ARB A new tool to prioritize candidate genes and characterize sample behavior in differential expression analysis of transcriptomic data

Danielle Novick[1], Michael B. Papah[2], M. Joseph Tomlinson IV[1], José Daniel Chazi Capelo[1], Behnam Abasht[1,2]

University of Delaware (1) Center for Bioinformatics and Computational Biology (2) Department of Animal and Food Sciences

What is ARB?

Differential gene expression studies can produce staggeringly long gene lists. Many researchers use fold-change (FC) and p-value to prioritize these genes, but this does not completely account for the ability of individual samples to skew a group's average. ARB is a Python-based tool that classifies genes based on the degree to which expression values between experimental groups overlap (see "How does ARB work?") and assesses the behavior of each sample across all genes. The graphic below shows how ARB fits into a typical differential expression (DE) analysis pipeline.

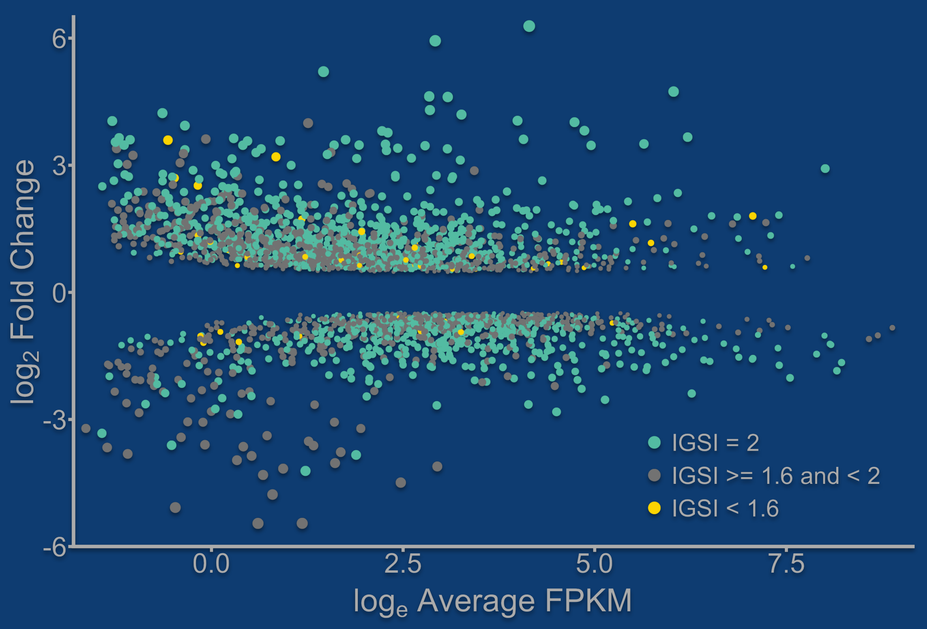

FIGURE 1. Relationship between FC and average FPKM for the differentially expressed genes from dataset 2. Genes with high IGSI (max is 2) and high |FC| should be considered "high priority" because the two conditions cluster independently and FC isn't driven by only a small number of samples.

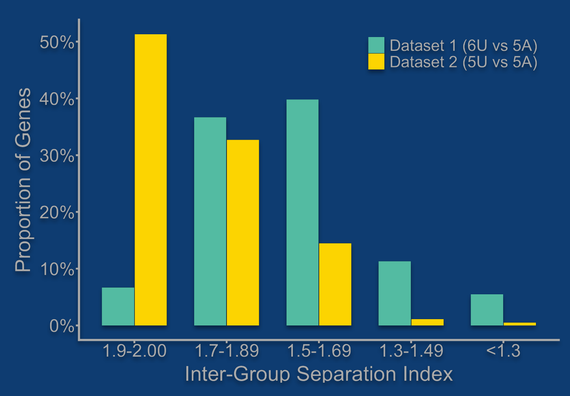

FIGURE 2. IGSI distributions for both datasets. Dataset 1 IGSIs are lower due to a misclassified sample.

