Contribution Write-up 1/13/2021-2/26/2021

University of Idaho Department of Statistical Science

Jarred Kvamme

2/25/2021

Incorporation of HiC Data To Verify Trans Mediated Regulation

• 2/12/2021 - 2/26/2021

HiC mapping quality thresholded chromatin interaction data was obtained from the ENCODE consortium for four tissues (lymphoblastoid cells, fibroblast cells, skin, and lung) to identify the presence of interaction enrichment for trios classified as trans-mediated cis regulation by **MRPC-ADDIS** (Davis et al. 2018; Consortium and others 2012). The binned interactions between genomic regions was extracted at 10,000 bp resolution using the *StrawR* package provided by Aiden Lab for use with the *.hic* file format (Durand et al. 2016). From this data, the total number of interactions between each trio's variant and it's associated trans gene was checked. This was done by summing all interaction counts for a 200,000 bp bin around the positions of both the variant and trans gene on their respective chromosomes. This was considered the observed number of interactions for a variant/trans-gene pair. Trios that had no observed interactions (or for which the data wasn't available) were desginated missing or "NA"

To identify enrichment, we conducted 10,000 resamplings of interactions between randomly selected and uniformly probable positions throughout the trans-gene and variant's respective chromosomes. As with the observed pairs, the total number of interactions was counted for a 200,000 bp bin around the gene and variant chromosomal positions. This served as a comparative unique null distribution of counts for each trio and was used to ascertain the upper-tail probability of a pair's observed number of interactions.

Many of the randomly selected 200,000 bp regions had either no interactions or no reads available. Without the ability to discern between unavailable data and 0 interaction values between two randomly selected regions, all empty interaction pairs were treated as missing data points and designated "NA". Therefore, the 10,000 resamples for each trio were partitioned into available data (non-NA) and unavailable data (NA's). In calculating the upper tail probability it is of significance to consider the quantity of unavailable data for a specific variant/trans-gene pair.

$$P_{obs} = P\big(A_i \geq A_{obs} \big| A_i \neq NA\big) = \frac{P(A_i \geq A_{obs} \ \cap A_i \neq NA)}{P(A_i \neq NA)}$$

using the following indicator functions:

$$\text{Let } f(A_i) = \begin{cases} 1, & \text{if } A_i \neq \text{NA } \forall \ i \in 1:N \\ 0, & \text{else} \end{cases}$$

$$\text{Let } g(B_j) = \begin{cases} 1, & \text{if } B_j \geq B_{obs} \ \forall \ j \in 1:n_1, \ B \subseteq A \\ 0, & \text{else} \end{cases}$$

Where n_1 is the number of resamples not in the set of NA's and n_2 be number of resamples in the complement event such that $n_1 + n_2 = N = 10,000$. Thus we have:

$$= \frac{\frac{\sum_{j=1}^{n_1} g(B_j)}{\sum_{i=1}^{N} f(A_i)} \times \frac{\sum_{i=1}^{N} f(A_i)}{N}}{\left(\frac{\sum_{j=1}^{n_1} g(B_j)}{\sum_{i=1}^{N} f(A_i)} \times \frac{\sum_{i=1}^{N} f(A_i)}{N}\right) + \left(\frac{n_1 - \sum_{j=1}^{n_1} g(B_j)}{\sum_{i=1}^{N} f(A_i)} \times \frac{\sum_{i=1}^{N} f(A_i)}{N}\right)}$$

$$= \frac{\frac{\sum_{j=1}^{n_1} g(B_j)}{N}}{\frac{\sum_{j=1}^{n_1} f(B_j)}{N} + \frac{n_1 - \sum_{j=1}^{n_1} f(B_j)}{N}}$$

$$\frac{\sum_{j=1}^{n_1} g(B_j)}{\sum_{j=1}^{n_1} g(B_j) + \left(n_1 - \sum_{j=1}^{n_1} g(B_j)\right)} = \frac{\sum_{j=1}^{n_1} g(B_j)}{n_1}$$

Note that a much simpler approach is to notice the exclusivity of the complement event $A_i = NA$ and the event $A_i \geq A_{obs}$ which leads to the realization that $P\big(A_i \geq A_{obs} \big| A_i \neq NA\big) \equiv P\big(B_j \geq B_{obs}\big)\ \forall\ j \in 1 : n_1$

Significant enrichment was defined as an observed probability less than the threshold α taken at the usual level of $\alpha = 0.05$. To control for the false detection of significant enrichment, two FWER and one FDR correction were applied to the observed probabilities which included: Holm-Bonferroni (FWER), BY method (FDR), and the BH method (FDR).

Some Notes On q-values, FWER and FDR:

• In the special case of all Null hypotheses being true the FWER and FDR are equivalent. In all other cases the FWER adjustments control the expected number of type I errors among all hypotheses/tests. The FDR controls the expected number of type I errors among significant tests.

• FWER methods are more stringent then FDR methods because of the consideration of all tests, therefore FDR methods are more powerful.

• In general, controling the FWER lowers the risk of a type I error at the expense of an increased risk of committing a type II error (failing to reject a null hypothesis). Holm-Bonferroni has a lower risk of type II error than the standard Bonferroni procedure and is therefore uniformly more powerful

• Just as the p-value gives the expected False Positive Rate (FPR) by rejecting any hypotheses with a p-value at or below the FPR, the q-value similarly gives the Positive False Discovery Rate (pFDR)/type I error by rejecting any hypothesis with a q-value at or below the pFDR

After adjusting for multiple comparisons none of the observed p-values remained significant for any of the correction methods. The histograms of the sampling distributions were retained for all trios and observed p-values were vetted against the sampling distribution.

Consortium, ENCODE Project, and others. 2012. "An Integrated Encyclopedia of Dna Elements in the Human Genome." *Nature* 489 (7414): 57.

Davis, Carrie A, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, et al. 2018. "The Encyclopedia of Dna Elements (Encode): Data Portal Update." *Nucleic Acids Research* 46 (D1): D794–D801.

Durand, Neva C, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. 2016. "Juicebox Provides a Visualization System for Hi-c Contact Maps with Unlimited Zoom." *Cell Systems* 3 (1): 99–101.