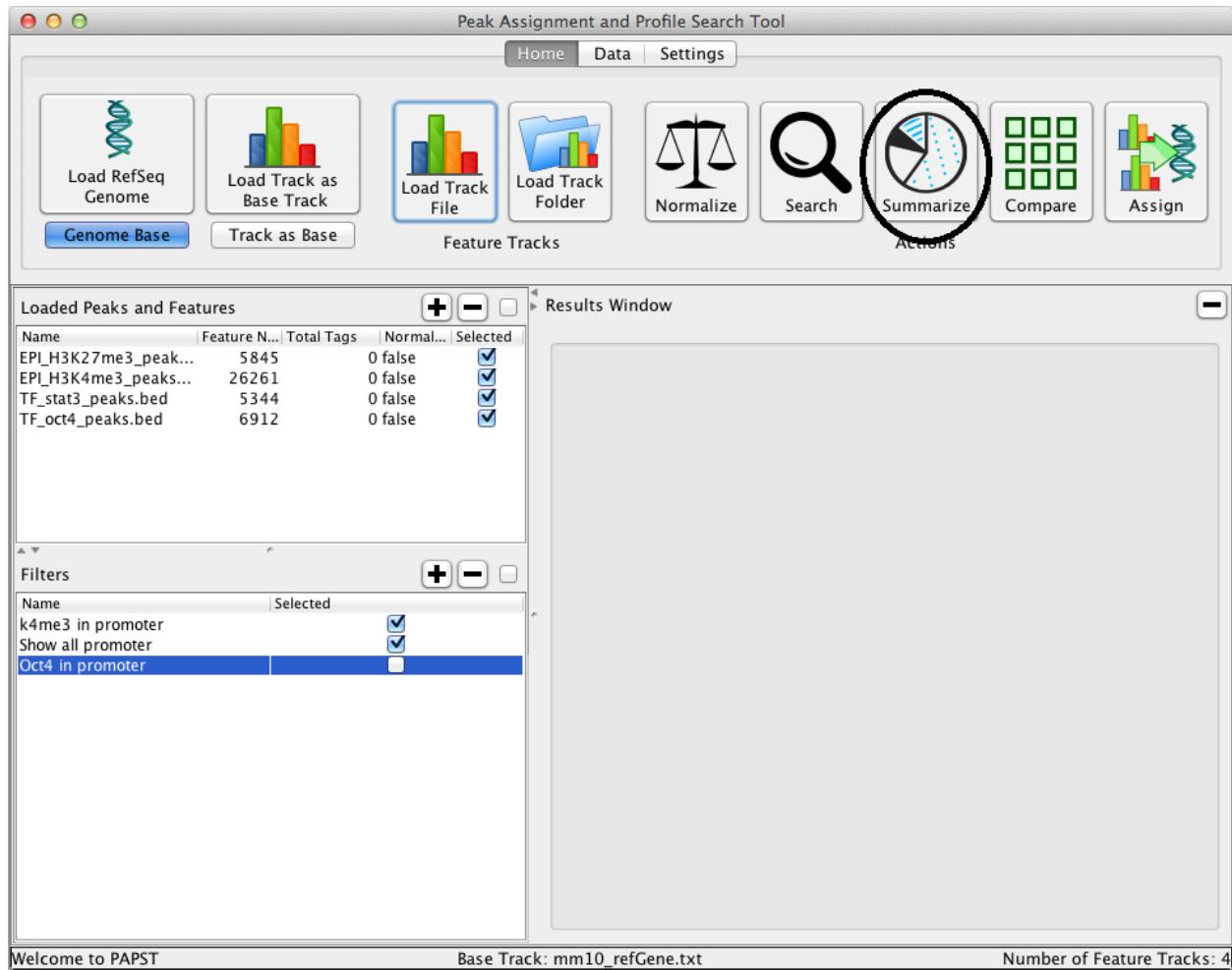
All results are now removed.

General Feature Analysis in PAPST

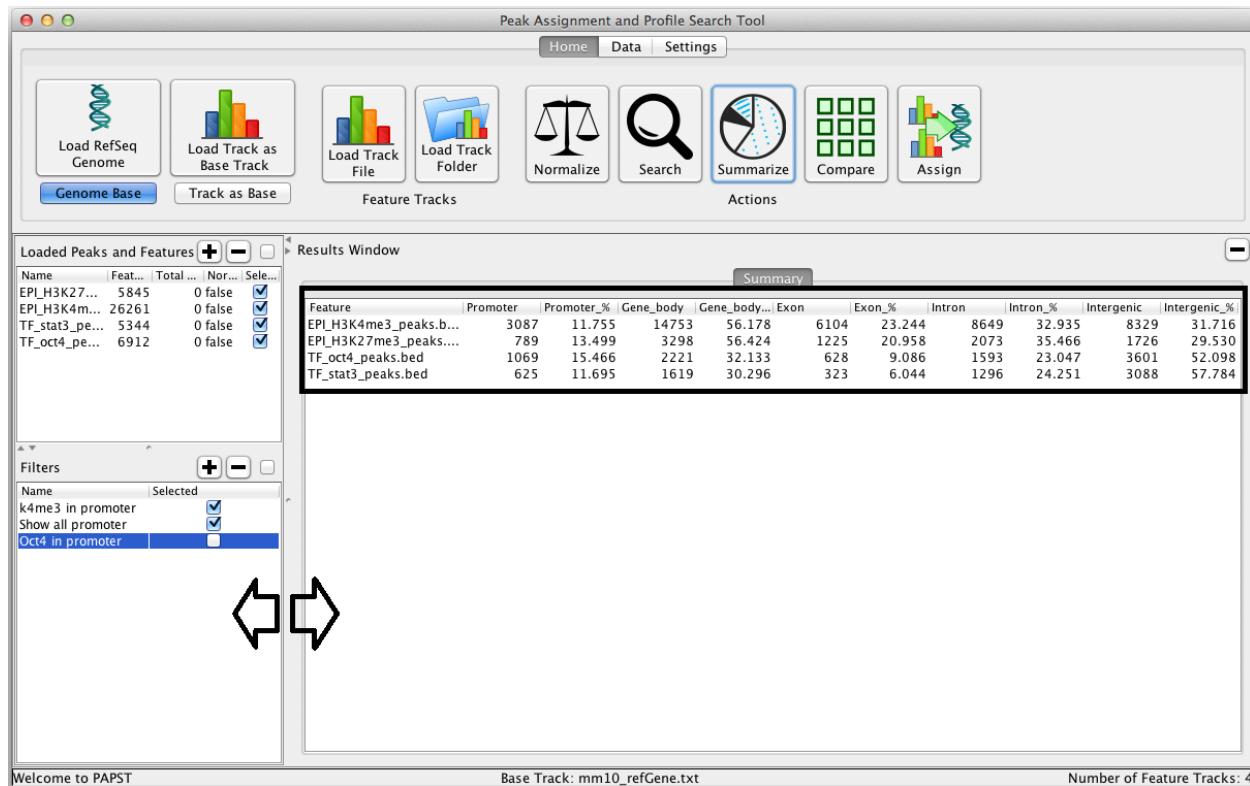
Peak Distributions

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 with 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.

Peak Assignment and Profile Search Tool

Settings (selected)

Default Genome Parser: refseq_subset

Show Gene Isoforms?:

Default Promoter: 10000 Upstream of TSS, 500 Downstream of TSS

Normalization Factor: 1000000

Parser Settings

Search Settings

Results Window (Summary)

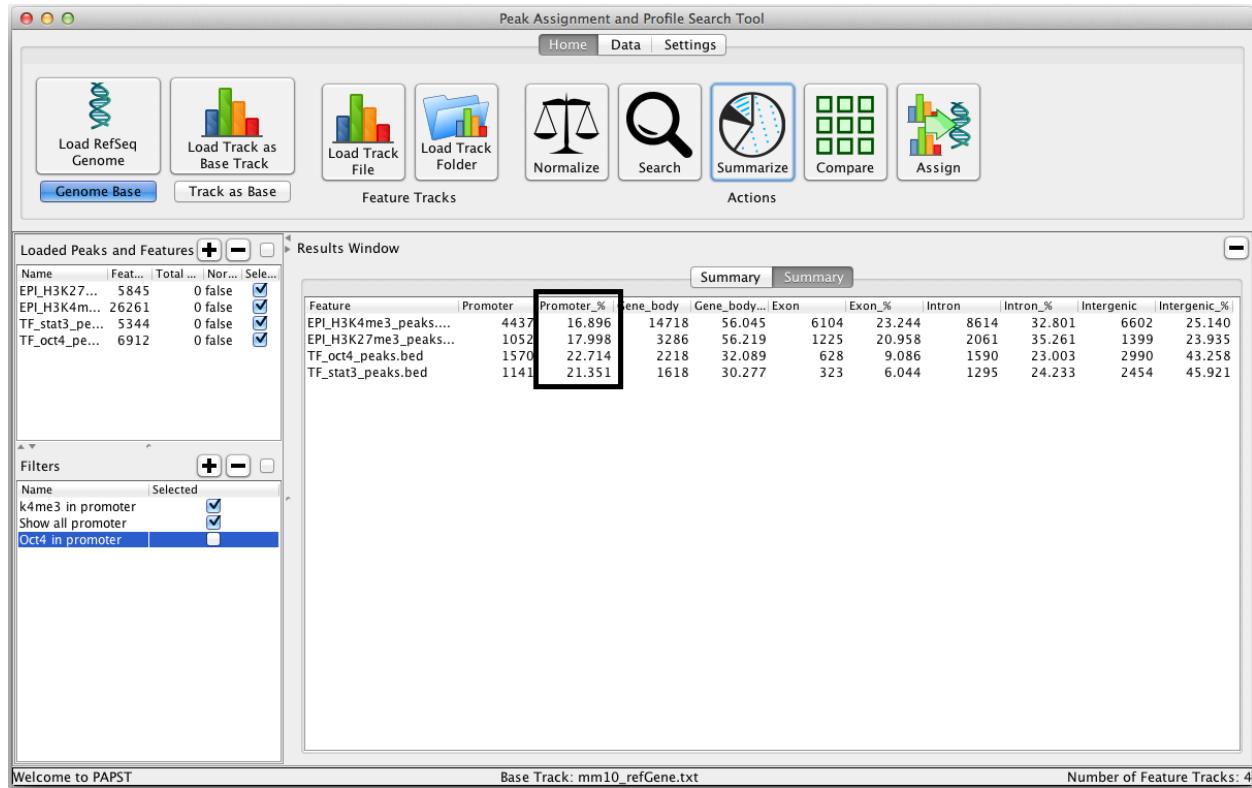
Feature	Promoter	Promoter_%	Gene_body	Gene_body..._Exon	Exon_%	Intron	Intron_%	Intergenic	Intergenic_%	
EPI_H3K27me3_peaks.b...	3087	11.755	14753	56.178	6104	23.244	8649	32.935	8329	31.716
EPI_H3K27me3_peaks....	789	13.499	3298	56.424	1225	20.958	2073	35.466	1726	29.530
TF_oct4_peaks.bed	1069	15.466	2221	32.133	628	9.086	1593	23.047	3601	52.098
TF_stat3_peaks.bed	625	11.695	1619	30.296	323	6.044	1296	24.251	3088	57.784

Filters

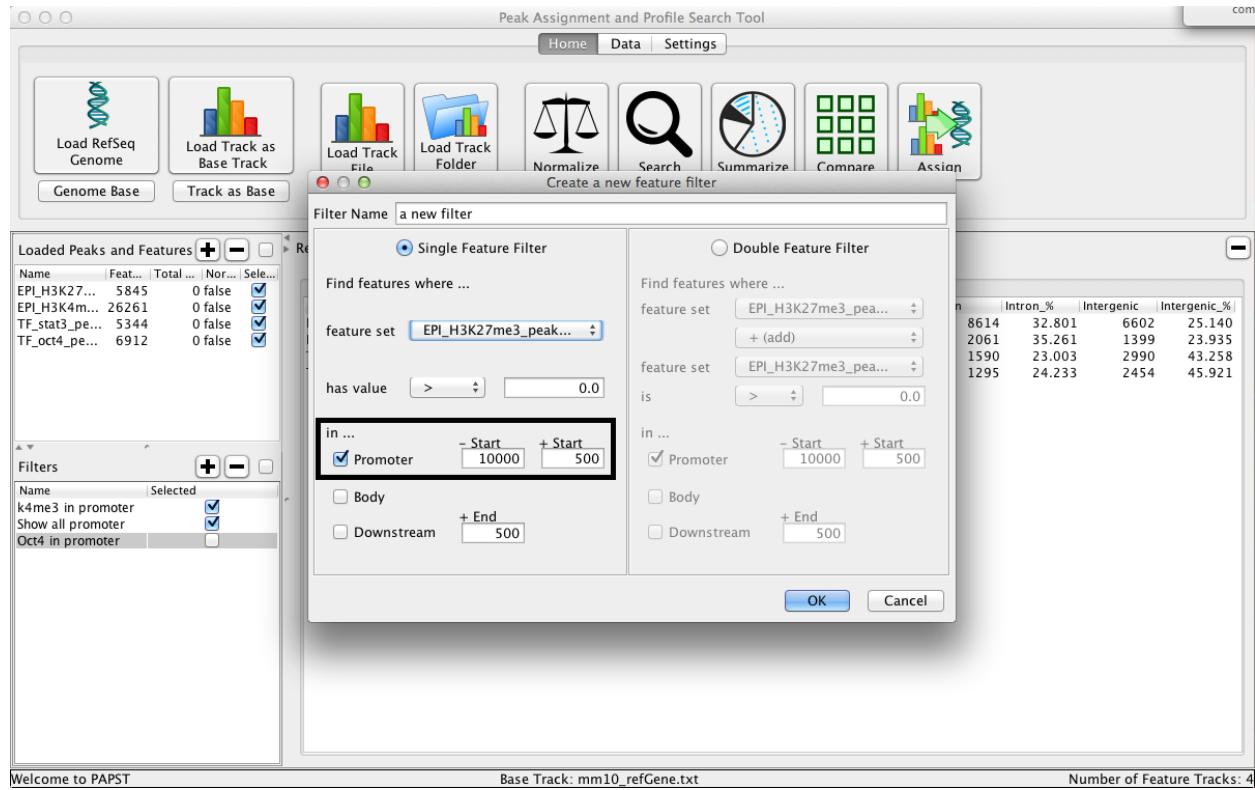
- Name: Selected
- k4me3 in promoter:
- Show all promoter:
- Oct4 in promoter:

Welcome to PAPST Base Track: mm10_refGene.txt Number of Feature Tracks: 4

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.

