Analysis of PacBio repeats through tweaked PfTools software suite

Thierry Schuepbach, Emmanuel Beaudoing

July 19, 2016

1 Description of the experiments

In an attemp to account tandem repeats inside PacBio Next Generation Sequencing Technology the following experiment was undertook. Two libraries L_1, L_2 were generated through the use of known Trinucleotide Tandem Repeats (TDR) taken from chosen regions of the human genome. The number of injected repeats is taken to be of length 50, 97, ..., 270. Those were cut of the genome keeping on each side part of the human genome. In addition a specific barcode was added prior to the smrtcell adapter to the sequence. Consequently, we should be in a position to identify each smrt cell hole provenance, and thus confirm the correctness of the potential detected number of TDR.

The difference between L_1 and L_2 still needs some explanation here?

The experiment was undertook with different initial conditions, that is

- MacBeads or diffusion smrt cell preparation
- DNA repair or not

Alltogether we have 8 different analysis to perform.

2 Raw data

Access to the raw data was granted by Emmanuel Beaudoing. It consists of:

- Library 1
  - Diffusion loading
    - m150603_150743_42182_c100829742550000001823181912311505
  - MagBead loading
    - m150812_171008_42182_c100858212550000001823192601241610
- Library 2
  - with DNA damage repair

    - m151124_111416_42182_c100906162550000001823201404301604
    - m151124_153718_42182_c100906162550000001823201404301605
  - without DNA damage repair
    - m151124_200325_42182_c100906162550000001823201404301606
    - m151125_002713_42182_c100906162550000001823201404301607

Barcodes used are:

1. ATCTGTGCGAGACTAC
2. AGACTCTACAGAGATA
3. TCTCTCACAGTCGAGC
4. GCTCGACTGTGAGAGA

3 Analysis performed by the GTF

-
-

4 Potential issues that could be cleaned by using PfTools software suite

The GTF analysis revealed that few sequences bearing 270 TDR were found. We have to be quite careful with such results as the method used to identify provenance is based upon exact barcode matching. Also, we need to check whether this is concensus or subread based. In the case of very long reads, as expected for numerous TDR, the concensus is not great. This could of course be checked through the quality given by the PacBio pipeline for each sequence. Anyway, it is of major importance to be able to identify barcode through PfTools profile simply to loose a bit on the exactness of the match.

In addition, Emmanuel believes those could be mainly be found within the subset of holes showing issues of two kinds:

- the polimerase was not able to cope with the first met smrt cell adapter and stopped there.
- the polimerase was not able to drive through the 3D structure of such reads due to their heavy number of TDR.

In any case, it is of main importance to be able to achieve a in depth analysis on where are those reads provided they exist. To that purpose, we propose to demultiplex the holes with PfTools software suite.

4.1 PfTools demultiplexing

Let us here compare the results obtained from perfect match over barcodes with the ones coming from profile methods. To that purpose, we have to generate foreach barcode a profile and run those over the sequencing holes. Hence, each hole shall have an average score in both direct and reverse complement based on its subreads. From there, we could check and provide to the best of the scoring system a provenance of the hole.

5