# Motif Detection



Motif detection identifies enriched sequence motifs in peak regions generated by ChIP-Seq and ATAC-Seq data. Partek Flow includes *Search for known motifs*, which allows users to search for known motifs from a user-specified set or a database, and *Detect de novo motifs*, which can identify novel motifs. These tasks can be invoked on data nodes with genomic regions as features (not genes or transcripts).

## Search for known motifs

Given a set of genomic regions, *Search for known motifs* can search for enrichment based on a sequence provided by user or using a sequence database like JASPAR.

- Click on a Peaks data node
- Click **Motif detection** in the toolbox
- Click **Search for known motifs**

The configuration dialog offers two search methods, By sequence and By database.

### By sequence

By sequence has two options (Figure 1):

Manually: the user can manually specify the sequences. Multiple sequences can be added by clicking the + button

From file: use can specify a text file (.txt) that contains a list of sequences, one row per sequence.

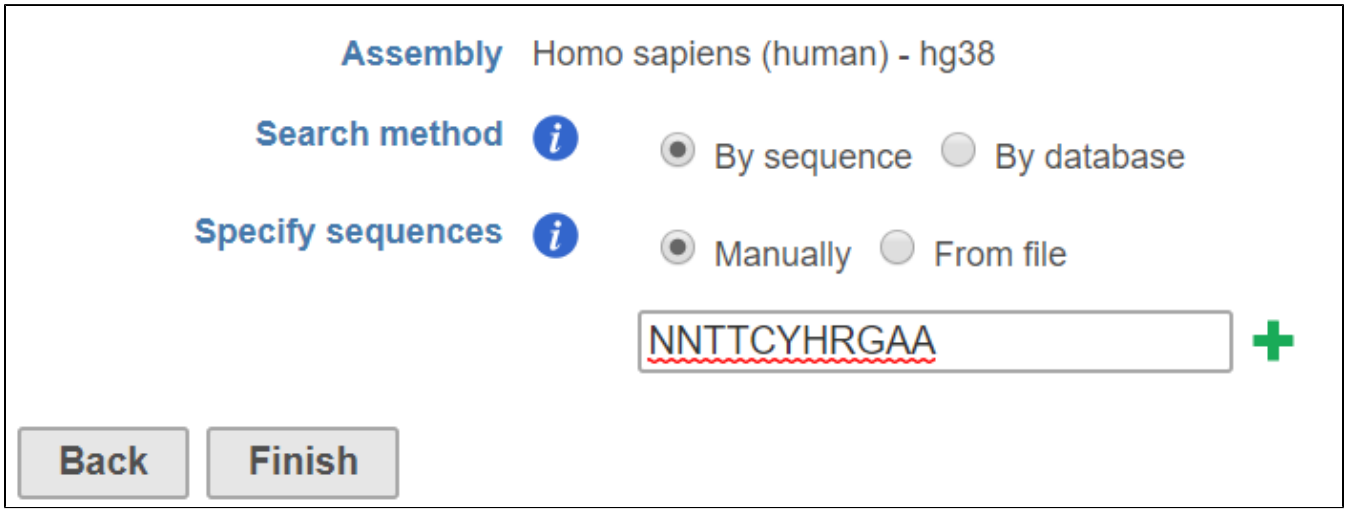



*Figure 7. Search for known motifs by specifying a sequence manually*

The *By sequence* option uses a string search tool to return all positions in the set of genomic regions that match the given string(s). The string match is case insensitive, meaning if you search for ATCG, you may get atcG as a match. Nucleotides that are lower case have been "repeat masked", meaning they are located in a repetitive region of the genome. Your search string may contain any of the characters from the IUPAC nucleotide code . For example, if you search for WAAA, you may get back AAAA or TAAA (or any variation of upper and lower cases), because W represents A or T.

### By database

The *By database* option uses an alignment matrix to match sequences against a motif database. We distribute the JASPAR database, but you can add any custom or public motif database.

