InSyBio
Intelligent Systems Biology

Analyze DNA Sequencing data with InSyBio DNASeq

User Manual

March 2022

Insybio Suite v3.0

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.

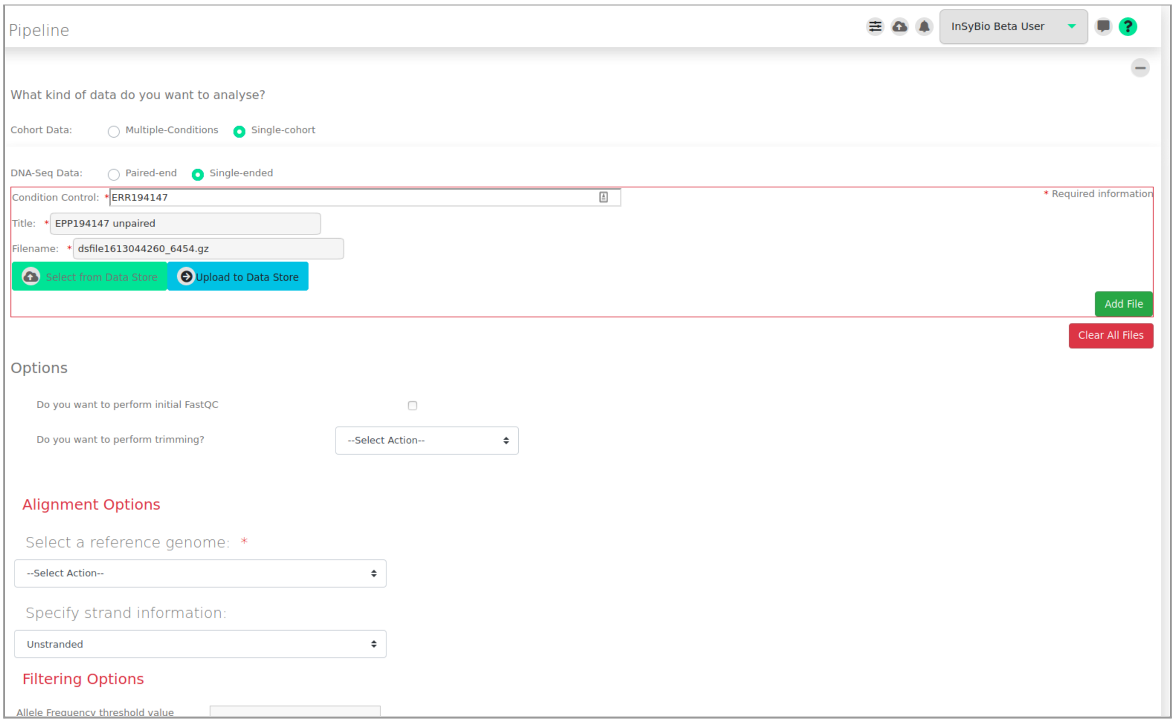
Firstly, the Pipeline uses Fastqc to create a report with the input sequences quality, then trimm the sequences accordingly using Trimmomatic and create new reports with Fastqc. Then using Bowtie2 it creates the alignment SAM files with the Genome files, we sort them using SAMtools and transform them 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

Select from Data Store | Upload to Data Store

Add File

Clear All Files

Options

Do you want to perform initial FastQC □

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

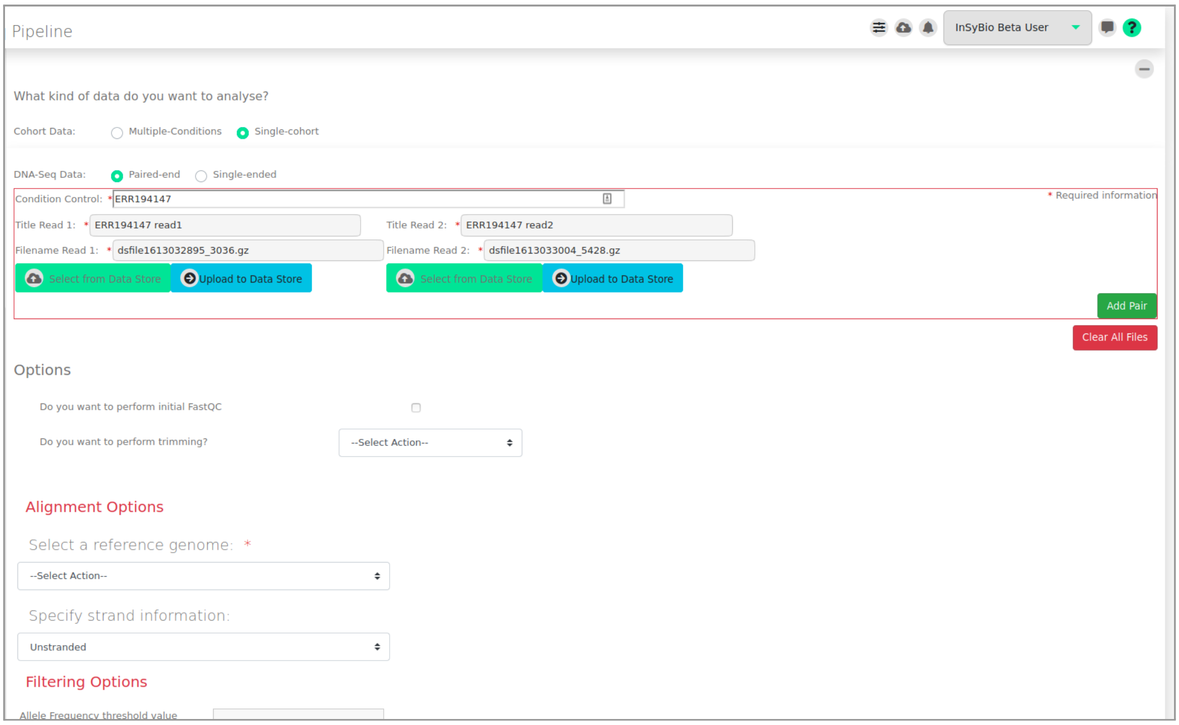
Alignment Options

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

Select from Data Store | Upload to Data Store | Select from Data Store | Upload to Data Store

