# m6aViewer Version 1.5.0

# Documentation

**Contents**

# 1. About

m6aViewer is a java application for m$^6$A peak detection and visualisation from m$^6$A-seq/Me-RIP sequencing data. m6aViewer is developed by Agne Antanaviciute, currently based at University of Leeds. For any queries, please email: umaan@leeds.ac.uk

# 2. Requirements

m6aViewer is a cross-platform application and only requires Java 1.7+ Runtime Environment which may be downloaded from the Oracle website (http://www.oracle.com/). m6aViewer can be downloaded as an executable from http://dna2.leeds.ac.uk/m6a/.

# 3. Launching m6aViewer

### Default launch:

m6aViewer can be launched directly from the executable .jar file. By default, the executable will attempt to start with the following JVM settings:

For 32bit systems, JVM will attempt to reserve 1.5G RAM

For 64bit systems, JVM will attempt to reserve 2G RAM

In order for this to work, **Java must be in your system path**.

### Custom JVM launch:

If you are running a non-standard Java installation, have less than default amount of usable RAM, want to allocate additional RAM or set custom JVM parameters, m6aViewer can be launched from command line with '—skipautoheap' parameter. This will use custom JVM settings. Example command:

java –Xmx4G –jar m6aViewer1.5.0.jar --skipautoheap

# 4. Running Time Estimates

Running times are estimated using a machine with 16GB RAM, i7 CPU @ 3.4GHz, with BAM files stored on SSD. Computation times will vary on different machines.

A single sample of ~30 million 36-base pair single-end reads will require approximately 20 minutes processing time using the default running mode.

Peak de-convolution running mode (which can be selected from the settings menu) is slower, requiring approximately twice as long to complete.

Additional samples will increase the running time linearly.

Peak calling speed can be increased substantially by:

a) Increasing block size parameter under **Settings -> Other -> Block Size.** This will increase RAM use, make sure enough is made available for JVM! This may also require restarting the application from command line to allocate additional heap space for the JVM.

b) Enabling multi-threading under **Settings -> Other -> Number of Parallel Processing Threads to use**

The default settings are conservative for current hardware.

## 5. Basic Peak Calling

Basic peak calling requires:

1. Sorted and indexed BAM from sample IP. Index must located in the same directory as the BAM file

2. Sorted and indexed BAM from sample INPUT control (RNA-Seq). Index must be located in the same directory as the BAM file

For SAM to BAM file conversions, sorting and indexing, please see samtools (http://www.htslib.org/). For users unfamiliar with command line use, Galaxy (https://usegalaxy.org/) may be used for alignment, BAM file sorting and indexing steps through a web-based interface.

If the average sequenced fragment length differs from our default (~100bp), set the expected peak width parameter in the Settings -> Peak Calling -> Expected Peak Width to twice the average sequenced fragment length. This setting refers to RNA fragment length, NOT sequencing read length – these may differ substantially.

Select matched IP and INPUT indexed BAM files and click 'ADD SAMPLE'. Repeat for multiple samples if needed.

Click 'Find Peaks' at the bottom to start peak detection process. A progress indicator will appear. Peak calling may take some time, depending on the number of samples and total number of reads per sample.

## 6. Running Modes

m6aViewer has two running modes – peak detection mode (default) and peak de-convolution mode.

**Peak detection mode:** the default setting. Peaks are detected by smoothing and scanning the sample IP coverage and detecting all local maxima in the distribution. These are then tested (Fisher's Exact) against the background INPUT for significance and adjusted for multiple testing. Peak enrichment is calculated as a ratio between (normalised) IP and INPUT coverage. Detected peak positions thus correspond to the 'summits' in the coverage distribution.

**Peak de-convolution mode**: can be selected using 'Settings'->'Peak Calling' menu. THIS REQUIRES REFERENCE SEQUENCES IN FASTA FORMAT AND GTF ANNOTATION FILE TO RUN. If these are not provided, the program will automatically run in default mode. A slower peak calling mode that attempts to determine the exact $m^6A$ positions within an enriched region using a likelihood maximisation approach. Each sequenced fragment is modelled as a data point arising from each putative $m^6A$ site and Expectation Maximisation is applied to estimate the likeliest model. See section below for further details.

## 7. Multiple Samples\Sample Replicates

Multiple samples can be selected through the user interface. Samples may be grouped to indicate biological replicates. m6aViewer provides support for basic replicate-based filtering. The options (accessible under Settings->Sample Replicates) available are:

**No Filtering** – all peaks in all samples are kept

**>50%** - peaks are kept only if they are detected in more than half the samples within a replicate group at a corresponding position

**100%** - peaks are kept only if they are detected in all the samples within a replicate group at a corresponding position.

As detected peak positions across samples may not correspond perfectly, a maximum allowed distance overlap (default: 100bp) can be specified.

### 7.1 Differential Methylation

If multiple samples are present, saving peaks to text file will produced an additional tab-delimited text file containing all sample pair-wise differential methylation combinations for detected peaks. The differential methylation is calculated as an enrichment log fold change between two samples at a given position. Fisher's Exact test together with multiple testing correction is used to obtain a significance p-value. This takes into account the gene expression changes, however is not sensitive to cases where methylation is not present in one sample due to gene expression falling so low it is undetectable as there is no reliable way to detect whether methylation is present.

Individual peak information is included in a tab-delimited file to easily allow alternative and/or more sophisticated differential methylation calculations in downstream analysis. All positions, regardless of significance, are provided in the output.

