



# Unipro UGENE Workflow Designer Manual

Version 1.14.0

July 28, 2014



# Workflow Designer Manual

- About the Workflow Designer
- Introduction
  - Launching Workflow Designer
  - Workflow Designer Window Components
  - Workflow Elements and Connections
  - Managing Parameters
  - UGENE Components and Workflow Designer
    - Task View, Notifications and Log View
    - Actions Menu
    - Toolbar
    - Context Menus
    - Application Settings
  - How to Create and Run Workflow
- Manipulating Element
  - Adding Element
  - Copying Element
  - Pasting Element
  - Cutting Element
  - Deleting Element
  - Selecting All Elements on Scene
- Manipulating Workflow
  - Creating New Workflow
  - Loading Workflow
  - Saving Workflow
  - Exporting Workflow as Image
  - Validating Workflow
  - Running Workflow
  - Dashboard
    - Dashboard Window Components
    - Using Dashboard
  - Stopping and Pausing Workflow
- Changing Appearance
- Custom Elements with Scripts
  - Functions Supported for Multiple Alignment Data
  - Functions Supported for Sequence Data
  - Functions Supported for Set of Annotations Data
  - Functions Supported for Files
  - Common Function
- Custom Elements with Command Line Tools
  - Creating Element
  - Editing Element
  - Adding Existent Element
  - Removing Element
- Using Script to Set Parameter Value
- Running Workflow from the Command Line
- Running Workflow in Debugging Mode
  - Creating Breakpoints
  - Manipulating Breakpoints
- Workflow File Format
  - Header
  - Body
    - Elements
    - Dataflow
    - Metainformation
- Workflow Elements
  - Data Readers
    - File List Element
    - Read Alignment Element
    - Read Annotations Element
    - Read Assembly Element
    - Read from DAS Element
    - Read from Remote Database Element
    - Read Plain Text Element
    - Read Sequence Element
    - Read Variations Element
  - Data Writers
    - Write Alignment Element
    - Write Annotations Element
    - Write Assembly Element
    - Write FASTA Element
    - Write Plain Text Element
    - Write Sequence Element
    - Write Variations Element
  - Data Flow
    - Filter Element
    - Grouper Element

- Multiplexer Element
- Sequence Marker Element
- Basic Analysis
  - Amino Translations Element
  - Annotate with DAS Element
  - Annotate with UQL Element
  - CD-Search Element
  - Collocation Search Element
  - Export PHRED Qualities Element
  - Fetch Sequences by ID From Annotation Element
  - Filter Annotation by Name Element
  - Filter Annotations by Qualifier
  - Find Pattern Element
  - Find Repeats Element
  - Gene-by-gene approach report
  - Get Sequences by Annotations Element
  - Import PHRED Qualities Element
  - Local BLAST Search Element
  - Local BLAST+ Search Element
  - Merge Annotations Element
  - ORF Marker Element
  - Remote BLAST Element
  - Remove Duplicates in BAM Files Element
  - Smith-Waterman Search Element
- Data Converters
  - Convert bedGraph Files to bigWig Element
  - Convert Text to Sequence Element
  - File Format Conversion Element
  - Reverse Complement Element
  - Split Assembly into Sequences Element
- DNA Assembly
  - Align reads with BWA-MEM
  - Assembly Sequences with CAP3
  - Extract Consensus from Assembly
- HMMER2 Tools
  - HMM Build Element
  - HMM Search Element
  - Read HMM Profile Element
  - Write HMM Profile Element
- HMMER3 Tools
  - HMM3 Build Element
  - HMM3 Search Element
  - Read HMM3 Profile
  - Write HMM3 Profile
- Multiple Sequence Alignment
  - Align Profile to Profile with MUSCLE Element
  - Align with ClustalO Element
  - Align with ClustalW Element
  - Align with Kalign Element
  - Align with MAFFT Element
  - Align with MUSCLE Element
  - Align with T-Coffee Element
  - Extract Consensus from Alignment
  - Join Sequences into Alignment Element
  - Split Alignment into Sequences Element
- NGS Basic
  - CASAVA FASTQ Filter Element
  - FASTQ Quality Trimmer Element
  - Filter BAM/SAM Files Element
  - Genome Coverage Element
  - Merge BAM Files Element
  - Slopbed Element
  - Sort BAM Files Element
- NGS: ChIP-Seq Analysis
  - Annotate Peaks with peak2gene Element
  - Build Conservation Plot Element
  - Collect Motifs with SeqPos Element
  - Conduct GO Element
  - Create CEAS Report Element
  - Find Peaks with MACS Element
- NGS: RNA-Seq Analysis
  - Assembly Transcripts with Cufflinks Element
  - Extract Transcript Sequences with gffread Element
  - Find Splice Junction with TopHat Element
  - Merge Assemblies with Cuffmerge Element
  - Test for Diff. Expression with Cuffdiff Element
- NGS: Variant Calling
  - Call Variants with SAMtools Element
  - Create VCF consensus
- SNP Annotation

- Annotate variations with SNPToolbox Element
  - Detect Transcription Factors with rSNP-Tools Element
  - Determine SNP effect on TATA-boxes Element
  - ProtStability1D Element
  - ProtStability3D Element
  - SNP Chip Tools Element
  - SNP Effect on PDB sites Element
  - Write SNP Report Element
- Transcription Factor
  - Build Frequency Matrix Element
  - Build SITECON Model Element
  - Build Weight Matrix Element
  - Convert Frequency Matrix Element
  - Read Frequency Matrix Element
  - Read SITECON Model Element
  - Read Weight Matrix Element
  - Search for TFBS with SITECON Element
  - Search for TFBS with Weight Matrix Element
  - Write Frequency Matrix Element
  - Write SITECON Model Element
  - Write Weight Matrix Element
- Utils
  - DNA Statistics Element
  - Generate DNA Element
- Custom Elements With Script
  - CASAVA FASTQ Filter Script Element
  - Dump Sequence Info Element
  - FASTQ Trimmer Element
  - LinkData Fetch Element
  - Quality Filter Element
- Workflow Samples
  - Alignment
    - Align sequences with MUSCLE
  - Conversions
    - Convert seq/qual pair to Fastq
    - Convert alignments to ClustalW
    - Convert UQL schema results to alignment
    - Convert sequence to Genbank
  - Custom elements
    - CASAVA FASTQ Filter
    - FASTQ Trimmer
    - Dump sequence info
    - LinkData fetch
    - Quality filter
  - Data Marking
    - Marking Sequences by Annotation Number
    - Marking Sequences by Length
  - Data Merging
    - Find Substrings at Sequences
    - Merge Sequences and Shift Corresponding Annotations
    - Search for TFBS
  - HMMER
    - Build HMM from alignment and test it
    - Search sequences with profile HMM
  - NGS
    - Call variants with SAMtools
    - ChIP-seq analysis with Cistrome tools
    - Extract Consensus
    - Extract transcript sequences
    - RNA-seq analysis with Tuxedo tools
  - Scenarios
    - Filter sequence that match a pattern
    - Find patterns
    - Gene-by-gene approach for characterization of genomes
    - Merge sequences and annotations
  - Transcriptomics
    - Search for transcription factor binding sites (TFBS) in genomic sequences



## About the Workflow Designer

UGENE Workflow Designer is a part of [UGENE](#) genome analysis suite that allows a molecular biologist to create and run complex computational workflows even if he or she is not familiar with any programming language.

The workflows comprise reproducible, reusable and self-documented research routines, with a simple and unambiguous visual representation suitable for publications.

The workflows can be run both locally and remotely, either using graphical interface or launched from the command line.

The elements that a workflow consists of corresponds to the bulk of algorithms integrated into [UGENE](#). Additionally you can create custom workflow elements.

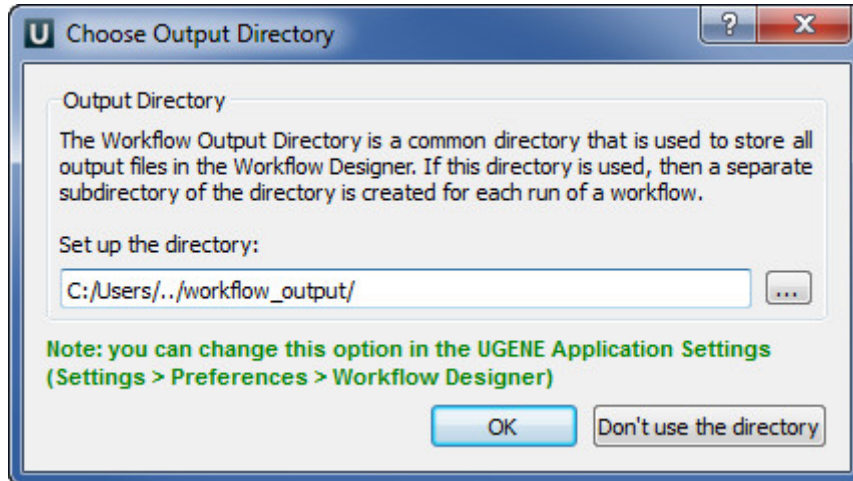
## Introduction

This chapter describes the Workflow Designer key elements and provides an example on how to create and run a simple [workflow](#).

- [Launching Workflow Designer](#)
- [Workflow Designer Window Components](#)
- [Workflow Elements and Connections](#)
- [Managing Parameters](#)
- [UGENE Components and Workflow Designer](#)
- [How to Create and Run Workflow](#)

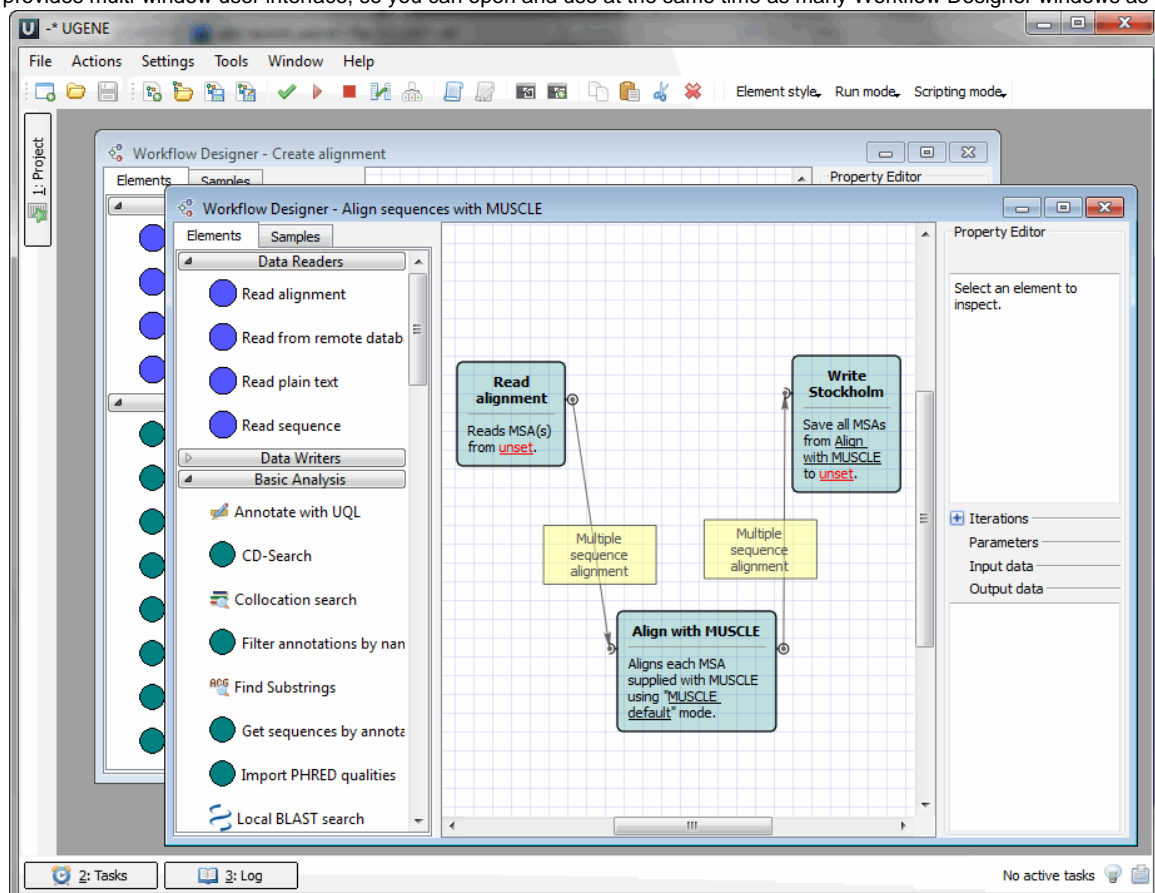
## Launching Workflow Designer

To launch the Workflow Designer select the *Tools Workflow Designer* item in the UGENE main menu. The following Choose Output Directory dialog appears:



The output directory is a common directory that is used to store all output files in the Workflow Designer. If this directory is used, then a separate subdirectory of the directory is created for each run of a workflow. You can change this option in the [Application Settings](#) dialog.

The tool provides multi-window user interface, so you can open and use at the same time as many Workflow Designer windows as you need.



## Workflow Designer Window Components

Each Workflow Designer window consists of:

### Palette

The *Elements* tab of the palette contains *workflow elements* for most algorithms intergrated in UGENE and sets of common input / output routines. The elements are grouped into categories that reflect their uses and features. The *Samples* tab of the palette contains examples of *workflow*.

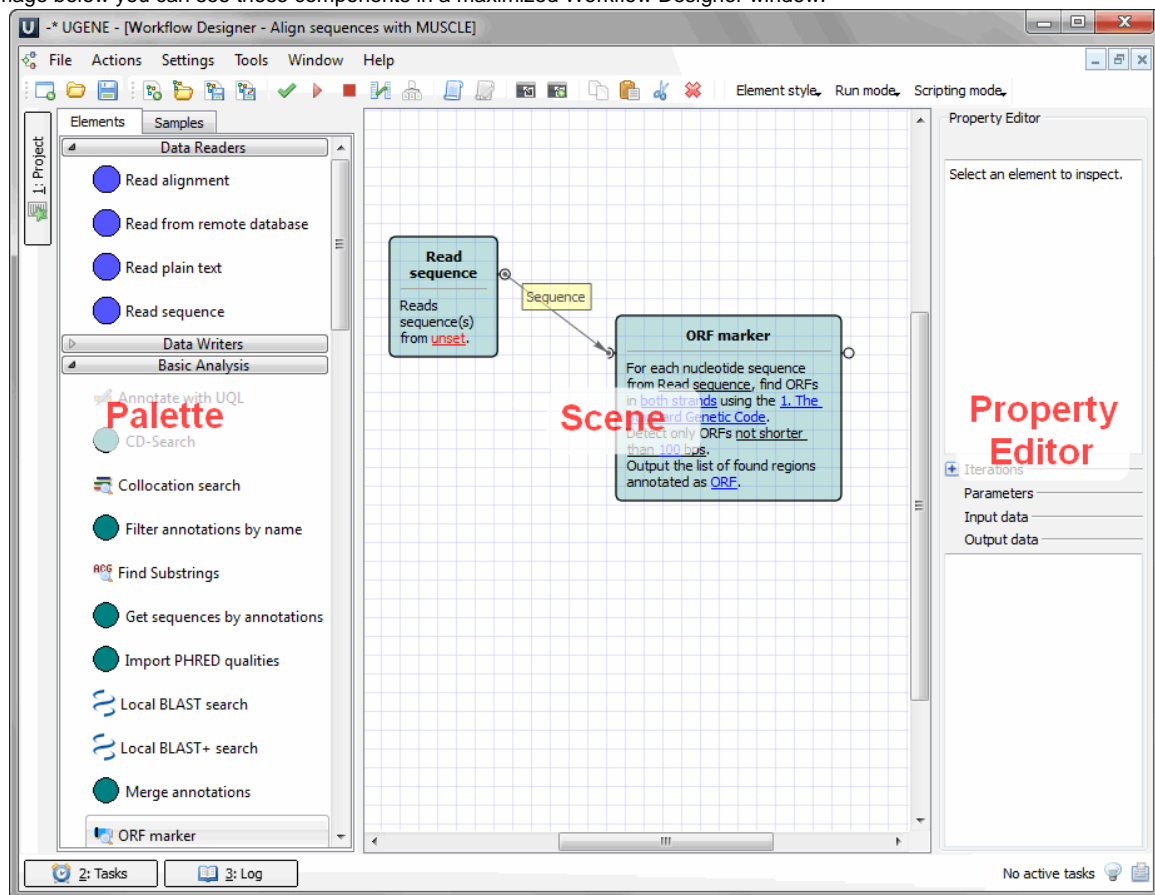
### Scene

The main drawing scene is the place where the workflow elements are constructed into a workflow.

### Property Editor

Provides information about a currently selected workflow element and allows configuring it.

On the image below you can see these components in a maximized Workflow Designer window:



All these components are resizable and can be adjusted to individual needs.

## Workflow Elements and Connections

The *Scene* is initially empty and you start with creating a workflow on it:

### workflow

A workflow is a visual representation of the dataflow. It consists of workflow elements and their connections.

### workflow element

An element of a workflow. Different elements are used to read data from files on disk, perform some algorithms and to write data to files on disk. Each element contains one or several input and output ports.

### element connection

Connection between two elements specifies that data in output port of one element should be passed to a matching input port of another element.

### input port

An input port of an element is used to collect data from another element. On the Scene it is displayed as prominent knob on an element with opened bubble.

### output port

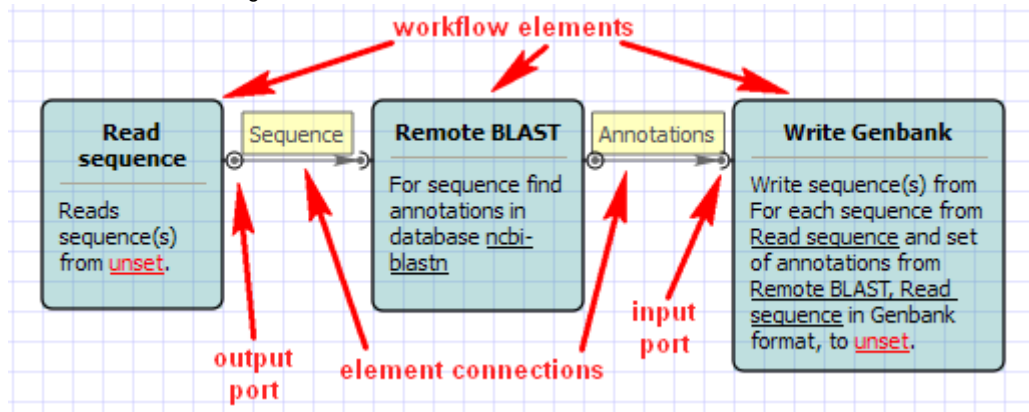
An output port of an element is used to provide data to another element. On the Scene it is displayed as prominent knob on an element with closed bubble.

### slot

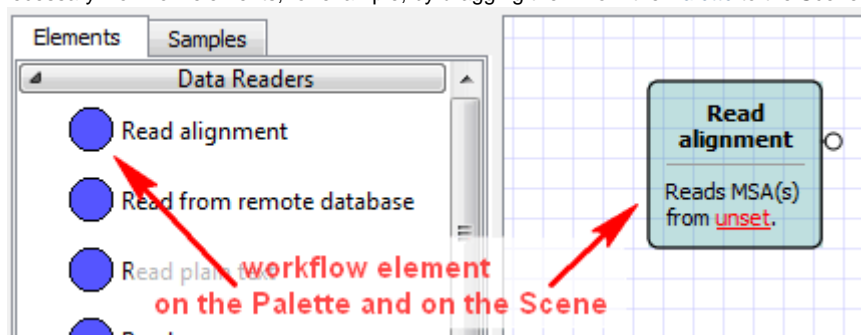
A slot specifies the kind of data that can be passed through it (for example Sequence, Set of annotations, etc.)

The Scene is initially empty and you start with creating a workflow on it:

See an example of a workflow on the image below:

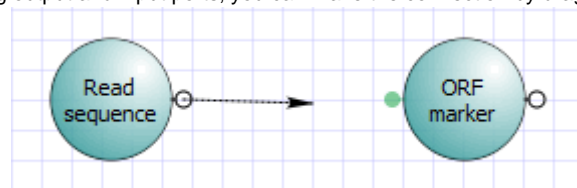


Your first step is to **add** necessary workflow elements, for example, by dragging them from the *Palette* to the Scene:



The added element can be moved around on the Scene by dragging it and can be resized by dragging its borders. Read chapter *Manipulating Element* to learn what else you can do with workflow elements.

If you have two elements with matching output and input ports, you can make the connection by dragging the arrow between the ports:



All matching ports of available processes are highlighted while you drag the arrow, besides the arrow sticks to a near match when you drag closer. If an element has a sole matching port, you can just drop the arrow on the element itself to create a correct connection.

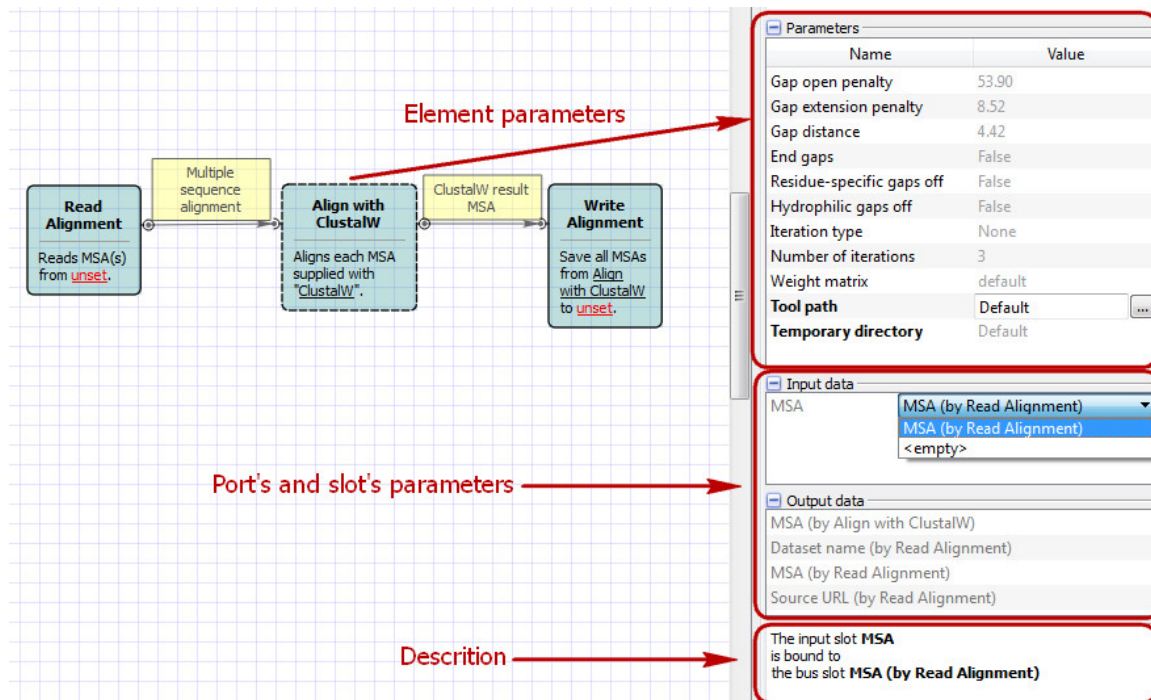
Once created, a connection will follow movements of the linked elements; you cannot redirect or reshape the connection arrow but only remove it. You can move the port around an element that it belongs to by dragging it and holding the Alt key at the same time. This is helpful to fine-tune visual layout of a workflow.

## Managing Parameters

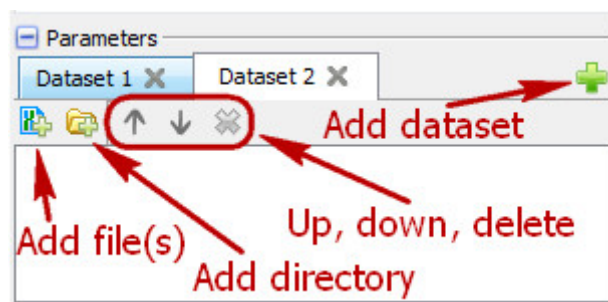
When you select an *element* on the *Scene* the *Property Editor* displays detailed information about it: it's name, description, parameters, *input* and *output* ports, etc. To change the name of the element displayed on the Scene edit the *Element name* value.

All the parameters available for the element are displayed in the *Parameters* area. Some parameters must have a value, they are displayed in bold. Notice, that when you select a parameter, it's description is shown below. To modify a value click on it. Depending on the parameter's type you may be required to either input a value or browse for a file(s). Also you can configure slots of a connected input port by

selecting different (matching) data available through the dataflow. More advanced users can use their own scripts to set a parameter's value, read chapter [Using Script to Set Parameter Value](#) to learn more. The image below shows the *Property Editor*:



For [Data Readers](#) you can manipulate with file(s) or directory(ies) with a help of dataset(s):



Also, to remove files from dataset you can select it and press the *Delete* button.

## UGENE Components and Workflow Designer

This paragraph provides an overview of UGENE components that affect your work with the Workflow Designer.

- [Task View, Notifications and Log View](#)
- [Actions Menu](#)
- [Toolbar](#)
- [Context Menus](#)
- [Application Settings](#)

### Task View, Notifications and Log View

When a workflow is executed in the Workflow Designer a **task** is created.

#### Task View

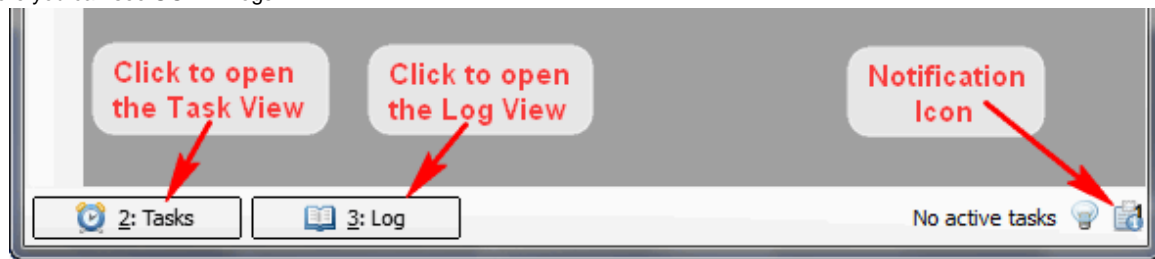
Here you can see the tasks currently executed in UGENE.

#### Notification Icon

When a task has finished its execution, a notification is pop up. At any time you can watch the last notifications by clicking the *Notification Icon*.

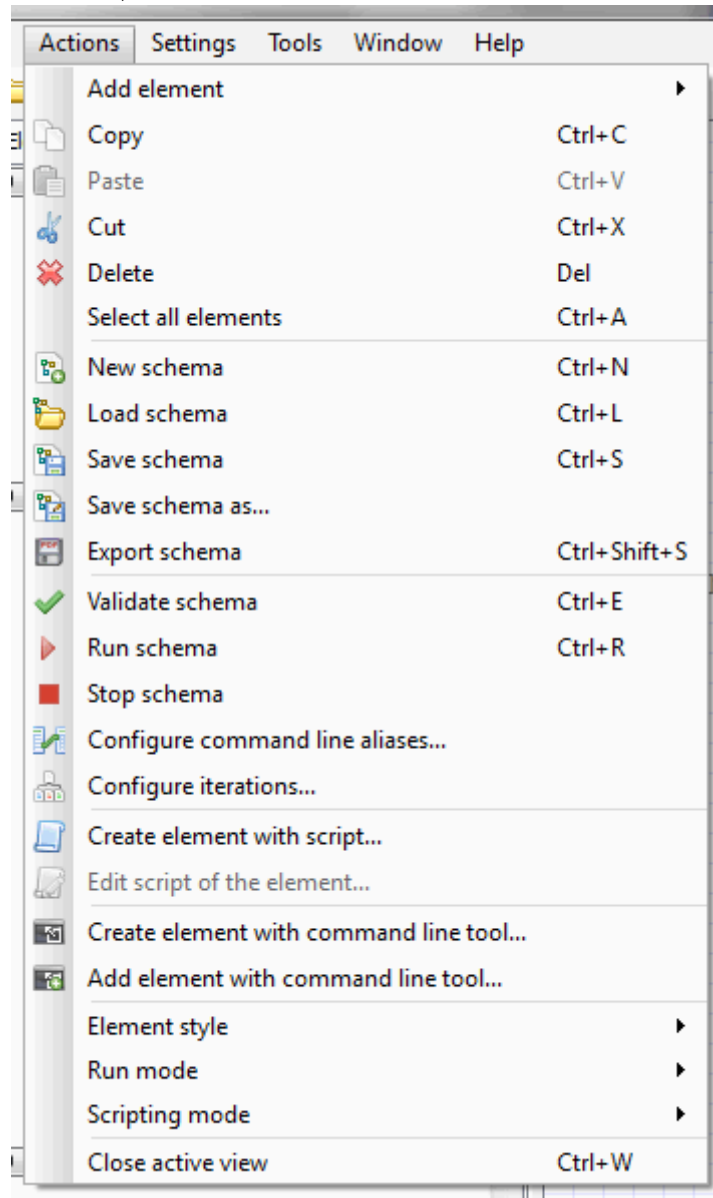
#### Log View

Here you can see UGENE logs.



## Actions Menu

When a Workflow Designer window is active, all standard actions to work with workflow are available from the *Action* main menu:



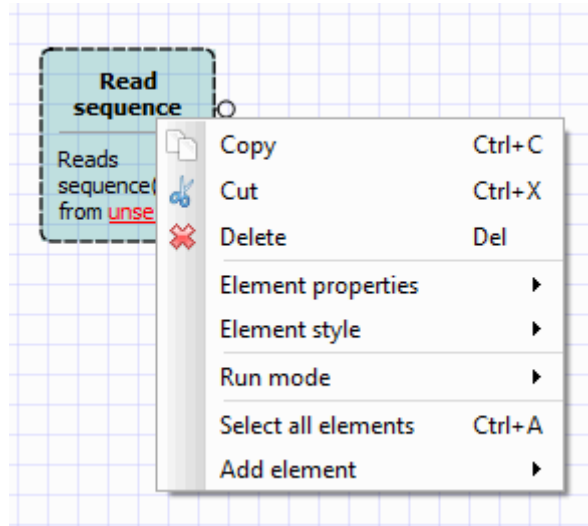
## Toolbar

Most common actions are available on the main toolbar:



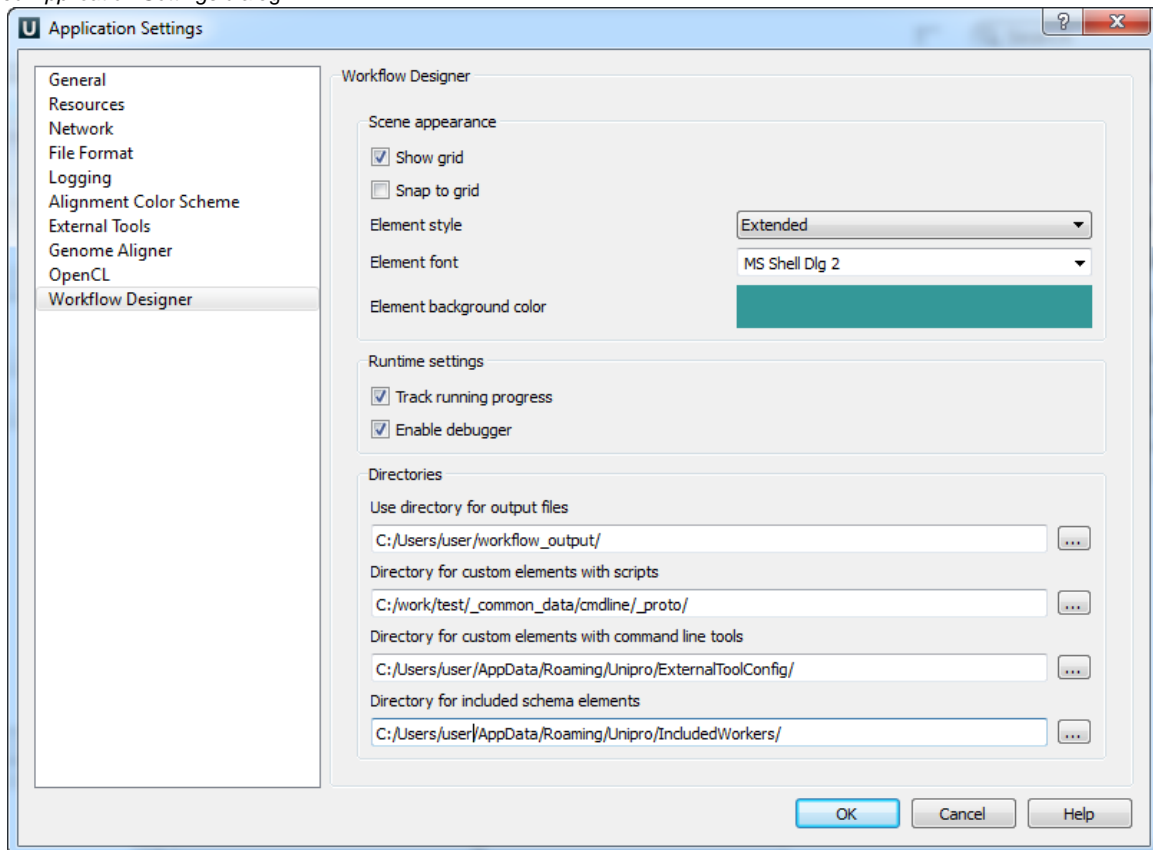
## Context Menus

Some features are also available through context menus over corresponding areas, e.g.:



## Application Settings

To change common Workflow Designer setting select the *Settings Preferences...* main menu item and select the *Workflow Designer* tab in the opened *Application Settings* dialog.



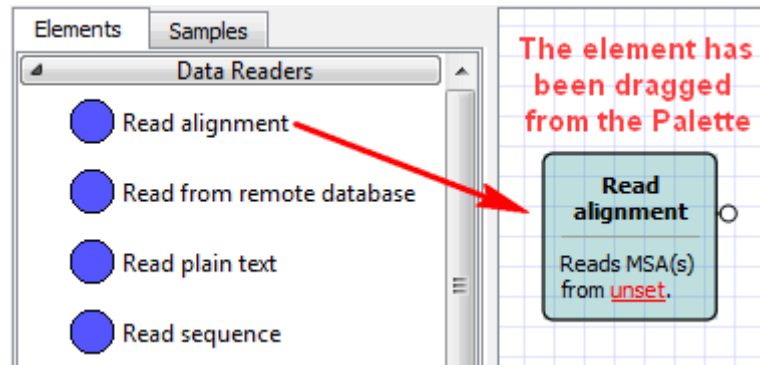
## How to Create and Run Workflow

- Select *Tools* → *Workflow Designer* in the main menu.

**Result:** The Workflow Designer window appears.

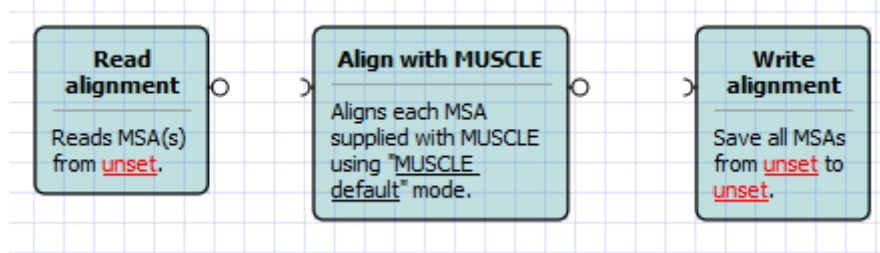
- On the *Elements* tab of the *Palette* find the *Read alignment* element. It is located in the *Data sources* group and drag it to the *Scene*.

**Result:** The element is shown on the Scene.



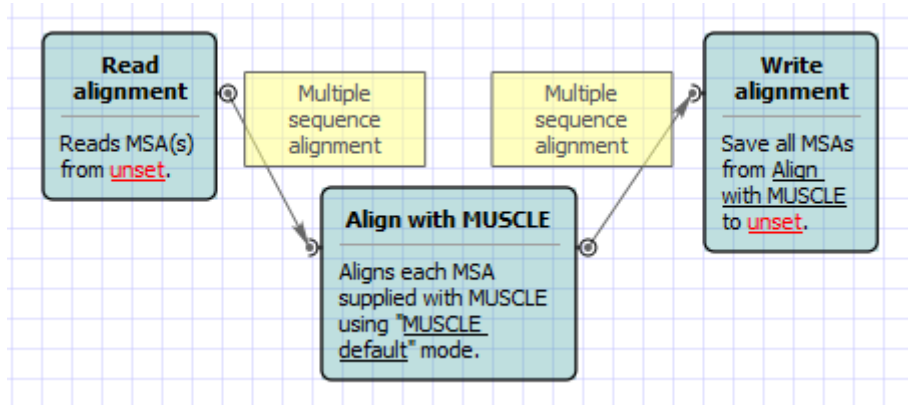
- Repeat the previous step for the *Write Alignment* element from the *Data sinks* group and for the *Align with MUSCLE* element from the *Multiple sequence alignment* group.

**Result:** All three elements are on the Scene.



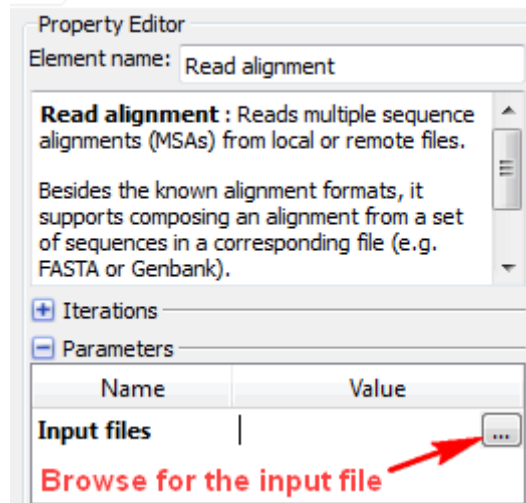
- Connect the elements:
  - Drag an arrow from the *output port* of the *Read alignment* element to the *Align with MUSCLE* element.
  - Drag an arrow from the output port of the *Align with MUSCLE* element to the *Write alignment* element.

**Result:** The elements are connected with arrows.



- Select the *Read alignment* element. In the *Parameters* area of the *Property Editor* click on the *Value* column of the *Input files* parameter:





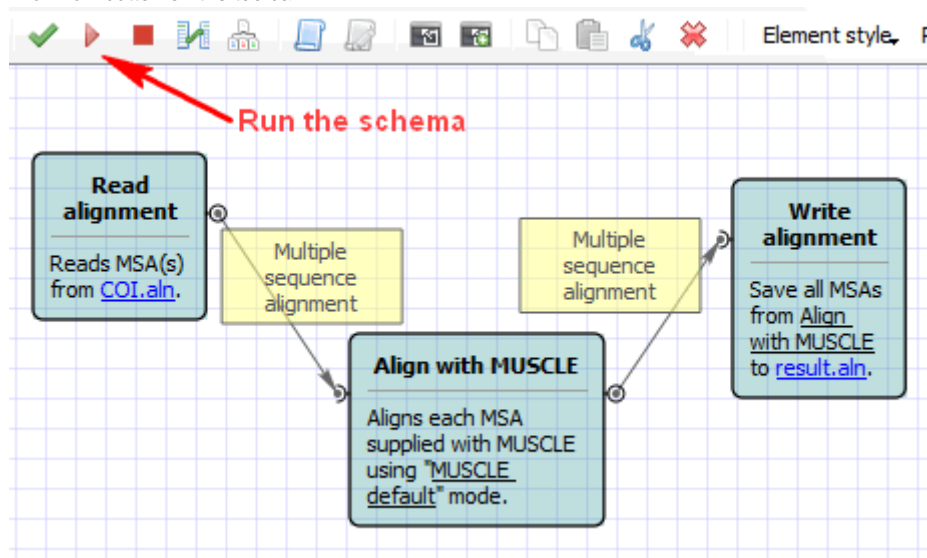
- And browse for an input file, e.g. Select the \$UGENE\data\samples\CLUSTALW\COI.aln file.

**Result:** The *Input files* value is set to the file's path.

- Select the *Write alignment* element and set the *Output file*, e.g. you can just enter result.aln.

**Result:** All required workflow parameters are set.

- Click the *Run workflow* button on the toolbar.



**Result:** After the workflow has run, a blue notification has pop up.

- Open the the result.aln file in UGENE.

**Result:** The file has been opened. It contains the result of the alignment with MUSCLE.

## Manipulating Element

You can add new *workflow element* to the *Scene*, copy, cut, paste or delete it. Also you can select all elements currently presented on the *Scene*.

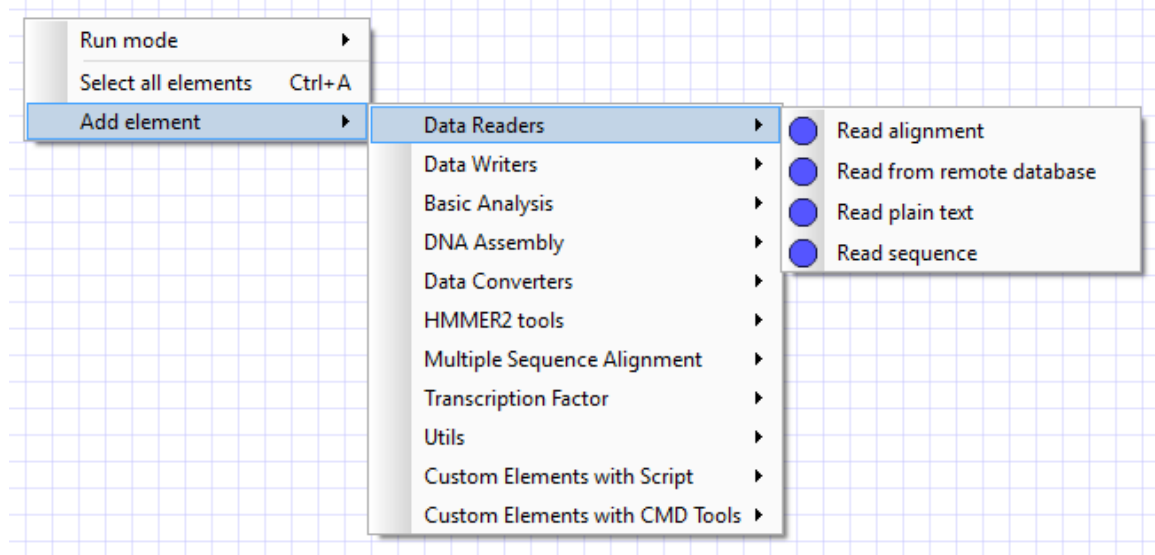
- Adding Element
- Copying Element
- Pasting Element
- Cutting Element
- Deleting Element
- Selecting All Elements on Scene

## Adding Element

There are several ways to add an *element* to the *Scene*.

The easiest way is to drag the required element from the *Palette* to the *Scene*. Or you can just click on the element on the *Palette* and then click somewhere on the *Scene*.

Also you can select an element in the *Add item* submenu of the *Actions* main menu or of the *Scene* context menu, for example:



When the required element is selected click somewhere on the *Scene* to insert it.

## Copying Element

To copy one or several *workflow elements* select them on the *Scene*. Note, that you can hold the Ctrl key to select several elements. Then choose the *Copy* item in the *Actions* main menu or in a selected element context menu.

The Ctrl+C hotkey is also available for this action.

Now you can *paste* these elements somewhere on the *Scene*.

## Pasting Element

You can paste *workflow elements* that have been *cut* or *copied*.

To do it choose the *Paste* item in the *Actions* main menu or in the *Scene* context menu.

Or use the Ctrl+V hotkey to paste the elements.

## Cutting Element

To cut one or several *workflow elements* select them on the *Scene*. Choose the *Cut* item in the *Actions* main menu or in a selected element context menu.

The Ctrl+X hotkey is also available for this action.

Now you can *paste* these elements.

## Deleting Element

Select one or several *workflow elements* on the *Scene* that you want to delete. Then choose the *Delete* item in the *Actions* main menu or in a selected element context menu.

The hotkey for this action is Del.

### Selecting All Elements on Scene

To select all *workflow elements* presented on the *Scene* choose the *Select all elements* in the *Actions* main menu or in the Scene context menu.

Or use the Ctrl+A hotkey.

## Manipulating Workflow

You can create a new [workflow](#), save it and then load it again.

The designed workflow can be displayed in a neat self-describing layout and exported to PDF document, raster or vector image with publication-ready quality.

You can validate created or modified workflow before running it.

If you need, you can stop a workflow execution.

- [Creating New Workflow](#)
- [Loading Workflow](#)
- [Saving Workflow](#)
- [Exporting Workflow as Image](#)
- [Validating Workflow](#)
- [Running Workflow](#)
- [Dashboard](#)
- [Stopping and Pausing Workflow](#)

## Creating New Workflow

To create a new [workflow](#) select the *Actions* *New workflow* item in the main menu or *New workflow* toolbar button.

Or press Ctrl+N.

## Loading Workflow

To load a workflow select the *Actions* *Load workflow* item in the main menu or *Load workflow* toolbar button.

Or press Ctrl+L.

### Hint

You can load a workflow by dragging the workflow file (e.g. with .uwl extension) to the UGENE window.

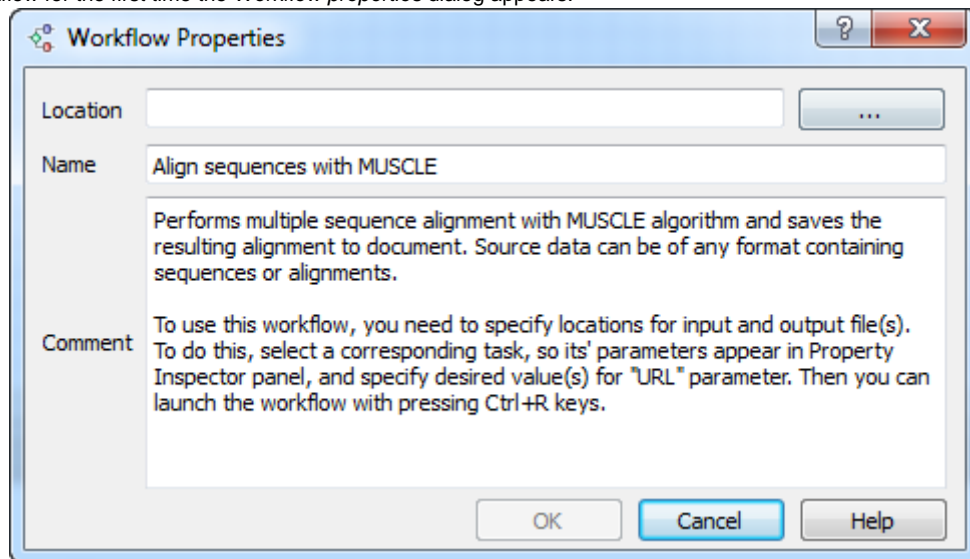
## Saving Workflow

Choose *Actions* *Save workflow* item in the main menu or *Save workflow* toolbar button to save a workflow. The workflow is saved to a file of native UGENE format (with .uwl extension).

The format is human-readable, you can find it's description in chapter [Workflow File Format](#).

There is Ctrl+S keyboard shortcut for this action.

If you save a workflow for the first time the *Workflow properties* dialog appears:



Here you can browse for the workflow file *Location* and specify the workflow *Name* and *Comment*.

Once a workflow has been saved, it can be [loaded](#). If you modify the loaded workflow and save changes, then corresponding .uwl file is modified.

To save the workflow with different properties choose the *Actions* *Save workflow as* item in the main menu and specify the required settings in the *Workflow properties* dialog.

## Exporting Workflow as Image

Workflow workflow can be exported as:

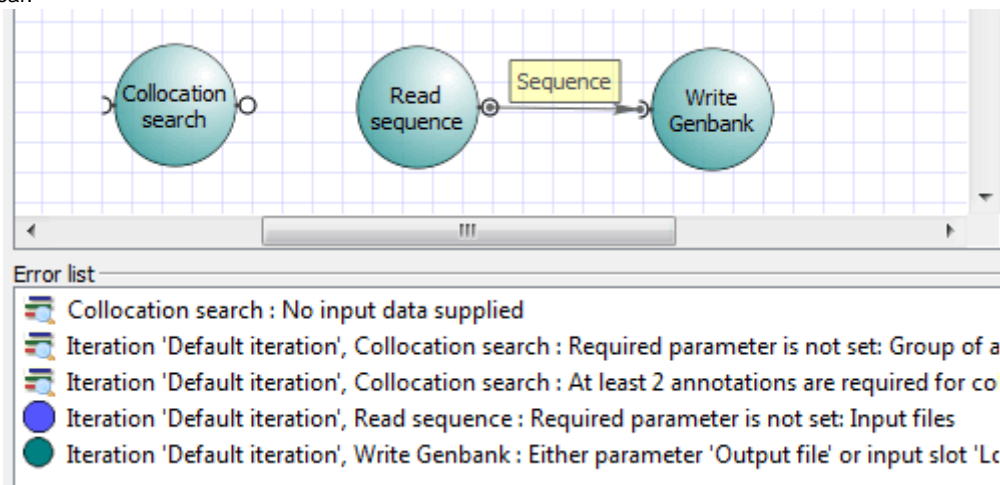
- Raster image (\*.png, \*.bmp, \*.jpg, \*.jpeg, \*.ppm, \*.xbm, \*.xpm)
- Vector image (\*.svg)
- Portable document (\*.pdf, \*.ps)

To export a workflow select the *Actions* *Export workflow as image* item in the main menu or use the Ctrl+Shift+S keyboard shortcut. *Export Image* dialog will appear. Enter a file name and choose the file type.

## Validating Workflow

Before a workflow can be actually executed, it should be verified by the Workflow Designer. During the process of verification the Workflow Designer checks if there are errors in the dataflow logic or unspecified parameters and can provide a user with optimization or layout hints. If no errors were found, the workflow is valid to be *run*.

You can request workflow validation at any stage of workflow design. To do it choose the *Actions* *Validate workflow* item in the main menu or *Validate workflow* toolbar button or invoke it by pressing Ctrl+E. A list of identified issues and warnings if any, or a notification of validation success will appear.

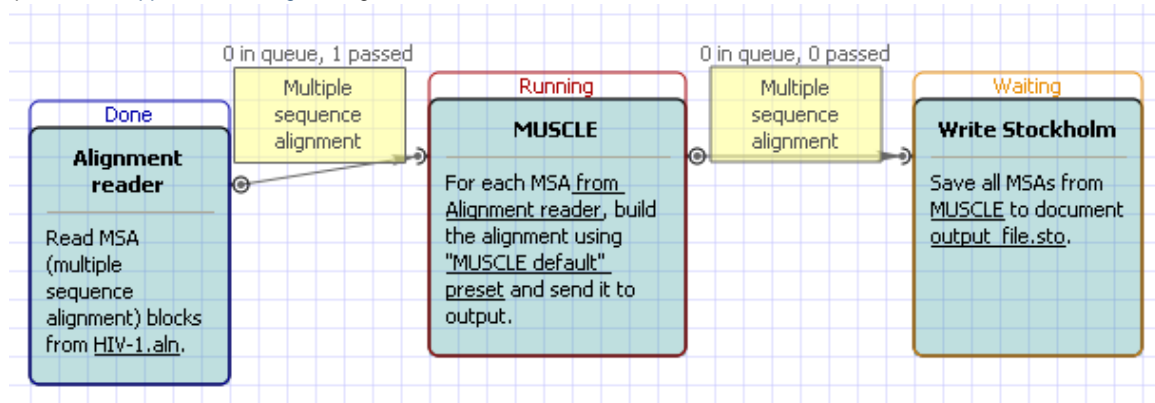


Double-clicking on items in the list selects the faulty element/iteration.

## Running Workflow

Once you are satisfied with the designed workflow and have it configured, click the *Run workflow* button on the toolbar (alternatively, you can select the *Actions* *Run workflow* item in the main menu or launch it by pressing Ctrl+R). The workflow gets verified and scheduled for background execution. If you continue editing the workflow, this will not affect the launched execution. You can control the workflow execution via the *Task View*: watch progress, cancel it, etc. Upon completion, the Workflow Designer produces a *dashboard* with a summary report. The report displays status of each iteration execution and provides other details.

Note, that you can see the progress of a workflow execution in a Workflow Designer window by checking the *Track running progress on diagram* option in the *Application Settings* dialog:



## Dashboard

The dashboard is a central place to view the overall progress of a single workflow. Every dashboard contains two tabs:

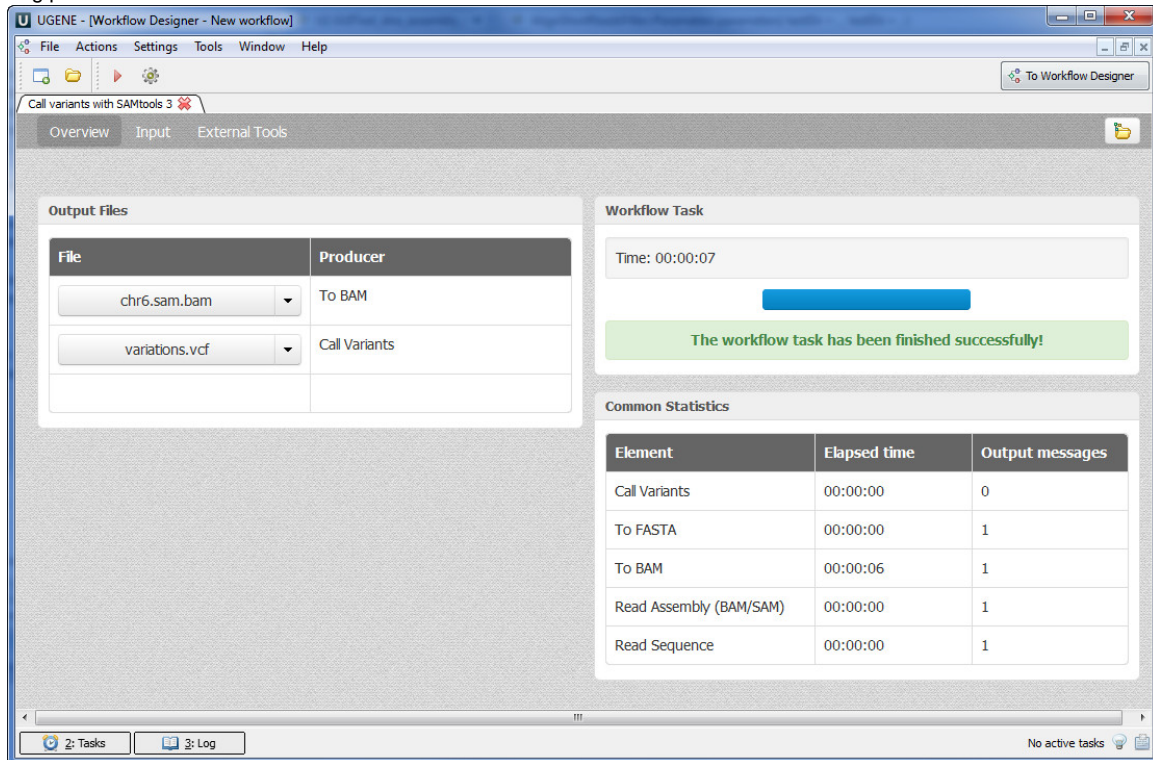
- Overview tab

- Input tab

If a workflow uses external tools the following tab appears on dashboard:

- External Tools tab

The following picture shows the sketch of the the dashboard:



- Dashboard Window Components
- Using Dashboard

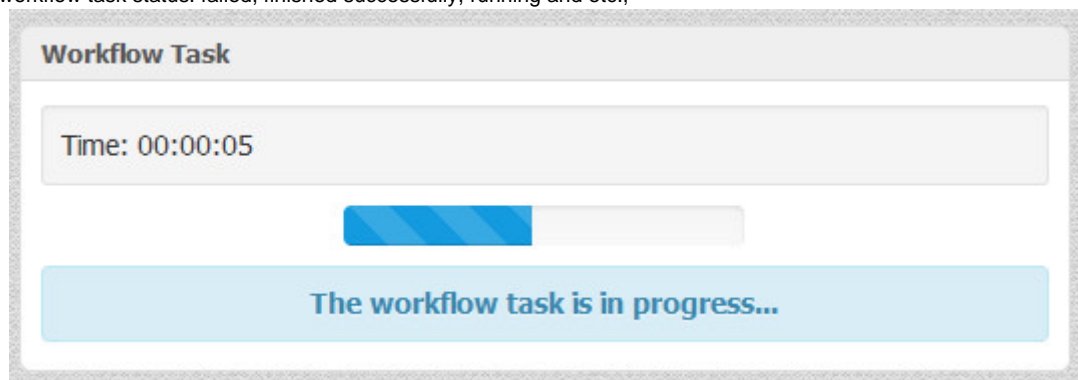
## Dashboard Window Components

### Overview tab

#### **"Workflow Task" widget**

It contains:

- the workflow working time;
- the workflow running progress;
- the workflow task status: failed, finished successfully, running and etc.;



#### **"Output Files" widget**

It contains a table with the information about all created output files. The table columns are:

- clickable file name (here you can open the file containing directory or open the file by operating system);
- the name of the workflow element that has produced the file;

Output Files	
File	Producer
Align_with_MUSCLE ▾	Write alignment

**"Common Statistics" widget**

It contains a table with common statistic for each workflow element in the workflow. The table columns are:


- name of the workflow element;
- time of the workflow element execution;
- the number of messages that has been retrieved;

Common Statistics		
Element	Elapsed time	Output messages
Align with MUSCLE	00:00:01	1
Read alignment	00:00:00	1
Write alignment	00:00:00	0

**"Problems" widget**

It contains a table with problems. The table columns are:

- problems type (warning, error and etc.)
- name of the element with problem
- error message

Problems		
Type	Element	Message
	Read Alignment	Unsupported document format

Input tab

**"Parameters" widget**



It contains a table with common statistic for each workflow element's parameter in the workflow. The table columns are:

- names of the workflow elements;
- names of the workflow parameters;
- values of the workflow parameters;
- clickable file name values of the workflow parameters (here you can open the file containing directory or open the file by operating system);

Parameters	
Read alignment	
Align with MUSCLE	
Write alignment	
Parameter	Value
Max iterations	-1
Mode	0
Region to align	Whole alignment
Stable order	True

## External Tools tab

### "External Tools" widget

It contains information about external tools. There are:

- names of the external tools;
- executable file of the external tool;
- arguments of the external tool;

The 'External Tools' widget displays a list of external tools used in the workflow. Each tool entry includes a 'Run info' button, an 'Executable file' field, an 'Arguments' field, and an 'Error log' button.

