## Reviewer 2

*I would have liked to have seen an example which did not involve just Venn-2, but I don’t know whether or not such diagrams would have any relevance from a biological point of view.*

In fact, they do, and frequently they are combined with an incorrect analysis such as the one described in the text. However, we think that the principle remains the same, and therefore we have decided to stick to a two by two design for simplicity and clarity.

*Or, put another way, the authors could expand on this point a bit, and describe more methods of visualising interaction effects, and their relative benefits.*

Initially, we kept this section short so as not to distract from the main – in our opinion – issue, that is, the notion that “significant in one comparison only” is the same as “specific for that comparison”. Also note that many of the visualization techniques which work for a single test – such as the box plots shown on Fig. 2 – are not helpful for visualizing the interaction in the case of thousands of variables. In other words, it is easy to visualize the interaction for a single gene, but what can one use to show the overall effects in 20,000 genes? VDs are meant as such a “grand overview” of the many thousands of statistical tests performed. The classical statistical visualizations of interactions are not a replacement for them.

We have now expanded the relevant paragraps of the Discussion.

*So, the real question I am left wondering about is how prevalent these incorrect Venn-diagram conclusions are? With that information, the reader could be confident that we are not reading about a strawman argument. (The only figure I could find was about a statistical error made by 50% of authors, from Nieuwenhuis et al. 2011)*

Unfortunately, we are not able to put a precise figure on this question. However, both personal experience and the informal literature survey we performed (see Discussion) indicate that it is a widely spread problem. We think that out of the papers which mention “differential expression” and “venn diagram”, at least a third – on average – combines the VDs with incorrect statistical reasoning.

## Reviewer 3

*This is a relevant paper describing an error in statistical reasoning when conducting transcriptomics and gene set enrichment analyses where several related comparisons are required. Indeed, the difference in significances is not itself guaranteed to be significant. However the presentation of the topic in the manuscript is misleading, including the title and the abstract. The real problem is not about the use of Venn diagrams, but the use of suboptimal statistical analyses where the Venn diagrams are just the final step (a limited step at that, given that non-proportional diagrams are commonly used). The more important matter, the lack of statistical interaction testing (such as ANOVA), remains buried deeper in the manuscript and more challenging to understand to a less computational audience.*

In principle, we do agree. The underlying problem is drawing conclusions from juxtaposing a significant result in one group with an insignificant in another. However, the use of Venn diagrams is – in our opinion – illustrative in this context. The erronous analyses are very frequently linked to the use of Venn diagrams and the very notion that a gene significant in one condition, and not significant in another is “specific” for that first condition. It is this particular notion that we are addressing. In other words, while the use of Venn Diagram is a symptom, not a cause, they are useful for diagnostic purposes. Therefore, we would like to sensitivize the readers to the “Venn diagram curse”.

There is also another reason why VDs play a central part in our paper. Once a bioinformatician creates a VD showing, for example, 456 genes up-regulated in the first condition, but not in others, she is frequently confronted with the question: what are these genes? Can’t you run a gene set enrichment on them? VDs practically beg the question what these “unique” genes may be.

We propose the following amendments: (i) we changed the title to “Venn diagrams may indicate erroneous statistical approaches leading to artifacts in transcriptomic analyses”, (ii) we updated the abstract to include the mention of interactions. We have also expanded the section on explaining the underlying statistical problem and (iii) we have expanded the Discussion to better explain our focus on VDs.

* The figures could be improved by adding a schematic of the case-control analysis design that leads to this challenge.

Agreed, we have now added a simple scheme of the example used in Table 1.

* Singling out one journal such as Science Immunology in the significance statement is likely a stretch. Especially given the comment below.

We agree; that was completely unnecessary and we have removed that statement. However, we have now included Sci Immunol in our literature survey (see below).

* The literature survey of how widespread the problem is is quite limited, since the only journal they consider seems to publish many studies with fairly wide variation in quality. More journals, especially the higher-impact ones, should be included in a comprehensive survey.

We have chosen Scientific Reports because of the large number of articles published there. We have now included “Nature Communications” (however, we only checked first 30 articles out of 127 found; 9 were incorrect) and “Science Immunology”, for the latter choosing the publications between 2015 and 2020 (notably, out of the 14 studies which fullfilled our search criteria, 6 were incorrect). We think that this is not a formal literature survey, and this is why we have included it in the discussion rather then elsewhere in the manuscript.

Moreover, while the quality of Sci Rep articles may vary, we think that they are nonetheless being treated as bona fide scientific papers, with many high-profile papers citing and building upon them, and their collective impact may be considerable. To wit, the papers in Sci Rep which were erronous have collectively gathered more than 450 citations in less then two years (median 5 citations per article, with a maximum of 29 citations).

* The study itself were more convincing if the extent of the problem would be covered first (how many studies fall into this trap), followed by the case study and the potential solutions.

We respectfully disagree. We feel that the case study demonstration is the most convincing argument. Also, we would not like to overstate the importance or quality of our informal survey.

* The potential solutions should distinguish statistical and visualization approaches. How would these extend to cases where up-regulated and down-regulated genes are considered?

Do you mean, where up-regulated and down-regulated genes are shown on different Venn diagrams? Given the nature of the interaction (which is not “up” or “down” in itself, in the sense that both a positive and negative interaction coefficient may correspond to a case where in both comparisons genes are down- or, respectively, up-regulated), such a division is of limited (real) applicability. Please also see response on visualizations to reviewer 2.

We have now expanded the relevant fragments of the Discussion.

* GSEA does not use hypergeometric tests. https://www.pnas.org/content/102/43/15545

The authors of the GSEA did use the rather unfortunate acronym for their algorithm, although gene set enrichment analysis can be performed with a number of different algorithms, GSEA being one of them. Hypergeometric test is definitely a gene set enrichment analysis method, see for example https://link.springer.com/chapter/10.1007/978-0-387-77240-0\_14. In our text, GSEA was used as an abbreviation for gene set enrichment analysis not referring to the particular algorithm called GSEA.

However, to avoid confusion, we have removed the GSEA abbreviation from the manuscript.