DNAFORM

Sequencing Guide of nAnT-iCAGE library with an Illumina 2000/2500 sequencer

Written by Yuzuki Manabe, June 27, 2014

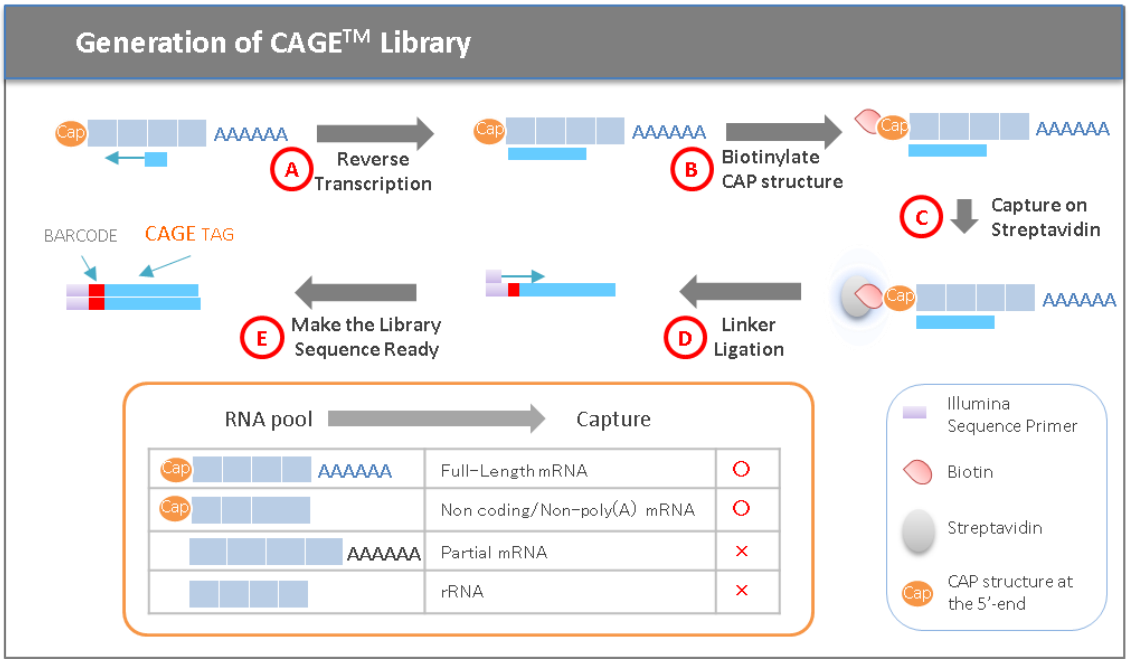
List of Contents:

1. Objective

This document is prepared for bioinformaticians who will be performing sequencing and data analysis of **n**on-**A**mplified **n**on-**T**agging **I**llumina **C**ap **A**nalysis of **G**ene **E**xpression (nAnT-iCAGE) libraries. The objective of this document is to provide an overview of nAnT-iCAGE technology and guidance for sequencing and data analysis of nAnT-iCAGE library.

2. nAnT-iCAGE technology

nAnT-iCAGE is a technology to analyze 5'-ends of fully matured mRNAs using an Illumina 2000/2500 sequencer. A nAnT-iCAGE library is prepared from total RNA extracted from Eukaryotic cells. We'd like to give you a quick overview of how the CAGE library is prepared. ⒶAt first, the cDNA is synthesized with random primers using total RNA as a template. A fully matured cDNA has a 'cap structure' at 5'-end. Ⓑ The 'cap structure' is biotinylated at oxidatied diol residues using biotin hydrazide. Un-hybridized single strand RNAs are then digested with RNase ONE. ⓒCapture the biotin on streptavidin beads. After releasing the cDNAs from the streptavidin beads, ⓓ linkers are ligated to the cDNAs. A 5'-linker contains a barcode sequence which marks the origin of a CAGE tag from each sample. ⓔ When the 2nd strand synthesis is completed, the library now contains Illumina primer sequence and the barcode and ready for sequencing with Illumina 2000/2500 sequences.



