SUPPLEMENTARY MATERIAL

for the manuscript:

Abundance estimation and differential testing on strain level in metagenomics data

Martina Fischer,* Benjamin Strauch, Bernhard Y. Renard

Research Group Bioinformatics (NG 4), Robert Koch-Institute, Nordufer 20, 13353 Berlin, Germany

1. part: Figures & Tables

2. part: Data set & model descriptions

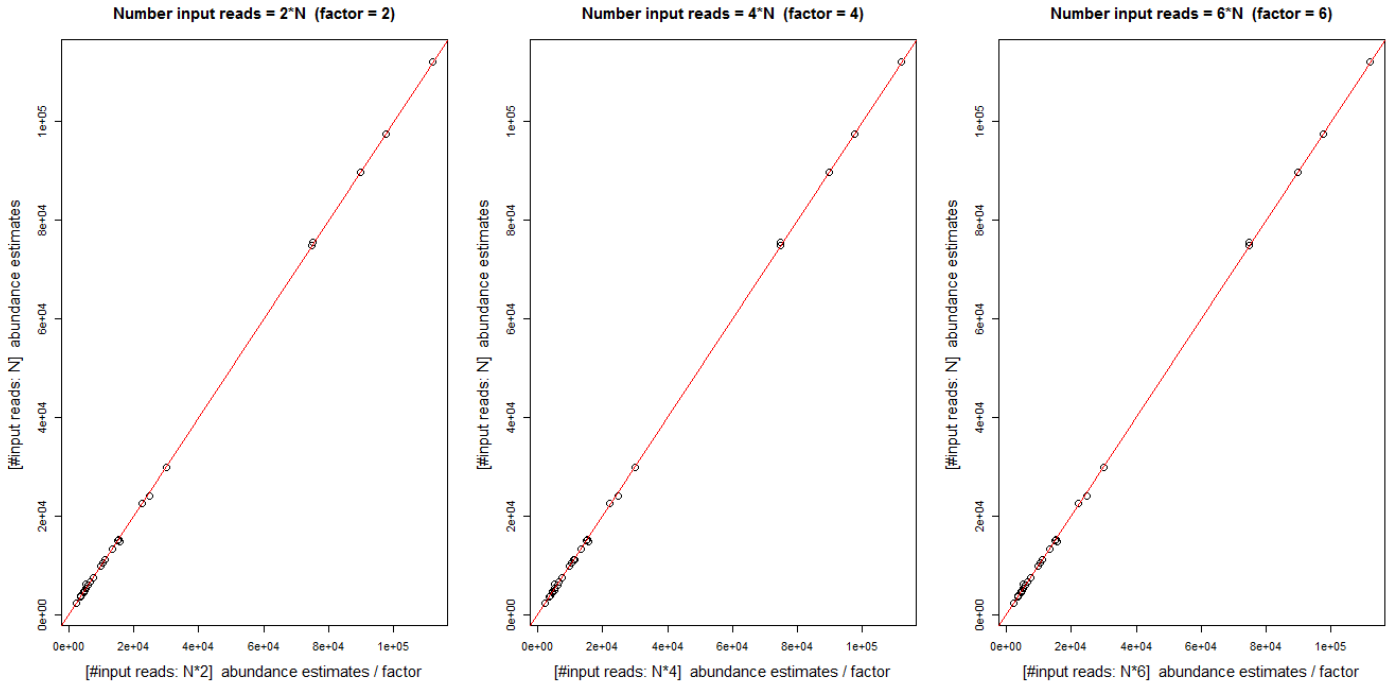
~~~~~~

Figures & Tables

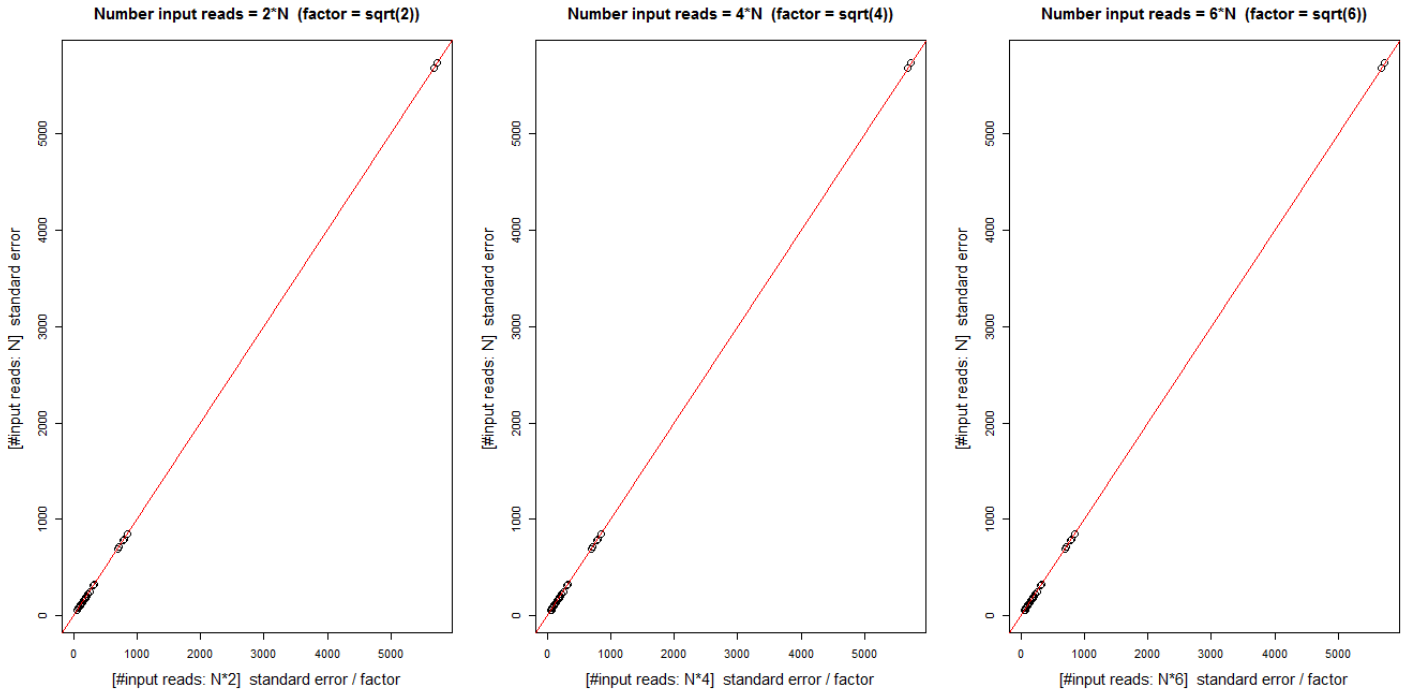
~~~~~~

Supplementary Figure 1:

(a)

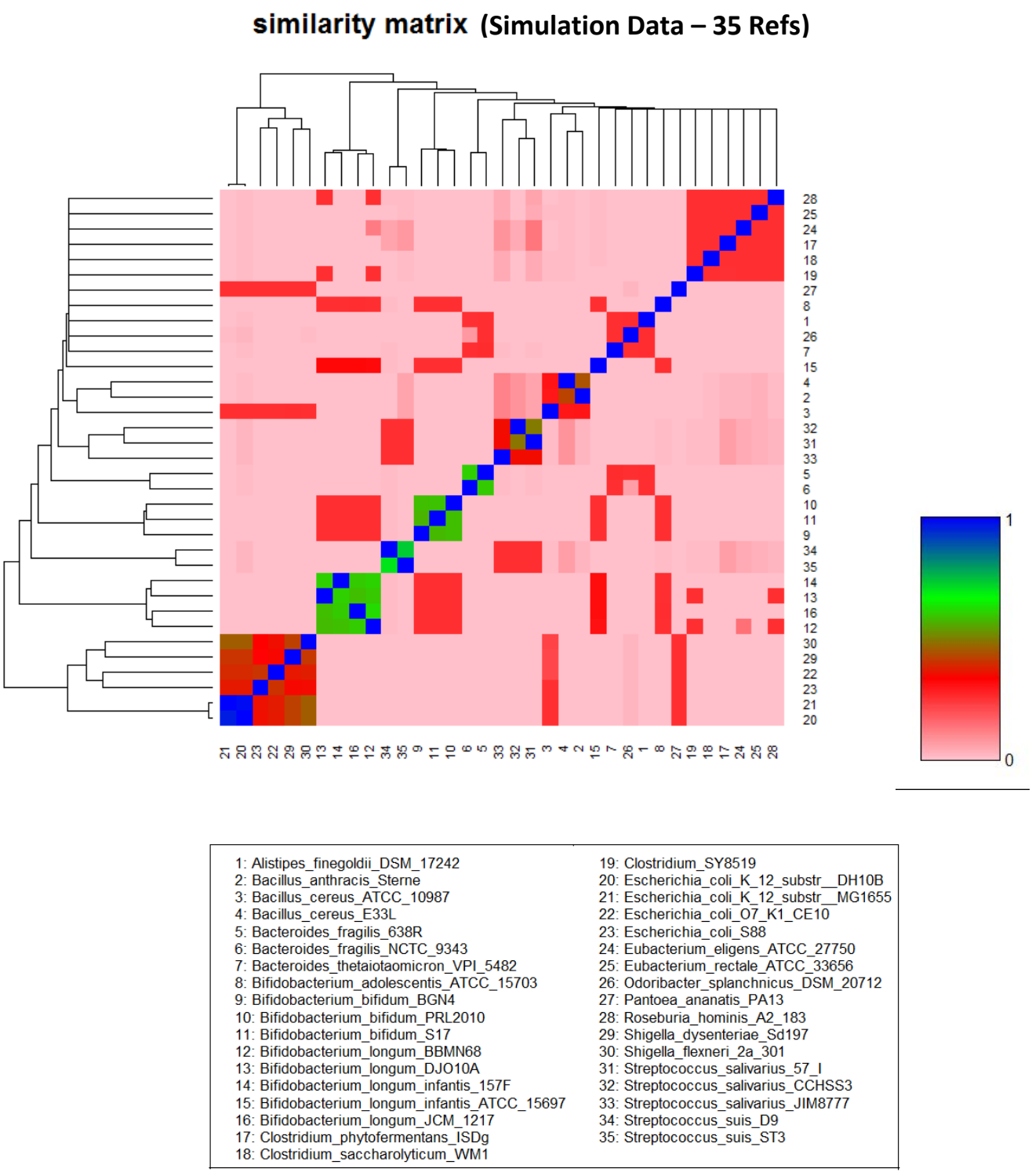


(b)