**Find Peaks with MACS**

- MACS run 1** (Run info button)
- Executable file**: C:\Python27/python.exe
- Arguments**
- Error log**

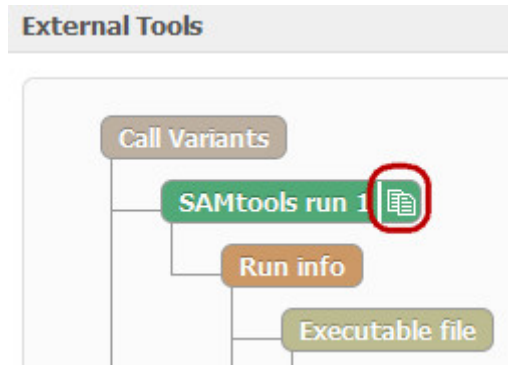
**Build Conservation Plot**

- conservation\_plot run 1** (Run info button)
- Executable file**: C:\Python27/python.exe
- Arguments**:
 

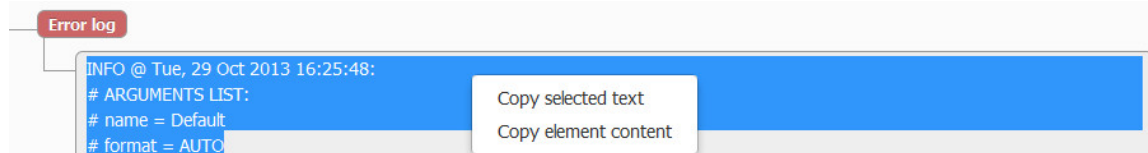
```
--phasdb=E:/UGENE/trunk/data/cistrome/phastCons/hg19
--height=1000
--width=1000 "-w 1000"
--title=""Average Phastcons around the Center of Sites""
--bed-label=Conservation_at_peak_summits C:/Users/yalgaer/AppData/Local/Temp/ugene_tmp/p54244/ConservationPlot_tmp/1383038905_0/Conservation_at_peak_summits.bed
```
- Error log**

To copy external tool run string click the following button:



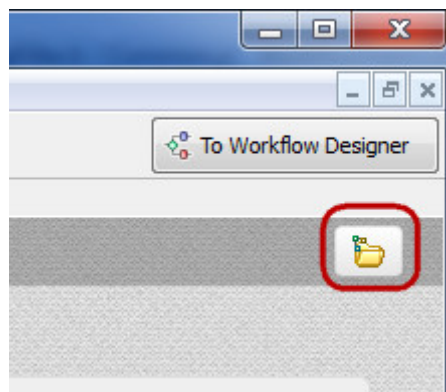


With a help of the context menu of this widget you can copy selected text from the dashboard or copy all text of the active element:

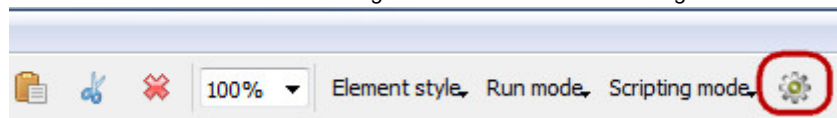


## Using Dashboard

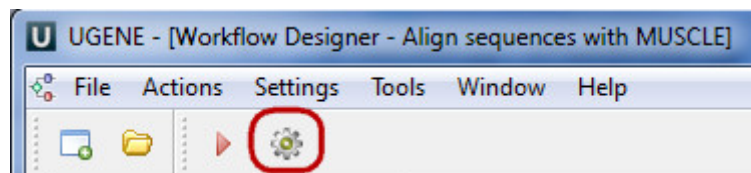
For each workflow which has been runned new dashboard will be opened. This dashboards will be saved in the *selected directory*. Also you will see this dashboard after UGENE will be runned again. Furthermore you can open the original workflow for your results by clicking on this button:



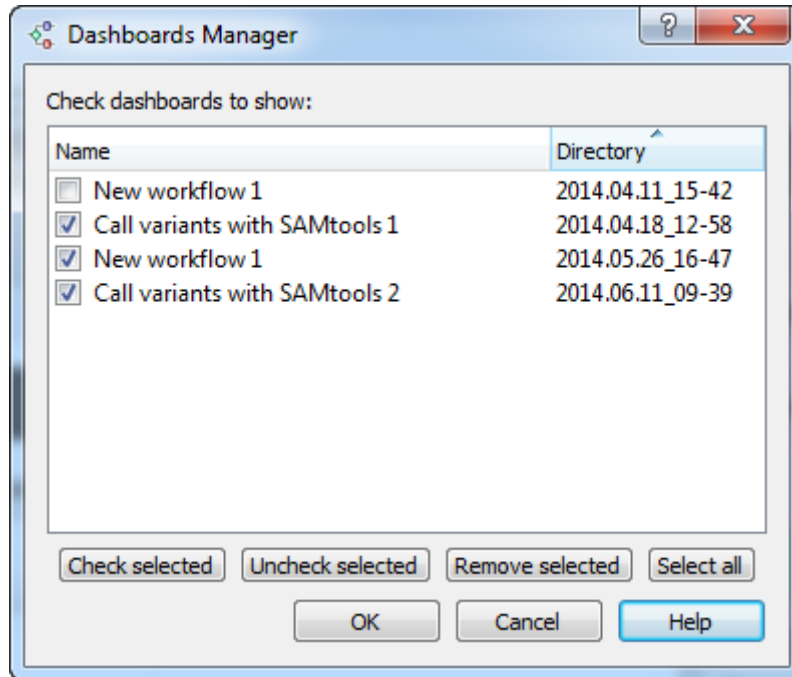
To remove or to load a dashboard click to the *Dashboards manager* button on the *Workflow Designer* main toolbar:



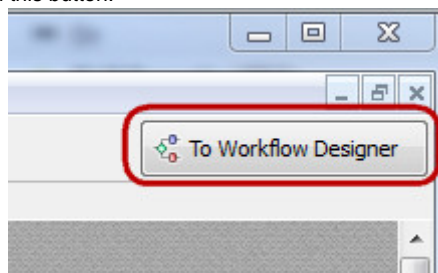
or on the *Dashboard* toolbar:



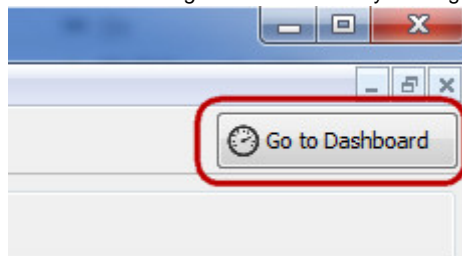
The following dialog appears:



To see a dashboard select it and check it's checkbox. To remove a dashboard select it and click the *Remove selected* button. Click OK button. The selected and checked dashboards appears in the *Dashboard* main window. You can go back to the *Workflow Designer* main window from *Dashboard* window by clicking on this button:

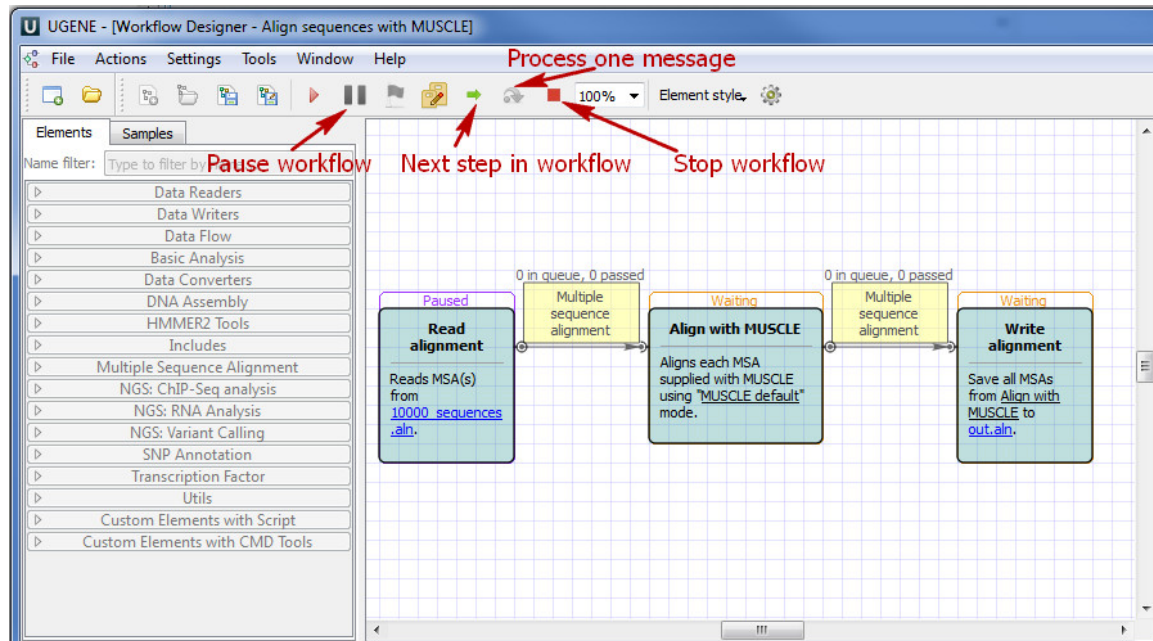


And go back to the *Dashboard* main window from *Workflow Designer* main window by clicking on this button:



## Stopping and Pausing Workflow

A workflow execution can be stopped, paused and run step by step. After you run workflow the following toolbar buttons appears:



With a help of these buttons you can:

*Pause workflow* - pause the runned workflow.

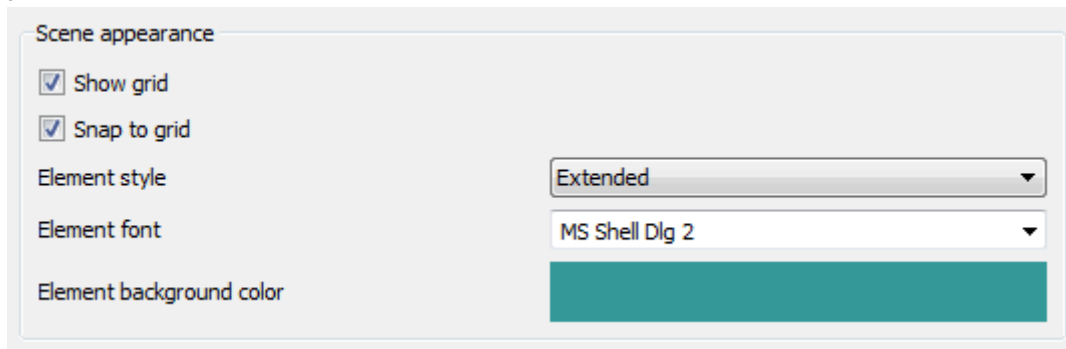
*Next step in workflow* - do the next step in workflow.

*Process one message* - do the first queue message step of the selected element in workflow. It is active if an element selected.

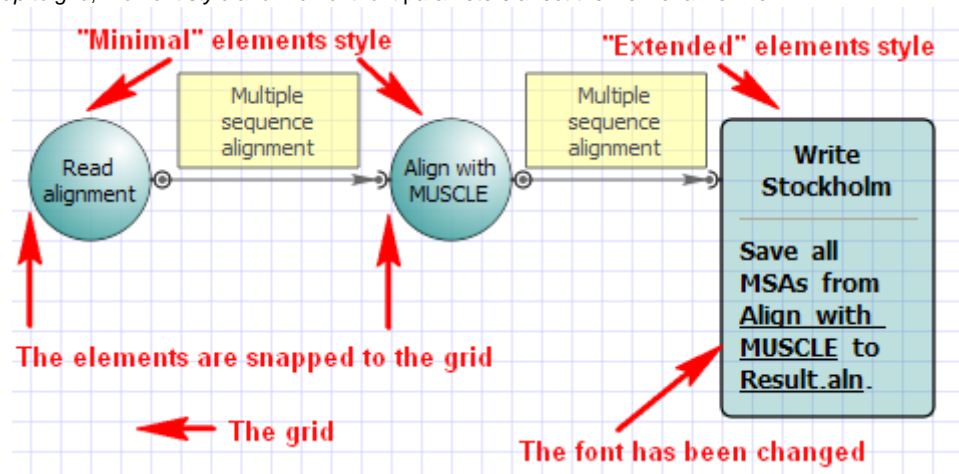
*Stop workflow* - cancel workflow process.

## Changing Appearance

Default setting that influence the Workflow Designer appearance can be set in the [Application Settings](#) dialog. The parameters are shown on the image below:

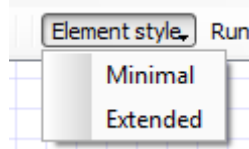


The *Show grid*, *Snap to grid*, *Element style* and *Element font* parameters affect the view of a workflow:



To change an appearance of a particular element use it's context menu submenus *Item properties* and *Item style*.

Another way to change an element style is to use the *Item style* submenu in the toolbar.



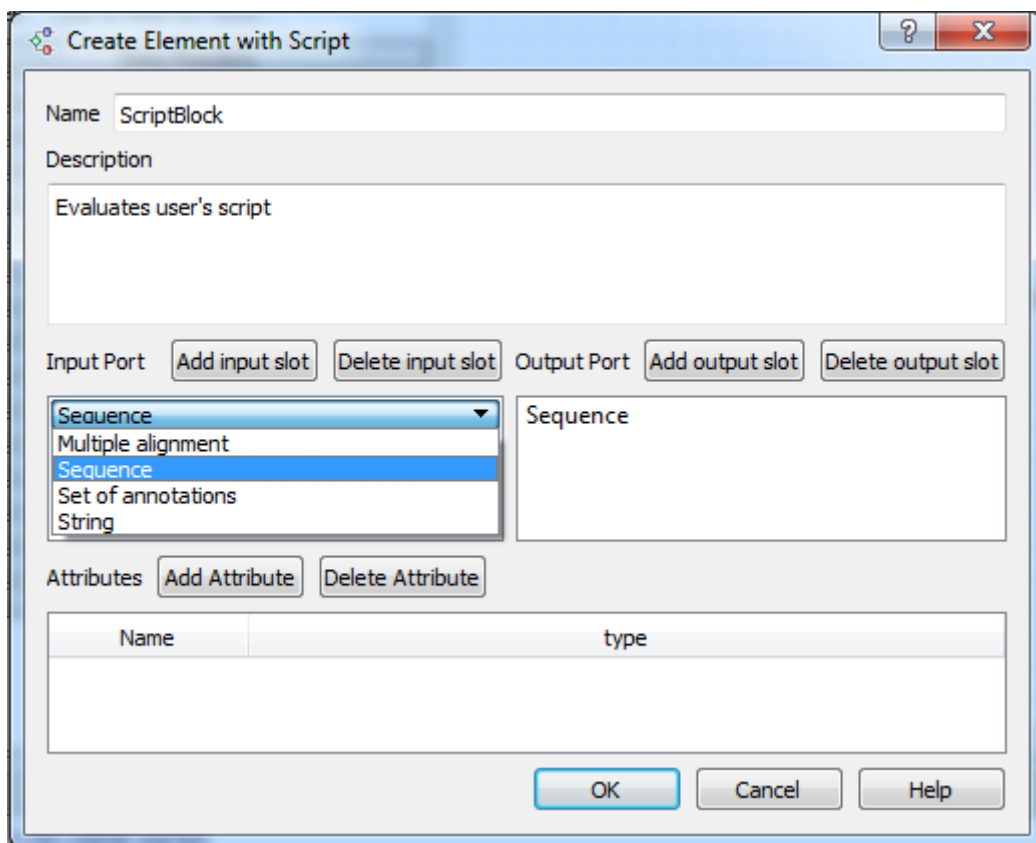
## Custom Elements with Scripts

It is possible to create custom algorithmic blocks using scripts in the Workflow Designer.

To create an element either select *Actions Create Script Object* in the main menu, select *Create element with script* in the context menu or click on the following button on the toolbar:



The *Create Element with Script* dialog will appear:



Here you should set the name of the element, its description and input / output ports of the element. It is possible to create a port with several input / output slots.

There are 4 types of data for a slot available:

- Multiple alignment
- Sequence
- Set of annotations
- Files

You can also add an attribute. The following types are supported for attributes:

- String
- Number
- Boolean

The element created is stored in a directory that can be set in the *Application Settings* dialog.

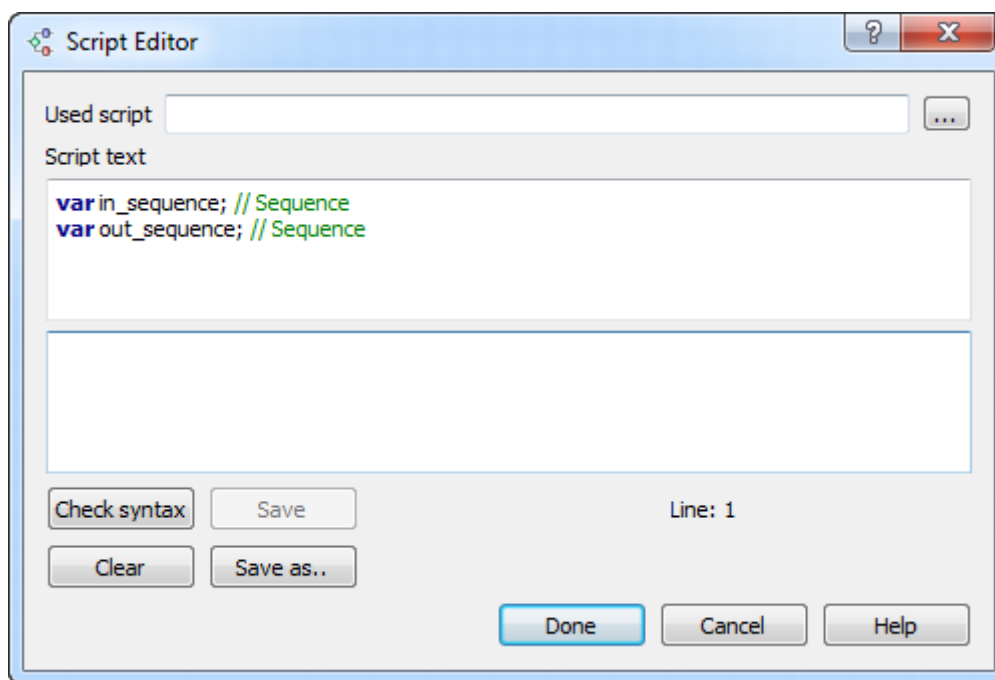
The element also becomes available in the *Custom Elements with Scripts* group on the *Palette*.

It is required to write a script for the element. Supported languages for the script are languages based on the ECMAScript (Javascript, QtScript).

To edit the script select the element on the *Scene* and either select *Actions Edit script of the element* in the main menu, use the *Edit script of the element* item in the context menu or click on the following button on the toolbar:



The *Script editor* dialog will appear:



As you can see there are predefined variables for the ports and the attributes in the script. The variables for the input slots begin with the "in\_" prefix, variables for the output slots begin with the "out\_" prefix. It is possible to load a script from a file (use the *Used script* field to do it).

For each supported data type UGENE provides a number of functions that can be used in the scripts.

- [Functions Supported for Multiple Alignment Data](#)
- [Functions Supported for Sequence Data](#)
- [Functions Supported for Set of Annotations Data](#)
- [Functions Supported for Files](#)
- [Common Function](#)

## Functions Supported for Multiple Alignment Data

- **createAlignment** (Sequence seq1, Sequence seq2, ...) — returns the alignment created from the sequences.
- **addToAlignment** (MAAlignment aln, Sequence seq, int row = -1) — adds the sequence to the specified row of the alignment. If the "row" parameter is not specified the sequence is added to the end of the alignment.
- **sequenceFromAlignment** (MAAlignment aln, int row) — returns the sequence from the specified row of the alignment.
- **findInAlignment** (MAAlignment aln, Sequence seq) — searches the alignment for the specified string. Return the number of the row if the sequence has been found or "-1" if it hasn't been found.
- **findInAlignment** (MAAlignment aln, QString name) — searches the alignment for a sequence with the specified name.
- **removeFromAlignment** (MAAlignment aln, int row) — removes a sequence from the specified row of the alignment.
- **rowNum** (MAAlignment aln) — returns the number of rows in the alignment.
- **columnNum** (MAAlignment aln) — returns the length of the alignment.
- **alignmentAlphabetType** (MAAlignment aln) — returns the alignment's alphabet.

## Functions Supported for Sequence Data

- **subsequence** (Sequence seq, int beg, int end) - returns the subsequence between the "beg" and "end" parameters.
- **complement** (Sequence seq) - returns the complement sequence.
- **translate** (Sequence seq, int offset = 0) - returns one of the three sequence translations. Which one is returned is determined by the "offset" parameter.
- **size** (Sequence seq) - returns the length of the sequence.
- **getName** (Sequence seq) - returns the name of the sequence.
- **alphabetType** (Sequence seq) - returns the alphabet of the sequence.
- **charAt** (Sequence seq, int ind) - returns the symbol located in the "ind" position of the sequence.
- **hasQuality** (Sequence seq) - determines whether the sequence has the "Quality" parameter.
- **getMinimumQuality** (Sequence seq) - returns the minimum value of the "Quality".
- **isAmino** (Sequence seq) - returns true if it is amino acid sequence.
- **concatSequence** (Sequence1 seq1, Sequence2 seq2,...) - returns the one sequence consists of the all input sequences.

- **sequenceFromText** (QString " ") - returns the sequence consists of the input text.

## Functions Supported for Set of Annotations Data

- **annotatedRegions** (Sequence seq, AnnotationTable anns, QString name) — returns subsequences of the annotations with the specified "name".
- **addQualifier** (AnnotationTable anns, QString qual, QString val, QString name = "") — sets the qualifier in the annotations with the specified "name" to the specified value. If the "name" is not specified, then all annotations are taken into account.
- **getLocation** (AnnotationTable anns, int ind) — returns the annotation location with the specified index.
- **filterByQualifier** (AnnotationsTable anns, QString qual, QString val) - returns the qualifier with the specified value.
- **hasAnnotationName** - (AnnotationsTable anns, QString " ") - returns the annotation with the specified name there is or there is not.

## Functions Supported for Files

- **writeFile** (QString url, QString " ") - writes the specified text data to the file with specified url.
- **appendFile** (QString url, QString " ") - appends the specified text data to the end of the file with the specified url.
- **readFile** (QString url) - reads the file with the specified url.

## Common Function

- **printToLog** (parameter) - prints the results to the [Log View](#).

## Custom Elements with Command Line Tools

In UGENE you can create a custom workflow *element* that would launch any command line tool.

- Creating Element
- Editing Element
- Adding Existent Element
- Removing Element

### Creating Element

To create an element for a command line tool select either *Actions* *Create element with command line tool* in the main menu or the following icon on the toolbar:



The *Create Element with Command Line Tool* wizard appears. On the first page of the wizard input a name and a description of the element in the *Property Editor*. Letters, numbers and underscores are allowed in the name.

On the second page add the required input and output data:



Input and output data for external tool. Name is a command line parameter for input/output data in external tool. Set data type and format in which external tool reads/writes input/output data. You also can set description for workflow designer. Each input data will be represented as port in workflow designer. Each output data will be represented as slot of single slot.

**Input data**

Name for command	Type	Read as	Description

Add input Delete

**Output data**

Name for command	Type	Write as	Description

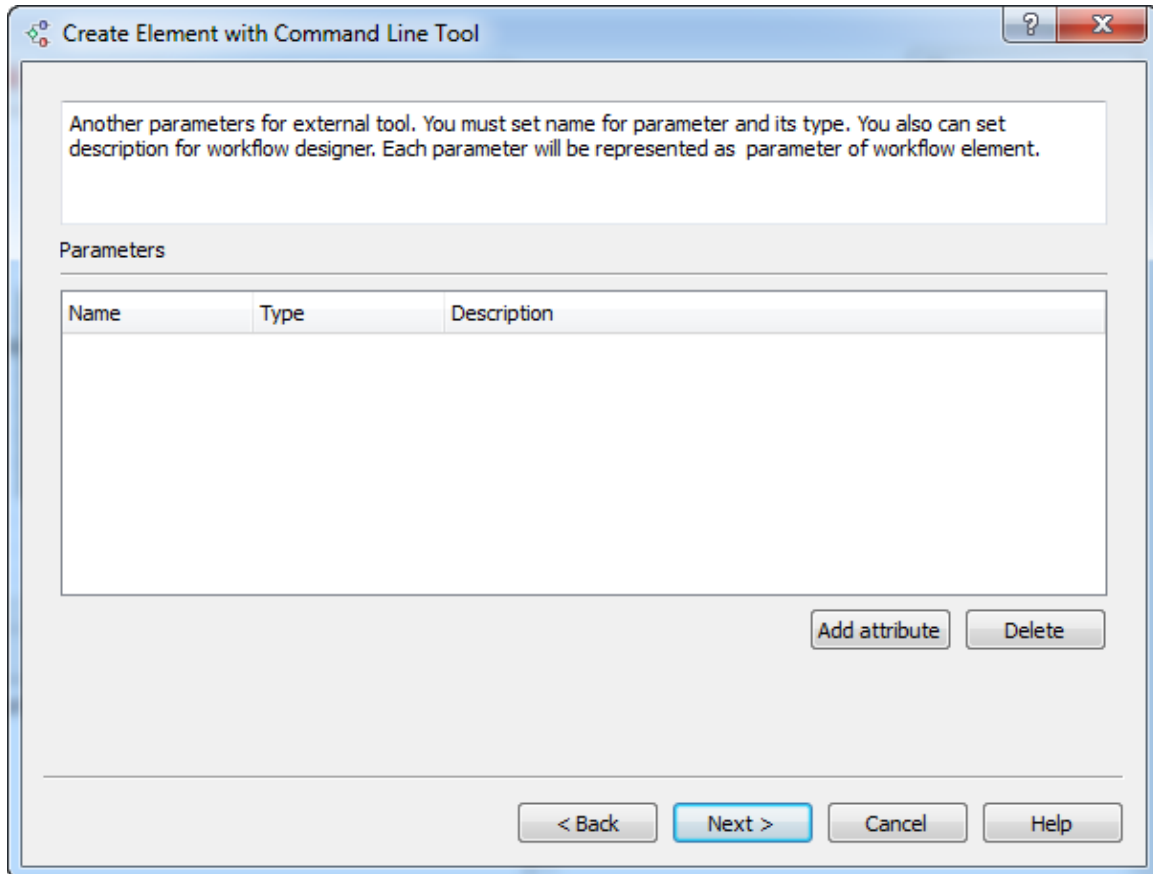
Add output Delete

< Back Next > Cancel Help

For each input or output you should:

- Input a name (letters, numbers and underscores are allowed in the name).
- Select a type: multiple alignment, sequence, sequence with annotations, a set of annotations or string.
- Specify how the input or output should be handled (for example, you can specify that a value of the input parameter should be handled as a FASTA file).
- Optionally input a description.

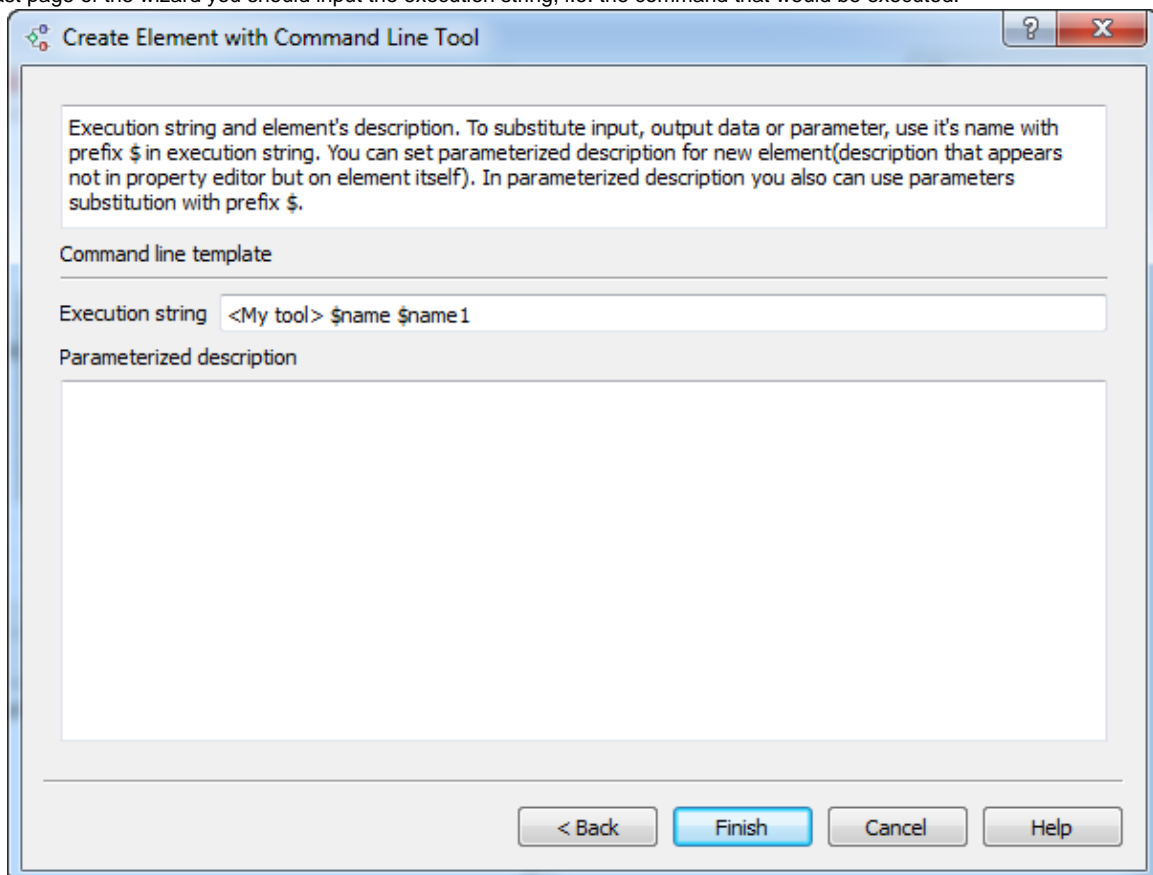
On the third page of the wizard you can add attributes for the command line tool. Later you would be able to set values for the attributes in the Property Editor, i.e. the attributes are actually the parameters of the new element.



For each attribute added you should:

- Input a name (letters, numbers and underscores are allowed in the name).
- Select it's type: boolean, number, string or URL.
- Optionally input the description.

On the last page of the wizard you should input the execution string, i.e. the command that would be executed.



The signature of the execution string depends on the command that is launched. But the general rule is that input/output data and attributes have prefix \$. For example let there be some perl script “myScript.pl” that accepts an input file as the first attribute and accepts the second attribute denoted as “param1”. The command may look as follows:

```
perl [path_to_script]myScript.pl $infile $param1 > $outfile
```

Here *infile* and *outfile* are input and output data set on the step 2, *param1* is an attribute set on the step 3.

On the same wizard page you can optionally input the description of the element. It would be shown on the element on the *Scene*. The description can be parameterized. This means that if you input e.g. an attribute name (with prefix \$), the name on the element would be substituted with the value of the corresponding parameter.

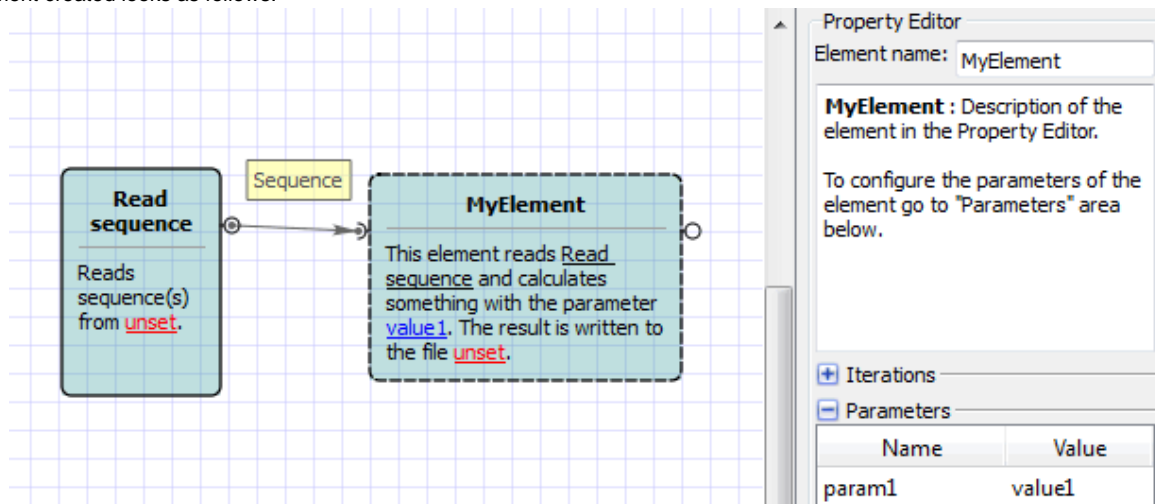
For example input the following parameters:

Execution string `perl C:\myScript.pl $infile $param1 > $outfile`

Parameterized description

This element reads \$infile and calculates something with the parameter \$param1. The result is written to the file \$outfile.

The element created looks as follows:



## Editing Element

The element created appears in the *Custom Elements with CMD Tools* group on the *Palette*.

To edit an element select the *Edit* item in it's context menu in the *Palette* or select the *Edit configuration* item in it's context menu on the *Scene*. The creation element wizard would appear.

## Adding Existent Element

The elements are stored in the files with the .etc extension.

The directory to store the elements can be set in the *Application Settings* dialog.

To add an element from a file to the *Workflow Designer* select either *Actions Add element with command line tool* in the main menu or the following icon on the toolbar:



In the appeared dialog select the required .etc file. The element is added to the group on the *Palette* and appears on the *Scene*.

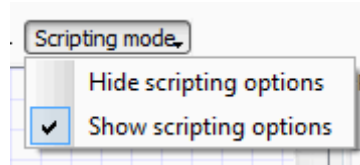
## Removing Element

To remove an element right-click on it and select the *Remove* item in the element's context menu. The corresponding .etc file is also removed in this case.

## Using Script to Set Parameter Value

When you select an element the *Parameters* area of the *Property Editor* displays two columns: *Name* and *Value*.

Select the *Show scripting options* item in the *Scripting mode* menu on the toolbar or in the *Actions* main menu.



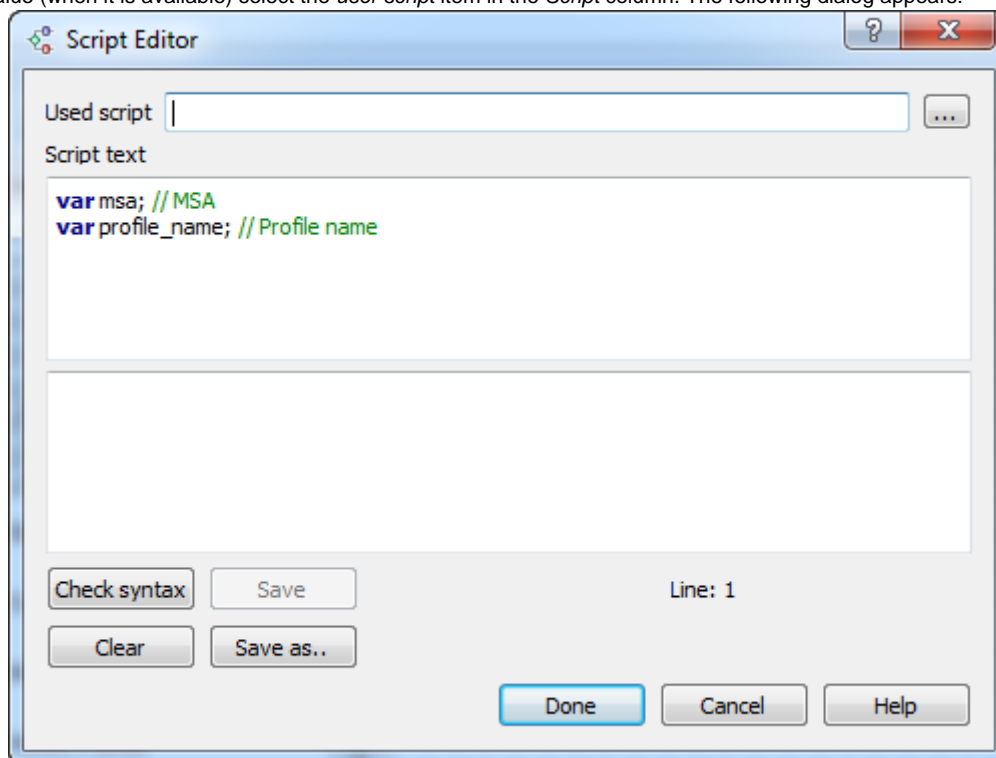
You can see that the third column *Script* has appeared in the *Parameters* area, for example:

Parameters		
Name	Value	Script
Accumulate objects	True	N/A
Document format	fasta	no script
Output file		no script
Existing file	Rename	no script

A script value can either be:

- not available for a parameter (*N/A* value)
- not set (*no script*)
- set by user (*user script*)

To set a script value (when it is available) select the *user script* item in the *Script* column. The following dialog appears:



Here you can see the variables available from the dataflow and can write your script. Supported languages for the script are languages based on the ECMAScript (JavaScript, QtScript).

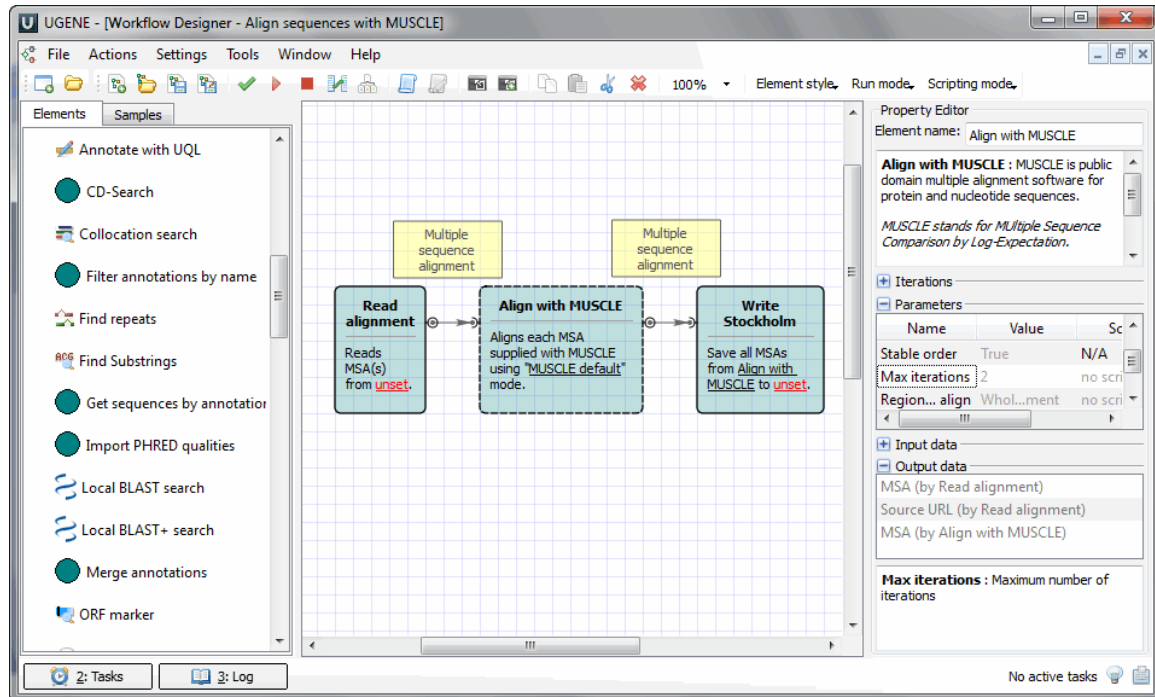
## Running Workflow from the Command Line

UGENE provides command line interface (CLI). To learn more about UGENE CLI and commands available read [main UGENE User Manual](#).

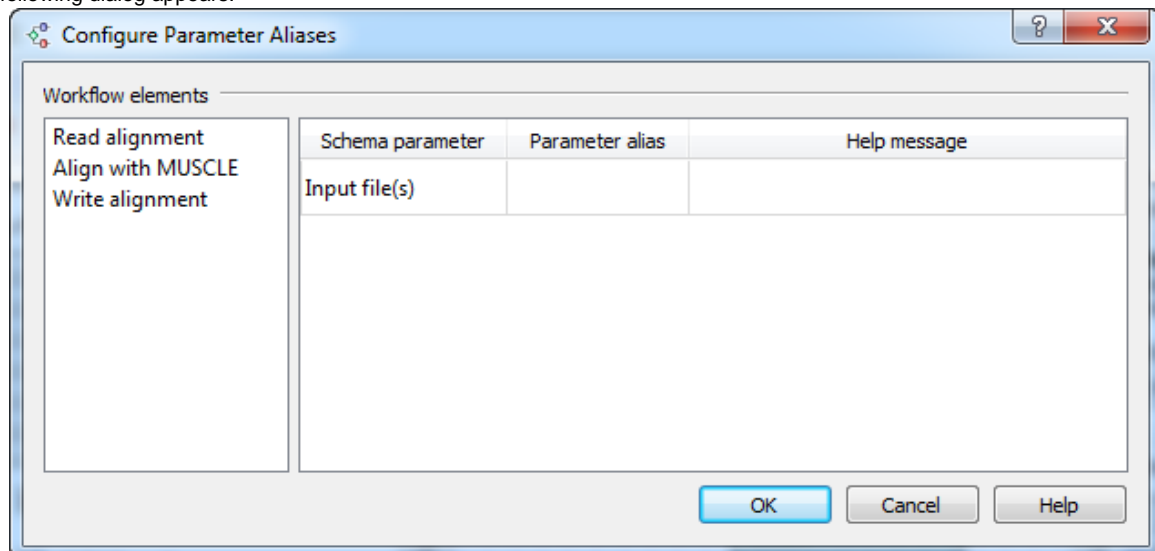
This chapter describes how you can create a new command using a *workflow*.

To run a workflow from the command line do the following:

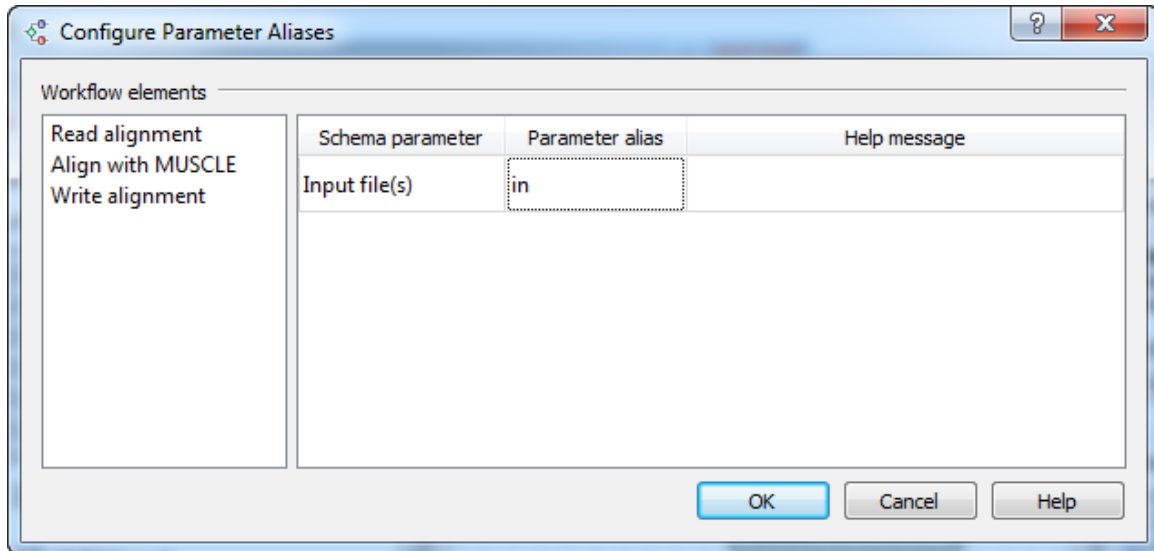
- Create the workflow in the Workflow Designer. For example on the image below the *Align sequences with MUSCLE* sample workflow is used:



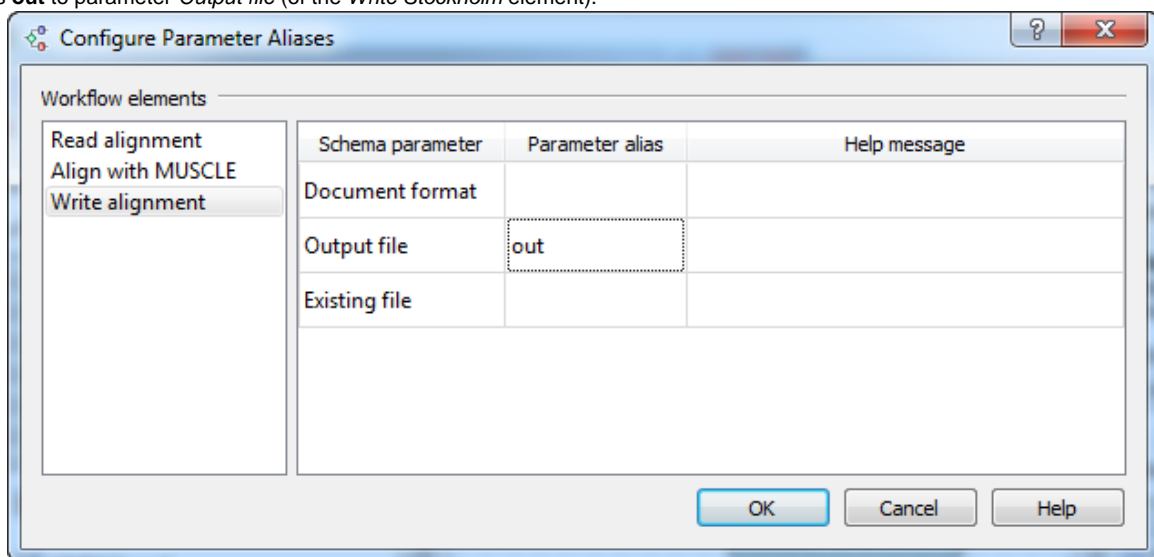
- Now you should configure aliases for those parameters and ports and slots that you are going to use from the command line. To do it select the *Actions* *Configure parameter aliases* item in the main menu or the *Configure parameter aliases* toolbar button. The following dialog appears:



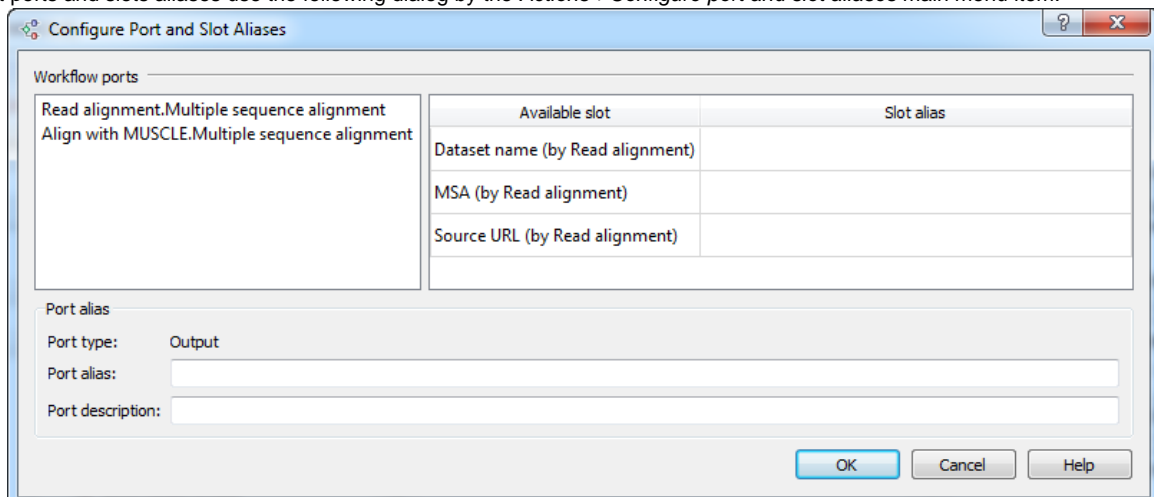
It contains the list of objects that corresponds to the *elements* of the workflow. For each object the list of parameters is available for which you can assign command line aliases. For example, assign alias **in** to parameter *Input file* (of the *Read alignment* element):



And alias **out** to parameter *Output file* (of the *Write Stockholm* element).



To select ports and slots aliases use the following dialog by the *Actions->Configure port and slot aliases* main menu item:



Press the *Ok* button to save aliases and close the dialog. When you create aliases you can import workflow to element by the *Actions->Import workflow to element* main menu item.

- *Save the workflow* to a file: if you follow the example, choose the *Actions Save workflow as...* item in the main menu, browse for the file location and enter **mySchema** as the workflow name. This name will be used to launch the workflow from the command line.
- Launch the workflow from the command line:

```
[path_to_ugene\]ugene --task={schema_name} [--{parameter1}={value1}  
[--{parameter2}={value2} ...]]
```

The run information will be saved into the text file. By default it is the working directory.

For example on Windows the command can look as follows:

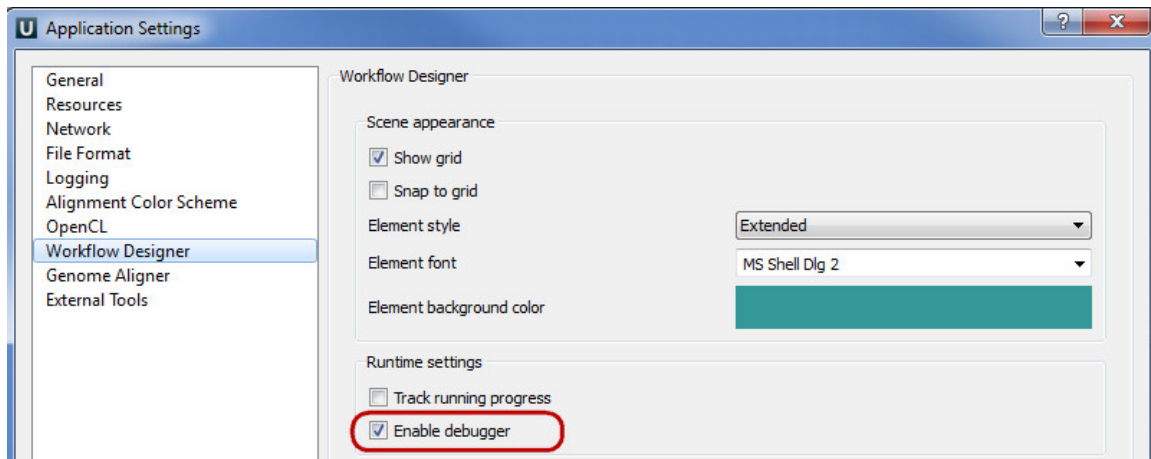
```
ugene --task=C:\mySchema --in=C:\COI.aln --out=C:\COI.sto
```



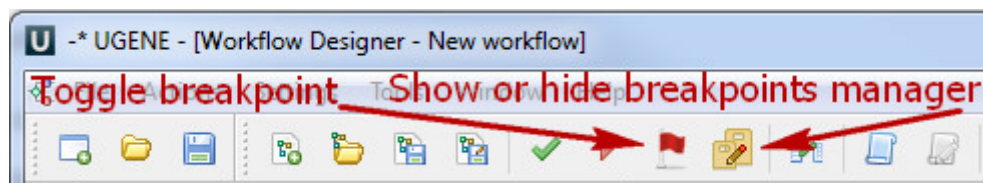
In this example the path to the directory with the UGENE executable is added to the system PATH variable.

## Running Workflow in Debugging Mode

By default a *workflow* runs without debugging settings. To use it go to the *Application Settings* (Settings→Preferences) and check the following checkbox and click *OK*:



After that the two new buttons appears on the main toolbar:

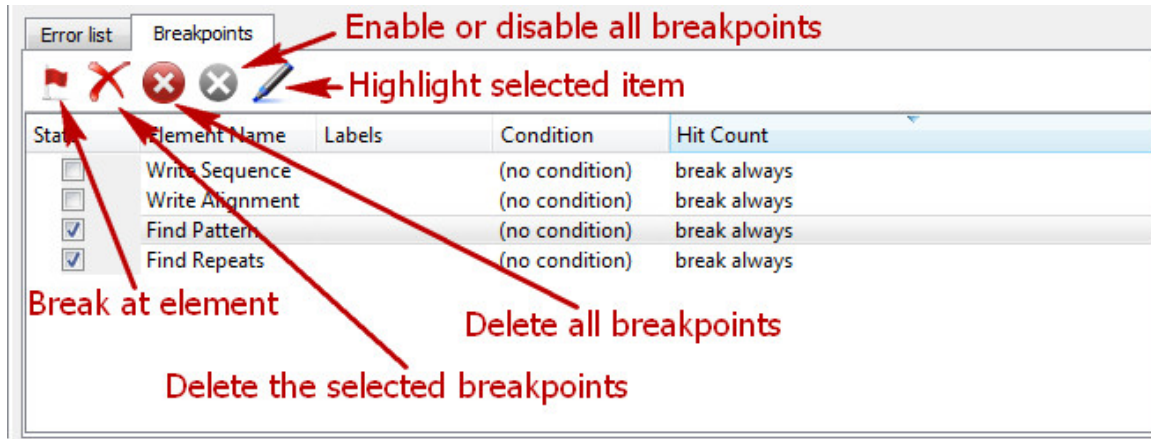


- Creating Breakpoints
- Manipulating Breakpoints

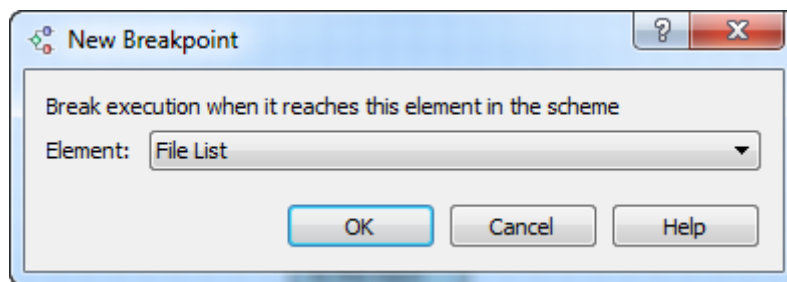
## Creating Breakpoints

You can create a pause element in a workflow with a help of the *Toggle breakpoint* button or by the *Ctrl+B* shortcut. To do it select the element and press this button. If you press the *Show or hide breakpoint manager* the breakpoint manager appears:





*Break at element* - creates new breakpoint. If you press on this button the following dialog will appear. Choose the breakpoint element and click OK button.



*Delete the selected breakpoints* - this button deletes the selected breakpoint.

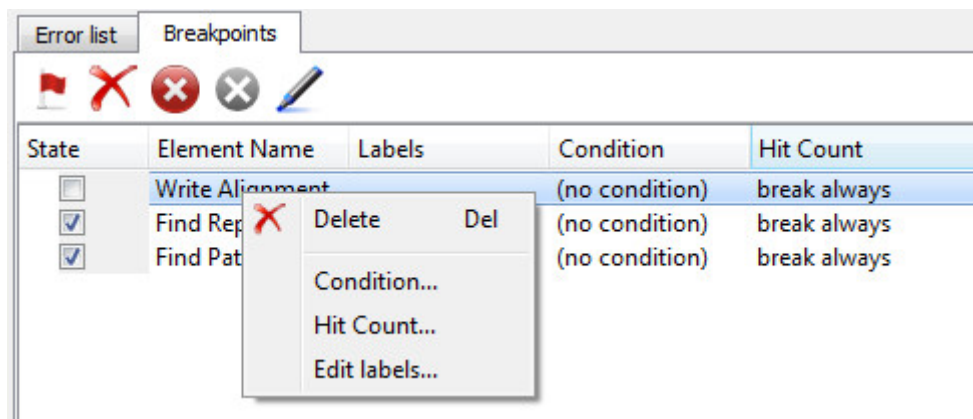
*Delete all breakpoints* - this button deletes all breakpoints.

*Enable or disable all breakpoints* - this button check or uncheck all breakpoints. Check on the breakpoint means that the breakpoint enable and will be used.

*Highlight selected item* - this button highlights the breakpoint element.

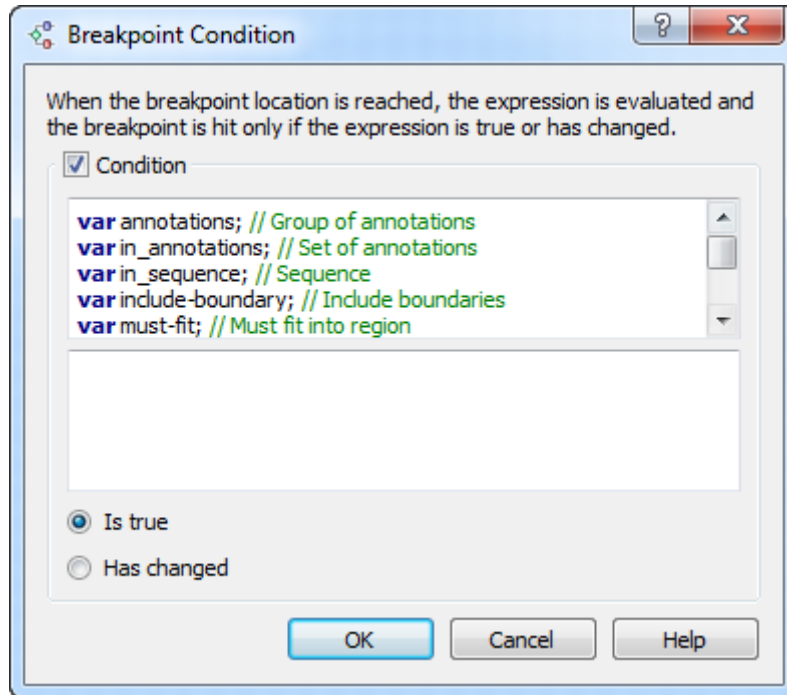
## Manipulating Breakpoints

The following operations are available for each breakpoint:



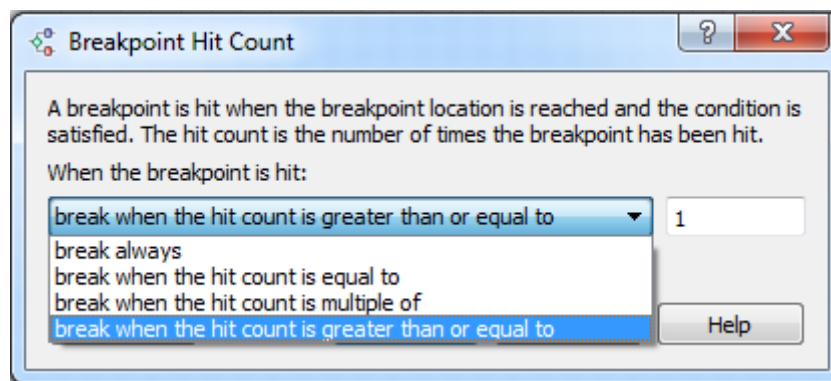
*Delete* - delete the selected breakpoint.

*Condition* - creates a breakpoint condition. Click on this menu item and the following dialog appears:



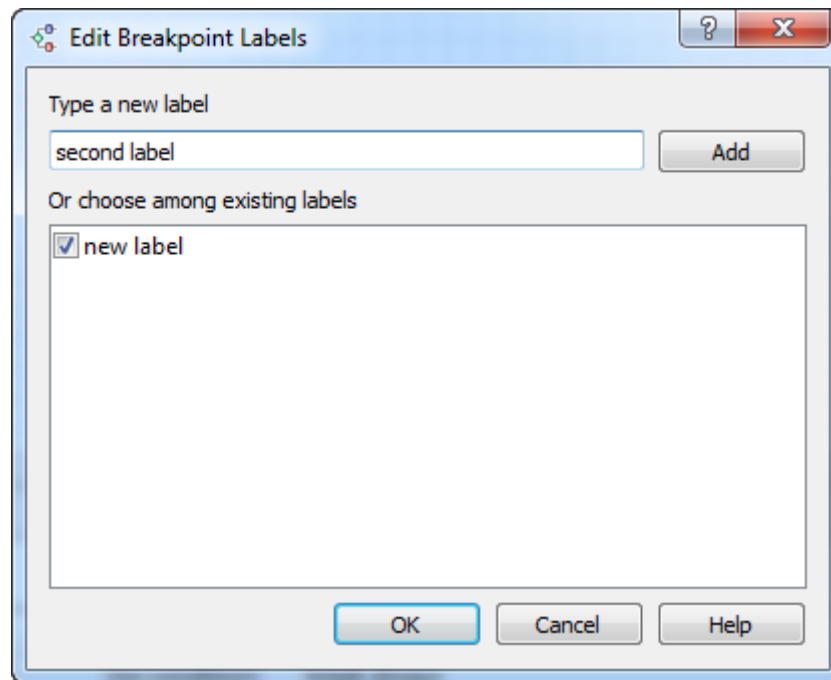
When the breakpoint location is reached, the expression is evaluated and the breakpoint is hit only if the expression is true or has changed.

