



InSyBio

Intelligent Systems Biology

User Manual

# Analyze non-coding RNAs with InSyBio ncRNASeq

March 2022

Insybio Suite v3.0

[www.insybio.com](http://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.

The screenshot displays the InSyBio ncRNASeq web interface. The sidebar on the left contains the following navigation options:

- InSyBio Interact
- InSyBio ncRNASeq
- non-coding RNA Analytics
  - ncRNA Feature Calculation (selected)
  - miRNA Prediction
  - miRNA Target site Feature Calculation
  - miRNA Target site Prediction
  - miRNA Target Prediction
  - ncRNASeq Knowledge Base
  - RNA-Seq Data Analysis

The main panel shows the 'ncRNA Feature Calculation' module. It includes input fields for 'Sequences' (containing 'ncrna15\_12\_') and 'File Title' (containing 'dsfile1639562377\_7291.txt'). Below these fields are two buttons: 'Select file from Data Store' and 'Go to Data Store to Upload File'. A 'Start calculation' button is located on the right side of the main panel.

Below the input fields is a table showing the status of various processes:

Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	35	ncrna15_12_	3/16/22 3:22 PM	3/16/22 3:22 PM	3/16/22 3:22 PM	View Results
Completed	34	ncrna15_12_	12/15/21 10:00 AM	12/15/21 10:00 AM	12/15/21 10:00 AM	View Results
Completed	33	ncrna 15_12	12/15/21 9:48 AM	12/15/21 9:48 AM	12/15/21 9:48 AM	View Results
Completed	32	ncrna14_12	12/14/21 10:01 AM	12/14/21 10:01 AM	12/14/21 10:01 AM	View Results
Completed	31	test	6/4/21 8:11 AM	6/4/21 8:11 AM	6/4/21 8:11 AM	View Results
Completed	29	75 sequences including pre-miRNAs, random cds and snoRNAs	3/4/21 4:43 PM	3/4/21 4:43 PM	3/4/21 4:43 PM	View Results
Completed	28	75 sequences including pre-miRNAs, random cds and snoRNAs	3/1/21 10:17 PM	3/1/21 10:17 PM	3/1/21 10:17 PM	View Results
Completed	27	75 sequences including pre-miRNAs, random cds and snoRNAs	1/4/21 6:06 PM	1/4/21 6:06 PM	1/4/21 6:06 PM	View Results

### 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	Actions	
		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	<a href="#">View Results</a>
Completed	9	test	11/15/18 8:59 PM	11/15/18 8:59 PM	11/15/18 8:59 PM	<a href="#">View Results</a>
Completed	8	sequences75_premiRNAs_cds_snoRNAs2222	11/8/18 2:35 PM	11/8/18 2:35 PM	11/8/18 2:35 PM	<a href="#">View Results</a>
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	<a href="#">View Results</a>
Completed	6	test	11/7/18 12:04 PM	11/7/18 12:04 PM	11/7/18 12:04 PM	<a href="#">View Results</a>
Pending	3	75 sequences including pre-miRNAs, random cds and snoRNAs	11/11/19 11:01 AM	-	-	<a href="#">View Details</a>

## 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.

Sequence	G+C	AU	AA	AC	AG	AU	CA	CC	CG	CU
> hsa-mir-26a-1 MI0000083 GUGGCCUCGUUCAAGUUAUCCAGGAUAGGCGUGGAGGUCACAAUGGGCCUAUUUUGGUUACUUGCACGGGGACGC	55.844	44.156	3.947	3.947	5.263	5.263	6.579	6.579	3.947	6.579
> random_sequence_from_cds_1 GAGGGCAGGGGGCACAGUCCAAUCACAGGCUUGUAGUCUGUCAGGGGCGUGGUGCCGCCCGGCAGCGCAGACUGUUCUGUGGGCCGUGCACA	69.072	30.928	1.042	4.167	8.333	0	10.417	9.375	4.167	6.25
> snoRNA_1 AAAGUGAGUGAUAGUUCUGUGGCAUAGUAAUCAAUUUUUUUAUUAAACCCUAAACUCUGAAGUCC	32.857	67.143	14.493	2.899	5.797	11.594	2.899	4.348	0	5.797
> hsa-mir-32 MI0000090 GGAGAUUUGCACAUUACUAGUUGCAUUGUUGUACGGCCUCAUGCAAUUUAGUGUGUGAUUUUUUUC	38.571	61.429	4.348	4.348	4.348	11.594	8.696	1.449	1.449	2.899
> hsa-mir-199a-1 MI0000242 GCCAACCCAGUUGUUCAGACUACCUUGUUCAGGAGGUCUCAUUGUUGUACAGUUGUCACAUUGGUUAGGC	50.704	49.296	2.857	7.143	10	2.857	11.429	5.714	0	7.143
> hsa-mir-148a MI0000253 GAGGCAAAGUUCUGAGACACUCCGACUCUGAGUAGUAGGAAGUCAGUCACUACAGAACUUUGUCUC	45.588	54.412	5.97	8.955	11.94	2.985	7.463	1.493	1.493	10.448

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 $\Sigma$ [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 = $dG/\text{number\_of\_stems}$ , where each stem is at least 3 continuous base pairs in the structure	MFE2
MFE Index 3 = $dG/\text{number\_of\_loops}$ , where number_of_loops is the number of the loops in the secondary structure	MFE3
MFE Index 4 = $dG/\text{total\_bases}$	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 $dP = \text{total\_bases}/L$ , 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
<b>Feature</b>	<b>ABBR</b>
$\text{Diff} =  MFE-EFE /L$ 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.

The screenshot displays the InSyBio ncrNASeq web interface. On the left is a navigation menu with options like 'ncRNA Feature Calculation', 'miRNA Prediction', 'miRNA Target site Feature Calculation', 'miRNA Target site Prediction', 'miRNA Target Prediction', 'ncRNASeq Knowledge Base', and 'RNA-Seq Data Analysis'. The main area shows a form for 'Sequences' with 'ncrna15\_12\_' entered, and 'File Title' with 'dsfile1639562377\_7291.txt'. Below the form are buttons for 'Select file from Data Store' and 'Go to Data Store to Upload File', and a 'Start calculation' button. Below the form is a table of process results:

Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	36	ncrna15_12_	3/16/22 3:26 PM	3/16/22 3:26 PM	3/16/22 3:26 PM	View Results
Completed	30	75 sequences including pre-miRNAs, random cds and snoRNAs	3/4/21 4:49 PM	3/4/21 4:50 PM	3/4/21 4:50 PM	View Results
Completed	14	75 sequences including pre-miRNAs, random cds and snoRNAs	11/11/19 11:36 AM	11/11/19 11:36 AM	11/11/19 11:36 AM	View Results
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

## 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.



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.

The screenshot displays the InSyBio ncRNASeq interface. On the left, a sidebar lists navigation options: InSyBio Interact, InSyBio ncRNASeq, non-coding RNA Analytics, ncrRNA Feature Calculation, miRNA Prediction, miRNA Target site Feature Calculation (highlighted), miRNA Target site Prediction, miRNA Target Prediction, ncrRNASeq Knowledge Base, RNA-Seq Data Analysis, and Single Cell RNA-Seq Data Analysis. The main area shows the 'miRNA Target site Feature Calculation' workflow. It includes input fields for mRNA Target Sequences (Filename: dsfile1444783391\_6577.fa) and miRNA Sequences (Filename: dsfile1444764074\_5421.fa). Below these are buttons for 'Select file from Data Store' and 'Go to Data Store to Upload File'. A 'Start calculation' button is visible on the right. At the bottom, a table lists completed processes:

Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	16	mRNAs: mrnas462, miRNAs: mirnas462	3/16/22 3:28 PM	3/16/22 3:28 PM	3/16/22 3:32 PM	View Results
Completed	14	mRNAs: mrnas462, miRNAs: mirnas462	3/4/21 5:22 PM	3/4/21 5:22 PM	3/4/21 5:44 PM	View Results
Completed	13	mRNAs: mirnas462, miRNAs: mrnas462	11/11/19 11:51 AM	11/11/19 11:51 AM	11/11/19 12:36 PM	View Results
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-324-50TCL1B-001.fa, miRNAs:	11/15/18 9:01	11/15/18 9:01	11/15/18 9:02	View Results

## 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	<a href="#">View Results</a>
Completed	9	mRNAs: targetshsa-miR-324-5pTCL1B-001.fa, miRNAs: miRNAs:hsa-miR-324-5pTCL1B-001.fa	11/15/18 9:01 PM	11/15/18 9:01 PM	11/15/18 9:02 PM	<a href="#">View Results</a>
Completed	8	mRNAs: mrnas462, miRNAs: mirnas462	11/8/18 1:45 PM	11/8/18 1:45 PM	11/8/18 5:51 PM	<a href="#">View Results</a>
Completed	3	mRNAs: mrnas462, miRNAs: mirnas462	9/26/18 11:21 AM	9/26/18 11:21 AM	9/26/18 12:00 PM	<a href="#">View Results</a>
Completed	1	mRNAs: genes_5_5_0_shuffled_targets, miRNAs: genes_5_5_0_miRNAs	8/17/18 7:13 AM	8/17/18 7:13 AM	8/17/18 7:33 AM	<a href="#">View Results</a>
Pending	13	mRNAs: mirnas462, miRNAs: mrnas462	11/11/19 11:51 AM	-	-	<a href="#">View Details</a>

### 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 Site Features Calculations. At completion of the calculation you can select the View Results at the Actions column and view the calculated features.

miRNA Sequence	Target Sequence	mats	matos	mat	gcmats	gcmatos	gcmat	aumats	aumatos	aumat	unps	unpos	unp	gus	guos	gu	miss	m
> [hsa-miR-101] Homo sapiens UACAGUACUGUAUACUGAA	> NM_004456EZ220478051 Homo sapiens TGAATTTGCAAAGTACTGTA	9	2	11	3	1	4	6	1	7	-2	22	20	0	0	0	-2	
> [hsa-miR-101] Homo sapiens UACAGUACUGUAUACUGAA	> NM_004456EZ220478051 Homo sapiens TTCAGGAACCTCGACTGTG	8	6	14	3	3	6	5	3	8	0	16	16	2	2	4	-2	
> [hsa-miR-101] Homo sapiens UACAGUACUGUAUACUGAA	> NM_181833NF217220301 Homo sapiens TACAAGAGATTCTCTGCTCA	4	3	7	2	2	4	2	1	3	8	22	30	0	0	0	8	
> [hsa-miR-101] Homo sapiens UACAGUACUGUAUACUGAA	> NM_001039111TRIM7117890240 Homo sapiens ACAACATTGCTTAAGTCTACCTCA	1	5	6	0	2	2	1	3	4	14	21	35	0	2	2	14	
> [hsa-miR-101] Homo sapiens	> NM_001039111TRIM7117890240 Homo sapiens	9	3	12	3	2	5	6	1	7	-2	25	23	0	0	0	-2	

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= $(-1 * \text{free energy of the target sequence}) / \log(\text{length of target} * \text{length of miRNA})$	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.

