Comment on Gasparetto et al. by Lyons PA, McKinney EF and Smith KGC. 6th June 2020

We were delighted to see the attached preprint from the Zilbauer group, incorporating data which we have been involved in helping analyse as academic collaborators. For transparency, we are also co-Founders of PredictImmune, which has an interest in the transcriptomic analysis of IBD. Though we did not see the paper before submission, we are in broad agreement with most of its findings, but thought it worth emphasising the problems that arise when adjusting 'omic datasets to account for systematic batch effects, particularly when study groups and their clinical covariates are not evenly distributed between such batches (Goh WWB et al. Trends Biotechnology, 2017;35:498-507).

When we analysed the data presented in Figure 1 of Gasparetto et al., we found that there was significant batch structure (Fig. 1A). While identifying and accounting for systematic variation between groups of samples due to technical factors is unquestionably important, doing so is complex when technical 'batches' overlap with differences in biological covariates (as is the case here, Fig1B). While the observed effect could be regressed out using ComBat (as described in the methods) the imbalance in event rate in each batch (Fig. 1B) risk both inappropriate deflation of 'true' biological associations and the introduction of distinct artefact (Nygaard V et al. Biostatistics 2016;17:29-39. Buhule OD et al. Frontiers in Genetics, 2014;5:1–11). The risk of the former is arguably greater, and we believe that it is not possible to use the dataset presented to exclude an association between transcriptional signatures and clinical outcome as is claimed: it remains possible that the exhaustion signature seen in adults may also exist in this dataset but has been inadvertently removed during the batch correction process. The optimal way to overcome such challenges is through careful study design (Goh WWB et al. Trends Biotechnology 2017;35(6):498-507), although this is more challenging when undertaking prospective studies with analysis occurring before all outcome measures are known. Within-batch analysis may also avoid the need for correction, although sample sizes typically then become prohibitively small, as is the case here. These observations underline the final point made by the authors: that findings such as these need to be replicated in independent datasets before conclusions drawn from them can be trusted. We know that the authors are aware of these issues, and plan to address them in a subsequent peer-reviewed publication.

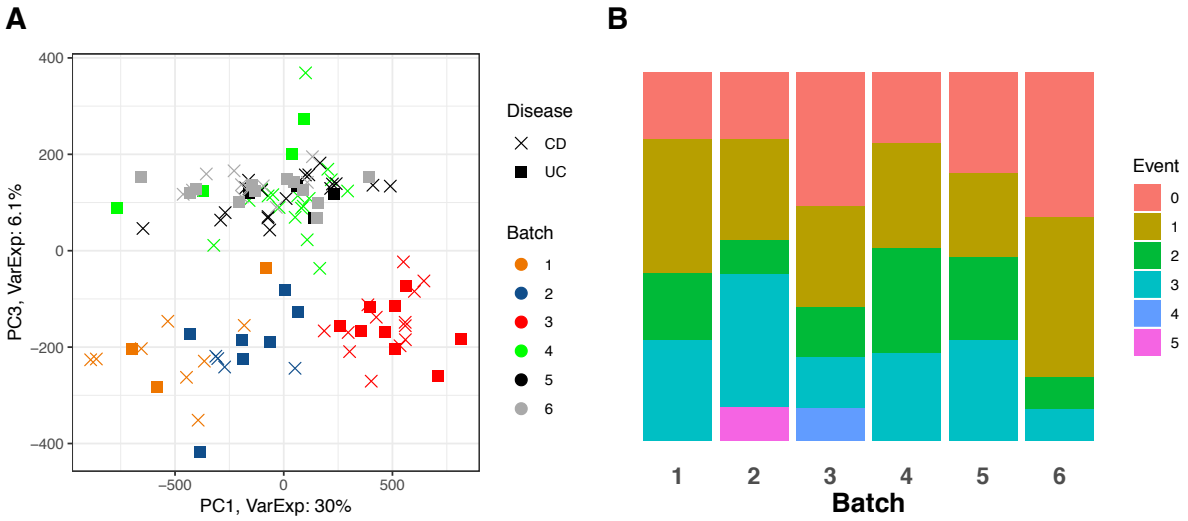

Figure 1. Principal component analysis identifies significant variation attributable to batch in the CD8 T cell microarray data (A); the rates of treatment escalation are unbalanced across the sample batches (B).