README_ProbeDealer

1. **System requirements**
   The ProbeDealer MATLAB application requires the installation of BLAST (version 2.10.1) and MATLAB (version R2019a or a more recent version) software on the computer. The ProbeDealer application also requires the MATLAB Bioinformatic Toolbox. The ProbeDealer application was developed with MATLAB version R2019a.

2. **Installation guide**
   Install BLAST
   Windows users should download and install ncbi-blast-2.10.1+-win64.exe from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. Mac users should download and install ncbi-blast-2.10.1+.dmg from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. Mac users then need to add the BLAST executable bin file to the PATH by following these instructions:

   In Terminal, type: sudo nano /etc/paths
   enter password
   scroll down to the bottom of the list
   add the path: /usr/local/ncbi/blast/bin
   exit and save the list
   restart the terminal

   Mac users may use the following command in Terminal to verify BLAST installation:

   type: echo $PATH
   expected outcome:
   /usr/local/bin:/usr/bin:/bin:/usr/sbin:/sbin:/usr/local/ncbi/blast/bin
   type: which blastn
   expected outcome: /usr/local/ncbi/blast/bin/blastn

   Typical install time on a “normal” computer takes several minutes.

   Install ProbeDealer
   ProbeDealer provides MATLAB add-on applications for Windows and Mac users and a standalone application for Windows users, available for download from https://campuspress.yale.edu/wanglab/probedealer.

To install the ProbeDealer MATLAB application, open MATLAB, and click “Install App” in APP tab. Select a .mlappinstall file according to your operating system. To use the ProbeDealer standalone application on Windows, install MATLAB Runtime version 9.8 (R2020a), and double click ProbeDealer_Win.exe to execute the standalone application.

Users of ProbeDealer MATLAB application need to install the MATLAB Bioinformatic Toolbox in order to use ProbeDealer.

Typical install time on a “normal” computer takes less than one minute.

3. Demo and instructions

3.1 Provide BLAST genome database

Input path to genome file folder containing necessary BLAST database subfolders. This path should not contain spaces.

For chromatin tracing probes, provide the genome file folder that contains the following sub-folder: “genome”, for the genome BLAST database. To enable the “only target antisense of genes” feature, also include a sub-folder “UnsplicedTx” as the unspliced transcriptome database. To enable the “avoid exon regions” feature, include another sub-folder “TxShortHeader” as the transcriptome database.

For sequential single-molecule RNA FISH and MERFISH probes, provide the genome file folder path that contains the following sub-folders: “TxShortHeader”, as the transcriptome database, and “Tx” which contains the Gencode transcriptome fasta file with full headers to match transcripts with their genes. Note that for all probe types, the sub-folders should be named as indicated in this manual.

We provide example “genome”, “UnsplicedTx”, “TxShortHeader”, and “Tx” subfolders that contain human and mouse databases at https://campuspress.yale.edu/wanglab/probedealer/humandatabase and https://campuspress.yale.edu/wanglab/probedealer/mousedatabase, respectively. Users may download these two folders, and use the folder paths as inputs for ProbeDealer to design human and mouse probes.

For additional user-defined databases, we recommend using UCSC genome fasta files to generate BLAST databases and UCSC GTF files to generate unspliced transcriptomes. ProbeDealer only accepts Gencode transcriptome files.

3.2 Choose probe type

Select a probe type in the application panel. Chromatin tracing users may choose additional features according to their needs. Users may also specify their preferred oligo parameters by editing oligoparameters.xlsx provided in the ProbeDealer package. For example, to allow designed probes to overlap each other

by up to 20 nucleotides, as previously demonstrated in chromatin tracing to trace a gene *cis*-regulatory region at 5-kb genomic resolution[3], users may change ProbeGap in oligoparameters.xlsx from 31 to 11. This overlapping design allows more probes to be designed for short genomic regions or RNA species.

oligoparameter.xlsx includes the following parameters that can be customized:

ProbeLength: desired length of oligos. Default is 30 nt.

MinTm: minimum Tm of oligos. Default is 66 ℃.

MaxTm: maximum Tm of oligos. Default is 100 ℃.

SecondaryStructureTm: maximum Tm of concatenated stems in the oligo. Default is 76 ℃.

CrossHybTm: maximum Tm of concatenated cross-hybridization regions in the oligo. Default is 72 ℃.

MinGC: minimum percentage of GC of oligos. Default is 30%.

MaxGC: maximum percentage of GC of oligos. Default is 90%.

ExcludeSeq: sequences that should be avoided in oligos; excluded sequences are separated by "|". Default is GGGGGG|CCCCCC|TTTTTT|AAAAAA.