Add Pair

Clear All Files

Options

Do you want to perform initial FastQC □

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

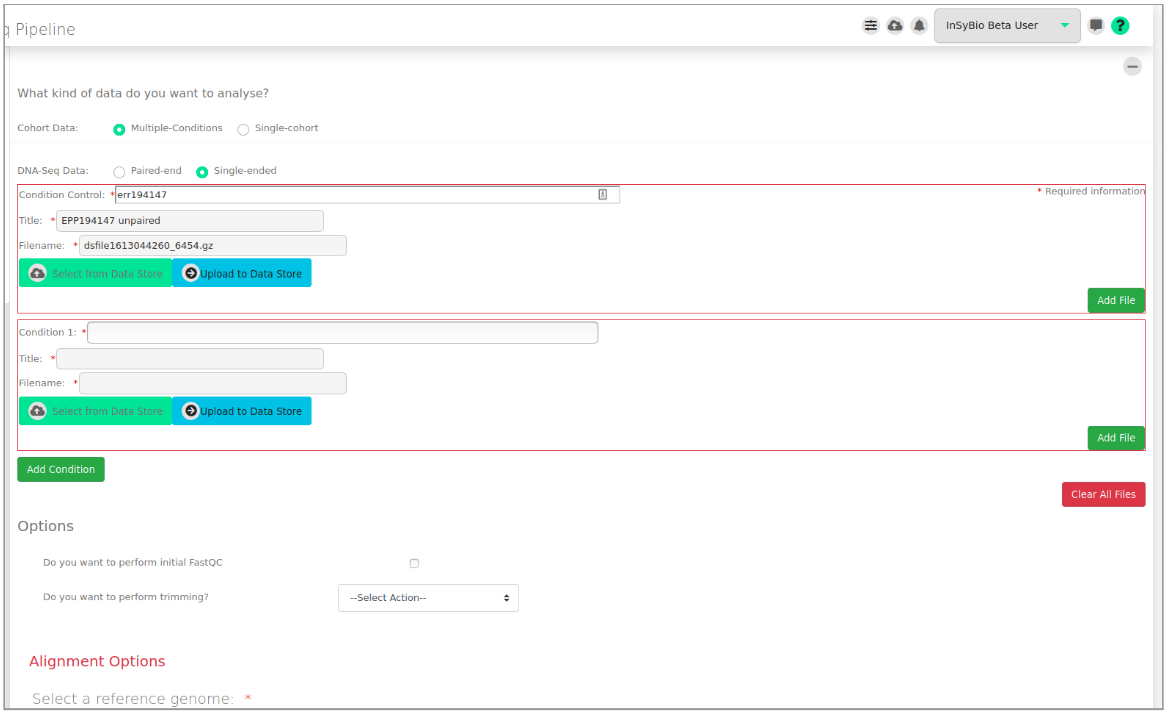
Alignment Options

Select a reference genome: * --Select Action--

Specify strand information: Unstranded

Filtering Options

Allele Frequency threshold value



What kind of data do you want to analyse?

Cohort Data: ● Multiple-Conditions ○ Single-cohort

DNA-Seq Data: ○ Paired-end ● Single-ended

* Required information

Condition Control: * err194147

Title: * EPP194147 unpaired

Filename: * dsfile1613044260_6454.gz

Condition 1: *

Title: *

Filename: *

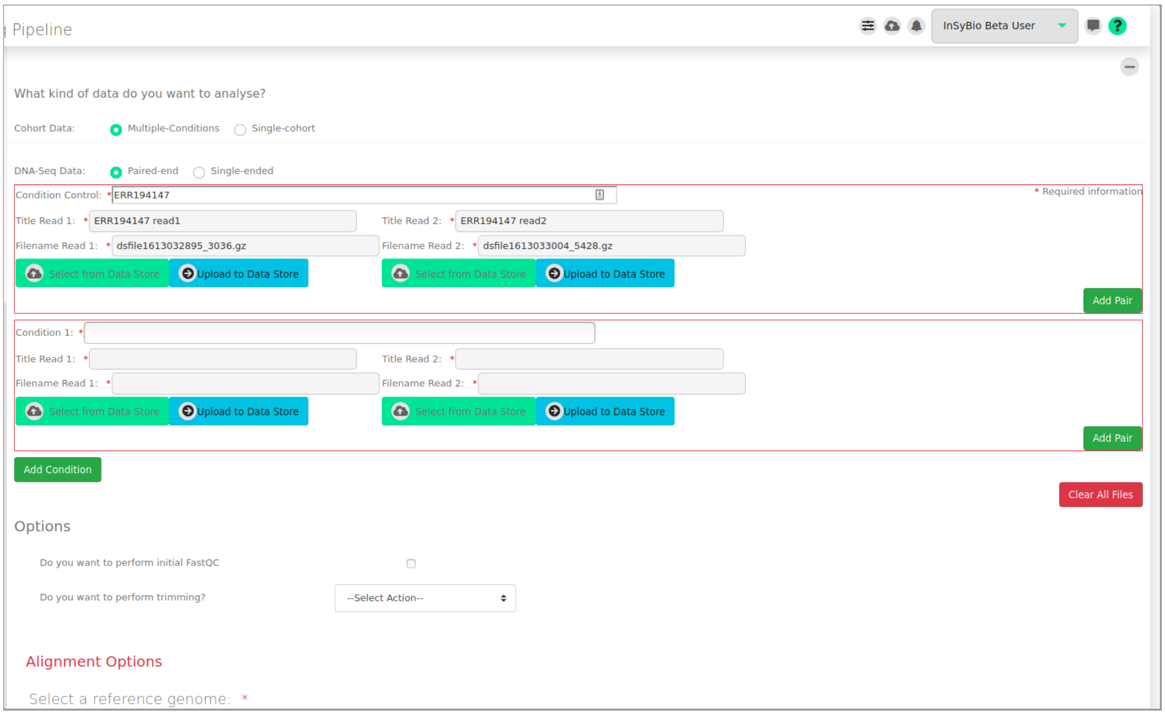
Options

Do you want to perform initial FastQC □

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

Alignment Options

Select a reference genome: *



What kind of data do you want to analyse?

Cohort Data: ● Multiple-Conditions ○ Single-cohort

DNA-Seq Data: ● Paired-end ○ Single-ended

* Required information

Condition Control: * ERR194147

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

Filename Read 1: * dsfile1613032895_3036.gz Filename Read 2: * dsfile1613033004_5428.gz

Condition 1: *

Title Read 1: * Title Read 2: *

Filename Read 1: * Filename Read 2: *

Options

Do you want to perform initial FastQC □

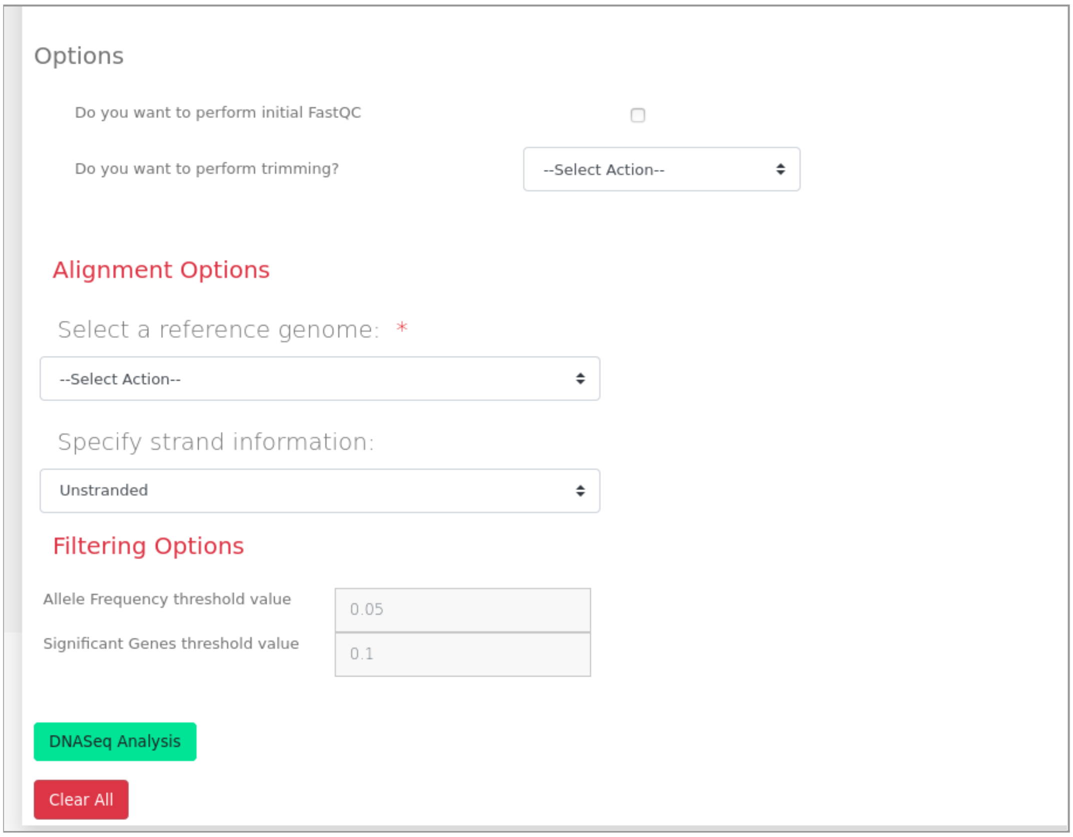
Do you want to perform trimming? --Select Action--