*Hit Count* - breakpoint hit count. Click on this menu item and the following dialog appears:



A breakpoint is hit when the breakpoint location is reached and the condition is satisfied. The hit count is the number of times the breakpoint has been hit.

*Edit labels* - allows to add breakpoint labels. Click on this menu item and the following dialog appears:



## Workflow File Format

Using the GUI is not the only way to create/edit a *workflow workflow*. A workflow is saved to a file with .uwl extension. The format of the file is human-readable. This chapter describes this format and explains how you can create/edit a workflow file using a text editor.

The best way to learn workflow workflow file format is to study an existent .uwl file. The file consists of the header and the body. Check the description of each part below.

- [Header](#)
- [Body](#)

### Header

The header consists of the following key string:

```
#!UGENE_WORKFLOW
```

And multiline description of the workflow:

```
# Write here the description
# of your workflow.
```

### Body

The body begins with the **workflow** keyword followed by the name of the workflow and curly braces:

```
workflow schema_name {

    # Description of the elements
    # Description of the dataflow
    # Description of the iterations
    # Metainformation (aliases and visual information)

}
```

- [Elements](#)
- [Dataflow](#)
- [Metainformation](#)

### Elements

Each *element* used in the *workflow* must be described inside the body. An element description consists of the element name and a set of parameters enclosed in curly braces. A parameter and the value are separated by ':', different parameters are separated by ';':

```
element_name {

    parameter1:value1;
    parameter2:value2;
    ...

}
```

See, for example, a description of the [Read alignment](#) element:

```
read-msa {
  type:read-msa;
  name:"Read alignment";
  url-in:/home/user/pkinase.sto;
}
```

Note, that the values of the parameters for an element can also be presented in the [iterations](#) block. For all elements the following parameters are defined:

- **type** - specifies the type of the element.
- **name** - specifies the name of the element. It corresponds to the element's name in the GUI
- **.validator** - validates the element by the input validator type's parameters:
  - **type** - specifies the type of the validator.

For example this validator validate that the read sequence element has two or three datasets:

```
read-sequence {
  type:read-sequence;
  name:"Read Sequence";
  .validator {
    type:datasets-count;
    min:2;
    max:3;
  }
}
```

For *custom elements* there is special parameter:

- **script** - sets the script text of the element, for example:

```
dump-info {
  type:"Script-Dump sequence info"
  name:"Dump sequence info"
  script {
    out_text=getName(in_sequence) + ": " + size(in_sequence);
  }
}
```

The list of parameters available depend on an element. Refer to the [Workflow Elements](#) chapter to find out the parameters for a particular element. To [set a script value for a parameter](#) use the following form:

```
parameter_name {
  a script value
};
```

## Dataflow

The description of the elements is followed by the description of their connections to each other, i.e. the dataflow. For ports connections the description starts with the **.actor-bindings** keyword and has the following format:

```
.actor-bindings {
  element1_name.output_port1_name->element2_name.input_port2_name;
}
```

This pair says that data from port 1 of *element1* will be transferred to *port2* of *element2*. For slots the following format without start keyword is used:

```
element1_name.slot1_name->element2_name.port2_name.slot2_name
```

This pair says that data from *slot1* of *element1* will be transferred to *slot2* of *port2* of *element2*. See, for example, the minimum description of a dataflow of a workflow, that aligns an input MSA and writes the result to a file in ClustalW format.

```
.actor-bindings {
    read-msa.out-msa->muscle.in-msa
    muscle.out-msa->write-msa.in-msa
}
read-msa.msa->muscle.in-msa.msa
muscle.msa->write-msa.in-msa.msa
```

## Metainformation

A metainformation block sets visual parameters of the workflow and aliases for running it from the command line.

Each block starts with **.meta** keyword and consists of the aliases and visual blocks:

```
.meta {
    aliases {
        # The workflow aliases
    }
    visual {
        # Visual data for element1
        # Visual data for element2
        # ...
    }
}
```

### Parameter Aliases

The block starts with the **parameter-aliases** keyword and has the following format:

```
parameter-aliases {
    element_name.parameter_name:value;
    ...
}
```

The value specified for an element parameter is used as the alias for this parameter when the workflow is *executed from the command line*.

See an example of setting workflow aliases:

```
.meta {
    parameter-aliases {
        read-msa.url-in:in;
        write-msa.url-out:out;
    }
    ...
}
```

## Visual

The block starts with the **visual** keyword. It describes the appearance of the workflow in a Workflow Designer window, i.e. appearance of the workflow *elements* and *connections*:

```
visual {

    # Elements appearance
    element_name1 {
        element_appearance_parameter1:value1;
        element_appearance_parameter2:value2;
        ...
    }
    element_name2 {
        ...
    }
    ...

    # Connections appearance
    element1_name.port1_name->element2_name.port2_name {
        connection_appearance_parameter1:value3;
        ...
    }
    ...
}
```

To describe an element appearance the following parameters are used:

- **description** — description of the element in the *Property Editor*. It is in HTML format.
- **tooltip** — tooltip shown on the element.
- **pos** — position of the element, assuming that bottom right corner of the window is (0, 0) position.
- **style** — style of the element. The following values are available:
  - **ext** — for extended element style
  - **simple** — for minimal element style
- **bounds** — defines the bounds of the element rectangle in the extended style.
- **bg-color-ext** — color of the element in the extended style. The color must be specified in the RGBA format.
- **bg-color-simple** — color of the element in the minimal style.
- **port\_name.angle** — position of the port on the element. Here the *port\_name* must be replaced by the name of the port.

For now, the only parameter that describes a connection appearance is:

- **text-pos** — position of the text near the connection arrow.

For example:

```
visual {  
  read-sequence {  
    description:"";  
    tooltip:"Reads sequences and annotations ...";  
    pos:"-930 -885";  
    style:ext;  
    bg-color-ext:"0 128 128 64";  
    bounds:"-30 -30 45 103";  
    out-sequence.angle:272.309;  
  }  
  write-sequence {  
    ...  
  }  
  read-sequence.out-sequence->write-sequence.in-sequence {  
    text-pos:"-27.5 -24";  
  }  
}
```



## Workflow Elements

This section contains detailed description of all workflow elements presented in the Workflow Designer.

For each element you can find:

- Description of the parameters used in the GUI
- Corresponding parameters names used in a workflow file
- Information about input and output ports

The type of a parameter can be one of the following:

### **string**

A string.

### **numeric**

A number.

### **boolean**

A boolean data type. Available values are: true / false, 0 / 1 and yes / no.

A port's slot type can be one of the following:

### **sequence**

Biological sequence

### **msa**

Multiple sequence alignment

### **text**

A text

### **annotation-table**

Table of annotations

### **annotation-table-list**

A list of different tables of annotations

### **ebwt-index**

Bowtie index

### **hmm2-profile**

A HMM profile of HMMER2 package

### **fmatrix**

Frequency matrix

### **wmatrix**

Weight matrix

### **sitecon-model**

SITECON model

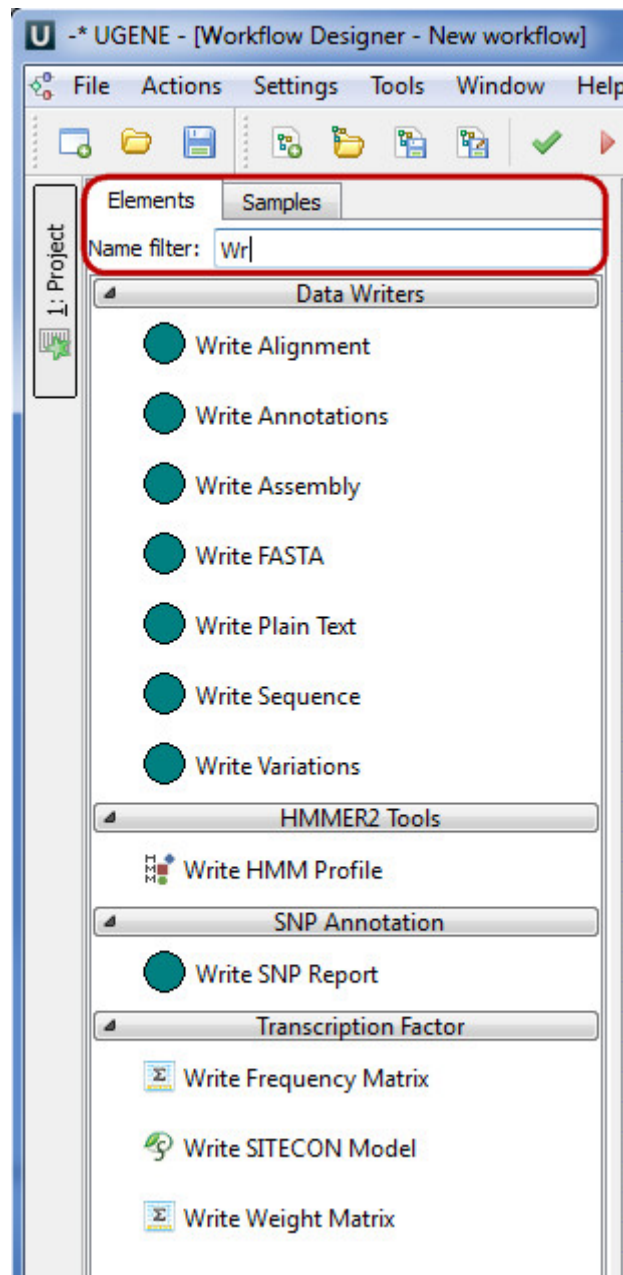
### **assembly**

Assembly

### **variation**

Variation track

To search an element use the name filter or press the *Ctrl+F* shortcut that moves you to the name filter also:



- Data Readers
  - File List Element
  - Read Alignment Element
  - Read Annotations Element
  - Read Assembly Element
  - Read from DAS Element
  - Read from Remote Database Element
  - Read Plain Text Element
  - Read Sequence Element
  - Read Variations Element
- Data Writers
  - Write Alignment Element
  - Write Annotations Element
  - Write Assembly Element
  - Write FASTA Element
  - Write Plain Text Element
  - Write Sequence Element
  - Write Variations Element
- Data Flow
  - Filter Element
  - Grouper Element
  - Multiplexer Element
  - Sequence Marker Element
- Basic Analysis
  - Amino Translations Element
  - Annotate with DAS Element
  - Annotate with UQL Element

- CD-Search Element
- Collocation Search Element
- Export PHRED Qualities Element
- Fetch Sequences by ID From Annotation Element
- Filter Annotation by Name Element
- Filter Annotations by Qualifier
- Find Pattern Element
- Find Repeats Element
- Gene-by-gene approach report
- Get Sequences by Annotations Element
- Import PHRED Qualities Element
- Local BLAST Search Element
- Local BLAST+ Search Element
- Merge Annotations Element
- ORF Marker Element
- Remote BLAST Element
- Remove Duplicates in BAM Files Element
- Smith-Waterman Search Element
- Data Converters
  - Convert bedGraph Files to bigWig Element
  - Convert Text to Sequence Element
  - File Format Conversion Element
  - Reverse Complement Element
  - Split Assembly into Sequences Element
- DNA Assembly
  - Align reads with BWA-MEM
  - Assembly Sequences with CAP3
  - Extract Consensus from Assembly
- HMMER2 Tools
  - HMM Build Element
  - HMM Search Element
  - Read HMM Profile Element
  - Write HMM Profile Element
- HMMER3 Tools
  - HMM3 Build Element
  - HMM3 Search Element
  - Read HMM3 Profile
  - Write HMM3 Profile
- Multiple Sequence Alignment
  - Align Profile to Profile with MUSCLE Element
  - Align with ClustalO Element
  - Align with ClustalW Element
  - Align with Kalign Element
  - Align with MAFFT Element
  - Align with MUSCLE Element
  - Align with T-Coffee Element
  - Extract Consensus from Alignment
  - Join Sequences into Alignment Element
  - Split Alignment into Sequences Element
- NGS Basic
  - CASAVA FASTQ Filter Element
  - FASTQ Quality Trimmer Element
  - Filter BAM/SAM Files Element
  - Genome Coverage Element
  - Merge BAM Files Element
  - Slopbed Element
  - Sort BAM Files Element
- NGS: ChIP-Seq Analysis
  - Annotate Peaks with peak2gene Element
  - Build Conservation Plot Element
  - Collect Motifs with SeqPos Element
  - Conduct GO Element
  - Create CEAS Report Element
  - Find Peaks with MACS Element
- NGS: RNA-Seq Analysis
  - Assembly Transcripts with Cufflinks Element
  - Extract Transcript Sequences with gffread Element
  - Find Splice Junction with TopHat Element
  - Merge Assemblies with Cuffmerge Element
  - Test for Diff. Expression with Cuffdiff Element
- NGS: Variant Calling
  - Call Variants with SAMtools Element
  - Create VCF consensus
- SNP Annotation
  - Annotate variations with SNPToolbox Element
  - Detect Transcription Factors with rSNP-Tools Element
  - Determine SNP effect on TATA-boxes Element
  - ProtStability1D Element
  - ProtStability3D Element
  - SNP Chip Tools Element

- SNP Effect on PDB sites Element
- Write SNP Report Element
- Transcription Factor
  - Build Frequency Matrix Element
  - Build SITECON Model Element
  - Build Weight Matrix Element
  - Convert Frequency Matrix Element
  - Read Frequency Matrix Element
  - Read SITECON Model Element
  - Read Weight Matrix Element
  - Search for TFBS with SITECON Element
  - Search for TFBS with Weight Matrix Element
  - Write Frequency Matrix Element
  - Write SITECON Model Element
  - Write Weight Matrix Element
- Utils
  - DNA Statistics Element
  - Generate DNA Element
- Custom Elements With Script
  - CASAVA FASTQ Filter Script Element
  - Dump Sequence Info Element
  - FASTQ Trimmer Element
  - LinkData Fetch Element
  - Quality Filter Element

## Data Readers

Data Readers *elements* read data (from files, remote databases, etc.) and provide them to other elements in a *workflows*.

- File List Element
- Read Alignment Element
- Read Annotations Element
- Read Assembly Element
- Read from DAS Element
- Read from Remote Database Element
- Read Plain Text Element
- Read Sequence Element
- Read Variations Element

## File List Element

Gets the list of files in the specified directories.

### Parameters in GUI

Parameter	Description	Default value
<b>Input directory</b>	Input directory.	
<b>Absolute output paths</b>	Specify whether to output absolute or relative paths of the files.	True
<b>Recursive reading</b>	Get files from all nested directories or just from the current one.	False
<b>Include name filter</b>	Filter files by the specified value. It can be, for example, a file name or a regular expression of the file name.	
<b>Exclude name filter</b>	Exclude files using the specified filter value. The value can be, for example, a file name or a regular expression of the file name.	

## Parameters in Workflow File

**Type:** get-file-list

Parameter	Parameter in the GUI	Type
<b>in-path</b>	<b>Input directory</b>	<i>string</i>

<b>absolute</b>	<b>Absolute output paths</b>	<i>boolean</i>
<b>recursive</b>	<b>Recursive reading</b>	<i>boolean</i>
<b>include-name-filter</b>	<b>Include name filter</b>	<i>string</i>
<b>exclude-name-filter</b>	<b>Exclude name filter</b>	<i>string</i>

## Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** *out-url*

**Name in Workflow File:** out-url

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	out-url	<i>string</i>

## Read Alignment Element

Reads multiple sequence alignments (MSAs) from local or remote files.

**Parameters in GUI**

Parameter	Description	Default value
<b>Input files</b> (required)	Semicolon-separated list of paths to the input files.	

## Parameters in Workflow File

**Type:** read-msa

Parameter	Parameter in the GUI	Type
url-in	Input files	<i>string</i>

## Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** *Multiple sequence alignment*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	<i>msa</i>
Source URL	url	<i>string</i>

## Read Annotations Element

Reads annotations from files.

**Parameters in GUI**

Parameter	Description	Default value
<b>Input file(s)</b>	Input files.	Dataset 1;

<b>Mode</b>	<p>If the file contains more than one annotation table, Split mode sends them "as is" to the output, while Merge appends all the annotation tables and outputs the sole merged annotation table.</p> <p>In Merge files is the same as Merge but it operates with all annotation tables from all files of one dataset.</p>	Merge
-------------	---	-------

## Parameters in Workflow File

**Type:** read-annotations

Parameter	Parameter in the GUI	Type
url-in	Input file(s)	string
mode	Mode	numeric

### Input/Output Ports

The element has 1 *output port*.

**Name in GUI:** Annotations

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table-list
Dataset name	dataset	string
Source URL	out-url	string

## Read Assembly Element

Reads assembly from files.

### Parameters in GUI

Parameter	Description	Default value
Input file(s)	Input files.	Dataset 1;

**Type:** read-assembly

Parameter	Parameter in the GUI	Type
url-in	Input file(s)	string

### Input/Output Ports

The element has 1 *output port*.

**Name in GUI:** Assembly

**Name in Workflow File:** out-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Assembly data	assembly	assembly

<b>Dataset name</b>	<b>dataset</b>	<i>string</i>
<b>Source URL</b>	<b>out-url</b>	<i>string</i>

## Read from DAS Element

Reads sequences and annotations if any from the Distributed Annotation System.

### Parameters in GUI

Parameter	Description	Default value
<b>Feature Sources</b>	The DAS sources to read features from.	InterPro-Matches-Overview, Pride DAS 1.6, UniProt, cbs_sort, signalp
<b>Reference Sources</b>	The DAS source to read reference from.	UniProt (DAS)
<b>Resource ID(s)</b>	Semicolon-separated list of resource ID's in the source.	
<b>Save file to directory</b>	The directory to store sequence files loaded from the source.	default

## Parameters in Workflow File

**Type:** fetch-das

Parameter	Parameter in the GUI	Type
<b>annotations</b>	<b>Feature Sources</b>	<i>string</i>
<b>database</b>	<b>Reference Sources</b>	<i>string</i>
<b>resource-id</b>	<b>Resource ID(s)</b>	<i>string</i>
<b>save-dir</b>	<b>Save file to directory</b>	<i>string</i>

### Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** Sequence

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table-list</i>
<b>Sequence</b>	<b>sequence</b>	<i>string</i>

## Read from Remote Database Element

Reads sequences and annotations if any from a remote database.

### Parameters in GUI

Parameter	Description	Default value
<b>Resource IDs (required)</b>	Semicolon-separated list of resource IDs in the database.	
<b>Database (required)</b>	Name of the database to read from.	NCBI Genbank (DNA sequence)
<b>Save file to directory</b>	Directory to store a file loaded from the database.	default

## Parameters in Workflow File

Type: fetch-sequence

Parameter	Parameter in the GUI	Type
resource-id	Resource IDs	string
database	Database	string  Available values are: <ul style="list-style-type: none"> <li>ncbi-dna (NCBI GenBank (DNA sequence))</li> <li>ncbi-protein (NCBI protein sequence database)</li> <li>pdb (PDB)</li> <li>swiss-plot (SWISS-PROT)</li> <li>uniprot-swiss-prot (UniProtKB/Swiss-Prot)</li> <li>uniprot-trembl (UniProtKB/TrEMBL)</li> </ul>
save-dir	Save file to directory	string

## Input/Output Ports

The element has 1 *output port*.

Name in GUI: *Sequence*

Name in Workflow File: out-sequence

Slots:

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Set of annotations	annotations	annotation-table

## Read Plain Text Element

Reads text from local or remote files.

Parameters in GUI

Parameter	Description	Default value
Input files (required)	Semicolon-separated list of paths to the input files.	
Read by lines (required)	Specifies to read the input file line by line.	false

## Parameters in Workflow File

Type: read-text

Parameter	Parameter in the GUI	Type
url-in	Input files	string
read-by-lines	Read by lines	boolean

## Input/Output Ports

The element has 1 *output port*.

Name in GUI: *Plain text*

Name in Workflow File: out-text



Slots:

Slot In GUI	Slot in Workflow File	Type
Plain text	text	string
Source URL	url	string

## Read Sequence Element

Reads sequences and annotations if any from local or remote files.

### Parameters in GUI

Parameter	Description	Default value
Input files	Semicolon-separated list of datasets to the input files.	
Mode	If the file contains more than one sequence, "split" mode sends them as is to output, while "merge" appends all the sequences and outputs the merged sequence.	Split
Merging gap	In the "merge" mode, inserts the specified number of gaps between the original sequences. This is helpful e.g. to avoid finding false positives at the merge boundaries.	10
Sequence count limit	Split mode only. Read only first N sequences from each file. Set 0 value for reading all sequences.	0
Accession filter	Only reports a sequence with the specified accession (id).	

## Parameters in Workflow File

Type: read-sequence

Parameter	Parameter in the GUI	Type
url-in	Input files	string
mode	Mode	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for split mode</li> <li>• 1 - for merge mode</li> </ul>
merge-gap	Merging gap	numeric
sequence-count-limit	Sequence count limit	numeric
accept-accession	Accession filter	string

## Input/Output Ports

The element has 1 *output port*.

Name in GUI: *Sequence*

Name in Workflow File: out-sequence

Slots:

Slot In GUI	Slot in Workflow File	Type
-------------	-----------------------	------

<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table</i>
<b>Source URL</b>	<b>url</b>	<i>string</i>

## Read Variations Element

Reads variations from files and produces variations tracks.

### Parameters in GUI

Parameter	Description	Default value
<b>Input file(s)</b>	Input file(s).	Dataset 1

## Parameters in Workflow File

**Type:** read-variations

Parameter	Parameter in the GUI	Type
<b>url-in</b>	<b>Input file(s)</b>	<i>string</i>

### Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** Variation track

**Name in Workflow File:** out-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Dataset name</b>	<b>dataset</b>	<i>string</i>
<b>Source url</b>	<b>url</b>	<i>string</i>
<b>Variation track</b>	<b>variation-track</b>	<i>variation</i>

## Data Writers

Data Writers *elements* write data supplied from other elements in a workflow to a file or files.

- [Write Alignment Element](#)
- [Write Annotations Element](#)
- [Write Assembly Element](#)
- [Write FASTA Element](#)
- [Write Plain Text Element](#)
- [Write Sequence Element](#)
- [Write Variations Element](#)

## Write Alignment Element

Writes all supplied alignments to file(s) in selected format.

### Parameters in GUI

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this parameter is set, then the “Location” slot is not taken into account.	

<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Document format</b>	Format of the output file.	clustal

## Parameters in Workflow File

Type: write-msa

Parameter	Parameter in the GUI	Type
url-out	Output file	string
write-mode	Existing file	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
document-format	Document format	string  Available values are: <ul style="list-style-type: none"> <li>• clustal</li> <li>• mega</li> <li>• msf</li> <li>• sam</li> <li>• srfasta</li> <li>• stockholm</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Multiple sequence alignment*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa
Location	url	string

## Write Annotations Element

Writes all supplied annotations to file(s) in the selected format.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b>	Location of the output data file. If this attribute is set, slot "Location" in port will not be used.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Document format</b>	Document format of output file.	genbank
<b>Annotations name</b>	Object name of the annotations.	unknown feature

<b>CSV separator</b>	String which separates values in CSV file(s).	"," (comma)
<b>Write sequence name</b>	Write sequence to CSV file(s).	False

## Parameters in Workflow File

Type: write-annotations

Parameter	Parameter in the GUI	Type
<b>url-out</b>	<b>Output file</b>	<i>string</i>
<b>write-mode</b>	<b>Existing file</b>	<i>numeric</i> Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
<b>document-format</b>	<b>Document format</b>	<i>string</i> Available values are: <ul style="list-style-type: none"> <li>• CSV</li> <li>• GenBank</li> <li>• GFF</li> </ul>
<b>annotations-name</b>	<b>Annotations name</b>	<i>string</i>
<b>separator</b>	<b>CSV separator</b>	<i>string</i>
<b>write_names</b>	<b>Write sequence name</b>	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input annotations*

**Name in Workflow File:** in-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table-list</i>
<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>
<b>Source URL</b>	<b>url</b>	<i>string</i>

## Write Assembly Element

Writes all supplied assemblies to file(s) in selected format.

**Parameters in GUI**

Parameter	Description	Default value
<b>Document format</b>	Document format of output file.	bam
<b>Build index (BAM only)</b>	Build BAM index for the target BAM file. The file .bai will be created in the same directory.	True
<b>Output file</b>	Location of output data file. If this attribute is set, slot "Location" in port will not be used.	

<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format). If Rename option is chosen existing file will be renamed.	Rename
----------------------	---	--------

### Parameters in Workflow File

**Type:** write-assembly

Parameter	Parameter in the GUI	Type
document-format	Document format	string
build-index	Build index (BAM only)	boolean
out-url	Output file	string
write-mode	Existing file	numeric

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Assembly

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Assembly data	assembly	assembly
Location	url	string

## Write FASTA Element

Writes all supplied sequences to file(s) in FASTA format.

### Parameters in GUI

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, then the "Location" slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Accumulate objects</b>	Accumulates all incoming data in one file or creates separate files for each input. In the latter case, an incremental numerical suffix is added to a file name.	True

## Parameters in Workflow File

**Type:** write-fastq

Parameter	Parameter in the GUI	Type
url-out	Output file	string

<b>write-mode</b>	<b>Existing file</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
<b>accumulate</b>	<b>Accumulate objects</b>	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>
<b>Location</b>	<b>url</b>	<i>string</i>
<b>FASTA header</b>	<b>fasta-header</b>	<i>string</i>

## Write Plain Text Element

Writes strings to a file.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, then the “Location” slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Accumulate objects</b>	Accumulates all incoming data in one file or creates separate files for each input. In the latter case, an incremental numerical suffix is added to a file name.	True

## Parameters in Workflow File

**Type:** write-text

Parameter	Parameter in the GUI	Type
<b>url-out</b>	<b>Output file</b>	<i>string</i>
<b>write-mode</b>	<b>Existing file</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
<b>accumulate</b>	<b>Accumulate objects</b>	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Plain text*

**Name in Workflow File:** in-text

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Plain text	text	string
Location	url	string

## Write Sequence Element

Writes all supplied sequences to file(s) in selected format.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, then the “Location” slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Document format</b>	Format of the output file.	fasta
<b>Accumulate objects</b>	Accumulates all incoming data in one file or creates separate files for each input. In the latter case, an incremental numerical suffix is added to a file name.	True

## Parameters in Workflow File

**Type:** write-sequence

Parameter	Parameter in the GUI	Type
url-out	<b>Output file</b>	string
write-mode	<b>Existing file</b>	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
document-format	<b>Document format</b>	string  Available values are: <ul style="list-style-type: none"> <li>• fasta</li> <li>• fastq</li> <li>• genbank</li> <li>• raw</li> </ul>
accumulate	<b>Accumulate objects</b>	boolean

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Location	url	string
Set of annotations	annotations	annotation-table-list

## Write Variations Element

Writes all supplied variations to file(s) in selected format.

**Parameters in GUI**

Parameter	Description	Default value
<b>Accumulate objects</b>	Accumulate all incoming data in one file or create separate files for each input. In the latter case, an incremental numerical suffix is added to the file name.	True
<b>Document format</b>	Document format of output file.	snp
<b>Output file</b>	Location of output data file. If this attribute is set, slot "Location" in port will not be used.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format). If Rename option is chosen existing file will be renamed.	Rename

**Parameters in Workflow File**

**Type:** write-variations

Parameter	Parameter in the GUI	Type
accumulate	Accumulate objects	boolean
document-format	Document format	string
out-url	Output file	string
write-mode	Existing file	numeric

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** Variation track

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Location	url	string
Variation track	variation-track	variation

## Data Flow

- [Filter Element](#)
- [Grouper Element](#)



- Multiplexer Element
- Sequence Marker Element

## Filter Element

This element passes through only data that matches the input filter value (or values).

### Parameters in GUI

Parameter	Description	Default value
Filter by value(s)	Semicolon-separated list of values used to filter the input data.	

## Parameters in Workflow File

**Type:** filter-by-values

Parameter	Parameter in the GUI	Type
text	Filter by value(s)	string

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Input values*

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Input values	text	string

The element has 1 *output port*.

**Name in GUI:** *Passing values (by Filter)*

**Name in Workflow File:** filtered-data

## Grouper Element

The element groups data supplied to the specified slot by the specified property (for example, by value). Additionally, it is possible to merge data from another slots associated with the specified one.

### Parameters in GUI

To use the *Grouper* element connect the *Grouper's* input port to the required workflow element. Select the *Grouper* element on the *Scene* and specify *Group slot* and *Group operation* parameters in the *Parameters* area in the *Property Editor*. To merge associated data, it is possible to create as many *Output slot(s)* as required (see details below).

### Group slot

The *Group slot* specifies a *slot* that is used to group the input data. The list of available values of the parameter depend on the slots of workflow elements which produce data in the workflow before the *Grouper* element. There is a special *Unset* value. When it is selected, only one group is created.

### Group operation

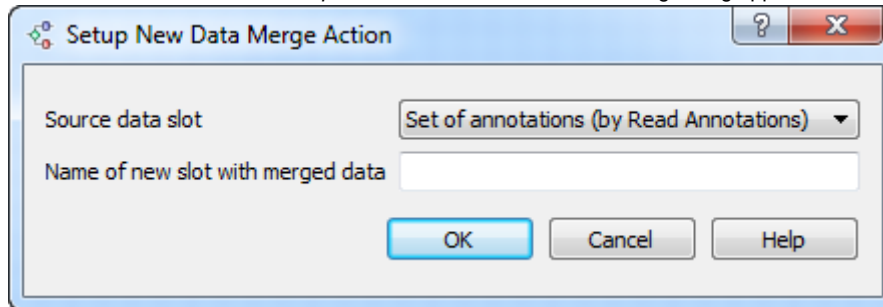
The *Group operation* specifies criteria to group data supplied to the *Group slot*. It can take the following values:

- *By value* - input data are compared by value (a group is created for each unique value, it can contain one or several identical values)
- *By identity* - input data are compared by internal data ID (all values are unique)
- *By name* - input data are compared by their names

*By value* group operation is available for group slots of types *Sequence*, *Set of annotations*, *MSA*, *Plain text*, *Source URL*. *By identity* and *By name* group operations are available for group slots of type *Sequence* only.

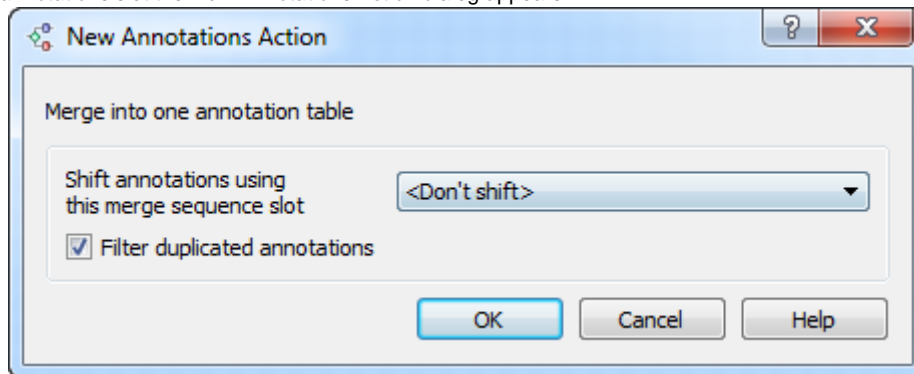
### Output slots

When data supplied to the *Group slot* are divided into different groups the associated data are also got into a group. The possible associated data depend on the workflow. For example, a *Sequence Reader* element contains slots *Sequence* and *Set of annotations*. These data are **as sociated** as annotations belong to a sequence. Another example of associated data are sequence markers created by the *Sequence Marker* element. The associated data, therefore, can be additionally handled (i.e. merged) by the *Grouperelement*. The action that can be performed on the associated data depends on their type. In any case to output handled associated data you need to create a new output slot in the *Grouper* element. To create it click the *Add* button in the *Grouper's Parameters* area. The following dialog appears:



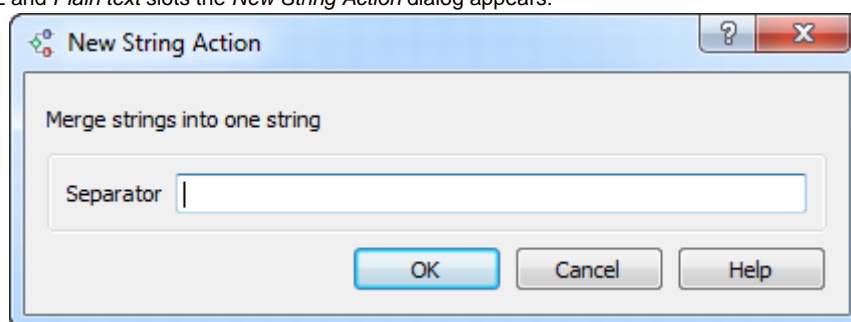
In the dialog you should select a *Source data slot* (i.e. a slot with the associated data) and input a name of the new slot. Click the *OK* button. A new dialog appears that specifies how the associated data should be merged. The view of the dialog and the available merge actions for different types of the *Source data slot* are the following:

- For a *Set of annotations* slot the *New Annotations Action* dialog appears:



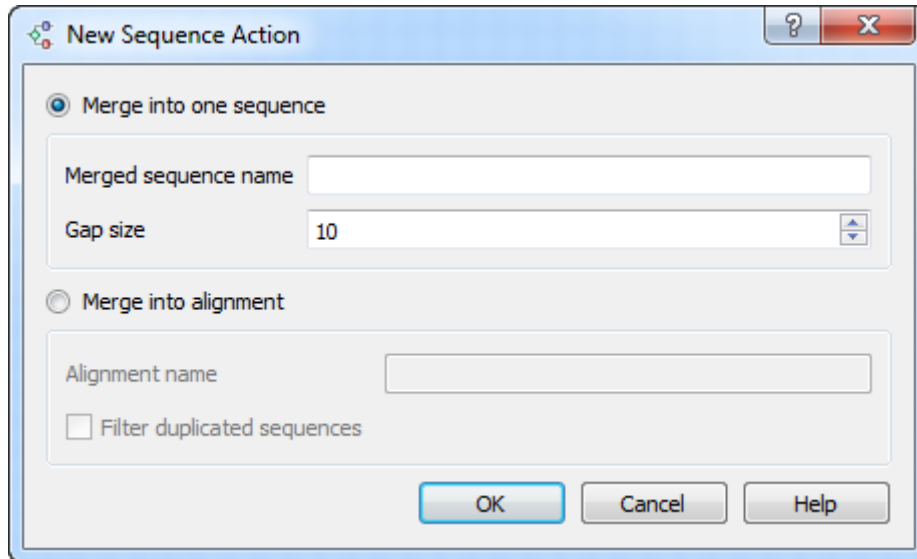
You can merge annotations into one annotation table and, optionally, filter duplicated annotations. Also, you can shift annotations. To do it, you need to create another output slot with type *Sequence* and *Merge into one sequence* option selected (see below). In other words you need to merge all sequences in a group into one sequence. In this case you select the corresponding sequence slot in the *New Annotations Action* dialog and each set of annotations in a group is shifted according to the corresponding sequence in the group. As the result you have one sequence and one set of annotations allocated on the whole sequence.

- For *Source URL* and *Plain text* slots the *New String Action* dialog appears:



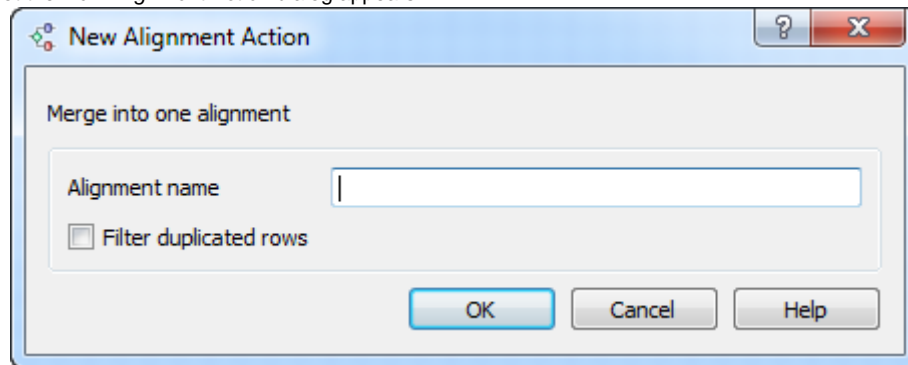
Using this dialog you can merge strings into one string. Optionally, you can specify an additional strings separator.

- For a *Sequence* slot the *New Sequence Action* dialog appears:



You can either merge all sequences in a group into one sequence or create a multiple sequence alignment. In the first case you need to specify the *Merged sequence name* and you can select the number of unknown characters between the merged sequences. In the second case you need to specify the alignment name. To filter duplicated sequence check the corresponding check box.

- For a MSA slot the *New Alignment Action* dialog appears:



Input the alignment name in this dialog. To filter duplicated rows check the corresponding check box.

To edit a created slot, select it in the *Parameters* area of the *Groupier* element and click the *Edit* button. To remove the slot, select it and click the *Remove* button.

### Parameters in Workflow File

**Type:** grouper

### Input/Output Ports

The element has 1 *input port* that can take any incoming data.

**Name in GUI:** *Input data flow*

**Name in workflow File:** input-data

The element has 1 *output port*.

**Name in GUI:** *Grouped output data flow*

**Name in workflow File:** output-data

### Slots:

Slot In GUI	Slot in workflow File	Type
Group size	group-size	string

Also the port has one default slot of the grouped data and it may also have one or several customized output slots (see above).

## Multiplexer Element

Construct an output data flow using two input data flows a multiplexing rule.

There are the following multiplexing rules:

- 1 to 1 – for every message from the first input data flow it gets only one message from the second input data flow and puts them to the output.
- 1 to many, Many to 1 – for every message from the first input data flow it gets every message from the second input data flow and puts them to the output.
- Streaming mode – puts every message from the first and the second input data flows to the output.

Also see the [Find Substrings at Sequences](#), [Search for TFBS](#) examples with *Multiplexer* element.

### Parameters in GUI

Parameter	Description	Default value
<b>Multiplexing rule</b>	How to multiplex the input data flows.  Available values are: <ul style="list-style-type: none"> <li>• 1 to 1</li> <li>• 1 to Many</li> <li>• Many to 1</li> <li>• Streaming mode</li> </ul>	1 to 1
<b>If empty input</b>	Specifies how to multiplex the data if one of input ports produces no data. It can be used for 1 to 1 multiplexing rule.  Available values are: <ul style="list-style-type: none"> <li>• Fill by empty values (if one of input ports produces no data, get data from another port only and put them to the output.)</li> <li>• Truncate (if one of input port produces no data, then do not output anything.)</li> </ul>	Fill by empty values

## Parameters in Workflow File

**Type:** multiplexer

Parameter	Parameter in the GUI	Type
<b>multiplexing-rule</b>	<b>Multiplexing rule</b>	<i>string</i>
<b>empty-input-action</b>	<b>If empty input</b>	<i>string</i>

## Input/Output Ports

The *Multiplexer* has ports but has not slots, because its use the whole data flow.

The element has 2 *input port*.

**Name in GUI:** *First input data flow*

**Name in Workflow File:** input-data-1

**Name in GUI:** *Second input data flow*

**Name in Workflow File:** input-data-2

The element has 1 *output port*.

**Name in GUI:** *Multiplexed output data flow*

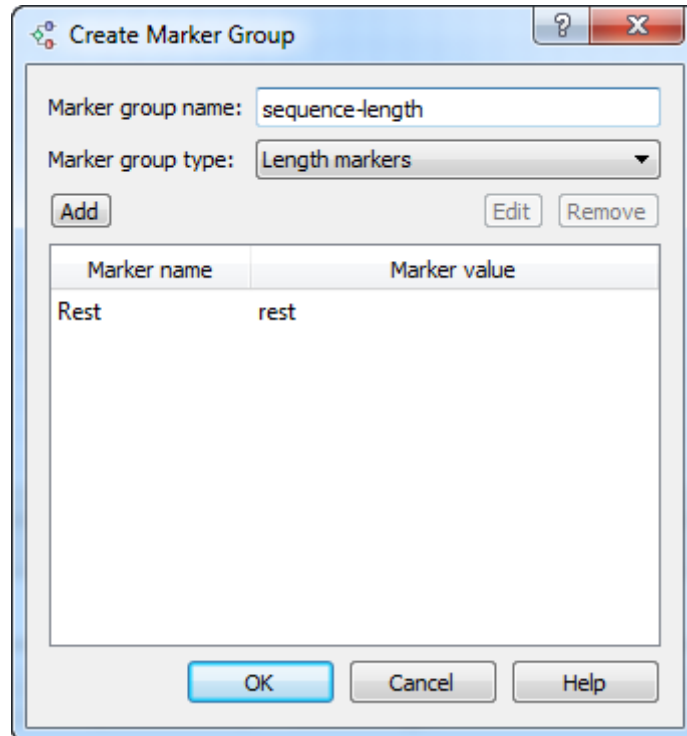
**Name in Workflow File:** output-data

## Sequence Marker Element

Adds one or several marks to the input sequence depending on the sequence properties. Use this element, for example, in conjunction with the *Filter* element.

## Parameters in GUI

To create a new marker group that would mark the input sequence, select the *Add* button in the *Parameters* area. The *Create Marker Group* dialog appears:



Choose a type of the marker group and input a marker group name. The following types are available:

*Length markers* — marks a sequence by length. The sequence is marked, for example, if its length is less or greater than the specified value.

*Sequence name markers* — marks a sequence by a sequence name.

*Annotations count markers* — marks a sequence by the number of annotations.

*Qualifier integer value markers* — marks a sequence by the number of integer qualifiers.

*Qualifier text value markers* — marks a sequence by the number of text qualifiers.

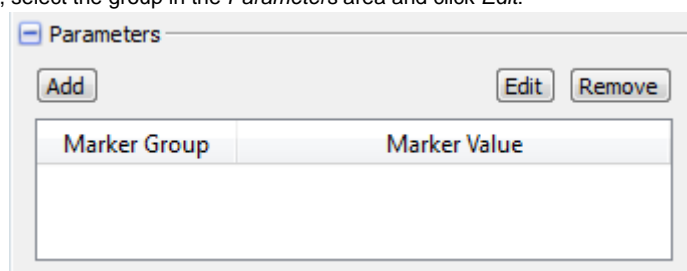
*Qualifier float value markers* — marks a sequence by the number of float qualifiers.

*Text markers* — marks a sequence by a file name. For example, if the name:

1. starts with the specified text;
2. ends with the specified text;
3. contains the specified text;
4. matches the specified regular expression .

Each marker group can contain more than one marker. Use the *Add*, *Edit* and *Remove* buttons in the dialog to create, modify and delete markers in the marker group.

To edit the created marker group, select the group in the *Parameters* area and click *Edit*.



To remove a marker group select it in the list and click *Remove*.

### Parameters in Workflow File

**Type:** mark-sequence

### Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Location	url	string
Set of annotations	annotations	annotation-table-list

The element has 1 *output port*.

**Name in GUI:** *Marked sequence*

**Name in Workflow File:** out-marked-seq

**Slots:**

Each created marker group adds a text slot with the following properties:

Slot In GUI	Slot in Workflow File	Type
Name of the marker group	Name of the marker group	string

## Basic Analysis

- Amino Translations Element
- Annotate with DAS Element
- Annotate with UQL Element
- CD-Search Element
- Collocation Search Element
- Export PHRED Qualities Element
- Fetch Sequences by ID From Annotation Element
- Filter Annotation by Name Element
- Filter Annotations by Qualifier
- Find Pattern Element
- Find Repeats Element
- Gene-by-gene approach report
- Get Sequences by Annotations Element
- Import PHRED Qualities Element
- Local BLAST Search Element
- Local BLAST+ Search Element
- Merge Annotations Element
- ORF Marker Element
- Remote BLAST Element
- Remove Duplicates in BAM Files Element
- Smith-Waterman Search Element

## Amino Translations Element

Translates a sequence into it's amino translation or translations.

### Parameters in GUI

Parameter	Description	Default value
Translate from	Specifies position that should be used to translate the sequence from: first, second, third or all (three output amino sequences would be generated).	all

<b>Auto selected genetic code</b>	Specifies that genetic code should be selected automatically.	True
<b>Genetic code</b>	Genetic code that should be used to translate the input nucleotide sequence.	The Standard Genetic Code

## Parameters in Workflow File

**Type:** sequence-translation

Parameter	Parameter in the GUI	Type
pos-2-translate	Translate from	string  Available values are: <ul style="list-style-type: none"> <li>• all</li> <li>• first</li> <li>• second</li> <li>• third</li> </ul>
auto-translation	Auto selected genetic code	boolean
genetic-code	Genetic code	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input Data*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Amino sequence*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Plain text	text	string

## Annotate with DAS Element

Finds similar protein sequence using remote BLAST. Using IDs of sequences found loads annotation for DAS sources. Nucleotide sequences are skipped if any supplied to input.

**Parameters in GUI**

Parameter	Description	Default value
<b>Max result IDs</b>	Use first IDs of similar sequences to load annotations.	5
<b>Database</b>	Database against which the search is performed: UniProtKB or clusters of sequences with 100%, 90% or 50% identity.	UniProtKB

<b>Min identity</b>	Minimum identity of a BLAST result and an input sequence.	90%
<b>Threshold</b>	The expectation value (E) threshold is a statistical measure of the number of expected matches in a random database. The lower the e-value, the more likely the match is to be significant.	10
<b>Matrix</b>	The matrix assigns a probability score for each position in an alignment.	Auto
<b>Filtering</b>	Low-complexity regions (e.g. stretches of cysteine in Q03751, or hydrophobic regions in membrane proteins) tend to producespurious, insignificant matches with sequences in the database which have the same kind of low-complexity regions, but are unrelated biologically. If 'Filter low complexity regions' is selected, the query sequence will be run through the program SEG, and all amino acids in low-complexity regions will be replaced by X's.	None
<b>Gapped</b>	This will allow gaps to be introduced in the sequences when the comparison is done.	true
<b>Hits</b>	Limits the number of returned alignments.	250
<b>Feature sources</b>	The DAS sources to read features from.	InterPro-Matches-Overview, Pride DAS 1.6, UniProt, cbs_sort, signalp

## Parameters in Workflow File

**Type:** dasannotation-search

Parameter	Parameter in the GUI	Type
idsnumber	<b>Max result IDs</b>	<i>numeric</i>
db	<b>Database</b>	<i>string</i>
identity	<b>Min identity</b>	<i>numeric</i>
threshold	<b>Threshold</b>	<i>string</i>
matrix	<b>Matrix</b>	<i>string</i>
filtering	<b>Filtering</b>	<i>string</i>
gapped	<b>Gapped</b>	<i>string</i>
maxres	<b>Hits</b>	<i>string</i>
fsources	<b>Feature sources</b>	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input sequences

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>string</i>



The element has 1 output *port*:

**Name in GUI:** DAS annotations

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	<i>annotation-table-list</i>

## Annotate with UQL Element

Analyzes a nucleotide sequence with a UGENE Query Language (UQL) workflow. The workflow specifies a set of features to search for and their positional relationship.

To learn more about UQL workflows read [UGENE Query Designer Manual](#).

## Parameters in GUI

Parameter	Description	Default value
<b>Workflow</b> (required)	UQL workflow file.	
<b>Merge</b>	Merges regions of each result into a single annotation.	False
<b>Offset</b>	If the <i>Merge</i> parameter is set to <i>True</i> , adds left and right offsets of the specified length to the annotation.	0

## Parameters in Workflow File

**Type:** query

Parameter	Parameter in the GUI	Type
schema	Workflow	<i>string</i>
merge	Merge	<i>boolean</i>
offset	Offset	<i>numeric</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequences*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	<i>sequence</i>

And 1 *output port*:

**Name in GUI:** *Result annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	<i>annotation-table</i>

## CD-Search Element

Finds conserved domains in protein sequences. In case conserved domains database is downloaded the search can be executed on local machine. The search can be submitted to the NCBI for remote execution.

### Parameters in GUI

Parameter	Description	Default value
Annotate as	Name of the result annotations marking found conserved domains.	CDD result

<b>Database</b>	<p>Currently, CD-Search is offered with the following search databases:</p> <ul style="list-style-type: none"> <li>• CDD - this is a superset including NCBI-curated domains and data imported from Pfam, SMART, COG, PRK, and TIGRFAM.</li> <li>• Pfam - a mirror of a recent Pfam-A database of curated seed alignments. Pfam version numbers do change with incremental updates. As with SMART, families describing very short motifs or peptides may be missing from the mirror. An HMM-based search engine is offered on the Pfam site.</li> <li>• SMART - a mirror of a recent SMART set of domain alignments. Note that some SMART families may be missing from the mirror due to update delays or because they describe very short conserved peptides and/or motifs, which would be difficult to detect using the CD-Search service. You may want to try the HMM-based search service offered on the SMART site. Note also that some SMART domains are not mirrored in CD because they represent “superfamilies” encompassing several individual, but related, domains; the corresponding seed alignments may not be available from the source database in these cases. Note also that SMART version numbers do not change with incremental updates of the source database (and the mirrored CD-Search database).</li> <li>• TIGRFAM - a mirror of a recent TIGRFAM set of domain alignments. An HMM-based search engine is offered on the TIGRFAM site.</li> <li>• COG - a mirror of the current COG database of orthologous protein families focusing on prokaryotes. Seed alignments have been generated by an automated process. An alternative search engine, “Cognitor”, which runs protein-BLAST against a database of COG-assigned sequences, is offered on the COG site.</li> <li>• KOG - a eukaryotic counterpart to the COG database. KOGs are not included in the CDD superset, but are searchable as a separate data set.</li> </ul>	<p>CDD Available values are:</p> <ul style="list-style-type: none"> <li>• CDD</li> <li>• Pfam</li> <li>• TIGRFAM</li> <li>• COG</li> <li>• KOG</li> <li>• Prk</li> <li>• SMART</li> </ul>
<b>Database directory</b>	Specifies database directory for local search.	
<b>Local search</b>	Perform the search on local machine or submit the search to NCBI for remote execution.	True

<b>Expect value</b>	Modifies the <b>E-value</b> threshold used for filtering results. False positive results should be very rare with the default setting of 0.01, results with E-values in the range of 1 and above should be considered putative false positives.	
---------------------	---	--

## Parameters in Workflow File

**Type:** cd-search

Parameter	Parameter in the GUI	Type
<b>result-name</b>	<b>Annotate as</b>	<i>string</i>
<b>db-name</b>	<b>Database</b>	<i>string</i>
<b>db-path</b>	<b>Database directory</b>	<i>string</i>
<b>local-search</b>	<b>Local search</b>	<i>boolean</i>
<b>e-val</b>	<b>Expect value</b>	<i>numeric</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>

And 1 *output port*:

**Name in GUI:** *Annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table</i>

## Collocation Search Element

Finds groups of specified annotations in each supplied set of annotations, stores found regions as annotations.

**Parameters in GUI**

Parameter	Description	Default value
<b>Result type</b>	Copy original annotations or annotate found regions with new ones.	Create new annotations
<b>Result annotation</b> (required)	Name of the result annotation to mark found collocations.	misc_feature
<b>Include boundaries</b>	Include most left and most right boundary annotations regions into result or exclude them.	True
<b>Group of annotations</b> (required)	List of annotation names to search. Found regions will contain all the named annotations.	

<b>Region size</b>	Effectively this is the maximum allowed distance between the interesting annotations in a group.	1000
<b>Must fit into region</b>	Specifies whether the interesting annotations should entirely fit into the specified region to form a group.	False

## Parameters in Workflow File

**Type:** colocated-annotation-search

Parameter	Parameter in the GUI	Type
result-type	Result type	string
result-name	Result annotation	string
annotations	Group of annotations	string
include-boundary	Include boundaries	boolean
region-size	Region size	numeric
must-fit	Must fit into region	boolean

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Input data*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Set of annotations	annotations	annotation-table-list

And 1 *output port*.

**Name in GUI:** *Group annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Export PHRED Qualities Element

Export corresponding PHRED quality scores from input sequences.

**Parameters in GUI**

Parameter	Description	Default value
PHRED output	Path to file with PHRED quality scores.	

## Parameters in Workflow File

**Type:** export-phred-qualities

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

url-out	PHRED output	string
---------	--------------	--------

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** DNA sequences

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	string

## Fetch Sequences by ID From Annotation Element

Parses annotations to find any IDs and fetches corresponding sequences.

**Parameters in GUI**

Parameter	Description	Default value
Save file to directory	The directory to store sequence files loaded from a database.	default
NCBI database	The database to read from.	nucleotide  Available values are: <ul style="list-style-type: none"> <li>nucleotide</li> <li>protein</li> </ul>

## Parameters in Workflow File

**Type:** fetch-sequence

Parameter	Parameter in the GUI	Type
save-dir	Save file to directory	string
database	NCBI database	string

The element has 1 *input port*:

**Name in GUI:** Input annotations

**Name in Workflow File:** in-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

And 1 *output port*:

**Name in GUI:** Sequence

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table
Sequence	sequence	sequence

## Filter Annotation by Name Element

Filters annotations by name.

#### Parameters in GUI

Parameter	Description	Default value
Annotation name	File with annotation names, separated with whitespaces or list of annotation names which will be accepted or filtered.	
Accept or filter	Selects the name filter: accept specified names or accept all except specified.	True

## Parameters in Workflow File

**Type:** filter-annotations

Parameter	Parameter in the GUI	Type
annotation-names	Annotation name	string
accept-or-filter	Accept or filter	boolean

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input annotations*

**Name in Workflow File:** in-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

And 1 *output port*:

**Name in GUI:** *Result annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Filter Annotations by Qualifier

Filters annotations by qualifier.

#### Parameters in GUI

Parameter	Description	Default value
Qualifier name	Name of the qualifier to use for filtering.	
Qualifier value	Text value of the qualifier to apply as filtering criteria.	
Accept or filter	Selects the name filter: accept specified names or accept all except specified.	True

#### Parameters in Workflow File

**Type:** filter-annotations-by-qualifier

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

<b>qualifier-name</b>	<b>Qualifier name</b>	<i>string</i>
<b>qualifier-value</b>	<b>Qualifier value</b>	<i>string</i>
<b>accept-or-filter</b>	<b>Accept or filter</b>	<i>boolean</i>

### Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Input annotations*

**Name in Workflow File:** in-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table</i>

And 1 *output port*.

**Name in GUI:** *Result annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Set of annotations</b>	<b>annotations</b>	<i>annotation-table</i>

## Find Pattern Element

Searches regions in a sequence similar to a pattern sequence. Outputs a set of annotations.

**Parameters in GUI**

Parameter	Description	Default value
<b>Annotate as</b>	Name of the result annotation.	misc_feature
<b>Pattern(s)</b>	Semicolon-separated list of patterns to search for.	
<b>Pattern file</b>	Load pattern from file in any sequence format or in newline-delimited format.	
<b>Use pattern name</b>	If patterns are loaded from a file, use names of pattern sequences as annotation names. The name from the parameters is used by default.	False
<b>Max Mismatches</b>	Maximum number of mismatches between a substring and a pattern.	0
<b>Search in</b>	Specifies which strands should be searched: direct, complementary or both.	both strands
<b>Allow Insertions/Deletions</b>	Takes into account possibility of insertions/deletions when searching. By default substitutions are only considered.	False
<b>Support ambiguous bases</b>	Performs correct handling of ambiguous bases. When this option is activated insertions and deletions are not considered.	False
<b>Search in Translation</b>	Translates a supplied nucleotide sequence to protein and searches in the translated sequence.	False



<b>Qualifier name for pattern name</b>	Name of qualifier in result annotations which is containing a pattern name.	pattern_name
--	---	--------------

## Parameters in Workflow File

Type: search

Parameter	Parameter in the GUI	Type
result-name	Annotate as	string
pattern	Pattern(s)	string
pattern_file	Pattern file	string
use-names	Use pattern name	boolean
max-mismatches-num	Max Mismatches	numeric
strand	Search in	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for searching in both strands</li> <li>• 1 - for searching in direct strand</li> <li>• 2 - for searching in complement strand</li> </ul>
allow-ins-del	Allow Insertions/Deletions	boolean
ambiguous	Support ambiguous bases	boolean
amino	Search in Translation	boolean
pattern-name-qual	Qualifier name for pattern name	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input data*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Plain text	text	string

And 1 *output port*:

**Name in GUI:** *Pattern annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Find Repeats Element

Finds repeats in each supplied sequence, stores found regions as annotations.

**Parameters in GUI**

Parameter	Description	Default value
-----------	-------------	---------------

<b>Annotate as</b> (required)	Name of the result annotation to mark found repeats.	repeat_unit
<b>Algorithm</b>	Control over variations of the algorithm.	Auto
<b>Filter nested</b>	Filters nested repeats.	True
<b>Identity</b>	Repeats identity in percents.	100
<b>Inverted</b>	Specifies to search for inverted repeats.	False
<b>Max distance</b>	Maximum distance between the repeats.	5000
<b>Min distance</b>	Minimum distance between the repeats.	0
<b>Min length</b>	Minimum length of the repeats.	5
<b>Parallel threads</b>	Number of parallel threads used for the task.	Auto

## Parameters in Workflow File

Type: repeats-search

Parameter	Parameter in the GUI	Type
result-name	Annotate as	string
algorithm	Algorithm	numeric Available values are: <ul style="list-style-type: none"> <li>0 - algorithm choosed automaticly</li> <li>1 - for diagonal algorithm</li> <li>2 - for suffix index algorithm</li> </ul>
filter-nested	Filter nested	boolean
identity	Identity	numeric
max-distance	Max distance	numeric
min-distance	Min distance	numeric
min-length	Min length	numeric
threads	Parallel threads	numeric 0 - for using autodetected threads number

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Repeat annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	<i>annotation-table</i>

## Gene-by-gene approach report

Output a table of genes found in a reference sequence.

### Parameters in GUI

Parameter	Description	Default value
Output file	File to store a report.	
Annotation name	Annotation name used to compare genes and reference genomes..	blast-result
Existing file	If a target report already exists you should specify how to handle that. Merge two table in one. Overwrite or Rename existing file..	Merge
Identity cutoff	Identity between gene sequence length and annotation length in per cent. BLAST identity (if specified) is checked after	90.0000%

### Parameters in Workflow File

**Type:** genebygene-report-id

Parameter	Parameter in the GUI	Type
output-file	Output file	<i>string</i>
annotation_name	Annotation name	<i>string</i>
existing	Existing file	<i>string</i>
identity	Identity cutoff	<i>numeric</i>

### Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** Gene by gene report data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Input annotations	gene-ann	<i>ann-table-list</i>
Input sequences	gene-seq	<i>seq</i>

## Get Sequences by Annotations Element

Extracts annotated regions from input sequence.