The screenshot displays the InSyBio ncrNASeq web interface. The sidebar on the left contains the following navigation options:

- InSyBio Interact
- InSyBio ncrNASeq
  - non-coding RNA Analytics
    - Prediction of ncRNAs and miRNA targets.
      - ncRNA Feature Calculation
        - Feature calculation module for 58 miRNA genes-related features.
      - miRNA Prediction
        - Prediction module for pre-miRNAs.
      - miRNA Target site Feature Calculation
        - Feature calculation module for 124 miRNA target features.
      - miRNA Target site Prediction
        - Prediction module for miRNA targets.
      - miRNA Target Prediction
        - Prediction module for miRNA targets.
    - ncRNASeq Knowledge Base
      - miRNA and transcript search.
    - RNA-Seq Data Analysis
      - Preprocessing and differential expression analysis of FASTQ files.
    - Single Cell RNA-Seq Data Analysis

The main form is titled "miRNA Target Site Prediction" and includes the following fields and buttons:

- mRNA Target Sequences:
- Filename: 
  - Select file from Data Store
  - Go to Data Store to Upload File
- miRNA Sequences:
- Filename: 
  - Select file from Data Store
  - Go to Data Store to Upload File
- Start calculation button

Below the form is a table showing the execution status of various jobs:

Status	Process ID	Information	Submission Date	Start Execution Date	Completion Date	Actions
Completed	17	mRNAs: mimas462, miRNAs: mimas462	3/16/22 4:18 PM	3/16/22 4:19 PM	3/16/22 4:22 PM	View Results
Completed	15	mRNAs: mimas462, miRNAs: mimas462	3/4/21 5:24 PM	3/4/21 5:24 PM	3/4/21 5:47 PM	View Results
Completed	12	mRNAs: , miRNAs: test	11/30/18 9:54 AM	11/30/18 9:54 AM	11/30/18 9:56 AM	View Results
Completed	10	mRNAs: test, miRNAs: test	11/15/18 9:29 PM	11/15/18 9:29 PM	11/16/18 12:00 AM	View Results
Error	7	mRNAs: , miRNAs: pseudosmi1848	9/27/18 9:36 AM	9/27/18 9:36 AM	11/30/18 10:11 AM	View Details

## 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	<a href="#">View Results</a>
Completed	9	mRNAs: targethsa-miR-324-5pTCL1B-001.fa, miRNAs: miRNAs:hsa-miR-324-5pTCL1B-001.fa	11/15/18 9:01 PM	11/15/18 9:01 PM	11/15/18 9:02 PM	<a href="#">View Results</a>
Completed	8	mRNAs: mrnas462, miRNAs: mirnas462	11/8/18 1:45 PM	11/8/18 1:45 PM	11/8/18 5:51 PM	<a href="#">View Results</a>
Completed	3	mRNAs: mrnas462, miRNAs: mirnas462	9/26/18 11:21 AM	9/26/18 11:21 AM	9/26/18 12:00 PM	<a href="#">View Results</a>
Completed	1	mRNAs: genes_5_5_0_shuffled_targets, miRNAs: genes_5_5_0_miRNAs	8/17/18 7:13 AM	8/17/18 7:13 AM	8/17/18 7:33 AM	<a href="#">View Results</a>
Pending	13	mRNAs: mirnas462, miRNAs: mrnas462	11/11/19 11:51 AM	-	-	<a href="#">View Details</a>

### 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.

**InSyBio Suite Beta - miRNA Target Site Prediction Results**

Job Status: **COMPLETED** | Job ID: 4 | Submission Date: Sep 26, 2018 11:29:30 AM | Execution Time: 01 hours, 10 minutes, 43 seconds

miRNA Sequence	Target Sequence	Prediction Score	Prediction	mats	matos	mat	gcmats	gcmatos	gcmat	aumats	aumatos	aumat	unps	unpos	unp
> [hsa-miR-101] Homo sapiens UACAGUACUGUGAUACUGAA	> NM_004456EH220478051 Homo sapiens TGAATTTGCAAAGTACTGTA	0.963256	Target	9	2	11	3	1	4	6	1	7	-2	22	20
> [hsa-miR-101] Homo sapiens UACAGUACUGUGAUACUGAA	> NM_004456EH220478051 Homo sapiens TTCAGGAACCTCGACTACTGTG	1.2725	Target	8	6	14	3	3	6	5	3	8	0	16	16
> [hsa-miR-101] Homo sapiens UACAGUACUGUGAUACUGAA	> NM_101833NF217220301 Homo sapiens TACAAGAGATTCTCCTGCCTCA	-0.786746	no Target	4	3	7	2	2	4	2	1	3	8	22	30
> [hsa-miR-101] Homo sapiens UACAGUACUGUGAUACUGAA	> NM_001039111TRIM7117890240 Homo sapiens ACAACATTGCTTAAGTCCTACCTCA	-0.880751	no Target	1	5	6	0	2	2	1	3	4	14	21	35

Showing 1 to 25 of 213,444 entries

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.



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	<a href="#">View Results</a>
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	<a href="#">View Results</a>
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,MHRN2,PRSS3,VEGFB,ADAM21,ADAMTSL4,C10TNF4,CCL3L3,COL4A2,LAMB1	11/29/18 3:40 PM	11/29/18 3:40 PM	11/29/18 3:52 PM	<a href="#">View Results</a>
Completed	86	miRNAs: hsa-miR-6126, hsa-miR-1200, hsa-let-7a-2-3p, hsa-miR-106b-3p targets: ZIK1, A18G-AS1, FGGY	11/29/18 3:39 PM	11/29/18 3:39 PM	11/29/18 3:39 PM	<a href="#">View Results</a>
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	<a href="#">View Results</a>
Error	84	miRNAs: targets: ZIK1	11/29/18 3:08 PM	11/29/18 3:08 PM	11/29/18 3:08 PM	<a href="#">View Details</a>

## 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.

Mirna Target Prediction Tool Results InSyBio Beta User

Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters	Actions																									
COMPLETED	89	Nov 11, 2019 3:02:12 PM	00 hours, 00 minutes, 02 seconds	<a href="#">Results Download all target sites found</a> <a href="#">Download miRNA-target genes scores</a>																										
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>miRNA</th> <th>Gene</th> <th>Score</th> <th>Actions</th> </tr> </thead> <tbody> <tr> <td>hsa-miR-6126</td> <td>ZIK1</td> <td>1.169</td> <td><a href="#">Details</a></td> </tr> </tbody> </table>						miRNA	Gene	Score	Actions	hsa-miR-6126	ZIK1	1.169	<a href="#">Details</a>																	
miRNA	Gene	Score	Actions																											
hsa-miR-6126	ZIK1	1.169	<a href="#">Details</a>																											
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>miRNA</th> <th>Gene</th> <th>Transcript</th> <th>Score</th> <th>Actions</th> </tr> </thead> <tbody> <tr> <td>hsa-miR-6126</td> <td>ZIK1</td> <td>ZIK1-002</td> <td>0.817</td> <td><a href="#">Utr Sequence</a></td> </tr> <tr> <td>hsa-miR-6126</td> <td>ZIK1</td> <td>ZIK1-001</td> <td>0.817</td> <td><a href="#">Utr Sequence</a></td> </tr> <tr> <td>hsa-miR-6126</td> <td>ZIK1</td> <td>ZIK1-004</td> <td>1.517</td> <td><a href="#">Utr Sequence</a></td> </tr> <tr> <td>hsa-miR-6126</td> <td>ZIK1</td> <td>ZIK1-003</td> <td>1.527</td> <td><a href="#">Utr Sequence</a></td> </tr> </tbody> </table>						miRNA	Gene	Transcript	Score	Actions	hsa-miR-6126	ZIK1	ZIK1-002	0.817	<a href="#">Utr Sequence</a>	hsa-miR-6126	ZIK1	ZIK1-001	0.817	<a href="#">Utr Sequence</a>	hsa-miR-6126	ZIK1	ZIK1-004	1.517	<a href="#">Utr Sequence</a>	hsa-miR-6126	ZIK1	ZIK1-003	1.527	<a href="#">Utr Sequence</a>
miRNA	Gene	Transcript	Score	Actions																										
hsa-miR-6126	ZIK1	ZIK1-002	0.817	<a href="#">Utr Sequence</a>																										
hsa-miR-6126	ZIK1	ZIK1-001	0.817	<a href="#">Utr Sequence</a>																										
hsa-miR-6126	ZIK1	ZIK1-004	1.517	<a href="#">Utr Sequence</a>																										
hsa-miR-6126	ZIK1	ZIK1-003	1.527	<a href="#">Utr Sequence</a>																										

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.

