

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

DETAILED DESCRIPTION:

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). For each trio classified as trans-mediated, 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 designated 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.

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

using the following indicator functions:

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

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

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:

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

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 $P(A_i \geq A_{obs} | A_i \neq NA) \equiv P(B_j \geq B_{obs}) \forall j \in 1 : n_1$

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 FDR and one FWER correction were applied to the empirical probabilities which included: Holm-Bonferroni (FWER), BY method (FDR), and the BH method (FDR). After Correction, none of the empirical probabilities were significant.

All parts of the HiC/Trans Mediated verification process were repeated for the trans mediated trios classified by **MRPC-LOND**. Under both methods none of the empirical p-values remained significant after adjusting for multiple testing.

MAIN POINTS:

(1) Identification of Trans Mediated Trios and Respective Meta-Data

- HiC data for four tissues (Lung, Skin, Lymphoblastoid Cells, Fibroblast Cells) was obtained from the ENCODE consortium
- All trans-mediated (M1 type-2) classified trios from both **ADDIS** and **LOND** methods were extracted using *ADDIS.M1.check()* function
- For each trio the SNP, Cis, and trans gene chromosome number and base pair positions on the chromosome were obtained for alignment with HiC data

(2) Extraction of HiC Interaction Data

- HiC data for each individual trio was extracted using *StrawR* package in *R* and at a 10,000 base pair resolution (the interaction bins were in 10,000 bp increments)
- Extraction was carried out by summing the number of interactions between 200,000 bp regions around both the SNP and the trans-gene's chromosomal positions
- if no interactions existed for a SNP/trans-gene pair, the count/sum was replaced by "NA" to indicate the information was not available
- The count of interactions between the SNP and trans-gene was carried out and stored for each M1 type-2 trio classified by both **MRPC** methods

(3) Calculation of Empirical Probabilities

- Empirical probabilities were used to determine if an observed number of interactions for a SNP/trans-gene pair was significant
- Empirical probabilities were calculated by using 10,000 re-samplings to construct a empirical null distribution of interactions between the SNP chromosome and the trans-gene chromosome
- Two empirical probabilities were calculated: (1) the ratio of the number of non-NA resamples larger than the observed number of interactions to of the total number of non-NA resamples and (2) the ratio of the number of non-NA resamples larger than the observed number of interactions to the total number of resamples (See detailed description or subsection (4) for more detail)

(4) Re-sampling Methodology

- we re-sampled the number of interactions between the chromosomes of the SNP and trans-gene by allowing a 200,000 bp region/bin to vary randomly along each chromosome and summing the number of interactions between the bins/regions
- at each iteration a bin mid-point was selected both the SNP and trans-gene chromosomes by a random draw from a uniform distribution specific to each chromosome and the number of interactions for a $\pm 100,000$ bp around each mid point were counted
- the boundaries of the uniform distribution for a given chromosome were the minimum and maximum extents of the chromosome after adjusting the ends by $\pm 100,000$ bp to account for the bin size
- this process was repeated for 10,000 re-samplings and for each individual trio

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.