



InSyBio

Intelligent Systems Biology

User Manual

# Analyze DNA Sequencing data with InSyBio DNaseq

February 2021

Insybio Suite v2.6

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

# Introduction

---

DNaseq is a new tool which enables the fast and accurate pre-processing and analysis of DNA-sequencing data by non-bioinformatics experts with optimized pipelines. This tool includes the following functionalities:

- Pre-processing of DNA-sequencing data with **optimized pipelines and a user-friendly interface**
- Population analysis of genetics data for the **identification of significant genomics biomarkers**
- Integration of genomic biomarkers with InSyBio Suite's knowledge base to allow the **biological interpretation of your data**
- **Integration of genomic biomarkers with other omics biomarkers and clinical data using statistical and machine learning** functionalities of InSyBio Biomarkers.

# DNA-Seq 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 Bowtie2, and sorting with Samtools.
- C.** Variant Calling using Freebayes.
- D.** Variant Annotation using known databases with Ensemble VEP.

Firstly, the Pipeline uses Fastqc to create a report with the input sequences quality, then trimm the sequences accordingly using Trimmomatic and create new reports with Fastqc. Then using Bowtie2 it creates the alignment SAM files with the Genome files, we sort them using SAMtools and transform them to BAM files. The BAM files are used as input of Freebayes, that creates VCF files with the variants that it detects. At the end, Variant Annotation with VEP is performed, extra information like allele frequency, SIFT variant score and the variant's id from dbSNP is annotated and some supplementary plots are created with a script using R.

We also offer a Significant Gene file creation, where if only one cohort is used we create a file with the variants with the lowest SIFT score or if multiple cohorts are used we create pairs of cohorts and calculate their significant gene variants..

## To start the DNA-Seq Pipeline:

Click in the menu "InSyBio DNA-Seq" and you will be redirected to the "DNA-Seq Pipeline Dashboard" , select the "Add new job" button and then:

- Select if you have Single-Cohort or Multiple-Conditions and if you have Paired or Single Ended data that you want to analyze.

Pipeline InSyBio Beta User

What kind of data do you want to analyse?

Cohort Data:  Multiple-Conditions  Single-cohort

DNA-Seq Data:  Paired-end  Single-ended

Condition Control: ERR194147 \* Required information

Title: EPP194147 unpaired

Filename: dsfile1613044260\_6454.gz

Options

Do you want to perform initial FastQC

Do you want to perform trimming? --Select Action--

**Alignment Options**

Select a reference genome: \*

--Select Action--

Specify strand information:

Unstranded

**Filtering Options**

Allele Frequency threshold value

Pipeline InSyBio Beta User

What kind of data do you want to analyse?

Cohort Data:  Multiple-Conditions  Single-cohort

DNA-Seq Data:  Paired-end  Single-ended

Condition Control: ERR194147 \* Required information

Title Read 1: ERR194147 read1 Title Read 2: ERR194147 read2

Filename Read 1: dsfile1613032895\_3036.gz Filename Read 2: dsfile1613033004\_5428.gz

Options

Do you want to perform initial FastQC

Do you want to perform trimming? --Select Action--

**Alignment Options**

Select a reference genome: \*

--Select Action--

Specify strand information:

Unstranded

**Filtering Options**

Allele Frequency threshold value

Pipeline InSyBio Beta User

What kind of data do you want to analyse?

Cohort Data:  Multiple-Conditions  Single-cohort

DNA-Seq Data:  Paired-end  Single-ended

Condition Control:  \* Required information

Title:

Filename:

---

Condition 1:

Title:

Filename:

Options

Do you want to perform initial FastQC:

Do you want to perform trimming?

**Alignment Options**

Select a reference genome: \*

Pipeline InSyBio Beta User

What kind of data do you want to analyse?

Cohort Data:  Multiple-Conditions  Single-cohort

DNA-Seq Data:  Paired-end  Single-ended

Condition Control:  \* Required information

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:

Options

Do you want to perform initial FastQC:

Do you want to perform trimming?