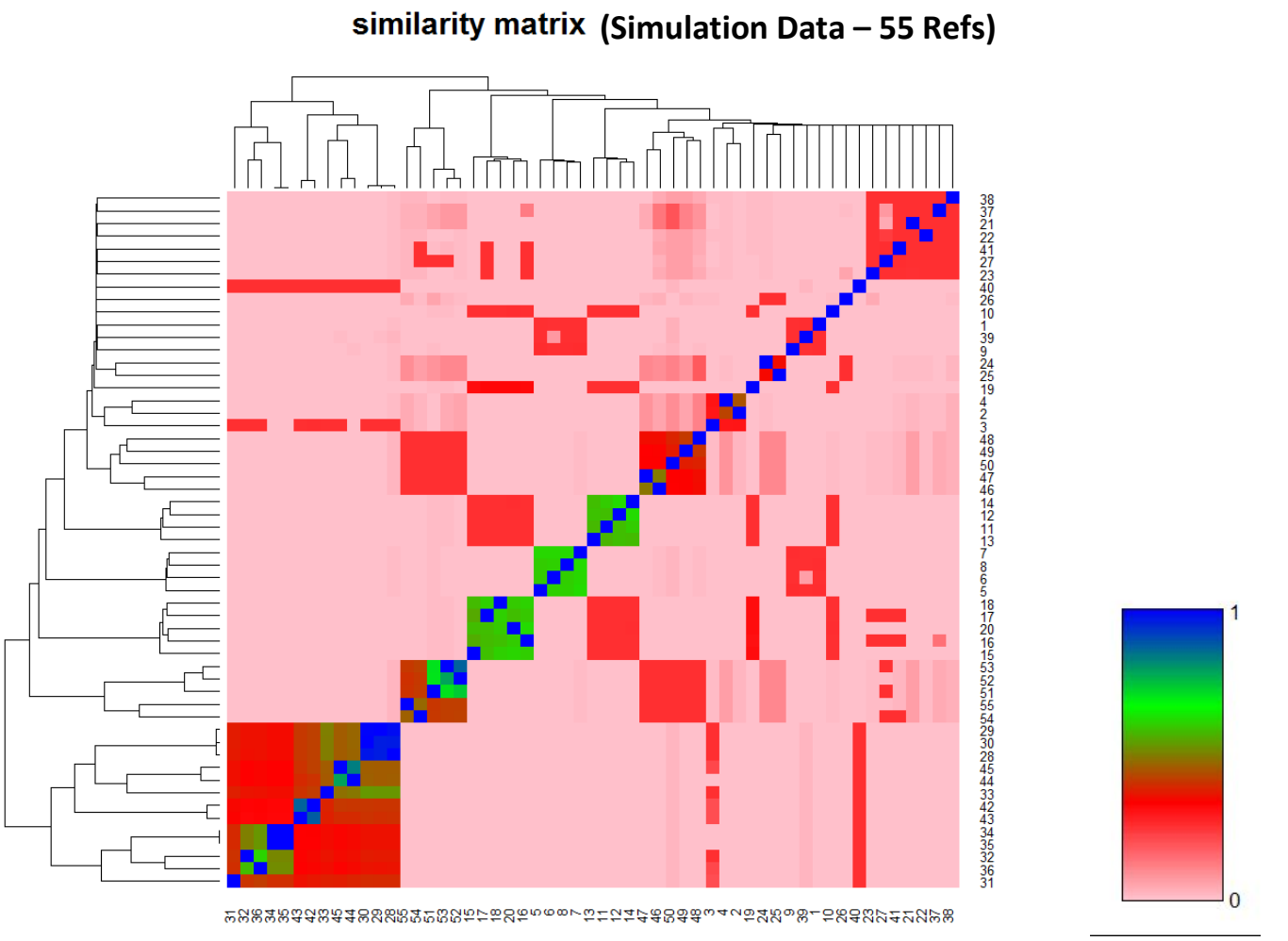
Supplementary Figure 1: **Impact of total number of input reads on (a) abundance estimates and (b) standard errors.** We conducted a study applying different total numbers of input reads (exemplary for the 'original' simulation set 4): increasing the original number of input reads N (N =750.000) by the factor of 2, 4, and 6, corresponding to total amounts of 1.5, 3, and 4.5 million input reads for the set. We conducted comparisons of abundance estimates and standard errors computed on the 'sets with increased read number' against the results obtained by the 'original' set. It can be observed that the abundance estimates scale linear with the number of reads, whereas the standard errors scale quadratic.

Supplementary Figure 2:



Supplementary Figure 2: **Similarity matrix of the simulation data sets comprising 35 reference genomes** (see list of taxa accession numbers in the subsequent section 'Data Set description'). The heatmap visualizes all pairwise reference sequence similarities ranging from 0 to 100% similarity (visualized from pink to dark blue). The diagonal of the matrix refers to the proportion of simulated reads mapping back to their reference of origin. Different clusters of strains exhibiting high reference sequence similarities can be observed. Notably is the first big cluster of diverse *Escherichia coli* strains (bottom left), comprising two sub-strains which share 98% sequence similarity, two more distant *E.coli* strains, and further two *Shigella* strains known to be closely related to *E.coli*. The second big cluster comprises four different strains of *Bifidobacterium longum*, followed by a cluster of *Bifidobacterium bifidum*, which expresses moderate similarities to the former. Further, various smaller clusters of strains are present.

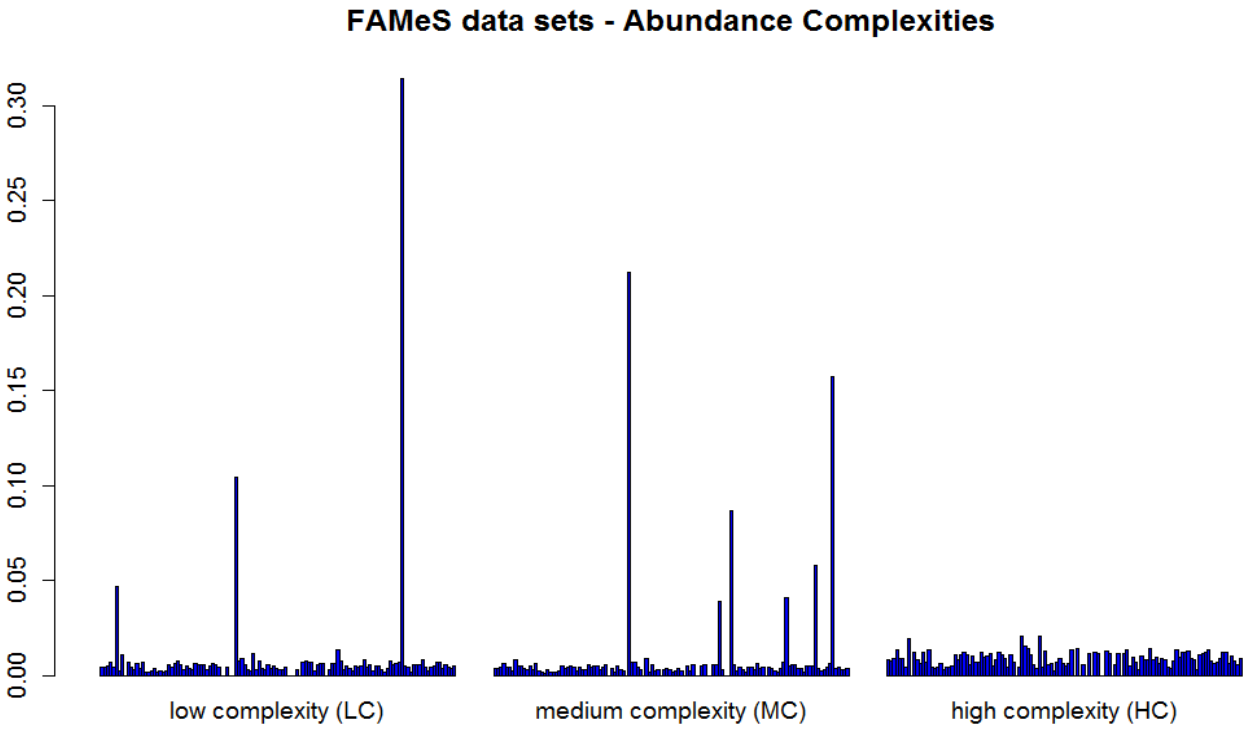
Supplementary Figure 3:



1: Alistipes_finegoldii_DSM_17242	20: Bifidobacterium_longum_JCM_1217	39: Odoribacter_splanchnicus_DSM_20712
2: Bacillus_anthracis_Sterne	21: Clostridium_phytofermentans_ISDg	40: Pantoea_ananatis_PA13
3: Bacillus_cereus_ATCC_10987	22: Clostridium_saccharolyticum_WM1	41: Roseburia_hominis_A2_183
4: Bacillus_cereus_E33L	23: Clostridium_SY8519	42: Shigella_dysenteriae_Sd197
5: Bacteroides_fragilis_638R	24: Clostridium_botulinum_A3_str_Loch_Maree	43: Shigella_dysenteriae_1617
6: Bacteroides_fragilis_NCTC_9343	25: Clostridium_botulinum_B1_str_Okra	44: Shigella_flexneri_5_str_8401
7: Bacteroides_fragilis_strain_BOB25	26: Clostridium_botulinum_B_str_Eklund_17B	45: Shigella_flexneri_2a_301
8: Bacteroides_fragilis_YCH46	27: Clostridium_cf_saccharolyticum_K10	46: Streptococcus_salivarius_57_I
9: Bacteroides_thetaiotaomicron_VPI_5482	28: Escherichia_coli_K_12_substr__DH10B	47: Streptococcus_salivarius_CCHSS3
10: Bifidobacterium_adolescentis_ATCC_15703	29: Escherichia_coli_K_12_substr__MG1655	48: Streptococcus_salivarius_JIM8777
11: Bifidobacterium_bifidum_BGN4	30: Escherichia_coli_str_K_12_substr_MC4100	49: Streptococcus_salivarius_strain_HSISS4
12: Bifidobacterium_bifidum_PRL2010	31: Escherichia_coli_O7_K1_CE10	50: Streptococcus_salivarius_strain_NCTC_8618
13: Bifidobacterium_bifidum_S17	32: Escherichia_coli_S88	51: Streptococcus_suis_D9
14: Bifidobacterium_bifidum_ATCC_29521	33: Escherichia_coli_O104_H4_str_2011C_3493	52: Streptococcus_suis_ST3
15: Bifidobacterium_longum_subsp_longum_44B	34: Escherichia_coli_O127_H6_str._E2348_69_substr._CVDNalr_genomic	53: Streptococcus_suis_05HAS68
16: Bifidobacterium_longum_BBMN68	35: Escherichia_coli_O127_H6_str._E2348_69_substr._UMD753_genomic	54: Streptococcus_suis_JS14
17: Bifidobacterium_longum_DJO10A	36: Escherichia_coli_O83_H1_str_NRG_857C	55: Streptococcus_suis_T15
18: Bifidobacterium_longum_infantis_157F	37: Eubacterium_eligens_ATCC_27750	
19: Bifidobacterium_longum_infantis_ATCC_15697	38: Eubacterium_rectale_ATCC_33656	

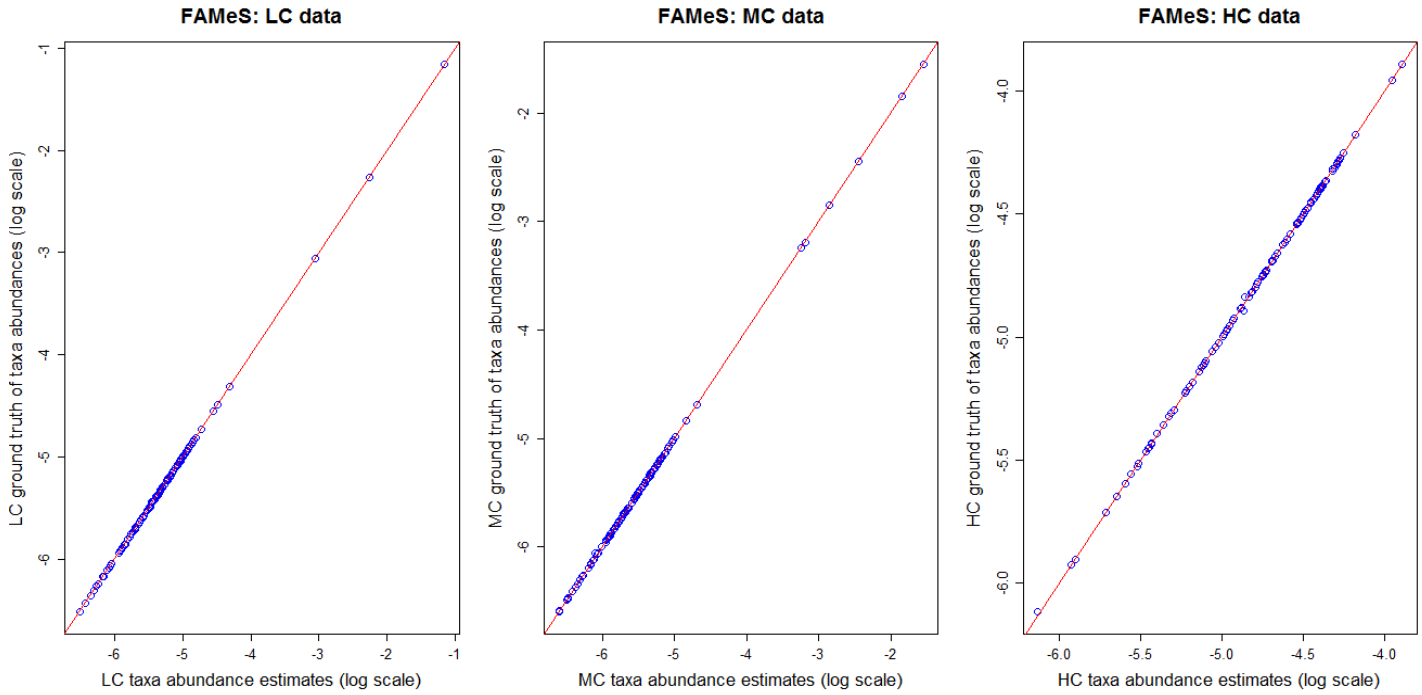
Supplementary Figure 3: **Similarity matrix of the simulation data sets comprising 55 reference genomes, exhibiting large similar strain clusters** (see list of taxa accession numbers in the subsequent section 'Data Set description'). The heatmap visualizes all pairwise reference sequence similarities ranging from 0 to 100% similarity (visualized from pink to dark blue). Additional strain and sub-strain sequences were added to the simulation set of 35 references to challenge the tools: a big cluster of overall 13 taxa of *Escherichia coli* strains containing three different sub-strain clusters with sequence similarities above 95%, mixed with diverse distant *E.coli* strains and closely related *Shigella* strains. Further, cluster of *Bifidobacterium longum*, *Bifidobacterium bifidum, Bacteroides fragilis* as wells as two different *Streptococcus* species cluster were largely extended to test the resolution performance of the tools within large and highly similar strain clusters.

Supplementary Figure 4:



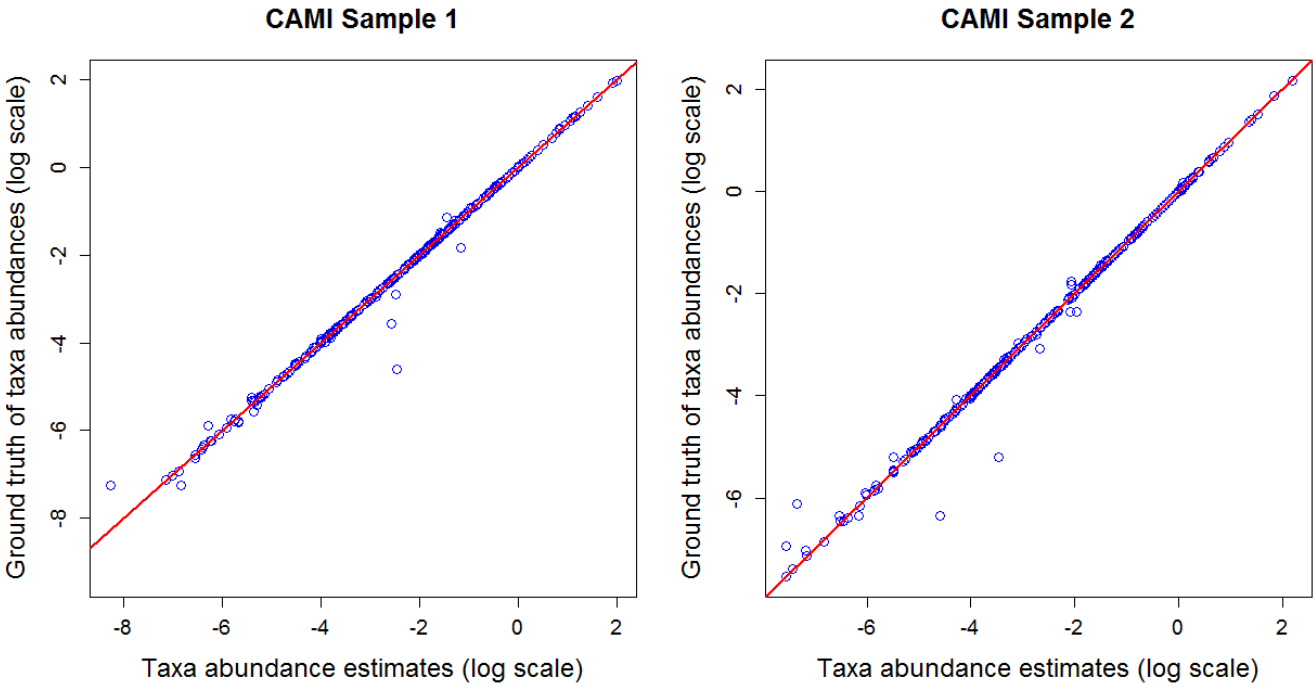
Supplementary Figure 4: The **FAMeS data** comprises three different samples with abundance profiles according to low (*LC*), medium (*MC*) and high complexity (*HC*), a common classification in metagenomics. Thereby, a low complexity sample may represent a bioreactor community with one dominant among low abundant genomes, while medium complexity refers to a moderately complex community with few dominating taxa. High complexity samples are frequently characterized by no dominating taxa present or also by very long tails of low abundant taxa.