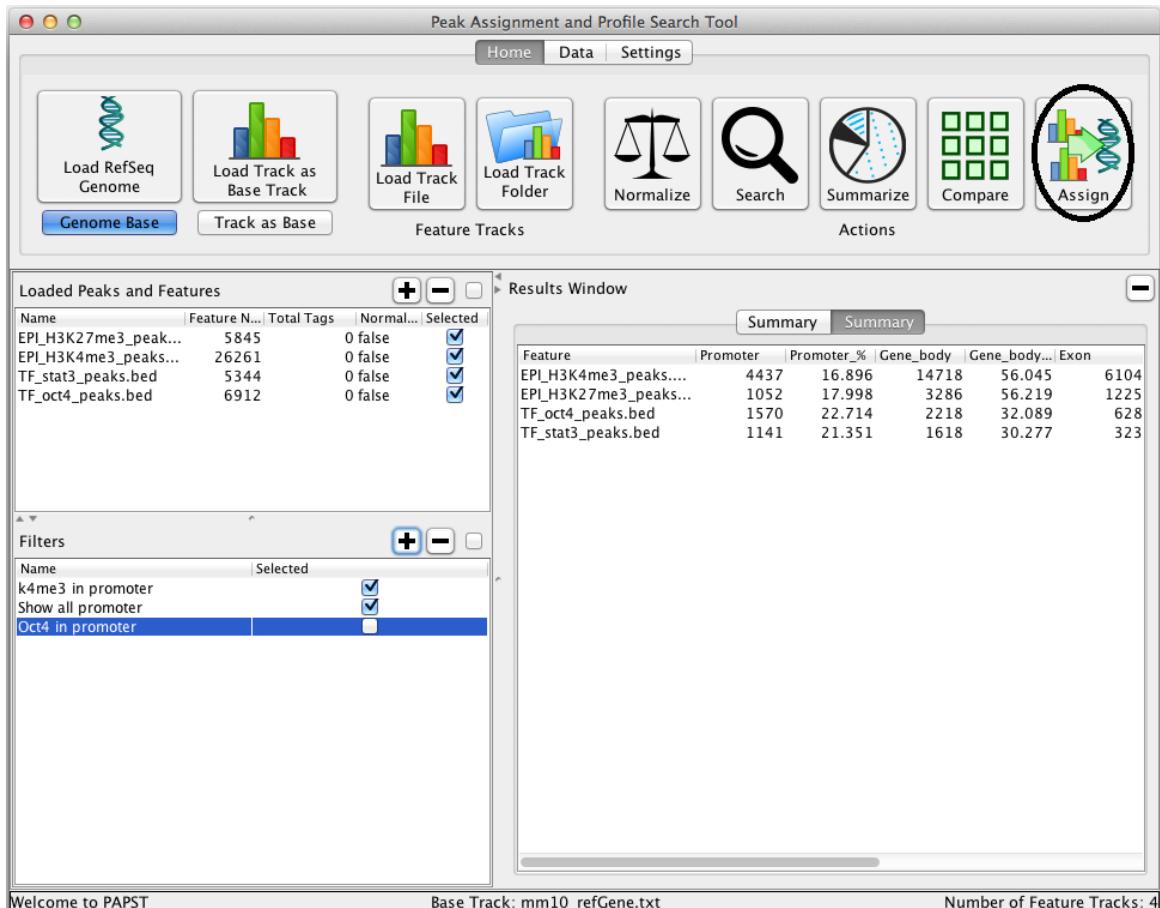


Now a new filter will use the newly defined promoter region.

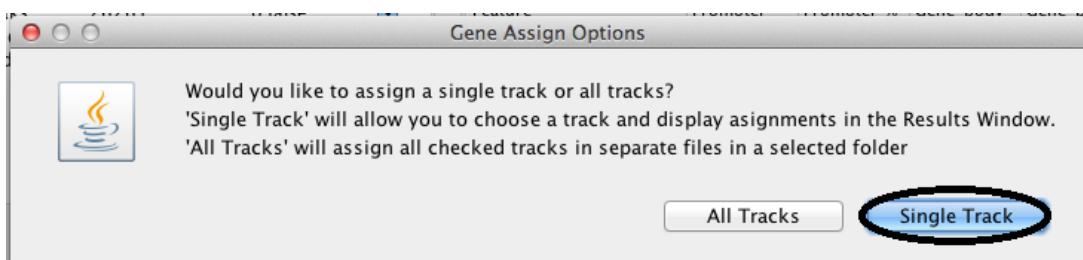
Assign Peaks to Their Nearest Genes

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

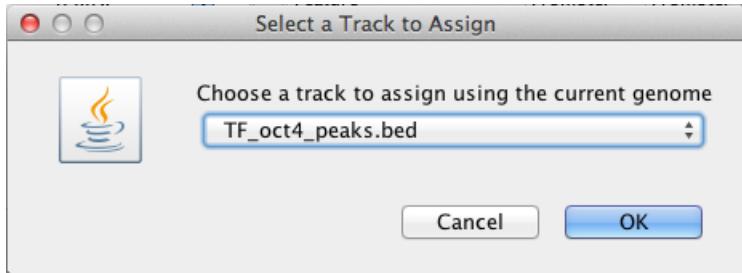
1. To assign a **single feature set** to genes, click on the **Assign** button on the **Home** tab.



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



3. Next select TF_oct4_peaks.bed. Click OK.

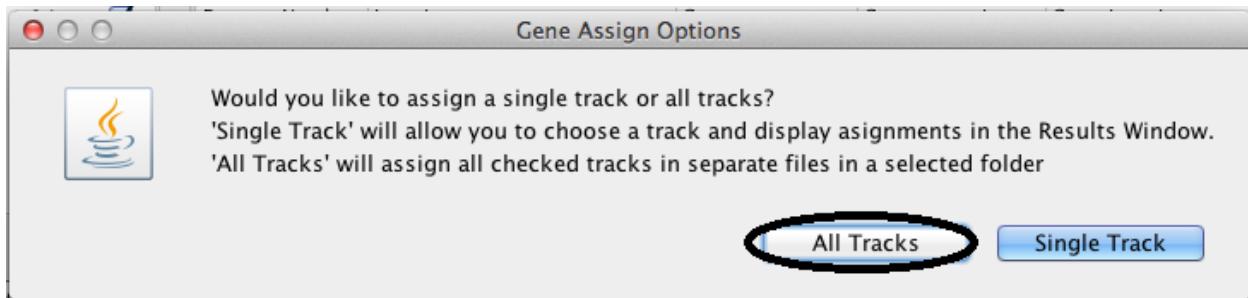


4. The results will be displayed in the **Results Window**. Each peak is given a feature number.

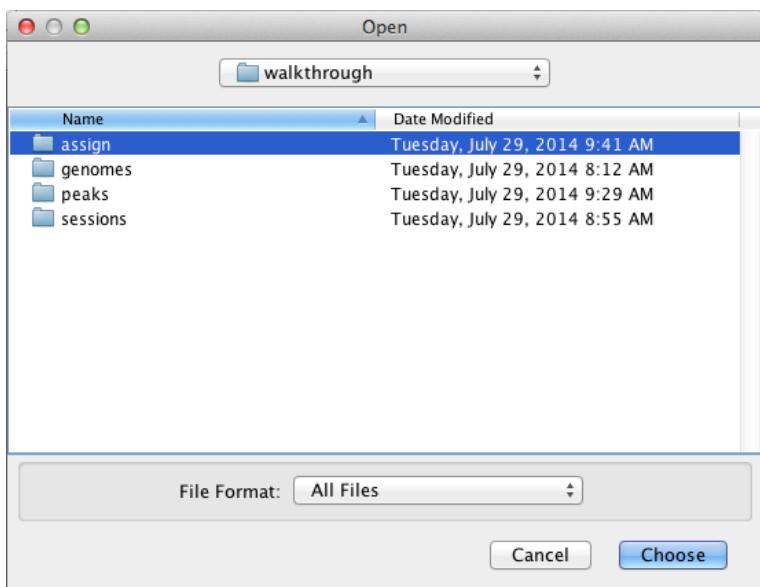
Assign reports peak location, gene location, and distance to TSS as well as Refseq identifier and gene name.

Feature_Number	Location	Gene_name	Gene_accession	Gene_Location	distance_to_TSS
0	chr10:3873582-3873883	Plekgh1	NM_001033253	chr10:3872666-3967302	1066
1	chr10:3891681-3892309	Plekgh1	NM_001033253	chr10:3872666-3967302	19329
2	chr10:3949248-3949627	Mthfd1l	NM_001170785	chr10:3973074-4167081	23637
3	chr10:4242233-4242645	Akap12	NM_031185	chr10:4266328-4359471	23889
4	chr10:4243564-4243943	Akap12	NM_031185	chr10:4266328-4359471	22575
5	chr10:4254369-4254819	Akap12	NM_031185	chr10:4266328-4359471	11734
6	chr10:4333204-4333753	Zbtb2	NM_001033466	chr10:4367073-4388108	54630
7	chr10:4375156-4375663	Zbtb2	NM_001033466	chr10:4367073-4388108	12699
8	chr10:4492965-4493341	Cddc170	NM_001195672	chr10:4509871-4561111	16718
9	chr10:6624023-6624170	Oprm1	NM_001039652	chr10:6788600-7038209	164504
10	chr10:6788645-6788969	Oprm1	NM_001039652	chr10:6788600-7038209	207
11	chr10:7589737-7590120	Lrp11	NM_172784	chr10:7589799-7625477	129
12	chr10:7662693-7663387	A630066F11Rik	NR_030698	chr10:7663370-7664623	330
13	chr10:8055155-8055579	Tab2	NM_138667	chr10:7905647-7956123	99244
14	chr10:8235161-8235602	Tab2	NM_138667	chr10:7905647-7956123	279258
15	chr10:8717202-8717511	Sash1	NM_175155	chr10:8722218-8886070	168714
16	chr10:8889245-8889838	Sash1	NM_175155	chr10:8722218-8886070	3471
17	chr10:9085524-9085863	Sash1	NM_175155	chr10:8722218-8886070	199623
18	chr10:9801654-9801946	Stxbp5	NM_001081344	chr10:9755546-9901040	99240
19	chr10:9803069-9803467	Stxbp5	NM_001081344	chr10:9755546-9901040	97772
20	chr10:11149348-11149712	Shprh	NM_001077707	chr10:11149429-11215273	101
21	chr10:11263529-11263856	Fbxo30	NM_027968	chr10:11281329-11297969	17637
22	chr10:11280765-11281131	Fbxo30	NM_027968	chr10:11281329-11297969	381
23	chr10:11491917-11492282	Epm2a	NM_010146	chr10:11343444-11457477	148655
24	chr10:12572691-12573117	Utrn	NM_011682	chr10:12382187-12861735	288831
25	chr10:12608630-12609080	Utrn	NM_011682	chr10:12382187-12861735	252880
26	chr10:13614282-13614678	Utrn	NM_011682	chr10:13614282-13614678	247080

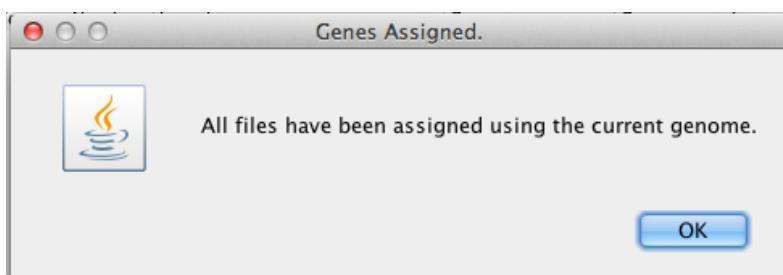
5. To assign all loaded peaks at once, click the **Assign** button, and then select **All Tracks**.



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.

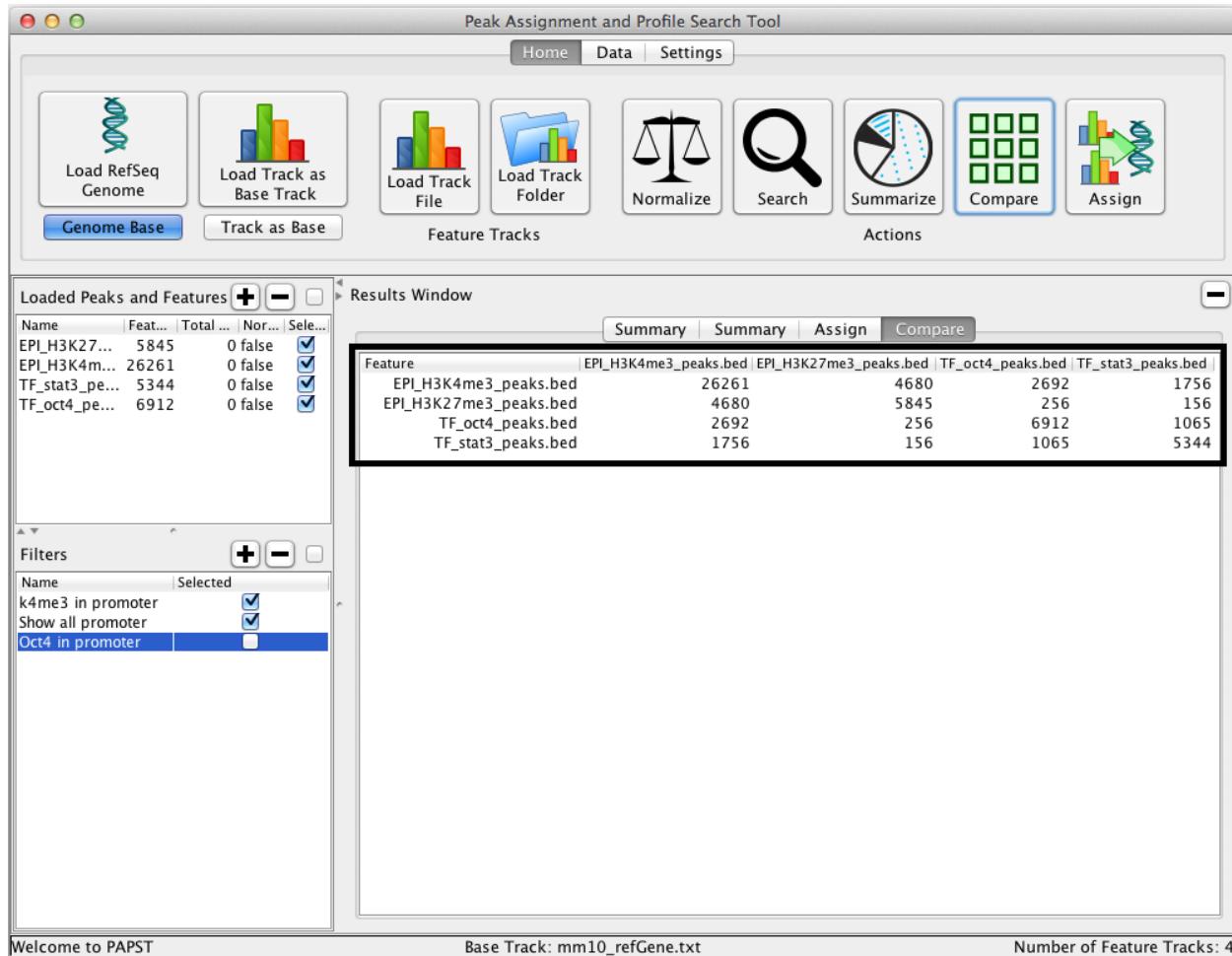
Comparing Peaks

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.

Feature_Number	Location	Gene_name	Gene_accession	Gene_Location
0	chr10:3873582-3873883	Plekhg1	NM_001033253	chr10:38:
1	chr10:3891681-3892309	Plekhg1	NM_001033253	chr10:38:
2	chr10:3949248-3949627	Mthfd1l	NM_001170785	chr10:39:
3	chr10:4242233-4242645	Akap12	NM_031185	chr10:42:
4	chr10:4243564-4243943	Akap12	NM_031185	chr10:42:
5	chr10:4254369-4254819	Akap12	NM_031185	chr10:42:
6	chr10:4333204-4333753	Zbtb2	NM_001033466	chr10:43:
7	chr10:4375156-4375663	Zbtb2	NM_001033466	chr10:43:
8	chr10:4492965-4493341	Ccdc170	NM_001195672	chr10:45:
9	chr10:6624023-6624170	Oprm1	NM_001039652	chr10:67:
10	chr10:6788645-6788969	Oprm1	NM_001039652	chr10:67:
11	chr10:7589737-7590120	Lrp11	NM_172784	chr10:75:
12	chr10:7662693-7663387	A630066F11Rik	NR_030698	chr10:76:
13	chr10:8055155-8055579	Tab2	NM_138667	chr10:79:
14	chr10:8235161-8235602	Tab2	NM_138667	chr10:79:
15	chr10:8717202-8717511	Sash1	NM_175155	chr10:87:
16	chr10:8889245-8889838	Sash1	NM_175155	chr10:87:
17	chr10:9085524-9085863	Sash1	NM_175155	chr10:87:
18	chr10:9801654-9801946	Stxbp5	NM_001081344	chr10:97:
19	chr10:9803069-9803467	Stxbp5	NM_001081344	chr10:97:
20	chr10:11149348-11149712	Shprh	NM_001077707	chr10:11:
21	chr10:11263529-11263850	Fbxo30	NM_027968	chr10:11:
22	chr10:11280765-11281131	Fbxo30	NM_027968	chr10:11:
23	chr10:11491917-11492282	Epm2a	NM_010146	chr10:11:
24	chr10:12572691-12573117	Utrn	NM_011682	chr10:12:
25	chr10:12608620-12608680	Utrn	NM_011682	chr10:12:

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:

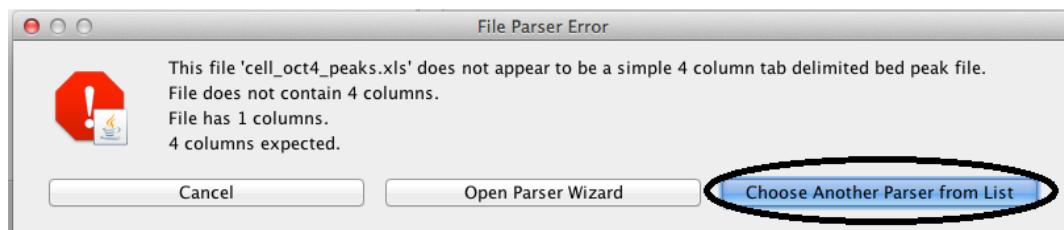
$$\text{Normalized Tag Score}_p = \frac{\# \text{ of tags for peak } p}{\text{Total # of tags in all peaks}} * 1,000,000$$

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.

