25 April 2023

These MATLAB scripts are designed to analyse Nanopore data described by Prajapati et al. (submitted):

**"The yeast genome is globally accessible in living cells"**

Hemant K. Prajapati, Peter R. Eriksson, Paul A. Elizalde, Christopher T. Coey and David J. Clark

Division of Developmental Biology, Eunice Kennedy-Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda MD 20892, USA.

The raw data (base-called fast5 files as tar.gz) are available in the GEO database: GSE230306.

**BACKGROUND**

The goal is to measure the fraction of DNA methylated at each genomic CG site in the budding yeast genome (*S. cerevisiae*). The M.SssI DNA cytosine methyltransferase methylates ‘C’ in CG sites.

m5C can be detected in Nanopore long sequence reads using Nanopolish software.

We induced expression of M.SssI and followed the time course of methylation from 0 to 240 min.

**Alignment of Nanopore sequence reads and estimation of the methylated fraction for each genomic CG site**

FAST5 files generated by the MinION instrument were basecalled with Guppy.

Base-called reads were indexed using Nanopolish software v.0.14.0 (https://github.com/jts/nanopolish).

Reads were mapped to sacCer3 with Minimap2 v.2.24 (https://github.com/lh3/minimap2).

Reads were sorted and indexed with samtools v.1.17 (https://github.com/samtools).

Reads were scored for methylation using Nanopolish (Simpson et al. 2017), providing a value of 'fmeth' for each genomic CG site.

'fmeth' is the fraction methylated of each genomic CG site (i.e., the fraction of reads in which the CG site is scored as methylated).

Output: An Excel file with one sheet for each time point:

**MATLAB scripts to plot median 'fmeth' and nucleosome phasing as a function of time**

The experiment has 5 time points: 0, 30, 60, 120 and 240 min.

Import and save the data from each sheet in the Excel file as a .mat file.

Example: 240 min time point: 'MSssI\_Rep2\_240m.mat'.

NOTE: Nanopolish estimates 'fmeth' for the majority of CG sites as single sites, but some CG sites are clustered and 'fmeth' is estimated for the group (this information is included in the Excel file). However, the number of called CG sites differs slightly for each time point.

1. Create a list of CG sites with data for each time point and quantile/median data for various genomic regions, as well as Pol II decile data. Use the sample-specific script 'script\_Get\_CpG\_site\_locations\_Rep2.m' to run MATLAB function 'Get\_CpG\_site\_locations.m'.

Outputs for each time point:

'CpG\_Site\_Locations\_Rep2\_---.mat': site data.

'Quantiles\_CpG\_Rep2\_---.mat': for various regions; median data is in the 10th quantile.

'Deciles\_ CpG\_Rep2\_---.mat': median and mean 'fmeth' for each gene transcription decile (Decile 1 contains the genes with the most Pol II, measured by ChIP-seq for the Rpb3 subunit).

'Phase\_Profiles\_CpG\_---.mat': mean 'fmeth' for all genes relative to the dyad of the +1 nucleosome.

2. Run sample-specific 'script\_Get\_Median\_plots\_CpG\_Rep2.m' .

Outputs:

Ten .eps files: 7 plots of 'fmeth' v. time as .eps files for various regions of the genome (CG sites in ORF, promoter NDR, ARS, TEL, Ty and tRNA regions), 2 plots for Pol II deciles (means and medians), and an exponential rate plot (ln(1 - fcut) v. time).

One .csv file containing the slopes, intercepts and correlation coefficients of determination for the rate plot.

3. Run 'script\_plot\_MSssI\_phasing\_Rep2.m' to obtain a metagene plot with all 5 time points and the negative control.

Output: two phasing plots: unsmoothed,and smoothed, with wild type MNase-seq data profile for comparison: "Avg\_dyad\_density\_WT\_A\_120\_160\_Ocampo\_NAR\_2016.mat".