**Alignment Options**

Select a reference genome: \*

- Name Conditions/Group of files you want to Analyze.
- For each condition add single or paired files by:

- Uploading a new file of DNA-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 DNA-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? --Select Action--

#### Alignment Options

Select a reference genome: \*

--Select Action--

Specify strand information:

Unstranded

#### Filtering Options

Allele Frequency threshold value

Significant Genes threshold value

DNaseq Analysis

Clear All

- 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? YES (Set Options ▾)

---

### Trimmomatic Options

Perform initial ILLUMINACLIP step? YES ▾

Select standard adapter sequences or provide custom? \* Standard ▾

Adapter sequences to use: \* TruSeq3 (single-ended, f ▾)

1. Trimmomatic Operation

Sliding window trimmi ▾

Number of bases to average across: 4 ▾

Average quality required: 15 ▾

Add Trimmomatic Operation

- Select the Genome the input files belong, from our 2 built-in options (HumanGRCh38 or MouseGRCm38).

**Alignment Options**

Select a reference genome: \*

Mouse GRCh38

Specify strand information:

Forward (FR)

**Filtering Options**

- Select the strandness of your input files, Unstranded, Forward or Reverse.
- Select Filtering Options, choose Allele Frequency threshold value (0.05 is recommended and the default value), and Significant Genes threshold value (0.1 is recommended and the default value)
- Last but not least select to perform the DNA-Seq Analysis.

