RNA-FISH TrueProbes User Manual

Table of Contents  
1. Introduction  
2. System Requirements  
3. Installation Guide  
4. Probe Design Workflow  
5. Running TrueProbes  
6. Input Files  
7. Parameter Configuration  
8. Output Files  
9. Custom Organism Setup  
10. Custom Application-Specific Setup

11. Appendices

# 1. Introduction

This manual provides comprehensive instructions for installing, configuring, and using the RNA-FISH TrueProbes software. It is intended for researchers and bioinformaticians involved in probe design for RNA-FISH experiments.

# 2. System Requirements

The following software packages are required to run TrueProbes:

* From MathWorks Inc:
* MATLAB version 2022b or higher
* Bioinformatics Toolbox
* Curve Fitting Toolbox
* Parallel Computing Toolbox
* Statistics and Machine Learning Toolbox
* Symbolic Math Toolbox
* Signal Processing Toolbox
* Image Processing Toolbox
* From NCBI:
* BLAST+

# 3. Installation Guide

## Install MATLAB and product packages

<https://www.mathworks.com/help/install/ug/install-products-with-internet-connection.html>

## If you are unable to get access to a MATLAB License, then install MATLAB Runtime 2025a

<https://www.mathworks.com/products/compiler/matlab-runtime.html>

<https://www.mathworks.com/help/compiler/install-deployed-application.html>

<https://www.mathworks.com/help/compiler/mcr-path-settings-for-run-time-deployment.html>

## Install NCBI-BLAST+

<https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html>

<https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

## Ensure Your Computer Security allows BLAST+ files to run

Compatibility: Ensure the blast+ folder files are allowed to run by your computer's security settings. Specifically, verify that permission to run is granted in macOS or that any security or virus scanners do not block the program.

For MacOS: Go to Privacy and Security to enable blastn, makeblastdb, and blastdbcmd to run on Mac

**4. Probe Design Workflow**

**The TrueProbes software operates by performing eight sequential steps.**

**1) Probe Generation.** It generates all possible probes within a set probe length range shared between RNA sequences located inside NCBI-RefSeq or ENSEMBL transcript accession numbers (RNA sequence identifiers) and RNA sequence inclusion text files, that are also not located inside a list of NCBI-RefSeq or ENSEMBL transcript accession numbers or RNA sequence files to exclude from probe design.

**2) BLAST Alignment.** BLAST is used to identify where probes bind in the target genome and/or transcriptome. These blast results are then recorded and stored for probe-target binding events whose number of nucleotide base pairs matches is greater than or equal to the minimum homology length target search size.

**3) BLAST Target Gene/Transcript Expression.** The gene expression and transcript expression levels are collected for reference in the specified expression databases, as defined in the TrueProbes settings, for all targets in the BLAST results.

**4) Binding Affinity Calculation**. Binding affinities are calculated for all pairs of probe and target sequences with homology matches in the BLAST results.

**5) Probe-Target Binding Site Mapping.** All blast hits and binding affinities are converted into a site-specific binding map to generate a formatted map by relative binding site position on each target gene, transcript, or chromosome.

**6) Probe-Target Statistics**. Generate statistics on blast hits, thermodynamics, and a comparison of probes sharing off-targets and relative trade-offs when quantifying off-targets by probe and comparing probes that bind them in a site-specific manner.

**7) Probe Design.** We first sort probes low to high based on the number of off-targets they have when either including or excluding reference off-target gene expression levels. Within each range of probes with a certain number of off-targets, the probes are then sorted from high to low on the difference between their on-target and off-target/secondary structure binding affinity. After this, we first design probes with no off-targets, prioritizing the packing of the most significant number of high-affinity probes, while iteratively re-prioritizing probes to minimize cross-binding to probes we have already designed. We then iteratively design probes with the fewest number of off-targets and most significant difference in on-target to off-target and secondary structure binding affinity until we reach the maximum number of desired probes or no more probes can be designed. We then print the designed probe sequences in an Excel Spreadsheet with a list of their properties and off-target information.

**8) Model Evaluation.** Final probe sets and reference expression are combined to compute equilibrium probe binding and statistics, including cumulative off-target binding and on-target binding, when reference values for cell size, probe concentration, and probe intensity are provided.

# 5. Running TrueProbes

## TrueProbes RNA-FISH Probe Design

**A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5** is the main script for running the software. The design file requires two inputs (id and cluster), as well as files that specify the settings that describe how the probe design will be executed. For a description and list of files for specifying the design settings, see **Section 6.**

### In MATLAB

The code is run via the command line.

**A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5(id,cluster)**

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

### In MATLAB Runtime

The code can also be run via the terminal in MATLAB Runtime for users who do not have a MathWorks MATLAB license. Users must download the MATLAB Runtime version created with the compiled MATLAB License-Free Version of TrueProbes software or use the TrueProbes Design Software Installer, which includes the required MATLAB Runtime for running the software.

|  |  |  |
| --- | --- | --- |
| **Operating System** | **MATLAB Runtime and TrueProbes Installer** | **Run Command** |
| Windows / WIN64 | TrueProbes\_ProbeDesignerInstaller\_win64.exe | .\A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.exe id cluster |
| Intel Mac / MACI64 | TrueProbes\_ProbeDesignerInstaller\_maci64.app | ./A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.app/Contents/  MacOS/A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5 id cluster |
| Silicon Mac/ MACA64 | TrueProbes\_ProbeDesignerInstaller\_maca64.app | ./A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.app/Contents/  MacOS/A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5 id cluster |
| Linux / GLNXA64 | TrueProbes\_ProbeDesignerInstaller\_glnxa64 | ./A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5 id cluster |

Additionally, on Mac or Linux, this code can be run via the included shell script for temporarily setting environment variables and executing the application.

**run\_A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.sh**

To run the shell script, type in the Linux or Mac command prompt

./run\_A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.sh <mcr\_directory> id cluster

Where <mcr\_directory> is the directory where MATLAB Runtime(R2024a) is installed or the directory where MATLAB is installed on the machine.

**In Linux**

**Set the Runtime Library Path**

export LD\_LIBRARY\_PATH="${LD\_LIBRARY\_PATH:+${LD\_LIBRARY\_PATH}:}\ /usr/local/MATLAB/MATLAB\_Runtime/R2024a/runtime/glnxa64:\ /usr/local/MATLAB/MATLAB\_Runtime/R2024a/bin/glnxa64:\ /usr/local/MATLAB/MATLAB\_Runtime/R2024a/sys/os/glnxa64:\ /usr/local/MATLAB/MATLAB\_Runtime/R2024a/extern/bin/glnxa64"

**Run Command**

./run\_A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.sh $LD\_LIBRARY\_PATH id cluster

**In Intel processor-based macOS**

**Set the Runtime Library Path**

export DYLD\_LIBRARY\_PATH="${DYLD\_LIBRARY\_PATH:+${DYLD\_LIBRARY\_PATH}:}\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/runtime/maci64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/bin/maci64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/sys/os/maci64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/extern/bin/maci64"

./run\_A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.sh $DYLD\_LIBRARY\_PATH id cluster

**In Apple Silicon-based macOS**

**Set the Runtime Library Path**

export DYLD\_LIBRARY\_PATH="${DYLD\_LIBRARY\_PATH:+${DYLD\_LIBRARY\_PATH}:}\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/runtime/maca64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/bin/maca64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/sys/os/maca64:\ /Applications/MATLAB/MATLAB\_Runtime/R2024a/extern/bin/maca64"

**Run Command**

./run\_A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5.sh $DYLD\_LIBRARY\_PATH id cluster







## Running TrueProbes Evaluation on Existing Probe Sets

**A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5** is the main script for running the software. The design file requires three inputs (id, cluster, id2), as well as files that specify the settings describing how the probe set evaluation will be performed. For a description and list of files used to define the probe evaluation settings, see **Section 6.**

### In MATLAB

The code is run via the command line.

**A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5(id,cluster,id2)**

**1st Input Argument id: id** is an integer and is the row of the input design table file and determines which target to search the designed probes against.

|  |  |
| --- | --- |
| **ID value type** | **Description** |
| integer | row of the input design table file to design probes for |

**2nd Input Argument cluster: cluster** is an integer that determines the software parallelization pool between local and remote servers when running the script via Slurm. The only difference between running on a cluster and running on a computer is the number of cores in the Slurm file.

|  |  |
| --- | --- |
| **Cluster value** | **Description** |
| 0 | Run Locally |
| 1 | Run Remotely with Slurm file environmental variables. |

**3rd Input Argument id2: id2** is an integer and determines which probe set from other software, or custom list, to use for any comparison of probe design metrics.

|  |  |
| --- | --- |
| **Id2 value** | **Output Software Name** |
| 2 | Stellaris |
| 3 | OligostanHT |
| 4 | PaintSHOP |
| 5 | Xiaowei Zhuang Lab MERFISH |
| 6 | Allen Institute MERFISH |
| 7 | Length Optimized |
| 8 | Custom Upload |

### In MATLAB Runtime

The code can also be run via the terminal in MATLAB Runtime for users who do not have a MathWorks MATLAB license. Users must download the MATLAB Runtime version created with the compiled MATLAB License-Free Version of TrueProbes software or use the TrueProbes Design Software Installer, which includes the required MATLAB Runtime for running the software.

|  |  |  |
| --- | --- | --- |
| **OS** | **MATLAB Runtime and TrueProbes Installer** | **Run Command** |
| Windows  WIN64 | TrueProbes\_ProbeComparisonInstaller\_win64.exe | .\A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.exe id cluster id2 |
| Intel Mac  MACI64 | TrueProbes\_ProbeComparisonInstaller\_maci64.app | ./A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.app/Contents/  MacOS/A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5 id cluster id2 |
| Silicon Mac  MACA64 | TrueProbes\_ProbeComparisonInstaller\_maca64.app | ./A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.app/Contents/  MacOS/A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5 id cluster id2 |
| GLNXA64 | TrueProbes\_ProbeComparisonInstaller\_glnxa64 | ./A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5 id cluster id2 |

Additionally, on Mac or Linux, this code can be run via the included shell script for temporarily setting environment variables and executing the application.

