

Identifying novel and known motifs

- [Discover de novo motifs](#)
- [Description of Motif Detection Output](#)
- [Search JASPAR for known motifs](#)
- [Generating a list of regions containing a motif](#)

With a list of enriched regions, you can now identify recurring patterns or motifs in these regions. Transcription factors bind sites throughout the genome, but each has a characteristic sequence it binds - a consensus sequence that appears in most of its binding sites. By searching for binding site motifs, you can determine the consensus sequence for a transcription factor and predict potential binding locations throughout the genome that may not have been found in your experiment.

Partek Genomics Suite detects *de novo* motifs using the Gibbs motif sampler (Neuwald et al., 1995) and can search for known transcription factor binding sites using a database such as [JASPAR](#).

Discover *de novo* motifs

- Select **Motif Discovery** from the *Peak Analysis* section of the *ChIP-Seq* workflow
- Select **Discover de novo motifs**
- Select **OK**

The *Detect Motifs* dialog will open to allow you to configure the search (Figure 1).

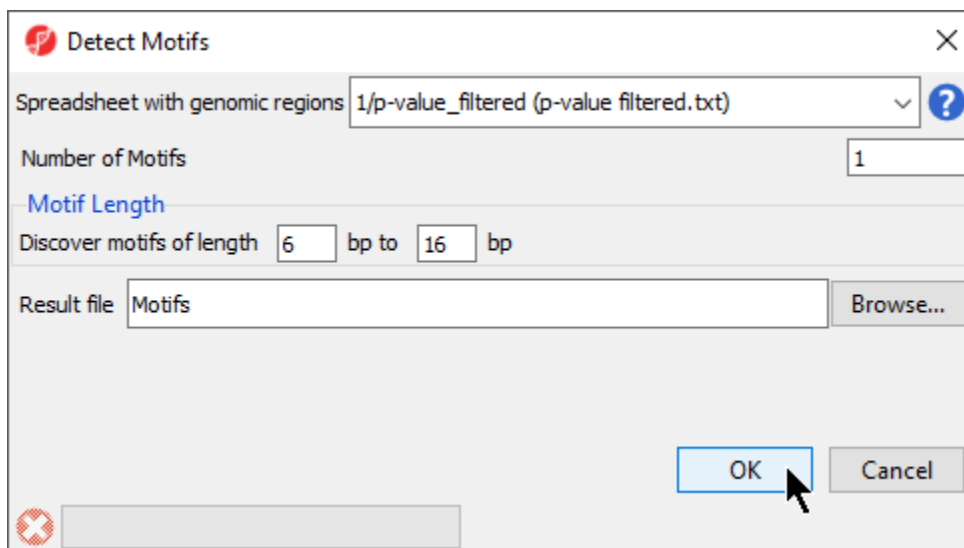


Figure 12. Configuring search parameters for *de novo* motifs

- Select **1/p-value_filtered** from the *Spreadsheet with genomic regions* drop-down menu
- Set *Number of Motifs* to **1**
- Set *Discover motifs of length* to **6 bp to 16 bp**
- Set *Result file* to **Motifs**
- Select **OK**

If you have not previously downloaded the reference genome on your computer, you may be asked if you would like to download the .2bit reference genome. If prompted, select **Automatically download a .2bit file** then select **OK**. If Partek Genomics Suite cannot connect to the internet, this option may not be available. If not, you will need to download the .2bit file from the UCSC Genome Browser and import it by selecting **Manually specify a .2bit file** and choosing the downloaded .2bit file. The reference genome map is required to determine which genes overlap the enriched peak regions and to display the aligned sequences in the *Genome Viewer*.

A motif visualization tab, *Sequence Logo*, will open and two spreadsheets will be generated. One spreadsheet, *motifs (Motifs)*, contains information about the motif. The other, *instances (Motifs_instances.txt)*, lists the genomic locations of the motif.

Description of Motif Detection Output

Sequence Logo Window

The *Sequence Logo* tab (Figure 2) opens after motif detection and displays the most significant motif found in the regions listed in the source spreadsheet.