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

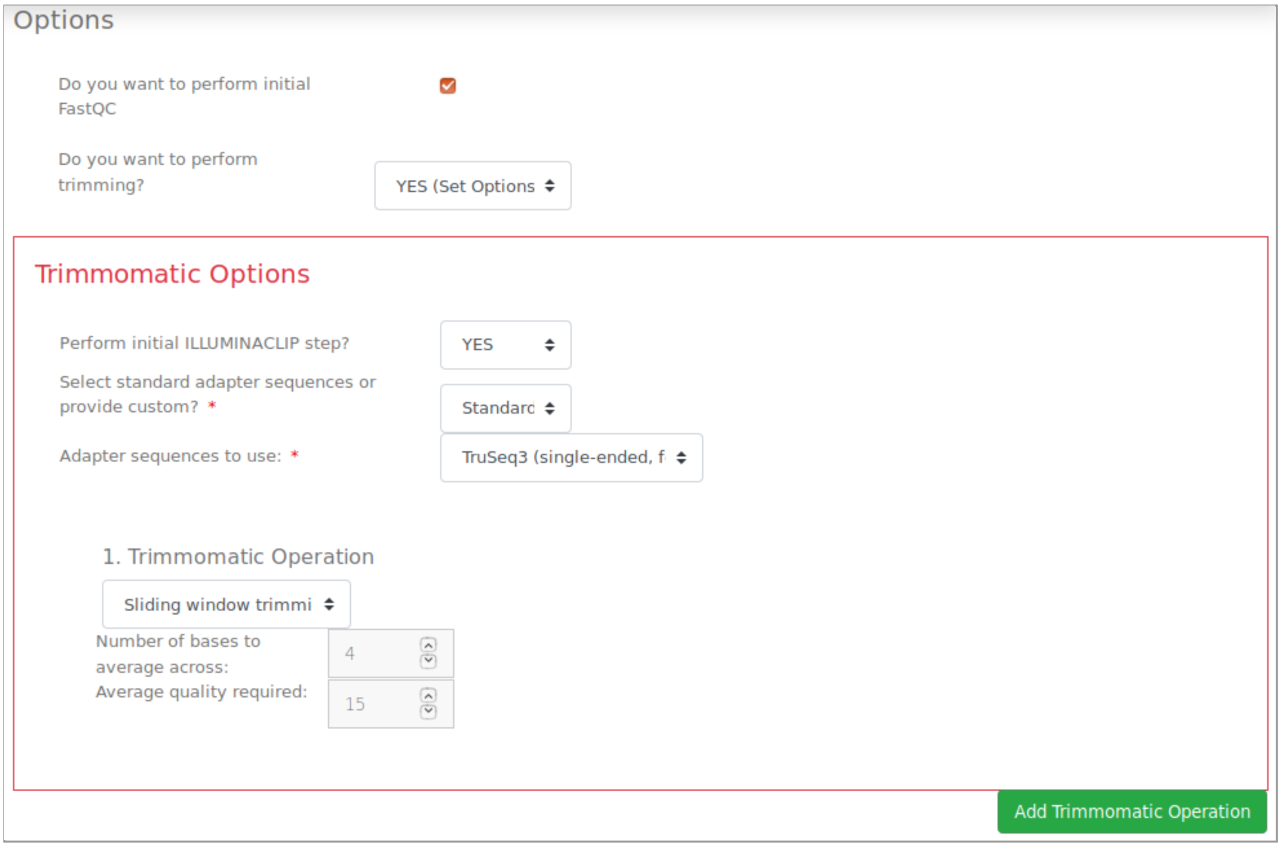
Filtering Options

Allele Frequency threshold value 0.05

Significant Genes threshold value 0.1

- 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? * Standarc

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

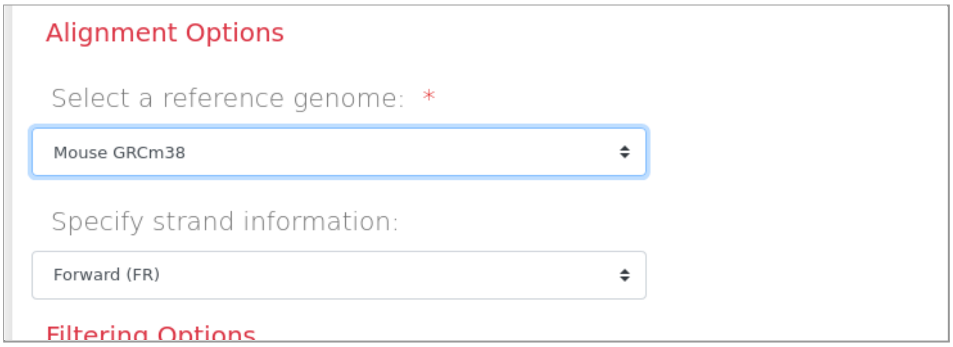
1. Trimmomatic Operation

Sliding window trimmi

Number of bases to average across: 4

Average quality required: 15

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



Alignment Options

Select a reference genome: *

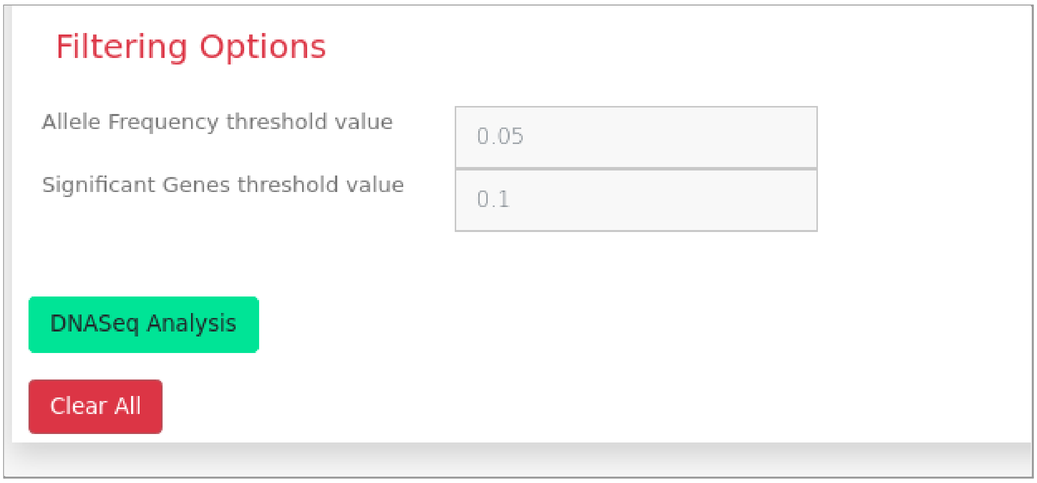
Mouse GRCm38

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.