**run\_A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.sh**

To run the shell script, type in the Linux or Mac command prompt

./run\_A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.sh <mcr\_directory> id cluster id2

Where <mcr\_directory> is the directory where MATLAB Runtime(R2024a) is installed or the directory where MATLAB is installed on the machine.

**In Linux**

./run\_A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.sh $LD\_LIBRARY\_PATH id cluster id2

**In MacOS**

./run\_A0\_BKJH\_ProbeComparison\_Wrapper\_cluster\_V5.sh $DYLD\_LIBRARY\_PATH id cluster id2

# 6. Input Files

TrueProbes Probe Design utilizes four main input files to specify the probe design.

1. **TrueProbes\_DesignTargets.csv:** CSV File that lists all design targets
2. **ProbeDesignSettings\_Parameters.xml:** XML File where all design settings are specified
3. **DatabaseLocations.xml:** XML File with the location of all database files needed in design when using the NCBI or ENSEMBL reference genome, for any potential organism designed against
4. **GeneExpressionDataFileLocation.xml:** XML File with the location of all gene expression files, schema, and sample label for all reference gene expressions desired in the design

|  |  |
| --- | --- |
| **Name** | **Description** |
| TrueProbes\_DesignTargets.csv | A file specifying all targets for probe design with information on the organism and which targets should be included and/or excluded from probe design. |
| DatabaseLocations.xml | A file specifying where all gene annotations and database files needed for designing probes are located |
| GeneExpressionDataFileLocation.xml | A file specifying the location of any reference gene expression data included for designing probes is located. |
| ProbeDesignSettings\_Parameters.xml | A file specifying how all probes for all targets in the design target file should be designed. |

# 7. Database Locations

The EMBL to NCBI table in DatabaseLocation.xml stores the location of files for mapping ENSEMBL annotations to and from NCBI annotations. These files contain columns that map ENSEMBL gene IDs, ENSEMBL transcript IDs, and NCBI RefSeq transcript accession numbers. These mappings are used when reference genome/transcriptome BLAST databases and reference gene expression data use different annotation formats. This enables TrueProbes to convert NCBI-RefSeq/ENSEMBL gene or transcript expression level data into inferred expression levels of off-targets in ENSEMBL/NCBI-RefSeq Blast hits. These files currently default to using ENSEMBL stable ID to RefSeq stable ID mapping done in ENSEMBL Other Annotation TSV RefSeq gene annotation data dumping (<https://useast.ensembl.org/info/data/ftp/index.html>).

## EMBL to NCBI

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | Name of organism to search for in TrueProbes\_DesignTargets.csv |
| Stable IDs | Location of the file with EMBL and NCBI transcripts paired to one another for that organism. |

The EMBL table in DatabaseLocation.xml stores the location of files for using ENSEMBL annotation in probe design. Each row of EMBL corresponds to an organism, containing information on the database location and files. This table in the Database Location xml file needs to be filled out for each organism against which the probes are designed, enabling the automatic use of GTF and GFF files to retrieve gene names, isoforms, and any other target information required when using only the accession number.

## EMBL

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | Name of organism to search for in TrueProbes\_DesignTargets.csv |
| Root\_FASTA | Location of folder with all ENSEMBL annotation DNA and RNA fasta files |
| BLASTDB\_DNA | Location of ENSEMBL blast+ genome DNA database files |
| BLASTDB\_RNA | Location of ENSEMBL blast+ transcript RNA database files |
| GTF | Location of ENSEMBL reference genome GTF file |
| GFF | Location of ENSEMBL reference genome GFF3 file |

The NCBI table in DatabaseLocation.xml stores the location of files for using the NCBI-RefSeq annotation in probe design. Each row of NCBI corresponds to an organism, containing information on the database location and files. This table in the Database Location xml file needs to be filled out for each organism against which the probes are designed, enabling the automatic use of GTF and GFF files to retrieve gene names, isoforms, and any other target information required when using only the accession number.

## NCBI

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | Name of organism to search for in TrueProbes\_DesignTargets.csv |
| Root\_FASTA | Location of folder with all NCBI RefSeq annotation DNA and RNA fasta files |
| BLASTDB\_DNA | Location of NCBI RefSeq blast+ genome DNA database files |
| BLASTDB\_RNA | Location of NCBI RefSeq blast+ transcript RNA database files |
| GTF | Location of the NCBI RefSeq reference genome GTF file |
| GFF | Location of the NCBI RefSeq reference genome GFF file |

# 8. Gene Expression Data Location

Stored for each organism, the location of all expression files with the identifier name. Rows for each organism's reference gene or transcript level expression files.

## Gene Expression File Locations

| **Name** | **Description** |
| --- | --- |
| Organism | Name of organism to search for in TrueProbes\_DesignTargets.csv |
| “Data Identifier Name” | Location of gene expression file with extension in TrueProbes Folder |

Stored for each organism is the location of all gene expression data file schema files, which list the different cell sample names associated with the expression values within each expression data file. This list is the names of each cell line or single cell identifier for each row or column of reference gene or transcript level expression values associated with each gene/transcript in each gene expression data file.

## Gene Expression Schema

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | Name of organism to search for in TrueProbes\_DesignTargets.csv |
| “Data Identifier Name” | Location of the schema for the gene expression file with extension |

The track table in the Gene Expression File Location XML file stores the name of the column entries for expression data files ending in each file extension. This is used to retrieve the rows corresponding to gene names, transcript IDs, and expression values when reading and mapping gene expression data to obtain gene expression data for blast+ on- and off-targets.

## Gene Expression File Column Names

| **Name** | **Description** |
| --- | --- |
| “File extension” | column-sorted list of variable names |

# 9. Parameter Configuration

Parameters are specified in TrueProbes\_ParameterSettings.xml and are grouped into distinct categories, which can be configured to determine how the probes are designed and evaluated. These categories are settings groups for Main Probe Design, Thermodynamics, Design Filtering, Parallelization, Gene Expression, BLASTDB, BLAST, and Model Simulations.

## Main Probe Design Parameters:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| minProbeSize | min nt length of potential probes | 20 |
| maxProbeSize | max nt length of potential probes | 20 |
| MinProbeSpacing | min spacing between probes | 3 |
| MaxNumberOfProbes | Max number of probes to design | 96 |
| targetStrand | Strand of the target to design probes against 1 for ‘plus’, 0 for ‘minus’ | 1 |
| MinHomologySearchTargetSize | Minimum homology length for BLAST alignments to be recorded and used in probe design and evaluation | 15 |
| BLASTrna | Decide to blast RNA sequences | 1 |
| BLASTdna | Decide to blast DNA sequences | 0 |
| ExpressionReferenceForProbeDesign | Which row across all expression reference files to use in probe design, with zero meaning not to use expression data to design probes | 0 |

## Thermodynamic Settings:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| Gibbs Model | Which thermodynamic model to use for probe design and evaluation | 4 |
| Hybridization Temperature | Hybridization temperature in Celsius (°C) | 37 |
| Heat Capacity Reference Temperature | Reference temperature for Cp measurement and Gibbs model in Celsius (°C) | 37 |
| Salt Concentration | Salt Concentration in M | 0.05 |
| Primer Concentration | Primer Concentration M | 50e-6 |
| Remove Mismatches | Remove sequence-mismatched base pairs before evaluating probe-target binding affinity (1). Setting to 0 will include adding flanking sequences to alignments and using Gibbs model 8 with mismatch base pair inclusion | 1 |

### Gibbs Free Energy Models

|  |  |  |
| --- | --- | --- |
| **Model** | **Abbreviation** | **Reference** |
| 1 | Breslauer86 | Breslauer K.J., Frank R., Blocker H., Marky L.A., (1986) Predicting DNA duplex stability from the base sequence Proc Natl Acad Sci U S A 83, 3746-3750 |
| 2 | SantaLucia96 | SantaLucia, J., Allawi, H. T., and Seneviratne, P. A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability Biochemistry 35, 3555-3562 |
| 3 | SantaLucia98 | SantaLucia, J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics Proc Natl Acad Sci U S A 95, 1460-1465 |
| 4 | Sugimoto96 | Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes Nucleic Acids Res 24, 4501-4505 |
| 5 | SantaLucia04 | SantaLucia Jr, J., and Hicks, D. (2004) The thermodynamics of DNA structural motifs Annu Rev Biophys Biomol Struct 33, 415-440 |
| 6 | Allawi97 | Allawi, H. T., and SantaLucia, J. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA Biochemistry 36, 10581-10594 |
| 7 | Rejali21 | Rejali, N. A., Ye, F. D., Zuiter, A. M., Keller, C. C., and Wittwer, C. T. (2021) Nearest-neighbour transition-state analysis for nucleic acid kinetics Nucleic Acids Res 49, 4574-4585 |
| 8 | Martins24 | de Oliveira Martins, E., and Weber, G. (2024) Nearest-neighbour parametrization of DNA single, double and triple mismatches at low sodium concentration Biophys Chem 306, 107156 |

## Design Filtering Settings:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| RemoveProbesBindingOffTargetRibosomalHits | Filter out probes with off-targets to rRNA | 1 |
| packOptimal\_ProbesWithNoOffTargets | When designing probes without off-target use, optimal packing to get as many probes with no off-targets as possible, as opposed to normal sequential selection | 1 |
| IncludeSelfHybridizationInProbeSelection | When designing probes, consider probe self-hybridization when ranking probes based on binding affinity | 1 |