## 8. Gene Functional Enrichment Analysis

m6aViewer provides in-built gene functional enrichment analysis (Gene Ontology terms, Reactome pathways) for detected peaks. This feature requires internet connection to work (this part of the application depends on a successful connection to Ensembl servers) and is accessible via 'Functions' button in the coverage browser interface (accessible from 'Peak Browser' button from main interface after peak-calling has been performed).

Functional enrichment analysis is performed on gene level. In each sample, we test individual Gene Ontology/Reactome terms for enrichment in the set of all transcripts which have been methylated (irrespective of a number of peaks) against the background of all expressed transcripts.

## 9. False Positive Peak Filter

Peaks in m6a-seq coverage data can often be a product of noise, such as DNA contamination, poor alignments or non-specific anti-body binding. These can represent a substantial proportion of all detected peaks. m6aViewer provides an optional false positive peak filter that aims to filter out these events while retaining as high proportion of true positive sites as possible.

This is achieved via a supervised learning model (a combination of Random Forest and an RNA sequenced-based Mixture Transition Distribution model) trained on data from RNA methyltransferase complex knockdown experiments together with matched controls. Peaks which exhibited little change in enrichment even under methyltransferase knockdown were considered false positive training examples, peaks which were no longer present under methyltransferase knockdown were considered true positive training examples.

Using default settings, an estimated 86% of false positive peaks can be filtered out at the loss of 9% of true positives. This trade-off can be set to custom precision/recall cut-offs using a slider from 'Settings->False Positive Peak Filter'.

By default, no filtering is performed. There are two settings that can be enabled under the settings menu – annotation only and annotation and filtering. Selecting annotation only will assess each peak and annotate it with a likelihood score of it being a genuine m6a site (0-1, with 0.5 confidence level used as a default cut-off). Selecting filtering will additionally remove peaks below the required confidence threshold.

## 10. Saving/Loading Sessions

To view peaks in the future without having to re-run peak detection (which can take some time), a session file (File -> Save Session) can be created. This creates a binary .m6aSession file that stores detected peak information.

When loading a session, all files (i.e. BAMs) that were used to create a session must be in exactly the same directory, or loading will fail (i.e., you cannot currently load a session file created on a different computer).

# 11. User Interface

## 11.1 Main Application Window

1. Contains various options and settings, such as saving peaks to file and saving and loading sessions and modifying peak detection settings. Save as file will produce an individual text file for each sample (your chosen file name will be appended with sample name), and if more than one sample was analysed, a file which contains differential methylation information.

2. Import two matched BAM files for each sample, an IP and INPUT, and add to sample list. Files must be sorted and indexed and indexes located in the same directory. (To sort and index BAM files, use a program such as samtools with sort and index commands).

3. Current sample selection

4. Import a GTF file containing annotations. This is optional, but will generate a gene track and annotate transcript to peaks.

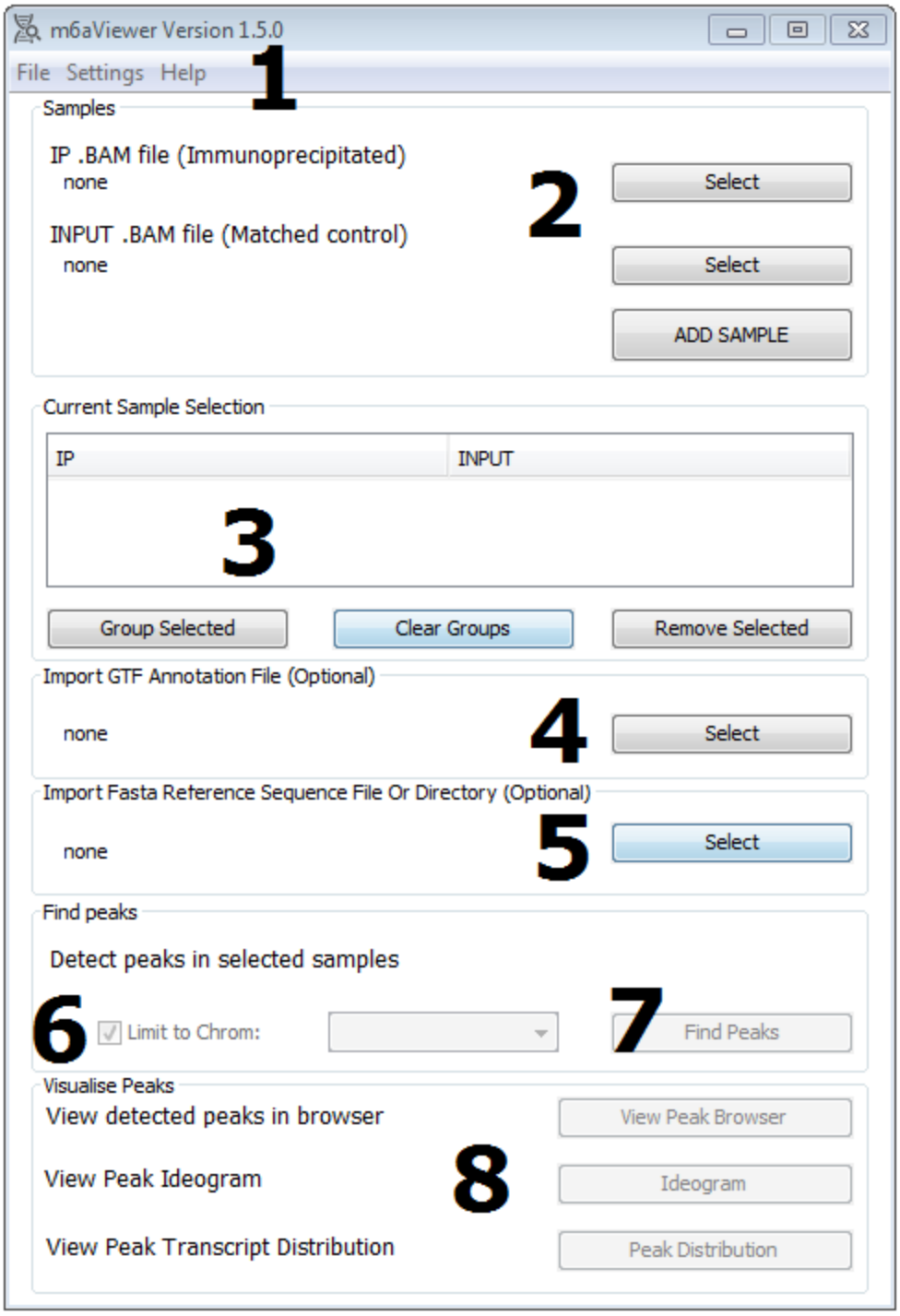
5. Import either :

