**Required Software Packages**

**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+**

**Modules of Probe Design**

**TrueProbe Probe Design Requirements by Design Step**

**• A0 Probe Generation**

**• MATLAB**

**• Bioinformatics Toolbox**

**• A1 Probe Blasting**

**• MATLAB**

**• Bioinformatics Toolbox**

**• Parallel Computing Toolbox**

**• A2 Gene Expression Information**

**• MATLAB**

**• Parallel Computing Toolbox**

**• Statistics and Machine Learning Toolbox**

**• A3 Thermodynamic Information**

**• MATLAB**

* **Bioinformatics Toolbox**
* **Parallel Computing Toolbox**
* **Statistics and Machine Learning Toolbox**
* **Symbolic Math Toolbox**

**• A4 Binding Site Mapping**

**• MATLAB**

* **Bioinformatics Toolbox**
* **Parallel Computing Toolbox**
* **Statistics and Machine Learning Toolbox**
* **Symbolic Math Toolbox**

**• A5 Probe Designer Stats**

**• MATLAB**

**• Parallel Computing Toolbox**

**• A6 Probe Selection**

**• MATLAB**

**• Bioinformatics Toolbox**

**• Parallel Computing Toolbox**

**• A7 Get Metrics**

**• MATLAB**

**• Symbolic Math Toolbox**

**• Statistics and Machine Learning Toolbox**

**• Bioinformatics Toolbox**

**• Signal Processing Toolbox**

**• A8 ProbeSpecificityFiltering**

**• MATLAB**

**• Bioinformatics Toolbox**

**• A9 MultiFluorphoreColocalizationConvolutionND**

**• MATLAB**

* **Bioinformatics Toolbox**
* **Symbolic Math Toolbox**

**• Image Processing Toolbox**

**How to Install TrueProbes**

1. **Install MATLAB and product packages.**

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

1. **Install NCBI-BLAST+**

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

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

1. **Ensure your computer security allows blast+ files to run**

**Compatibility** **Make sure the blast+ folder files are allowed to run by your computer security, i.e., permission to run in MacOS or any security or virus scanners allow the program to run and do not block it.**

**For MacOSX: Go To Privacy and Security to enable blastn, makeblastdb, and blastdbcmd to run on Mac**

**How to Run the TrueProbes Software.**

* 1. **A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5** is the main script for running the software. The design file needs to be run with two inputs (**id and cluster**), with a table (**inputs1**) describing each target gene (inputs1), and a set of settings describing how the probe design will be run (settings)

The code is run via the command line via

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

**1st Input Argument id: id** is an integer and is the row of the input design table file to run and design probes against.

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

* 1. **The TrueProbes software runs by performing eight steps sequentially.**

**1) Probe Generation.** It generates all possible probes shared between a list of inclusion IDs and inclusion text files, but not in exclusion text files or exclusion IDs, within a set probe length range.

**2) BLAST Alignment.** All probes target hits in the reference genome and/or transcriptome are identified at least as long as the minimum homology length.

**3) BLAST Target Gene/Transcript Expression.** The gene expression and transcript expression levels are collected for reference expression databases specified 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 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.** Sort probes by with/without expression data, the number of off-targets, and then by difference in on-target binding to off-target binding and secondary structure binding affinity to iteratively design probes. Print out the list in an Excel spreadsheet.

**8) Model Evaluation.** The final probe set and reference expression are combined to compute equilibrium probe binding and statistics, cumulative off-target binding, on-target binding, etc., when reference values for cell size, probe concentration, and probe intensity are given.