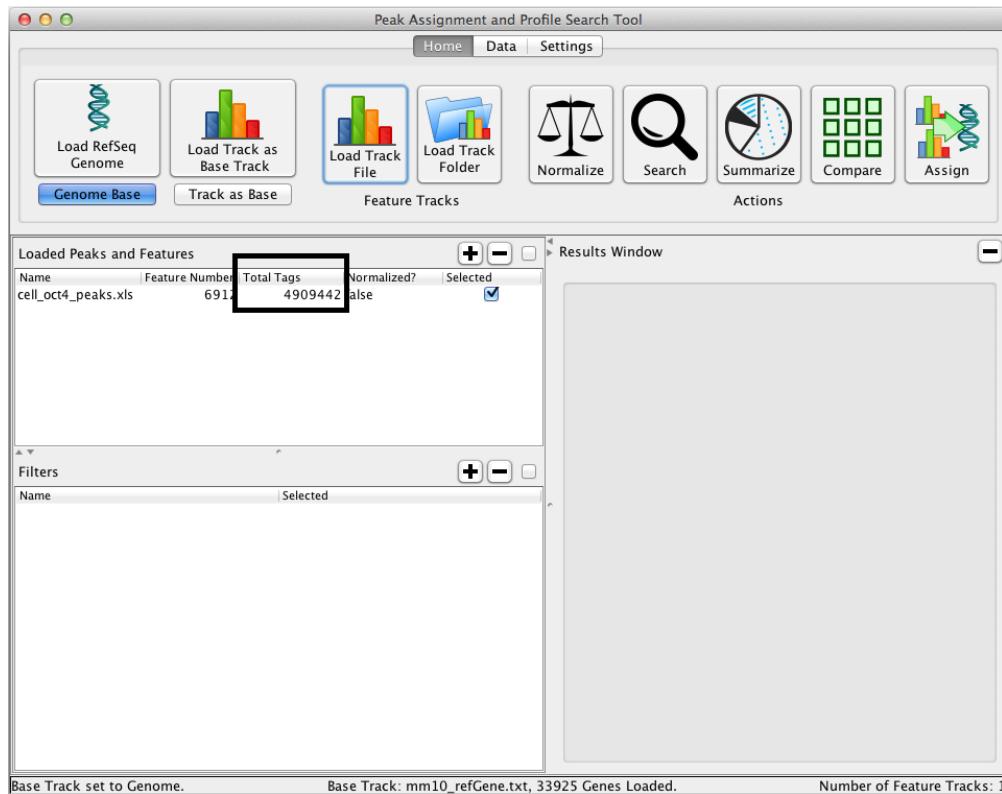
1. Load the **mm10_refGene.txt** into papst.
2. Load the peak file **cell_oct4_peaks.xls** under peaks/macs.xls/ in the walkthrough folder. The MACS 1.4 parser is not enabled by default. Click **Choose Another Parser from List**.



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.

Filter Name: Oct4 in promoter

Single Feature Filter

Find features where ...

feature set: cell_oct4_peaks.xls

has value: > 0.0

in ...

Promoter - Start: 2000 + Start: 200

Body + End: 500

Downstream + End: 500

Double Feature Filter

Find features where ...

feature set: cell_oct4_peaks.xls

+ (add)

feature set: cell_oct4_peaks.xls

is: > 0.0

in ...

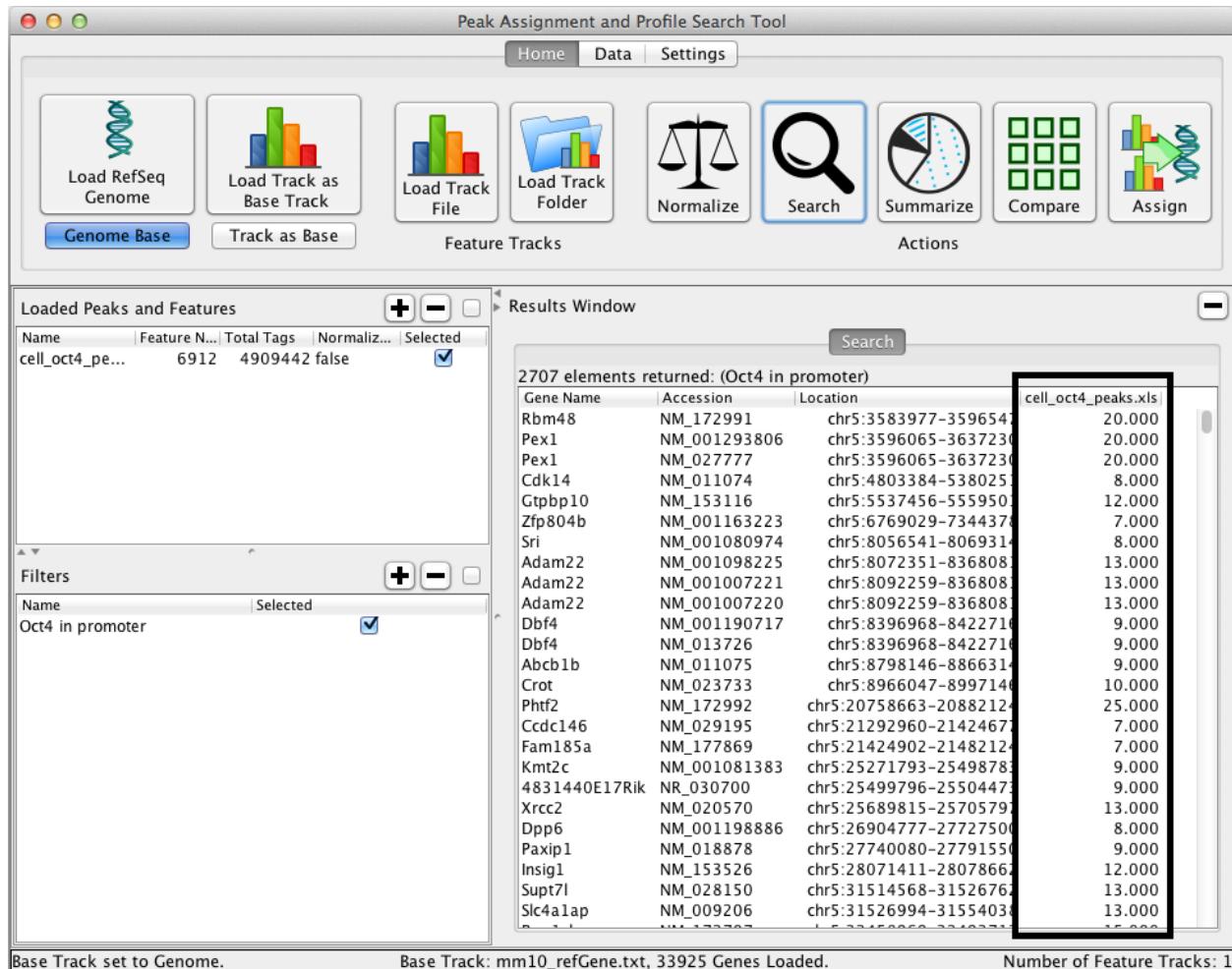
Promoter - Start: 2000 + Start: 200

Body + End: 500

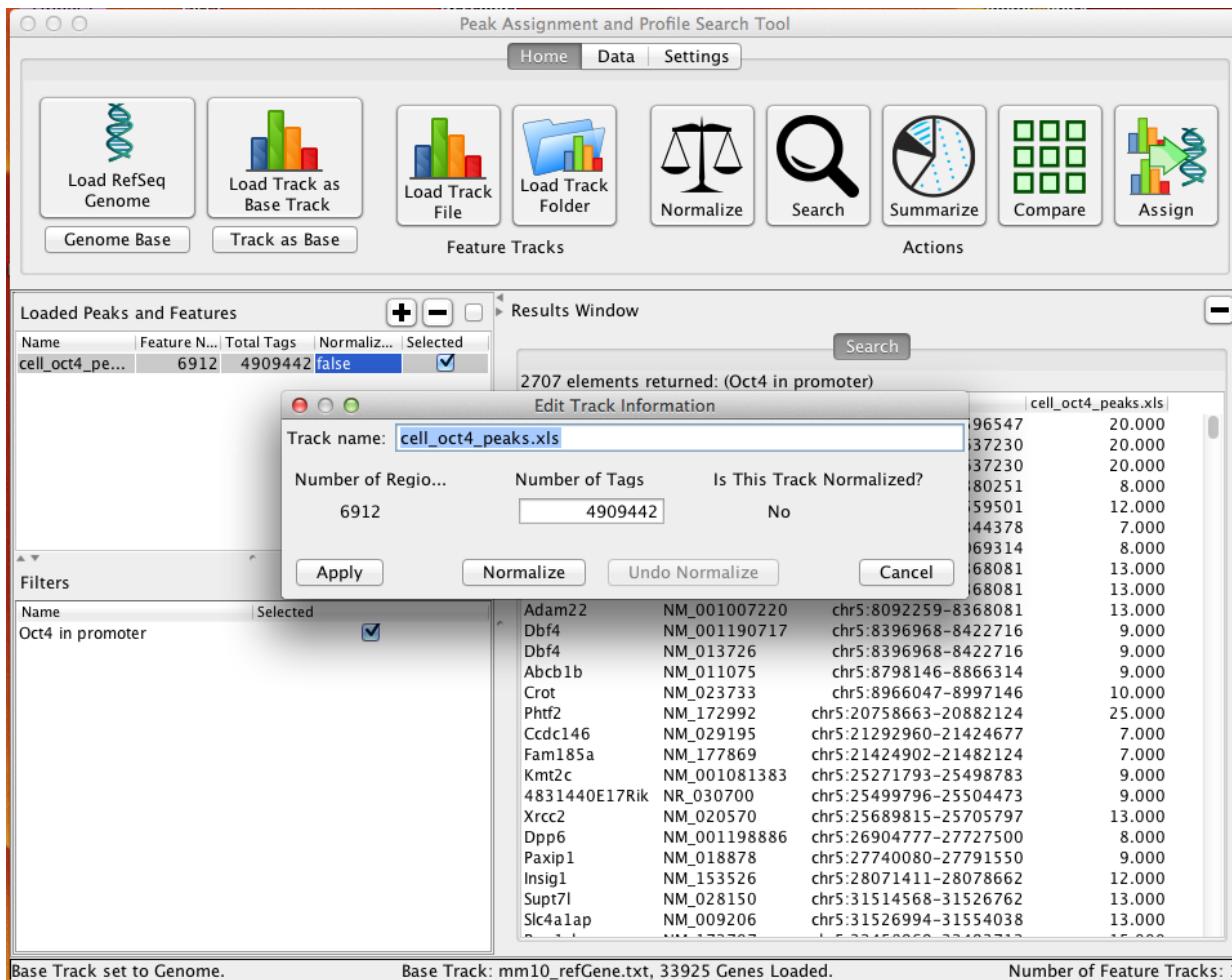
Downstream + End: 500

OK Cancel

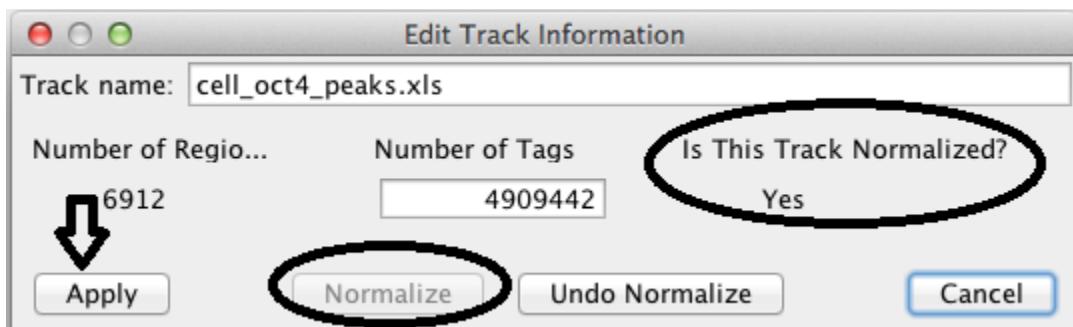
6. Click the **Search** button to show genes with Oct4 binding their promoters. Notice the range of values.



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.

Results Window			
2707 elements returned: (Oct4 in promoter)			
Gene Name	Accession	Location	Normalized
Rbm48	NM_172991	chr5:3583977-3596547	20.000
Pexl	NM_001293806	chr5:3596065-3637230	20.000
Pexl	NM_027777	chr5:3596065-3637230	20.000
Cdk14	NM_011074	chr5:4803384-5380251	8.000
Gtpbp10	NM_153116	chr5:5537456-5559501	12.000
Zfp804b	NM_001163223	chr5:6769029-7344378	7.000
Sri	NM_001080974	chr5:8056541-8069314	8.000
Adam22	NM_001098225	chr5:8072351-8368081	13.000
Adam22	NM_001007221	chr5:8092259-8368081	13.000
Adam22	NM_001007220	chr5:8092259-8368081	13.000
Dbf4	NM_001190717	chr5:8396968-8422716	9.000
Dbf4	NM_013726	chr5:8396968-8422716	9.000
Abcb1b	NM_011075	chr5:8798146-8866314	9.000
Crot	NM_023733	chr5:8966047-8997146	10.000
Phtf2	NM_172992	chr5:20758663-20882124	25.000
Ccd146	NM_029195	chr5:21292960-21424677	7.000
Fam185a	NM_177869	chr5:21424902-21482124	7.000
Kmt2c	NM_001081383	chr5:25271793-25498783	9.000
4831440E17Rik	NR_030700	chr5:25499796-25504473	9.000
Xrcc2	NM_020570	chr5:25689815-25705797	13.000
Dpp6	NM_00119886	chr5:26904777-27727500	8.000
Paxip1	NM_018878	chr5:27740080-27791550	9.000
Insig1	NM_153526	chr5:28071411-28078662	12.000
Supt7l	NM_028150	chr5:31514568-31526762	13.000
Slc4a1ap	NM_009206	chr5:31526994-31554038	13.000

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

Results Window			
2707 elements returned: (Oct4 in promoter)			
Gene Name	Accession	Location	Normalized
Rbm48	NM_172991	chr5:3583977-3596547	4.074
Pexl	NM_001293806	chr5:3596065-3637230	4.074
Pexl	NM_027777	chr5:3596065-3637230	4.074
Cdk14	NM_011074	chr5:4803384-5380251	1.630
Gtpbp10	NM_153116	chr5:5537456-5559501	2.444
Zfp804b	NM_001163223	chr5:6769029-7344378	1.426
Sri	NM_001080974	chr5:8056541-8069314	1.630
Adam22	NM_001098225	chr5:8072351-8368081	2.648
Adam22	NM_001007221	chr5:8092259-8368081	2.648
Adam22	NM_001007220	chr5:8092259-8368081	2.648
Dbf4	NM_001190717	chr5:8396968-8422716	1.833
Dbf4	NM_013726	chr5:8396968-8422716	1.833
Abcb1b	NM_011075	chr5:8798146-8866314	1.833
Crot	NM_023733	chr5:8966047-8997146	2.037
Phtf2	NM_172992	chr5:20758663-20882124	5.092
Ccd146	NM_029195	chr5:21292960-21424677	1.426
Fam185a	NM_177869	chr5:21424902-21482124	1.426
Kmt2c	NM_001081383	chr5:25271793-25498783	1.833
4831440E17Rik	NR_030700	chr5:25499796-25504473	1.833
Xrcc2	NM_020570	chr5:25689815-25705797	2.648
Dpp6	NM_00119886	chr5:26904777-27727500	1.630
Paxip1	NM_018878	chr5:27740080-27791550	1.833
Insig1	NM_153526	chr5:28071411-28078662	2.444
Supt7l	NM_028150	chr5:31514568-31526762	2.648
Slc4a1ap	NM_009206	chr5:31526994-31554038	2.648

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.

