Erroneous use of Venn diagram leads to artifacts in transcriptomic analyses

January Weiner1,†, Benedikt Obermayer1 and Dieter Beule1

2021-11-19

1 Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Unit Bioinformatics, Charitéplatz 1, 10117 Berlin, Germany

† To whom the correspondence should be addressed

## Abstract

Venn diagrams are frequently used to illustrate the results of differential expression analysis and further downstream 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. Moreover, the combination of incorrect statistical reasoning and gene set enrichment analysis may result in particularly misleading artifacts.

## 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 (wildtype, WT and knockout, 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 ([Nieuwenhuis, Forstmann, and Wagenmakers 2011](#ref-nieuwenhuis2011erroneous)), since lack of statistical significance is not the same as testing for 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 itself significantly different from the difference in the KO. This “difference of differences” is known is statistics as interaction ([Blalock Jr 1965](#ref-blalock1965theory)). Here, it is the interaction between strain and treatment. In fact, the obtained p-values might be just over the assumed p-value 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 in over 150 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 to counting the times a comparison for a gene was statistically significant in one condition, but not significant in another. More disturbingly, VDs 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 numbers corresponding to features (such as genes), for which there was a significant test result in one, but not in another comparison. The sets of genes in a Venn diagram cell may then be subsequently analysed to test whether they share a particular characteristics. For example, gene set enrichment analysis may be used to interpret the biological function of genes that are considered “specific” to one condition. It turns out that due to the fallacious nature of this procedure, it is likely to produce results that appear to make sense in the given biological context.

In this paper, we illustrate this rather simple statistical statement with a real world 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 discuss alternative approaches.

## Results

### Transcriptomic changes due to Sars-Cov-2 infection

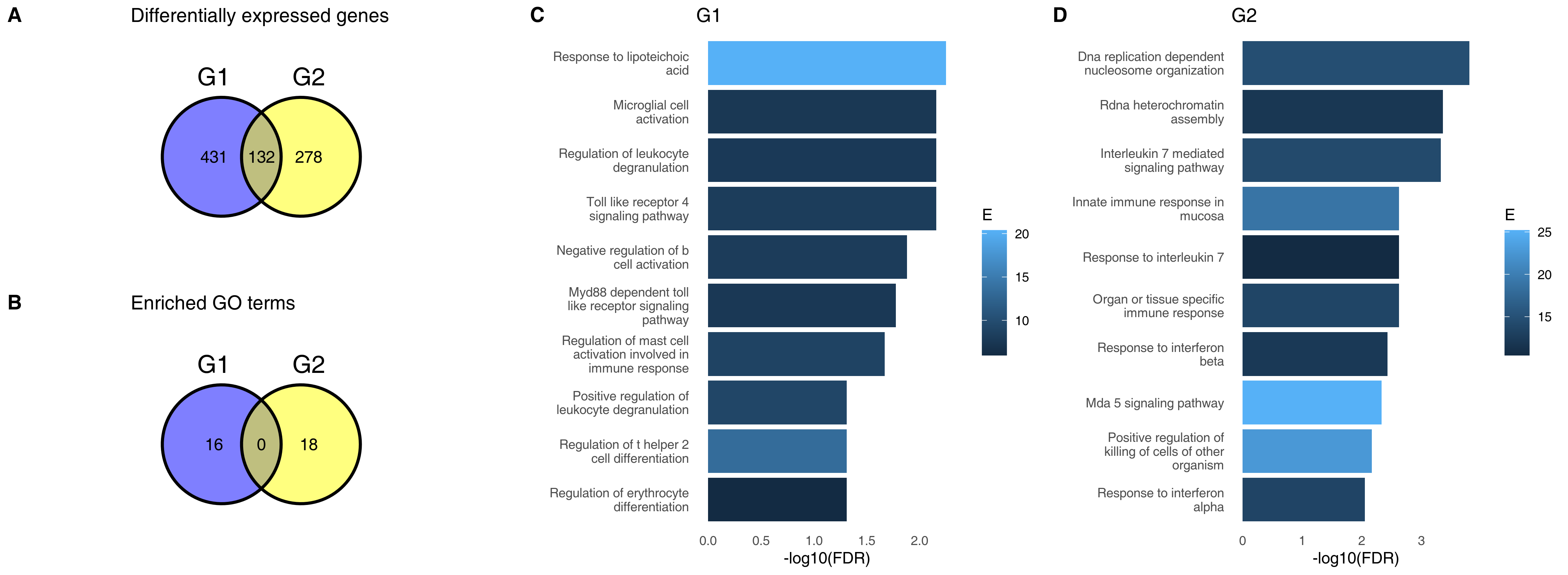
Consider two group of patients, G1 and G2. Each group contains 40 individuals. In both groups, there is an equal number of healthy individuals (labeled “Ctrl” on figures below) or patients infected with Sars-Cov-2 (labeled “SC2”). Our aim is to understand the differences between G1 and G2 in the response to infection. For example, we ask which genes or pathways are specifically upregulated by SC2 infection in G1 as compared to G2, and vice versa. In the following, we used the data set GSE156063 ([Mick et al. 2020](#ref-mick2020upper)) in two approaches (an incorrect and the correct one) which arrive at opposite conclusions.