Matching is determined using an alignment matrix. Alignment matrices are often used in literature to model transcription factor binding sites, alignment matrices are matrices of nucleotide counts per position [1]. Each instance of the motif is aligned to each other and the number of nucleotides at each position is counted and summarized in an alignment matrix. All positions from the set of genomic regions are scored against the alignment matrix. The score represents how likely the position is an instance of the motif. A quality cutoff is used to determine which sequences in the regions are instances of the motif. The scoring scheme and quality cutoff are similar to [2] and it briefly described below:

Let M be a motif of length L consisting of N motif instances.Let A be a 4XL alignment matrix such that $a_{i,j}$ is the count of letter i at position j. Let $B_i$ be the background frequency of letter i (calculated as the number of nucleotides i in the regions divided by the total oligonucleotides in the regions). Let S be a sequence of length L. The score of S given the alignment matrix is

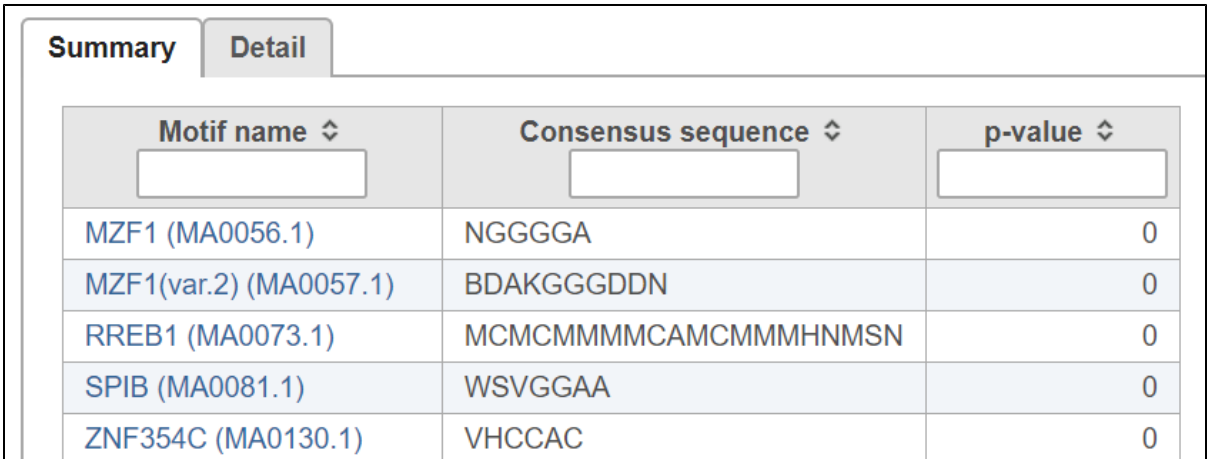
$$L_A(S) = \sum_j \left[ \ln\left(\frac{a_{S(j),j} + b_{S(j)}}{N+1}\right) - \ln(b_{S(j)}) \right]$$

Equation 1

Let *h* be the maximum of $L_A$. The quality score of a sequence is calculated as $Q_A(S) = L_A(S)/h$. A quality score of 1 corresponds to a sequence with the most likely base at each position of the alignment matrix. The user will specify a threshold $Q_A$. All sequences that have a score $T_A > Q_A$ *h will be reported.

### Task report

The *Search for known motifs* task report contains summary and detail tabs.



