Analyze non-coding RNAs with InSyBio ncRNASeq

InSyBio
Intelligent Systems Biology

User Manual

March 2022

Insybio Suite v3.0

www.insybio.com

Introduction

ncRNASeq is a RNA analysis tool for the prediction and analysis of:

- Coding RNAs
- non-coding RNAs
- miRNA target genes
- Bulk RNA-sequencing data
- Single Cell RNA-sequencing data

Non-coding RNA genes are RNA sequences transcribed from DNA, but not translated to proteins. Their identification as well as the identification of the genes they regulate is a promising research area.

InSyBio ncRNASeq enables users to analyze non-coding RNAs. Users can search and analyze the RNA sequence of their interest. They can also analyze a full sequences dataset derived from online available databases, experimental sequencing techniques or computational in silico techniques.

With InSyBio ncRNASeq you can predict and analyze RNA genes and miRNA target genes combining a variety of sequential, structural and functional information, and using a high performance machine learning technique. The RNA analysis is conducted by the calculation of the 58 most informative features described in the literature, and the miRNA-miRNA targets analysis is conducted by the calculation of the 124 most informative ones. InSyBio ncRNASeq also provides results storage in its knowledge base, equipped with information retrieval tools, to allow users to produce and extract their own datasets.

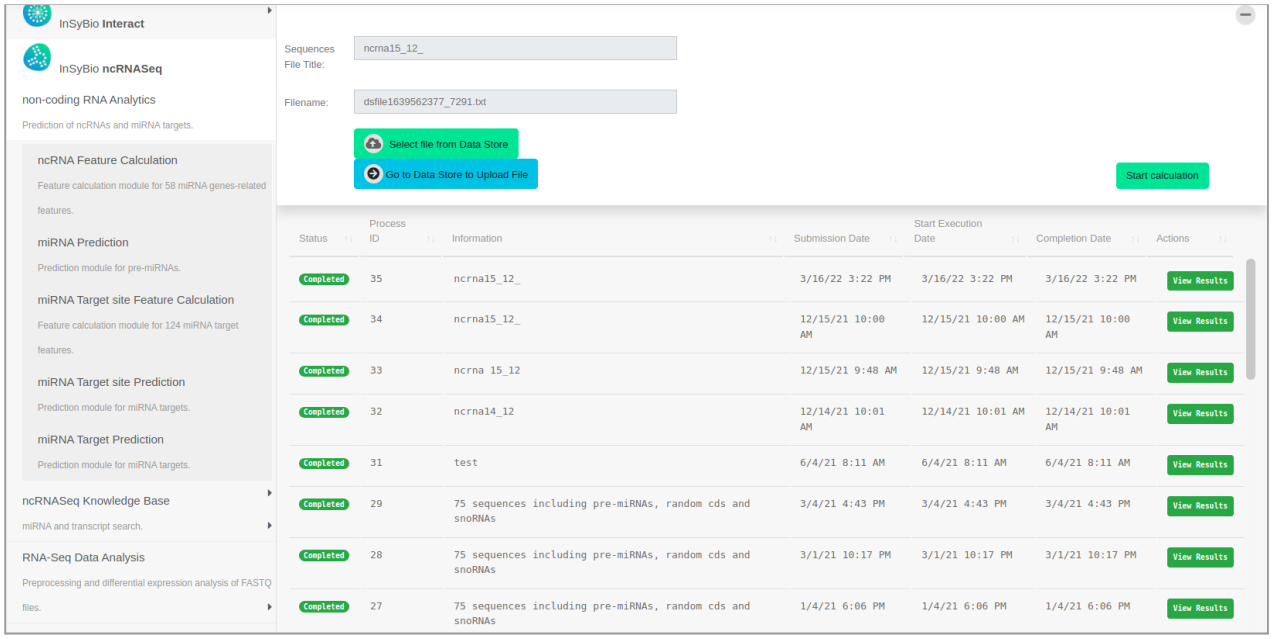
With InSyBio ncRNASeq you can:

a) Calculate 58 RNA genes-related features
b) Predict miRNAs
c) Calculate 124 miRNA target sites features
d) Predict miRNA target sites
e) Search stem-loop and mature miRNAs

f) Search transcripts and genes
g) Search transcripts and genes for potential miRNA targets
h) Predict miRNA targets
i) Apply our pipeline on your RNASeq data and perform Differential Expression Analysis
j) Apply our pipeline on your Single Cell RNASeq data and perform Differential Expression Analysis
k) Identify different types of novel small non-coding RNAs (e.g. snoRNAs, miRNAs, tRNA fragments etc) from your raw RNA-sequencing data

ncRNA Feature Calculation

You can calculate 58 informative features for non-coding RNAs by supplying their sequence in fasta format. These features include sequential, thermodynamical and structural properties of the RNA sequences.

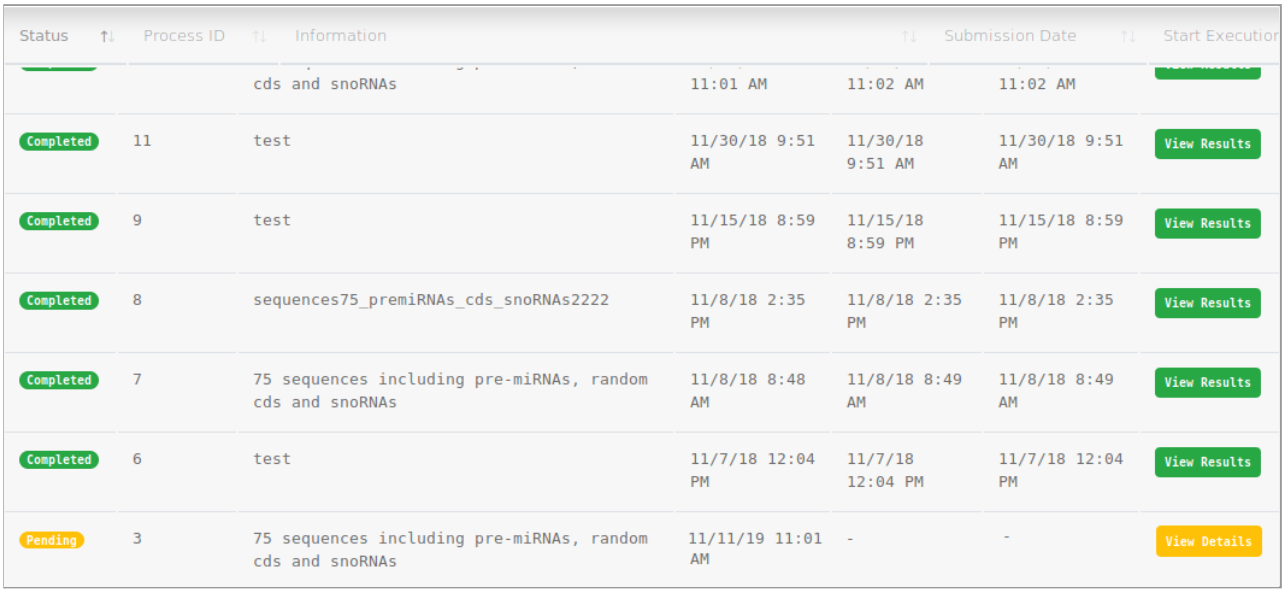


To start the calculation:

Select from the menu “Insybio ncRNASeq” → “non-coding RNA Analytics” → “ncRNA Feature Calculation”:

- Upload a new file of sequences in fasta format. You are redirected to the Data Store where step by step instructions guide you.
- Or Select a file from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.

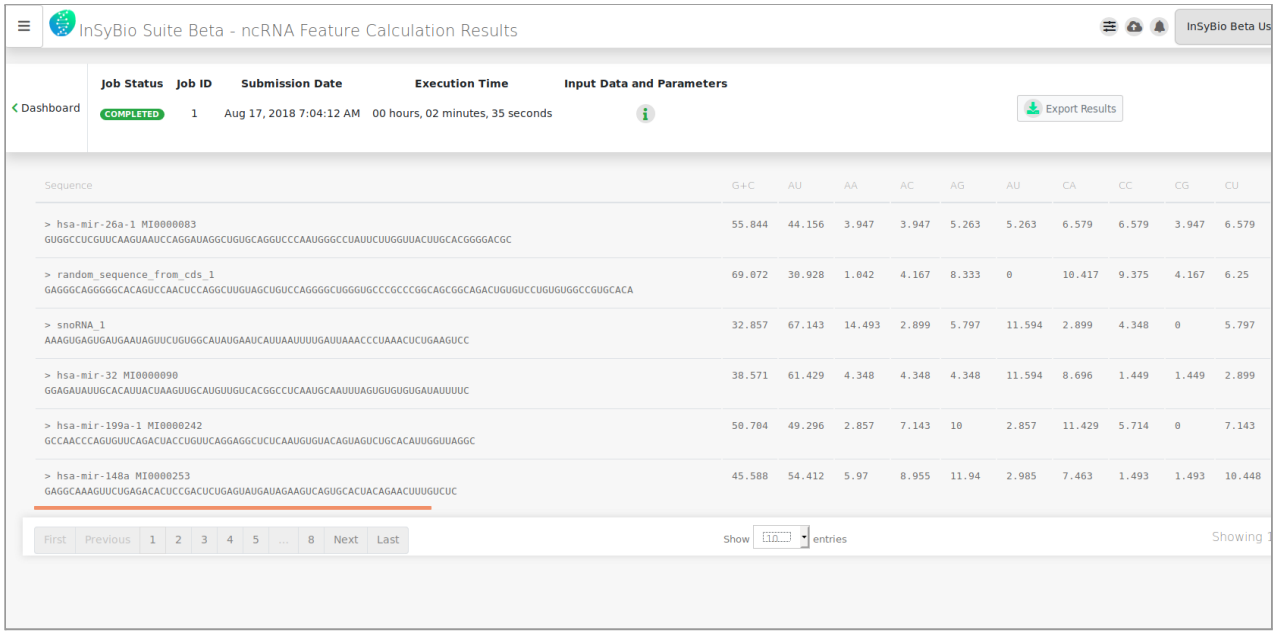
Batch calculations of many sequences are allowed. Just put the sequences in one file in fasta format.



Status	Process ID	Information			Submission Date	Start Execution
		cds and snoRNAs	11:01 AM	11:02 AM	11:02 AM	
Completed	11	test	11/30/18 9:51 AM	11/30/18 9:51 AM	11/30/18 9:51 AM	View Results
Completed	9	test	11/15/18 8:59 PM	11/15/18 8:59 PM	11/15/18 8:59 PM	View Results
Completed	8	sequences75_premiRNAs_cds_snoRNAs2222	11/8/18 2:35 PM	11/8/18 2:35 PM	11/8/18 2:35 PM	View Results
Completed	7	75 sequences including pre-miRNAs, random cds and snoRNAs	11/8/18 8:48 AM	11/8/18 8:49 AM	11/8/18 8:49 AM	View Results
Completed	6	test	11/7/18 12:04 PM	11/7/18 12:04 PM	11/7/18 12:04 PM	View Results
Pending	3	75 sequences including pre-miRNAs, random cds and snoRNAs	11/11/19 11:01 AM	-	-	View Details

To view the results:

By starting a calculation the ncRNA Feature Calculation dashboard is updated with the submitted job, there you can view the status of your current and previous ncRNA feature calculations. At completion of the calculation you can select the View Details at the Actions column and view the calculated features.



The results are presented on your screen in a browse-able table or you can download them as a TAB delimited txt file.

For each non-coding RNA, its sequence and its 58 features are presented.

The description of the supported features for the characterization of the non coding RNAs is the following:

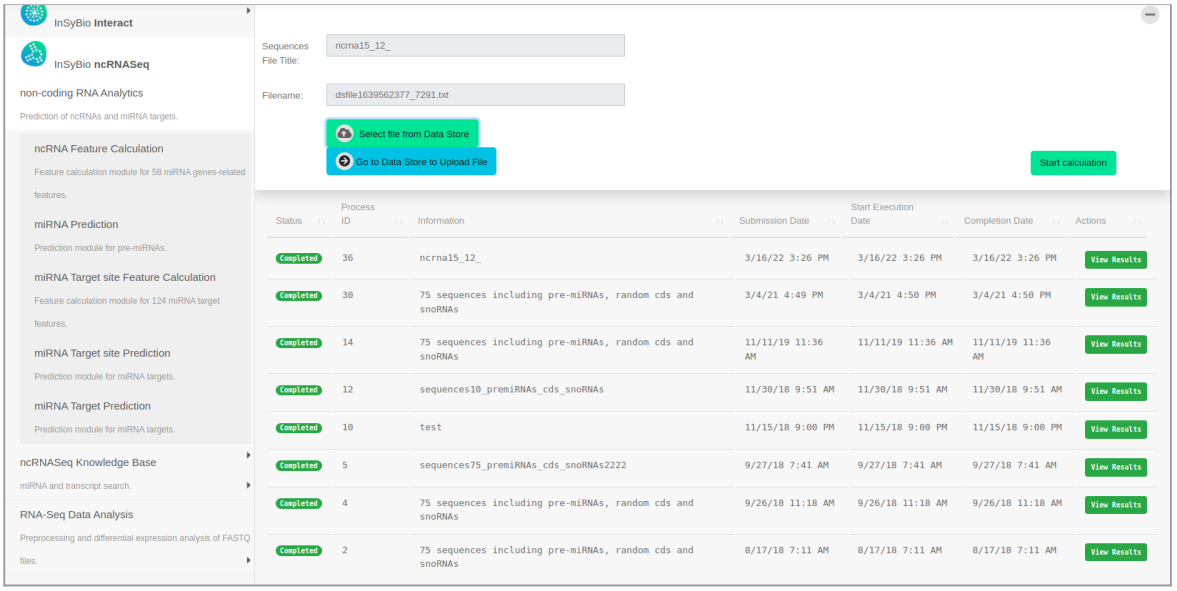
Feature	ABBR
2 Aggregate Dinucleotide Frequencies (%G+C ratio, %A+U ratio)	G + C, A + U
16 dinucleotide frequencies (%XY) such that X,Y e Σ[A,C,G,U]	AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU
MFE Index 1 = dG/%(C+G)	MFE1
MFE Index 2 = dG/number_of_stems, where each stem is at least 3 continuous base pairs in the structure	MFE2
MFE Index 3 = dG/number_of_loops , where number_of_loops is the number of the loops in the secondary structure	MFE3
MFE Index 4 = dG/total_bases	MFE4
MFE Index 5 = dG/%(A+U) ratio	MFE5

Adjusted Minimum Free Energy of folding dG = MFE/L, where MFE is the minimum free energy of the structure as calculated by the Vienna fold routine	dG
Adjusted base pairing propensity dP = total_bases/L, where L is the length of the structure and total_bases the number of base pairs in the structure	dP
Adjusted base pair distance dD	dD
Adjusted shannon entropy dQ	dQ
Positional Entropy dPs: a new introduced attribute which estimates the structural volatility of the secondary structure	PosEntropy
Normalized Ensemble Free Energy	EAFE
Structural Diversity	Div/ty
Frequency of MFE structure	Freq
Feature	**ABBR**
Diff = |MFE-EFE|/L where, EFE is the ensemble free energy	Diff
Structure Enthalpy dH	dH
Normalized Structure Enthalpy dH/L	dH/L
Structure Entropy dS	dS
Normalized Structure Entropy dS/L	dS/L
Melting Temperature Tm	Tm
Normalized Structure Enthalpy TH/L	Tm/L
|X-Y| is the number of (X-Y) base pairs in the secondary structure	|A-U|/L, |G-C|/L, |G-U|/L
Average base pair per stem	Avg_BP_stems
%(A-U)/n_stems, %(G-C)/n_stems, %(G-U)/n_stems.	(A-U)/n_stems, (G-C)/n_stems, (G-U)/n_stems
Ratio G/C ,where G,C is the number of G,C bases	G/C
BP is the total number of base pairs and GC,GU,AU the number of respective base pairs	BP/GC, BP/GU, BP/AU

Length of the sequence	Len
Centroid Energy: RNA folding related attribute calculated by the Vienna RNA package	DE/L
Centroid Distance: RNA folding related attribute calculated by the Vienna RNA package	CE_dist
5 statistical features	zG, zP, zD, zQ, zSP
Topological descriptor dF	dF

miRNA Prediction

You can predict pre-miRNAs and discriminate them between pseudo-hairpins and other molecules providing RNA sequences in fasta format. The prediction of pre-miRNAs and pseudo-hairpins is accomplished through the application of a novel methodology which combines Genetic Algorithms with epsilon-SVR techniques. Genetic Algorithms were used to optimize the feature subset which should be used as inputs and the parameters C, sigma and epsilon of epsilon SVR models. The accuracy of this technique in predicting pre-miRNAs is 95%. A sequence is predicted as other, if the minimum free energy is more than -15 kcal/mol or the number of base pairs is less than 18.

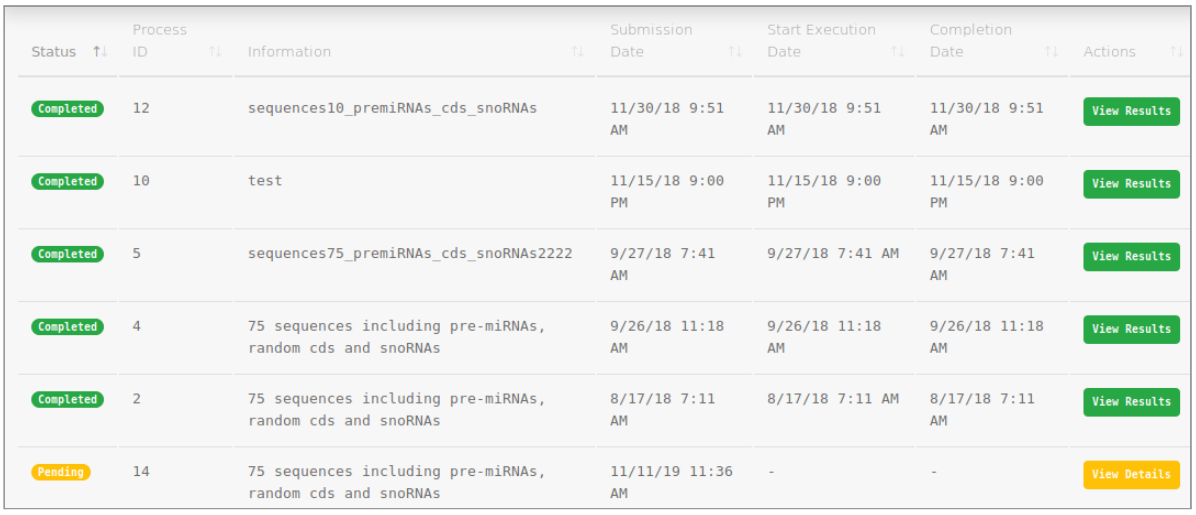


To start the calculation:

Select from the menu “Insybio ncRNASeq” → “non-coding RNA Analytics” → “miRNA Prediction”:

- Upload a new file of sequences in fasta format. You are redirected to the Data Store where step by step instructions guide you.
- Or Select a file from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.

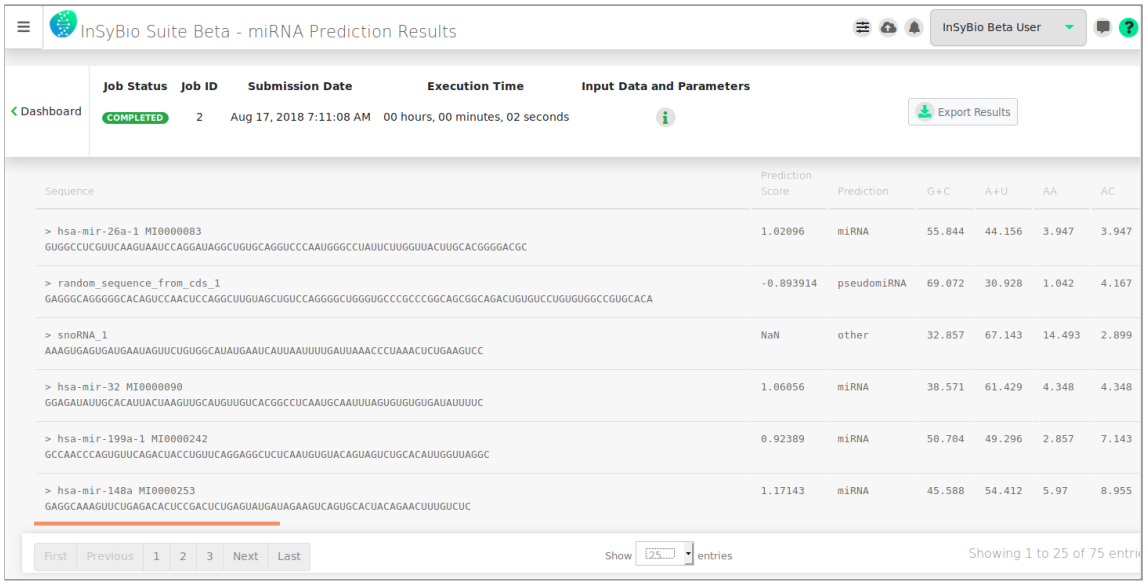
Batch calculations of many sequences are allowed. Just put the sequences in one file in fasta format.



Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	12	sequences10_premiRNAs_cds_snoRNAs	11/30/18 9:51 AM	11/30/18 9:51 AM	11/30/18 9:51 AM	View Results
Completed	10	test	11/15/18 9:00 PM	11/15/18 9:00 PM	11/15/18 9:00 PM	View Results
Completed	5	sequences75_premiRNAs_cds_snoRNAs2222	9/27/18 7:41 AM	9/27/18 7:41 AM	9/27/18 7:41 AM	View Results
Completed	4	75 sequences including pre-miRNAs, random cds and snoRNAs	9/26/18 11:18 AM	9/26/18 11:18 AM	9/26/18 11:18 AM	View Results
Completed	2	75 sequences including pre-miRNAs, random cds and snoRNAs	8/17/18 7:11 AM	8/17/18 7:11 AM	8/17/18 7:11 AM	View Results
Pending	14	75 sequences including pre-miRNAs, random cds and snoRNAs	11/11/19 11:36 AM	-	-	View Details

To view the results:

By starting a calculation the “miRNA Prediction” dashboard is updated with the status of the new job, there you can view the status of your current and previous miRNA prediction. At completion of the prediction you can select the View Results at the Actions column and view the calculated features.

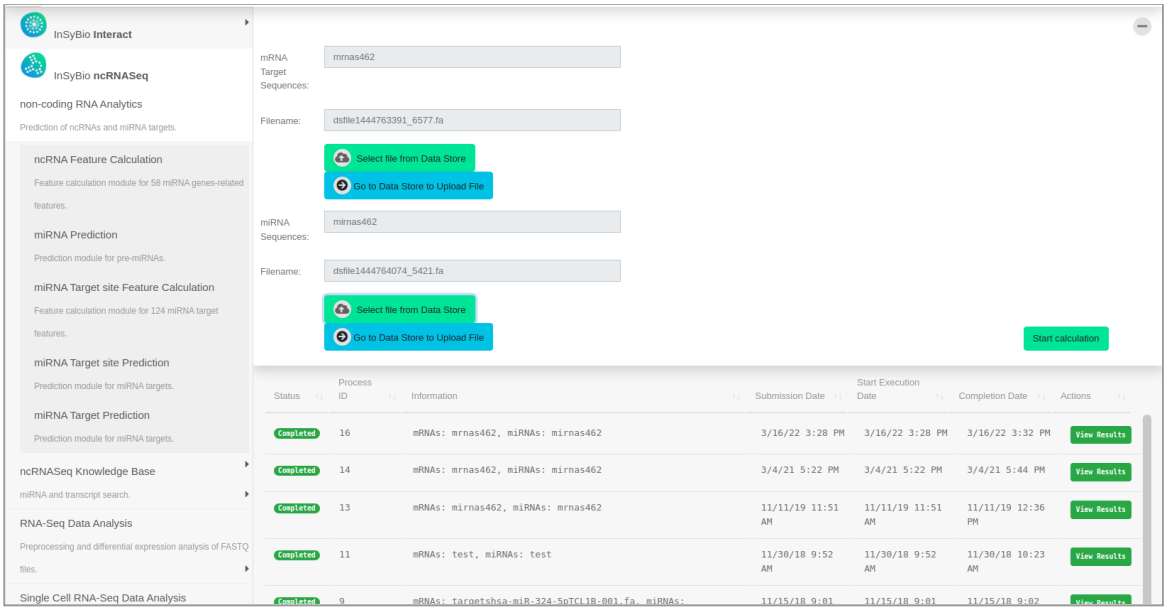


The results are presented on your screen in a browseable table or you can download them as a TAB delimited txt file.