* 1. TrueProbes Probe Design uses four main input files for specifying probe design

**TrueProbes\_DesignTargets.csv:** [CSV File with list of all design targets]

**ProbeDesignSettings\_Parameters.xml:** [XML File where all design settings are specified]

**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]

**GeneExpressionDataFileLocation.xml:** [XML File with location of all gene expression files, schema, and sample label for all reference gene expression desired in design]

* 1. Targets for probe design are specified in input file TrueProbes\_DesignTargets.csv

Input table CSV file has five columns:

1. Organism to design probes for.

2. Included target accession IDs. Designs probes shared across all accession numbers

3. Excluded target accession IDs. Removes probes in exclusion accession numbers.

3. Text Sequence Files to Include (files). Default empty

4. Text Sequence Files to Exclude (files). Default empty

* 1. **Parameters are specified in TrueProbes\_ParameterSettings.xml and are grouped into different categories**
     1. Main Probe Settings

**minProbeSize**:[min nt length of potential probes, default 20]

**maxProbeSize**: [max nt length of potential probes, default 20]

**MinProbeSpacing** [min spacing between probes, default 3]

**MaxNumberOfProbes:** [Max number of probes to design, default 96]

**targetStrand:** [which strand to of target to design probes against, default 1 for ‘plus’, with 0 for ‘minus’]

**MinHomologySearchTargetSize:**[Minimum homology length for BLAST alignments to be recorded and used in probe design and evaluation, default 15]

**BLASTrna**:[decide to blast RNA sequences, default 1]

**BLASTdna**:[decide to blast DNA sequences, default 0]

**ExpressionReferenceForProbeDesign**:[which row across all expression reference files to use in probe design, with 0 meaning to not use expression data to design probes, default 0]

* + 1. Thermodynamic Settings

**Gibbs\_model:**[Which thermodynamic model to use for probe design and evaluation, default 4]

Model 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

Model 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

Model 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

Model 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

Model 5:SantaLucia04. SantaLucia Jr, J., and Hicks, D. (2004) The thermodynamics of DNA structural motifs Annu Rev Biophys Biomol Struct 33, 415-440

Model 6: Allawi97. Allawi, H. T., and SantaLucia, J. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA Biochemistry 36, 10581-10594

Model 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

Model 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

**HybridizationTemperature**:[Hybridization temperature, default 37C]

**HeatCapacityReferenceTemperature**:[Reference temperature for Cp measurement and Gibbs model, default 37C]

**SaltConcentration:**[Salt Concentration M, default 0.05]

**PrimerConcentration:**[Primer Concentration M, default 50e-6]

**RemoveMisMatches:**[Remove sequence mismatched base pairs before evaluating probe-target binding affinity, default 1, including adds flanking sequences to alignments and uses Gibbs model 8 with mismatch base pair inclusion]

* + 1. Design Filtering Settings

**RemoveProbesBindingOffTargetRibosomalHits**:[Filter out probes with targets to off-target ribosomal proteins, default 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, default 1]

**IncludeSelfHybridizationInProbeSelection:[**When designing probes consider probe self-hybridization when ranking probes based on binding affinity, default 1]

* + 1. Parallelization Settings

**Parallization\_probeBatchSize:** [number of probes to evaluate in a single batch when performing parallelized calculations, default 20]

**Parallization\_targetBatchSize:** [number of targets to evaluate in a single batch when performing parallelized calculations, default 200]

**ParsingPreference**:[blast simultaneously in parallel(1) or blast probes sequentially (0), default 1]

* + 1. Make Blast Database Settings

**Parse\_seqids:** [when making blastdb in software decision to parse sequence ids from fasta files as blastdb ids, useful for blastdbcmd sequence retrieval, default true]

**Hash\_index:**[when making blastdb make sequence hash indexes leads to faster exact match retrieval but less accurate range matches, default false]

* + 1. Blastn settings

**evalue:**[Expectation value cutoff, default 1000]

**Gapextend:** [Cost to extend a gap (integer), default 2]

**Gapopen:**[cost to open a gap (integer), default 5]

**Num\_alignments:**[number of database sequences to show num\_alignments for, default 1000]

**Penalty:**[penalty for a nucleotide mismatch, default -3, at most zero]

**Reward:[**reward for a nucleotide match, default 1, at least zero]

**Word\_size:** [word size for wordfinder algorithm, default 7, at least two]

**Dust:**[filter query sequences with DUST, default no]

* + 1. Gene Expression Settings

**DoAllGenesHaveSameExpression:** [decide to assume equal expression for all genes (1) or to use gene expression reference (0) , default 1]