**Filtering Options**

Allele Frequency threshold value	0.05
Significant Genes threshold value	0.1

**DNaseq Analysis**

**Clear All**

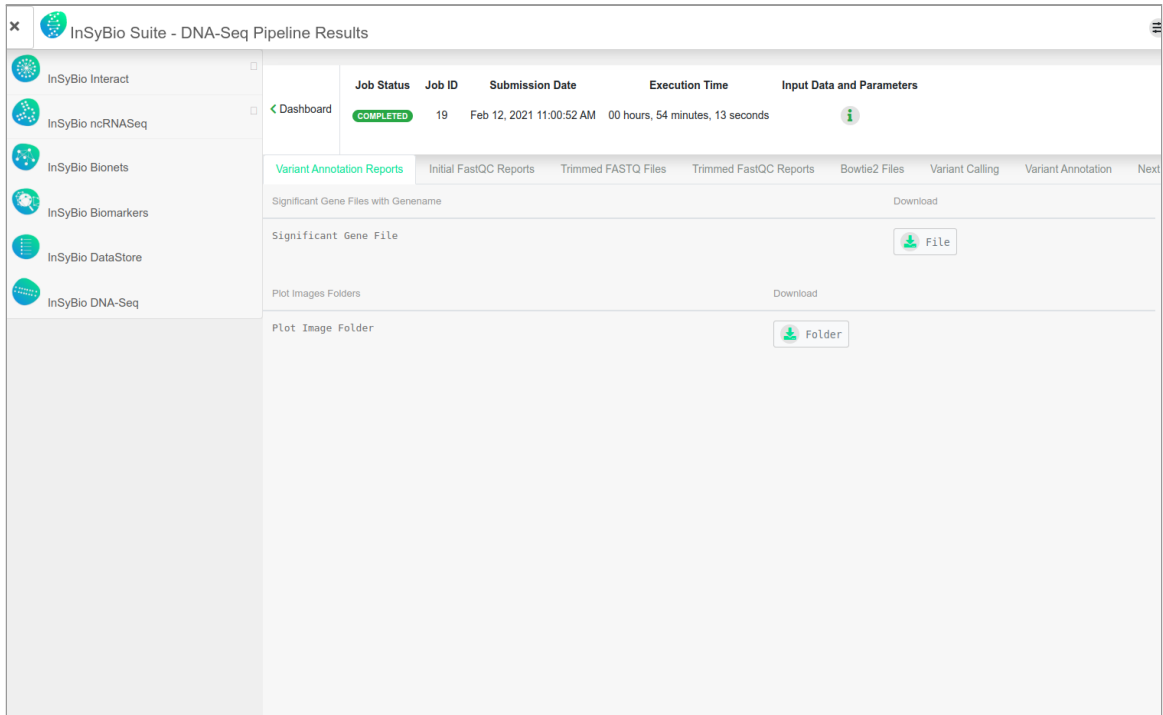


## To view the results:

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

Status	Job ID	Input File(s)	Submission Date	Start Execution Date	Completion Date	Current Step	Actions
Completed	19	err194147 fastqc: 1. EPP194147 unpaired	2/12/21 11:00 AM	2/12/21 11:00 AM	2/12/21 11:55 AM	Plot Creation	View Results
Completed	18	err194147: 1. EPP194147 unpaired	2/11/21 12:19 PM	2/12/21 8:32 AM	2/12/21 9:05 AM	Plot Creation	View Results
Error	17	ERR194147: 1. ERR194147 read1 , ERR194147 read2	2/11/21 8:53 AM	2/11/21 9:04 AM	2/11/21 9:17 AM	Alignment	View Details
Completed	16	Control7: 1.	6/24/20 8:04 AM	6/24/20 8:47 AM	6/24/20 8:54 AM	Plot Creation	View Results
Error	15	control6: 1.	6/23/20 11:53 AM	6/23/20 11:53 AM	6/23/20 11:53 AM	Alignment	View Details
Error	14	control5: 1. Job-9 trimmend paired file of HBR repl read1. Job-9 trimmend paired file of HBR repl read2	6/23/20 8:17 AM	6/23/20 8:17 AM	6/23/20 8:17 AM	Alignment	View Details

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 - DNA-Seq Pipeline Results

Job Status: **COMPLETED** Job ID: 19 Submission Date: Feb 12, 2021 11:00:52 AM Execution Time: 00 hours, 54 minutes, 13 seconds

Variant Annotation Reports | Initial FastQC Reports | Trimmed FASTQ Files | Trimmed FastQC Reports | Bowtie2 Files | Variant Calling | Variant Annotation | Next