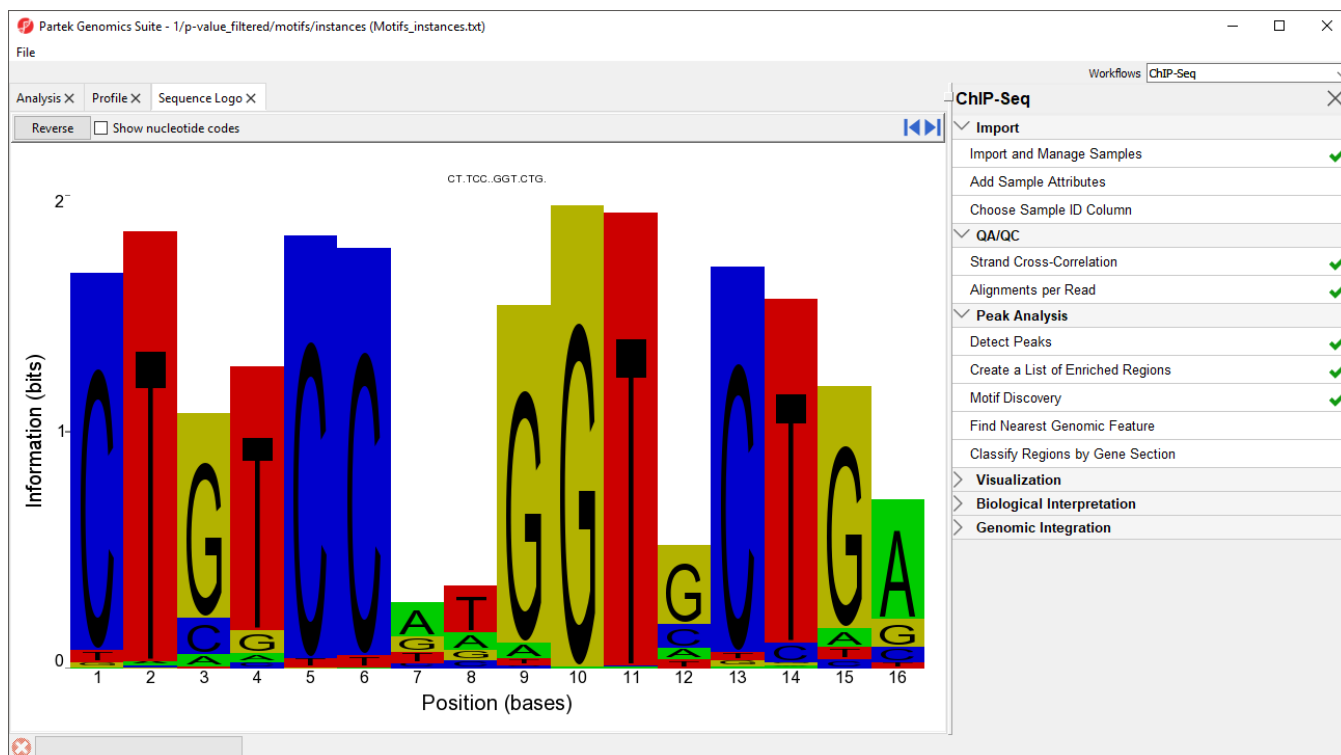


Figure 13. Viewing the binding site for NRSF. Use the blue arrows to cycle through views of all motif found (if there are more than one). Select Reverse to view the reverse complement sequence.

In this case, the motif finder discovered a motif in the NRSF-enriched regions that is 16 base pairs in length. The height of each position is the relative entropy (in bits) and indicates the importance of a base at a particular location in the binding site.

The title *CT.TCC..GGT.CTG.* is the consensus sequence for the sequence logo. Dots represent positions that contain more than one significant base across all reads in the motif. The dots can be replaced with characters representing the possible bases at each location by selecting **Show nucleotide codes**. A description of the IUPAC nucleotide codes is available at the [UCSC Genome Browser](#).

To view the reverse complement of the motif, select **Reverse**.

Motifs spreadsheet

The motif information spreadsheet (Figure 3), *Motifs*, lists the information about all motifs discovered during *de novo Motif Detection*. This includes five columns describing each motif.