**UseGeneOverTranscLevelExpression:**[use gene level (1) or transcript isoform level (0) gene expression values, default 0]

**UseRegularDNAExpression:**[(0) use DNA expression from gene expression track in expression data, (1) set expression to 2 for DNA, default 1]

**nullRNAcopynumber:**[number of RNA copy when not using reference expression levels, default 100]

**nullDNAcopynumber**:[number of DNA copy number when not using reference expression levels, default 2]

**TMM\_LogRatioTrim**:[when normalizing TPM expression data using TMM set log ratio trim threshold cutoff, default 0.3]

**TMM\_SumTrim**:[when normalizing TPM expression data using TMM set sum trim threshold cutoff, default 0.05]

**TMM\_Acutoff**:[when normalizing TPM expression data using TMM set A cutoff value, default -1e10]

**TMM\_doWeighting**:[when normalizing TPM expression data weight terms using inverse of approximate asymptotic variance of the M-values to account for genes with higher read counts having lower variance on log scale and more reliable mean estimation, default 1]

* + 1. Model Simulation Settings

**removeUndesiredIsoformsFromPrediction**: [when evaluating main on/off-target binding should alternate isoforms of desired targets be removed and not included in off-target quantification, default 1]

**ProbeConcentration\_MicroMolar**:[probe concentration in uM, default 5e-6]

**CellRadius\_Micron: [**cell radius in microns for converting RNA molecule counts into concentration for solving binding equilibrium, default 10]

**Dilution\_Vector:** [vector of probe dilutions to evaluate binding equilibrium and predictions at, default 1,1e-2,1e-4]

**Gibbs\_Model\_Vector:** [vector of gibbs thermodynamic models to use for evaluating binding equilibrium and predictions at, default 1,2,3,4]

**Temperature\_Celsius\_Model\_Vector:** [vector of hybridization temperatures in Celsius to use for evaluating binding equilibrium and predictions at, default 37,42,50,60]

**InitialFreeSolutionGuessConcentration\_MicroMolar**:[initial solution guess for steady state free probe concentration in uM, default 1e-10]

**SolutionErrorTolerance:** [total tolerance level for error in final equilibrium solution, default 1]

**MaxRecursiveEquilibriiumIterations:** [max number of equilibrium equation iterations stopping at final calculated steady state, overrides solution error tolerance]

**CellDiameter\_Pixels:** [cell pixel diameter for background predictions, default 50]

**SpotRadius\_Pixels:** [ spot pixel radius for predictions, default 5]

**NumberOfReferenceZStacks:[** number of z-stacks to spread background across for intensity predictions, default 67]

**SignalStepSize:** [step size for signal intensity bins in signal predictions, default 1e-1]

**SignalMaxValue:** [max intensity value for ranges in solution, default 3000]

**AutoFluoresenceBackground\_MEAN**:[reference mean autofluorescence, default 278]

**AutoFluoresenceBackground\_STD**: [reference autofluorescence standard deviation, default 33]

**NumberOfProbesInReferenceSpots**:[ number of probes in reference spot intensity for calibrating intensity predictions, default 48]

**ReferenceSpotIntensity\_MEAN:** [mean reference spot intensity , default 827 ]

**ReferenceSpotIntensity\_STD:** [reference spot intensity standard deviation, default 28]

* + 1. Custom Organism Settings

**Custom Root FASTA:**[location of custom organism fasta files, default N/A]

**Custom BLAST Databse DNA:**[location of custom DNA Blast database, default N/A]

**Custom BLAST Databse RNA:**[location of custom RNA Blast database, default N/A]

**Custom GTF:**[location of custom GTF File, default N/A]

**Custom GFF:**[location of custom GFF File, default N/A]

* 1. Locations of Input Database files are put in DatabaseLocations.xml
     1. **EMBL\_to\_NCBI** stores the location of files for mapping ENSEMBL gene and transcript accession numbers to NCBI RefSeq gene and accession numbers