## Parallelization Parameters: (Usually only changed when using longer genes)

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| Parallization\_probeBatchSize | The number of probes to evaluate in a single batch when performing parallelized calculations | 20 |
| Parallization\_targetBatchSize | The number of targets to evaluate in a single batch when performing parallelized calculations | 200 |
| ParsingPreference | Parse blast results for different probes simultaneously in parallel (1) or sequentially (0) | 1 |

## Gene Expression Parameters:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| UseGeneOverTranscLevelExpression | Use gene level (1) or transcript isoform level (0) gene expression values | 0 |
| DoAllGenesHaveSameExpression | Decide to assume equal expression for all genes (1) or to use gene expression reference (0) | 0 |
| UseRegularDNAExpression | (0) use DNA expression from the gene expression track in the expression data, (1) set expression to 2 for DNA | 1 |
| nullRNAcopynumber | RNA copy number when not using reference expression levels | 100 |
| nullDNAcopynumber | DNA copy number when not using reference expression levels | 2 |
| TMM\_LogRatioTrim | When normalizing TPM expression data using TMM, set the log ratio trim threshold cutoff | 0.3 |
| TMM\_SumTrim | When normalizing TPM expression data using TMM, set the sum trim threshold cutoff | 0.05 |
| TMM\_Acutoff | When normalizing TPM expression data using TMM, set A cutoff value | -1e10 |
| TMM\_doWeighting | When normalizing TPM expression data, weigh the terms using the inverse of the approximate asymptotic variance of the M-values to account for genes with higher read counts having lower variance on the log scale, and more reliable mean estimation | 1 |

## Make Blast Database Settings:

| **Name** | **Description** | **Default** |
| --- | --- | --- |
| Parse\_seqids | When making blast databases in the software parse sequence IDs from fasta files as blast database sequence IDs | false |
| Hash\_index | When making a blast database, creating sequence hash indexes leads to faster exact match retrieval, but less accurate range matches | false |

## BLAST Parameters:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| evalue | Expectation value cutoff | 1000 |
| Dust | Filter query sequences with DUST | no |
| Gapextend | Cost to extend a gap (integer) | 2 |
| Gapopen | Cost to open a gap (integer) | 5 |
| Num\_alignments | Max number of database sequence alignments to report per query | 1000 |
| Penalty | Penalty for a nucleotide mismatch | -3 |
| Reward | reward for a nucleotide match | 1 |
| Word size | word size for the wordfinder algorithm | 7 |
| Task | Which task to use when evaluating BLASTN | megablast |

## Model Simulation Settings:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Default** |
| removeUndesiredIsoformsFromPrediction | When evaluating main on/off-target binding, should alternate isoforms of desired targets be removed and not included in off-target quantification | 1 |
| ProbeConcentration\_MicroMolar | probe concentration in uM | 5e-6 |
| CellRadius\_Micron | cell radius in microns for converting RNA molecule counts into concentration for solving the binding equilibrium | 10 |
| Dilution\_Vector | comma-separated vector of probe dilutions to evaluate binding equilibrium and predictions at | 1 |
| Gibbs\_Model\_Vector | comma-separated vector of Gibbs thermodynamic models to use for evaluating binding equilibrium and predictions at | 1,2,3,4 |
| Temperature\_Celsius\_Model\_Vector | comma-separated vector of hybridization temperatures in Celsius to use for evaluating binding equilibrium and predictions at | 37 |
| InitialFreeSolutionGuessConcentration  \_MicroMolar | initial solution guess for steady state free probe concentration in uM | 1e-10 |
| SolutionErrorTolerance | Total tolerance level for error in the final equilibrium solution | 1 |
| MaxRecursiveEquilibriumIterations | max number of equilibrium equation iterations, stopping at the final calculated steady state, overrides solution error tolerance | 40 |
| CellDiameter\_Pixels | Cell pixel diameter for background predictions | 50 |
| SpotRadius\_Pixels | spot pixel radius for predictions | 5 |
| NumberOfReferenceZStacks | Number of z-stacks to spread background across for intensity predictions | 67 |
| SignalStepSize | Step size for signal intensity bins in signal predictions | 1e-1 |
| SignalMaxValue | max intensity value for ranges in the solution | 3000 |
| AutoFluoresenceBackground\_MEAN | reference mean autofluorescence | 278 |
| AutoFluoresenceBackground\_STD | reference autofluorescence standard deviation | 33 |
| NumberOfProbesInReferenceSpots | number of probes in the reference spot intensity for calibrating intensity predictions | 48 |
| ReferenceSpotIntensity\_MEAN | mean reference spot intensity | 827 |
| ReferenceSpotIntensity\_STD | reference spot intensity standard deviation | 28 |

# 8. Output Files

Around a dozen results files are generated during the probe design process. Files First start with the GeneName in parentheses, with up to four parameter values recorded in the output file names. In the following table, the following parameter values are highlighted using placeholders in place of their actual values: the design target (**AccessionID)**, the maximum number of desired probes (**MaxProbes**), the Hybridization Temperature (**T**), and the software (**SW**) if running TrueProbes Probe Designer or other software if running Probe Design Comparison.

**SW** options include: TrueProbes, Stellaris, OligostanHT, MERFISH, PaintSHOP, LengthOptimized, Custom, Random, and any other “user-defined names.”

For Example:

**GeneName** = ARF4, **SW** = TrueProbes, **T** = 37, **MaxProbes** = 96, and **AccessionID** = NM\_001660.4

(**GeneName**)\_**AccessionID**\_probes\_**SW**.mat -> (ARF4)\_NM\_001660.4\_probes\_TrueProbes.mat

(**GeneName**)\_**AccessionID**\_Tm**T**\_BindingEnergyMatrix\_**SW**.mat ->

(ARF4)\_NM\_001660.4\_Tm37\_BindingEnergyMatrix\_TrueProbes.mat

(**GeneName**)\_**AccessionID**\_probes\_final\_**MaxProbes**max.xlsx -> (ARF4)\_NM\_001660.4\_probes\_final\_96max.xlsx

## TrueProbes Design Software Output Files:

|  |  |
| --- | --- |
| **Name** | **Description** |
| (**GeneName**)\_**AccessionID**\_probes\_**SW**.mat | Cell array with probe tile |
| (**GeneName**)\_**AccessionID**\_hits\_table\_**SW**.mat | A table with information on reported BLAST hits |
| (**GeneName**)\_**AccessionID**\_ExpressionInfo\_**SW**.mat | Matrix with BLAST target expression data |
| (**GeneName**)\_**AccessionID**\_dCpInfo\_**SW**.mat | Matrix with heat capacity for all target binding reactions |
| (**GeneName**)\_**AccessionID**\_dHInfo\_**SW**.mat | Matrix with the enthalpy for all target binding reactions |
| (**GeneName**)\_**AccessionID**\_dSInfo\_**SW**.mat | Matrix with entropy for all target binding reactions |
| (**GeneName**)\_**AccessionID**\_dKbInfo\_**SW**.mat | Matrix with equilibrium binding constant for all target binding reactions at the Hybridization Temperature |
| (**GeneName**)\_**AccessionID**\_TmInfo\_**SW**.mat | Structure with melting temperature for all target binding reactions |
| (**GeneName**)\_**AccessionID**\_Tm**T**\_OnOffThermoInfo\_**SW**.mat | Matrix with non-site-specific binding energy of all on/off-targets |
| (**GeneName**)\_binding\_hits\_map\_**SW**.mat | Probe-target binding site map |
| (**GeneName**)\_**AccessionID**\_Tm**T**\_BindingEnergyMatrix\_**SW**.mat | Equilibrium binding energy in binding site map format |
| (**GeneName**)\_**AccessionID**\_Tm**T**\_\_BindingEnergyMatrix2\_**SW**.mat | Complementary DNA strand Equilibrium Binding Energy in binding site map format |
| (**GeneName**)\_BindingMatrices\_**SW**.mat | Entropy, Enthalpy, and heat capacity in binding site map format for RNA |
| (**GeneName**)\_BindingMatrices2\_**SW**.mat | Entropy, Enthalpy, and heat capacity in binding site map format for complementary DNA strand binding |
| (**GeneName**)\_**AccessionID**\_Tm**T**\_BasicDesignerStats\_**SW**.mat | Statistics used for designing probes |
| (**GeneName**)\_**AccessionID**\_chosen.mat | List of chosen probe indexes in the probe designer tile |
| (**GeneName**)\_**AccessionID**\_probes\_final\_**MaxProbes**max.xlsx | An Excel spreadsheet with final sequences of probes, thermodynamic information, off-target information, and info detailing why it was designed. |
| (**GeneName**)\_**AccessionID**\_Tm**T**\_ModelMetrics\_**SW**.mat | Structure with designed probe binding affinity calculations and probe design metrics |

## TrueProbes Design Software Output Files:

## Outputs in (GeneName)\_AccessionID\_probes\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **probes** | Cell array storing target probe tiling |

### Entries in probes

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Probe Num | The unique number ID is given to each potential probe sequence in the target tile. probes{probe ID,1} = Probe Num. |
| Probe Sequence | Probe Tile Sequence, which is the matching sequence, will bind (i.e., the reverse complement of the probe that will be ordered).  probes{probe ID,2} = Sequence of Target that the probe binds. |
| Probe Start Position | The nucleotide position in the target sequence from which the probe sequence starts 5’>3’  probes{probe ID,3} = Probe Start Position on Target. |

## Outputs in (GeneName)\_AccessionID\_hits\_table\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Gene\_hit\_table | Table storing all the blast hits for all probes in the target probe tiling. |

