Analyze DNA Sequencing data with InSyBio DNASeq

InSyBio
Intelligent Systems Biology

User Manual

October 2025

Insybio Suite v3.4

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.

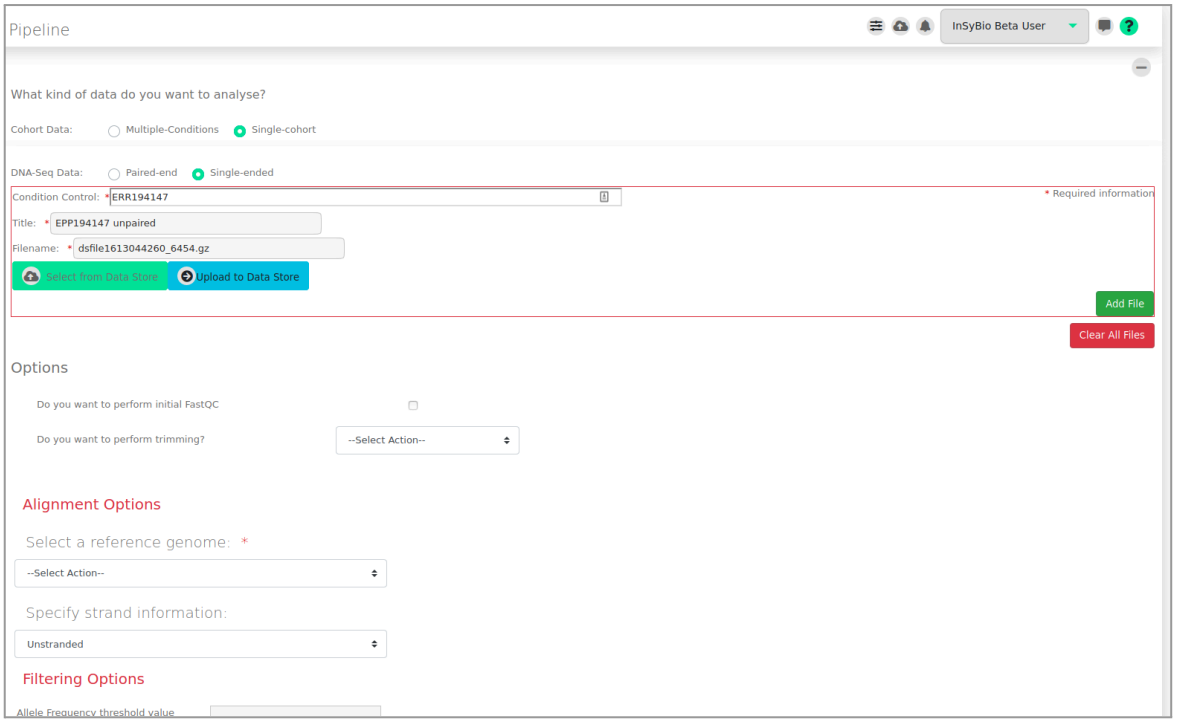
Firstly, the Pipeline uses Fastqc to create a report with the input sequences quality, then trimm the sequences accordingly using Trimmomatic and create new reports with Fastqc. Then using Bowtie2 it creates the alignment SAM files with the Genome files, we sort them using SAMtools and transform them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

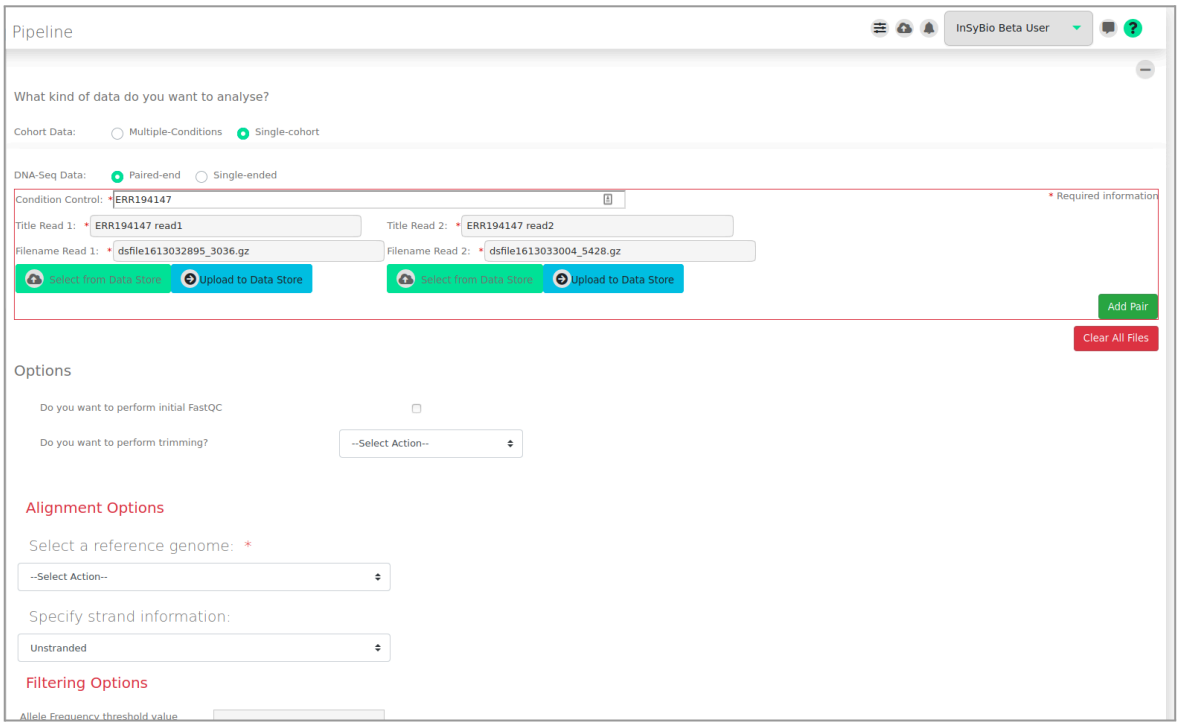
We also offer a Significant Gene file creation, where if only one cohort is used we create a file with the variants with the lowest SIFT score or if multiple cohorts are used we create pairs of cohorts and calculate their significant gene variants..