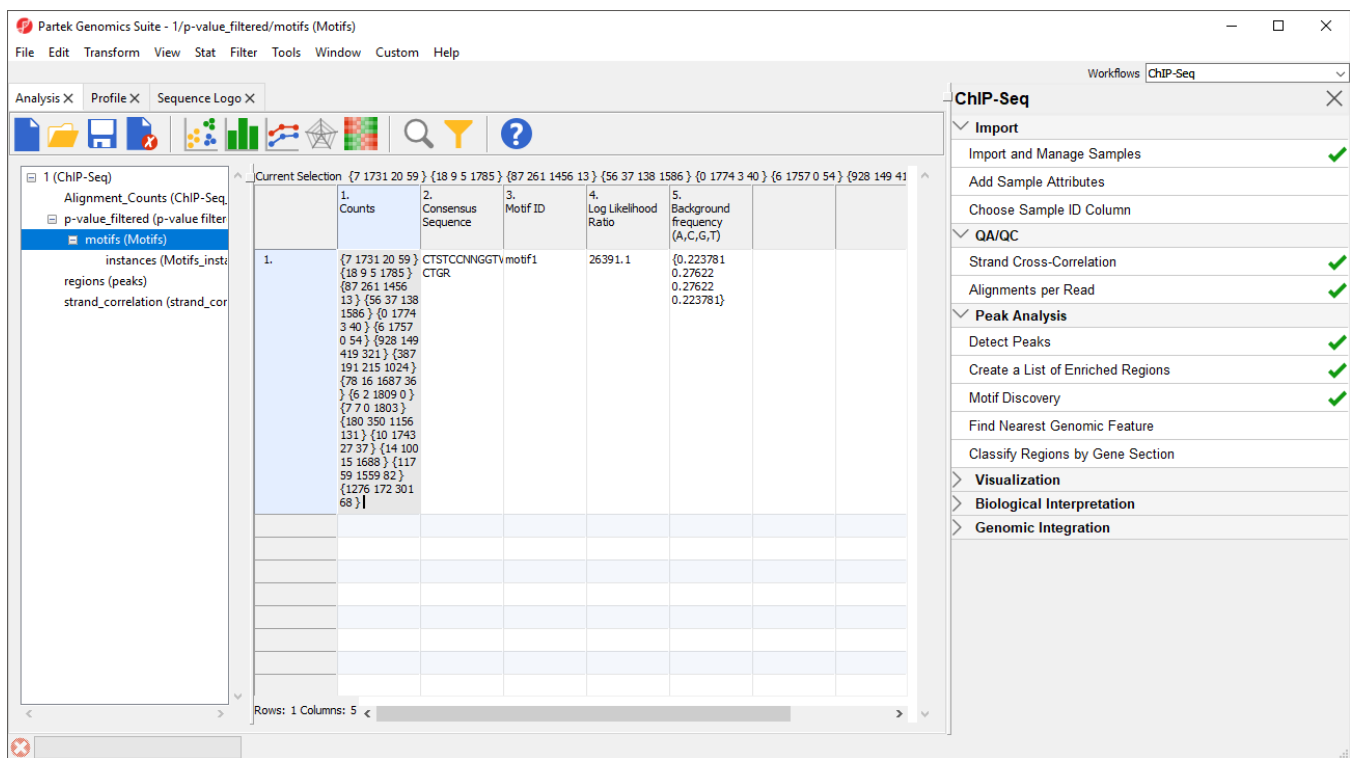


Figure 14. Viewing the Motifs spreadsheet

1. *Counts* gives the summed counts for each base call across all occurrences of the motif in the region list as {A, C, G, T}
2. *Consensus Sequence* gives the consensus sequence of the motif in IUPAC nucleotide codes
3. *Motif ID* gives a unique ID to each discovered motif using its row in the *Motifs* spreadsheet
4. *Log Likelihood Ratio* scores the relative likelihood that the pattern did not occur by chance, with larger numbers indicating that it is less likely to have occurred by chance
5. *Background frequency (A,C,G,T)* gives the frequency of each of the bases in all the sequences of that motif

You can bring up the Sequence Logo visualization of a listed motif by right-clicking on the row header and selecting **Logo View** from the pop-up menu.

Motif_instances spreadsheet

The *instances (Motif_instances)* spreadsheet (Figure 4) is a child spreadsheet of the *Motifs* spreadsheet. It details all the locations of the motif(s) detected in the enriched regions. Each row lists a putative binding site for a motif. The columns give detailed information about the putative binding sites.