### Entries in gene\_hit\_table

|  |  |
| --- | --- |
| **Row Name** | **Description** |
| Score | BLAST+ Alignment Score. Gene\_hit\_table(:,1) = Alignment Scores. |
| Expect | BLAST+ Alignment E-value. Gene\_hit\_table(:,2) = Alignment E-values. |
| Strand | Which target strand alignment is on with: 'Plus/Plus’ for 5’->3’ and ‘Plus/Minus’ for reverse complement. Gene\_hit\_table(:,3) = Strand |
| Probe Alignment | Is the probe alignment formatted with three lines: Probe Sequence  Alignment, a line of spaces, followed by the target sequence alignment.  Gene\_hit\_table(:,4) = Probe Alignment = {‘Probe Alignment Sequence’  ‘Space lines’  ‘Target Alignment Sequence’}. |
| Name | The Target Accession Number and description from the FASTA file header. Gene\_hit\_table(:,5) = Target Name and Definition. |
| Probe Sequence | Sequence of Probe in Probe Tiling. Gene\_hit\_table(:,6) = Probe Tile Sequence, which is the oligonucleotide sequence the probe binds. |
| Probe Number | The Probe ID for the probe in the Probe Tiling. Gene\_hit\_table(:,7) = Probe Tile ID. |
| Match | The number of homology base pair matches. Gene\_hit\_table(:,8) = number of homology base pair matches. |
| Possible | The number of base pairs of the probe that could possibly match the target. Gene\_hit\_table(:,9) = Number of probe base pairs that could match the target. |
| Percentage | The percentage of base pair matches relative to the number of probe nucleotides that could bind the target. Gene\_hit\_table(:,10) = The Percentage of possible target base pairs that match the target. |

## Outputs in (GeneName)\_AccessionID\_ExpressionInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| ExpressionMatrix | Matrix with expression values for each unique target name in the gene hit table in any number of cell lines or expression conditions. Targets are ordered in the table by unique sorting of all hit names in the gene\_hit\_table after filtering out all hits under the minimum homology setting size. [Targets x Cell Lines] |

## Outputs in (GeneName)\_AccessionID\_dCpInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| dCpeq\_Match | 2D Matrix with thermodynamic-kinetic equilibrium heat capacity correction values for all unique pairs of probe target sequence alignments calculated for all nearest-neighbor models in TrueProbes software. Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models] |
| dCon\_eq | 2D Matrix with thermodynamic-kinetic heat capacity correction values for each probe tile’s on-target binding calculated for all nearest-neighbor models in TrueProbes software. Matrix is stored as [ Probe Tile Position x Number of Gibbs Models] |

## Outputs in (GeneName)\_AccessionID\_dHInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| dHeq\_Match | 2D Matrix with thermodynamic-kinetic equilibrium enthalpy for all unique pairs of probe target sequence alignments calculated for all nearest-neighbor models in TrueProbes software. Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models] |
| dHf\_Match | 2D Matrix with thermodynamic-kinetic transition-state binding/forward reaction enthalpy for all unique pairs of probe target sequence alignments calculated for the Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [Unique Probe-Target Sequence x 3] |
| dHr\_Match | 2D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction enthalpy for all unique pairs of probe target sequence alignments calculated for the Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [Unique Probe-Target Sequence x 3] |
| dHon\_eq | 2D Matrix with thermodynamic-kinetic enthalpy for each probe tile’s on-target binding calculated for all nearest-neighbor models in TrueProbes software. Matrix is stored as [ Probe Tile Position x Number of Gibbs Models] |
| dHon\_f | 2D Matrix with thermodynamic-kinetic transition-state binding/forward reaction enthalpy for each probe tile’s on-target binding calculated for the Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x 3 ] |
| dHon\_r | 2D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction enthalpy for each probe tile’s on-target binding calculated for the Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x 3 ] |

## Outputs in (GeneName)\_AccessionID\_dSInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| dSeq\_Match | 2D Matrix with thermodynamic-kinetic equilibrium entropy for all unique pairs of probe target sequence alignments calculated for all nearest-neighbor models in TrueProbes software. Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models] |
| dSf\_Match | 2D Matrix with thermodynamic-kinetic transition-state binding/forward reaction entropy for each probe tile’s on-target binding calculated for the Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [ Probe Tile Position x 3 ] |
| dSr\_Match | 2D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction entropy for all unique pairs of probe target sequence alignments calculated for the Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [Unique Probe-Target Sequence x 3] |
| dSon\_eq | 2D Matrix with thermodynamic-kinetic entropy for each probe tile’s on-target binding calculated for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Probe Tile Position x Number of Gibbs Models] |
| dSon\_f | 2D Matrix with thermodynamic-kinetic transition-state binding/forward reaction entropy for each probe tile’s on-target binding calculated for the Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [ Probe Tile Position x 3 ] |
| dSon\_r | 2D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction entropy for each probe tile’s on-target binding calculated for the Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties. Matrix is stored as [ Probe Tile Position x 3 ] |

## Outputs in (GeneName)\_AccessionID\_TmInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Tm\_Match | 2D Matrix with thermodynamic-kinetic equilibrium melting temperature for all unique pairs of probe target sequence alignments calculated for a basic non-nearest model Tm equation and Tm equations for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models + 1] |
| Tm\_on | 2D Matrix with thermodynamic-kinetic equilibrium melting temperature for each probe tile’s on-target binding calculated for a basic non-nearest model Tm equation and Tm equations for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [ Probe Tile Position x Number of Gibbs Models+1] |

## Outputs in (GeneName)\_AccessionID\_dKbInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Kb\_Match | 2D Matrix with thermodynamic-kinetic equilibrium binding constant for all unique pairs of probe target sequence alignments calculated for all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature.  Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models] |

## Outputs in (GeneName)\_AccessionID\_TmT\_OnOffThermoInfo\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Kb\_Match | 2D Matrix with thermodynamic-kinetic equilibrium binding constant for all unique pairs of probe target sequence alignments calculated for all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature.  Matrix is stored as [Unique Probe-Target Sequence x Number of Gibbs Models] |
| Kon | 2D Matrix with thermodynamic-kinetic binding equilibrium constant for each probe tile’s on-target binding calculated for all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature.  Matrix is stored as [ Probe Tile Position x Number of Models] |
| Koff | 2D Matrix with thermodynamic-kinetic binding equilibrium constant for the strongest off-target binding interaction a probe in the probe tile has for each potential unique off-target RNA or DNA at the desired Hybridization Temperature.  Matrix is stored as [ Unique Target IDs x Number of Models] |

## Outputs in (GeneName)\_binding\_hits\_map\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| DoesProbeBindSite | The unfiltered Binding Site Map is zero if the probe does not bind the target site and is one if the probe binds the target at the site.  DoesProbeBindSite is a 3D Matrix with Size  [Probe Tile Position x Target x Binding Site]. |
| DoesProbeBindSite2 | Filtered DoesProbeBindSite Binding Site Map, making sure probes do not bind overlapping across adjacent binding sites.  DoesProbeBindSite2 is a 3D Matrix with Size  [Probe Tile Position x Target x Binding Site]. |
| MolN\_ProbesAtEvent | Cell array where each cell entry is the number of probes binding the target at each site.  MolN\_ProbesAtEvents{target}=Number of Probes Binding Targets at Each Site. |
| MolProbesAtEvent | Cell Array where one cell per target, with each cell containing another cell array with probes binding the target at site 1 to N.  MolProbesAtEvent{target}{site} = Probes Binding Target Site. |
| Mol\_ProbesAtEvents\_ID | Cell Array where one cell per target, with each cell containing another cell array with ID row of the Match List corresponding to the binding events at each target site.  MolProbesAtEvent{target}{site} = Row of binding event in Kb\_Match, dH\_Match, dS\_Match, and dCp\_Match. |
| Num\_of\_Molecule\_Sites | The number of distinct binding sites at each unique target probe in the tile could bind. |

## Outputs in (GeneName)\_AccessionID\_TmT\_BindingEnergyMatrix\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Kb\_mod | 4D Matrix with thermodynamic-kinetic equilibrium binding constant for all probes binding targets at any site for all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature.  Matrix is stored as [Probe Tile Position x Target x Binding Site x Number of Gibbs Models] |

## Outputs in (GeneName)\_AccessionID\_TmT\_\_BindingEnergyMatrix2\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Kb\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium binding constants for all DNA target complementary strand binding reactions at any DNA target probe binding site using all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature  Matrix is stored as [Target x Binding Site x Number of Gibbs Models] |
| POGmod\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium Gibbs Free Energy for all DNA target complementary strand binding reactions at any DNA target probe binding site using all nearest-neighbor models in TrueProbes software at the desired Hybridization Temperature  Matrix is stored as [Target x Binding Site x Number of Gibbs Models] |

## Outputs in (GeneName)\_BindingMatrices\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| dCp\_mod | 4D Matrix with thermodynamic-kinetic equilibrium heat capacity correction values for all probes binding targets at any binding site for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Probe Tile Position x Target x Binding Site x Number of Gibbs Models] |
| dHeq\_mod | 4D Matrix with thermodynamic-kinetic equilibrium enthalpy for all probes binding targets at any site for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Probe Tile Position x Target x Binding Site x Number of Gibbs Models] |
| dHf\_mod | 4D Matrix with thermodynamic-kinetic transition-state binding/forward reaction enthalpy for all probes binding targets at any site using Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x Target x Binding Site x 3 ] |
| dHr\_mod | 4D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction enthalpy for all probes binding targets at any site using Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x Target x Binding Site x 3 ] |
| dSeq\_mod | 2D Matrix with thermodynamic-kinetic equilibrium entropy for all probes binding targets at any site for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Probe Tile Position x Target x Binding Site x Number of Gibbs Models] |
| dSf\_mod | 4D Matrix with thermodynamic-kinetic transition-state binding/forward reaction entropy for all probes binding targets at any site using Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x Target x Binding Site x 3 ] |
| dSr\_mod | 4D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction entropy for all probes binding targets at any site using Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [ Probe Tile Position x Target x Binding Site x 3 ] |
| Tm\_mod | 4D Matrix with thermodynamic-kinetic equilibrium melting temperature for all probes binding targets at any site, all calculated for a basic non-nearest model Tm equation and Tm equations for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Probe Tile Position x Target x Binding Site x Number of Gibbs Models+1] |

