Methods for populating the gene sets tested

The method used to determine the gene set types/databases and their assigned genes is species-dependent:

Human: GO data is from the R package GO.db. All other data is from ConceptGen (http://conceptgen.ncibi.org).

Mouse and Rat: Cytoband data is from the R package org.Mm.eg.db for mouse or org.Rn.eg.db for rat. GO gene sets is from the R package GO.db. For all other concept types, Homologene from NCBI is used to obtain the human homologs and the ConceptGen database (http://conceptgen.ncibi.org).

ChIP-Enrich Statistical Method

ChIP-Enrich uses a logistic regression approach to simultaneously 1) adjust for the gene locus length and mappability, and 2) test for gene set enrichment. Our model is shown below:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 geneset + f\left(log_{10}(locus\ length * mappability + 1)\right)$$

The probabilities π_i are defined as the probability of gene *i* being assigned a peak given the data. The π_i values are not observed directly, instead we observe only whether each gene was assigned a peak. Only those genes annotated in GO and present within our locus definitions are used. Our dependent variable is then a binary vector with 1 if the gene has a peak assigned to it, and 0 otherwise. The parametric term *geneset* is also a binary vector, where 0 denotes that the gene does not belong in the set of genes being tested and 1 otherwise. The function *f(log$_{10}$ (locus length * mappability+1))* is a cubic smoothing spline term that takes into account both the locus length and the average mappability of each gene locus. We apply the log transformation to locus length as this greatly improves the model fit. The constant of 1 is added to avoid taking the log of 0. The spline is fit with 10 knots, distributed evenly throughout the data. The

model is fit using a penalized likelihood maximization approach, where the smoothing penalty is the conventional squared second derivative penalty, and where the smoothing parameters are estimated using generalized cross-validation. **More detailed methods are provided in the original publication.**

Adjusting for mappability is optional. We have pre-calculated mappability values for each gene in each pre-defined locus definition. We define base pair mappability as the average read mappability of all possible reads of size K that encompass a specific base pair location, b. Mappability files from UCSC Genome Browser mappability track were used to calculate base pair mappability. The mappability track provides values for theoretical read mappability, or the number of places in the genome that could be mapped by a read that begins with the base pair location b. For example, a value of 1 indicates a Kmer read beginning at b is mappable to one area in the genome. A value of 0.5 indicates a Kmer read beginning at b is mappable to two areas in the genome. For our purposes, we are only interested in uniquely mappable reads; therefore, all reads with mappability less than 1 were set to 0 to indicate non-unique mappability. Then, base pair mappability is calculated as:

$$M_i = \left(\frac{1}{2K-1}\right) \sum_{j=i-K+1}^{i+(K-1)} M_{readj}$$

where M_i is the mappability of base pair *i*, and M_{readj} is mappability (from UCSC's mappability track) of read *j* where *j* is the start position of the *K* length read. We define gene locus mappability as the average of all base pair mappability values for a gene locus. We calculated gene locus mappability, which we simply refer to as 'mappability' on the ChIP-Enrich webpage, for reads of lengths 24, 36, 40, 50, 75, and 100 base pairs for *Homo sapiens* (build hg19) and for reads of lengths 36, 40, 50, 75, and 100 base pairs for *Mus musculus* (build mm9). Currently, no mappability is available for *Rattus norvegious* (build rn4).

Advanced Analysis Options

Filter: The gene sets tested can be filtered by number of genes. Decreasing the maximum number of genes required for a gene set to be tested may result in faster completion of test.

Enrichment Method: In addition to ChIP-Enrich, Fisher's exact test is also provided. We recommend using Fisher's exact test only with the 1kb, 5kb, or a locus definition with uniform loci lengths.

Locus Definition: The 1kb, 5kb, Exon, Nearest Gene, and Nearest TSS locus definitions are predefined. The 1kb and 5kb locus definitions are equivalent to assigning peaks within 1kb or 5kb of the transcription start site respectively. The Exon locus definition is equivalent to using only peaks that occur within an annotated exon. The Nearest Gene and the Nearest TSS locus definitions are equivalent to assigning peaks to the nearest gene or TSS respectively. We also provide the user to define their own locus definition, in which case the format should be have the following columns: entrez gene ID, chromosome, start of locus, end of locus.

Adjust for the mappability of the gene locus regions: If true, locus length is adjusted for mappability. The associated pre-calculated mappability values for the chosen locus definition will be used in the enrichment test. In the case where the locus definition is 'User Defined,' user defined mappability values should also be provided. If false, locus length is not adjusted for mappability, which is equivalent to setting all mappability values to equal 1, i.e. as if all loci are equally mappable.

References

R.P. Welch, C. Lee, R.A. Smith, Patil S, P. Imbriano, L.J. Scott, M.A. Sartor. "ChIP-Enrich: gene set enrichment testing for ChIP-seq data." In preparation