Significant Gene Files with Genename Download

Significant Gene File File

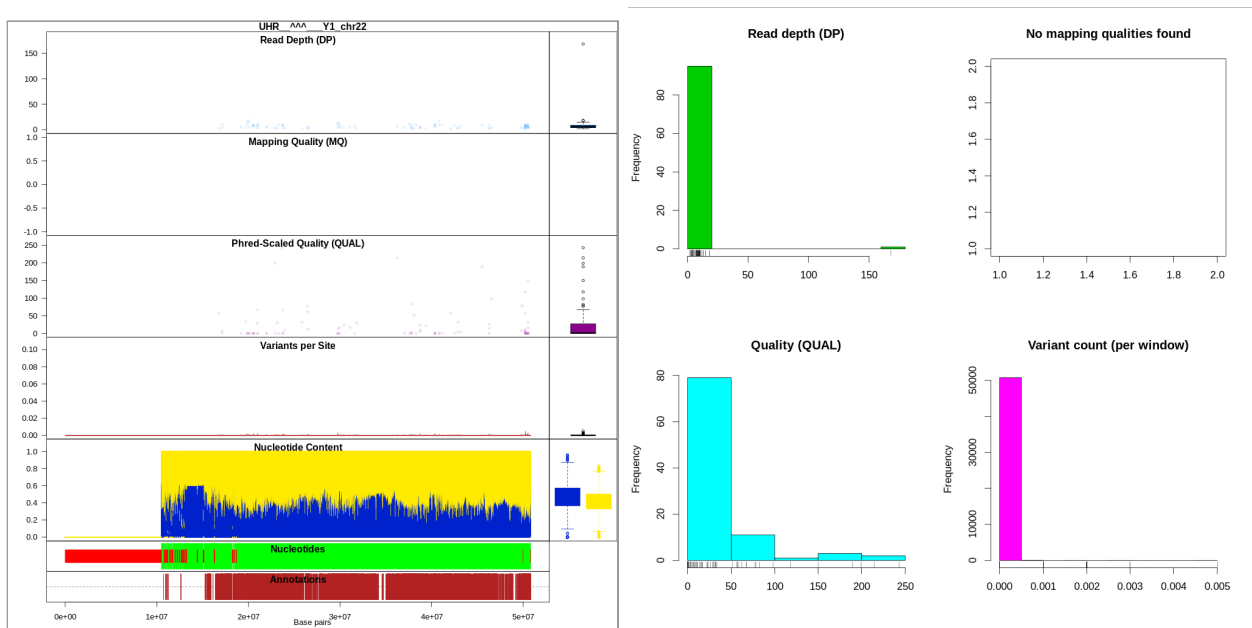
Plot Images Folders Download

Plot Image Folder Folder

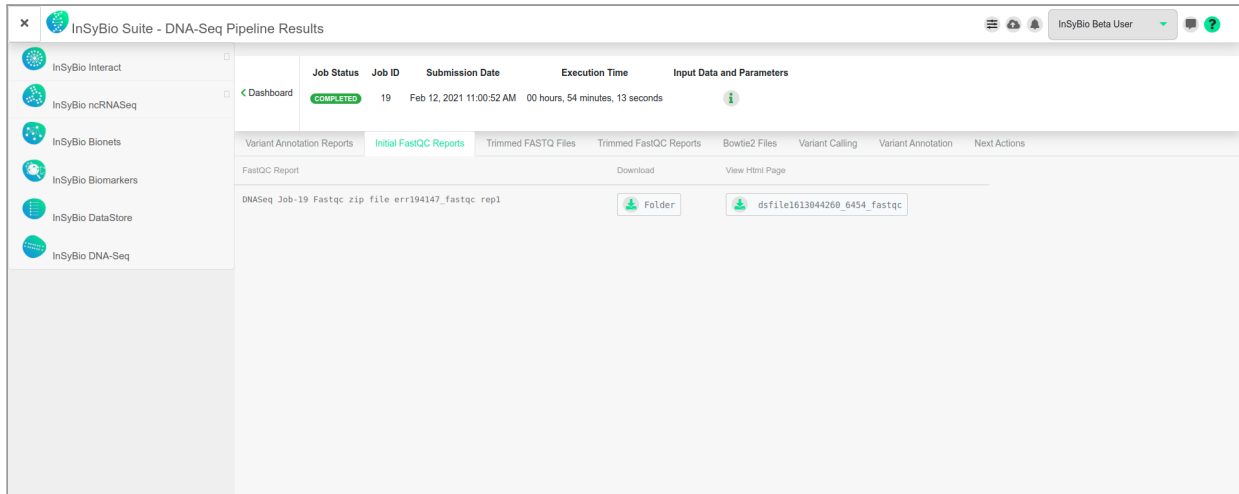
In the Variant Annotations reports tab you can download visual information and the Significant Gene Files with Genename notation, and some variant alignment images.

	A	B	C	D	E
1	Ensemble Gene id	Sift score	Associated Genename	Associated Uniprot IDs	
2	ENSG00000139055	0.0	ERP27	Q96DN0,F5GYS6	
3	ENSG00000198888	0.0	MT-ND1	P03886	
4	ENSG00000198840	0.0	MT-ND3	P03897	
5					
6					
7					
8					
9					
10					
11					