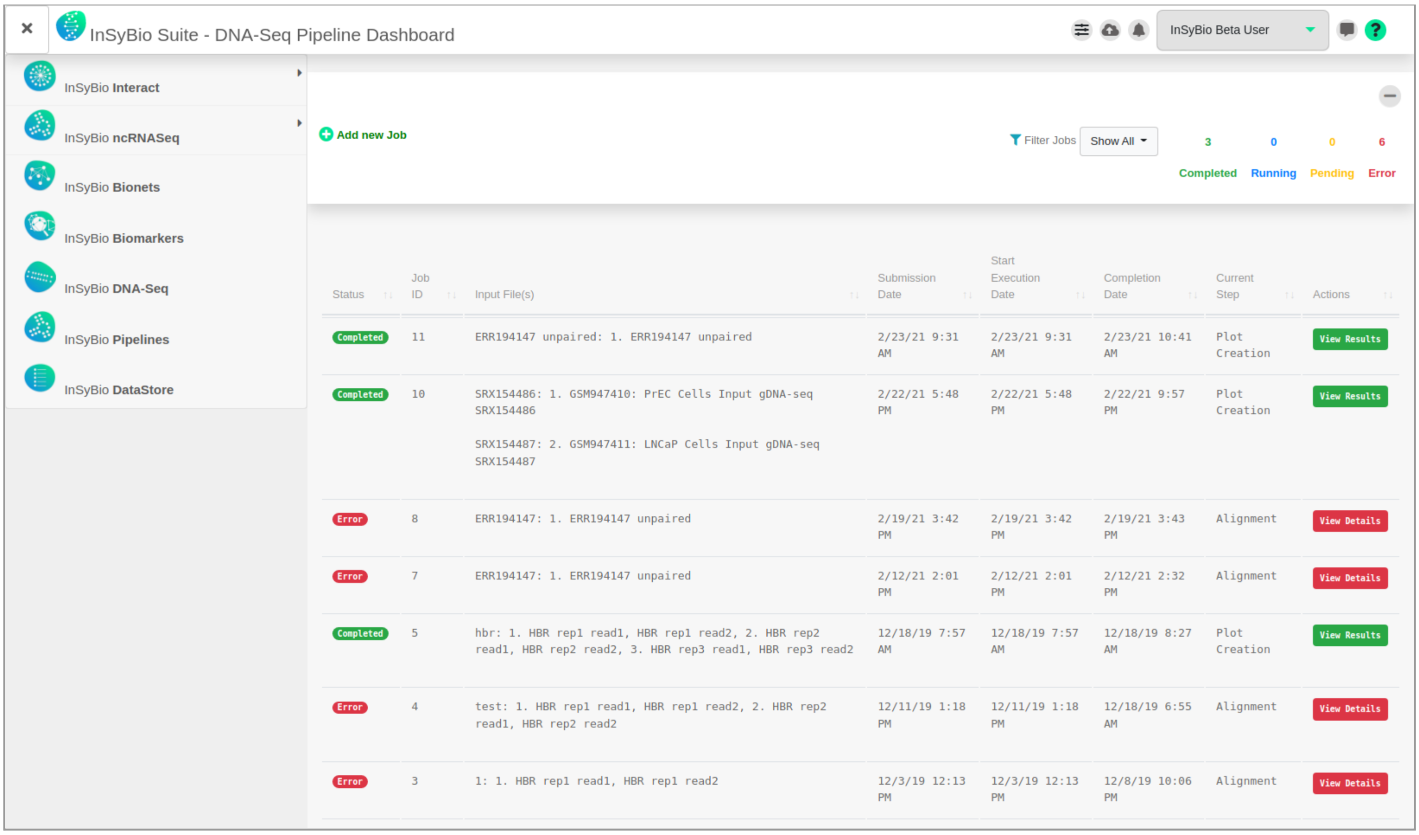
Filtering Options

Allele Frequency threshold value 0.05

Significant Genes threshold value 0.1

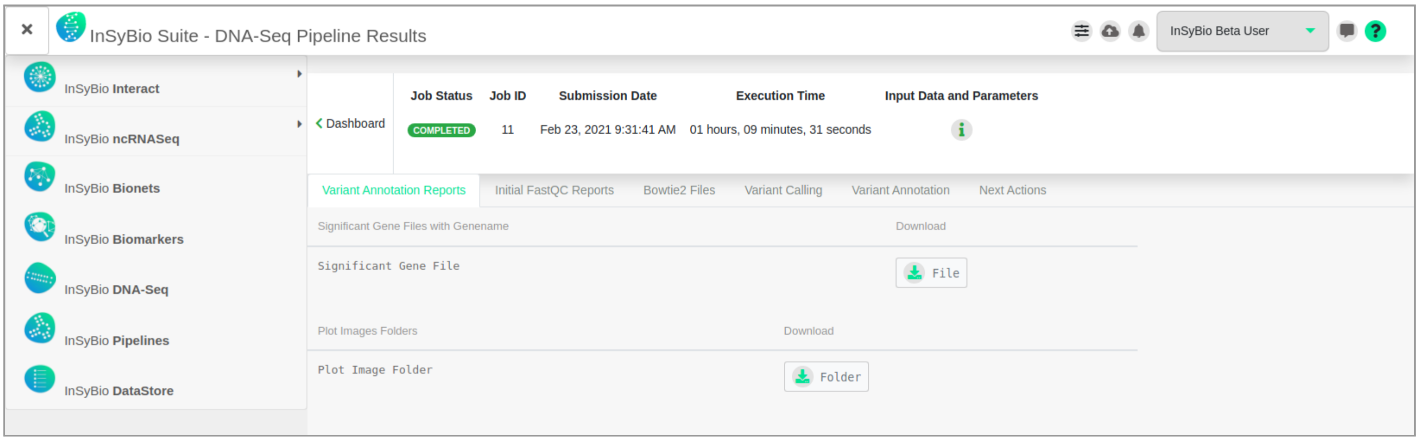
To view the results:

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	11	ERR194147 unpaired: 1. ERR194147 unpaired	2/23/21 9:31 AM	2/23/21 9:31 AM	2/23/21 10:41 AM	Plot Creation	View Results
Completed	10	SRX154486: 1. GSM947410: PrEC Cells Input gDNA-seq SRX154486 SRX154487: 2. GSM947411: LNCaP Cells Input gDNA-seq SRX154487	2/22/21 5:48 PM	2/22/21 5:48 PM	2/22/21 9:57 PM	Plot Creation	View Results
Error	8	ERR194147: 1. ERR194147 unpaired	2/19/21 3:42 PM	2/19/21 3:42 PM	2/19/21 3:43 PM	Alignment	View Details
Error	7	ERR194147: 1. ERR194147 unpaired	2/12/21 2:01 PM	2/12/21 2:01 PM	2/12/21 2:32 PM	Alignment	View Details
Completed	5	hbr: 1. HBR rep1 read1, HBR rep1 read2, 2. HBR rep2 read1, HBR rep2 read2, 3. HBR rep3 read1, HBR rep3 read2	12/18/19 7:57 AM	12/18/19 7:57 AM	12/18/19 8:27 AM	Plot Creation	View Results
Error	4	test: 1. HBR rep1 read1, HBR rep1 read2, 2. HBR rep2 read1, HBR rep2 read2	12/11/19 1:18 PM	12/11/19 1:18 PM	12/18/19 6:55 AM	Alignment	View Details
Error	3	1: 1. HBR rep1 read1, HBR rep1 read2	12/3/19 12:13 PM	12/3/19 12:13 PM	12/8/19 10:06 PM	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.

