**Required Packages**

**MATLAB 23.2+**

**Bioinformatics Toolbox 23.2+**

**Curve Fitting Toolbox 23.2+**

**Image Processing Toolbox 23.2+**

**Parallel Computing Toolbox 23.2+**

**Statistics and Machine Learning Toolbox 23.2+**

**Symbolic Math Toolbox 23.2+**

**Signal Processing Toolbox 23.2+**

**Main TrueProbe Probe Design**

* **A0 Probe Generation** 
  + **MATLAB 23.2**
  + **Bioinformatics Toolbox 23.2**
* **A1 Probe Blasting**
  + **MATLAB 9.10**
  + **Bioinformatics Toolbox 23.2**
  + **Parallel Computing Toolbox 23.2**
* **A2 GeneExpressionInformation**
  + **MATLAB 23.2**
  + **Statistics and Machine Learning Toolbox 23.2**
  + **Parallel Computing Toolbox 23.2**
* **A3 ThermodynamicInformation**
  + **MATLAB 23.2**
  + **Symbolic Math Toolbox 23.2**
  + **Statistics and Machine Learning Toolbox 23.2**
  + **Parallel Computing Toolbox 23.2**
  + **Bioinformatics Toolbox 4.15.1**
* **A4 BindingSiteMapping**
  + **MATLAB 23.2**
  + **Symbolic Math Toolbox 23.2**
  + **Statistics and Machine Learning Toolbox 23.2**
  + **Parallel Computing Toolbox 23.2**
  + **Bioinformatics Toolbox 23.2**
* **A5 ProbeDesignerStats**
  + **MATLAB 23.2**
  + **Parallel Computing Toolbox 23.2**
* **A6 ProbeSelection**
  + **MATLAB 23.2**
  + **Bioinformatics Toolbox 23.2**
  + **Parallel Computing Toolbox 23.2**
* **A7 GetMetrics**
  + **MATLAB 23.2**
  + **Symbolic Math Toolbox 23.2**
  + **Statistics and Machine Learning Toolbox 23.2**
  + **Bioinformatics Toolbox 23.2**
* **A8 ProbeSpecificityFiltering**
  + **MATLAB 23.2+**
  + **Bioinformatics Toolbox 23.2+**
* **A9 MultiFluorphoreColocalizationConvolutionND**
  + **MATLAB 23.2+**
  + **Symbolic Math Toolbox 23.2+**
  + **Image Processing Toolbox 23.2+**
  + **Bioinformatics Toolbox 23.2+**

**Run Other Probe Design**

* **GenerateIndividualGeneFasta**
  + **MATLAB 23.2+**
  + **Bioinformatics Toolbox 23.2+**
* **Run OtherProbeSets GetMetrics**
  + **MATLAB 23.2+**
  + **Statistics and Machine Learning Toolbox 23.2+**
  + **Bioinformatics Toolbox 23.2+**

**Compare Software Probe Design**

* **IntegrateResults**
  + **MATLAB 23.2+**
  + **Bioinformatics Toolbox 23.2+**
* **Experimental Analysis**
  + **Curve Fitting Toolbox 23.2+**
* **PrintProbeSets**
  + **MATLAB 23.2+**
  + **Bioinformatics Toolbox 23.2+**
* **CalculateTemperatureOrNewConditionSolution**
  + **MATLAB 23.2+**
* **Main PlotResults**
  + **MATLAB 23.2+**
  + **Symbolic Math Toolbox 23.2+**
  + **Image Processing Toolbox 23.2+**
  + **Statistics and Machine Learning Toolbox 23.2+**
  + **Curve Fitting Toolbox 23.2+**
  + **Parallel Computing Toolbox 23.2+**

**How to Install TrueProbes**

**Install MATLAB and product packages.**

**How to add new genome or transcriptome.**

**Create New Folder in data/DatabaseData/BLASTDatabase/OrganismName.**

**For NCBI Go to:**

**<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/> then species reference**

<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Homo_sapiens/reference/GCF_000001405.40_GRCh38.p14/>

**Download GTF Files to** DatabaseData/GTF\_Databases/OrganismName/NCBI\_RefSeq/

GTF Files (.gtf)

**Download GFF Files to**

DatabaseData/GFF3\_Databases/OrganismName/NCBI\_RefSeq/

GFF Files (.gff)

**Download Reference Genome & Transcriptome Files to** DatabaseData/BLAST\_Databases/OrganismName/NCBI\_RefSeq/

**Reference Genome**

DNA FASTA (\*.rna.fna) at

cDNA FASTA (\*.rna\_from\_genomic.fna) at

DatabaseData/BLAST\_Databases/OrganismName/NCBI\_RefSeq/

**For ENSEMBL, Go to FTP Download Page <https://useast.ensembl.org/info/data/ftp/index.html>**

**Download GTF Files to** DatabaseData/GTF\_Databases/OrganismName/EMBL-EBI/

GTF File (\*.gtf)