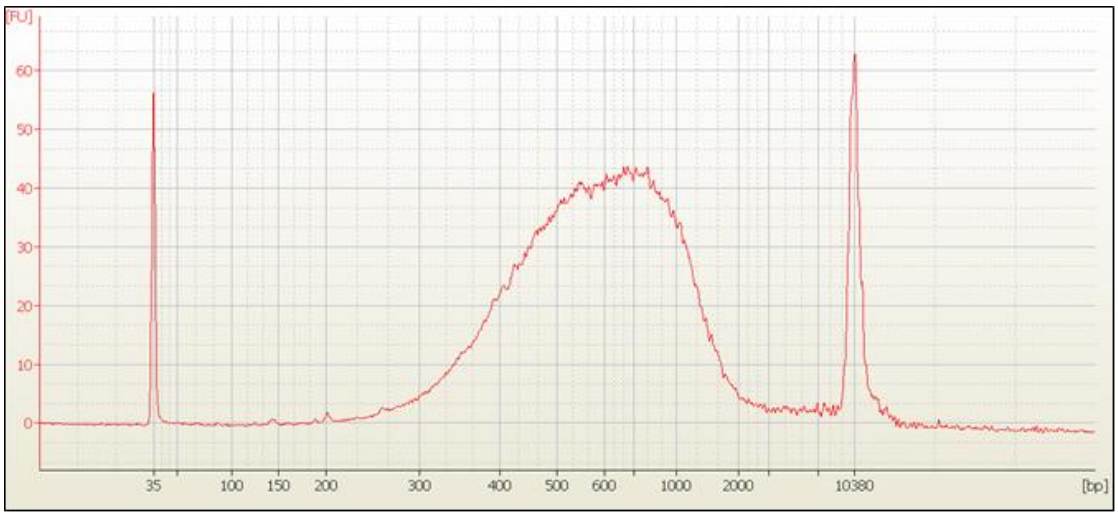
3. Starting material

Starting material is a nAnT-iCAGE library prepared according to CAGE library preparation kit manual. The sample will be arrived as dried pellet in a screw cap tube. Minimum 6 ng of library will be provided. Please spin the tube down before opening it and add _____µl of water to achieve final concentration of 1 ng / µl. Generally, 3 ng should be sufficient for sequencing with an Illumina sequencer 2000/2500.

Result of QC : Highsensitivity DNA kit

We will provide a result of final QC performed with High Sensitivity DNA kit and pico green analysis.

Example: A result from analysis with High Sensitivity DNA kit



4. Sample preparation for cluster generation (1hr)

Purpose: Prepare sample for cluster generation.

It is confirmed that the protocol provided from illumine, Inc. does not provide sufficient number of read. We strongly recommend changing the dilution rate for the sample and the amount of sample to be used for cluster generation referring to the following steps.

Table 1 : Required reagents, consumables and equipment for sample preparation for cluster generation

Reagents not included in the kit	Consumable	Equipment
☐ H_2O (R1)	☐ 10 µl filter tip	☐ P2 pipette
☐ 2 M NaOH	☐ 20 µl filter tip	☐ P20 pipette
☐ 1 M Tris-HCl(pH7.0)	☐ 200µl filter tip	☐ P200 pipette
☐ Hybridization buffer HT1 (Included in sample preparation for cluster generation(1hr))	☐ 1.5 ml tube	☐ Ice bucket ☐ Tabletop centrifuge

Preparation

・Dissolve the sample prepared in the Step2<i> 8 on ice.

・Stand 1 M Tris-HCl (pH7.0) and Hybridization buffer HT1 for more than 30 min on ice.

・Prepare a 1.5 ml tube and label with the sample name. (This is sample tube ⑬)

<Sample preparation>

1.☐ Based on the concentration that is measured in the QC3, dispense 3ng of the sample into the sample tube ⑬. (P2/P20)

Note: 3ng is preferable to acquire sufficient number of read. However, if the acquired amount of sample from the library preparation is less than 3ng, use the whole sample (7µl) for the sequencing.