Supplementary Figure 5:



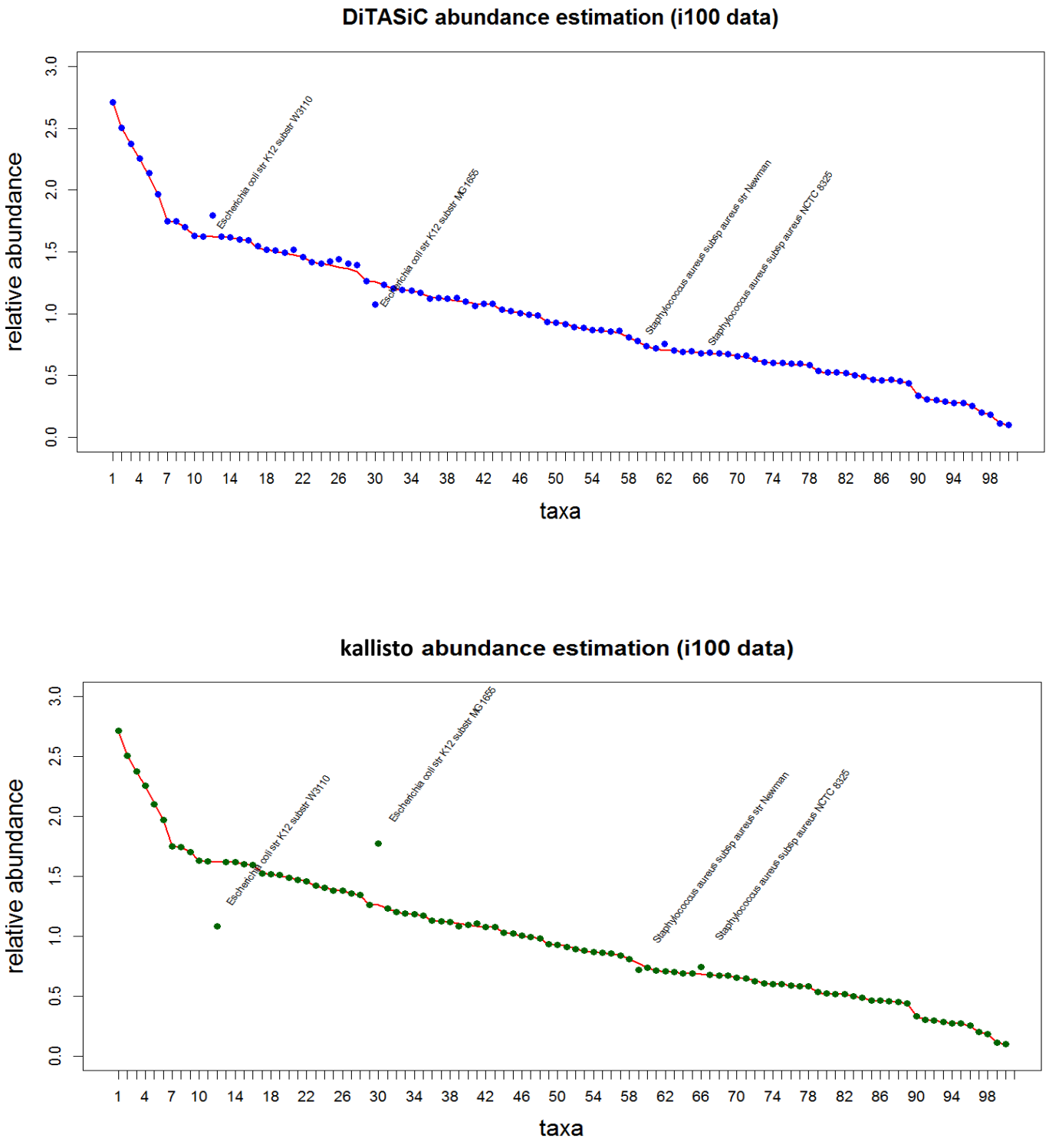
Supplementary Figure 5: **Accuracy of abundance estimates by DiTASiC for the FAMeS data sets.** For all three samples of LC, MC, and HC, abundance estimates exhibit only tiny divergences from the ground truth. High accuracy is depicted by the points found on the diagonal. Hence, highly accurate abundance estimates of the considered 122 taxa are achieved across all three different abundance complexity profiles.

Supplementary Figure 6:



Supplementary Figure 6: **Accuracy of abundance estimates by DiTASiC for samples of the CAMI benchmark data set.** For both samples, abundance estimates of the 255 taxa show high accuracy apart from very few outliers. High accuracy is depicted by points found on the diagonal. Notably, accurate estimates are also achieved for very low relative abundances below 0.01%.

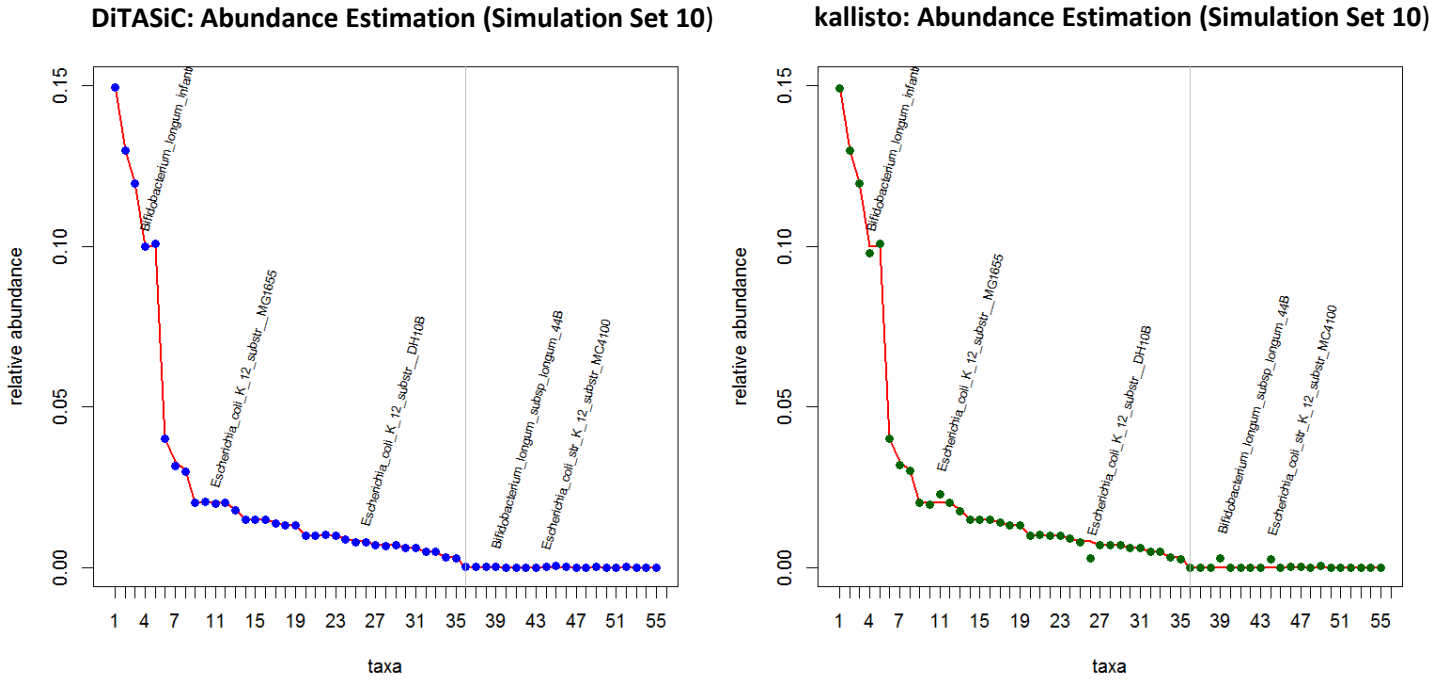
Supplementary Figure 7:



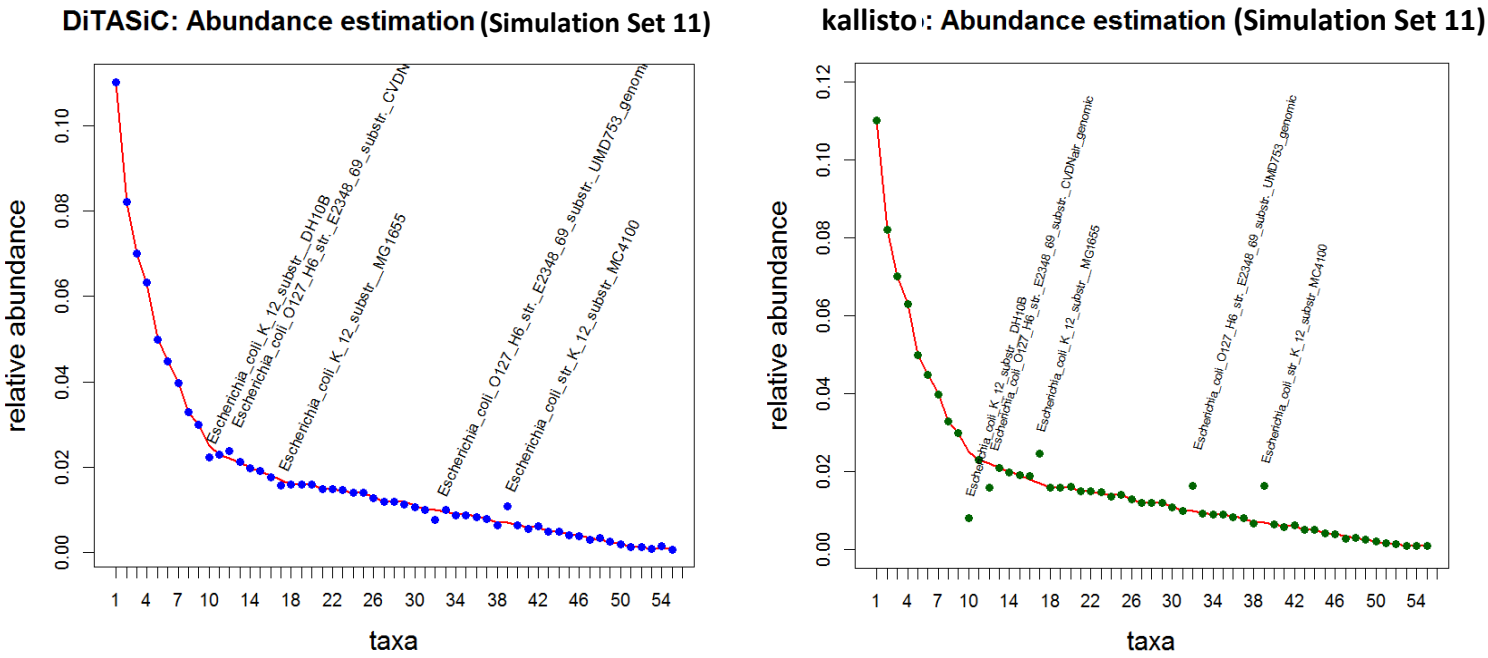
Supplementary Figure 7: **Accuracy of abundance estimates by DiTASiC and kallisto for the Illumina 100 benchmark data set (i100)** *(Mende et al., 2012).* The red line refers to the ground truth values and the points show the abundance estimates obtained by the corresponding tool. Overall, a high accuracy of abundance estimates is achieved for the 100 taxa by both tools across the entire abundance range. A bias in abundance estimation is observed for some strains of high sequence similarity, namely for the *Escherichia coli* sub-strains and for two *Staphylococcus aureus* strains. A more accurate abundance resolution of these strain clusters is obtained by DiTASiC in comparison to kallisto (also refer to Figure S 9C). Further, results of DiTASiC can be related to a recent benchmark study of different abundance profiling tools tested on the *i100* data set by Schaffer *et al*. (2017): see second part 'Data& Models' at the end of this Supplemental Material.

Supplementary Figure 8:

(A)



(B)

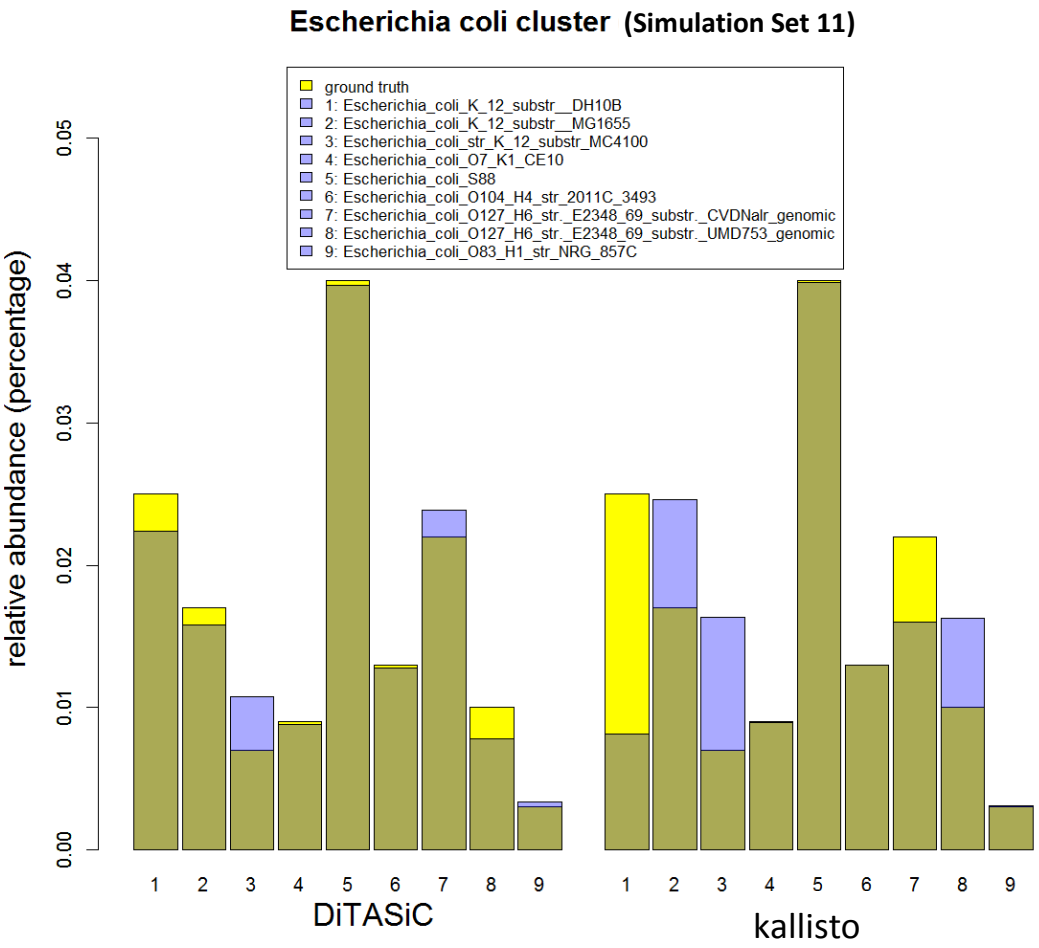


Supplementary Figure 8: **Accuracy of abundance estimates by DiTASiC and kallisto for (A) data set 10 and (B) data set 11 of simulation group (3).** The red line refers to the ground truth values and the points show the abundance estimates obtained by the corresponding tools.

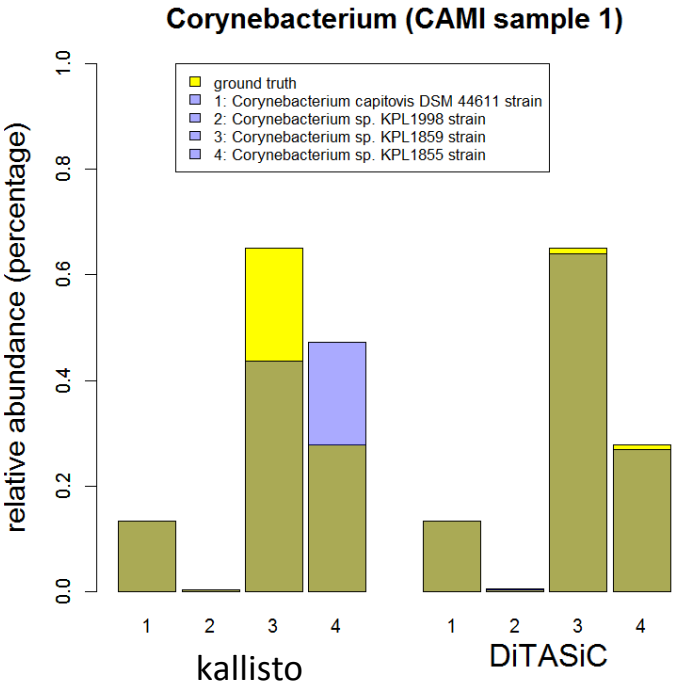
(A) Set 10 serves to study the impact of absent strains from highly similar clusters (gray vertical line in the plot to mark the section of absent strains). Overall, highly accurate abundance estimates are obtained by DiTASiC. Hence an un-biased estimation of strains of the clusters affected by absent strains is achieved. kallisto exhibits difficulties with some strains of

high sequence similarity, here concerning the *Escherichia coli K12* sub-strain cluster and the *Bifidobacterium longum* strain group, causing a bias of abundance estimations and calling two of the absent strains abundant.
(B) Set 11 focuses on the resolution of large and highly similar strain clusters, having all 55 taxa abundant in the data set (refer also to the matrix of reference similarities in Figure S3). Overall accurate abundance estimations are obtained and also an accurate resolution within the diverse strain clusters is achieved by DiTASiC. The large *E.coli* cluster causes some abundance biases for both tools, especially for the sub-strain sequences of sequence similarities above 95%. Here, DiTASiC proves more accurate estimations and an overall better resolution within the considered cluster (see also Figure S 9A).