## Outputs in (GeneName)\_BindingMatrices2\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| dCp\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium heat capacity for all DNA target complementary strand binding reactions at any DNA target probe binding site using all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Target x Binding Site x Number of Gibbs Models] |
| dHeq\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium enthalpy for all DNA target complementary strand binding reactions at any DNA target probe binding site using all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Target x Binding Site x Number of Gibbs Models] |
| dHf\_Complement | 3D Matrix with thermodynamic-kinetic transition-state binding/forward reaction enthalpy for all DNA target complementary strand binding reactions at any DNA target probe binding site using Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [Target x Binding Site x 3 ] |
| dHr\_Complement | 3D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction enthalpy for all DNA target complementary strand binding reactions at any DNA target probe binding site using Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [Target x Binding Site x 3 ] |
| dSeq\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium entropy for all DNA target complementary strand binding reactions at any DNA target probe binding site using all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Target x Binding Site x Number of Gibbs Models] |
| dSf\_Complement | 3D Matrix with thermodynamic-kinetic transition-state binding/forward reaction entropy for all DNA target complementary strand binding reactions at any DNA target probe binding site using Gibbs nearest neighbor model 7, transition-state forward reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [Target x Binding Site x 3 ] |
| dSr\_Complement | 3D Matrix with thermodynamic-kinetic transition-state unbinding/reverse reaction entropy for all DNA target complementary strand binding reactions at any DNA target probe binding site using Gibbs nearest neighbor model 7, transition-state reverse reaction at the lower bound, mean, and upper bound of model parameter uncertainties.  Matrix is stored as [Target x Binding Site x 3 ] |
| Tm\_Complement | 3D Matrix with thermodynamic-kinetic equilibrium melting temperature for all DNA target complementary strand binding for any targets at any DNA target binding site, all calculated for a basic non-nearest model Tm equation and Tm equations for all nearest-neighbor models in TrueProbes software.  Matrix is stored as [Target x Binding Site x Number of Gibbs Models+1] |

## Outputs in (GeneName)\_AccessionID\_TmT\_BasicDesignerStats\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Off\_Score | Off-Target Score combining the number of off-targets with/without expression level times log10(binding affinities of probes binding off-targets). Higher is worse. |
| Specificity\_Score | Off/On-Target Specificity Score combining the number of off-targets with/without expression level  times relative binding affinity of log10(probe off-target binding affinity/probe on-target binding affinity). Higher is worse. |
| NumRNAOffTargetOptions | List of the unique numbers of RNA off-target species that tiled probes could bind. |
| Probes\_WithNRNAOFF | Cell array with a list of probes that bind each unique number of RNA off-target species. Probes\_WithNRNAOFF{nth} = probes with the nth lowest number of RNA off-target species they bind. |
| NumDNAOffTargetOptions | List of the unique numbers of DNA off-target species that tiled probes could bind. |
| Probes\_WithNDNAOFF | Cell array with a list of probes that bind each unique number of DNA off-target species. Probes\_WithNDNAOFF{nth} = probes with the nth lowest number of DNA off-target species they bind. |
| Nvec\_RNAsingle | A list for each probe ID, the number of off-target RNA species it binds, excluding those that bind more than once.  Nvec\_RNAsingle(probe ID) = number of off-target RNA species that the probe binds to once. |
| Nvec\_RNAmulti | A list for each probe ID, the number of off-target RNA species it binds, including more than once.  Nvec\_RNAmulti(probe ID) = number of off-target RNA species that the probe binds to any number of times. |
| Svec\_RNA | A Cell Array for each probe ID that records the list of sites on RNA off-target IDs that the probe binds.  Svec\_RNA{probe ID} = Target Sites Probe Binds. |
| Tvec\_RNA | A Cell Array for each probe ID that records the list of RNA off-target IDs the probe hits, with multiple hits to the same target repeating for each site within that target they bind. Tvec\_RNA{probe ID} = Targets Probe Binds. |
| Tvec\_logKOFF\_RNA | A Cell Array for each probe ID that records the log10(binding affinity) of probes binding RNA off-target IDs, with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKOFF\_RNA{probe ID} = log10(Probe off-target Binding Affinity). |
| Tvec\_logKOFFdivON\_RNA | A Cell Array for each probe ID that records the log10(binding affinity) of probes binding RNA off-target IDs relative to the probe log10(on-target binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKOFFdivON\_RNA{probe ID} = log10(Probe off-target Binding Affinity/Probe on-target Binding Affinity). |
| Tvec\_logKONdivOFF\_RNA | A Cell Array for each probe ID that records the log10(on-target binding affinity) relative to the probes' log10(RNA off-target IDs binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKONdivOFF\_RNA{probe ID} = log10(Probe off-target Binding Affinity/Probe on-target Binding Affinity). |
| NTPvec\_RNAsingle | A list for each off-RNA target ID, the number of unique probe IDs it binds, excluding those that bind more than once.  NTPvec\_RNAsingle(off-target RNA id) = number of probe IDs that the probe binds to once. |
| NTPvec\_RNAmulti | A list for each off-RNA target ID, the number of unique probe IDs it binds, including more than once.  NTPvec\_RNAmulti(off-target RNA id) = number of probe IDs that the probe binds to any number of times. |
| TSvec\_RNA | A Cell Array for each RNA off-target ID that records the list of binding sites on RNA off-target IDs that probes that bind it will bind, with multiple hits to the same target repeating for each site within that target they bind.  TSvec\_RNA{RNA off-target id} = Target Sites Probe Binds. |
| TPvec\_RNA | A Cell Array for each RNA off-target ID that records the list of probe IDs, with multiple hits to the same target repeating for each site within that target they bind.  TPvec\_RNA{RNA off-target id} = Probe IDs that Bind Target. |
| TPvec\_logKOFF\_RNA | A Cell Array for each RNA off-target ID that records the log10(binding affinity) of probes binding that ID, with multiple hits to the same target repeating for each site within that RNA off-target they bind. In a matrix form, where each row is the binding affinity of the probe binding event 1xN, where N is the total number of probe binding events on the RNA off-target ID.  TPvec\_logKOFF\_RNA{RNA off-target ID} = log10(probe ID RNA off-target Binding Affinity). |
| TPvec\_logKOFFdivON\_RNA | A Cell Array for each RNA off-target ID that records the log10(binding affinity) of probes binding that ID relative to log10(probe on-target binding affinity), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 binding affinity off/on-target ratio of the probe binding event 1xN, where N is the total number of probe binding events on the RNA off-target ID.  TPvec\_logKOFFdivON\_RNA{RNA off-target ID} = log10(probe ID RNA off-target Binding Affinity/probe ID on-target Binding affinity for all probe IDs that bind RNA off-target ID). |
| TPvec\_logKONdivOFF\_RNA | A Cell Array for each RNA off-target ID that records the for probes that bind that RNA off-target the log10(on-target binding affinity) of probes binding that ID relative to log10(probe off-target binding affinity at RNA off-target ID), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 binding affinity on/off-target ratio of the probe binding event 1xN, where N is the total number of probe binding events on the RNA off-target ID.  TPvec\_logKONdivOFF\_RNA{RNA off-target ID} = log10(probe ID on-target Binding Affinity/probe ID RNA off-target Binding affinity for all probe IDs that bind RNA off-target ID). |
| Nvec\_DNAsingle | A list for each probe ID, the number of off-target DNA species it binds, excluding those that bind more than once.  Nvec\_DNAsingle(probe ID) = number of off-target DNA species that the probe binds to once. |
| Nvec\_DNAmulti | A list for each probe ID, the number of off-target DNA species it binds to, including more than once.  Nvec\_DNAmulti(probe ID) = number of off-target DNA species that the probe binds to any number of times. |
| Svec\_DNA | A Cell Array for each probe ID that records the list of sites on DNA off-target IDs that the probe binds.  Svec\_DNA{probe ID} = Target Sites Probe Binds |
| Tvec\_DNA | A Cell Array for each probe ID that records the list of DNA off-target IDs the probe hits, with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_DNA{probe ID} = Targets Probe Binds. |
| Tvec\_logKOFF\_DNA | A Cell Array for each probe ID that records the log10(binding affinity) of probes binding DNA off-target IDs, with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKOFF\_DNA{probe ID} = log10(Probe off-target Binding Affinity). |
| Tvec\_logKOFFdivON\_DNA | A Cell Array for each probe ID that records the log10(binding affinity) of probes binding DNA off-target IDs relative to the probe log10(on-target binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKOFFdivON\_DNA{probe ID} = log10(Probe off-target Binding Affinity/Probe on-target Binding Affinity). |
| Tvec\_logKONdivOFF\_DNA | A Cell Array for each probe ID that records the log10(on-target binding affinity) relative to the probes' log10(DNA off-target IDs binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKONdivOFF\_DNA{probe ID} = log10(Probe off-target Binding Affinity/Probe on-target Binding Affinity). |
| Tvec\_logKOFFdivCOMP\_DNA | A Cell Array for each probe ID that records the log10(binding affinity) of probes binding DNA off-target IDs relative to the DNA off-target complementary strand log10(off-target complementary strand binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logKOFFdivCOMP\_DNA{probe ID} = log10(Probe off-target Binding Affinity/off-target complementary strand Binding Affinity). |
| Tvec\_logCOMPdivKOFF\_DNA | A Cell Array for each probe ID that records the log10(DNA off-target complementary strand binding affinity) relative to the probes' log10(DNA off-target IDs binding affinity), with multiple hits to the same target repeating for each site within that target they bind.  Tvec\_logCOMPdivKOFF\_DNA{probe ID} =  log10(off-target complementary strand Binding Affinity/Probe off-target Binding Affinity).  It is empty if the probe has no DNA off-targets. |
| NTPvec\_DNAsingle | A list for each off-DNA target ID, the number of unique probe IDs it binds, excluding those that bind more than once.  NTPvec\_DNAsingle(off-target DNA ID) = number of probe IDs that the probe binds to once. |
| NTPvec\_DNAmulti | A list for each off-DNA target ID, the number of unique probe IDs it binds, including more than once.  NTPvec\_DNAmulti(off-target DNA ID) = number of probe IDs that the probe binds to any number of times. |
| TSvec\_DNA | A Cell Array for each DNA off-target ID that records the list of binding sites on DNA off-target IDs that probes that bind it will bind, with multiple hits to the same target repeating for each site within that target they bind.  TSvec\_DNA{DNA off-target ID} = Target Sites Probe Binds |
| TPvec\_DNA | A Cell Array for each DNA off-target ID that records the list of probe IDs, with multiple hits to the same target repeating for each site within that target they bind.  TPvec\_DNA{DNA off-target ID} = Probe IDs that Bind Target. |
| TPvec\_logKOFF\_DNA | A Cell Array for each DNA off-target ID that records the log10(binding affinity) of probes binding that ID, with multiple hits to the same target repeating for each site within that DNA off-target they bind. In a matrix form, where each row is the binding affinity of the probe binding event 1xN, where N is the total number of probe binding events on the DNA off-target ID.  TPvec\_logKOFF\_DNA{DNA off-target ID} =  log10(probe ID DNA off-target Binding Affinity). |
| TPvec\_logKOFFdivON\_DNA | A Cell Array for each DNA off-target ID that records the log10(binding affinity) of probes binding that ID relative to log10(probe on-target binding affinity), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 binding affinity off/on-target ratio of the probe binding event 1xN, where N is the total number of probe binding events on the DNA off-target ID.  TPvec\_logKOFFdivON\_DNA{DNA off-target ID} =  log10(probe ID DNA off-target Binding Affinity/probe ID on-target Binding affinity for all probe IDs that bind DNA off-target ID). |
| TPvec\_logKONdivOFF\_DNA | A Cell Array for each DNA off-target ID that records the for probes that bind that DNA off-target the log10(on-target binding affinity) of probes binding that ID relative to log10(probe off-target binding affinity at DNA off-target ID), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 binding affinity on/off-target ratio of the probe binding event 1xN, where N is the total number of probe binding events on the DNA off-target ID.  TPvec\_logKONdivOFF\_DNA{DNA off-target ID} =  log10(probe ID on-target Binding Affinity/probe ID DNA off-target Binding affinity for all probe IDs that bind DNA off-target ID). |
| TPvec\_logKOFFdivCOMP\_DNA | A Cell Array for each DNA off-target ID that records the log10(binding affinity) of probes binding that ID relative to log10(target complementary strand binding affinity), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 of the off-target binding affinity to the complementary-target binding affinity ratio of the probe binding event 1xN, where N is the total number of probe binding events on the DNA off-target ID.  TPvec\_logKOFFdivCOMP\_DNA{DNA off-target ID} =  log10(probe ID DNA off-target Binding Affinity/DNA off-target complementary strand Binding affinity for all probe IDs that bind DNA off-target ID). |
| TPvec\_logCOMPdivKOFF\_DNA | A Cell Array for each DNA off-target ID that records the for probes that bind that DNA off-target the log10(off-target complementary strand DNA off-target binding affinity) of probes binding that DNA off-target ID relative to log10(probe off-target binding affinity at DNA off-target ID), with multiple hits to the same target repeating for each site within that off-target they bind. In a matrix form, where each row is the log10 binding affinity complementary/off-target ratio of the probe binding event 1xN, where N is the total number of probe binding events on the DNA off-target ID.  TPvec\_logKONdivOFF\_DNA{DNA off-target ID} =  log10(DNA off-target complementary strand Binding Affinity/probe ID DNA off-target Binding affinity for all probe IDs that bind DNA off-target ID). |