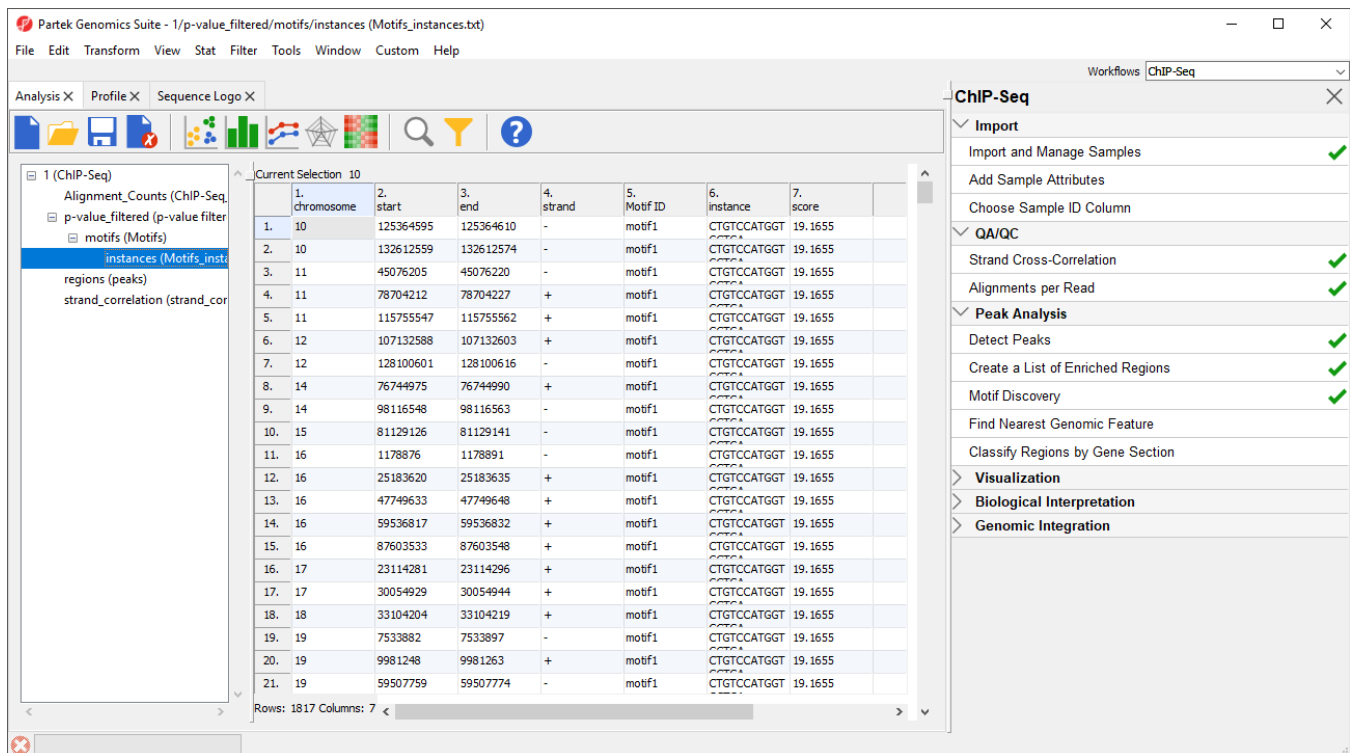


Figure 15. Viewing the instances spreadsheet

1-4. *chromosome*, *start*, *stop*, *strand* give the position

5. *Motif ID* gives the identity of the motif

6. *instance* gives the sequence of this instance of the motif

7. *score* gives the log ratio of the probability that this sequence was generated by the motif versus the background distribution. A higher number indicates a better chance that the sequence is an instance of the motif.

Search JASPAR for known motifs

- Select **Motif discovery** from the *Peak Analysis* section of the *ChIP-Seq* workflow
- Select **Search for known motifs**
- Select **OK**

Search for known motifs will search the JASPAR database for motifs that are over-represented in the list of sequences in the significant regions list. The JASPAR database will download automatically if needed during the *Search for known motifs* step. Downloading the JASPAR database will create a spreadsheet in your experiment named *JASPAR.txt* that contains all of the species-specific motifs in the database. To visualize the motifs, right-click on a row in the *JASPAR.txt* spreadsheet and select *Logo View*.

Before *Search for known motifs* runs, we need to configure the search (Figure 5).

Search for Motif(s) in Sequences

Description of Motif Search
This dialog will search the reference sequences of the given regions for the selected motifs. Each of the subsequences in the regions is scored against the motif model and is called a motif instance if the score exceeds the specified fraction of the best possible score.

Choose Region Spreadsheet
Search for motifs in: **1/p-value_filtered (p-value filtered.txt)**

Choose Motifs to Search
☐ Search for motif:
☒ Search using motifs specified in: **2 (JASPAR.txt.bin)** Search for **All Motifs**
☐ Import motifs from text file: **Browse...**

Sequence Quality Threshold
Sequence Quality >= **0.7**

Result file **MotifSearch** **Browse...**

OK **Cancel**

Figure 16. Configuring a search for known motifs in the JASPAR database

- Select **1/p-value_filtered (p-value filtered.txt)** from the *Choose Region Spreadsheet* drop-down menu
- Select **Search using motifs specified in:** for *Choose Motifs to Search*
- Set *Search using motifs specified in:* to **2 (JASPAR.txt)** using the drop-down menu
- Set *Search for* to **All Motifs** using the drop-down menu
- Set *Sequence Quality* >= to **0.7**
- Name the result file **MotifSearch**
- Select **OK**

Because we are searching for around 1200 motifs, the process will take some time to complete. Progress is displayed in the progress bar in the lower left-hand side of the *Search for Motif(s) in Sequences* dialog (Figure 6).