Acronyms

ARB Analysis of RNA-seq Data Behavior

FPKM Fragments Per Kilobase of transcript per Million mapped reads

FC Fold-Change

IGSI Inter-Group Separation Index

Demo Data

Data is from a study of Wooden Breast Disease (WBD) in broiler chickens where affected (A) chickens showed clinical signs of WBD and unaffected (U) did not. Dataset 1 has 11 samples, one of which is misclassified. Dataset 2 has the misclassified sample removed.

Why use ARB?

1. Prioritize candidate genes

IGSI and FC are used together to identify genes where all samples support a high-magnitude difference between conditions. ARB's prioritization strategy is more holistic than using FC alone because it incorporates the relative expression of each sample rather than relying solely on average FPKM values between conditions (see Figure 1).

2. Identify misclassified samples

If a sample is consistently removed from a large proportion of genes in order to achieve complete separation of expression values by condition, that sample might be misclassified. This high exclusion rate will be reflected in a lower average IGSI for the dataset containing a misclassified sample (see Figure 2).

3. Identify outlier samples

ARB uses the interquartile range method to identify samples that are moderate or extreme outliers based on a sample's expression across all genes. An "outlier" value is relative to all samples, not just those in the same condition.

4. Rank samples by relative expression across all genes

Rank-order analysis is useful for recognizing trends in expression for each sample across all genes. Samples are ranked by FPKM with the order of ranking adjusted by the direction of the FC (see Figure 3).

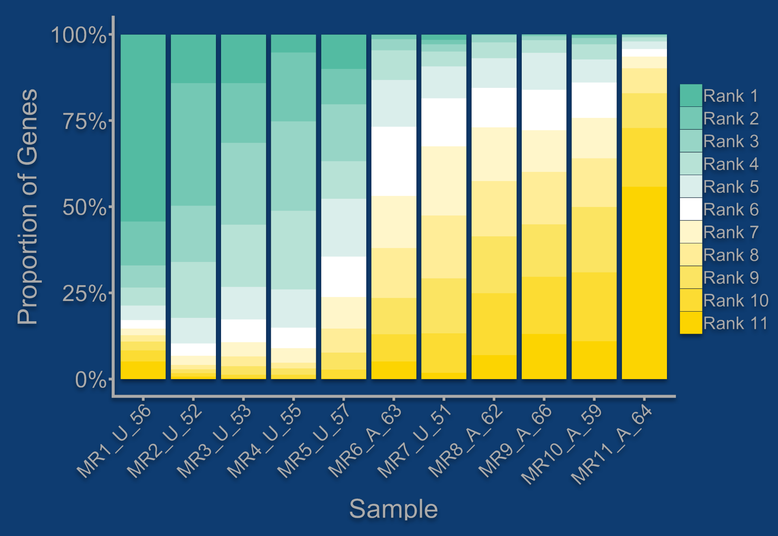

FIGURE 3. Rank-order analysis of samples from dataset 1 showing the distribution of rankings for each sample based on FPKM values. Notice that MR7_U_51 is classified as unaffected (U), but has a mean-rank (MR) that places it among the affected (A) samples. This is another indicator that the sample is misclassified.