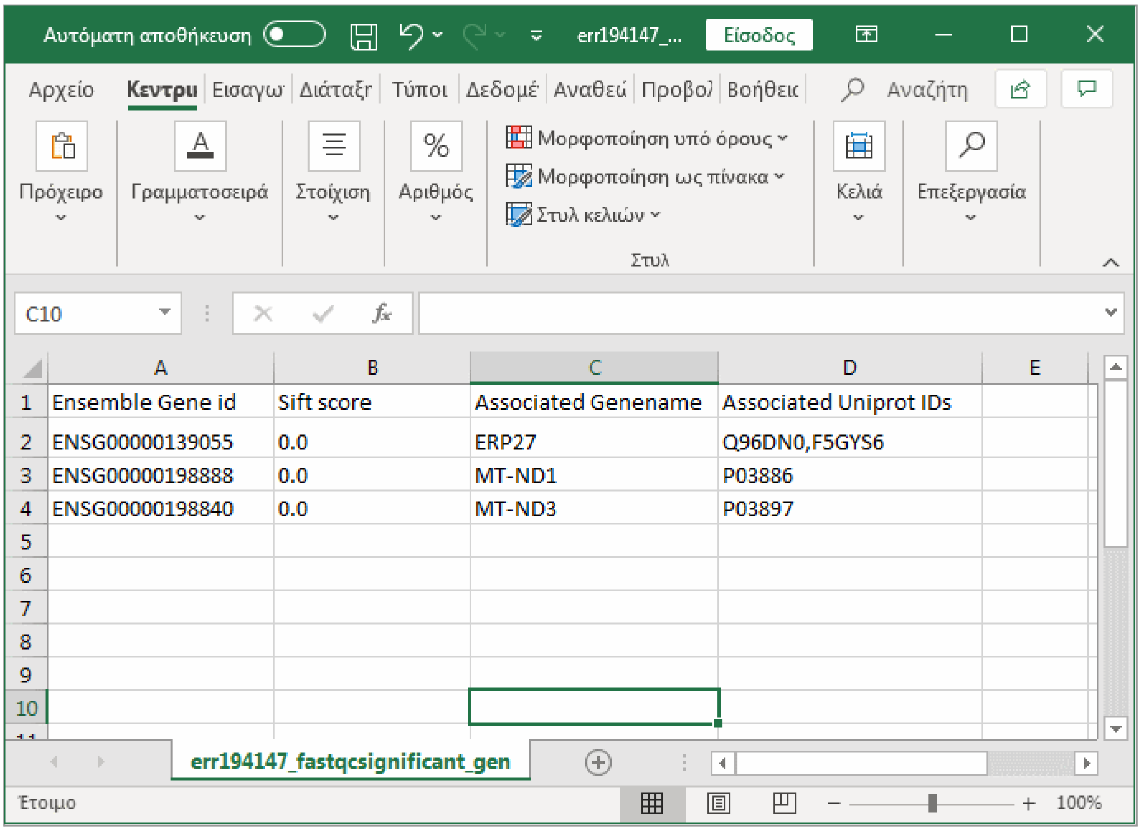


Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	11	Feb 23, 2021 9:31:41 AM	01 hours, 09 minutes, 31 seconds	i

Significant Gene Files with Genename	Download
Significant Gene File	File

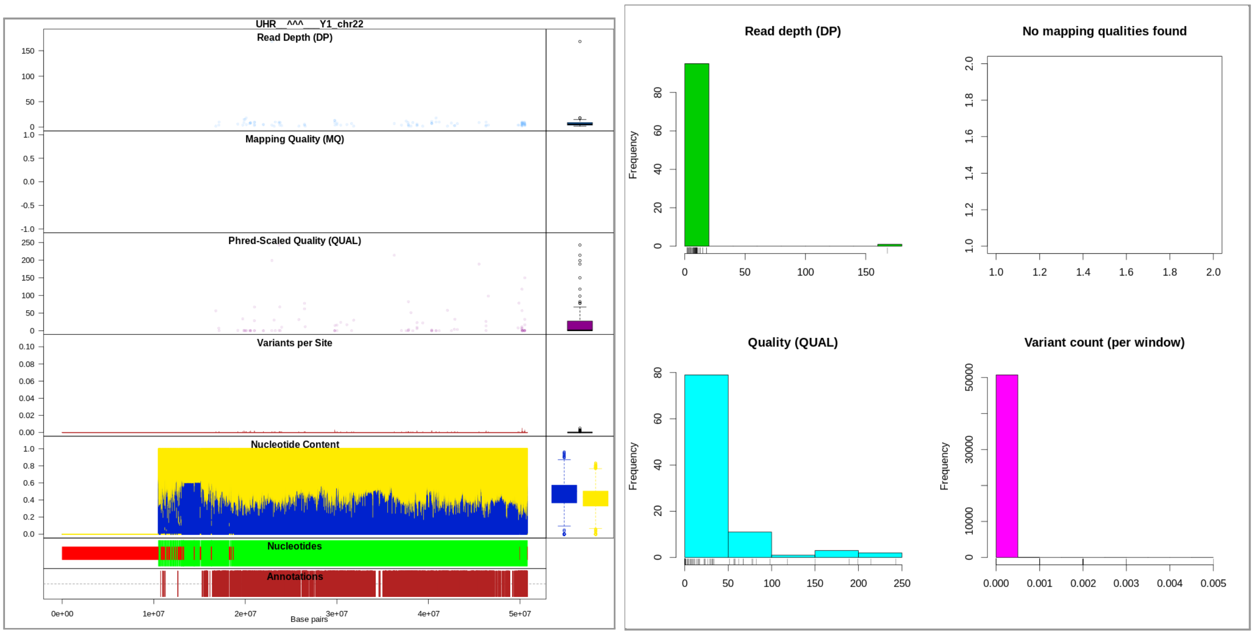
Plot Images Folders	Download
Plot Image Folder	Folder

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

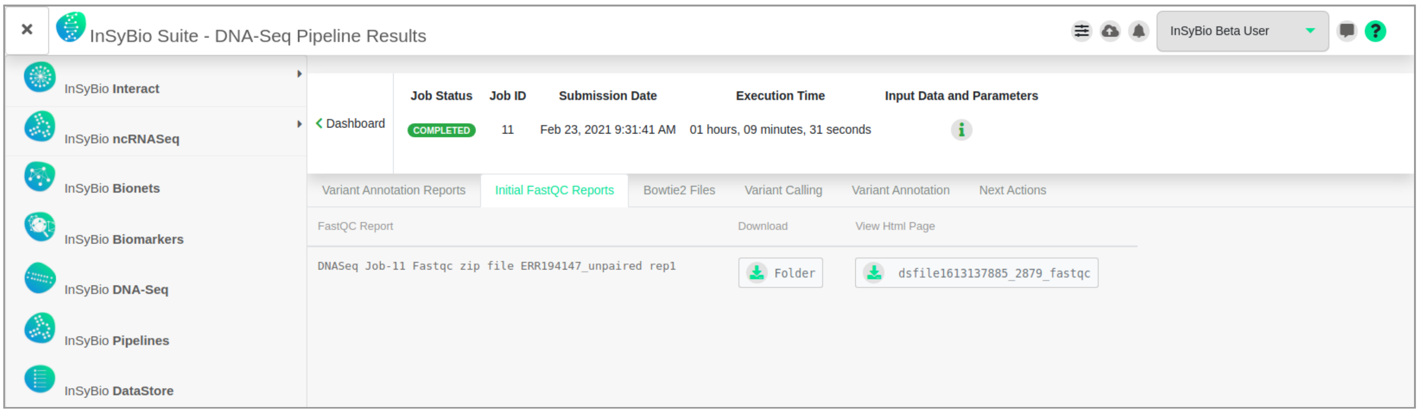


Ensemble Gene id	Sift score	Associated Genename	Associated Uniprot IDs
ENSG00000139055	0.0	ERP27	Q96DN0,F5GYS6
ENSG00000198888	0.0	MT-ND1	P03886
ENSG00000198840	0.0	MT-ND3	P03897