For each non-coding RNA, its sequence, its calculated confidence score, the prediction whether it is a miRNA, a pseudo-hairpin or other and its 58 features are presented.

miRNA Target Site Feature Calculation

You can calculate 124 features for every pair of a miRNA and its potential target site within an mRNA. These features include sequential, thermodynamical and structural properties of the miRNA:mRNA pair.

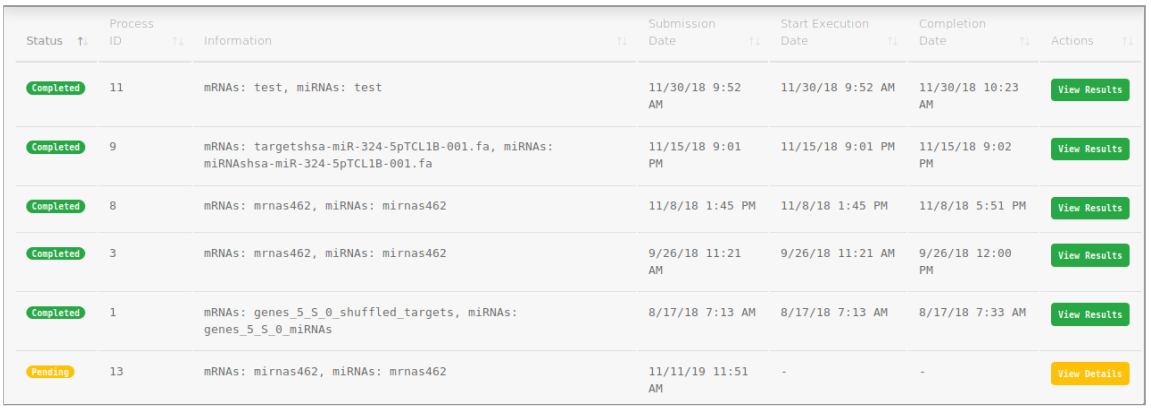


To start the calculation:

Select from the menu “InSyBio ncRNASeq” → “non-coding RNA Analytics” → “miRNA Target Features Calculation” and then:

- Upload a new file of mRNA binding sites sequences and a new file of miRNA sequences, both in fasta format. The mRNA target site of the first file and every miRNA of the second file are considered as a miRNA:mRNA pair. You are redirected to the Data Store where step by step instructions guide you for both files uploading.
- Or Select a file of mRNA binding sites sequences and a file of miRNA sequences, both in fasta format from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.

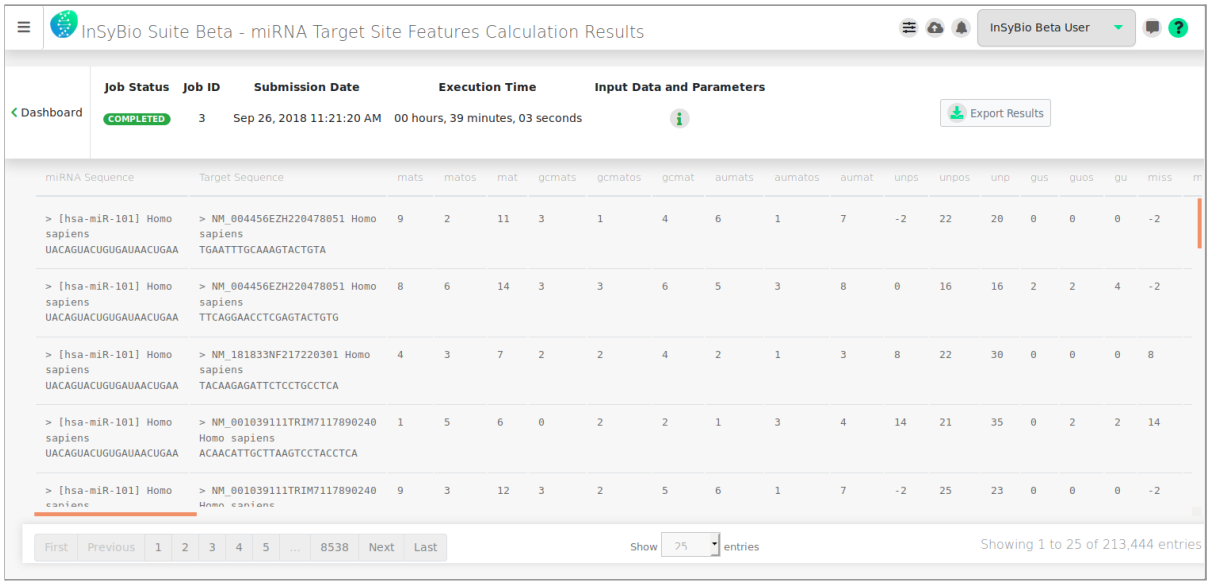
Batch feature calculation of many miRNA:mRNA pairs with a single run is allowed. Just put the mRNA binding sites sequences in the first file and miRNA sequences in the second file in fasta format.



Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	11	mRNAs: test, miRNAs: test	11/30/18 9:52 AM	11/30/18 9:52 AM	11/30/18 10:23 AM	View Results
Completed	9	mRNAs: targetshsa-miR-324-5pTCL1B-001.fa, miRNAs: miRNAshsa-miR-324-5pTCL1B-001.fa	11/15/18 9:01 PM	11/15/18 9:01 PM	11/15/18 9:02 PM	View Results
Completed	8	mRNAs: mrnas462, miRNAs: mirnas462	11/8/18 1:45 PM	11/8/18 1:45 PM	11/8/18 5:51 PM	View Results
Completed	3	mRNAs: mrnas462, miRNAs: mirnas462	9/26/18 11:21 AM	9/26/18 11:21 AM	9/26/18 12:00 PM	View Results
Completed	1	mRNAs: genes_5_S_0_shuffled_targets, miRNAs: genes_5_S_0_miRNAs	8/17/18 7:13 AM	8/17/18 7:13 AM	8/17/18 7:33 AM	View Results
Pending	13	mRNAs: mirnas462, miRNAs: mrnas462	11/11/19 11:51 AM	-	-	View Details

To view the results:

By starting a new calculation the “miRNA Target Site Feature Calculation” dashboard is updated with the new job, there you can view the status of your current and previous miRNA Target Features Calculations. At completion of the calculation you can select the View Results at the Actions column and view the calculated features.



The results are presented on your screen in a browse-able table or you can download them as a TAB delimited txt file.

For each miRNA:mRNA pair, the miRNA sequence, the mRNA binding site sequence and the 124 miRNA::mRNA pair features are presented.

The description of the supported features for the characterization of the miRNA::mRNA pair is the following:

Feature	ABBR	Category
number of matches in seed part	mats	structural
number of matches in out-seed part	matos	structural
total number of matches	mat	structural
number of GC matches in seed part	gcmats	structural
number of GC matches in out-seed part	gcmatos	structural
total number of GC matches	gcmat	structural
number of AU matches in seed part	aumats	structural
number of AU matches in out-seed part	aumatos	structural
total number of AU matches	aumat	structural
number of mismatches in seed part	unps	structural
number of mismatches in out-seed part	unpos	structural
total number of mismatches	unp	structural
number of GU wobble pairs in seed part	gus	structural
number of GU wobble pairs in out-seed part	guos	structural
total number of GU wobble pairs	gu	structural
number of other mismatches in seed part	miss	structural
number of other mismatches in out-seed part	misos	structural
total number of other mismatches	mis	structural
number of bulges in seed part	buls	structural

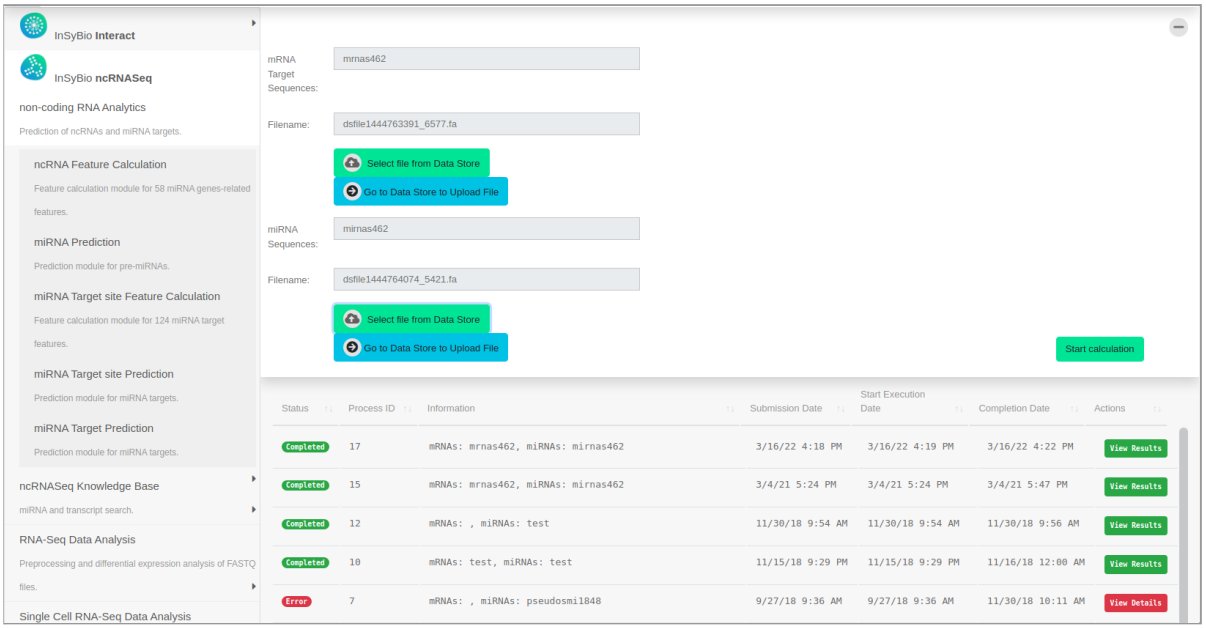
Feature	ABBR	Category
number of bulges in out-seed part	bulos	structural
total number of bulges	bul	structural
number of loops in seed part	symls	structural
number of loops in out-seed part	symlos	structural
total number of loops	syml	structural
number of asymmetric loops in seed part	asymls	structural
number of asymmetric loops in out-seed part	asymlos	structural
total number of asymmetric loops	asyml	structural
length of largest bulge	maxbul	structural
number of bulges of length 1-7 and greater than 7 in seed part (8 features)	cbul1s, cbul2s, cbul3s, cbul4s, cbul5s, cbul6s, cbul7s, cbul8s	structural
number of bulges of length 1-7 and greater than 7 in out-seed part (8 features)	cbul1os, cbul2os, cbul3os, cbul4os, cbul5os, cbul6os, cbul7os, cbul8os	structural
number of symmetric loops of length 1-7 and greater than 7 in seed part (8 features)	csl1s, csl2s, csl3s, csl4s, csl5s, csl6s, csl7s, csl8s	structural
number of symmetric loops of length 1-7 and greater than 7 in out-seed part (8 features)	csl1os, csl2os, csl3os, csl4os, csl5os, csl6os, csl7os, csl8os	structural
number of asymmetric loops of length 1-7 and greater than 7 in seed part (8 features)	casl1s, casl2s, casl3s, casl4s, casl5s, casl6s, casl7s, casl8s	structural
number of asymmetric loops of length 1-7 and greater than 7 in out-seed part (8 features)	casl1os, casl2os, casl3os, casl4os, casl5os, casl6os, casl7os, casl8os	structural
proportion of A, C, G, U in the target sequence (4	aper, cper, gper,	structural

features)	upper	
distance from the start of the seed part to the last match of the out-seed part	dist	structural
seed score obtained by the sum of pair scores in the seed region. GC and AU with 5, GU with 2 and the others with -3	scores	structural
out-seed score obtained by the sum of pair scores in the out-seed region. GC and AU with 5, GU with 2 and the others with -3	scoreos	structural
free energy of the seed part	mfes	thermodynamic
free energy of the out-seed part	mfeos	thermodynamic
free energy of the total miRNA-mRNA alignment structure	mfe	thermodynamic
free energy of the target sequence	mfet	thermodynamic
normalized free energy of the target sequence=(-1 * free energy of the target sequence)/log(length of target * length of miRNA)	nmfe	thermodynamic
difference in the free energies of the total miRNA-perfect target alignment structure and the total miRNA-mRNA alignment structure	dmfe	thermodynamic
positions from 1 to 20 with a GC match, an AU match, a GU match or a mismatch (20 features)	pos1, pos2, pos3, pos4, pos5, pos6, pos7, pos8, pos9, pos10, pos11, pos12, pos13, pos14, pos15, pos16, pos17, pos18, pos19, pos20	positional
terminal (position 8) base match	match8	positional
positional pair score obtained by the sum of the product of the weight and the corresponding pair score throughout the total miRNA-mRNA alignment structure. G:C and A:U are awarded with 5, G:U with 1, all other mismatches with -3 and the mismatches containing gaps with -1. Positional weight is 1 for all non-seed positions and 2 for all seed positions.	s106	positional

Feature	ABBR	Category
matrix score obtained by the sum of the diagonal elements in the matrix formed by the miRNA and its target. WC pairs: 5, Wobble pairs: 2, Inserts: -1, Deletes: -1, Symmetric mismatches: -3, Mismatches: -2	score	positional
deviation of the positional pair score with the score obtained with a perfect target	ds108	positional
deviation of the matrix score with the score obtained with a perfect target	ds109	positional
existence of the 10 most frequent nucleotide sequence 'words' with lengths 4, 5, 6, 7, 8 from the seed sequence of the miRNAs of our dataset	ugag, cagu, agug, agguag, aggua, aggu, gguag, ggua, guag, ugcu	'motif'

miRNA Target Site Prediction

You can computationally validate miRNA targets. The computational intelligent technique, which was applied for the prediction of miRNAs (hybrid combination of Genetic Algorithms and epsilon-SVRs), and 124 informative features are used.

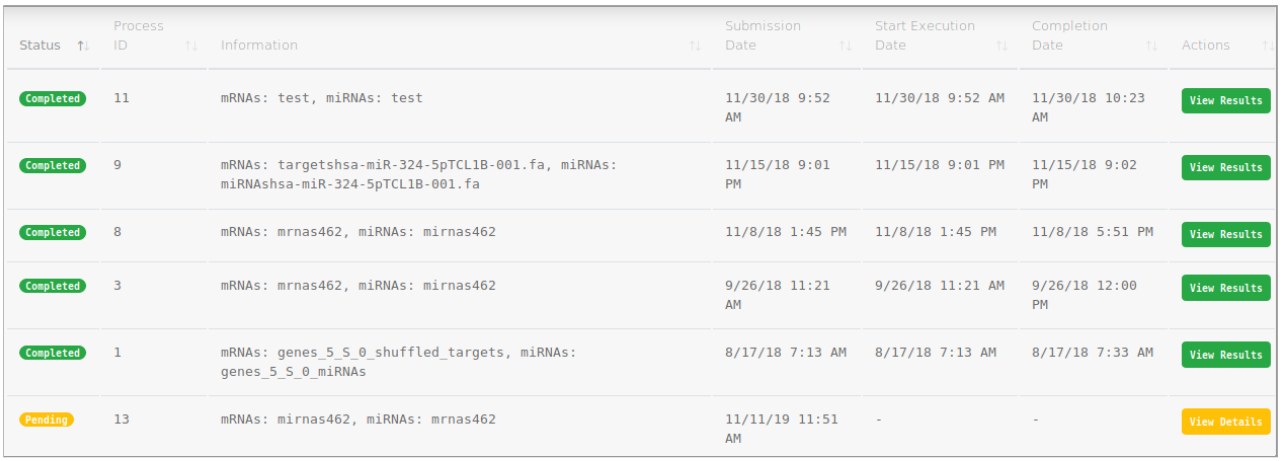


To start the prediction:

Select from the menu “InSyBio ncRNASeq” → “non-coding RNA Analytics” → “miRNA Target Site Prediction” and then:

- Upload a new file of candidate mRNA target binding sites sequences and a new file of miRNA sequences, both in fasta format. The mRNA target site of the first file and every miRNA of the second file are considered as a miRNA:mRNA pair. You are redirected to the Data Store where step by step instructions guide you for both files uploading.
- Or Select a file of candidate mRNA target binding sites sequences and a file of miRNA sequences, both in fasta format from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.

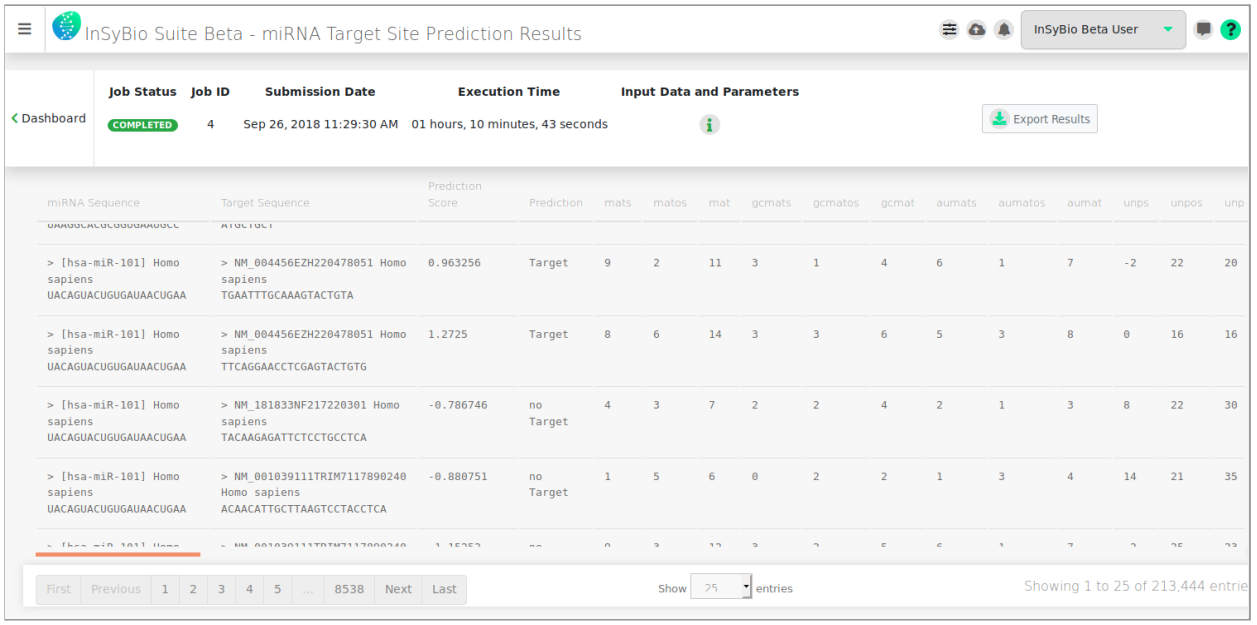
Batch predictions of many miRNA:mRNA pairs with a single run are allowed. Just put the candidate mRNA target binding sites sequences in the first file and miRNA sequences in the second file in fasta format.



Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	11	mRNAs: test, miRNAs: test	11/30/18 9:52 AM	11/30/18 9:52 AM	11/30/18 10:23 AM	View Results
Completed	9	mRNAs: targetshsa-miR-324-5pTCL1B-001.fa, miRNAs: miRNAshsa-miR-324-5pTCL1B-001.fa	11/15/18 9:01 PM	11/15/18 9:01 PM	11/15/18 9:02 PM	View Results
Completed	8	mRNAs: mrnas462, miRNAs: mirnas462	11/8/18 1:45 PM	11/8/18 1:45 PM	11/8/18 5:51 PM	View Results
Completed	3	mRNAs: mrnas462, miRNAs: mirnas462	9/26/18 11:21 AM	9/26/18 11:21 AM	9/26/18 12:00 PM	View Results
Completed	1	mRNAs: genes_5_S_0_shuffled_targets, miRNAs: genes_5_S_0_miRNAs	8/17/18 7:13 AM	8/17/18 7:13 AM	8/17/18 7:33 AM	View Results
Pending	13	mRNAs: mirnas462, miRNAs: mrnas462	11/11/19 11:51 AM	-	-	View Details

To view the results:

By starting a calculation the “miRNA Target Site Prediction” dashboard is updated with the new job, where you can view the status of your current and previous miRNA Target Site Prediction. At completion of the calculation you can select the View Results at the Actions column and view the predictions and calculated features.

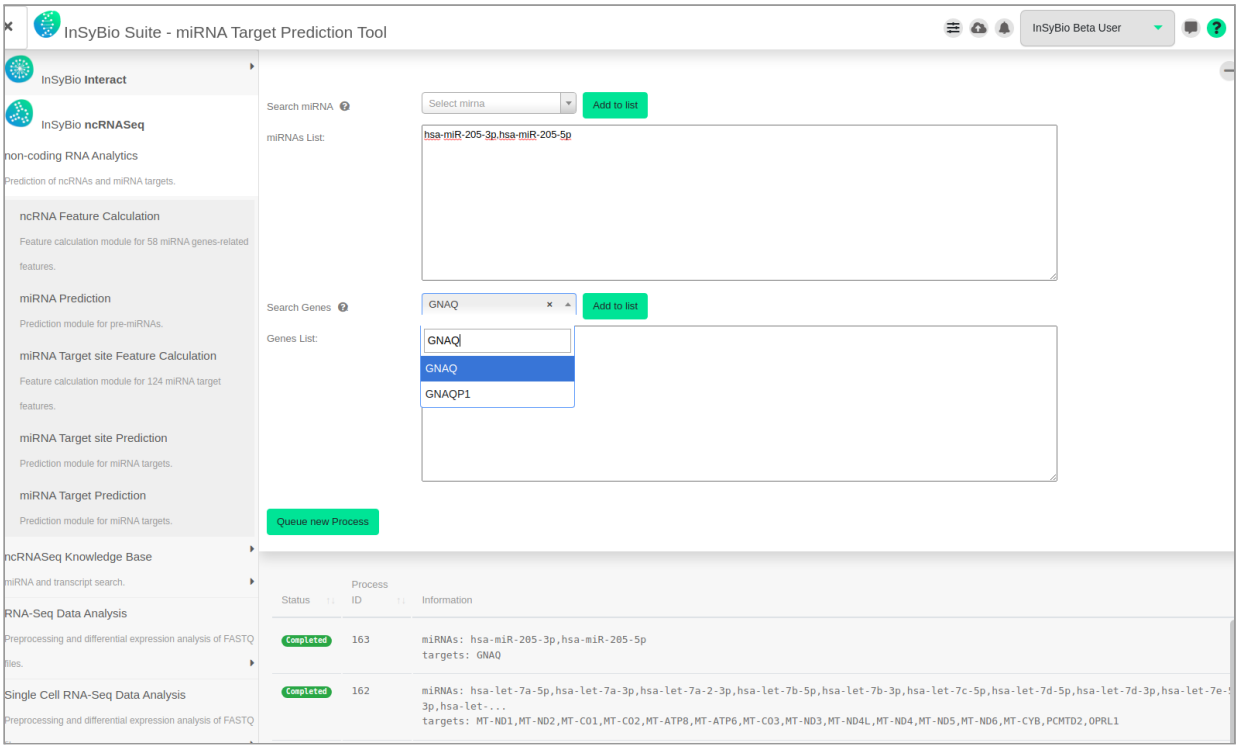


The results are presented on your screen in a browseable table or you can download them as a TAB delimited txt file.

