README: Single-molecule microscopy image data and analysis files for "Ultra-specific and Amplification-free Quantification of Mutant DNA by Single-molecule Kinetic Fingerprinting."

Authors: Stephen L. Hayward, Paul E. Lund, Qing Kang, Alexander Johnson-Buck, Muneesh Tewari, and Nils G. Walter.

Contact: Alexander Johnson-Buck [alebuck@med.umich.edu]

Contents

Research Overview

These data were generated in the course of developing and demonstrating an analytical technique to detect and quantify the presence of small DNA allele fragments that contain mutations that are associated with non-small lung cell cancers. The two mutations in the *EGFR* gene chosen for use in this study were an in-frame deletion in exon 19 (COSMIC ID: COSM6225; c.2236_2250del15 [p.E746_A750delELREA]), and a single point mutation in exon 20 that results in the missense mutation T790M (COSMIC ID: COSM6240, c.2369C>T [p.T790M]). This work also investigated the influence of DNA damage (*e.g.*, cytosine deamination) as it relates to the kinetics of hybridization for small DNA oligonucleotides.

The DNA allele fragments used in this study are 28 base pairs in length (22-bases of the relevant *EGFR* gene sequence plus a 6-base barcode, TAGGAC) and were prepared by chemical synthesis (IDT), or by restriction digestion from a DNA plasmid carrying one or more copies of the allele fragments.

This work was conducted at the University of Michigan in the Department of Chemistry and Department of Internal Medicine, Division of Hematology/Oncology and was funded in large part by the Michigan Economic Development Corporation and the University of Michigan with MTRAC and other pilot grants.

Methods

Data Collection

Single-Molecule Recognition through Equilibrium Poisson Sampling (SiMREPS) experiments were performed on an Olympus IX-81 objective-type TIRF microscope equipped with a 60X oil-immersion objective (APON 60XOTIRF, 1.49 NA) with both Cell^TIRF and *z*-drift control modules, and an EMCCD camera (IXon 897, Andor), using MetaMorph acquisition software (Molecular Devices). Transient binding of a fluorescent probe oligonucleotide to DNA molecules immobilized to the surface of a custom-built sample cell was monitored for 10 min under TIRF illumination by 640 nM laser light with a 500 ms exposure time (1200 total frames), camera EM gain=150, and recorded as a stack of TIF images (movie).

Data Analysis

Movie files were analyzed using custom scripts written in MATLAB and the QuB software suite (State University of New York at Buffalo) to:

1) Identify the locations of immobilized candidate DNA molecules and extract a fluorescence-intensity versus time trace for each,

2) Determine the number of binding and dissociation events (Nb+d) and the median fluorescent probe bound (τbound,median) and unbound (τunbound,median) time for each candidate molecule, which together comprise the kinetic fingerprint of the candidate molecule, and

3) Evaluate the kinetic fingerprint and data quality of each candidate to arrive at the final number of immobilized DNA molecules detected for a specific sequence.

The MATLAB scripts are available upon request from Alex Johnson-Buck. The detailed steps for analysis of the data using the diffraction-limited workflow are provided below.

Diffraction-limited Analysis Workflow

Matlab 1) Generate_time_traces_TIF_SiMREPS_v3a.m
Does peak finding on movies and extracts fluorescence-time info for candidate molecules (requires Image Processing Toolbox).

Typically we use the following parameters:

```
channel = 'whole';
stdfactor = 2.5;
straight_intensity_cutoff = 0;
selectionbyflucts = 1;
analysis = auto;
driftcorr = 'no';
```

Output:
background corrected average image for movie, `~_avg_bgsub.tif`
fluorescence vs time info for identified coordinates, `~_traces.dat`
candidate molecule coordinates, `~_coords.dat, ~_coords.txt`

Matlab 2) traces_2_qub_v2.m
Converts trace data into format readable in QuB
Output: `~_traces.txt`, one file per movie