### Parameters in GUI

Parameter	Description	Default value
Annotation names (required)	List of annotation names which will be accepted or filtered. Use space as the separator.	
Accept of filter	Selects the name filter: accept specified names or accept all except specified.	Accept

<b>Complement</b>	Complements the annotated regions if the corresponding annotation is located on the complement strand.	True
<b>Translate</b>	Translates the annotated regions if the corresponding annotation marks a protein subsequence.	True
<b>Extend left</b>	Extends the resulted regions to left.	0
<b>Extend right</b>	Extends the resulted regions to right.	0
<b>Gap length</b>	Inserts a gap of a specified length between the merged locations of the annotation.	1

## Parameters in Workflow File

**Type:** extract-annotated-sequence

Parameter	Parameter in the GUI	Type
annotation-names	Annotation names	<i>string</i>
accept-or-filter	Accept or filter	<i>boolean</i> Available values are: <ul style="list-style-type: none"> <li>• true - for accept</li> <li>• false - for filter</li> </ul>
complement	Complement	<i>boolean</i>
translate	Translate	<i>boolean</i>
extend-left	Extend left	<i>numeric</i>
extend-right	Extend right	<i>numeric</i>
merge-gap-length	Gap length	<i>numeric</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	<i>sequence</i>
Set of annotations	annotations	<i>annotation-table</i>

And 1 *output port*:

**Name in GUI:** *Annotated regions*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	<i>sequence</i>
Set of annotations	annotations	<i>annotation-table</i>

## Import PHRED Qualities Element

Adds corresponding PHRED quality scores to the sequences. Use this element to convert .fasta and .qual pair to fastq format.

### Parameters in GUI

Parameter	Description	Default value
<b>PHRED input</b> (required)	Path to a file with PHRED quality scores.	
<b>Quality format</b>	Format to encode quality scores.	Sanger

## Parameters in Workflow File

**Type:** import-phred-qualities

Parameter	Parameter in the GUI	Type
url-in	PHRED input	string
quality-format	Quality format	string Available values are: <ul style="list-style-type: none"> <li>Sanger</li> <li>Illumina 1.3+</li> <li>Solexa/Illumina 1.0</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** DNA sequences

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** DNA sequences with imported qaualities

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## Local BLAST Search Element

Finds annotations for the supplied DNA sequence in local BLAST database.



BLAST is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

## Parameters in GUI

Parameter	Description	Default value
-----------	-------------	---------------

<b>Search type</b>	Selects the type of the BLAST searches.	blastn
<b>Database path</b>	Path to the database files.	
<b>Database name</b>	Base name for BLAST DB files.	
<b>Tool path</b>	Path to the BLAST executable.	default
<b>Temporary directory</b>	Directory for temporary files.	default
<b>Expected value</b>	Expectation threshold value.	10
<b>Annotate as</b>	Name of the result annotations.	blast_result
<b>Gapped alignment</b>	Perform gapped alignment.	use
<b>Gap costs</b>	Cost to create and extend a gap in an alignment.	2 2
<b>Match scores</b>	Reward and penalty for matching and mismatching bases.	1 -3
<b>BLAST output</b>	Location of BLAST output file.	
<b>BLAST output type</b>	Type of BLAST output file.	XML (-m 7)

## Parameters in Workflow File

Type: blast

Parameter	Parameter in the GUI	Type
blast-type	Search type	string Available values are: <ul style="list-style-type: none"> <li>blastn</li> <li>blastp</li> <li>blastx</li> <li>tblastn</li> <li>tblastx</li> </ul>
db-path	Database path	string
db-name	Database name	string
tool-path	Tool path	string
temp-dir	Temporary directory	string
e-val	Expected value	numeric
result-name	Annotate as	string
gapped-aln	Gapped alignment	boolean
gap-costs	Gap costs	string
match-scores	Match scores	string
blast-output	BLAST output	string
type-output	BLAST output type	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** Annotations

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Local BLAST+ Search Element

Finds annotations for DNA sequence in a local BLAST database.

BLAST+ is a newer version of the BLAST package and is recommended to use by the NCBI.



BLAST+ is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

## Parameters in GUI

Parameter	Description	Default value
Search type	Selects the type of the BLAST searches.	blastn
Database path	Path to the database files.	
Database name	Base name for BLAST DB files.	
Tool path	Path to the BLAST executable.	default
Temporary directory	Directory for temporary files.	default
Expected value	Expectation threshold value.	10
Annotate as	Name of the result annotations.	blast_result
Gapped alignment	Perform gapped alignment.	use
Gap costs	Cost to create and extend a gap in an alignment.	2 2
Match scores	Reward and penalty for matching and mismatching bases.	1 -3
BLAST output	Location of BLAST output file.	
BLAST output type	Type of BLAST output file.	XML (-outfmt 5)

## Parameters in Workflow File

**Type:** blast-plus

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

<b>blast-type</b>	<b>Search type</b>	<i>string</i>  Available values are: <ul style="list-style-type: none"><li>• blastn</li><li>• blastp</li><li>• blastx</li><li>• tblastn</li><li>• tblastx</li></ul>
<b>db-path</b>	<b>Database path</b>	<i>string</i>
<b>db-name</b>	<b>Database name</b>	<i>string</i>
<b>tool-path</b>	<b>Tool path</b>	<i>string</i>
<b>temp-dir</b>	<b>Temporary directory</b>	<i>string</i>
<b>e-val</b>	<b>Expected value</b>	<i>numeric</i>
<b>result-name</b>	<b>Annotate as</b>	<i>string</i>
<b>gapped-aln</b>	<b>Gapped alignment</b>	<i>boolean</i>
<b>gap-costs</b>	<b>Gap costs</b>	<i>string</i>
<b>match-scores</b>	<b>Match scores</b>	<i>string</i>
<b>blast-output</b>	<b>BLAST output</b>	<i>string</i>
<b>type-output</b>	<b>BLAST output type</b>	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Merge Annotations Element

Writes all supplied sequences to file(s) in FASTQ format.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, then the "Location" slot is not taken into account.	



<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename
<b>Accumulate objects</b>	Accumulates all incoming data in one file or creates separate files for each input. In the latter case, an incremental numerical suffix is added to a file name.	True

## Parameters in Workflow File

**Type:** write-fastq

Parameter	Parameter in the GUI	Type
url-out	Output file	string
write-mode	Existing file	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>
accumulate	Accumulate objects	boolean

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence
Location	url	string

## ORF Marker Element

Finds Open Reading Frames (ORFs) in each supplied nucleotide sequence, stores found regions as annotations.

**Parameters in GUI**

Parameter	Description	Default value
<b>Annotate as</b> (required)	Name of the result annotations.	ORF
<b>Search in</b>	Specifies which strands should be searched: direct, complement or both.	both strands
<b>Min length</b>	Ignores ORFs shorter than the specified length.	100
<b>Genetic code</b>	Specifies which genetic code should be used for translating the input nucleotide sequence.	The Standard Genetic Code
<b>Require init codon</b>	Allows or not ORFs starting with any codon other than terminator.	True
<b>Require stop codon</b>	Ignores boundary ORFs which last beyond the search region (i.e. have no stop codon within the range).	False

<b>Allow alternative codons</b>	Allows ORFs starting with alternative initiation codons, accordingly to the current translation table.	False
---------------------------------	--	-------

## Parameters in Workflow File

Type: orf-search

Parameter	Parameter in the GUI	Type
result-name	Annotate as	string
strand	Search in	numeric Available values are: <ul style="list-style-type: none"> <li>• 0 - for searching in both strands</li> <li>• 1 - for searching in direct strand</li> <li>• 2 - for searching in complement strand</li> </ul>
min-length	Min length	numeric
genetic-code	Genetic code	string Available values are: <ul style="list-style-type: none"> <li>• NCBI-GenBank #1</li> <li>• NCBI-GenBank #2</li> <li>• etc.</li> </ul>
require-init-codon	Require init codon	boolean
require-stop-codon	Require stop codon	boolean
allow-alternative-codons	Allow alternative codons	boolean

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *ORF annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Remote BLAST Element

Finds annotations for the supplied DNA sequence in the NCBI remote database.

**Parameters in GUI**

Parameter	Description	Default value
-----------	-------------	---------------

<b>Database</b>	Selects the database to search through. Available databases are blastn, blastp and cdd.	ncbi-blastn
<b>Expected value</b>	This parameter specifies the statistical significance threshold of reporting matches against the database sequences.	10
<b>Max hits</b>	Maximum number of hits. The maximum available number is 5000.	10
<b>Short sequence</b>	Optimizes search for short sequences.	False
<b>Entrez query</b>	Enter an Entrez query to limit search.	
<b>Annotate as</b>	Name of the result annotations.	
<b>BLAST output</b>	Location of the BLAST output file. This parameter insignificant for cdd search.	
<b>Gap costs</b>	Cost to create and extend a gap in an alignment.	2 2
<b>Match scores</b>	Reward and penalty for matching and mismatching bases.	1 -3

## Parameters in Workflow File

Type: blast-ncbi

Parameter	Parameter in the GUI	Type
db	Database	string Available values are: <ul style="list-style-type: none"> <li>ncbi-blastn</li> <li>ncbi-blastp</li> <li>ncbi-cdd</li> </ul>
e-val	Expected value	string
max-hits	Max hits	numeric
short-sequence	Short sequence	boolean
entrez-query	Entrez query	string
result-name	Annotate as	string
blast-output	BLAST output	string
gap-costs	Gap costs	string
match-scores	Match scores	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	<i>annotation-table</i>

## Remove Duplicates in BAM Files Element

Remove PCR duplicates of BAM files using SAMTools rmdup.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	
<b>Output BAM name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	
<b>Remove for single-end reads</b>	Remove duplicate for single-end reads. By default, the command works for paired-end reads only (-s).	False
<b>Treat as single-end</b>	Treat paired-end reads and single-end reads (-S).	False

### Parameters in Workflow File

**Type:** rmdup-bam

Parameter	Parameter in the GUI	Type
out-mode	Output directory	<i>numeric</i>
custom-dir	Custom directory	<i>string</i>
out-name	Output file name	<i>string</i>
remove-single-end	Remove for single-end reads	<i>boolean</i>
treat_reads	Treat as single-end	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	<i>string</i>

And 1 *output port*:

**Name in GUI:** Output File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

## Smith-Waterman Search Element

Searches regions in a sequence similar to a pattern sequence. Outputs a set of annotations.

Under the hood is the well-known Smith-Waterman algorithm for performing local sequence alignment.

### Parameters in GUI

Parameter	Description	Default value
<b>Substitution Matrix</b>	Describes the rate at which one character in a sequence changes to other character states over time.	Auto
<b>Algorithm</b>	Version of the Smith-Waterman algorithm. You can use the optimized versions of the algorithm (SSE, CUDA and OpenCL) if your hardware supports these capabilities.	OPENCL
<b>Filter Results</b>	Specifies either to filter the intersected results or to return all the results.	filter-intersections
<b>Min Score</b>	Minimal percent similarity between a sequence and a pattern.	90%
<b>Search in</b>	Specifies which strands should be searched: direct, complementary or both.	both strands
<b>Search in Translation</b>	Translates a supplied nucleotide sequence to protein and searches in the translated sequence.	False
<b>Gap Open Score</b>	Penalty for opening a gap.	-10.0
<b>Gap Extension Score</b>	Penalty for extending a gap.	-1.0
<b>Use Pattern Names</b>	Use a pattern name as an annotation name.	True
<b>Annotate as</b>	Name of the result annotations.	misc_feature
<b>Qualifier name for pattern name</b>	Name of qualifier in result annotations which is containing a pattern name.	pattern name

## Parameters in Workflow File

**Type:** ssearch

Parameter	Parameter in the GUI	Type
matrix	<b>Substitution Matrix</b>	string  Available values are: <ul style="list-style-type: none"> <li>• Auto - for auto detecting matrix</li> <li>• blosum60</li> <li>• dna</li> <li>• rna</li> <li>• ...</li> </ul>

<b>algorithm</b>	<b>Algorithm</b>	<i>string</i> Available values are: <ul style="list-style-type: none"><li>• Classic 2</li><li>• SSE2</li><li>• OpenCL</li><li>• CUDA</li></ul>
<b>filter-strategy</b>	<b>Filter Results</b>	<i>string</i> Available values are: <ul style="list-style-type: none"><li>• filter-intersections</li><li>• none</li></ul>
<b>min-score</b>	<b>Min Score</b>	<i>numeric</i>
<b>strand</b>	<b>Search in</b>	<i>numeric</i> Available values are: <ul style="list-style-type: none"><li>• 0 - for searching in both strands</li><li>• 1 - for searching in direct strand</li><li>• 2 - for searching in complement strand</li></ul>
<b>amino</b>	<b>Search in Translation</b>	<i>boolean</i>
<b>gap-open-score</b>	<b>Gap Open Score</b>	<i>numeric</i>
<b>gap-ext-score</b>	<b>Gap Extension Score</b>	<i>numeric</i>
<b>use-names</b>	<b>Use Pattern Names</b>	<i>boolean</i>
<b>result-name</b>	<b>Annotate as</b>	<i>string</i>
<b>pattern-name-qual</b>	<b>Qualifier name for pattern name</b>	<i>string</i>

## Input/Output Ports

The element has 2 *input ports*. The first input port:

**Name in GUI:** *Input data*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The second input port:

**Name in GUI:** *Pattern data*

**Name in Workflow File:** pattern

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Pattern annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	<i>annotation-table</i>

## Data Converters

- [Convert bedGraph Files to bigWig Element](#)
- [Convert Text to Sequence Element](#)
- [File Format Conversion Element](#)
- [Reverse Complement Element](#)
- [Split Assembly into Sequences Element](#)

### Convert bedGraph Files to bigWig Element

Convert bedGraph files to bigWig.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	
<b>Genome</b>	File with genome length.	human.hg18
<b>Output name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	
<b>Block size</b>	Number of items to bundle in r-tree (-blockSize).	256
<b>Items per slot</b>	Number of data points bundled at lowest level (-itemsPerSlot).	1024
<b>Uncompressed</b>	If set, do not use compression.(-unc).	False

### Parameters in Workflow File

**Type:** bgfbw-bam

Parameter	Parameter in the GUI	Type
out-mode	Output directory	<i>numeric</i>
custom-dir	Custom directory	<i>string</i>
genome	Genome	<i>string</i>
out-name	Output name	<i>string</i>
bs	Block size	<i>numeric</i>
its	Items per slot	<i>numeric</i>
unc	Uncompressed	<i>boolean</i>

## Input/Output Ports

The element has 1 *input* port:

**Name in GUI:** BedGrapgh files

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

And 1 *output* port:

**Name in GUI:** BigWig files

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

## Convert Text to Sequence Element

Converts the input text to a sequence.

**Parameters in GUI**

Parameter	Description	Default value
<b>Sequence name</b> (required)	Result sequence name.	<i>Sequence</i>
<b>Sequence alphabet</b>	Alphabet of the sequence. Choose <i>Auto</i> to auto-detect the alphabet or one of the following values: <ul style="list-style-type: none"> <li><i>All symbols</i></li> <li><i>Extended DNA</i></li> <li><i>Extended RNA</i></li> <li><i>Standard DNA</i></li> <li><i>Standard RNA</i></li> <li><i>Standard amino</i></li> </ul>	<i>Auto</i>
<b>Skip unknown symbols</b>	If <i>True</i> , ignores all symbols that are not presented in the sequence alphabet selected.	<i>True</i>
<b>Replace unknown symbols with</b>	Replaces all unknown symbols with the specified symbol.	<i>N</i>

## Parameters in Workflow File

**Type:** convert-text-to-sequence

Parameter	Parameter in the GUI	Type
sequence-name	Sequence name	string
alphabet	Alphabet	string
skip-unknown	Skip unknown symbols	boolean
replace-unknown-with	Replace unknown symbols with	string (1 character)



## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input text*

**Name in Workflow File:** in-text

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Plain text	text	string

And 1 *output port*:

**Name in GUI:** *Output sequence*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## File Format Conversion Element

Converts the file to selected format if it is not excluded.

**Parameters in GUI**

Parameter	Description	Default value
Document format	Document format of output file.	
Excluded formats	Input file won't be converted to any of selected formats.	

**Parameters in Workflow File**

**Type:** files-conversion

Parameter	Parameter in the GUI	Type
document-format	Document format	string
excluded-formats	Excluded formats	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	input-url	string

And 1 *output port*:

**Name in GUI:** File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	output-url	string

## Reverse Complement Element

Converts input sequence into its reverse, complement or reverse-complement counterpart.

### Parameters in GUI

Parameter	Description	Default value
Operation type	Selects either to produce the reverse, complement, or reverse-complement sequence.	Reverse Complement

## Parameters in Workflow File

**Type:** reverse-complement

Parameter	Parameter in the GUI	Type
op-type	Operation type	string  Available values are: <ul style="list-style-type: none"> <li>reverse-complement</li> <li>complement</li> <li>reverse</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Output sequence*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## Split Assembly into Sequences Element

Splits assembly into sequences(reads).

**Type:** reverse-complement

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** in-assembly

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Assembly data	assembly	assembly

And 1 *output port*.

**Name in GUI:** out-sequence

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	seq	string

## DNA Assembly

- [Align reads with BWA-MEM](#)
- [Assembly Sequences with CAP3](#)
- [Extract Consensus from Assembly](#)

### Align reads with BWA-MEM

Performs alignment of short reads with BWA-MEM.

## Parameters in GUI

Parameter	Description	Default value
Output directory	Directory to save BWA-MEM output files.	
Reference genome	Path to indexed reference genome.	
Output file name	Base name of the output file. 'out.sam' by default.	out.sam
Number of threads	Number of threads (-t).	1
Min seed length	Path to indexed reference genome (-k).	19
Band width	Band width for banded alignment (-w).	100
Dropoff	Off-diagonal X-dropoff (-d).	100
Internal seed length	Look for internal seeds inside a seed longer than {-k} (-r).	1.50000
Skip seed threshold	Skip seeds with more than INT occurrences (-c).	10000
Drop chain threshold	Drop chains shorter than FLOAT fraction of the longest overlapping chain (-D).	0.5
Rounds of made rescues	Perform at most INT rounds of mate rescues for each read (-m).	100
Skip mate rescue	Skip mate rescue (-S).	False
Skip pairing	Skip pairing; mate rescue performed unless -S also in use (-P).	False
Mismatch penalty	Score for a sequence match (-A).	1
Mismatch penalty	Penalty for a mismatch (-B).	4
Gap open penalty	Gap open penalty (-O).	6

<b>Gap extension penalty</b>	Gap extension penalty; a gap of size k cost {-O} {-E}.	1
<b>Penalty for clipping</b>	Penalty for clipping (-L).	5
<b>Penalty unpaired</b>	Penalty for an unpaired read pair (-U).	17
<b>Score threshold</b>	Minimum score to output (-T).	30

### Parameters in Workflow File

**Type:** bwamem-id

Parameter	Parameter in the GUI	Type
output-dir	Output directory	string
reference	Reference genome	string
outname	Output file name	string
threads	Number of threads	numeric
min-seed	Min seed length	numeric
band-width	Band width	numeric
dropoff	Dropoff	numeric
seed-lookup	Internal seed length	numeric
seed-threshold	Skip seed threshold	numeric
drop-chains	Drop chain threshold	numeric
mate-rescue	Rounds of made rescues	numeric
skip-mate-rescues	Skip mate rescue	boolean
skip-pairing	Skip pairing	boolean
match-score	Mismatch penalty	numeric
mismatch-penalty	Mismatch penalty	numeric
gap-open-penalty	Gap open penalty	numeric
gap-ext-penalty	Gap extension penalty	numeric
clipping-penalty	Penalty for clipping	numeric
inpaired-penalty	Penalty unpaired	numeric
score-threshold	Score threshold	numeric

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** BWA data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
URL of a file with mate reads	readsurl	string
URL of a file with reads	readspairedurl	string

And 1 *output port*:

Name in GUI: BWA-MEM output data

Name in Workflow File: out-data

Slots:

Slot In GUI	Slot in Workflow File	Type
Assembly URL	assembly-out	string

## Assembly Sequences with CAP3

CAP3 is a contig assembly program. It allows to assembly long DNA reads (up to 1000 bp). Binaries can be downloaded from <http://seq.cs.ia.state.edu/cap3.html> Huang, X. and Madan, A. (1999) CAP3: A DNA Sequence Assembly Program, Genome Research, 9: 868-877.

## Parameters in GUI

Parameter	Description	Default value
Output file	Write assembly results to this output file in ACE format..	result.ace
Quality cutoff for clipping	Base quality cutoff for clipping (-c).	12
Clipping range	Set a number which unit is base. It will get the refGenes in n bases from peak center. (--distance).	100
Quality cutoff for differences	Base quality cutoff for differences (-b).	20
Maximum difference score	Max qscore sum at differences (-d). If an overlap contains lots of differences at bases of high quality, then the overlap is removed. The difference score is calculated as follows. If the overlap contains a difference at bases of quality values q1 and q2, then the score at the difference is $\max(0, \min(q1, q2) - b)$ , where b is Quality cutoff for differences. The difference score of an overlap is the sum of scores at each difference.	200
Match score factor	Match score factor (-m) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	2
Mismatch score factor	Mismatch score factor (-n) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	-5
Gap penalty factor	Gap penalty factor (-g) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	6

<b>Overlap similarity score cutoff</b>	If the similarity score of an overlap is less than the overlap similarity score cutoff (-s), then the overlap is removed. The similarity score of an overlapping alignment is defined using base quality values as follows. A match at bases of quality values q1 and q2 is given a score of $m * \min(q1, q2)$ , where m is Match score factor. A mismatch at bases of quality values q1 and q2 is given a score of $n * \min(q1, q2)$ , where n is Mismatch score factor. A base of quality value q1 in a gap is given a score of $-g * \min(q1, q2)$ , where q2 is the quality value of the base in the other sequence right before the gap and g is Gap penalty factor. The score of a gap is the sum of scores of each base in the gap minus a gap open penalty. The similarity score of an overlapping alignment is the sum of scores of each match, each mismatch, and each gap.	900
<b>Overlap length cutoff</b>	An overlap is taken into account only if the length of the overlap in bp is no less than the specified value (parameter -o of CAP3).	40
<b>Overlap percent identity cutoff</b>	An overlap is taken into account only if the percent identity of the overlap is no less than the specified value (parameter -p of CAP3).	90
<b>Max number of word matches</b>	This parameter allows one to trade off the efficiency of the program for its accuracy (parameter -t of CAP3). For a read f, CAP3 computes overlaps between read f and other reads by considering short word matches between read f and other reads. A word match is examined to see if it can be extended into a long overlap. If read f has overlaps with many other reads, then read f has many short word matches with many other reads. This parameter gives an upper limit, for any word, on the number of word matches between read f and other reads that are considered by CAP3. Using a large value for this parameter allows CAP3 to consider more word matches between read f and other reads, which can find more overlaps for read f, but slows down the program. Using a small value for this parameter has the opposite effect.	300
<b>Band expansion size</b>	CAP3 determines a minimum band of diagonals for an overlapping alignment between two sequence reads. The band is expanded by a number of bases specified by this value (parameter -a of CAP3).	20
<b>Max gap length in an overlap</b>	The maximum length of gaps allowed in any overlap (-f). I.e. overlaps with longer gaps are rejected. Note that a small value for this parameter may cause the program to remove true overlaps and to produce incorrect results. The parameter may be used to split reads from alternative splicing forms into separate contigs.	20

<b>Assembly reverse reads</b>	Specifies whether to consider reads in reverse orientation for assembly (originally, parameter -r of CAP3).	True
<b>CAP3 tool path</b>	The path to the CAP3 external tool in UGENE.	default
<b>Temporary directory</b>	The directory for temporary files.	default

### Parameters in Workflow File

Type: cap3

Parameter	Parameter in the GUI	Type
out-file	Output file	string
clipping-cutoff	Quality cutoff for clipping	numeric
clipping-range	Clipping range	numeric
diff-cutoff	Quality cutoff for differeneecs	numeric
diff-max-qscore	Maximum difference score	numeric
match-score-factor	Match score factor	numeric
mismatch-score-factor	Mismatch score factor	numeric
gap-penalty-factor	Gap penalty factor	numeric
overlap-sim-score-cutoff	Overlap similarity score cutoff	numeric
overlap-length-cutoff	Overlap length cutoff	numeric
overlap-perc-id-cutoff	Overlap percent identity cutoff	numeric
max-num-word-matches	Max number of word matches	numeric
band-exp-size	Band expansion size	numeric
max-gap-in-overlap	Max gap length in an overlap	numeric
assembly-reverse	Assembly reverse reads	boolean
path	CAP3 tool path	string
tmp-dir	Temporary directory	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input sequences

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Dataset name	dataset	string
Input URL(s)	in.url	string

## Extract Consensus from Assembly

Extract the consensus sequence from the incoming assembly.

## Parameters in GUI

Parameter	Description	Default value
<b>Algorithm</b>	The algorithm of consensus extracting.	Default
<b>Keep gaps</b>	Set this parameter if the result consensus must keep the gaps.	True

### Parameters in Workflow File

**Type:** extract-consensus

Parameter	Parameter in the GUI	Type
<b>algorithm</b>	<b>Algorithm</b>	<i>string</i>
<b>keep-gaps</b>	<b>Keep gaps</b>	<i>boolean</i>

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** in-assembly

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Assembly data</b>	<b>assembly</b>	<i>assembly</i>

And 1 *outut port*:

**Name in GUI:** out-sequence

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>string</i>

## HMMER2 Tools

- [HMM Build Element](#)
- [HMM Search Element](#)
- [Read HMM Profile Element](#)
- [Write HMM Profile Element](#)

### HMM Build Element

Builds a HMM profile from a multiple sequence alignment. The HMM profile is a statistical model which captures position-specific information about how conserved each column of the alignment is, and which residues are likely.

#### Parameters in GUI

Parameter	Description	Default value
<b>Profile name</b>	Descriptive name of the HMM profile.	
<b>HMM strategy</b>	Specifies the kind of alignments you want to allow.	hmmls



<b>Calibrate profile</b>	Enables/disables optional profile calibration. An empirical HMM calibration costs time but it only has to be done once per model, and can greatly increase the sensitivity of a database search.	True
<b>Parallel calibration</b>	Number of parallel threads that the calibration will run in.	1
<b>Standard deviation</b>	Standard deviation of the synthetic sequence length. A positive number. Note that the Gaussian is left-truncated so that no sequences have lengths.	200.0
<b>Fixed length of samples</b>	Fixes the length of the random sequences to, where is a positive (and reasonably sized) integer. The default is instead to generate sequences with a variety of different lengths, controlled by a Gaussian (normal) distribution.	0
<b>Mean length of samples</b>	Mean length of the synthetic sequences, positive real number.	325
<b>Number of samples</b>	Number of synthetic sequences. If is less than about 1000, the fit to the EVD may fail. Higher numbers of will give better determined EVD parameters. The default is 5000; it was empirically chosen as a tradeoff between accuracy and computation time.	5000
<b>Random seed</b>	The random seed, where is a positive integer. The default is to use time() to generate a different seed for each run, which means that two different runs of hmmcalibrate on the same HMM will give slightly different results. You can use this option to generate reproducible results for different hmmcalibrate runs on the same HMM.	0

## Parameters in Workflow File

Type: hmm2-build

Parameter	Parameter in the GUI	Type
<b>profile-name</b>	<b>Profile name</b>	<i>string</i>
<b>strategy</b>	<b>HMM strategy</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for hmms</li> <li>• 1 - for hmmls</li> <li>• 2 - for hmmsfs</li> <li>• 3 - for hmmsw</li> </ul>
<b>calibrate</b>	<b>Calibrate profile</b>	<i>boolean</i>
<b>calibration-threads</b>	<b>Parallel calibration</b>	<i>numeric</i>
<b>deviation</b>	<b>Standard deviation</b>	<i>numeric</i>
<b>fix-samples-length</b>	<b>Fixed length of samples</b>	<i>numeric</i>
<b>mean-samples-length</b>	<b>Mean length of samples</b>	<i>numeric</i>

<b>samples-num</b>	<b>Number of samples</b>	<i>numeric</i>
<b>seed</b>	<b>Random seed</b>	<i>numeric</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>msa</i>

And 1 *output port*:

**Name in GUI:** *HMM profile*

**Name in Workflow File:** out-hmm2

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>HMM profile</b>	<b>hmm2-profile</b>	<i>hmm2-profile</i>

## HMM Search Element

Searches each input sequence for significantly similar sequence matches to all specified HMM profiles. In case several profiles were supplied, searches with all profiles one by one and outputs united set of annotations for each sequence

**Parameters in GUI**

Parameter	Description	Default value
<b>Result annotation</b>	Name of the result annotations.	hmm_signal
<b>Filter by high E-value</b>	E-value filtering can be used to exclude low-probability hits from result.	1e-1
<b>Number of seqs</b>	Calculates the E-value scores as if we had seen a sequence database of sequences.	1
<b>Filter by low score</b>	Score based filtering is an alternative to E-value filtering to exclude low-probability hits from result.	-1000000000.0

## Parameters in Workflow File

**Type:** hmm2-search

Parameter	Parameter in the GUI	Type
<b>result-name</b>	<b>Result annotation</b>	<i>string</i>
<b>e-val</b>	<b>Filter by high E-value</b>	<i>numeric</i>
<b>seqs-num</b>	<b>Number of seqs</b>	<i>numeric</i>
<b>score</b>	<b>Filter by low score</b>	<i>numeric</i>

## Input/Output Ports

The element has 2 *input port*. The first gets the input sequence:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The second input port gets the HMM profile:

**Name in GUI:** *HMM profile*

**Name in Workflow File:** in-hmm2

**Slots:**

Slot In GUI	Slot in Workflow File	Type
HMM profile	hmm2-profile	hmm2-profile

And 1 *output port*:

**Name in GUI:** *HMM annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Read HMM Profile Element

Reads HMM profiles from file(s). The files can be local or Internet URLs.

**Parameters in GUI**

Parameter	Description	Default value
Input files (required)	Semicolon-separated list of paths to the input files.	

## Parameters in Workflow File

**Type:** hmm2-read-profile

Parameter	Parameter in the GUI	Type
url-in	Input files	string

## Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** *HMM profile*

**Name in Workflow File:** out-hmm2

**Slots:**

Slot In GUI	Slot in Workflow File	Type
HMM profile	hmm2-profile	hmm2-profile

## Write HMM Profile Element

Saves all input HMM profiles to specified location.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, the “Location” slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename

## Parameters in Workflow File

**Type:** hmm2-write-profile

Parameter	Parameter in the GUI	Type
<b>url-out</b>	<b>Output file</b>	<i>string</i>
<b>write-mode</b>	<b>Existing file</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *HMM profile*

**Name in Workflow File:** in-hmm2

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>HMM profile</b>	<b>hmm2-profile</b>	<i>hmm2-profile</i>
<b>Location</b>	<b>url</b>	<i>string</i>

## HMMER3 Tools

- [HMM3 Build Element](#)
- [HMM3 Search Element](#)
- [Read HMM3 Profile](#)
- [Write HMM3 Profile](#)

### HMM3 Build Element

Builds a HMM3 profile from a multiple sequence alignment. The HMM3 profile is a statistical model which captures position-specific information about how conserved each column of the alignment is, and which residues are likely.

**Parameters in GUI**

Parameter	Description	Default value
<b>Random seed</b>	Random generator seed. 0 - means that one-time arbitrary seed will be used.	0

**Parameters in Workflow File**

**Type:** hmm3-build

Parameter	Parameter in the GUI	Type
<b>seed</b>	<b>Random seed</b>	<i>numeric</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

**Name in GUI:** *HMM3 profile*

**Name in Workflow File:** out-hmm3

**Slots:**

Slot In GUI	Slot in Workflow File	Type
HMM profile	hmm3-profile	hmm3-profile

## HMM3 Search Element

Searches each input sequence for significantly similar sequence matches to all specified HMM profiles. In case several profiles were supplied, searches with all profiles one by one and outputs united set of annotations for each sequence.

**Parameters in GUI**

Parameter	Description	Default value
<b>Result annotation</b>	Name of the result annotations.	hmm_signal
<b>Seed</b>	Random generator seed. 0 - means that one-time arbitrary seed will be used.	0
<b>Filter by high E-value</b>	E-value filtering can be used to exclude low-probability hits from result.	1e-1
<b>Filter by low score</b>	Score based filtering is an alternative to E-value filtering to exclude low-probability hits from result.	0.01

**Parameters in Workflow File**

**Type:** hmm3-search

Parameter	Parameter in the GUI	Type
result-name	Result annotation	string
seed	Seed	numeric
seqs-num	Number of seqs	numeric
score	Filter by low score	numeric

## Input/Output Ports

The element has 2 *input port*. The first gets the input sequence:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The second input port gets the HMM profile:

**Name in GUI:** *HMM3 profile*

**Name in Workflow File:** in-hmm3

**Slots:**

Slot In GUI	Slot in Workflow File	Type
HMM profile	hmm3-profile	hmm3-profile

And 1 *output port*:

**Name in GUI:** *HMM3 annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Read HMM3 Profile

Reads HMM3 profiles from file(s). The files can be local or Internet URLs.

**Parameters in GUI**

Parameter	Description	Default value
Input files (required)	Semicolon-separated list of paths to the input files.	

**Parameters in Workflow File**

**Type:** hmm3-read-profile

Parameter	Parameter in the GUI	Type
url-in	Input files	string

### Input/Output Ports

The element has 1 *output port*:

**Name in GUI:** *HMM3 profile*

**Name in Workflow File:** out-hmm3

**Slots:**

Slot In GUI	Slot in Workflow File	Type
HMM profile	hmm3-profile	hmm3-profile

## Write HMM3 Profile

Saves all input HMM3 profiles to specified location.

**Parameters in GUI**

Parameter	Description	Default value
-----------	-------------	---------------

<b>Output file</b>	Location of the output data file. If this attribute is set, the "Location" slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format). If Rename option is chosen existing file will be renamed.	Rename

### Parameters in Workflow File

**Type:** hmm3-write-profile

Parameter	Parameter in the GUI	Type
<b>url-out</b>	<b>Output file</b>	<i>string</i>
<b>write-mode</b>	<b>Existing file</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *HMM3 profile*

**Name in Workflow File:** in-hmm3

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>HMM profile</b>	<b>hmm3-profile</b>	<i>hmm3-profile</i>
<b>Location</b>	<b>url</b>	<i>string</i>

## Multiple Sequence Alignment

- [Align Profile to Profile with MUSCLE Element](#)
- [Align with ClustalO Element](#)
- [Align with ClustalW Element](#)
- [Align with Kalign Element](#)
- [Align with MAFFT Element](#)
- [Align with MUSCLE Element](#)
- [Align with T-Coffee Element](#)
- [Extract Consensus from Alignment](#)
- [Join Sequences into Alignment Element](#)
- [Split Alignment into Sequences Element](#)

### Align Profile to Profile with MUSCLE Element

Aligns second profile to master profile with MUSCLE aligner.

**Type:** align-profile-to-profile

#### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** in-profiles

**Name in Workflow File:** in-profiles

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Master profile	master-msa	<i>malignment</i>
Second profile	second-msa	<i>malignment</i>

And 1 *output port*:

**Name in GUI:** out-msa

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	<i>malignment</i>

## Align with ClustalO Element

Aligns multiple sequence alignments (MSAs) supplied with ClustalO.

**Parameters in GUI**

Parameter	Description	Default value
Number of iterations	Number of (combined guide-tree/HMM) iterations.	1
Number of guidetree iterations	Maximum number guidetree iterations.	0
Number of HMM iterations	Maximum number of HMM iterations.	0
Set auto options	Set options automatically (might overwrite some of your options).	False
Tool path	Path to the ClustalO tool.  The default path can be set in the UGENE application settings.	Default
Temporary directory	Directory to store temporary files.	Default

**Parameters in Workflow File**

**Type:** ClustalO

Parameter	Parameter in the GUI	Type
num-iterations	Number of iterations	<i>numeric</i>
max-guidetree-iterations	Number of guidetree iterations	<i>numeric</i>
max-hmm-iterations	Number of HMM iterations	<i>numeric</i>
set-auto	Set auto options	<i>boolean</i>
path	Tool path	<i>string</i>
temp-dir	Temporary directory	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input MSA

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
-------------	-----------------------	------



<b>MSA</b>	<b>msa</b>	<i>malignment</i>
------------	------------	-------------------

And 1 *output port*:

**Name in GUI:** ClustalO result MSA

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>malignment</i>

## Align with ClustalW Element

Aligns multiple sequence alignments (MSAs) supplied with ClustalW.

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. Visit <http://www.clustal.org/> to learn more about it.



Clustal is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

## Parameters in GUI

Parameter	Description	Default value
<b>Weight matrix</b>	For proteins it is a scoring table which describes the similarity of each amino acid to each other. For DNA it is the scores assigned to matches and mismatches.	default
<b>End gaps</b>	The penalty for closing a gap.	False
<b>Gap distance</b>	The gap separation penalty. Tries to decrease the chances of gaps being too close to each other.	4.42
<b>Gap extension penalty</b>	The penalty for extending a gap.	8.52
<b>Gap open penalty</b>	The penalty for opening a gap.	53.90
<b>Hydrophilic gaps off</b>	Hydrophilic gap penalties are used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids.	False
<b>Residue-specific gaps off</b>	Residue-specific penalties are amino specific gap penalties that reduce or increase the gap opening penalties at each position in the alignment.	False
<b>Iteration type</b>	Alignment improvement iteration type.	None
<b>Number of iterations</b>	The maximum number of iterations to perform.	3
<b>Tool path (required)</b>	Path to the ClustalW tool. The default path can be set in the UGENE Application Settings.	default
<b>Temporary directory</b>	Directory to store temporary files.	default

## Parameters in Workflow File

**Type:** clustalw

Parameter	Parameter in the GUI	Type
matrix	Weight matrix	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for IUB</li> <li>• 1 - for ClustalW</li> <li>• 2 - for BLOSUM</li> <li>• 3 - for PAM</li> <li>• 4 - for GONNET</li> <li>• 5 - for ID</li> <li>• -1 - for default matrix</li> </ul>
close-gap-penalty	End gaps	<i>boolean</i>
gap-distance	Gap distance	<i>numeric</i>
gap-ext-penalty	Gap extension penalty	<i>numeric</i>
gap-open-penalty	Gap open penalty	<i>numeric</i>
no-hydrophilic-gaps	Hydrophilic gaps off	<i>boolean</i>
no-residue-specific-gaps	Residue-specific gaps off	<i>boolean</i>
iteration-type	Iteration type	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for None</li> <li>• 1 - for Tree</li> <li>• 2 - for Alignment</li> </ul>
iterations-max-num	Number of iterations	<i>numeric</i>
path	Tool path	<i>string</i>
temp-dir	Temporary directory	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	<i>msa</i>

And 1 *output port*:

**Name in GUI:** *ClustalW result MSA*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	<i>msa</i>

## Align with Kalign Element

Aligns multiple sequence alignments (MSAs) supplied with Kalign. Kalign is a fast and accurate multiple sequence alignment tool. The

original version of the tool can be found on <http://msa.sbc.su.se>.

### Parameters in GUI

Parameter	Description	Default value
Gap extension penalty	The penalty for extending a gap.	8.52
Gap open penalty	The penalty for opening/closing a gap. Half the value will be subtracted from the alignment score when opening, and half when closing a gap.	54.90
Terminal gap penalty	The penalty to extend gaps from the N/C terminal of protein or 5'/3' terminal of nucleotide sequences.	4.42
Bonus score	A bonus score that is added to each pair of aligned residues.	0.02

## Parameters in Workflow File

Type: kalign

Parameter	Parameter in the GUI	Type
gap-ext-penalty	Gap extension penalty	numeric
gap-open-penalty	Gap open penalty	numeric
terminal-gap-penalty	Terminal gap penalty	numeric
bonus-score	Bonus score	numeric

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

**Name in GUI:** *Kalign result MSA*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

## Align with MAFFT Element

Originally, MAFFT is a multiple sequence alignment program for unix-like operating systems. Currently, Windows version is also available.



MAFFT is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

MAFFT is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

#### Parameters in GUI

Parameter	Description	Default value
Offset	Works like gap extension penalty.	0
Gap open penalty	Gap open penalty.	1.53
Max iteration	Maximum number of iterative refinement.	0
Tool path (default)	Path to the ClustalW tool. The default path can be set in the UGENE application settings.	default
Temporary directory	Directory to store temporary files.	default

## Parameters in Workflow File

Type: mafft

Parameter	Parameter in the GUI	Type
gap-ext-penalty	Offset	numeric
gap-open-penalty	Gap open penalty	numeric
iterations-max-num	Max iteration	numeric
path	Tool path	string
temp-dir	Temporary directory	string

## Input/Output Ports

The element has 1 *input port*:

Name in GUI: *Input MSA*

Name in Workflow File: in-msa

Slots:

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

Name in GUI: *Multiple sequence alignment*

Name in Workflow File: out-msa

Slots:

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

## Align with MUSCLE Element

MUSCLE is public domain multiple alignment software for protein and nucleotide sequences. MUSCLE stands for Multiple Sequence Comparison by Log-Expectation.

#### Parameters in GUI

Parameter	Description	Default value
-----------	-------------	---------------

<b>Mode</b>	Selector of preset configurations, that give you the choice of optimizing accuracy, speed, or some compromise between the two. The default favors accuracy.	MUSCLE default
<b>Stable order</b>	Do not rearrange aligned sequences (-stable switch of MUSCLE). Otherwise, MUSCLE re-arranges sequences so that similar sequences are adjacent in the output file. This makes the alignment easier to evaluate by eye.	True

## Parameters in Workflow File

Type: muscle

Parameter	Parameter in the GUI	Type
<b>mode</b>	<b>Mode</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for MUSCLE default</li> <li>• 1 - for Large alignment</li> <li>• 2 - for Refine only</li> </ul>
<b>stable</b>	<b>Stable order</b>	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>msa</i>

And 1 *output port*:

**Name in GUI:** *Multiple sequence alignment*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>msa</i>

## Align with T-Coffee Element

T-Coffee is a multiple sequence alignment package.



T-Coffee is used as an external tool from UGENE and it must be installed on your system. To learn more about the external tools, please, read main [UGENE User Manual](#).

## Parameters in GUI

Parameter	Description	Default value
-----------	-------------	---------------

<b>Gap extension penalty</b>	Gap Extension Penalty. Positive values give rewards to gaps and prevent the alignment of unrelated segments.	0
<b>Gap open penalty</b>	Gap open penalty. Must be negative, best matches get a score of 1000.	-50
<b>Max iteration</b>	Number of iteration on the progressive alignment. 0 - no iteration, -1 - Nseq iterations.	0
<b>Tool path (required)</b>	Path to the ClustalW tool. The default path can be set in the UGENE Application Settings.	default
<b>Temporary directory</b>	Directory to store temporary files.	default

## Parameters in Workflow File

Type: tcoffee

Parameter	Parameter in the GUI	Type
gap-ext-penalty	Offset	numeric
gap-open-penalty	Gap open penalty	numeric
iterations-max-num	Max iteration	numeric
path	Tool path	string
temp-dir	Temporary directory	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input MSA*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

**Name in GUI:** *Multiple sequence alignment*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

## Extract Consensus from Alignment

Extract the consensus sequence from the incoming multiple sequence alignment.

**Parameters in GUI**

Parameter	Description	Default value
<b>Algorithm</b>	The algorithm of consensus extracting.	
<b>Threshold</b>	The threshold of the algorithm.	100

<b>Keep gaps</b>	Set this parameter if the result consensus must keep the gaps.	True
------------------	--	------

### Parameters in Workflow File

**Type:** extract-msa-consensus

Parameter	Parameter in the GUI	Type
algorithm	Algorithm	string
threshold	Threshold	numeric
keep-gaps	Keep gaps	boolean

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *in-msa*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

**Name in GUI:** *out-sequence*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	seq

## Join Sequences into Alignment Element

Creates a multiple sequence alignment from sequences.

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequences*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Result alignment*

**Name in Workflow File:** out-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

## Split Alignment into Sequences Element

Splits an input alignment into sequences.

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input alignment*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa

And 1 *output port*:

**Name in GUI:** *Output sequences*

**Name in Workflow File:**

**Slots:** out-sequence

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## NGS Basic

- [CASAVA FASTQ Filter Element](#)
- [FASTQ Quality Trimmer Element](#)
- [Filter BAM/SAM Files Element](#)
- [Genome Coverage Element](#)
- [Merge BAM Files Element](#)
- [Slopbed Element](#)
- [Sort BAM Files Element](#)

### CASAVA FASTQ Filter Element

Reads in FASTQ file produced by CASAVA 1.8 contain 'N' or 'Y' as a part of an identifier. 'Y' if a read is filtered, 'N' if the read is not filtered. The workflow cleans up the filtered reads. For example: @HWI-ST880:181:D1WRUACXX:8:1102:4905:2125 1:N:0:TAAGGG CTTACATAACTACTGACCATGCTCTCTCTTGTCTGTCTCTTATACACATCT + 11144222322324232AAFFHIJJJJJIHIF111CGGFHIG???FGB @HWI-ST880:181:D1WRUACXX:8:1102:7303:2101 1:Y:0:TAAGGG TCCTTACTGTCTGAGCAATGGGATTCCATCTTTACGATCTAGACATGGCT + 11++4222322.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	
<b>Output file name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extension.	

## Parameters in Workflow File

**Type:** CASAVAFilter



Parameter	Parameter in the GUI	Type
out-mode	Output directory	numeric
custom-dir	Custom directory	string
out-name	Output file name	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

And 1 *output port*:

**Name in GUI:** Output File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

## FASTQ Quality Trimmer Element

The workflow scans each input sequence from the end to find the first position where the quality is greater or equal to the minimum quality threshold. Then it trims the sequence to that position. If a the whole sequence has quality less than the threshold or the length of the output sequence less than the minimum length threshold then the sequence is skipped.

## Parameters in GUI

Parameter	Description	Default value
Output directory	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
Custom directory	Specify the output directory.	
Output file name	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	
Quality threshold	Quality threshold for trimming.	30
Min Length	Too short reads are discarded by the filter.	0

**Parameters in Workflow File**

**Type:** QualityTrim

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

<b>out-mode</b>	<b>Output directory</b>	<i>numeric</i>
<b>custom-dir</b>	<b>Custom directory</b>	<i>string</i>
<b>out-name</b>	<b>Output file name</b>	<i>string</i>
<b>qual-id</b>	<b>Quality threshold</b>	<i>numeric</i>
<b>len-id</b>	<b>Min Length</b>	<i>numeric</i>

Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	<i>string</i>

And 1 *output port*:

**Name in GUI:** Output File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	<i>string</i>

## Filter BAM/SAM Files Element

Filters BAM/SAM files using SAMTools view.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	
<b>Custom directory</b>	Custom output directory.	
<b>Output name</b>	A name of an output BAM/SAM file. If default of empty value is provided the output name is the name of the first BAM/SAM file with .filtered extention.	
<b>Output format</b>	Format of an output assembly file.	bam
<b>Region</b>	Regions to filter. For BAM output only. chr2 to output the whole chr2. <a href="#">chr2:1000</a> to output regions of chr 2 starting from 1000. <a href="#">chr2:1000-2000</a> to ouput regions of chr2 between 1000 and 2000 including the end point. To input multiple regions use the space separator (e.g. chr1 chr2 <a href="#">chr3:1000-2000</a> ).	
<b>MAPQ threshold</b>	Minimum MAPQ quality score.	0

<b>Skip flag</b>	Skip alignment with the selected items. Select the items in the combobox to configure bit flag. Do not select the items to avoid filtration by this parameter.	
------------------	---	--

### Parameters in Workflow File

**Type:** filter-bam

Parameter	Parameter in the GUI	Type
<b>out-mode</b>	<b>Output directory</b>	<i>numeric</i>
<b>custom-dir</b>	<b>Custom directory</b>	<i>string</i>
<b>out-name</b>	<b>Output name</b>	<i>string</i>
<b>out-format</b>	<b>Output format</b>	<i>string</i>
<b>region</b>	<b>Region</b>	<i>string</i>
<b>mapq</b>	<b>MAPQ threshold</b>	<i>numeric</i>
<b>flag</b>	<b>Skip flag</b>	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** BAM/SAM File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Source URL</b>	<b>input-url</b>	<i>string</i>

And 1 *output port*:

**Name in GUI:** Filtered BAM/SAM files

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Source URL</b>	<b>output-url</b>	<i>string</i>

## Genome Coverage Element

Calculates genome coverage using bedtools genomecov.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	

<b>Output file name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	
<b>Genome</b>	In order to prevent the extension of intervals beyond chromosome boundaries, bedtools slop requires a genome file defining the length of each chromosome or contig (-g).	human.hg18
<b>Report mode</b>	<p>Histogram () - Compute a histogram of coverage.</p> <p>Per-base (0-based) (-dz) - Compute the depth of feature coverage for each base on each chromosome (0-based).</p> <p>Per-base (1-based) (-d) - Compute the depth of feature coverage for each base on each chromosome (1-based)</p> <p>BEDGRAPH (-bg) - Produces genome-wide coverage output in BEDGRAPH format.</p> <p>BEDGRAPH (including uncovered) (-bga) - Produces genome-wide coverage output in BEDGRAPH format (including uncovered).</p>	Histogram
<b>Split</b>	Treat âsplitâ BAM or BED12 entries as distinct BED intervals when computing coverage. For BAM files, this uses the CIGAR âNâ and âDâ operations to infer the blocks for computing coverage. For BED12 files, this uses the BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10,11,12) (-split).	False
<b>Strand</b>	Calculate coverage of intervals from a specific strand. With BED files, requires at least 6 columns (strand is column 6) (-strand).	False
<b>5 prime</b>	Calculate coverage of 5â positions (instead of entire interval) (-5).	False
<b>3 prime</b>	Calculate coverage of 3â positions (instead of entire interval) (-3).	False
<b>Max</b>	Combine all positions with a depth >= max into a single bin in the histogram (-max).	2147483647
<b>Scale</b>	Scale the coverage by a constant factor.Each coverage value is multiplied by this factor before being reported. Useful for normalizing coverage by, e.g., reads per million (RPM). Default is 1.0; i.e., unscaled (-scale).	1.00000
<b>Trackline</b>	Adds a UCSC/Genome-Browser track line definition in the first line of the output (-trackline).	False
<b>Trackopts</b>	Writes additional track line definition parameters in the first line (-trackopts).	

## Parameters in Workflow File

**Type:** genomecov

Parameter	Parameter in the GUI	Type
out-mode	Output directory	numeric
custom-dir	Custom directory	string
out-name	Output file name	string
genome	Genome	string
mode-id	Report mode	numeric
split-id	Split	boolean
strand-id	Strand	boolean
prime5-id	5 prime	boolean
prime3-id	3 prime	boolean
max-id	Max	numeric
scale-id	Scale	numeric
trackline-id	Trackline	boolean
trackopts-id	Trackopts	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

And 1 *output port*:

**Name in GUI:** Output File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

## Merge BAM Files Element

Merge BAM files using SAMTools merge.

**Parameters in GUI**

Parameter	Description	Default value
Output directory	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	
Custom directory	Custom output directory.	

<b>Output BAM name</b>	A name of an output BAM file. If default of empty value is provided the output name is the name of the first BAM file with .merged.bam extention.	
------------------------	---	--

### Parameters in Workflow File

**Type:** merge-bam

Parameter	Parameter in the GUI	Type
<b>out-mode</b>	<b>Output directory</b>	<i>numeric</i>
<b>custom-dir</b>	<b>Custom directory</b>	<i>string</i>
<b>out-name</b>	<b>Output name</b>	<i>string</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** BAM File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Source URL</b>	<b>input-url</b>	<i>string</i>

And 1 *output port*:

**Name in GUI:** Merged BAM files

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Source URL</b>	<b>output-url</b>	<i>string</i>

## Slopped Element

Increases the size of each feature in files using bedtools slop.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	
<b>Output file name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	

<b>Genome</b>	In order to prevent the extension of intervals beyond chromosome boundaries, bedtools slop requires a genome file defining the length of each chromosome or contig (-g).	human.hg18
<b>Each direction increase</b>	Increase the BED/GFF/VCF entry by the same number base pairs in each direction. If this parameter is used -l and -l are ignored. Enter 0 to disable (-b).	0
<b>Subtract from start</b>	The number of base pairs to subtract from the start coordinate. Enter 0 to disable (-l).	0
<b>Add to end</b>	The number of base pairs to add to the end coordinate. Enter 0 to disable (-r).	0
<b>Strand-based</b>	Define -l and -r based on strand. For example. if used, -l 500 for a negative-stranded feature, it will add 500 bp to the end coordinate (-s).	False
<b>As fraction</b>	Define -l and -r as a fraction of the feature's length. E.g. if used on a 1000bp feature, -l 0.50, will add 500 bp 'upstream' (-pct).	False
<b>Print header</b>	Print the header from the input file prior to results (-header).	False

### Parameters in Workflow File

**Type:** sloped

Parameter	Parameter in the GUI	Type
out-mode	Output directory	numeric
custom-dir	Custom directory	string
out-name	Output file name	string
genome-id	Genome	string
b-id	Each direction increase	numeric
l-id	Subtract from start	numeric
r-id	Add to end	numeric
s-id	Strand-based	boolean
pct-id	As fraction	boolean
header-id	Print header	boolean

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

And 1 *output port*:

**Name in GUI:** Output File

**Name in Workflow File:** out-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	<i>string</i>

## Sort BAM Files Element

Sort BAM Files using SAMTools Sort.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Select an output directory. Custom - specify the output directory in the 'Custom directory' parameter. Workflow - internal workflow directory. Input file - the directory of the input file.	Input file
<b>Custom directory</b>	Specify the output directory.	
<b>Output BAM name</b>	A name of an output file. If default of empty value is provided the output name is the name of the first file with additional extention.	
<b>Build index</b>	Build index for the sorted file with SAMTools index.	human.hg18

## Parameters in Workflow File

**Type:** Sort-bam

Parameter	Parameter in the GUI	Type
<b>out-mode</b>	<b>Output directory</b>	<i>numeric</i>
<b>custom-dir</b>	<b>Output BAM name</b>	<i>string</i>
<b>out-name</b>	<b>Output file name</b>	<i>string</i>
<b>index</b>	<b>Build index</b>	<i>boolean</i>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** BAM File

**Name in Workflow File:** in-file

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source URL	url	<i>string</i>

And 1 *output port*:

**Name in GUI:** Sorted BAM File

**Name in Workflow File:** out-file



Slots:

Slot In GUI	Slot in Workflow File	Type
Source URL	url	string

## NGS: ChiP-Seq Analysis

- Annotate Peaks with peak2gene Element
- Build Conservation Plot Element
- Collect Motifs with SeqPos Element
- Conduct GO Element
- Create CEAS Report Element
- Find Peaks with MACS Element

### Annotate Peaks with peak2gene Element

Gets refGenes near the ChIP regions identified by a peak-caller.

#### Parameters in GUI

Parameter	Description	Default value
Genome file	Select a genome file (sqlite3 file) to search refGenes. (--genome).	hg19
Output file	Select which type of genes need to output. up for genes upstream to peak summit, down for genes downstream to peak summit, all for both up and down. (--op).	all
Official gene symbols	Output official gene symbol instead of refseq name. (--symbol).	False
Distance	Set a number which unit is base. It will get the refGenes in n bases from peak center. (--distance).	3000

#### Parameters in Workflow File

Type: peak2gene-id

Parameter	Parameter in the GUI	Type
genome	Genome file	string
outpos	Output file	string
symbol	Official gene symbols	boolean
distance	Distance	numeric

#### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Peak2gene data

**Name in Workflow File:** in-data

Slots:

Slot In GUI	Slot in Workflow File	Type
Treatment features	_treat-ann	ann-table-list

And 1 *output port*:

**Name in GUI:** Peak2gene output data

**Name in Workflow File:** out-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Gene regions	gene-annotation	<i>ann-table-list</i>
Peak regions	peak-annotation	<i>ann-table-list</i>

**Build Conservation Plot Element**

Plots the PhastCons scores profiles.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b>	File to store phastcons results (BMP).	
<b>Title</b>	Title of the figure (--title).	Average Phastcons around the Center of Sites
<b>Label</b>	Label of data in the figure (--bed-label).	Conservation_at_peak_summits
<b>Assembly version</b>	The directory to store phastcons scores (--phasdb).	hg19
<b>Window width</b>	Window width centered at middle of regions (-w).	1000
<b>Height</b>	Height of plot (--height).	1000
<b>Width</b>	Width of plot (--width).	1000

**Parameters in Workflow File**

**Type:** conservation\_plot-id

Parameter	Parameter in the GUI	Type
output-file	<b>Output file</b>	<i>string</i>
title	<b>Title</b>	<i>string</i>
label	<b>Label</b>	<i>string</i>
assembly_version	<b>Assembly version</b>	<i>string</i>
windows_s	<b>Window width</b>	<i>numeric</i>
height	<b>Height</b>	<i>numeric</i>
width	<b>Width</b>	<i>numeric</i>

**Input/Output Ports**

The element has 1 *input port*.

**Name in GUI:** conservation\_plot data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Input regions	cp_treat-ann	<i>ann-table-list</i>

**Collect Motifs with SeqPos Element**

Finds motifs enriched in a set of regions.

#### Parameters in GUI

Parameter	Description	Default value
Output directory	The directory to store seqpos results.	
Genome assembly version	UCSC database version (GENOME).	hg19
Output file name	Name of the output file which stores new motifs found during a de novo search (-n).	Default
De novo motifs	Run de novo motif search (-d).	False
Motif database	Known motif collections. (-m). Warning: computation time increases with selecting additional databases. It is recommended to use cistrome.xml. It is a comprehensive collection of motifs from the other databases with similar motifs deleted.	cistrome.xml
Region width	Width of the region to be scanned for motifs; depends on a resolution of assay (-w).	600
Pvalue cutoff	Pvalue cutoff for the motif significance (-p).	0.001

#### Parameters in Workflow File

Type: seqpos-id

Parameter	Parameter in the GUI	Type
output-dir	Output directory	string
assembly	Genome assembly version	string
out_name	Output file name	string
de_novo	De novo motifs	boolean
motif_db	Motif database	string
reg_width	Region width	numeric
p_val	Pvalue cutoff	numeric

#### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** SeqPos data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Input regions	cp_treat-ann	ann-table-list

## Conduct GO Element

Given a list of genes, using Bioconductor (GO, GOSTats) and DAVID at NIH.

#### Parameters in GUI

Parameter	Description	Default value
Output directory	The directory to store Conduct GO results.	

<b>Title</b>	Title is used to name the output files - so make it meaningful.	Default
<b>Gene Universe</b>	Select a gene universe.	hgu133a.db

### Parameters in Workflow File

**Type:** conduct-go-id

Parameter	Parameter in the GUI	Type
output-dir	Output directory	string
title	Title	string
gene-universe	Gene Universe	string

### Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** Conduct GO data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Target genes	in-ann	ann-table-list

## Create CEAS Report Element

Provides summary statistics on ChIP enrichment in important genomic regions such as individual chromosomes, promoters, gene bodies or exons, and infers the genes most likely to be regulated by the binding factor under study.

### Parameters in GUI

Parameter	Description	Default value
<b>Output report file</b>	Path to the report output file. Result for CEAS analysis.	
<b>Output annotations file</b>	Name of tab-delimited output text file, containing a row of annotations for every RefSeq gene. (file is not generated if no peak location data is supplied).	
<b>Gene annotations table</b>	Path to gene annotation table (e.g. a refGene table in sqlite3 db format (--gt)).	hg19
<b>Span size</b>	Span from TSS and TTS in the gene-centered annotation (base pairs). ChIP regions within this range from TSS and TTS are considered when calculating the coverage rates in promoter and downstream (--span).	3000
<b>Wiggle profiling resolution</b>	Wiggle profiling resolution. WARNING: Value smaller than the wig interval (resolution) may cause aliasing error. (--pf-res).	50
<b>Promoter/downstream interval</b>	Promoter/downstream intervals for ChIP region annotation are three values or a single value can be given. If a single value is given, it will be segmented into three equal fractions (e.g. 3000 is equivalent to 1000,2000,3000) (--rel-dist).	3000

<b>BiPromoter ranges</b>	Bidirectional-promoter sizes for ChIP region annotation. It's two values or a single value can be given. If a single value is given, it will be segmented into two equal fractions (e.g. 5000 is equivalent to 2500,5000) (--bisizes).	5000
<b>Relative distance</b>	Relative distance to TSS/TTS in WIGGLE file profiling (--rel-dist).	3000
<b>Gene group files</b>	Gene groups of particular interest in wig profiling. Each gene group file must have gene names in the 1st column. The file names are separated by commas (--gn-groups).	
<b>Gene group names</b>	Set this parameter empty for using default values. The names of the gene groups from "Gene group files" parameter. These names appear in the legends of the wig profiling plots. Values range: comma-separated list of strings. Default value: 'Group 1, Group 2,...Group n' (--gn-group-names).	