Summary | Detail

| Motif name | Consensus sequence | p-value |
|---|---|---|
| MZF1 (MA0056.1) | NGGGGA | 0 |
| MZF1(var.2) (MA0057.1) | BDAKGGGDDN | 0 |
| RREB1 (MA0073.1) | MCMCMMMMCAMCMMMHNMSN | 0 |
| SPIB (MA0081.1) | WSVGGAA | 0 |
| ZNF354C (MA0130.1) | VHCCAC | 0 |

| | | |
|---|---|---|
| EWSR1-FLI1 (MA0149.1) | GGAAGGAAGGAAGGAAGG | 0 |
| Pparg::Rxra (MA0065.2) | NNRGRNSARRGGBBA | 0 |
| NFATC2 (MA0152.1) | TTTTCCH | 0 |
| NR4A2 (MA0160.1) | RAGRHCAV | 0 |
| Stat5a::Stat5b (MA0519.1) | NNTTCYHRGAA | 0 |
| Stat4 (MA0518.1) | BTKMHRRGAAVNNN | 0 |
| SP2 (MA0516.1) | NBYCCDCCYHYNNNN | 0 |
| NRF1 (MA0506.1) | GCGCVTGCGCV | 0 |
| Nr5a2 (MA0505.1) | NNNBYCAAGGHCANN | 0 |
| Nkx2-5(var.2) (MA0503.1) | NDSCWCTCARV | 0 |
| Myog (MA0500.1) | VRCAGSTGNNN | 0 |
| Klf1 (MA0493.1) | NRCCMCRCCCW | 0 |
| HNF4G (MA0484.1) | NNRGDNCARAGKBCW | 0 |
| E2F6 (MA0471.1) | DVGMGGGARVN | 0 |
| Tcf12 (MA0521.1) | VRCAGCTGNNN | 0 |
| TCF7L2 (MA0523.1) | NNDSWTSAAAGVNN | 0 |
| ZNF263 (MA0528.1) | RRRGGAGGRNDRDVDDRRRRR | 0 |
| SP1 (MA0079.3) | NYYCCDCCYHY | 0 |
| STAT1 (MA0137.3) | NTTCYRGGAAN | 0 |
| STAT3 (MA0144.2) | NTKCYDGGAAD | 0 |

Rows per page 25 (1 of 24)

*Figure 8. Search for known motif task report summary tab*

In the summary tab (Figure 2), each row is a motif in the search database specified. Clicking on the motif name opens the JASPAR database page with detailed information about the motif.

The p-value indicates whether instances of the motif are enriched in the input regions.

The p-value is calculated as follows:

The probability $P_{Expected}$ of a sequence having a score above $T_A$ is calculated under the assumption that the base are i.i.d. according to the background distribution B. Let $N_{Trials}$ be the number of sequences compared to the alignment matrix. The expected number of occurrences of the motif in the regions is $P_{Expected} * N_{Trials}$. The p-value of observing $N_{Actual}$ instances with a score above $T_A$ is calculated based on the binomial distribution, where $NT_{rials}$ is the number of trials and $P_{Expected}$ is the probability of success.

The detail tab lists motif sequence locations on rows and includes the quality score for each instance (Figure 3).

Summary | Detail

| Chromosome | Start | End | Strand | Motif ID | Instance sequence | Score |
|---|---|---|---|---|---|---|
| 22 | 49,097,099 | 49,097,107 | + | Smad4 (MA1153.1) | TGTCTAGA | 1.00 |
| 22 | 44,577,055 | 44,577,063 | - | Smad4 (MA1153.1) | TGTCTAGA | 1.00 |
| 22 | 36,234,524 | 36,234,532 | - | Smad4 (MA1153.1) | TGTCTAGA | 1.00 |
| 22 | 29,694,970 | 29,694,978 | + | Smad4 (MA1153.1) | TGTCTAGA | 1.00 |
| 22 | 24,087,618 | 24,087,626 | + | Smad4 (MA1153.1) | TGTCTAGA | 1.00 |
| 22 | 49,385,110 | 49,385,118 | + | VDR (MA0693.2) | TGAGTTCA | 1.00 |
| 22 | 46,222,348 | 46,222,356 | - | VDR (MA0693.2) | TGAGTTCA | 1.00 |
| 22 | 42,401,436 | 42,401,444 | - | VDR (MA0693.2) | TGAGTTCA | 1.00 |
| 22 | 42,071,164 | 42,071,172 | - | VDR (MA0693.2) | TGAGTTCA | 1.00 |
| 22 | 37,754,380 | 37,754,388 | + | VDR (MA0693.2) | TGAGTTCA | 1.00 |

Rows per page 10 (1 of 3539)

*Figure 9. Search for known motifs task report detail tab*

## Detect de novo motifs

*Detect de novo motifs* can be used to identify novel motifs that are enriched in the input regions.