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



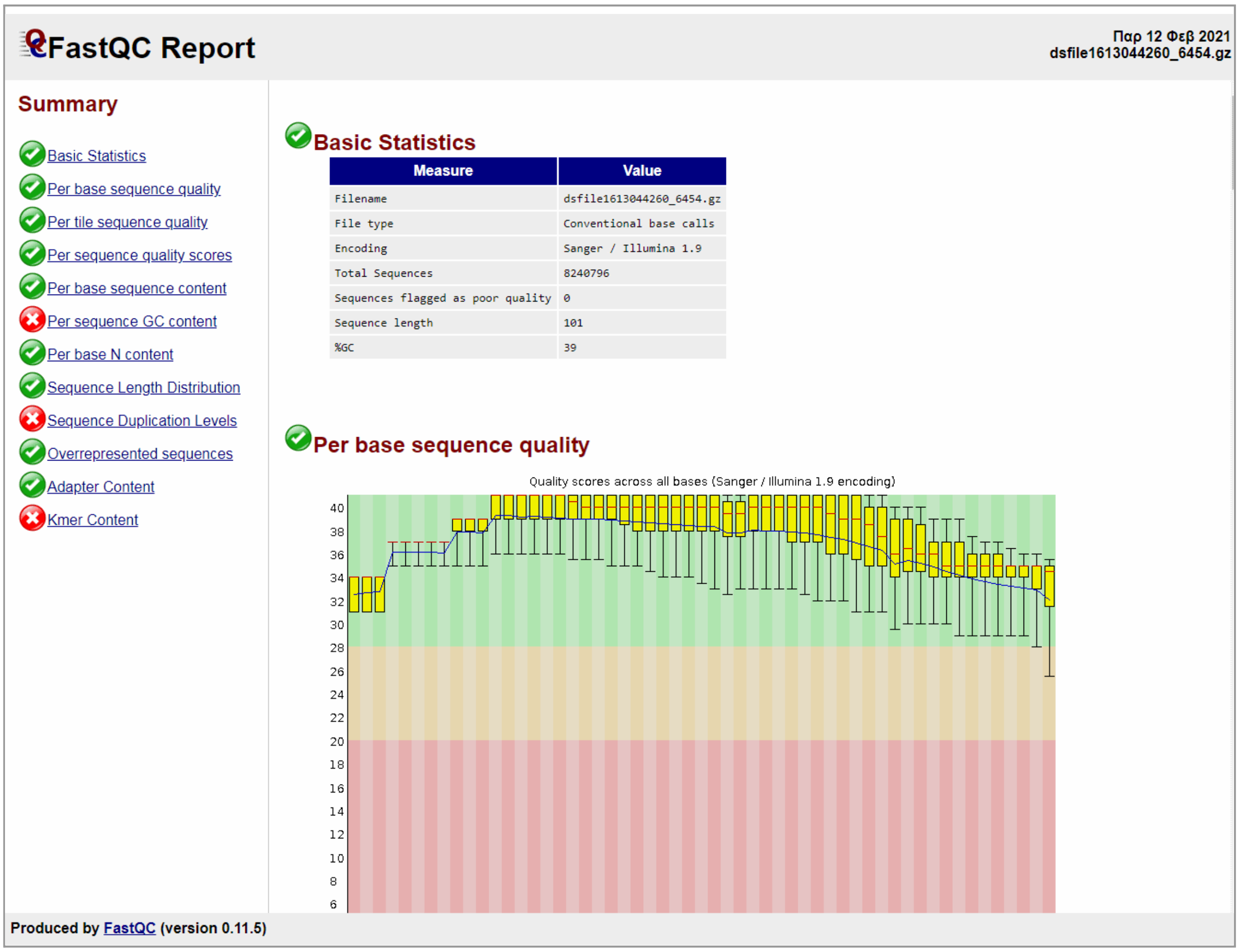
Example of the produced images and plots, (if there are enough data per chromosome).



Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	11	Feb 23, 2021 9:31:41 AM	01 hours, 09 minutes, 31 seconds	i

FastQC Report	Download	View Html Page
DNASeq Job-11 Fastqc zip file ERR194147_unpaired rep1	Folder	dsfile1613137885_2879_fastqc

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



FastQC Report

Παρ 12 Φεβ 2021
dsfile1613044260_6454.gz

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

✓ Basic Statistics

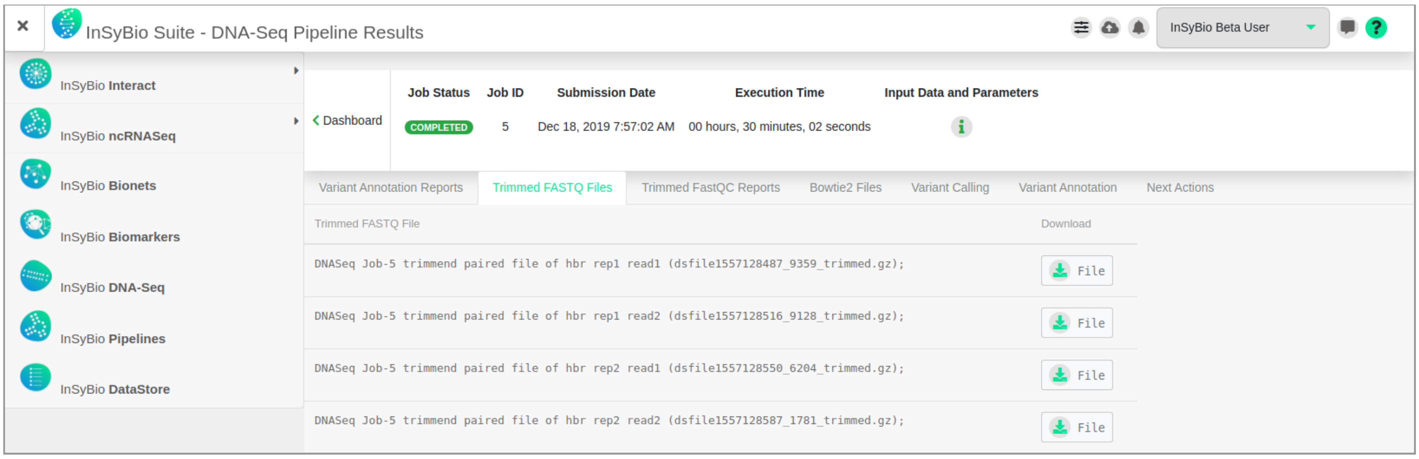
Measure	Value
Filename	dsfile1613044260_6454.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8240796
Sequences flagged as poor quality	0
Sequence length	101
%GC	39

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6

Produced by FastQC (version 0.11.5)

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

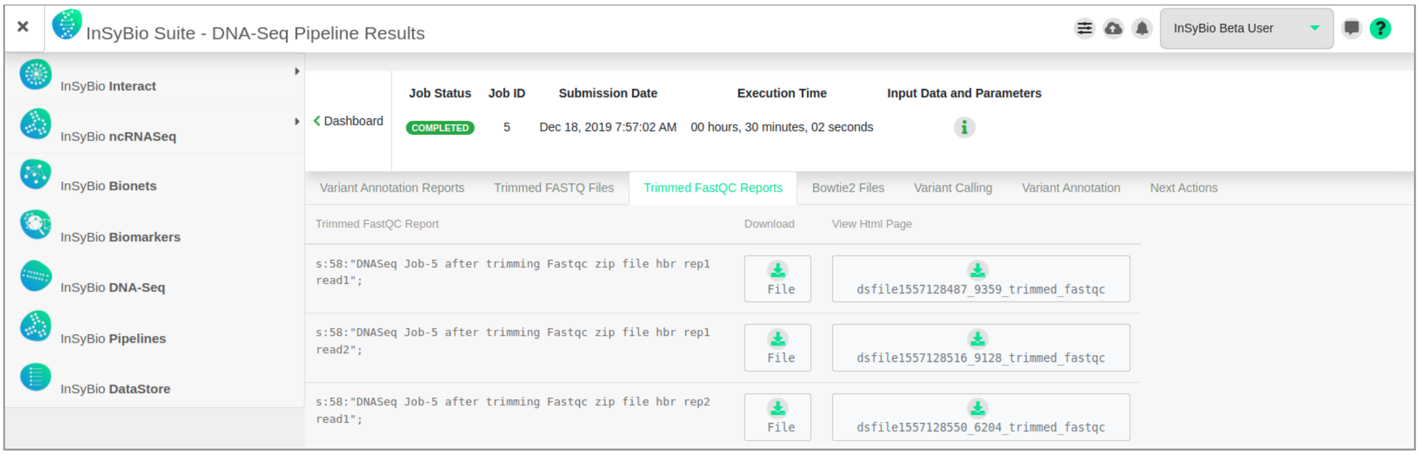


InSyBio Suite - DNA-Seq Pipeline Results

Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	5	Dec 18, 2019 7:57:02 AM	00 hours, 30 minutes, 02 seconds	ℹ

