RNA Imaging with MERFISH - Design of Oligonucleotide Probes

Jeffrey R. Moffitt,Xiaowei Zhuang

Abstract

The first step in any MERFISH experiment is the design of the oligonucleotide probes that will be used to label individual RNA species. In our current implementation of MERFISH, each oligonucleotide encoding probe consists of three basic components as illustrated in Figure 2. The first region is a 30-nt targeting region that is complementary to a portion of the sequence of the RNA to which it is designed to bind. The second region is a set of sequences that are called readout sequences, which were designed to be complementary and hence only bind to MERFISH readout probes and not other nucleic acid in the cell. Finally, the third region is a set of priming regions used in the construction of these probes, which will be discussed in detail in Probe Construction. In addition to the nucleotide sequences for each of these components, a codebook —the specific set of binary barcodes that will be used and their association with different RNA species of interest—must also be designed. In this section, we provide protocols to design these sequences and to build a codebook. Example code to perform these steps can be found at http://zhuang.harvard.edu/merfish/ .

Citation: Jeffrey R. Moffitt,Xiaowei Zhuang RNA Imaging with MERFISH - Design of Oligonucleotide Probes. **protocols.io** dx.doi.org/10.17504/protocols.io.menc3de
Published: 29 Mar 2018

Guidelines

Design of Target Regions

Functionally, the goal of a target region is to direct the binding of each encoding probe to its target RNA of interest with high binding efficiency and specificity. The central challenge in the design of target regions for MERFISH (and smFISH, in general), is to design a set of target regions where these properties are optimized for all probes under a constant set of hybridization conditions, e.g. incubation temperature. To accomplish this goal target regions are designed to cover a relatively narrow range of GC content and melting temperatures (TM) with their target. In addition, a good target region should have limited homology to other RNAs in the transcriptome, reducing the probability that it will bind to the wrong RNA. Finally, the typical smFISH measurement does not bind each RNA with a single probe but rather tiles that RNA with multiple probes, each of which targets a different portion of the sequence of the RNA. See Figure 2. For many smFISH measurements, the

number of unique probes per RNA is often ~50; however, this number can be lowered with a corresponding reduction in the brightness of the individual RNA spots (Raj et al., 2008). For initial MERFISH work, we recommend having at least 50 encoding probes per RNA.