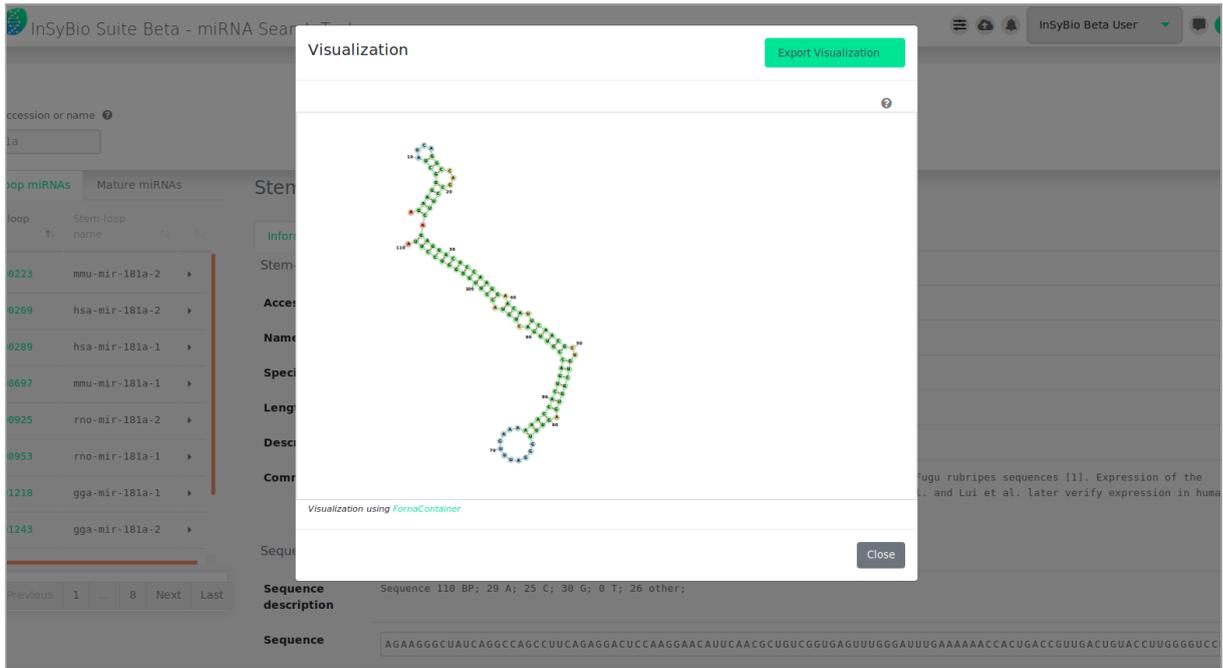
- Gene show page InSyBio Beta User

< Mirna Target Prediction Tool Results

<b>miRNA</b>	hsa-miR-6126
<b>Gene</b>	ZIK1
<b>Transcript</b>	ZIK1-001
<b>miRNA-Gene Score</b>	1.169
<b>miRNA-Transcript Score</b>	0.817
<b>Number of target sites</b>	11
<b>3'UTR sequence</b>	<div style="display: flex; justify-content: space-between;"> <div style="width: 80%;"> <pre style="font-family: monospace; font-size: 8px; margin: 0;"> 1  AGGCCTCATGAATGCAGCAAAATGTGGAAAGCCCTTCAACTCAAGATCTATCATCATTTAGCTCCTGAAAGTCCACACTTA      80 81  AGTAGAGCCTTAGACCTACAGGAAAGTGCCTCTCTGTAGTATTGTAGCAGTAGAGAGCCTTTGTGAGGGAGCCATCTG      160 161  CCTGAAGTTGAACCTCATCTTCCCTTGTCTCTGGTAGAAACCATCTACCCTTACCACCTTGCACAGTGGGCACCTGGT      240 241  CACTCCTATGTGCTAAGACAAGGCAGACATCTGTGTCTCTTAAGTCTTTGGAGGAAATCTTGAGCAGTCTAAGCCTT      320 321  TAGAGAAAAATCATTCTTTTTCTGACTGATCACAGCATACGTGTGACCCAGTTTGGGTCAGGAGGGCCAGCCTTGGTT      400 401  CTGCTGGACACTTATGTGCAAGGATTCCTCTCATGTAAATCTTTGGTCTCACAAGACACTTGGTCATCTTCCAGCCTCC      480 481  ATGTCACCACAGTGGTGAATGGCTGCCTCACATTGCTCCAGTTTGTGCACATAAAGCCCTTATATTTGAATCACTCTGT      560 561  AGCTTTGGGGTCTGTCTTACTGTGTGGGGTGGCTGGGAGACAGACTTCAACTCTATATGAAGGAATGGATGGCTTTTGTG      640 641  GGCCTCTGCAGGAAAGTAAAGTACAGAGTAATCTAATCTGGTTTGGTCATCTTGTCTTGTACCTAAATCTTCTCTG      720 721  AGGAAAAAATGCAAGGTTTGGTTATTCTAATTTGGCCCTGGATCCCTATTCTTCTGTGAGACTAGAGGTCATCCTGA      800 801  GGAGAGGCCAGCTGTTATGACAAGCATGTGTGCTTCAGGGAATAGGACAATTTATTCATTTGTTCCAGAGGATGTCAT      880 881  ATGATGCCAGTCTGCTGASAAAGCTTTTCATGGGGTCTATAAGGAGGCATGCCCTGATATCAAAACATCCATAGGCCG      960 961  ATGTCAAGCAGAAAGACAACCGGAGTCAACATGTGAAGTGAATTTGGTACAGAAATACCTGGGATTTTCTGTACTGTGTGTA      1040 1041  CTGTAGCAAACTAGTTGGAATGTGCCTCTATAAAAGTACATTTACAATCTTCCCCTGACTGGCTTTGAGCAGTCA      1120 1121  AGGGACCTAGAAATCTGTGTATGTCCAATAGCTGAGGTTATTTTCAGCAAAAATAATTAAGGGTTTATTTTTTAATCT      1200 1201  GTTGGTTTTCTAGGTTGTTACCTCAAGTGCATTTGCTGAGAGCCAGAAAAGGAGGATAAAGATAACAGAAAGTCCAT      1280 1281  AGGCCAGGGATGATTGATAGCTCTTGTGATTTCCACCAGTGTGCTGTTGCTCAAATGGCCACAGCCTTCATTGCTTG      1360 1361  CCAACTTTCTGTCATGAGAGGACTCATGGTTGCCCTTCCCAGGCCTGAAGAGAGAGTGCAGTCAACATGAGATTGCTA      1440 1441  GGCATTCTGGTTCTGAAAGTGGGTGATCAGACTACTTTATTGAAACATGTTTACAACATTTCTTGATGTGAAGTG      1520 1521  ACATGCCATAGTTTACATCCATTTATGGGTATAAATTTGAAGAGTTTGTCTACAAGCCTGTGAACCATAAATCATGATC      1600 1601  ATGAAACATATTCATGATTCACCTCTTGCCTTTTACAATCTCTGCTGTACTTTCCAGGCCTTCAGGAGTCTGTCTATT      1680 1681  ACTTCTCCCTACAGGAGAATAGTTTGTGTTTTCTAGGATTTTATGTGAATTGAACGTAAATACTTACTCATTTTTCTCT      1760                     </pre> </div> <div style="width: 15%; border: 1px solid #ccc; padding: 5px; font-size: 8px;"> <p>Score : 1.7294313303229796</p> <p>TTCCCTCATGTAATTTCTTGTCT-CACAT</p> <p>                   </p> <p>---AGAGG-----CGGCCGGAAGUG--</p> <hr/> <p>Score : 1.5224538611539185</p> <p>TGACACTTGGTCATTTCCAGCCTCCATG</p> <p>                   </p> <p>-----AGAGCCGGCCCGGAAGUG</p> </div> </div>

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





## Mature miRNAs and references

miRNA accession or name  Show results

Stem-loop miRNAs | Mature miRNAs | Stem-loop: MI0000269 hsa-mir-181a-2

Stem-loop id	Stem-loop name	Information	Mature miRNAs	References			
MI0000223	mmu-mir-181a-2	Accession	Name	Sequence	FASTA	Evidence	Experiment
MI0000269	hsa-mir-181a-2	MIMAT0000256	hsa-miR-181a-5p	39 aacaucaacgcgucgugugagu 61	<a href="#">Download</a>	Experimental	cloned [2,4-6]
MI0000289	hsa-mir-181a-1	MIMAT0004558	hsa-miR-181a-2-3p	77 accacugaccguugacugucc 98	<a href="#">Download</a>	Experimental	cloned [4]

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.

miRNA accession or name 

 Show results

Stem-loop miRNAs Mature miRNAs **Stem-loop: MI0000269 hsa-mir-181a-2**

Stem-loop id	Stem-loop name	ti	ti
MI0000223	mmu-mir-181a-2		
MI0000269	hsa-mir-181a-2		
MI0000289	hsa-mir-181a-1		
MI0000697	mmu-mir-181a-1		
MI0000925	rno-mir-181a-2		
MI0000953	rno-mir-181a-1		
MI0001218	gga-mir-181a-1		
MI0001243	gga-mir-181a-2		

Information Mature miRNAs **References**

Links to external database entries

Database	External Link
	MI0000269
	mir-181
	MIR181A2
	MIR181A2

Publications

- Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. **Vertebrate microRNA genes**. Science. 299:1540(2003). [\[PubMed\]](#)
- Dostie J, Mourelatos Z, Yang M, Sharma A, Dreyfuss G. **Numerous microRNPs in neuronal cells containing novel microRNAs**; RNA. 9:180-186(2003). [\[PubMed\]](#)
- Weber MJ. **New human and mouse microRNA genes found by homology search**; FEBS J. 272:59-73(2005). [\[PubMed\]](#)
- Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Jj, Sander C, Zavolan M, Tuschl T. **A mammalian microRNA expression atlas based on small RNA library sequencing**; Cell. 129:1401-1414(2007). [\[PubMed\]](#)
- Lui WO, Pourmand N, Patterson BK, Fire A. **Patterns of known and novel small RNAs in human cervical cancer**; Cancer Res. 67:6031-6043(2007). [\[PubMed\]](#)
- Marton S, Garcia MR, Robello C, Persson H, Trajtenberg F, Pritsch O, Rovira C, Naya H, Dighiero G, Cayota A. **Small RNAs analysis in CLL reveals a deregulation of miRNA**

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

miRNA accession or name 

 Show results

Stem-loop miRNAs **Mature miRNAs** **Mature: MIMAT0000210 mmu-miR-181a-5p**

Mature id	Mature name	ti	ti
MIMAT0000210	mmu-miR-181a-5p		
MIMAT0000210	mmu-miR-181a-5p		
MIMAT0000256	hsa-miR-181a-5p		
MIMAT0000256	hsa-miR-181a-5p		
MIMAT0000270	hsa-miR-181a-3p		
MIMAT0000660	mmu-miR-181a-1-3p		

**Information** Stem-loop miRNAs References

**Accession** MIMAT0000210

**Name** mmu-miR-181a-5p

**Sequence** 14 aacauucaacgcugucggugagu 36

**FASTA** Download

**Evidence** Experimental

**Experiment** cloned [2,4], Illumina [5-6]

**Similarity** MI0000223

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



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

The screenshot shows the InSyBio ncRNASeq interface. At the top, there is a search bar with 'zik1' entered and a 'Show results' button. Below the search bar, there are two tabs: 'Transcripts' (selected) and 'Genes'. The 'Transcripts' tab displays a table of search results:

Ensemble Transcript id	Transcript name	TI	TI
ENST00000307468	ZIK1-004		
ENST00000456074	ZIK1P1-001		
ENST00000536878	ZIK1-002		
ENST00000597219	ZIK1-006		
ENST00000597850	ZIK1-001		
ENST00000598689	ZIK1-007		
ENST00000598726	ZIK1-008		
ENST00000599456	ZIK1-003		

Below the table are navigation buttons: 'First', 'Previous', '1', 'Next', 'Last'. The main content area displays details for the selected transcript: 'Transcript: ZIK1-004 ENST00000307468'. The 'Information' tab is active, showing the following details:

- Name - Source:** ZIK1-004 (HGNC transcript name)
- Gene:** This transcript is a product of gene [ZIK1 - ENSG00000171649](#)
- Protein:** This transcript corresponds to protein [ENSP00000303820](#).
- Location:** Chromosome 19: 57584260-57592390 forward strand
- Transcription Start Site (TSS):** 57584260
- Length:** 2510
- Transcript Support Level (TSL):** TSL:1
- Gencode annotation:** GENCODE basic
- GC content:** 47.45 %
- Biotype:** protein\_coding
- Status:** Known
- Annotation method:** Havana
- Version:** ENST00000307468.4
- Description:** zinc finger protein interacting with K protein 1 [Source:HGNC Symbol;Acc:HGNC:33104 [External Link to HGNC](#)]

At the bottom, there are two buttons: '3'UTR Visualization' (with a 'Visualization' sub-button) and 'Download'.