For each miRNA:mRNA pair, the miRNA sequence, the mRNA binding site sequence, whether the miRNA:mRNA pairs share a targeting relation or not, the confidence score of the prediction and the all 124 miRNA::mRNA are presented.

miRNA Target Prediction

You can computationally predict potential miRNA targets at given Genes or Transcripts and given miRNAs. BLAST is performed in order to find potential target sites, and then the computational intelligent technique, which was applied for the prediction of miRNAs (hybrid combination of Genetic Algorithms and epsilon-SVRs), and 124 informative features are used in order to calculate a prediction score.



To start the prediction:

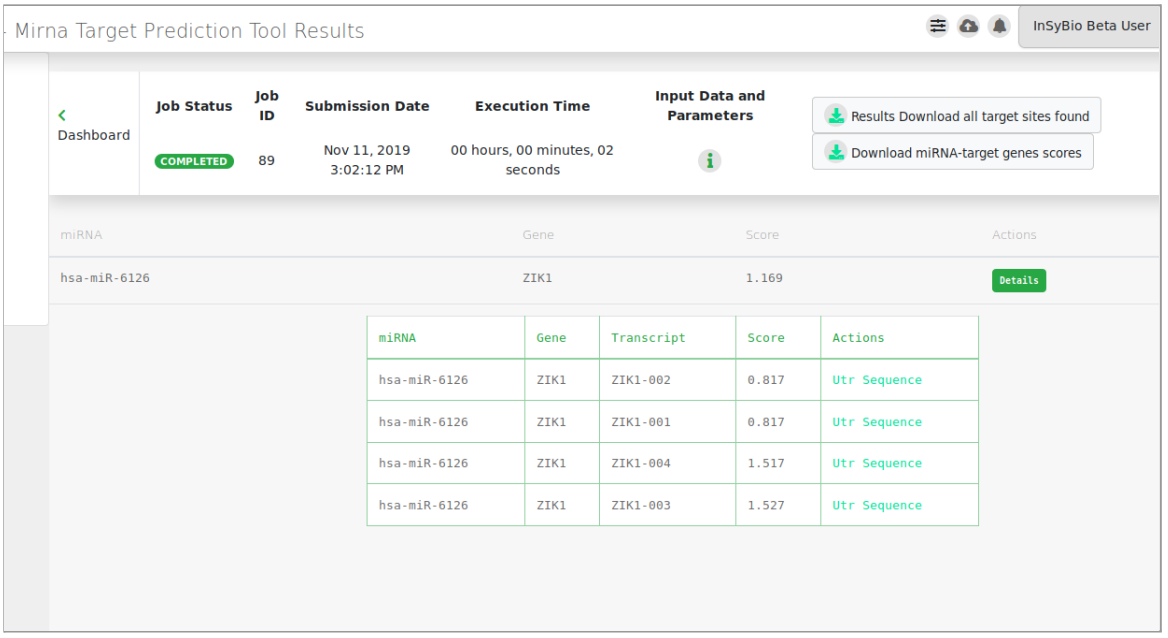
Select from the menu “InSyBio ncRNASeq” → “non-coding RNA Analytics” → “miRNA Target Prediction” field and then:

- Select the miRNAs and the Genes you want to search for potential targets by searching in our Database and adding them to the miRNA List and Genes List or add them manually to their Lists and separating them with commas.

Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	89	miRNAs: hsa-miR-6126 targets: ZIK1	11/11/19 3:02 PM	11/11/19 3:02 PM	11/11/19 3:02 PM	View Results
Completed	88	miRNAs: mmu-miR-3072-3p,mmu-miR-7051-3p,mmu-miR-3968,mmu-miR-8106,mmu-miR-99a-3p,mmu-miR-21a-5p,mmu-miR-3110-5p,mmu-miR-505-3p,mmu-miR-7091-5p,mmu-miR-337-5p,mmu-miR-18a-3p,mmu-miR-1949,mm... targets: ZIK1	2/11/19 12:11 PM	6/6/19 11:21 AM	6/6/19 3:39 PM	View Results
Completed	87	miRNAs: hsa-miR-576-3p,hsa-miR-140-5p,hsa-miR-522-5p,hsa-miR-1298-5p,hsa-miR-133a-3p,hsa-miR-4743-3p,hsa-miR-557,hsa-miR-548ao-3p,hsa-miR-5088-5p,hsa-miR-4649-5p,hsa-miR-665,hsa-miR-3622b-... targets: NELL2,SERPINI1,SMOC1,FGF2,MMRN2,PRSS3,VEGFB,ADAM21,ADAMTSL4,C1QTNF4,CCL3L3,COL4A2,LAMB1	11/29/18 3:40 PM	11/29/18 3:40 PM	11/29/18 3:52 PM	View Results
Completed	86	miRNAs: hsa-miR-6126, hsa-miR-1200, hsa-let-7a-2-3p, hsa-miR-106b-3p targets: ZIK1, A1BG-AS1, FGGY	11/29/18 3:39 PM	11/29/18 3:39 PM	11/29/18 3:39 PM	View Results
Completed	85	miRNAs: hsa-miR-6126 targets: ZIK1	11/29/18 3:09 PM	11/29/18 3:09 PM	11/29/18 3:09 PM	View Results
Error	84	miRNAs: targets: ZIK1	11/29/18 3:08 PM	11/29/18 3:08 PM	11/29/18 3:08 PM	View Details

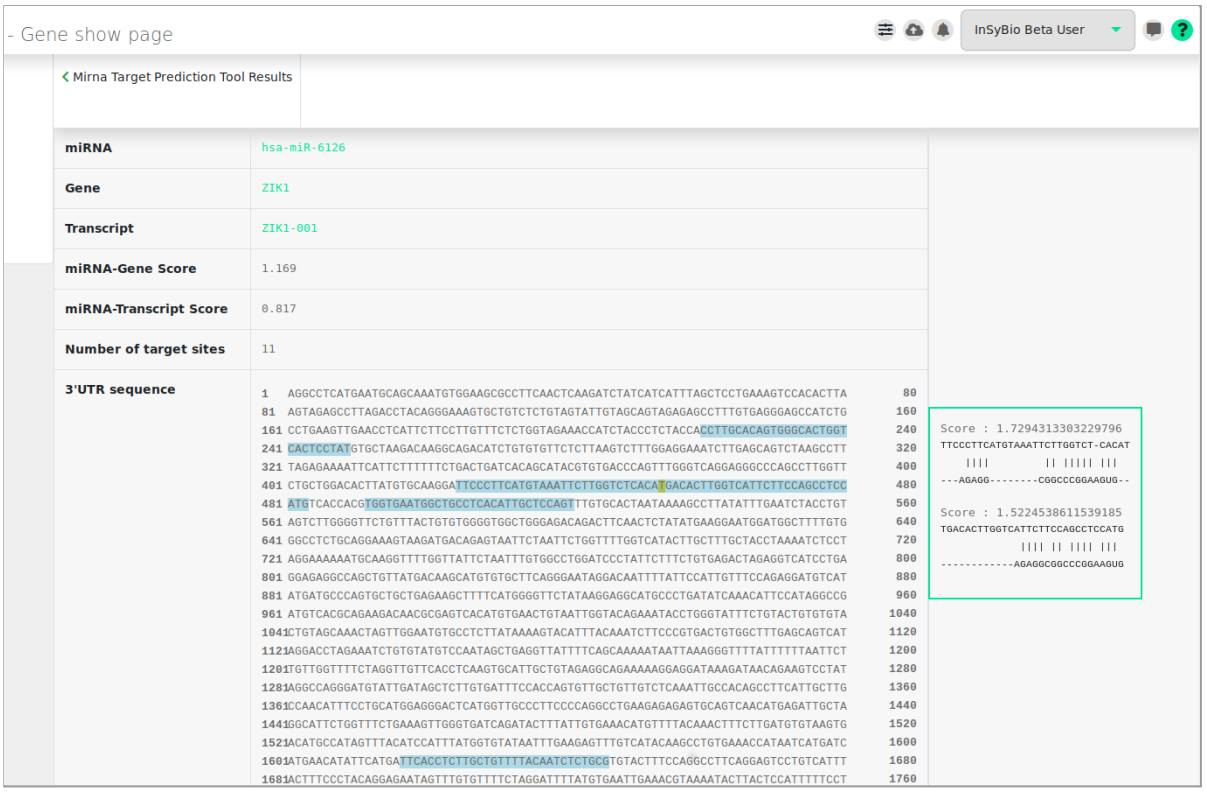
To view the results:

By starting a calculation the “miRNA target Prediction” dashboard is updated with the new job’s information, there you can view the status of your current and previous miRNA Target Predictions. At completion of the calculation you can select the View Results at the Actions column and view the results.



The results are presented on your screen in a browse-able table, with each miRNA and gene pair in a row with their confidence score. By pressing Details at the Actions Column the specific scores between the miRNA and the gene’s transcripts can be

viewed. If no target sites are found “No targets found!” is presented at the score column. If one or more target sites are found you can view its UTR sequence, with the target sites of the miRNA highlighted. Multiple target sites are marked with green color and unique target sites are marked with light blue.

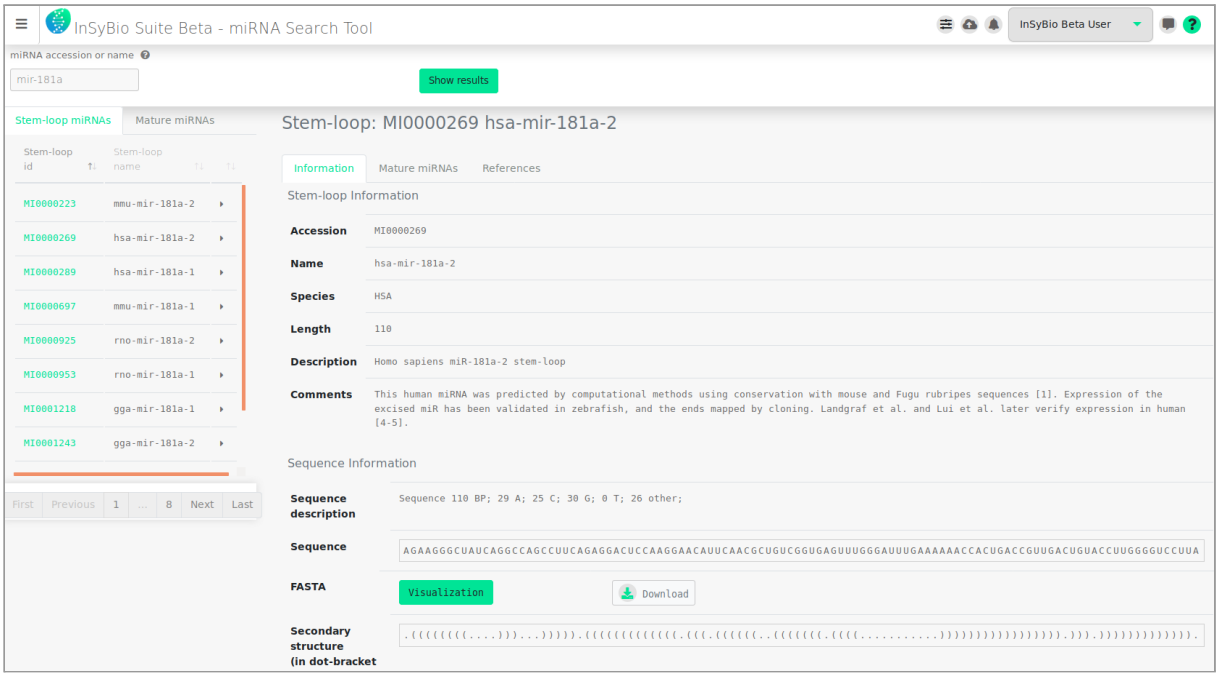


You can download all target sites found as a txt file.

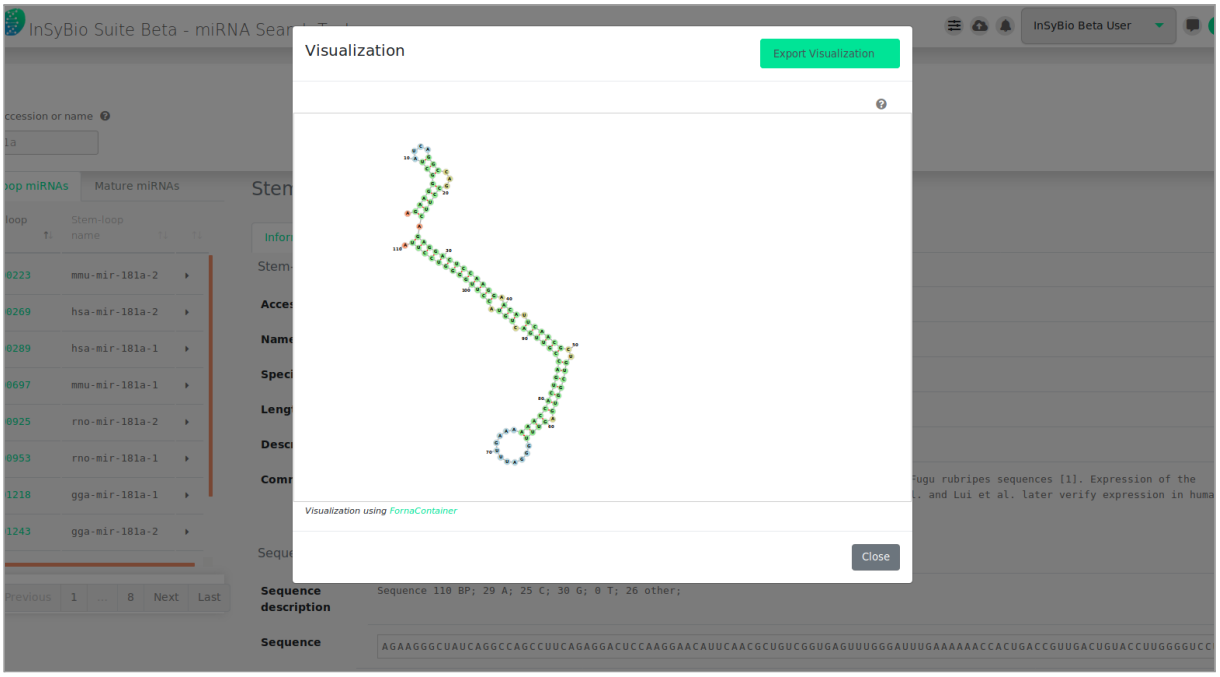
miRNA Search

You can search stem-loop and mature miRNAs giving a miRNA accession or name or part of them. Choosing the stem-loop or mature miRNA of those returned, its show page is shown.

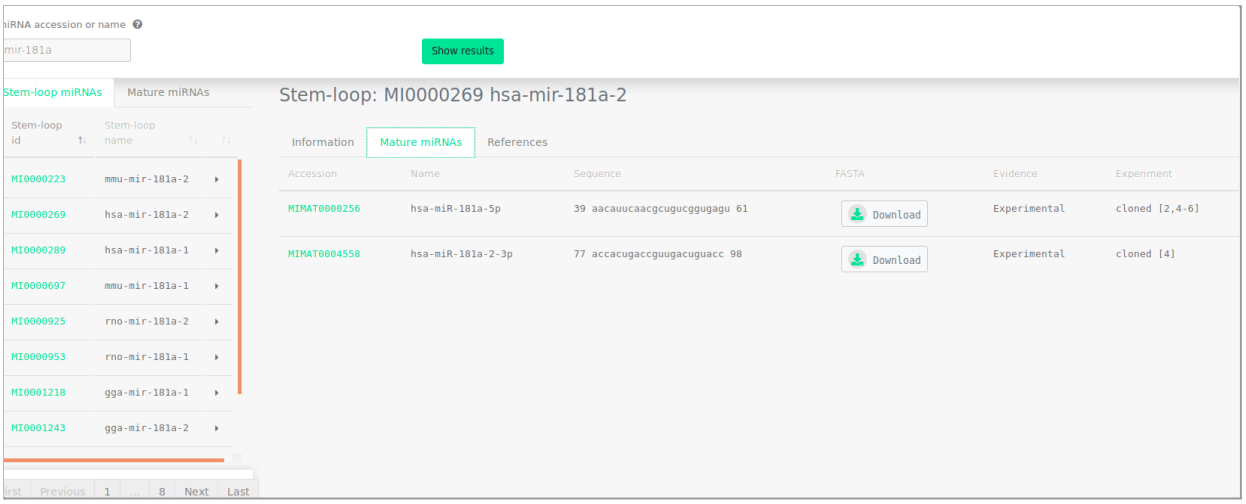
Stem-loop information



For the stem-loop you can view its accession, name, species, length, description and comments. Concerning its sequence, you can download the fasta format and view the sequence description, the sequence and the secondary structure in dot-bracket notation. You can view the visualization of the miRNA by clicking the “Visualization” button, this visualization of the secondary structure is performed with FornaContainer. It is the Minimum Free Energy (MFE) structure.



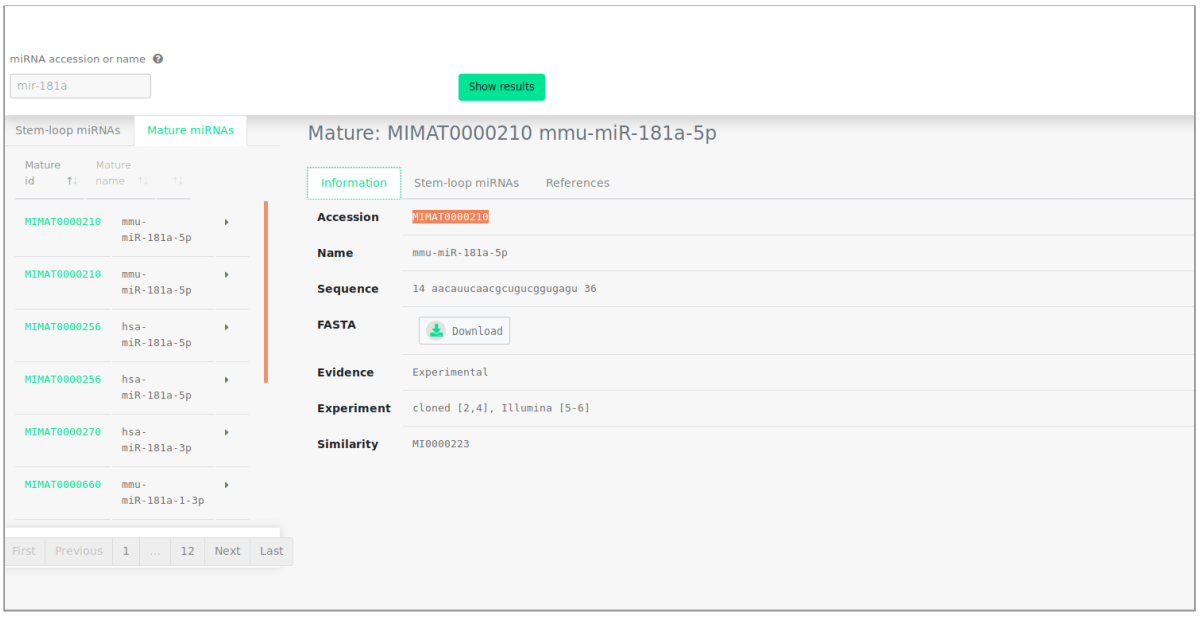
Mature miRNAs and references



For the mature miRNAs related to the stem-loop of interest you can view their accession, name and sequence. Concerning the sequence, you can download the fasta format. You can also view the evidence of each mature miRNA, which can be experimental, or by similarity of the related stem-loop to another stem-loop or found in literature.

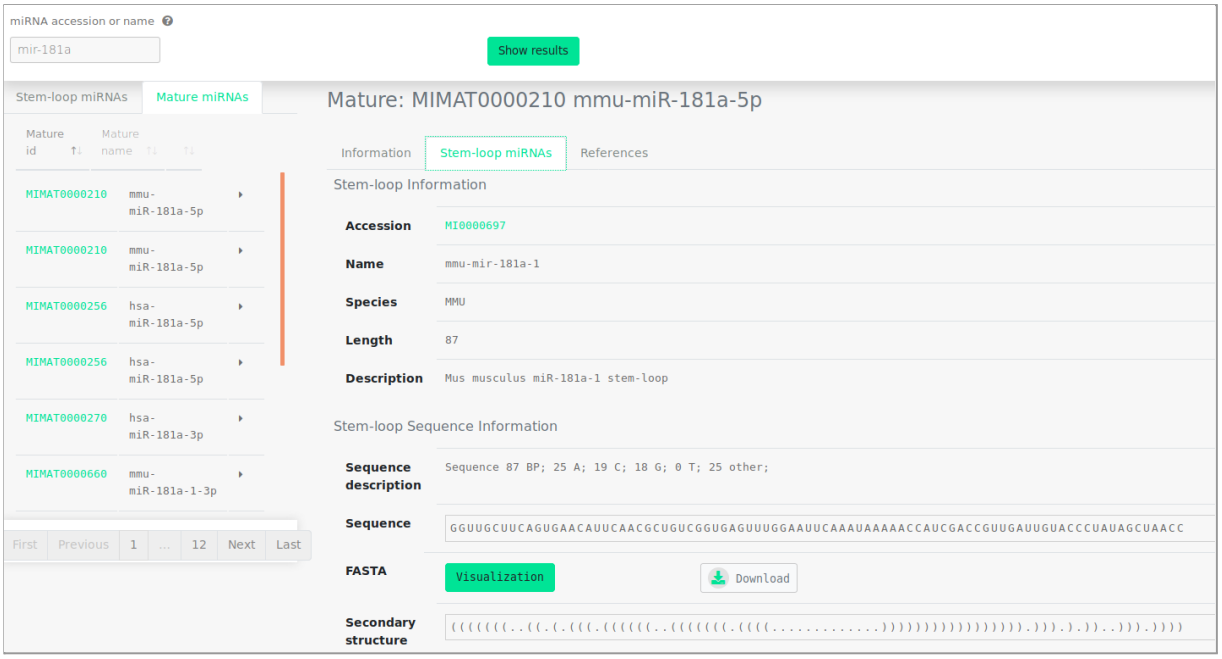
You can also view references for the miRNA of interest. There are external links to other databases (MIRBASE, ENTEZGENE, HGNC, RFAM, MGI, and WORMABASE) and publications.

Mature miRNA information

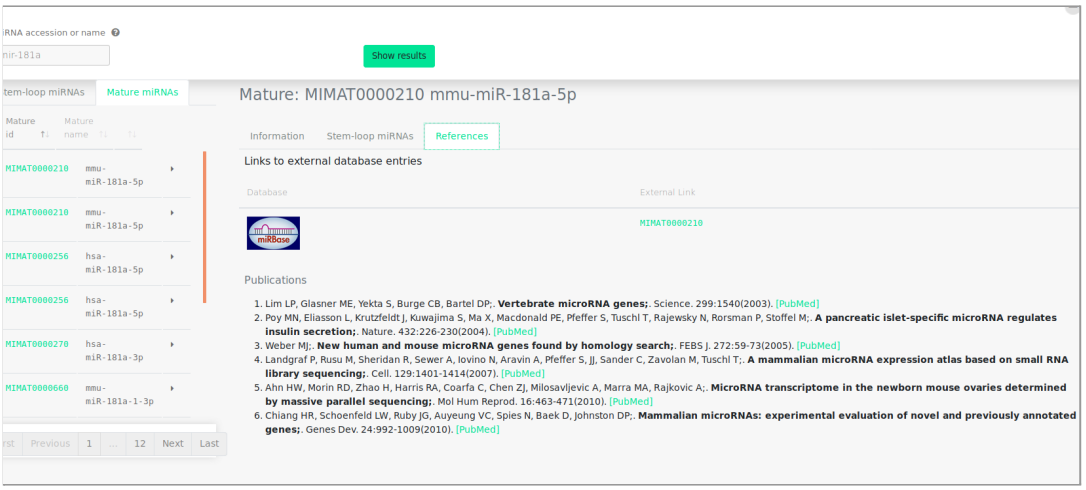


For the Mature miRNA you can view their accession, name and sequence. Concerning the sequence, you can download the fasta format. You can also view the

evidence of each mature miRNA, which can be experimental, or by similarity of the related stem-loop to another stem-loop or found in literature.