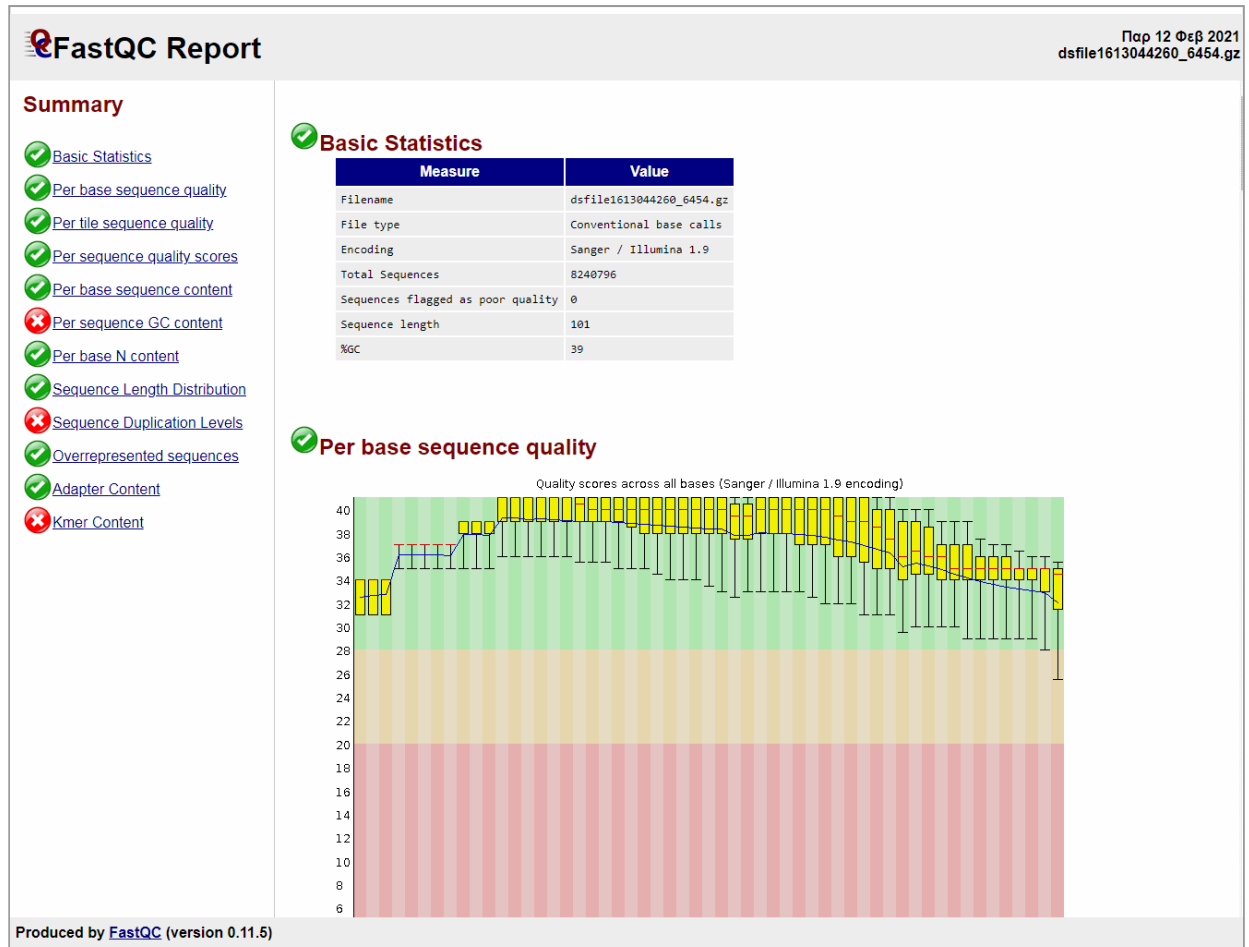
Example of the Significant Gene File being viewed with Microsoft Excel.



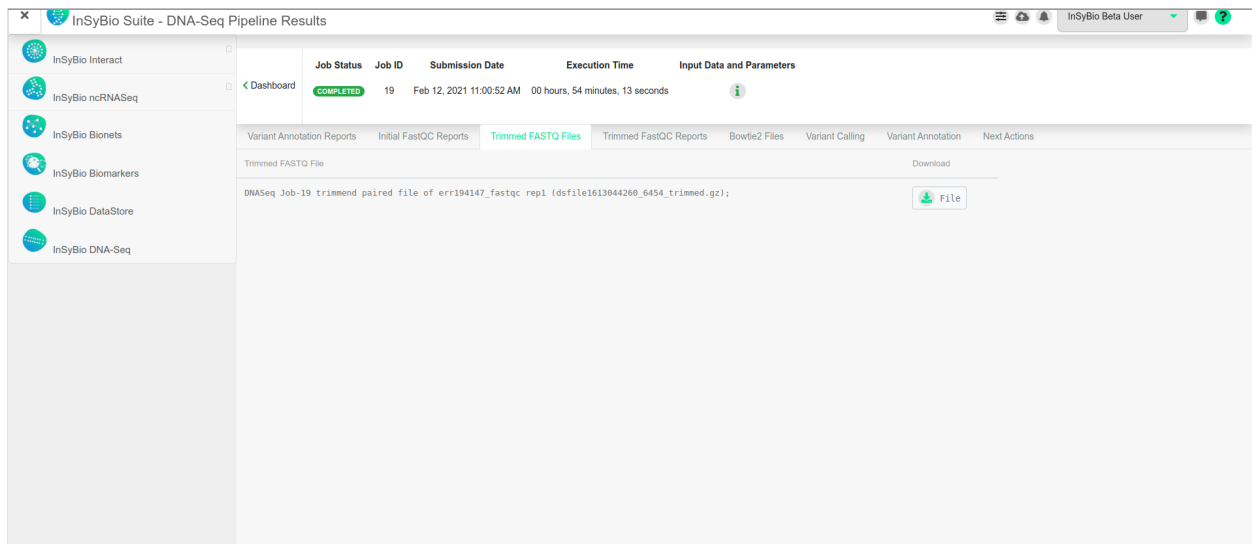
Example of the produced images and plots, (if there are enough data per chromosome).