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.



Ensemble Transcript id	Transcript name	Information
ENST00000307468	ZIK1-004	<p>3'UTR sequence</p> <p>GAGTGTACAGTCAAAGGCAGGTTTCATCCACACAGAAGACTCAATCCTGTGAGATGTGTGCCAGTCCGAAAGATATT                      TTGCATCTAGCTGATCTCCCTGGGCAGAAACCATACTTGGTTGGAGAATGTACAAACCATCACCAGCACAGAAAGCATCA                      CAGTGCAAAGAAATCCTTGAAGAGGGACATGGACAGAGCCATATGTGAAGTGCTGCCTATTCTGTATGTCATTGAAGC                      CCTTCGAAATGGGAGGTTGAAAGGACCTCCAGCCATGTTGCGGCTTCTGAGGTCCTGGTCTTTCTGGAGGCAAG                      AAACCCGGCACAATTACTGAATGTGGGAGGACATTGCGAGTCAAAAAAGTCATTACAAGTCAGGTGAATGTGGAAAGGC                      TTCCAGGCACAAACACACTCCTGTTTACCATCCAAGAGTCTACACTGGAAAAAGCTTTATGAGTGTAGCAAAATGTGGGA                      AAGCCTTCCGTGGCAAGTACTCACTTGTTCAGCACAGAGAGTCCATACTGGAGAAAGGCCTGGGAGTGAATGAATGT                      GGAAATTTCTTAGCCAAACCTCCACCTGAATGATCATCGGAGAATCCACACCGGAGAAAGGCCTTATGAGTGCAGCGA                      ATGTGGAAAATTTTAGCAAAACTCCAGCCTTGTGACCACAGAAAAATACACACTGGAGCAAGGCCTTATGAGTGTGA                      GCCAGTGTGGGAAATCCTTTAGCCAAAAAGCCACCTTGTAAACACCAAAGAGTTCACACTGGAGAAAGGCCTTATAAG                      TGTGGTGAATGTGGGAATCCTTTAGTCAAAGTGCATTCTTAATCAACACCGAAGAATTCACTGGAGCAAAAGCCTTA                      TGAGTGTGGCCAGTGTGGGAAATCCTTTAGTCAAAAAGCTACCTCATTAAACACAGAGAGTTCACACTGGAGAAAGGC                      CTTATAAGTGTGGTGAAGTGTGGGAAATCCTTTAGTCAAAGCTCCATCCTTATTCAACACCGGAGAAATTCATACTGGAGCA                      CACACTTAAGTAGAGCCTTAGACCTACAGGGAAAGTGTCTCTGTAGTATTTAGCAGTAGAGAGCCTTTGTGAGGGGA                      GCCATCTGCCGTGAAGTTGAACCTCATTCTCTTGTCTCTGGTAGAAACCATCTACCCTTACCACCTTGACAGTGG                      GCACTGGTCACTCCTATGTGCTAAGACAAGGCAGACATCTGTGTCTCTTAAGTCTTTGGAGGAAATCTTGAGCAGTC                      TAAGCCTTTAGAGAAAATTCATTCTTTTTCTGACTGATCACAGCATACGTGTGACCCAGTTTGGGTCAAGGAGGCCAG                      CCTTGGTTCTGCTGGACACTATGTGCAAGGATCCCTTCAATGAAATTTCTGGTCTCACATGACACTTGGTCACTTCTTC                      CAGCCTCCATGTCACCACGTGGTGAATGGCTGCCTCACATTGCTCCAGTTTGTGCACTAATAAAAGCCTTATATTTGAAT                      CTACCTGATGCTTGGGTTCTGTTTACTGTGTGGGTGGCTGGGAGACAGACTTCAACTCTATATGAAGGAATGGATGG                      CTTTTGTGGCCTCTGCAAGGAAAGTAAGATGACAGAGTAATCTAATTTCTGGTTTTGGTCACTTTGCTTGTACCTAA                      AATCCTAGGAAAAAATGCAAGGTTTTGGTATTCTAATTTGTGGCTGGATCCCTATCTTTCTGTGAGACTAGAGGT                      CATCTGAGGAGAGGCAGCTGTTATGACAAGCATGTGTCTCAGGGAATAGGACAATTTATTCCATTGTTTCCAGAG                      CATTCTGAGGAGAGGCAGCTGTTATGACAAGCATGTGTCTCAGGGAATAGGACAATTTATTCCATTGTTTCCAGAG</p>
ENST00000456074	ZIK1P1-001	
ENST00000536878	ZIK1-002	
ENST00000597219	ZIK1-006	
ENST00000597850	ZIK1-001	
ENST00000598689	ZIK1-007	
ENST00000598726	ZIK1-008	
ENST00000599456	ZIK1-003	
ENST00000600000	ZIK1-005	
ENST00000600000	ZIK1-005	

## Genes information

The screenshot shows the InSyBio Gene Search Tool interface. The search term 'zik1' is entered in the search bar, and the 'Show results' button is highlighted. The 'Genes' tab is selected, displaying a list of genes with their Ensemble Gene IDs and Official Gene Symbols. The gene ZIK1 (ENSG00000171649) is highlighted. The 'Information' tab is selected for ZIK1, showing the following details:

- Name - Source:** ZIK1 (HGNC Symbol)
- Description:** zinc finger protein interacting with K protein 1 [Source:HGNC Symbol;Acc:HGNC:33104 External Link to HGNC]
- Location:** Chromosome 19: 57578456-57593777 forward strand
- Transcript count:** 8
- Biotype:** protein\_coding
- Status:** Known
- Annotation method:** Annotation for this gene includes both automatic annotation from Ensembl and Havana External Link manual curation, see article External Link
- Version:** ENSG00000171649.11

At the bottom of the page, there are navigation buttons: First, Previous, 1, Next, Last.

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.

The screenshot shows the InSyBio Gene Search Tool interface. The search term 'zik1' is entered in the search bar, and the 'Show results' button is highlighted. The 'Genes' tab is selected, displaying a list of genes with their Ensemble Gene IDs and Official Gene Symbols. The gene ZIK1 (ENSG00000171649) is highlighted. The 'Transcript Table' tab is selected for ZIK1, showing a table of transcripts associated with the gene:

#	Ensemble id	Name
1	ENST00000536878	ZIK1-002
2	ENST00000597219	ZIK1-006
3	ENST00000597850	ZIK1-001
4	ENST00000598689	ZIK1-007
5	ENST00000598726	ZIK1-008
6	ENST00000599456	ZIK1-003
7	ENST00000600053	ZIK1-005
8	ENST00000307468	ZIK1-004

At the bottom of the page, there are navigation buttons: First, Previous, 1, Next, Last.

# 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.

InSyBio Suite - RNA-Seq Differential Expression Pipeline

RNA-Seq  Paired-end  Single-ended

Data:

Condition Control:  \* Required information

Title Read 1:  Title Read 2:

Filename Read 1:  Filename Read 2:

Title Read 1:  Title Read 2:

Filename Read 1:  Filename Read 2:

Condition 1:

Title Read 1:  Title Read 2:

Filename Read 1:  Filename Read 2:

Title Read 1:  Title Read 2:

Filename Read 1:  Filename Read 2:

Options

Do you want to perform initial FastQC?

Do you want to perform trimming?

InSyBio Suite - RNA-Seq Differential Expression Pipeline

RNA-Seq  Paired-end  Single-ended

Data:

Condition Control:  \* Required information

Title:

Filename:

Title:

Filename:

Condition 1:

Title:

Filename:

Title:

Filename:

Options

Do you want to perform initial FastQC

Do you want to perform trimming?

- 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.

### Options

Do you want to perform initial FastQC

Do you want to perform trimming?

### Alignment Options

Source for the reference genome \*

Specify strand information:

- 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

**Options**

Do you want to perform initial FastQC

Do you want to perform trimming?

---

**Trimmomatic Options**

Perform initial ILLUMINACLIP step?

Select standard adapter sequences or provide custom? \*

Adapter sequences to use: \*

**1. Trimmomatic Operation**

Number of bases to average across:

Average quality required:

- 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.

### Alignment Options

Source for the reference genome \*

Use a genome from Data Store ▾

Select the reference genome (FASTA): \*

Title: chr22 fasta

Filename: dsfile1573556494\_9916.fa

Select the reference genome (GTF): \*

Title: chr22 GTF

Filename: dsfile1573556655\_8832.gtf

### Alignment Options

Source for the reference genome \*

Use a built-in genome ▾

Select a reference genome: \*

HumanGRCh38 ▾

Specify strand information:

Forward (FR) ▾

- 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.