    - A FOLDER containing reference sequences in fasta format. One .fa or .fasta file per sequence.

    - A single .fa or .fasta file containing sequence of one chromosome

    - A single .fa or .fasta file containing multiple chromosome reference sequences. Will be indexed running run time for faster access.

    Optional, but allows to search for sequence motifs.




6. Limit peak search to a single chromosome at a time if you want to view results faster (takes a minute or so for each chromosome, but this adds up when looking at them all), otherwise untick and perform peak search for all chromosomes.

7. Find peaks.

8. Visualisation options.

View Peak Browser:

Opens the peak viewer for detected peaks. Loading up initial data may take a while, as the viewer needs to read indexes for each BAM file. Pan, zoom, quickly jump between peaks.

Ideogram:

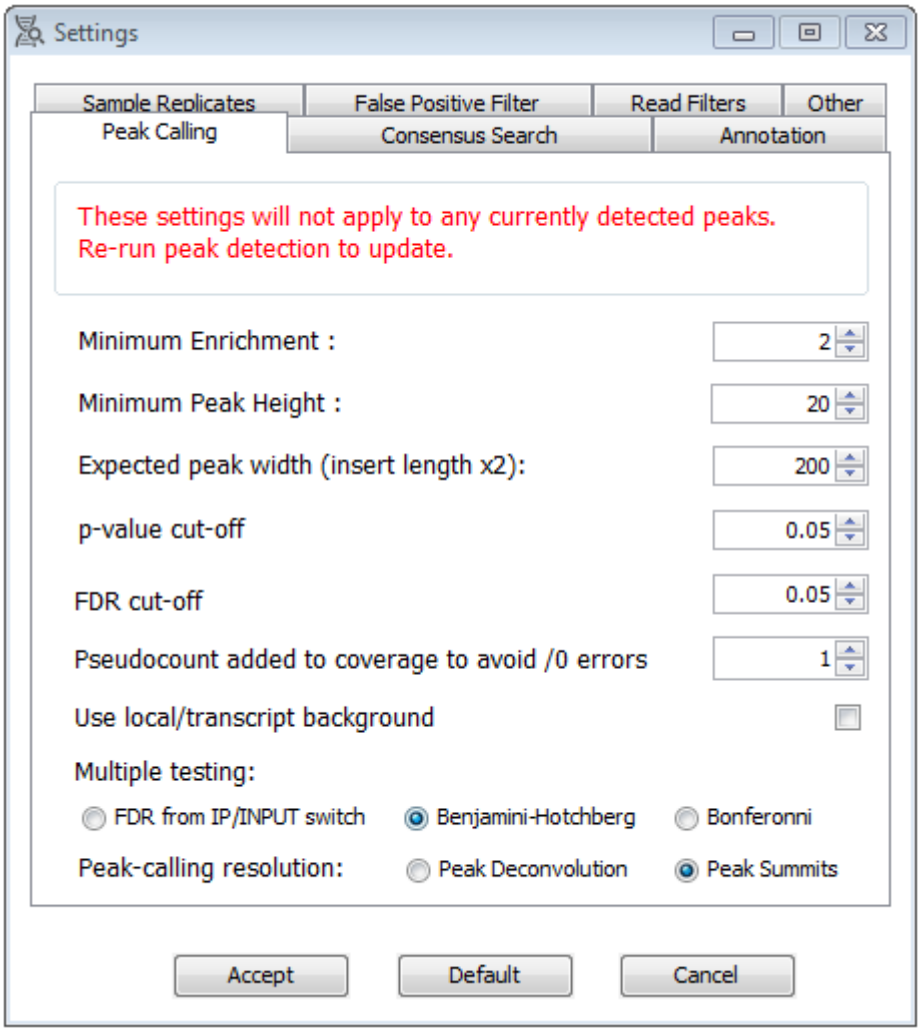
Interactive, zoomable ideogram showing all peaks for a given chromosome. For non-human data, please supply custom cytoband data under Settings->Annotation-> Ideogram

Peak Distribution:

Provides summary data, peak distribution ( % in CDS, introns, UTR,) data and if multiple samples are present, sample-to-sample heat chart.

