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

University of Idaho Department of Statistical Science

Jarred Kvamme

2/25/2021

MAIN POINTS:

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

• Mapping Quality thresholded chromatin interaction data for four tissues: Lung, Skin, Lymphoblastoid Cells, Fibroblast Cells (See table 1).

• 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

• We used the *StrawR* package in *R* to extract reads between specific regions on the SNP and trans-gene chromosomes.

• The extracted reads are returned as counts per bin for all bins nested within the desired region and at a desired resolution.

• In *StrawR* resolution refers to the bin-size applied to the counts between chromosomal regions. Therefore the number of interactions for a region is the sum of the counts for all bins nested in that region

• The resolution was arbitrarily set to 10,000 base pairs which was the smallest resolution shared by all tissues.

• To ensure coverage of the regulatory elements of both the SNP and trans-gene, we extracted and totaled all interactions between 200,000 base pair regions on each chromosome and centered at the SNP's/trans-gene's starting position on that chromosome

• if no interactions existed for a SNP/trans-gene pair, the count/sum was replaced by "NA" to indicate the information was not available

Table 1: File and dataset ID numbers for mapping quality thresholded chromatin interactions obtained from the ENCODE consortium

Tissue	**File ID**	**Dataset ID**
Lung	$ENCFF366ERB$	$ENCSR645ZPH$
Skin	$ENCFF569RJM$	$ENCSR912RAV$
Lymphoblastoid Cells	$ENCFF355OWW$	$ENCSR423SYP$
Fibroblast Cells	$ENCFF768UBD$	$ENCSR220MNC$

• 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 P-values By MC Sampling

• Empirical p-values were used to determine if an observed number of interactions for a SNP/trans-gene pair was significant

• Empirical p-values were calculated by randomly sampling the interactions between the SNP chromosome and the trans-gene chromosome to construct a empirical null distribution of interactions (see section (4) for more detail)

• Two empirical p-values were calculated:

(1) the ratio of the number of non-NA samples larger than the observed number of interactions to of the total number of non-NA samples

(2) the ratio of the number of non-NA samples larger than the observed number of interactions to the total number of samples

(4) MC Sampling Procedure

• A single sample point was defined as the total number of interactions between two randomly positioned 200,000 bp regions; one on the SNP chromosome and one on the trans chromosome.

• At each iteration in the sampling procedure, we first selected a single position on each chromosomes by a random draw from a uniform distribution with parameters tailored to the respective chromosome

• The randomly selected position was then treated as the center of a 200,000 bp region at that position. Again this was done for both the SNP and trans-gene chromosomes

• each position was sampled without the possibility of replacement, but regions had the potential to overlap on smaller chromosomes

• the min/max parameters of the uniform distribution used to select a position on a given chromosome were set to be the minimum base pair number represented by the data plus 100,000 bp and the maximum base pair number represented by the data minus 100,000 bp

• the parameters were constrained to the above values so that any randomly selected 200,000 bp region would lie completely within the available data for the respective chromosome.

• this sampling process was repeated 10,000 times for each SNP/trans-gene pair to produce a unique empirical null distribution representative of the number of interactions between the SNP chromosome and trans-gene chromosome

(5) Multiple Comparisons Adjustments

• Significant interaction enrichment was defined as an empirical p-value less than the threshold α taken at the usual level of $\alpha = 0.05$

• To account for multiple testing we employed 3 correction techniques to the empirical p-values:

(1) the Holm-Bonferroni (HB) Family Wise Error Rate (FWER) adjustment

(2) Benjamin Yekutieli (BY) control of the false discovery rate

(3) Benjamin Hochberg (BH) control of the false discovery rate

- the HB corrected p-values were compared at the original rejection threshold $\alpha = 0.05$
- The resulting q-values for the BH/BY methods were compared the desired false discovery rate $\beta = 0.05$
- After applying the corrections, none of the empirical p-values remained significant