*Table 5: Results for hypotheses of aim 2b.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hypothesis Number** | **Result from reference study** | **Hypothesis** | **Specification of  DA-test or method** | **Outcome  threshold** | **Simulation tool:  metaSPARSim** | **Simulation tool: sparseDOSSA2** |
| **H1** | "We found that in both the filtered and unfiltered analyses the percentage of significant ASVs identified by each DA method varied widely across datasets, with means ranging from 3.8–32.5% and 0.8–40.5%, respectively." | For FILTERED and UNFILTERED data, the percentage of significant features identified by each DA method varies widely across datasets.     Dataset: 10 Simulations Project: 38 templates DA\_method: 14 DA tests | UNFILTERED: % significant ~ dataset \* project + DA\_method | p-value  DA\_method  < 0.05 | p-value  DA\_method = 0 | p-value  DA\_method = 0 |
| **H1.1** | UNFILTERED: % significant ~ dataset \* project + DA\_method \* project | p-vlaue  project:DA\_method  < 0.05 | p-value  project:DA\_method = 0 | p-value  project:DA\_method = 0 |
| **H1.2** | FILTERED: % significant ~ dataset \* project + DA\_method | p-value  DA\_method  < 0.05 | p-value  DA\_method = 0 | p-value  DA\_method = 0 |
| **H1.3** | FILTERED % significant ~ dataset \* project +DA\_method \* project | p-vlaue  project:DA\_method  < 0.05 | p-value  project:DA\_method = 0 | p-value  project:DA\_method = 0 |
| **H2** | "Many tools behaved differently between datasets. Some tools identified the most features in one dataset while identifying only an intermediate number in other datasets." | For FILTERED and UNFILTEREDdata, rankings of the DA methods with respect to the proportion of identified features depend on the data template. | UNFILTERED: value ~ DA\_method \* project | p-value  DA\_method:project  < 0.05 | p-value  DA\_method:project = 0 | p-value  DA\_method:project = 0 |
| **H2.1** | FILTERED: value ~ DA\_method \* project | p-value  DA\_method:project  < 0.05 | p-value  DA\_method:project = 0 | p-value  DA\_method:project = 0 |
| **H5** | "[For UNFILTERED data] In a few datasets, such as the Human-ASD and Human-OB (2) datasets, edgeR found a higher proportion of significant ASVs than any other tool." | For UNFILTERED data, there are datasets, where edgeR identifies the largest proportion of significant features. | EdgeR | CI.LB > 0 | CI.LB= 7e-5 | CI.LB= 0,1275 |
| **H8** | "[For UNFILTERED data ] Such extreme findings [as in H7] were also seen in the Wilcoxon (CLR) output, where more than 90% of ASVs were called as significant in eight separate datasets." | For UNFILTERED data, there are datasets, where Wilcoxon CLR identifies more than 90% of features as significant | Wilcoxon (CLR) | CI.LB > 0 | CI.LB= 0,0102 | CI.LB= 0,0458 |
| **H14** | "Over all 38 exp. datasets] edgeR, LEfSe, limma voom, and Wilcoxon tended to output the highest numbers of significant ASVs." | For FILTERED data, there is no method other than EdgeR, LEfSe, limma voom TMMwsp, limma voom TMM, or Wilcoxon CLR that identifies the largest number of significant features in total, i.e. when considering ranks of DA methods over all 38 templates. | EdgeR, LEfSe, limma voom (TMM), limma voom (TMMwsp) or Wilcoxon (CLR) | Estimated proportion > 0.95 | Estimated proportion= 1 | Estimated proportion= 1 |
| **H3** | "This [H2] was especially evident in unfiltered data." | Rankings of the DA methods with respect to the proportion of identified features depend stronger on the data template in unfiltered data than in filtered datasets. | value ~ DA\_method + DA\_method:project | Comparing interaction Mean-Squares DA\_method:project: unfiltered - filtered > 0 | Interaction MSQ DA\_method:project: unfiltered - filtered 12,06 | Interaction MSQ DA\_method:project: unfiltered - filtered -13,612 |
| **H4** | "In the unfiltered datasets, we found that limma voom, Wilcoxon, LEfSe, and edgeR tended to find the largest number of significant ASVs compared with other methods." | In unfiltered data, either limma voom TMMwsp, limma voom TMM, Wilcoxon, or LEfSe identify the largest proportion of significant features. | limma voom (TMM) or limma voom (TMMwsp) or Wilcoxon (CLR) or LEfSe | CI.LB > 0.95 | CI.LB= 0,6187 | CI.LB= 0,9191 |
| **H10** | "[For UNFILTERED data ] We found that two of the three compositionally aware methods we tested identified fewer significant ASVs than the other tools tested. Specifically, ALDEx2 and ANCOM-II identified the fewest significant ASVs. We found the conservative behavior of these tools to be consistent across all 38 datasets we tested." | In UNFITLERED data, either ALDEx2 or ANCOM-II identify the fewest significant features. | ALDEx2 or  ANCOM-II | CI.LB > 0.95 | CI.LB= 0,7851 | CI.LB= 0,7503 |
| **H11** | "The results [from H10] based on the filtered tables showed a similarly conservative behavior." | In FILTERED data, ALDEx2 or ANCOM-II do not identify significantly more features than the most conservative tests. | ALDEx2 or  ANCOM-II | CI.LB > 0.95 | CI.LB= 0,7503 | CI.LB= 0,4461 |
| **H6** | "[For UNFILTERED data ] We found that limma voom (TMMwsp) found the majority of ASVs to be significant (73.5%) in the Human-HIV (3) dataset." | For UNFILTERED data, Limma voom TMMwsp identifies the largest proportion of features in the Human-HIV (3) dataset. | limma voom (TMMwsp) | Estimated proportion > 0.95 | Estimated proportion= 0 | Not testable |
| **H7** | "[For UNFILTERED data ] We found that both limma voom methods identified over 99% of ASVs to be significant in several cases such as the Built-Office and Freshwater-Arctic datasets." | For UNFILTERED data, there are datasets, where both limma voom methods identify more than 99% of features as significant | limma voom (TMM) limma voom (TMMwsp) | CI.LB > 0 | CI.LB= 0 | CI.LB= 0 |
| **H9** | "[For UNFILTERED data ] We found similar [as in H7 and H8], although not as extreme, trends with LEfSe where in some datasets, such as the Human-T1D (1) dataset, the tool found a much higher percentage of significant hits compared with all other tools." | For UNFILTERED data, there are datasets, where LEfSe identifies more taxa as significant compared with all other tools. | LEfSe | CI.LB > 0 | CI.LB= 0 | CI.LB=0 |
| **H12** | "For filtered data, there was a smaller range in the number of significant features identified by each tool. All tools except for ALDEx2 found a lower number of total significant features when compared with the unfiltered dataset." | No tool (except ALDEx2) identifies a smaller number of features for unfiltered data (than for filtered). | ALDEx2 | CI.LB > 0.95 | CI.LB= 0,0118 | CI.LB= 0,0136 |
| **H13** | As with the unfiltered data, ANCOM-II was the most stringent method [over all 38 exp. datasets]. | For FILTERED and UNFILTERED data, ANCOM-II identifies the least significant features in total, i.e. when summing ranks of DA methods over all 38 templates. | UNFILTERED: ANCOM-II | Estimated proportion > 0.95 | Esitmated proportion= 0 | Estimated proportion= 0 |
| **H13.1** | FILTERED: ANCOM-II | Estimated proportion > 0.95 | Estimated proportion= 0 | Estimated proportion= 0 |