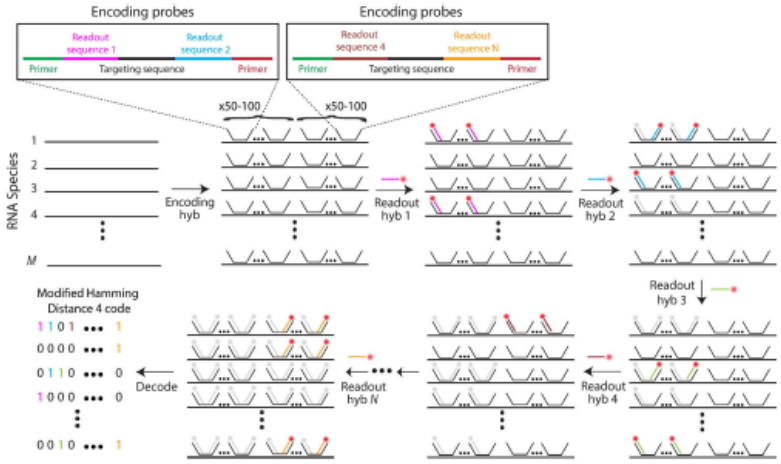

Figure 2.
Schematic depiction of the hybridization process used for MERFISH. Cellular RNAs are hybridized with a set of oligonucleotide probes, which we term *encoding probes*. These encoding probes contain a targeting sequence which directs their binding to the specific RNA. They also contain two readout sequences. For an experiment utilizing N-bit binary barcodes, N different readout sequences will be used with each bit assigned a different unique readout sequence. The specific readout sequences contained by an encoding probe to a given RNA are determined by the binary barcode assigned to that RNA: only the readout sequences assigned to bits for which this barcode contains a `1' are used. Each encoding probe also contains PCR priming regions used in its construction. To increase the signal from each copy of the RNA, multiple encoding probes, each with a different target region, are bound to the same RNA. The length of this *tile* of probes is typically between 50–100 probes. To identify the readout sequences contained on the encoding probes bound to each RNA, N rounds of hybridization and imaging are performed. Each round uses a unique, fluorescently labeled probe whose sequence is complementary to the readout sequence for that round. The binding of these fluorescent probes determines the bits which contain a `1', allowing the measurement of the specified binary code. Modified with permission from (Chen et al., 2015).

As described by others (Raj et al., 2008), smFISH probes can be designed online using a web interface (https://www.biosearchtech.com/support/tools/design-software/stellaris-probedesigner). In principle, probes designed with this software should work well as target regions for MERFISH. However, there is no batch processing option for such software and submitting hundreds to thousands of genes individually would be very labor intensive. Thus, we describe the alternative approach that we have used.

We design our target regions with the software package, OligoArray 2.0, which was developed for the design of microarrays (Rouillard, 2003). This software has been used previously in the design of FISH

probes for DNA (Beliveau et al., 2012). In addition to the ability to batch process mRNAs, one advantage of this software is the additional stringency OligoArray applies to the design of its probes. Specifically, internal secondary structure and off-target binding are assessed not via the number of complementary bases but by the thermodynamic stability of these structures. Of course, these extra calculations come at a computational cost, and calculation of target regions for a large number of genes can take days of computing resources on a desktop computer.

Design of Readout Probes

There are several important considerations in the design of readout probes. First, to improve the binding efficiency of these probes, it is desirable to select probes that have similar TM and GC content so that their hybridization properties are similar under a given hybridization condition. Second, to limit the number of potential off-target binding sites, potential sequences should be screened for homology to RNAs in the transcriptome of interest. Third, these sequences must be orthogonal, in that they should have limited homology with one another to prevent binding of one readout probe to the wrong readout sequence.

We have already screened and validated several readout probes for human samples, and we recommend the use of these sequences, which are provided in Table 1. The steps can be taken if new or additional readout probes are required.

Table 1

The sequence of the readouts probes that have been validated.

CGCAACGCTTGGGACGGTTCCAATCGGATC	CGCGAAATCCCCGTAACGAGCGTCCCTTGC
CGAATGCTCTGGCCTCGAACGAACGATAGC	GCATGAGTTGCCTGGCGTTGCGACGACTAA
ACAAATCCGACCAGATCGGACGATCATGGG	CCGTCGTCTCCGGTCCACCGTTGCGCTTAC
CAAGTATGCAGCGCGATTGACCGTCTCGTT	GGCCAATGGCCCAGGTCCGTCACGCAATTT
GCGGGAAGCACGTGGATTAGGGCATCGACC	TTGATCGAATCGGAGCGTAGCGGAATCTGC
AAGTCGTACGCCGATGCGCAGCAATTCACT	CGCGCGGATCCGCTTGTCGGGAACGGATAC
CGAAACATCGGCCACGGTCCCGTTGAACTT	GCCTCGATTACGACGGATGTAATTCGGCCG
ACGAATCCACCGTCCAGCGCGTCAAACAGA	GCCCGTATTCCCGCTTGCGAGTAGGGCAAT