**Each row of EMBL\_to\_NCBI is row for each organism with**

**Organism:**[name of organism to search for in input table]

**StableIDs:**[location of file with EMBL and NCBI transcripts paired to one another for that organism]

* + 1. **EMBL** stores the location of files for using ENSEMBL annotation in probe design

**Each row of EMBL is row for each organism with**

**Organism:**[name of organism to search for in input table]

**Root\_FASTA:**[Location of folder with all ENSEMBL annotation dna and rna fasta files]

**BLASTDB\_RNA:**[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 gff file]

* + 1. **NCBI** stores location of files for using NCBI RefSeq annotation in probe design

**Each row of NCBI is row for each organism with**

**Organism:**[name of organism to search for in input table]

**Root\_FASTA:**[Location of folder with all refseq annotation dna and rna fasta files]

**BLASTDB\_RNA:**[Location of refseq blast genome DNA database files]

**BLASTDB\_RNA:**[Location of refseq blast transcript RNA database files]

**GTF:**[Location of refseq reference genome gtf file]

**GFF:**[Location of refseq reference genome gff file]

* 1. Gene Expression Data File Locations are in GeneExpressionDataLocations.xml
     1. Stored for each organism location of all expression files with identifier name

**Row for each organism's reference gene or transcript level expression file**

**Organism:**[name of organism to search for in input table]

**[“Data Identifier Name”]:**[Location of gene expression file with extension]

* + 1. Stored for each organism location of all expression schema files listing different cell sample names associated with each expression file

**Schema for each organism's reference gene or transcript level expression file**

**Organism:**[name of organism to search for in input table]

**[“Data Identifier Name”]:**[Location of schema for gene expression file with extension]

* + 1. Stored also is the list of columns for each file output extension to use when finding expression data, and gene or target ids to map it to blast results

**tracks for output file types list of column information when reading the expression data**

**[“file extension”]:**[column sorted list of variable names]

1. Several out files are made during probe design software execution

**(GeneName)\_AccessionID\_probes\_TrueProbes.mat** [structure with probe sequences, location on on-target]

**(GeneName)\_AccessionID\_hits\_table\_TrueProbes.mat** [structure with information on BLAST hits]

**(GeneName)\_AccessionID\_ExpressionInfo\_TrueProbes.mat** [Structure with expression data]

**(GeneName)\_AccessionID\_Tm\_T\_OnOffThermoInfo\_TrueProbes.mat**

[Structure with binding energy of all hits]

**(GeneName)\_AccessionID \_dCpInfo\_TrueProbes.mat**

[Structure with heat capacity for all target binding reactions]

**(GeneName\_AccessionID\_dHInfo\_TrueProbes.mat**

[Structure with enthalpy for all target bindings reactions]

**(GeneName\_AccessionID\_dSInfo\_TrueProbes.mat**

[Structure with entropy for all target binding reactions]

**(GeneName\_binding\_hits\_map\_TrueProbes**

[binding site map]

**(GeneName)\_AccessionID\_Tm T \_BindingEnergyMatrix\_TrueProbes.mat**

[Equilibrium Binding Energy in binding site map format]

**(GeneName)\_AccessionID\_BindingMatricies\_TrueProbes.mat**

[Entropy, Enthalphy, and heat capacity in binding site map format for RNA]

**(GeneName)\_AccessionID\_BindingMatricies\_TrueProbes.mat**

[Entropy, Enthalphy, and heat capacity in binding site map format for complementary strand DNA binding]

**(GeneName)\_AccessionID\_Tm37\_BasicDesignerStats\_TrueProbes.mat**

[Index information on stats used for design probes]

**(GeneName)\_AccessionID\_chosen.mat**

[List of chosen probe indexes]

**(GeneName)\_AccessionID\_probes\_final\_96max.xlsx**

[Excel spreadsheet with final probes, and some stats]

**(GeneName)\_AccessionID\_Tm T \_ModelMetrics\_TrueProbes.mat**

[Structure with binding affinity calculations and probe design metrics]

* 1. Individual file variables saved

**(GeneName)\_AccessionID\_probes\_TrueProbes.mat** [structure with probe sequences, location on on-target]

-Cell array named probes for the probe tile Nx3 cell array where N is each probe tile.