### Parameters in Workflow File

**Type:** ceas-report

Parameter	Parameter in the GUI	Type
image-file	Output report file	string
anns-file	Output annotations file	string
anns-table	Gene annotations table	string
span	Span size	numeric
profiling-resolution	Wiggle profiling resolution	numeric
promoter-sizes	Promoter/downstream interval	numeric
promoter-bisizes	BiPromoter ranges	string
relative-distance	Relative distance	string
group-files	Gene group files	string
group-names	Gene group names	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** CEAS data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Enrichment signal	enrichment-signal	ann-table-list
Peak regions	peak-regions	string

## Find Peaks with MACS Element

Performs peak calling for ChIP-Seq data.

## Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Directory to save MACS output files.	
<b>Name</b>	The name string of the experiment. MACS will use this string NAME to create output files like 'NAME_peaks.xls', 'NAME_negative_peaks.xls', 'NAME_peaks.bed', 'NAME_summits.bed', 'NAME_model.r' and so on. So please avoid any confliction between these filenames and your existing files (--name).	
<b>Wiggle output</b>	If this flag is on, MACS will store the fragment pileup in wiggle format for the whole genome data instead of for every chromosomes (--wig) (--single-profile).	hg19
<b>Wiggle space</b>	By default, the resolution for saving wiggle files is 10 bps,i.e., MACS will save the raw tag count every 10 bps. You can change it along with '--wig' option (--space).	3000
<b>Genome size (Mbp)</b>	Homo sapience - 2700 Mbp Mus musculus - 1870 Mbp Caenorhabditis elegans - 90 Mbp Drosophila melanogaster - 120 Mbp It's the mappable genome size or effective genome size which is defined as the genome size which can be sequenced. Because of the repetitive features on the chromosomes, the actual mappable genome size will be smaller than the original size, about 90% or 70% of the genome size (--gsize).	50
<b>P-value</b>	P-value cutoff. Default is 0.00001, for looser results, try 0.001 instead (--pvalue).	3000
<b>Tag size (optional)</b>	Length of reads. Determined from first 10 reads if not specified (input 0) (--tsize).	5000
<b>Keep duplicates</b>	It controls the MACS behavior towards duplicate tags at the exact same location -- the same coordination and the same strand. The default auto option makes MACS calculate the maximum tags at the exact same location based on binomial distribution using 1e-5 as pvalue cutoff; and the all option keeps every tags. If an integer is given, at most this number of tags will be kept at the same location (--keep-dup).	3000
<b>Use model</b>	Whether or not to use MACS paired peaks model (--nomodel).	

<b>Model fold</b>	Select the regions within MFOLD range of high-confidence enrichment ratio against. Model fold is available when Use model is true, which is the foldchange to chose paired peaks to build paired peaks model. Users need to set a lower(smaller) and upper(larger) number for fold change so that MACS will only use the peaks within these foldchange range to build model (--mfold).	
<b>Shift size</b>	An arbitrary shift value used as a half of the fragment size when model is not built. Shift size is available when Use model is false, which will represent the HALF of the fragment size of your sample. If your sonication and size selection size is 300 bps, after you trim out nearly 100 bps adapters, the fragment size is about 200 bps, so you can specify 100 here (--shiftsize).	
<b>Band width</b>	The band width which is used to scan the genome for model building. You can set this parameter as the sonication fragment size expected from wet experiment. Used only while building the shifting model (--bw).	
<b>Use lambda</b>	Whether to use local lambda model which can use the local bias at peak regions to throw out false positives (--nolambda).	
<b>Small nearby region</b>	The small nearby region in basepairs to calculate dynamic lambda. This is used to capture the bias near the peak summit region. Invalid if there is no control data (--slocal).	
<b>Large nearby region</b>	The large nearby region in basepairs to calculate dynamic lambda. This is used to capture the surround bias (--llocal).	
<b>Auto bimodal</b>	Whether turn on the auto pair model process.If set, when MACS failed to build paired model, it will use the nomodelsettings, the Shift size parameter to shift and extend each tags (--on-auto).	
<b>Scale to large</b>	When set, scale the small sample up to the bigger sample.By default, the bigger dataset will be scaled down towards the smaller dataset,which will lead to smaller p/qvalues and more specific results.Keep in mind that scaling down will bring down background noise more (--to-large).	

#### Parameters in Workflow File

Type: macs-id

Parameter	Parameter in the GUI	Type
output-dir	Output directory	string
file-names	Name	string
wiggle-output	Wiggle output	boolean

wiggle-space	Wiggle space	numeric
genome-size	Genome size (Mbp)	numeric
p-value	P-value	numeric
tag-size	Tag size (optional)	numeric
keep-duplicates	Keep duplicates	string
use-model	Use model	boolean
model-fold	Model fold	string
shift-size	Shift size	numeric
band-width	Band width	numeric
use-lambda	Use lambda	boolean
small-nearby	Small nearby region	numeric
large-nearby	Large nearby region	numeric
auto_bimodal	Auto bimodal	boolean
scale_large	Scale to large	boolean

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** MACS data

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Treatment features	_treatment-ann	ann-table-list
Control features	control-ann	ann-table-list

And 1 *output port*:

**Name in GUI:** MACS output data

**Name in Workflow File:** out-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Peak regions	peak-regions	ann-table-list
Peak summits	peak-summits	ann-table-list
Treatment fragments pileup	wiggle-treat	string

## NGS: RNA-Seq Analysis

- [Assembly Transcripts with Cufflinks Element](#)
- [Extract Transcript Sequences with gffread Element](#)
- [Find Splice Junction with TopHat Element](#)
- [Merge Assemblies with Cuffmerge Element](#)
- [Test for Diff. Expression with Cuffdiff Element](#)

### Assembly Transcripts with Cufflinks Element



Cufflinks accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

#### Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Directory to save MACS output files.	
<b>Reference annotation</b>	Tells Cufflinks to use the supplied reference annotation to estimate isoform expression. Cufflinks will not assemble novel transcripts and the program will ignore alignments not structurally compatible with any reference transcript.	
<b>RABT annotation</b>	Tells Cufflinks to use the supplied reference annotation to guide Reference Annotation Based Transcript (RABT) assembly. Reference transcripts will be tiled with faux-reads to provide additional information in assembly. Output will include all reference transcripts as well as any novel genes and isoforms that are assembled.	
<b>Library type</b>	Specifies RNA-Seq protocol.	Standart Illumina
<b>Mask file</b>	Ignore all reads that could have come from transcripts in this file. It is recommended to include any annotated rRNA, mitochondrial transcripts other abundant transcripts you wish to ignore in your analysis in this file. Due to variable efficiency of mRNA enrichment methods and rRNA depletion kits, masking these transcripts often improves the overall robustness of transcript abundance estimates.	
<b>Multi-read correct</b>	Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.	False
<b>Min isoform fraction</b>	After calculating isoform abundance for a gene, Cufflinks filters out transcripts that it believes are very low abundance, because isoforms expressed at extremely low levels often cannot reliably be assembled, and may even be artifacts of incompletely spliced precursors of processed transcripts. This parameter is also used to filter out introns that have far fewer spliced alignments supporting them.	0.1
<b>Frag bias correct</b>	Providing Cufflinks with a multifasta file via this option instructs it to run the bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates.	

<b>Pre-mRNA fraction</b>	Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored.	0.15
<b>Cufflinks tool path</b>	The path to the Cufflinks external tool in UGENE.	default
<b>Temporary directory</b>	The directory for temporary files.	default

### Parameters in Workflow File

**Type:** cufflinks

Parameter	Parameter in the GUI	Type
out-dir	Output directory	string
ref-annotation	Reference annotation	string
rabt-annotation	RABT annotation	string
library-type	Library type	numeric
mask-file	Mask file	string
multi-read-correct	Multi-read correct	boolean
min-isoform-fraction	Min isoform fraction	numeric
frag-bias-correct	Frag bias correct	string
pre-mrna-fraction	Pre-mRNA fraction	numeric
path	Cufflinks tool path	string
tmp-dir	Temporary directory	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input reads

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Assembly data	assembly	assembly
Source url	url	string

And 1 *output port*:

**Name in GUI:** Output annotations

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Isoform-level expression values	isolevel.slot	<i>ann_table</i>

## Extract Transcript Sequences with gffread Element

Extract transcript sequences from the genomic sequence(s) with gffread.

### Parameters in GUI

Parameter	Description	Default value
Output sequences	The url to the output file with the extracted sequences.	

### Parameters in Workflow File

Type: gffread

Parameter	Parameter in the GUI	Type
url-out	Output sequences	<i>string</i>

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input transcripts

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Genomic sequence url	genome	<i>string</i>
Transcripts url	transcripts	<i>string</i>

And 1 *output port*:

**Name in GUI:** Extracted sequences url

**Name in Workflow File:** extracted-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
sequences	sequences	<i>string</i>

## Find Splice Junction with TopHat Element

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

### Parameters in GUI

Parameter	Description	Default value
Output directory	Directory to save MACS output files.	
Bowtie index directory	The directory with the Bowtie index for the reference sequence.	
Bowtie index basename	The basename of the Bowtie index for the reference sequence.	
Mate inner distance	The expected (mean) inner distance between mate pairs.	200

<b>Mate standard deviation</b>	The standard deviation for the distribution on inner distances between mate pairs.	20
<b>Library type</b>	Specifies RNA-Seq protocol.	Standard Illumins
<b>No novel junctions</b>	Only look for reads across junctions indicated in the supplied GFF or junctions file. This parameter is ignored if Raw junctions or Known transcript file is not set.	False
<b>Raw junctions</b>	The list of raw junctions.	
<b>Known transcript file</b>	A set of gene model annotations and/or known transcripts.	
<b>Max multihits</b>	Instructs TopHat to allow up to this many alignments to the reference for a given read, and suppresses all alignments for reads with more than this many alignments.	20
<b>Segment length</b>	Each read is cut up into segments, each at least this long. These segments are mapped independently.	25
<b>Fusion search</b>	Turn on fusion mapping.	False
<b>Transcriptome only</b>	Only align the reads to the transcriptome and report only those mappings as genomic mappings.	False
<b>Transcriptome max hits</b>	Maximum number of mappings allowed for a read, when aligned to the transcriptome (any reads found with more than this number of mappings will be discarded).	60
<b>Prefilter multihits</b>	When mapping reads on the transcriptome, some repetitive or low complexity reads that would be discarded in the context of the genome may appear to align to the transcript sequences and thus may end up reported as mapped to those genes only. This option directs TopHat to first align the reads to the whole genome in order to determine and exclude such multi-mapped reads (according to the value of the Max multihits option).	False
<b>Min anchor length</b>	The anchor length. TopHat will report junctions spanned by reads with at least this many bases on each side of the junction. Note that individual spliced alignments may span a junction with fewer than this many bases on one side. However, every junction involved in spliced alignments is supported by at least one read with this many bases on each side.	8
<b>Splice mismatches</b>	The maximum number of mismatches that may appear in the anchor region of a spliced alignment.	0
<b>Read mismatches</b>	Final read alignments having more than these many mismatches are discarded.	2
<b>Segment mismatches</b>	Read segments are mapped independently, allowing up to this many mismatches in each segment alignment.	2

<b>Solexa 1.3 quals</b>	As of the Illumina GA pipeline version 1.3, quality scores are encoded in Phred-scaled base-64. Use this option for FASTQ files from pipeline 1.3 or later.	False
<b>Bowtie version</b>	Specifies which Bowtie version should be used.	Bowtie2
<b>Bowtie -n mode</b>	TopHat uses -v in Bowtie for initial read mapping (the default), but with this option, -n is used instead. Read segments are always mapped using -v option.	Use -v mode
<b>Bowtie tool path</b>	The path to the Bowtie external tool.	default
<b>SAMtools tool path</b>	The path to the SAMtools tool. Note that the tool is available in the UGENE External Tool Package.	default
<b>TopHat tool path</b>	The path to the TopHat external tool in UGENE.	default
<b>Temporary directory</b>	The directory for temporary files.	default

### Parameters in Workflow File

Type: tophat

Parameter	Parameter in the GUI	Type
out-dir	Output directory	string
bowtie-index-dir	Bowtie index directory	string
bowtie-index-basename	Bowtie index basename	string
mate-inner-distance	Mate inner distance	numeric
mate-standard-deviation	Mate standard deviation	numeric
library-type	Library type	numeric
no-novel-junctions	No novel junctions	boolean
raw-junctions	Raw junctions	string
known-transcript	Known transcript file	string
max-multihits	Max multihits	numeric
segment-length	Segment length	numeric
fusion-search	Fusion search	boolean
transcriptome-only	Transcriptome only	boolean
transcriptome-max-hits	Transcriptome max hits	numeric
prefilter-multihits	Prefilter multihits	boolean
min-anchor-length	Min anchor length	numeric
splice-mismatches	Splice mismatches	numeric
read-mismatches	Read mismatches	numeric
segment-mismatches	Segment mismatches	numeric
solexa-1-3-quals	Solexa 1.3 quals	boolean
bowtie-version	Bowtie version	numeric

<b>bowtie-n-mode</b>	<b>Bowtie -n mode</b>	<i>numeric</i>
<b>bowtie-tool-path</b>	<b>Bowtie tool path</b>	<i>string</i>
<b>samtools-tool-path</b>	<b>SAMtools tool path</b>	<i>string</i>
<b>path</b>	<b>TopHat tool path</b>	<i>string</i>
<b>temp-dir</b>	<b>Temporary directory</b>	<i>string</i>

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input reads

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Dataset name</b>	<b>dataset</b>	<i>string</i>
<b>Input reads</b>	<b>first.in</b>	<i>assembly</i>
<b>Input reads url</b>	<b>in-url</b>	<i>string</i>
<b>Input paired reads url</b>	<b>paired-url</b>	<i>string</i>
<b>Input paired reads</b>	<b>second.in</b>	<i>assembly</i>

And 1 *output port*:

**Name in GUI:** TopHat output

**Name in Workflow File:** out-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Accepted hits</b>	<b>accepted.hits</b>	<i>assembly</i>
<b>Accepted hits url</b>	<b>hits-url</b>	<i>string</i>

## Merge Assemblies with Cuffmerge Element

Cuffmerge merges together several assemblies. It also handles running Cuffcompare for you, and automatically filters a number of transfrags that are probably artifacts. If you have a reference file available, you can provide it to Cuffmerge in order to gracefully merge input (e.g. novel) isoforms and known isoforms and maximize overall assembly quality.

### Parameters in GUI

Parameter	Description	Default value
<b>Output directory</b>	Directory to save MACS output files.	
<b>Reference annotation</b>	Merge the input assemblies together with this reference annotation.	
<b>Reference sequence</b>	The genomic DNA sequences for the reference. It is used to assist in classifying transfrags and excluding artifacts (e.g. repeats). For example, transcripts consisting mostly of lower-case bases are classified as repeats.	
<b>Minimum isoform fraction</b>	Discard isoforms with abundance below this.	0.05

<b>Cuffcompare tool path</b>	The path to the Cuffcompare external tool in UGENE.	default
<b>Cuffmerge tool path</b>	The path to the Cuffmerge external tool in UGENE.	default
<b>Temporary directory</b>	The directory for temporary files.	default

### Parameters in Workflow File

**Type:** cuffmerge

Parameter	Parameter in the GUI	Type
out-dir	Output directory	string
ref-annotation	Reference annotation	string
ref-seq	Reference sequence	string
min-isoform-fraction	Minimum isoform fraction	numeric
cuffcompare-tool-path	Cuffcompare tool path	string
path	Cuffmerge tool path	string
tmp-dir	Temporary directory	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Set of annotations

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	in-annotations	ann_table

And 1 *output port*:

**Name in GUI:** Set of annotations

**Name in Workflow File:** out-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	out-annotations	ann_table

## Test for Diff. Expression with Cuffdiff Element

Cuffdiff takes a transcript file as input, along with two or more fragment alignments (e.g. in SAM format) for two or more samples. It produces a number of output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene. Tracking the former allows one to see changes in splicing, and the latter lets one see changes in relative promoter use within a gene.

### Parameters in GUI

Parameter	Description	Default value
Output directory	Directory to save MACS output files.	

<b>Time series analysis</b>	If set to True, instructs Cuffdiff to analyze the provided samples as a time series, rather than testing for differences between all pairs of samples. Samples should be provided in increasing time order.	False
<b>Upper quartile norm</b>	If set to True, normalizes by the upper quartile of the number of fragments mapping to individual loci instead of the total number of sequenced fragments. This can improve robustness of differential expression calls for less abundant genes and transcripts.	False
<b>Hits norm</b>	Instructs how to count all fragments. Total specifies to count all fragments, including those not compatible with any reference transcript, towards the number of mapped fragments used in the FPKM denominator. Compatible specifies to use only compatible fragments. Selecting Compatible is generally recommended in Cuffdiff to reduce certain types of bias caused by differential amounts of ribosomal reads which can create the impression of falsely differentially expressed genes..	Compatible
<b>Frag bias correct</b>	Providing the sequences your reads were mapped to instructs Cuffdiff to run bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates..	
<b>Multi read correct</b>	Do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.	False
<b>Library type</b>	Specifies RNA-Seq protocol.	Standard Illumina
<b>Mask file</b>	Ignore all reads that could have come from transcripts in this file. It is recommended to include any annotated rRNA, mitochondrial transcripts other abundant transcripts you wish to ignore in your analysis in this file. Due to variable efficiency of mRNA enrichment methods and rRNA depletion kits, masking these transcripts often improves the overall robustness of transcript abundance estimates..	
<b>Min alignment count</b>	The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples. If no testing is performed, changes in the locus are deemed not significant, and the locus' observed changes don't contribute to correction for multiple testing..	10
<b>FDR</b>	The allowed false discovery rate used in testing.	0.05
<b>Max MLE iterations</b>	Sets the number of iterations allowed during maximum likelihood estimation of abundances.	5000



<b>Emit count tables</b>	Include information about the fragment counts, fragment count variances, and fitted variance model into the report.	False
<b>Cuffdiff tool path</b>	The path to the Cuffdiff external tool in UGENE.	default
<b>Temporary directory</b>	The directory for temporary files.	default

### Parameters in Workflow File

**Type:** cuffdiff

Parameter	Parameter in the GUI	Type
out-dir	Output directory	string
time-series-analysis	Time series analysis	boolean
upper-quartile-norm	Upper quartile norm	boolean
hits-norm	Hits norm	numeric
frag-bias-correct	Frag bias correct	string
multi-read-correct	Multi read correct	boolean
library-type	Library type	numeric
mask-file	Mask file	numeric
min-alignment-count	Min alignment count	string
fdr	FDR	numeric
max-mle-iterations	Max MLE iterations	numeric
emit-count-tables	Emit count tables	boolean
path	Cuffdiff tool path	string
temp-dir	Temporary directory	string

### Input/Output Ports

The element has 2 *input port*:

**Name in GUI:** Annotations

**Name in Workflow File:** in-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	in-annotations	ann_table

**Name in GUI:** Assembly

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Assembly data	assembly	assembly
Source url	url	string

## NGS: Variant Calling

- [Call Variants with SAMtools Element](#)
- [Create VCF consensus](#)

## Call Variants with SAMtools Element

Calls SNPs and INDELS with SAMtools mpileup and bcftools.

### Parameters in GUI

Parameter	Description	Default value
<b>Illumina-1.3+ encoding</b>	Assume the quality is in the Illumina 1.3+ encoding (mpileup)(-6).	False
<b>Count anomalous read pairs</b>	Do not skip anomalous read pairs in variant calling (mpileup)(-A).	False
<b>Disable BAQ computation</b>	Disable probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being misaligned. Applying this option greatly helps to reduce false SNPs caused by misalignments (mpileup)(-B).	False
<b>Mapping quality downgrading coefficient</b>	Coefficient for downgrading mapping quality for reads containing excessive mismatches. Given a read with a phred-scaled probability q of being generated from the mapped position, the new mapping quality is about $\sqrt{(INT-q)/INT} \cdot INT$ . A zero value disables this functionality; if enabled, the recommended value for BWA is 50 (mpileup)(-C).	0
<b>Max number of reads per input BAM</b>	At a position, read maximally the number of reads per input BAM (mpileup)(-d).	250
<b>Extended BAQ computation</b>	Extended BAQ computation. This option helps sensitivity especially for MNPs, but may hurt specificity a little bit (mpileup)(-E).	False
<b>BED or position list file</b>	BED or position list file containing a list of regions or sites where pileup or BCF should be generated. (mpileup)(-l).	
<b>Pileup region</b>	Only generate pileup in region STR (mpileup)(-r).	
<b>Minimum mapping quality</b>	Minimum mapping quality for an alignment to be used (mpileup)(-q).	0
<b>Minimum base quality</b>	Minimum base quality for a base to be considered (mpileup)(-Q).	13
<b>Gap extension error</b>	Phred-scaled gap extension sequencing error probability. Reducing INT leads to longer indels (mpileup)(-e).	20
<b>Homopolymer errors coefficient</b>	Coefficient for modeling homopolymer errors. Given an l-long homopolymer run, the sequencing error of an indel of size s is modeled as $INT \cdot s/l$ . (mpileup)(-h).	100
<b>No INDELS</b>	Do not perform INDEL calling (mpileup)(-l).	False
<b>Max INDEL depth</b>	Skip INDEL calling if the average per-sample depth is above INT (mpileup)(-L).	250

<b>Gap open error</b>	Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls (mpileup)(-o).	40
<b>List of platforms for indels</b>	Comma delimited list of platforms (determined by @RG-PL) from which indel candidates are obtained. It is recommended to collect indel candidates from sequencing technologies that have low indel error rate such as ILLUMINA. (mpileup)(-P).	
<b>Retain all possible alternate</b>	Retain all possible alternate alleles at variant sites. By default, the view command discards unlikely alleles. (bcf view)(-A).	False
<b>Indicate PL</b>	Indicate PL is generated by r921 or before (ordering is different) (bcf view)(-F).	False
<b>No genotype information</b>	Suppress all individual genotype information (bcf view)(-G).	False
<b>A/C/G/T only</b>	Skip sites where the REF field is not A/C/G/T (bcf view)(-N).	False
<b>List of sites</b>	List of sites at which information are outputted (bcf view)(-l).	
<b>QCALL likelihood</b>	Output the QCALL likelihood format (bcf view)(-Q).	False
<b>List of samples</b>	List of samples to use. The first column in the input gives the sample names and the second gives the ploidy, which can only be 1 or 2. When the 2nd column is absent, the sample ploidy is assumed to be 2. In the output, the ordering of samples will be identical to the one in FILE (bcf view)(-s).	
<b>Min samples fraction</b>	skip loci where the fraction of samples covered by reads is below FLOAT (bcf view)(-d).	0
<b>Per-sample genotypes</b>	Call per-sample genotypes at variant sites. (bcf view)(-g).	True
<b>INDEL-to-SNP Ratio</b>	Ratio of INDEL-to-SNP mutation rate. (bcf view)(-i).	-1
<b>Max P(ref D)</b>	A site is considered to be a variant if P(ref D)	0.5
<b>Prior allele frequency spectrum</b>	If STR can be full, cond2, flat or the file consisting of error output from a previous variant calling run (bcf view)(-P).	full
<b>Mutation rate</b>	Scaled mutation rate for variant calling (bcf view)(-t).	0.001

<b>Pair/trio calling</b>	Enable pair/trio calling. For trio calling, option -s is usually needed to be applied to configure the trio members and their ordering. In the file supplied to the option -s, the first sample must be the child, the second the father and the third the mother. The valid values of STR are pair, trioauto, trioxd and trioxs, where pair calls differences between two input samples, and trioxd (trioxs) specifies that the input is from the X chromosome non-PAR regions and the child is a female (male) (bcf view)(-T).	
<b>N group-1 samples</b>	Number of group-1 samples. This option is used for dividing the samples into two groups for contrast SNP calling or association test. When this option is in use, the following VCF INFO will be outputted: PC2, PCHI2 and QCHI2 (bcf view)(-1).	0
<b>N permutations</b>	Number of permutations for association test (effective only with -1) (bcf view)(-U).	0
<b>Min P(chi^2)</b>	Only perform permutations for P(chi^2).	0.01
<b>Minimum RMS quality</b>	Minimum RMS mapping quality for SNPs (varFilter) (-Q).	10
<b>Minimum read depth</b>	Minimum read depth (varFilter) (-d).	2
<b>Maximum read depth</b>	Maximum read depth (varFilter) (-D).	10000000
<b>Alternate bases</b>	Minimum number of alternate bases (varFilter) (-a).	2
<b>Gap size</b>	SNP within INT bp around a gap to be filtered (varFilter) (-w).	3
<b>Window size</b>	Window size for filtering adjacent gaps (varFilter) (-W).	10
<b>Strand bias</b>	Minimum P-value for strand bias (given PV4) (varFilter) (-1).	0.0001
<b>BaseQ bias</b>	Minimum P-value for baseQ bias (varFilter) (-2).	1e-100
<b>MapQ bias</b>	Minimum P-value for mapQ bias (varFilter) (-3).	0
<b>End distance bias</b>	Minimum P-value for end distance bias (varFilter) (-4).	0.0001
<b>HWE</b>	Minimum P-value for HWE (plus F).	0.0001
<b>Log filtered</b>	Print filtered variants into the log (varFilter) (-p).	False

## Parameters in Workflow File

**Type:** call\_variants

Parameter	Parameter in the GUI	Type
illumina13-encoding	Illumina-1.3+ encoding	<i>boolean</i>
use_orphan	Count anomalous read pairs	<i>boolean</i>

disable_baq	Disable BAQ computation	<i>boolean</i>
capq_thres	Mapping quality downgrading coefficient	<i>numeric</i>
max_depth	Max number of reads per input BAM	<i>numeric</i>
ext_baq	Extended BAQ computation	<i>boolean</i>
bed	BED or position list file	<i>string</i>
reg	Pileup region	<i>string</i>
min_mq	Minimum mapping quality	<i>numeric</i>
min_baseq	Minimum base quality	<i>numeric</i>
extQ	Gap extension error	<i>numeric</i>
tandemQ	Homopolymer errors coefficient	<i>numeric</i>
no_indel	No INDELs	<i>boolean</i>
max_indel_depth	Max INDEL depth	<i>numeric</i>
openQ	Gap open error	<i>numeric</i>
pl_list	List of platforms for indels	<i>string</i>
keepalt	Retain all possible alternate	<i>boolean</i>
fix_pl	Indicate PL	<i>boolean</i>
no_geno	No genotype information	<i>boolean</i>
acgt_only	A/C/G/T only	<i>boolean</i>
bcf_bed	List of sites	<i>string</i>
qcall	QCALL likelihood	<i>boolean</i>
samples	List of samples	<i>string</i>
min_smpl_frac	Min samples fraction	<i>numeric</i>
call_gt	Per-sample genotypes	<i>boolean</i>
indel_frac	INDEL-to-SNP Ratio	<i>numeric</i>
pref	Max P(ref D)	<i>numeric</i>
potype	Prior allele frequency spectrum	<i>string</i>
theta	Mutation rate	<i>numeric</i>
ccall	Pair/trio calling	<i>string</i>
n1	N group-1 samples	<i>numeric</i>
n_perm	N permutations	<i>numeric</i>
min_perm_p	Min P(chi^2)	<i>numeric</i>
min-qual	Minimum RMS quality	<i>numeric</i>
min-dep	Minimum read depth	<i>numeric</i>
max-dep	Maximum read depth	<i>numeric</i>
min-alt-bases	Alternate bases	<i>numeric</i>
gap-size	Gap size	<i>numeric</i>
window"	Window size	<i>numeric</i>

min-strand	Strand bias	<i>numeric</i>
min-baseQ	BaseQ bias	<i>string</i>
min-mapQ	MapQ bias	<i>numeric</i>
min-end-distance	End distance bias	<i>numeric</i>
min-hwe	HWE	<i>numeric</i>
print-filtered	Log filtered	<i>boolean</i>

### Input/Output Ports

The element has 2 *input ports*:

**Name in GUI:** Input assembly

**Name in Workflow File:** in-assembly

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Dataset name	dataset	<i>string</i>
Source url	url	<i>string</i>

**Name in GUI:** Input sequences

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Source url	url	<i>string</i>

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	<i>variation</i>

## Create VCF consensus

Apply VCF variants to a fasta file to create consensus sequence.

### Parameters in GUI

Parameter	Description	Default value
Output FASTA consensus	The url to the output file with the result consensus.	

### Parameters in Workflow File

**Type:** vcf-consensus

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

consensus-url	Output FASTA consensus	string
---------------	------------------------	--------

## Input/Output Ports

The element has 1 *input ports*:

**Name in GUI:** Input FASTA and VCF

**Name in Workflow File:** in-data

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Fasta url	fasta	string
VCF url	vcf	string

And 1 *output port*:

**Name in GUI:** Fasta consensus url

**Name in Workflow File:** out-consensus

**Slots:**

Slot In GUI	Slot in Workflow File	Type
out-consensus	out-consensus	string

## SNP Annotation

- [Annotate variations with SNPToolbox Element](#)
- [Detect Transcription Factors with rSNP-Tools Element](#)
- [Determine SNP effect on TATA-boxes Element](#)
- [ProtStability1D Element](#)
- [ProtStability3D Element](#)
- [SNP Chip Tools Element](#)
- [SNP Effect on PDB sites Element](#)
- [Write SNP Report Element](#)

### Annotate variations with SNPToolbox Element

Assess damage effect and find intersected genes with SNPToolbox algorithms.

**Parameters in GUI**

Parameter	Description	Default value
Database path	Path to SNPToolbox database with sequences, features and damage effect data.	

**Parameters in Workflow File**

**Type:** SNPToolbox-id

Parameter	Parameter in the GUI	Type
db_path	Database path	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	<i>variation</i>

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

## Detect Transcription Factors with rSNP-Tools Element

Identification of transcription factor binding sites in the DNA which have been modified by polymorphic mutation.

### Parameters in GUI

Parameter	Description	Default value
Database path	Path to SNPToolbox database with sequences, features and damage effect data.	
First site state	First sequence TFBS state.	Normal (1.0)
Second site state	Second sequence TFBS state.	Weakened (0.5)
SNP significance	Significance value for the SNP.	0.00025

### Parameters in Workflow File

**Type:** rSnp-tools

Parameter	Parameter in the GUI	Type
db_path	Database path	<i>string</i>
first_site_state	First site state	<i>numeric</i>
second_site_state	Second site state	<i>numeric</i>
snp_significance	SNP significance	<i>numeric</i>

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	<i>variation</i>

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

## Determine SNP effect on TATA-boxes Element

Define the influence of SNP on TATA-boxes belonging to the sequence.



## Parameters in GUI

Parameter	Description	Default value
Database path	Path to SNP database.	

## Parameters in Workflow File

Type: tata-box-snp

Parameter	Parameter in the GUI	Type
db_path	Database path	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

**ProtStability1D Element**

Identification of the SNP influence on protein primary structure thermodynamic stability.

## Parameters in GUI

Parameter	Description	Default value
Database path	Path to SNP database.	

## Parameters in Workflow File

Type: prot-stability-1d

Parameter	Parameter in the GUI	Type
db_path	Database path	string

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

**ProtStability3D Element**

Identification of the SNP influence on protein tertiary structure thermodynamic stability.

#### Parameters in GUI

Parameter	Description	Default value
Database path	Path to SNP database.	

#### Parameters in Workflow File

Type: prot-stability-3d

Parameter	Parameter in the GUI	Type
db_path	Database path	string

#### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

### SNP Chip Tools Element

Assess the SNP impact on regulatory regions.

Type: prot-stability-3d

#### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

### SNP Effect on PDB sites Element

Identification of the SNP influence on PDB sites.

#### Parameters in GUI

Parameter	Description	Default value
Database path	Path to SNP database.	

#### Parameters in Workflow File

**Type:** snp2pdb-site

Parameter	Parameter in the GUI	Type
db_path	Database path	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

And 1 *output port*:

**Name in GUI:** Output variations

**Name in Workflow File:** out-variations

## Write SNP Report Element

Use variations and their effects to write a report.

### Parameters in GUI

Parameter	Description	Default value
In gene report path	Path to save in-gene SNP effects reports.	
Regulatory report path	Path to save regulatory SNP effects reports.	
Database path	Path to SNP database.	

### Parameters in Workflow File

**Type:** snp-report-writer-id

Parameter	Parameter in the GUI	Type
report_path	In gene report path	
regulatory_report_path	Regulatory report path	
db_path	Database path	string

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input variations

**Name in Workflow File:** in-variations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Variation track	variation-track	variation

## Transcription Factor

- [Build Frequency Matrix Element](#)
- [Build SITECON Model Element](#)
- [Build Weight Matrix Element](#)

- Convert Frequency Matrix Element
- Read Frequency Matrix Element
- Read SITECON Model Element
- Read Weight Matrix Element
- Search for TFBS with SITECON Element
- Search for TFBS with Weight Matrix Element
- Write Frequency Matrix Element
- Write SITECON Model Element
- Write Weight Matrix Element

## Build Frequency Matrix Element

Builds a frequency matrix. Frequency matrices are used for probabilistic recognition of transcription factor binding sites.

### Parameters in GUI

Parameter	Description	Default value
<b>Matrix type</b>	Dinucleic matrices are more detailed, while mononucleic one are more useful for small input data sets.	Mononucleic

## Parameters in Workflowa File

**Type:** fmatrix-build

Parameter	Parameter in the GUI	Type
<b>type</b>	<b>Matrix type</b>	<i>boolean</i>  Available values are: <ul style="list-style-type: none"> <li>• true - for Dinucleic</li> <li>• false - for Mononucleic</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input alignment*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>msa</i>

And 1 *output port*:

**Name in GUI:** *Frequency matrix*

**Name in Workflow File:** out-fmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Frequency matrix</b>	<b>fmatrix</b>	<i>fmatrix</i>

## Build SITECON Model Element

Builds statistical profile for SITECON. The SITECON is a program for probabilistic recognition of transcription factor binding sites.

### Parameters in GUI

Parameter	Description	Default value
-----------	-------------	---------------

<b>Weight algorithm</b>	Optional feature, in most cases applying no weight will fit. In some cases choosing algorithm 2 will increase the recognition quality.	None
<b>Window size, bp</b>	Window is used to pick out the most important alignment region and is located at the center of the alignment. Must be: windows size is not greater than TFBS alignment length, recommended: windows size is not greater than 50 bp.	40
<b>Calibration length</b>	Length of random synthetic sequences used to calibrate the profile. Should not be less than window size.	1M
<b>Random seed</b>	The random seed, where is a positive integer. You can use this option to generate reproducible results for different runs on the same data.	0

## Parameters in Workflow File

Type: sitecon-build

Parameter	Parameter in the GUI	Type
weight-algorithm	Weight algorithm	boolean  Available values are: <ul style="list-style-type: none"> <li>0 - for None</li> <li>1 - for Algorithm2</li> </ul>
window-size	Window size, bp	numeric
calibrate-length	Calibration length	numeric
seed	Random seed	numeric

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input alignment*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
MSA	msa	msa
Origin	url	string

And 1 *output port*:

**Name in GUI:** *Sitecon model*

**Name in Workflow File:** out-sitecon

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sitecon model	sitecon-model	sitecon-model

## Build Weight Matrix Element

Builds weight matrix. Weight matrices are used for probabilistic recognition of transcription factor binding sites.

#### Parameters in GUI

Parameter	Description	Default value
<b>Matrix type</b> (required)	Dinucleic matrices are more detailed, while mononucleic one are more useful for small input data sets.	Mononucleic
<b>Weight algorithm</b>	Different weight algorithms uses different functions to build weight matrices. It allows us to get better precision on different data sets. Log-odds, NLG and Match algorithms are sensitive to input matrices with zero values, so some of them may not work on those matrices.	Berg and Von Hippel

## Parameters in Workflow File

Type: wmatrix-build

Parameter	Parameter in the GUI	Type
<b>type</b>	<b>Matrix type</b>	<i>boolean</i> Available values are: <ul style="list-style-type: none"> <li>• true - for Dinucleic</li> <li>• false - for Mononucleic</li> </ul>
<b>weight-algorithm</b>	<b>Weight algorithm</b>	<i>string</i> Available values are: <ul style="list-style-type: none"> <li>• Berg and von Hippel</li> <li>• Log-odds</li> <li>• Match</li> <li>• NLG</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input alignment*

**Name in Workflow File:** in-msa

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>MSA</b>	<b>msa</b>	<i>msa</i>

And 1 *output port*:

**Name in GUI:** *Weight matrix*

**Name in Workflow File:** out-wmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Weight matrix</b>	<b>wmatrix</b>	<i>wmatrix</i>

## Convert Frequency Matrix Element

Converts a frequency matrix to a weight matrix. Weight matrices are used for probabilistic recognition of transcription factor binding sites.

#### Parameters in GUI

Parameter	Description	Default value
<b>Matrix type</b> (required)	Dinucleic matrices are more detailed, while mononucleic one are more useful for small input data sets.	Mononucleic
<b>Weight algorithm</b>	Different weight algorithms uses different functions to build weight matrices. It allows us to get better precision on different data sets. Log-odds, NLG and Match algorithms are sensitive to input matrices with zero values, so some of them may not work on those matrices.	Berg and Von Hippel

## Parameters in Workflow File

**Type:** fmatrix-to-wmatrix

Parameter	Parameter in the GUI	Type
<b>type</b>	<b>Matrix type</b>	<i>boolean</i>  Available values are: <ul style="list-style-type: none"> <li>• true - for Dinucleic</li> <li>• false - for Mononucleic</li> </ul>
<b>weight-algorithm</b>	<b>Weight algorithm</b>	<i>string</i>  Available values are: <ul style="list-style-type: none"> <li>• Berg and von Hippel</li> <li>• Log-odds</li> <li>• Match</li> <li>• NLG</li> </ul>

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Frequency matrix*

**Name in Workflow File:** in-fmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Frequency matrix	fmatrix	fmatrix

And 1 *output port*:

**Name in GUI:** *Weight matrix*

**Name in Workflow File:** out-wmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Weight matrix	wmatrix	wmatrix

## Read Frequency Matrix Element

Reads frequency matrices from file(s). The files can be local or Internet URLs.

**Parameters in GUI**

Parameter	Description	Default value
<b>Input files</b> (required)	Semicolon-separated list of paths to the input files.	

## Parameters in Workflow File

**Type:** fmatrix-read

Parameter	Parameter in the GUI	Type
url-in	Input files	string

## Input/Output Ports

The element has 1 *output port*.

**Name in GUI:** *Frequency matrix*

**Name in Workflow File:** out-fmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Frequency matrix	fmatrix	fmatrix

## Read SITECON Model Element

Reads SITECON profiles from file(s). The files can be local or Internet URLs.

**Parameters in GUI**

Parameter	Description	Default value
<b>Input files</b> (required)	Semicolon-separated list of paths to the input files.	

## Parameters in Workflow File

**Type:** sitecon-read

Parameter	Parameter in the GUI	Type
url-in	Input files	string

## Input/Output Ports

The element has 1 *output port*.

**Name in GUI:** *Sitecon model*

**Name in Workflow File:** out-sitecon

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sitecon model	sitecon-model	sitecon-model

## Read Weight Matrix Element

Reads weight matrices from file(s). The files can be local or Internet URLs.

**Parameters in GUI**

Parameter	Description	Default value
-----------	-------------	---------------



<b>Input files</b> (required)	Semicolon-separated list of paths to the input files.	
-------------------------------	---	--

## Parameters in Workflow File

**Type:** wmatrix-read

Parameter	Parameter in the GUI	Type
url-in	Input files	string

## Input/Output Ports

And 1 *output port*:

**Name in GUI:** *Weight matrix*

**Name in Workflow File:** out-wmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Weight matrix	wmatrix	wmatrix

## Search for TFBS with SITECON Element

Searches each input sequence for transcription factor binding sites significantly similar to specified SITECON profiles. In case several profiles were supplied, searches with all profiles one by one and outputs merged set of annotations for each sequence.

**Parameters in GUI**

Parameter	Description	Default value
<b>Result annotation</b>	Name of the result annotations.	misc_feature
<b>Search in</b>	Specifies which strands should be searched: direct, complement or both.	both strands
<b>Min score</b>	Recognition quality threshold, should be less than 100%. Choosing too low threshold will lead to recognition of too many TFBS recognised with too low trustworthiness. Choosing too high threshold may result in no TFBS recognised.	85
<b>Min err1</b>	Alternative setting for filtering results, minimal value of Error type I. Note that all thresholds (by score, by err1 and by err2) are applied when filtering results.	0.0
<b>Max err2</b>	Alternative setting for filtering results, max value of Error type II. Note that all thresholds (by score, by err1 and by err2) are applied when filtering results.	0.001

## Parameters in Workflow File

**Type:** sitecon-search

Parameter	Parameter in the GUI	Type
result-name	Result annotation	string

<b>strand</b>	<b>Search in</b>	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for searching in both strands</li> <li>• 1 - for searching in direct strand</li> <li>• 2 - for searching in complement strand</li> </ul>
<b>min-score</b>	<b>Min score</b>	<i>numeric</i>
<b>err1</b>	<b>Min err1</b>	<i>numeric</i>
<b>err2</b>	<b>Max err2</b>	<i>numeric</i>

## Input/Output Ports

The element has 2 *input ports*. The first port:

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The second input port gets the SITECON model:

**Name in GUI:** *Sitecon model*

**Name in Workflow File:** in-sitecon

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sitecon model	sitecon-model	sitecon-model

And there is 1 *output port*:

**Name in GUI:** *Sitecon annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Search for TFBS with Weight Matrix Element

Searches each input sequence for transcription factor binding sites significantly similar to specified weight matrices. In case several profiles were supplied, searches with all profiles one by one and outputs merged set of annotations for each sequence.

**Parameters in GUI**

Parameter	Description	Default value
<b>Result annotation</b>	Name of the result annotations.	misc_feature
<b>Search in</b>	Specifies which strands should be searched: direct, complement or both.	both strands
<b>Min score</b>	Minimum score to detect transcription factor binding site in percents.	85

## Parameters in Workflow File

**Type:** wmatrix-search

Parameter	Parameter in the GUI	Type
result-name	Result annotation	string
strand	Search in	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for searching in both strands</li> <li>• 1 - for searching in direct strand</li> <li>• 2 - for searching in complement strand</li> </ul>
min-score	Min score	numeric

## Input/Output Ports

The element has 2 *input ports*. The first port:

**Name in GUI:** *Sequence*

**Name in Workflow File:** in-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The second input port gets the SITECON model:

**Name in GUI:** *Weight matrix*

**Name in Workflow File:** in-wmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Weight matrix	wmatrix	wmatrix

And there is 1 *output port*:

**Name in GUI:** *Weight matrix annotations*

**Name in Workflow File:** out-annotations

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table

## Write Frequency Matrix Element

Saves all input frequency matrices to specified location.

**Parameters in GUI**

Parameter	Description	Default value
Output file (required)	Location of the output data file. If this attribute is set, the "Location" slot is not taken into account.	
Existing file	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename

## Parameters in Workflow File

**Type:** fmatrix-write

Parameter	Parameter in the GUI	Type
url-out	Output file	string
write-mode	Existing file	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Frequency matrix*

**Name in Workflow File:** in-fmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Frequency matrix	fmatrix	fmatrix
Source URL	url	string

## Write SITECON Model Element

Saves all input SITECON profiles to specified location.

**Parameters in GUI**

Parameter	Description	Default value
Output file (required)	Location of the output data file. If this attribute is set, the "Location" slot is not taken into account.	
Existing file	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename

## Parameters in Workflow File

**Type:** sitecon-write

Parameter	Parameter in the GUI	Type
url-out	Output file	string
write-mode	Existing file	numeric  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Sitecon model*

**Name in Workflow File:** in-sitecon

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sitecon model	sitecon-model	<i>sitecon-model</i>
Source URL	url	<i>string</i>

## Write Weight Matrix Element

Saves all input weight matrices to specified location.

**Parameters in GUI**

Parameter	Description	Default value
<b>Output file</b> (required)	Location of the output data file. If this attribute is set, the "Location" slot is not taken into account.	
<b>Existing file</b>	If a target file already exists, you can specify how it should be handled: either overwritten, renamed or appended (if supported by file format).	Rename

## Parameters in Workflow File

**Type:** wmatrix-write

Parameter	Parameter in the GUI	Type
url-out	Output file	<i>string</i>
write-mode	Existing file	<i>numeric</i>  Available values are: <ul style="list-style-type: none"> <li>• 0 - for overwrite</li> <li>• 1 - for append</li> <li>• 2 - for rename</li> </ul>

## Input/Output Ports

The element has 1 *input port*.

**Name in GUI:** *Weight matrix*

**Name in Workflow File:** in-wmatrix

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Weight matrix	wmatrix	<i>wmatrix</i>
Source URL	url	<i>string</i>

## Utils

- [DNA Statistics Element](#)
- [Generate DNA Element](#)

## DNA Statistics Element

Evaluates statistic for DNA sequences.

**Parameters in GUI**

Parameter	Description	Default value
GC-content	Evaluate GC-content.	True
GC1-content	Evaluate GC1-content.	True
GC2-content	Evaluate GC2-content.	True
GC3-content	Evaluate GC3-content.	True

## Parameters in Workflow File

Type: dna-stats

Parameter	Parameter in the GUI	Type
gc-content	GC-content	boolean
gc1-content	GC1-content	boolean
gc2-content	GC2-content	boolean
gc3-content	GC3-content	boolean

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input sequence*

**Name in Workflow File:** in-sequence

Slots:

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

The element has 1 *output port*:

**Name in GUI:** *Result annotation*

**Name in Workflow File:** out-annotations

Slots:

Slot In GUI	Slot in Workflow File	Type
Set of annotations	annotations	annotation-table-list

## Generate DNA Element

Generates random DNA sequences with given nucleotide content that can be specified manually or evaluated from the reference file.

**Parameters in GUI**

Parameter	Description	Default value
Length	Length of the resulted sequence or sequences.	1000 bp
Count	Number of sequences to generate.	1
Seed	Value to initialize the random generator. By default (seed = -1) the generator is initialized with the system time.	-1
Content	Specifies how the nucleotide content of the sequence(s) should be generated. It can be either taken from the reference file (see the <i>Referenceparameter</i> ), or input manually.	manual

<b>Algorithm</b>	Algorithm for generating random sequence(s). Two algorithms are available: GC Content and GC Skew. If you choose GC Content, then parameters <i>A</i> , <i>C</i> , <i>G</i> , <i>T</i> are used to generate the sequence. Otherwise, the <i>GC Skew</i> parameter is used to generate the sequence(s).	GC Content
<b>Window size</b>	The DNA sequence generation is divided into windows of the specified size. In each window the bases ratio, defined by other parameters, is kept.	1000
<b>Reference</b>	Path to the reference file (could be a sequence or an alignment).	
<b>A</b>	Adenine content.	25%
<b>C</b>	Cytosine content.	25%
<b>G</b>	Guanine content.	25%
<b>T</b>	Thymine content.	25%
<b>GC Skew</b>	GC Skew is calculated as $(G - C) / (G + C)$ , where <i>G</i> is the number of G's in the window, and <i>C</i> is the number of C's.	0.25

## Parameters in Workflow File

Type: generate-dna

Parameter	Parameter in the GUI	Type
<b>length</b>	<b>Lenght</b>	<i>numeric</i>
<b>count</b>	<b>Count</b>	<i>numeric</i>
<b>seed</b>	<b>Seed</b>	<i>numeric</i>
<b>content</b>	<b>Countent</b>	<i>string</i>
<b>algorithm</b>	<b>Algorithm</b>	<i>string</i> Available values are: <ul style="list-style-type: none"> <li>gc-content</li> <li>gc-skew</li> </ul>
<b>window-size</b>	<b>Window size</b>	<i>numeric</i>
<b>reference-url</b>	<b>Reference</b>	<i>string</i> Available values are: <ul style="list-style-type: none"> <li>manual</li> <li>reference</li> </ul>
<b>percent-a</b>	<b>A</b>	<i>numeric</i>
<b>percent-c</b>	<b>C</b>	<i>numeric</i>
<b>percent-g</b>	<b>G</b>	<i>numeric</i>
<b>percent-t</b>	<b>T</b>	<i>numeric</i>
<b>gc-skew</b>	<b>GC Skew</b>	<i>numeric</i>

## Input/Output Ports

The element has 1 *output port*.

**Name in GUI:** *Sequences*

**Name in Workflow File:** out-sequence

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## Custom Elements With Script

- CASAVA FASTQ Filter Script Element
- Dump Sequence Info Element
- FASTQ Trimmer Element
- LinkData Fetch Element
- Quality Filter Element

### CASAVA FASTQ Filter Script Element

Filters FASTQ reads generated by CASAVA 1.8.

The element works on the basis of the following script:

```
var seqName = getName(in_sequence);
if(seqName.search(".*[^:]*:N:[^:]*:") == 0){
  out_sequence = in_sequence;
}else{
  out_sequence = null;
}
```

## Parameters in GUI

The element has no parameters.

## Parameters in Workflow File

**Type:** "Script-CASAVA-FASTQ-filter"

The element has no parameters.

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** input data

**Name in Workflow File:** in

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** output data

**Name in Workflow File:** out

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

### Dump Sequence Info Element

For each incoming sequences, dumps to output the sequence name and the sequence size.



The element works on the basis of the following script:

```
out_text=getName(in_sequence) + ": " + size(in_sequence);
```

## Parameters in GUI

The element has no parameters.

### Parameters in Workflow File

**Type:** "Script-Dump sequence info"

The element has no parameters.

### Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** *Input data*

**Name in Workflow File:** in

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** *Output data*

**Name in Workflow File:** out

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Plain text	text	string

## FASTQ Trimmer Element

Trim input sequence from the end, using the quality threshold.

The element works on the basis of the following script:

```
var trimmedSequence = getTrimmedByQuality(in_sequence, min_quality,
min_sequence_length);
if(size(trimmedSequence) != 0){
out_sequence = trimmedSequence;
}else{
out_sequence = null;
}
```

### Parameters in GUI

Parameter	Description	Default value
min_quality	Number.	
min_sequence_length	Number.	

### Parameters in Workflow File

**Type:** "Script-FASTQ Trimmer"

Parameter	Parameter in the GUI	Type
-----------	----------------------	------

<b>min_quality</b>	<b>min_quality</b>	<i>numeric</i>
<b>min_sequence_length</b>	<b>min_sequence_length</b>	<i>numeric</i>

**Input/Output Ports**

The element has 1 *input port*:

**Name in GUI:** input data

**Name in Workflow File:** in

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>

And 1 *output port*:

**Name in GUI:** output data

**Name in Workflow File:** out

**Slots:**

Slot In GUI	Slot in Workflow File	Type
<b>Sequence</b>	<b>sequence</b>	<i>sequence</i>

**LinkData Fetch Element**

Fetches sequence from LinkData service.

The element works on the basis of the following script:

```
out_sequence =
sequenceFromText(LinkData.getObjects(workId,filename,subject,property));
```

**Parameters in GUI**

Parameter	Description	Default value
<b>workId</b>	String.	
<b>filename</b>	String.	
<b>subject</b>	String.	
<b>property</b>	String.	

**Parameters in Workflow File**

**Type:** "Script-LinkData Fetch"

Parameter	Parameter in the GUI	Type
<b>workId</b>	<b>workId</b>	<i>string</i>
<b>filename</b>	<b>filename</b>	<i>string</i>
<b>subject</b>	<b>subject</b>	<i>string</i>
<b>property</b>	<b>property</b>	<i>string</i>

**Input/Output Ports**

The element has 1 *output port*:

**Name in GUI:** output data

**Name in Workflow File:** out

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

## Quality Filter Element

Filters sequences by their qualities.

The element works on the basis of the following script:

```
var qual;

if(hasQuality(in_sequence)) {
    qual = getMinimumQuality(in_sequence);
    if(qual >= quality) {
        out_sequence = in_sequence;
    }
}
```

## Parameters in GUI

Parameter	Description	Default value
quality	Quality used to filter.	

## Parameters in Workflow File

**Type:** "Script-Quality filter example"

Parameter	Parameter in the GUI	Type
quality	quality	numeric

## Input/Output Ports

The element has 1 *input port*:

**Name in GUI:** Input data

**Name in Workflow File:** in

**Slots:**

Slot In GUI	Slot in Workflow File	Type
Sequence	sequence	sequence

And 1 *output port*:

**Name in GUI:** Output data

**Name in Workflow File:** out

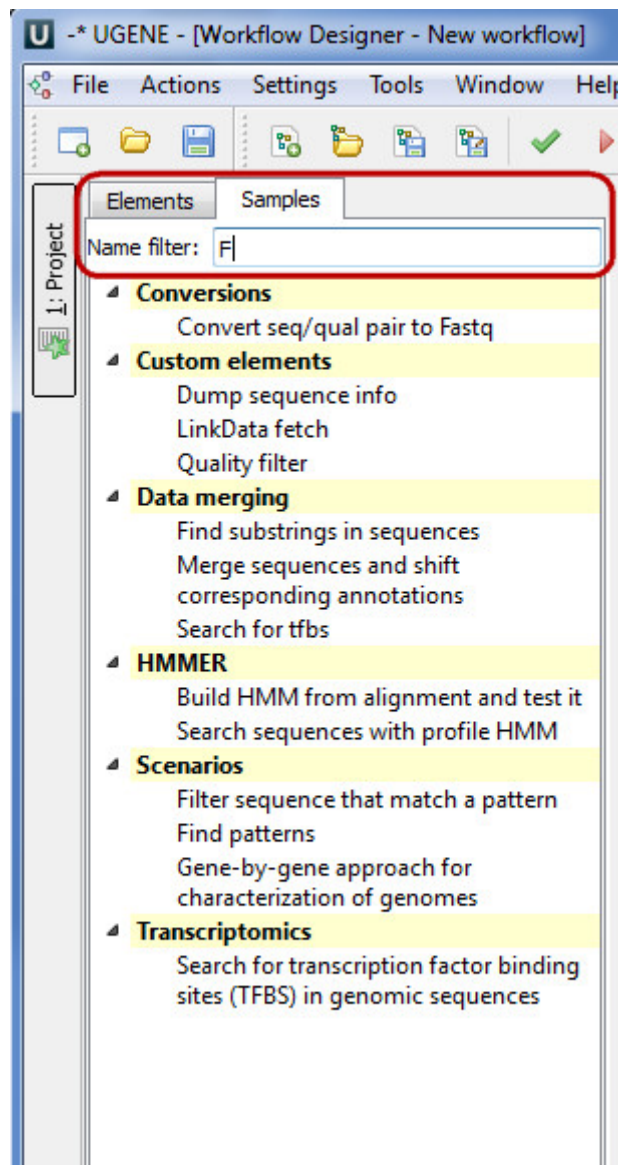
**Slots:**

Slot In GUI	Slot in Workflow File	Type
-------------	-----------------------	------

Sequence	sequence	<i>sequence</i>
----------	----------	-----------------

## Workflow Samples

This section contains detailed description of workflow samples presented in the Workflow Designer. To search a sample use the name filter or press the *Ctrl+F* shortcut that moves you to the name filter also:



- Alignment
  - Align sequences with MUSCLE
- Conversions
  - Convert seq/qual pair to Fastq
  - Convert alignments to ClustalW
  - Convert UQL schema results to alignment
  - Convert sequence to Genbank
- Custom elements
  - CASAWA FASTQ Filter
  - FASTQ Trimmer
  - Dump sequence info
  - LinkData fetch
  - Quality filter
- Data Marking
  - Marking Sequences by Annotation Number
  - Marking Sequences by Length
- Data Merging
  - Find Substrings at Sequences
  - Merge Sequences and Shift Corresponding Annotations
  - Search for TFBS
- HMMER
  - Build HMM from alignment and test it
  - Search sequences with profile HMM
- NGS

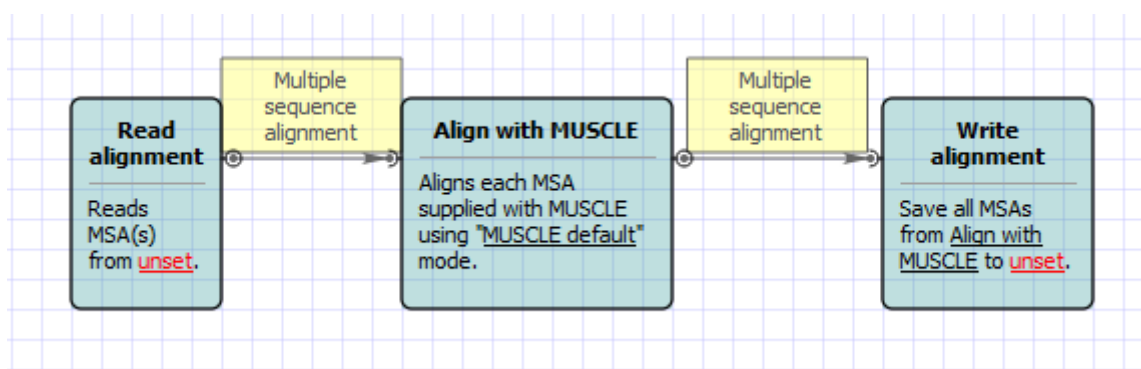
- Call variants with SAMtools
- ChIP-seq analysis with Cistrome tools
- Extract Consensus
- Extract transcript sequences
- RNA-seq analysis with Tuxedo tools
- Scenarios
  - Filter sequence that match a pattern
  - Find patterns
  - Gene-by-gene approach for characterization of genomes
  - Merge sequences and annotations
- Transcriptomics
  - Search for transcription factor binding sites (TFBS) in genomic sequences

## Alignment

- Align sequences with MUSCLE

### Align sequences with MUSCLE

This workflow performs multiple sequence alignment with MUSCLE algorithm and saves the resulting alignment to Stockholm document. Source data can be of any format containing sequences or alignments. To use this workflow, you need to specify locations for input and output file(s). To do this, select a corresponding task, so its' parameters appear in Property Inspector panel, and specify desired value(s) for "URL" parameter. Then you can launch the workflow with pressing Ctrl+R keys.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Conversions

- Convert seq/qual pair to Fastq
- Convert alignments to ClustalW
- Convert UQL schema results to alignment
- Convert sequence to Genbank

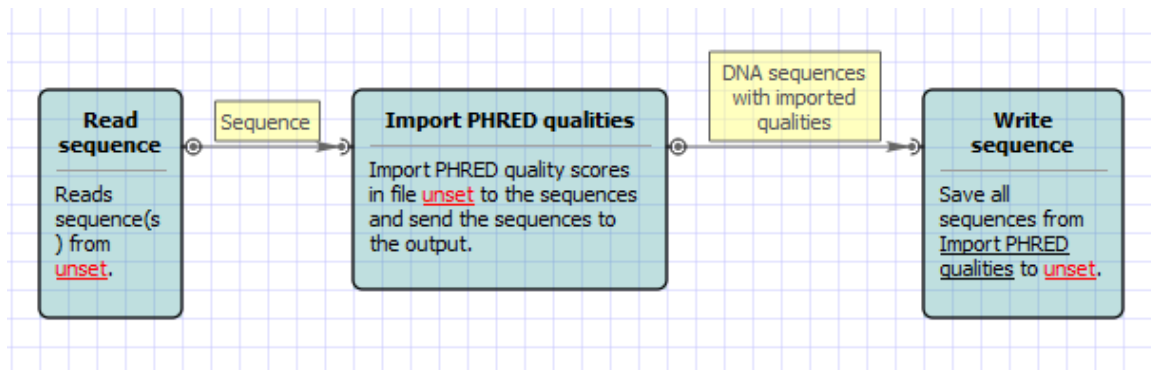
### Convert seq/qual pair to Fastq

This workflow allows to add PHRED quality scores to the sequence and save output to Fastq. For example, one can read a Fasta file, import PHRED quality values from corresponding qualities file and export the result to Fastq.

