Venn diagrams indicate erroneous analyses in transcriptomics

2021-10-26

Venn diagrams are frequently used to illustrate the results of differential expression analysis, and are an illustration of further downstream process of functional analysis. Here we show that both the use of Venn diagrams to find genes which are thought to be specific for a certain comparison as well as gene set enrichment analysis applied to such subsets is a fallacy and results in artifacts.

Table of Contents

## Introduction: Venn diagrams are a showcase of erroneous interpretation of interactions

Venn diagrams (VDs) are commonly used to visualize high throughput data such as transcriptomic profiles. Frequently, a Venn diagram serves the purpose of comparing the results from two experiments or two comparisons. For example, a VD may illustrate up- or down-regulated genes in two strains of mice upon infection, showing which genes are regulated in both strains, and which are regulated only in one of them. The VD serves as a basis for the statement that the regulation of certain genes is “specific” for one strain or another. We argue that not only is this inference incorrect, but it may also lead to misleading – although appealing – artifacts when combined with downstream analyses.

Consider a gene for which the expression has been analysed in four groups: two different mouse strains (WT and KO) and two different experimental conditions (naive versus infected). We find that the gene expression significantly differs between infected and naive KO animals, but that there is no significant difference in the WT strain. Such a gene may be incorrectly considered as “specific” to KO, and will be accordingly entered in a Venn diagram.

However, this is a common fallacy [@nieuwenhuis2011erroneous], since lack of statistical significance is not the same as a significant lack of difference. In other words, the fact that we failed to detect a significant difference in the WT does not mean that the difference is absent and significantly different from the difference in the KO. This “difference of differences” is known is statistics as interaction (here, interaction between strain and treatment). In fact, the obtained p-values might be just over the assumed threshold in one, but just under it in the other strain (e.g. 0.009 vs 0.011), with the difference in both strains being essentially the same. The correct analysis is to test the significance of interaction using an ANOVA model. Notably, in 2011 Nieuwenhuis et al. found that out of half papers in top neuroscience journals where the authors could make this particular statistical error, half of them did.

Using VDs to show genes “specific” for a condition amounts, therefore, to counting the times a comparison for a gene was statistically significant in one condition, but not significant in another.

More disturbingly, Venn diagrams are a visual illustration of a procedure that exacerbates this problem by applying a downstream analysis to supposedly specific sets of genes. Several cells in a Venn diagram contain variables that were selected based on the fact that they are significant in one, but not significant in another comparison. Variables in a cell may then subsequently analysed to test whether a share a particular characteristics. For example, gene set enrichment analysis may be used to interpret the biological function of genes that are “specific” to one condition, but - supposedly - not to another. It turns out that due to the fallacious nature of this procedure, this is likely to produce results that appear reasonably in the context of the analysed biology. This will be illustrated in the following with an example.

In this paper, we illustrate this rather simple statistical statement with a simulated data set, demonstrating how choosing this sort of approach results in apparently sound gene set enrichment results, which are in fact artifacts. Next, we dissect the underlying mechanism of how these artifacts are generated. Finally, we explore and demonstrate alternative approaches.

# Results

## Example 1: Transcriptomic changes due to Sars-Cov-2 infection

Consider two group of patients, G1 and G2. Each group contains 20 individuals. In both groups, there are patients who are either healthy (“no infection”) or infected with Sars-Cov-2 (“SC2”). Our aim is to understand which genes or pathways are specifically upregulated by SC2 infection in G1 compared to G2, and vice versa. In the following, we used the data set XXX [ ] to perform two such analyses, which arrive to opposite conclusions.

First, we have performed differential gene expression analysis for each of the groups G1 and G2 separately. For each comparison, we defined genes with significantly different gene expression (DE genes) by using an FDR threshold of 0.05 and absolute log2 fold change (LFC) threshold of 1. There were 114 DE genes in the G1 group, and 1491 in the G2 group.

Fig. IncorrectAnalysis. **Results of differential gene expression analysis and gene set enrichment analysis using a Venn Diagram driven approach.**

## Example 1: group specific responses to tuberculosis (TB)

The data in the following example comes from a study of transcriptomes of TB patients and their comparison with healthy individuals [@kaforou2013detection]. The transcriptomes of TB patients were here compared with healthy controls in two groups, each containing 20 individuals, and the experimental question was whether the regulation in TB is different between these two groups labeled here “A” and “B”. The results of a Venn diagram driven analysis is shown in Fig. X. In short, for both groups, the transcriptomes of TB patients were compared to the transcriptomes of the healthy individuals. Differentially expressed genes were acquired by using a cutoff of 1 for absolute log2 fold change and 0.01 for q-value. The results are shown on a Venn diagram: 68 out of 107 differentially regulated genes in group A were only found in this comparison, while 49 genes were “specific” for group B.

Gene set enrichment analysis applied to the sets of “specific” genes shows a stunning picture (Fig. ): group A shows several gene sets related to T cell modules, while most of the enriched gene sets in group B correspond to innate immunity. The researcher might then conclude that group A is stronger in adaptive, and group B - in innate immune response. The main problem with this interpretation is that groups A and B were randomly sampled from a single population. Why then are there statistically significant changes? Why are the genes that are only significant in the group A comparison enriched in functions of the immune system?

Consider the gene

## “Specific” genes aren’t

## Example 2: TB specific responses

In the same study there are three real groups of individuals defined: healthy (but *M.tb.* infected) controls (“LTBI”), TB patients and patients suffering from diseases other than TB (“OD”). In the example below, the differences between TB patients and healthy controls (LTBI) are compared to differences between OD and LTBI, and for this, the complete data set with HIV- individuals from South Africa is used. Again, both approaches will be used: (i) comparing the results of the two comparisons, and (ii) testing the “difference of differences”. It is easy to see that since the group of reference (LTBI) is common in both comparisons, the interaction is actually equivalent to difference between TB and OD.

## Artifacts arise because of false negatives

To understand how gene set enrichment actually gives any significant results in such randomly generated gene sets and despite no genes with significant interaction, it is first necessary to consider the definition of a differentially expressed gene in this context. Most often than not, DEGs are defined by a threshold in p-value adjusted for multiple testing, possibly combined with a threshold in log2 fold change. The commonly used Benjamini-Hochberg procedure ensures that among genes for which the adjusted padj < 0.05 there are at most 5% false positives.

# Methods

## Data

The data has been downloaded from GEO, accession GSE37250.

## Statistical analyses

## Methods availability

This document has been written as an Rmarkdown file, containing all statistical calculations required to replicate the findings.