First, we have performed differential gene expression analysis for each of the groups G1 and G2 separately using standard bioinformatic tools. For each comparison, we defined genes with significantly different gene expression (differentially expressed genes, DEGs) by using an FDR threshold of 0.05 and absolute log2 fold change (LFC) threshold of 1. There were 563 DE genes in the G1 group, and 410 in the G2 group. In total, 132 DEGs were common for G1 and G2, 431 DEGs were significant in G1 only (“specific” for G1), and 278 were significant in G2 only (“specific” for G2; see Fig. 1, panel *A*). A naive interpretation of these results implies that there is a substantial difference between these two groups of individuals, as evidenced by a small overlap in commonly regulated genes, while the majority of DEGs is significant in one comparison only.

To understand which pathways are upregulated in each of the two groups, we used a standard generation I gene set enrichment analysis (GSEA) – a hypergeometric test – on the DEGs in each group. Gene sets for GSEA were taken from the Gene Ontology (GO) database. Gene sets with more than 50 or fewer than 10 genes were removed. For each group, we have selected only genes which are DEGs in that group, but not the other, mimicking a naive approach for finding “specifically” regulated pathways. Here, a similar picture emerged. Just 16 gene sets were significantly enriched in G1, and 18 gene sets were significantly enriched in G2. Again, both the Venn diagram (Fig. 1, B) and the results of enrichments (Fig. 1, C and D) suggest that there is a fundamental difference between the groups, and that the groups have little in common in their response to the virus.

Importantly, the different GO terms enriched in the two groups were related to infectious diseases, and may tempt to speculate about the underlying biological differences between these two groups (e.g. significance of Toll like receptor 4 pathway in G1, but not G2; and, vice versa, significance of response to interleukin 7 in G2, but not in G1).