Design of the Codebook

Before readout sequences and target sequences can be assembled to form the sequence of encoding probes, a codebook must be designed. The first step in codebook design is the choice of an encoding scheme that is appropriate for the experimental goals. We have utilized two different encoding schemes for MERFISH. The first scheme is a constant Hamming Weight code generated by keeping all binary barcodes from an existing HD4 code known as the Extended Hamming Code that contain only four `1' bits. We call this code the Modified Hamming Distance 4 code (MHD4). The 16-bit version of this encoding scheme contains 140 binary barcodes and is capable of identifying all two-bit errors and correcting any individual single-bit error. Alternatively, if the error correcting abilities of this code are not utilized, it can also detect three-bit errors. The advantage of this encoding scheme is the very high calling rate, ~80%, and low misidentification rates that it provides, as discussed in the full manuscript. In the second encoding scheme, we utilized a constant Hamming Weight code generated from all possible combinations of four `1' bits. This code has a minimum HD of 2 between all barcodes, so we refer to it as a Modified HD 2 code (MHD2). The benefit of this encoding scheme is that the reduced HD allows a much higher level of multiplexing for a given number of a bits—a 14-bit code contains 1,001 barcodes. However, the reduced HD produces a lower calling rate and a slight reduction in accuracy. Again, see the attached manuscript.

Here we describe algorithms to generate both encoding schemes; however, we would encourage users to consider alternative encoding schemes, e.g. versions of those provided below with fewer or more bits or schemes with different Hamming Weight or HD, in order to find a scheme that best suits their experimental needs. Again, examples of algorithms to generate these codes as well as the barcodes for the published 16-bit MHD4 and 14-bit MHD2 codes are provided at: http://zhuang.harvard.edu/merfish/.

Generation of the MHD4 Encoding Scheme—Here we present a method for generating sets of constant-weight-4 HD-4 (MHD4) barcodes for any number of bits. As an interesting aside, we note that the task of creating the optimal constant weight coding scheme — the encoding scheme that has the maximum possible number of barcodes given a specific HD — remains an unsolved problem. Thus, the algorithm that we present here likely does not produce a constant weight HD-4 code with the maximum possible number of barcodes given a specific number of bits. However, comparison of estimates of the upper bound on the number of barcodes possible in such codes (Brouwer and Etzion, 2011) and the number of barcodes generated by this algorithm suggests that it performs fairly well, typically producing ~75% of the possible number of barcodes. The 16-bit MHD4 code that we have produced is the rare exception: It contains 140 barcodes which is the current known upper limit for an MHD4 code of this length (Brouwer and Etzion, 2011).

Step 1: Generate the Extended Hamming Code that contains the desired number of bits.

Details on how to generate this code are discussed many places online, and software to

perform this calculation is included at http://zhuang.harvard.edu/merfish/.

Step 2: Select only barcodes with the desired Hamming Weight.

Generation of the MHD2 Encoding Scheme—Constant weight HD2 codes are straightforward to generate, and the algorithm that we provide below will produce the optimal code, i.e. the maximum number of barcodes with four `1's and a minimum HD of 2.

Step 1: Generate a barcode of the desired length with only four `1's.

Step 2: Generate all possible permutations of the bits in this barcode.

Barcode Assignment—Once the desired barcodes have been designed, the codebook is designed by randomly associating each barcode with a specific RNA of interest. We recommend leaving 5-10% of the possible barcodes unassigned. These `blank' or `control' barcodes will serve as internal controls and will provide estimates of the frequency with which barcodes can be generated by background or spurious signals. The specific use of these `blank' measurements in discussed in the attached full manuscript.

Assembly and Screening of Encoding Probes

Once the target regions, the readout regions, and the barcodes associated with the desired encoding scheme are designed, the sequence of the encoding probes can be assembled. Each encoding probe will contain multiple readout sequences. However, given length restrictions in synthesis of these sequences, it is typically not the case that all of the readout sequences for each RNA will fit into each encoding probe. We use two readout sequences per encoding probe.

Design of Priming Regions

The protocol that we use to make encoding probes that contain the sequences designed above require the addition of two priming regions to each probe. The optimal regions should have similar TM , no contiguous stretches of the same nucleotide longer than three, and relatively narrow GC content. They should also have limited homology to each other and to non-priming regions of the encoding probes.

Protocol

Design of Target Regions

Step 1.