Which conditions do you want to compare? Set manually ▾

	Control ▾	Tumor ▾	-
	Control ▾	Treated ▾	-
Condition Pairs:	Tumor ▾	Treated ▾	-

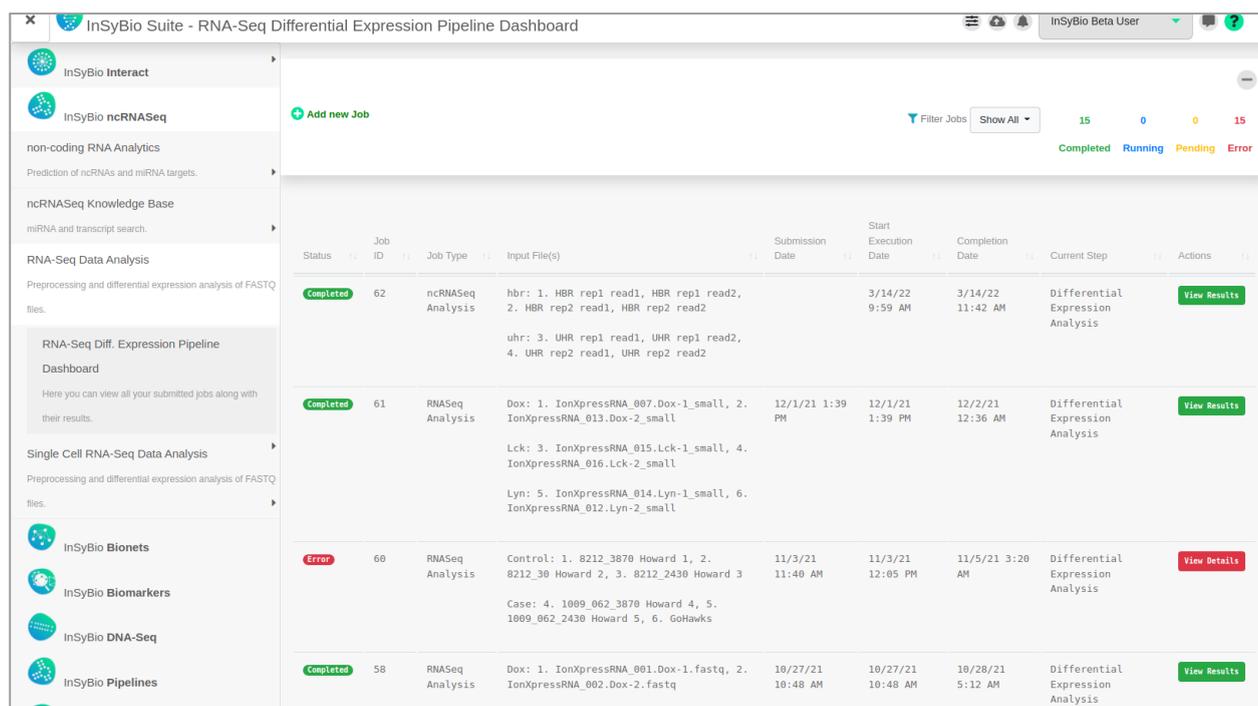
+

**RNASeq Analysis**   **ncRNASeq Analysis**

**Clear All**

## 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.



Status	Job ID	Job Type	Input File(s)	Submission Date	Start Execution Date	Completion Date	Current Step	Actions
Completed	62	ncRNASeq Analysis	hbr: 1. HBR rep1 read1, HBR rep1 read2, 2. HBR rep2 read1, HBR rep2 read2 uhr: 3. UHR rep1 read1, UHR rep1 read2, 4. UHR rep2 read1, UHR rep2 read2	3/14/22 9:59 AM	3/14/22 11:42 AM	Differential Expression Analysis	View Results	
Completed	61	RNASeq Analysis	Dox: 1. IonXpressRNA_007.Dox-1_small, 2. IonXpressRNA_013.Dox-2_small Lck: 3. IonXpressRNA_015.Lck-1_small, 4. IonXpressRNA_016.Lck-2_small Lyn: 5. IonXpressRNA_014.Lyn-1_small, 6. IonXpressRNA_012.Lyn-2_small	12/1/21 1:39 PM	12/1/21 1:39 PM	12/2/21 12:36 AM	Differential Expression Analysis	View Results
Error	60	RNASeq Analysis	Control: 1. 8212_3870 Howard 1, 2. 8212_38 Howard 2, 3. 8212_2430 Howard 3 Case: 4. 1089_062_3870 Howard 4, 5. 1089_062_2430 Howard 5, 6. GoHawks	11/3/21 11:40 AM	11/3/21 12:05 PM	11/5/21 3:20 AM	Differential Expression Analysis	View Details
Completed	58	RNASeq Analysis	Dox: 1. IonXpressRNA_001.Dox-1.fastq, 2. IonXpressRNA_002.Dox-2.fastq Lck: 3. IonXpressRNA_003.Lck-1.fastq, 4. IonXpressRNA_004.Lck-2.fastq Lyn: 5. IonXpressRNA_005.Lyn-1.fastq, 6. IonXpressRNA_006.Lyn-2.fastq	10/27/21 10:48 AM	10/27/21 10:48 AM	10/28/21 5:12 AM	Differential Expression Analysis	View Results

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.

**InSyBio Suite Beta - RNA-Seq Differential Expression Pipeline Results**

**Job Status** **Job ID** **Submission Date** **Execution Time** **Input Data and Parameters**

COMPLETED 1 May 6, 2019 7:55:09 AM 00 hours, 15 minutes, 49 seconds

**Deseq2 Reports** Initial FastQC Reports Trimmed FASTQ Files Trimmed FastQC Reports Alignment Files Read Count Files Next Actions

**HBR\_UHR**

Deseq2 Report File (.pdf) Download

Job-1 DESeq2 pdf output File

Deseq2 Report File (.png) Download

HBR\_UHRimages.zip Image Folder

Deseq2 Report File (.csv) Download

Job-1 DESeq2 output HBR\_UHR\_diffexpr-results-with-counts.csv (HBR\_UHR\_diffexpr-results-with-counts.csv); File

Job-1 DESeq2 output HBR\_UHR\_diffexpr-results.csv (HBR\_UHR\_diffexpr-results.csv); File

Job-1 DESeq2 output HBR\_UHR\_diffexpr-resultssignificant\_pvalues.csv (HBR\_UHR\_diffexpr-results\_significant\_pvalues.csv); File

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

**Deseq2 Reports** **Initial FastQC Reports** Trimmed FASTQ Files Trimmed FastQC Reports Alignment Files Read Count Files Next Actions

FastQC Report Download View Html Page

Job-1 Fastqc zip file HBR rep1 read1 Folder dsfile1557128487\_9359\_fastqc

Job-1 Fastqc zip file HBR rep1 read2 Folder dsfile1557128516\_9128\_fastqc

Job-1 Fastqc zip file HBR rep2 read1 Folder dsfile1557128550\_6204\_fastqc

Job-1 Fastqc zip file HBR rep2 read2 Folder dsfile1557128587\_1781\_fastqc

Job-1 Fastqc zip file HBR rep3 read1 Folder dsfile1557128617\_6024\_fastqc

Job-1 Fastqc zip file HBR rep3 read2 Folder dsfile1557128647\_9984\_fastqc

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

Deseq2 Reports	Initial FastQC Reports	Trimmed FASTQ Files	Trimmed FastQC Reports	Alignment Files	Read Count Files	Next Actions
Trimmed FASTQ File					Download	
Job-1 trimmend paired file of HBR rep1 read1 (dsfile1557128487_9359_trimmed.gz);					 File	
Job-1 trimmend paired file of HBR rep1 read2 (dsfile1557128516_9128_trimmed.gz);					 File	
Job-1 trimmend paired file of HBR rep2 read1 (dsfile1557128550_6204_trimmed.gz);					 File	
Job-1 trimmend paired file of HBR rep2 read2 (dsfile1557128587_1781_trimmed.gz);					 File	
Job-1 trimmend paired file of HBR rep3 read1 (dsfile1557128617_6024_trimmed.gz);					 File	
Job-1 trimmend paired file of HBR rep3 read2 (dsfile1557128647_9984_trimmed.gz);					 File	
Job-1 trimmend paired file of UHR rep1 read1 (dsfile1557128760_6526_trimmed.gz);					 File	

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

Deseq2 Reports	Initial FastQC Reports	Trimmed FASTQ Files	Trimmed FastQC Reports	Alignment Files	Read Count Files	Next Actions
Trimmed FastQC Report			Download	View Html Page		
s:51:"Job-1 after trimming Fastqc zip file HBR rep1 read1";			 File	 dsfile1557128487_9359_trimmed_fastqc		
s:51:"Job-1 after trimming Fastqc zip file HBR rep1 read2";			 File	 dsfile1557128516_9128_trimmed_fastqc		
s:51:"Job-1 after trimming Fastqc zip file HBR rep2 read1";			 File	 dsfile1557128550_6204_trimmed_fastqc		
s:51:"Job-1 after trimming Fastqc zip file HBR rep2 read2";			 File	 dsfile1557128587_1781_trimmed_fastqc		
s:51:"Job-1 after trimming Fastqc zip file HBR rep3 read1";			 File	 dsfile1557128617_6024_trimmed_fastqc		
s:51:"Job-1 after trimming Fastqc zip file HBR rep3 read2";			 File	 dsfile1557128647_9984_trimmed_fastqc		

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

The screenshot displays the 'Alignment Files' tab in the InSyBio interface. It is organized into three main sections: SAM Files, BAM Files, and Run Info. Each section has a 'Download' link. The SAM Files section lists six files: HBR\_1.sam, HBR\_2.sam, HBR\_3.sam, UHR\_1.sam, UHR\_2.sam, and UHR\_3.sam. The BAM Files section lists six files: HBR\_1.bam, HBR\_2.bam, HBR\_3.bam, UHR\_1.bam, UHR\_2.bam, and UHR\_3.bam. The Run Info section contains a single file named 'hisat2\_report.txt'. Each file entry includes a green download icon and the text 'File'.