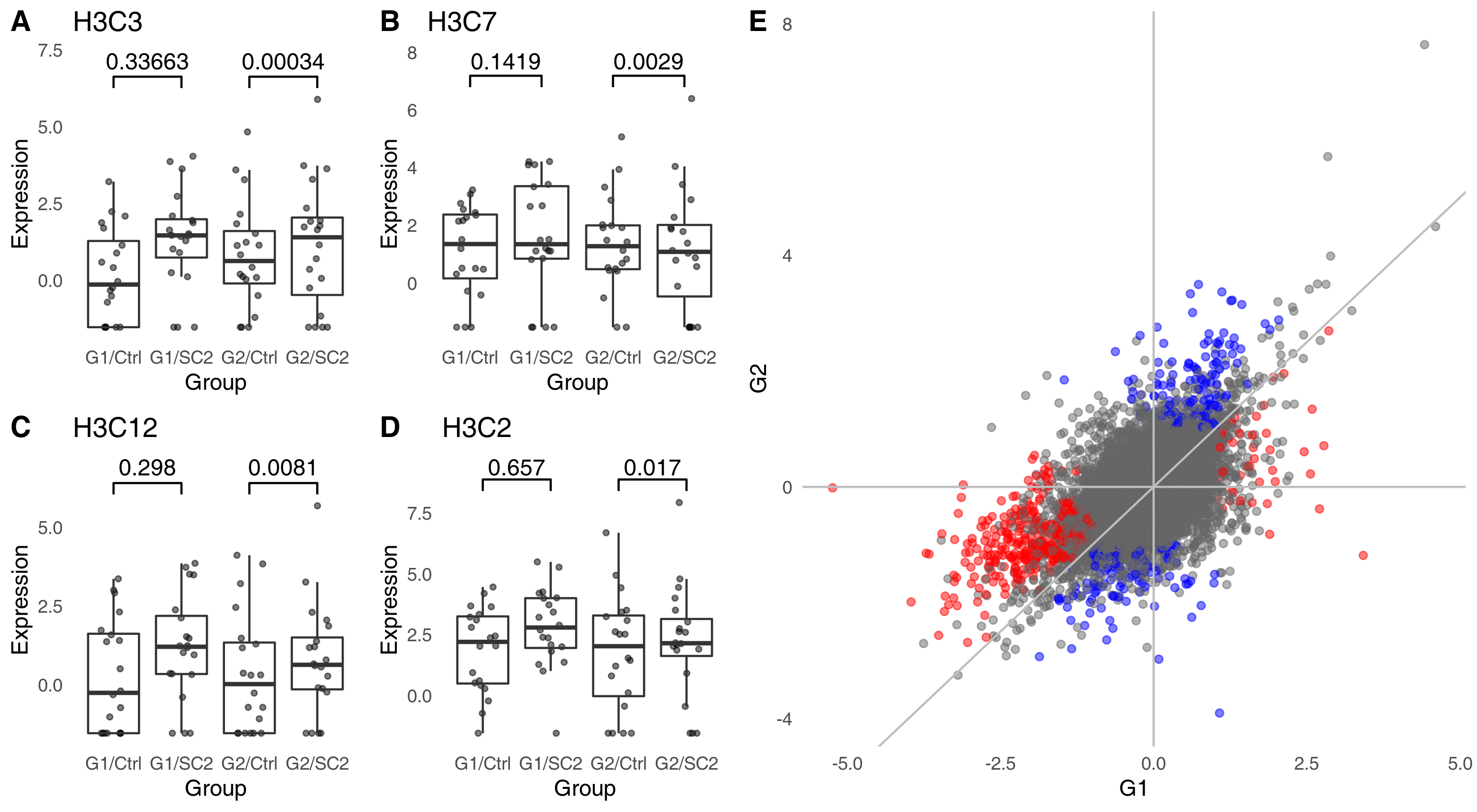
**Fig. 1. Results of differential gene expression analysis and gene set enrichment analysis using an incorrect Venn diagram driven approach.** **A**, Venn diagram showing numbers of differentially expressed genes (DEG) in each of the two groups, G1 and G2; **B**, Venn diagram showing numbers of significantly enriched GO terms in each of the two groups; **C** results of gene set enrichment analysis for genes “specific” to group G1; **D**, results of gene set enrichment analysis for genes “specific” to group G2 (only top 10 terms are shown).