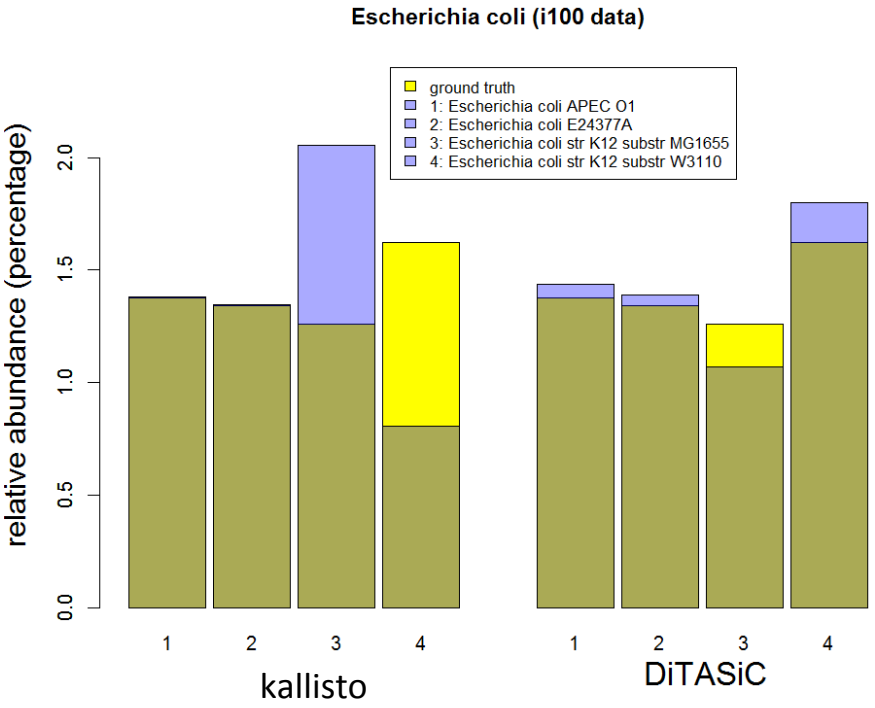
Supplementary Figure 9:
(A)

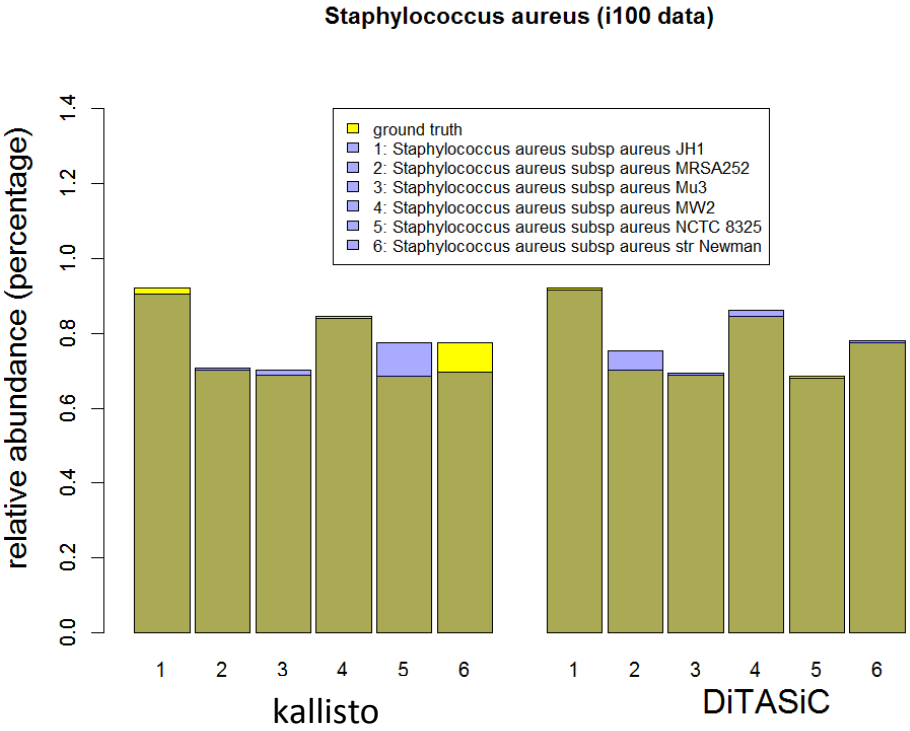


Supplementary Figure 9 (B): see caption on page (14)

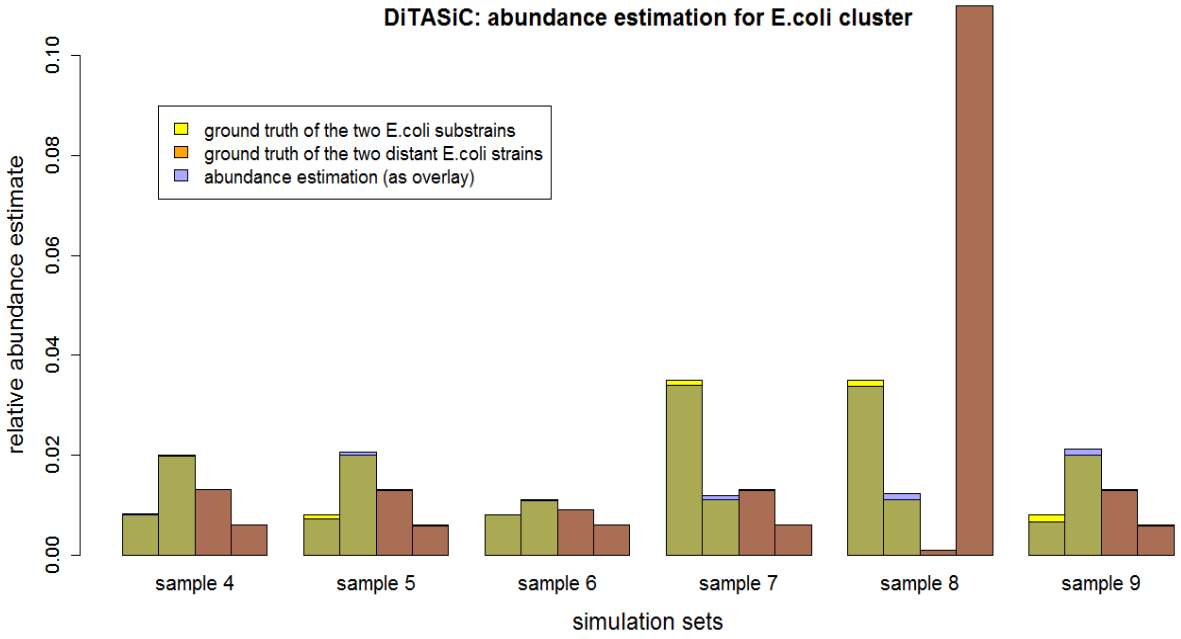


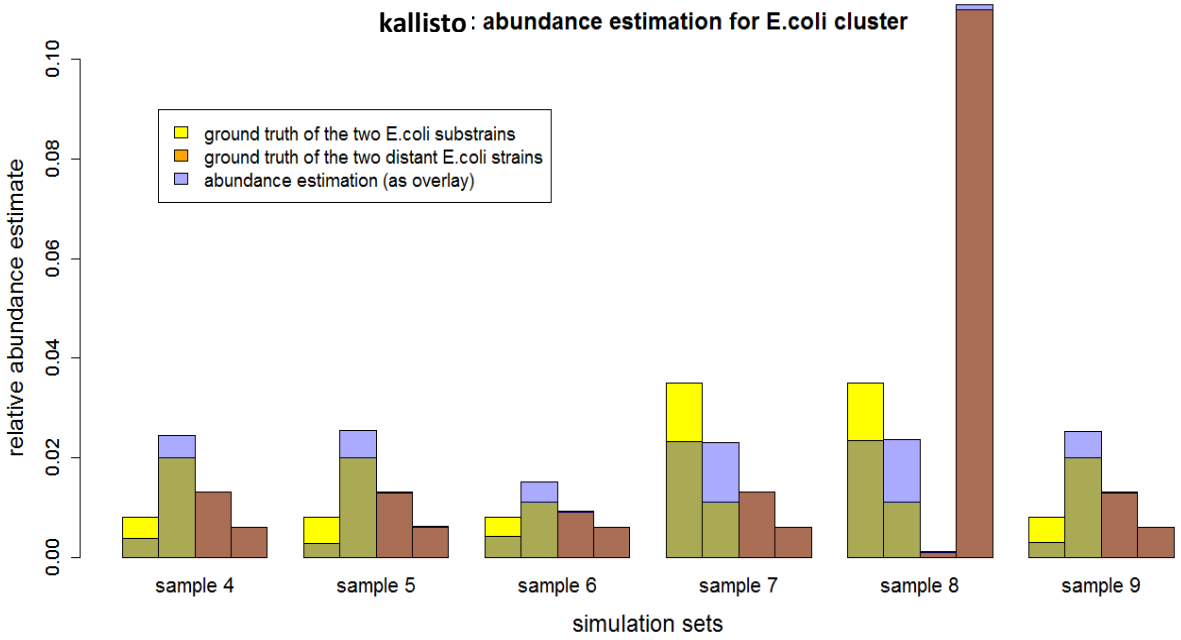
Supplementary Figure 9 (C): see caption on page (14)





Supplementary Figure 9 (D): see caption on page (14)





