21 April 2023

These scripts are for analysis of qDA-seq 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 (fastq.gz) are available in the GEO database: GSE229797.

qDA-seq data: Illumina paired-end data with 50-nt reads (fastq.gz files).

**BACKGROUND**

The goal is to measure the fraction of DNA methylated at each genomic GATC site in the budding yeast genome (*S. cerevisiae*).

The *E. coli* *dam* DNA adenine methyltransferase (Dam) methylates ‘A’ in GATC sites. If the ‘A’ is methylated on both strands, the GATC site can be cut by the restriction enzyme DpnI.

DpnI was expected to yield blunt ended DNA fragments (ending on GA and beginning with TC).

However, we found that a significant fraction of ends were missing the last ‘A’ or the first ’T’.

To calculate the methylated fraction for each GATC site ‘fcut’, we counted the number of ‘GA’ and ‘G’ ends and divided by the fragment coverage (not the read coverage), to obtain a measure of the total number of DNA molecules containing a specific GATC site.

Similarly, we calculated ‘fcut’ for the other side of the GATC site as number of TC and C ends/number of molecules containing this GATC site.

The result is two separate measures of ‘fcut’ for each GATC site.

However, we noticed that GATC sites within 200 bp of a neighbouring GATC site show reduced fragment coverage, presumably due to loss of short DNA fragments during DNA purification.

To correct for this issue, we ignored data for GATC sites within 200 bp of a neighbouring GATC site on both sides. If a neighbouring GATC site is too close on only one side, we set the ‘fcut’ value for the close side to the value for the other side.

**Alignment, end and fragment coverage counts**

These steps were performed using the NIH HPC Biowulf cluster:

1. Adjust fastq.gz file names: ./batch\_rename\_fastq.gz\_files.sh

2. Align to the sacCer3 genome using Bowtie 2 and strict parameters:

./align\_PE\_reads\_sacCer3.sh

3. Count the number of DNA molecules ending at each nucleotide in every chromosome "Cuts\_filename.mat" and the number of times each nucleotide in every chromosome appears in a sequenced DNA fragment "Occupancy\_filename.mat". Run:

./compute\_Cuts\_and\_Occ\_profiles\_sacCer3.sh 0 5000

The numbers indicate the minimum and maximum distances allowed between aligned read pairs.

The Cuts and Occupancy files contain the raw data (with no corrections).

**MATLAB scripts to plot median 'fcut' as a function of time**

Data for all 35,830 GATC sites are included in: sacCer3\_GATC\_Site\_Data.csv

This file was converted for use in MATLAB: sacCer3\_GATC\_Site\_Data.mat

The basic experiment has 5 time points (other experiments have up to 8).

Cuts and Occupancy files for experiment "808H\_2" are included here, as examples.

1. Use the sample-specific script ('script\_Get\_fcut\_quantiles\_808H\_2.m') to run MATLAB function 'Get\_fcut\_quantiles.m' to obtain quantile data as .mat files for each time point.

Outputs:

A "Quant" file for each time point (the median fcut value is quantile 10).

An "Rpb3\_Deciles" file for each time point, containing the median or mean 'fcut' for each decile (Decile 1 contains the genes with the most Pol II, measured by ChIP-seq for the Rpb3 subunit).

2. Run sample-specific 'script\_Get\_Median\_plots\_808H\_2.m' .

Outputs:

Ten .eps files, including 7 plots of 'fcut' v. time as .eps files for various regions of the genome (GATC 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 .mat file containing the median 'fcut' data': 808H\_2\_medians.mat

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

**MATLAB scripts for nucleosome phasing metagene plots: mean 'fcut' at different time points v. the +1 nucleosome dyad location**

1. Use the sample-specific script ('script\_Get\_phasing\_data\_808H\_2.m') to run MATLAB function 'Get\_fcut\_phasing.m' to obtain phasing data as a .mat file for each time point.

This script requires 'Phase\_Coords.mat', which contains the data needed to align all genes correctly on the +1 nucleosome.

Output = 'Mean\_phase\_fcut' .mat file for each time point.

2. Run 'script\_Get\_phasing\_plots\_808H\_2.m'.

Output: three phasing plots: unsmoothed, and smoothed with and without the specified MNase-seq data profile.

NOTE: This example uses the wild type MNase-seq data profile for comparison:

"Avg\_dyad\_density\_WT\_A\_120\_160\_Ocampo\_NAR\_2016.mat"

For the remodeler degron plots, replace this file in the script with one of the following:

rsc8: Phase\_Dyads\_rsc8\_A\_120\_160.mat

isw1: Phase\_Dyads\_isw1\_A\_120\_160.mat

chd1: Phase\_Dyads\_chd1\_A\_120\_160.mat

isw1 chd1: Phase\_Dyads\_isw1\_chd1\_A\_120\_160.mat