The groups G1 and G2 were randomly sampled from the same data set. In fact, repeated re-sampling always results in some genes being found to be significantly different in one group, but not the other, despite the fact that one does not expect any major differences between sets of individuals randomly drawn from one population. Thus, the conclusions drawn from a Venn diagram-driven gene set enrichment analysis are artefactual. Closer inspection of genes which belong to DEG in one group, but not the other reveals the underlying statistical fallacy (Fig. 2, *A* – *D*), that is, that difference between significant and non-significant is, in itself, not statistically significant ([Gelman and Stern 2006](#ref-gelman2006difference)). This does not necessarily mean that there are no differences at all between these two groups, but that lack of significance in one group and significance in the other group does not correctly identify differences between groups.

To find genes which are differentially regulated in the two groups, the correct statistical approach is to calculate interaction between groups (G1, G2) and disease status (no disease vs. COVID). While it may be argued that a test for interaction has lower power than a test for a simple contrast, no genes show a significant interaction even at FDR < 0.1. In fact, this is not surprising. The log2 fold changes for comparisons withing G1 and G2 are strongly correlated (Fig. 2, *E*). For all significant genes, the Pearson correlation coefficient is 0.72, while for genes exclusively significant in G1 or G2 (genes “specific” to G1 or G2), it is 0.7 and 0.73, respectively. Thus, genes which are significant in one, but not in the other comparison often have have similar log2 fold changes (e.g. Fig. 2, *A*, *C* and *D*).

**Fig. 2. Genes which are significant in one comparison, but not the other do not show a statistically significant interaction.** **A** - **D**, examples of genes which are DEG in one group, but are not significantly different in the other group. “Ctrl,” healthy individuals; “SC2,” Sars-Cov-2 infected patients. Values above the plot indicate FDR (p-values corrected for multiple testing). **E**, correlation between log2 fold changes in G1 and G2. Color indicates genes which are are significant in one, but not significant in the other comparison; red indicates genes significant in G1, blue indicates genes significant in G2. The overall Pearson correlation coefficient between log2 fold changes is 0.55.

 Consequently, it is not possible to calculate gene set enrichment for the interaction using a hypergeometric test, as there are no DEGs for the interaction contrast. Gene set enrichment using a second generation algorithm (CERNO), relying on the ordering of genes according to their raw p-values from the interaction contrast rather than selecting a set of DEGs ([Zyla et al. 2019](#ref-zyla2019gene)), does not show any significant enrichment.

### Artifacts arise because of false negatives

It is worth noting that in the gene set enrichment analysis of the genes “specific” for a given comparison – i.e., genes which are significant in that comparison, but not significant in others – we have observed a number of terms associated with immune response. It is a crucial point of this manuscript to note that the spurious enrichments not only show significant p-values, but also that the terms or pathways which appear in them are relevant to the research hypothesis being tested. Below, we will show why these terms (rather than random terms which have no obvious relevance for an infectious disease) appear in the results.

