**Document of SuperDecode toolkit v1.0**

South China Agricultural University, Yao-Guang Liu’s Lab

July 2024

SuperDecode is an integrated toolkit for mutation analysis. It comprises three modules, DsDecodeMS, HiDecode, and LaDecode, each designed to automatically decode mutations from Sanger, high-throughput short-read (next-generation sequencing), and long-read sequencing data (third-generation sequencing), respectively, from target-site-containing amplicons. DSDecodeMS is an updated version of our previously developed web-based tools DSDecode, which identify mutations directly from Sanger sequencing chromatograms of target-site-containing PCR amplicons (2015; 2017). DSDecodeMS is used to identify mutant alleles from superimposed sequencing peaks derived from biallelic and heterozygous mutations in diploid organisms. HiDecode is designed for large-scale mutation identification from one or two target-site-containing amplicons in a large number of samples using NGS. Various small mutations, including chimeric mutations, can be analyzed by HiDecode in diverse materials, such as diploid and polyploid organisms, culture cell lines, protoplasts, and calli. LaDecode is a dedicated tool for decoding multiple and complex mutations induced by multiplex target-site editing within a certain region by long-read TGS of long-range PCR amplicons, including PacBio and Nanopore sequencing. Leveraging the advantages of long sequencing reads, LaDecode is capable of identifying complex mutations, including complex SVs, and distinguishing all haplotypes in the samples.

Considering the time and cost, it is recommended to use Sanger sequencing and DsDecodeMS analysis for fewer samples; NGS and HiDecode analysis are used for large numbers of samples (close to or more than 96 samples).

* **Software download and installation**

1. Software download:

We provide GUI version for Windows, MacOS systems, and a command-line version for Linux system. You can obtain the packages, including this document, from GitHub (<https://github.com/xiexr/SuperDecode>). Alternatively, your can directly download the packages from Cowtransfer.

Cowtransfer link: <https://tbtools.cowtransfer.com/s/d69cdec128f64e>

SuperDecode can also be accessed from the plugin store of TBtools.

2. Software installation:

The packages can be directly used in your own PC without installation. After you download the bianary file, unzip it to a local directory on your disk (Note: no space allowed in the directory path). For GUI-versions, clicking the "SuperDecode.exe" in the unziped directory to open the program. For convenience, you can send a shortcut of "SuperDecode.exe" to Desktop. It will take a few minutes to open the program of MacOS for the first time.

For Linux version, Unzip the tar.gz file to your local disk (using command: tar -zxf superdecode\_linux.tar.gz). Ensure no space in the path. An example script is provided in the directory for you to run the test example data of each module.

3. Download the example file

SuperDecode provides example file for each module, which can be accessed in the folder of SuperDecode package. For convenience, the example files can be loaded automatically by clicking the “Load example files” from the “Help” menu.

* **DSDecodeMS module**

DSDecodeMS decodes mutations directly from the Sanger sequencing chromatograms of target-site(s) contained amplicons. It is majorly used for analyzing simple mutation types (such as homozygous, heterozygous and biallelic) of diploid organisms.

**1. Amplification of target-site(s) contained fragments**

Due to the limitation of read length of Sanger sequencing, the size of amplified fragment is suggested to be less than ~700 bp. Site-specific primers T#-F and T#-R are designed about 200~350 bp upsteam and downstream of the target site, repectively. The length of T#-F and T#-R is about 18~21 nt. *Tm* value of T#-F and T#-R is 58℃~60℃, which is calculated according to the formula *Tm* = (GC% × 41 + 69.3) - 650/L (L = primer length). We highly suggest to check the specificity of the primers using using the “Primer-BLAST” tool of NCBI. The amplicon fragment can contain 1~2 target sites. If it has highly homologue sequences in the genome, the length of amplified fragment can be extended appropriately. The recommended PCR mixture is shown in Table 1-1.

Table 1-1 The PCR reaction mixture of the DSDecodeMS (25 μL)

|  |  |
| --- | --- |
| 2 × Taq mix | 12.5 μL |
| T#-F (10 μM) | 0.5 μL (0.2 M final concentration) |
| T#-R (10 μM) | 0.5 μL |
| gDNA | 20~30 ng |
| ddH2O | up to 25 μL |

The PCR program is shown in Table 1-2.

Table 1-2 PCR program for amplying target-site contained fragment

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 94℃, 3 min |  |
| Denaturing | 96℃, 15 s | 30~32 cycles |
| Annealing | 58℃, 15 s |
| Extension | 65℃, 10s,  68℃, 10s,  72℃, 10s,  65℃, 10s  (If the target fragment is >1 kb, increase the temperature by 5 s per section) |
| Final extension | 72℃, 5 min |  |

(Note: Extension uses variable temperature to improve the amplification efficiency of sequences with uneven GC distribution)

Taking 3~4 μL of the PCR products for agarose gel (1%) electrophoresis to confirm the size and specificity. If the concentration of products is too low, 2~3 cycles can be appropriately added of extension to increase the concentration.

The PCR products (~20 μL) are sent to the company for Sanger sequencing (most companies provide the service of PCR product purification). Alternatively, user can purify the final PCR products using a purification kit and send the purified products to the company for Sanger sequencing. To obtain high-quality sequencing chromatograms, We highly recommend to use an inner primer which is designed 150~300 bp upstream or downstream of the target site for sequencing, but not use the PCR primers. If two target sites contained in the amplicon, a forwad primer upstream of the first target site and a reverse primer downstream the second target site can be designed for sequencing each site.

**2. Decode Sanger sequencing file using DSDecodeMS**

The DSDecodeMS module in SuperDecode was used to detect mutation on the ab1 file obtained by Sanger sequencing.

(1) The interface of DSDecodeMS

The interface of DSDecodeMS mainly includes a reference panel, a sequencing file input panel, a viewer for chromatogram traces, and a result panel. (Figure 1-1).

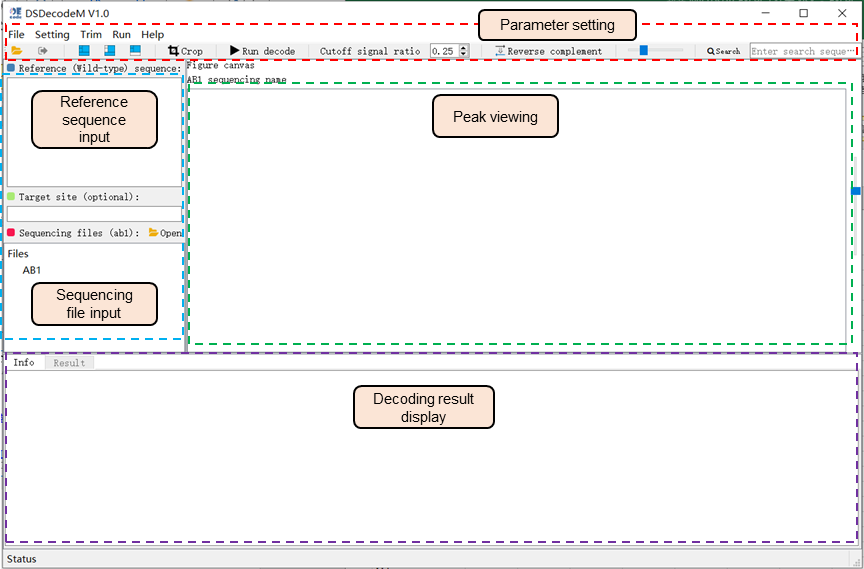


Figure 1-1 The interface of DSDecodeMS

(2) Decoding operation of DSDecodeMS

The use of DSDecodeMS is mainly includes following three steps:

i) Paste the wild-type reference sequence corresponding to the amplicon in the sequence input module box (the wild-type reference sequence is between the T#-F primer and the T#-R primer). The input type is common text or Fasta format (same direction as the sequencing chain or reverse complementary). Next, enter the target sequence (with or without PAM), which will be serviced for locating mutations close to the target. This is an optional steps in most cases. Clicking the “Open” button to open the sequencing file input dialog. User can input the sequencing files (in ab1 format) corresponding to the same reference sequence and target in this dialog. By double-clicking each sequencing file name, user can view the peak on the right viewer track; checking the peak quality one by one is recommended before starting the analysis.

ii) Generally, the preset parameters of DSDecodeMS is enough for analysis, and no parameter setting adjustment is required. If some sequencing files of slightly poor quality cannot produce correct results, user can adjust the sequencing cutoff signal threshold (default value is 0.25). Other optional settings, including the length of anchor sequenc and length of degenerate sequence can be adjuted through the “Setting” from the menu.

iii) Click the “Run decode” button on the tool bar or form “Run” menu to start the mutation analysis.