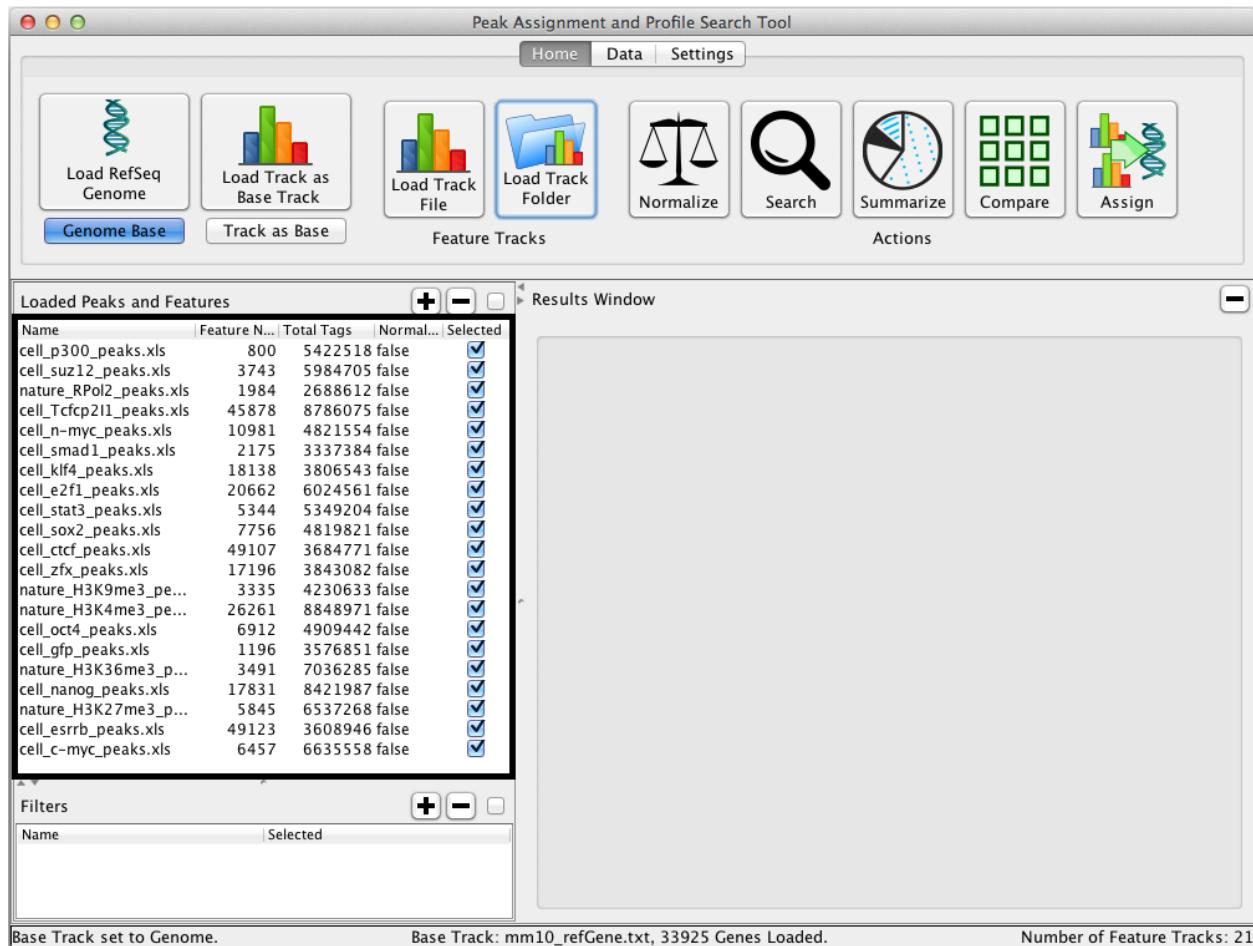
1. Load the **mm10_refGene.txt** gene models.
2. Click '**Load Track Folder**' and select the folder **/peaks/macs.xls**
3. If prompted, click '**Choose Parser from List**'



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



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



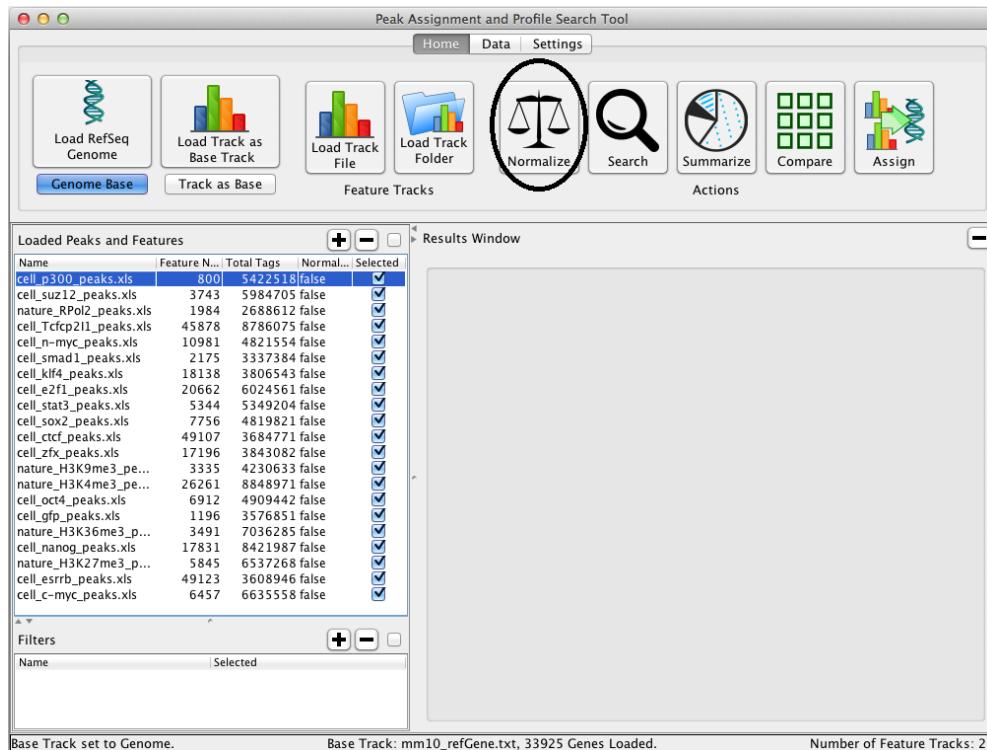
6. Create a tab delimited file with the peak set names followed by the tag count. Use one line per peak set like below. The example file, **tag_counts.txt**, is provided in the walkthrough folder.

tag_counts.txt

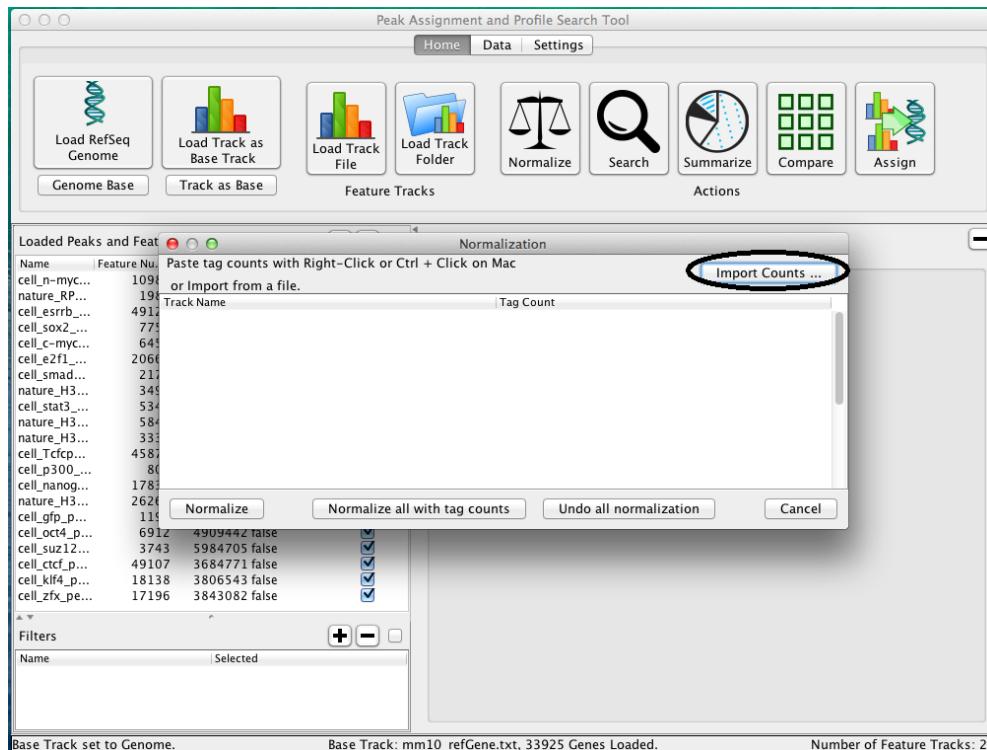
```
cell_Tcfcp2l1_peaks.xls 8786075
cell_c-myc_peaks.xls 6635558
cell_ctcf_peaks.xls 3684771
cell_e2f1_peaks.xls 6024561
cell_esrrb_peaks.xls 3608946
cell_gfp_peaks.xls 3576851
```

...

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.

Normalization	
Paste tag counts with Right-Click or Ctrl + Click on Mac or Import from a file.	
Track Name	Tag Count
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
cell_klf4_peaks.xls	3806543
cell_n-myc_peaks.xls	4821554
cell_nanog_peaks.xls	8421987
cell_oct4_peaks.xls	4909442
cell_p300_peaks.xls	5422518

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

The screenshot shows the 'Peak Assignment and Profile Search Tool' window. At the top, there are tabs for Home, Data, and Settings. Below the tabs is a toolbar with icons for Load RefSeq Genome, Load Track as Base Track, Load Track File, Load Track Folder, Normalize (which is highlighted with a blue border), Search, Summarize, Compare, and Assign. The main area has two panes: 'Loaded Peaks and Features' on the left and 'Results Window' on the right. In the 'Loaded Peaks and Features' pane, there is a table with columns: Name, Feature Nu..., Total Tags, and Normali... (with a dropdown menu). A vertical column of checkboxes labeled 'Selected' is to the right of the table. Most checkboxes are checked. In the 'Results Window' pane, there is a large empty area. At the bottom, there are status messages: 'Base Track set to Genome.', 'Base Track: mm10_refGene.txt, 33925 Genes Loaded.', and 'Number of Feature Tracks: 21'.

Name	Feature Nu...	Total Tags	Normali...	Selected
cell_p300_peaks.xls	800	542251	true	<input checked="" type="checkbox"/>
cell_suz12_peaks....	3743	598470	true	<input checked="" type="checkbox"/>
nature_RPol2_pea...	1984	268861	true	<input checked="" type="checkbox"/>
cell_Tcfcp2l1_pea...	45878	878607	true	<input checked="" type="checkbox"/>
cell_n-myc_peaks...	10981	482155	true	<input checked="" type="checkbox"/>
cell_smad1_peaks...	2175	333738	true	<input checked="" type="checkbox"/>
cell_klf4_peaks.xls	18138	380654	true	<input checked="" type="checkbox"/>
cell_e2f1_peaks.xls	20662	602456	true	<input checked="" type="checkbox"/>
cell_stat3_peaks.xls	5344	534920	true	<input checked="" type="checkbox"/>
cell_sox2_peaks.xls	7756	481982	true	<input checked="" type="checkbox"/>
cell_ctcf_peaks.xls	49107	368477	true	<input checked="" type="checkbox"/>
cell_zfx_peaks.xls	17196	384308	true	<input checked="" type="checkbox"/>
nature_H3K9me3_...	3335	423063	true	<input checked="" type="checkbox"/>
nature_H3K4me3_...	26261	884897	true	<input checked="" type="checkbox"/>
cell_oct4_peaks.xls	6912	490944	true	<input checked="" type="checkbox"/>
cell_gfp_peaks.xls	1196	357685	true	<input checked="" type="checkbox"/>
nature_H3K36me...	3491	703628	true	<input checked="" type="checkbox"/>
cell_nanog_peaks.xls	17831	842198	true	<input checked="" type="checkbox"/>
nature_H3K27me...	5845	653726	true	<input checked="" type="checkbox"/>
cell_esrrb_peaks.xls	49123	360894	true	<input checked="" type="checkbox"/>
cell_c-myc_peaks....	6457	663555	true	<input checked="" type="checkbox"/>

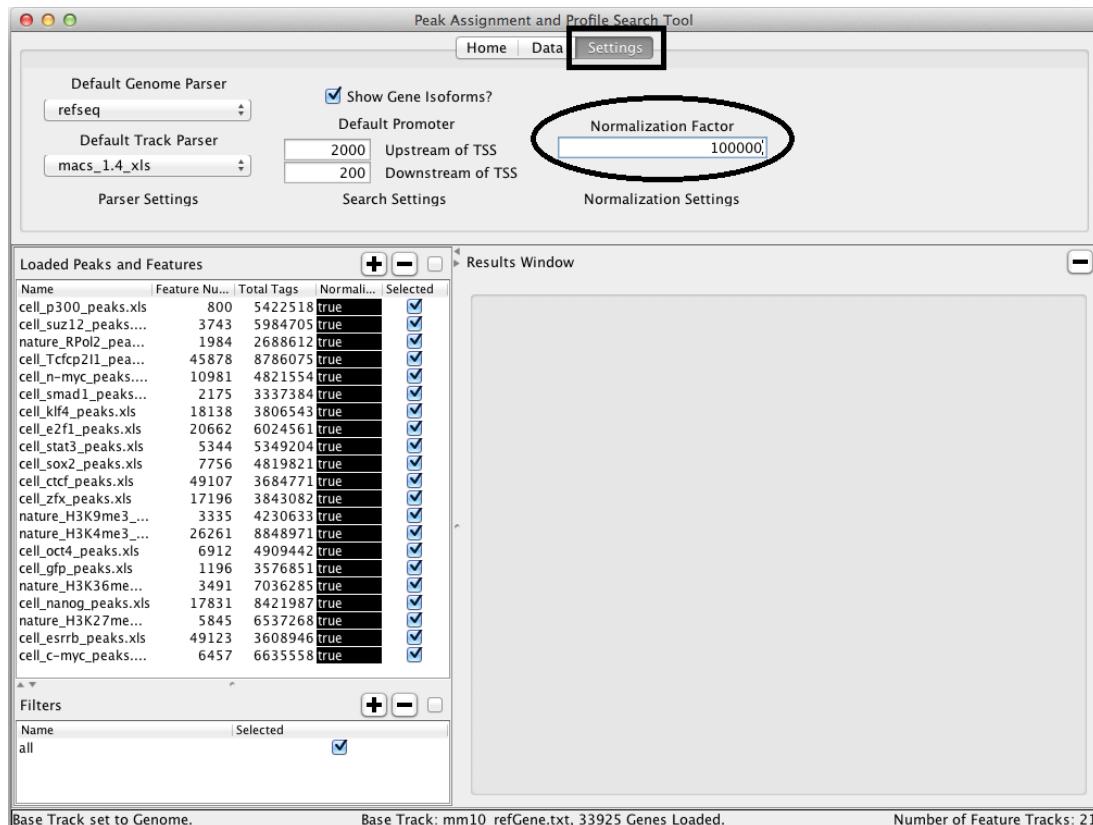
Filters

Base Track set to Genome. Base Track: mm10_refGene.txt, 33925 Genes Loaded. Number of Feature Tracks: 21

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.

