**Supplementary Materials**

**Table S1** - List of the fungal strains of the mock community and their ITS Sanger sequences. The following properties of the mock strains are also indicated: the uniqueness of their ITS1 sequence (Unique\_ITS1\_Sanger\_Sequence), whether the exact ITS1 sequence was present or not in the raw MiSeq data (Present\_in\_Raw\_MiSeq\_Data) and the putative cause of absence (Putative\_Cause\_Absence). Absences were attributed either to the presence of ambiguous bases in the Sanger sequence preventing a perfect match with the MiSeq sequence (AB), or to sequencing errors preventing a perfect match between MiSeq and Sanger sequences (SE), or to lack of amplification due to an imperfect match with the ITS2 primer sequence (LA) or to the loss of the strain during DNA pooling (DP). In the case of AB and SE, the percentage of similarity between the Sanger sequence and the closest MiSeq sequence is given into brackets. To assign the strain to the LA category, the sequence of the ITS2 primer was searched for in both the Sanger sequence and the closest NCBI sequence (given into brackets). Strains that were absent but did not fall into the three first categories were classified into the DP category.

(Attached as a .csv file)

**Table S2** - Sequence summary statistics for the raw data and the assembled reads, before quality filtering. Values were averaged over the 3 replicates for each bioinformatic approach.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Assembly*** | **Mean number of reads** | **Mean quality** | **Median length** | **Min/Max length** | **% reads**  **[0,100 nt[** | **% reads**  **[100, 200 nt[** | **% reads**  **[200, 300 nt[** | **% reads**  **≥ 300 nt** |
| RAW\_R1 | 47958 | 35.59 | 250 | 250/250 | 0 | 0 | 100 | 0 |
| RAW\_R2 | 47958 | 34.31 | 250 | 250/250 | 0 | 0 | 100 | 0 |
| FASTQJOIN\_50 | 26209 | 37.77 | 267 | 250/444 | 0 | 0 | 97.85 | 2.15 |
| FASTQJOIN\_100 | 26132 | 37.77 | 267 | 250/400 | 0 | 0 | 97.85 | 2.15 |
| FASTQJOIN\_150 | 23972 | 37.78 | 266 | 250/350 | 0 | 0 | 97.85 | 2.15 |
| PEAR\_50 | 47439 | 39.01 | 267 | 50/447 | 0.04 | 0.1 | 97.78 | 2.08 |
| PEAR\_100 | 47283 | 39.02 | 267 | 100/400 | 0 | 0.1 | 97.82 | 2.09 |
| PEAR\_150 | 43835 | 39.13 | 265 | 150/350 | 0 | 0.06 | 97.85 | 2.09 |

**Table S3** - Sequence summary statistics for the raw data and the assembled reads, after quality filtering. The first two columns correspond to sequence with a Phred score higher than 30 on 75% of length (Q30). The last two columns correspond to sequences with less than one expected error according to error probabilities obtained from quality scores (EE). For the CUTADAPT\_MERGED assembly method, reads with less than two expected errors were retained. Percentages of reads passing quality filter were computed over the total number of raw reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Assembly*** | **Mean number**  **of reads (Q30)** | **% of reads (Q30)** | **Mean number**  **of reads (EE)** | **% of reads (EE)** |
| QUALITY\_R1 | 38356 | 80% | 37441 | 78.1% |
| FASTQJOIN\_50 | 24499 | 93.5% | 25857 | 98.7% |
| FASTQJOIN\_100 | 24425 | 93.5% | 25821 | 98.8% |
| FASTQJOIN\_150 | 22277 | 92.9% | 23832 | 99.4% |
| PEAR\_50 | 45283 | 95.5% | 44707 | 94.2% |
| PEAR\_100 | 45162 | 95.5% | 44657 | 94.4% |
| PEAR\_150 | 42080 | 96% | 41970 | 95.7% |
| CUTADAPT\_MERGED | NA | NA | 38058 | 79.4% |