# 112 Settings

## 11.2.1 Peak Calling



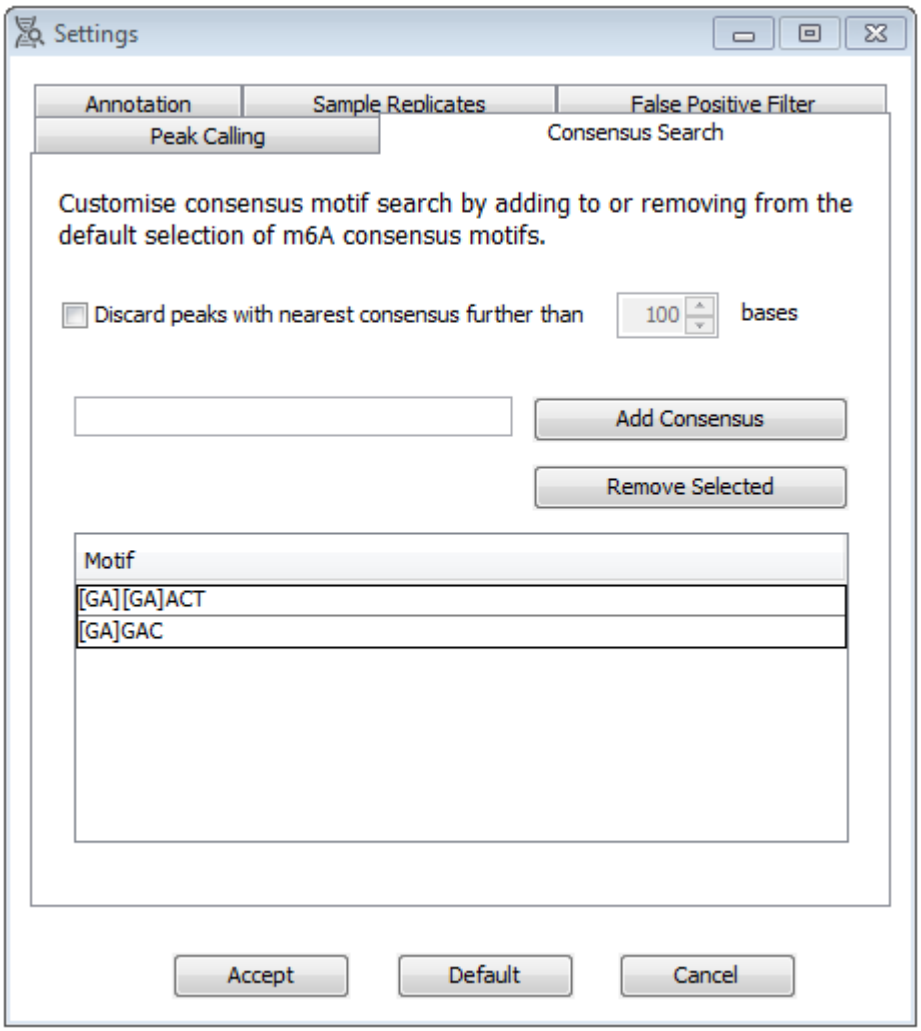
Minimum enrichment and minimum peak height values increase stringency of peak detection. These are an absolute cut-off and ensure peaks are not called from very low coverage regions.

Expected peak width refers to the twice the average sequenced fragment (insert) length. This refers to the RNA fragment size, rather than sequencing read length, which may differ. In the case of single-end reads, reads shorter than fragment length will be extended to avoid a shift.

Use local/transcript background uses the number of reads in the transcript as background counts for Fisher's Exact test.

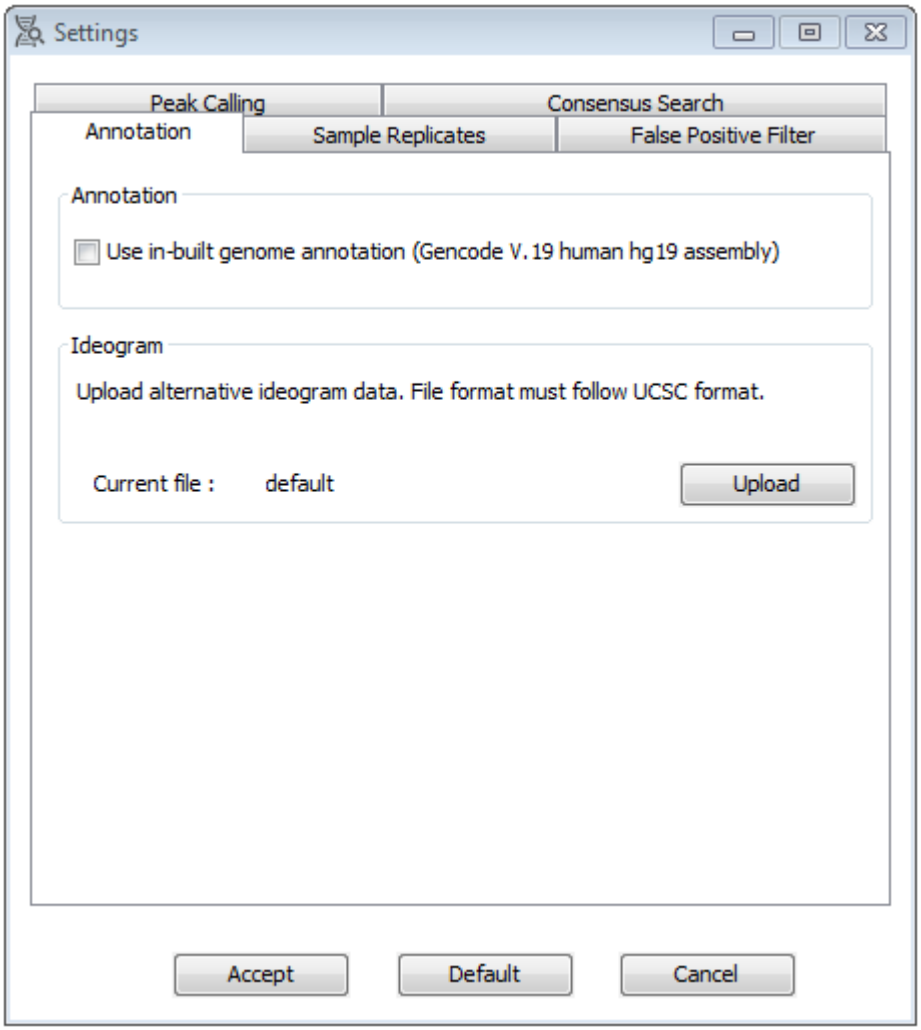
Selecting 'Peak Deconvolution' runs peak calling in a slower alternative mode, which attempts to deconvolute multiple peaks in a region.