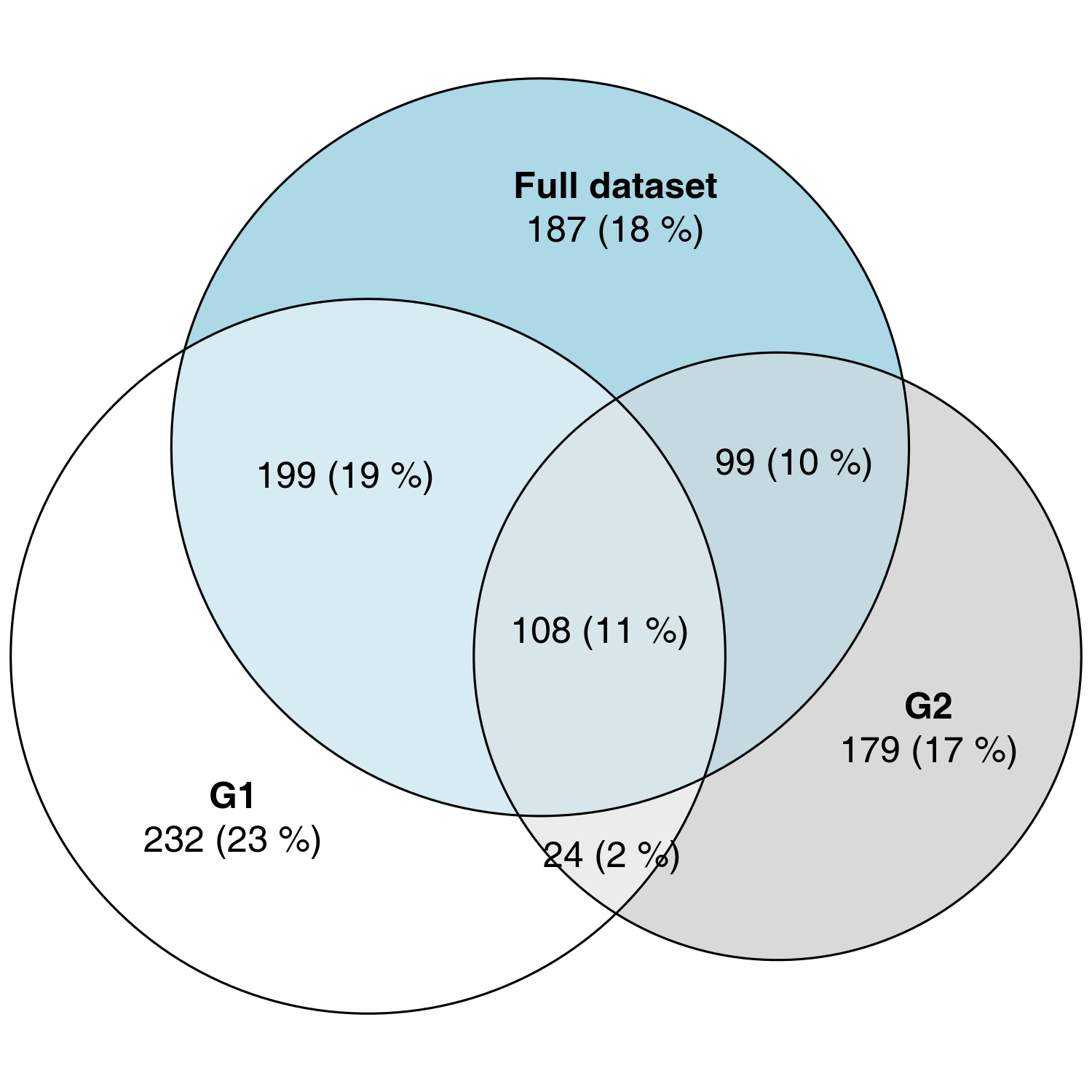
To understand how gene set enrichment actually gives significant results in such randomly generated gene sets, despite absence of genes with significant interaction, it is first necessary to consider the definition of a differentially expressed gene in this context. More often than not, DEGs are defined 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 ([Benjamini and Hochberg 1995](#ref-benjamini1995controlling)) ensures that among genes for which the adjusted padj < 0.05 there are at most 5% false positives.

This way, we can exert control over the false positive rate (FPR, type I errors), keeping it at a relatively low level. However, we do not control the false negative rate (FNR, type II errors). In a powerful statistical test (such as a t-test), the test power in a typical application will rarely achieve more than 80%. For example, even for large effects (Cohen’s d > 0.8) and type I error rate of 0.05, a t-test only achieves 80% power with at least 25 samples per group. For small effects (Cohen’s d > 0.2), the required number of samples is at least 393 per group. Even assuming a test power of 80%, the FNR is 20%. Clearly, false negatives (FNs) occur at much higher rates than false positives (FPs). In case of high throughput data sets, where the FPR is controlled by Bejnamini-Hochberg procedure or a similar technique, the FNR may be even as high as 80% ([White, Ende, and Nichols 2019](#ref-white2019beyond)).