Download and install all necessary software. OligoArray2.0 can be downloaded from:

http://berry.engin.umich.edu/oligoarray2_1/ .

This software requires OligoArrayAux which can be downloaded from:

http://unafold.rna.albany.edu/q=DINAMelt/OligoArrayAux.

OligoArray2.0 will look for this software in a specific directory (C:\Program Files \OligoArrayAux\), so it must be installed there. Finally, OligoArray2.0 also requires several legacy BLAST functions, which can be downloaded and installed from: http://
blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download .

Design of Target Regions

Step 2.

Download a fasta file containing the transcriptome of interest. These files can be found from a variety of repositories, such as those hosted by NCBI, UCSC, or Ensembl. For example, the human transcriptome can be downloaded from Ensembl using the cDNA link found on this page: http://www.ensembl.org/info/data/ftp/index.html.

Design of Target Regions

Step 3.

Create a BLAST database of the transcriptome for OligoArray2.0 to identify potential off-target binding. Instructions on how to create this database using the legacy BLAST functionality can be found here:

http://blast.ncbi.nlm.nih.gov/Blast.cgiCMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download.

Design of Target Regions

Step 4.

Run OligoArray2.0 on the desired transcripts. OligoArray2.0 is a java based program and a full description of its use can be found here: http://berry.engin.umich.edu/oligoarray2_1/.

When run, it must be supplied with a variety of parameters that place limits on the properties of the designed target regions, e.g. the length of possible probes, the suitable GC and melting temperature ranges, as well as the maximum T_M for potential off-target sequences. For the design of our target regions, we have used the following parameters: target region length of 30 nt; theT_M of the properly hybridized probe greater than 70°C; a lower bound on theT_M of hybridization to potential off-target sequences of 72°C; no internal secondary structure with a T_M lower than 76°C; and no contiguous run of the same nucleotide longer than six. These parameter ranges were selected to balance the demands for high stringency of probe binding with the design of enough distinct target regions to

label each RNA.

Design of Target Regions

Step 5.

Parse the OligoArray output. OligoArray2.0 generates an output file, the details of which are described here: http://berry.engin.umich.edu/oligoarray2_1/.

If a potential target region has homology to other transcripts, it will be indicated in this file. We only use potential target regions for which no potential off-targets were discovered.

Design of Target Regions

Step 6.

Repeat this process for all desired genes. As mentioned above, OligoArray2.0 is computationally intensive; thus, we recommend only using this software to design target regions for the desired subset of the transcriptome. Moreover, because computational cost does not grow linearly with the length of a transcript, we recommended that longer transcripts be split into smaller fragments and processed individually. We split transcripts into 1-kb increments.

Design of Readout Probes

Step 7.

Utilize existing sets of orthogonal nucleic acid sequences to design readout probes. It is important that one readout probe has little homology with a sequence of another probe to prevent potential off-target binding. Fortunately, a set of 25-mer nucleotide sequences designed to have limited cross homology exists, and we recommend using this resource to design readout sequences (Xu et al., 2009). These sequences can be downloaded from:

http://elledgelab.med.harvard.edu/?page_>.

We have found good performance with readout sequences of 30-nt length, which can be created from these 25-nt sequences by either concatenating portions of them or adding 5 random nucleotides to either end.

Design of Readout Probes

Step 8.

Remove potential probes with homology to members of the transcriptome of interest. To remove probes with significant homology to members of the transcriptome, we create a BLAST library to the transcriptome as described in Section 3.1 Step 3. We then BLAST each potential readout probe sequence against this library and remove any probe which contains a contiguous stretch of homology longer than 14 nt. This length was selected to balance the desire for shorter regions of homology with the increased frequency with which such shorter regions appear in the transcriptome.

Design of Readout Probes

Step 9.