Results Window						
33925 elements returned: (all)						
Gene Name	Accession	Location	cell_Tc...	cell_c-myc...	cell_ctcf_p...	cell_e2f1...
Rn45s	NR_046233	chr17:39...	354.652	423.175	112.626	329.153
Mapk4	NM_1726...	chr18:73...	87.183	0.000	0.000	3.154
Slc43a2	NM_1733...	chr11:75...	86.159	3.165	15.740	16.599
Slc43a2	NM_0011...	chr11:75...	86.159	3.165	15.740	16.599
Slc43a2	NM_0011...	chr11:75...	86.159	3.165	15.740	16.599
Slc16a6	NM_0010...	chr11:10...	83.769	9.946	2.442	19.918
Slc16a6	NM_1340...	chr11:10...	83.769	9.946	2.442	19.918
Hmgb2	NM_0082...	chr8:575...	78.192	2.261	0.000	80.006
Etv4	NM_0088...	chr11:10...	74.208	0.000	0.000	27.554
Nxnl1	NM_1455...	chr8:715...	69.542	0.000	0.000	0.000
Slc27a1	NM_0119...	chr8:715...	69.542	0.000	0.000	0.000

Results using 100k as the normalization factor.

Results Window								
33925 elements returned: (all)								
Gene Name	Accession	Location	cell_Tc...	cell_c-myc...	cell_ctcf_p...	cell_e2f1...	cel...	cel...
Rn45s	NR_046233	chr17:39...	35.465	42.317	11.263	32.915		
Mapk4	NM_1726...	chr18:73...	8.718	0.000	0.000	0.315		
Slc43a2	NM_1733...	chr11:75...	8.616	0.316	1.574	1.660		
Slc43a2	NM_0011...	chr11:75...	8.616	0.316	1.574	1.660		
Slc43a2	NM_0011...	chr11:75...	8.616	0.316	1.574	1.660		
Slc16a6	NM_0010...	chr11:10...	8.377	0.995	0.244	1.992		
Slc16a6	NM_1340...	chr11:10...	8.377	0.995	0.244	1.992		
Hmgb2	NM_0082...	chr8:575...	7.819	0.226	0.000	8.001		
Etv4	NM_0088...	chr11:10...	7.421	0.000	0.000	2.755		
Nxnl1	NM_1455...	chr8:715...	6.954	0.000	0.000	0.000		
Slc27a1	NM_0119...	chr8:715...	6.954	0.000	0.000	0.000		
Slc29a1	NM_0011...	chr17:45...	6.806	0.000	0.000	27.803		

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.

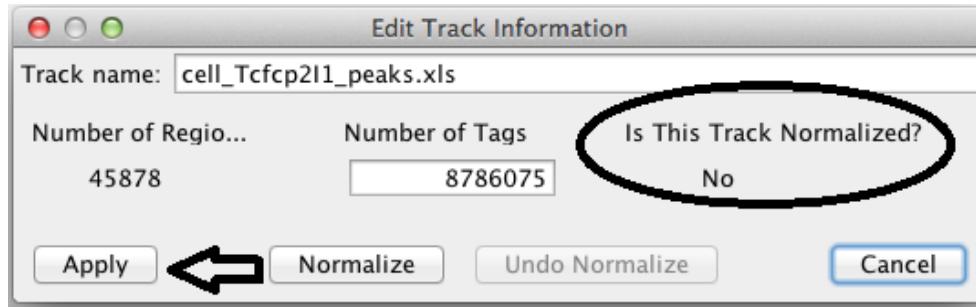
The screenshot shows the PAPST software interface. At the top, there are tabs for Home, Data, and Settings. Below the tabs are several icons: Load RefSeq Genome, Load Track as Base Track, Load Track File, Load Track Folder, Normalize, Search, Summarize, Compare, and Assign. The main window has two panes: 'Loaded Peaks and Features' on the left and 'Results Window' on the right. The 'Loaded Peaks and Features' pane lists various peak sets with columns for Name, Feature Num., Total Tags, Normalized, and Selected. A specific entry, 'cell_Tcfcp2l1_peaks.xls', is selected. The 'Results Window' pane shows a table of results with columns for Gene Name, Accession, Location, and several expression values. An 'Edit Track Information' dialog is open over the Results Window, containing fields for Track name, Number of Tags, and Is This Track Normalized? (set to Yes). At the bottom of this dialog is a 'Normalize' button and a circled 'Undo Normalize' button. The status bar at the bottom of the application window displays: 'Base Track set to Genome.', 'Base Track: mm10_refGene.txt, 33925 Genes Loaded.', and 'Number of Feature Tracks: 21'.

Name	Feature Num.	Total Tags	Normalized	Selected
cell_Tcfcp2l1_peaks.xls	45878	8786075	true	<input checked="" type="checkbox"/>
cell_c-myc_peaks.xls	6457	6635558	true	<input checked="" type="checkbox"/>
cell_ctcf_peaks.xls	49107	31	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_e2f1_peaks.xls	20662	6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_esrrb_peaks.xls	49123	3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_gfp_peaks.xls	1196	3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_klf4_peaks.xls	18138	3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_n-myc_peaks.xls	10981	4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_nanog_peaks.xls	17831	8	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_oct4_peaks.xls	6912	4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_p300_peaks.xls	800	5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_smad1_peaks.xls	2175	3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
cell_sox2_peaks.xls	7756	4819821	true	<input checked="" type="checkbox"/>
cell_stat3_peaks.xls	5344	5349204	true	<input checked="" type="checkbox"/>
cell_suz12_peaks.xls	3743	5984705	true	<input checked="" type="checkbox"/>
cell_zfx_peaks.xls	17196	3843082	true	<input checked="" type="checkbox"/>
nature_H3K4me3...	26261	8848971	true	<input checked="" type="checkbox"/>
nature_H3K9me3...	3335	4230633	true	<input checked="" type="checkbox"/>
nature_H3K27me...	5845	6537268	true	<input checked="" type="checkbox"/>
nature_H3K36me...	3491	7036285	true	<input checked="" type="checkbox"/>
nature_RPol2_pea...	1984	2688612	true	<input checked="" type="checkbox"/>

Results Window

33925 elements returned: (all)								
Edit Track Information								
Track name:	cell_Tcfcp2l1_peaks.xls	Number of Tags	8786075	Is This Track Normalized?	Yes	Normalize	Undo Normalize	Cancel
myc...	2.317	11.263	32.915					
c-myc...	0.000	0.000	0.315					
ctcf...	0.316	1.574	1.660					
e2f1...	0.316	1.574	1.660					
gfp...	0.316	1.574	1.660					
klf4...	0.995	0.244	1.992					
nanog...	0.995	0.244	1.992					
oct4...	0.995	0.244	1.992					
p300...	0.226	0.000	8.001					
smad1...	0.000	0.000	2.755					
sox2...	0.000	0.000	0.000					
stat3...	0.000	0.000	0.000					
suz12...	0.000	0.000	27.803					
zfx...	0.000	0.000	27.803					
nature_H3K4me3...	0.000	0.000	27.803					
nature_H3K9me3...	0.000	0.000	27.803					
nature_H3K27me...	0.000	0.000	1.992					
nature_H3K36me...	0.000	0.000	0.000					
nature_RPol2_pea...	0.000	0.000	4.448					
Base Track set to Genome.	Base Track: mm10_refGene.txt, 33925 Genes Loaded.	Number of Feature Tracks: 21						

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.

Name	Feature Nu...	Total Tags	Normali...	Selected
cell_Tcfcp2l1_pea...	45878	87860	false	<input checked="" type="checkbox"/>
cell_c-myc_peaks...	6457	6635558	true	<input checked="" type="checkbox"/>
cell_ctcf_peaks.xls	49107	3684771	true	<input checked="" type="checkbox"/>
cell_e2f1_peaks.xls	20662	6024561	true	<input checked="" type="checkbox"/>
cell_esrrb_peaks.xls	49123	3608946	true	<input checked="" type="checkbox"/>
cell_gfp_peaks.xls	1196	3576851	true	<input checked="" type="checkbox"/>
cell_klf4_peaks.xls	18138	3806543	true	<input checked="" type="checkbox"/>
cell_n-myc_peaks....	10981	4821554	true	<input checked="" type="checkbox"/>
cell_nanog_peaks.xls	17831	8421987	true	<input checked="" type="checkbox"/>
cell_oct4_peaks.xls	6912	40909442	true	<input checked="" type="checkbox"/>
cell_p300_peaks.xls	800	5422518	true	<input checked="" type="checkbox"/>
cell_smad1_peaks...	2175	3337384	true	<input checked="" type="checkbox"/>
cell_sox2_peaks.xls	7756	4819821	true	<input checked="" type="checkbox"/>
cell_stat3_peaks.xls	5344	5349204	true	<input checked="" type="checkbox"/>
cell_suz12_peaks....	3743	5984705	true	<input checked="" type="checkbox"/>
cell_zfx_peaks.xls	17196	3843082	true	<input checked="" type="checkbox"/>
nature_H3K4me3....	26261	8848971	true	<input checked="" type="checkbox"/>
nature_H3K9me3....	3335	4230633	true	<input checked="" type="checkbox"/>
nature_H3K27me...	5845	6537268	true	<input checked="" type="checkbox"/>
nature_H3K36me...	3491	7036285	true	<input checked="" type="checkbox"/>
nature_RPol2_pea...	1984	2688612	true	<input checked="" type="checkbox"/>

33925 elements returned: (all)							
Gene Name	Accession	Location	cell_Tc...	cell_c-myc...	cell_ctcf_p...	cell_e2f1...	cel...
Rn45s	NR_046233	chr17:39...	35.465	42.317	11.263	32.915	
Mapk4	NM_1726...	chr18:73...	8.718	0.000	0.000	0.315	
Slc43a2	NM_1733...	chr11:75...	8.616	0.316	1.574	1.660	
Slc43a2	NM_0011...	chr11:75...	8.616	0.316	1.574	1.660	
Slc16a6	NM_0010...	chr11:10...	8.377	0.995	0.244	1.992	
Slc16a6	NM_1340...	chr11:10...	8.377	0.995	0.244	1.992	
Hmgb2	NM_0082...	chr8:575...	7.819	0.226	0.000	8.001	
Etv4	NM_0088...	chr11:10...	7.421	0.000	0.000	2.755	
Nxnl1	NM_1455...	chr8:715...	6.954	0.000	0.000	0.000	
Slc27a1	NM_0119...	chr8:715...	6.954	0.000	0.000	0.000	
Slc29a1	NM_0011...	chr17:45...	6.806	0.000	0.000	27.803	
Gm7325	NM_0011...	chr17:45...	6.806	0.000	0.000	27.803	
Gm7325	NM_0011...	chr17:45...	6.806	0.000	0.000	27.803	
Gm7325	NM_0011...	chr17:45...	6.806	0.000	0.000	27.803	
Arsg	NM_0287...	chr11:10...	6.544	0.573	0.000	1.992	
Ccdc141	NM_0010...	chr2:770...	6.488	0.000	0.000	0.000	
Nfe2l2	NM_0109...	chr2:756...	6.431	0.452	0.000	4.448	
Sfl1	NM_0302...	chr11:31...	6.271	6.510	7.680	12.051	
Actg1	NM_0096...	chr11:12...	6.203	0.000	0.000	9.777	
Sema4b	NM_0136...	chr7:801...	6.123	0.000	0.000	8.631	
Malat1	NR_002847	chr19:57...	6.101	0.482	0.000	0.000	
Gpcpd1	NM_0288...	chr2:132...	5.941	0.000	0.000	0.581	
Gpcpd1	NM_0012...	chr2:132...	5.941	0.000	0.000	0.581	

Base Track set to Genome. Base Track: mm10_refGene.txt, 33925 Genes Loaded. Number of Feature Tracks: 21

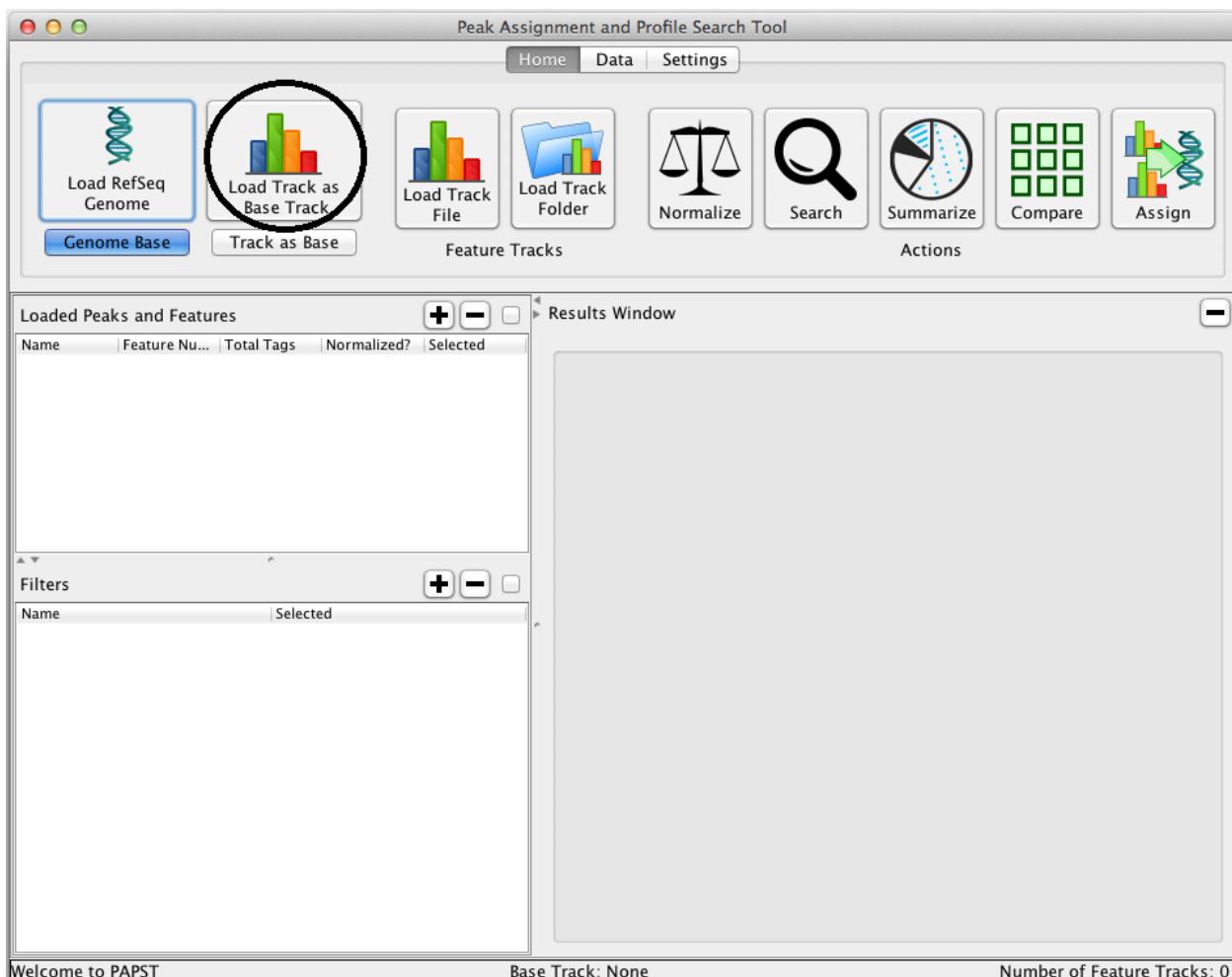
Alternative Base Tracks

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.

