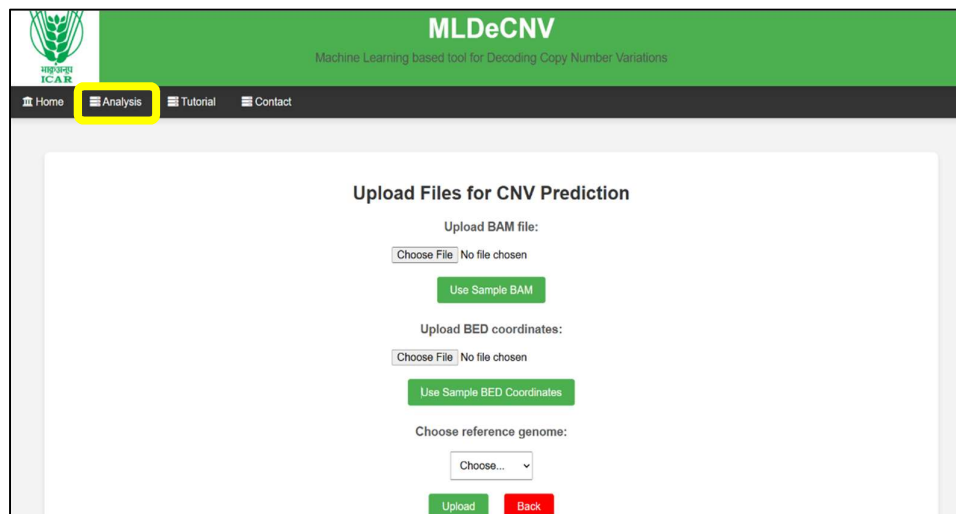


MLDeCNV User Manual

Welcome to the MLDeCNV User Manual. This guide will help you navigate through the analysis process using our tool to detect deletions and duplications in plant genomes.

Uploading Data

1. **Navigate to the Analysis Page:** Go to <http://login1.cabgrid.res.in:5106/analysis>.

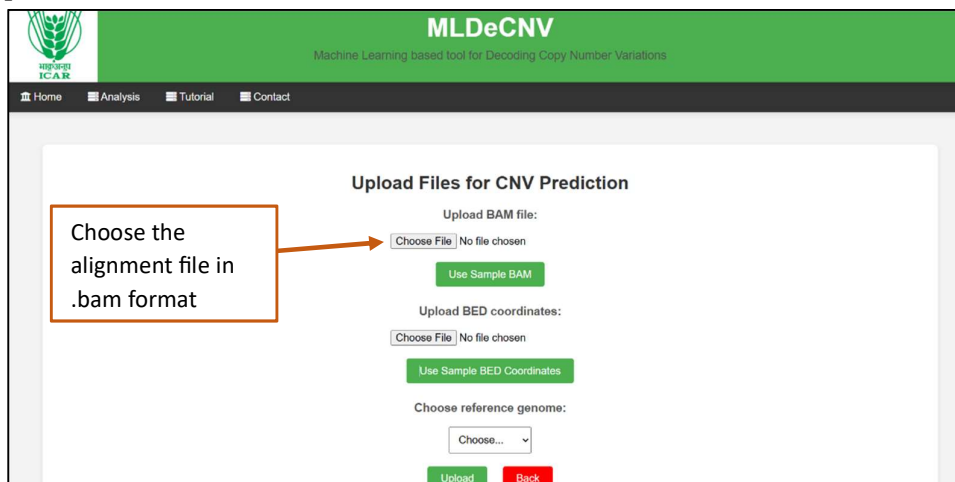


The screenshot shows the MLDeCNV web application interface. The top header is green with the MLDeCNV logo and the tagline 'Machine Learning based tool for Decoding Copy Number Variations'. Below the header is a dark navigation bar with links for Home, Analysis (highlighted with a yellow box), Tutorial, and Contact. The main content area is titled 'Upload Files for CNV Prediction' and contains three sections: 'Upload BAM file:' with a 'Choose File' button and a 'Use Sample BAM' button; 'Upload BED coordinates:' with a 'Choose File' button and a 'Use Sample BED Coordinates' button; and 'Choose reference genome:' with a 'Choose...' dropdown menu. At the bottom of the form are 'Upload' and 'Back' buttons.

2. **Select Your Data File:**

One needs to obtain the following items to successfully run this tool:

- BAM File (*.bam):** This is your binary alignment map file in .bam format, which contains the sequence alignment data. For that you need to run BWA alignment of your sample WGS or WES data to the reference genome of that plant.



This screenshot is similar to the previous one but includes an annotation. A text box on the left side of the form contains the text 'Choose the alignment file in .bam format'. An orange arrow points from this text box to the 'Choose File' button in the 'Upload BAM file:' section of the form.

- ii. **Target BED Coordinate File (*.bed):** This file specifies the genomic regions of interest. It should be formatted in the standard .bed file format, which includes chromosome number (as specified in the reference genome index file) and start & end coordinate regions.