-probe{N,1} = stores the probe number

-probe{N,2} = probe sequence on the target (convention is the same as match, so order reverse complement of this sequence)

-probe{N,3}= stores probe start index on target

**(GeneName)\_AccessionID\_hits\_table\_TrueProbes.mat** [structure with information on BLAST hits]

-Table of all blast hits recorded with 12 rows

-gene\_table(:,1) is E-value Score

-gene\_table(:,2) is Expect

-gene\_table(:,3) is Strand either Plus/Plus or Plus/Minus

-gene\_table(:,4) is probe alignment (Probe Sequence Alignment, space, Target Sequence)

-gene\_table(:,5) is starting and ending indices of alignment in the probe Query

-gene\_table(:,6) is the starting and ending indices of alignment in the target Subject

-gene\_table(:,7) is the target Name (Name and Definition, i.e., fasta Header)

-gene\_table(:8) is the probe sequence

-gene\_table(:,9) is the probe number

-gene\_table(:,10) is the number of base pair homology Matches

-gene\_table(:,11) is Number of possible base pairs

-gene\_table(:,12) is Percentage of homology match

**(GeneName)\_AccessionID\_ExpressionInfo\_TrueProbes.mat** [Structure with expression data]

Matrix with expression for each unique target name by cell-line or single-cells.

Matrix(NonDNA\_IDs,: ) is for RNA

Matrix(DNA\_IDs,: ) for any DNA expression

**(GeneName)\_AccessionID\_Tm\_HybridizationTemperature\_OnOffThermoInfo\_TrueProbes.mat** [Structure with binding energy of all hits]

Kb\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

Kon(probe\_number,1:N\_Models)

Koff(probe\_number,target,1:N\_Models)

* Off is sum or max off-target interaction probe has on target.

**(GeneName)\_AccessionID \_dCpInfo\_TrueProbes.mat** [Structure with heat capacity for all target binding reactions]

Binding energy heat capacity correction values for each unique pair entry in the gene table

Stored for all 8 models evaluated. Is a Nx8 double. N is the set of unique target and probe binding alignment sequences. Is default to zero as NN models included do not use a heat capacity in their fitting.

dCpeq\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dCon\_eq(probe\_number,1:N\_Models)

**(GeneName\_AccessionID\_dKbInfo\_TrueProbes.mat** [Structure with enthalpy for all target bindings reactions]

Binding constant values for each unique pair entry in the gene table at temperature specified in settings. Stored for all 8 models evaluated. Is a Nx8 double. N is the set of unique target and probe binding alignment sequences.

**(GeneName\_AccessionID\_dHInfo\_TrueProbes.mat** [Structure with enthalpy for all target probe target binding reactions]. Binding enthalpy values for each unique pair entry in the gene table. Stored for all eight models evaluated. It is an NxM double. N is the set of unique target and probe binding alignment sequences, and M is the number of models included.

dHeq\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dHf\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dHr\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dHon\_eq(probe\_number,1:N\_Models)

dHon\_f(probe\_number,1:N\_Models)

dHon\_r(probe\_number,1:N\_Models)

**(GeneName\_AccessionID\_dSInfo\_TrueProbes.mat** [Structure with entropy for all target probe target binding reactions]. Binding enthalpy values for each unique pair entry in the gene table. Stored for all eight models evaluated. It is an NxM double. N is the set of unique target and probe binding alignment sequences, and M is the number of models included.

dSeq\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dSf\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dSr\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models)

dSon\_eq(probe\_number,1:N\_Models)

dSon\_f(probe\_number,1:N\_Models)

dSon\_r(probe\_number,1:N\_Models)

**(GeneName\_AccessionID\_TmInfo\_TrueProbes.mat** [Structure with melting temperature for all probe target binding reactions]. Melting temperature values for each unique pair entry in the gene table are stored for all basic calculations and used with all eight models evaluated. Is an Nx9 double. N is the set of Tm\_Match(unique\_probe\_target\_sequence\_Index,1:N\_Models+1)

Tm\_on(probe\_number,1:N\_Models+1)

unique target and probe binding alignment sequences.