### 11.2.2 Consensus Search




A list of expected m6A consensus motifs that is used for annotation and filtering.

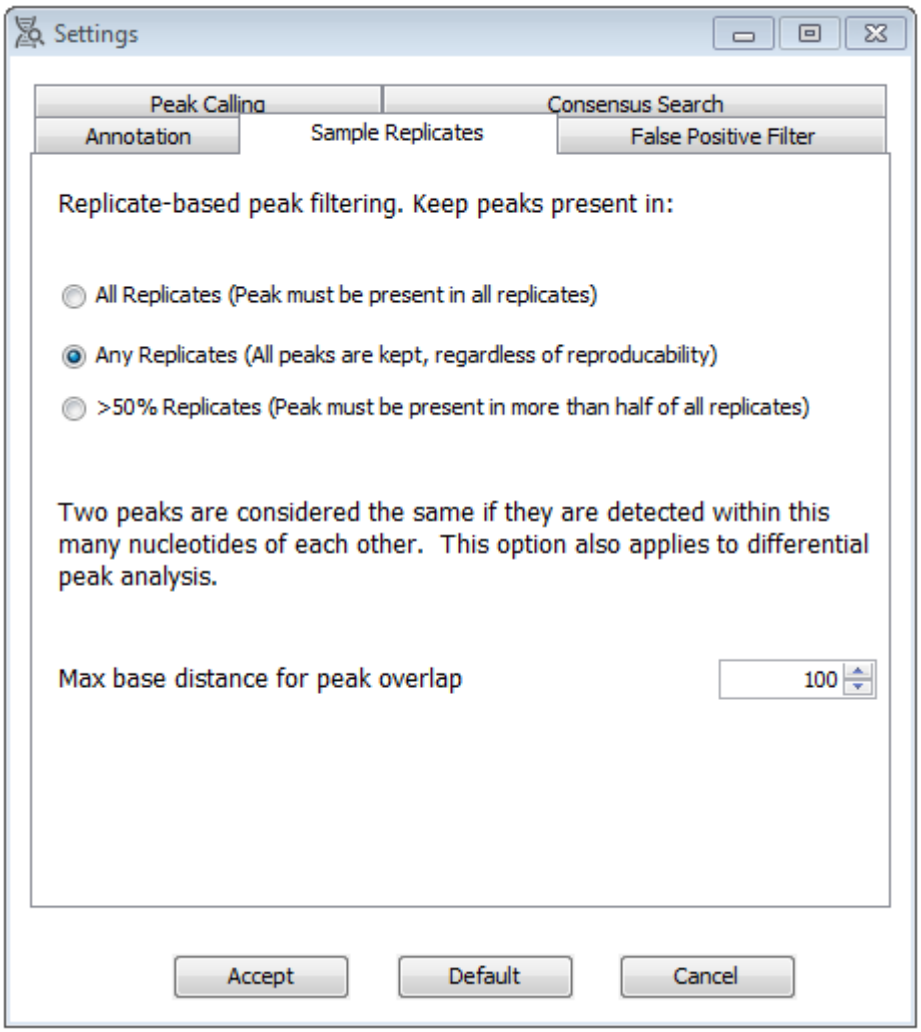
### 11.2.3 Annotation




Allows using in-built gene annotations instead of uploading a GTF file. The data is from human hg19 gencode annotation, and therefore can only be used with BAM files aligned to hg19 reference genome.

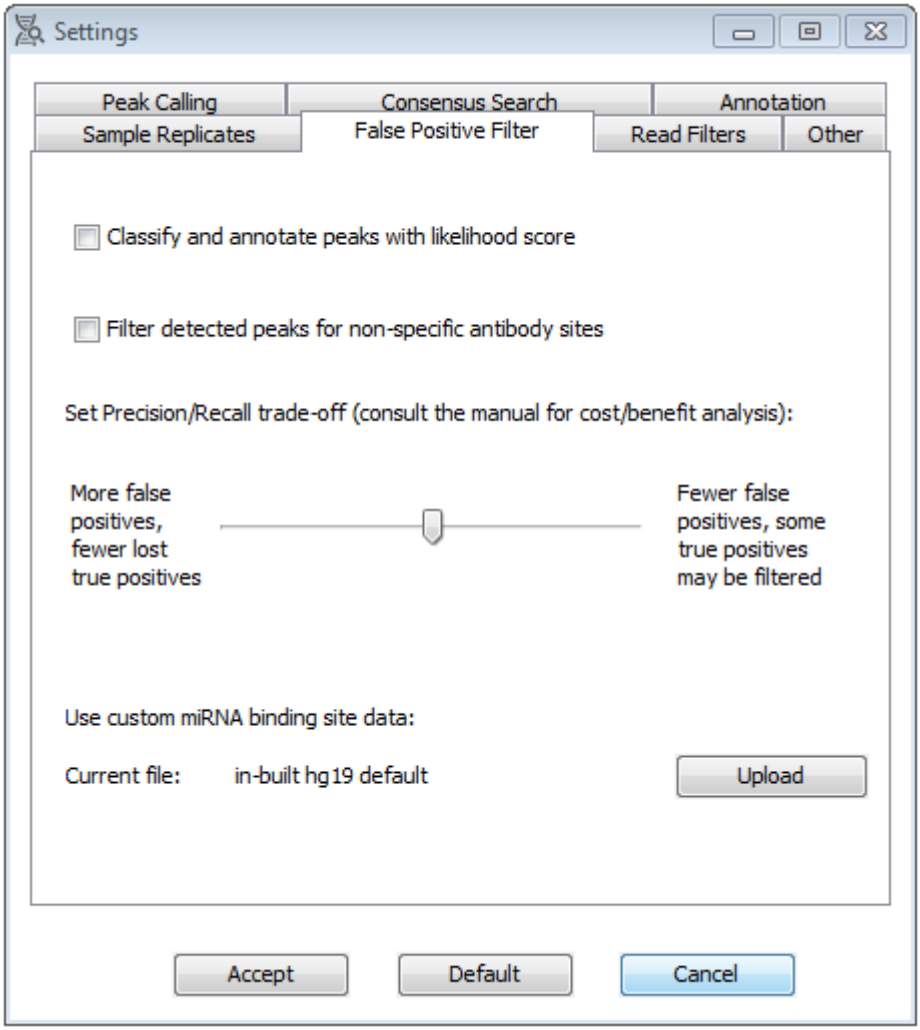
Allows the user to upload alternative ideogram information. In-built ideogram is drawn for human data only.