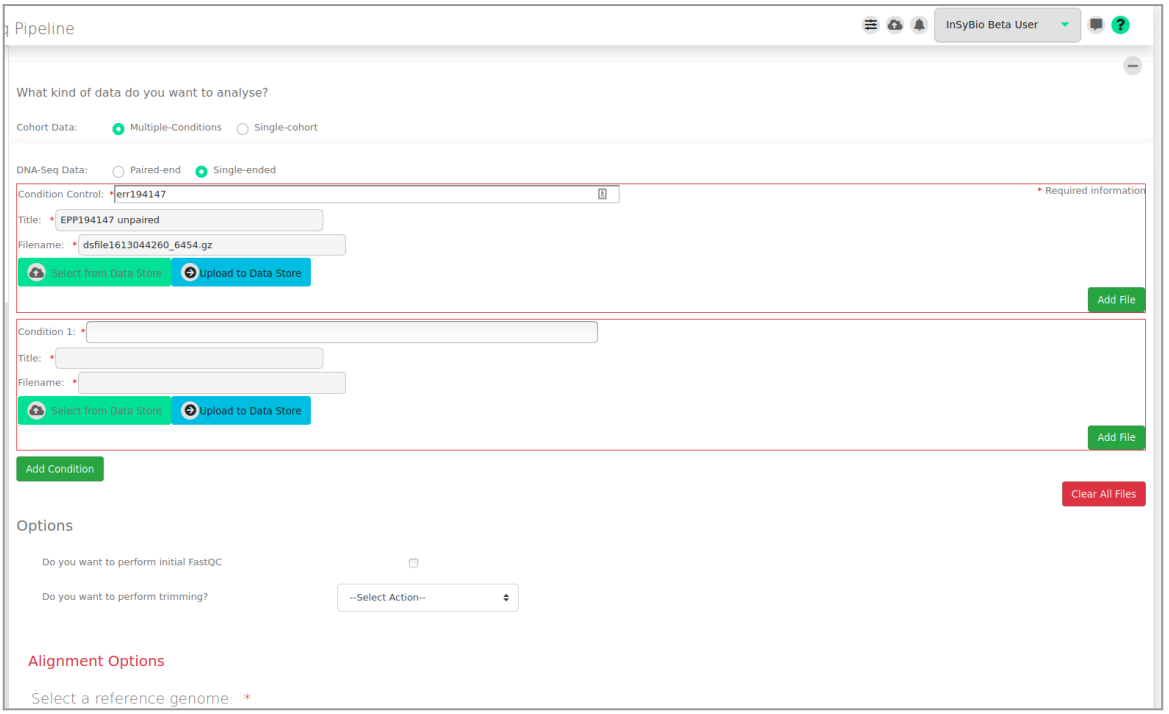
To start the DNA-Seq Pipeline:

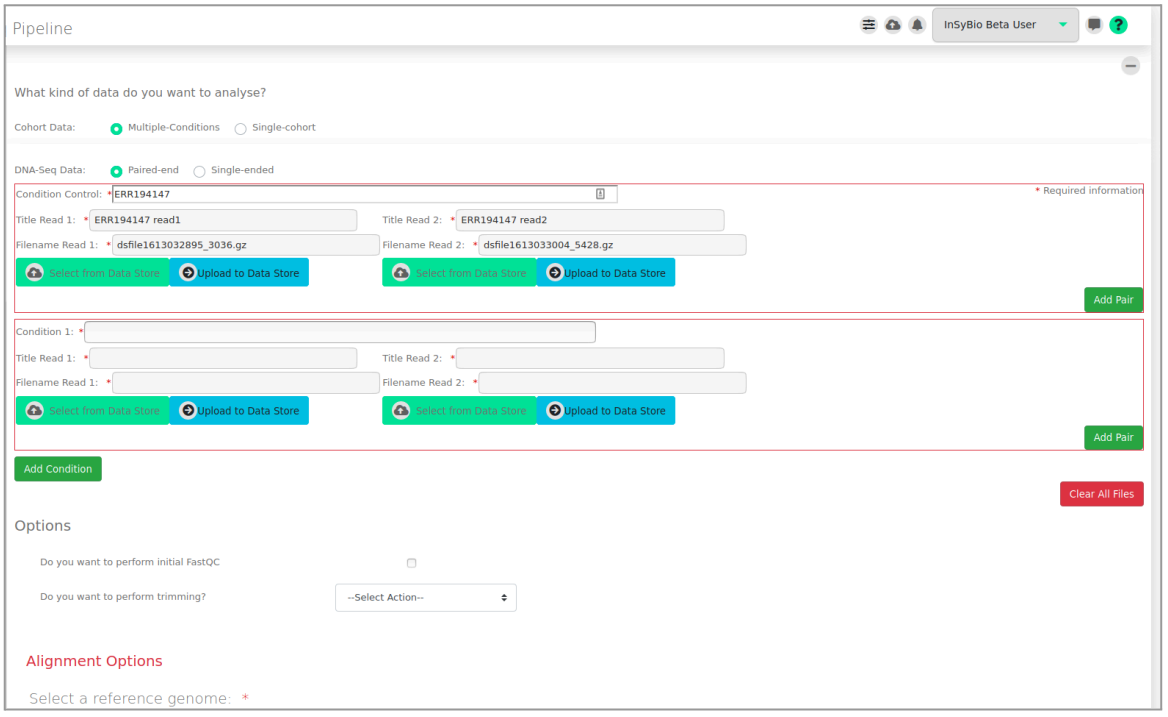
Click in the menu "InSyBio DNA-Seq" and you will be redirected to the "DNA-Seq Pipeline Dashboard" , select the "Add new job" button and then:

- 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

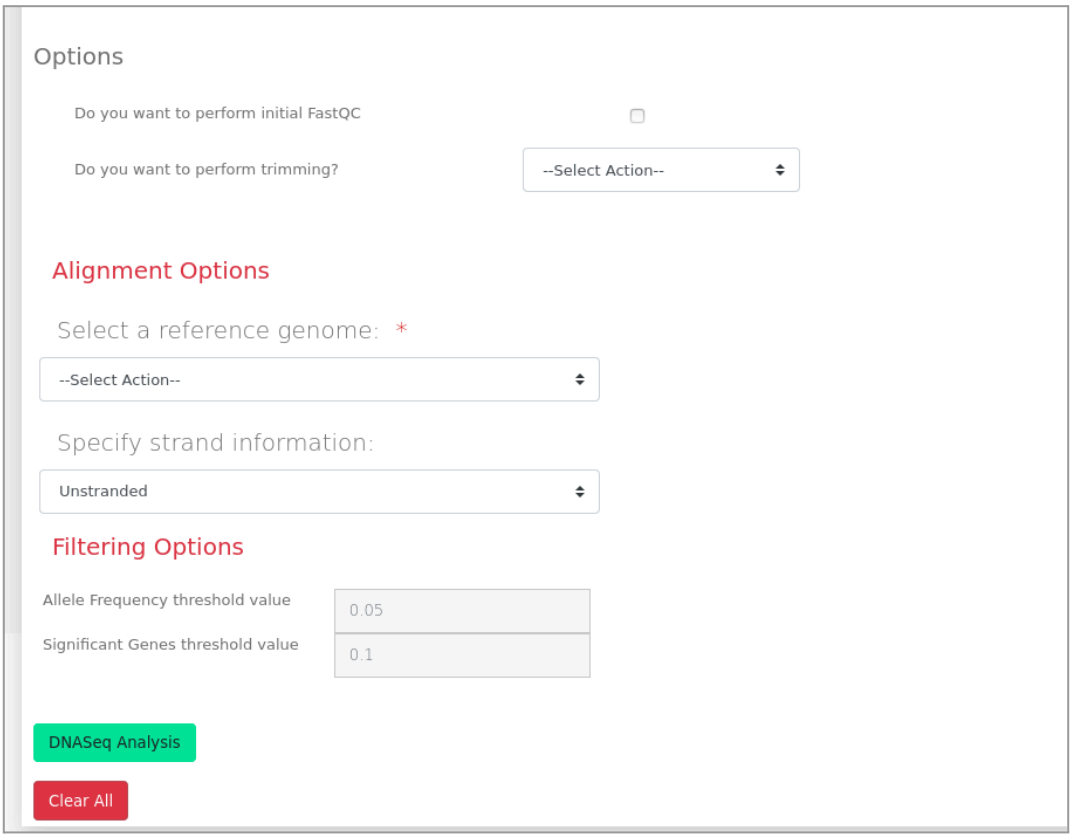




- Name Conditions/Group of files you want to Analyze.
- For each condition add single or paired files by:

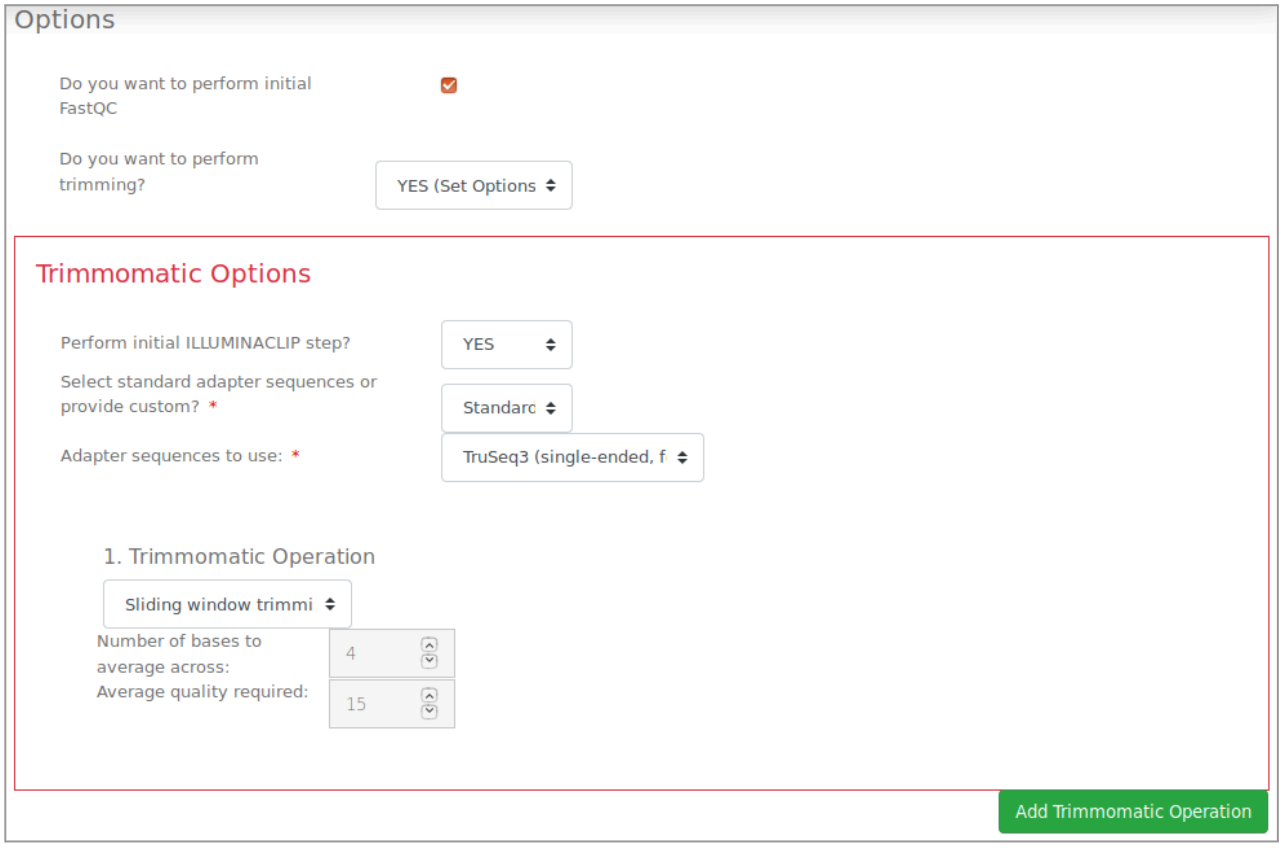
- Uploading a new file of DNA-Seq Experiments in fastq format. You are redirected to the Data Store where step by step instructions guide you for both files uploading.
- Or Selecting a file of DNA-Seq Experiments in fastq format from the Data Store. There you can find your previously uploaded files or InSyBio pre-uploaded sample datasets.

- Select if you want to perform FastQC Quality Control to the initial Data.



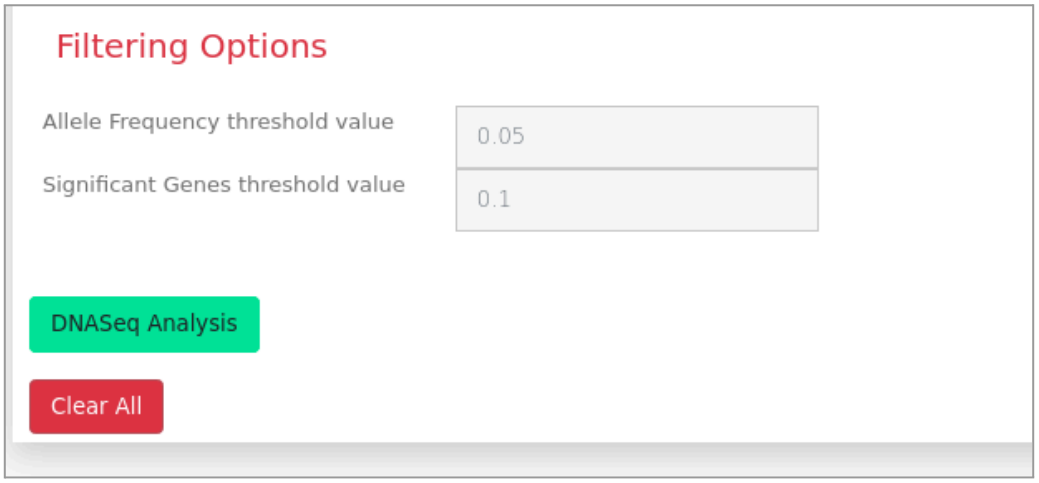
- Select if you want to perform trimming of the data with Trimmomatic, either with our Default Options or add your own (If trimming is selected FastQC will be performed to the trimmed data). Possible manual options are to:
  - Perform initial ILLUMINACLIP step
    - With Standard adapters (TrueSeq2,TrueSeq3 or Nextera for paired or single ended)
    - Or With Custom adapters in fasta format
  - Perform sliding window trimming
  - Drop reads below a specific length

- Cut bases off the start of a read, if below a threshold quality
- Cut bases off the end of a read, if below a threshold quality
- Cut the read to a specified length
- Cut the specified number of bases from the start of the read
- Drop the read if the average quality is below a specified value
- Trim reads adaptively, balancing read length and error rate to maximise the value of each read