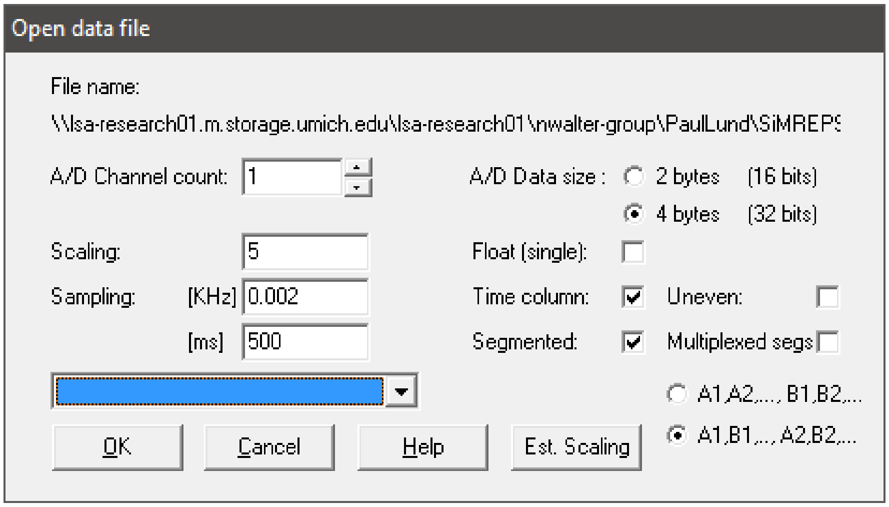
2-columns, tab-separated; frame number + fluorescence intensity

Trace data for each molecule candidate are concatenated vertically in the file, separated by \n

QuB)

Fit fluorescence intensity data to two state model using hidden markov modeling

- Open ~_traces.txt files in **QuB classic v2.0.0.22** (https://www.qub.buffalo.edu/)
- Set integration time to match framerate (500 ms), Time column, Segmented, 4 bytes.

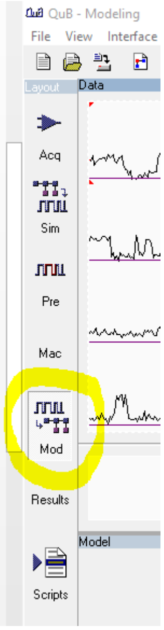


Open data file

File name:
\\lsa-research01.m.storage.umich.edu\lsa-research01\nwalter-group\PaulLund\SiMREPS

A/D Channel count: 1
A/D Data size : ○ 2 bytes (16 bits) ● 4 bytes (32 bits)
Scaling: 5
Float (single): □
Sampling: [KHz] 0.002 [ms] 500
Time column: ☑ Uneven: □
Segmented: ☑ Multiplexed segs □
○ A1,A2,..., B1,B2,...
● A1,B1,..., A2,B2,...

- Hit **Est. Scaling** button for each data file as it opens.
- Switch to Modeling interface "Mod" on left hand side

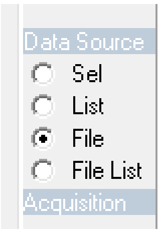


- Create a new model File>New Model…

- Set rates in model to something reasonable

e.g., 0.2 s^-1 (state 1-->2) and 0.1 s^-1 (2-->1)
- When working with multiple movies files at once:

- Select "File" radio button from Data Source panel on right hand side



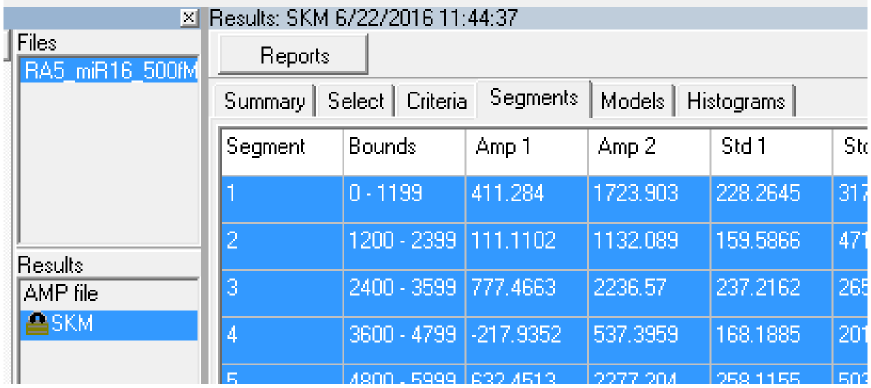
Data Source
○ Sel
○ List
● File
○ File List
Acquisition