DSDecodeMS will quickly outputs the decoding results in the result panel. Users can copy the results to a Word or PPT file; or save the results by clicking the the “Save” button at the upper right corner

(3) Other handy functions

i) Triming the low-quality ends in the sequencing chromatograms

The quality of the Sanger sequencing chromatograms is usually poor at the beginning and end (about 30-60 bp). Low-quality ends in the sequencing chromatograms can also hamper decoding. DSDecodeMS provides the function to remove low-quality ends of the chromatograms using the Richard Mott algorithm or user-defined parameters from the “Trim” menu (Figure 1-2).

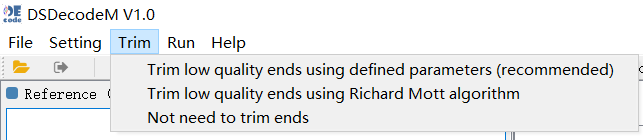


Figure 1-2 Trim the low-quality ends using DSDecodeMS

ii) Capturing figures using the “Crop” function

Each module of SuperDecode provides a screenshot function. User can open the screenshot function by clicking the “Crop” button on the toolbar. After selecting the screenshot area by dragging the left mouse button, double-clicking to copy the screenshot or right-clicking to save the screenshot.

iii) Subsequence search

Enter a subsequence in the “Search” box on the toolbar for searching it in the opened chromatograms. After pressing the “Enter” key, the program will automatically searche for the subsequence. This function supports the search for degenerate sequences.

iv) Export fasta format sequence files from ab1 files

By clicking “Export fasta” in the “Run” menu, the sequence of each ab1 file can be exported into a local file in fasta format.

(4) Causes and solutions for decoding failure

i) The PCR product may contain non-specific products. Repeated use of amplification primers for sequencing may result in poor-quality sequencing chromatograms. It is highly recommended to use an inner primer for Sanger sequencing.

ii) The amplified sequence has homologous sequence(s) in the genome. Reset the site-specific primers for amplifying the target sequence by extending a certain distance to the outside of the target to design the specific primers. Users can also design two sets of specific primers for two rounds of nested PCR to improve the specificity of the product.

iii) The mutation of the sample is a complex mutation such as chimeric mutations, or there are multiple targets in the target region that always inducing structural variations such as fragment deletion or inversion among the targets. For chimeric mutations, it is recommended to use HiDecode.For multiplex editing, it is recommended to use LaDecode.

iv) The input wild-type reference sequence does not match the sequencing file. Check whether the input reference sequence corresponds to the target amplification sequence.

v) Unenve GC or polyA(T) in the sequence region between the primer and target site causing low-quality chromatogram. Trying to use another sequencing primer on the other side of the target site that do not contain such complicated regions.

* **NGS library preparation and analysis of HiDecode**

When large numbers (>50) of the total number of samples to be analyzed, HiDecode is recommended to reduce the cost. This method is particularly suitable for detecting one- or two-targets contained sequence within a certain region (<150 bp) in samples such as diploid chimeric mutation, polyploids, cell lines, protoplasts, or callus tissue.

**1. NGS library construction of HiDecode**

The NGS library for HiDecode is constructed mainly through two rounds of PCR. A pair of unique barcodes are added to both ends of the amplicons and then mixed into a sequencing library for NGS (Figure 2-1).



Figure 2-1 PCR-based construction of sequencing libraries for HiDecode

(1) Primer design of HiDecode

HiDecode library construction involves three pairs of primers, including site-specific primers T#-F and T#-R for amplifying target sequences, universal primers posN-F and plateN-R for labeling plate-well positions (from A1 to H12) and plate numbers, and primers Lib-F and Lib-R for adding library adapters. The sequence structure of the NGS library is shown in Figure 2-2.



Figure 2-2 Sequence structure of NGS library

i) Site-specific primers T#-F and T#-R

Considering that the common read length of current paired-end NGS tools is ca. 150 nt for each side, the fragment size of first-round PCR is designed within ca. 200–260 bp (including the two bridge sequences). After checking the specificity, the primers T#-F and T#-R are added with different bridge sequences (16 nt) at their 5′ ends.

Site-specific forward primer (T#-F): 5′-ctcggagtgatcgcacNNNNNN…NNNNN-3′ (17 to 19 N are the genomic site-specific region, *Tm* = 56℃~58℃; the 16 nt at the 5′ end is the bridge sequence that match the posN-F);

Site-specific reverse primer (T#-R): 5′-ctgagaggctggatggNNNNNN…NNNNN-3′ (17 to 19 N are the genomic site-specific region, *Tm* = 56℃~58℃; the 16 nt at the 5′ end is the bridge sequence that match the plateN-R).

ii) Universal primers posN-F and plateN-R

posN-F contains sequences (from 5′ to 3′) matching the library adapter, position barcode, and sequences (16 nt) binding to the 5′ sequence of the first round of products. plateN-F contains sequences (from 5′ to 3′) matching the library adapter, plate barcode, and sequences (16 nt) binding to the 3′ sequence of the first round of products. A total of 96 position barcodes (6 nt) and 96 plate barcodes (6 nt) for labeling the positions (A01 to H12) of samples in 96-well plates and the corresponding plate number, respectively, thus creating uniquely barcoded amplicons for up to 9,216 samples (96 × 96). Users can synthesize a certain number of plate-barcoded primers based on the number of samples.

posN-F: 5′-CGCTCTTCCGATCTGTCANNNNNNctcggagtgatcgcac-3′ (the underlined sequence is the adapter sequence match the library primer Lib-F, the 6 N are the position barcode, and the lowercase are the sequence binding to the 5′ sequence of the first round of products).

plateN-R: 5′-CAGACGTGTGCTCTTCCGATCTCTGTNNNNNNctgagaggctggat gg-3′ (the underlined sequence is the adapter sequence match the library primer Lib-F, the 6 N are the position barcode, and the lowercase are the sequence binding to the 3′ sequence of the first round of products).

For detailed sequences, see Appendix 1-HiDecode all primer sequences.xlsx

For convenience, the 96 position-barcoded primers can be prepared in a 96-well PCR plate and stored at -20℃ for future use.

iii) Library adapter primers Lib-F and Lib-R

NGS sequencing adapters can be introduced into the amplicons by PCR. The adapter contains a laboratory-specific barcode (Lab barcode) to distinguish libraries from different customers. Sequencing companies demultiplex raw sequencing reads by this barcode. A single- or dual-barcodes are supported. Following primers are designed with single Lab barcode.

Lib-F: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCA-3′ (the underlined sequence is the adapter sequence bingding posN-F)

Lib-R: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGA GTTCAGACGTGTGCTCTTCC-3′ (8 N sequence is the Lab barcode; the underlined sequence is the adapter sequence binding to plateN-R primer). The Lab barcode needs to be reverse complement in the library information form when send it for sequencing. If there are multiple independent sequencing libraries in the same batch (the company may mix these libraries into the same batch), different Lab barcodes should be used according to the suggestions from sequencing company. In addition, if dual-Lab barcodes are used, the sequences of Lib-F and Lib-R should be modified according to the recommendations of company.

Note: This is designed based on the Illumina platform. If other sequencing platforms are used, posN-F and plateN-R, Lib-F and Lib-R are needed to be re-designed based on the platform adapter sequence.

(2) Steps of amplification

i) Preparation of template DNA and various primers

All primers and genomic DNA working solutions are prepared with 0.3×TE (do not use 1x TE, in case too much EDTA will inhibit enzyme activity).

Site-specific primers T#-F, T#-R, and plateN-R: diluting to 10 M for use.

posN-F working solution: synthesize 96 posN-F (from posA1-F to posH12-F) with 5 OD each, 1 OD/tube. Dissolve one tube of each primer with 1×TE into 10 M storage solution, and freeze the rest as dry powder at -20℃ for storage. Take 2.0 L of each posN-F (10 M) and add it to the corresponding wells (A1~H12) of a 96-well PCR plate. Add 18 L of 0.3×TE to each well, which is 1.0 M posN-F working solution. Store at -20℃ (discard when the remaining 4~5 L is less than used below).

Template DNA working solution: Take about 200 ng of template DNA from each sample and add it to the 96-well PCR plate (A1~H12). Add 0.3×TE to each well to 10 L (20 ng/L). Store at -20℃.

ii) First round of specific amplification

The HiDecode library is constructed via two rounds of PCR reaction. In the first-round PCR, site-specific primers T#-F and T#-R are used to amplify the target-site contained fragment. The total volumn of PCR mixture is prepared according to the number of samples and then dispensed into a 96-well PCR plate. The PCR reaction mixture is shown in Table 2-1 (the number of samples is *n*). 15 L of the reaction mixture is added to each well.

Table 2-1 Total reaction mixture of first-round PCR (15 L per reaction)

|  |  |
| --- | --- |
| 2 × Taq mix | 7.5 μL × *n* |
| T#-F (10 μM) | 0.3 μL × *n* (0.2 μM final concentration) |
| T#-R (10 μM) | 0.3 μL × *n* (0.2 μM final concentration) |
| ddH2O | 7.5 μL × *n* |

Use a 96-pin replicator (or multichannel pipette) to add ~1μL of DNA template into the wells that contain the PCR reaction mixture. The program of PCR reaction is shown in Table 2-2.