For the stem-loop related to the mature mi-RNA of interest you can view its accession, name, species, length, description and comments. Concerning its sequence, you can download the fasta format and view the sequence description, the sequence and the secondary structure in dot-bracket notation. You can view the visualization of the secondary structure by clicking the “Visualization” button, this visualization is performed with FornaContainer. It is the Minimum Free Energy (MFE) structure.

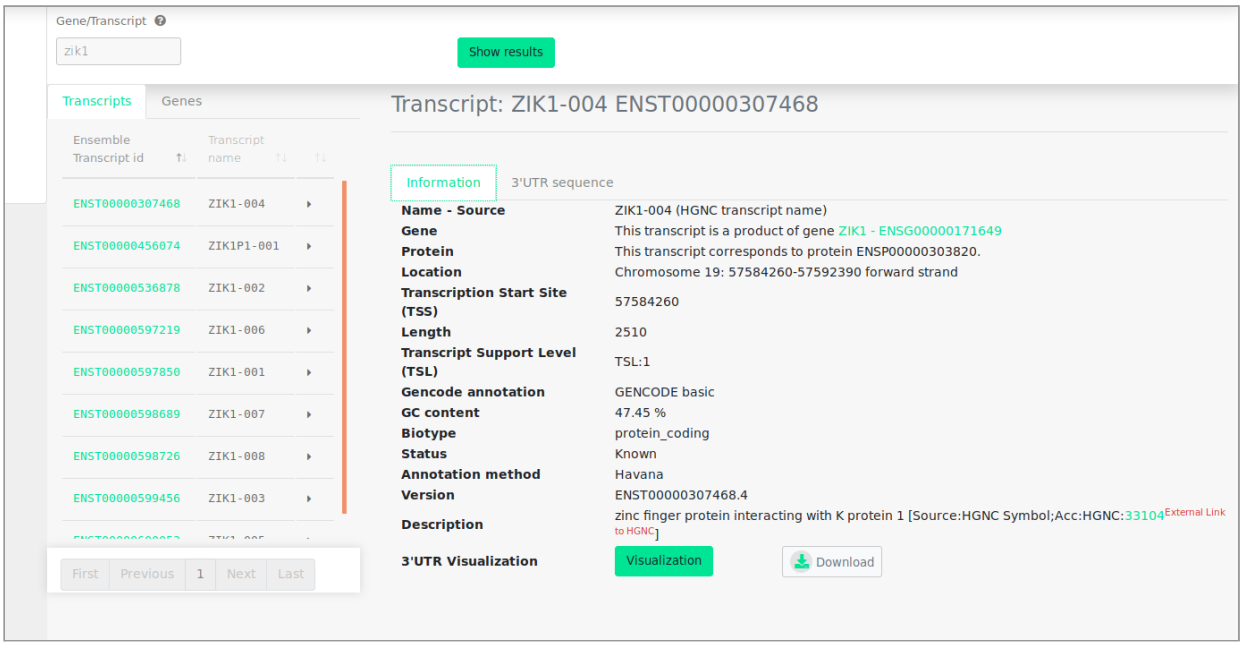


You can also view references for the mature miRNA of interest. There are external links to other databases (MIRBASE, ENTEZGENE, HGNC, RFAM, MGI, and WORMABASE) and publications.

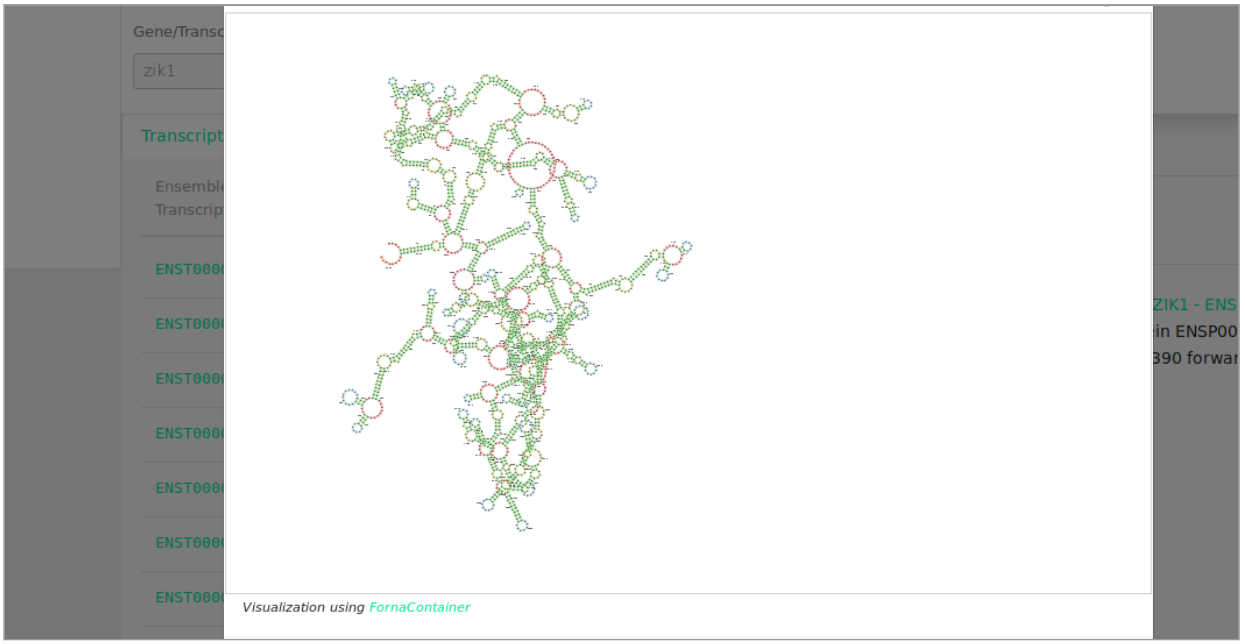
Transcript Search

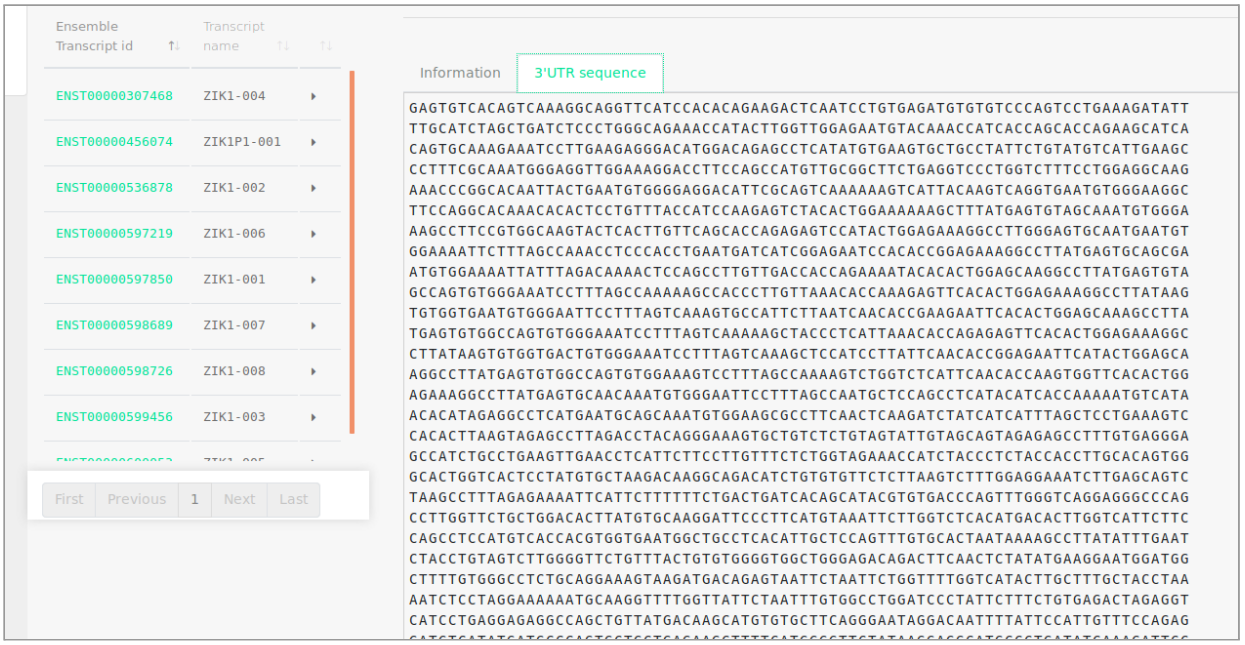
You can search transcripts and genes giving a transcript accession or name or part of them. Choosing the transcript or gene of those returned, its show page is shown.

Transcripts information

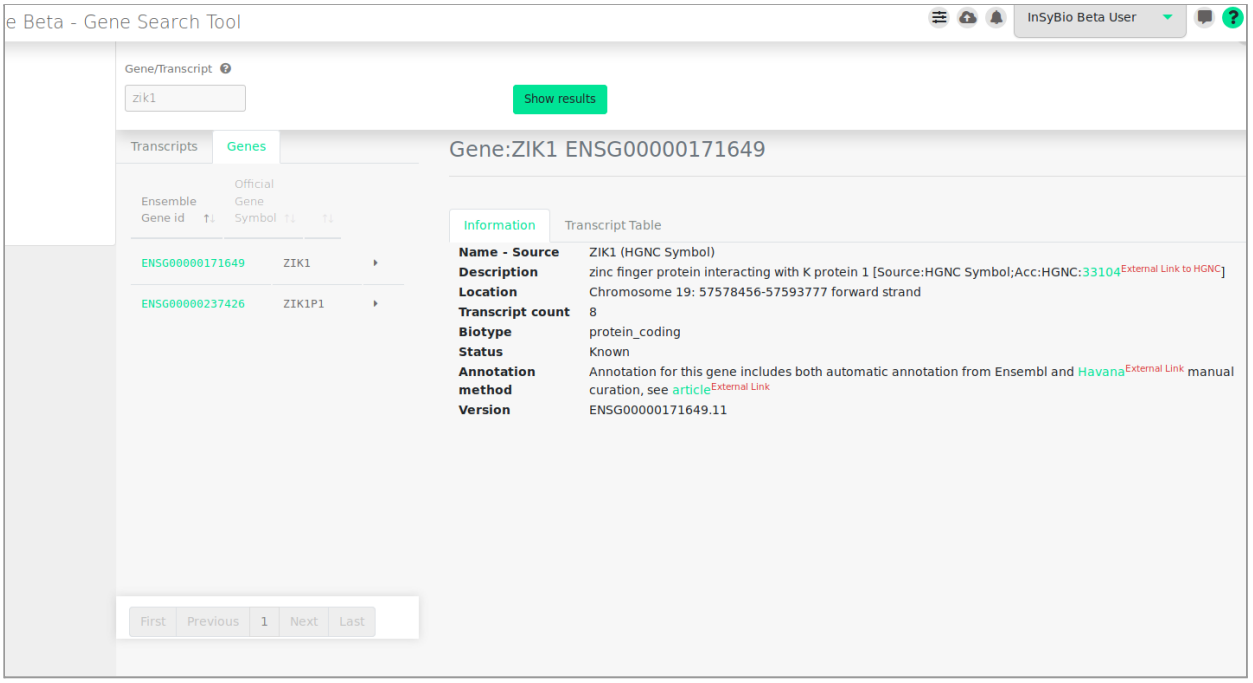


For the Transcript you can view its name-source, gene, protein, location, transcription start site (TSS), length, transcription support level (TSL), Gencode annotation, GC content, biotype, status, annotation method and version description. Concerning its 3'UTR sequence, you can download the fasta format and view the sequence description, the sequence and the secondary structure in dot-bracket notation. You can view the visualization of the secondary structure by clicking the "Visualization" button, this visualization of the secondary structure is performed with FornaContainer. It is the Minimum Free Energy (MFE) structure.

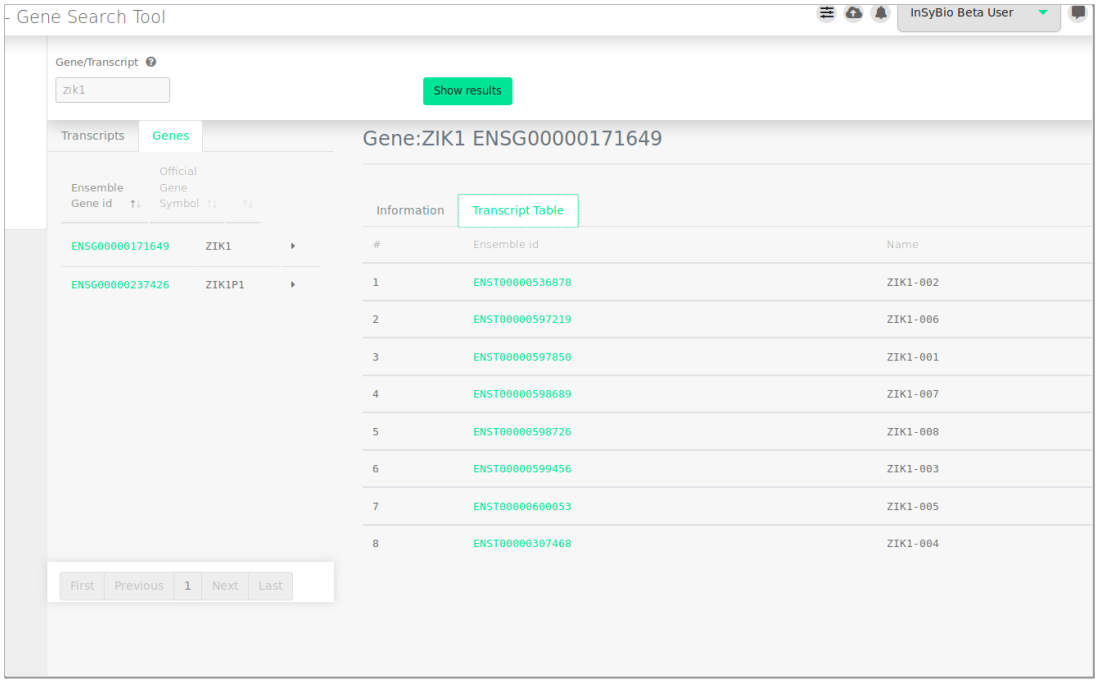

Gene/Transc
zik1
Transcript
Visualization using FornaContainer


Ensemble Transcript id
Transcript name
ENST00000307468 ZIK1-004
ENST00000456074 ZIK1P1-001
ENST00000536878 ZIK1-002
ENST00000597219 ZIK1-006
ENST00000597850 ZIK1-001
ENST00000598689 ZIK1-007
ENST00000598726 ZIK1-008
ENST00000599456 ZIK1-003
First Previous 1 Next Last
Information
3'UTR sequence
GAGTGTCACAGTCAAAGGCAGGTTCATCCACACAGAAGACTCAATCCTGTGAGATGTGTGTCCCAGTCCTGAAAGATATT
TTGCATCTAGCTGATCTCCCTGGGCAGAAACCATACTTGGTTGGAGAATGTACAAACCATCACCAGCACCAGAAGCATCA
CAGTGCAAAGAAATCCTTGAAGAGGGACATGGACAGAGCCTCATATGTGAAGTGCTGCCTATTCTGTATGTCATTGAAGC
CCTTTCGCAAATGGGAGGTTGGAAAGGACCTTCCAGCCATGTTGCGGCTTCTGAGGTCCCTGGTCTTTCCTGGAGGCAAG
AAACCCGGCACAATTACTGAATGTGGGGAGGACATTCGCAGTCAAAAAAGTCATTACAAGTCAGGTGAATGTGGGAAGGC
TTCCAGGCACAAACACACTCCTGTTTACCATCCAAGAGTCTACACTGGAAAAAAGCTTTATGAGTGTAGCAAATGTGGGA
AAGCCTTCCGTGGCAAGTACTCACTTGTTCAGCACCAGAGAGTCCATACTGGAGAAAGGCCTTGGGAGTGCAATGAATGT
GGAAAATTCTTTAGCCAAACCTCCCACCTGAATGATCATCGGAGAATCCACACCGGAGAAAGGCCTTATGAGTGCAGCGA
ATGTGGAAAATTATTTAGACAAAACTCCAGCCTTGTTGACCACCAGAAAATACACACTGGAGCAAGGCCTTATGAGTGTA
GCCAGTGTGGGAAATCCTTTAGCCAAAAAGCCACCCTTGTTAAACACCAAAGAGTTCACACTGGAGAAAGGCCTTATAAG
TGTGGTGAATGTGGGAATTCCTTTAGTCAAAGTGCCATTCTTAATCAACACCGAAGAATTCACACTGGAGCAAAGCCTTA
TGAGTGTGGCCAGTGTGGGAAATCCTTTAGTCAAAAAGCTACCCTCATTAAACACCAGAGAGTTCACACTGGAGAAAGGC
CTTATAAGTGTGGTGACTGTGGGAAATCCTTTAGTCAAAGCTCCATCCTTATTCAACACCGGAGAATTCATACTGGAGCA
AGGCCTTATGAGTGTGGCCAGTGTGGAAAGTCCTTTAGCCAAAAGTCTGGTCTCATTCAACACCAAGTGGTTCACACTGG
AGAAAGGCCTTATGAGTGCAACAAATGTGGGAATTCCTTTAGCCAATGCTCCAGCCTCATACATCACCAAAAATGTCATA
ACACATAGAGGCCTCATGAATGCAGCAAATGTGGAAGCGCCTTCAACTCAAGATCTATCATCATTTAGCTCCTGAAAGTC
CACACTTAAGTAGAGCCTTAGACCTACAGGGAAAGTGCTGTCTCTGTAGTATTGTAGCAGTAGAGAGCCTTTGTGAGGGA
GCCATCTGCCTGAAGTTGAACCTCATTCTTCCTTGTTTCTCTGGTAGAAACCATCTACCCTCTACCACCTTGCACAGTGG
GCACTGGTCACTCCTATGTGCTAAGACAAGGCAGACATCTGTGTGTTCTCTTAAGTCTTTGGAGGAAATCTTGAGCAGTC
TAAGCCTTTAGAGAAAATTCATTCTTTTTTCTGACTGATCACAGCATACGTGTGACCCAGTTTGGGTCAGGAGGGCCCAG
CCTTGGTTCTGCTGGACACTTATGTGCAAGGATTCCCTTCATGTAAATTCTTGGTCTCACATGACACTTGGTCATTCTTC
CAGCCTCCATGTCACCACGTGGTGAATGGCTGCCTCACATTGCTCCAGTTTGTGCACTAATAAAAGCCTTATATTTGAAT
CTACCTGTAGTCTTGGGGTTCTGTTTACTGTGTGGGGTGGCTGGGAGACAGACTTCAACTCTATATGAAGGAATGGATGG
CTTTTGTGGGCCTCTGCAGGAAAGTAAGATGACAGAGTAATTCTAATTCTGGTTTTGGTCATACTTGCTTTGCTACCTAA
AATCTCCTAGGAAAAAATGCAAGGTTTTGGTTATTCTAATTTGTGGCCTGGATCCCTATTCTTTCTGTGAGACTAGAGGT
CATCCTGAGGAGAGGCCAGCTGTTATGACAAGCATGTGTGCTTCAGGGAATAGGACAATTTTATTCCATTGTTTCCAGAG

Genes information



For the Genes you can view its name-source, description, location, transcript count, biotype, status, annotation method and version. Also a Transcript Table is provided with the genes associated transcripts and links to their information.



Rna-Seq Differential Expression Pipeline

You can calculate the differential expression between two RNA-Seq experiments. It uses FastQC and Trimmomatic for Quality Control, HISAT2 for Alignment, FeatureCounts for Quantification and DESeq2 for Differential Expression analysis. The Rna-Seq Differential Expression we have implemented consists of 4 steps:

A. Quality Control using FastQC and Filtering using Trimmomatic (Optional step).
B. Alignment using HISAT2, and sorting with Samtools.
C. Quantification using FeatureCounts.
D. Differential Expression using Deseq2.

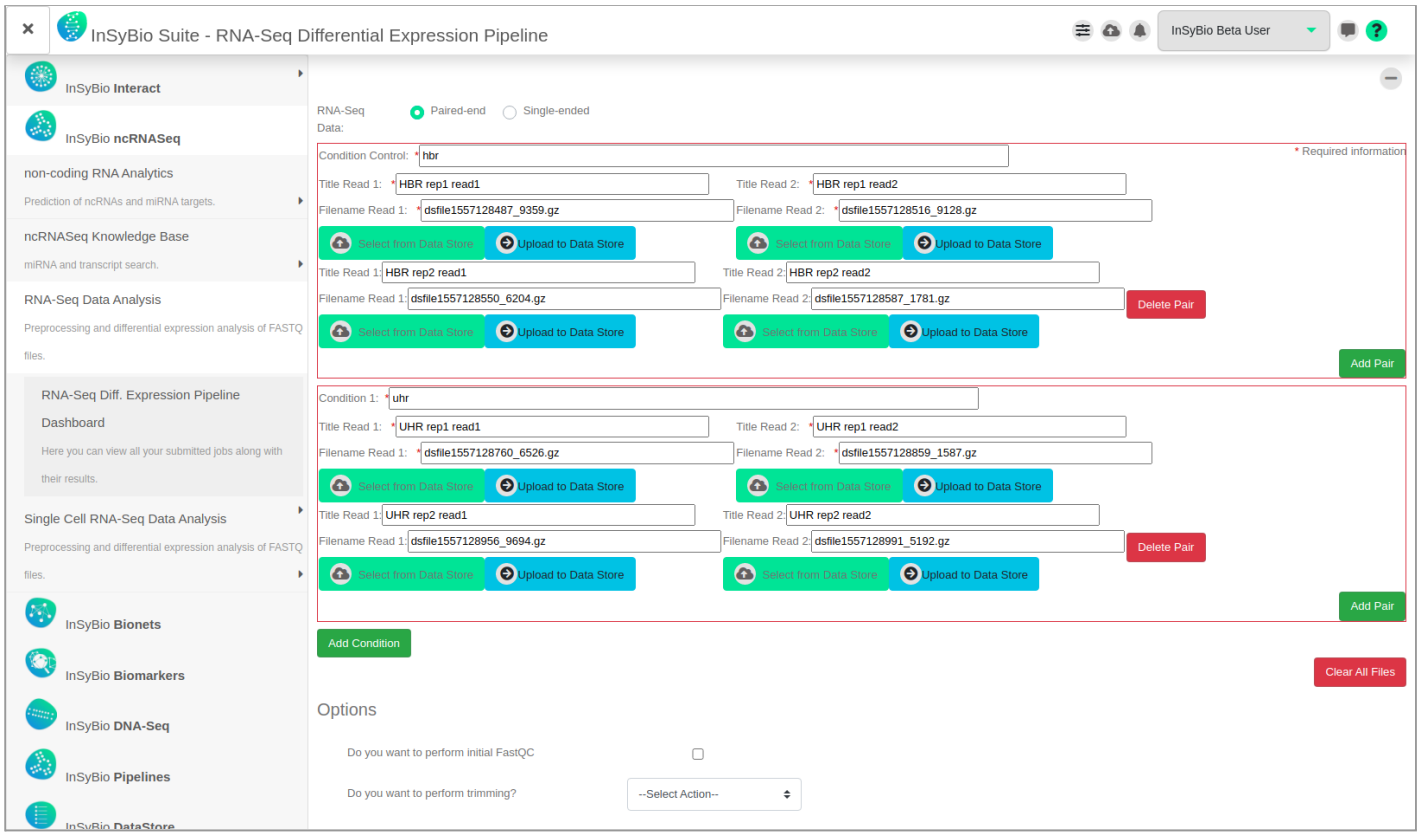
Firstly, the Pipeline uses Fastqc to create a report with the sequences quality, then trimm the sequences accordingly using Trimmomatic and create new reports with Fastqc. Then using HISAT2 it creates the alignment SAM files, we sort them using SAMtools and transform them to BAM files. The BAM files are used as input of FeatureCounts, that creates text files with the quantity of each gene. At the end, DESeq2 performs Differential Expression Analysis for all the pairs of conditions using R.