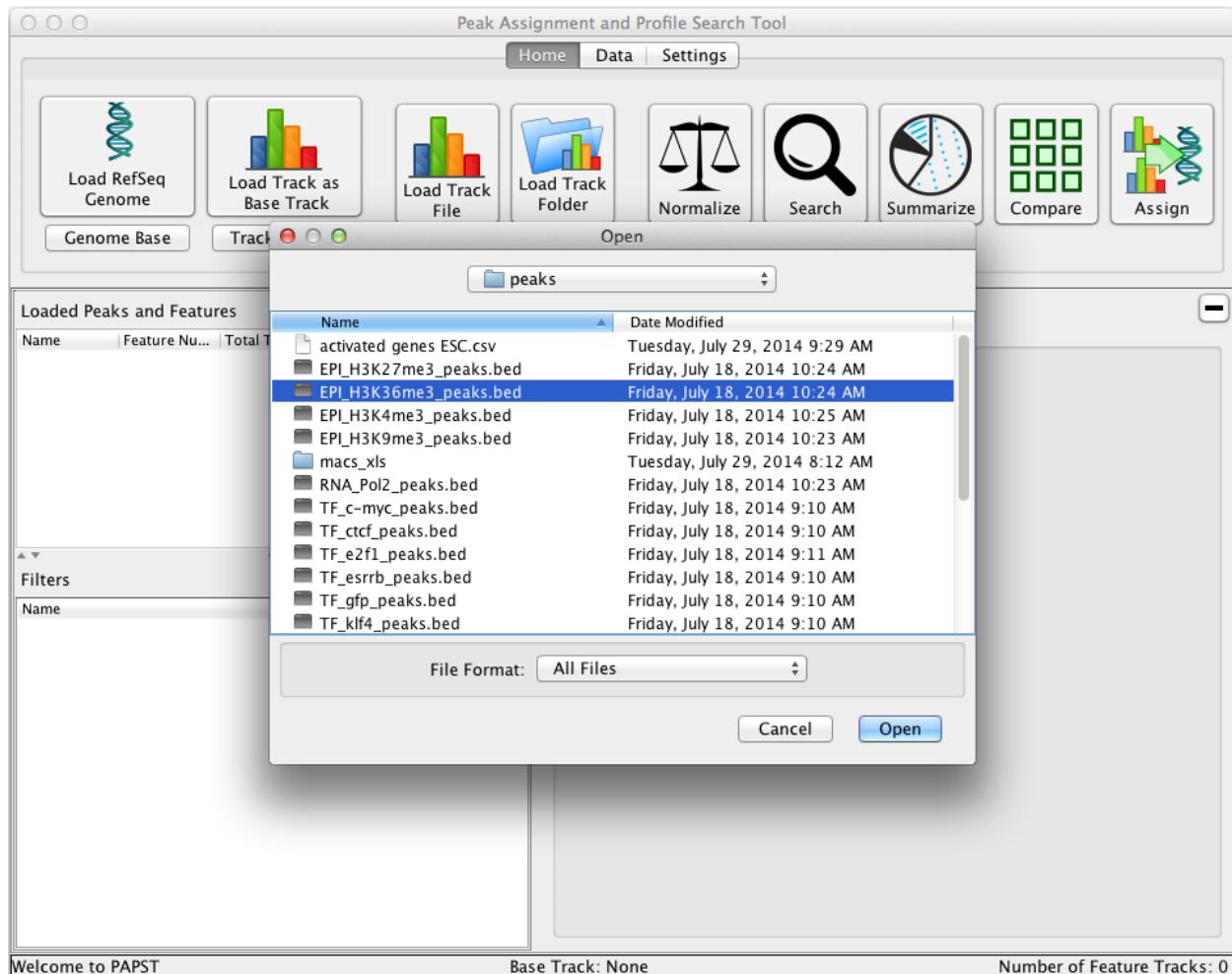
Search Example using K36me3 as Base

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.

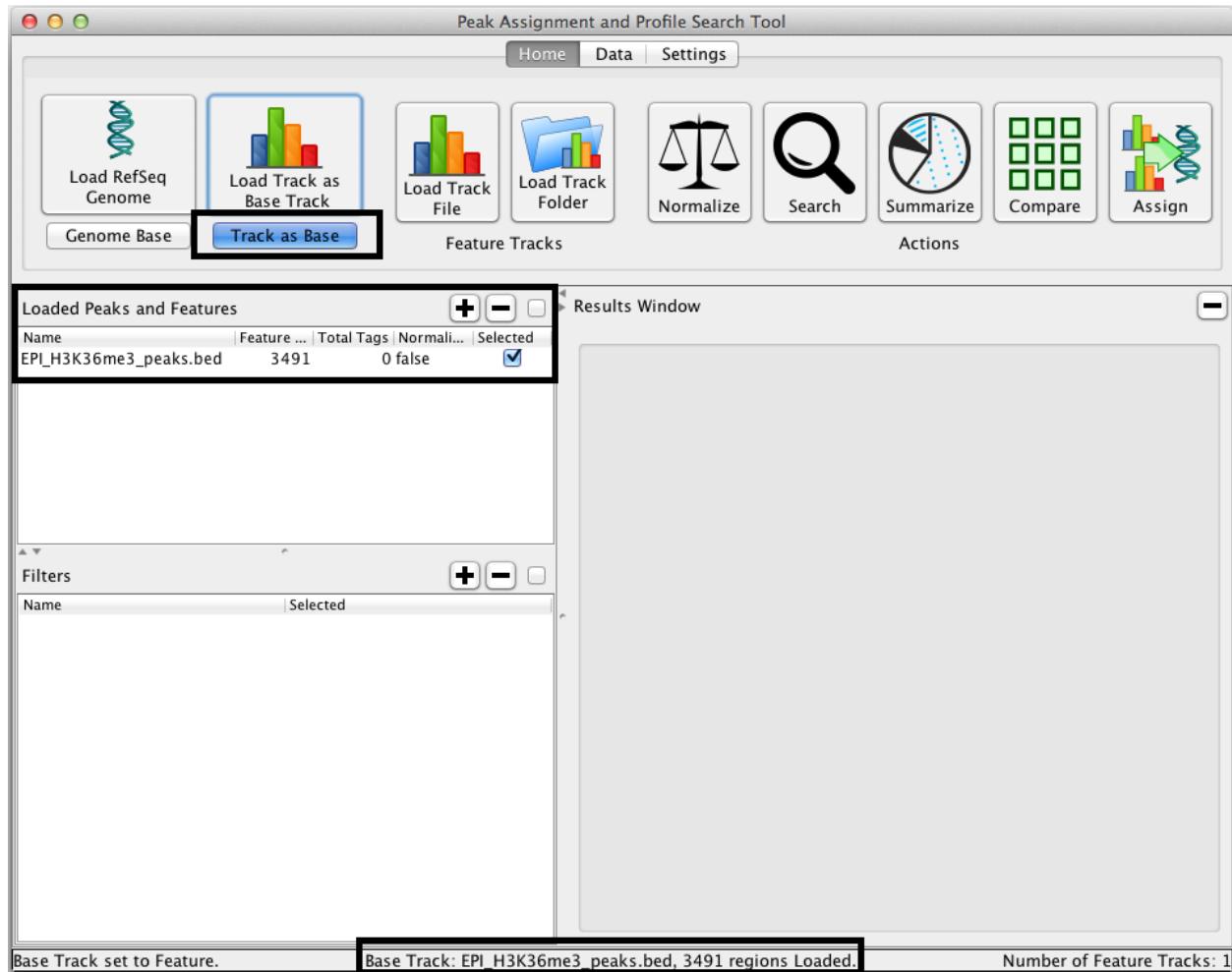
1. To load an alternative base track, click the **Load Track as Base Track** button on the **Home** tab.



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

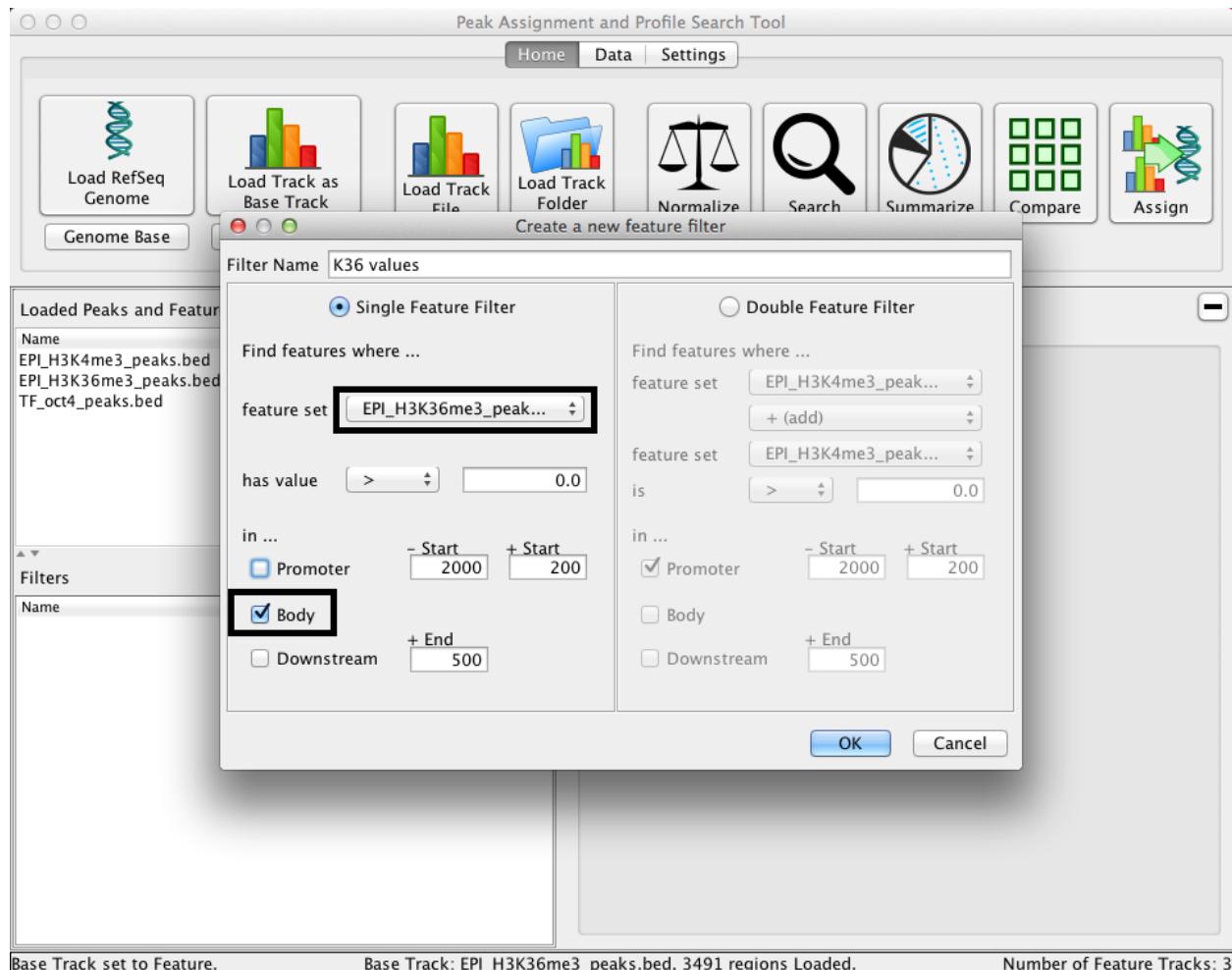


3. The peak track will be loaded and selected as the current base track. The **Track as Base** toggle button will be selected. The status bar at the bottom of the application window shows that the current base is **EPI_H3K36me3_peaks.bed**.



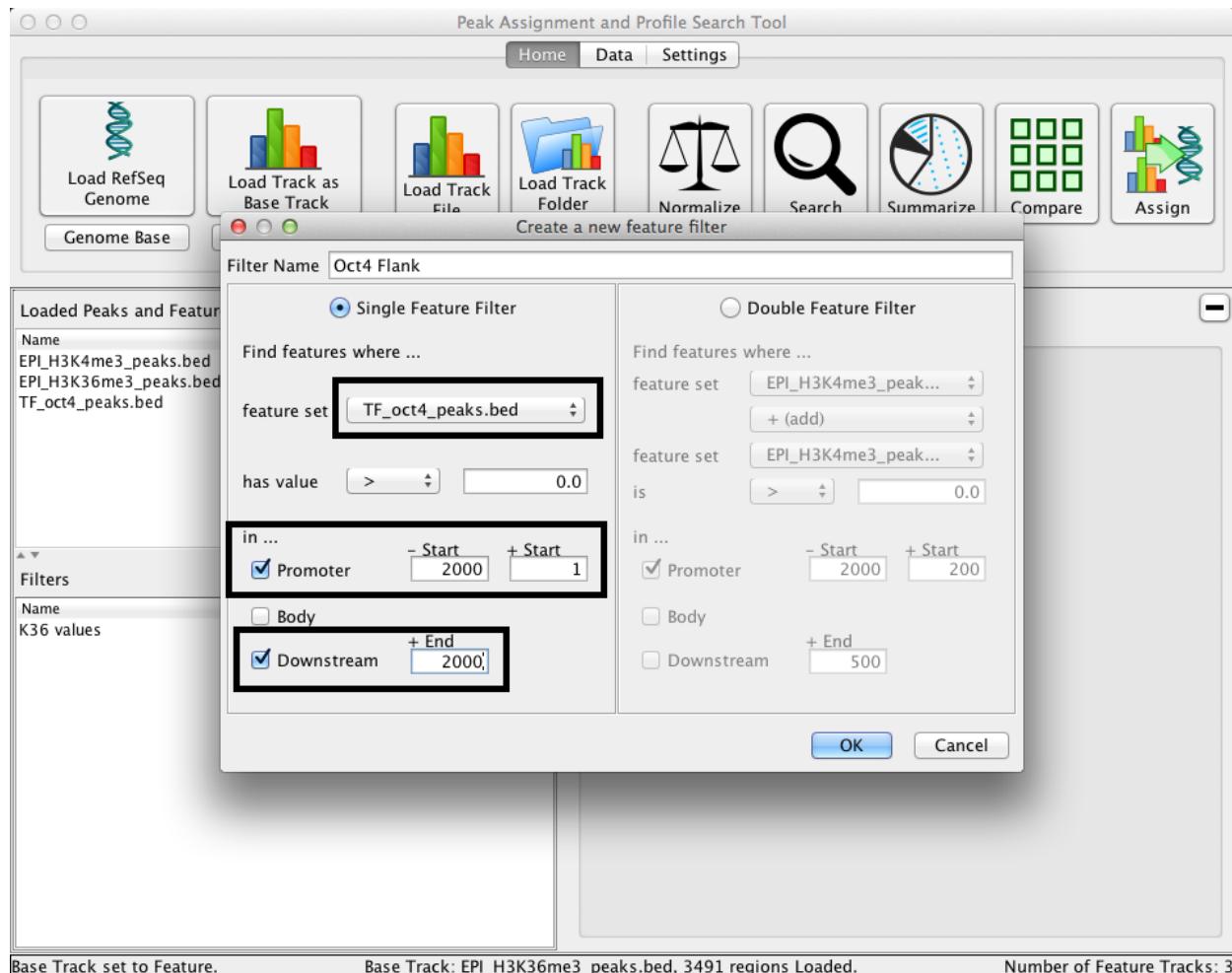
4. Load the histone modification track, **EPI_K4me3_peaks.bed**.
5. Load the transcription factor track, **TF_oct4_peaks.bed**. We will search for regions with K4 and Oct4 binding in the flanking regions of K36 sites.