**Table S4** – Number and length of sequences after extraction of the ITS1 region with ITSx. Percentages of reads extracted were computed over the total number of quality reads (Q30) (Table S3).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Assembly*** | **Number of reads** | **% of reads** | **% reads**  **[0, 100 nt[** | **% reads**  **[100, 200 nt[** | **% reads**  **[200, 300 nt[** | **% reads**  **≥ 300 nt** |
| QUALITY\_R1 | 22622 | 59% | 0.84 | 99.12 | 0.04 | 0 |
| FASTQJOIN\_50 | 24155 | 98.60% | 0 | 82.91 | 17.09 | 0 |
| FASTQJOIN\_150 | 22144 | 99.40% | 0 | 90.36 | 9.64 | 0 |
| FASTQJOIN\_100 | 24121 | 98.80% | 0 | 82.95 | 17.05 | 0 |
| PEAR\_50 | 44109 | 97.40% | 0.49 | 86.47 | 13.04 | 0 |
| PEAR\_100 | 44065 | 97.60% | 0.49 | 86.5 | 13.01 | 0 |
| PEAR\_150 | 41443 | 98.50% | 0.52 | 91.97 | 7.51 | 0 |

**Table S5** - Number of OTUs (or ASVs) and sequences in the final OTU (ASV) table according to the method of sequence variation analysis and the filtering option, when QUALITY\_R1 reads without ITS1 extraction and without chimeric sequences were used. Values were averaged over the 3 replicates for each bioinformatic approach. LULU was applied with default settings.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Variation*** | ***Filtering*** | **Mean number of OTUs (or ASVs)** | **Mean number of sequences** |
| DADA2 | All | 157 | 34991 |
| DADA2 | 1 | 157 | 34991 |
| DADA2 | 10 | 148 | 34950 |
| DADA2 | RA | 152 | 34978 |
| DADA2 | LULU | 105 | 34991 |
| USEARCH | All | 878 | 38245 |
| USEARCH | 1 | 660 | 38027 |
| USEARCH | 10 | 329 | 37408 |
| USEARCH | RA | 437 | 37688 |
| USEARCH | LULU | 298 | 38245 |
| VSEARCH | All | 577 | 37760 |
| VSEARCH | 1 | 428 | 37611 |
| VSEARCH | 10 | 257 | 37293 |
| VSEARCH | RA | 311 | 37444 |
| VSEARCH | LULU | 253 | 37760 |