## Outputs in (GeneName)\_AccessionID\_chosen.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| chosenProbes | The probe tile positions of all final designed probes in the order in which TrueProbes designs them. |

## Outputs in (GeneName)\_AccessionID\_probes\_final\_MaxProbes\_max.xlsx

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Probe Target | Target Accession ID |
| Transcript Name | Name of Target |
| Base Position Start | First nt position where the sequence starts on the target |
| Base Position End | Last nt position where the sequence occurs on the target |
| Probe Sequence | Probe 5’->3’ Oligonucleotide Sequence |
| Probe Length | Number of nucleotides in the probe |
| GC Fraction | Probe GC Percentage |
| Probe Tm | Probe Melting Temperature in Celsius |
| Number of Isoform-Agnostic On-Target Matches | The number of Target Isoforms the probe binds |
| Number of Isoform-Specific On-Target Matches | The number of times the probe binds the desired target Accession ID |
| Off-Score | Off-Target Score combining the number of off-targets with/without expression level and binding affinities of off-targets relative to on-targets. |
| Total Off-Target Matches | The total number of off-target Matches probe has. |
| Number of X Off-Target Matches for X max probe length to Min homology length | Columns for the number of off-targets at each length from the max probe length to the min homology search size. |

## Outputs in (GeneName)\_AccessionID\_TmT\_ModelMetrics\_SW.mat

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| PackingEfficiency | The Number of Probes Designed Relative to the Maximum Number of Probes that Could be Designed Given Probe Length Range, Min Spacing Between Probes, and Max Probes Cutoff. |
| ProbeSetMetrics.Ns\_Config | The maximum number of distinct self-hybridization binding sequences a designed probe has. |
| ProbeSetMetrics.Nc\_Config | The maximum number of distinct cross-dimerization binding sequences a probe has with another probe. |
| ProbeSetMetrics.Js\_RNA | A Handle function that, when given probe IDs for a list of tiled probes, returns the unique target IDs of all RNA targets they bind.  Js\_RNA(Probe Set) = IDs of RNA Targets |
| ProbeSetMetrics.Js\_DNA | A Handle function that, when given probe IDs for a list of tiled probes, returns the unique target IDs of all DNA targets they bind.  Js\_RNA(Probe Set) = IDs of RNA Targets |
| ProbeSetMetrics.Js\_Sites | A Handle function that, when given probe IDs for a list of tiled probes, returns the binding site numbers across the targets they bind. Not ordered by each target, but the unique list of binding site positions.  Js\_Sites(Probe Set) = Unique Target Binding Sites |
| ProbeSetMetrics. SolutionModels | The vector for which Gibbs Model results are evaluated. |
| ProbeSetMetrics.  SolutionTemperatures | The vector for which Hybridization Temperatures (°C) results are evaluated. |
| ProbeSetMetrics.  SolutionDilutions | The vector for the probe concentration dilutions results is evaluated. |
| ProbeSetMetrics.  SolutionCellLines | The vector of which Cell Line rows in the Expression Matrix all predictions and on/off-target quantifications are generated for. |
| ProbeSetMetrics.  ModelSolverFunctions\_linIndex | List of Handle Functions used in solving binding equilibrium and probe binding quantification with only indexing binding sites that exist to be most efficient quantification. |
| ProbeSetMetrics.  ModelSolverFunctions\_5D | List of Handle Functions used in solving binding equilibrium and probe binding quantification with vector variable inputs of [Probe x Target x Site x Model x Cell Line] |
| ProbeSetMetrics.  ModelSolverFunctions\_7D | List of Handle Functions used in solving binding equilibrium and probe binding quantification with vector variable inputs of [Probe x Target x Site x Model x Temperature x Dilution x Cell Line]. |
| ProbeSetMetrics.iter | The number of iterations of the binding equilibrium model solver needed to get the binding equilibrium at the specified tolerance level. |
| ProbeSetMetrics.err | The total error of the model equilibrium solution. |
| ProbeSetMetrics.varSSE | The variance in the sum of squares errors at the model equilibrium solution |
| ProbeSetMetrics.eqSSE | The sum of squares error at the model equilibrium solution. |
| ProbeSetMetrics.CProbes\_Free | The equilibrium solution free probe concentration for a cell array of {gibbs\_model,temperature,dilution}. |