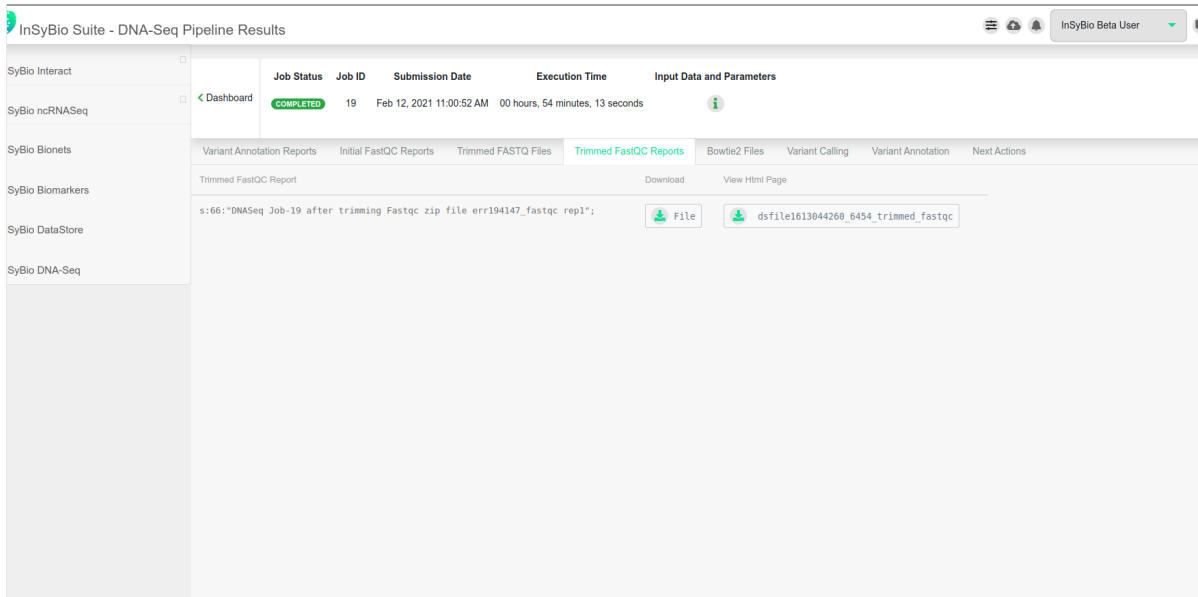
If Initial FastQC is selected, in the Initial FastQC reports the FastQC reports of the input files can be downloaded.



Example of a FastQC Report html file, one for each experiment is produced.



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



The screenshot displays the InSyBio Suite - DNA-Seq Pipeline Results interface. The top navigation bar includes the InSyBio logo, the title 'InSyBio Suite - DNA-Seq Pipeline Results', and a user profile 'InSyBio Beta User'. The main content area is divided into a left sidebar with navigation options (SyBio Interact, SyBio ncRNASeq, SyBio Bionets, SyBio Biomarkers, SyBio DataStore, SyBio DNA-Seq) and a main panel. The main panel shows a table of job results with columns for Job Status, Job ID, Submission Date, Execution Time, and Input Data and Parameters. A job with ID 19 is shown as 'COMPLETED' on Feb 12, 2021 at 11:00:52 AM, with an execution time of 00 hours, 54 minutes, and 13 seconds. Below the table, there are tabs for 'Variant Annotation Reports', 'Initial FastQC Reports', 'Trimmed FASTQ Files', 'Trimmed FastQC Reports', 'Bowtie2 Files', 'Variant Calling', 'Variant Annotation', and 'Next Actions'. The 'Trimmed FastQC Reports' tab is active, showing a 'Trimmed FastQC Report' with a 'Download' button and a file link: 'dsfile1613044260\_6454\_trimmed\_fastqc'.