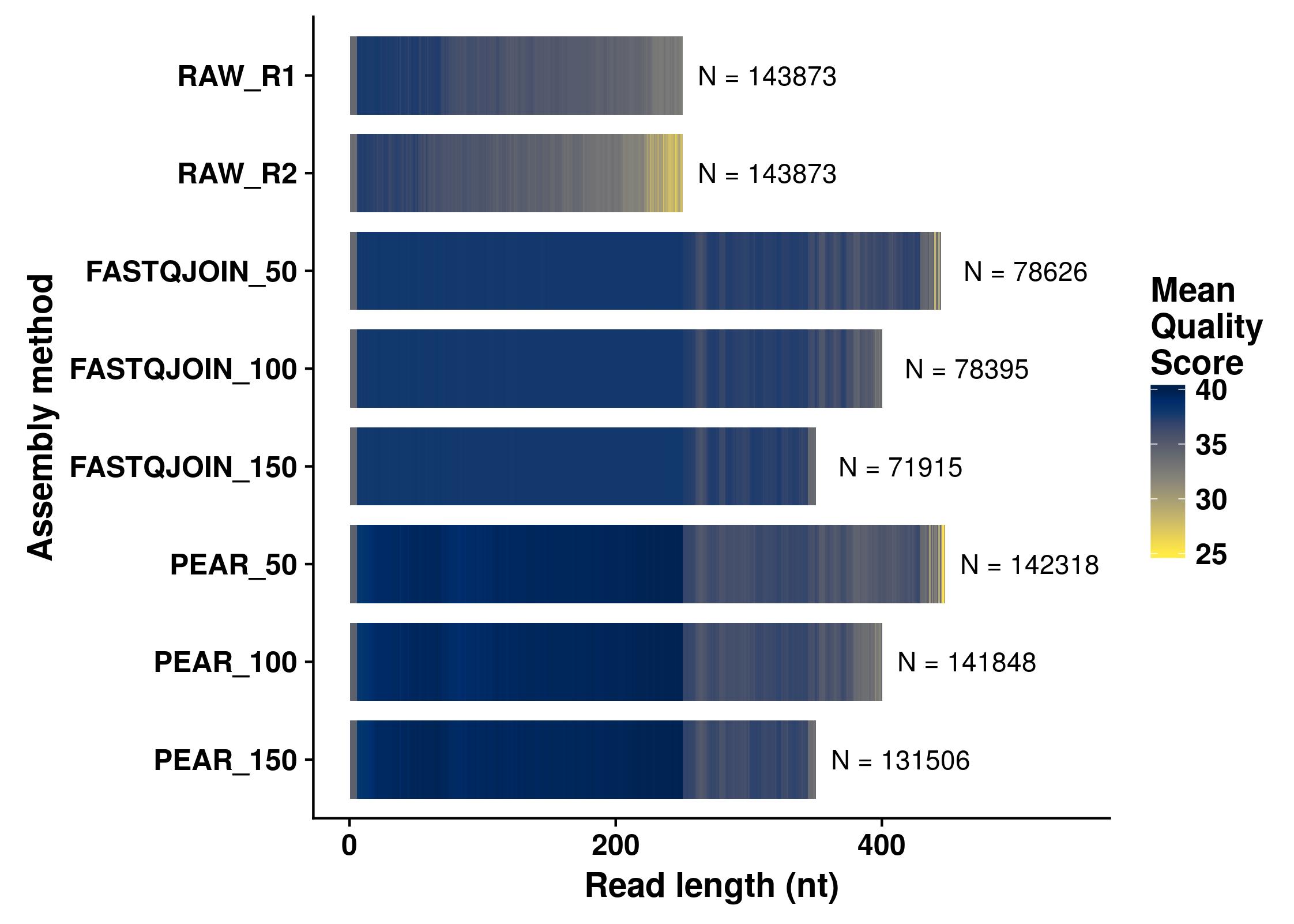
**Table S6** - Sensitivity, precision and similarity values for all bioinformatic approaches.

(Attached as a .csv file)