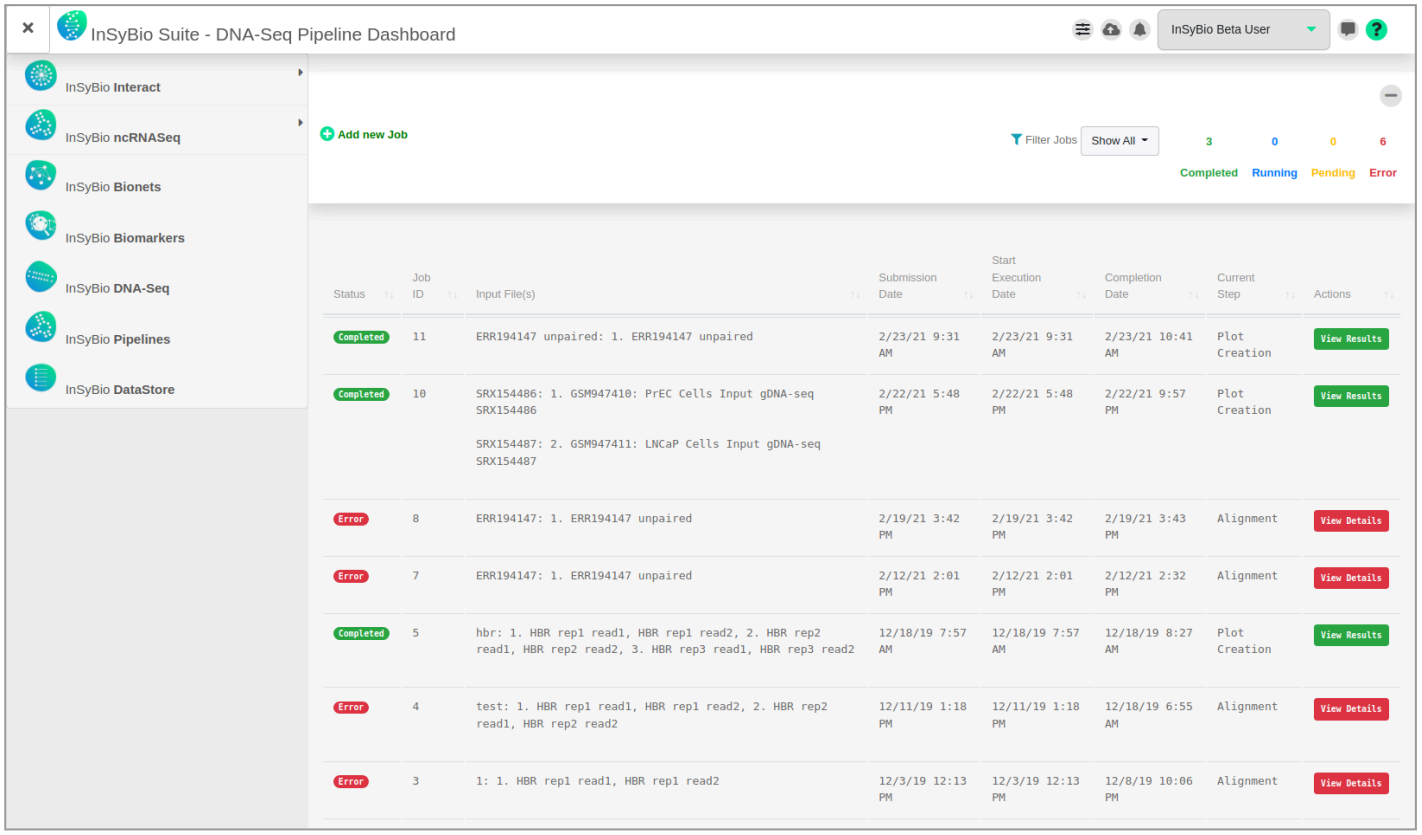
- Select the Genome the input files belong, from our 2 built-in options (HumanGRCh38 or MouseGRCm38).

- 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.

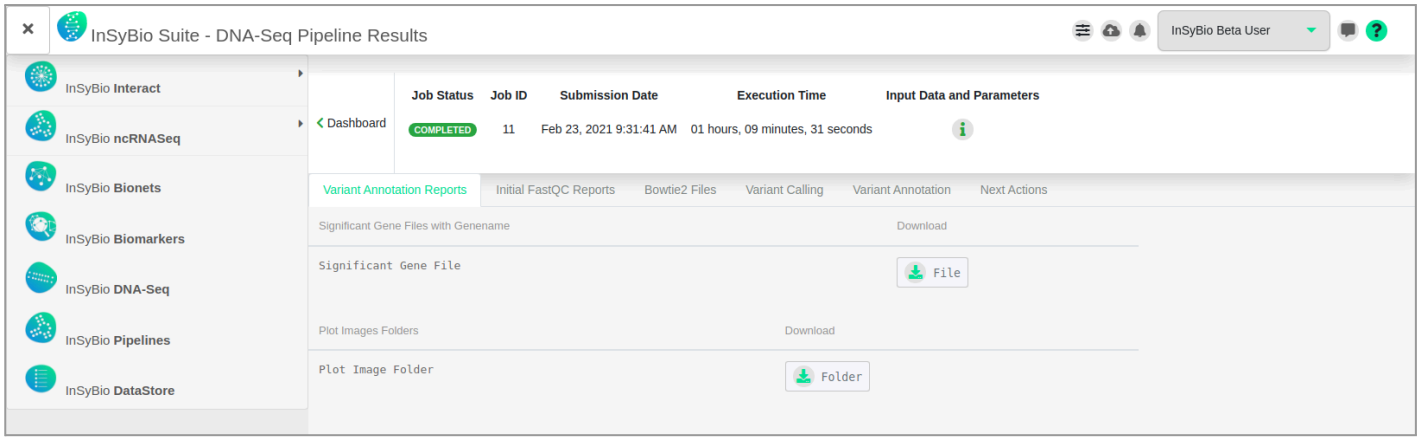


To view the results:

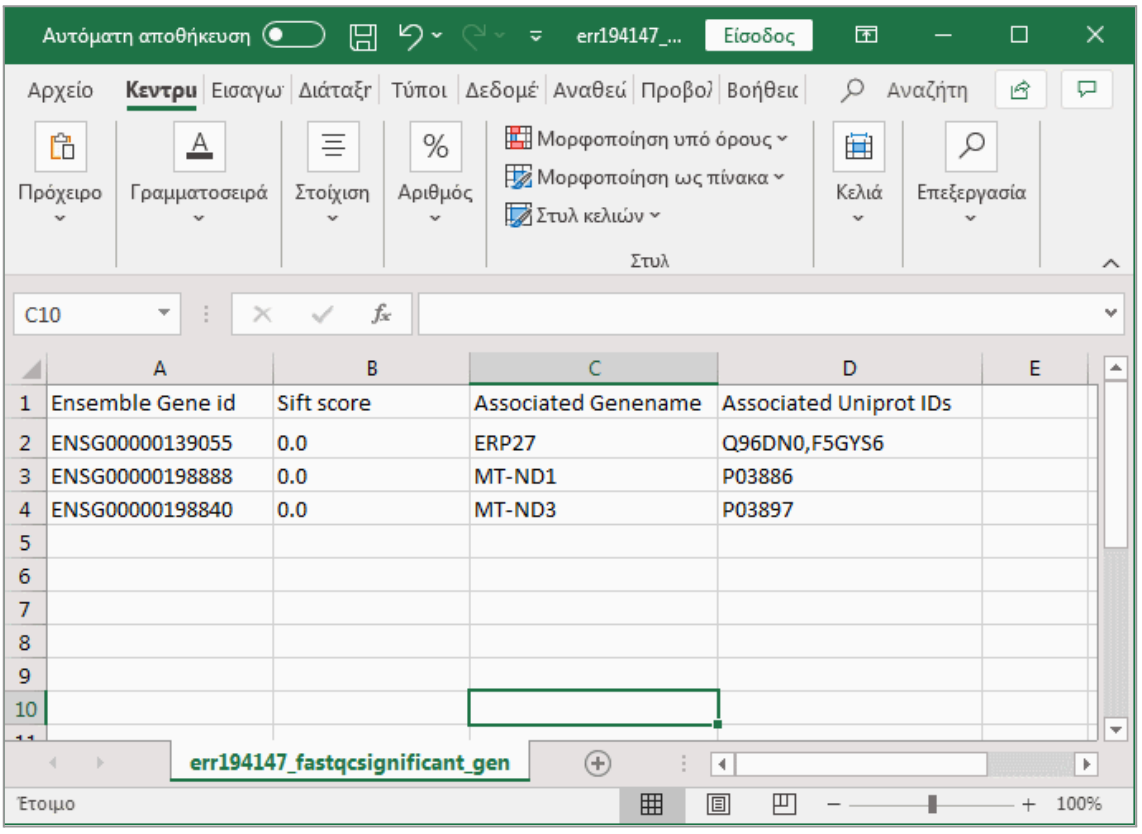
By starting a calculation you are informed if it was submitted successfully. Then you can move to the DNA-Seq Pipeline and view the Dashboard, where you can view the status of your current and previous DNA-Seq Pipeline jobs.



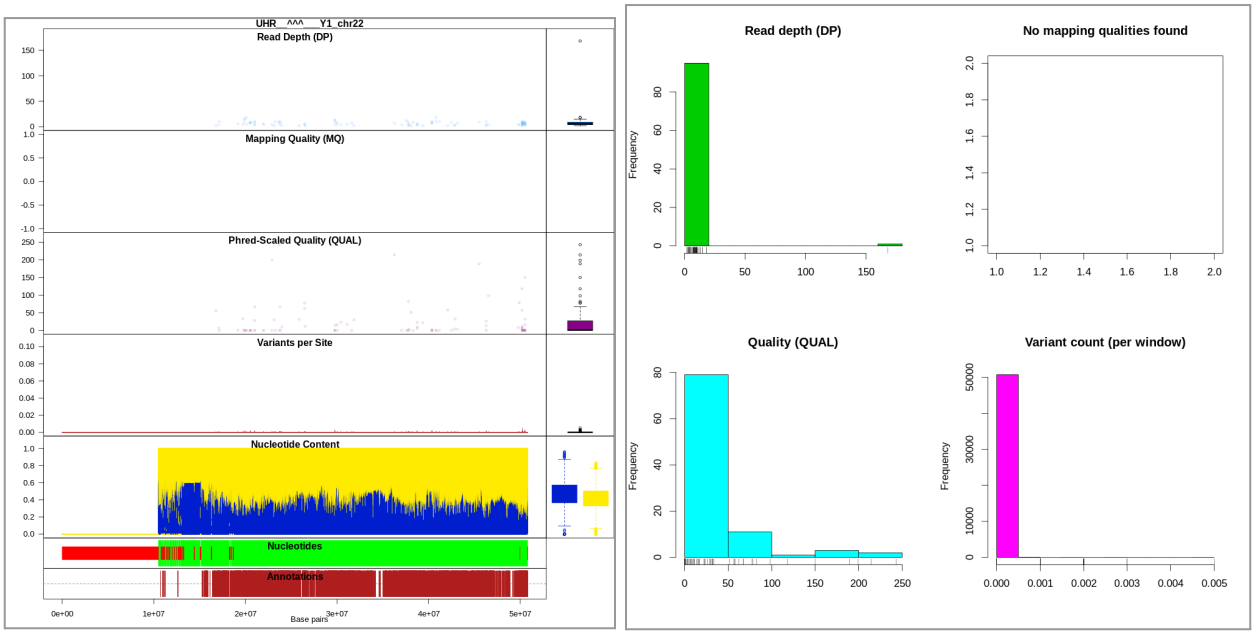
At completion of the Analysis you can select the View Results at the Actions column and view the produced files, that are separated according to the step they were produced.



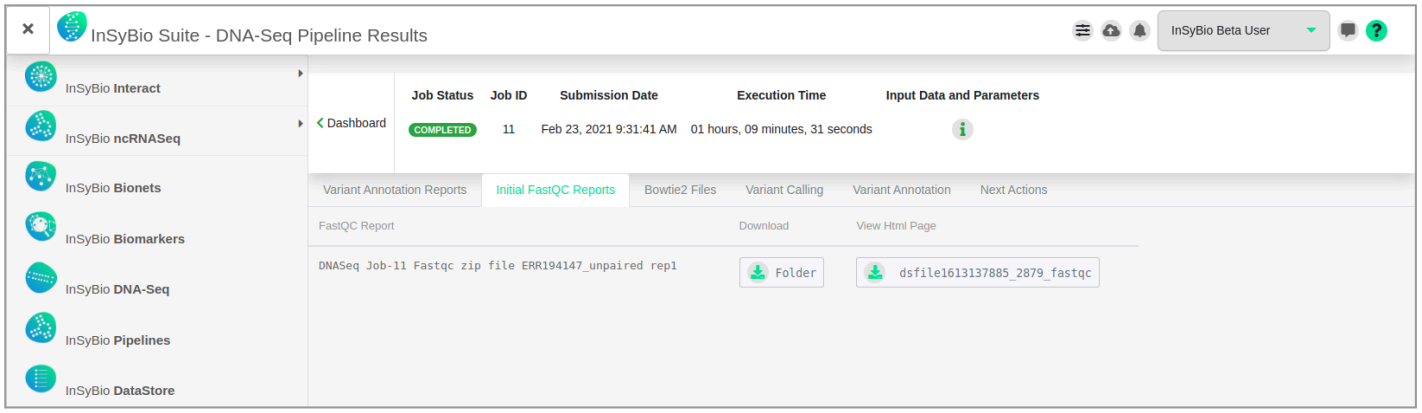
In the Variant Annotations reports tab you can download visual information and the Significant Gene Files with Genename notation, and some variant alignment images.



