1 spike-in normalization for ChIP-seq libraries

N.B. I use 'experimental' and 'spike-in' throughout to refer to the two genomes, e.g. experimental signal and spike-in signal.

The point of including spike-ins in a ChIP-seq experiment is to end up with a normalized experimental signal which is proportional to the absolute abundance of the factor being IP'd. Let's say we're going to do the normalization by linearly scaling the experimental signal of a library by a normalization factor α. We can find α for each library by using the fact that a normalized 'spike-in signal' should be the same for all libraries, since the biological state of the spike-in cells is the same for all libraries. The key to finding α is defining exactly what this spike-in signal is for a library.

The measurement we start with to determine the spike-in signal is the number of reads in the library which map uniquely to the spike-in genome. This value will vary based on two factors: the sequencing depth of the library, and the proportion of cells which were spike-in cells. Let's assign variable names to the values mentioned:

$$R_{\text{spike}} \coloneqq \text{the number of reads in the library mapping uniquely to the spike-in genome}$$
$$\phi \coloneqq \text{the proportion of spike-in cells in the sample}$$

It will be easier to understand the derivation of α in terms of absolute cell numbers rather than ϕ, so let's also define those variables:

$$C_{\text{exp}} \coloneqq \text{the number of experimental cells used to prepare a library}$$
$$C_{\text{spike}} \coloneqq \text{the number of spike-in cells used to prepare a library}$$

We can express the **number of spike-in reads per spike-in cell** by simply taking the fraction $\frac{R_{\text{spike}}}{C_{\text{spike}}}$. We know that the biological state of a spike-in cell is the same no matter which sample it belongs to, so we *could* set $\frac{R_{\text{spike}}}{C_{\text{spike}}}$ equal to all samples in order to calculate α. However, this would not account for differences in ϕ between samples. Two libraries representing the same condition and sequenced to the same depth should have equivalent values of $\frac{R_{\text{spike}}}{C_{\text{spike}}}$ but they would not if they differed in the amount of spike-in added.

The metric for 'spike-in signal' that leads us to the correct form for α is the **number of spike-in reads per spike-in cell *per experimental cell***:

$$\frac{\frac{R_{\text{spike}}}{C_{\text{spike}}}}{C_{\text{exp}}} = \frac{R_{\text{spike}} C_{\text{exp}}}{C_{\text{spike}}}$$

From here, it's simple to calculate α by setting this value to be the same for all samples. Since the actual value of the spike-in signal doesn't matter as long as it's the same for all libraries, we can arbitrarily set it to 1 for convenience.

$$\alpha \frac{R_{\text{spike}} C_{\text{exp}}}{C_{\text{spike}}} = 1$$
$$\alpha = \frac{C_{\text{spike}}}{R_{\text{spike}} C_{\text{exp}}}$$

Notice that you only need to know the ratio of spike-in to experimental cells to calculate α, and not their absolute values. We can also rewrite this in terms of ϕ, the proportion of the sample that was spike-in cells:

$$\phi = \frac{C_{\text{spike}}}{C_{\text{spike}} + C_{\text{exp}}}$$
$$C_{\text{spike}} = \phi\left(C_{\text{spike}} + C_{\text{exp}}\right)$$
$$C_{\text{spike}}\left(1 - \phi\right) = \phi C_{\text{exp}}$$
$$\frac{C_{\text{spike}}}{C_{\text{exp}}} = \frac{\phi}{1 - \phi}$$

$$\alpha = \frac{C_{\text{spike}}}{R_{\text{spike}} C_{\text{exp}}}$$
$$\alpha = \frac{\phi}{R_{\text{spike}}\left(1 - \phi\right)}$$

Notice how this form for α is different than the one derived in Orlando et al. (2014):

$$\alpha = \frac{\phi}{R_{\text{spike}}\left(1 - \phi\right)} \qquad \alpha_{\text{orlando}} = \frac{\phi}{R_{\text{spike}}}$$

Let's do some examples with both versions to make sure the α values we get make sense, and compare them to α_{orlando}. First, we'll vary sequencing depth, keeping everything else constant. Consider a single ChIP library prep in which 20% of the cells were spike-in cells (i.e., $\phi = 0.2$). The library is then unevenly split into two aliquots and sequenced. One library has four times the reads of the other library.