### 11.2.4 Sample Replicates



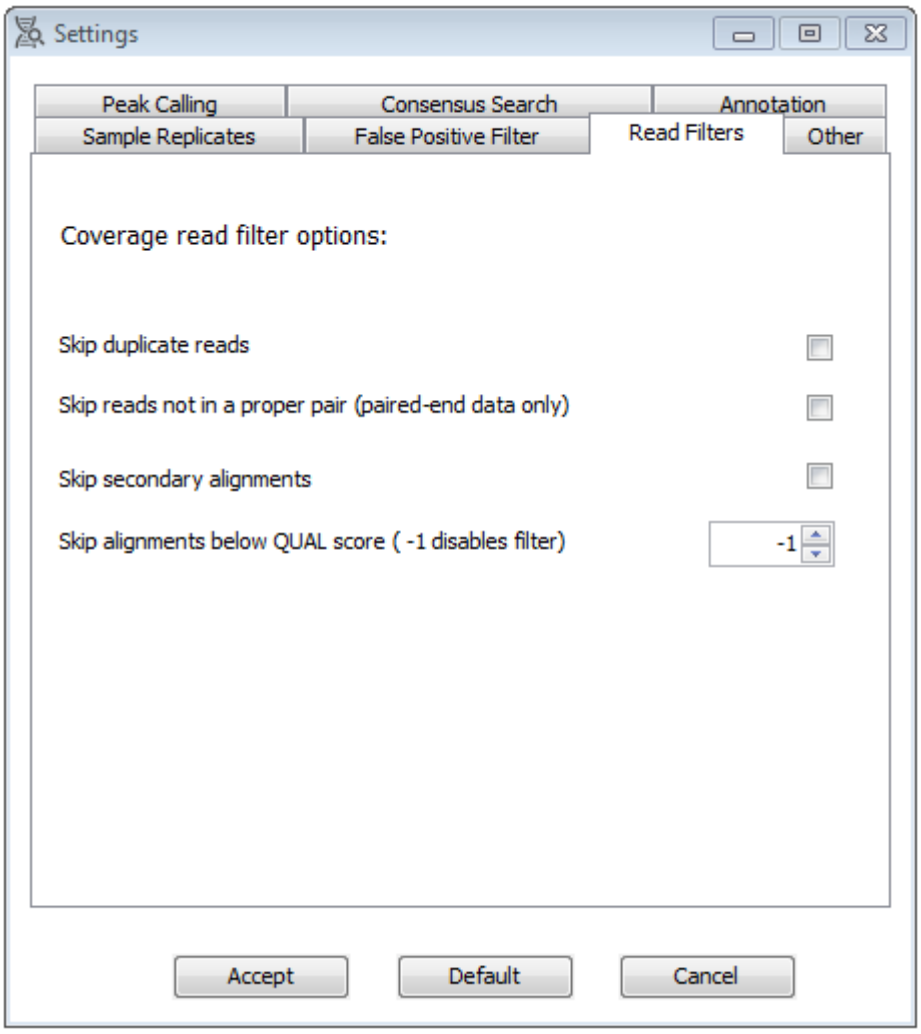

Allows filtering out peaks based on reproducibility. Please see section **6. Multiple Samples/Sample Replicates** for more details.