File Name	Download Link
Job-1 Hisat2 alignment file HBR_1.sam (HBR_1.sam);	File
Job-1 Hisat2 alignment file HBR_2.sam (HBR_2.sam);	File
Job-1 Hisat2 alignment file HBR_3.sam (HBR_3.sam);	File
Job-1 Hisat2 alignment file UHR_1.sam (UHR_1.sam);	File
Job-1 Hisat2 alignment file UHR_2.sam (UHR_2.sam);	File
Job-1 Hisat2 alignment file UHR_3.sam (UHR_3.sam);	File
BAM File	
Job-1 BAM file HBR_1.bam (HBR_1.bam);	File
Job-1 BAM file HBR_2.bam (HBR_2.bam);	File
Job-1 BAM file HBR_3.bam (HBR_3.bam);	File
Job-1 BAM file UHR_1.bam (UHR_1.bam);	File
Job-1 BAM file UHR_2.bam (UHR_2.bam);	File
Job-1 BAM file UHR_3.bam (UHR_3.bam);	File
Run Info	
Alignment Info	hisat2_report.txt

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

Read Count File	Download	Download Run Info File
Job-1 Feature counts file (HBR_1.counts);	 HBR_1.counts	 HBR_1.features.summary
Job-1 Feature counts file (HBR_2.counts);	 HBR_2.counts	 HBR_1.features.summary
Job-1 Feature counts file (HBR_3.counts);	 HBR_3.counts	 HBR_1.features.summary
Job-1 Feature counts file (UHR_1.counts);	 UHR_1.counts	 HBR_1.features.summary
Job-1 Feature counts file (UHR_2.counts);	 UHR_2.counts	 HBR_1.features.summary
Job-1 Feature counts file (UHR_3.counts);	 UHR_3.counts	 HBR_1.features.summary

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

Job Status	Job ID	Submission Date	Execution Time	Input Data and F
<b>COMPLETED</b>	79	Oct 2, 2019 8:56:41 AM	00 hours, 01 minutes, 56 seconds	

Predicted ncRNAs	Download
Predicted ncRNAs file	 File

If ncrRNASeq 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.

The screenshot displays the 'Next Actions' tab in the InSyBio Suite for the HBR\_UHR dataset. The interface is organized into two main sections: 'Molecule Quantification Files per Condition' and 'Full Molecule Quantification File and Associated Labels'. Each section contains a list of files with associated actions.

File Name	Download	Next Action
Job-1 MQ file HBR_UHR_diffexpr-MQHBR.csv (HBR_UHR_diffexpr-MQHBR.csv);	 File	--Select Action--
Job-1 MQ file HBR_UHR_diffexpr-MQUHR.csv (HBR_UHR_diffexpr-MQUHR.csv);	 File	--Select Action--
Full Molecule Quantification File and Associated Labels		
Job-1 MQ file HBR_UHR_diffexpr-MQ.csv (HBR_UHR_diffexpr-MQ.csv);	 File	--Select Action--
Job-1 label file HBR_UHR_diffexpr-labels.txt (HBR_UHR_diffexpr-labels.txt);	 File	--Select Action--

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.

InSyBio Suite - Single Cell RNA-Seq Differential Expression Pipeline

Workflow: One Sample, One GEM Well, One Flowcell

**Input Data Files**

Choose or upload to input your Fastq files to InSyBio Single Cell RNA-Seq Differential Expression Pipeline tool following the rules:

- 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

**Fastq File 1**

Title1:

Filename 1:

[Select file from Data Store](#) [Go to Data Store to Upload File](#)

**Options**

Transcriptome: Human

Cluster annotation

Species: --Select Action--

Tissue: --Select Action--

InSyBio Suite - Single Cell RNA-Seq Differential Expression Pipeline

Workflow: One Sample, Multiple GEM Wells, One Flowcell

**Input Data Files**

Choose or upload to input your Fastq files to InSyBio Single Cell RNA-Seq Differential Expression Pipeline tool following the rules:

- 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

**Fastq File 1**

Title1:

Filename 1:

[Select file from Data Store](#) [Go to Data Store to Upload File](#)

**Fastq File 2**

Title2:

Filename 2:

[Select file from Data Store](#) [Go to Data Store to Upload File](#)

[Add File](#)

**Options**

- 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

**Advanced Options** +

Expected number of recovered cells

BAM file generation

First filtering

Minimum cells:

Minimum features:

Secondary filtering

nFeature\_RNA  Lower limit:

Upper limit:

nCount\_RNA

Feature Extraction Method

Shared Nearest Neighbor (SNN) Graph

k parameter (k-nearest-neighbor):

Clustering

Resolution parameter

Differentially expressed genes criteria

Threshold (logfc):

Minimum Pct:

Plot for the top differentially expressed genes for each cluster

Number of top markers per cluster:

Average log2FC

Genes for visualization  All  Custom

- Submit your job pressing the respective button.

[Submit Job](#)

## 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

InSyBio Suite - Single Cell RNA-Seq Differential Expression Pipeline Dashboard

InSyBio **Interact**

InSyBio **ncRNASeq** [Add new job](#)

Filter Jobs Show All - 1 1 0 2

Completed Running Pending Error

Status	Job ID	Job Type	Input File(s)	Submission Date	Start Execution Date	Completion Date	Current Step	Actions
Error	1	RNASeq Analysis		2/9/22 1:10 PM	2/28/22 9:56 AM	2/27/22 7:20 PM	Single Cell Alignment	<a href="#">View Details</a>
Completed	2	RNASeq Analysis		2/23/22 1:21 PM	2/28/22 6:51 AM	2/28/22 8:04 AM	Secondary Single Cell Analysis	<a href="#">View Results</a>
Error	3	RNASeq Analysis		3/9/22 8:24 PM	2/28/22 5:58 PM		Single Cell Alignment	<a href="#">View Details</a>
Running	4	RNASeq Analysis		3/15/22 10:08 AM			Secondary Single Cell Analysis	<a href="#">View Details</a>

First Previous 1 Next Last

Show 50 entries Showing 1 to 4 of 4 entries

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.

InSyBio Suite - RNA-Seq Single Cell Pipeline Differential Expression Pipeline Results
InSyBio Beta User

Job Status	Job Type	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	RNASeq Single Cell Analysis	2	Feb 23, 2022 1:21:53 PM	01 hours, 13 minutes, 17 seconds	<span style="color: green;">i</span>

**Report** | Summary | Additional Cell Statistics | Dot Plots Visualization | Feature Plots Visualization | Ridge Plots Visualization | Umap Plots Visualization | All Results Download

### Single Cell Pipeline Report

Alignment of the sequencing reads in the provided FASTQ files to the selected reference transcriptome and read counts computation are performed with the Cellranger count pipeline. The outs folder contains the outputs of this step and includes the web\_summary.html file which summarizes the results.

### Secondary Single Cell Analysis

For the secondary single cell analysis quality control checks and filtering criteria are applied to the single cell data. With the Seurat Object the data are filtered using min.cells = 0 and min.features = 0.

min.cells: Include features detected in at least this many cells.  
min.features: Include cells where at least this many features are detected.

An additional filtering step is performed with Seurat, keeping only cells that have unique feature counts and total number of molecules detected within a cell with the following limits:  
nFeature\_RNA = unique feature counts, lower limit: 100, upper limit: 3000  
nCount\_RNA = total number of molecules detected within a cell, lower limit: , upper limit:

The data are then normalized using the LogNormalize method, which normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10.000) and log-transforms the result.  
2000 highly variable features that exhibit high cell-to-cell variation in our data are identified. Scaling is subsequently performed scaled, so that the mean expression across cells is 0 and variance across cells is 1. This last step is necessary for performing PCA on the data

The cells are clustered using a modularity optimization technique called Louvain algorithm with a resolution parameter of 1 (it sets the granularity of the downstream clustering) having firstly constructed the KNN graph (with k=30) based on the Euclidean distance in PCA space and using Jaccard similarity. Using the clustered data, non-linear dimensionality reduction is performed, producing the Umap plot.

In scRNA seq data analysis, differentially expressed features that define the clusters are called markers. To identify these markers, we firstly used the FindAllMarkers() function of the Seurat package, which identifies these markers for all clusters by comparing all clusters with each other. For this function we used parameters min.pct (a feature to be detected at a minimum percentage in either of the groups of cells) with value 0.1 and logfc.threshold (Limit testing to genes which show, on average, at least X-fold difference (log-scale) between the two groups of cells) with value 0.25. The matrices produced by these functions contain the genes as rows and these specific associated statistics for each gene as columns: P value, Average log2 Fold Change, Percentage of cells 1, Percentage of cells 2 and Adjusted P value.

The Dotplots include the differentially expressed genes that are only differentially expressed in one cluster of cells while sorting them by their p value.

The scCATCH package, a single cell Cluster-based annotation Toolkit for Cellular Heterogeneity is finally used to identify the cluster marker genes and creates the cluster annotations. We used the scCATCH() function which does the cluster annotation by matching the potential marker genes with known cell marker genes in a tissue-specific cell taxonomy reference database (CellMatch). We used the cancer type: and tissue types: Blood, Bladder.

The selected species was Human.

### Results files description

Outs folder: The output files of the cellranger platform.

Web\_summary.html: Variety of metrics such as Mean Reads per Cell, Median Genes per Cell, Valid Barcodes etc. At the analysis tab, t-SNE projection can be seen with UMI Counts or Clustered. Also, info about the Top features by cluster can be found.

Results of the secondary single cell analysis:

RidgePlots folder: Folder containing a Ridge Plot per gene you selected.

FeaturePlots folder: Folder containing a Feature Plot per gene you selected.

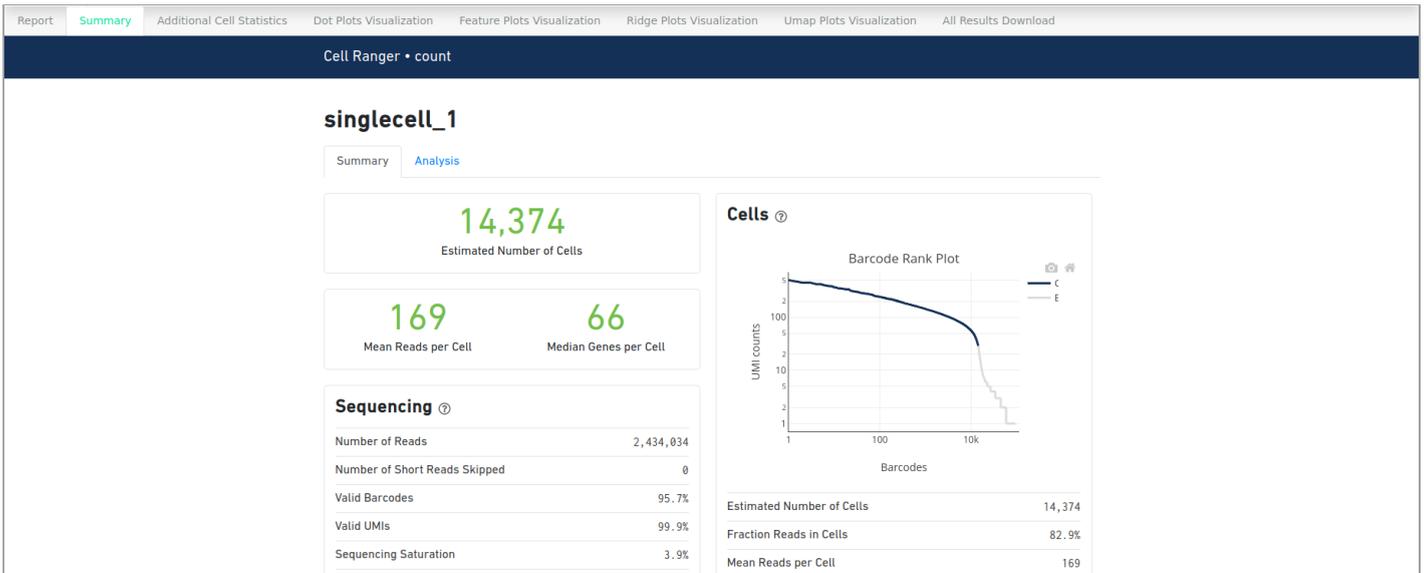
Dimensionality Reduction Plot folder:  
Contains Umap.png: Umap projection plot of the clusters.