Table 2-2 PCR reaction program of first-round PCR

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 94℃, 3 min |  |
| Denaturing | 96℃, 10 s | 10 cycles |
| Annealing | 58℃, 20 s |
| Extension | 65℃, 5 s |
| Extension | 72℃, 10 s |
| Extension | 65℃, 5 s |
| Denaturing | 96℃, 10 s | 18~20 cycles |
| Annealing/ Extension | 65℃, 15 s |
| Extension | 72℃, 10 s |
| Extension | 65℃, 5 s |
| Final extension | 68℃, 1 min | 1 cycle |

(Note: Extension uses variable temperature to improve the amplification efficiency of sequences with uneven GC distribution)

If the target fragment has homologous sequence(s) in genome, causing T#-F and T#-R unable to specifically amplify the target fragment, the specific primers exT#-F and exT#-R (*Tm* = ~60℃) can be designed to an extended region. Perform the first round of nested PCR to obtain highly specific amplification products. The PCR reaction mixture is shown in Table 2-3.

Table 2-3 First-round nested PCR reaction mixture (15 μL per reaction)

|  |  |
| --- | --- |
| 2 × Taq mix | 7.5 μL × *n* |
| exT#-F (10 μM) | 0.15 μL × *n* (0.1 μM final concentration) |
| exT#-R (10 μM) | 0.15 μL × *n* (0.1 μM final concentration) |
| ddH2O | 7.5 μL × *n* |

The program of PCR reaction is shown in Table 2-4:

Table 2-4 The reaction program of first-round nested PCR

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 93℃, 3 min |  |
| Denaturing | 96℃, 10 s | 15 cycles |
| Annealing | 60℃, 20 s |
| Extension | 65℃, 10 s |
| Extension | 72℃, 30 s |
| Extension | 65℃, 10 s |

Mix T#-F and T#-R primers (10 μM) in equal amounts, and then dilute with ddH2O to 3 μM each (total > *n* μL). After completing the first-round nested PCR reaction, add 1 μL T#-F/T#-R mixture (3 μM) to each well (final concentration is 0.2 μM). Continue the PCR amplification reaction to obtain the specific amplification product of the target fragment. The PCR reaction program is shown in Table 2-5.

Table 2-5 The reaction program of second-round nested PCR

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 96℃, 10 s | 17-18 cycles |
| Annealing / Extension | 58℃, 20 s |
| Extension | 65℃, 5 s |
| Extension | 72℃, 10 s |
| Extension | 65℃, 5 s |
| Final extension | 68℃, 1 min |  |

The fragment size of the first-round PCR products is 200~260 bp. The amplified products of several samples can be randomly selected from each plate for agarose gel (1.5%) electrophoresis to check the specificity and product size.

ii) Second-round PCR for barcoding

The purpose of the second-round of PCR reaction is to add the barcodes and sequencing adapters. Prepare the second-round of PCR total mixture according to the number of samples. The PCR reaction total mixture is shown in Table 2-6 (the number of samples is *n*):

Table 2-6 Second-round PCR reaction mixture (15 μL per reaction)

|  |  |
| --- | --- |
| 2 × Taq mix | 7.5 μL × *n* |
| plateN-R (10 μM) | 0.08 μL × *n* (0.05 μM final concentration) |
| Lib-F, Lib-R (10 μM) | Per 0.3 μL × *n* (0.2 μM final concentration) |
| ddH2O | 7.5 μL × *n* |

Using a 96-pin replicator (or multichannel pipette) to take ~1μL of the posN-F (1.0 μM) into a 96-well plate and add it to the reaction mixture in each PCR well (the final concentration of each posN-F is about 0.06 μM).

Add 30 μL ddH2O to dilute the first-round of PCR products for three times. And then use a 96-pin replicator (or multichannel pipette to) take ~1μL of the diluted products and add them to the reaction mixture in each PCR well (the total dilution of the products is about 45×).

The program of second-round PCR is shown in Table 2-7.

Table 2-7 PCR reaction program of second-round PCR

|  |  |  |
| --- | --- | --- |
| Denaturing | 96℃, 10s | 8 cycles |
| Annealing | 55℃, 15 s |
| Extension | 65℃, 5 s |
| Extension | 68℃, 5 s |
| Extension | 72℃, 5 s |
| Denaturing | 96℃, 10 s | 10~12 cycles |
| Annealing / Extension | 65℃, 15 s |
| Extension | 68℃, 5 s |
| Extension | 72℃, 5 s |
| Final extension | 68℃, 1 min | 1 cycle |

The length of the final products after the second-round PCR is about 340~400 bp (including adapters and barcodes). Several products can be randomly selected from each plate for agarose gel (1.3%) electrophoresis. If the concentration of products is low, 2~3 cycles can be increased of extension to increase the concentration.

Pool 3 μL of each sample from each 96-well plate to generate a group (nearly 300 μL per group) for purification using 1.3% agarose gel electrophoresis followed by a DNA purification kit. After quantifying the concentration of each group of purified products, mix the purified products of each group in equal amounts (the group with less than 96 samples should reduce the mixing amount accordingly). The final concentration of the total mixed product is at least 50 ng/μL, and the product volume is at least 60 μL.

**2. NGS sequencing**

The mixed library is sent to the company for sequencing through Illumina X-Ten, or Nova-seq platforms (if other platforms are used, posN-F and plateN-R, Lib-F and Lib-R are needed to be redesigned according to the adapter sequences provided by the platform). Since the minimum sequencing amount required by the sequencing company is 1 Gb, it is recommended that the sequencing amount of 1 to 10 96-well PCR plates is 1 Gb. Based on the product length of about 350 bp, the average depth of each sample is 30000× ~ 3000× in theory.

Some companies in China (such as Wuhan Kinstar Global, HaploX, and Geneplus) can complete the sequencing within 3~5 workdays after receiving a customer’s sequencing library samples.

**3. Mutation analysis using HiDecode**

(1) The interface of HiDecode

The main interface of HiDecode includes the menu, toolbar, file input panel, plate canvas, and decoding results panel (Figure 2-3). The menu bar provides commonly used setting functions and plugin. The bartool provides some setting options such as setting mutation threshold and CPU thread number. The file input module is used to input wild-type reference sequence file, target sequence file, and paired-end sequencing files. In the barcode settings, user can select the used position- and plate-barcodes. The plate canvas is used to interactively display the analysis results of each sample.

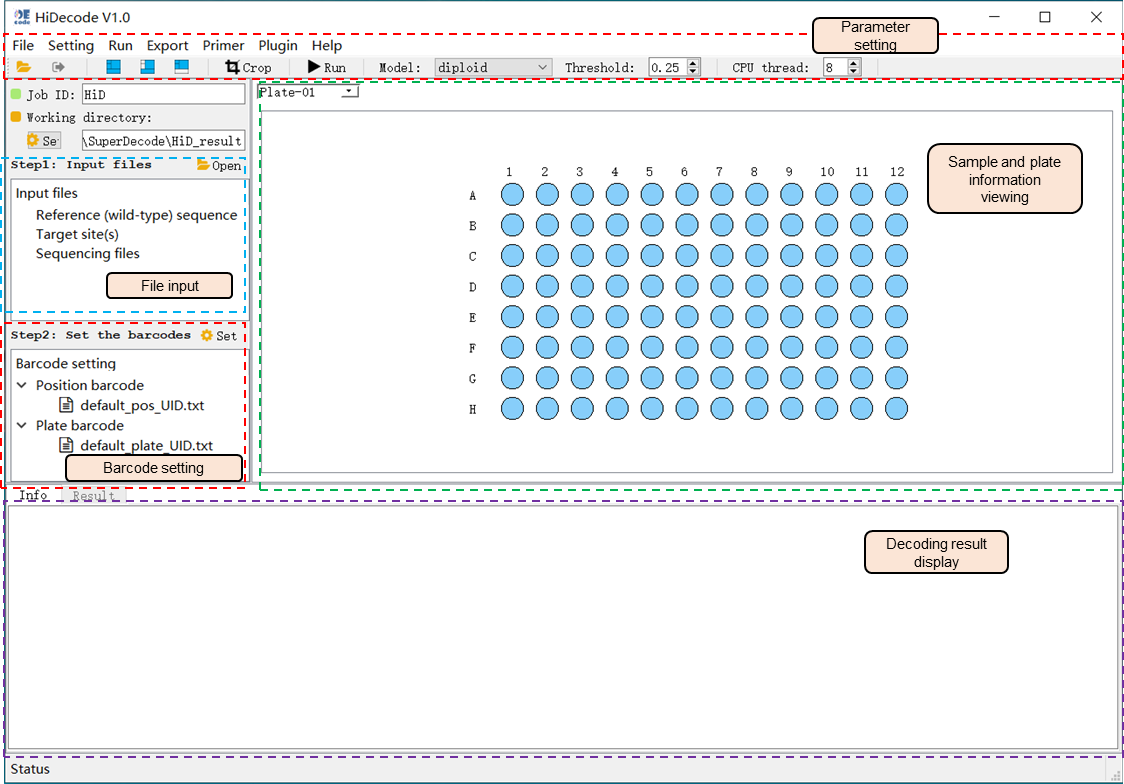


Figure 2-3 The interface of HiDecode

(2) Decoding operation of HiDecode

The use of HiDecode for decoding mutations from NGS reads of pooled amplicons is mainly includes following three steps:

i) File iniput: Click the “File” button on the toolbar or from the file input area to open the file input dialog box (Figure 2-4). Next, input the files, including: the reference file in Fasta format (covering the wild-type genome reference sequence between T#-F and T#-R), optional target sequence file (with or without PAM sequence), and sequencing file (paired-end data, which requires clean data after filtering and removing adapters). The format of reference and target site files can be viewed by clicking the provided example files.