- Click **Amps** button to populate model with amplitude estimates for state 1 and state 2
- Click **Idealize**, then inspect fitting+ histograms for each trace (segment). SKM should appear in Results list

- If fitting is bad, adjust model parameters (rates, amplitudes, etc). If good:

  - Save idealized trace data: File> Idealized Data >save Idealized Data …

  - Copy contents of **Segments** tab from SKM results to excel workbook and rename sheet with file name



Results: SKM 6/22/2016 11:44:37

Files: RA5_miR16_500fM

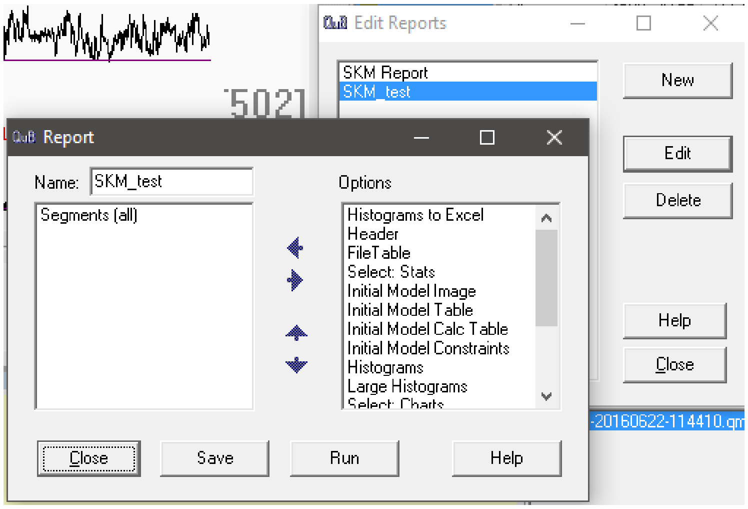
Results: AMP file, SKM

Segment	Bounds	Amp 1	Amp 2	Std 1	Stc
1	0 - 1199	411.284	1723.903	228.2645	317
2	1200 - 2399	111.1102	1132.089	159.5866	471
3	2400 - 3599	777.4663	2236.57	237.2162	265
4	3600 - 4799	-217.9352	537.3959	168.1885	201
5	4800 - 5999	632.4513	2277.204	258.1155	503

  - QuB can also export the results to Excel but is VERY SLOW

    - Click the Reports tab > Edit Reports …

    - Click New and set up a new report Macro with just the Segments (all) command



Edit Reports

SKM Report
SKM_test

Report

Name: SKM_test

Segments (all)

Options

Matlab 3) SiMREPS_QuB_Analysis_v_0_2d.m

Filters “genuine’ molecules from candidate list based on thresholds

Set appropriate thresholds using negative control movies (excel workbook) such that non-specific or off-target candidate molecules are excluded

Adjust threshold values for filtering

After establishing appropriate thresholds, filter data from entire experiment

Output:
Nb+d histogram plot, `~_hist.jpg`
median bound and unbound dwell time scatter plot, `~_tauPlot.jpg`

accepted candidate molecules = "Counts"

Optional:

Excel workbook with summary data for each trace, `~_valuesOut.xlsx`

Matlab variable list of accepted candidate numbers, `~_acceptedMols.mat`

Matlab variable list of rejected candidate numbers `~_rejectedMols.mat`

File Organization

This work contains multiple ***zip files***, each of which represents one of the ***principal experiment groups*** presented in the publication. Each experiment group contains files corresponding to various experimental conditions related to that experiment group. File trees in PDF format are included within each zip file and at the top level of the Deep Blue deposit.