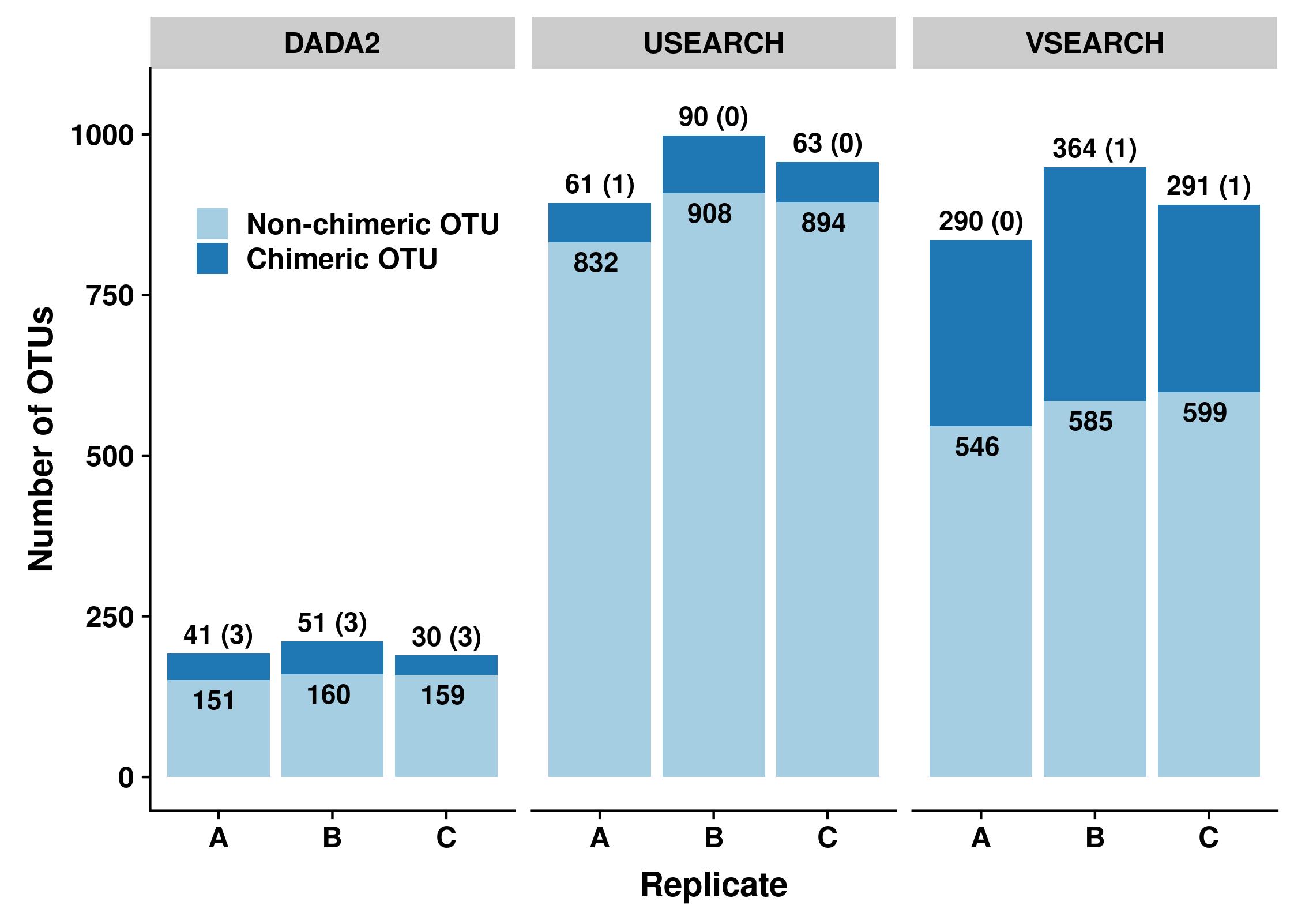
**Table S7** - Effect of primer removal on the sensitivity, precision and similarity values for the Si1 and Si5 approaches.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Approach** | **Primers** | **Richness** | **Sensitivity** | **Precision** | **Similarity** |
| Si1 | Retained | 197 | 0.774 | 0.743 | 0.396 |
| Si1 | Removed | 200 | 0.774 | 0.735 | 0.396 |
| Si5 | Retained | 157 | 0.758 | 0.915 | 0.393 |
| Si5 | Removed | 157 | 0.758 | 0.913 | 0.393 |