**(GeneName\_binding\_hits\_map\_TrueProbes** [binding site map]

Is where binding site map is stored and related variabls

DoesProbeBindSite(probe,target,site) pre-filtering for overlapping site binding

DoesProbeBindSite2(probe,target,site) post filtering for overlapping site binding

MolN\_ProbesArEvent: The number of probes that bind to the target at a particular site.

MolProbesAtEvents which probes are at binding each target at each binding site

Mol\_ProbesAtEvents\_ID tells position of binding site in binding match matrix and gene\_hit\_table

Num\_of\_Molecule\_Sites has the number of distinct binding sites probes are bound to on each unique target

**(GeneName)\_AccessionID\_Tm\_BindingMatrices\_TrueProbes.mat**

[Entropy, Enthalphy, and heat capacity in binding site map format for RNA]Tm\_mod(probe,target,binding site number,1:N\_Models+1)

dCp\_mod(probe,target,binding site number,1:N\_Models)

dHeq\_mod(probe,target,binding site number,1:N\_Models)

dHf\_mod(probe,target,binding site number,1:3)

dHr\_mod(probe,target,binding site number,1:3)

dSeq\_mod(probe,target,binding site number,1:N\_Models)

dSf\_mod(probe,target,binding site number,1:3)

dSr\_mod(probe,target,binding site number,1:3)

**(GeneName)\_AccessionID\_Tm\_HybridizationTempBindingEnergyMatrix\_TrueProbes.mat**

[Equilibrium Binding Energy in binding site map format]

Kb\_mod(probe,target,binding site number,1:N\_Models)

**GeneName)\_AccessionID\_Tm\_HybridizationTemp**

**\_BindingEnergyMatrix2\_TrueProbes.mat**

[Equilibrium Binding Energy in binding site map format]

Kb\_Complement(target,binding site number,1:N\_Models)

POGmod\_Complement(target,binding site number, 1:N\_Models)

**GeneName)\_AccessionID\_Tm**

**\_BindingMatrices\_TrueProbes.mat**

[Entropy, Enthalphy, and heat capacity in binding site map format for complementary strand DNA binding]

Tm\_Complement(target,binding site number,1:N\_Models+1)

dCp\_Complement(target,binding site number,1:N\_Models)

dHeq\_Complement(target,binding site number,1:N\_Models)

dHf\_Complement(target,binding site number,1:3)

dHr\_Complement(target,binding site number,1:3)

dSeq\_Complement(target,binding site number,1:N\_Models)

dSf\_Complement(target,binding site number,1:3)

dSr\_Complement(target,binding site number,1:3)

**(GeneName)\_AccessionID\_Tm37\_BasicDesignerStats\_TrueProbes.mat**

[Index information on stats used for design probes]

**NumDNAOffTargetOptions**

**NumRNAOffTargetOptions**

**Nvec\_RNAmulti**

**Off\_Socre**

**Probes\_WithNDNAOFF**

**Probes\_WithNRNAOFF  
Specificity\_Score**

**Svec\_DNA**

**Svec\_RNA**

**TPvec\_DNA**

**TPvec\_RNA**

**TPvec\_logKCOMPdivOFF\_DNA**

**TPvec\_logKOFF\_DNA**

**TPvec\_logKOFF\_RNA**

**TPvec\_logKOFFdivCOMP\_DNA**

**TPvec\_logKOFFdivON\_DNA**

**TPvec\_logKOFFdivON\_RNA**

**TPvec\_logKONdivOFF\_DNA**

**TPvec\_logKON\_divOFF\_RNA**

**TSvec\_DNA**

**TSvec\_RNA**

**Tvec\_DNA**

**Tvec\_RNA**

**(GeneName)\_AccessionID\_chosen.mat**

[List of chosen probe indexes]

**(GeneName)\_AccessionID\_probes\_final\_96max.xlsx**

[Excel spreadsheet with final probes, and some stats]

**(GeneName)\_AccessionID\_Tm T \_ModelMetrics\_TrueProbes.mat**

[Structure with binding affinity calculations and probe design metrics]