Remove potential readout probes with significant homology to one another. While these oligos were originally designed to have limited cross homology, changing their length may introduce new regions

of homology. We recommend selecting a subset of possible readout probes, building a BLAST database for these sequences, and then use BLAST to identify regions of homology, as described in Step 2. We exclude probes that contain a region of homology to another potential probe longer than 10 nt.

Design of Readout Probes

Step 10.

Order these probes. We order probes from IDT (www.idtdna.com) tagged on the 3' end with a Cy5. We have them synthesized on the 100 nmole synthesis scale and HPLC purified.

Design of the Codebook

Step 11.

Please see Guidelines for Design of the Codebook.

Assembly and Screening of Encoding Probes

Step 12.

Select the readout sequences for each encoding probe. For each possible target region, use the barcode assigned to that RNA to determine which readout sequences to use. We select the readout sequences that we use for each encoding probe randomly, without repeating a readout sequence more than once in a single encoding probe. We find that random assignment rarely produces issues with an imbalance in the usage of some readout sequences and negates concerns over potential biases that may arise from the position of the readout sequence in the encoding probe.

Assembly and Screening of Encoding Probes

Step 13.

Assemble the encoding probe. We place one readout sequence on either side of the targeting sequence though there is no reason why alternative arrangements could not be used. See Figure 2. Care must be taken to insure that the proper orientation of these sequences is used: the readout sequences are the reverse complement of the readout probes designed in the attached full manuscript. Thus, the encoding probe sequence must contain the reverse complement of these probe sequences. Similarly, the target region portion of the encoding probe must contain the reverse complement of the corresponding sequence of the target RNA.

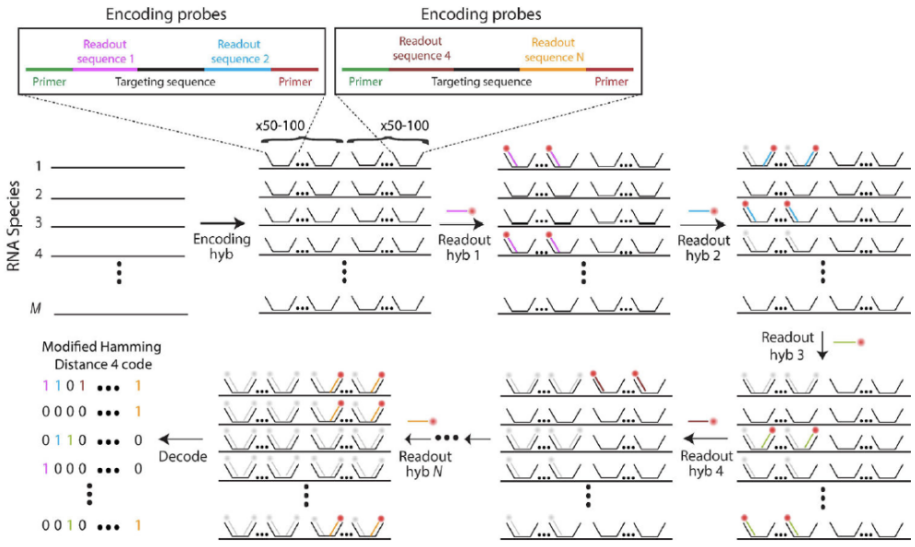

Figure 2.
Schematic depiction of the hybridization process used for MERFISH. Cellular RNAs are hybridized with a set of oligonucleotide probes, which we term *encoding probes*. These encoding probes contain a targeting sequence which directs their binding to the specific RNA. They also contain two readout sequences. For an experiment utilizing N-bit binary barcodes, N different readout sequences will be used with each bit assigned a different unique readout sequence. The specific readout sequences contained by an encoding probe tc a given RNA are determined by the binary barcode assigned to that RNA: only the readout sequences assigned to bits for which this barcode contains a `1' are used. Each encoding probe also contains PCR priming regions used in its construction. To increase the signal from each copy of the RNA, multiple encoding probes, each with a different target region, are bound to the same RNA. The length of this *tile* of probes is typically between 50–100 probes. To identify the readout sequences contained on the encoding probes bound to each RNA, N rounds of hybridization and imaging are performed. Each round uses a unique, fluorescently labeled probe whose sequence is complementary to the readout sequence for that round. The binding of these fluorescent probes determines the bits which contain a `1', allowing the measurement of the specified binary code. Modified with permission from (Chen et al., 2015).

