from Gene set enrichment analysis of RNA-Seq data:
integrating differential expression and splicing
Xi Wang, Murray J Cairns, BMC Bioinformatics 2013

Enrichment score (ES) is the main concept in GSEA. Given an *a priori* defined gene set X, the ES is a Kolmogorov-Smirnov-like statistic and reflects the degrees how the genes in X overrepresented at the top of the gene list sorted by gene scores in a descending order. To compute ES, we first rank gene scores decreasingly, forming an ordered gene list $L = \{g_1, g_2, \ldots, g_G\}$. By walking down the list, we introduce two vectors to assist ES computation: one is for the fraction of genes in X up to a position i in list L, weighted by a factor depending on their gene scores, denoted by $P_{\text{in}}(X, i)$; the other for the fraction of genes not in X, denoted by $P_{\text{out}}(X, i)$.

$$P_{\text{in}}(X, i) = \sum_{g_j \in X, j \le i} \frac{S_{g_j}^p}{W}, \quad P_{\text{out}}(X, i) = \sum_{g_j \notin X, j \le i} \frac{1}{G - G_X} \tag{16}$$

where $W = \sum_{g_j \in X} S_{g_j}^p$, representing a normalization factor for genes in set X, while $(G - G_X)$ is for genes outside and G_X denotes the total number of genes in gene set X. Based on the two vectors, ES can be computed as (Supplementary Figure S1d)

$$E_X = \max_i [P_{\text{in}}(X, i) - P_{\text{out}}(X, i)] \tag{17}$$

We denote the position where the enrichment score is achieved as i_0, and the leading set X_0 is defined as the subset of X whose positions in list L not behind i_0. Notably, when $p = 0$, E_X reduces to the standard Kolmogorov-Smirnov statistic; when $p = 1$, genes in X are weighted just by their gene scores; when $p > 1$, genes in X with large gene scores will be weighted exponentially more. We set $p = 1$ for the analyses in this study.

To estimate the significance level of ES, we perform empirical permutation tests by shuffling samples' group labels. The permutation serves a null distribution for the observed ES, so the empirical p-values can be calculated according to this null distribution. For the adjustment of multiple hypothesis testing when multiple gene sets evaluated simultaneously, E_X are normalized and made comparable across the whole datasets of gene sets. Similar to the normalization of DE/DS scores, the normalized score $E_{X,\text{norm}}$ equals to E_X divided by the mean value of permutation ES for the same gene set X. Then, *FDR* is defined as the ratio of the number of normalized permutation ESs exceeding $E_{X,\text{norm}}$, to the number of normalized observed ESs no less than $E_{X,\text{norm}}$. We set the number of permutations to be 1,000 to generate results throughout this study.