Math behind **gchromVAR**

C. A. Lareau, J.C. Ulirsch, E.L. Bao

December 23, 2017

gchromVAR

The bias-corrected enrichment statistics for T traits and a set of S samples (chromatin cell type profiles) with P peaks computed by **gchromVAR** is a generalization of the **chromVAR** algorithm. Specifically, our implementation of **gchromVAR** relaxes the requirement that previously described[1] methods enforce where peak annotations be binary, allowing for uncertainty in annotations (such as transcription factor binding or localization of GWAS variants as we show here). Specifically, other implementations, including, `chromVAR`, implementation requires a binarized matrix $\boldsymbol{M}$ (dimension P by S) where $m_{i,k}$ is 1 if annotation k is present in peak i and 0 otherwise.[1] However, our examination of this assumption for binarized annotations performed poorly for GWAS enrichment applications due in part to the wide-range of uncertainties surrounding association results per-variant (see main text figures).

Instead, our methodology, **gchromVAR**, uses a matrix of variant posterior probabilities $\boldsymbol{G}$, where $g_{i,k}$ is the sum of the posterior probabilities of the variants contained in the genomic coordinates of peak i for each trait k. Using the matrix of fragment counts in peaks $\boldsymbol{X}$, where $x_{i,j}$ represents the number of accessible reads from peak i in sample j, the product $\boldsymbol{X}^T\boldsymbol{G}$ yields the total number of fragments weighted by the causal variant posterior probabilities for S samples (rows) and T traits (columns). To compute a raw weighted accessibility deviation, we compute the expected number of fragments per peak per sample in $\boldsymbol{E}$, where $e_{i,j}$ is computed as the proportion of all fragments mapping to the specific peak multiplied by the total number of fragments in peaks for that cell:

$$e_{i,j} = \frac{\sum_j x_{i,j}}{\sum_j \sum_i x_{i,j}} \sum_i x_{i,j}$$

Analogously, $\boldsymbol{X}^T\boldsymbol{E}$ yields the expected number of fragments weighted by the fine mapped variant posterior probabilities for S samples (rows) and T traits (columns). Using the $\boldsymbol{G}$, $\boldsymbol{X}$, and $\boldsymbol{E}$ matrices, we then compute the raw weighted accessibility deviation matrix $\boldsymbol{Y}$ for each sample j and trait k ($y_{j,k}$) as follows:

$$y_{j,k} = \frac{\sum_{i=1}^{P} x_{i,j} g_{i,k} - \sum_{i=1}^{P} e_{i,j} g_{i,k}}{\sum_{i=1}^{P} e_{i,j} g_{i,k}}$$

To correct for technical confounders present in assays (differential PCR amplification or variable Tn5 tagmentation conditions), **gchromVAR** borrows the strategy suggested previously[1] by generating a background set of peaks intrinsic to the set of epigenetic data examined. We note that other GWAS enrichment tools such as `LDSR` or `goShifter` ignore biases prevalent in epigenomic assays that are explicitly corrected by **gchromVAR**. In particular, variance in PCR or Tn5 tagmentation quality can lead to substantial differences in the number of counts between cells based on an individual peak's GC content or average accessibility, leading to errant GWAS enrichments. To correct for these technical confounders, each peak is assigned a background set of peaks that are matched in mean nucleotide GC content and average fragment accessibility between the sums of the cell types. An inverse Cholesky transformation is applied to a P by 2 matrix containing these variables to generate two uncorrelated dimensions describing the per-peak confounding. This two-dimensional space is divided into a pre-defined number of equally spaced bins where bin i is indicated β_i. Each peak q is assigned a bin from the shortest Euclidean distance between the bin's centroid and the individual peak in this transformed space. The probability that a peak q' in bin j is selected as a background peak for peak q is proportional to the distance between bins i and j over the total number of peaks in bin j ($|\beta_j|$):

$$P(q' \in \beta_j | q \in \beta_i) \propto \frac{d(\beta_i, \beta_j)}{|\beta_j|}$$

where the distance function d contains hyperparameters which, along with the total number of bins, have been verified using simulations previously.[1]

By default, the framework used to sample a background set uses 50 background elements per peak, which we've verified to be robust (Supplemental Result). The matrix $\boldsymbol{B}^{(b)}$ encodes this background peak mapping where $b_{i,j}^{(b)}$ is 1 if peak i has peak j as its background peak in the b background set ($b \in \{1, 2, ..., 50\}$) and 0 otherwise. The matrices $\boldsymbol{B}^{(b)}\boldsymbol{X}$ and $\boldsymbol{B}^{(b)}\boldsymbol{E}$ thus give an intermediate for the observed and expected counts also of dimension P by S. For each background set b, sample j, and trait k, the elements $y_{j,k}^{(b)}$ of the background weighted accessibility deviations matrix $\boldsymbol{Y}^{(b)}$ can be computed:

$$y_{j,k}^{(b)} = \frac{\sum_{i=1}^{P} (\boldsymbol{B}^{(b)}\boldsymbol{X})_{i,k} g_{i,k} - \sum_{i=1}^{P} (\boldsymbol{B}^{(b)}\boldsymbol{E})_{i,k} g_{i,k}}{\sum_{i=1}^{P} (\boldsymbol{B}^{(b)}\boldsymbol{E})_{i,k} g_{i,k}}$$

After the background deviations are computed over the 50 sets, the bias-corrected matrix $\boldsymbol{Z}$ for sample j and trait k, ($z_{j,k}$) is computed as follows:

$$z_{j,k} = \frac{y_{j,k} - \text{mean}(y_{j,k}^{(b)})}{\text{sd}(y_{j,k}^{(b)})}$$

where the mean and variance of $y_{j,k}^{(b)}$ is taken over all values of b ($b \in \{1, 2, ..., 50\}$). Sample-trait p-values can then be computed from the one-tailed normal distribution of these z-scores using the `pnorm` function in `R`. Our implementation of **gchromVAR** utilizes efficient sparse-matrix operations for each step and can compute pairwise trait-cell type enrichments in 1 minute on a modern laptop computer for these results.

References

[1] Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. *Nature Methods.* 2017 Oct 1;14:975.