ModelMetrics.PackingEfficiency

1. **How To Get Working for A New Organism** 
   1. **Folder Creation**
      1. **Create a BLAST Database new folder for where the reference genome or transcriptome file is stored (under NCBI or ENSEMBL)**

TrueProbes/data/DatabaseData/Blast\_Databases/Organism/NCBI\_RefSeq

TrueProbes/data/DatabaseData/Blast\_Databases/Organism/EMBL\_EBI

* + 1. **Create a new folder for where the reference GTF file is stored (under NCBI or ENSEMBL)**

TrueProbes/data/DatabaseData/GTF\_Databases/Organism/NCBI\_RefSeq

TrueProbes/data/DatabaseData/GTF\_Databases/Organism/EMBL\_EBI

* + 1. **Create a new folder for where the reference GFF file is stored (under NCBI or ENSEMBL)**

TrueProbes/data/DatabaseData/GFF3\_Databases/Organism/NCBI\_RefSeq

TrueProbes/data/DatabaseData/GFF3\_Databases/Organism/EMBL\_EBI

* + 1. **Create a new folder for where any reference gene or transcript expression data is stored**

TrueProbes/data/DatabaseData/Gene\_Expression\_Data/Organism

* 1. **Acquire files needed for probe design**
     1. **For ENSEMBL Format**

**First, go to the FTP Site for the most current release and search for your organism**

<https://useast.ensembl.org/info/data/ftp/index.html>

* + - 1. **For Genome, get DNA (FASTA)**

<https://ftp.ensembl.org/pub/release-114/fasta/Organism/dna/>

Get file ending in .dna.primary\_assembly.fa.gz or if that is not there .dna.toplevel.fa.gz

Add to TrueProbes/data/DatabaseData/Blast\_Databases/Organism/EMBL\_EBI

* + - 1. **For Transcriptome, get cDNA (FASTA) and ncRNA (FASTA)**

<https://ftp.ensembl.org/pub/release-114/fasta/Organism/cdna/>

Get File ending in .cdna.all.fa.gz

Add to TrueProbes/data/DatabaseData/Blast\_Databases/Organism/EMBL\_EBI

<https://ftp.ensembl.org/pub/release-114/fasta/Organism/ncrna/>

Get File ending in .ncrna.fa.gz

Add to TrueProbes/data/DatabaseData/Blast\_Databases/Organism/EMBL\_EBI

* + - 1. **For GTF Go To Gene Sets (GTF)**

<https://ftp.ensembl.org/pub/release-114/gtf/Organism/>

Get File ending in (release #).gtf.gz

Add to TrueProbes/data/DatabaseData/GTF\_Databases/Organism/EMBL\_EBI

* + - 1. **For GFF Go To Gene Sets (GFF3)**

<https://ftp.ensembl.org/pub/release-114/gff3/Organism/>

Get File ending in (release #).gff3.gz

Add to TrueProbes/data/DatabaseData/GFF3\_Databases/Organism/EMBL\_EBI

* + 1. **For NCBI Format,**

**First, go to NCBI Genomes Page Search for your Organism in Genome and find Assembly with NCBI RefSeq Annotation**

<https://www.ncbi.nlm.nih.gov/datasets/genome/>

<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

* + - 1. **For Genome, Get DNA (FASTA)**

[https://ftp.ncbi.nlm.nih.gov/genomes/refseq/taxonomy group /Organism/reference/](https://ftp.ncbi.nlm.nih.gov/genomes/refseq/taxonomy%20group%20/Organism/reference/)

Get File ending in \_genomic.fna.gz

Add to TrueProbes/data/DatabaseData/Blast\_Databases/Organism/NCBI\_RefSeq

* + - 1. **For Transcriptome, Get RNA (FASTA)**

Get File ending in \_rna.fna.fna.gz

Add to TrueProbes/data/DatabaseData/Blast\_Databases/Organism/NCBI\_RefSeq

* + - 1. **Get Genome Assemblies GTF File**

Get File ending in \_genomic.gtf.gz

Add to TrueProbes/data/DatabaseData/GTF\_Databases/Organism/NCBI\_RefSeq

* + - 1. **Get Genome Assemblies GFF File**

Get File ending in \_genomic.gff.gz

Add to TrueProbes/data/DatabaseData/GTF\_Databases/Organism/NCBI\_RefSeq

* + 1. **For Converting between Gene Expression Data Annotated using NCBI or ENSEMBL when using a different reference type get conversion file from ENSEMBL**