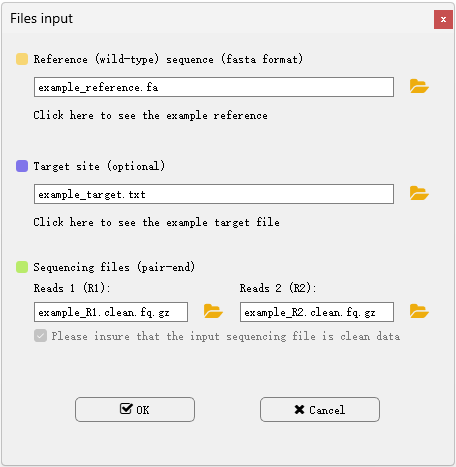


Figure 2-4 The interface of file input of HiDecode

ii) Double-click “default\_pos\_UID.txt” and “default\_plate\_UID.txt” in the barcode setting panel to select the used position barcode and plate barcode in amplification. After opening the setting box, check the barcode to be used.

iii) Set the parameters for data analysis. HiDecode provides some built-in threshold values for different analysis models, including the diploid, polyploid, and low-frequency models, based on our laboratory tests. The low-frequency model is mainly used to identify mutations in cell lines, calli, or plant protoplasts; the output is all possible variations at the threshold. In addition to the preset threshold, users can adjust the threshold as needed. When the proportion of reads of a certain variant site in the sample is greater than the set threshold, the program considers the genotype as a realiable mutation and outputs the mutation site as well the frequency.

iv) Click the “Run” button to start the analysis task.

(3) Analysis results

The analysis results of each well will be displayed in the result panel. Yellowed position(s) indicate no result was produced for the sample(s) (Figure 2-5). When a sample is selected, the mutation result will be displayed in the result panel at the bottom, including the mutation type, total number of reads covered, sequence alignment, and supported read count of each allele. By draging the slider above, user can adjust the sequence range to be displayed.

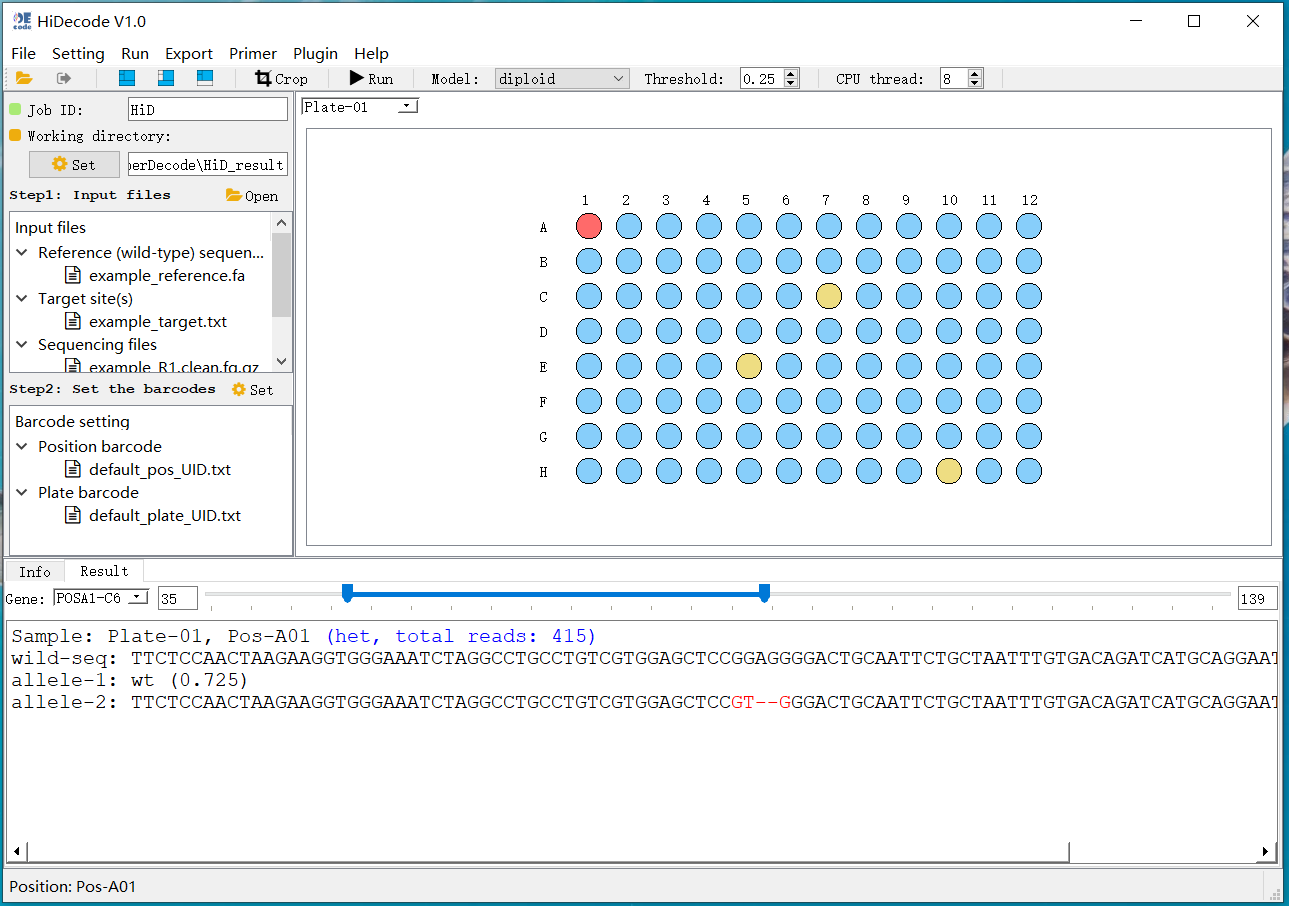


Figure 2-5 The decoding result of HiDecode

The results can be exported to an Excel file via the “Export” menu. Two types of result files can be exported. The function of “Export all results to excel file” is to export the detailed decoding results of all wells and plates (including the number of reads, mutant sequences, and the sequence alignments) (Figure 2-6), and “Fetch mutations to a single file” is to export simplified information, including mutant sequences, the count of supported reads, and mutant types.

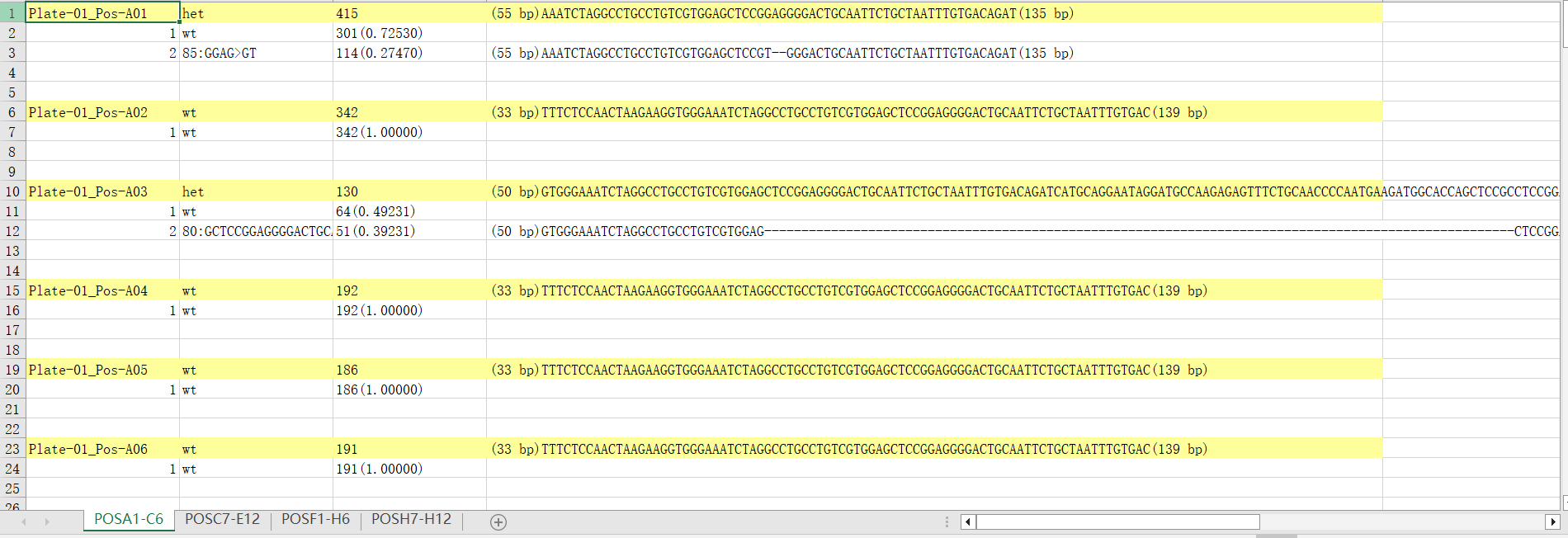


Figure 2-6 Detailed mutation result table of HiDecode

In addition, the “Plot mutations of all samples” in “Export” can output the genotype plot of all samples, allowing users to intuitively view the mutation results of the samples (Figure 2-7).