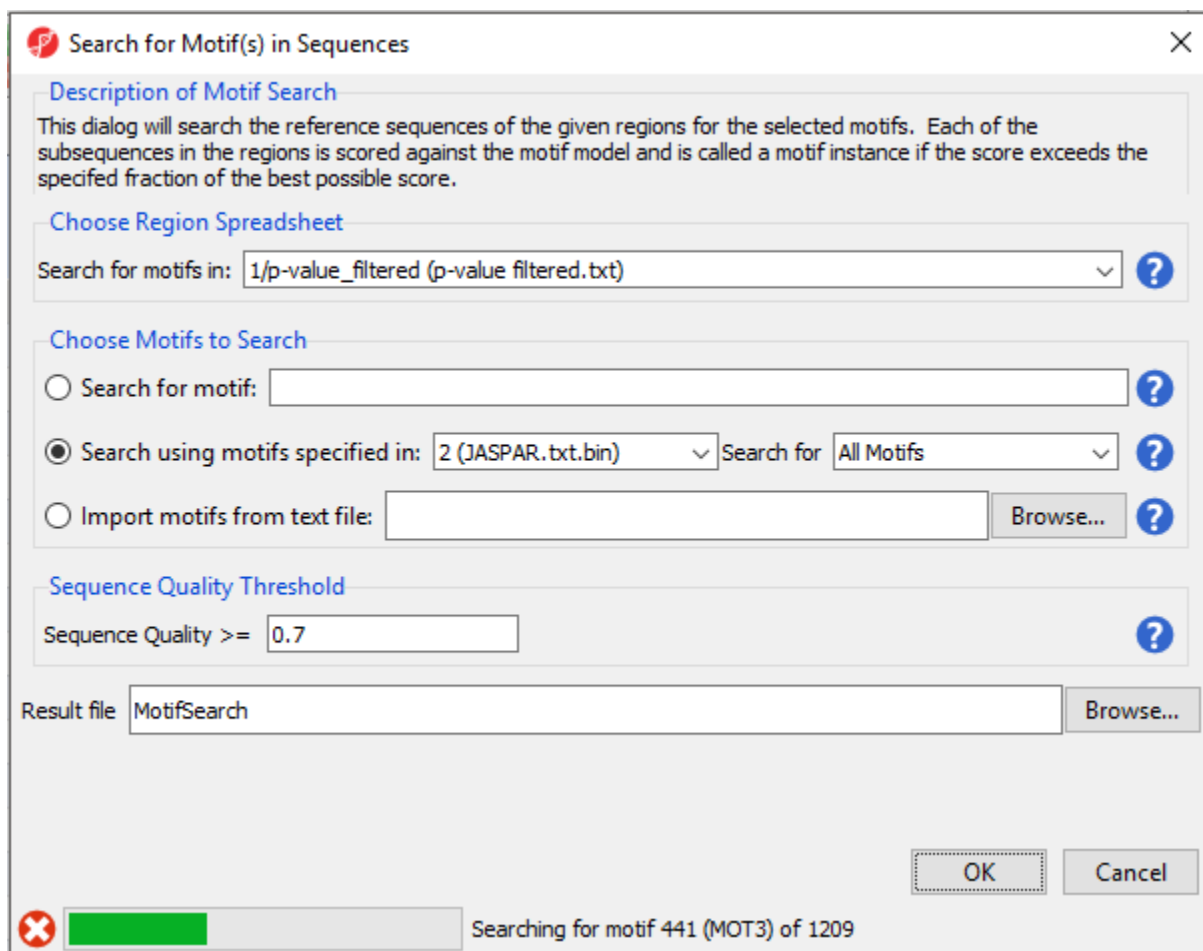


Figure 17. Progress in the motif search will display in the progress bar

Two spreadsheets are created, similar to the spreadsheets in the *de novo* motif discovery, the *motif_summary* (*MotifSearch*) spreadsheet (Figure 7) and the *motif_instances* (*MotifSearch.instance*) spreadsheet.