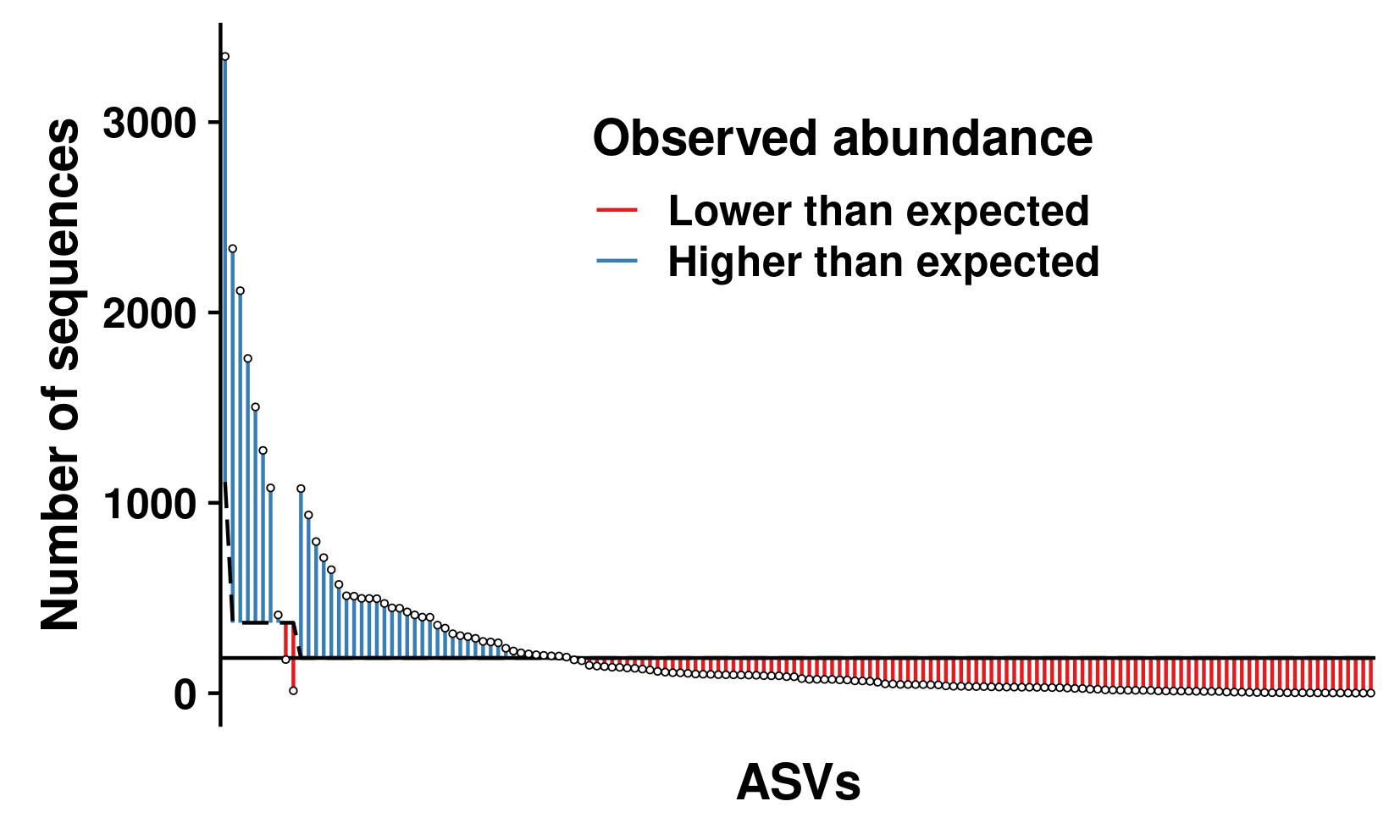
**Figure S1** - Mean quality score along raw reads and assembled reads before quality filtering. The total number of sequences (N) is indicated for each category of reads.



**Figure S2 -** Total number of OTUs (or ASVs) per replicate as a function of the method for analyzing sequence variation, when using QUALITY\_R1 reads without ITS1 extraction. The number of True Positives OTUs present in the Chimeric OTUs is indicated in brackets.

****

**Figure S3 -** Variability of sequence counts among the true positive ASVs generated by the Si5 approach. Dots correspond to the observed number of sequences per ASV (averaged over the 3 replicates). The horizontal black plain line indicates the expected number of reads per ASV, under the assumption of a uniform distribution of sequences between fungal strains of the mock community. Under this assumption, the expected number of reads per ASV is defined as the final number of sequences obtained with the bioinformatic approach Si5 divided by the total number of strains in the mock community. The black dashed line indicates the expected number of reads per ASV multiplied by the total number of fungal strains (with identical ITS1 sequences) corresponding to an ASV. Deviations from this expectation are indicated with blue and red bars.



**Figure S4 –** Sensitivity, precision and compositional similarity obtained for all the bioinformatic approaches using USEARCH and VSEARCH, as a function of the extraction of the ITS1 region with ITSx. Values were averaged over the 3 replicates for each bioinformatic approach and compared using paired *t*-tests.