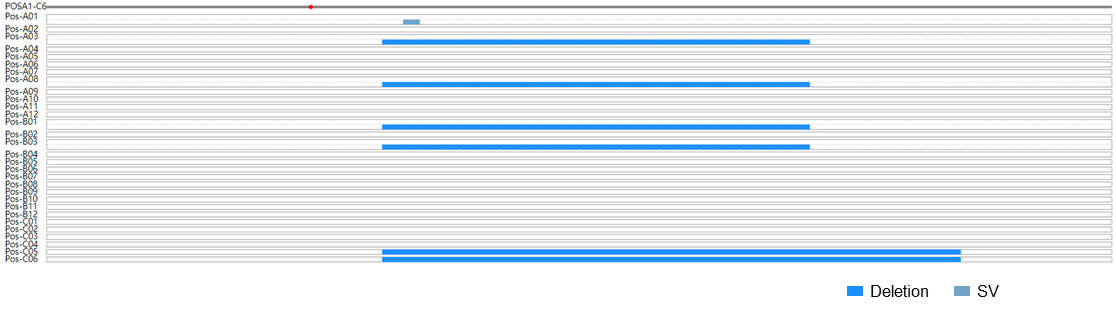


Figure 2-7 The genotype plot of each sample in HiDecode

(4) Other handy functions

i) Generation of site-specific primers

HiDecode can automatically add the bridge sequences to the site-specific primers for first-round PCR. The program allows users to modify the bridge sequences by clicking the “edit” button (Figure 2-8). After entering the bridge sequence, the name of the sample target site, and the specific sequence (5′-3′) of the target fragment, click the “Generate primers” button to generate target sequence specific amplification primers containing bridge sequences for the first-round PCR reaction. (Note: this function does not have the function of detecting the specificity of the primer. It is recommended that use other tools such as NCBI “Primer-BLAST” or CRISPR-GE “primerDesign” to detect the specificity of the primer.)

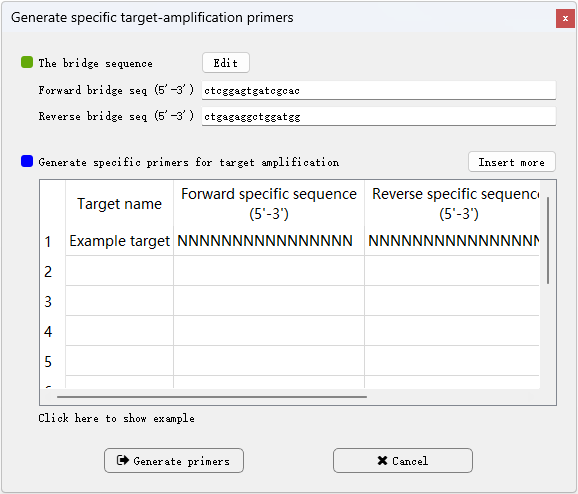


Figure 2-8 The interface of primer design

ii) Barcode setting

Insteading using the default postion and plate barcodes, users can define their own barcodes according to actual use during PCR. Open the barcode setting dialog by clicking the “Setting the default barcodes sequence” from the “Setting” menu to define new barcodes in batch. Alternatively, one by one modification is allowed via the barcode selection dock (Figure 2-9).

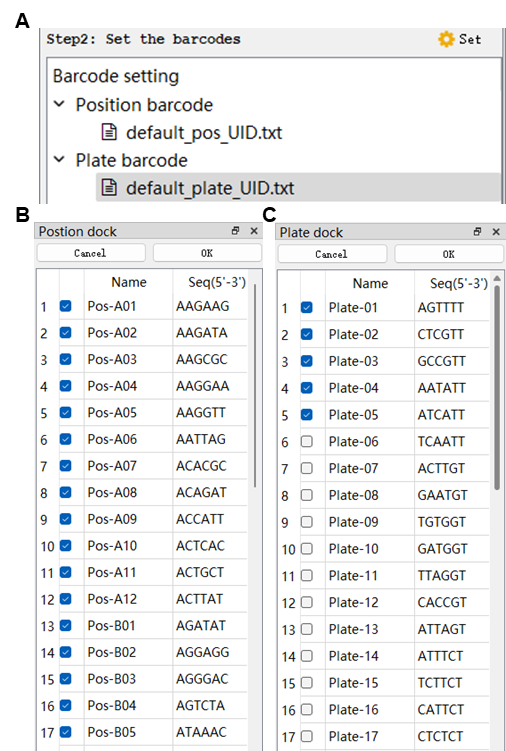


Figure 2-9 Setting the barcode sequence through the barcode selection box

ii) Setting the position and length of the barcodes in amplicons

To improve the running speed of the program, HiDecode extracts the barcode sequences according to their position in the amplicon for distinguishing different samples. If the user modifies the position of the barcodes in posN-F and plateN-R, the barcode position and length can be reset in the “Setting the barcodes position and length” dialog from the “Setting” menu (Figure 2-10).

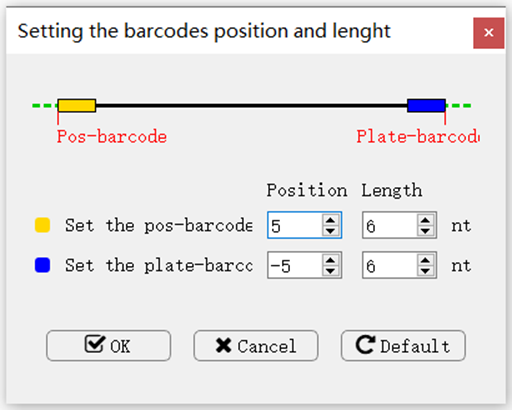


Figure 2-10 The interface of barcode position and length setting in HiDecode

iii) Removing sequencing adapters and filtering reads from raw sequencing files

Most of companies provide clean data that the sequencing adapters have been removed and low-quality rads have been filtered. In case some companies just provide raw data, HiDecode provides the function for removing adapter sequence and filtering raw quality reads via the “Quality control for sequencing files” from the “Plugin” menu (Figure 2-11). The program will automatically remove sequencing adapters and filter low-quality reads for the raw sequencing reads, and output a quality control reports for users.

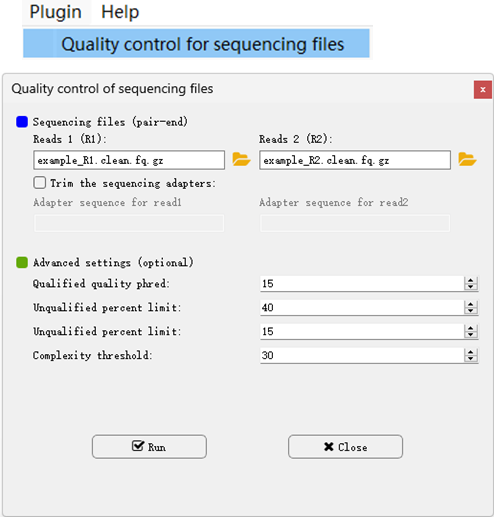


Figure 2-11 The quality control function of HiDecode for removing the sequencing adapters and low-quality reads of raw sequencing files

iv) Mutation analysis from NGS data for a single sample

In addition to handling sequencing data from pooled amplicons, HiDecode is capable of analyzing NGS data from a single sample in which the target-site(s)-containing amplicons are directly subjected to sequencing without the barcoding step using the “Mutation analysis for single sample” function via the “Plugin” menu. In the dialog (Figure 2-12), input the reference sequence, target sequence (optional), and sequencing read files for starting the analysis.