To execute the workflow do the following:

1. Select "Sequence Reader" task and specify source file(s) at "URL" field in the Property Editor.
2. Select "Import PHRED qualities" task and specify URL to the quality file. Usually such files have .qual extension.
3. Launch the schema with pressing Ctrl+R shortcut.

After running the workflow, target fastq file will appear in the same folder as the source file, with the same name but different extension.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Convert alignments to ClustalW

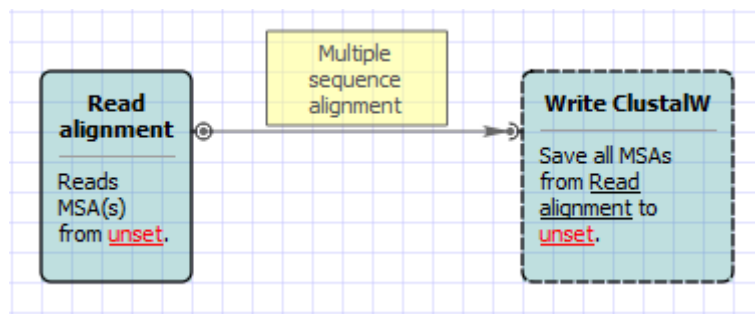
This workflow converts multiple alignment file(s) of any format to ClustalW document(s). If source file is a sequence format (e.g. FASTA), all contained sequences are added to the result alignment. Yet no real alignment is performed, this particular workflow illustrates pure data format conversion.

To get this workflow working, you only need to select "Alignment Reader" task, so its' parameters appear in Property Inspector panel, and specify source file(s) at "URL" field. Launch the workflow with pressing Ctrl+R shortcut.

After running the workflow, target clustal file will appear in the same folder as the source file, with the same name but different extension (".aln").

If several input files were selected, several clustal files will be generated accordingly.

You can override the target file location by editing "URL" parameter of "Write ClustalW" task.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

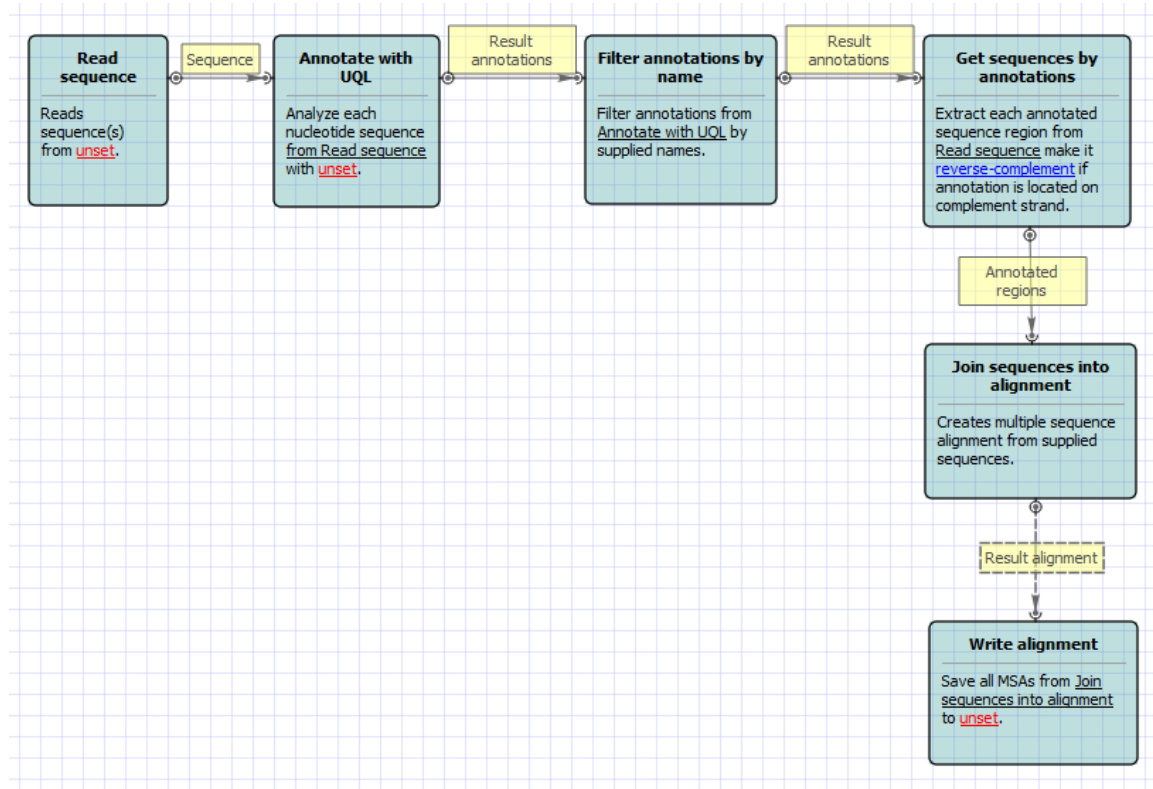


## Convert UQL schema results to alignment

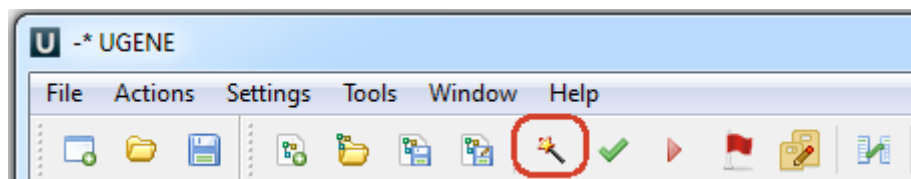
This schema allows to analyze sequence with Query and save results as alignment of selected features.

To execute the workflow do the following:

1. Select "Sequence Reader" task and specify source file at "URL" field in the Property Editor.
2. Select "Annotate with UQL" task and specify the URL of the UQL schema file.
3. Select "Filter annotations by name" task and specify the name of features to be joined into alignment.
4. Select "Join sequences into alignment" task and specify the URL of the result file.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



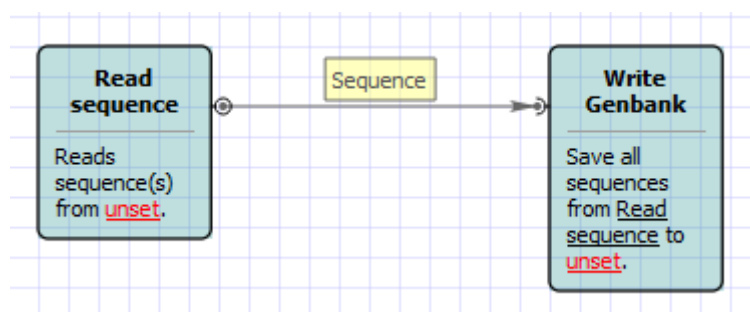
## Convert sequence to Genbank

This workflow converts sequence file(s) of any format (including PDB, alignments etc) to Genbank document(s). If source format supports annotations, they are also saved as feature tables in target file. Sequence meta-information (accessions etc) is preserved as well.

To get this workflow working, you only need to select "Sequence Reader" task, so its' parameters appear in Property Inspector panel, and specify source file(s) at "URL" field. Launch the workflow with pressing Ctrl+R keys.

After running the workflow, target genbank file will appear in the same folder as the source file, with the same name but different extension (".gb" by default). If several input files were selected, several output files will be generated.

You can override the target file location by editing "URL" parameter of "Write Genbank" task. In this case all data from different sources will be saved to the single location (unless you change "Accumulate objects" parameter, see related docs).



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



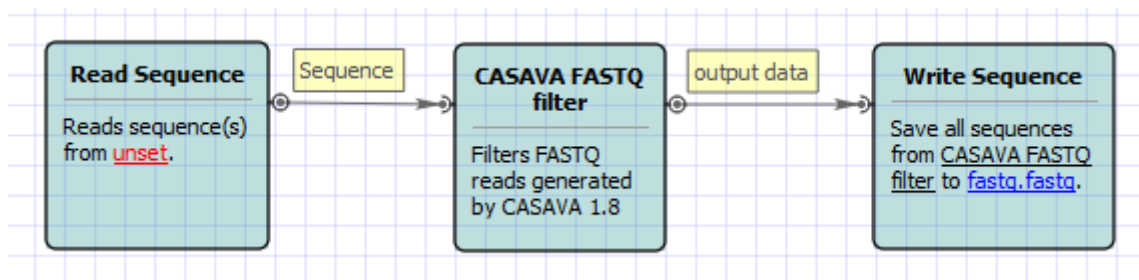


## Custom elements

- CASAVA FASTQ Filter
- FASTQ Trimmer
- Dump sequence info
- LinkData fetch
- Quality filter

## CASAVA FASTQ Filter

Reads in FASTQ file produced by CASAVA 1.8 contain 'N' or 'Y' as a part of an identifier. 'Y' if a read is filtered, 'N' if the read is not filtered. The workflow cleans up the filtered reads.

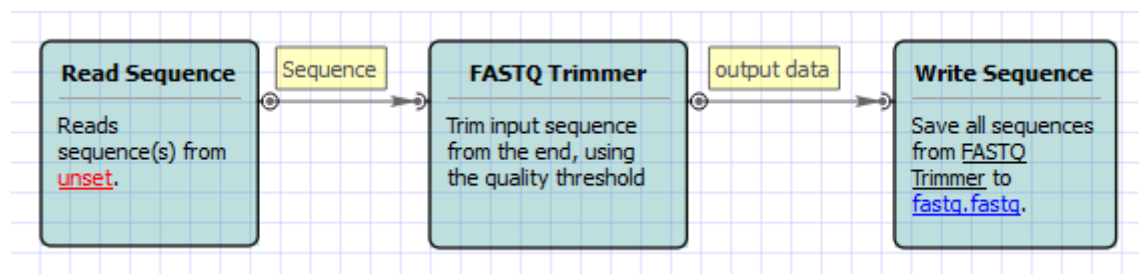


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## FASTQ Trimmer

The workflow scans each input sequence from the end to find the first position where the quality is greater or equal to the minimum quality threshold. Then it trims the sequence to that position. If the whole sequence has quality less than the threshold or the length of the output sequence less than the minimum length threshold then the sequence is skipped.

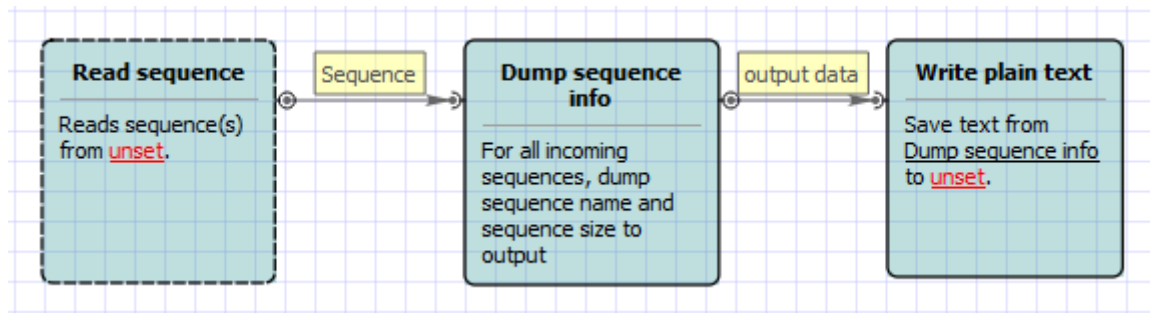


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Dump sequence info

This workflow dump sequence name and sequence size to output for all incoming sequences,.

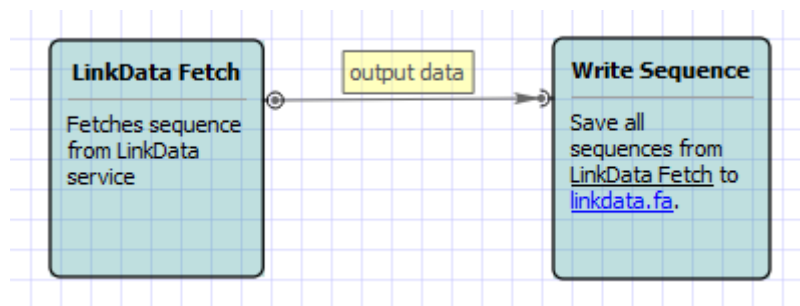


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

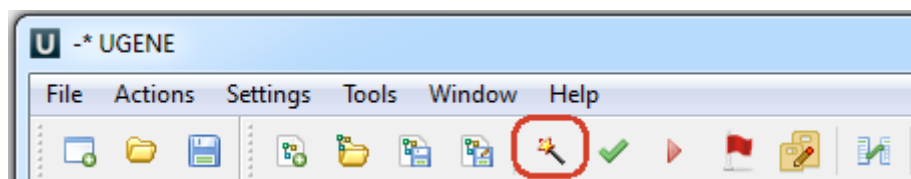


## LinkData fetch

This workflow fetches sequence from LinkData by specified work ID, filename, subject ID, property ID and writes result in file in FASTA format

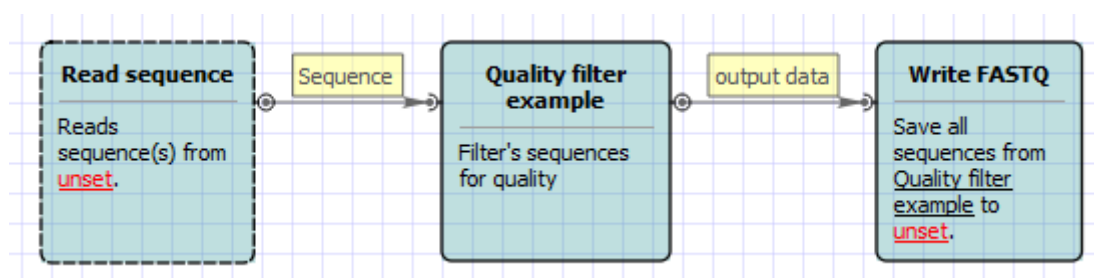


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Quality filter

This workflow filters sequences with quality  $\geq$  than parameter "quality" and writes result in file in FASTQ format.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Data Marking

- [Marking Sequences by Annotation Number](#)
- [Marking Sequences by Length](#)

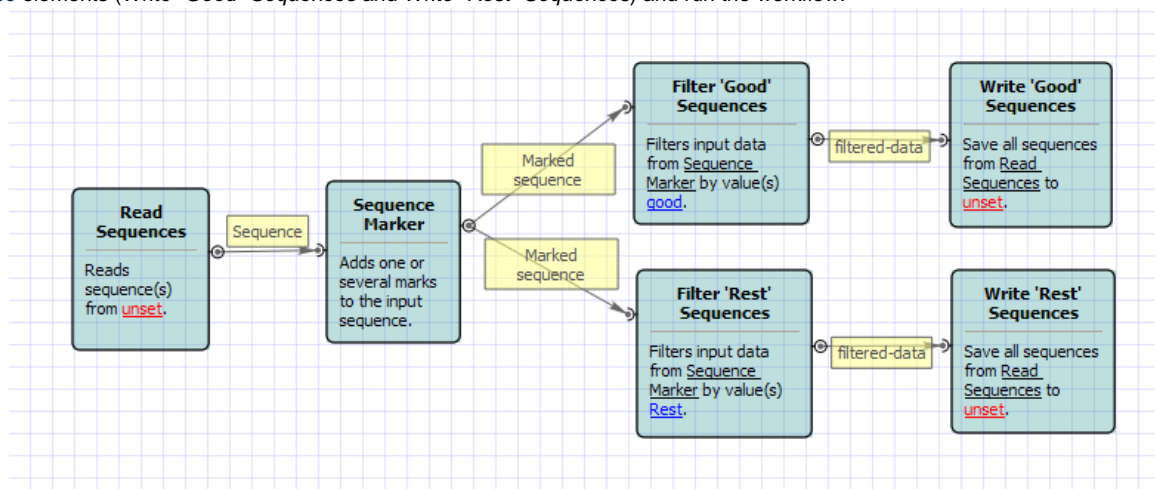
### Marking Sequences by Annotation Number

This sample describes how to identify sequences with the specified number of annotations.

First, the workflow reads sequences input by a user. Then, each sequence is marked either with the “Good” or with the “Rest” mark, depending on the number of the sequence annotations. After marking, the sequences are filtered by the marks. And finally, the filtered sequences are written into files, specified by a user.

By default, a sequence with 1 or more annotations is marks as “Good”. You can configure this value in the [Sequence Marker](#) element parameters. Also, it is possible to set up the annotation names that should be taken into account.

To try out this sample, add the input files to the [Read Sequence](#) element, select the name and location of the output files in the [Write Sequence](#) elements ([Write “Good” Sequences](#) and [Write “Rest” Sequences](#)) and run the workflow.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



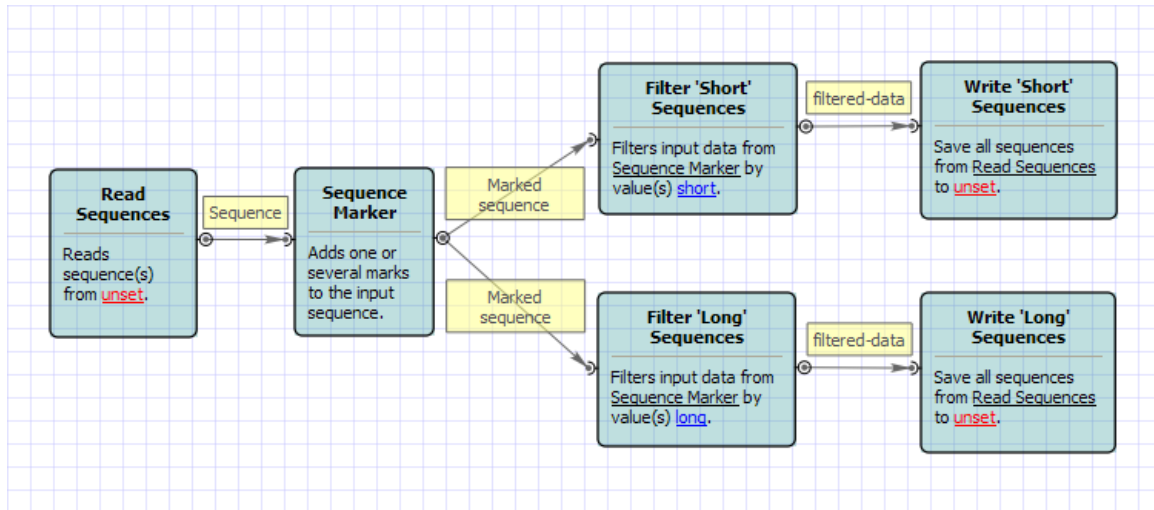
### Marking Sequences by Length

This sample describes how to identify sequences with the specified length.

First, the workflow reads sequences input by a user. Then, each sequence is marked either with the “Short” or with the “Long” mark, depending on the sequence length. After marking, the sequences are filtered by the marks. And finally, the filtered sequences are written into files, specified by a user.

By default, a sequence with a length 200 or less bp is marks as “Short”. A sequence with a length more than 200 bp is marks as “Long”. You can configure this value in the [Sequence Marker](#) element parameters.

To try out this sample, add the input files to the [Read Sequence](#) element, select the name and location of the output files in the [Write Sequence](#) elements ([Write “Short” Sequences](#) and [Write “Long” Sequences](#)) and run the workflow.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Data Merging

- Find Substrings at Sequences
- Merge Sequences and Shift Corresponding Annotations
- Search for TFBS

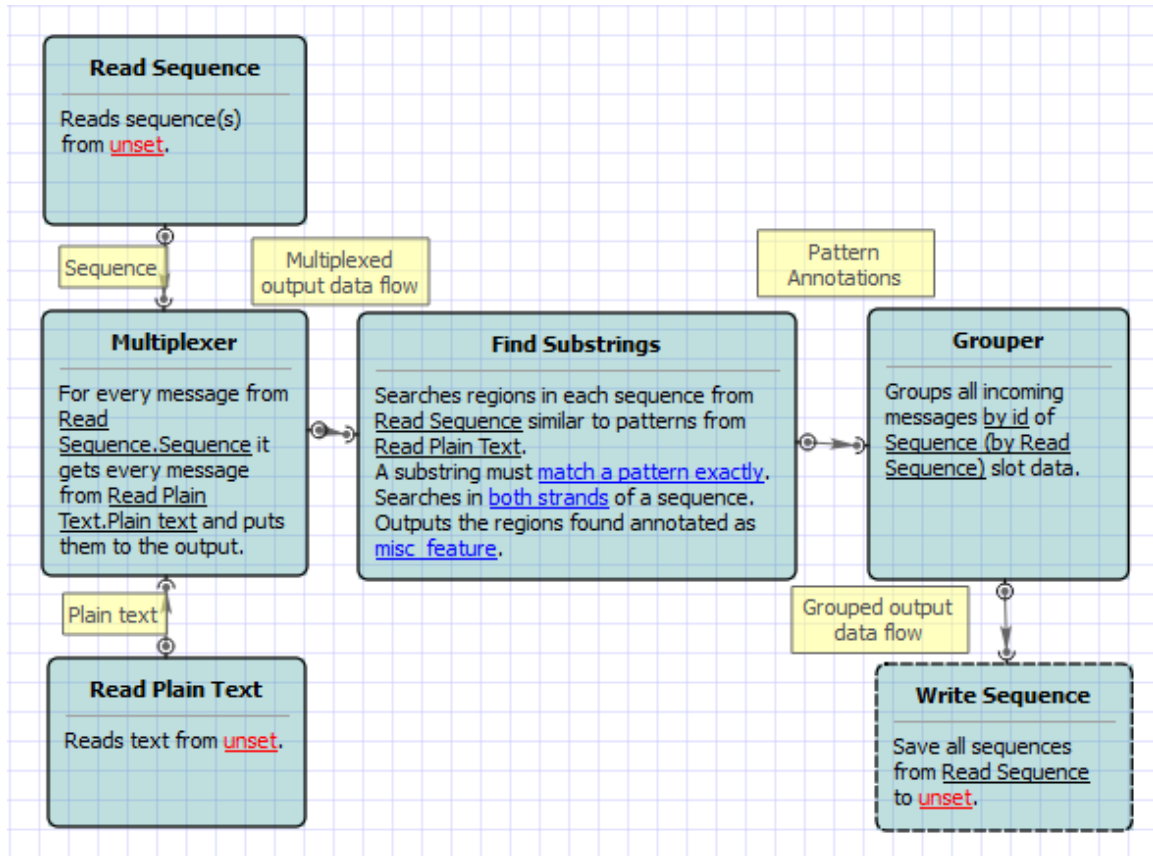
### Find Substrings at Sequences

This scheme describes how to find substrings in sequences and group these sequences by different parameters.

First, the workflow reads sequences and text strings (patterns) from files. Then, these data sets are multiplexed using this rule: every sequence is united with every pattern. After multiplexing these united data sets are transported to the find patterns element. The results of patterns searching are grouped by id of a sequence: original and find patterns annotations are united into two new grouped annotations sets. And finally, the grouped data are written into file, specified by a user.

By default, sequence multiplexed using the rule "1 to 1". You can configure this value in the *Multiplexer* element parameters. Also, you can configure the *Pattern* element parameters and *Grouper* element parameters.

To try out this sample, add the input files to the *Read Sequence* and *Read Plain Text* elements, select the name and location of the output files in the *Write Sequence* element and run the workflow.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



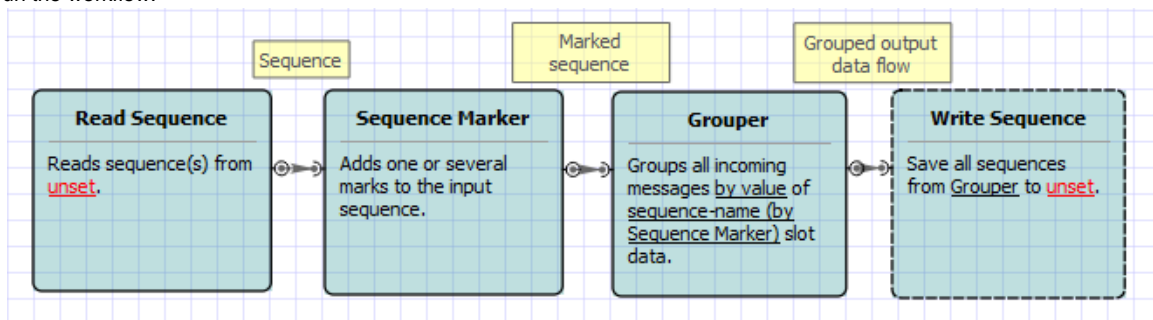
## Merge Sequences and Shift Corresponding Annotations

This scheme describes how to merge sequences and manipulate with its annotations.

First, the workflow reads sequence(s) from file(s). Then, marks the input sequences with the sequence name marker. After marking the sequences are grouped by the marker. Sequences with equal markers are merged into one sequence. Annotations are shifted using the position of the corresponding sequence at the merged sequence. And finally, the grouped data are written into file, specified by a user.

By default, sequence is marked using the sequence name marker. You can configure this value in the *Marker* element parameters. Also, you can configure the *Grouper* element parameters.

To try out this sample, add the input files to the *Read Sequence*, select the name and location of the output files in the *Write Sequence* element and run the workflow.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



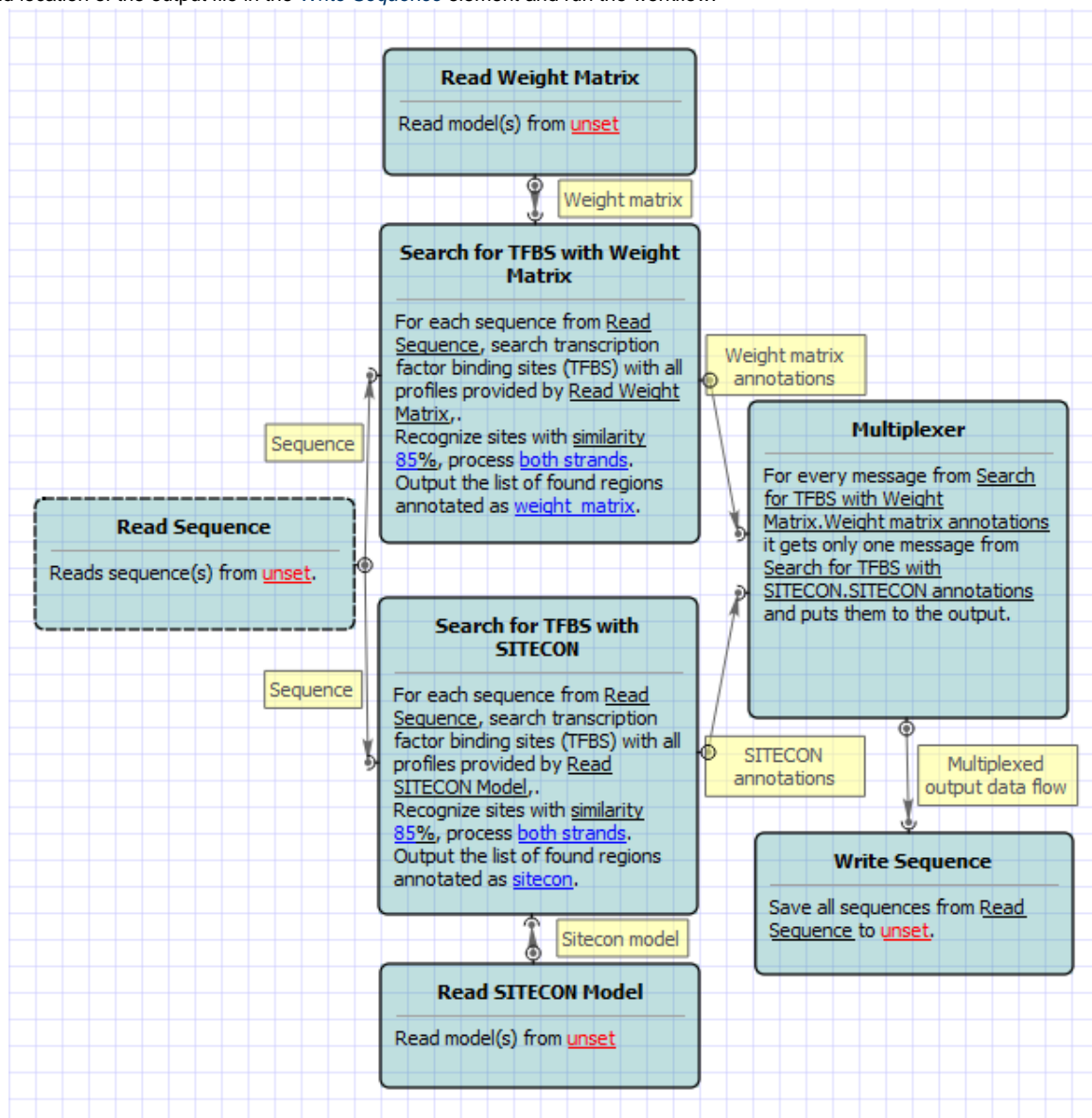
## Search for TFBS

This sample describes how to search for TFBS with a different methods and how to write the results into one output file.

First, the workflow reads sequences input by a user. Then, each sequence goes to searching TFBS elements. At that time two reading elements reads the matrix and model for TFBS searching and transferring this data into TFBS searching elements. After that the TFBS searching elements searches TFBS in the input sequences. After that the two data flows multiplexes into one output data flow. And finally, the multiplexed data are written into file, specified by a user.

You can configure the parameters of *Search for TFBS with Weight Matrix*, *Search TFBS with SITECON* and *Multiplexer* elements.

To try out this sample, add the input files to the *Read Sequence* element, select the *Read Weight Matrix*, *Read SITECON model* and select name and location of the output file in the *Write Sequence* element and run the workflow.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

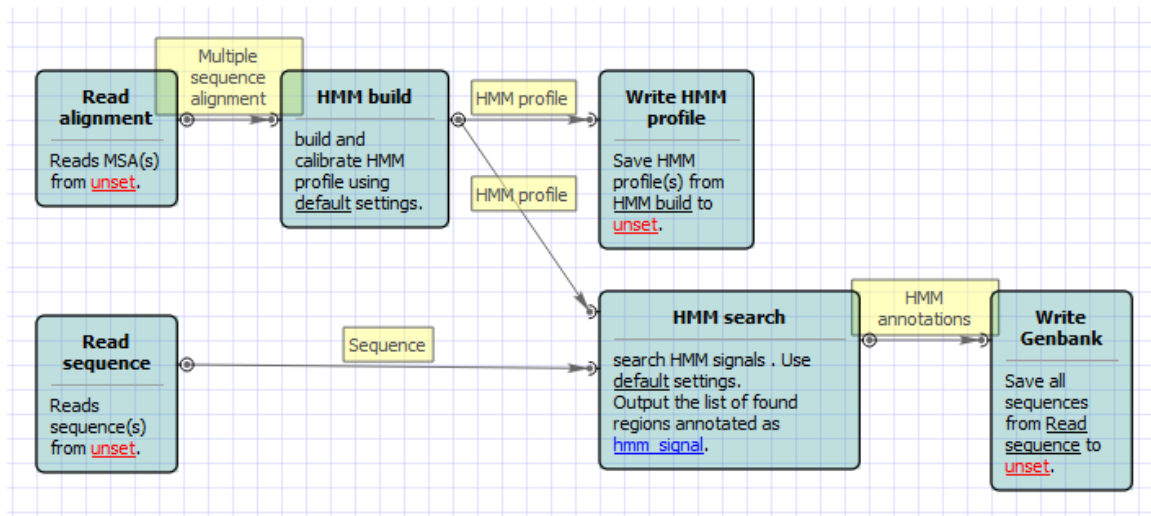


## HMMER

- Build HMM from alignment and test it
- Search sequences with profile HMM

### Build HMM from alignment and test it

This workflow builds a new profile HMM from input alignment, calibrates the HMM and saves to a file. Then runs a test HMM search over sample sequence and saves test results to Genbank file. To run this workflow, you need to specify appropriate locations for input/output files. This is achieved by selecting a task and editing interesting parameters in Property Inspector panel. Optionally, fine tune the build/search parameters as you see fit. Then schedule the workflow for execution by pressing CTRL+R shortcut. You can watch its" progress in a Task View of UGENE.



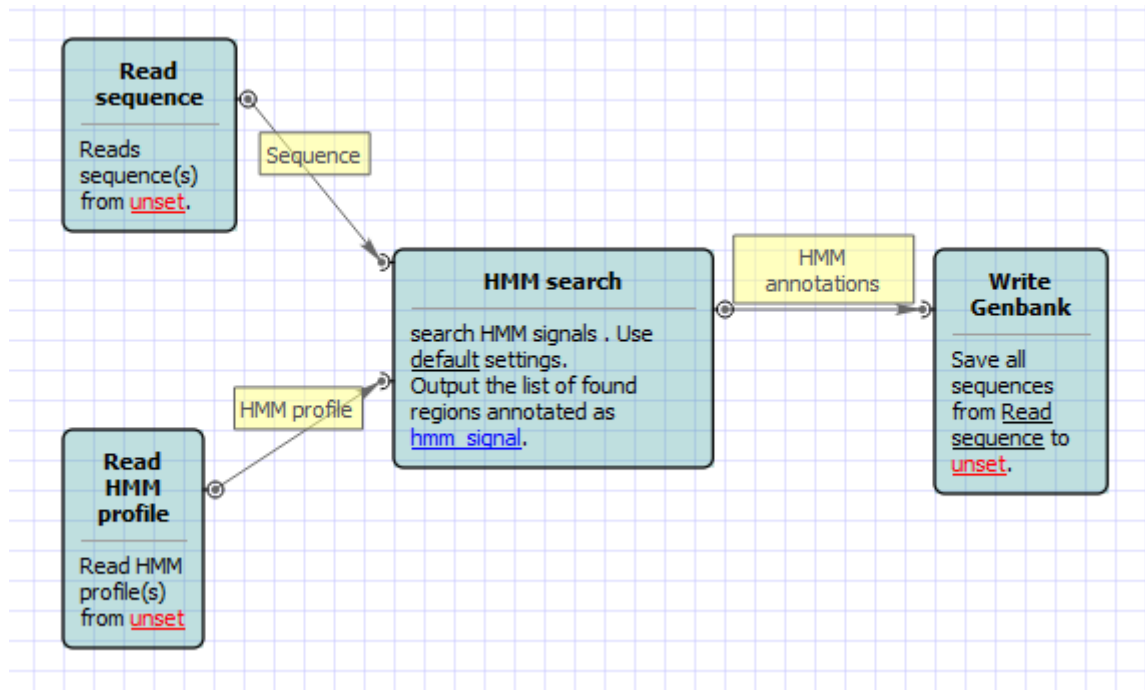
Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



### Search sequences with profile HMM

This workflow reads an HMM from a file and searches input sequences for significantly similar matches, saves found signals to a file. You can specify several input files for both HMM and sequences, the workflow will process Cartesian product of inputs. That is, each sequence will be searched with all specified HMMs in turn. To specify task parameters, select it and edit interesting fields in table "Parameters" of Property Inspector panel. Schedule the workflow for execution by pressing CTRL+R shortcut. You can watch its" progress in Task View of UGENE.





Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

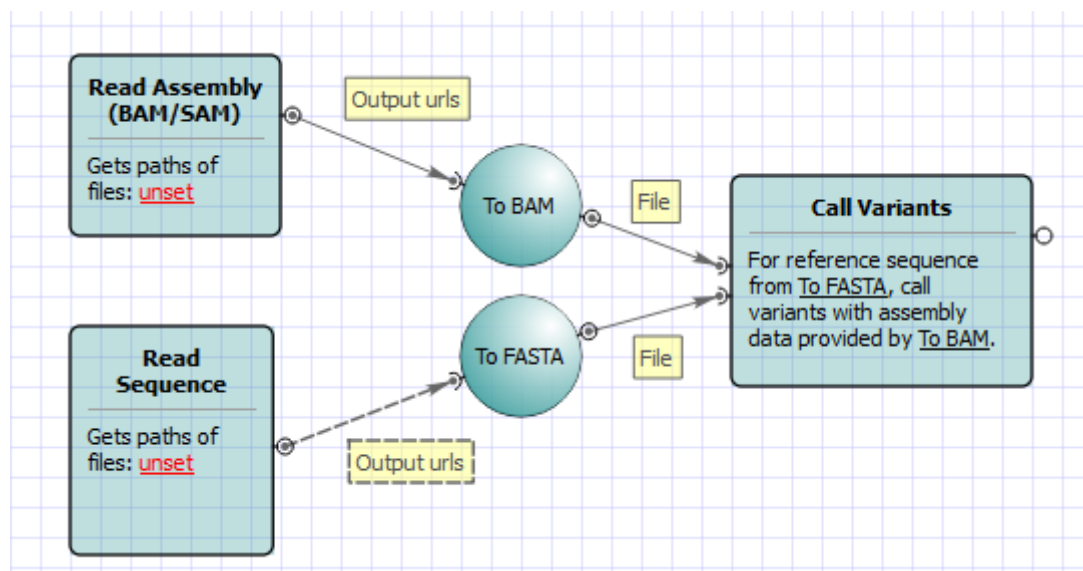


## NGS

- Call variants with SAMtools
- ChIP-seq analysis with Cistrome tools
- Extract Consensus
- Extract transcript sequences
- RNA-seq analysis with Tuxedo tools

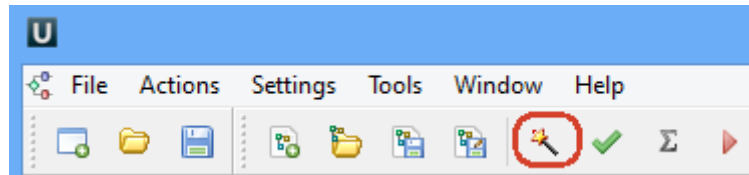
### Call variants with SAMtools

Call variants in UGENE can be done using SAMtools mpileup and bcftools view utilities. To read additional information about SAMtools and its utilities visit [SAMTools homepage](#). Both utilities are embedded into UGENE and there is no need in additional configuration.

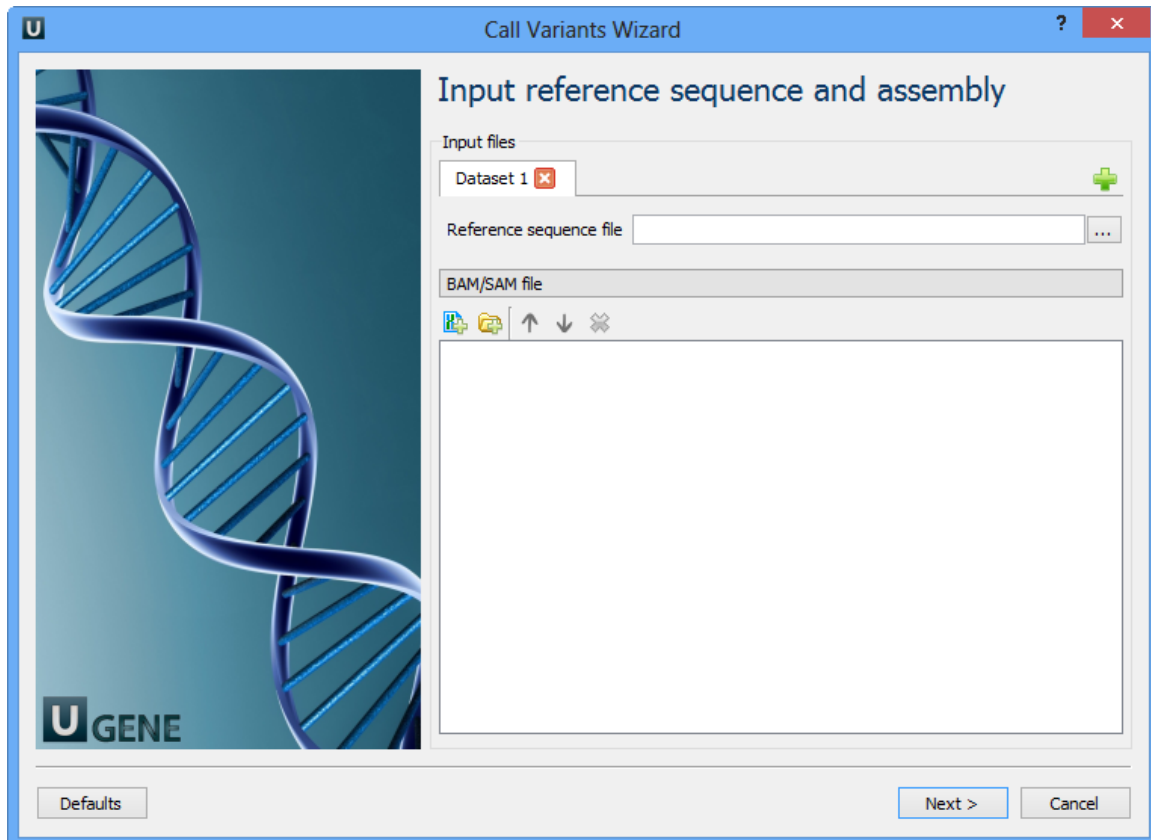




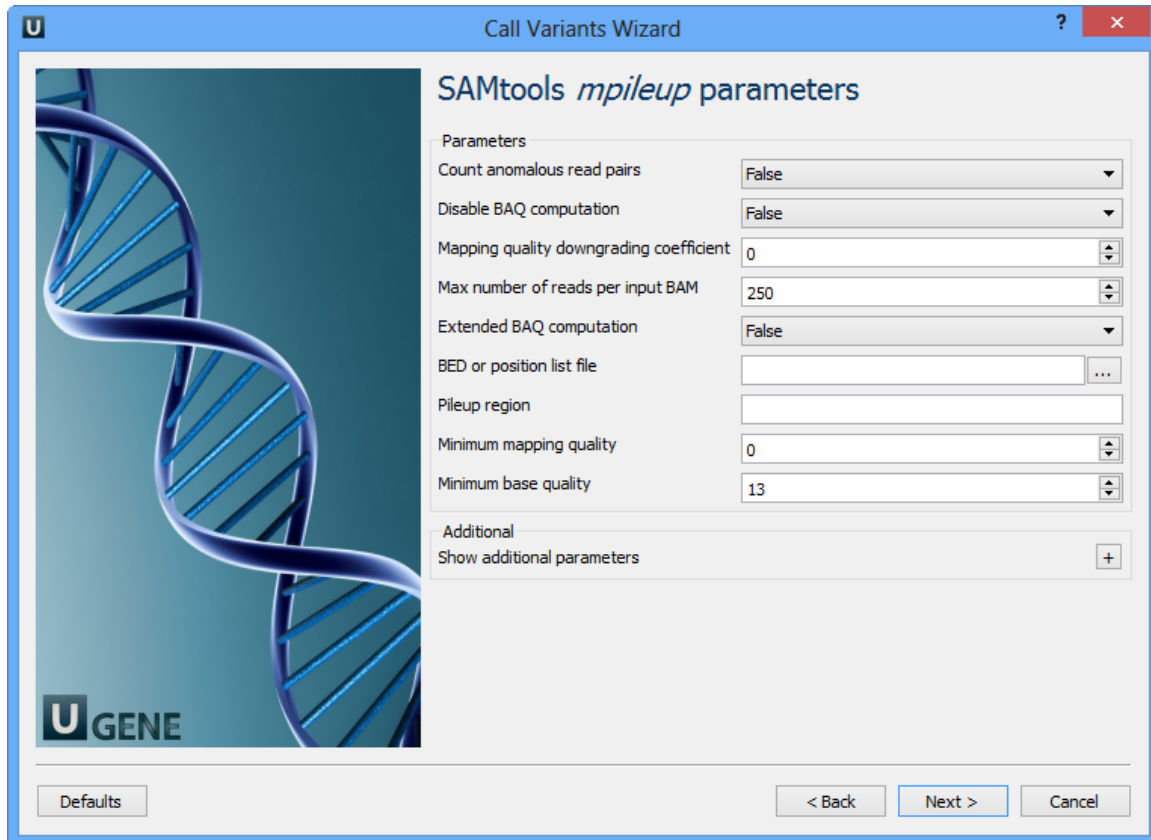
To run the workflow you need to select an input reference sequence, a BAM or SAM file and an output file with variations. Optionally, you can change other parameters, for example, set additional parameters of the SAMtools mpileup and bcftools view utilities. Use the workflow wizard to guide you through the parameters setup process. Click Show wizard button on the Workflow Designer toolbar to open it:



The first wizard page appears:



Here you need to input a file with a reference sequence and a sorted BAM or SAM file. Note that the input BAM or SAM file may be unsorted. Click the Next button. The next page appears:

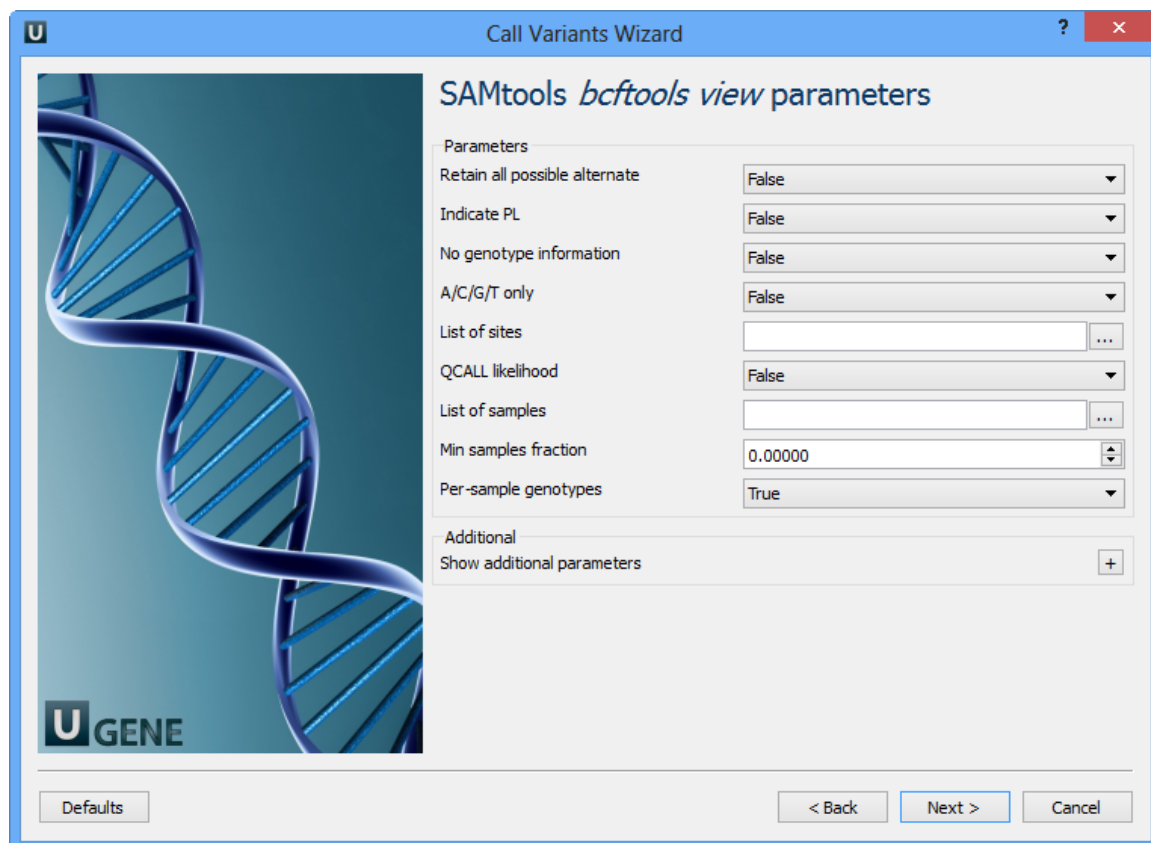


Here you can change default parameters of the SAMtools mpileup utility. To show additional parameters click the + button. The following parameters are available:

Count anomalous read pairs	Do not skip anomalous read pairs in variant calling.
Disable BAQ computation	Disable probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being misaligned. Applying this option greatly helps to reduce false SNPs caused by misalignments.
Mapping quality downgrading coefficient	Coefficient for downgrading mapping quality for reads containing excessive mismatches. Given a read with a phred-scaled probability $q$ of being generated from the mapped position, the new mapping quality is about $\sqrt{(\text{INT}-q)/\text{INT}} \cdot \text{INT}$ . A zero value disables this functionality; if enabled, the recommended value for BWA is 50.
Max number of reads per input BAM	At a position, read maximally INT reads per input BAM.
Extended BAQ computation	Extended BAQ computation. This option helps sensitivity especially for MNPs, but may hurt specificity a little bit.
BED or position list file	BED or position list file containing a list of regions or sites where pileup or BCF should be generated.
Pileup region	Only generate pileup in region STR.
Minimum mapping quality	Minimum mapping quality for an alignment to be used.
Minimum base quality	Minimum base quality for a base to be considered.
Illumina-1.3+encoding	Assume the quality is in the Illumina 1.3+ encoding.
Gap extension error	Phred-scaled gap extension sequencing error probability. Reducing INT leads to longer indels.

Homopolymer errors coefficient	Coefficient for modeling homopolymer errors. Given an I-long homopolymer run, the sequencing error of an indel of size s is modeled as $INT^s/I$ .
No INDELS	Do not perform INDEL calling.
Max INDEL depth	Skip INDEL calling if the average per-sample depth is above INT.
Gap open error	Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls.
List of platforms for indels	Comma delimited list of platforms (determined by @RG-PL) from which indel candidates are obtained. It is recommended to collect indel candidates from sequencing technologies that have low indel error rate such as ILLUMINA.

Choose these parameters and click the Next button. The next page appears:

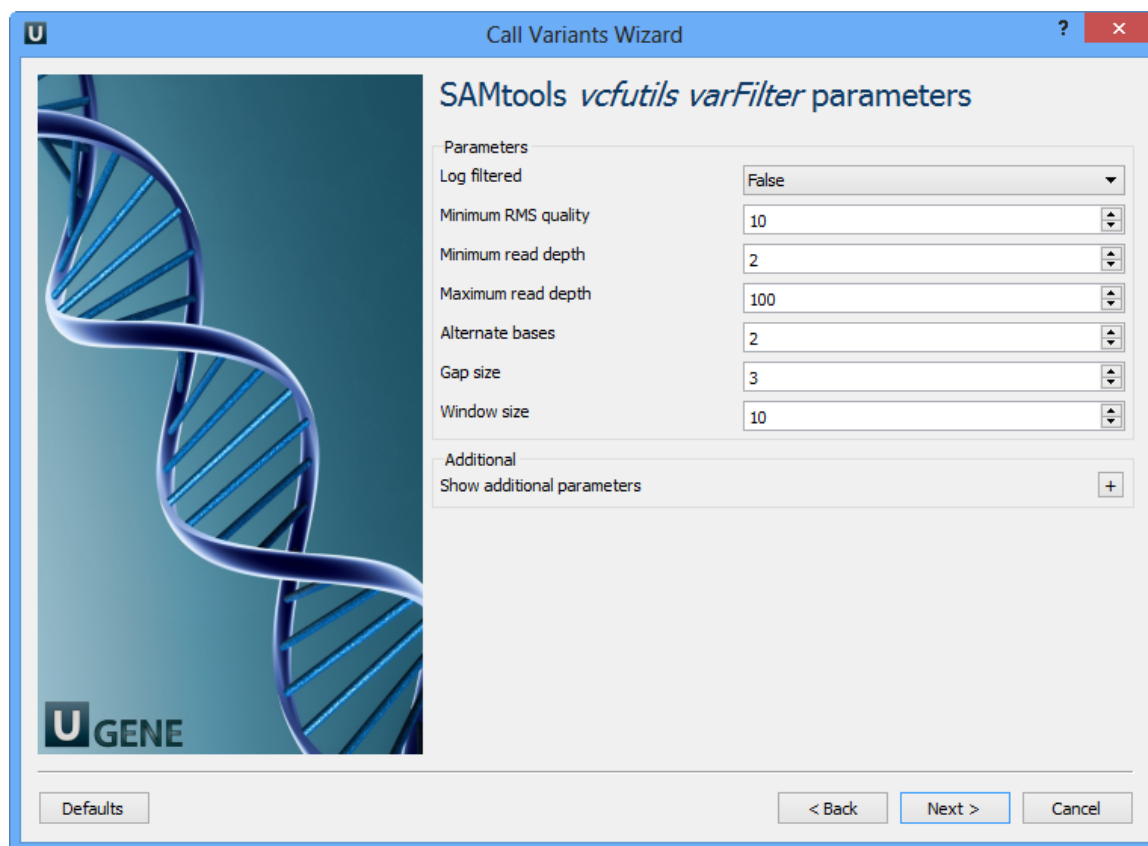


The next page allows one to configure SAMtools bcftools view utility parameters:

Retain all possible alternative	Retain all possible alternate alleles at variant sites. By default, the view command discards unlikely alleles.
Indicate PL	Indicate PL is generated by r921 or before (ordering is different).
No genotype information	Suppress all individual genotype information.
A/C/G/T only	Skip sites where the REF field is not A/C/G/T.
List of sites	List of sites at which information are outputted.
QCALL likelihood	Output the QCALL likelihood format.
List of samples	List of samples to use. The first column in the input gives the sample names and the second gives the ploidy, which can only be 1 or 2. When the 2nd column is absent, the sample ploidy is assumed to be 2. In the output, the ordering of samples will be identical to the one in FILE.

Min samples fraction	Skip loci where the fraction of samples covered by reads is below FLOAT.
Per-sample genotypes	Call per-sample genotypes at variant sites.
INDEL-to-SNP Ratio	Ratio of INDEL-to-SNP mutation rate.
Gap open error	Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls.
Max P(ref D)	A site is considered to be a variant if P(ref D).
Pair/trio calling	Enable pair/trio calling. For trio calling, option -s is usually needed to be applied to configure the trio members and their ordering. In the file supplied to the option -s, the first sample must be the child, the second the father and the third the mother. The valid values of STR are "pair", "trioauto", "trioxd" and "trioxs", where "pair" calls differences between two input samples, and "trioxd" ("trioxs") specifies that the input is from the X chromosome non-PAR regions and the child is a female (male).
N group-1 samples	Number of group-1 samples. This option is used for dividing the samples into two groups for contrast SNP calling or association test. When this option is in use, the following VCF INFO will be outputted: PC2, PCHI2 and QCHI2.
N permutations	Number of permutations for association test (effective only with -1).
Max P(chi^2)	Only perform permutations for P(chi^2).

Choose these parameters and click the Next button. The next page of the wizard appears:

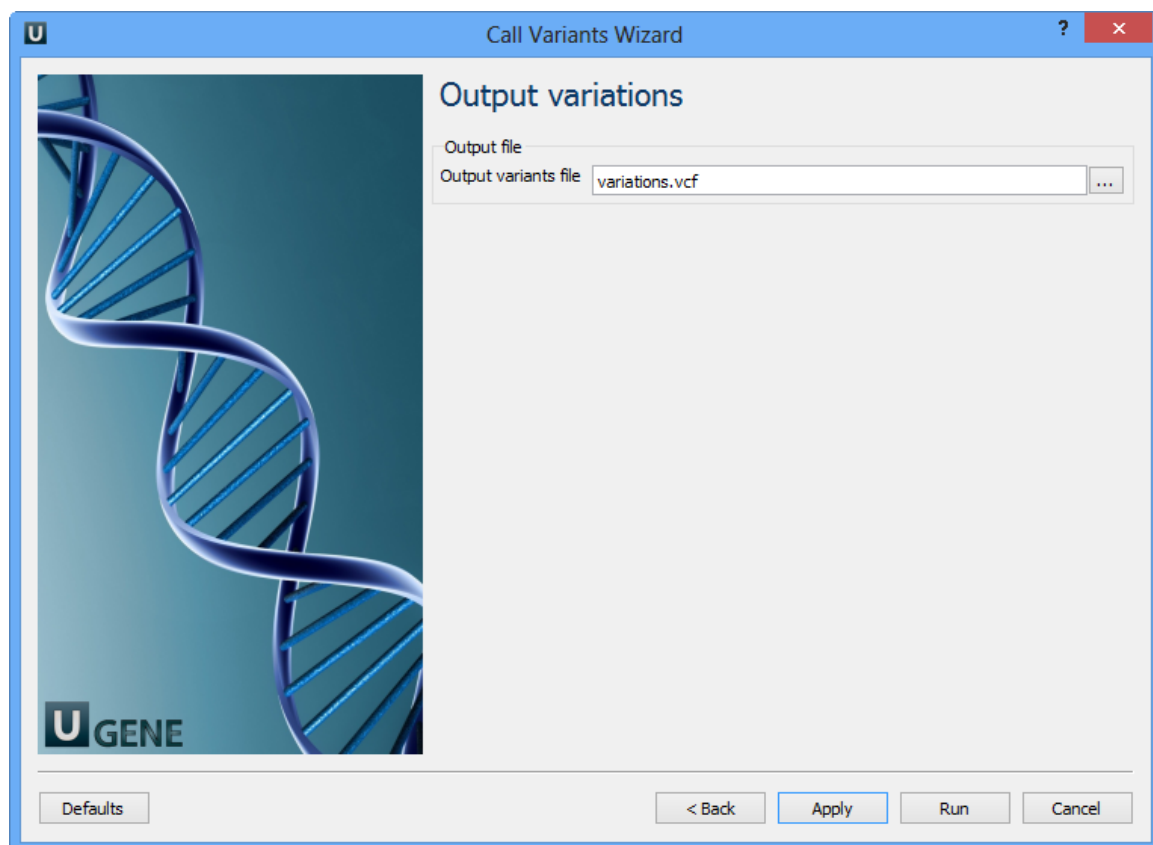


The next page allows one to configure SAMtools vcfutils parameters:

Log filtered	Print filtered variants into the log (varFilter) (-p).
Minimum RMS quality	Minimum RMS mapping quality for SNPs (varFilter) (-Q).

Minimum read depth	Minimum read depth (varFilter) (-d).
Maximum read depth	Maximum read depth (varFilter) (-D).
Alternate bases	Minimum number of alternate bases (varFilter) (-a).
Gap size	SNP within INT bp around a gap to be filtered (varFilter) (-w).
Window size	Window size for filtering adjacent gaps (varFilter) (-W).
Strand bias	Minimum P-value for strand bias (given PV4) (varFilter) (-1).
BaseQ bias	Minimum P-value for baseQ bias (varFilter) (-2).
MapQ bias	Minimum P-value for mapQ bias (varFilter) (-3).
End distance bias	Minimum P-value for end distance bias (varFilter) (-4).
HWE	Minimum P-value for HWE (plus F<0) (varFilter) (-e).