- Click on a Peaks data node
- Click **Motif detection** in the toolbox
- Click **Detect de novo motifs**
- Specify the number of motifs to report (Figure 4)
- Specify the range of motif length to search

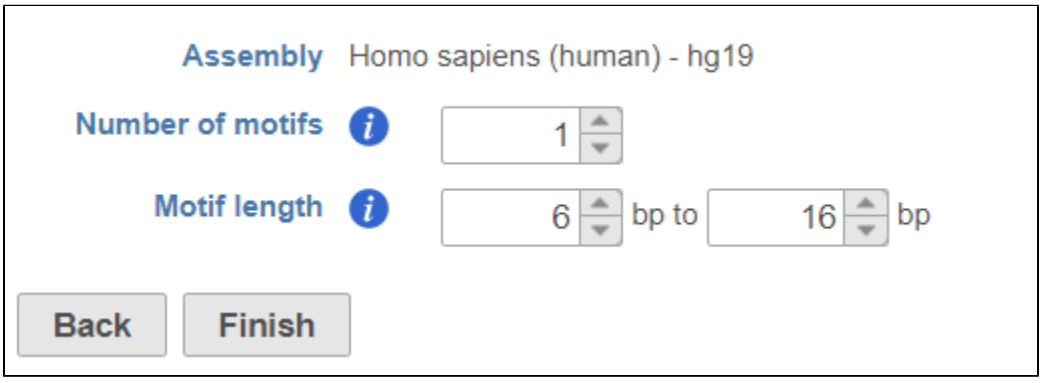



*Figure 10. Detect novo motifs configuration dialog*

Motif discovery uses Gibbs motif sampling. The implementation of the Gibbs motif sampling in Partek Flow is based on Neuwal, et all [3]. The Gibbs sampling method is a stochastic procedure that attempts to find the subset of sequences within the regions that maximizes the log likelihood ratio (LLR)

$$LLR = \sum_i \sum_j \left[ \ln\left(\frac{a_{i,j} + b_i}{N+1}\right) - \ln(b_i) \right]$$

Equation 2

This is done by repeating the following two steps until convergence:

A. Given the alignment matrix from step B, search for location in the set of regions that score highly compared to the alignment matrix using Equation 1

B. Create a new alignment matrix from the set of high scoring positions from step A and return to step A.

The Gibbs sampler is run on a range of the motif sizes specified by the user. The motif with the greatest average LLR (LLR /length) is returned. To find N motifs in a set of regions, the Gibbs sampling method is run N times. The motif instances found from the previous run of the Gibbs sampler are removed before performing the next run.

### Task report

The task report includes summary and detail tabs.

The summary tab gives the consensus sequence and sequence logo for each detected motif (Figure 5).