These FNs will occur at a much higher rate within the sets of DEGs defined by the non-overlaping areas of the VDs, that is DEGs considered to be “specific” for one group or other in a naive approach. To illustrate this phenomenon, we have analysed the full data set from which G1 and G2 were drawn (Fig. 3), comparing the 100 healthy controls to 93 COVID-19 patients. Of the 431 genes significant in G1, but not in G2, 199 (46%) are significant in the full data set; of the 278 genes significant in G2, but not in G1, 99 (36%) are significant in the full data set. Given that G1 and G2 were sampled from the total population, and since the FDR was set to 0.05, we can assume that at least between a third and a half of the genes that appeared to be “specific” in the initial analysis were, in fact, false negatives in one of the comparisons.

In other words, a substantial fraction of the “specific” genes are genes that are in reality differentially expressed in both groups alike. Therefore, if one is to perform a gene set enrichment analysis on one of these “specific” groups of genes, then the enriched functions will be related to the pathways and processes up- or down-regulated in both groups due to the common factor (in this example, the COVID-19 disease), but which are not related to differences between the two groups.

**Fig. 3. Area-proportional Venn diagram showing overlaps in DEGs between G1, G2 and the full data set.** The majority of genes which have been labeled as DEGs in only one of the groups G1, G2 are DEGs when all data were analysed.



## Discussion

In an informal survey of articles from the journal “Scientific reports” (chosen for its broad scope, large number of published articles and open access), of the 238 articles from 2021 which used the term “venn diagram” and “differential expression,” at least 73 were using Venn diagrams to compare statistical significance with lack thereof by referring to “unique,” “specific,” “solely regulated” or “exclusive” DEGs. Out of these, at least 31 coupled the VDs with some form of gene set enrichment analysis on the set of supposedly “specific” DEGs.

Drawing conclusions from comparing significance with lack thereof is a common statistical fallacy ([Gelman and Stern 2006](#ref-gelman2006difference)). Just as absence of evidence is not evidence of absence, the failure to reject the null hypothesis does not consitute the same level of evidence as rejecting it. However, when such an incorrect analysis is combined with downstream functional analysis – i.e., gene set enrichments of genes “specific” to one or another comparison – the resulting pathways or gene ontologies are misleadingly relevant, yielding gene sets associated with immune answer for research hypotheses involving an infectious disease, or cancer pathways if the underlying research hypothesis involved cancer treatment. Such results may be hard to resist for an involved researcher, especially if the correct analysis of interactions does not show any significant differences.

That is not to say that VDs are not a useful tool, even in the context of transcriptomics and gene set enrichments, if used correctly. For example, analysis of an intersection of DEGs (i.e., by considering genes from the overlap in a VD) is not an incorrect procedure. Genes in the overlapping part of a VD are significant in both (or all) comparisons, hence no comparison between significance and non-significance is made.

As an alternative to Venn diagrams two approaches can be considered. Firstly, as shown above, a test for interaction can reveal genes for which the impact of treatment significantly differ between the groups. As a mean of visualization, we recommend plotting the log2 fold changes in one comparison versus the log2 fold changes in the other comparison. In addition, a concordance / discordance analysis may provide additional insight in genes which are regulated either in the same direction or in the opposite directions in both groups ([Domaszewska et al. 2017](#ref-domaszewska2017concordant)).

While the use of Venn diagrams in combination with the incorrect gene set enrichment analysis is common, it is not the only way of achieving artifactual and incorrect conclusions by means of comparing significant with non-significant results in transcriptomics and other high-throughput applications. Our literature survey showed that at least two other ways are common. Firstly, the direct comparison of gene set enrichment results: that is, drawing conclusions from the fact that a gene set enrichment result was significant in one comparison only. Second, while Venn diagrams are often used to illustrate the numbers of “specific” DEGs and so present a mean to find examples of this fallacy in scientific literature, researches often test for enrichment these “specific” genes without using the phrase “Venn diagram” or even clearly stating how the lists of “specific” genes were derived. In all these cases, the analysis boils down to comparing results significant in one, but not in another comparison.

The use of Venn diagrams to illustrate specific differences between comparisons should therefore be abandoned in favor of statistically correct approaches. Furthermore, gene set enrichment analysis must never be applied to sets of genes defined as significant in one comparison, but not the other.

## Methods

### Methods availability

This document has been written as an Rmarkdown file, containing all statistical calculations required to replicate the findings and figures. The markdown file, along with additional files required to recreate this manuscript have been uploaded to <https://github.com/bihealth/manuscript_venn_diagrams>.

### Data

The expression data as a count matrix has been downloaded from GEO, accession GSE156063.

### Statistical analyses

Power calculation was done using the R package pwr, version 1.3.0. For differential gene expression, the R package DESeq2, version 1.32.0 has been used. Gene set enrichments were done using either hypergeometric test (where stated) or the CERNO test using the package tmod ([Zyla et al. 2019](#ref-zyla2019gene)), version 0.46.2. GO terms have been sourced from the R package msigdbr, version 7.4.1.

### Literature survey

An informal literature survey was performed using Google Scholar to estimate the frequency of the incorrect use of Venn diagrams. We searched for articles from 2020 mentioning the phrases “differential expression” and “venn diagram” in the Scientific Reports journal. The journal was selected because it represents a wide spectrum of research areas, the articles are available through an open access license and there is a large number of publications per year. For each of the papers identified, we checked whether (i) the authors used the VD to show differentially expressed transcripts significant in one comparison, but not another, (ii) the authors discussed “unique,” “non-overlapping” or “specific” regions of the Venn diagram and (iii) whether this was coupled to gene set enrichment analysis is any form. Articles which (i) focused only on the intersections of the Venn diagrams (genes common to all conditions), or (ii) which used the Venn diagrams for a purpose other than to compare genes significant in one condition, but not significant in another condition or (iii) for which a clear-cut error could not be indicated past any reasonable doubt were not considered incorrect.

## Bibliography

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Blalock Jr, Hubert M. 1965. “Theory Building and the Statistical Concept of Interaction.” *American Sociological Review*, 374–80.

Domaszewska, Teresa, Lisa Scheuermann, Karin Hahnke, Hans Mollenkopf, Anca Dorhoi, Stefan HE Kaufmann, and January Weiner. 2017. “Concordant and Discordant Gene Expression Patterns in Mouse Strains Identify Best-Fit Animal Model for Human Tuberculosis.” *Scientific Reports* 7 (1): 1–13.

Gelman, Andrew, and Hal Stern. 2006. “The Difference Between ‘Significant’ and ‘Not Significant’ Is Not Itself Statistically Significant.” *The American Statistician* 60 (4): 328–31.

Mick, Eran, Jack Kamm, Angela Oliveira Pisco, Kalani Ratnasiri, Jennifer M Babik, Gloria Castañeda, Joseph L DeRisi, et al. 2020. “Upper Airway Gene Expression Reveals Suppressed Immune Responses to SARS-CoV-2 Compared with Other Respiratory Viruses.” *Nature Communications* 11 (1): 1–7.

Nieuwenhuis, Sander, Birte U Forstmann, and Eric-Jan Wagenmakers. 2011. “Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance.” *Nature Neuroscience* 14 (9): 1105.

White, Tonya, Jan van der Ende, and Thomas E Nichols. 2019. “Beyond Bonferroni Revisited: Concerns over Inflated False Positive Research Findings in the Fields of Conservation Genetics, Biology, and Medicine.” *Conservation Genetics* 20 (4): 927–37.

Zyla, Joanna, Michal Marczyk, Teresa Domaszewska, Stefan HE Kaufmann, Joanna Polanska, and January Weiner 3rd. 2019. “Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms.” *Bioinformatics* 35 (24): 5146–54.