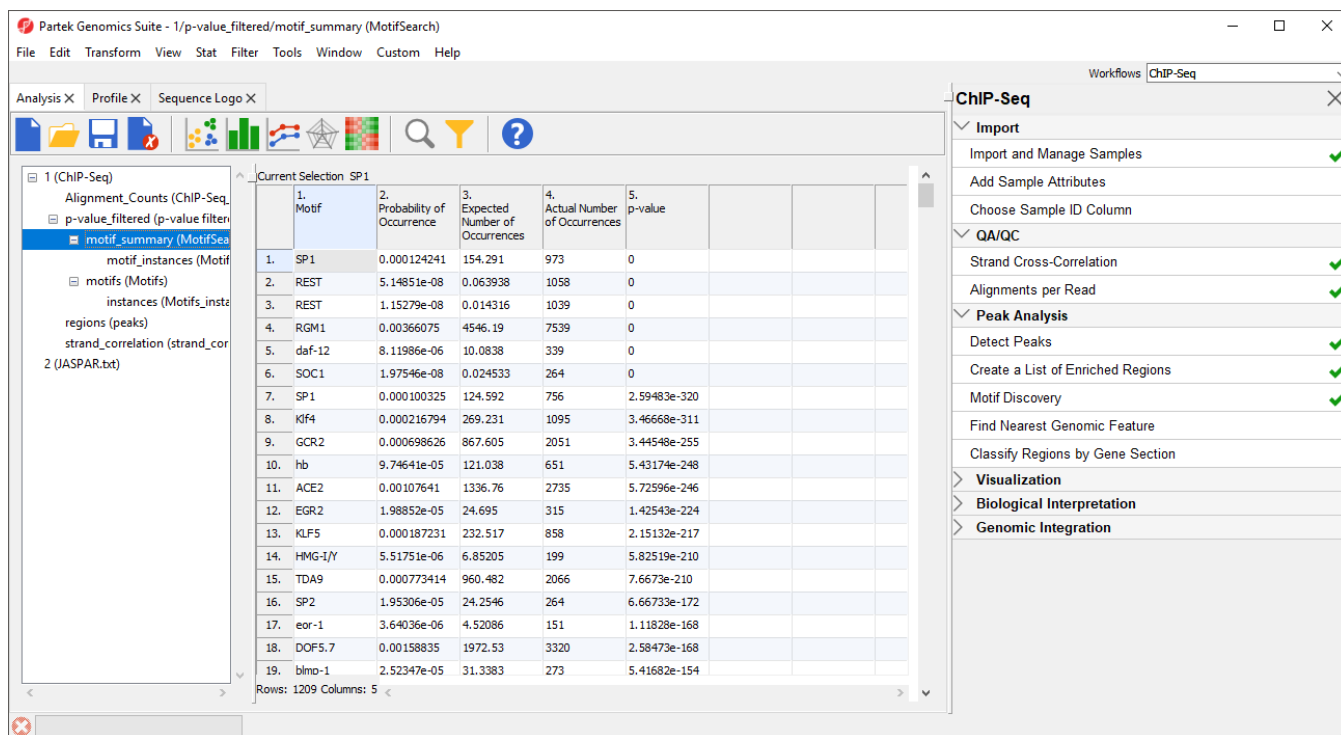


Figure 18. Viewing the results of motif search

In the *MotifSearch* spreadsheet, each motif used in the motif search is shown. The columns detail the results of the search for each motif that was found in the reads.

1. *Motif* this is the name or ID of the motif
2. *Probability of Occurrence* gives the probability of detecting a false positive for this motif in a random DNA sequence
3. *Expected Number of Outcomes* gives the Probability of Occurrence multiple by the summed length of the reads
4. *Actual Number of Occurrences* gives a count of sequences that match the known motif in the reads
5. *p-value* is the uncorrected p-value (binomial test)

As you can see, REST, which is another name for NRSF, is near the top of the list as one of the most significantly over-represented motifs (Figure 7). This motif agrees with the motif found in the *de novo* motif detection step. Interestingly, other motifs appear a significant number of times in the ChIP-Seq peaks and may represent possible co-factors or regulators.

The *motif_instances* spreadsheet contains all instances of the motifs from the *motif_summary* spreadsheet in a format identical to the *instances* spreadsheet from *de novo* motif detection.

Generating a list of regions containing a motif

While the *motif_instances* spreadsheet contains every instance of every motif, it may be useful to create a spreadsheet with just instances of one motif or a select group of motifs. Let's do this for both REST motifs.