$$R_{\text{spike}_1} = 1 \qquad R_{\text{spike}_2} = 4$$
$$R_{\text{exp}_1} = 4 \qquad R_{\text{exp}_2} = 16$$

$$\alpha_1 = \frac{\phi}{R_{\text{spike}_1}\left(1 - \phi\right)} \qquad \alpha_2 = \frac{\phi}{R_{\text{spike}_2}\left(1 - \phi\right)} \qquad \alpha_{\text{orlando}_1} = \frac{\phi}{R_{\text{spike}_1}} \qquad \alpha_{\text{orlando}_2} = \frac{\phi}{R_{\text{spike}_2}}$$
$$\alpha_1 = \frac{0.2}{1\left(0.8\right)} \qquad \alpha_2 = \frac{0.2}{4\left(0.8\right)} \qquad \alpha_{\text{orlando}_1} = \frac{0.2}{1} \qquad \alpha_{\text{orlando}_2} = \frac{0.2}{4}$$
$$\alpha_1 = \frac{4}{16} \qquad \alpha_2 = \frac{1}{16} \qquad \alpha_{\text{orlando}_1} = \frac{4}{20} \qquad \alpha_{\text{orlando}_2} = \frac{1}{20}$$

The total levels of spike-in normalized experimental signal can be found for each library by multiplying α by R_{exp}:

$$\text{signal}_1 = \alpha_1 R_{\text{exp}_1} \qquad \text{signal}_2 = \alpha_2 R_{\text{exp}_2} \qquad \text{signal}_{\text{orlando}_1} = \alpha_{\text{orlando}_1} R_{\text{exp}_1} \qquad \text{signal}_{\text{orlando}_2} = \alpha_{\text{orlando}_2} R_{\text{exp}_2}$$
$$\text{signal}_1 = \frac{4}{16}\left(4\right) \qquad \text{signal}_2 = \frac{1}{16}\left(16\right) \qquad \text{signal}_{\text{orlando}_1} = \frac{4}{20}\left(4\right) \qquad \text{signal}_{\text{orlando}_1} = \frac{1}{20}\left(16\right)$$
$$\text{signal}_1 = 1 \qquad \text{signal}_2 = 1 \qquad \text{signal}_{\text{orlando}_1} = 0.8 \qquad \text{signal}_{\text{orlando}_1} = 0.8$$

Only the relative abundances within normalization methods matter, so in this case both calculations correctly normalized for library size and say that the normalized signal in the two libraries are the same.

Now let's consider two libraries from two different conditions with $\phi = 0.1$. In condition 2, there is a known global decrease in experimental signal expected. I'll skip the algebra this time:

$$R_{\text{spike}_1} = 1 \qquad R_{\text{spike}_2} = 4$$
$$R_{\text{exp}_1} = 9 \qquad R_{\text{exp}_2} = 6$$

$$\alpha_1 = \frac{4}{36} \qquad \alpha_2 = \frac{1}{36} \qquad \alpha_{\text{orlando}_1} = \frac{4}{40} \qquad \alpha_{\text{orlando}_2} = \frac{1}{40}$$

$$\text{signal}_1 = 1 \qquad \text{signal}_2 = 1/6 \qquad \text{signal}_{\text{orlando}_1} = 0.9 \qquad \text{signal}_{\text{orlando}_1} = 0.15$$

Both methods correctly detect that experimental signal levels in library 2 are 1/6th that of library 1.

Finally, let's consider two libraries from the same condition which were spiked in with different amounts of spike-in cells. Both libraries are sequenced to the same depth. Since the libraries are from the same condition, we expect their total experimental signal to be the same after normalization, even though they had different amounts of spike-in added.

$$\begin{aligned} \phi_1 &= 0.2 \\ R_{\text{spike}_1} &= 2 \\ R_{\text{exp}_1} &= 8 \end{aligned} \qquad \begin{aligned} \phi_2 &= 0.4 \\ R_{\text{spike}_2} &= 4 \\ R_{\text{exp}_2} &= 6 \end{aligned}$$

$$\alpha_1 = \frac{\phi_1}{R_{\text{spike}_1}(1-\phi_1)} \qquad \alpha_2 = \frac{\phi_2}{R_{\text{spike}_2}(1-\phi_2)} \qquad \alpha_{\text{orlando}_1} = \frac{\phi_1}{R_{\text{spike}_1}} \qquad \alpha_{\text{orlando}_2} = \frac{\phi_2}{R_{\text{spike}_2}}$$

$$\alpha_1 = \frac{0.2}{2(0.8)} \qquad \alpha_2 = \frac{0.4}{4(0.6)} \qquad \alpha_{\text{orlando}_1} = \frac{0.2}{2} \qquad \alpha_{\text{orlando}_2} = \frac{0.4}{4}$$

$$\alpha_1 = \frac{3}{24} \qquad \alpha_2 = \frac{4}{24} \qquad \alpha_{\text{orlando}_1} = \frac{1}{10} \qquad \alpha_{\text{orlando}_2} = \frac{1}{10}$$