### 11.2.5 False Positive Filter



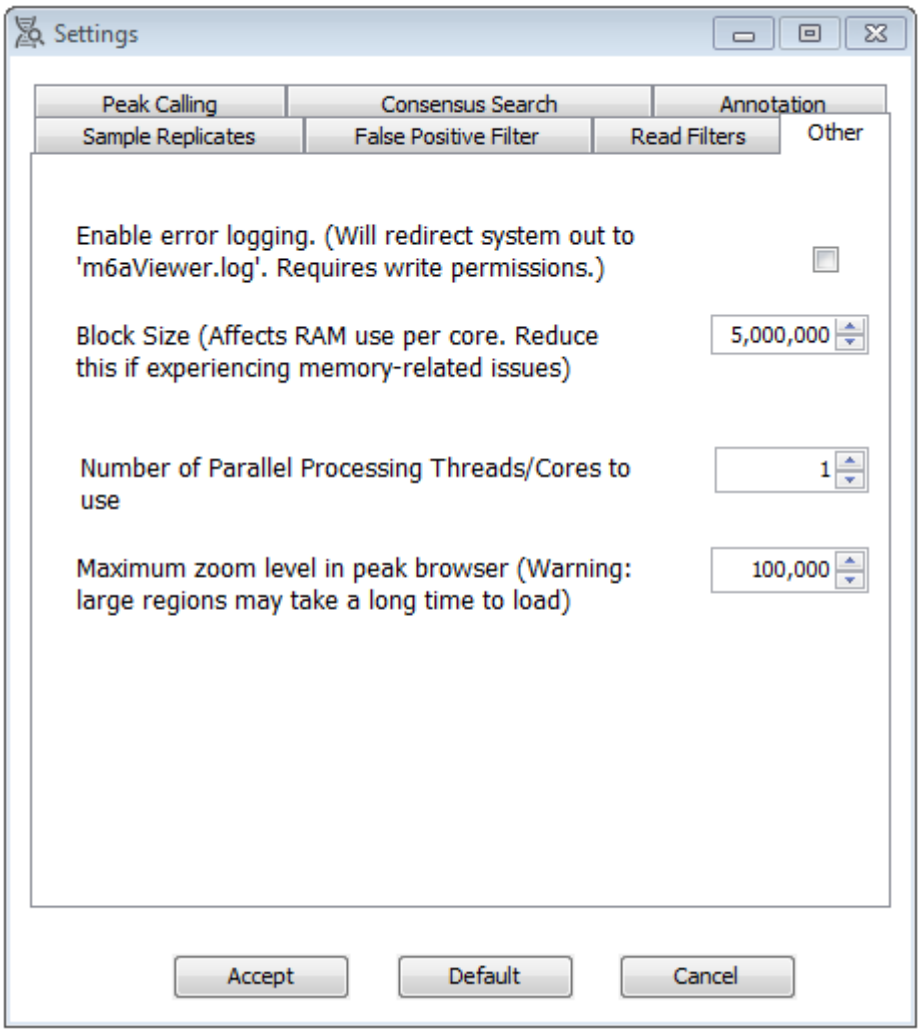

Allows to filter out peaks which are likely to be false positives. By default, this does nothing. The top checkbox will only annotate peaks with a confidence score, while the bottom checkbox will filter out peaks below a certain confidence score – default is 0.5, and can be increased or decreased using the slider.

### 11.2.6 Read Filters




Options to exclude poor quality reads/alignments from coverage calculations. Appropriate SAM flags must be set prior to analysis, e.g. by running Picard for duplicate reads.