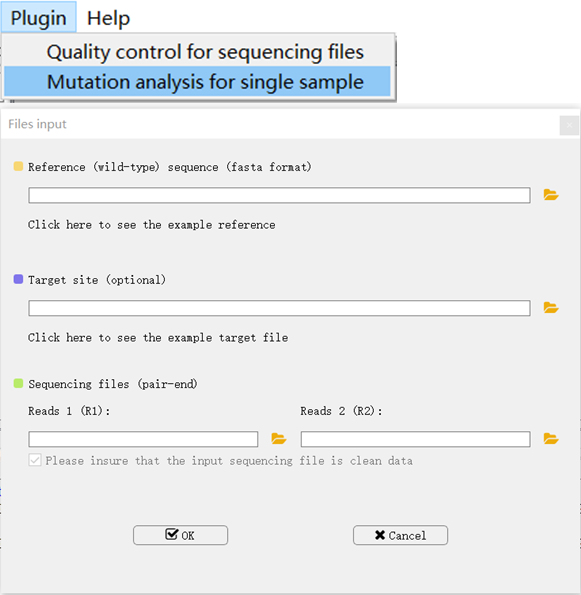


Figure 2-12 The input dialog of Mutation analysis from NGS data for a single sample

* **TGS library preparation and analysis of LaDecode**

Multiplex editing within a certain region can induce complex structural variations (SVs), such as fragment deletions, inversions, and even duplications among target sites, making it impossible to design primers to amplify each target site. LaDecode is a dedicated tool for decoding multiple and complex mutations induced by multiplex target-site editing within a certain region by long-read TGS of long-range PCR amplicons, including PacBio and Nanopore sequencing.

**1. TGS library construction of LaDecode**

The library construction procedure of LaDecode is similar to HiDecode. Through two rounds of PCR reactions, a pair of unique long-barcodes are added to the ends of amplicons (Figure 3-1). Considering the higher native error rates of Nanopore and PacBio reads, a set of longer barcodes (12 nt or longer) is introduced into the second-round PCR primers to confidently assign the reads to their template molecules for each sample. In addition, the sequencing adapter is added by sequencing company.

(1) LaDecode primer design

Similar to HiDecode library construction, library construction of LaDecode involves two pairs of primers, including site-specific primers (Target-F and Target-R) for amplifying the target sequence, and barcoding primers (posN-F and plateN-R) for labeling samples.

i) Design of site-specific primers Target-F and Target-R

The site-specific forward primer (Target-F) is designed ~100 bp upstream of the first target site, and the site-specific reverse primer (Target-R) is designed ~100 bp downstream of the last target site. The amplified fragment is controlled to be 1 to 10 kb. The site-specific pairing parts of the primers (18~20 nt, *Tm* value is 60℃~63℃) need to be checked for specificity using the Blast tool in NCBI, and then add the bridge sequence required for the second-round of PCR to the 5' end of the primer. The primer sequences are as following:

Target-F: 5'-atcgcctggctccacgctccgagttNNNNNNNNNNNNNNNNNNNN-3' (NNN... is the site-specific binding sequences, with a length of 18~20 nt and *Tm* value of 58℃~62℃; lowercase letters are the bridge sequences; the underlined sequence can complement the other end of the amplified short fragment to form a stem-loop structure to inhibit the amplification of primer-dimers and shorter non-specific products).

Target-R: 5'-tacgcctggctccacgctccctagaNNNNNNNNNNNNNNNNNNNN-3' (NNN… is the site-specific binding sequence,s with a length of 18~20 nt and a *Tm* value of 58℃~62℃; lowercase letters are the bridge sequences).

ii) Design of universal primers posN-F and plateN-R

posN-F: 5'-cttgNNNNNNNNNNNNNNNNNNNNNNNNatcgcctggctccacgctccga gtt -3' (NNN…is a 24 nt position barcode sequence, indicating the position of each sample on the 96-well PCR plate)

plateN-R: 5'- cgacNNNNNNNNNNNNtacgcctggctccacgctccctaga -3' (NNN… is a 12 nt plate barcode sequence, indicating different 96-well PCR plates, one 96-well PCR plate corresponds to one plateN-R primer)

For detailed sequences, see Appendix 2-LaDecode all primer sequences.xlsx

For convenience, the 96 position-barcoded primers can be prepared in a 96-well PCR plate and stored at -20℃ for future use.



Figure 3-1 Construction process of TGS library for LaDecode

(2) Steps of amplification

Prepare the working solution of posN-F into the concentration of 3 M, and then take 20 L of each primer into positions (from A1 to H12) of a 96-well plate. Store the plate at -20℃.

i) First-round of site-specific amplification

Use site-specific primers Target-F and Target-R to amplify the target-sites contained fragment, and then prepare the PCR reaction mixture according to the number of samples. The mixture was dispensed into 96-well PCR plates. The PCR reaction mixture is shown in Table 3-1 (the number of samples is *n*).

Table 3-1 Total reaction mixture of first-round PCR for LaDecode (15 L)

|  |  |
| --- | --- |
| 2 × Phanta Max Buffer | 7.5 μL × *n* |
| Target-F (10 μM) | 0.3 μL × *n* (0.2 μM final concentration) |
| Target-R (10 μM) | 0.3 μL × *n* (0.2 μM final concentration) |
| dNTPs Mix (10 mM) | 0.3 μL × *n* |
| Phanta Max Super-Fidelity DNA Polymerase (1U/μL) | 0.3 μL × *n* |
| ddH2O | 6.3 μL × *n* |

(Note: It is recommended to use a high-fidelity polymerase to amplify fragments larger than 2 kb)

Dispense the DNA template of each sample (20 ng/μL) into a 96-well PCR plate, and use a 96-pin replicator (or a multi-channel pipette) to add ~1 μL of sample DNA (approximately 20 ng) into the PCR reaction mixture to amplify the target amplicon of the sample. The PCR program of the PCR reaction is shown in Table 3-2.

Table 3-2 First-round PCR reaction program of LaDecode

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 95℃, 2 min |  |
| Denaturing | 95℃, 30 s | 30~32 cycles |
| Annealing | 60℃, 20 s |
| Extension | 65℃, 10 s,  68℃, 10 s,  72℃, 15 s × *n*  (Set the number of cycles *n* according to the fragment length, 1 kb/n) |
| Final extension | 70 ℃, 2 min |  |

(Note: If the sequence fragment is large and the GC content is unevenly distributed, the extension variable temperature inner cycle of the STI PCR method (Zhao et al. 2022) is used for fragment amplification)

After the first-round of PCR reaction, randomly select a few samples of the products from each plate for agarose gel (0.8%) electrophoresis to check the concentration and specificity of the amplified products. If the concentration of the amplified product is too low, 2~3 cycles can be increased of extension to increase the concentration.

ii) Second-round of barcoding PCR

The purpose of the second-round PCR reaction is to add the barcode sequence. Dilute the first round of PCR amplification products 3 times by adding 30 μL ddH2O to each first-round product. Prepare the second-round PCR total mixture according to the number of samples (Table 3-3), and dispense 15 μL of the total PCR reaction mixture into a 96-well PCR plate.

Table 3-3 LaDecode second-round PCR reaction total mixed solution system

|  |  |
| --- | --- |
| 2 × Phanta Max Buffer | 7.5 μL × *n* |
| plateN-R (10 μM) | 0.3 μL × *n* (0.2 μM final concentration) |
| dNTPs Mix (10 mM) | 0.3 μL × *n* |
| Phanta Max Super-Fidelity DNA Polymerase (1U/μL) | 0.3 μL × *n* |
| ddH2O | 6.8 μL × *n* |

Use a 96-pin replicator (or multichannel pipette) to add ~1 μL of posN-F (3 μM) from the prepared 96-well plate and add it to the PCR reaction mixture (final concentration 0.2 μM), and add ~1 μL of the diluted first-round PCR product to the PCR reaction mixture (final dilution is about 45 ×). The PCR reaction program is shown in Table 3-4.

Table 3-4 Second-round PCR reaction program of LaDecode

|  |  |  |
| --- | --- | --- |
| Pre-denaturing | 95℃, 2 min |  |
| Denaturing | 95℃, 30 s | 15~18 cycles |
| Annealing | 60℃, 30 s |
| Extension | 65℃, 10 s,  68℃, 10 s,  72℃, 15 s × *n*  (Set the number of cycles *n* according to the fragment length, 1 kb/n) |
| Final extension | 70 ℃, 2 min |  |

After the second-round of PCR reaction, randomly select a few samples of the products from each plate for agarose gel (0.8%) electrophoresis for checking. If the concentration of the amplified product is low, 2~3 cycles can be increased of extension to increase the concentration.

Take 3 μL of the product of each sample from a same plate to be mixed as a group. The PCR products are purified using 0.8% agarose gel electrophoresis and a DNA purification kit. Various types of mutation are might be generated by multiplex editing, including large fragment deletions. In order to ensure that all fragments are collected, the time of electrophoresis should not be too long. The primary purpose of electrophoresis is to remove primers and products that are not fully extended. The products of each group are then mixed in equal amounts and are subjected to TGS. The concentration is required to be >20 ng/μL, and the total amount is >2 μg.