<https://useast.ensembl.org/info/data/ftp/index.html>

<https://ftp.ensembl.org/pub/release-114/tsv/Organism/>

Get File ending in (release #).refseq.tsv.gz

Add to TrueProbes/data/DatabaseData/ENSEMBL\_NCBI\_StableIDs/

* 1. **Update Parameter Files with locations of newly added file**
     1. **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]
     2. **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]

**Compatibility**

**Make sure blast+ folder files are allowed to run by your computer security i.e. permission to run in MacOS or any security or virus scanners allow the program to run and do not block it.**

**Make sure blast+ is installed using installer and can be queried in system terminal.**

**Outputs Files. TrueProbe Design Software.**

**• (GeneName)\_RefSeqID\_probes\_TrueProbes.mat**

• [structure with probe sequences, location on on-target]

**• (GeneName)\_RefSeqID \_hits\_table\_TrueProbes.mat**

• [structure with information on BLAST hits]

**• (GeneName)\_RefSeqID\_ExpressionInfo\_TrueProbes.mat**

• [Structure with expression in TCGA and GTEX]

**• (GeneName)\_RefSeqID\_Tm T\_OnOffThermoInfo\_TrueProbes.mat**

• [Structure with binding energy of all hits]

**• (GeneName) \_RefSeqID \_dCpInfo\_TrueProbes.mat**

• [Structure with heat capacity for all target binding reactions]

**• (GeneName)\_RefSeqID\_dHInfo\_TrueProbes.mat**

• [Structure with enthalpy for all target bindings reactions]

**• (GeneName)\_RefSeqID\_dSInfo\_TrueProbes.mat**

• [Structure with entropy for all target binding reactions]

**• (GeneName)\_binding\_hits\_map\_TrueProbes**

• [binding site map]

**• (GeneName)\_RefSeqID\_Tm T \_BindingEnergyMatrix\_TrueProbes.mat**

• [Equilibrium Binding Energy in binding site map format]

**• (GeneName)\_RefSeqID\_BindingMatricies\_TrueProbes.mat**

• [Entropy, Enthalphy, and heat capacity in binding site map format for RNA]

**• (GeneName)\_RefSeqID\_BindingMatricies\_TrueProbes.mat**

• [Entropy, Enthalphy, and heat capacity in binding site map format for complementary strand DNA binding]

**• (GeneName)\_RefSeqID\_Tm37\_BasicDesignerStats\_TrueProbes.mat**

• [Index information on stats used for design probes]

**• (GeneName)\_RefSeqID\_chosen.mat**

• [List of chosen probe indexes[

**• (GeneName)\_RefSeqID\_probes\_final\_96max.xlsx**

• [Excel spreadsheet with final probes, and some stats]

**• (GeneName)\_RefSeqID\_Tm T \_ModelMetrics\_TrueProbes.mat**

• [Structure with binding affinity calculations and probe design metrics]

Add order information.

Documentation.

Think someone wants to test.

(human gene, probe target id, number plug into software they get probes, number of off-targets and cell type)

(list of probes they order, off-targets and on-targets

Run same metrics on figure).

(show do it new organism,)

Change gene expression information (work the same).

Stress better (make comparisons,

Time probes (against multiple organisms)

Sequence alignment then design probes

Get from list of sequences to probe designed.

Subchapter for reporter strains.

**Direct fusion, and localization, different from promoter.**

**Plasmid has promoter with all 3 colors, together.**

**Copy in plasmid.**

**Folder (Option current vs New) (If want to update or not).**

**Reference Genome (Dataset, use reference.**

**% version on cluster the same.**