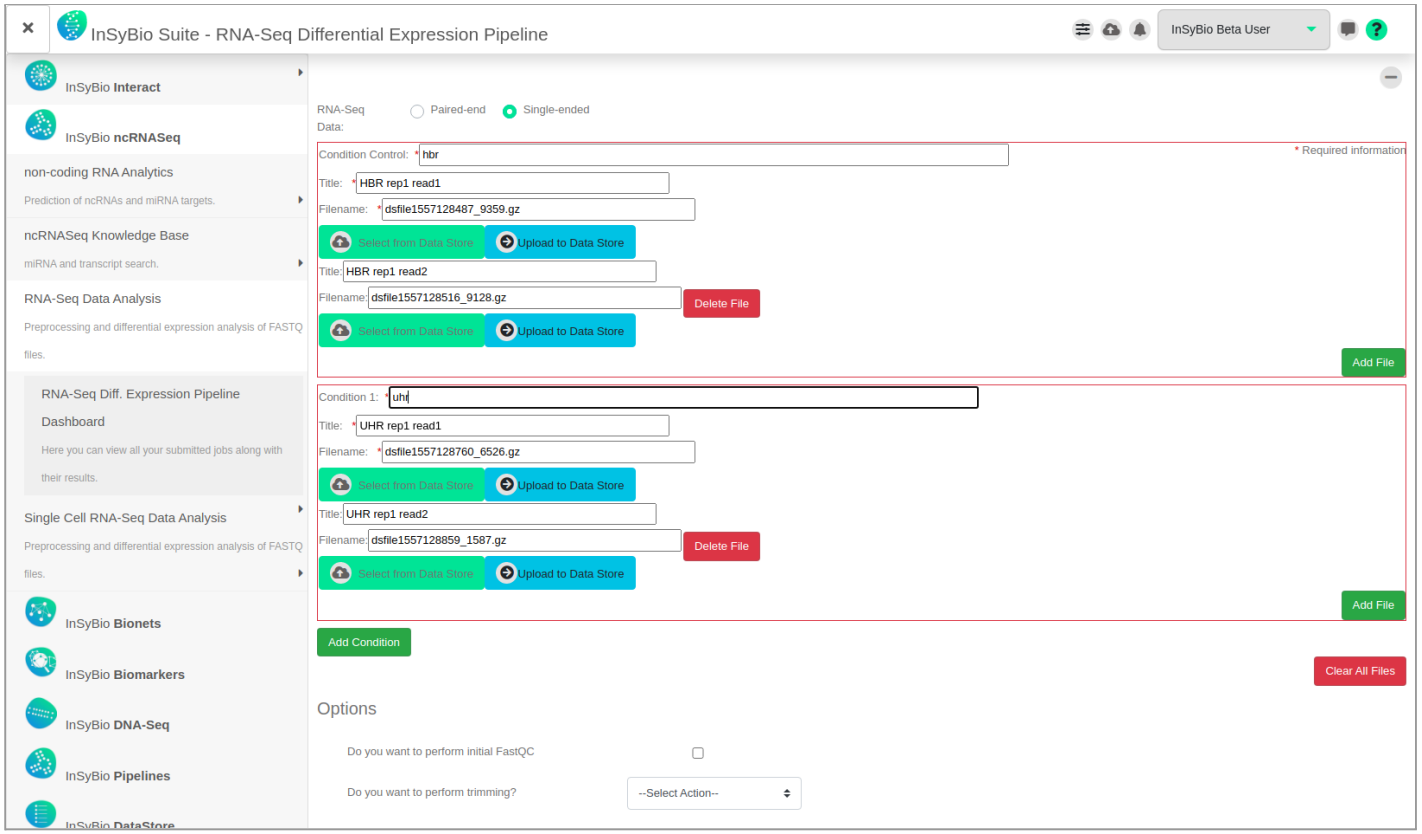
We also offer a modification to the above pipeline, called ncRNA-Seq Differential Expression Pipeline, where the unaligned reads from the Alignment step are used to enhance the quantification files with known or predicted ncRNAs. This is done by finding all the contigs of the unaligned reads files using the AbySS Assembler, and then either check if these contigs are known ncRNAs (from a list of 6 ncRNA types: miRNA, pre-miRNA, tRNA, rRNA, snoRNA and tRf) or use our novel method of an EnsembleGASVR Classifier to predict if the contigs are possible ncRNAs. Then the quantity of the known and predicted ncRNAs is used to enhance the quantification files produced by featureCounts and continue the pipeline as described above.

To start the differential expression:

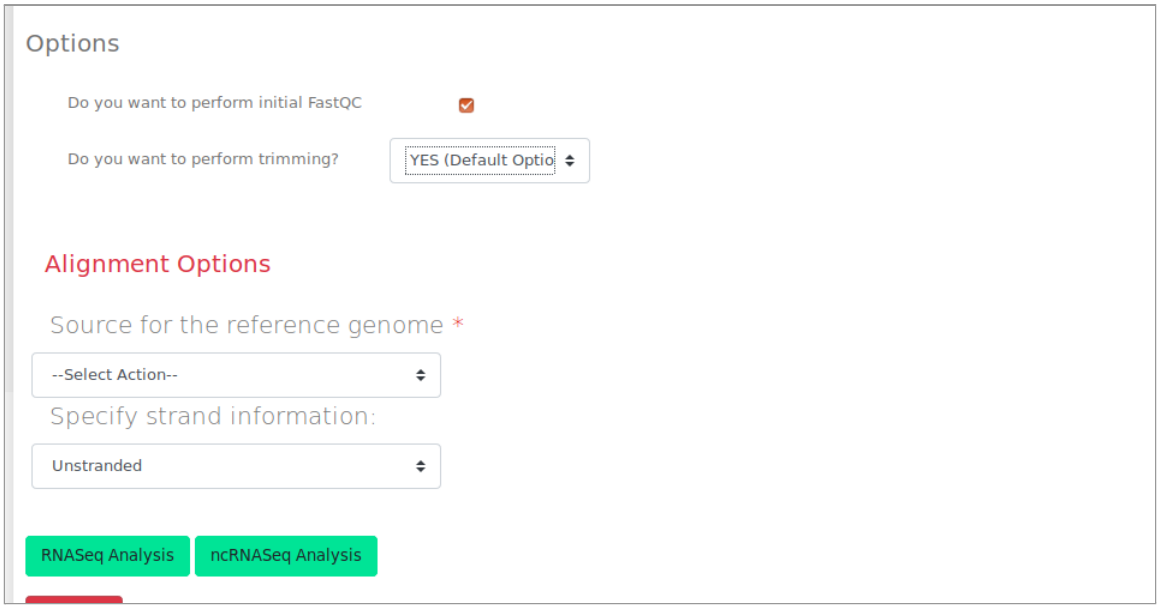
Click in the menu “InSyBio ncRNASeq” → “RNA-Seq Data Analysis” → “RNA-Seq Diff. Expression Pipeline Dashboard” , select the “Add new job” button and then:

- Select if you have Paired or Single Ended data.

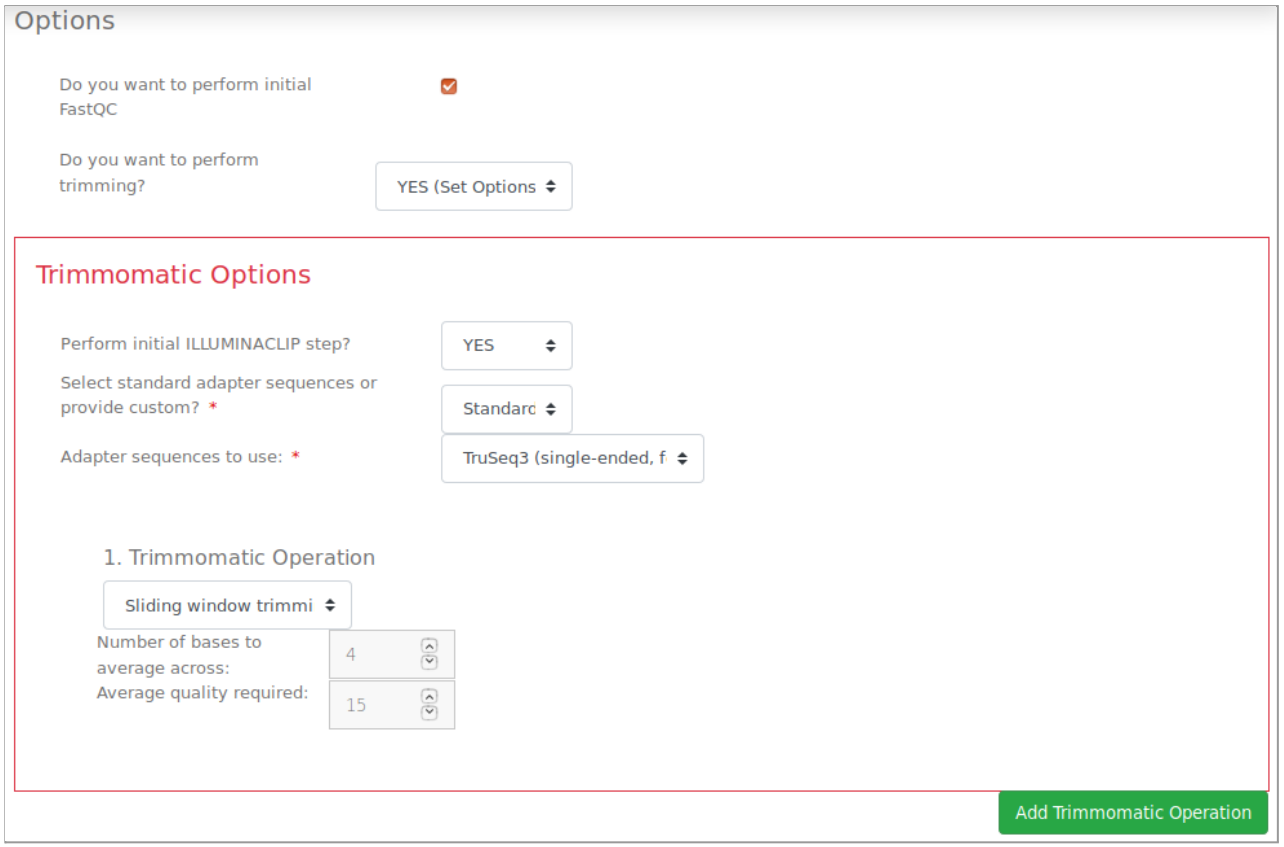




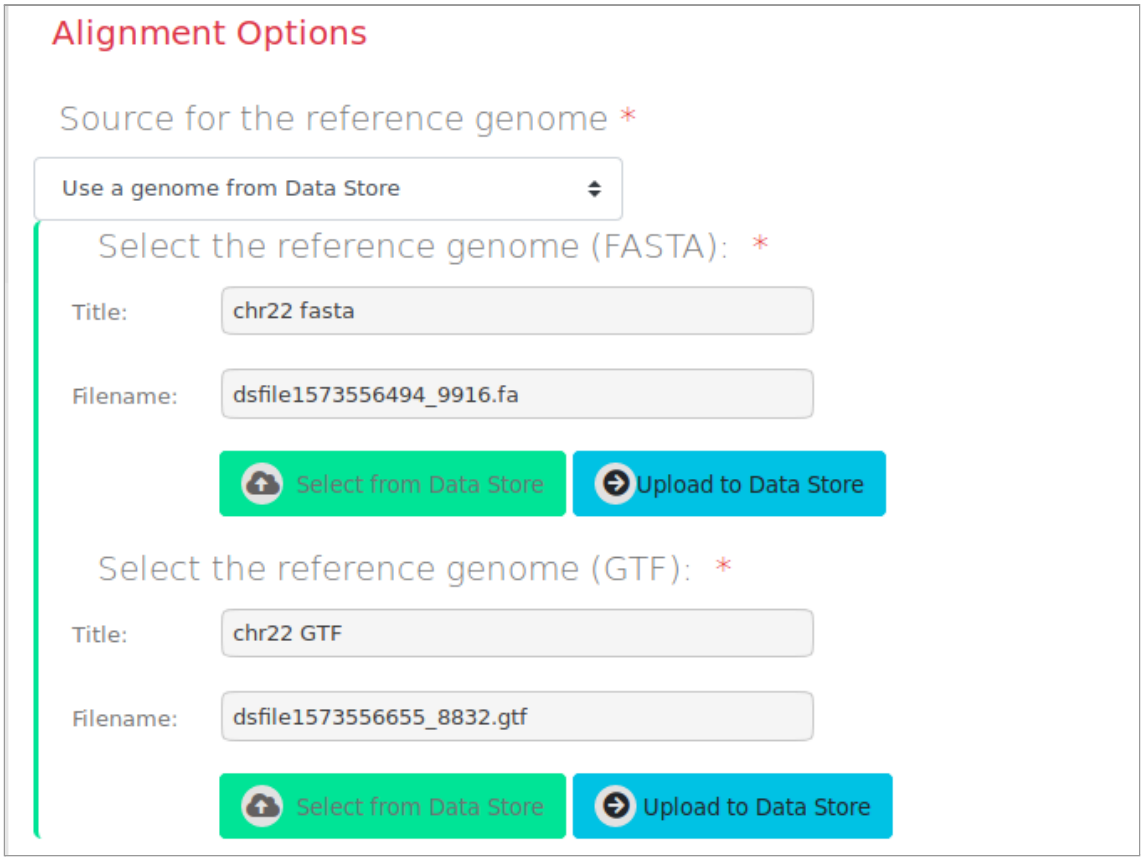
- Name Conditions/Group of files you want to compare.
- For each condition add single or paired files by:
    - Uploading a new file of Rna-Seq Experiments in fastq format. You are redirected to the Data Store where step by step instructions guide you for both files uploading.
    - Or Selecting a file of Rna-Seq Experiments in fastq format from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.
- Select if you want to perform FastQC Quality Control to the initial Data.

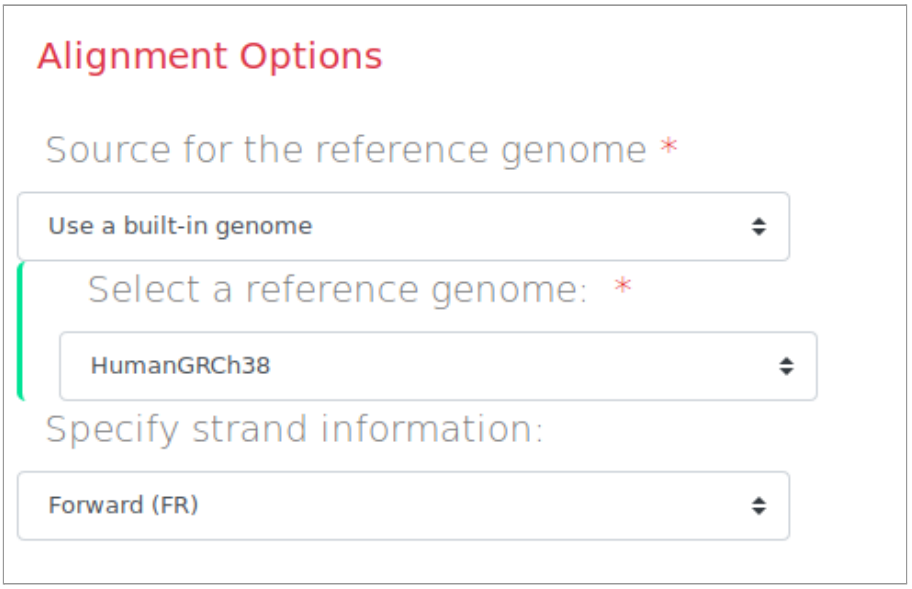


- Select if you want to perform trimming of the data with Trimmomatic, either with our Default Options or add your own (If trimming is selected FastQC will be performed to the trimmed data). Possible manual options are to:
    - Perform initial ILLUMINACLIP step
        - With Standard adapters (TrueSeq2,TrueSeq3 or Nextera for paired or single ended)
        - Or With Custom adapters in fasta format
    - Perform sliding window trimming
    - Drop reads below a specific length
    - Cut bases off the start of a read, if below a threshold quality
    - Cut bases off the end of a read, if below a threshold quality
    - Cut the read to a specified length
    - Cut the specified number of bases from the start of the read
    - Drop the read if the average quality is below a specified value
    - Trim reads adaptively, balancing read length and error rate to maximise the value of each read



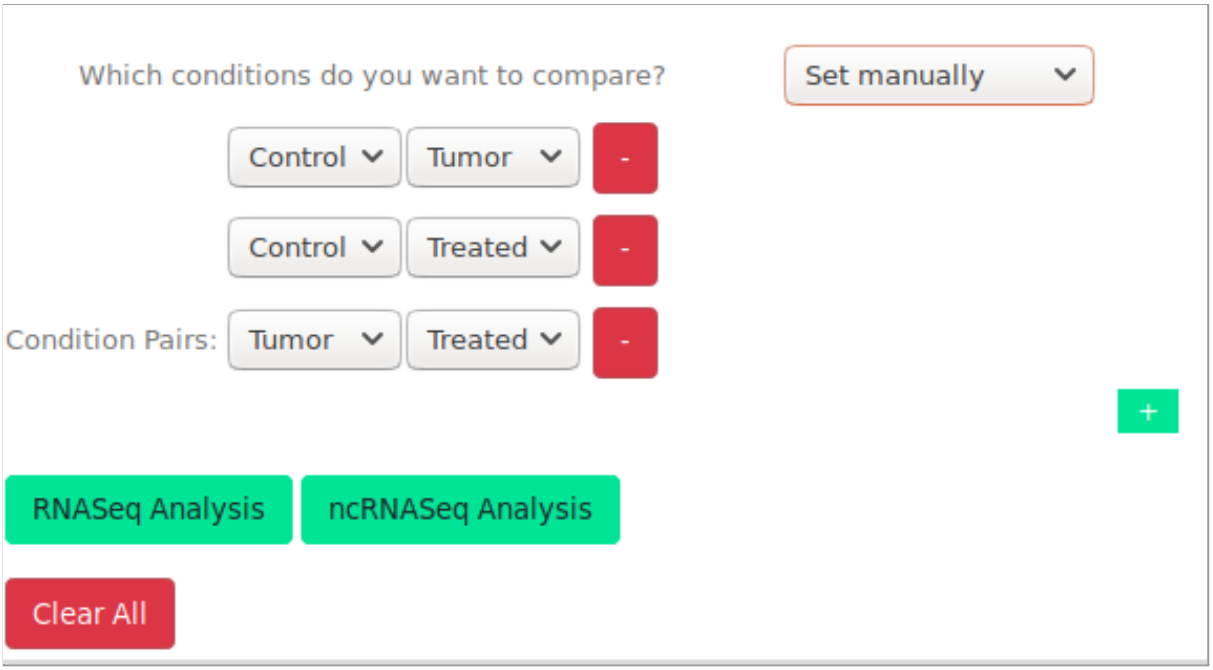
- Select the Genome the input files belong, either from our 4 built-in options (HumanGRCh37, HumanGRCh38, MouseGRCm38 and ZebrafishGRCz11), or
    - Upload new reference Genome files in fasta and gtf format. You are redirected to the Data Store where step by step instructions guide you for both files uploading.
    - Or Select two reference Genome files one in fasta and one in gtf format from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.





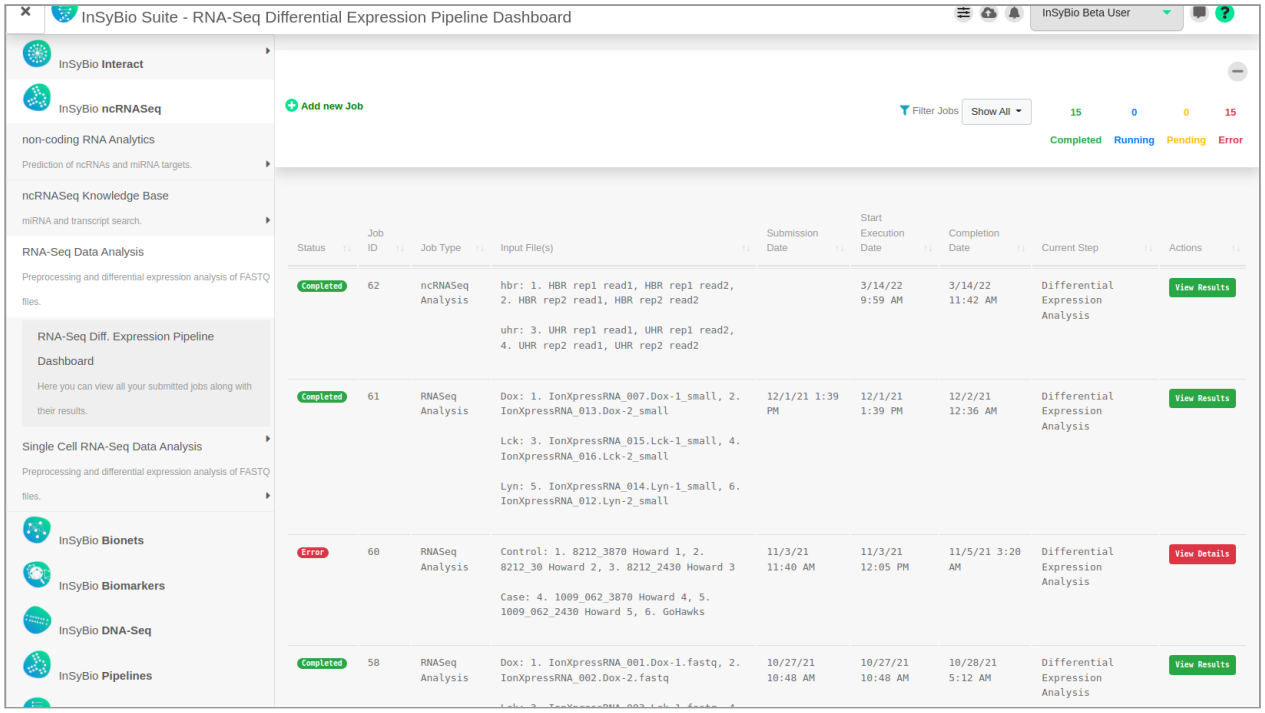
- Select the strandness of your input files, Unstranded, Forward or Reverse.
- If more than 2 Conditions are selected, you can select which pairs of conditions to Differentially Express (all versus Control, all versus all or assign manually).

- Last but not least select either to perform the regular RNASeq Differential Expression Pipeline or the enhanced ncRNASeq Differential Expression Pipeline.