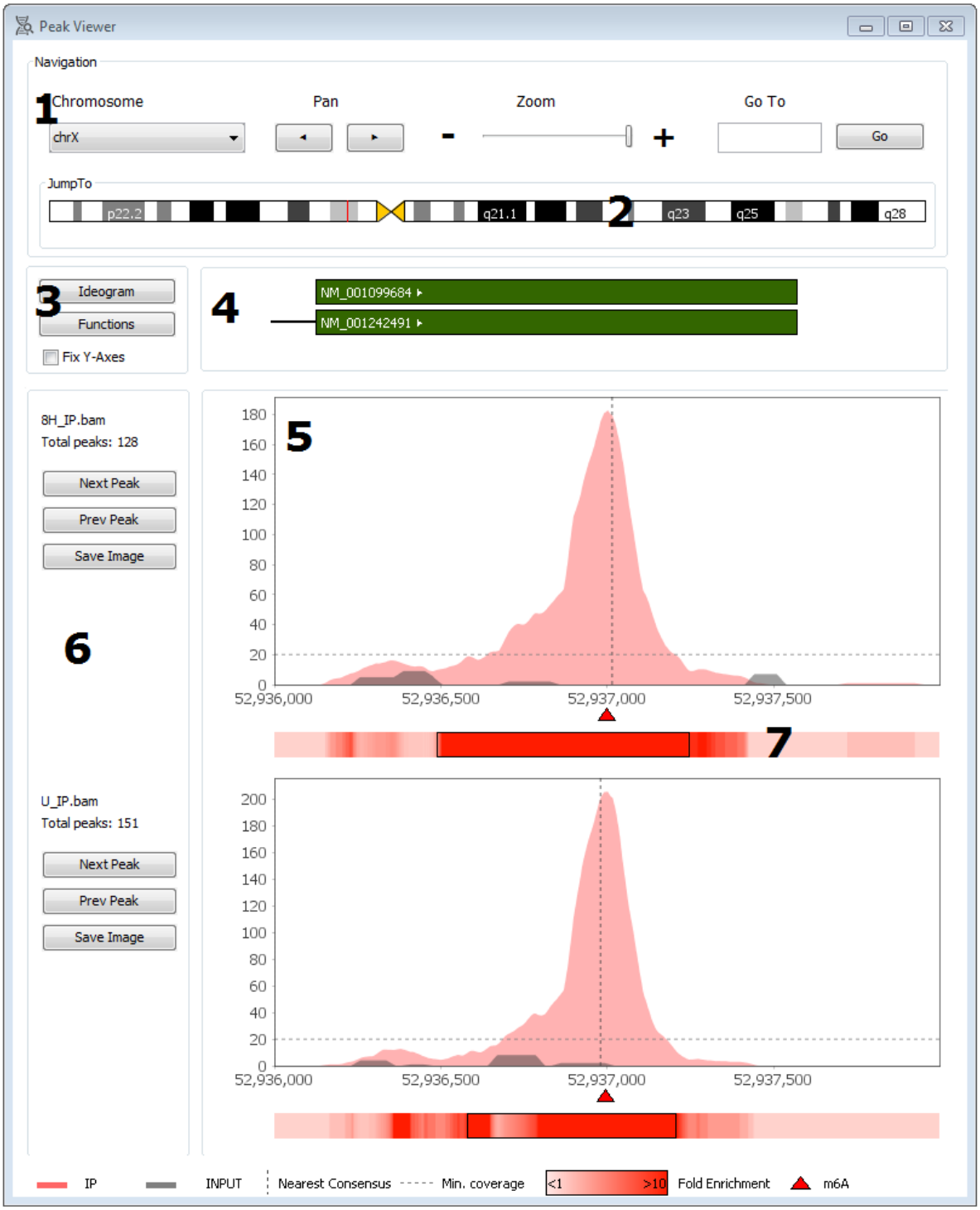
### 11.2.7 Other software options




Block size affects how much of a sequence is processed at once by each thread. Setting this too high can result in out-of-memory errors. However, higher values result in faster peak-calling.

Number of parallel processing cores to use – for multi-core processors, will perform peak calling on X blocks in parallel.

## 11.3 Viewer




1. Basic navigation, zooming, panning left and right, jump to a location and chromosome selection.

2. Drag a square to jump to a relative point in the chromosome

3. Buttons to view ideogram of current chromosome and perform functional analysis of peaks. These will open a separate dialog box. Additionally, a checkbox for when viewing more than one sample at the time, will lock the Y axis scale of all samples to that of the highest Y value in all samples.

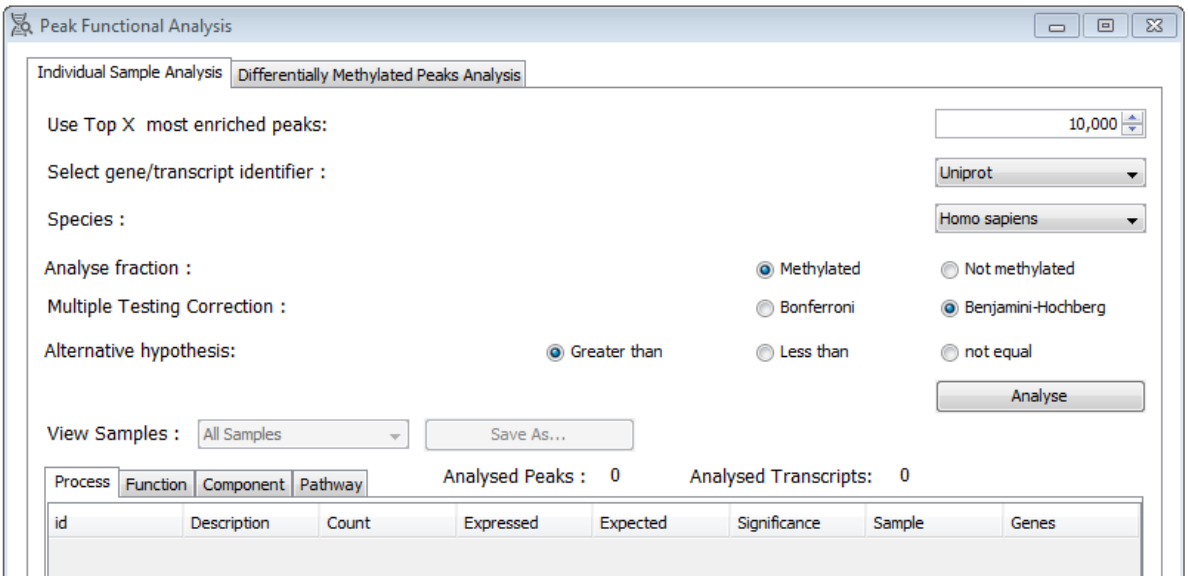
4. Annotation track. Will not be present if GTF file was not imported. Arrow indicates forward ( >) or reverse ( <) strand. Similarly, green indicates forward strand CDS, while dark green forward strand UTR; red reverse strand CDS and dark red reverse strand UTR. When zooming out, name labels might no longer fit onto exons and will not be displayed; however a mouse-over tooltip is available. This provides the name of the gene the exon is in and the exon start and end positions on the chromosome.

5. Sample tracks showing peaks in the data. Red triangles represent m6a sites and will provide tooltips with additional information such as gene name, enrichment value and nearest consensus sequence annotations (If relevant options were selected).Immunoprecipitated sample is shown in salmon pink, while the control is in grey. Nearest consensus sequence to the peak (if within less than 100bp) is indicated with a dotted line.

6. This displays how many peaks were detected for the sample for the current chromosome with current options. Next and Previous peak buttons will jump to the next peak either to the left or right of current view. This will move the image for all samples. Save image will save the current graph as an image.

7. Displays individual position enrichment. Significantly enriched regions that were detected are represented by a black outline.

## 11.4 Peak Functional Analysis




Reports on enriched gene ontology functions or reactome pathways in detected peaks.

This feature requires internet connection to work as Gene Ontology terms and Reactome Pathways are retrieved from Ensembl databases. This analysis may take some time, depending on the number of peaks analysed. This feature also requires GTF file to be uploaded (or in-built annotation selected prior to peak calling).

Select species and identifier type used in the GTF file.

Analyse fraction – select to a set of genes which are expressed and methylated or genes which are expressed but not methylated for gene functional enrichment.

Select the alternative hypothesis – look for terms which appear with greater than expected or lower than expected frequency in the gene set.