Cloned_DNA_enzymeTreatments_denatureTemp.zip	Experiments related to optimizing detection of the T790M mutant allele in the presence of competing wild-type T790 allele
UDG_treatment_in-situ	effect of treatment of surface-immobilized DNA with uracil DNA glycosylase on the number of false-positives detected
Synthetic-vs-Cloned_DNA _UDGTreatement	comparison of false-positives detected with and without uracil DNA glycosylase treatment in synthetic DNA alleles and DNA produced through restriction enzyme digestion from plasmid
DenaturingTemp_and_UDG_in-situ	testing the effect of different denaturation temperatures and incubation time on the number of false positive counts
Synthetic-vs-Cloned_DNA	comparison of false-positives detected in synthetic DNA alleles and DNA produced through restriction enzyme digestion from plasmid

Cloned_T790_WT_DNA_EnzymeTreated	comparison of false-positives detected after different enzymatic treatments to remove DNA damage
Effect_of_FP_Length_on_DNA_Detection.zip	Comparison of 8 nucleotide and 9 nucleotide fluorescent probe binding kinetics
FP-binding_kinetics_Deaminated_Synthetic_Oligo.zip	Characterization of fluorescent probe binding kinetics to various DNA alleles differing by a single nucleobase
Kinetic_Analysis_for_MUT-specific_FP	Binding of *mutant-specific* probe to T790M mutant allele, T790 wild-type allele containing a deoxyuracil representing deaminated cytosine, and T790 wild-type allele
Kinetic_Analysis_for_WT-specific_FP	Binding of *wild-type-specific* probe to T790 wild-type allele
Limit-of-Detection_EGFR_Exon19_Deletion_Data.zip	Limit-of-detection experiments for Exon 19 deletion mutant allele in the presence of wild-type Exon 19 allele
Limit-of-Detection_T790M_Data.zip	Limit-of-detection experiments for T790M mutant allele in the presence of wild-type T790 allele
Standard_Curve_Exon19_Deletion.zip	Number of detected molecules (Accepted Counts) as a function of solution DNA concentration for Exon 19 deletion mutant allele
Standard_Curve_T790M.zip	Number of detected molecules (Accepted Counts) as a function of solution DNA concentration T790M mutant allele
T790M_Spiked_into_Urine.zip	Detection of the T790M mutant allele in healthy donor urine

File Types

Below is a description of the various file types included in this deposit. A ~ represents the base filename.

File extension	**Description**
~.tif	TIF image stack (movie) recorded from the microscope
~_avg_bgsub.tif	single background-corrected TIF image generated from the original movie file, used for peak finding
~_coords.dat	pixel coordinates for candidate molecules, readable in MATLAB
~_coords.txt	pixel coordinates for candidate molecules in text format
~_traces.dat	fluorescence vs time data for identified candidate molecules, readable in MATLAB

~_traces.txt	fluorescence vs time data for identified candidate molecules in text format; data are arranged in two-columns, frame number and fluorescence intensity
*.qsf	QuB session file
*.dwt	2-state idealization from hidden Markov modeling, generated from QuB
*QuB Output.xlsx	matrix of values from QuB, derived from hidden Markov idealization for candidate molecules, where each row is one candidate, each worksheet in workbook represents data from a single movie file
~_valuesOut.xlsx	Summary output for molecule candidates with the data quality and kinetic criteria used for filtering
~_acceptedMols.mat	MATLAB variable containing a list of accepted candidate numbers

Within a given experiment group, excel files with additional experimental details are included where relevant.

Expt_plan_~.xlsx	Experimental protocol notes, including naming conventions used for movie files corresponding to various experimental conditions
Summary_~.xlsx	Summarized results from subsets of experiments

Definition of Common Terms and Abbreviations

Term	**Meaning**
bp	Base pair
Count	Candidate molecule that has passed data quality and kinetic filtering; number thereof
dsDNA	double-stranded DNA
EndoVIII	Endonuclease VIII
Exon 19	wild-type allele control for Exon 19 deletion
Exon 19 deletion	In-frame deletion mutant allele in the human *EGFR* gene, COSMIC ID: COSM6225; c.2236_2250del15 [p.E746_A750delELREA]
FP	fluorescent probe
FpG	formamidopyrimidine [fapy]-DNA glycosylase
MUT	mutant allele
nt	nucleotide
ssDNA	single-stranded DNA
T790	wild-type allele (control for T790M)
T790M	single point mutation in the human EGFR gene resulting in the missense mutation T790M, (COSMIC ID: COSM6240, c.2369C>T [p.T790M])
UDG	uracil DNA glycosylase
WT	wild-type allele