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

q Pipeline Results

	Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
<a href="#">Dashboard</a>	COMPLETED	18	Feb 11, 2021 12:19:48 PM	00 hours, 32 minutes, 31 seconds	<a href="#">i</a>
<a href="#">Variant Annotation Reports</a> <a href="#">Bowtie2 Files</a> <a href="#">Variant Calling</a> <a href="#">Variant Annotation</a> <a href="#">Next Actions</a>					
SAM File					<a href="#">Download</a>
DNaseq Job-18 Bowtie2 alignment file err194147_1.sam (err194147_1.sam);					<a href="#">File</a>
BAM File					<a href="#">Download</a>
DNaseq Job-18 BAM fileerr194147_1.bam (err194147_1.bam);					<a href="#">File</a>
Run Info					<a href="#">Download</a>
Alignment Info					<a href="#">bowtie2_report.txt</a>

In the Bowtie2 files tab, the Bowtie2 alignment sam and bam files can be downloaded.

Example of Alignment information inside the bowtie2\_report.txt:

8131633 reads; of these:

8131633 (100.00%) were unpaired; of these:

34333 (0.42%) aligned 0 times

4183088 (51.44%) aligned exactly 1 time

3914212 (48.14%) aligned >1 times

99.58% overall alignment rate

The screenshot shows the 'Bio Suite - DNA-Seq Pipeline Results' interface. The top navigation bar includes a hamburger menu, a user profile icon, and the text 'InSyBio Beta User'. Below this is a table with columns: Job Status (COMPLETED), Job ID (19), Submission Date (Feb 12, 2021 11:00:52 AM), Execution Time (00 hours, 54 minutes, 13 seconds), and Input Data and Parameters. A left sidebar contains a list of pipeline steps: Interact, DNaseSeq, Bowtie2, Variant Calling (selected), Variant Annotation, and Next Actions. The main content area shows 'Variant Call Files' with a 'Download' button and a 'File' download icon. The text below the download button reads: 'DNaseSeq Job-19 Variant Annotation file (err194147\_fastqc\_1.vcf);'.

In the Variant Calling tab the unfiltered VCF file is provided as created by Freebayes and is available to be downloaded.

The screenshot shows the 'Bio Suite - DNA-Seq Pipeline Results' interface with the 'Variant Annotation' tab selected. The top navigation bar and job status table are identical to the previous screenshot. The left sidebar highlights 'Variant Annotation'. The main content area displays 'Missense Variant Vep Files' with a 'Download' button and a 'File' download icon. Below this, the text reads: 'DNaseSeq Job-19 Filtered missense\_variants Variant Annotation file (err194147\_fastqc\_1\_missense\_annotations.vcf);'. Further down, 'Protein Altering Variants' are listed with another 'Download' button and 'File' icon, with the text: 'DNaseSeq Job-19 Filtered protein\_altering\_variants and AF < 0.05 Variant Annotation file (err194147\_fastqc\_1\_filtered\_annotations.vcf);'. At the bottom, 'All Variants' are listed with a 'Download' button and 'File' icon, with the text: 'DNaseSeq Job-19 Variant Annotation file (err194147\_fastqc\_1\_annotations.vcf);'.



In the Variant Annotation tab the different Annotated Variant vcf files for each sample can be downloaded., Missense Variant Vep files, Protein Altering Variants and All Variants are available.

The screenshot displays the InSyBio Suite interface for DNA-Seq Pipeline Results. The top navigation bar includes the InSyBio logo, the title 'InSyBio Suite - DNA-Seq Pipeline Results', and a user profile 'InSyBio Beta User'. A sidebar on the left lists various analysis modules. The main content area features a table of pipeline jobs and a set of tabs for different analysis stages. The 'Next Actions' tab is currently selected, showing options for downloading significant gene files and performing further actions.

Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	19	Feb 12, 2021 11:00:52 AM	00 hours, 54 minutes, 13 seconds	

In the Next Action tab, Significant Genes files, with the provided threshold (default 10%) most significant genes, for each cohort are provided. They can be downloaded or used as input in **InSyBio Interact**, to **Create Networks** from that set of significant genes based on the protein-protein interactions knowledge base of InSyBio Interact, or to perform GO Term **Enrichment Analysis** from that set of biomarkers based on the protein-go term correlation knowledge base of InSyBio Interact..

## How to get InSyBio DNaseq

---

To request a free one month license of InSyBio Suite please email us at [info@insybio.com](mailto:info@insybio.com).

To purchase InSyBio DNaseq commercial version 2.6 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 2019 InSyBio Ltd. Reproduction without written permission is completely forbidden.