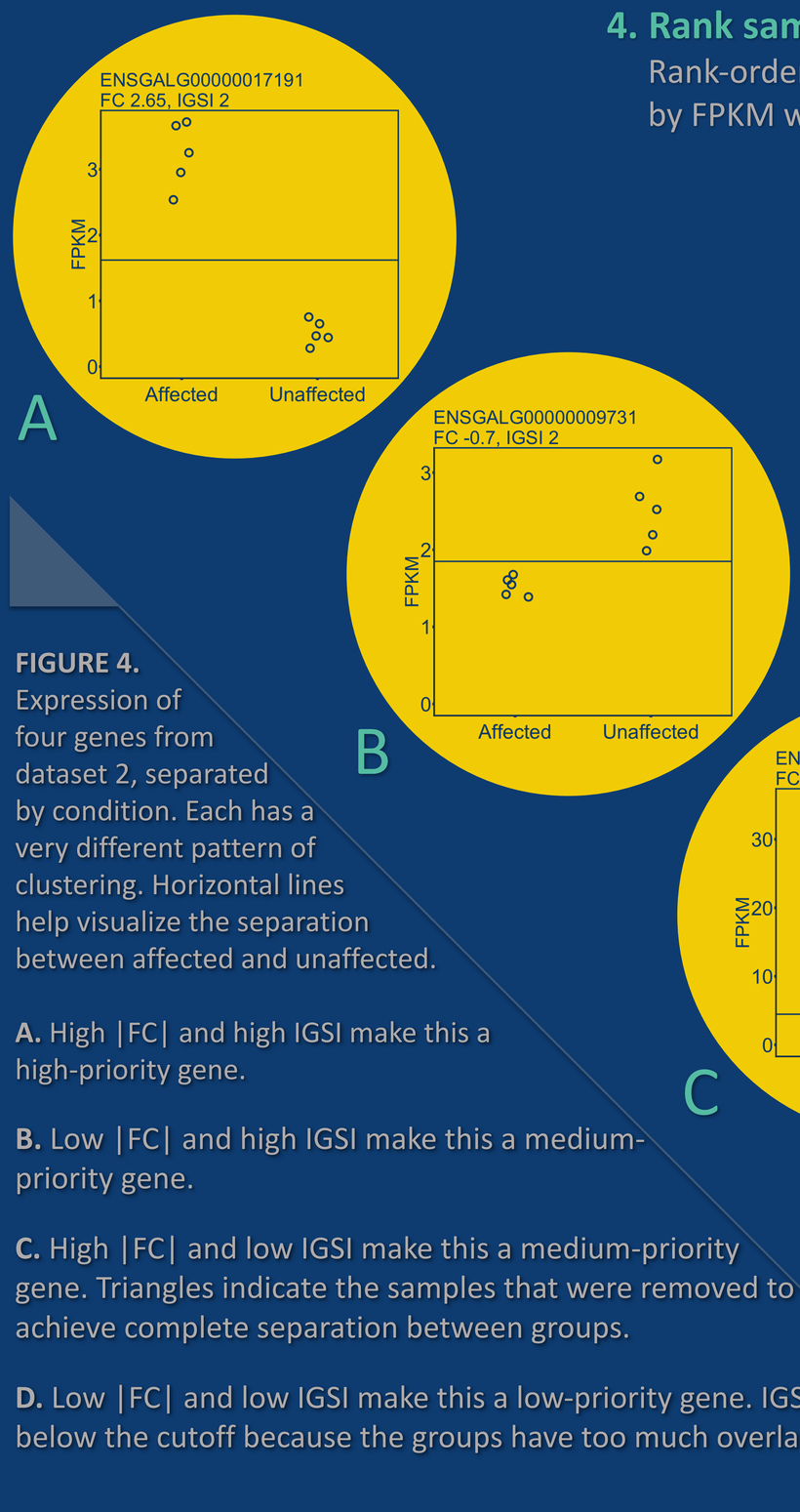

FIGURE 4. Expression of four genes from dataset 2, separated by condition. Each has a very different pattern of clustering. Horizontal lines help visualize the separation between affected and unaffected.

A. High |FC| and high IGSI make this a high-priority gene.

B. Low |FC| and high IGSI make this a medium-priority gene.

C. High |FC| and low IGSI make this a medium-priority gene. Triangles indicate the samples that were removed to achieve complete separation between groups.

D. Low |FC| and low IGSI make this a low-priority gene. IGSI is below the cutoff because the groups have too much overlap.

How does ARB work?

ARB examines each differentially expressed gene and applies a supervised subsetting algorithm that incrementally excludes samples up to a cutoff set by the user to achieve complete separation of expression values between experimental conditions. If a subset can be assembled that meets the user's constraints, then the following is calculated: (i) inter-group separation index (IGSI); 0-2 scale where a value of 2 corresponds to a subset where no samples have been removed (ii) FC using only that subset of samples, and (iii) the identities of the samples that have been excluded from the subset. With this information, researchers can easily select the genes with the highest IGSI and FC for further scrutiny. ARB also has functions to identify outliers, rank samples based on relative expression across genes, and identify misclassified samples based on frequency of exclusion.

Find ARB along with sample data at:
https://github.com/AbashtLaboratory/ARB