Derive the HiTS-EQ direct fitting Equation

Hsuan-Chun Lin Michael E. Harris
Case Western Reserve University

1.14.2015

1 Things before PCR

In the beginning, let us define some terms in this article. Before PCR The original substrate number for individual sequence in your sample without any binding S_0 The substrate number for individual sequence in the binding fraction $frac$ is S_1 The total sample counts before PCR is S_t and the sample counts before PCR in the binding fraction $frac$ is S_{t1}.

2 Things after PCR

In PCR, the amplification fold for S_t is n_t, and n_{t1} for the binding fraction S_{t1}. The total counting number after deep sequencing is D_t and D_{t1}. Also the counting number for S_0 and S_1 after PCR are D_0 and D_1. Because different primer or PCR for each sequence may have bias, we set the error to adjust the amplification fold as e_0 and e_1. For each sequence the amplification rate is $e_0 * n_t$ and $e_1 * n_{t1}$.

3 Equation derivation

The E means enzyme concentration The binding fraction could be written:

$$f = \frac{S_0 - S_1}{S_0} = \frac{E}{E + K_D}$$

And we could write the Deep sequecing read as

$$D_0 = S_0 \times n_0 e_0$$

$$D_1 = S_1 \times n_1 e_1$$

When we get our reads from High through-put sequencing, we can convert our reads to the original number of substrates as:

$$S_0 = \frac{D_0}{n_0 e_0}$$

$$S_1 = \frac{D_1}{n_1 e_1}$$

Now we can rewrite the fitting equation as:

$$f = \frac{\frac{D_0}{n_0 e_0} - \frac{D_1}{n_1 e_1}}{\frac{D_0}{n_0 e_0}} = \frac{E}{E + K_D}$$

By rearranging the equation we can get the following equation:

$$f = 1 - (\frac{n_0}{n_1}\frac{e_0}{e_1}\frac{D_1}{D_0})$$

We also know that

$$n_0 = \frac{D_t}{S_t}$$

$$n_1 = \frac{D_{t1}}{S_{t1}}$$

So we can get n_1/n_0 by:

$$\frac{n_0}{n_1} = \frac{D_t S_t 1}{D_t 1 S_t} = \frac{D_t}{D_t 1} \times (1 - frac)$$

We also make the basic assumption that for each individual sequence, the error from sample preparation between samples is the same. which means

$$\frac{e_0}{e_1} = 1$$

So our final equation for direct fitting is:

$$f = 1 - (\frac{D_t}{D_t 1} \times (1 - frac) \times \frac{D_1}{D_0}) = \frac{E}{E + K_D}$$

Here $\frac{D_{t1}}{D_t}$ is the ratio of total counts in sample with bounded fraction and the total counts in the unbound sample. The $\frac{D_1}{D_0}$ could be explained as mole fraction for each substrate.

4 Experimental approach

To make sure our basic postulate could establish, we need to focus on some details in the experiment. The most important one is PCR and primer design. Our first hypothesis is that the errors from PCR and primers for a individual sequence in different binding fractions are the same ($e_0/e_1 = 1$). Which means we need to make sure the amplification in PCR should be well controlled. The real-time PCR could help us to check the amplification curve for each sample, primer and barcode.