$$\text{signal}_1 = \alpha_1 R_{\text{exp}_1} \qquad \text{signal}_2 = \alpha_2 R_{\text{exp}_2} \qquad \text{signal}_{\text{orlando}_1} = \alpha_{\text{orlando}_1} R_{\text{exp}_1} \qquad \text{signal}_{\text{orlando}_2} = \alpha_{\text{orlando}_2} R_{\text{exp}_2}$$

$$\text{signal}_1 = \frac{3}{24}(8) \qquad \text{signal}_2 = \frac{4}{24}(6) \qquad \text{signal}_{\text{orlando}_1} = \frac{1}{10}(8) \qquad \text{signal}_{\text{orlando}_1} = \frac{1}{10}(6)$$

$$\text{signal}_1 = 1 \qquad \text{signal}_2 = 1 \qquad \text{signal}_{\text{orlando}_1} = 0.8 \qquad \text{signal}_{\text{orlando}_1} = 0.6$$

Here, our method correctly normalizes the two samples to the same total experimental signal while using the Orlando α results in an apparent decrease in signal in library 2. This is because the Orlando α fails to account for the fact that when you add more spike-in to a sample, you necessarily decrease the proportion of the sample that is experimental. In most experiments with spike-ins, this isn't really a problem because we assume that ϕ is the same for all samples. However, with ChIP-seq experiments that include input samples, if we assume that the experimental and spike-in input sample read counts are proportional to the amounts of experimental and spike-in cells mixed, we can plug these values in for values of ϕ to get a more reliable estimation of experimental signal levels. In this case, it becomes important to use the correct equation for α.

So, putting everything together, here's how I use the spike-in to normalize an IP ChIP-seq library paired with an input ChIP-seq library.

As stated above, we assume that the experimental and spike-in read counts in the input sample are proportional to the numbers of experimental and spike-in cells used to prepare the library:

$$\begin{aligned} R_{\text{input}_{\text{exp}}} &\propto C_{\text{exp}}, \\ R_{\text{input}_{\text{spike}}} &\propto C_{\text{spike}} \end{aligned}$$

Therefore, we can plug these values in for C for both the input and IP libraries (using the form of α without ϕ):

$$\alpha_{\text{input}} = \frac{C_{\text{input}_{\text{spike}}}}{R_{\text{input}_{\text{spike}}} C_{\text{input}_{\text{exp}}}} \qquad \alpha_{\text{IP}} = \frac{C_{\text{input}_{\text{spike}}}}{R_{\text{IP}_{\text{spike}}} C_{\text{input}_{\text{exp}}}}$$

$$\alpha_{\text{input}} \propto \frac{R_{\text{input}_{\text{spike}}}}{R_{\text{input}_{\text{spike}}} R_{\text{input}_{\text{exp}}}} \qquad \alpha_{\text{IP}} \propto \frac{R_{\text{input}_{\text{spike}}}}{R_{\text{IP}_{\text{spike}}} R_{\text{input}_{\text{exp}}}}$$

$$\alpha_{\text{input}} \propto \frac{1}{R_{\text{input}_{\text{exp}}}}$$

Notice how α_{input} reduces down to normalizing by the experimental library size, with no dependence at all on the spike-in. This makes sense because the input always represents the same state, regardless of how much spike-in is added to it. The function of the spike-in in the input is only to allow us to estimate abundances in the

corresponding IP library. Rewriting α_{IP} in the form below shows that α_{IP} will basically scale the experimental IP signal to the same scale as the experimental input signal, using the spike-in as a link between the two samples. This makes it natural to subtract the normalized input signal from the normalized IP signal: since they are on the same scale, the resulting coverage can be interpreted as reporting how much more IP signal was detected than was expected based on the input.

$$\alpha_{\mathsf{IP}} \propto \frac{1}{R_{\mathsf{IP}_{\mathsf{spike}}} \frac{R_{\mathsf{input}_{\mathsf{exp}}}}{R_{\mathsf{input}_{\mathsf{spike}}}}}$$

References

David A. Orlando, Mei Wei Chen, Victoria E. Brown, Snehakumari Solanki, Yoon J. Choi, Eric R. Olson, Christian C. Fritz, James E. Bradner, and Matthew G. Guenther. Quantitative chip-seq normalization reveals global modulation of the epigenome. *Cell Reports*, 9(3):1163 – 1170, 2014. ISSN 2211-1247. URL https://sci-hub.tw/10.1016/j.celrep.2014.10.018.