Markers folder:  
markers\_from\_FindAllMarkers and markers\_from\_scCATCH: Markers(differentially expressed genes) and associated statistics (p-values, avg\_log2FC etc) from FindAllMarkers and scCATCH functions respectively.  
average\_expression\_of\_genes.csv: Averaged expression values for every gene for every cluster.  
Barcode-cluster.csv: Barcode-cluster matrix.

Dotplots Folder: Folder containing all the dotplots needed. (Dotplot\_unique, Dotplot\_only\_specific\_genes)  
DotPlot\_unique.pdf: Top 5 unique differentially expressed genes for each cell cluster based on the p-value and log2fc value.  
Dotplot\_only\_specific\_genes.pdf: Same dotplot as the previous ones but for the specific genes you selected.

You can find all these files compressed at their respective zip file.

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.

Report	Summary	Additional Cell Statistics	Dot Plots Visualization	Feature Plots Visualization	Ridge Plots Visualization	Umap Plots Visualization	All Results Download
Total Markers		Markers with Cluster Annotation	Average Expression of genes	Barcode Cluster			
Total Markers Results							
<a href="#">Download Total Markers CSV</a>							
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	1.850e-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.

Total Markers	Markers with Cluster Annotation	Average Expression of genes	Barcode Cluster			
Markers with Cluster Annotation Results						
<a href="#">Download Markers with Cluster Annotation CSV</a>						
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, UBE25, RPL34, TFI27, RPL36A, SUB1, PFN1, RPL18, MT-ND5, RPS6, HNRNPA2B1, COTL1, S100A6, TRAC, HMGB1, TXN, RPL29, S100A11, RPS2, TP11, RPL14, SNHG29, RPL28, TUBB, BNIP3, ACTB, VIM, MYL12A	Dendritic Cell	0.65	FTL, S100A11, S100A4, TXN	28428369.0		
UBE2C, CALM2, UBE25, TUBA1B, TURB, 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, TM6IM6, 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, UQCRCB, 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	135.876	145.658	145.848	158.907	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	131.086	128.572	143.985	155.222	139.74	133.718	141.753
MT-CO2	125.888	123.558	128.09	143.324	137.983	127.243	126.882	137.135	136.51
B2M	116.847	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	97.826	97.815	75.135	186.436	96.976	112.873	187.371	85.869
MT-ATP6	47.766	60.765	67.886	80.226	69.185	77.966	68.456	61.839	83.244
MT-ND4	69.789	59.943	63.879	61.415	73.183	63.785	69.986	62.776	74.848
RPS18	74.245	53.424	65.431	64.848	64.979	64.94	70.533	71.987	66.842
RPL41	72.988	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	67.383
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.

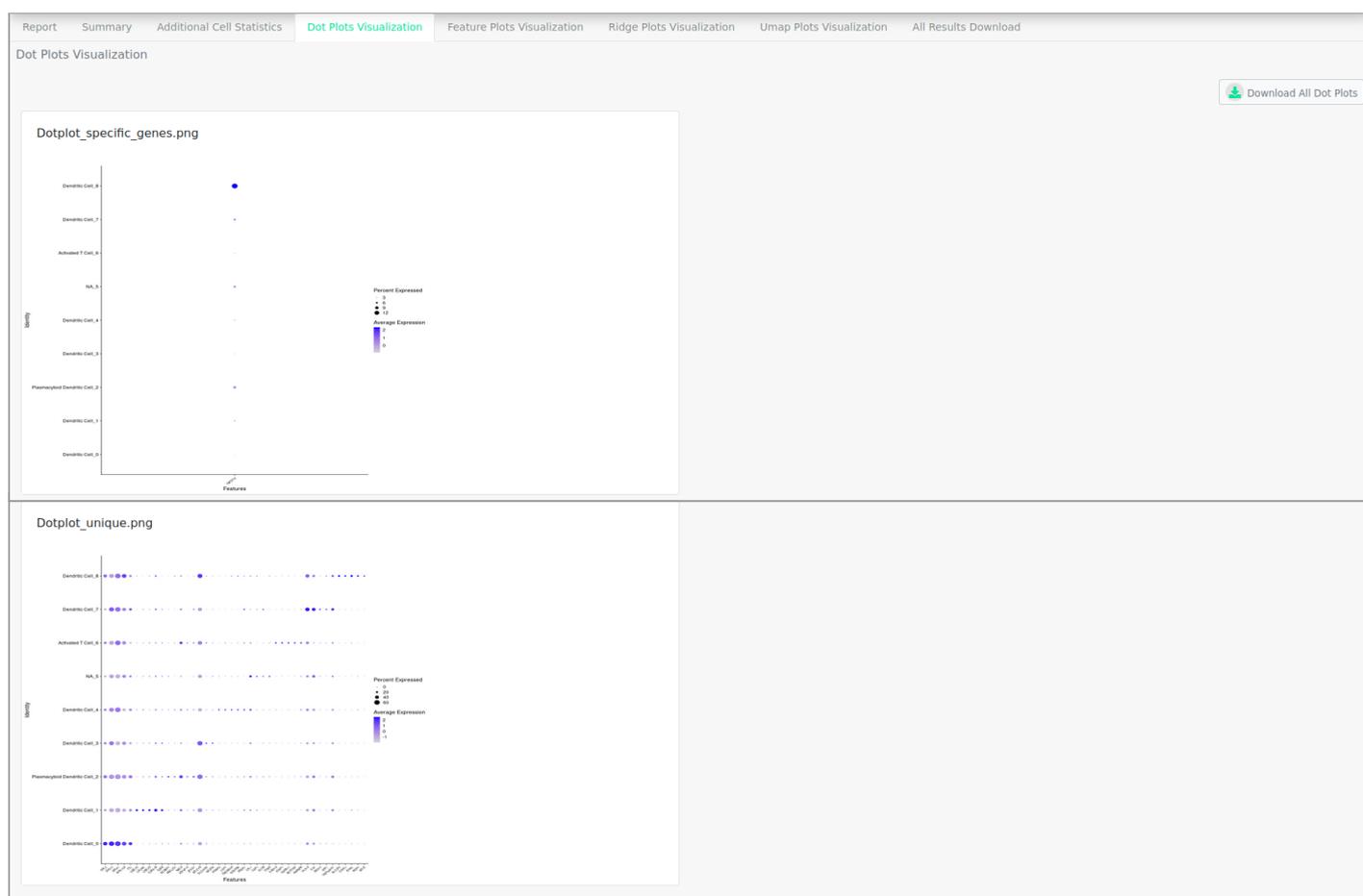
Total Markers   Markers with Cluster Annotation   Average Expression of genes   **Barcode Cluster**

Barcode Cluster Results

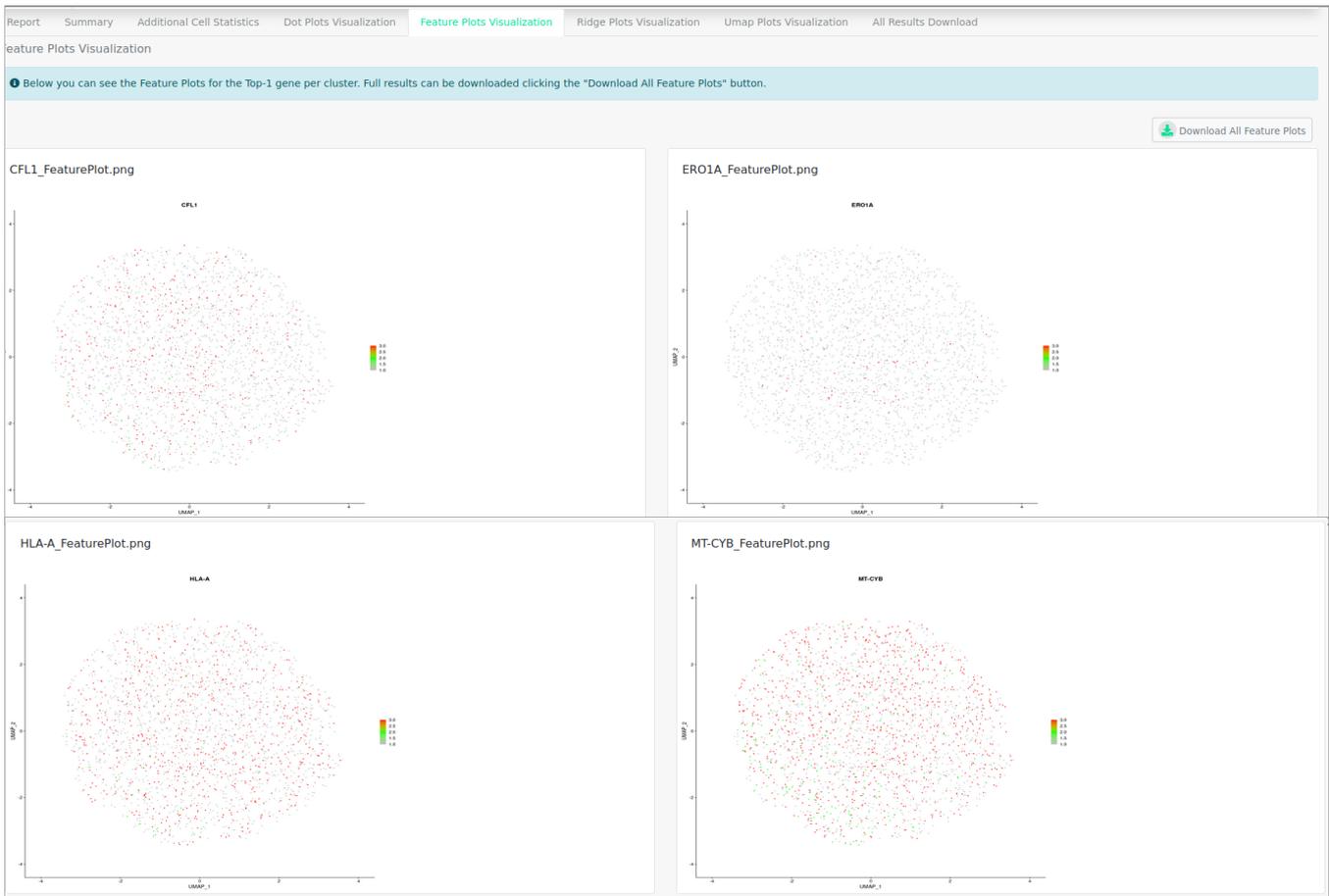
[Download Barcode Cluster CSV](#)