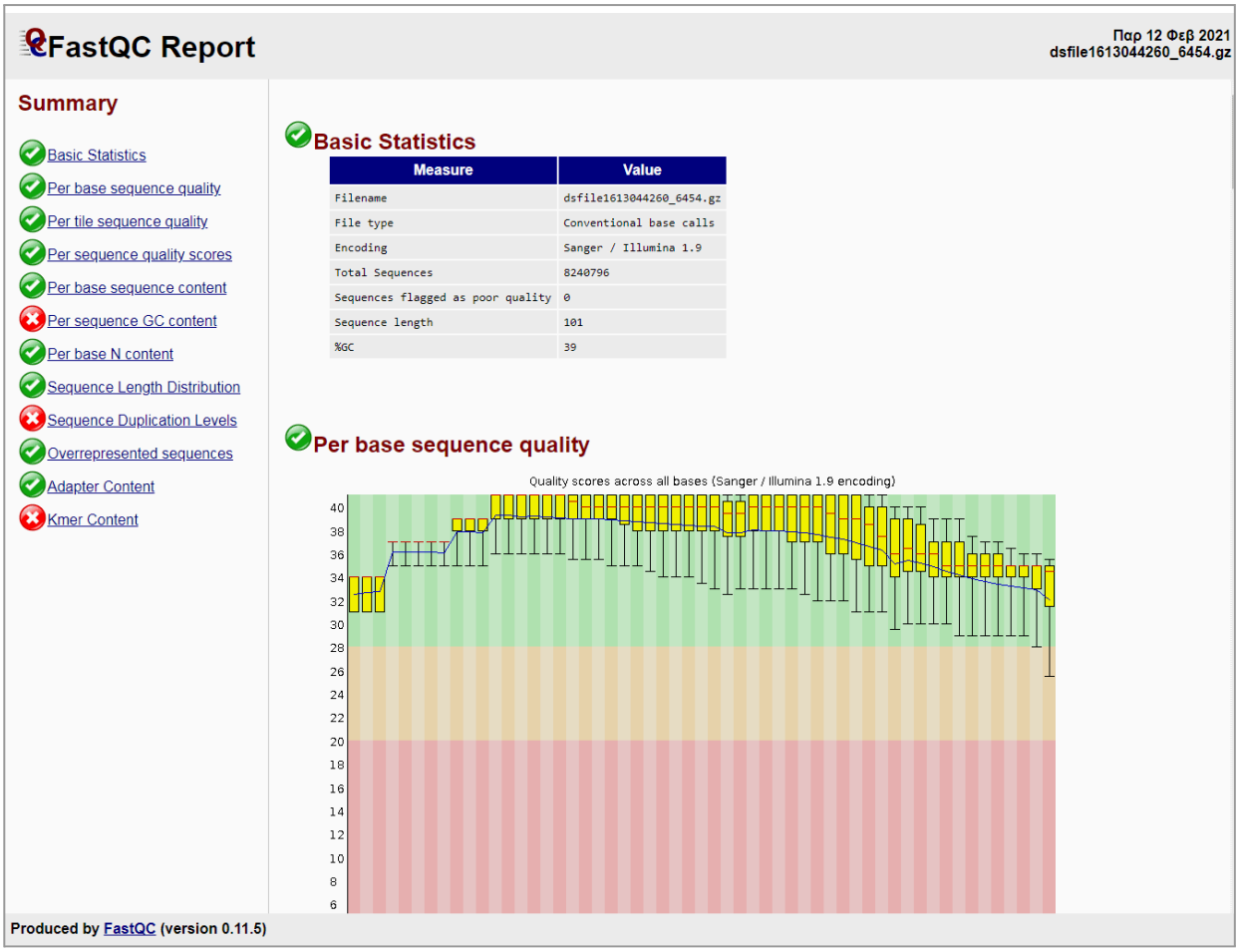
Example of the Significant Gene File being viewed with Microsoft Excel.



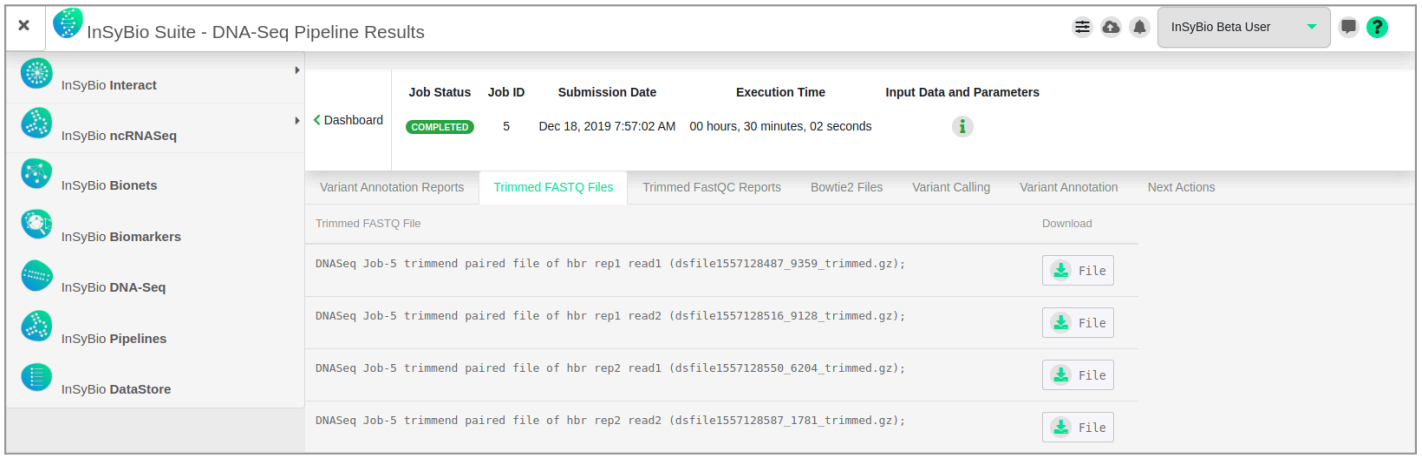
Example of the produced images and plots, (if there are enough data per chromosome).