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

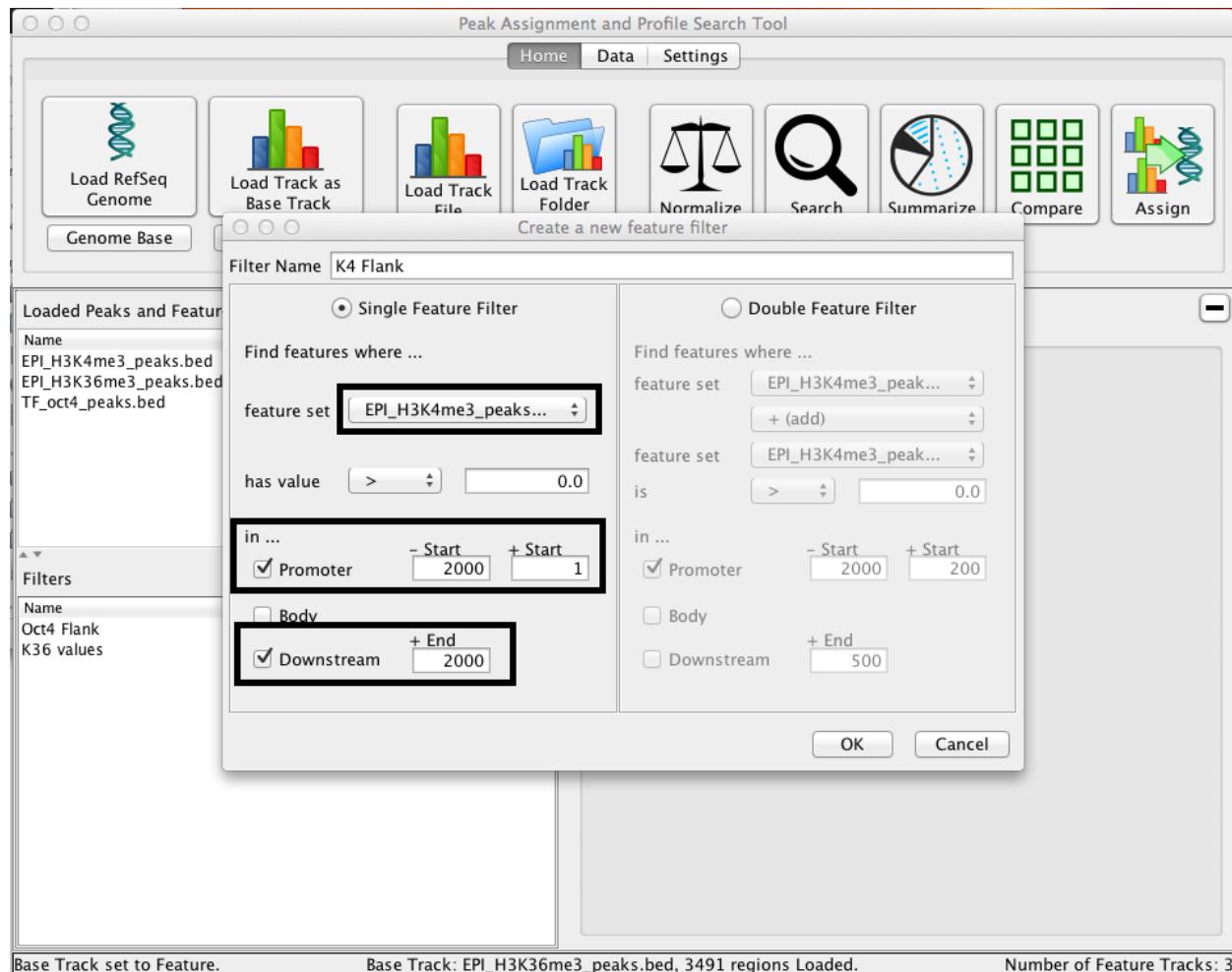


This filter will always accept because K36me3 is the current Base Track. This filter is necessary to report the K36me3 values to our results table.

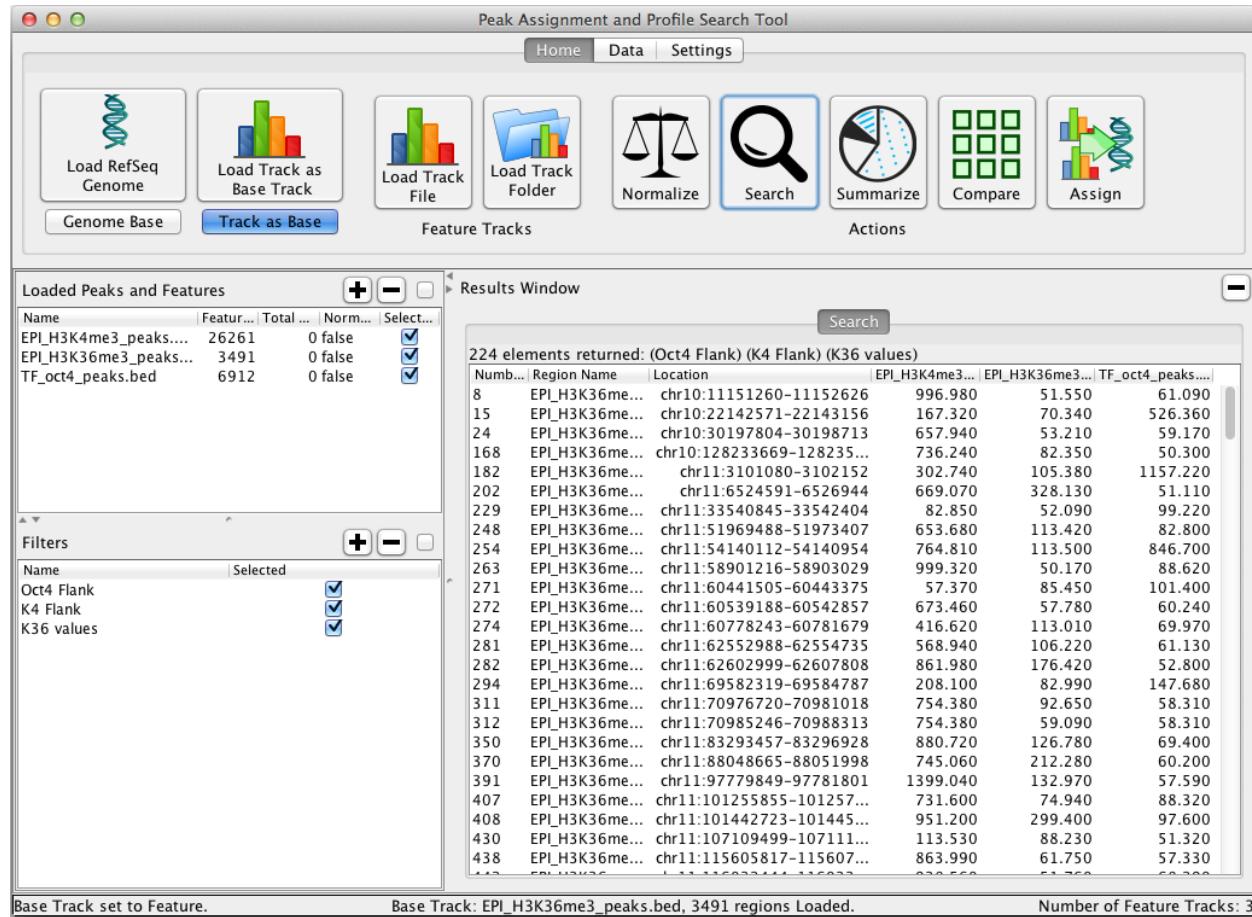
7. Create a new filter for Oct4. Select the Oct4 peak set. Select the promoter and downstream regions and set them to 2000. **When using Track as Base, the promoter represents the 5' flank and downstream represents the 3' flank.** Peaks are strand-less. Set the + Start to 1. Name the file 'Oct4 Flank'.



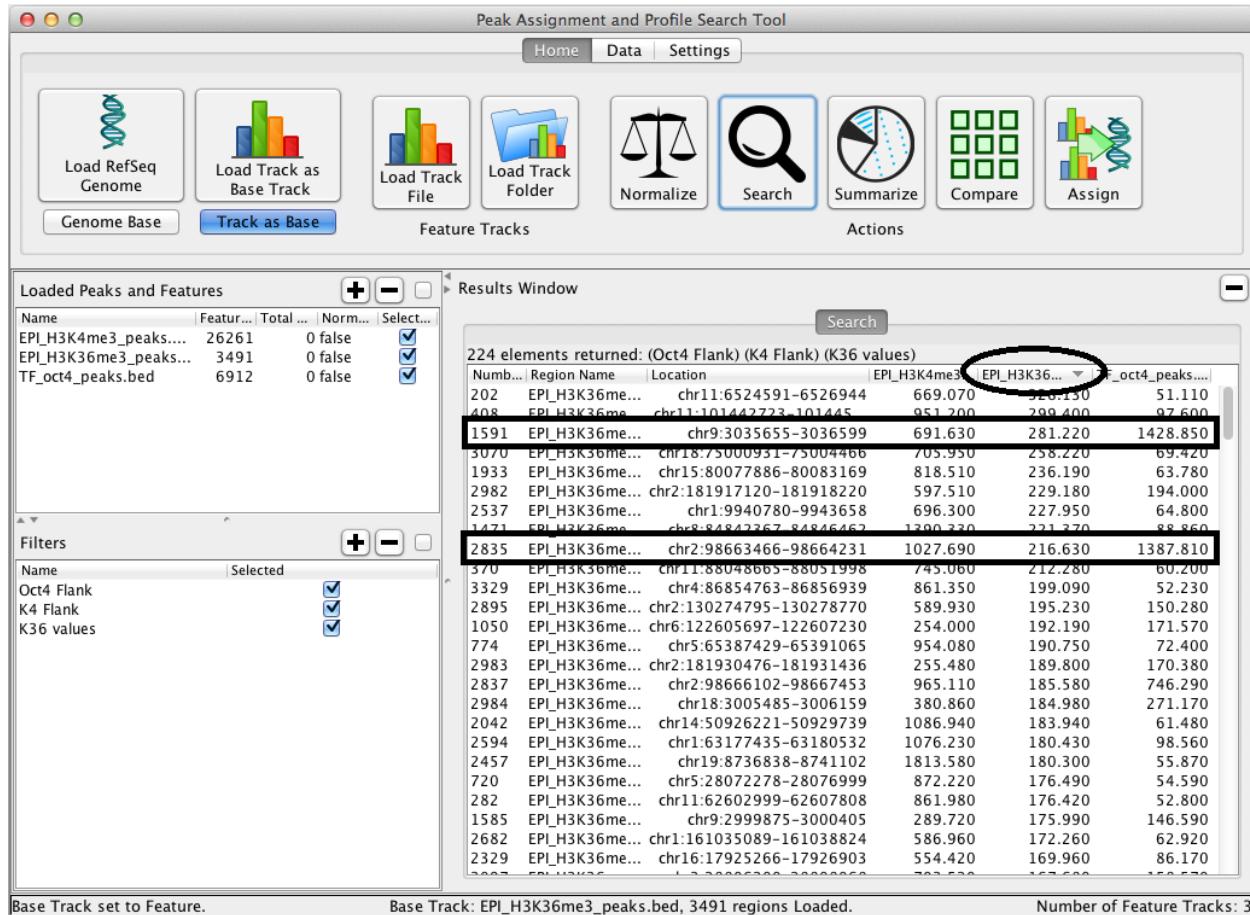
8. As in step 7 above, create a new filter for K4 with the same settings. Name the filter K4 Flank.



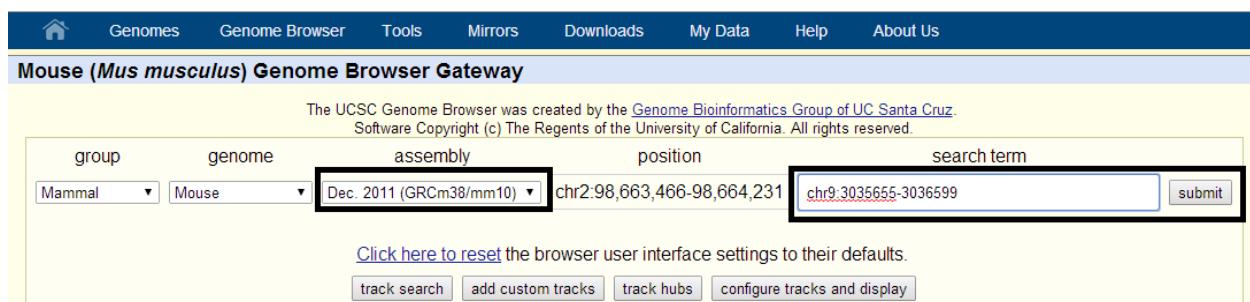
9. After creating the filters, click the **Search** button. PAPST will return all genomic regions that have Oct4 and K4me3 peaks in their flanking regions centered on K36me3 peaks. These regions may also include genes.



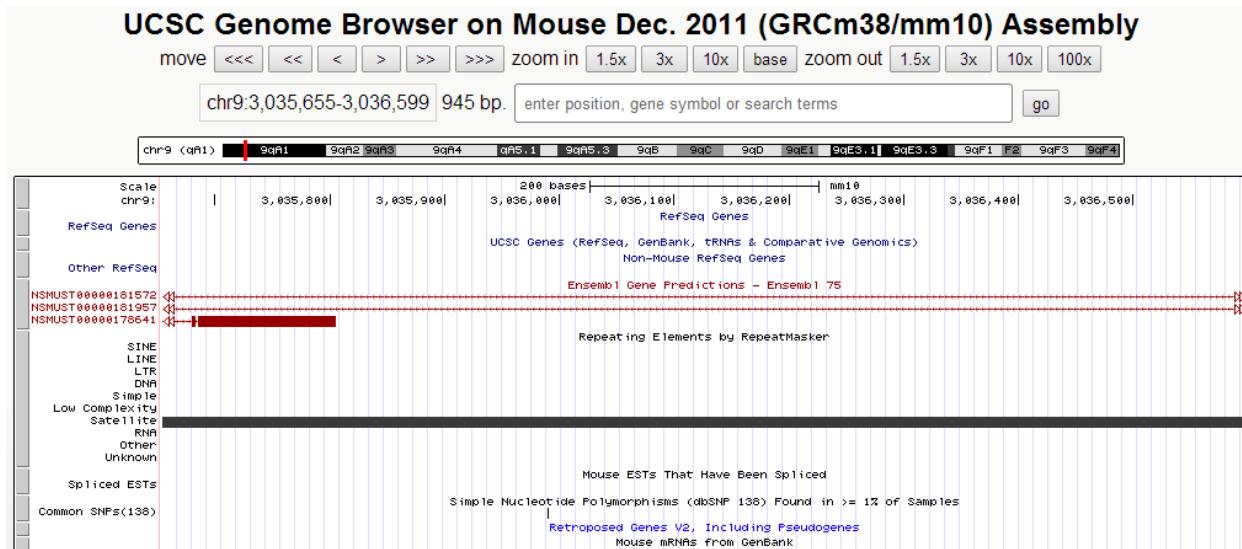
10. Sort the results table by the K36me3 value in descending order by clicking its column twice.
- Notice there are two regions with extremely high Oct4 signal among the highest K36me3 peaks.



11. Let's view these regions in the UCSC Genome Browser to learn more. Click on the **Location** cell of feature number 1591. Press **Command (⌘) + 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**'.

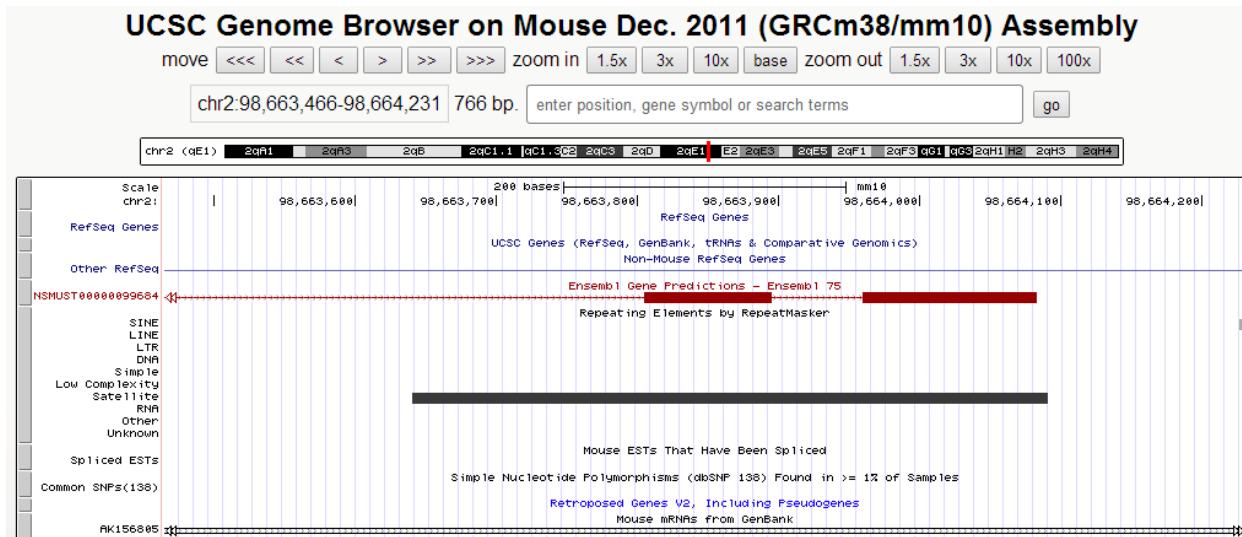


12. A region of Chr9 is return. A satellite region and two predicted genes occupy this region. There are no recognized RefSeq genes in this region though.



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.