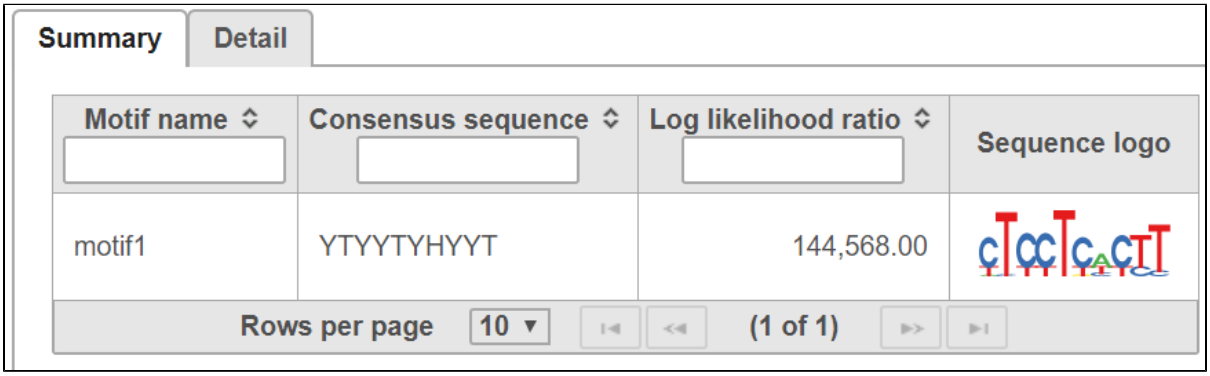



*Figure 11. Detect de novo motifs task report summary tab*

The detail tab is similar to the detail tab in *Search for known motifs* and lists the location of each motif sequence (Figure 6).

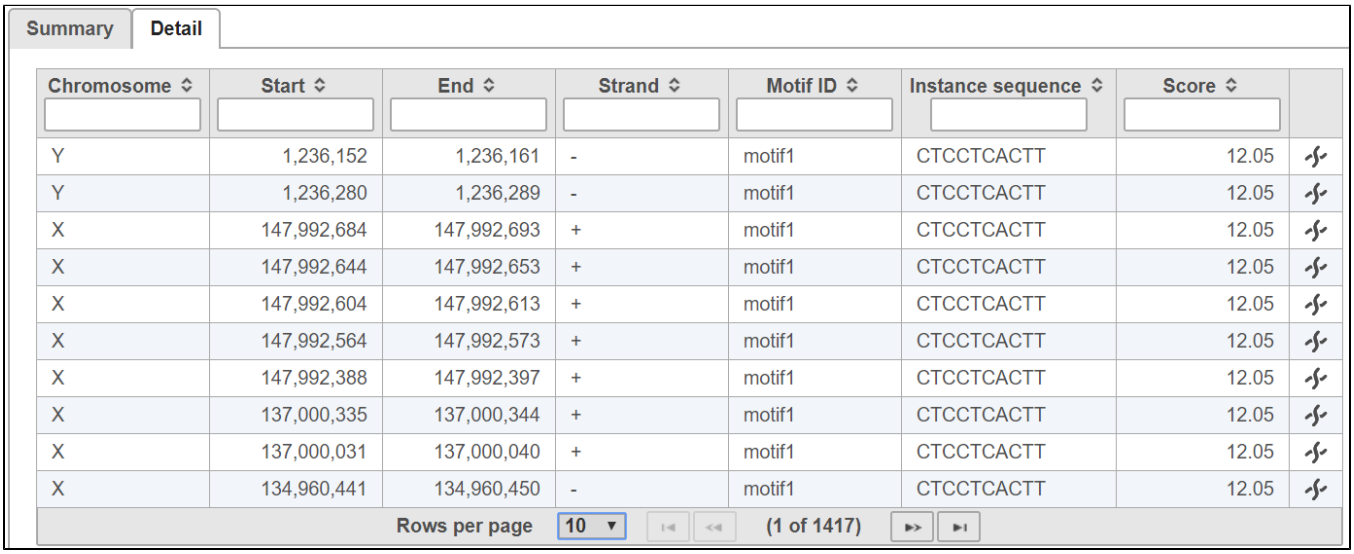



*Figure 12. Detect de novo motifs task report detail tab*

## References


1. Hertz, GZ., & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999 , 15, 563-577
2. Schug, J., & Overton, C.G. TESS, Transcription Element Search Software on WWW. Technical Report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, 1997
3. Neuwald, A.F., Liu, J.S., & Lawrence, C.E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science* 1995, 4: 1618-1632.