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

**Download GFF3 Files to** DatabaseData/GFF3\_Databases/OrganismName/EMBL-EBI/

GFF File (\*.gff3)

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

**Download Reference Genome & Transcriptome Files to** DatabaseData/BLAST\_Databases/OrganismName/EMBL-EBI/

**Reference Genome**

DNA FASTA (\*.dna.primary\_assembly.fa) at <https://ftp.ensembl.org/pub/release-114/fasta/ organism /dna/>

**Reference Transcriptome**

cDNA FASTA (\*.cdna.all.fa) at

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

ncRNA FASTA (\*.ncrna.fa) at

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

**How to Run the TrueProbes Software.**

**A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5** is the main script for running 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 command line via **A0\_BKJH\_ProbeDesign\_Wrapper\_cluster\_V5(id,cluster)**

**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. **Binding Affinity Calculation**. Binding affinities are calculated for all pairs of probe and target sequences homology matches in the BLAST results.
4. **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.
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 given reference values for cell size, probe concentration, and probe intensity.

**Input Argument 1: id**

id is an integer and is the row of the input design table at the top of the script to run and design probes against.

**Input Argument 2: 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 is the number of cores in the Slurm file.

cluster

The table row has nine column entries. Example: {{NM\_000805.5'},{},{}, 'Human','(GAST)','(GAST)','17',{},1 ;}...

Input table columns:

1. 1. Included target accession IDs. Designs probes shared across all accession numbers
2. 2. Text Sequence Files to Include (files). Default empty
3. 3. Text Sequence Files to Exclude (files). Default Empty
4. 4. Organism to design probes for.
5. 5. Gene Name 1.

6. Gene Name 2. Second potential gene name to use instead of the first.

1. 7. Chromosome. The chromosome number.
2. 8. Excluded target accession IDs. Removes probes in exclusion accession numbers.

9. Strand. Which strand to design probes against for RNA default is ‘plus’.

**Settings.**

Below the inputs1 table is a list of primary and secondary settings which can be configured to determine how the probes are designed and evaluated.

**Main Input Parameters.**

* **Locations**
  + **SaveRoot**
    - [Location where files are to be saved]
  + **customBlastDatabase\_DNA**
    - [Location of user-custom DNA blast database if user wants to use their own custom database]
  + **customBlastDatabase\_RNA**
    - [Location of user-custom RNA blast database if user wants to use their own custom database]
* **Main Probe Design Parameters**
  + **max\_probes**
    - [Max number of probes to design, default 96]
  + **minProbeSize**
    - [min nt length of potential probes, default 20]
  + **maxProbeSize**
    - [max nt length of potential probes, default 20]
  + **MininumProbeSpacing**
    - [min spacing between probes, default 3]
  + **HybridizationTemperature** 
    - [Hybridization temperature, default 37C]
  + **BLASTrna**
    - [decide to blast RNA sequences, default 1]
  + **BLASTdna**
    - [decide to blast DNA sequences, default 0]
* **Secondary Parameters**
  + **Nmodel** 
    - [Which thermodynamic model to use for probe design and evaluation, default 4]
  + **SaltConcentration** 
    - [Salt Concentration mM, default 0.05]
  + **RemoveRibosomalHits**
    - [Filter out probes with targets to ribosomal proteins, default 1]
  + **MinHomologySearchTargetSize**
    - [Minimum homology length for BLAST alignments to be recorded and used in probe design and evaluation, default 15]
  + **packOptimal**
    - [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]
  + **UseSelfProb**
    - [When designing probes consider probe self-hybridization when ranking probes based on binding affinity, default 1]
  + **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]
  + **UseGeneOverTranscLevelExpression** 
    - [use gene level (1) or transcript isoform level (0) gene expression values, default 0]
  + **DoAllGenesHaveSameExpression**
    - [decide to assume equal expression for all genes (1) or to use gene expression reference (0) , 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]
* **BLAST Settings Parameters**
  + **evalue** 
    - [Expectation value cutoff, default 1000]
  + **Dust**
    - [filter query sequences with DUST, default no]
  + **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]
* **Expression Parameters**
  + **MouseExpressionFile**
    - [location of mouse gene expression bed file]
  + **HumanTCGA\_TranscriptExpressionFile**
    - [location of human TCGA transcript expression bed file]
  + **HumanTCGA\_GeneExpressionFile**
    - [location of human TCGA gene level expression bed file]
  + **HumanGTEX\_TranscriptExpressionFile**
    - [location of human GTEx transcript expression bed file]
  + **HumanGTEX\_GeneExpressionFile**
    - [location of human GTEx gene level expression bed file]

**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\_Tm37\_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\_Tm37\_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\_Tm37\_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.**