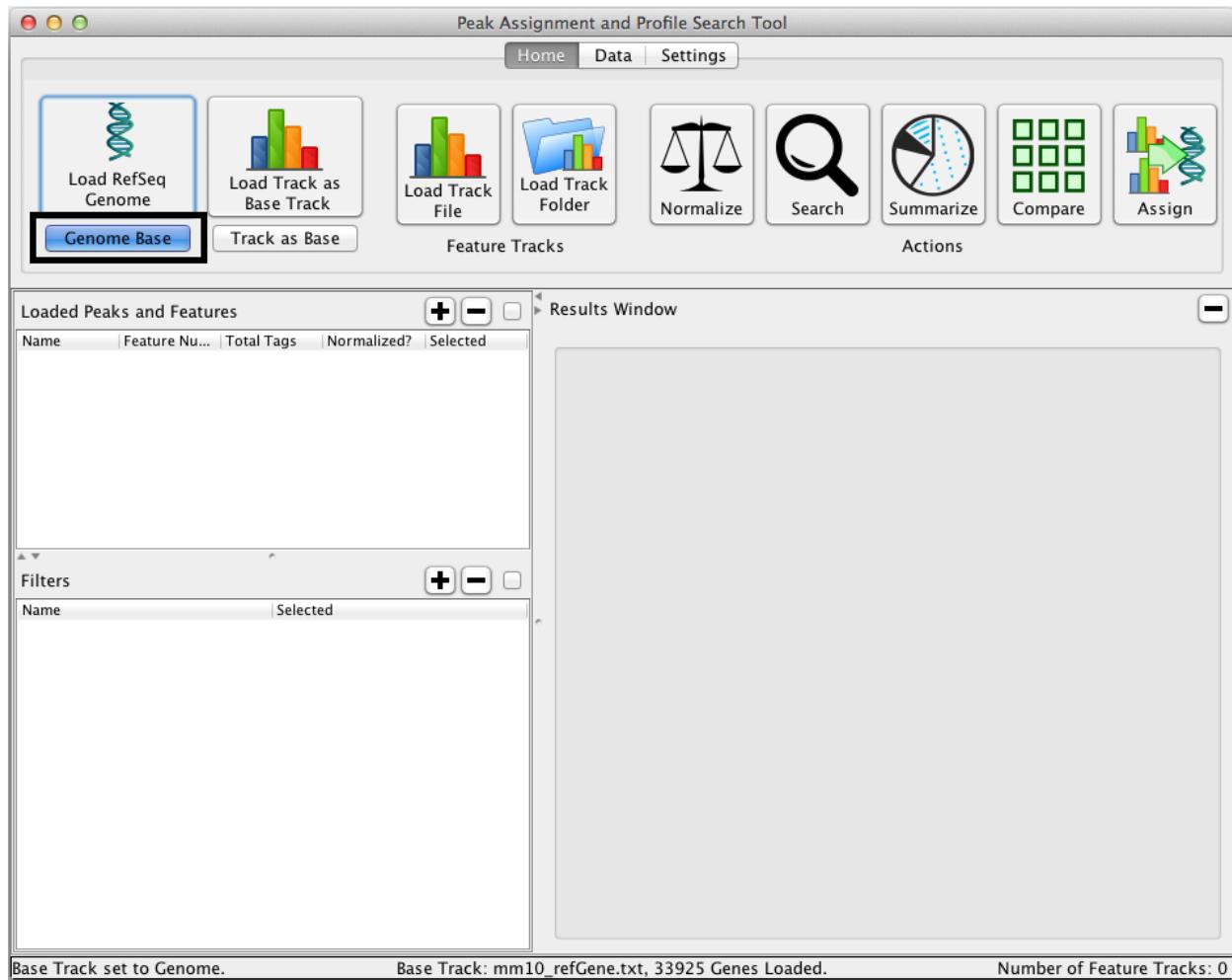
These regions have a particular epigenetic profile, specifically K36me3 flanked by K4me3 and Oct4 within 2000bp.

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

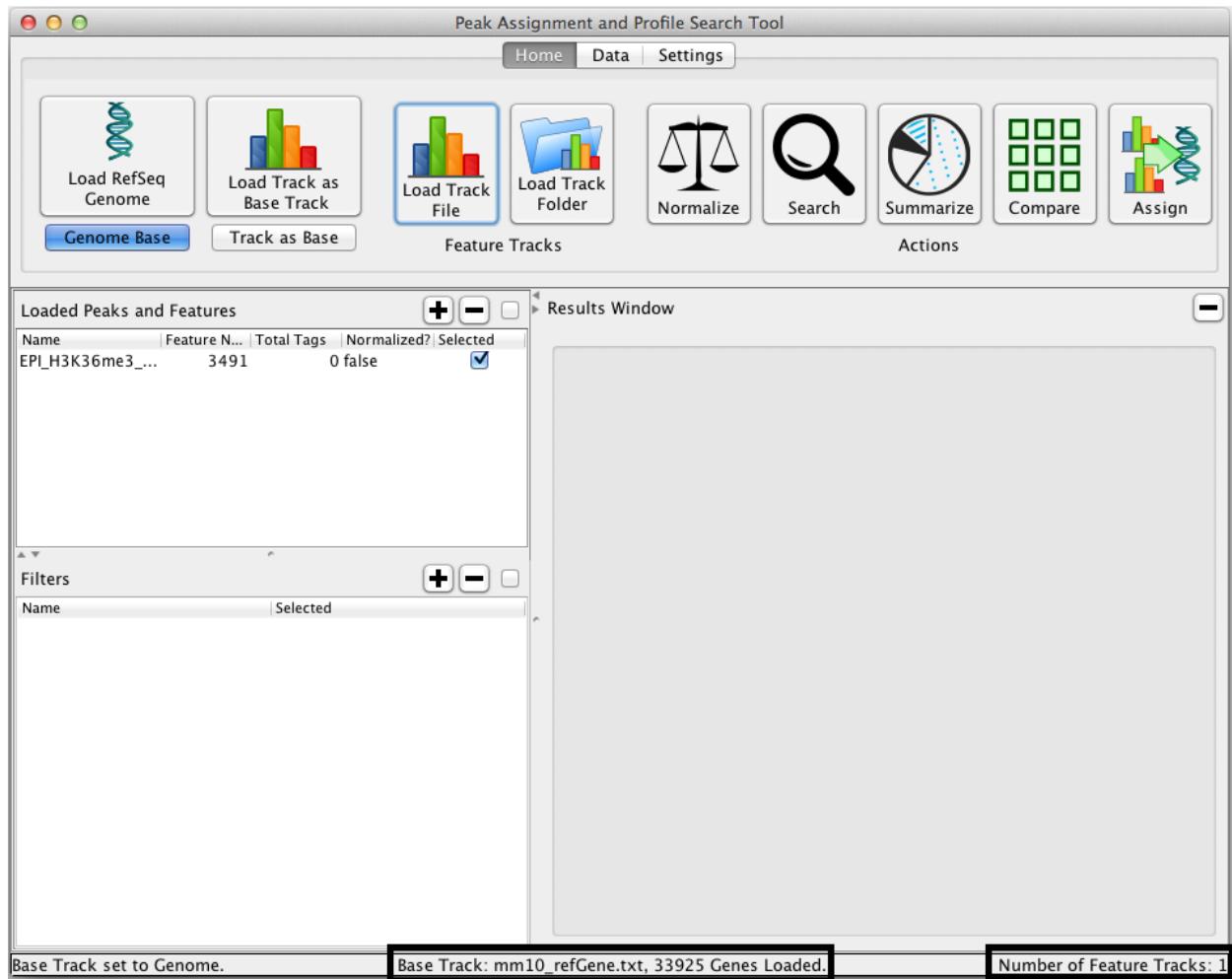
Switching between Genome and Custom Base Tracks

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.

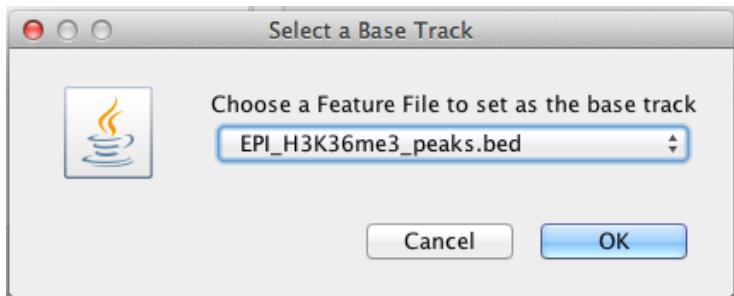


2. Add a peak track to PAPST by clicking **Load Track File**. Let's select **EPI_H3K36me3_peaks.bed**.

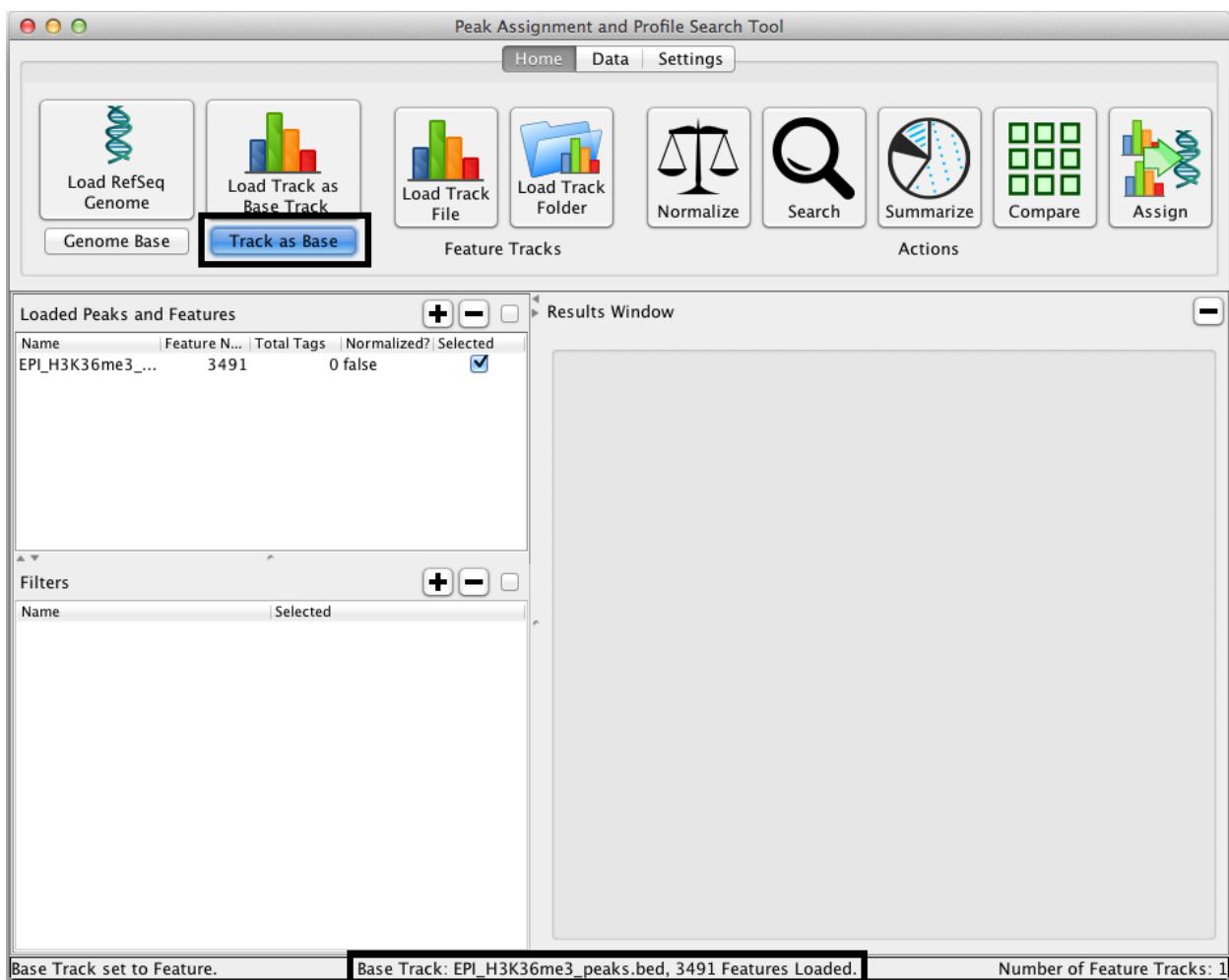


Notice that we have one peak set loaded and **mm10_refGene.txt** is the current base track.

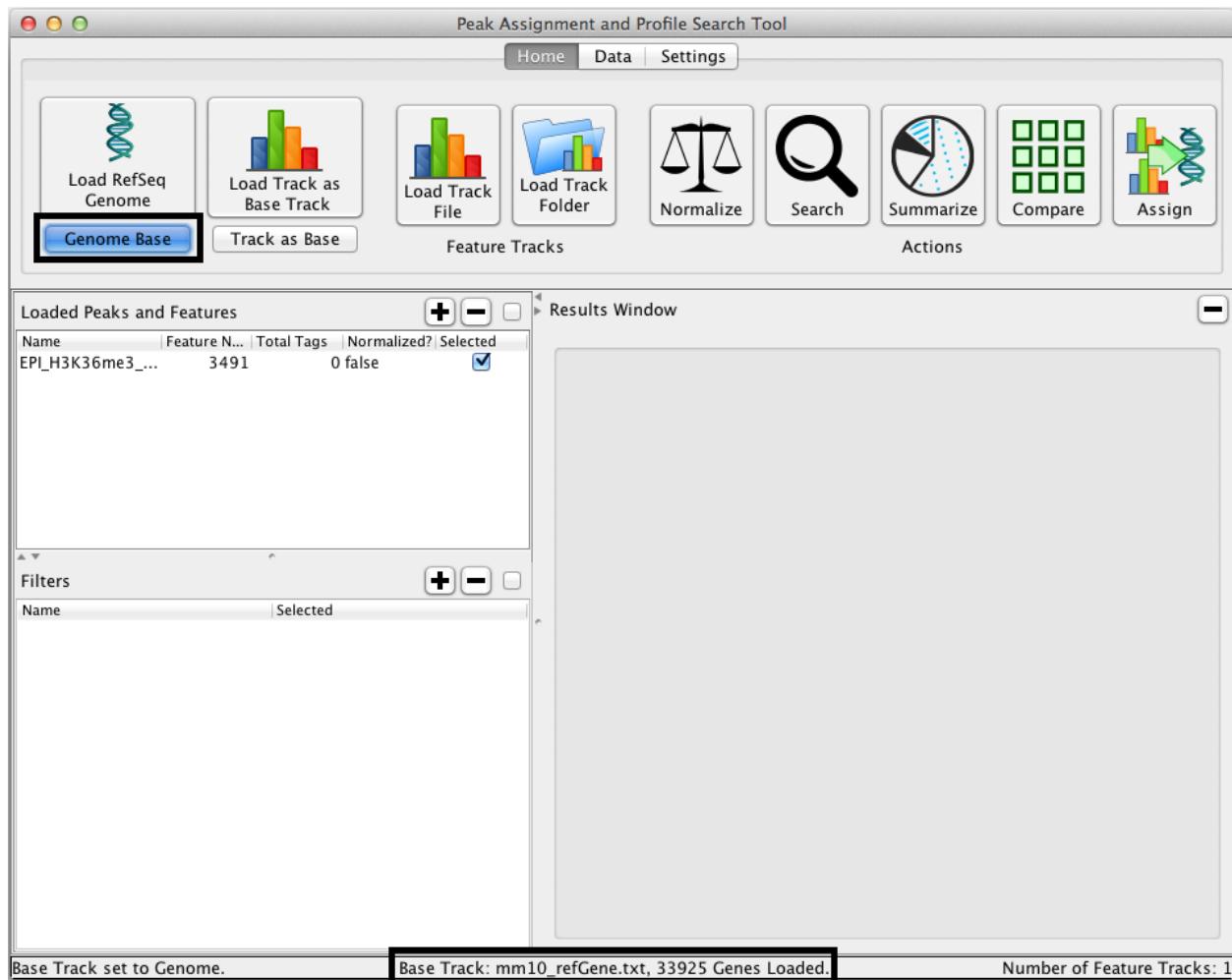
3. Let's change the current base track to K36me3. Click the **Track as Base** button. A track selection dialog will appear. Select **EPI_H3K36me3_peaks.bed** from the dropdown menu. Click OK.



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