### ModelMetrics.ProbeSetMetrics.BindingPredictions

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| P\_TargetSites\_Bound  \_ModelTemperature  DilutionVector | For a cell array of {Gibbs model, temperature, dilution, the Probability that each probe in the final probe set binds a target site in the format of [Probe x Target x Site x 1 x Cell Line]. |
| C\_TargetSites\_Bound  \_ModelTemperature  DilutionVector | For a cell array of {Gibbs model, temperature, dilution, the concentration at which each probe in the final probe set binds to a target site is given in the format [Probe x Target x Site x 1 x Cell Line]. |
| C\_OnOtherOff\_nBound  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the concentration that on-targets, other on-target isoforms, and off-targets have probes bound across all target sites in the format of [Target Type x Number of Site x Cell Line] with TargetType = [desired on-target isoform, other on-target isoforms, off-targets] |
| Con\_Distribution  \_ModelTemperature  DilutionVector | For a cell array of {Gibbs model, temperature, dilution, the concentration of the desired on-target isoform with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Coff\_Distribution  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the concentration of off-targets with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Cother\_Distribution  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the concentration of other on-target isoforms with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Pon\_Distribution  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of the desired on-target isoform with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Poff\_Distribution  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of off-targets with n probes bound for each cell line.  It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Pother\_Distribution  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of other on-target isoforms with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Non\_Count  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the total RNA Spot Count of the desired on-target isoform for each cell line. It is a 2D Matrix with dimensions [1 x Number of Cell Lines]. |
| Noff\_Count  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the total RNA Spot Count of off-targets for each cell line. It is a 2D Matrix with dimensions [1 x Number of Cell Lines]. |
| Nother\_Count  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the total RNA Spot Count of other on-target isoforms. It is a 2D Matrix with dimensions [1 x Number of Cell Lines]. |
| Non\_history  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of the desired on-target isoform with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Nother\_history  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of other on-target isoforms with n probes bound for each cell line. It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Noff\_history  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of off-targets with n probes bound for each cell line.  It is a 2D Matrix with dimensions [Number of Target Sites x Number of Cell Lines]. |
| Noff\_tot\_history  \_ModelTemperature  DilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the total number of probes bound across all off-target RNA for each cell line. It is a 2D Matrix with dimensions [1 x Number of Evaluated Cell Lines]. |
| IsoIgnorantConfusion\_Probe\_P\_ModelTemperature  DilutionVector | Confusion matrix metrics at all probe thresholds for different methods when only considering the desired on-target isoform (on) and off-targets, excluding other isoforms of the on-target (off) in units of the number of probes simultaneously bound to individual targets. |
| IsoSpecificConfusion\_Probe\_P\_ModelTemperature  DilutionVector | Confusion matrix metrics at all probe thresholds for different methods, considering only the desired on-target isoform (on) and off-targets, including other isoforms of the on-target (off) in units of the number of probes simultaneously bound to individual targets. |
| IsoAgnosticConfusion\_Probe\_P\_ModelTemperature  DilutionVector | Confusion matrix metrics at all probe thresholds for different methods, considering only on-target isoforms (on) and off-targets (off) in units of the number of probes simultaneously bound to individual targets. |

#### Confusion Matrix Outputs

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| T | True Positive + True Negative |
| F | False Positive + False Negative |
| AP | Actual Positive = True Positive + False Negative |
| AN | Actual Negative = False Positive + True Negative |
| PP | Predicted Positive = True Positive + False Positive |
| PN | Predicted Negative = True Negative + False Negative |
| TP | True Positive = hit |
| FN | False Negative = miss = underestimation |
| TN | True Negative = correct rejection |
| FP | False Positive = false alarm = overestimation |
| TPR | True Positive Rate = Sensitivity = recall = TP/AP |
| FNR | False Negative Rate = Miss rate = type II error = FN/AP |
| TNR | True Negative Rate = Specificity = Selectivity = TN/AN |
| FPR | False Positive Rate = Fallout = type I error = FP/AN |
| PPV | Positive Predictive Value = Precision = TP/(TP+FP) |
| FDR | False Discovery Rate = FP/(TP+FP) |
| NPV | Negative Predictive Value = TN/(TN+FN) |
| FOR | False Omission Rate = FN/(TN+FN) |
| PLR | Positive Likelihood Ratio = TPR/FPR |
| NLR | Negative Likelihood Ratio = FNR/TNR |
| DOR | Diagnostic Odds Ratio = PLR/NLR |
| Markedness | PPV+NPV-1 |
| Informedness | Bookmakers Informedness = TPR+TNR-1 |
| PREV\_Threshold | Prevalence Threshold  = (sqrt(TPR\*FPR)-FPR)/(TPR-FPR) |
| Prevalence | (TP+FN)/(TP+TN+FP+FN) |
| Accuracy | (TP+TN)/(TP+TN+FP+FN) |
| Balanced Accuracy | (TPR+TNR)/2 |
| Fowlkes Mallows Index | Sqrt(PPV\*TPR) = |
| Jaccard Index | Critical success index = TP/(TP+FN+FP) |
| F1 | F1 Score = 2\*TP/(2\*TP+FP+FN)= |
| Fbeta | Fbeta Score = (1+beta^2)\*TP/((1+beta^2)\*TP+FP+beta^2\*FN) |
| P4 | P4 Score = = |
| MCC | Matthews Correlation Coefficient = |
| CKC | Cohen’s Kappa Coefficient = |
| TOC\_AUC | Total Operators Curve Area Under the Curve |
| ROC\_AUC | Receiver Operator Curve Area Under the Curve |
| PR\_AUC | Precision Recall Curve Area Under the Curve |
| Confusion Matrix Function | Outputs confusion matrix [TP FP;TN FP] for all cell-lines at any given threshold vector. |

### ModelMetrics.ProbeSetMetrics.IntensityPredictions

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| P\_TargetSites\_Bound\_MTDvector | For a [Gibbs model, temperature, dilution] 3D cell array, the Probability that each probe in the final probe set binds a target site in the format of [Probe x Target x Site x 1 x Number of Evaluated Cell Line]. |
| Ioff\_history  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the per-pixel predicted increase in cell background due to off-target probe intensity for each cell line. It is a 2D Matrix with dimensions [1 x Number of Evaluated Cell Lines]. |
| P\_Non\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of the desired on-target isoform with intensity I for each cell line.  It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| P\_Nother\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of other on-target isoforms with intensity I for each cell line.  It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| P\_Noff\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the probability of off-targets with intensity I for each cell line.  It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| Non\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of the desired on-target isoform with intensity I for each cell line.  It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| Nother\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of other on-target isoforms with intensity I for each cell line.  It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| Noff\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the RNA Spot Count of off-targets with intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| QzSignal  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the on-target spot signal predicted cumulative distribution function in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| PzSignal  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the on-target spot signal predicted probability density function in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| QzCellBkg  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the predicted cell autofluorescence background intensity cumulative distribution function, when accounting for probe off-target signal intensity in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| PzCellBkg  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the predicted cell autofluorescence background intensity probability density function, when accounting for probe off-target signal intensity in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| QzSignalMinusBackgd  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the spot signal minus background predicted cumulative distribution function in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| PzSignalMinusBackgd  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the spot signal minus background predicted probability density function in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| P\_Signal\_wAuto\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the desired on-target isoform spot signal predicted probability density function in units of intensity I, accounting for cell autofluorescence intensity for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| P\_SignalOther\_wAuto\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, other on-target isoforms spot signal predicted probability density function in units of intensity I, accounting for cell autofluorescence intensity for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| P\_SignalOffNonAverage\_wAuto\_I  \_ModelTemperatureDilutionVector | For a [Gibbs model, temperature, dilution] 3D cell array, the non-per-pixel averaged predicted off-target spot signal probability density function in units of intensity I for each cell line. It is a 2D Matrix with dimensions [Number of Evaluated Cell Lines x Intensity Value Options]. |
| IsoIgnorantConfusion\_Intensity\_I  \_ModelTemperatureDilutionVector | Confusion matrix metrics at all intensity thresholds for different methods when only considering the desired on-target isoform (on) and off-targets, excluding other isoforms of the on-target (off) in units of spot intensity. |
| IsoSpecificConfusion\_Intensity\_I  \_ModelTemperatureDilutionVector | Confusion matrix metrics at all intensity thresholds for different methods, considering only the desired on-target isoform (on) and off-targets, including other isoforms of the on-target (off) in units of spot intensity. |
| IsoAgnosticConfusion\_Intensity\_I  \_ModelTemperatureDilutionVector | Confusion matrix metrics at all intensity thresholds for different methods, considering only on-target isoforms (on) and off-targets (off) in units of spot intensity. |

### ModelMetrics.ProbeSetMetrics.CountPredictions

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| IsoIgnorant\_SpotCountMetrics\_ModelTemperatureDilutionVector | Spot count metrics at given thresholds for different methods when only considering the desired on-target isoform (on) and off-targets, excluding other isoforms of the on-target (off). |
| IsoSpecific\_SpotCountMetrics\_ModelTemperatureDilutionVector | Spot count metrics at given thresholds for different methods when only considering the desired on-target isoform (on) and off-targets, including other isoforms of the on-target (off). |
| IsoAgnostic\_SpotCountMetrics\_ModelTemperatureDilutionVector | Spot count metrics at given thresholds for different methods when only considering all on-target isoforms (on) and off-targets (off). |

#### Entries in SpotCountMetrics

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| TrueSpot | Spot Count Metrics when using a threshold derived using TrueSpot with the cell-line predicted spot count curve. |
| F1 | Spot Count Metrics when using a threshold that maximizes the predicted F1 Score. |
| F2 | Spot Count Metrics when using a threshold that maximizes the predicted Fbeta Score for a beta value of two. |
| Fhalf | Spot Count Metrics when using a threshold that maximizes the predicted Fbeta Score for a beta value of 1/2. |

##### **Single Cell/Cell-Line Stats Info**

|  |  |
| --- | --- |
| **Variable Names** | **Description** |
| Threshold | The selected threshold Value for each cell line |
| AP | The True Positive Spot Count for each cell line |
| PP | The Predicted Positive Spot Count for each cell line at the picked threshold value |
| TP | The True Positive Spot Count for each cell line at the picked threshold value |
| FN | The False Negative Spot Count for each cell line at the picked threshold value |
| FP | The False Positive Spot Count for each cell line at the picked threshold value |
| FPR | The False Positive Rate for each cell line at the picked threshold value |
| FNR | The False Negative Rate for each cell line at the picked threshold value |
| TPR | The True Positive Rate for each cell line at the picked threshold value |
| TNR | The True Negative Rate for each cell line at the picked threshold value |
| FDR | The False Discovery Rate for each cell line at the picked threshold value |
| FOR | The False Omission Rate for each cell line at the picked threshold value |
| PPV | The Positive Predictive Value for each cell line at the picked threshold value |
| NPV | The Negative Predictive Value for each cell line at the picked threshold value |
| Accuracy | (TP+TN)/(TP+TN+FP+FN) for each cell line at the picked threshold value |
| Balanced Accuracy | The (TPR+TNR)/2 for each cell line at the picked threshold value |
| F1 | The F1 Score for each cell line at the picked threshold |
| P4 | The P4 Score for each cell line at the picked threshold |
| MCC | The Matthews Correlation Coefficient for each cell line at the picked threshold |
| CKC | Cohen’s Kappa Coefficient for each cell line at the picked threshold |
| Fowlkes  Mallow  Index | The Fowlkes-Mallow Index for each cell line at the picked threshold |

# 9. How to Set Up TrueProbes for a New Custom Organism

## Folder Creation

First, create folders where all the database data for the new organism will be stored within the TrueProbes data folder. You will need to create a BLAST Database, GTF, GFF, and Expression Data folder for your organism under the name of ENSEMBL or NCBI annotation. For this, the Organism is highlighted in **Red** and serves as a placeholder for your organism's name.