2.☐ Add required amount of H_2O to make the total volume 19 µl to the sample tube ⑬ according to Table 81. Mix well by tapping. (P20)

Table 2 : Dilution of library solution

Reagent	1 sample	Equipment
☐ CAGE Library	☐ x µl (3ng)	P2/P20
☐ H_2O	☐ (19-x) µl	P20
Total volume	19 µl	

3.☐ Spindown in the tabletop centrifuge to collect the solution to the bottom of the tube.
4.☐ Add 1 µl of 2 M NaOH to the sample tube ⑬. Mix well by tapping. (P2)
5.☐ Spindown in the tabletop centrifuge to collect the solution to the bottom of the tube.
6.☐ Incubate for 5 min at room temperature.
7.☐ Transfer the sample tube ⑬ to ice. Add 20 µl/sample of 1 M Tris-HCl(pH7.0). Set the scale of the pipette at 30 and mix well by pipetting 10 times. (P200)
8.☐ Add 110 µl/sample of Hybridization buffer HT1 to the sample tube ⑬ on ice. Set the scale of the pipette at 100 and mix well by pipetting 10 times. (P200)
9.☐ Spindown in the tabletop centrifuge to collect the solution to the bottom of the tube.
10.☐ Load 120µl (out of the total amount 150 µl) of the sample from the sample tube ⑬ into the cluster generation(c-Bot).

5. Application to the Illumina HiSeq2000/2500 sequencer

Please follow Illumina protocols except for the index, which is three letter barcodes provided in the table below.

Label Number	Oligo name	Barcode
L1-1	2.5µM 5'CAGE Linker #01	**ACC**
L1-2	2.5µM 5'CAGE Linker #04	**CAC**
L1-3	2.5µM 5'CAGE Linker #05	**AGT**
L1-4	2.5µM 5'CAGE Linker #07	**GCG**
L1-5	2.5µM 5'CAGE Linker #09	**ATG**
L1-6	2.5µM 5'CAGE Linker #10	**TAC**
L1-7	2.5µM 5'CAGE Linker #15	**ACG**
L1-8	2.5µM 5'CAGE Linker #25	**GCT**

The amount of CAGE library recommended to apply: 3 ng
Read length: 50 base single read　(This is a minimum length required for the mapping, and the CAGE kit has been optimized for this condition.)
Primers: Illumina sequence primers can be used except for primers from v1 or v1.5 small RNA kits. The primers from TruSeq kit or Paired End kit can be used.

6. Expected Sequence result

We usually get apporx 80 - 120 million reads on HiSeq, We do not expect more than 120 million reads, due to the different size composition of the fragments in the CAGE sequence.

7. Software available for Data analysis

There are 2 pipelines regarding the analysis of CAGE data.

A) CAGE R

Programmed by Vania Haberle from the Lenhard lab,
http://www.bioconductor.org/packages/release/bioc/html/CAGEr.html

CAGEr package performs identification of transcription start sites and
frequency of their usage from input CAGE sequencing data,
normalization of raw CAGE tag count, clustering of TSSs into tag clusters (TC) and
their aggregation across multiple CAGE experiments to construct the promoterome.
It manipulates multiple CAGE experiments at once,
performs expression profiling across experiments both at level of individual TSSs
and clusters of TSSs, exports several different types of track files for
visualization in the UCSC Genome Browser, performs analysis of promoter
width and detects differential usage of TSSs (promoter shifting) between samples.
Multicore option for parallel processing is supported on Unix-like platforms.

B) Moirai

This software is developed by RIKEN.

* IDR paraclu ver.3.1 pipeline
http://cell-innovation.nig.ac.jp/wiki2/tiki-index.php?page=P000001286

* CAGE workflow
http://cell-innovation.nig.ac.jp/wiki2/tiki-index.php?page=5.+CAGE

To access to the site on NIG, you will have to first send the message to them that you wish to use MOIRAI.

They will soon ask you several questions, commercial use or not... etc,
and give you the ID with the password.

There are some additional linker sequences,
however, if you use MOIRAI, it should automatically detect the barcode sequence,
and will give you back the mapping data for each barcode.

8. Reference

1. Hasegawa et al., 2014 **Moirai**
   - ✧ BMC Bioinformatics. 2014 May 16;15(1):144.
   - ✧ PubMed PMID: 24884663; PubMed Central PMCID: PMC4033680.
2. Ohmiya et al., 2014 **RECLU**
   - ✧ BMC Genomics. 2014 Apr 25;15(1):269.
   - ✧ PubMed PMID: 24779366; PubMed Central PMCID: PMC4029093.
3. Dimont et al., **CAGExploreR: R package**
   - ✧ Bioinformatics. 2014 Mar 27. [Epub ahead of print]
   - ✧ PubMed PMID: 24675730.