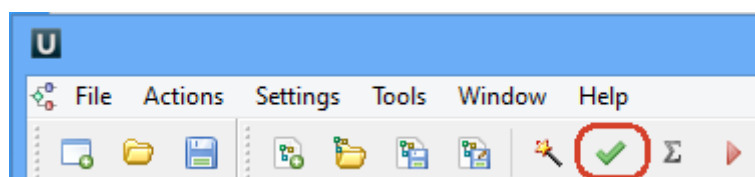
Choose these parameters and click the Next button. The last page of the wizard appears:



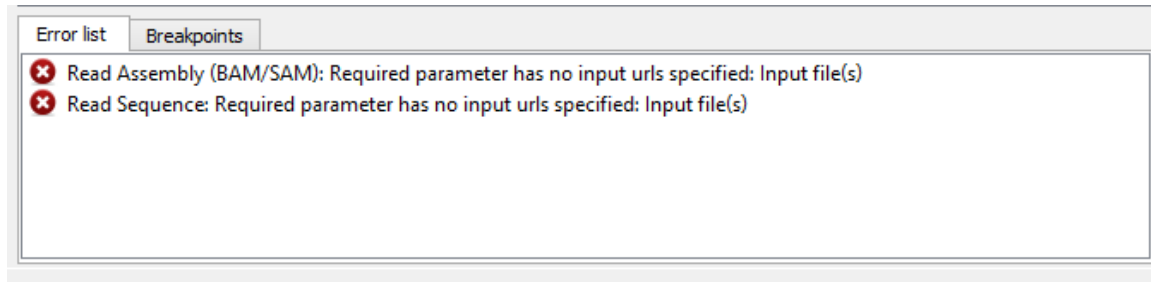
On this page you should select an output file. Set required parameters and click the Finish button.

Note that default button reverts all parameters to default settings.

Now let's validate and run the workflow. To validate that the workflow is correct and all parameters are set properly click the Validate workflow button on the Workflow Designer toolbar:



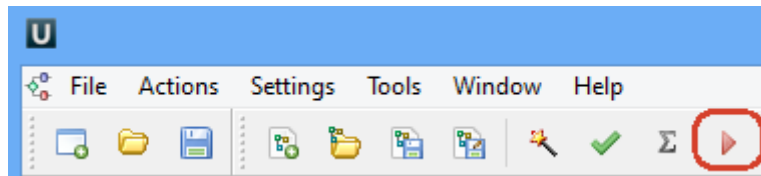
If there are some errors, they will be shown in the Error list at the bottom of the Workflow Designer window, for example:



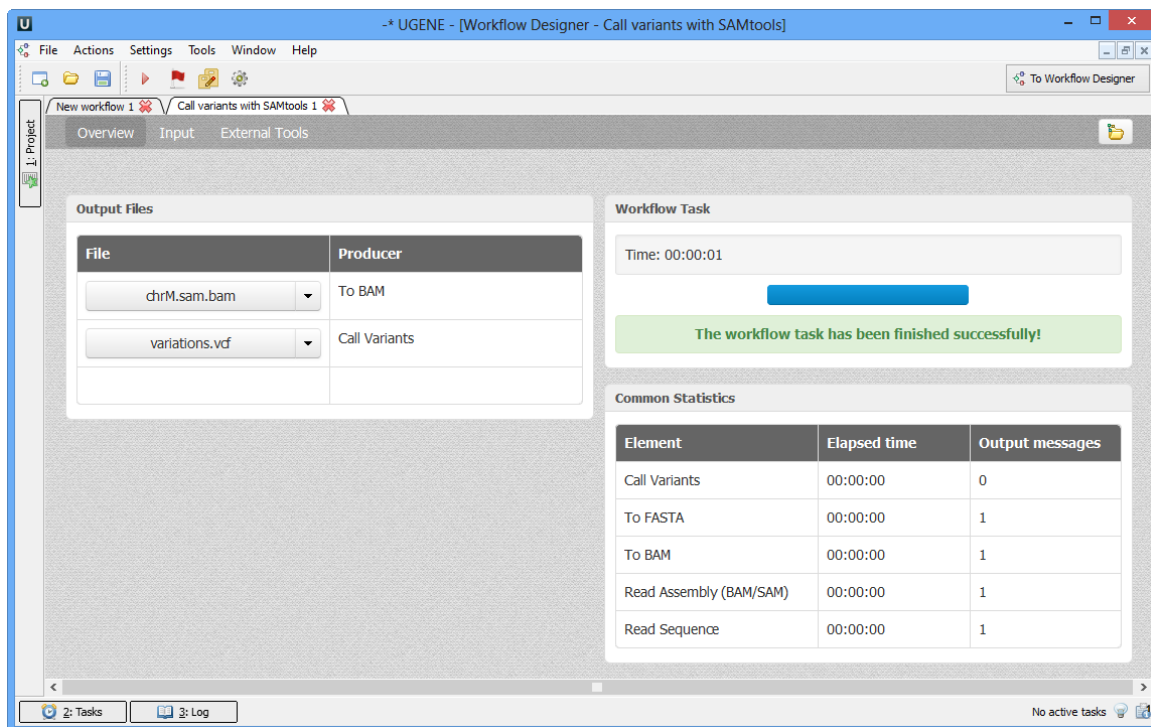
However, if you have set all the required parameters, then there shouldn't be errors. After that you can estimate the workflow. To run estimation click the *Estimate workflow* button:



To run a valid workflow, click the *Run workflow* button on the *Workflow Designer* toolbar:



As soon as the variants calling task is finished, a notification and dashboard will appear.



The dashboard will contain information about workflow: input and output files, all information about task. .


## ChIP-seq analysis with Cistrome tools

The ChIP-seq pipeline “Cistrome” integrated into UGENE allows one to do the following analysis steps: peak calling and annotating, motif search and gene ontology. ChIP-seq analysis is started from MACS tool. CEAS then takes peak regions and signal wiggle file to check which chromosome is enriched with binding/modification sites, whether bindings events are significant at gene features like promoters, gene

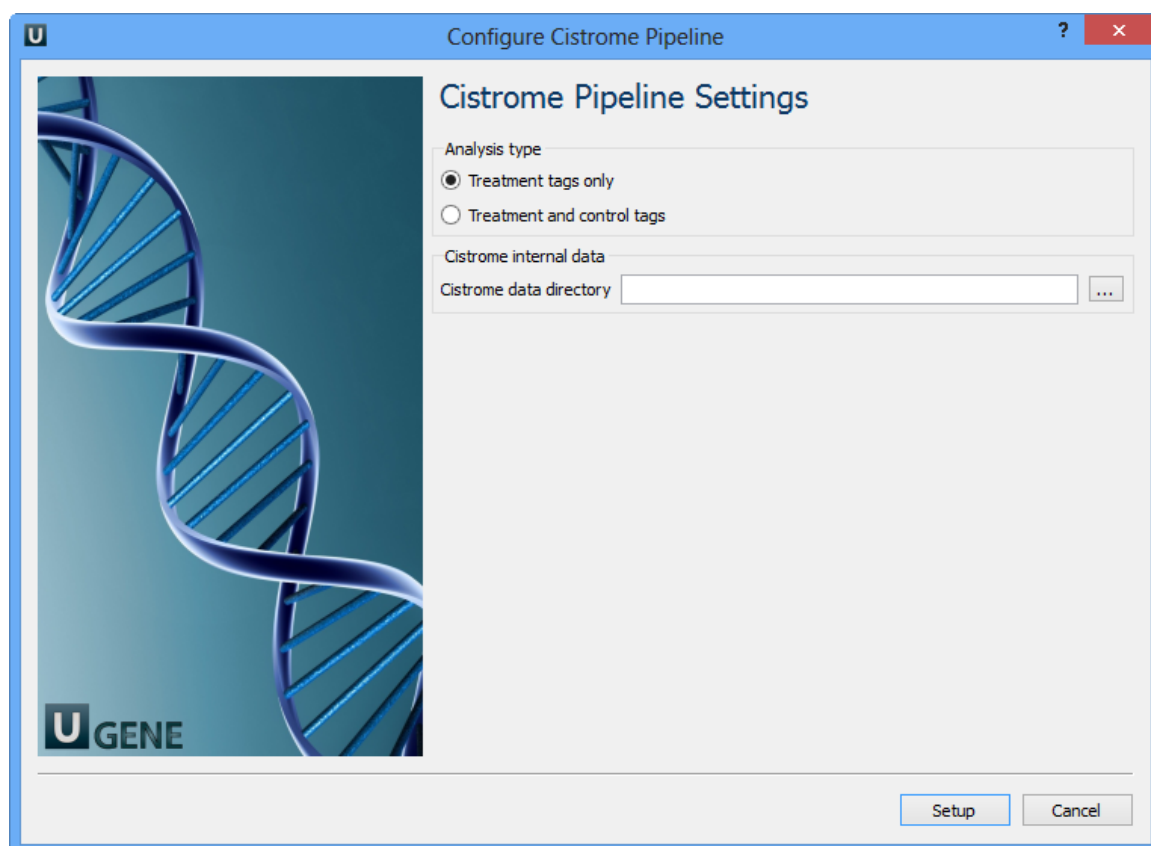
bodies, exons, introns or UTRs, and the signal aggregation at gene transcription start/end sites or meta-gene bodies (average all genes). Then peaks are investigated in these ways:

1. to check which genes are nearby so can be regarded as potential regulated genes, then perform GO analysis;
2. to check the conservation scores at the binding sites;
3. the DNA motifs at binding sites.

Note that it is originally based on the General ChIP-seq pipeline from the public [Cistrome installation](#) on the Galaxy workflow platform.

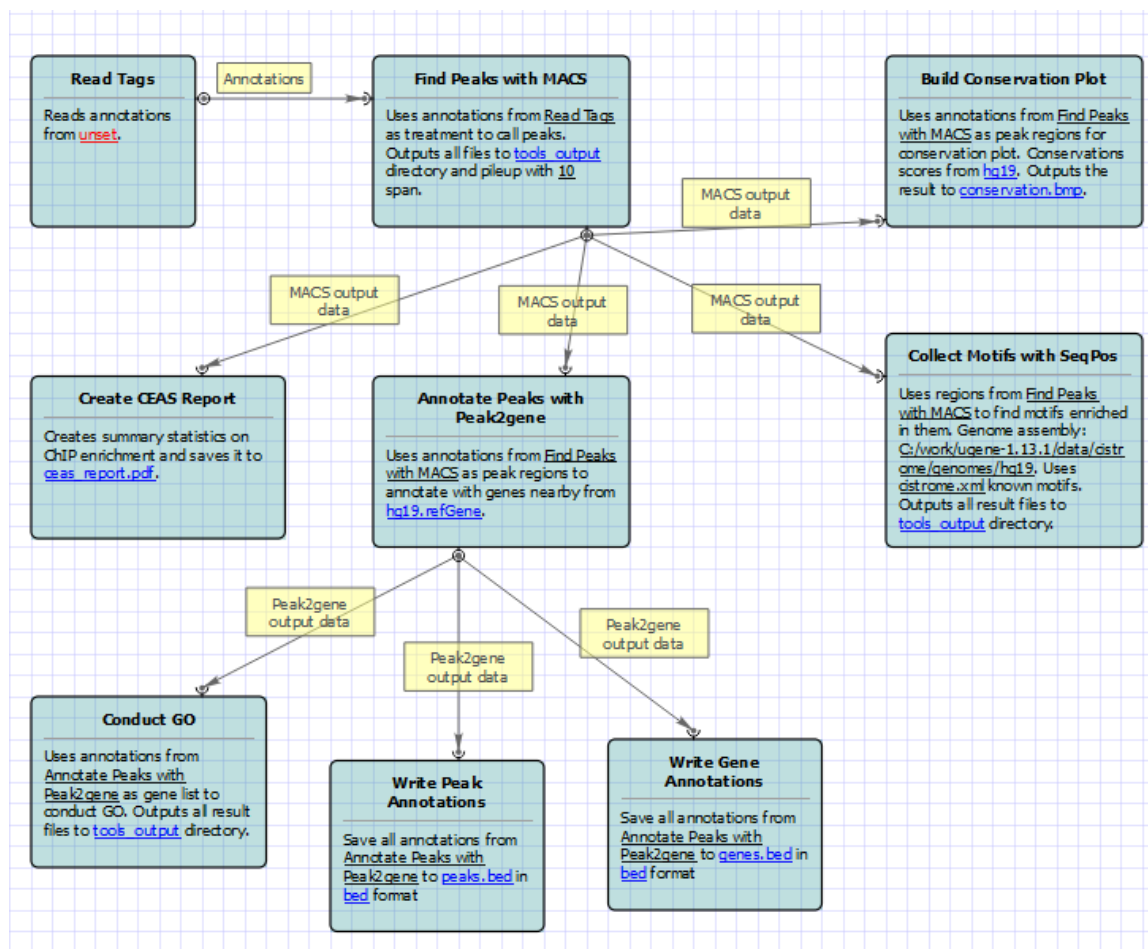
 Download and install the UGENE [NGS package](#) to use this pipeline.

Select Samples tab on the Workflow Designer Palette and double-click on the ChIP-seq analysis with Cistrome tools sample. The following configure wizard appears:



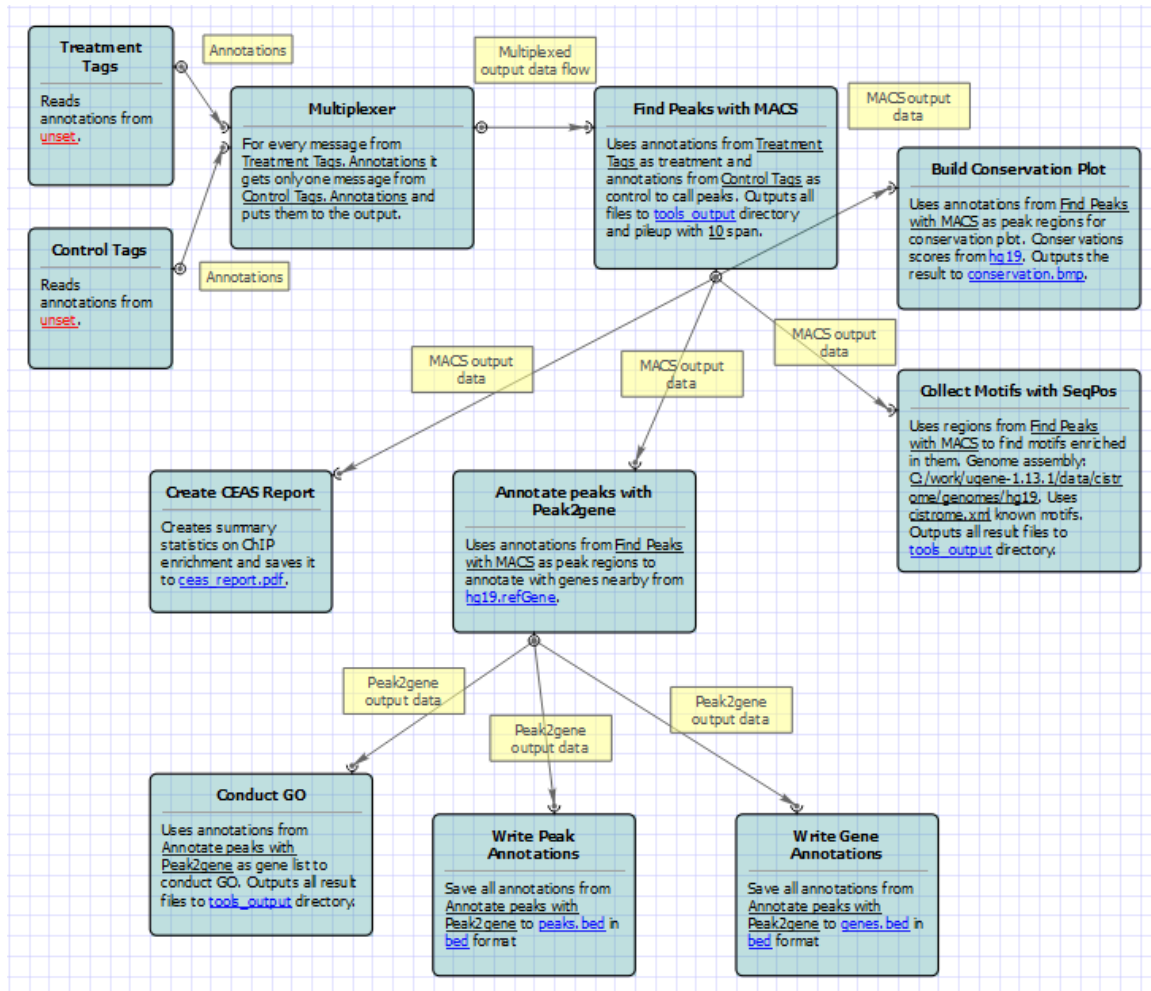
Here you need to choose analysis type and cistrome internal data and click Setup. For treatment tags only analysis type the following workflow appears:





For treatment and control tags analysis type the following workflow appears:

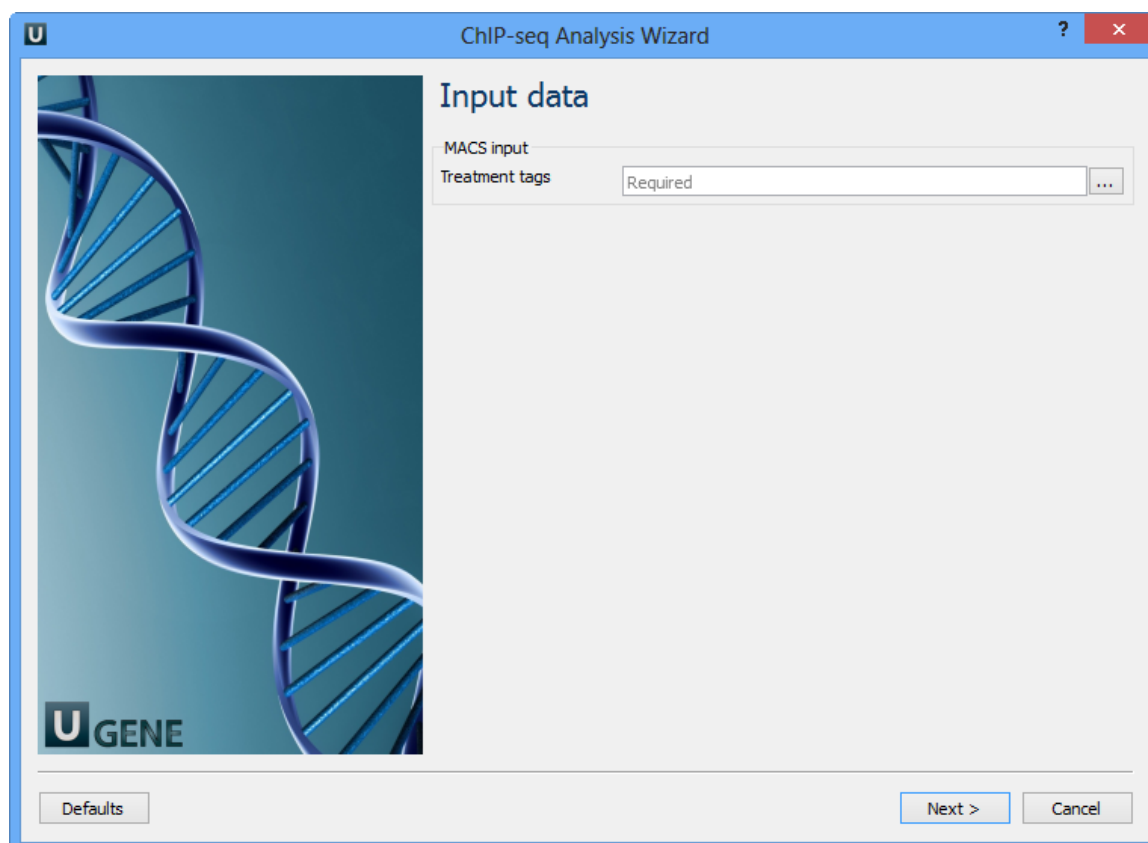




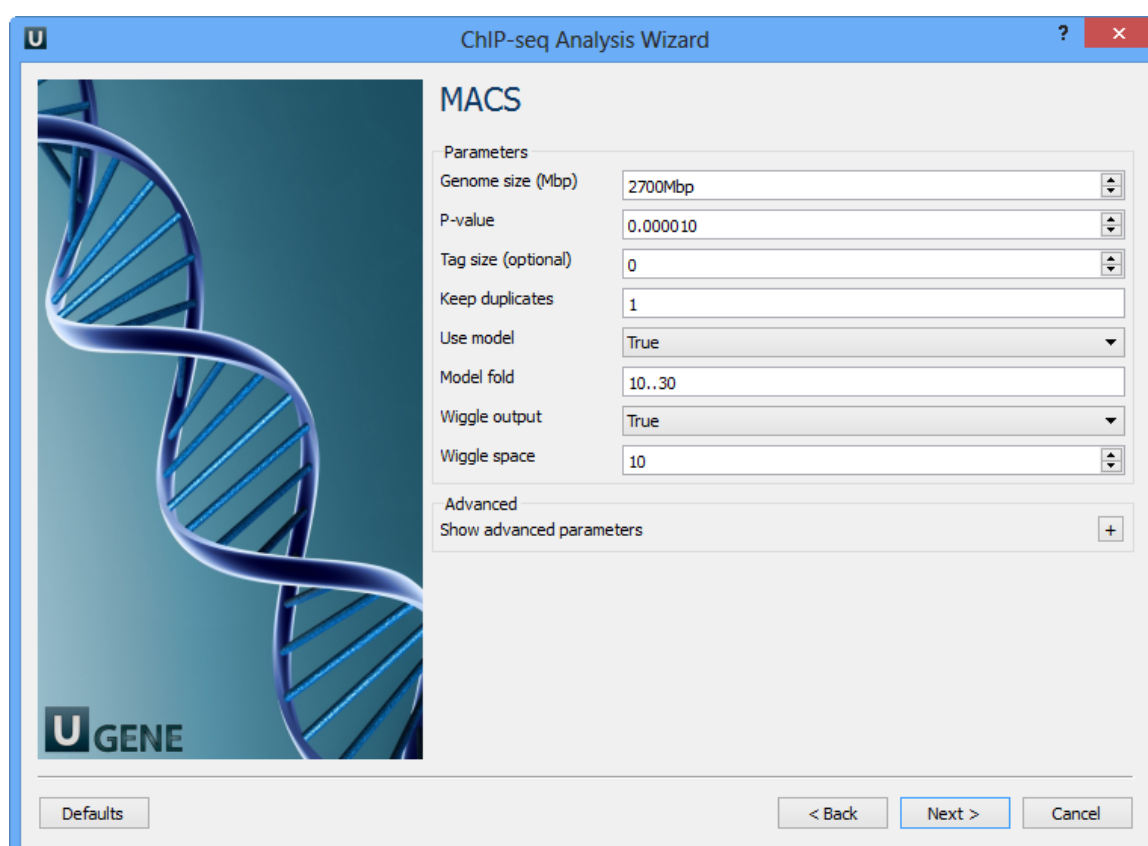
To run these workflows you need to select input annotations, and output files and directories. Also, if required, you can change parameters of MACS, CEAS, Conservation Plot, SeqPos, Peak2Gene, and Gene Ontology. Use the workflow wizard to guide you through the parameters setup process. The first wizard page appears automatically after the Setup button has been pressed or click Show wizard button on the Workflow Designer toolbar to open it:



The first wizard page:



Here you need to input a file with treatment annotations for MACS. Select a file and click Next. The next wizard page allows you to configure MACS parameters:



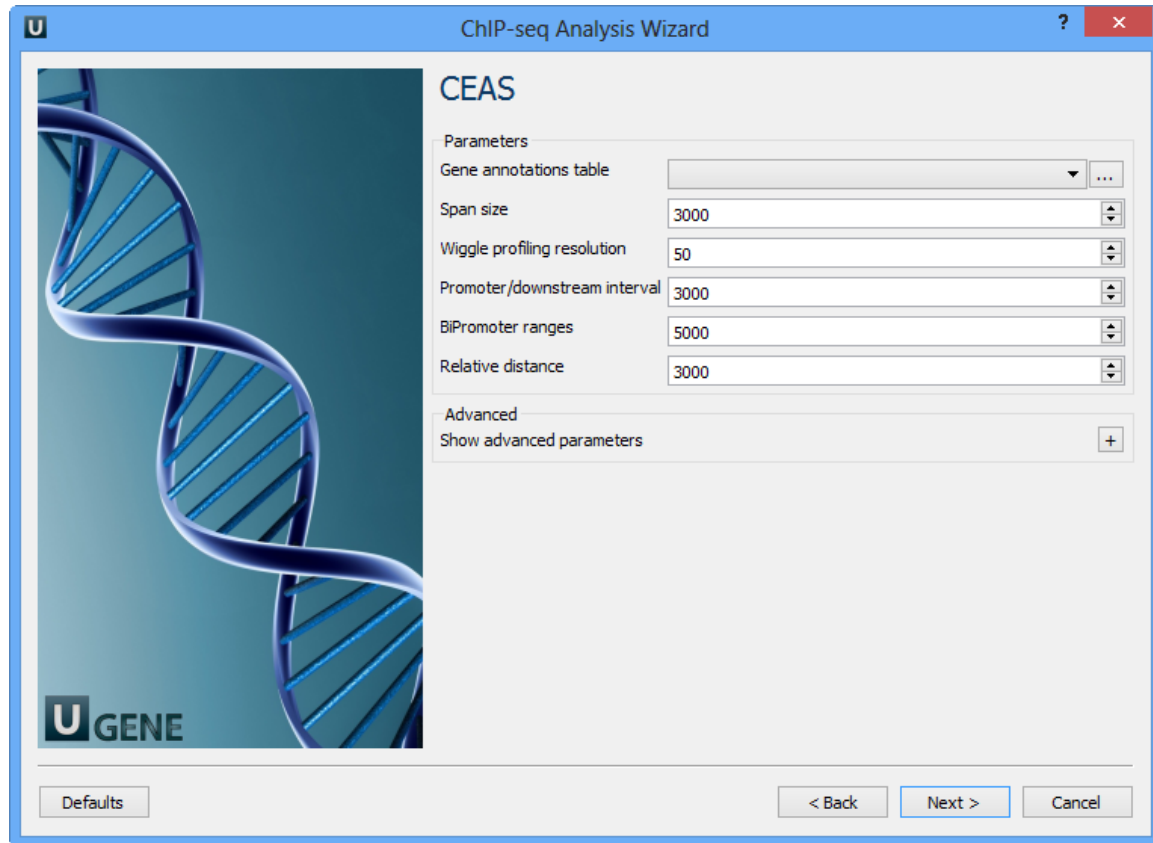
The following parameters are available:

Genome size (Mbp)	<p>Homo sapience - 2700 Mbp</p> <p>Mus musculus - 1870 Mbp</p> <p>Caenorhabditis elegans - 90 Mbp</p> <p>Drosophila melanogaster - 120 Mbp</p> <p>It's the mappable genome size or effective genome size which is defined as the genome size which can be sequenced. Because of the repetitive features on the chromosomes, the actual mappable genome size will be smaller than the original size, about 90% or 70% of the genome size.</p>
P-value	P-value cutoff. Default is 0.00001, for looser results, try 0.001 instead.
Tag size (optional)	Length of reads. Determined from first 10 reads if not specified (input 0).
Keep duplicates	It controls the MACS behavior towards duplicate tags at the exact same location -- the same coordination and the same strand. The default auto option makes MACS calculate the maximum tags at the exact same location based on binomal distribution using 1e-5 as pvalue cutoff; and the all option keeps every tags. If an integer is given, at most this number of tags will be kept at the same location.
Use model	Whether or not to use MACS paired peaks model.
Model fold	Select the regions within MFOLD range of high-confidence enrichment ratio against. Model fold is available when Use Model is true, which is the foldchange to chose paired peaks to build paired peaks model. Users need to set a lower(smaller) and upper(larger) number for fold change so that MACS will only use the peaks within these foldchange range to build model.
Wiggle output	If this flag is on, MACS will store the fragment pileup in wiggle format for the whole genome data instead of for every chromosomes.
Wiggle space	By default, the resolution for saving wiggle files is 10 bps, i.e., MACS will save the raw tag count every 10 bps. You can change it along with Wiggle output parameter.
Shift size	An arbitrary shift value used as a half of the fragment size when model is not built. Shift size is available when Use Model is false, which will represent the HALF of the fragment size of your sample. If your sonication and size selection size is 300 bps, after you trim out nearly 100 bps adapters, the fragment size is about 200 bps, so you can specify 100 here.
Band width	The band width which is used to scan the genome for model building. You can set this parameter as the sonication fragment size expected from wet experiment. Used only while building the shifting model.
Use lambda	Whether to use local lambda model which can use the local bias at peak regions to throw out false positives.
Small nearby region	The small nearby region in basepairs to calculate dynamic lambda. This is used to capture the bias near the peak summit region. Invalid if there is no control data.
Auto bimodal	Whether turn on the auto pair model process.If set, when MACS failed to build paired model, it will use the nomodelsettings, the Shift size parameter to shift and extend each tags.

Scale to large

When set, scale the small sample up to the bigger sample. By default, the bigger dataset will be scaled down towards the smaller dataset, which will lead to smaller p/q-values and more specific results. Keep in mind that scaling down will bring down background noise more.

Configure the parameters, if required, and click Next. The next page appears:

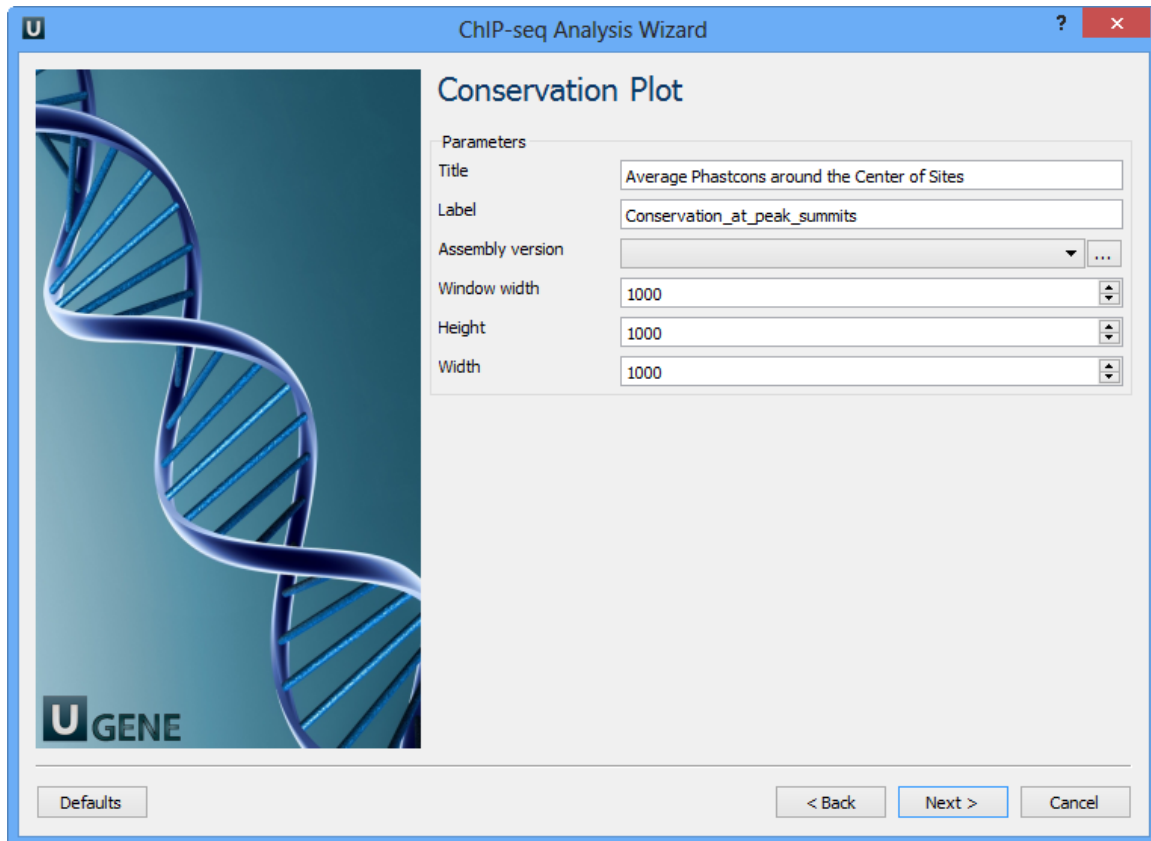


On this page you can configure CEAS parameters:

Gene annotations table	Path to gene annotation table (e.g. a refGene table in sqlite3 db format).
Span size	Span from TSS and TTS in the gene-centered annotation (base pairs). ChIP regions within this range from TSS and TTS are considered when calculating the coverage rates in promoter and downstream.
Wiggle profiling resolution	Wiggle profiling resolution. WARNING: Value smaller than the wig interval (resolution) may cause aliasing error.
Promoter/downstream interval	Promoter/downstream intervals for ChIP region annotation are three values or a single value can be given. If a single value is given, it will be segmented into three equal fractions (e.g. 3000 is equivalent to 1000,2000,3000).
BiPromoter ranges	Bidirectional-promoter sizes for ChIP region annotation. It's two values or a single value can be given. If a single value is given, it will be segmented into two equal fractions (e.g. 5000 is equivalent to 2500,5000).
Relative distance	Relative distance to TSS/TTS in WIGGLE file profiling.
Gene group files	Gene groups of particular interest in wig profiling. Each gene group file must have gene names in the 1st column. The file names are separated by commas.

Gene group names	<p>Set this parameter empty for using default values.</p> <p>The names of the gene groups from "Gene group files" parameter. These names appear in the legends of the wig profiling plots.</p> <p>Values range: comma-separated list of strings. Default value: 'Group 1, Group 2,...Group n'.</p>
------------------	--

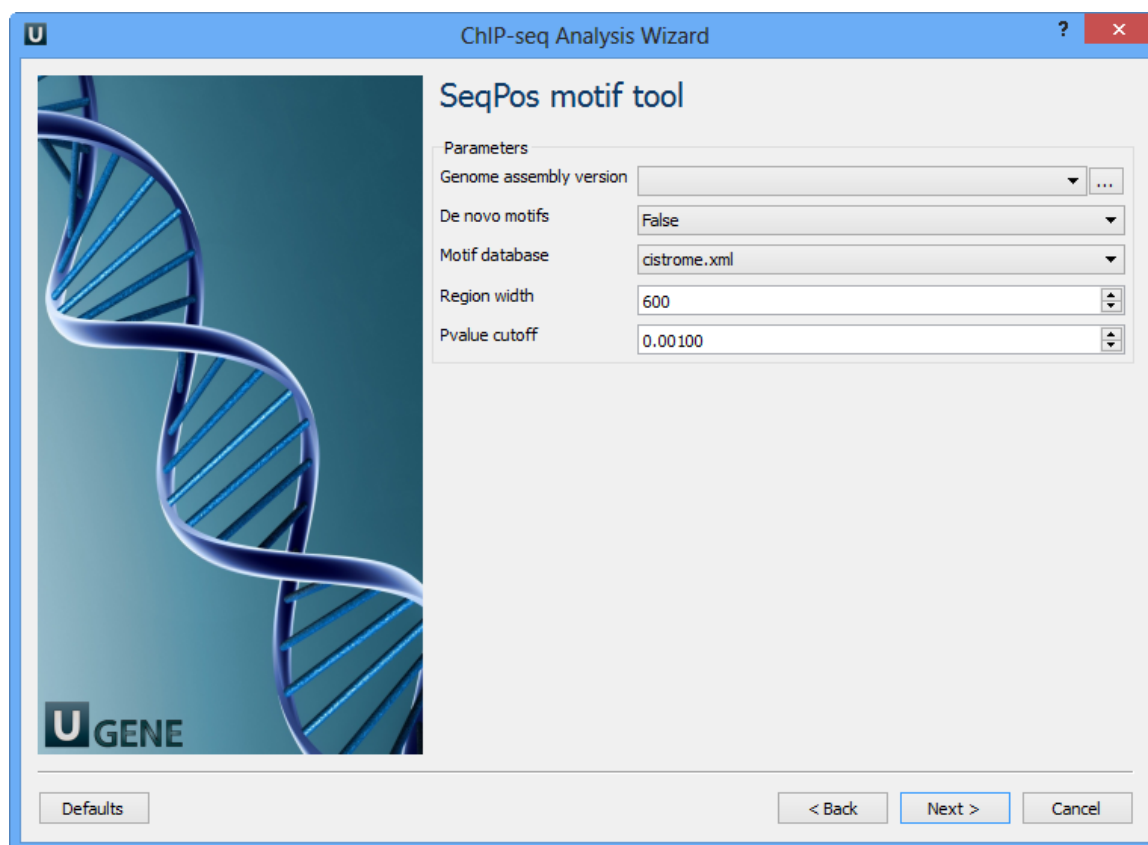
Click Next. The next page allows you to configure Conservation Plot parameters:



The following parameters are available:

Title	Title of the figure.
Label	Label of data in the figure.
Assembly version	The directory to store phastcons scores.
Window width	Window width centered at middle of regions.
Height	Height of plot.
Width	Width of plot.

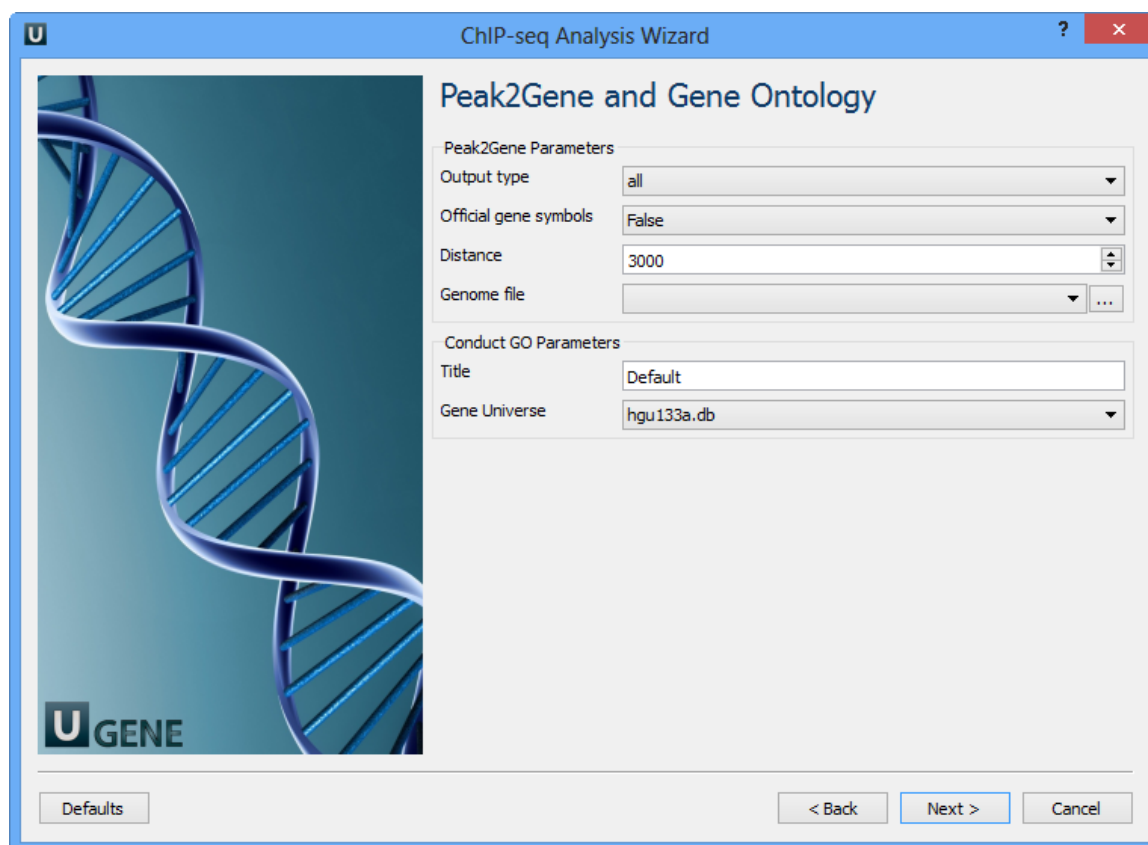
Optionally choose these parameters and click Next. The next page contains SeqPos motif parameters:



The following parameters are available:

Genome assembly version	UCSC database version.
De novo motifs	Run de novo motif search.
Motif database	Known motif collections.
Region width	Width of the region to be scanned for motifs; depends on a resolution of assay.
Pvalue cutoff	Pvalue cutoff for the motif significance.

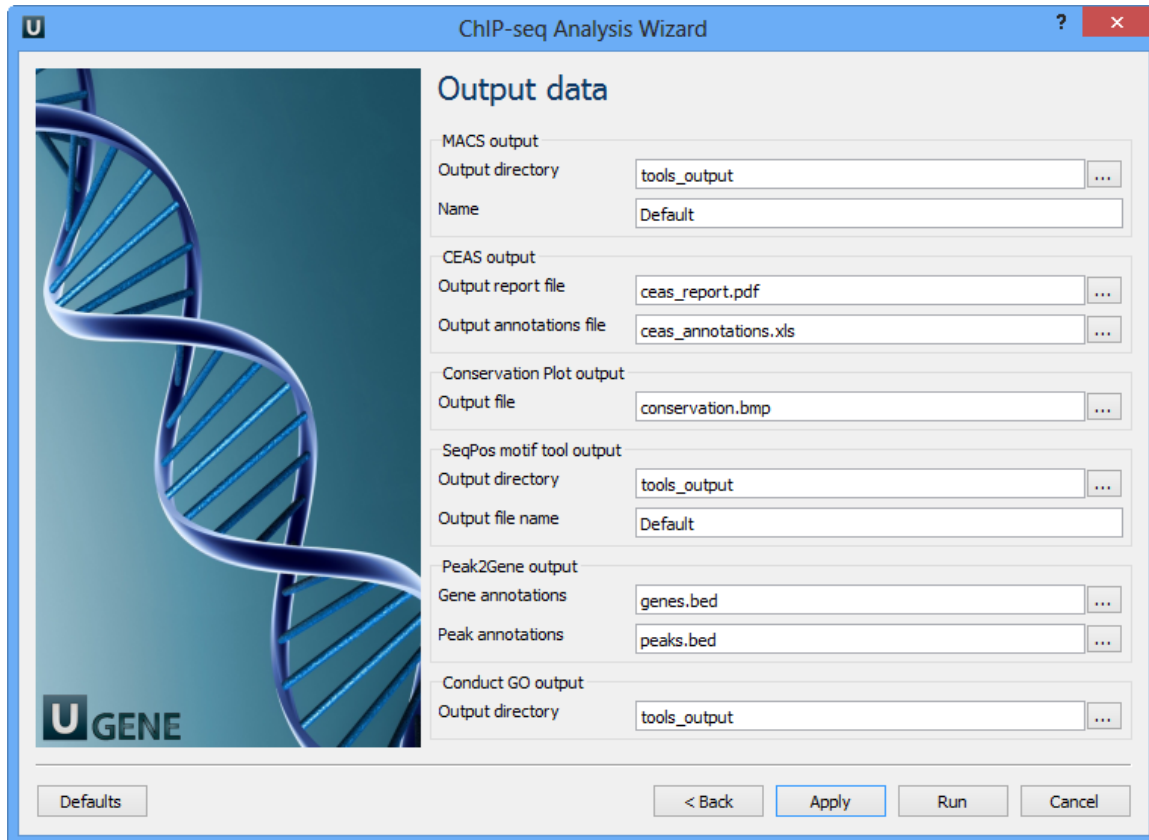
Optionally, modify these parameters and click Next. The next page contain Peak2Gene and Gene Ontology parameters:



You can configure the following parameters:

Output type	The directory to store Conduct GO results.
Official gene symbols	Output official gene symbol instead of refseq name.
Distance	Set a number which unit is base. It will get the refGenes in n bases from peak center.
Genome file	Select a genome file (sqlite3 file) to search refGenes.
Title	Title is used to name the output files - so make it meaningful.
Gene Universe	Select a gene universe.

The last wizard page:



**ChIP-seq Analysis Wizard**

**Output data**

MACS output

Output directory: tools\_output

Name: Default

CEAS output

Output report file: ceas\_report.pdf

Output annotations file: ceas\_annotations.xls

Conservation Plot output

Output file: conservation.bmp

SeqPos motif tool output

Output directory: tools\_output

Output file name: Default

Peak2Gene output

Gene annotations: genes.bed

Peak annotations: peaks.bed

Conduct GO output

Output directory: tools\_output

Defaults < Back Apply Run Cancel

Here you need to input output files and directories for all tools.

#### MACS output:

Output directory	Directory to save MACS output files.
Name	Name string of the experiment. MACS will use this string NAME to create output files like 'NAME_peaks.xls', 'NAME_negative_peaks.xls', 'NAME_peaks.bed', 'NAME_summits.bed', 'NAME_model.r' and so on. So please avoid any confliction between these filenames and your existing files.

#### CEAS output:

Output report file	Path to the report output file. Result for the CEAS analysis.
Output annotations file	Name of tab-delimited output text file, containing a row of annotations for every RefSeq gene. Note that the file is not generated if there is no peak regions input.

#### Conservation Plot output:

Output file	File to store phastcons results (BMP).
-------------	--

#### SeqPos motif tool output:

Output directory	Directory to store seqpos results.
------------------	------------------------------------



Output file name	Name of the output file which stores new motifs found during a de novo search.
------------------	--

#### Peak2Gene output:

Gene annotations	Location of peak2gene gene annotations data file.
Peak annotations	Location of peak2gene peak annotations data file.

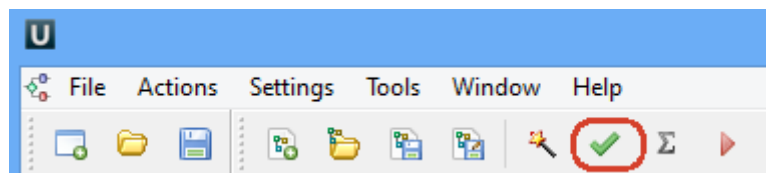
#### Conduct GO output:

Output directory	Directory to store Conduct GO results.
------------------	--

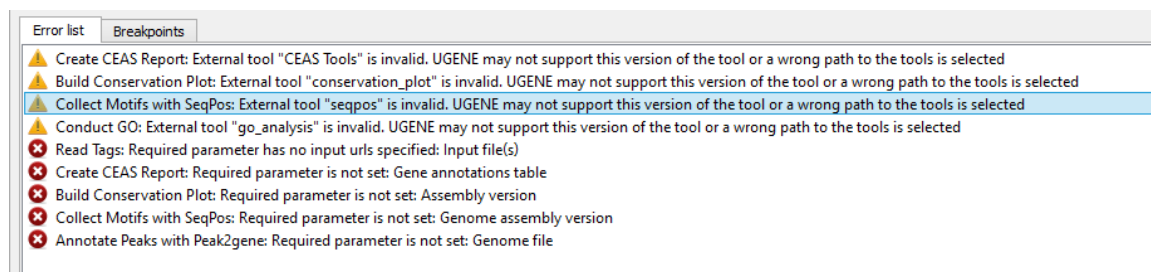
Choose these output directories click on the Finish button.

Note that default button reverts all parameters to default settings.

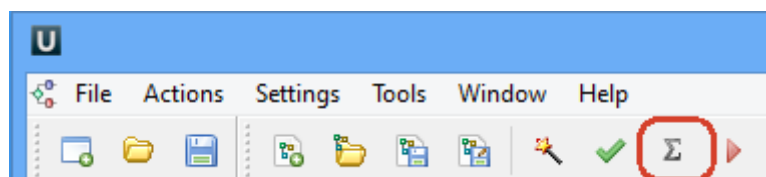
Now let's validate and run the workflow. To validate that the workflow is correct and all parameters are set properly click the Validate workflow button on the Workflow Designer toolbar:



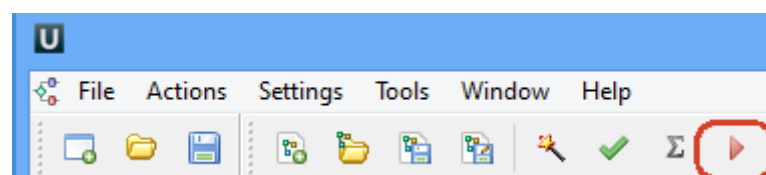
If there are some errors, they will be shown in the Error list at the bottom of the Workflow Designer window, for example:



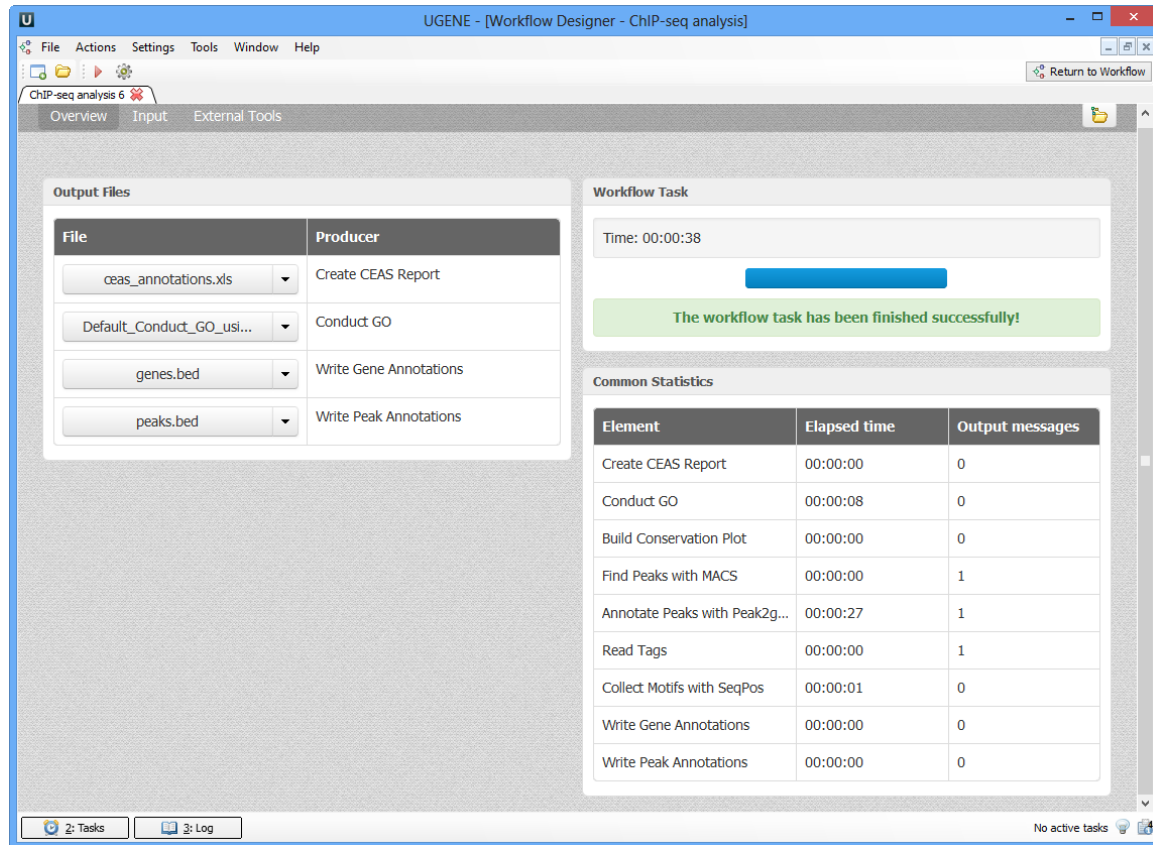
However, if you have set all the required parameters, then there shouldn't be errors. After that you can estimate the workflow. To run estimation click the *Estimate workflow* button:



To run a valid workflow, click the Run workflow button on the Workflow Designer toolbar:



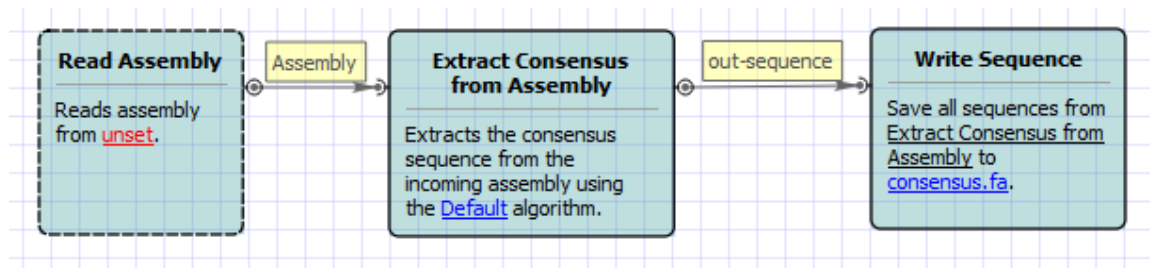
As soon as the variants calling task is finished, a notification and dashboard will appear.



The dashboard will contain information about workflow: input and output files, all information about task.

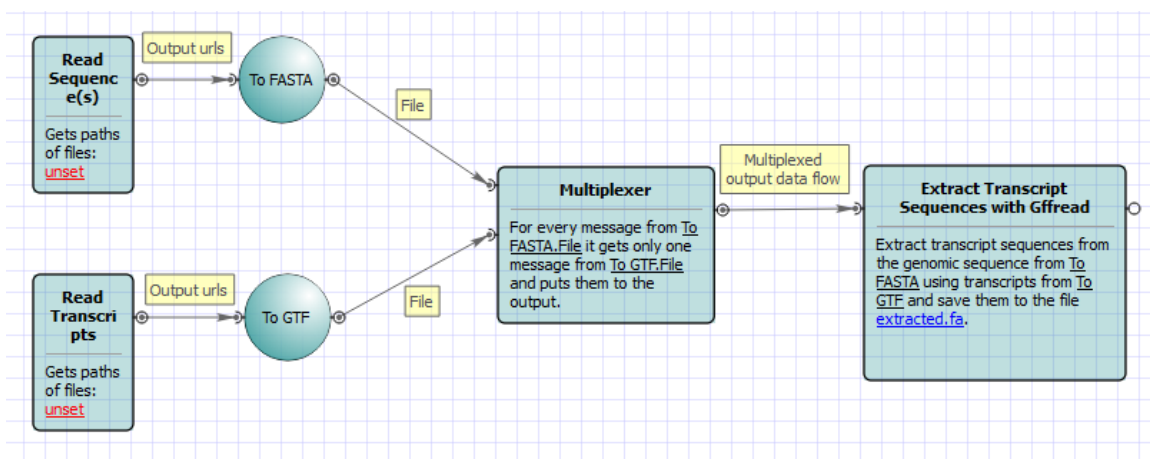
## Extract Consensus

Uses input assemblies to extract the consensus sequences and save them to a FASTA.



## Extract transcript sequences

This workflow uses input transcripts and genomic sequences to generate a FASTA file with the DNA sequences for the transcripts. Please make sure that contig or chromosome names in the transcript file(s) have corresponding entries in the input sequence(s).



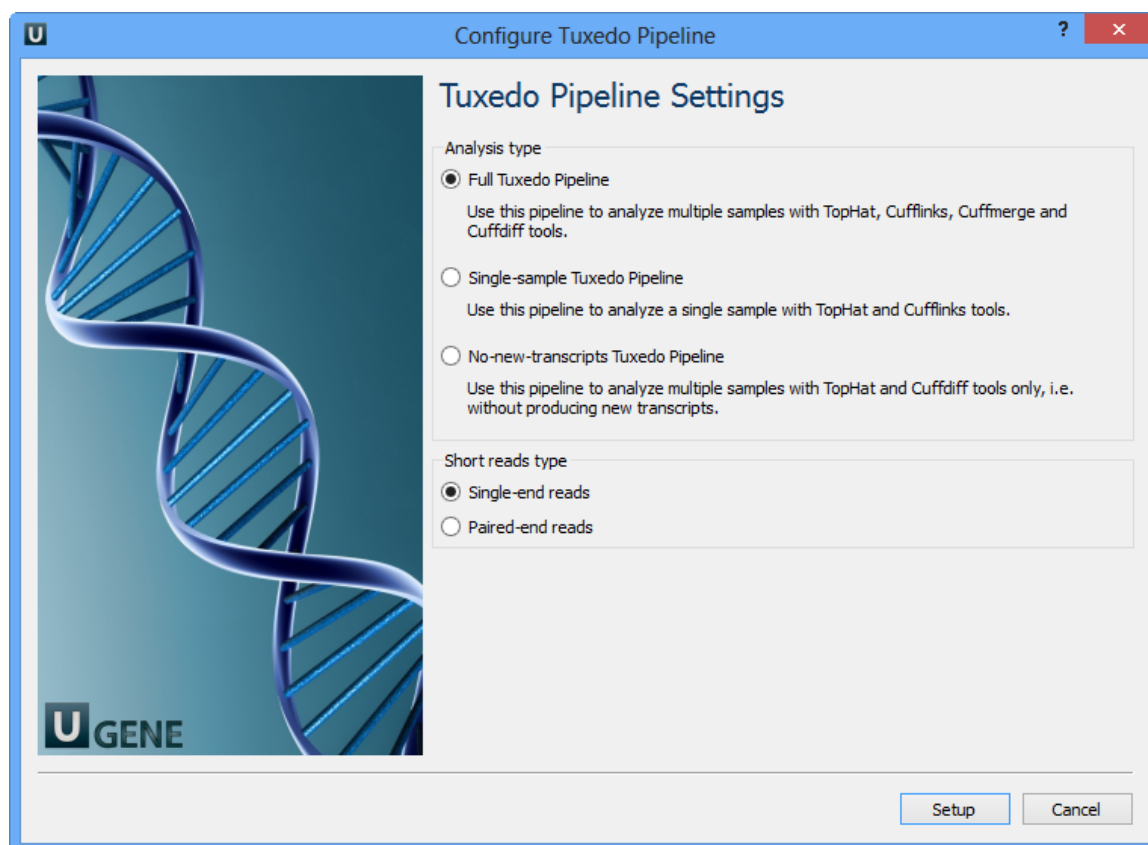
## RNA-seq analysis with Tuxedo tools

The RNA-seq pipeline “Tuxedo” consists of the **TopHat** spliced read mapper, that internally uses **Bowtie** or **Bowtie 2** short read aligners, and several **Cufflinks** tools that allows one to assemble transcripts, estimate their abundances, and tests for differential expression and regulation in RNA-Seq samples.

**Environment requirements:**

- The pipeline is currently available on Linux and Mac OS X only.