The format of the target file is as follows:

Chromosome	Start Position	End Position
ENA CP002684 CP002684.1	16822085	16822889
ENA CP002685 CP002685.1	3607303	3627774
ENA CP002685 CP002685.1	5412309	5412684
ENA CP002686 CP002686.1	11588609	11589356
ENA CP002688 CP002688.1	17364512	17367343

There should not any header row in the coordinate file. The first field denotes chromosome, the second and third fields define the start and the end positions of a target region, respectively. The columns of the target file must be separated by tabs.

The screenshot shows the MLDeCNV web interface. The header is green with the MLDeCNV logo and tagline 'Machine Learning based tool for Decoding Copy Number Variations'. Below the header is a navigation bar with links: Home, Analysis, Tutorial, and Contact. The main content area is titled 'Upload Files for CNV Prediction'. It contains three sections: 'Upload BAM file:' with a 'Choose File' button (labeled 'No file chosen') and a 'Use Sample BAM' button; 'Upload BED coordinates:' with a 'Choose File' button (labeled 'No file chosen') and a 'Use Sample BED Coordinates' button; and 'Choose reference genome:' with a 'Choose...' dropdown menu. At the bottom are 'Upload' and 'Back' buttons. An orange box with an arrow points to the 'Choose File' button for 'Upload BED coordinates:', with the text 'Choose the target coordinate file in .bed format'.

- iii. **Genome Reference File (*.fasta):** This is the reference genome sequence in FASTA format. It should match the genome version used for generating the BAM file.

The reference genome of the plant of interest can be obtained from NCBI genome browser (<https://www.ncbi.nlm.nih.gov/datasets/genome/>) or any other genomic databases available online. Select the corresponding genome reference for which sequenced reads were mapped.

MLDeCNV
Machine Learning based tool for Decoding Copy Number Variations

Home Analysis Tutorial Contact

Upload Files for CNV Prediction

Upload BAM file:

Choose File | No file chosen

Use Sample BAM

Upload BED coordinates:

Choose File | No file chosen

Use Sample BED Coordinates

Choose reference genome:

Choose...

Upload Back

Choose the reference genome file in .fasta format

3. **Upload the File:** Once selected all the 3 file, click the "Upload" button to upload your data.

login1.cabgrid.res.in:5106 says
Files uploaded successfully. Please click OK for prediction.

OK

Click on the "OK" button once the data are loaded.

Upload Files for CNV Prediction

Upload BAM file:

Use Sample BAM Sample BAM loaded.

Upload BED coordinates:

Use Sample BED Coordinates Sample BED Coordinates loaded.

Choose reference genome:


Arabidopsis

Arabidopsis Reference Genome loaded.

Uploading: 100.00%

Running the Analysis

Click on the "Ok" button on the pop-up once all the data are uploaded successfully. Then the analysis will start and the page will look like given below.


[Home](#)
[Analysis](#)
[Tutorial](#)
[Contact](#)

Upload Files for CNV Prediction

Upload BAM file:

[Use Sample BAM](#)
Sample BAM loaded.

Upload BED coordinates:

[Use Sample BED Coordinates](#)
Sample BED Coordinates loaded.

Choose reference genome:

Arabidopsis

Arabidopsis Reference Genome loaded.

The analysis is going on

Viewing Results

Once the analysis is done the page will automatically redirected to the results and the result table will be shown as below where the CNV type i.e whether “Deletion”, “Duplication” or “No CNV” will be predicted in each coordinate provided anda prediction probability and the other parameters of the coordinates will be also shown as below. The user can download the prediction result and also all the features used in the model in the backend.

Prediction Results

Region	PredictionType	Prediction Probability	Coverage	Mean Depth	Mean Base Quality	Mean Map Quality	Insert Size Average	Insert Size Standard Deviation	Percentage Properly Paired Reads
ENA CP002684 CP002684.1:270471-270571	Deletion	0.860	90.09900	27.41580	34.4	13.7	98.5	16.5	25.0
ENA CP002684 CP002684.1:493601-498900	Deletion	0.835	100.00000	58.10750	36.7	59.8	451.2	140.8	99.8
ENA CP002684 CP002684.1:792299-889155	Deletion	0.585	99.99900	78.86350	36.6	59.6	467.2	141.3	99.3
ENA CP002684 CP002684.1:21232801-21233500	Deletion	0.960	4.28571	2.41143	37.6	52.0	164.7	95.1	71.9
ENA CP002686 CP002686.1:22527370-22568227	Duplication	0.805	83.60420	64.68120	36.7	52.7	665.1	1180.0	95.4
ENA CP002686 CP002686.1:22615765-22616131	Duplication	0.835	100.00000	181.93700	35.4	27.5	315.3	138.6	97.8
ENA CP002686 CP002686.1:22616201-22618900	Duplication	0.920	99.37040	118.96100	36.6	58.3	479.0	146.8	99.7
ENA CP002686 CP002686.1:17540651-17542051	No CNV	0.985	100.00000	82.36900	36.6	59.3	463.2	123.7	99.6
ENA CP002684 CP002684.1:22536172-22540564	No CNV	0.920	99.74960	65.06060	36.7	60.0	462.4	134.9	99.8