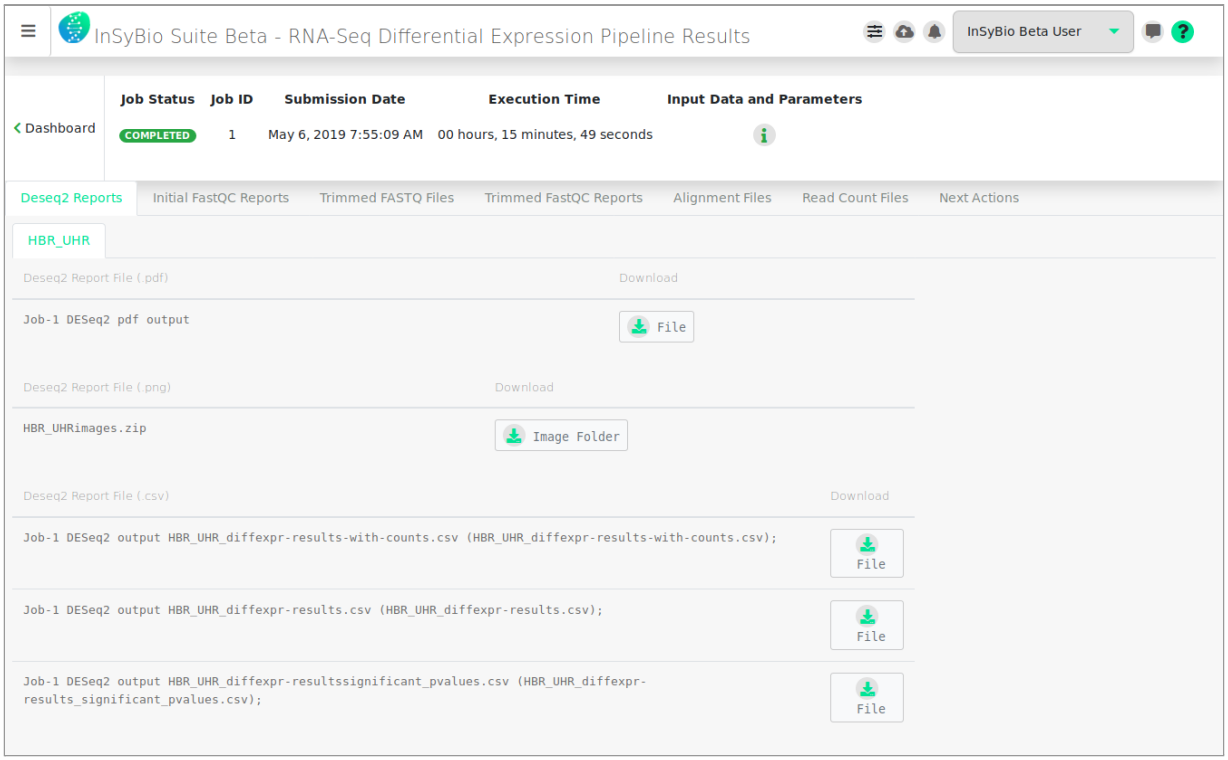


To view the results:

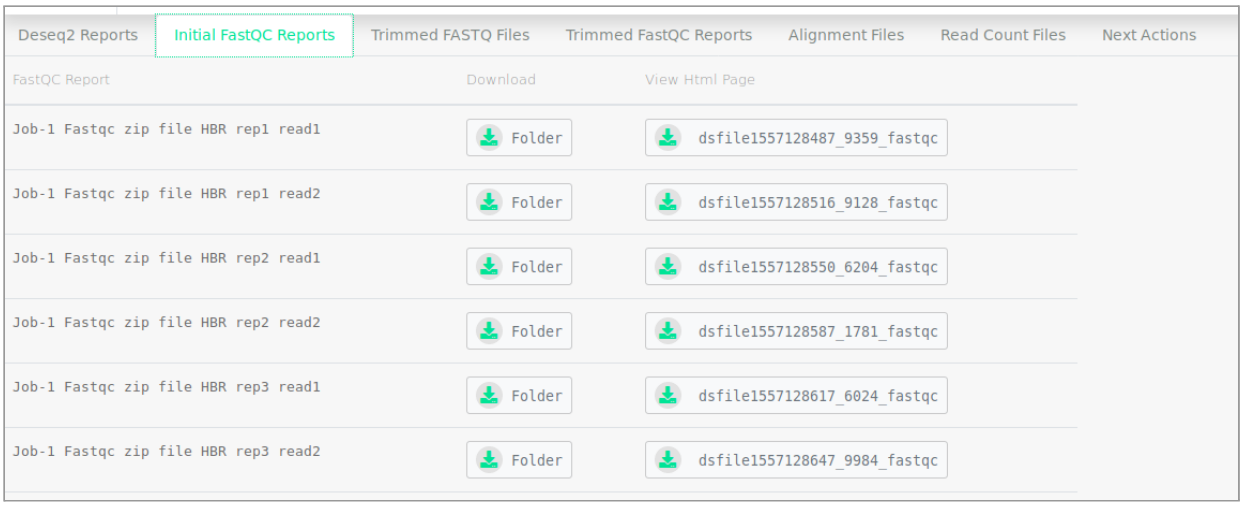
By starting a calculation you are informed if it was submitted successfully. Then you can move to the Rna-Seq Differential Expression Pipeline and view the Dashboard, where you can view the status of your current and previous Rna-Seq Differential Expression jobs.



At completion of the Analysis you can select the View Results at the Actions column and view the produced files, that are separated according to the step they were produced.

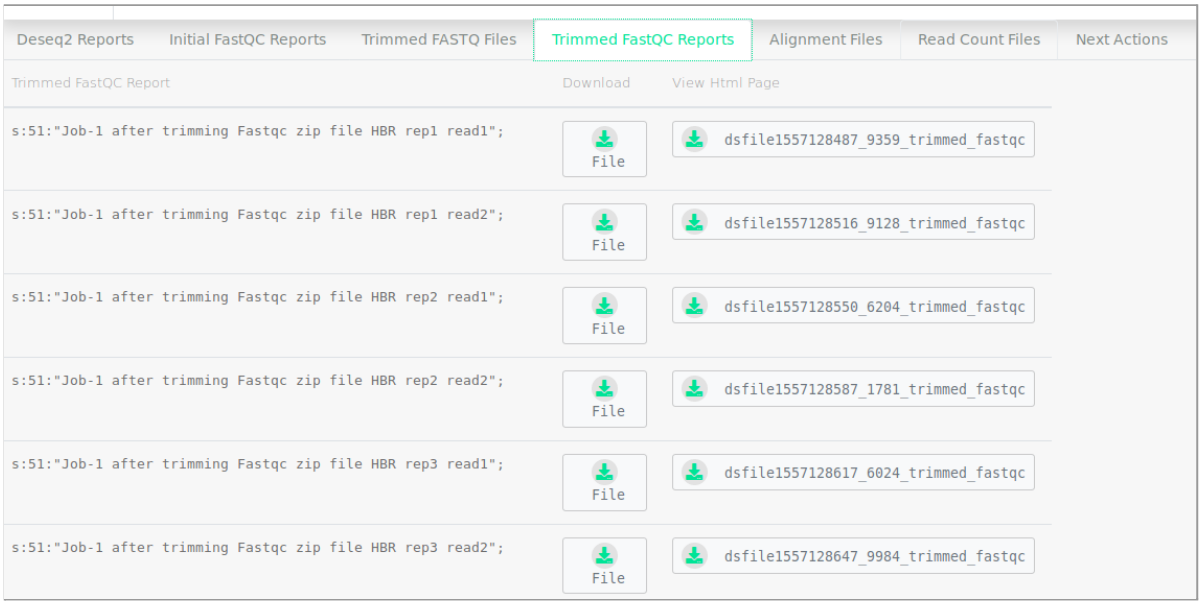


In Deseq2 reports tab you can download visual information and the Differential Expression calculated values for each pair compared.

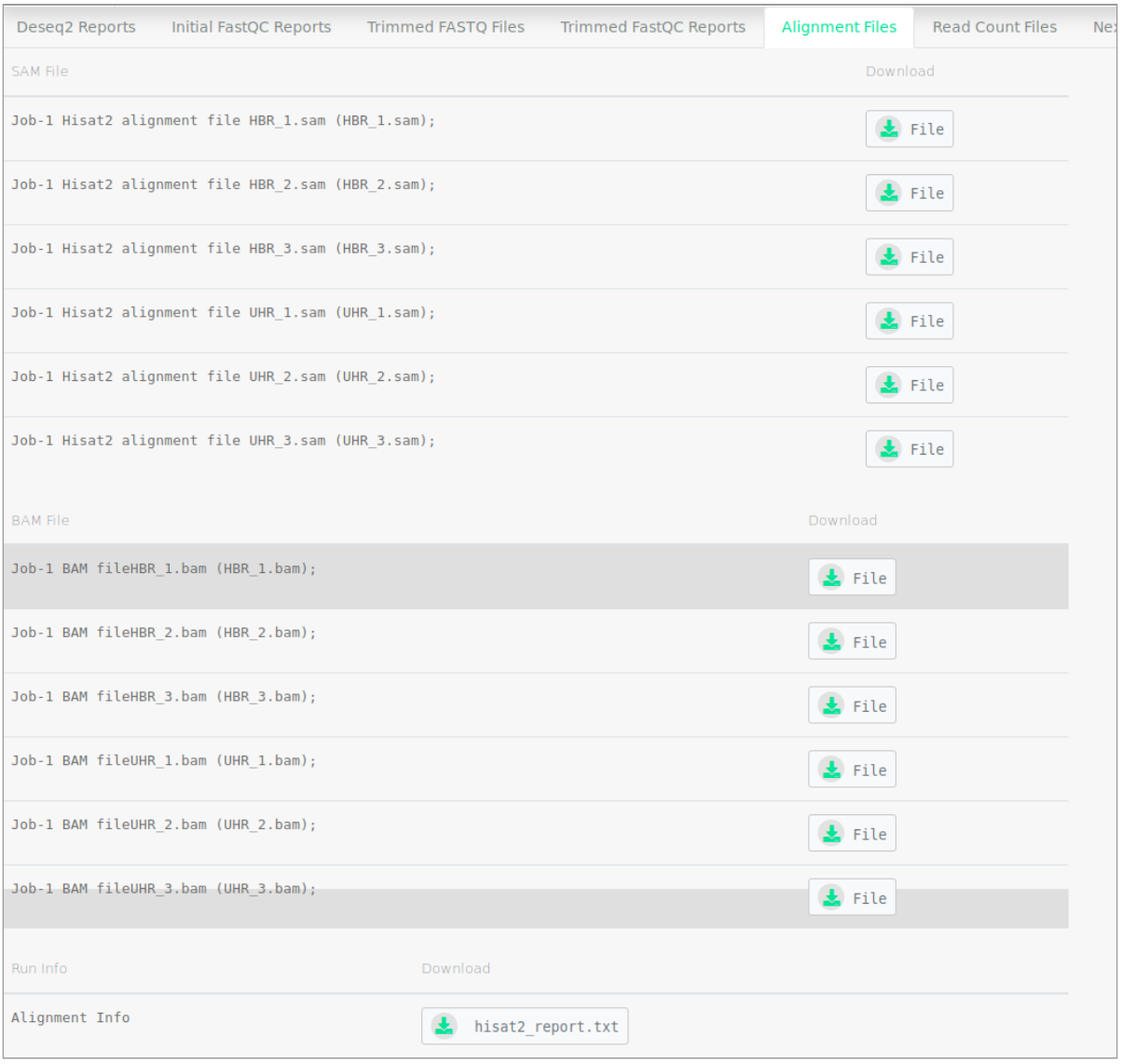


In the Initial FastQC reports the FastQC reports of the input files can be downloaded.

In the Trimmed FASTQ Files, the output Fastq files after trimming can be downloaded.

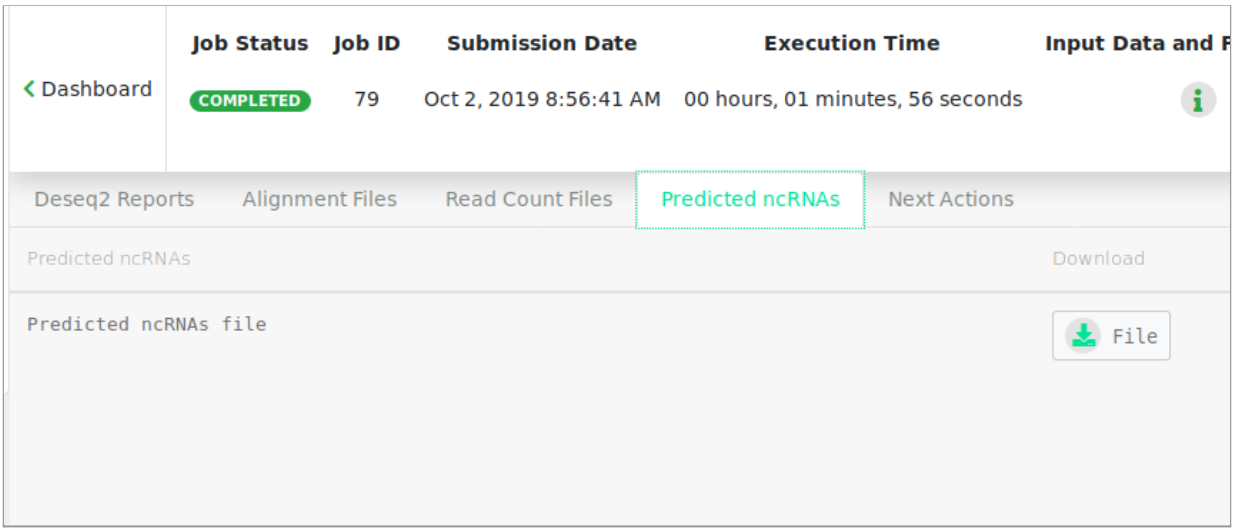


In the Trimmed FastQC reports the FastQC reports of the trimmed files can be downloaded.

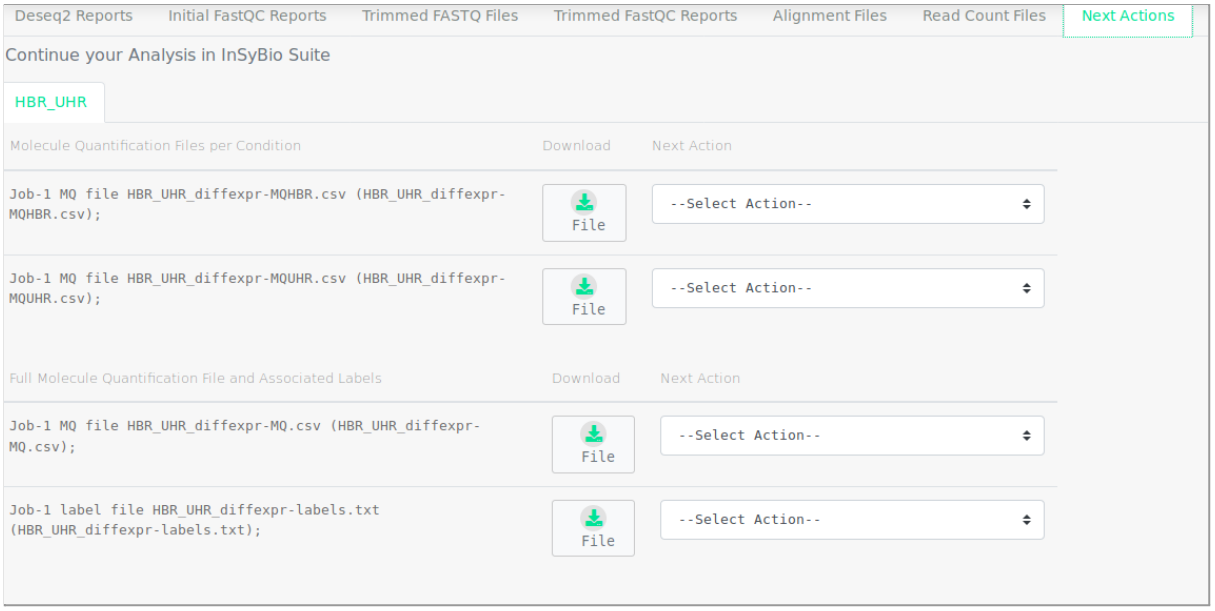


In the Alignment files tab, the HISAT2 alignment sam and bam files can be downloaded.

In the Read Count Files tab the Count files for each sample can be downloaded.



If ncRNASeq Analysis is selected in the Predicted ncRNAs tab a tsv file with the found ncRNAs in the unaligned file is provided, with its name and predicted labels can be downloaded.



In the Next Action tab, Molecule Quantifications files, with the 10% most significant genes, for each comparison are provided. They can be downloaded or used as input in **InSyBio Bionets,** to construct gene correlation networks with the gene expressions of the genes found as statistically significantly differential expressed, and in **InSyBio Biomarkers,** to perform additional statistical analysis and built a classification model able to predict to which of the two conditions a potential new sample belongs.

Single Cell Rna-Seq Differential Expression Pipeline

You can analyze Single Cell RNA-Seq experiments. Alignment, read counts computation and additional secondary analysis are all performed in one job. Depending on the selected workflow, the Single Cell Rna-Seq Differential Expression pipeline consists of the following 2 or 3 steps:

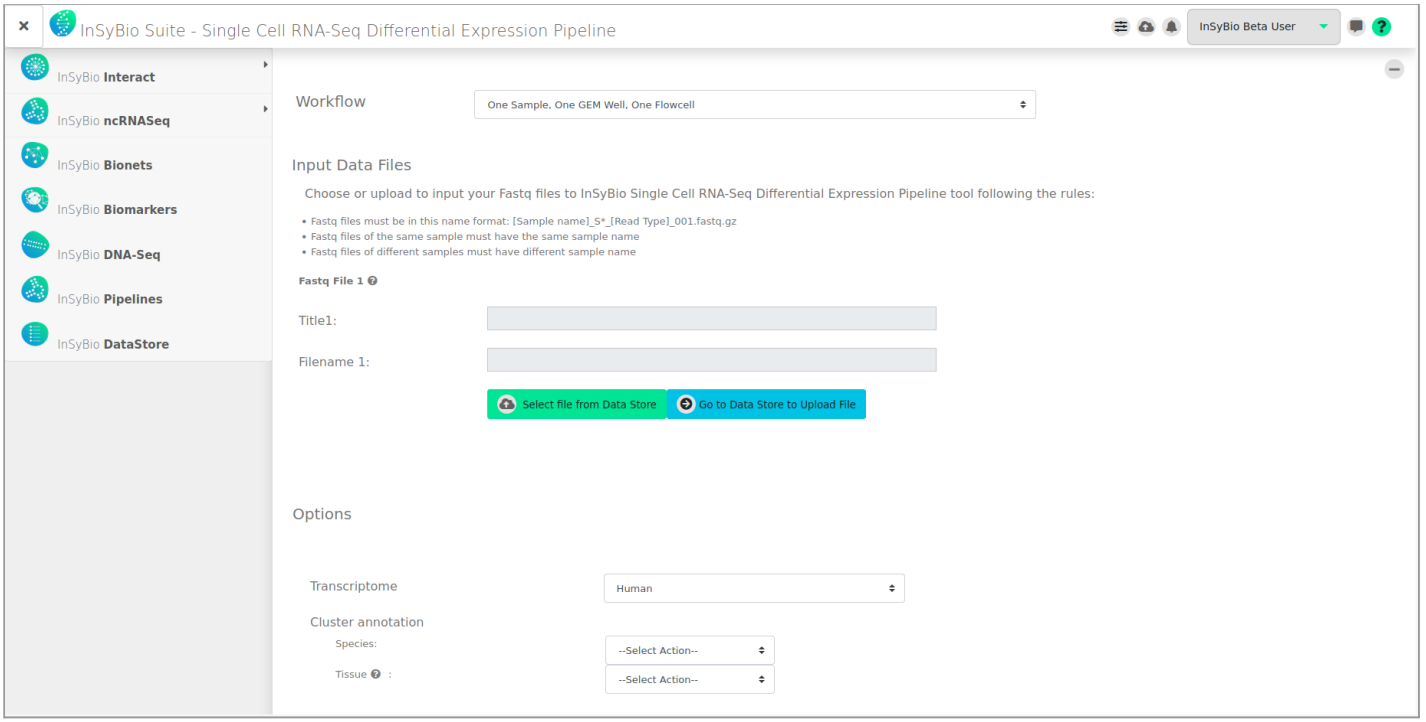
- Workflow 0 or 1:
  - Alignment and read counts computation using Cellranger count.
  - Further analysis using our Single Cell Analysis.
- Workflow 2 or 3:
  - Alignment and read counts computation using Cellranger count pipeline for each different sample or different GEM well.
  - Aggregation of the Cellranger count runs using Cellranger aggr pipeline.
  - Further analysis using our Single Cell Analysis.

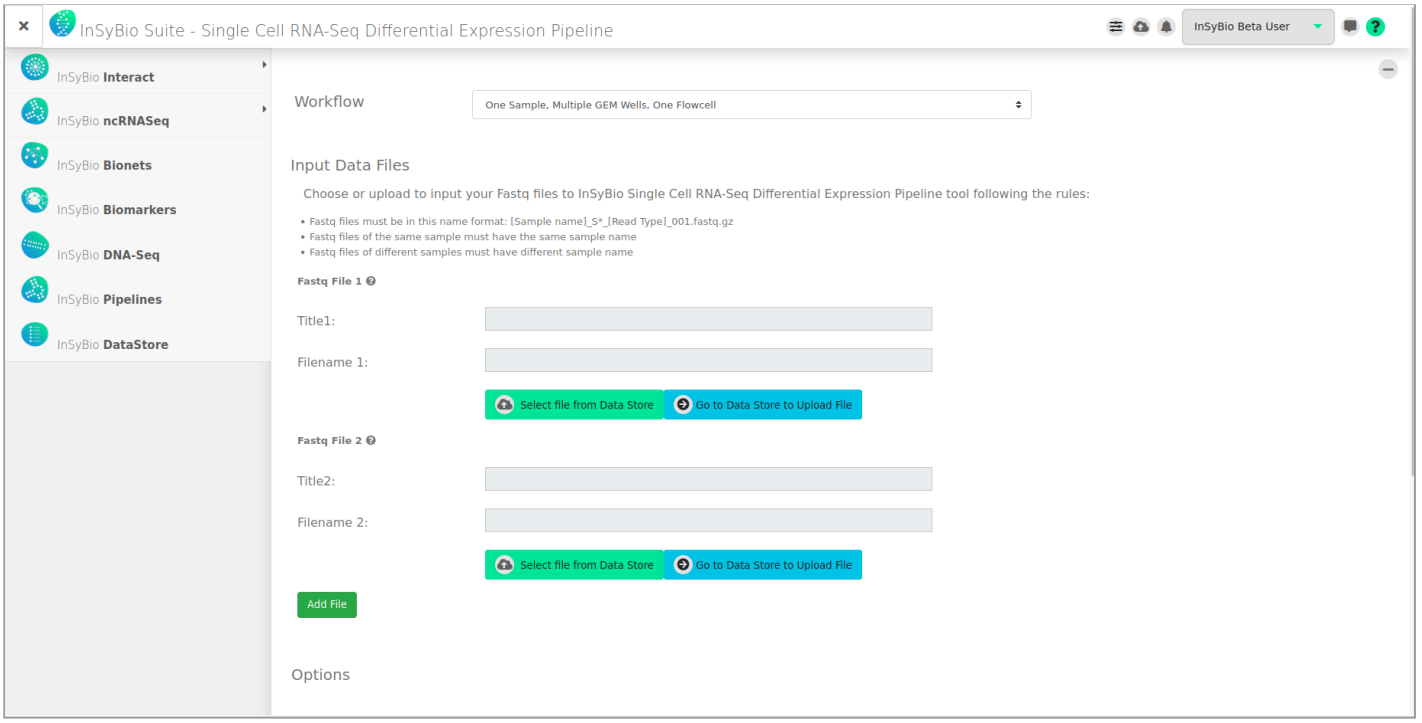
Firstly, the Pipeline uses the Cellranger count pipeline to perform the alignment and the read counts computation of the input fastq files. If the input fastq files are generated from different samples or different GEM wells, an extra step is performed. Specifically, the Cellranger aggr pipeline is used to aggregate the cellranger count runs for the creation of a single feature-barcode matrix and analysis. At the end, our Single Cell Analysis script is used to perform additional secondary differential expression analysis.