- Select the **motif_instances** spreadsheet in the spreadsheet tree
- Right-click the **5. Motif Name** column
- Select **Find / Replace / Select...** from the pop-up menu (Figure 8)

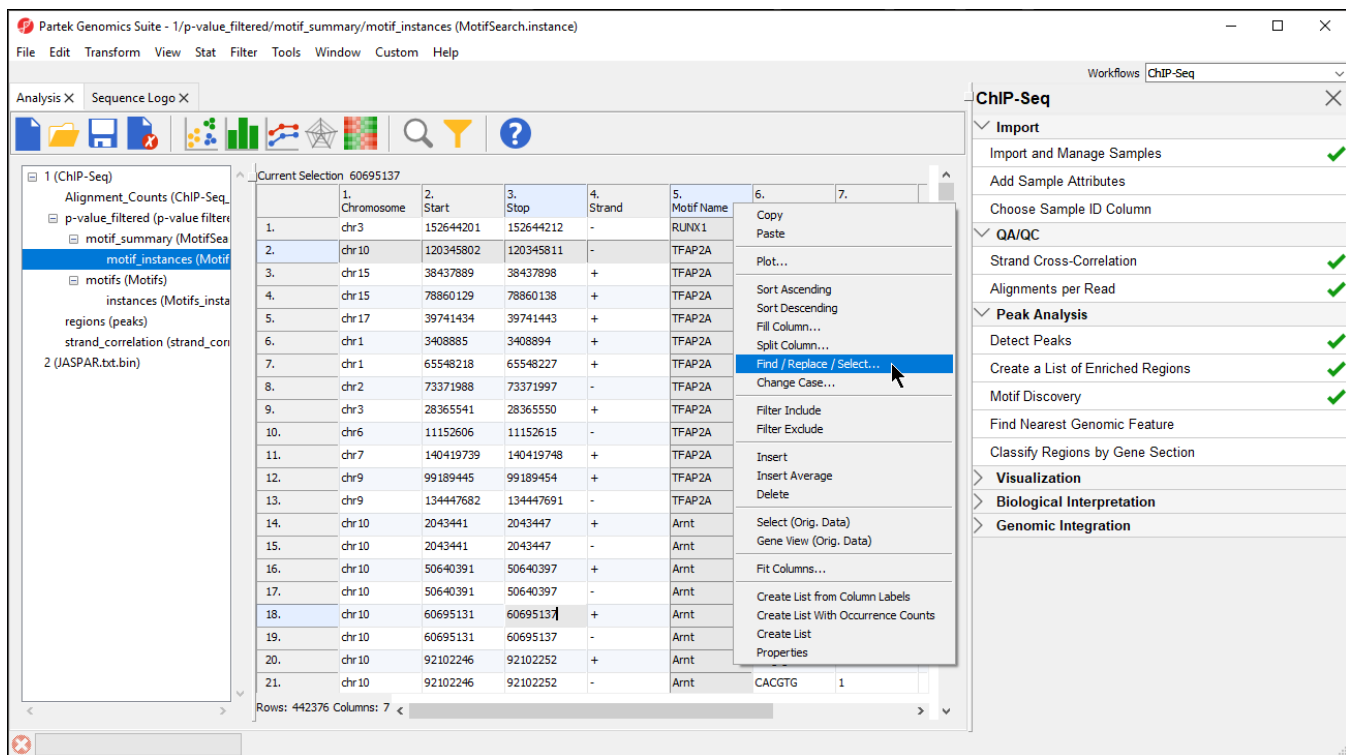


Figure 19. Finding all REST peaks (step 1)

- Set *Find What:* to **REST**
- Select **By Columns** for *Search:*
- Select **Only in column** with **5. Motif Name** selected from the drop-down menu
- Select **Select All** (Figure 9)