ProbeGap: minimum distance between the 5' end of two adjacent oligos. Default is 31, for generating oligos of 30-nt with no overlap.

3.3 Provide target sequences

ProbeDealer accepts two target input file formats: fasta file (.fasta, .fa, .fas) and spreadsheet file (*.xls, *.xlsx, *.csv) without headers. Chromatin tracing and sequential single-molecule RNA FISH accept both types of input, while MERFISH only accepts spreadsheet files. Example target input files are available for download from https://campuspress.yale.edu/wanglab/probedealer/exampletargetfiles.

The spreadsheet file for chromatin tracing should contain three columns. The first column indicates the chromosome of current target sequence, e.g. chr1 (note that "chr" should be present). The second and third column indicates the start and end points of the target sequence on the indicated chromosome, and coordinates should start from 1, inclusive (i.e. same as UCSC GTF coordinates, not UCSC BED coordinates). As a simplified example, for a sequence on a chromosome:

>chr1

ATCTATTTGGGCG

To design chromatin tracing probe for TATTT, the three columns should be:

chr1 4 8

To use our default human and mouse databases, the genome coordinates need to be from hg38 and mm10, respectively.

For RNA FISH, ProbeDealer will draw target information from the first column of the input spreadsheet. The first column should contain the Ensembl transcript IDs

(without version) of target sequences (e.g. ENST00000544455, ENSMUST00000106216).

The spreadsheet file for RNA MERFISH should contain at least two columns. The first column should contain the Ensembl transcript IDs (without version) of target sequences, and the second column should contain the corresponding gene FPKM values or other measures of relative transcript expression levels.

3.4 Secondary sequence choices

ProbeDealer allows freeform appending of customizable secondary sequences to the probes. For the convenience of users, we provide 50 default secondary sequences for chromatin tracing (DNA secondaries.xlsx) and 16 default secondary sequences for sequential single-molecule RNA FISH (RNA secondaries.xlsx, Sheet "sequential RNA FISH"). We also provide 16 default secondary sequences for RNA MERFISH (RNA secondaries.xlsx, Sheet "MERFISH"). Users can freely edit the provided secondary sequences in the spreadsheets.

Users should ensure the number of input sequences does not exceed the number of available secondary sequences (or secondary sequence combinations in MERFISH) in the provided spreadsheets. Using the default secondary sequences provided in the spreadsheets, the chromatin tracing target number should be ≤ 50, and the sequential single-molecule RNA FISH target number should be ≤ 16. If needed, users can provide more secondary sequences in the spreadsheets for more targets. By default, MERFISH accepts up to 140 target sequences.

In chromatin tracing and sequential single-molecule RNA FISH, users may also append different secondary sequences to the 5' end and 3' end of the primary targeting region of the template probe, or only append to one end. The "DNA secondary.xlsx" file and the sheet named "sequential RNA FISH" of the "RNA secondaries.xlsx" file each contain two columns named "sequence at 5' end" and "sequence at 3'end". The sequences in these two columns correspond to sequences to be appended to the 5' end and 3' end of the primary targeting regions in the template probe library. By default, the sequences provided in the two columns are the same, so that the same secondary sequences are appended to both ends, but users may offer different sequences for the two ends by directly editing the sequence entries in the spreadsheets. Users may also opt to append secondary sequences to only one end by deleting sequence entries in the other column and leaving them blank.

The default RNA MERFISH design generates 140 combinations (codes) of the 16 secondary sequences to encode 140 RNA targets. To use other (expanded) codebooks for RNA or DNA MERFISH (or seqFISH) designs, users may generate the corresponding secondary sequence combinations according to their codebooks, and use the "chromatin tracing" or "sequential RNA FISH" option to append each

secondary sequence combination to the corresponding primary targeting region designs for each target.

3.5 Output

Users may also specify how many probes they want for each input sequence, or choose to retain all probes. If an input sequence does not have as many probes as specified by the user, some of its probes will be duplicated to meet the requirement. We recommend at least 48 probes for each RNA target in MERFISH, at least 36 probes for each RNA target in sequential single-molecule RNA FISH, and at least 150 probes for each genomic target in chromatin tracing[3].

Choose output type and provide a path for output files. Output files include FinalOligos.fasta and FinalOligos.xlsx. If some sequences do not have as many probes as specified by the user, a log file will be generated to record those input sequences. A codebook will be generated as an Excel spreadsheet for MERFISH users to match each transcript with its MHD4 code.

Expected run time for a single gene and a 140-gene MERFISH library takes around 2 min and 15 min respectively.