**2. TGS sequencing**

The purified products are sent to the company for TGS library construction (including adding sequencing adapters and purification) and sequencing. It is recommended to use Pacbio HiFi sequencing platform for the sequencing. The sequencing company will provide the clean TGS read file(s) in Fasta, Fastq, or bam format that the sequencing adapter have been removed. The amount of sequencing data is determined by the fragment size and the number of samples. It is recommended that 1-4 kb, 4-7 kb and 7-10 kb correspond to ~1 Gb, 2 Gb and 3 Gb of data/96 samples respectively.

**3. Mutation analysis using LaDecode**

(1) The interface of LaDecode

The main interface of HiDecode includes the menu, toolbar, file input panel, plate canvas, and decoding results panel (Figure 3-2). The menu bar provides commonly used setting functions and plugin. The bartool provides some setting options such as setting mutation threshold and CPU thread number. The file input module is used to input wild-type reference sequence file, target sequence file, and paired-end sequencing files. In the barcode settings, user can select the used position- and plate-barcodes. The plate canvas is used to interactively display the analysis results of each sample.

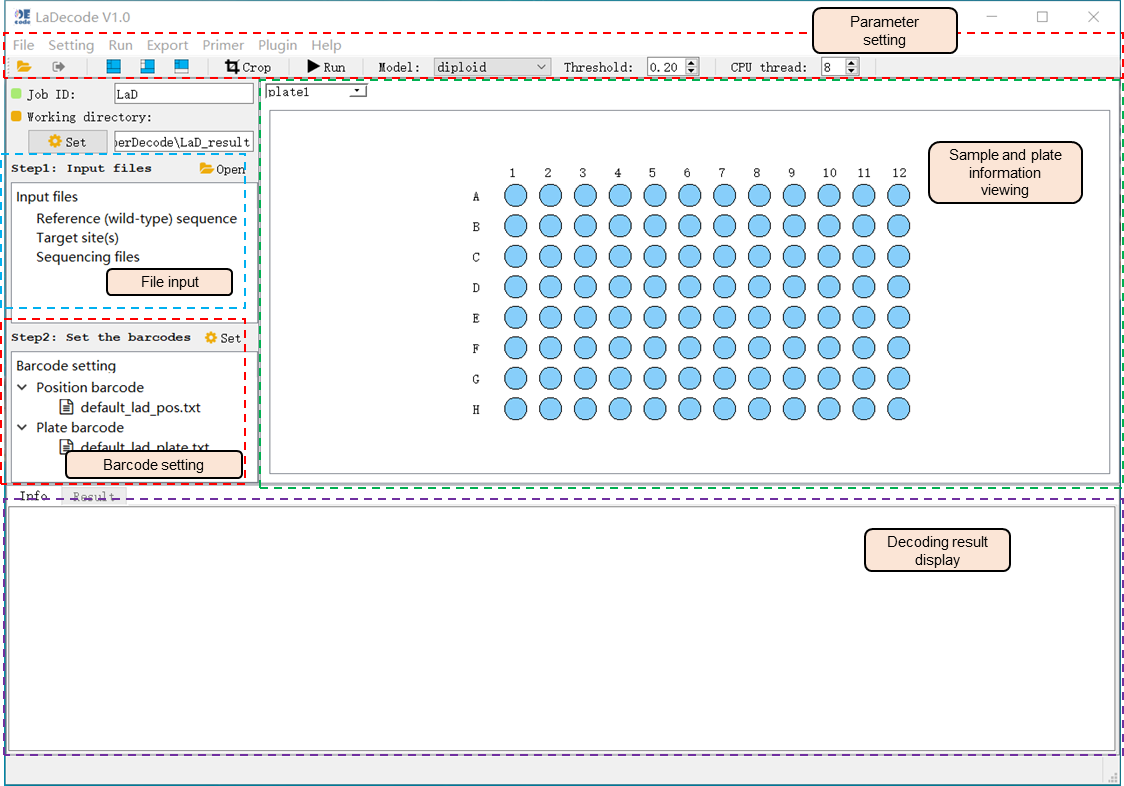


Figure 3-2 The interface of LaDecode

(2) Decoding operation of LaDecode

The use of LaDecode mainly includes three main steps: file input, parameter setting and decoding. The detailed operations are as follows:

i) Open the file input dialog box through the menu bar or the “Open” button in the file input panel, and enter the files required for analysis, including the reference sequence file in Fasta format (covering the wild-type genome sequence between Target-F and Target-R), optional target sequence file (with or without PAM sequence), and sequencing file (Fasta, Fastq, or bam format file) (Figure 3-3). The format of reference and target site files can be viewed by clicking the provided example files.

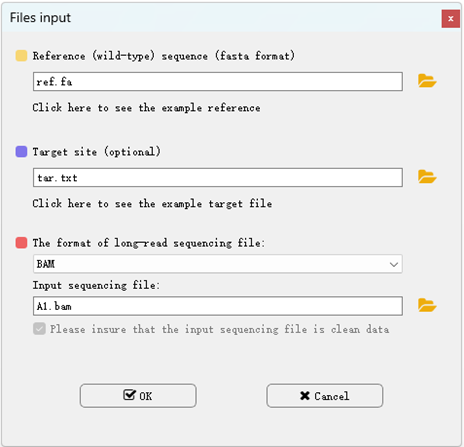


Figure 3-3 The interface of file input in LaDecode

ii) Double-click the “default\_lad\_pos.txt” and “default\_lad\_plate.txt” in the barcode setting panel to select the used position barcode and plate barcode.

iii) Set the parameters for data analysis. LaDecode provides some built-in threshold values for different analysis models, including the diploid, polyploid, and low-frequency models, based on our laboratory tests. The low-frequency model is mainly used to identify mutations in cell lines, calli, or plant protoplasts; the output is all possible variations at the threshold. In addition to the preset threshold, users can adjust the threshold as needed. When the proportion of reads of a certain variant site in the sample is greater than the set threshold, the program considers the genotype as a realiable mutation and outputs the mutation site as well the frequency.

iv) Click the “Run” button on the toolbar or from “Run” menu to start the analysis task. It will take a few minutes to run the decoding.

(3) Analysis results

The analysis results of each well will be displayed in the result panel. Yellowed position(s) indicate no result was produced for the sample(s) (Figure 3-4). When a sample is selected, the mutation result will be displayed in the result panel at the bottom, including the mutation type, total number of reads covered, sequence alignment, and supported read count of each allele. By draging the slider above, user can adjust the sequence range to be displayed.