### NCBI

|  |  |
| --- | --- |
| **Category** | **Description** |
| BLAST Database | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq |
| GTF Database | TrueProbes/data/DatabaseData/GTF\_Databases/**Organism**/NCBI\_RefSeq |
| GFF Database | TrueProbes/data/DatabaseData/GFF3\_Databases/**Organism**/NCBI\_RefSeq |
| Expression  Folder | TrueProbes/data/DatabaseData/GeneExpressionData/**Organism** |

### ENSEMBL

|  |  |
| --- | --- |
| **Category** | **Description** |
| BLAST Database | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI |
| GTF Database | TrueProbes/data/DatabaseData/GTF\_Databases/**Organism**/EMBL\_EBI |
| GFF Database | TrueProbes/data/DatabaseData/GFF3\_Databases/**Organism**/EMBL\_EBI |
| Expression Data Folder | TrueProbes/data/DatabaseData/GeneExpressionData/**Organism** |

## File Acquisition

To design probes, several files will need to be acquired or added to the newly created database folders.

### NCBI

For NCBI, first go to <https://www.ncbi.nlm.nih.gov/datasets/genome/> and search for your organism, then click on the genome assembly that is denoted as Annotation NCBI RefSeq. After that, either click on the FTP link and download each type of file listed in the table below or click 'Download' and check the boxes for Genome sequences (FASTA), Annotation Features (GTF), Annotation features (GFF), and Transcripts (FASTA) for RefSeq or GenBank. Ensure that you unzip each file you receive. In the following table, three placeholders are used: **taxonomic\_group** standsfor the taxonomic classification the species falls under, **#** stands for the ENSEMBL release number for mapping between NCBI and ENSEMBL transcripts, and **Organism** for the species you want to design probes against.

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Where to Acquire the Files** | **File To Acquire** | **Where To Store a File** |
| Reference Genome | <https://www.ncbi.nlm.nih.gov/datasets/genome/> Or go to  https://ftp.ncbi.nlm.nih.gov/genomes/refseq/**taxonomic\_group**/**Organism** | File ending in  \*\_genomic.fna.gz | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq |
| Reference  Transcriptome | <https://www.ncbi.nlm.nih.gov/datasets/genome/> Or go to  https://ftp.ncbi.nlm.nih.gov/genomes/refseq/**taxonomic\_group**/**Organism** | File ending in  \*\_rna.fna.gz | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq |
| GTF Database | <https://www.ncbi.nlm.nih.gov/datasets/genome/> Or go to  https://ftp.ncbi.nlm.nih.gov/genomes/refseq/**taxonomic\_group**/**Organism** | File ending in  \*\_genomic.gtf.gz | TrueProbes/data/DatabaseData/GTF\_Databases/**Organism**/NCBI\_RefSeq |
| GFF Database | <https://www.ncbi.nlm.nih.gov/datasets/genome/> Or go to  https://ftp.ncbi.nlm.nih.gov/genomes/refseq/**taxonomic\_group**/**Organism** | File ending in  \*\_genomic.gff.gz | TrueProbes/data/DatabaseData/GFF3\_Databases/**Organism**/NCBI\_RefSeq |
| EMBL to NCBI RefSeq | Click the link to other annotations (TSV)  Or go to  https://ftp.ensembl.org/pub/**release-#**/tsv/**Organism** | File ending in  \*.**#**.refseq.tsv.gz | TrueProbes/data/DatabaseData/ENSEMBL\_NCBI\_StableIDs/ |

### ENSEMBL

For ENSEMBL, first visit the FTP site for the current release at <https://useast.ensembl.org/info/data/ftp/index.html> and search for your organism. Then, click on the link for each type of file needed on the table below. Ensure that you unzip each file you receive. In the following table, two placeholders are used: **#** stands for the release number, and **Organism** for the species you want to design probes against.

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Where to Acquire the Files** | **File To Acquire** | **Where To Store a File** |
| Reference Genome | Click the link DNA (FASTA)  Or go to  https://ftp.ensembl.org/pub/**release-#**/fasta/**Organism**/dna/  Or go to  https://ftp.ensembl.org/pub/current/fasta/**Organism**/dna | File ending in \*.dna.primary\_assembly.fa.gz,  or if that is not there  \*.dna.toplevel.fa.gz | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI |
| Reference  Coding  Transcriptome | Click the link cDNA (FASTA)  Or go to  https://ftp.ensembl.org/pub/**release-#**/fasta/**Organism**/cdna  Or go to  https://ftp.ensembl.org/pub/current/fasta/**Organism**/cdna | File ending in  \*.cdna.all.fa.gz | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI |
| Reference Non-coding Transcriptome | Click the link ncRNA (FASTA)  Or go to  https://ftp.ensembl.org/pub/**release-#**/fasta/**Organism**/ncrna  Or go to  https://ftp.ensembl.org/pub/current/fasta/**Organism**/ncrna | File ending in  \*.ncrna.fa.gz | TrueProbes/data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI |
| GTF Database | Click the link Gene Sets (GTF)  Or go to  https://ftp.ensembl.org/pub/**release-#**/gtf/**Organism**  Or go to  https://ftp.ensembl.org/pub/current/gtf/**Organism** | File ending in  \*.**#**.gtf.gz | TrueProbes/data/DatabaseData/GTF\_Databases/**Organism**/EMBL\_EBI |
| GFF Database | Click the link Gene Sets (GFF3)  Or go to  https://ftp.ensembl.org/pub/**release-#**/gff3/**Organism**  Or go to  https://ftp.ensembl.org/pub/current/gff3/**Organism** | File ending in  \*.**#**.gff3.gz | TrueProbes/data/DatabaseData/GFF3\_Databases/**Organism**/EMBL\_EBI |
| EMBL to NCBI RefSeq | Click the link to other annotations (TSV)  Or go to  https://ftp.ensembl.org/pub/**release-#**/tsv/**Organism**  Or go to  https://ftp.ensembl.org/pub/current/tsv/**Organism** | File ending in  \*.**#**.refseq.tsv.gz | TrueProbes/data/DatabaseData/ENSEMBL\_NCBI\_StableIDs/ |

## Settings Specification

Update Parameter Files with the locations of the newly added files.

Add any organism reference gene expression files, schema, and column names to GeneExpressionDataFileLocation.xml: [XML File with location of all gene expression files, schema, and sample label for all reference gene expression desired in design]

### Update Database Locations XML File

Add organism file locations to DatabaseLocations.xml: [XML File with location of all database files needed in design when using NCBI or ENSEMBL reference genome, for any potential organism designed against] and information for converting between Gene Expression Data Annotated using NCBI or ENSEMBL, when using a different reference type, get the conversion file from ENSEMBL Other Annotations RefSeq TSV (<https://useast.ensembl.org/info/data/ftp/index.html>)

#### EMBL to NCBI

|  |  |
| --- | --- |
| **Name** | **Entry** |
| Organism | **Organism** |
| Stable IDs | Data/DatabaseData/ENSEMBL\_Stable\_IDs/**\***.refseq.tsv |

#### EMBL

|  |  |
| --- | --- |
| **Name** | **Entry** |
| Organism | **Organism** |
| Root\_FASTA | Data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI/ |
| BLASTDB\_DNA | Data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI/**Organism**\_ENSEMBL\_genomic |
| BLASTDB\_RNA | Data/DatabaseData/Blast\_Databases/**Organism**/EMBL\_EBI/**Organism**\_ENSEMBL\_transcript |
| GTF | Data/DatabaseData/GTF\_Databases/**Organism**/EMBL\_EBI/**\***.gtf |
| GFF | Data/DatabaseData/GFF3\_Databases/**Organism**/EMBL\_EBI/**\***.gff3 |

#### NCBI

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | **Organism** |
| Root\_FASTA | Data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq/ |
| BLASTDB\_DNA | Data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq/**Organism**\_NCBI\_genomic |
| BLASTDB\_RNA | Data/DatabaseData/Blast\_Databases/**Organism**/NCBI\_RefSeq/**Organism**\_NCBI\_transcript |
| GTF | Data/DatabaseData/GTF\_Databases/**Organism**/NCBI\_RefSeq/**\***.gtf |
| GFF | Data/DatabaseData/GFF3\_Databases/**Organism**/NCBI\_RefSeq/**\***.gff |

### Update Gene Expression File Locations XML File

Add any organism reference gene expression files, schema, and column names to GeneExpressionDataFileLocation.xml: [XML File with location of all gene expression files, schema, and sample label for all reference gene expression desired in design]

#### Gene Expression File Locations

| **Name** | **Description** |
| --- | --- |
| Organism | **Organism** |
| “Data Identifier Name” | Data/DatabaseData/Gene\_Expression\_Data/**Organism**/**\*** |

#### Gene Expression Schema

|  |  |
| --- | --- |
| **Name** | **Description** |
| Organism | **Organism** |
| “Data Identifier Name” | Data/DatabaseData/Gene\_Expression\_Data/**Organism**/**\*** |

# 10. Custom Application-Specific Setup

# 11. Appendices