Barcode	Cluster
"AAACCCACATATAGCC-1"	"Activated T Cell_6"
"AAACCCATCAGTCCT-1"	"Dendritic Cell_8"
"AAACCCATCGCATGAT-1"	"Dendritic Cell_3"
"AAACGAAACAATAGGAT-1"	"Plasmacytoid Dendritic Cell_2"
"AAACGAAACAAAGTA-1"	"NA_5"
"AAACGAAACATCTATCT-1"	"Dendritic Cell_3"
"AAACGCTAGCTACTGT-1"	"Dendritic Cell_8"
"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"
"AAAGGATCAGGAAAC-1"	"Dendritic Cell_4"
"AAAGGATCACTCATAG-1"	"Dendritic Cell_8"
"AAAGGATGTCCTATA-1"	"Dendritic Cell_1"

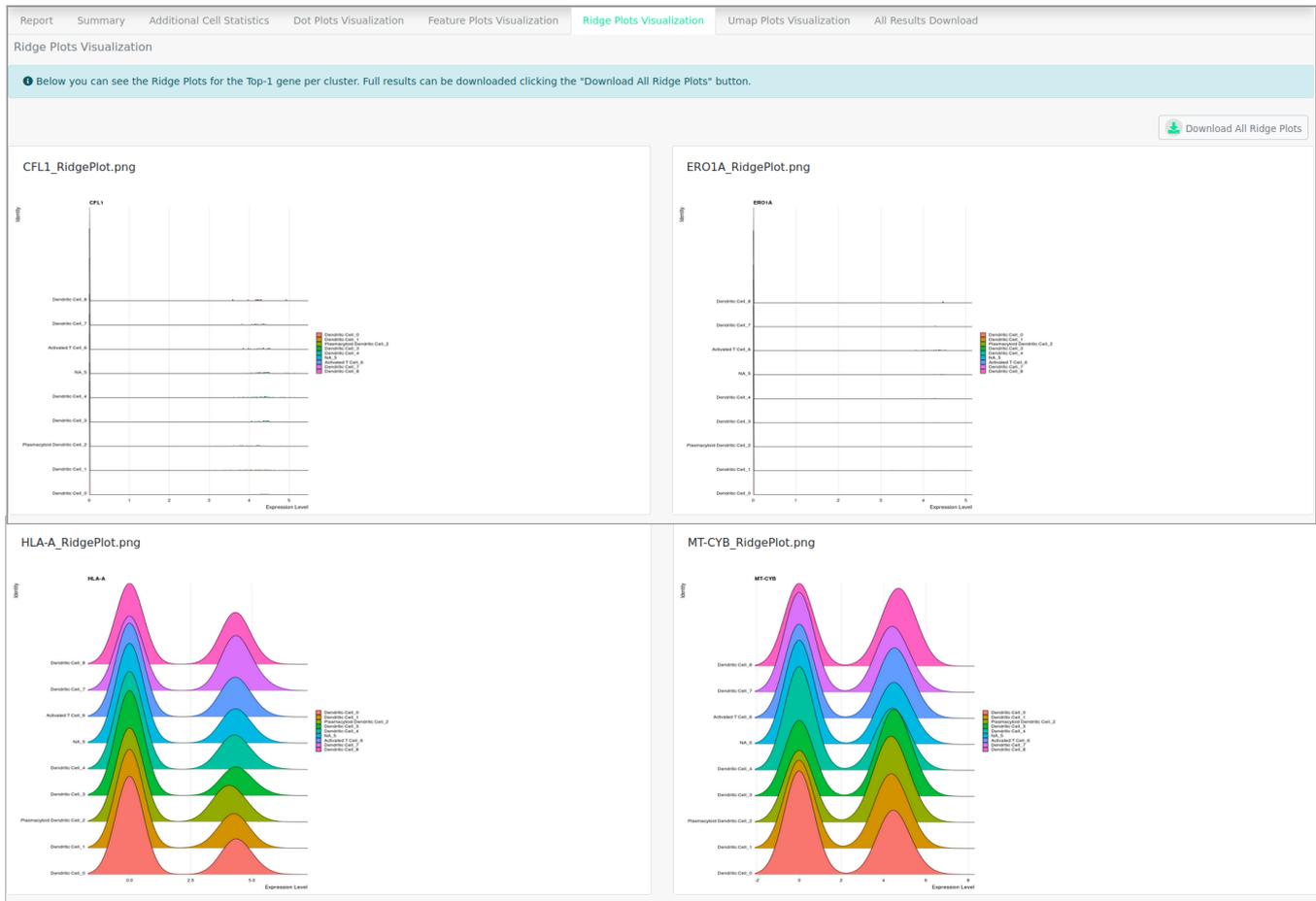
At the Barcode Cluster tab the Barcode-Cluster matrix can be found.



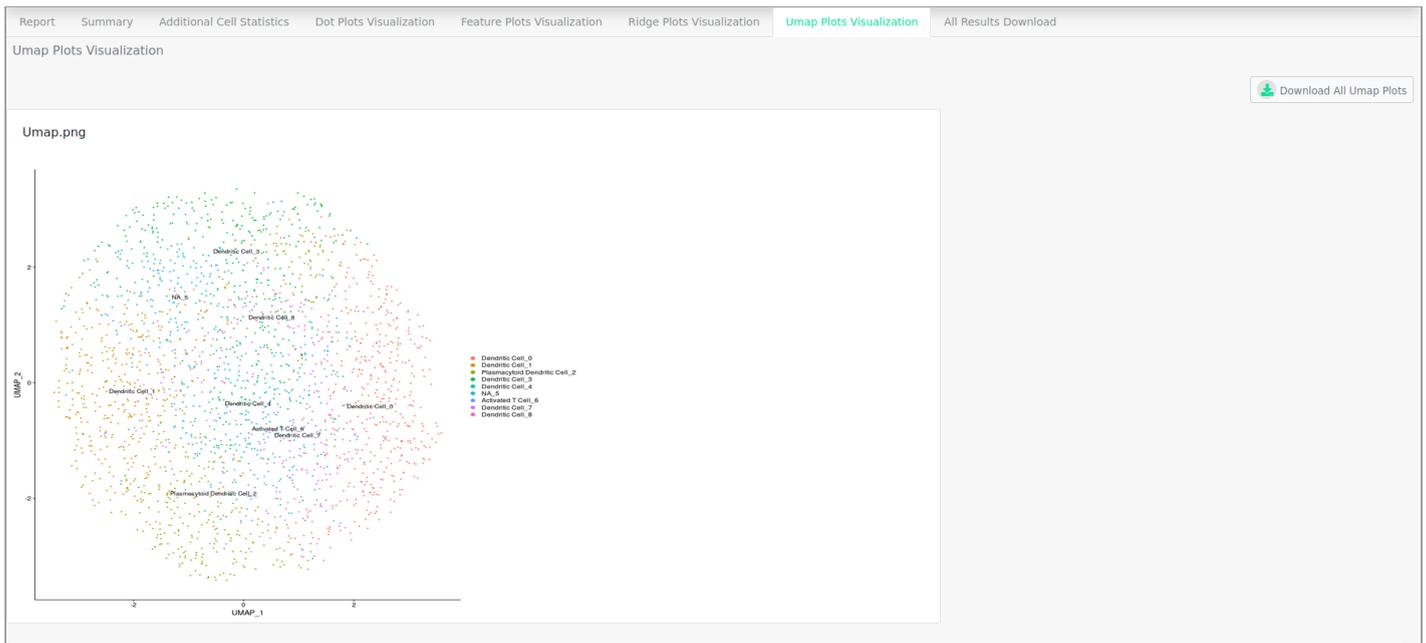
At the Dot Plots Visualization tab you can see the two Dot plots that are created. The first one is a Dot Plot with only the genes you specified at the manual parameters and second one is a Dot Plot that shows the Top 5 unique differentially expressed genes for each cell cluster based on the p-value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



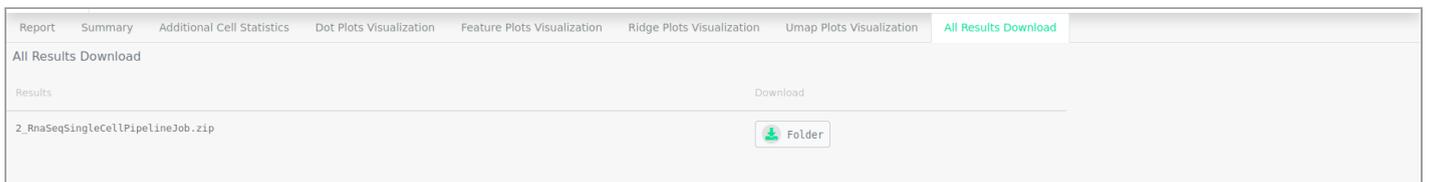
At the Feature Plots Visualization tab the Feature Plots for the Top-1 gene per cluster can be found. The Feature Plots of all the genes can be downloaded.



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](mailto:info@insybio.com).

To purchase InSyBio ncRNASeq commercial version 3.0 please contact us at [sales@insybio.com](mailto:sales@insybio.com).

## About Us

---

InSyBio Ltd is a bioinformatics pioneer company ([www.insybio.com](http://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**

External Publication of InSyBio Ltd - Any InSyBio information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the InSyBio Ltd. A draft of the proposed document should accompany any such request. InSyBio Ltd reserves the right to deny approval of external usage for any reason.

Copyright 2022 InSyBio Ltd. Reproduction without written permission is completely forbidden.