Design of Priming Regions

Step 14.

Truncate an orthogonal set of 25-nt long oligonucleotides to 20-nt. To generate possible primers that have limited possibility of binding to one another, we again start with the set of orthogonal 25mers designed by the Elledge lab as described in Step 7. We reduce the length of these priming sequences to 20-nt to decrease the overall length of the final encoding probes.

Design of Priming Regions

Step 15.

Screen oligonucleotides for ideal primer properties. We remove any potential primer with a T_M outside of the range of 70°C to 80°C as well as any oligonucleotide with any consecutive repeat of 3 or more identical nucleotides, which can create problems in the synthesis of the oligo. We also remove any oligonucleotide that does not contain a G or C in the final 2 nucleotides at the 3' end. The presence of this so-called GC clamp helps improve the efficiency and specificity of PCR primers.

Design of Priming Regions

Step 16.

Screen the final set of possible primers for homology against the encoding probes. Finally, we use BLAST to identify stretches of homology within these primers and the encoding probes designed in Section 'Assembly and Screening of Encoding Probes'. Any potential primer with a homology region of 11 nt or more is excluded.

Design of Priming Regions

Step 17.

Add these primers to each encoding probe to form the template sequence for each oligonucleotide. The sequence of the primer that is added to the 3' end of the encoding probe should be the reverse complement of the sequence of the primer designed in Step 16. Table 2 contains an example of a template molecule for a single encoding probe.

Table 2

An example template molecule for an encoding probe. The target region is marked with bold and italics, the readout sequences are marked with italics only, and the priming regions are marked with bold only. This encoding probe is to the VCAN RNA.

Encoding probe template	**CGCGGGCTATATGCGAACCG** *TTAGTCGTCGCAACGCCAGGCAACTCATGC* ***TAAAGAAATTAGATAGGCTGGAAATGCTTA*** *AAATTGCGTGACGGACCTGGGCCATTGGCC* **GCGTTGTATGCCCTCCACGC**

Design of Priming Regions

Step 18.

Screen assembled template sequences for homology to abundant RNAs. While individual components of these template molecules have been screened for homology against the transcriptome, it is possible that concatenation of the different component sequences will produce regions of homology that will allow the probe to bind to other RNAs. We recommend using BLAST to screen the set of probes designed in Step 17 for homology to abundant RNA species like rRNA and tRNA. For the human transcriptome these sequences can be found within the ncRNA fasta file at the Ensemble website as described in Step 2. Similarly, highly abundant mRNAs can also be included in this screen if such abundance information exists. To identify regions of homology, generate a BLAST database for these sequences, and use BLAST to screen the potential encoding probes. We remove any encoding probe with homology regions longer than 14 nt.

Design of Priming Regions

Step 19.

(Optional). Combine multiple oligo sets to fill a single pool of oligonucleotides. We typically purchase these oligonucleotides as complex oligonucleotide pools generated by array-based synthesis often from CustomArray (http://www.customarrayinc.com/). This company sells two different sizes of complex oligonucleotide pools—a pool with 12,472 unique sequences or a pool with 92,918 unique sequences. There are situations in which one set of encoding probes for one experiment does not fill an entire pool. In this case it is possible to combine multiple sets of encoding probes intended for different experiments in the same oligopool simply by assigning each a unique set of primers.

However, when primers are designed for such combined pools, we recommend performing Step 16 with a BLAST database comprising all encoding probes that will be present in the single oligonucleotide pool not just those in each set of encoding probes.