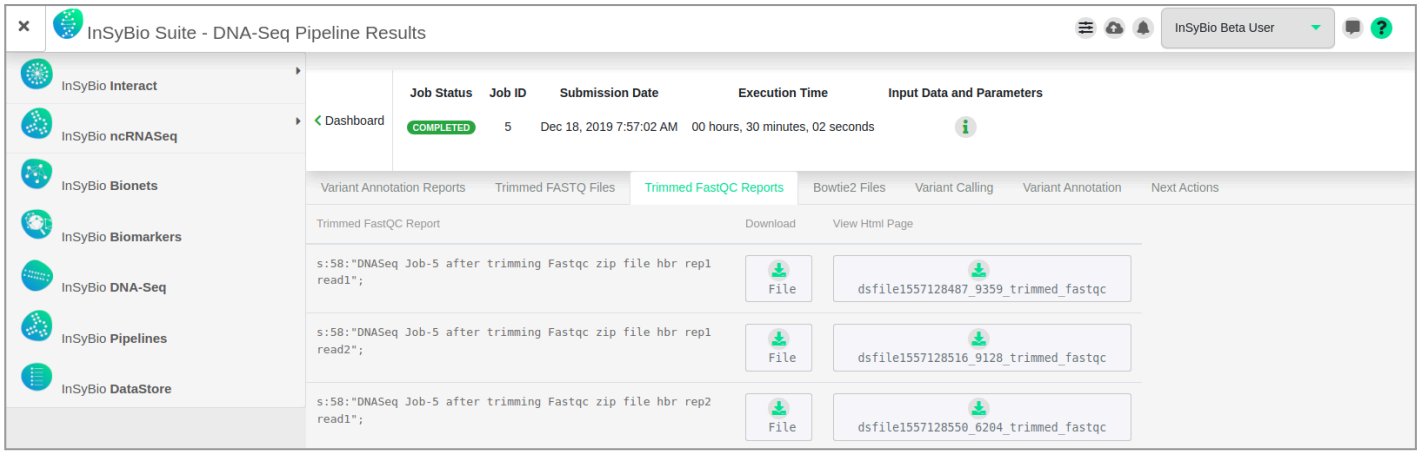
If Initial FastQC is selected, in the Initial FastQC reports the FastQC reports of the input files can be downloaded.



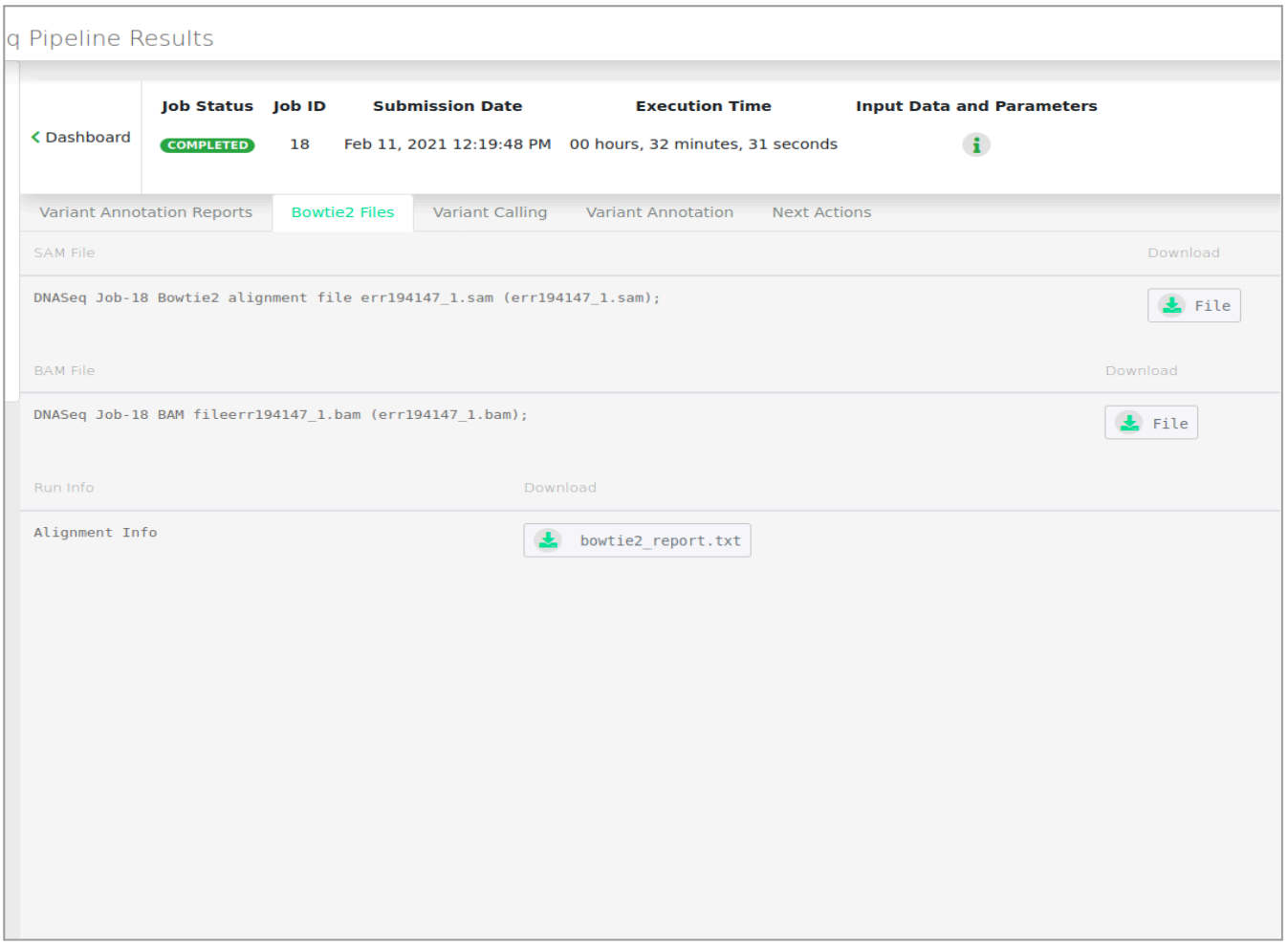
Example of a FastQC Report html file, one for each experiment is produced.



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



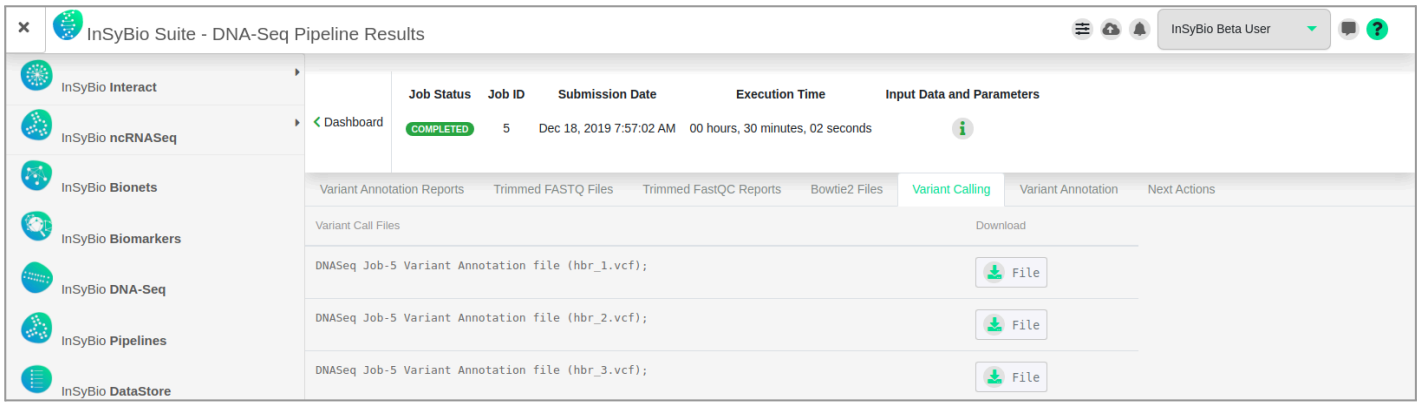
In the Trimmed FastQC reports the FastQC reports of the trimmed files can be downloaded.