To start the single cell differential expression:

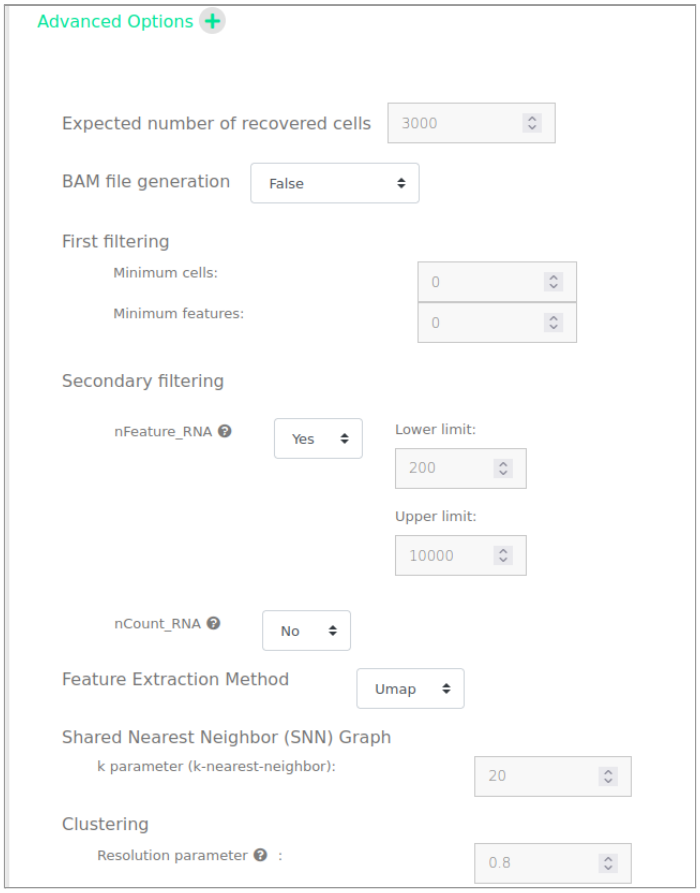
Click in the menu “InSyBio ncRNASeq” → “Single Cell RNA-Seq Data Analysis” → “RNA-Seq Single Cell Pipeline Dashboard” , select the “Add new job” button and then:

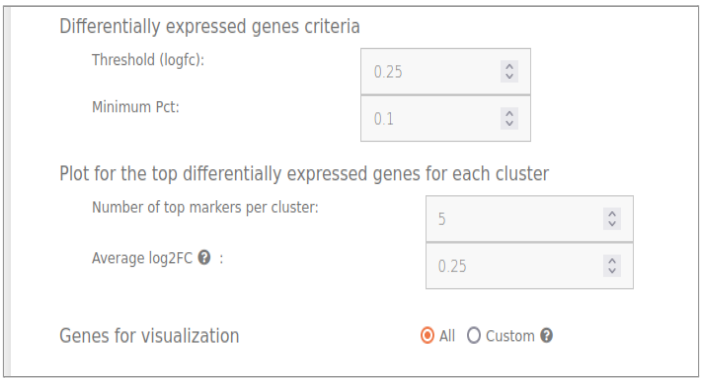
- Select your workflow.

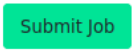




- Upload your files of Single Cell Rna-Seq Experiments in the following format:
    - Fastq files must be in this name format: [Sample name]_S*_[Read Type]_001.fastq.gz
    - Fastq files of the same sample must have the same sample name
    - Fastq files of different samples must have different sample name
- Select the transcriptome the input files belong to from our 3 built-in options (Human, Mouse, Human-mouse mixture).
- Select the species and tissue type of your sample for cluster annotation to be performed.
- Select if you want to manually configure the parameters of the pipeline. If you don't, our Default Options will be applied. Possible manual options are:
    - Expected number of recovered cells
    - BAM file generation
    - First filtering:
        - Minimum cells
        - Minimum features
    - Secondary filtering:
        - nFeature_RNA with lower and upper limits
        - nCount_RNA with lower and upper limits
    - Feature Extraction Method
    - Shared Nearest Neighbor (SNN) Graph
    - Clustering
    - Differentially expressed genes criteria
    - Plot for the top differentially expressed genes for each cluster
    - Genes for visualization



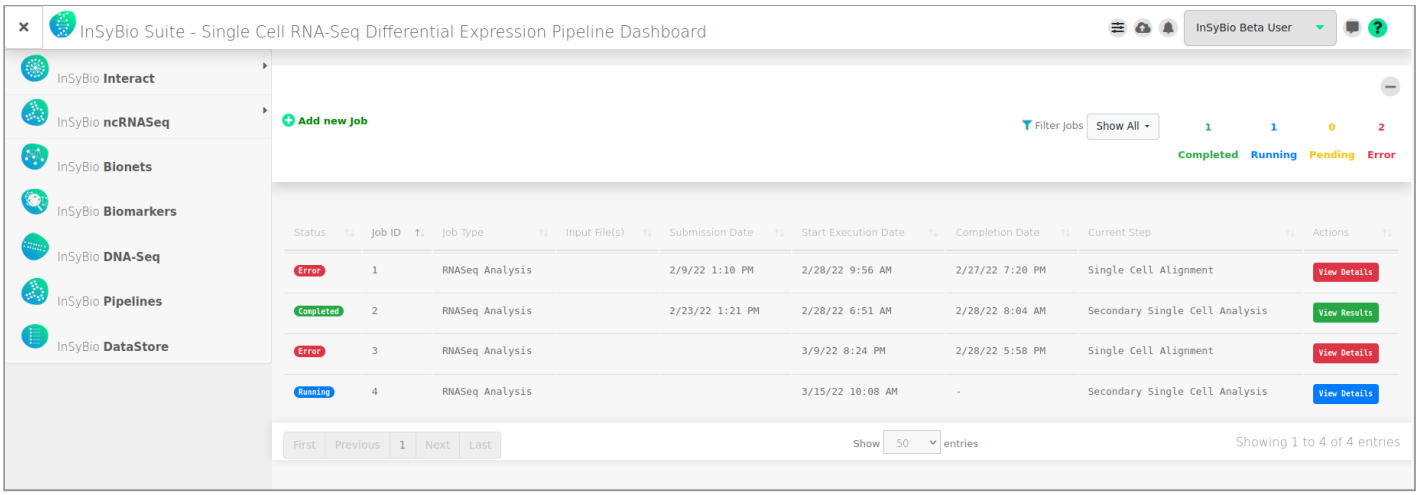


- Submit your job pressing the respective button. 

To view the results:

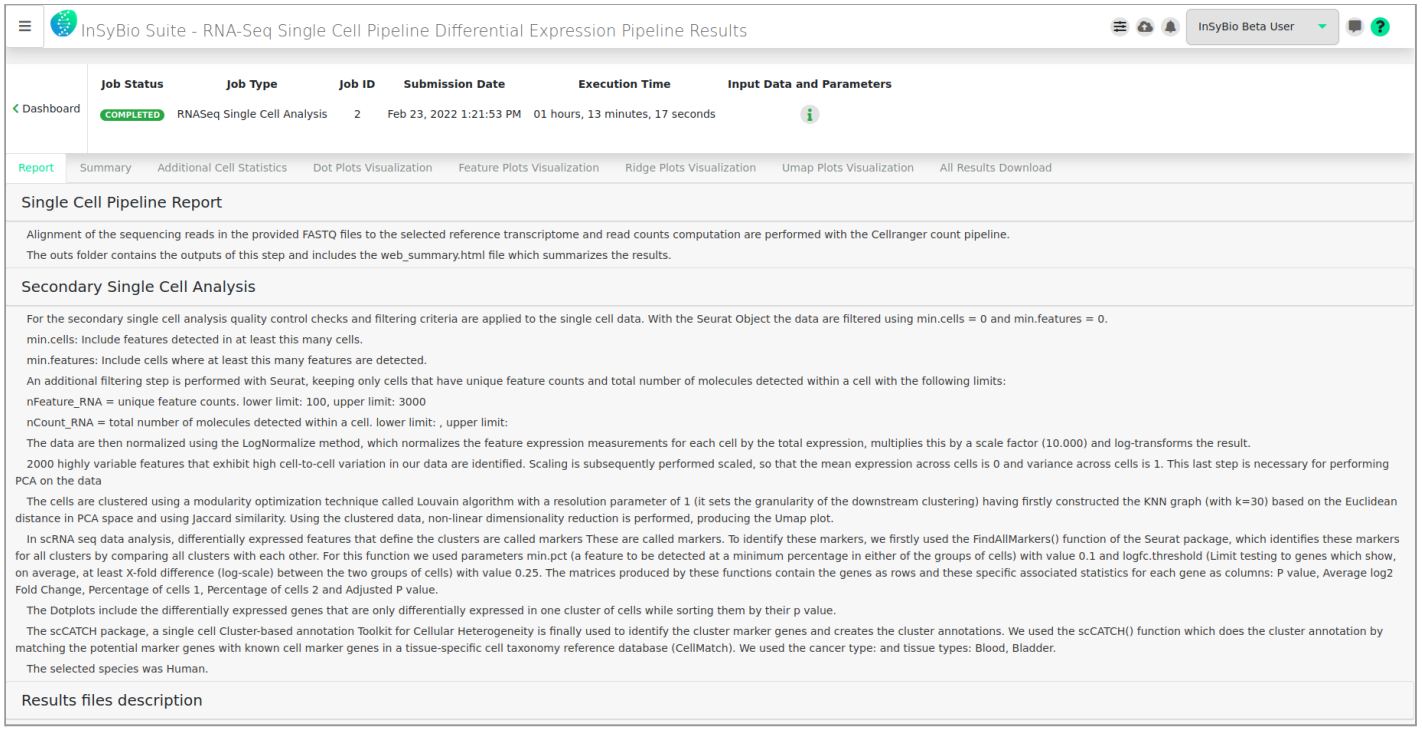
By starting a calculation you are informed if it was submitted successfully. Then you can move to the Single Cell Rna-Seq Differential Expression Pipeline and view the

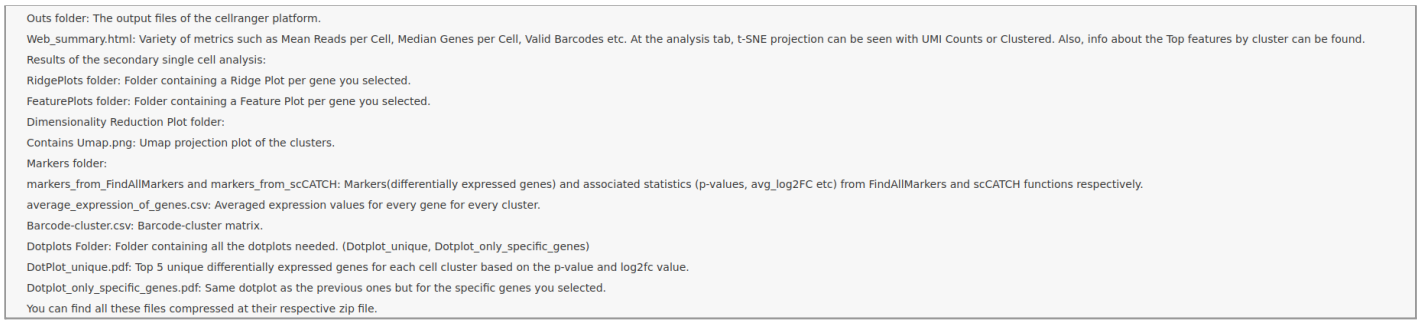
Dashboard, where you can view the status of your current and previous Single Cell



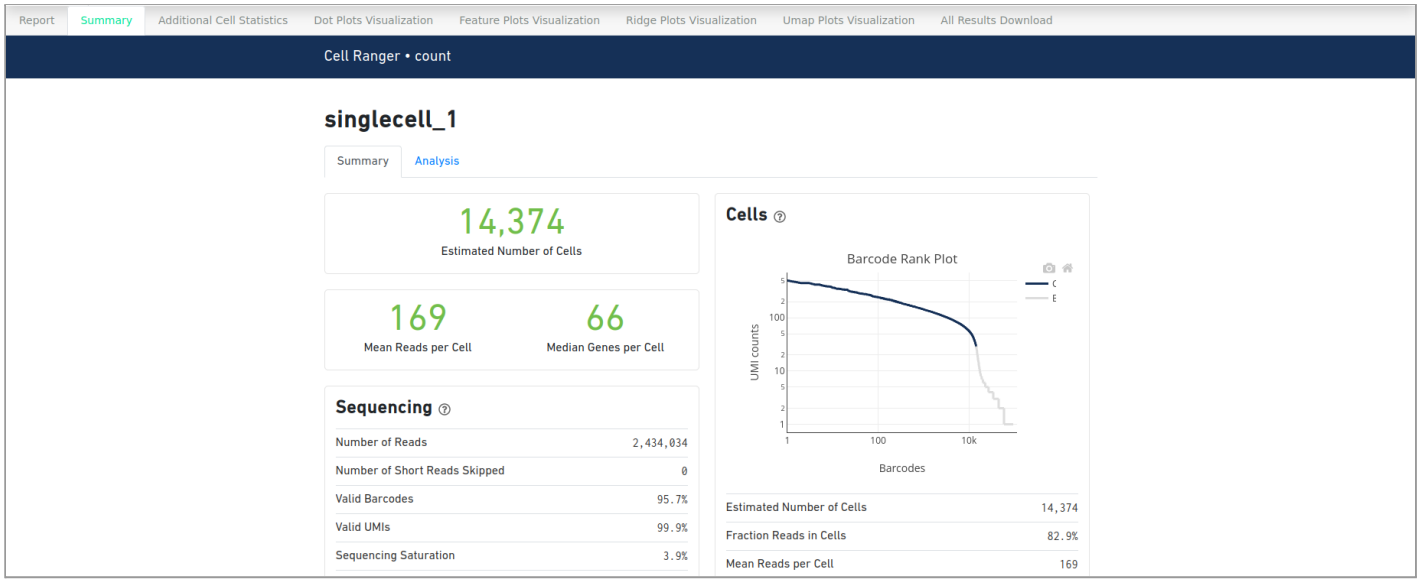
Rna-Seq Differential Expression jobs.

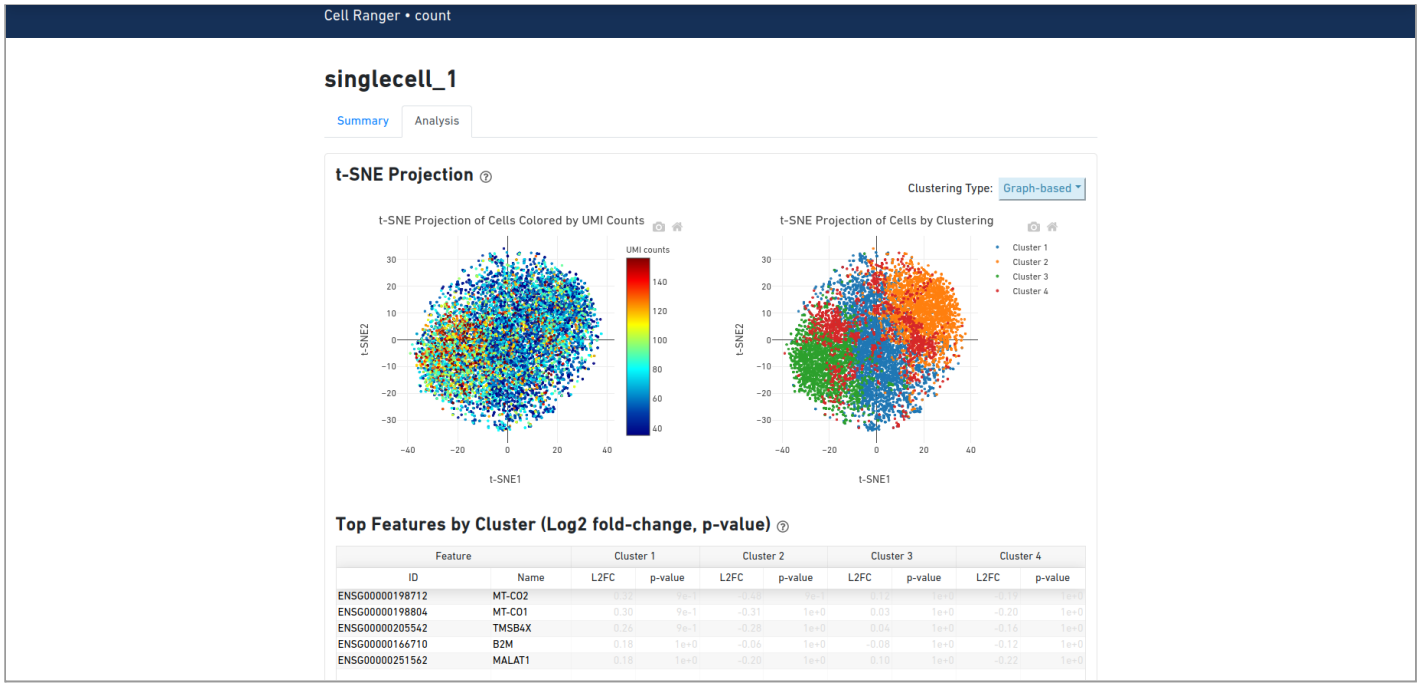
At completion of the Analysis you can select the View Results at the Actions column and view the produced files, that are separated according to the step that they were produced.



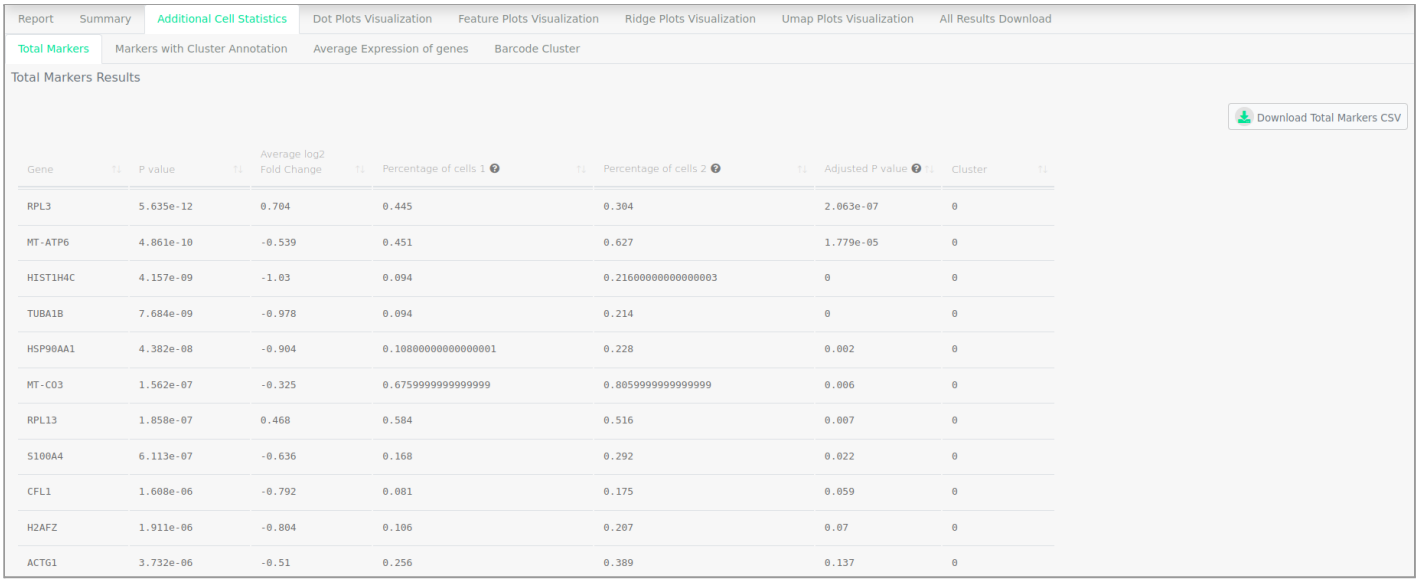


In the Report tab you can see a generated report that includes descriptions for every step and every parameter of the Single Cell Rna-Seq Differential Expression Pipeline for your job.



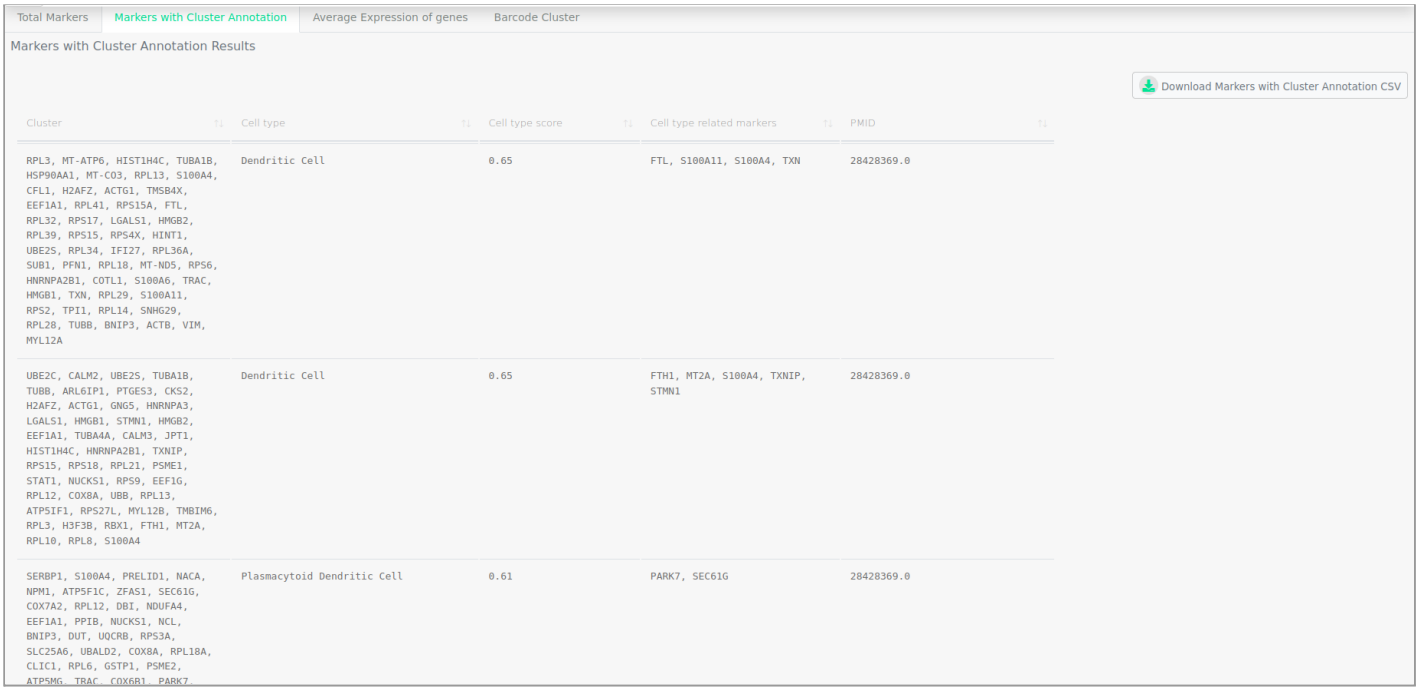


In the Summary tab you can see a summary of a variety of metrics from the first step of the Single Cell Rna-Seq Differential Expression Analysis and some T-SNE plots and information about the Top features by Cluster.