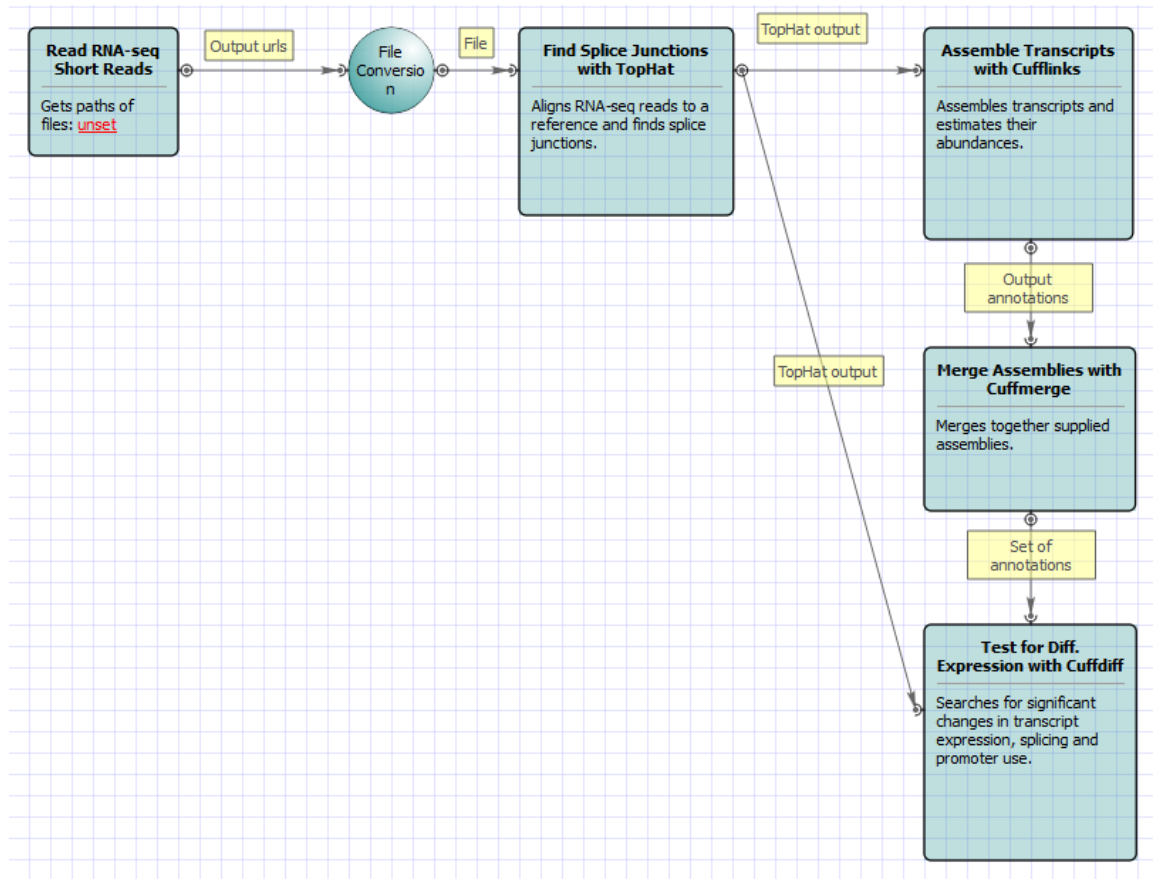
Select Samples tab on the Workflow Designer Palette and double-click on the ChIP-seq analysis with Cistrome tools sample. The following configure wizard appears:



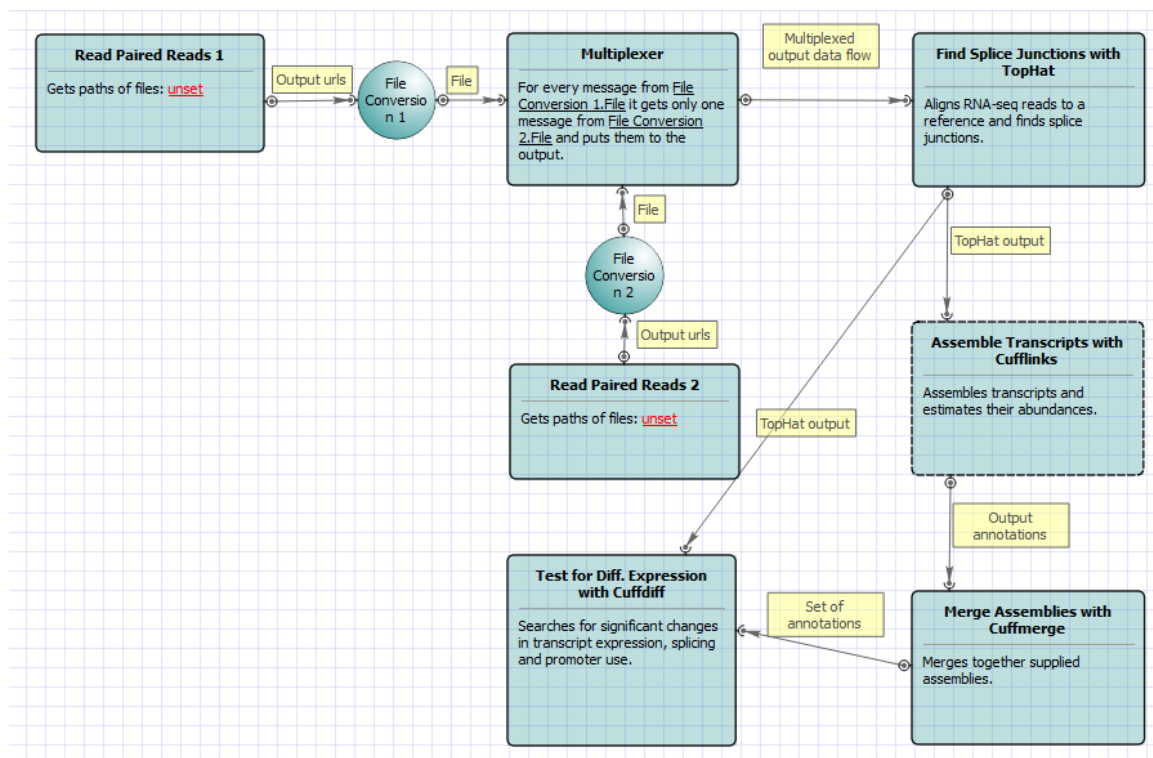
Here you need to choose analysis type and short reads type and click Setup. There are two short reads type: single-end and paired-end reads. For both of them there are three analysis type:

1. Full Tuxedo Pipeline - use this pipeline to analyze multiple samples with TopHat, Cufflinks, Cuffmerge and Cuffdiff tools.
2. Single-sample Tuxedo Pipeline - use this pipeline to analyze a single sample with TopHat and Cufflinks tools.
3. No-new-transcripts Tuxedo Pipeline - use this pipeline to analyze multiple samples with TopHat and Cuffdiff tools only, i.e. without producing new transcripts.

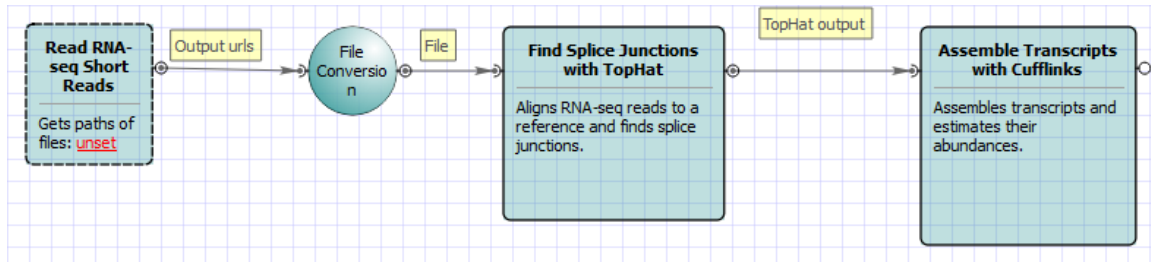
For **Full Tuxedo Pipeline** analysis type and **single-end reads** type the following workflow appears:



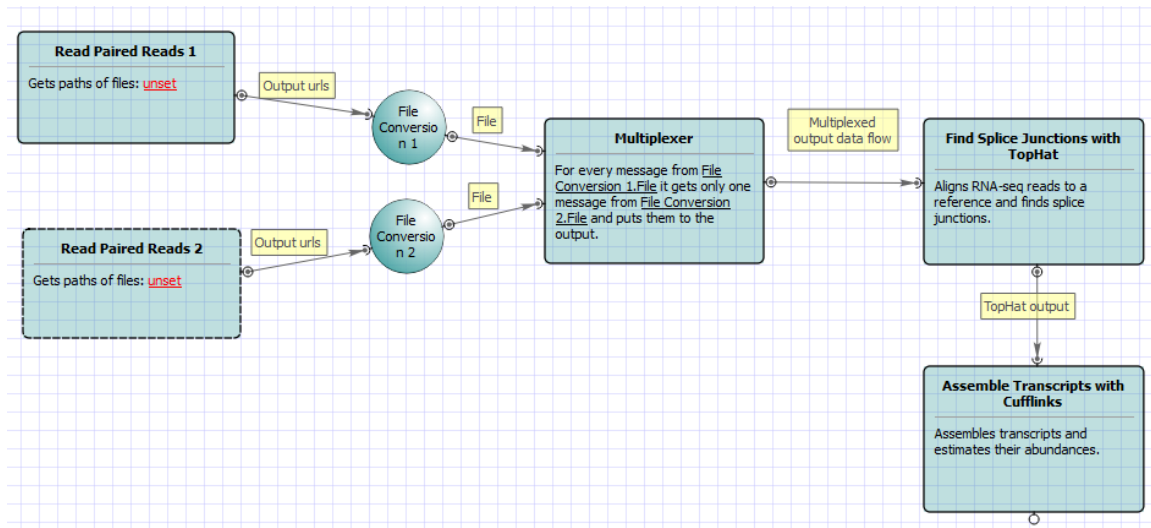
For **Full Tuxedo Pipeline** analysis type and **paired-end reads** type the following workflow appears:



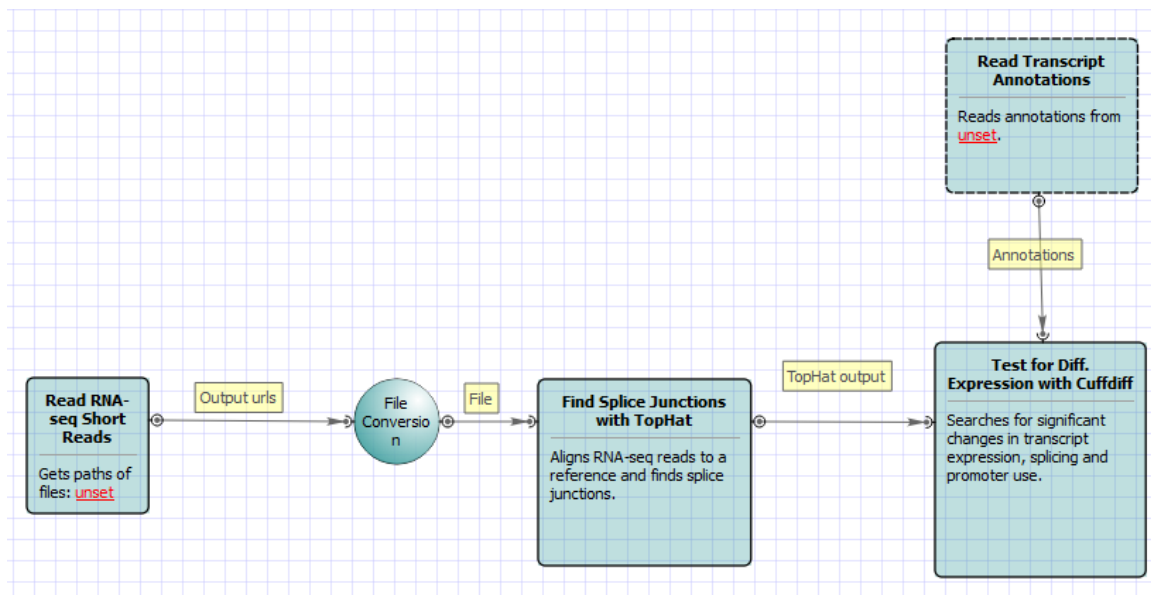
For **Single-sample Tuxedo Pipeline** analysis type and **single-end reads** type the following workflow appears:



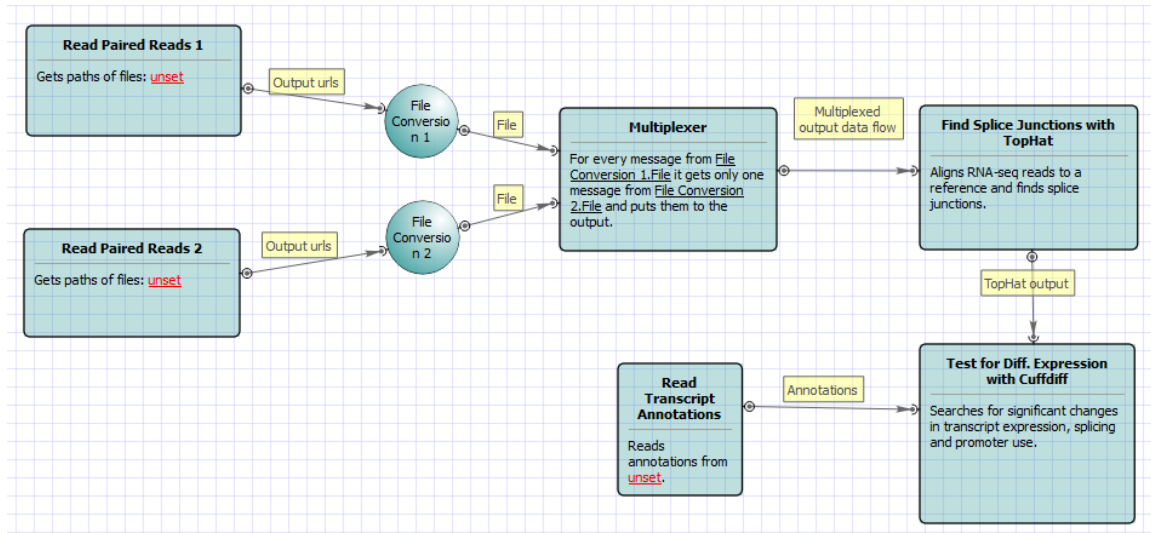
For **Single-sample Tuxedo Pipeline** analysis type and **paired-end reads** type the following workflow appears:



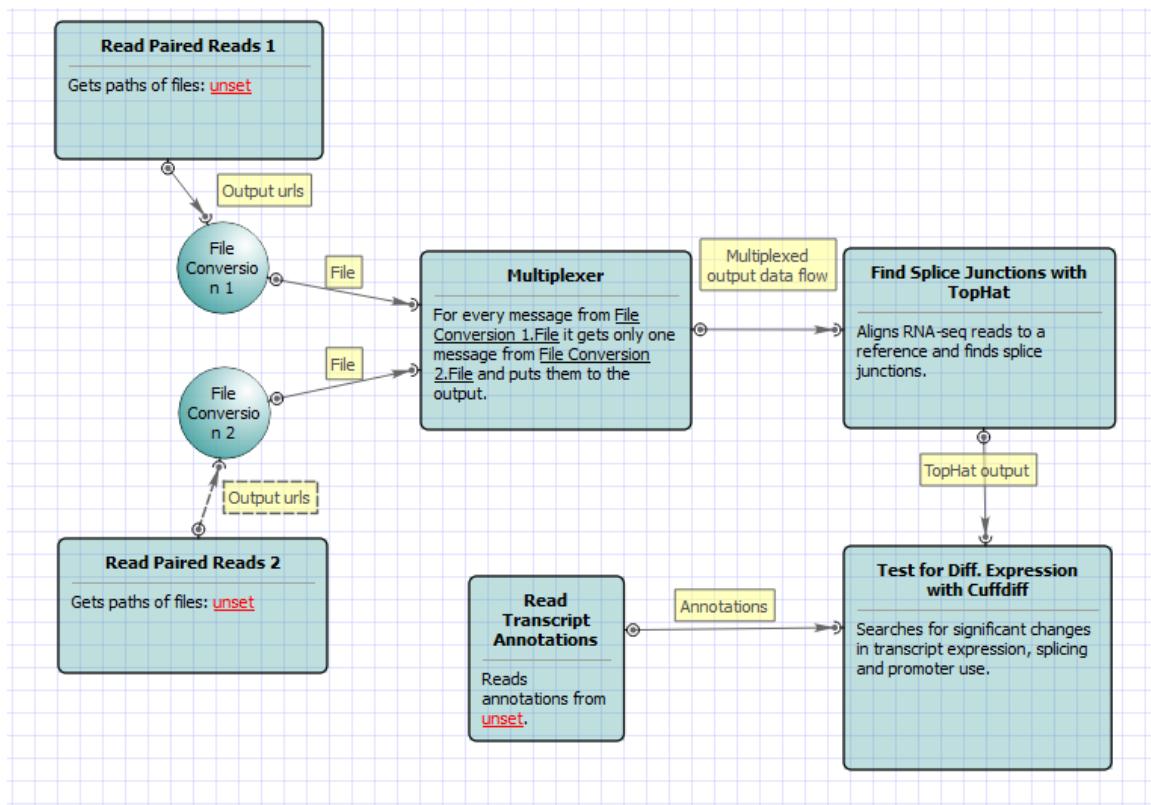
For **No-new-transcripts Tuxedo Pipeline** analysis type and **single-end reads** type the following workflow appears:



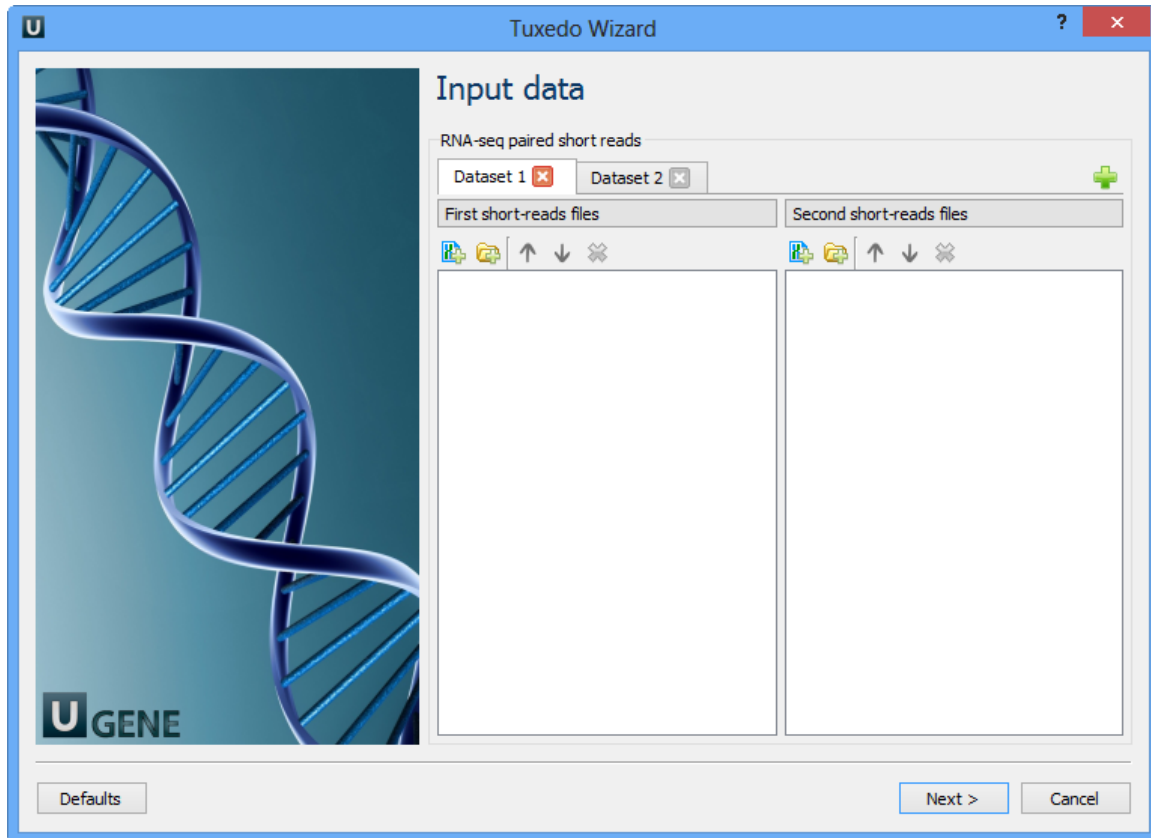
For **No-new-transcripts Tuxedo Pipeline** analysis type and **paired-end reads** type the following workflow appears:



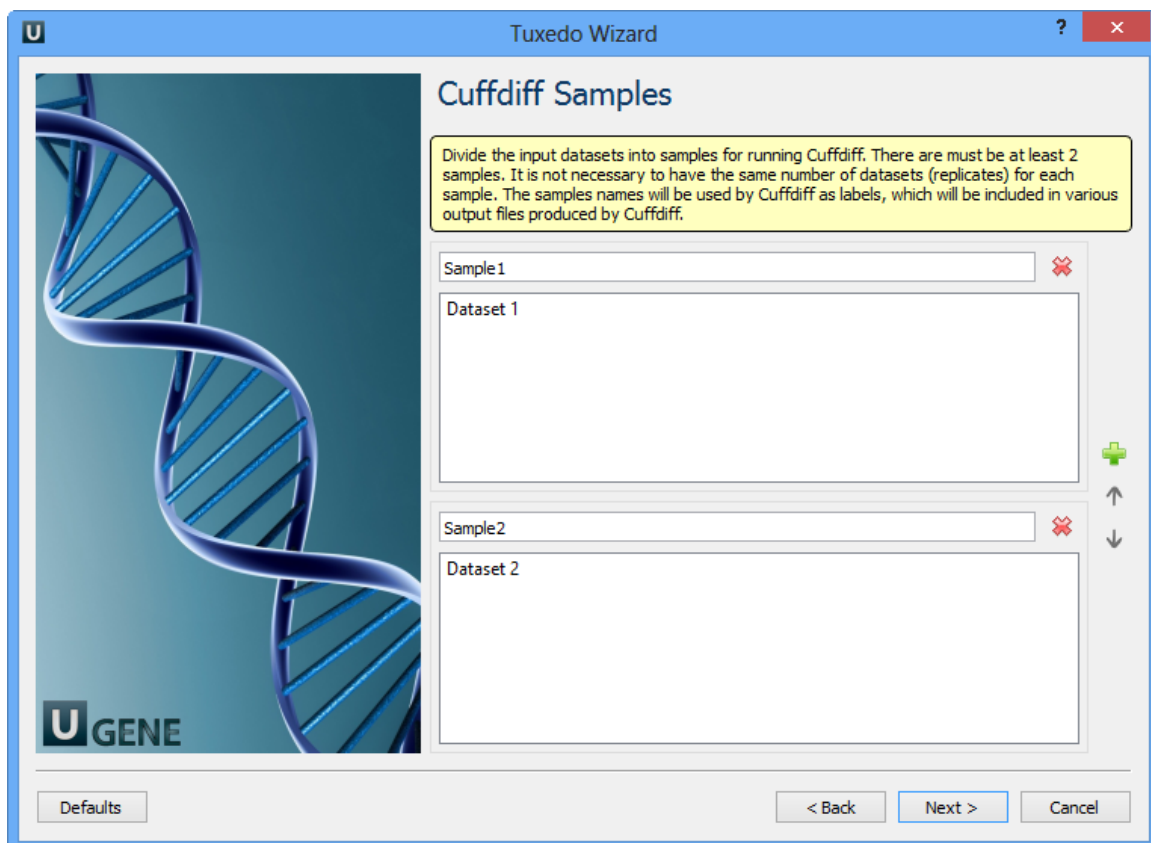
Use the workflow wizard to guide you through the parameters setup process. Click Show wizard button on the Workflow Designer toolbar to open it:



All of these workflows have the similar wizards. For **Full Tuxedo Pipeline** analysis type and **paired-end reads** type the following first wizard page appears:

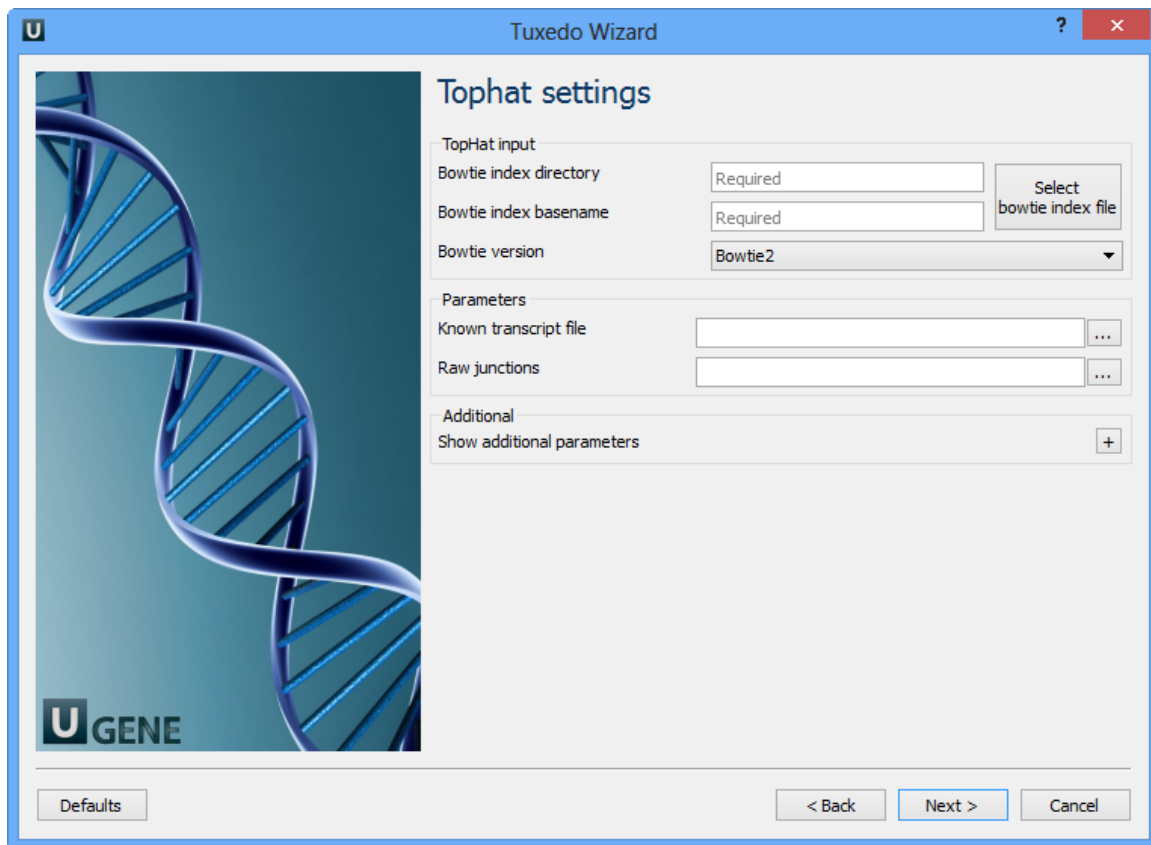


Here you need to input RNA-seq short reads in FASTA or FASTQ formats. Many datasets with different reads can be added. Click the Next button. The next page appears:



Here you need to divide the input datasets into samples for running Cuffdiff. There are must be at least 2 samples. It is not necessary to have the same number of datasets (replicates) for each sample. The samples names will be used by Cuffdiff as labels, which will be included in various output files produced by Cuffdiff. Click the Next button. The next page appears:





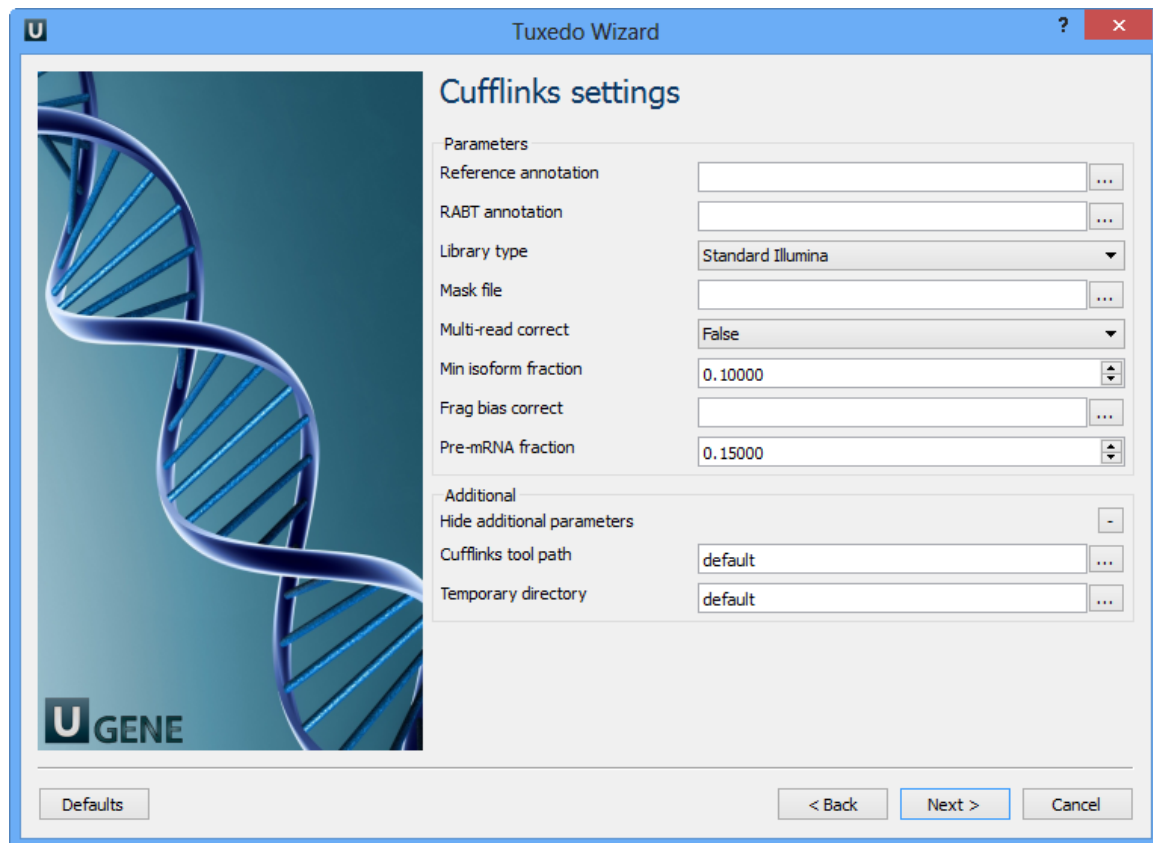
Here you can configure TopHat settings. To show additional parameters click on the + button. The following parameters are available:

Bowtie index directory	The directory with the Bowtie index for the reference sequence.
Bowtie index basename	The basename of the Bowtie index for the reference sequence.
Bowtie version	Specifies which Bowtie version should be used.
Known transcript file	A set of gene model annotations and/or known transcripts.
Raw junctions	The list of raw junctions.
Mate inner distance	Expected (mean) inner distance between mate pairs.
Mate standard deviation	Standard deviation for the distribution on inner distances between mate pairs.
Library type	Specifies RNA-seq protocol.
No novel junctions	Only look for reads across junctions indicated in the supplied GFF or junctions file. This parameter is ignored if Raw junctions or Known transcript file is not set.
Max multihints	Instructs TopHat to allow up to this many alignments to the reference for a given read, and suppresses all alignments for reads with more than this many alignments.
Segment length	Each read is cut up into segments, each at least this long. These segments are mapped independently.
Fusion search	Turn on fusion mapping.
Transcriptome max hits	Only align the reads to the transcriptome and report only those mappings as genomic mappings.



Prefilter multihints	When mapping reads on the transcriptome, some repetitive or low complexity reads that would be discarded in the context of the genome may appear to align to the transcript sequences and thus may end up reported as mapped to those genes only. This option directs TopHat to first align the reads to the whole genome in order to determine and exclude such multi-mapped reads (according to the value of the Max multihits option).
Min anchor length	The anchor length. TopHat will report junctions spanned by reads with at least this many bases on each side of the junction. Note that individual spliced alignments may span a junction with fewer than this many bases on one side. However, every junction involved in spliced alignments is supported by at least one read with this many bases on each side.
Splice mismatches	The maximum number of mismatches that may appear in the anchor region of a spliced alignment.
Read mismatches	Final read alignments having more than these many mismatches are discarded.
Segment mismatches	Read segments are mapped independently, allowing up to this many mismatches in each segment alignment.
Solexa 1.3 quals	As of the Illumina GA pipeline version 1.3, quality scores are encoded in Phred-scaled base-64. Use this option for FASTQ files from pipeline 1.3 or later.
Bowtie version	specifies which Bowtie version should be used.
Bowtie -n mode	TopHat uses -v in Bowtie for initial read mapping (the default), but with this option, -n is used instead. Read segments are always mapped using -v option.
Bowtie tool path	The path to the Bowtie external tool.
SAMtools tool path	The path to the SAMtools tool. Note that the tool is available in the UGENE External Tool Package.
TopHat tool path	The path to the TopHat external tool in UGENE.
Temporary directory	The directory for temporary files.

Choose these parameters and click the Next button. The next page allows one to configure Cufflinks settings:

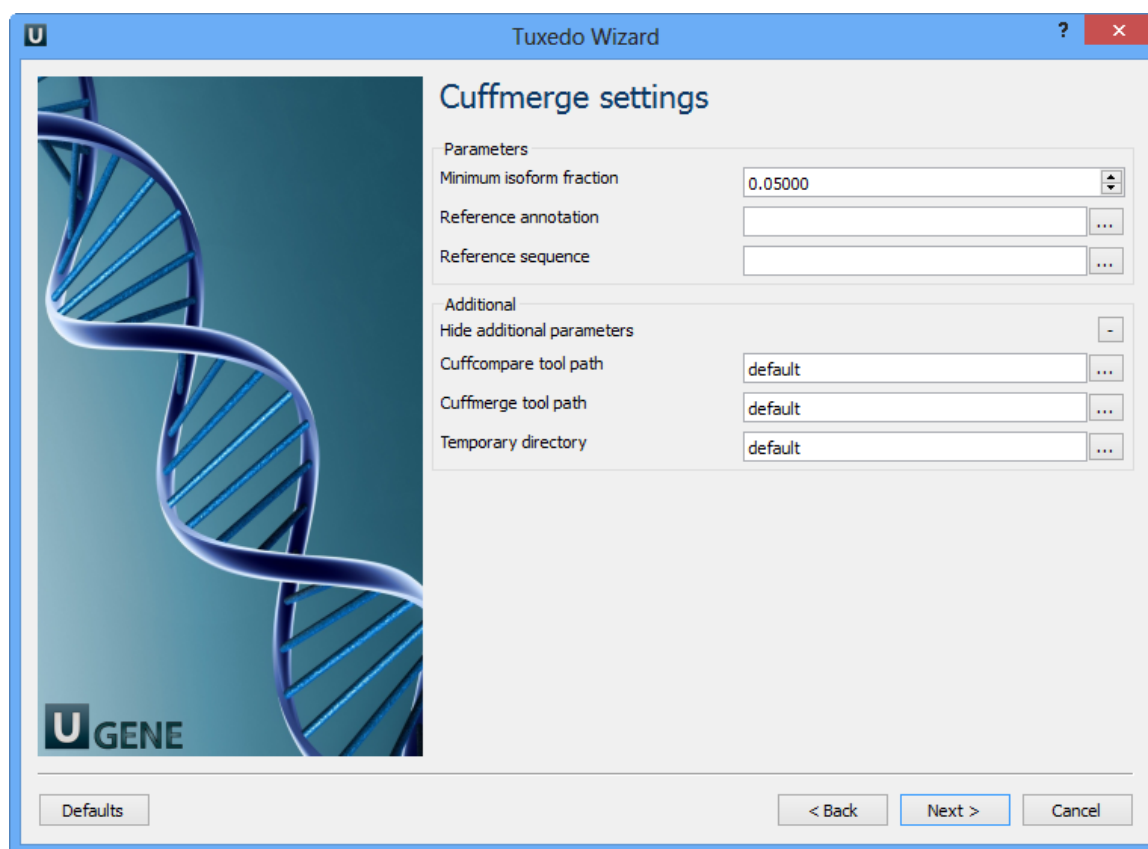


The following parameters are available:

Reference annotation	Tells Cufflinks to use the supplied reference annotation to estimate isoform expression. Cufflinks will not assemble novel transcripts and the program will ignore alignments not structurally compatible with any reference transcript.
RABT annotation	Tells Cufflinks to use the supplied reference annotation to guide Reference Annotation Based Transcript (RABT) assembly. Reference transcripts will be tiled with faux-reads to provide additional information in assembly. Output will include all reference transcripts as well as any novel genes and isoforms that are assembled.
Library type	Specifies RNA-seq protocol.
Mask file	Ignore all reads that could have come from transcripts in this file. It is recommended to include any annotated rRNA, mitochondrial transcripts other abundant transcripts you wish to ignore in your analysis in this file. Due to variable efficiency of mRNA enrichment methods and rRNA depletion kits, masking these transcripts often improves the overall robustness of transcript abundance estimates.
Multi-read correct	Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.
Min isoform fraction	After calculating isoform abundance for a gene, Cufflinks filters out transcripts that it believes are very low abundance, because isoforms expressed at extremely low levels often cannot reliably be assembled, and may even be artifacts of incompletely spliced precursors of processed transcripts. This parameter is also used to filter out introns that have far fewer spliced alignments supporting them.

Frag bias correct	Providing Cufflinks with a multifasta file via this option instructs it to run the bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates.
Pre-mRNA fraction	Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored.
Cufflinks tool path	The path to the Cufflinks external tool in UGENE.
Temporary directory	The directory for temporary files.

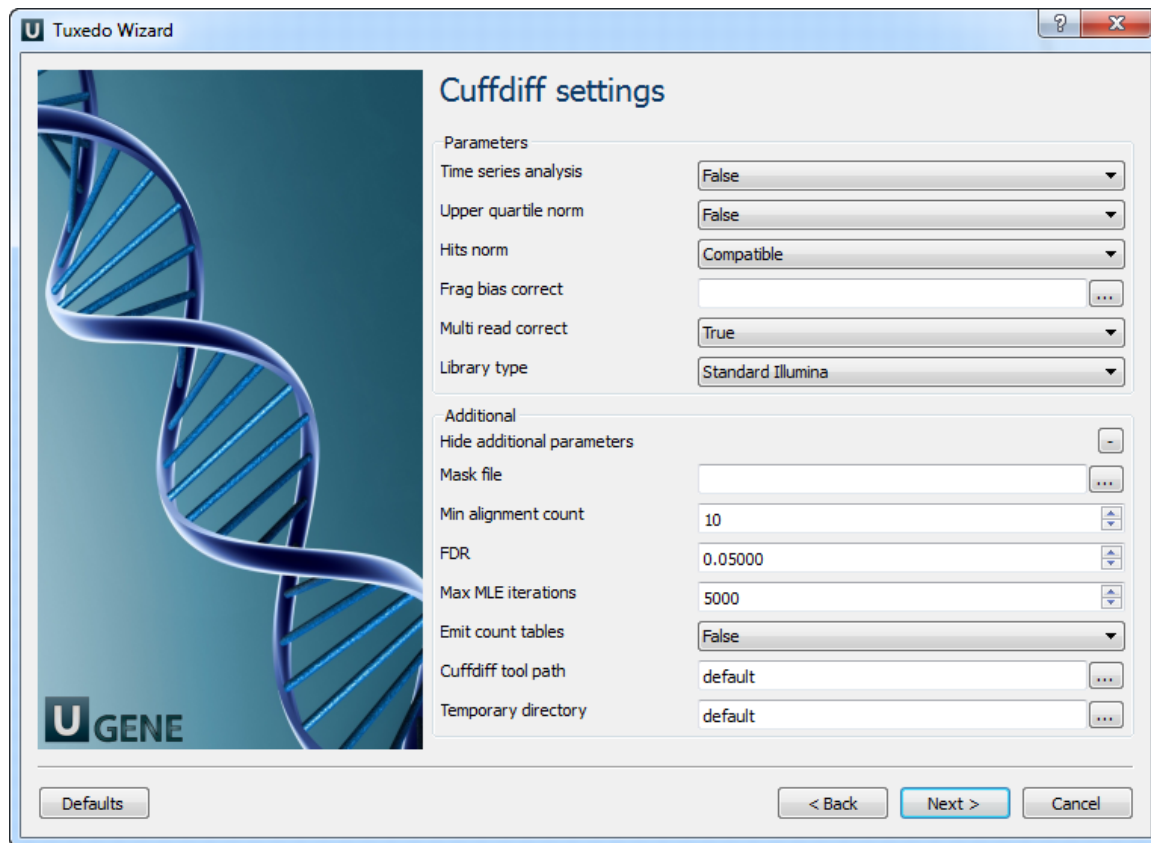
Configure parameters, if necessary, and click Next. On the next page you may configure Cuffmerge settings:



The following parameters are available:

Minimum isoform fraction	Discard isoforms with abundance below this.
Reference annotation	Merge the input assemblies together with this reference annotation.
Reference sequence	The genomic DNA sequences for the reference. It is used to assist in classifying transfrags and excluding artifacts (e.g. repeats). For example, transcripts consisting mostly of lower-case bases are classified as repeats.
Cuffcompare tool path	The path to the Cuffcompare external tool in UGENE.
Cuffmerge tool path	The path to the Cuffmerge external tool in UGENE.
Temporary directory	The directory for temporary files.

Configure parameters, if necessary, and click Next. On the next page you may configure Cuffdiff settings:

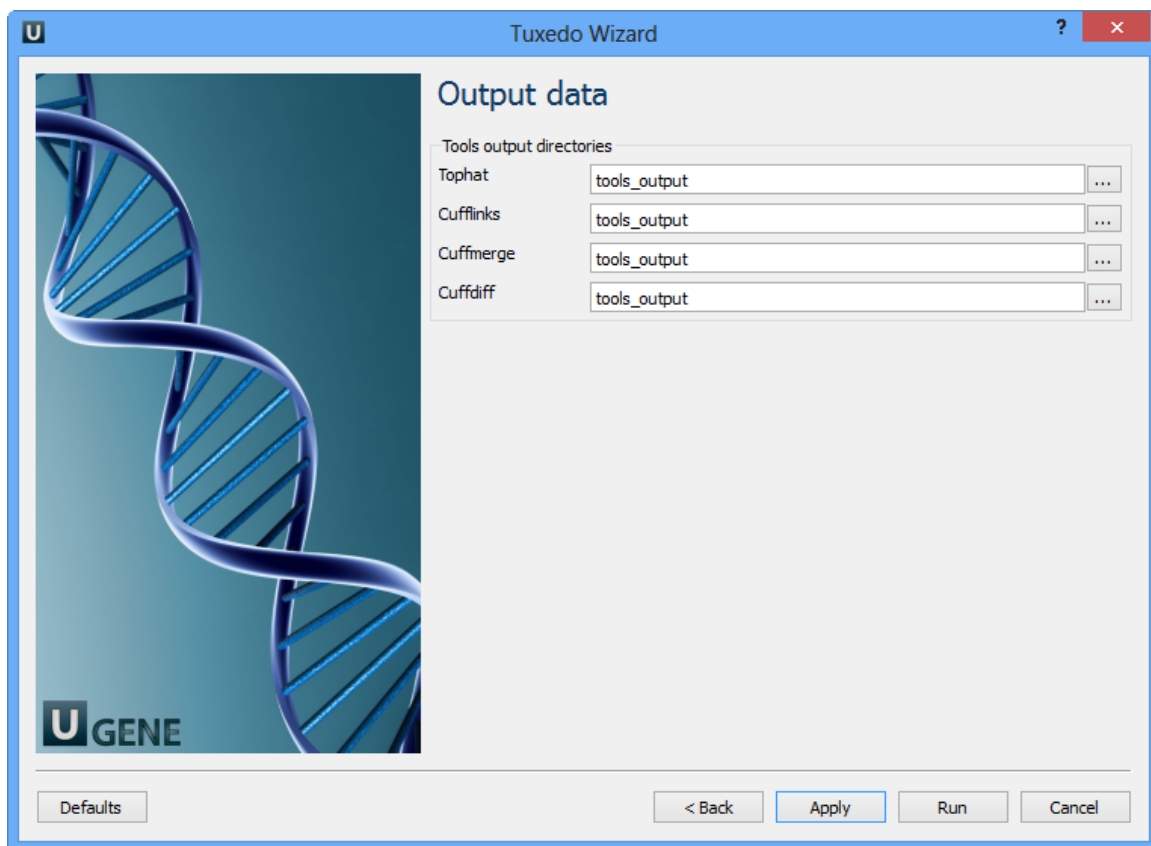


The following parameters are available:

Time series analysis	If set to True, instructs Cuffdiff to analyze the provided samples as a time series, rather than testing for differences between all pairs of samples. Samples should be provided in increasing time order.
Upper quartile norm	If set to True, normalizes by the upper quartile of the number of fragments mapping to individual loci instead of the total number of sequenced fragments. This can improve robustness of differential expression calls for less abundant genes and transcripts.
Hits norm	Instructs how to count all fragments. Total specifies to count all fragments, including those not compatible with any reference transcript, towards the number of mapped fragments used in the FPKM denominator. Compatible specifies to use only compatible fragments. Selecting Compatible is generally recommended in Cuff diff to reduce certain types of bias caused by differential amounts of ribosomal reads which can create the impression of falsely differentially expressed genes.
Frag bias correct	Providing the sequences your reads were mapped to instructs Cuffdiff to run bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates.
Multi read correct	Do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.
Library type	Specifies RNA-Seq protocol.
Mask file	Ignore all reads that could have come from transcripts in this file. It is recommended to include any annotated rRNA, mitochondrial transcripts other abundant transcripts you wish to ignore in your analysis in this file. Due to variable efficiency of mRNA enrichment methods and rRNA depletion kits, masking these transcripts often improves the overall robustness of transcript abundance estimates.

Min alignment count	The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples. If no testing is performed, changes in the locus are deemed not significant, and the locus' observed changes don't contribute to correction for multiple testing.
FDR	Allowed false discovery rate used in testing.
Max MLE iterations	Sets the number of iterations allowed during maximum likelihood estimation of abundances.
Emit count tables	Include information about the fragment counts, fragment count variances, and fitted variance model into the report.
Cuffdiff tool path	The path to the Cuffdiff external tool in UGENE.
Temporary directory	The directory for temporary files.

Configure parameters, if necessary, and click Next. The last page of the wizard appears:



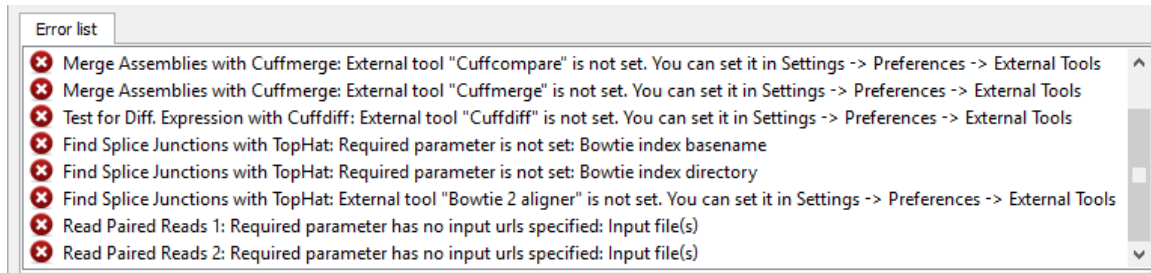
Choose output directories for each tools and click Finish.

Note that default button reverts all parameters to default settings.

Now let's validate and run the workflow. To validate that the workflow is correct and all parameters are set properly click the Validate workflow button on the Workflow Designer toolbar:



If there are some errors, they will be shown in the Error list at the bottom of the Workflow Designer window, for example:



However, if you have set all the required parameters, then there shouldn't be errors. After that you can estimate the workflow. To run estimation click the *Estimate workflow* button:



To run a valid workflow, click the Run workflow button on the Workflow Designer toolbar:



As soon as the variants calling task is finished, a notification and dashboard will appear. The dashboard will contain information about workflow: input and output files, all information about task.

## Scenarios

- Filter sequence that match a pattern
- Find patterns
- Gene-by-gene approach for characterization of genomes
- Merge sequences and annotations

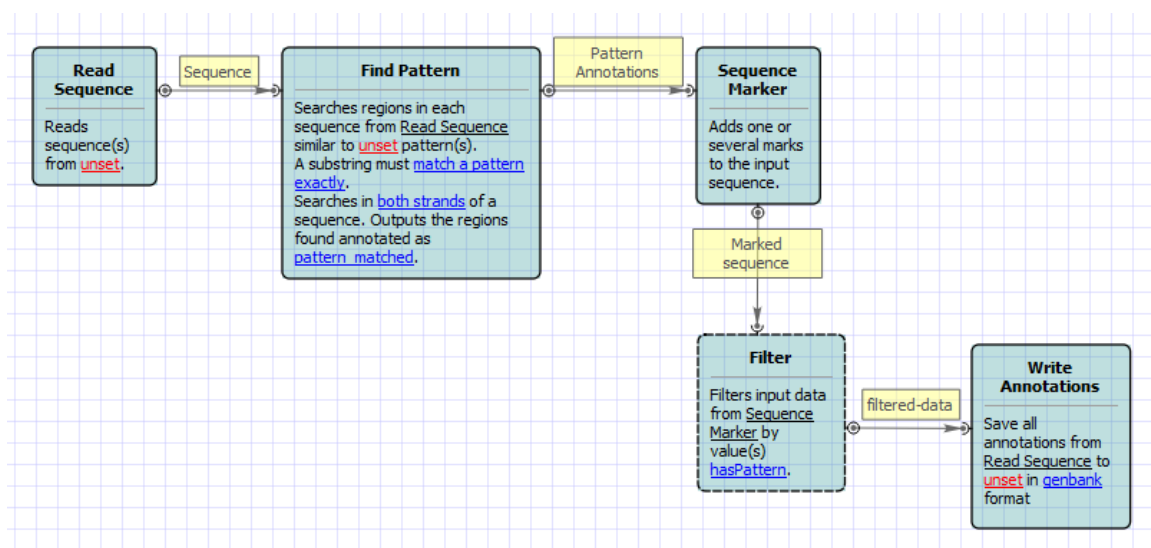
### Filter sequence that match a pattern

Using this workflow you can select (or reject) only those sequence that match any pattern you input. To find sequences matching a pattern:

1. In Read Sequence element specify a list of sequences you need to filter
2. In Find Pattern element input you pattern(s) or file with pattern in any sequence or newline-delimited format.
3. In Write Sequence element specify an output file

To find sequences that DO NOT match a pattern:

1. Put "Rest" instead "hasPattern" in Filter element.

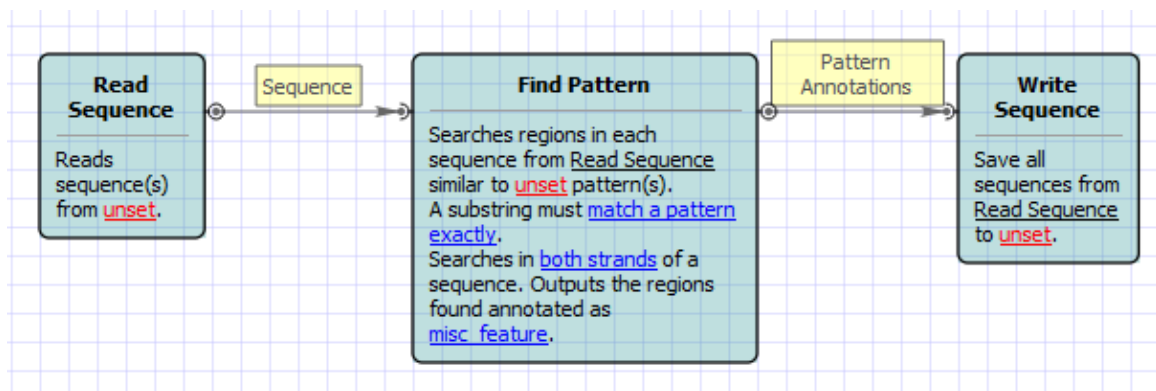


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Find patterns

This simple workflow finds patterns in you sequences and save them as annotations. You can use the workflow to map primers, regulatory signals, genes, etc. It loads any set of sequences from your files or folders and finds patterns in them. Just specify a dataset for the algorithm in the "Read sequence" element. Patterns are entered in comma-delimited format in the corresponding field of the "Find Pattern" element. Also you can load patterns from a file. In that case names of patterns can be saved as names of annotations. Files with patterns can be in any sequence format or in newline-delimited format.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

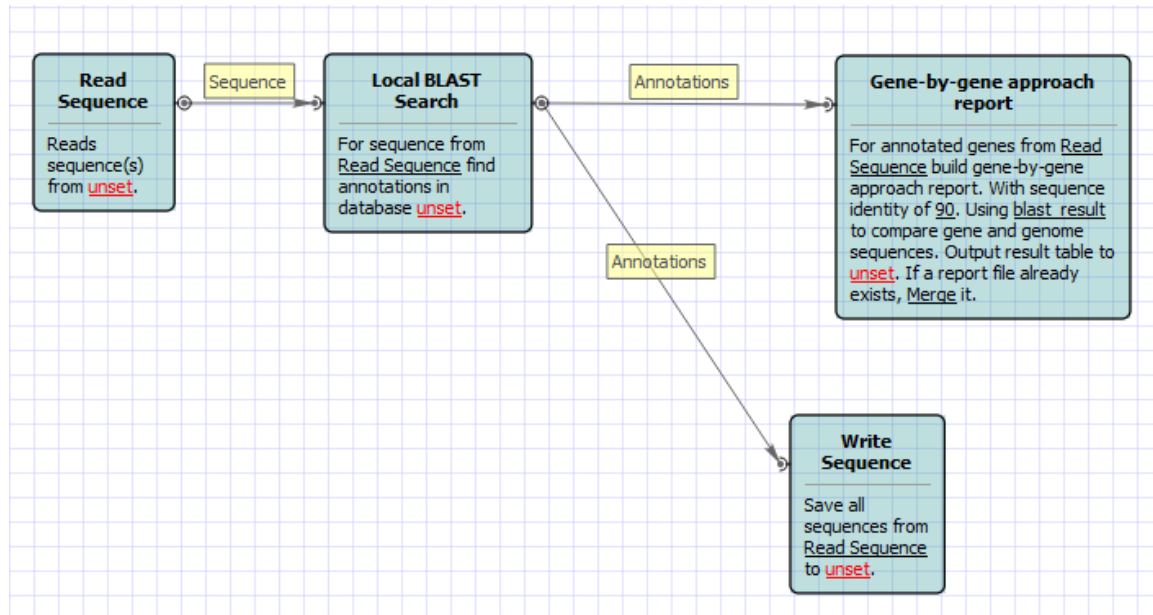


## Gene-by-gene approach for characterization of genomes

Suppose you have genomes and you want to characterize them. One of the ways to do that is to build a table of what genes are in each genome and what are not there.

1. Create a local BLAST db of your genome sequence/contigs. One db per one genome.
2. Create a file with sequences of genes you want to explore. This file will be the input file for the workflow.
3. Setup location and name of BLAST db you created for the first genome.
4. Setup output files: report location and output file with annotated (with BLAST) sequence. You might want to delete the "Write Sequence" element if you do not need output sequences.
5. Run the workflow.
6. Run the workflow on the same input and output files changing BLAST db for each genome that you have.

As the result you will get the report file. With "Yes" and "No" field. "Yes" answer means that the gene is in the genome. "No" answer MIGHT mean that there is no gene in the genome. It is a good idea to analyze all the "No" sequences using annotated files. Just open a file and find a sequence with a name of a gene that has "No" result.

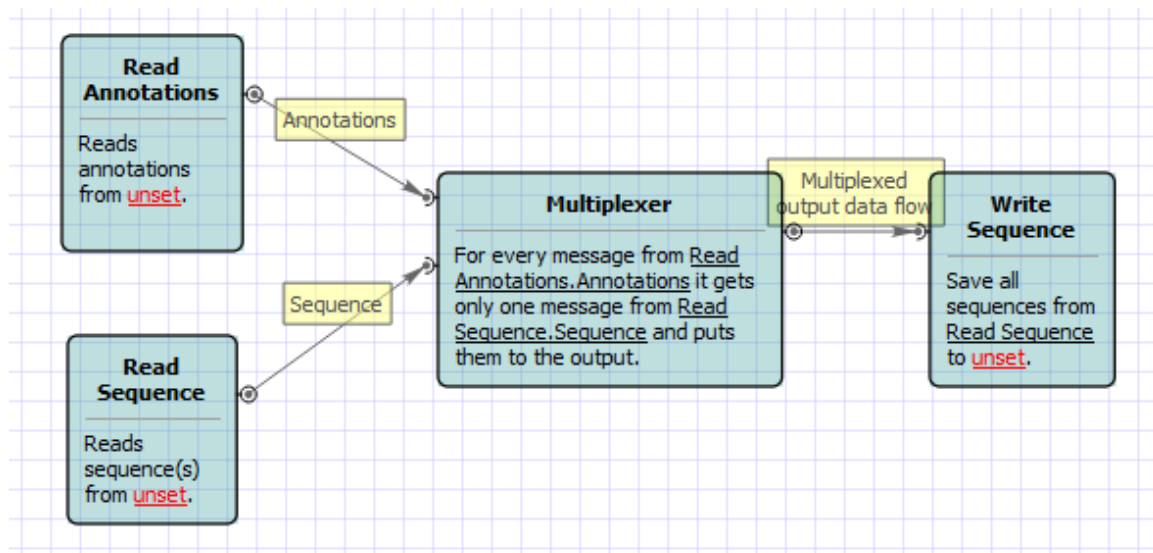


Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



## Merge sequences and annotations

If you have a list of files with sequences and separate files with annotation and you want to merge sequences and annotation, this workflow might help you. For instance, you have sequence in FASTA format and separate annotation in GFF. You want to merge them and write annotated sequences into Genbank files. By default, multiplexer takes sequences and annotation one by one, sticks one annotation to one sequence and passes it to the output. But you may change that behavior in parameters of Multiplexer.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:



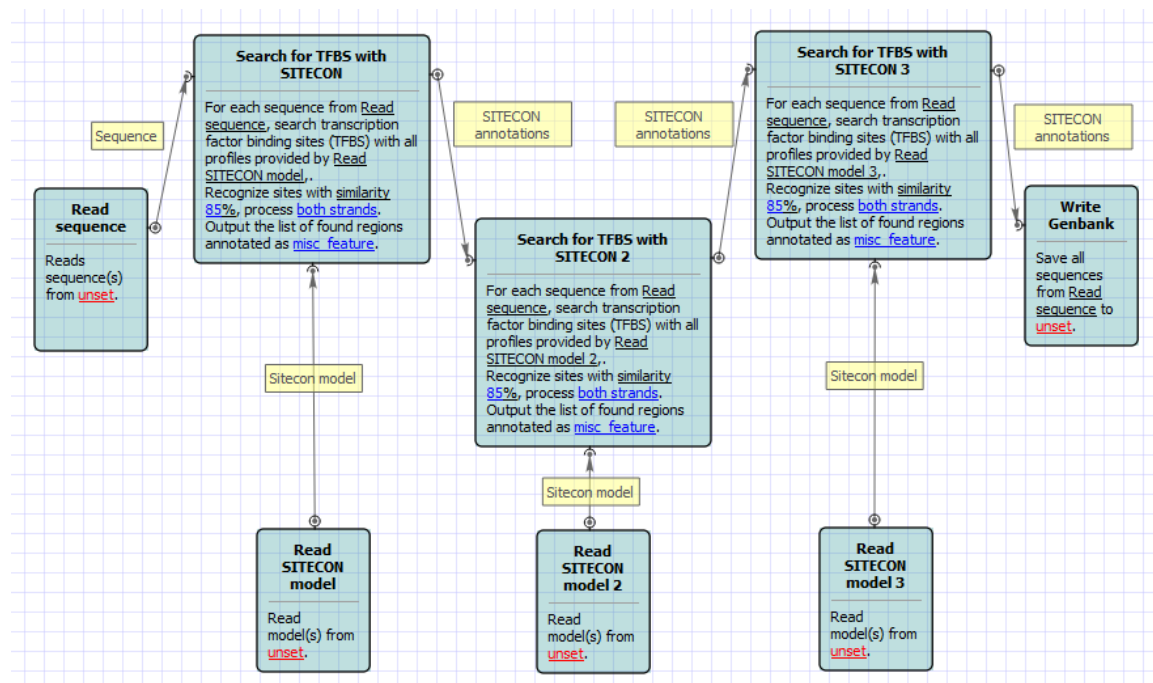


## Transcriptomics

- Search for transcription factor binding sites (TFBS) in genomic sequences

### Search for transcription factor binding sites (TFBS) in genomic sequences

This workflow predicts binding sites for number of transcription factors of interest using SITECON algorithm. The present workflow sample is designed for simultaneous recognition of binding sites for 3 different transcription factor types, you can expand it for recognition of any desired number of transcription factor types. SITECON - is a program package for recognition of potential transcription factor binding sites basing on the data about conservative conformational and physicochemical properties revealed on the basis of the binding sites sets analysis. Citing SITECON Please cite: Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for siterecognition. // Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W208-12.



Also, if required, you can change parameters. Use the workflow wizard to guide you through the parameters setup process. The first wizard page will appear when you click on the Show wizard button on the Workflow Designer toolbar:

