*Table 4: Results for hypotheses of aim 2a.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Hypothesis Number** | **Result from reference study** | **Hypothesis** | **Specification of DA-test or method:** | **metaSPARSim** | **sparseDOSSA2** |
| **H1** | "Based on the unfiltered data, limma voom methods identified similar sets of significant ASVs that were different from those of most other tools." | For UNFILTERED data the proportion of features jointly found as significant by limma voom TMM and limma voom TMMwsp but by less than 50% of the other methods, is larger than the overlap with more than 50% of the other methods | limma voom (TMM) & limma voom (TMMwsp) | CI.LB =  0,9888 | CI.LB =  0,9864 |
| **H4** | "[For UNFILTERED data] Both MaAsLin2 approaches had similar consistency profiles." | For UNFILTERED data, MaAsLin2 and MaAsLin2-rare have a more similar profile (larger test statistic D) than a randomly selected pair of methods. | MaAsLin2 & MaAsLin2 (rare) | CI.LB =  0,6615 | CI.LB =  0,9645 |
| **H7** | "[For UNFILTERED data] Corncob, metagenomeSeq, and DESeq2 identified ASVs at more intermediate consistency profiles." | For UNFILTERED data, for corncob, metagenomeSeq, and DESeq2, there are always multiple other methods (i.e. at least 2 out of 10 other DA methods) that have a more extreme consistency profile. | corncob | CI.LB =  0,8083 | CI.LB =  0,6212 |
| **H7.1** | DESeq2 | CI.LB =  0,8140 | CI.LB =  0,7472 |
| **H7.2** | metagenomeSeq | CI.LB =  0,4782 | CI.LB =  0,9191 |
| **H8** | "The overlap in significant ASVs based on the prevalence-filtered data was similar overall to the unfiltered data results." | The shape of the overlap profiles for all methods except both limma voom approaches is mainly determined by the exp. dataset and the DA method but only little of the fact whether data has been filtered. Filt vs unfilt (dataset[i], DA-test[j]) compared to ... i: 1..380, j:1..14 | … other unfiltered datasets | CI.LB =  0,8299 | CI.LB =  0,7313 |
| **H8.1** | … other filtered datasets | CI.LB =  0,8231 | CI.LB =  0,7195 |
| **H8.2** | … other DA tests in unfiltered data | CI.LB =  0,7929 | CI.LB =  0,8079 |
| **H8.3** | … other dA tests in filtered data | CI.LB =  0,8206 | CI.LB =  0,7956 |
| **H9** | "For prevalence-filtered data, the limma voom approaches identified a much higher proportion of ASVs that were also identified by most other tools, compared with the unfiltered data." | For filtered data, for both limma voom approaches the proportion of identified features that are also identified by the majority of other tests is larger than for un-filtered data. | limma voom (TMM) | CI.LB =  0,6421 | CI.LB =  0,6376 |
| **H9.1** | limma voom (TMMwsp) | CI.LB =  0,6239 | CI.LB =  0,6454 |
| **H10** | "[For FILTERED data] The Wilcoxon (CLR) significant ASVs displayed a bimodal distribution and a strong overlap with limma voom methods." | For FILTERED data, the overlap profile of Wilcoxon CLR is bimodal. | Wilcoxon (CLR) | CI.LB =  0,7070 | CI.LB =  0,7707 |
| **H5** | "[For UNFILTERED data] ALDEx2 and ANCOM-II primarily identified features that were also identified by almost all other methods." | For UNFILTERED data, ALDEx2 and ANCOM-II identify more features that were also identified by all except 3 (i.e. 10 out of 13) other methods | ALDEx2 | CI.LB =  0,5300 | CI.LB =  0,5998 |
| **H5.1** | ANCOM-II | CI.LB =  0 | CI.LB =  0,0633 |
| **H3** | "[For UNFILTERED data] The two Wilcoxon test approaches had highly different consistency profiles." | For UNFILTERED data, the Kolmogorov-Smirnov test statistic D when comparing the profile for Wilcoxon CLR and Wilcoxon rare is larger than for other pairs of methodson average. | Wilcoxon (CLR)  & Wilcoxon (rare) | CI.LB =  0,4404 | CI.LB =  0,7205 |
| **H2** | "[For UNFILTERED data]Many of the ASVs die ntified by the limma voom methods were also identified as significant based on the Wilcoxon (CLR) approach." | For UNFILTERED data the overlap of features jointly found as significant by limma voom TMM and limma voom TMMwsp with features found by Wilcoxon CLR is larger than the overlap with all other DA methods | both limma vooms  & Wilcoxon (CLR) | CI.LB =  0,1481 | CI.LB =  0,2149 |
| **H6** | "[For UNFILTERED data] EdgeR and LEfSe, two tools that often identified the most significant ASVs, output the highest percentage of ASVs that were not identified by any other tool." | For UNFILTERED data, EdgeRand LEfSe identify a larger percentage of features that are not identified by any other tool, compared to the same percentage for all other methods. | EdgeR | CI.LB =  0 | CI.LB =  0,1736 |
| **H6.1** | LEfSe | CI.LB =  0 | CI.LB =  0 |
| **H11** | "Overall, the proportion of ASVs consistently identified as significant by more than 12 tools was much higher in the filtered data (mean: 38.6%; SD: 15.8%) compared with the unfiltered data (mean: 17.3%; SD: 22.1%)." | The proportion of features identified by all except one DA method is larger for prevalence-filtered data. |  | CI.LB =  0,5161 | CI.LB =  0,5501 |
| **H12** | "In contrast with the unfiltered results, corncob, metagenomeSeq, and DESeq2 had lower proportions of ASVs at intermediate consistency ranks." | For filtered data, the consistency profiles of corncob**,** metagenomeSeq, and DESeq2 are more similar to the more extreme methods than for unfiltered data. | corncob | CI.LB =  0,0276 | CI.LB =  0,0043 |
| **H12.1** | DESeq2 | CI.LB =  0,0066 | CI.LB =  0,0066 |
| **H12.2** | metagenomeSeq | CI.LB =  0,0245 | CI.LB =  0,0191 |
| **H13** | "[For FILTERED data] ALDEx2 and ANCOM-II produce significant ASVs that largely overlapped with most other tools." | For FILTERED data, ALDEx2 and ANCOM-II identify more features that were also identified by all except 3 (i.e. 10 out of 13) other methods. | ALDEx2 | CI.LB =  0,4412 | CI.LB =  0,4571 |
| **H13.1** | ANCOM-II | CI.LB =  0 | CI.LB =  0,0578 |