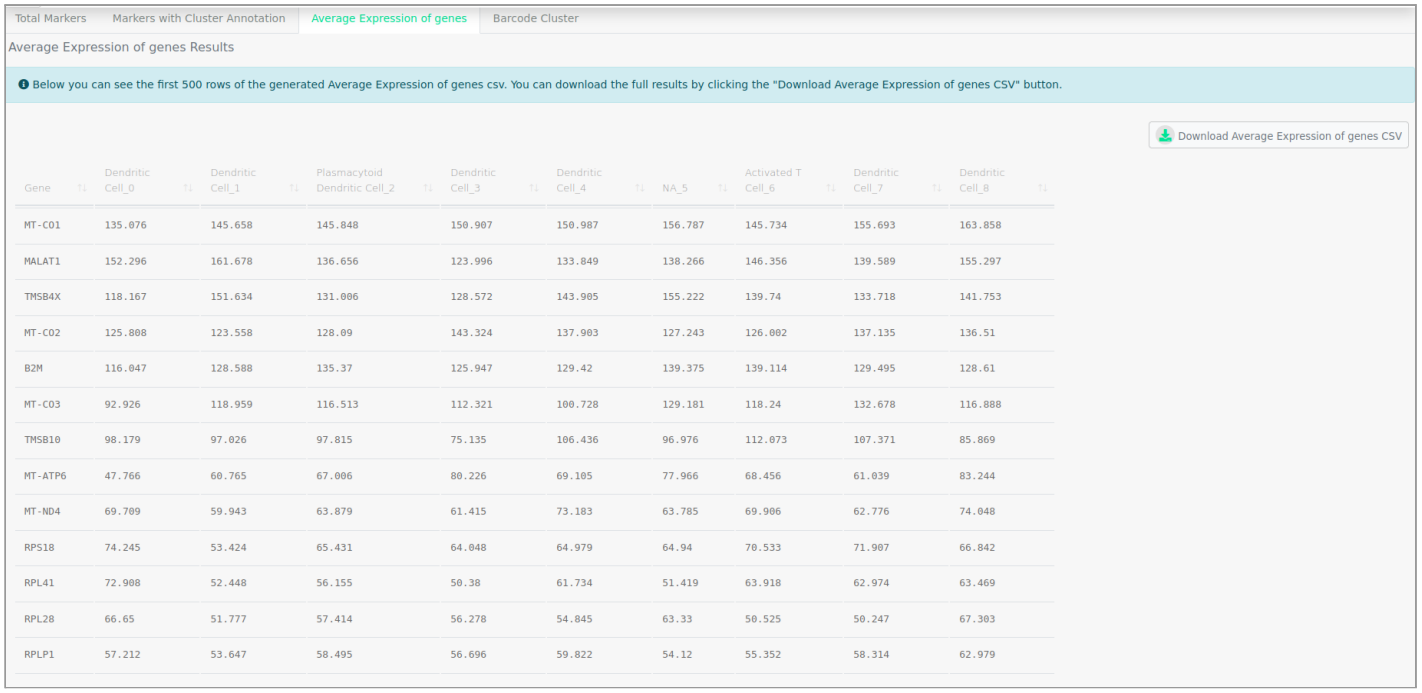
Gene	P value	Average log2 Fold Change	Percentage of cells 1	Percentage of cells 2	Adjusted P value	Cluster
RPL3	5.635e-12	0.704	0.445	0.304	2.063e-07	0
MT-ATP6	4.861e-10	-0.539	0.451	0.627	1.779e-05	0
HIST1H4C	4.157e-09	-1.03	0.094	0.21600000000000003	0	0
TUBA1B	7.684e-09	-0.978	0.094	0.214	0	0
HSP90AA1	4.382e-08	-0.904	0.10800000000000001	0.228	0.002	0
MT-CO3	1.562e-07	-0.325	0.6759999999999999	0.8059999999999999	0.006	0
RPL13	1.858e-07	0.468	0.584	0.516	0.007	0
S100A4	6.113e-07	-0.636	0.168	0.292	0.022	0
CFL1	1.608e-06	-0.792	0.081	0.175	0.059	0
H2AFZ	1.911e-06	-0.804	0.106	0.207	0.07	0
ACTG1	3.732e-06	-0.51	0.256	0.389	0.137	0

In the Additional Cell Statistics tab the user can view four different tabs that represent different information for the genes of the input files. The results for these four different tabs can be downloaded at the respective tab. At the Total Markers tab, markers (differentially expressed genes) and associated statistics (p-values, average log2 Fold change etc) can be found.



Cluster	Cell type	Cell type score	Cell type related markers	PMID
RPL3, MT-ATP6, HIST1H4C, TUBA1B, HSP90AA1, MT-CO3, RPL13, S100A4, CFL1, H2AFZ, ACTG1, TMSB4X, EEF1A1, RPL41, RPS15A, FTL, RPL32, RPS17, LGALS1, HMGB2, RPL39, RPS15, RPS4X, HINT1, UBE2S, RPL34, IFI27, RPL36A, SUB1, PFN1, RPL18, MT-ND5, RPS6, HNRNPA2B1, COTL1, S100A6, TRAC, HMGB1, TXN, RPL29, S100A11, RPS2, TPI1, RPL14, SNHG29, RPL28, TUBB, BNIP3, ACTB, VIM, MYL12A	Dendritic Cell	0.65	FTL, S100A11, S100A4, TXN	28428369.0
UBE2C, CALM2, UBE2S, TUBA1B, TUBB, ARL6IP1, PTGES3, CKS2, H2AFZ, ACTG1, GNG5, HNRNPA3, LGALS1, HMGB1, STMN1, HMGB2, EEF1A1, TUBA4A, CALM3, JPT1, HIST1H4C, HNRNPA2B1, TXNIP, RPS15, RPS18, RPL21, PSME1, STAT1, NUCKS1, RPS9, EEF1G, RPL12, COX8A, UBB, RPL13, ATP5IF1, RPS27L, MYL12B, TMBIM6, RPL3, H3F3B, RBX1, FTH1, MT2A, RPL10, RPL8, S100A4	Dendritic Cell	0.65	FTH1, MT2A, S100A4, TXNIP, STMN1	28428369.0
SERBP1, S100A4, PRELID1, NACA, NPM1, ATP5F1C, ZFAS1, SEC61G, COX7A2, RPL12, DBI, NDUFA4, EEF1A1, PPIB, NUCKS1, NCL, BNIP3, DUT, UQCRB, RPS3A, SLC25A6, UBALD2, COX8A, RPL18A, CLIC1, RPL6, GSTP1, PSME2, ATP5MG, TRAC, COX6B1, PARK7,	Plasmacytoid Dendritic Cell	0.61	PARK7, SEC61G	28428369.0

At the Markers with Cluster Annotation tab, the results of the Cluster Annotation step can be found.



Total Markers | Markers with Cluster Annotation | Average Expression of genes | Barcode Cluster

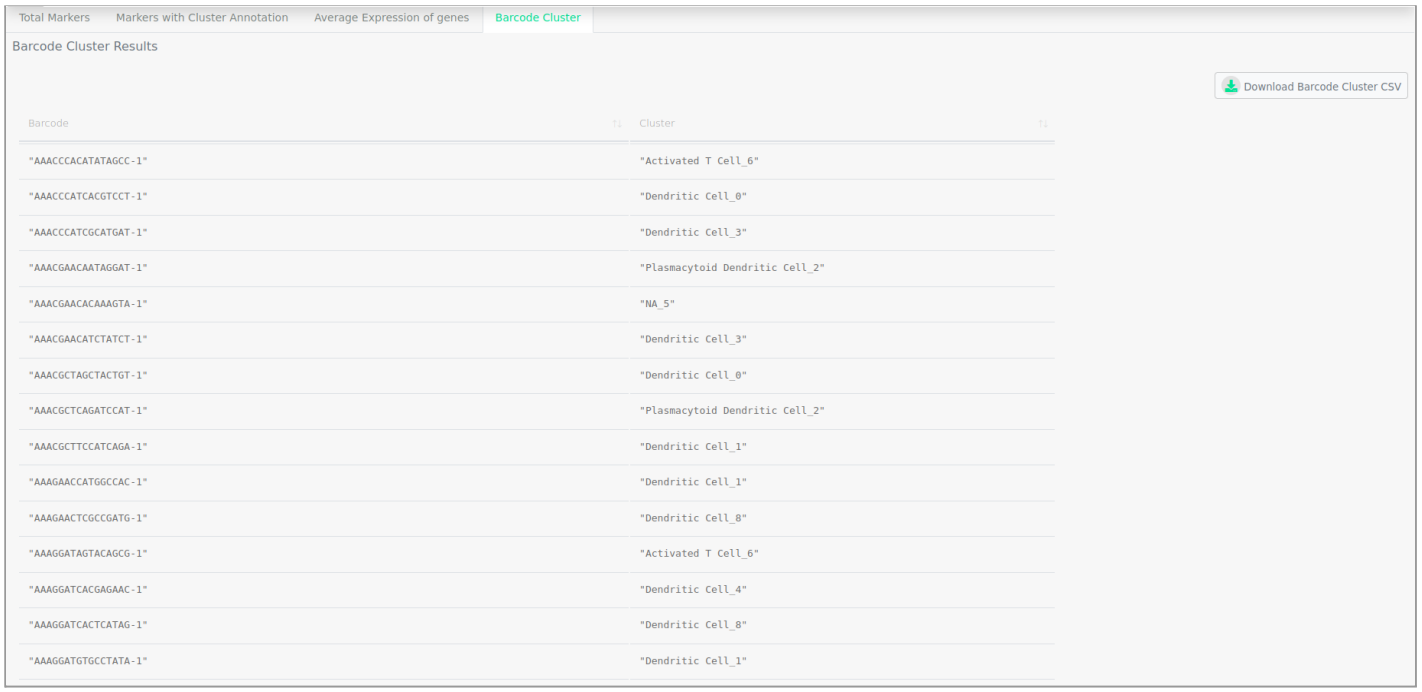
Average Expression of genes Results

Below you can see the first 500 rows of the generated Average Expression of genes csv. You can download the full results by clicking the "Download Average Expression of genes CSV" button.

Download Average Expression of genes CSV

Gene	Dendritic Cell_0	Dendritic Cell_1	Plasmacytoid Dendritic Cell_2	Dendritic Cell_3	Dendritic Cell_4	NA_5	Activated T Cell_6	Dendritic Cell_7	Dendritic Cell_8
MT-CO1	135.076	145.658	145.848	150.907	150.987	156.787	145.734	155.693	163.858
MALAT1	152.296	161.678	136.656	123.996	133.849	138.266	146.356	139.589	155.297
TMSB4X	118.167	151.634	131.006	128.572	143.905	155.222	139.74	133.718	141.753
MT-CO2	125.808	123.558	128.09	143.324	137.903	127.243	126.002	137.135	136.51
B2M	116.047	128.588	135.37	125.947	129.42	139.375	139.114	129.495	128.61
MT-CO3	92.926	118.959	116.513	112.321	100.728	129.181	118.24	132.678	116.888
TMSB10	98.179	97.026	97.815	75.135	106.436	96.976	112.073	107.371	85.869
MT-ATP6	47.766	60.765	67.006	80.226	69.105	77.966	68.456	61.039	83.244
MT-ND4	69.709	59.943	63.879	61.415	73.183	63.785	69.906	62.776	74.048
RPS18	74.245	53.424	65.431	64.048	64.979	64.94	70.533	71.907	66.842
RPL41	72.908	52.448	56.155	50.38	61.734	51.419	63.918	62.974	63.469
RPL28	66.65	51.777	57.414	56.278	54.845	63.33	50.525	50.247	67.303
RPLP1	57.212	53.647	58.495	56.696	59.822	54.12	55.352	58.314	62.979

At the Average Expression of genes Results tab the first 500 rows of the generated Average Expression of genes file can be found and it contains the expression levels of every gene for every cluster.



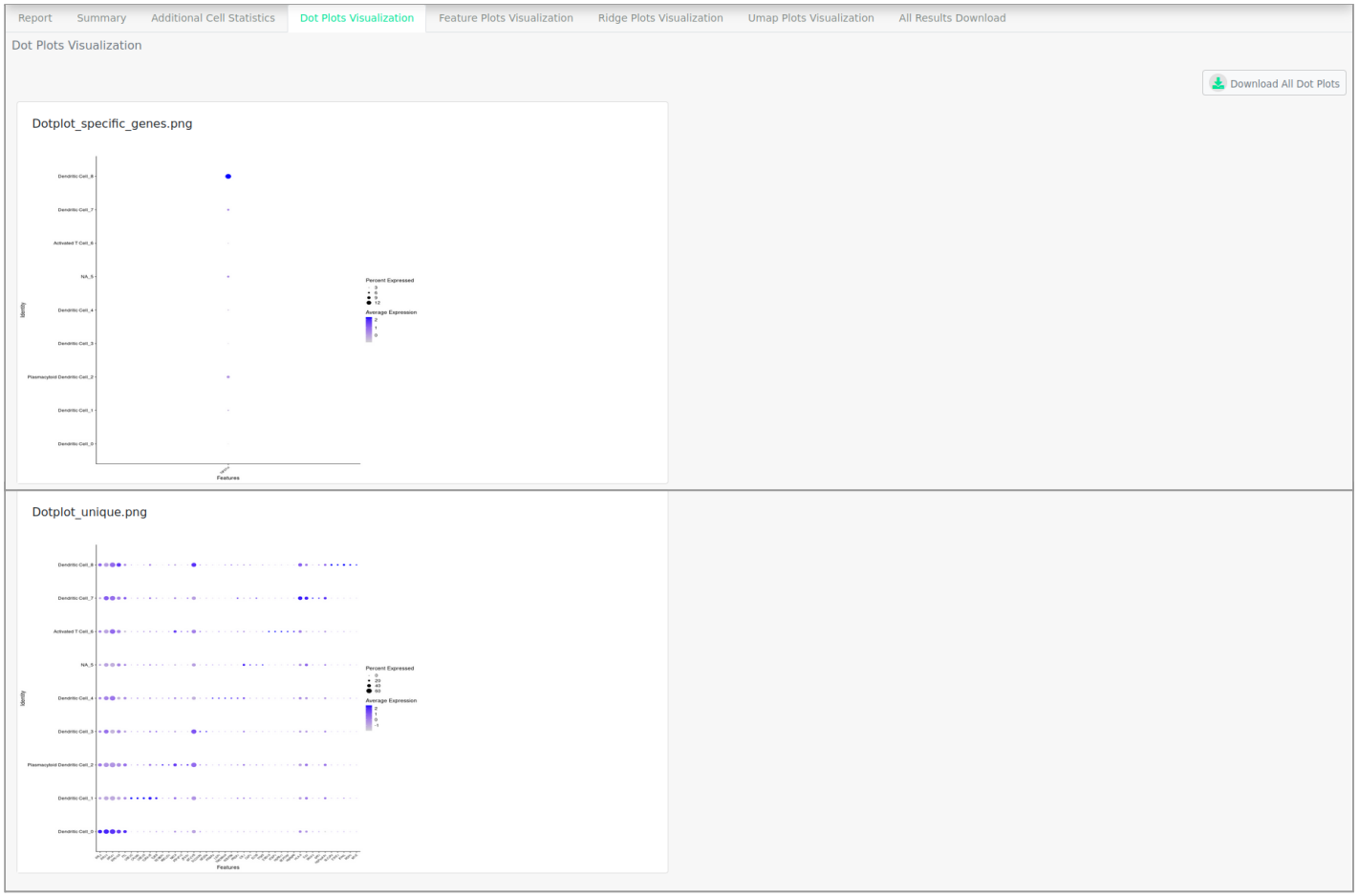
Total Markers | Markers with Cluster Annotation | Average Expression of genes | Barcode Cluster

Barcode Cluster Results

Download Barcode Cluster CSV

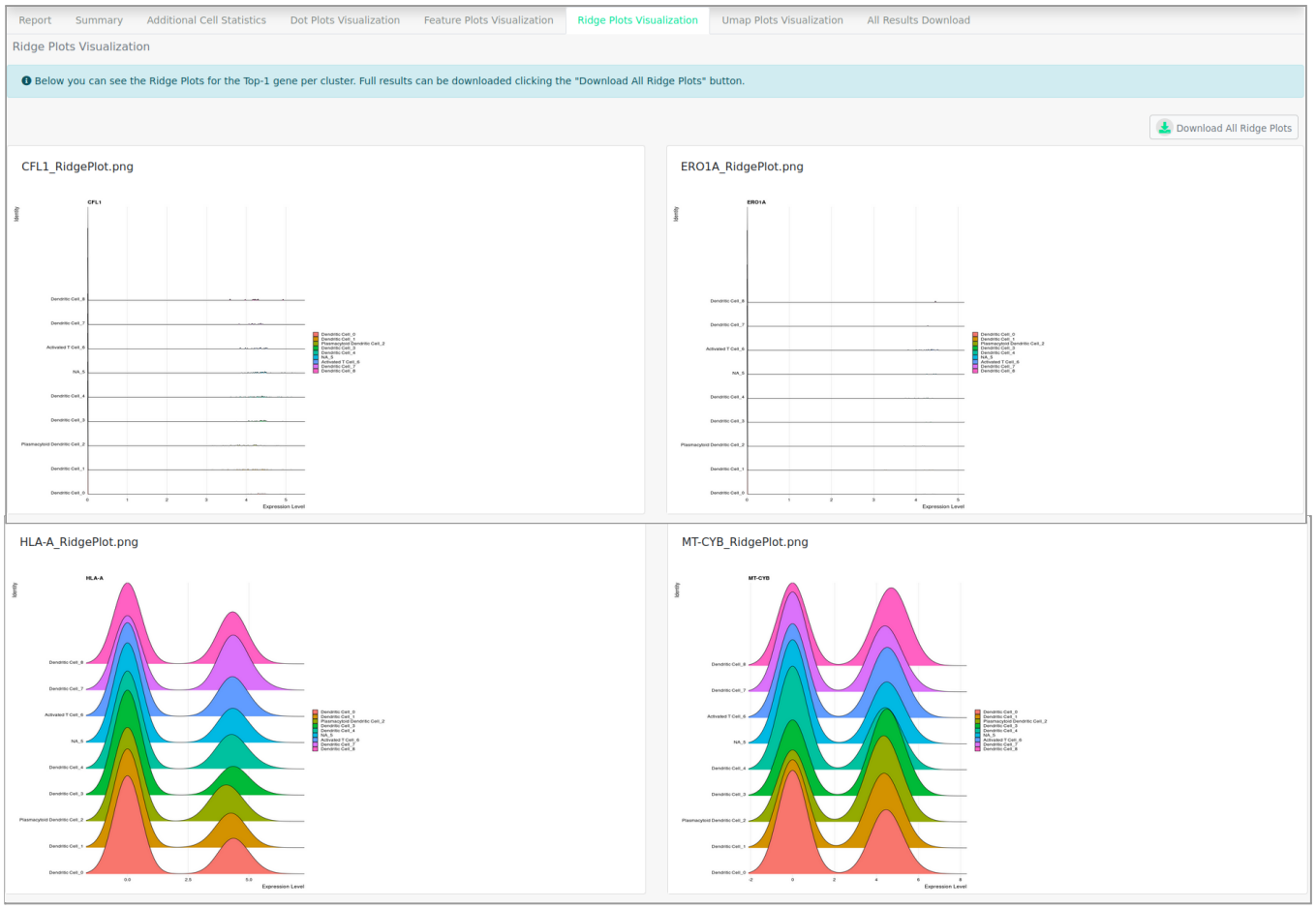
Barcode	Cluster
"AAACCCACATATAGCC-1"	"Activated T Cell_6"
"AAACCCATCACGTCCT-1"	"Dendritic Cell_0"
"AAACCCATCGCATGAT-1"	"Dendritic Cell_3"
"AAACGAACAATAGGAT-1"	"Plasmacytoid Dendritic Cell_2"
"AAACGAACACAAAGTA-1"	"NA_5"
"AAACGAACATCTATCT-1"	"Dendritic Cell_3"
"AAACGCTAGCTACTGT-1"	"Dendritic Cell_0"
"AAACGCTCAGATCCAT-1"	"Plasmacytoid Dendritic Cell_2"
"AAACGCTTCCATCAGA-1"	"Dendritic Cell_1"
"AAAGAACCATGGCCAC-1"	"Dendritic Cell_1"
"AAAGAACTCGCCGATG-1"	"Dendritic Cell_8"
"AAAGGATAGTACAGCG-1"	"Activated T Cell_6"
"AAAGGATCACGAGAAC-1"	"Dendritic Cell_4"
"AAAGGATCACTCATAG-1"	"Dendritic Cell_8"
"AAAGGATGTGCCTATA-1"	"Dendritic Cell_1"

At the Barcode Cluster tab the Barcode-Cluster matrix can be found.

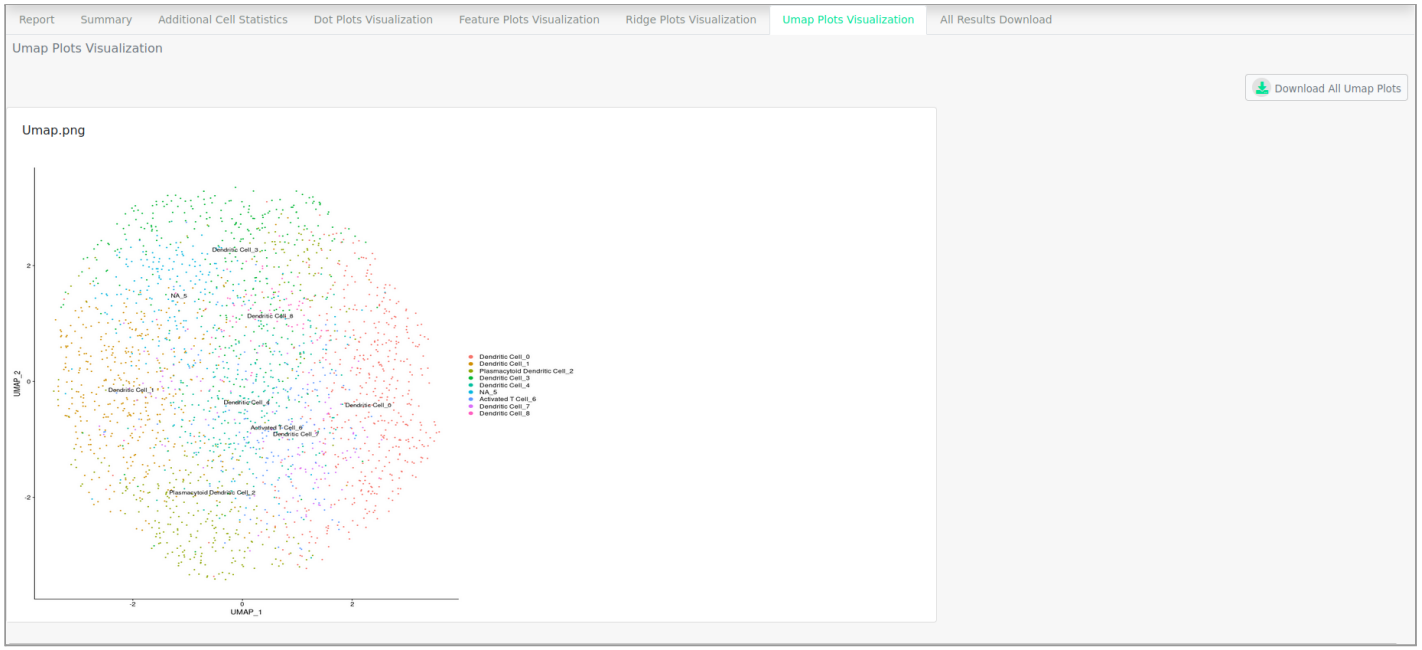


At the Dot Plots Visualization tab you can see the two Dot plots that are created. The first one is a Dot Plot with only the genes you specified at the manual parameters and second one is a Dot Plot that shows the Top 5 unique differentially expressed genes for each cell cluster based on the p-value and log2 fold change value. These plots can be downloaded.

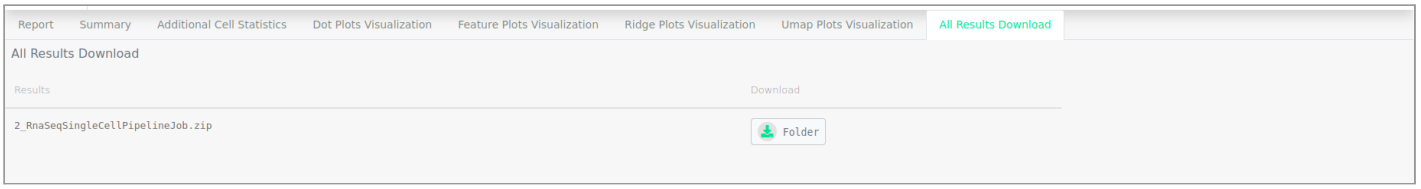
At the Feature Plots Visualization tab the Feature Plots for the Top-1 gene per cluster can be found. The Feature Plots of all the genes can be downloaded.



At the Ridge Plots Visualization tab the Ridge Plots for the Top-1 gene per cluster can be found. The Ridge Plots of all the genes can be downloaded.



At the Umap Plots Visualization tab the Umap Plots can be found. The Umap Plot can be downloaded.



At the All Results Download tab all the results of your job can be downloaded.

How to get InSyBio ncRNASeq

To request a free one month license of InSyBio Suite please email us at info@insybio.com .

To purchase InSyBio ncRNASeq commercial version 3.0 please contact us at sales@insybio.com.

About Us

InSyBio Ltd is a bioinformatics pioneer company (www.insybio.com) in personalized healthcare, that focuses on developing computational frameworks and tools for the analysis of complex life-science and biological data in order to develop predictive integrated biomarkers (biomarkers of various categories) with increased prognostic and diagnostic aspects for the personalized Healthcare Industry.

InSyBio Suite consists of tools for providing integrated biological information from various sources, while at the same time it is empowered with robust, user-friendly and installation-free bioinformatics tools based on intelligent algorithms and methods.

COPYRIGHT NOTICE