Trimmed FASTQ File	Download
DNASeq Job-5 trimmend paired file of hbr rep1 read1 (dsfile1557128487_9359_trimmed.gz);	File
DNASeq Job-5 trimmend paired file of hbr rep1 read2 (dsfile1557128516_9128_trimmed.gz);	File
DNASeq Job-5 trimmend paired file of hbr rep2 read1 (dsfile1557128550_6204_trimmed.gz);	File
DNASeq Job-5 trimmend paired file of hbr rep2 read2 (dsfile1557128587_1781_trimmed.gz);	File

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

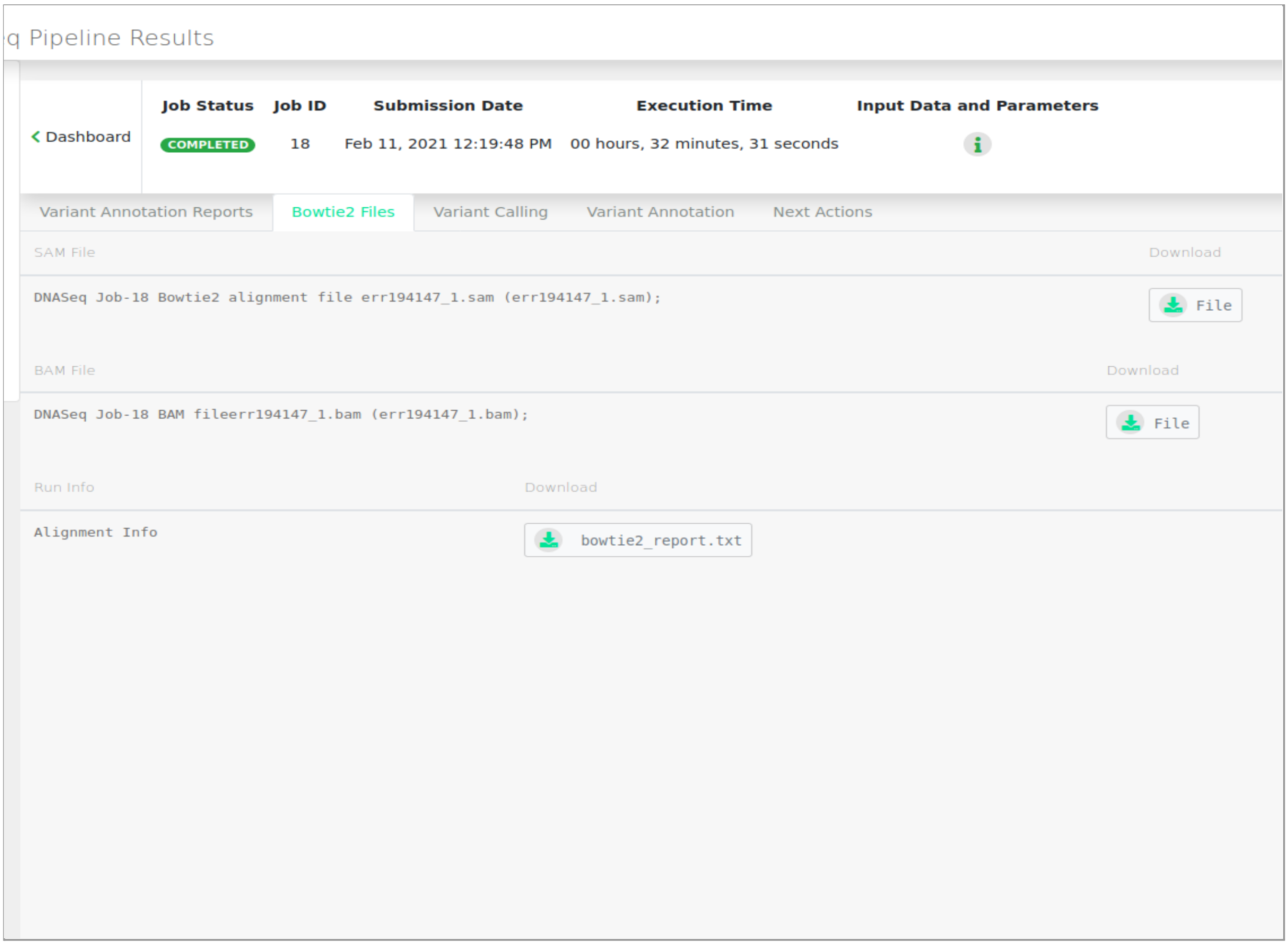


InSyBio Suite - DNA-Seq Pipeline Results

Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	5	Dec 18, 2019 7:57:02 AM	00 hours, 30 minutes, 02 seconds	ℹ

Trimmed FastQC Report	Download	View Html Page
s:58:"DNASeq Job-5 after trimming Fastqc zip file hbr rep1 read1";	File	dsfile1557128487_9359_trimmed_fastqc
s:58:"DNASeq Job-5 after trimming Fastqc zip file hbr rep1 read2";	File	dsfile1557128516_9128_trimmed_fastqc
s:58:"DNASeq Job-5 after trimming Fastqc zip file hbr rep2 read1";	File	dsfile1557128550_6204_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
COMPLETED	18	Feb 11, 2021 12:19:48 PM	00 hours, 32 minutes, 31 seconds	ℹ

SAM File	Download
DNASeq Job-18 Bowtie2 alignment file err194147_1.sam (err194147_1.sam);	File

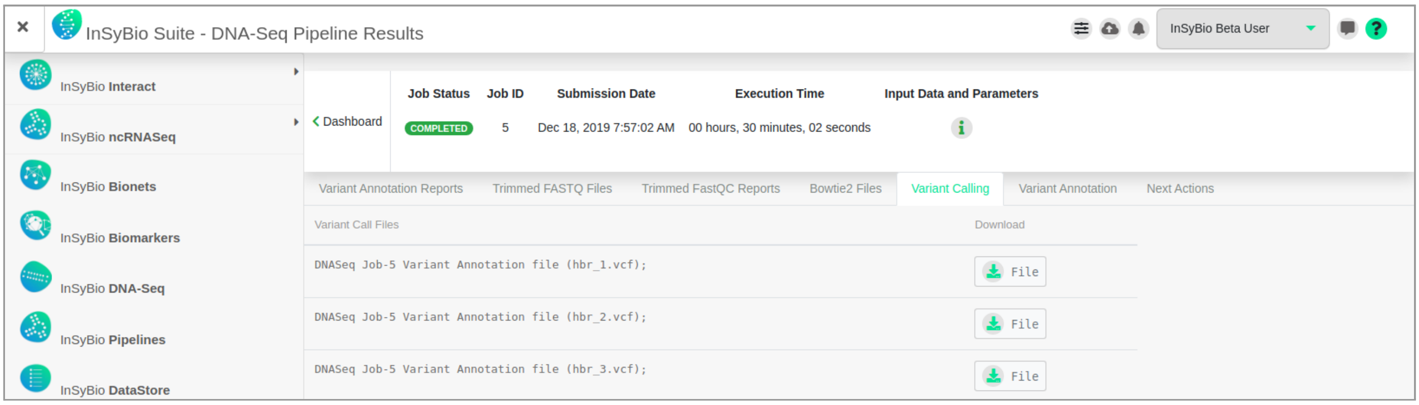
BAM File	Download
DNASeq Job-18 BAM fileerr194147_1.bam (err194147_1.bam);	File

Run Info	Download
Alignment Info	bowtie2_report.txt

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

Example of Alignment information inside the bowtie2_report.txt:

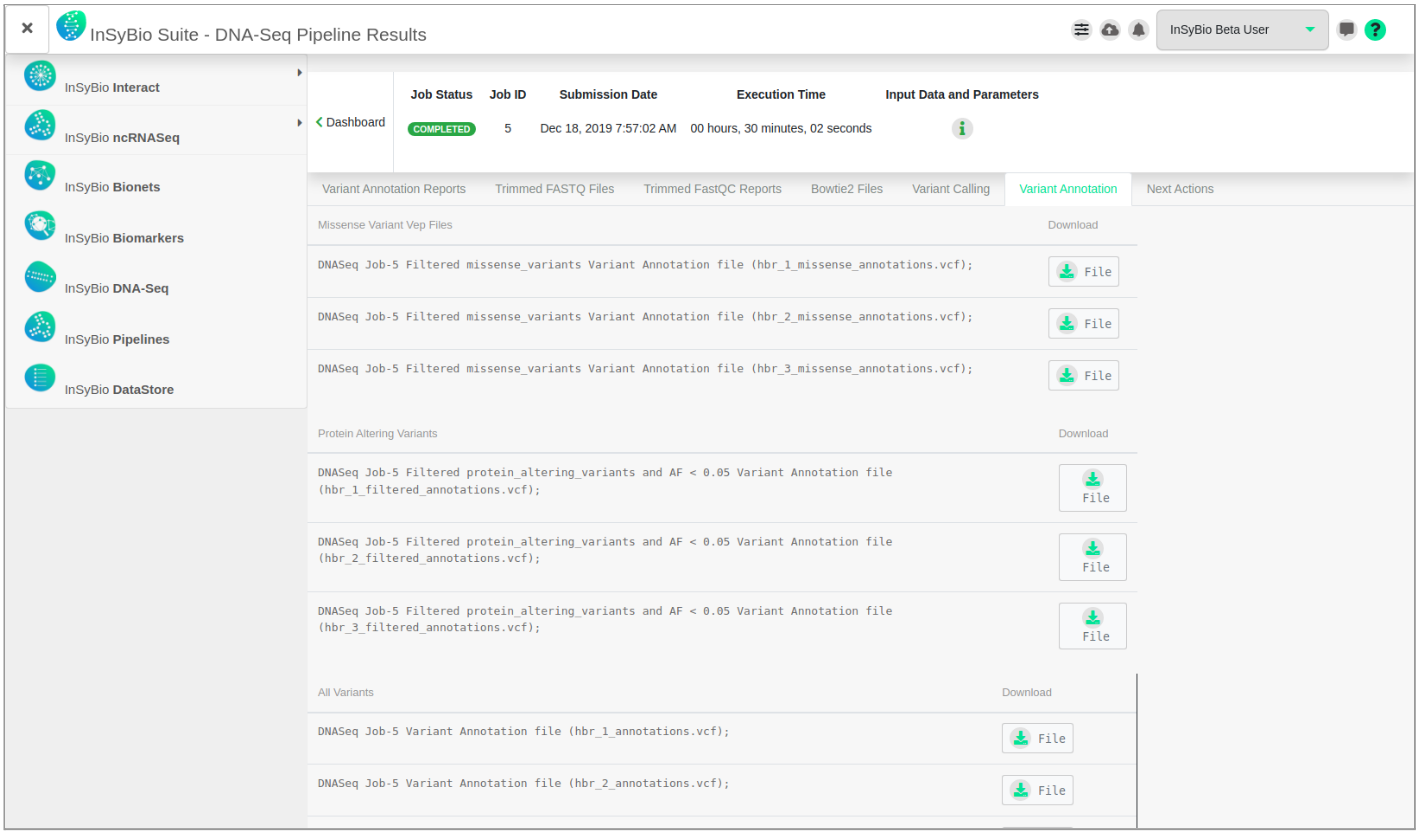
```
8131633 reads; of these:
      8131633 (100.00%) were unpaired; of these:
      34333 (0.42%) aligned 0 times
      4183088 (51.44%) aligned exactly 1 time
       3914212 (48.14%) aligned >1 times
99.58% overall alignment rate
```



Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	5	Dec 18, 2019 7:57:02 AM	00 hours, 30 minutes, 02 seconds	ℹ

Variant Call Files	Download
DNASeq Job-5 Variant Annotation file (hbr_1.vcf);	File
DNASeq Job-5 Variant Annotation file (hbr_2.vcf);	File
DNASeq Job-5 Variant Annotation file (hbr_3.vcf);	File

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



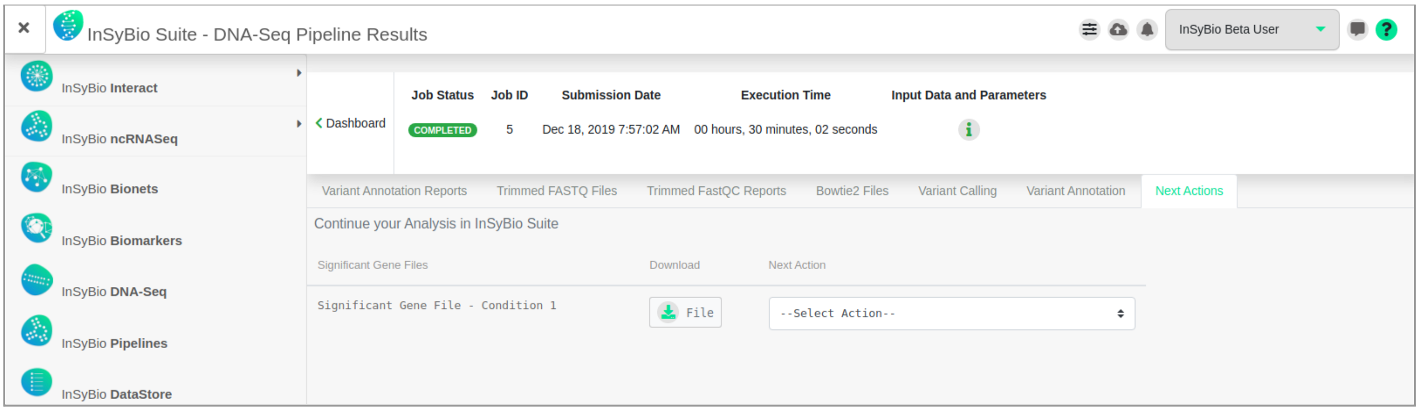
Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	5	Dec 18, 2019 7:57:02 AM	00 hours, 30 minutes, 02 seconds	ℹ

Missense Variant Vep Files	Download
DNASeq Job-5 Filtered missense_variants Variant Annotation file (hbr_1_missense_annotations.vcf);	File
DNASeq Job-5 Filtered missense_variants Variant Annotation file (hbr_2_missense_annotations.vcf);	File
DNASeq Job-5 Filtered missense_variants Variant Annotation file (hbr_3_missense_annotations.vcf);	File

Protein Altering Variants	Download
DNASeq Job-5 Filtered protein_altering_variants and AF < 0.05 Variant Annotation file (hbr_1_filtered_annotations.vcf);	File
DNASeq Job-5 Filtered protein_altering_variants and AF < 0.05 Variant Annotation file (hbr_2_filtered_annotations.vcf);	File
DNASeq Job-5 Filtered protein_altering_variants and AF < 0.05 Variant Annotation file (hbr_3_filtered_annotations.vcf);	File

All Variants	Download
DNASeq Job-5 Variant Annotation file (hbr_1_annotations.vcf);	File
DNASeq Job-5 Variant Annotation file (hbr_2_annotations.vcf);	File

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.



Job Status	Job ID	Submission Date	Execution Time	Input Data and Parameters
COMPLETED	5	Dec 18, 2019 7:57:02 AM	00 hours, 30 minutes, 02 seconds	ℹ

Continue your Analysis in InSyBio Suite

Significant Gene Files	Download	Next Action
Significant Gene File - Condition 1	File	--Select Action--

In the Next Action tab, Significant Genes files, with the provided threshold (default 10%) the 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 .

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

About Us

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

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

COPYRIGHT NOTICE

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.