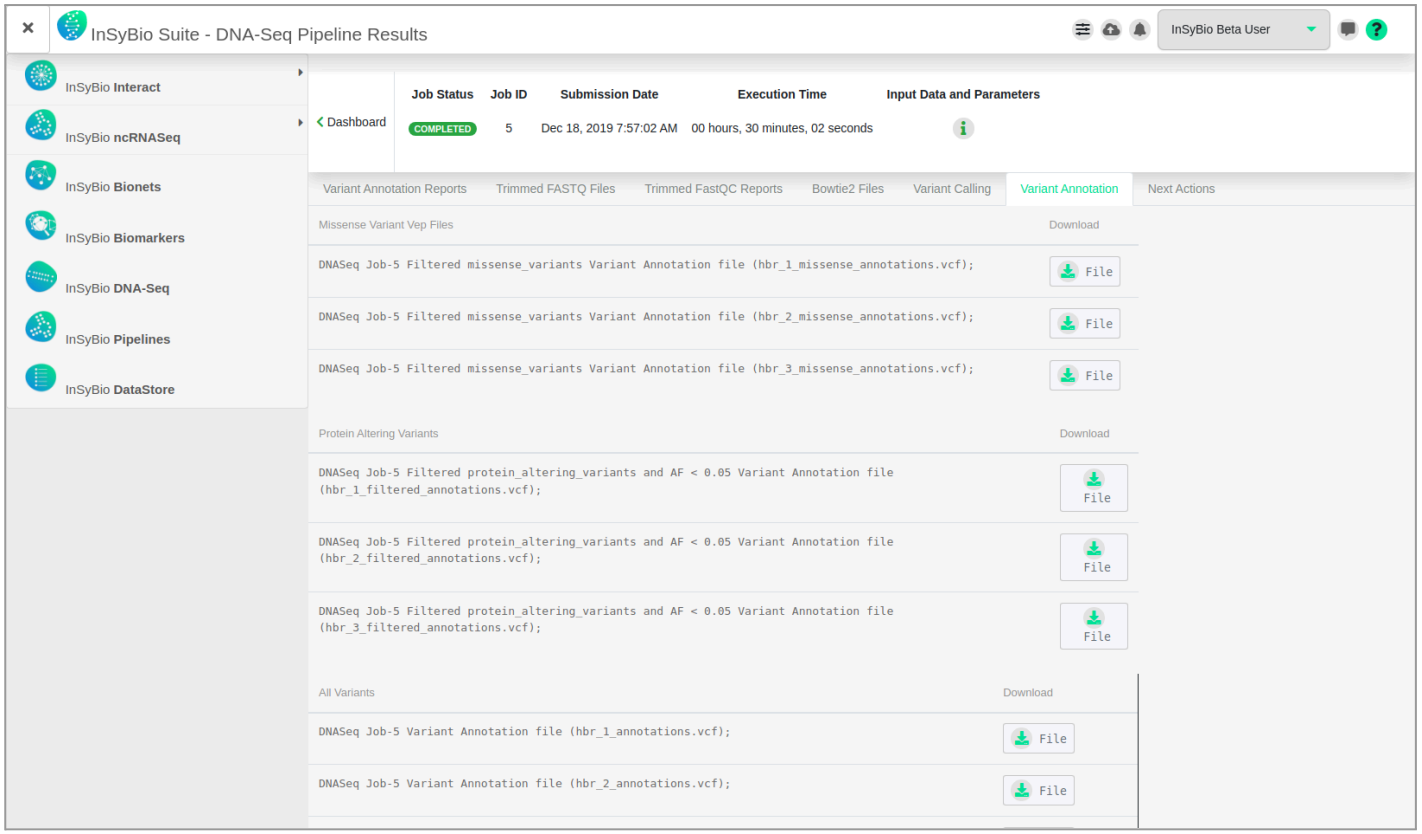
In the Bowtie2 files tab, the Bowtie2 alignment sam and bam files can be downloaded.

Example of Alignment information inside the bowtie2_report.txt:

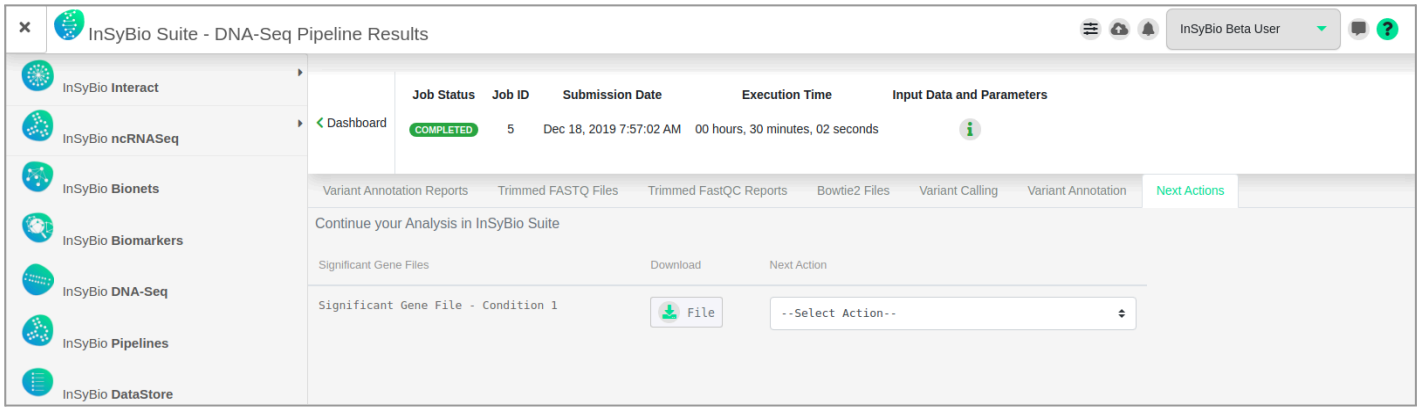
```
8131633 reads; of these:
      8131633 (100.00%) were unpaired; of these:
      34333 (0.42%) aligned 0 times
      4183088 (51.44%) aligned exactly 1 time
       3914212 (48.14%) aligned >1 times
99.58% overall alignment rate
```



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



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.



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..

Cloud computing Infrastructurer and Security Certifications

InSyBio Suite and all its tools are running over the cloud computing as a service infrastructure of Vultr (https://www.vultr.com), at the Amsterdam (Netherlands) facilities, offering the following security attestations and certifications (SOC 2+ (HIPAA), PCI (Merchant), CSA Star Level 1, ISO/IEC 20000-1:2018, ISO/IEC 27001:2022, ISO/IEC 27017:2015, ISO/IEC 27018:2019).

How to get InSyBio Interact

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

To purchase InSyBio Interact commercial version 3.3 please contact us at sales@insybio.com.

About Us

InSyBio Inc 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