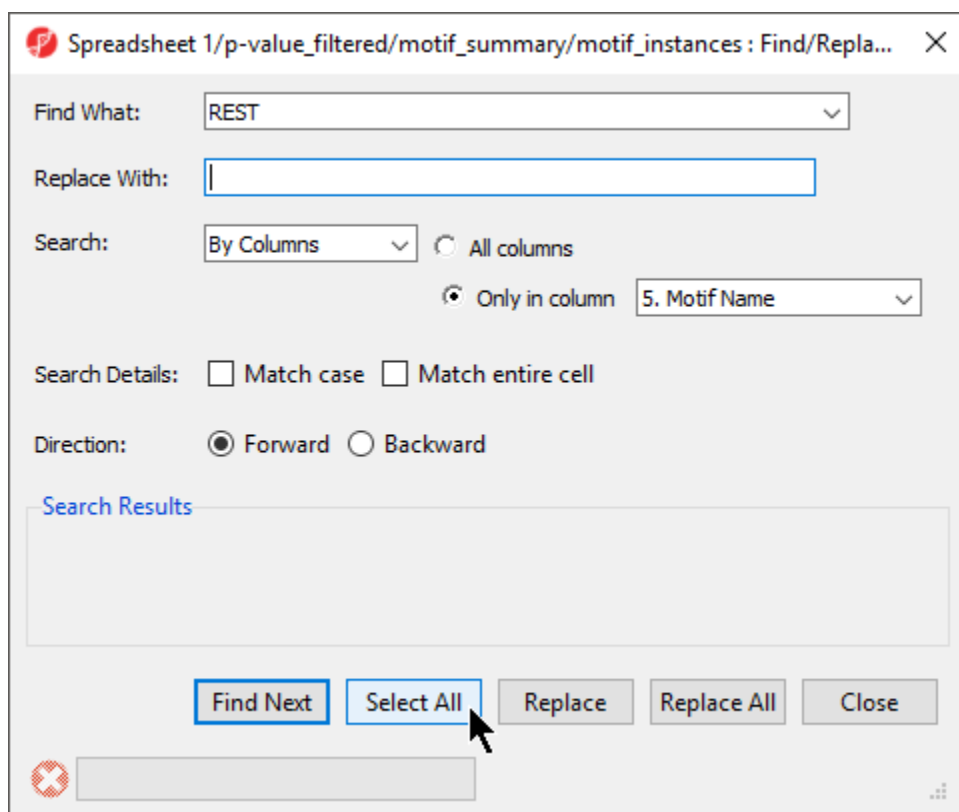


Figure 20. Selecting all REST instances in motif_instances spreadsheet (step 2)

This finds and selects every instance of REST in column 5. Motif Name.

- Select **Close**

In the motif_instances spreadsheet the selected columns are highlighted.

- Right-click on the first highlighted row visible; in this example, we see row 13196

- Select **Filter Include** from the pop-up menu (Figure 10)

The screenshot shows the Partek Genomics Suite interface. The main window displays a spreadsheet titled '1 (ChIP-Seq)' with columns: 1. Chromosome, 2. Start, 3. Stop, 4. Strand, 5. Motif Name, 6. Instance, 7. Quality Score. The spreadsheet contains 442376 rows. A context menu is open over row 13191, with 'Filter Include' selected. The right sidebar shows the 'ChIP-Seq' workflow with steps like 'Import and Manage Samples', 'QA/QC', 'Peak Analysis', 'Visualization', 'Biological Interpretation', and 'Genomic Integration'.

Figure 21. Filtering for selected rows

The spreadsheet will now include 2098 rows and a black and yellow bar will appear on the right-hand side of the spreadsheet (Figure 11). The black and yellow bar is a filter indicator showing the fraction of the spreadsheet currently visible as yellow and the filtered fraction as black.


The screenshot shows the same Partek Genomics Suite interface, but the spreadsheet now contains 2098 rows, all of which are REST motifs. The context menu is no longer visible. The right sidebar remains the same, showing the 'ChIP-Seq' workflow steps.

Figure 22. Filtered motif_instances spreadsheet containing 2098 instances of the REST motifs

To create a spreadsheet that contains only the REST instances, we can clone the *motif_instances* spreadsheet while the filter is applied.

- Right-click on *motif_instances* in the spreadsheet navigator
- Select **Clone...** from the pop-up menu
- Set the *Name of resulting copy* as **REST**
- Select **1/p-value_filtered/motif_summary (MotifSearch)** from the *Create as a child of spreadsheet* drop-down menu
- Select **OK**

This creates a temporary spreadsheet *rest* from the filtered *motif_instances* spreadsheet. We can now save the new spreadsheet.

- Select **rest** from the spreadsheet tree
- Select  from the command bar
- Name the file **REST**
- Select **Save**

We can now remove the filter from the source *motif_instances* spreadsheet.

- Select **motif_instances** from the spreadsheet tree
- Right-click the filter bar
- Select **Clear Filter**

References

Neuwald, A. F., Liu, J.S., & Lawrence, C.E. (1995). Gibbs motif sampling: detection of outer membrane repeats (Vol. 4). Protein Science.

[« Creating a list of enriched regions](#) [Finding nearest genomic features »](#)

Additional Assistance

If you need additional assistance, please visit [our support page](#) to submit a help ticket or find phone numbers for regional support.



✶

Your Rating:  Results:  36 rates