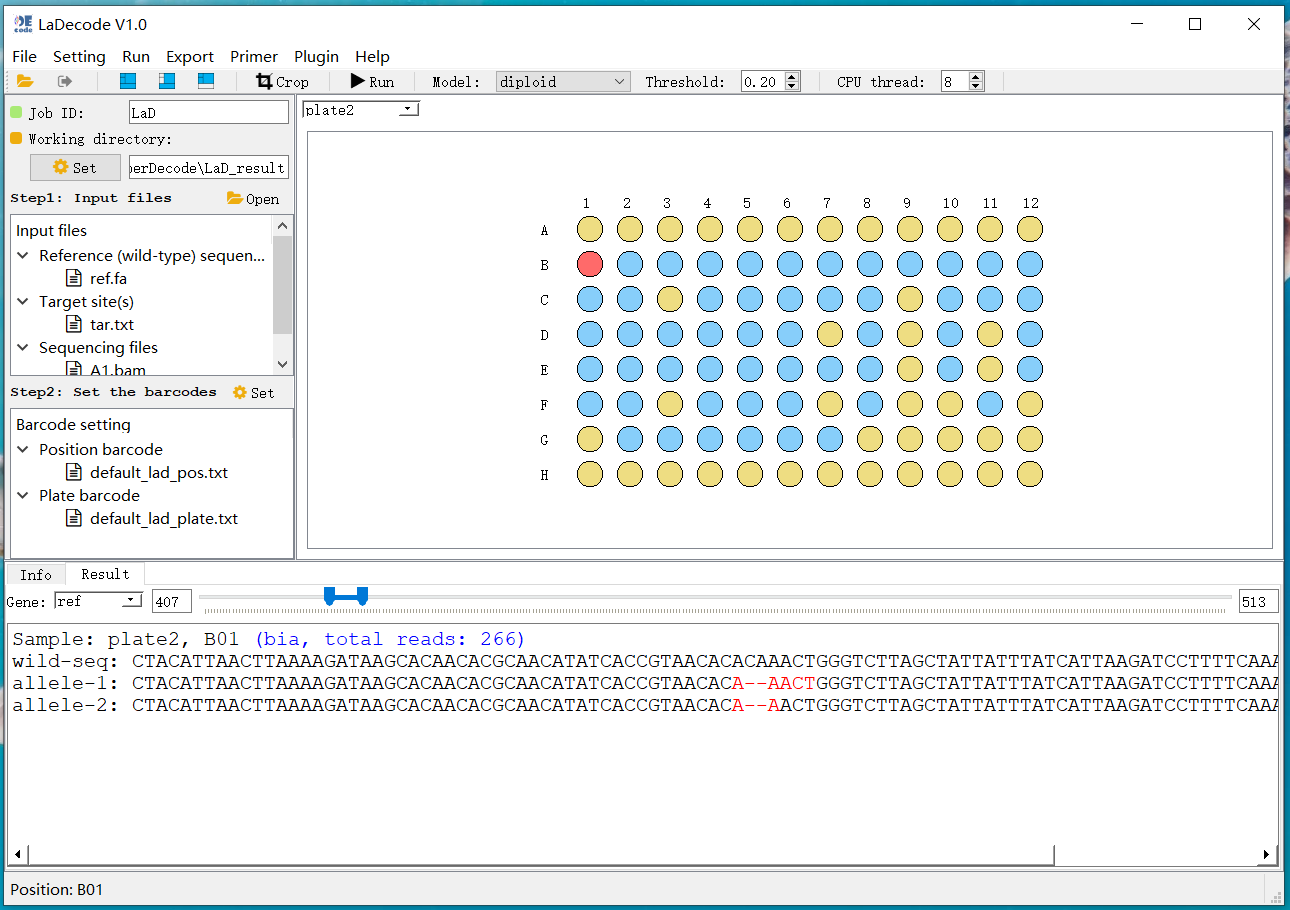


Figure 3-4 The result of LaDecode

The results can be exported to an Excel file via the “Export” menu. Two types of result files can be exported. The function of “Export all results to excel file” is to export the detailed decoding results of all wells and plates (including the number of reads, mutant sequences, and the sequence alignments) (Figure 3-5), and “Fetch mutations to a single file” is to export simplified information, including mutant sequences, the count of supported reads, and mutant types.



Figure 3-5 Detailed mutation result table of LaDecode

In addition, the “Plot mutations of all samples” in “Export” can output the genotype plot of all samples, allowing users to intuitively view the mutation results of the samples (Figure 3-6).

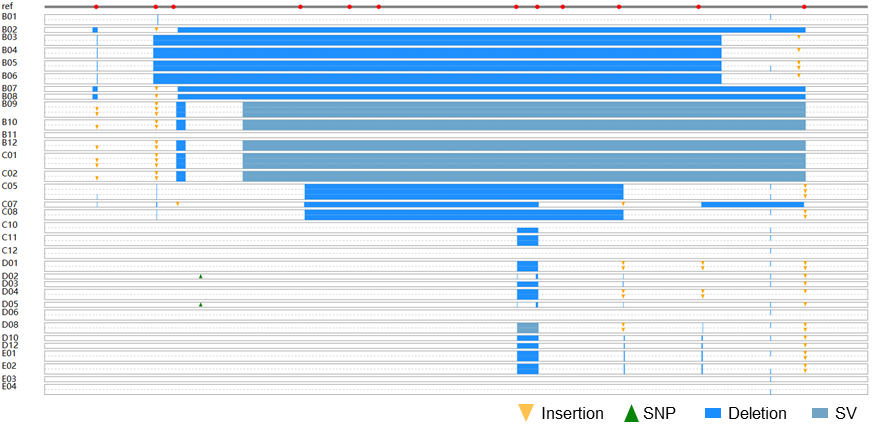


Figure 3-6 The genotype plot of each sample in LaDecode

(4) Other functions

i) Mutation analysis from demultiplexed sequencing reads

Sometimes, sequencing companies can provide demultiplexed sequencing reads from pooled amplicon samples according to the user’s barcode sequence. To analyze the mutations directly from the demultiplex files, LaDecode provides a plugin tool “Muataion analysis from demultiplexed fasta” from the “Plugin” menu. In the opened dialog, enter the wild-type reference sequence file, the optional target sequence file, and the directory path of the demultiplexed file (usually in Fasta format). Click the “OK” button for decoding analysis (Figure 3-7). The output result file is automatically saved to the same directory of the read files.

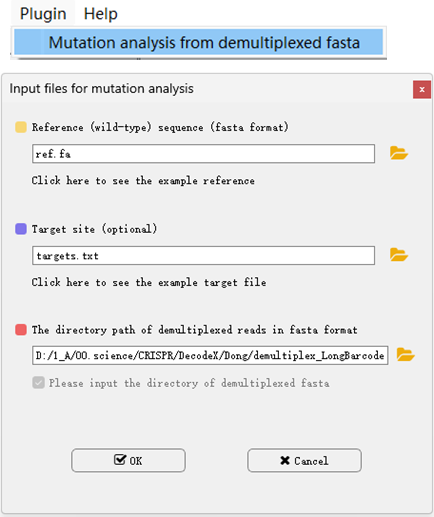


Figure 3-7 The interface of demultiplexed data analysis

ii) Generation of site-specific primers

HiDecode can automatically add the bridge sequences to the site-specific primers for first-round PCR. The program allows users to modify the bridge sequences by clicking the “edit” button (Figure 3-8). After entering the bridge sequence, the name of the sample target site, and the specific sequence (5′-3′) of the target fragment, click the “Generate primers” button to generate target sequence specific amplification primers containing bridge sequences for the first-round PCR reaction. (Note: this function does not have the function of detecting the specificity of the primer. It is recommended that use other tools such as NCBI “Primer-BLAST” or CRISPR-GE “primerDesign” to detect the specificity of the primer.)

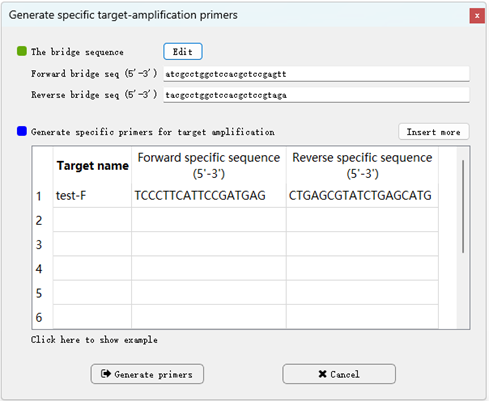


Figure 3-8 The interface of primer design in LaDecode

iii) Barcode setting

Insteading using the default postion and plate barcodes, users can define their own barcodes according to actual use during PCR. Open the barcode setting dialog by clicking the “Setting the default barcodes sequence” from the “Setting” menu to define new barcodes in batch. Alternatively, one by one modification is allowed via the barcode selection dock (Figure 3-9).

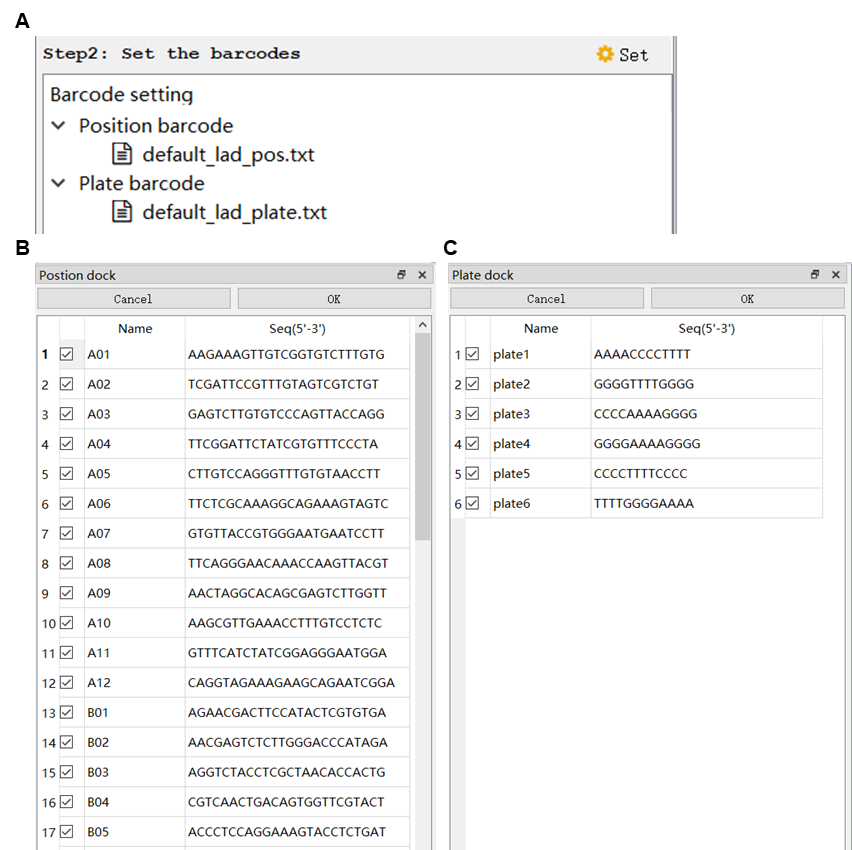


Figure 3-9 Barcode setting interface in LaDecode

* **Contact**

We highly appreciate your valuable suggestions for improving the experimental procedures and software packages. If you have any questions or suggestions, please contact Xianrong Xie (xiexianrong@scau.edu.cn)