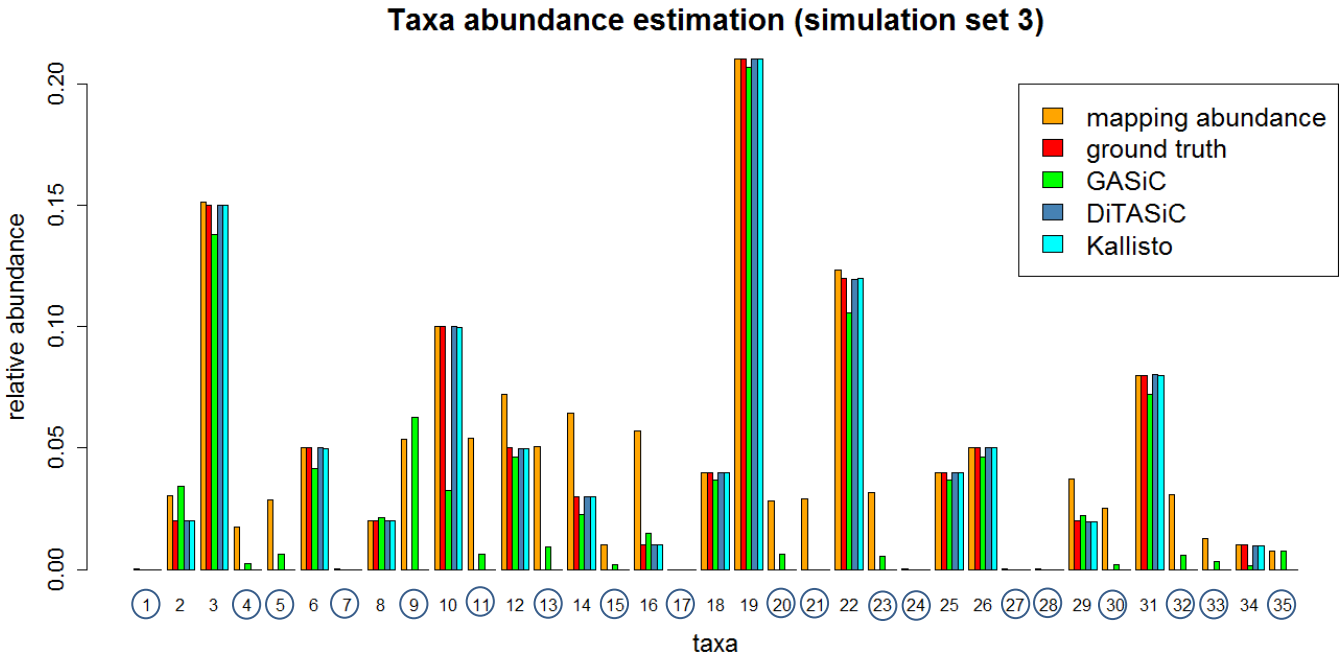
Supplementary Figure 9: **Accuracy of abundance resolution within strain clusters by DiTASiC and kallisto shown for different examples in the CAMI, i100 and simulation data (A-D).** The ground truth abundance is displayed in yellow and the estimated abundances by the corresponding tools are overlayed with purple colour.
(A) The largest cluster in simulation set 11 comprises nine different *Escherichia coli* strains. Challenging is the resolution of the present sub-strains which share sequence similarities above 98%. DiTASiC enables a more accurate abundance resolution in comparison to kallisto. **(B)** An example of a *Corynebacterium* cluster in the CAMI set reveals a perfect resolution by DiTASiC, again, two of the strains are characterized by high sequence similarity. **(C)** An increased error in abundance estimation in the i100 data was shown in Fig S7 for the *Escherichia coli* sub-strains and for two *Staphylococcus aureus* strains. A more accurate abundance resolution of these strain clusters is obtained by DiTASiC. **(D)** Here, we consider the six different simulations sets of group (2) focusing on the abundance estimates obtained for the 4 strains of the *E.coli* cluster (*E. coli K-12 substr. DH10B* , *E. coli K-12 substr. MG1655, E.coli O7:K1 str. CE10, E.coli S88)* (visualized only for group

(2), as in group (1) not all strains of the *E.coli* cluster are abundant). The *E.coli* cluster consists of two sub-strains, which share 98% sequence similarity, and two more distant strains. DiTASiC enables a highly accurate abundance resolution of the entire strain cluster, as is shown by an almost perfect abundance estimation overlay in the plot across all samples. kallisto exhibits problems in the resolution of the two sub-strains, which is shown by a consistent abundance underestimation of *E. coli K-12 substr. DH10B* and abundance overestimation of *E. coli K-12 substr. MG1655,* while the two distant strains receive accurate estimations.

Overall, it can be observed that a common error in the resolution of a strain cluster is an abundance interchange or equalization of abundances of similar sub-strains. In the resolution by DiTASiC these errors are shown to be avoided.

Supplementary Figure 10:

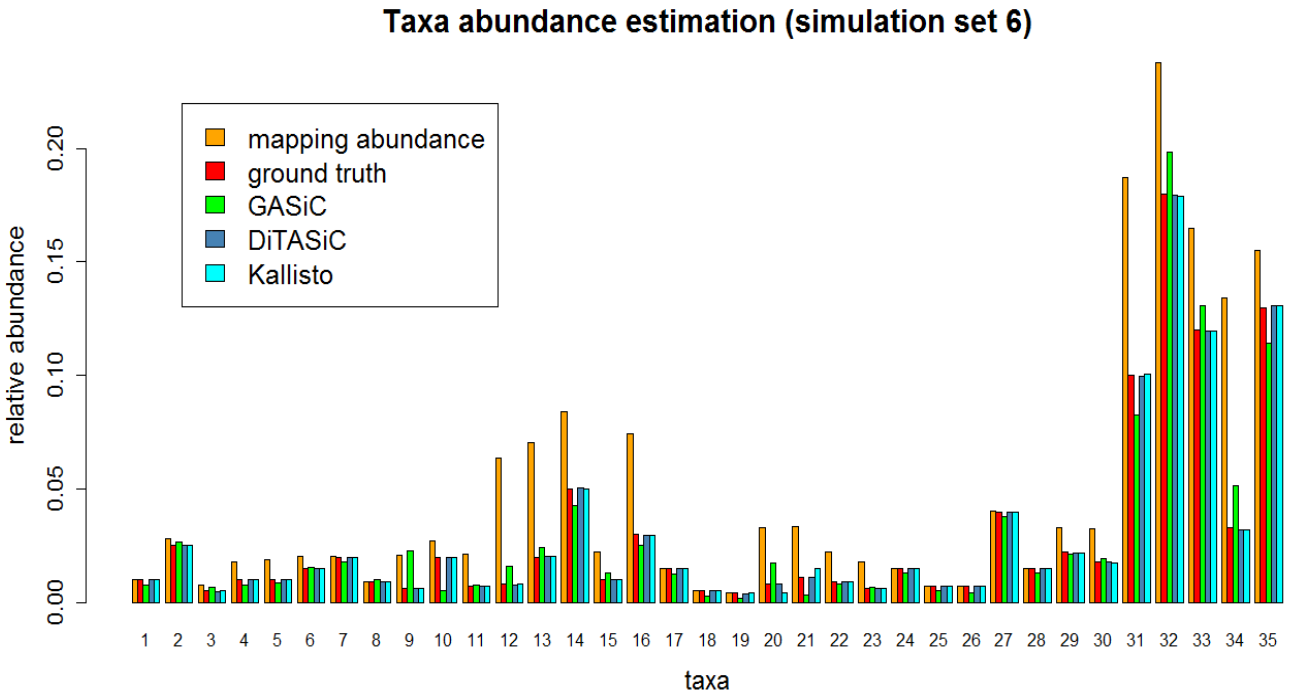
(a)



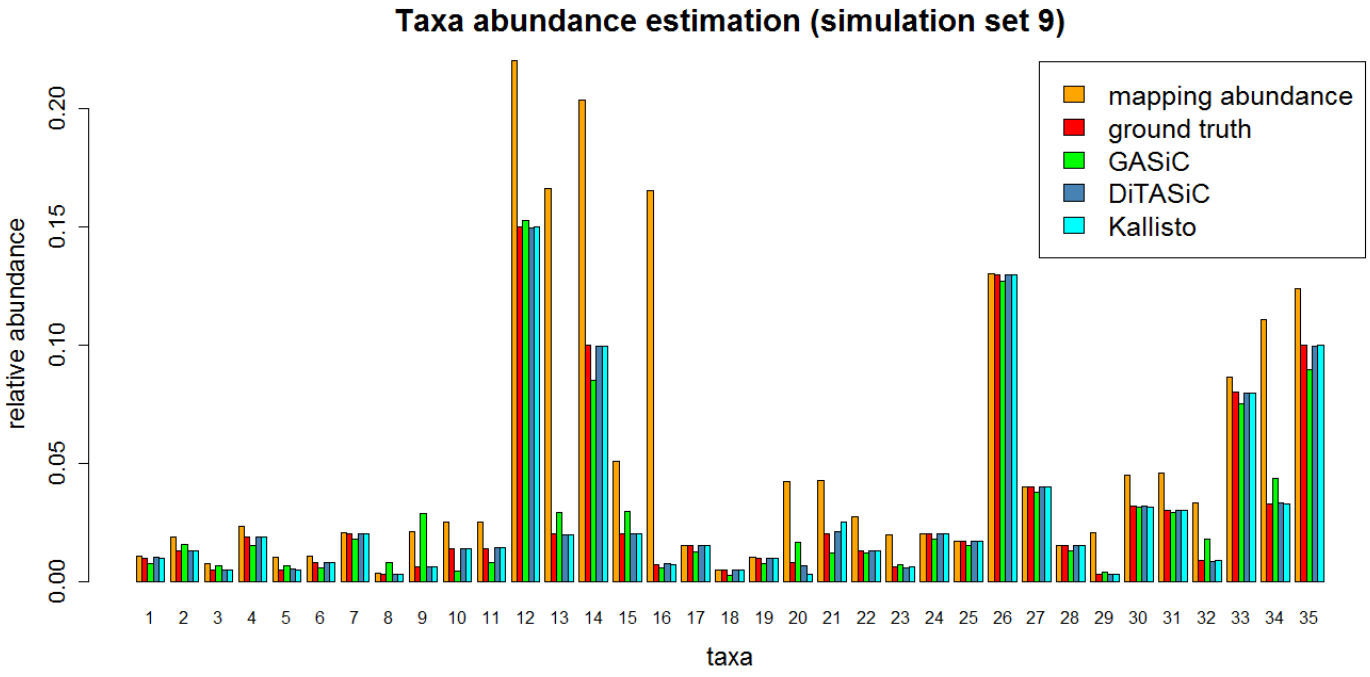
Taxa, shown in the plot according to the numbers:

1: Alistipes_finegoldii_DSM_17242
2: Bacillus_anthracis_Sterne
3: Bacillus_cereus_ATCC_10987
4: Bacillus_cereus_E33L
5: Bacteroides_fragilis_638R
6: Bacteroides_fragilis_NCTC_9343
7: Bacteroides_thetaiotaomicron_VPI_5482
8: Bifidobacterium_adolescentis_ATCC_15703
9: Bifidobacterium_bifidum_BGN4
10: Bifidobacterium_bifidum_PRL2010
11: Bifidobacterium_bifidum_S17
12: Bifidobacterium_longum_BBMN68
13: Bifidobacterium_longum_DJO10A
14: Bifidobacterium_longum_infantis_157F
15: Bifidobacterium_longum_infantis_ATCC_15697
16: Bifidobacterium_longum_JCM_1217
17: Clostridium_phytofermentans_ISDg
18: Clostridium_saccharolyticum_WM1
19: Clostridium_SY8519
20: Escherichia_coli_K_12_substr__DH10B
21: Escherichia_coli_K_12_substr__MG1655
22: Escherichia_coli_O7_K1_CE10
23: Escherichia_coli_S88
24: Eubacterium_eligens_ATCC_27750
25: Eubacterium_rectale_ATCC_33656
26: Odoribacter_splanchnicus_DSM_20712
27: Pantoea_ananatis_PA13
28: Roseburia_hominis_A2_183
29: Shigella_dysenteriae_Sd197
30: Shigella_flexneri_2a_301
31: Streptococcus_salivarius_57_I
32: Streptococcus_salivarius_CCHSS3
33: Streptococcus_salivarius_JIM8777
34: Streptococcus_suis_D9
35: Streptococcus_suis_ST3

(b)



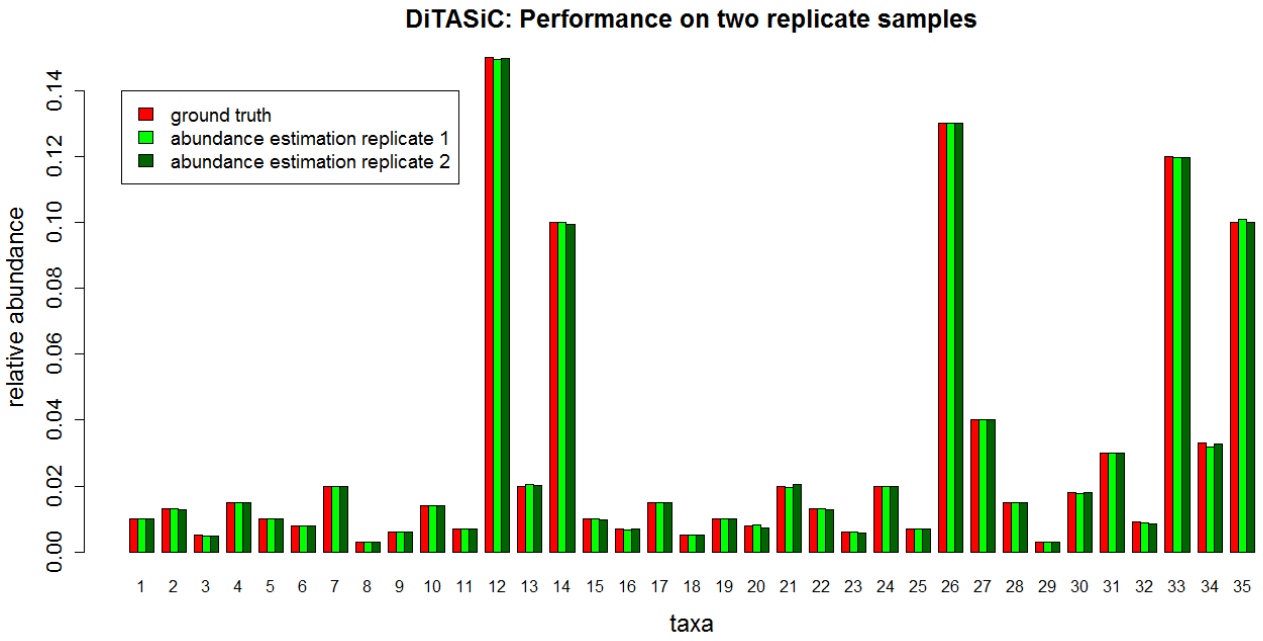
(c)



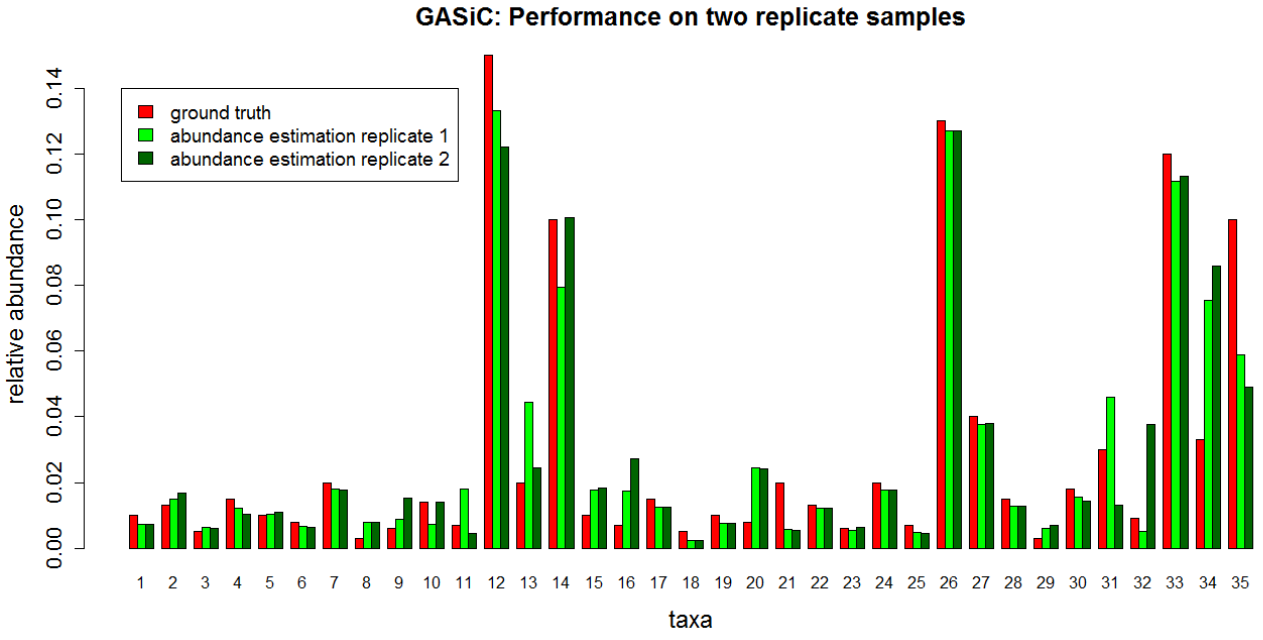
Supplementary Figure 10: **Taxa abundance estimates exemplary for three simulation data set of various abundance profiles (a-c), presenting the different tools DiTASiC, GASiC and kallisto in comparison to the ground truth and observed mapping abundances**. Mapping abundances are biased due to read ambiguities which causes overestimation or assignment of reads to absent taxa (absent taxa are marked with a circle around the taxa number). DiTASiC as well as kallisto exhibit highly accurate estimations and a clear improvement over GASiC.

Supplementary Figure 11:

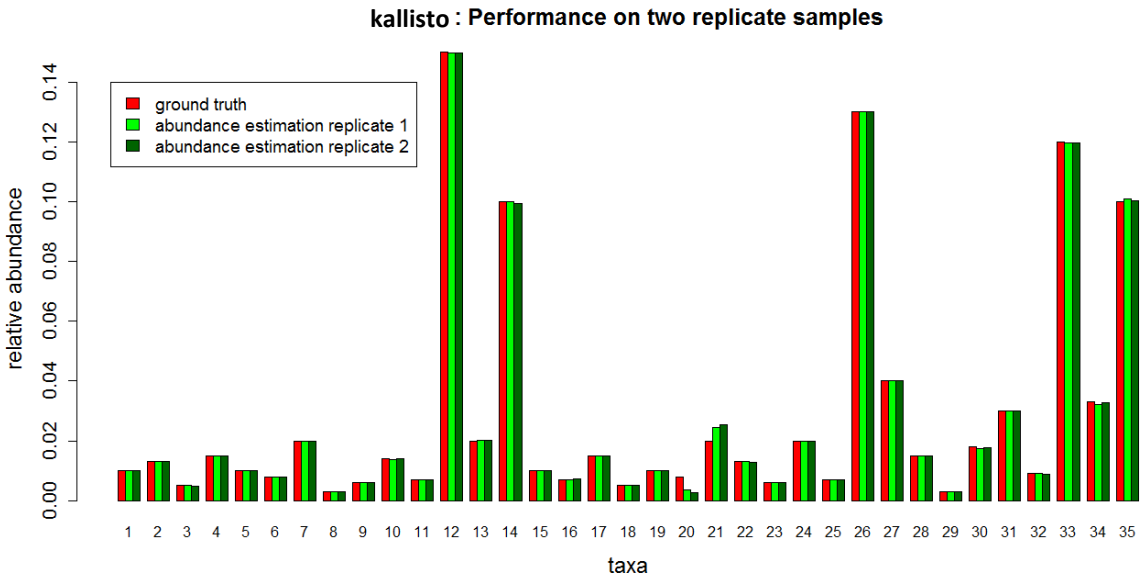
(a)



(b)



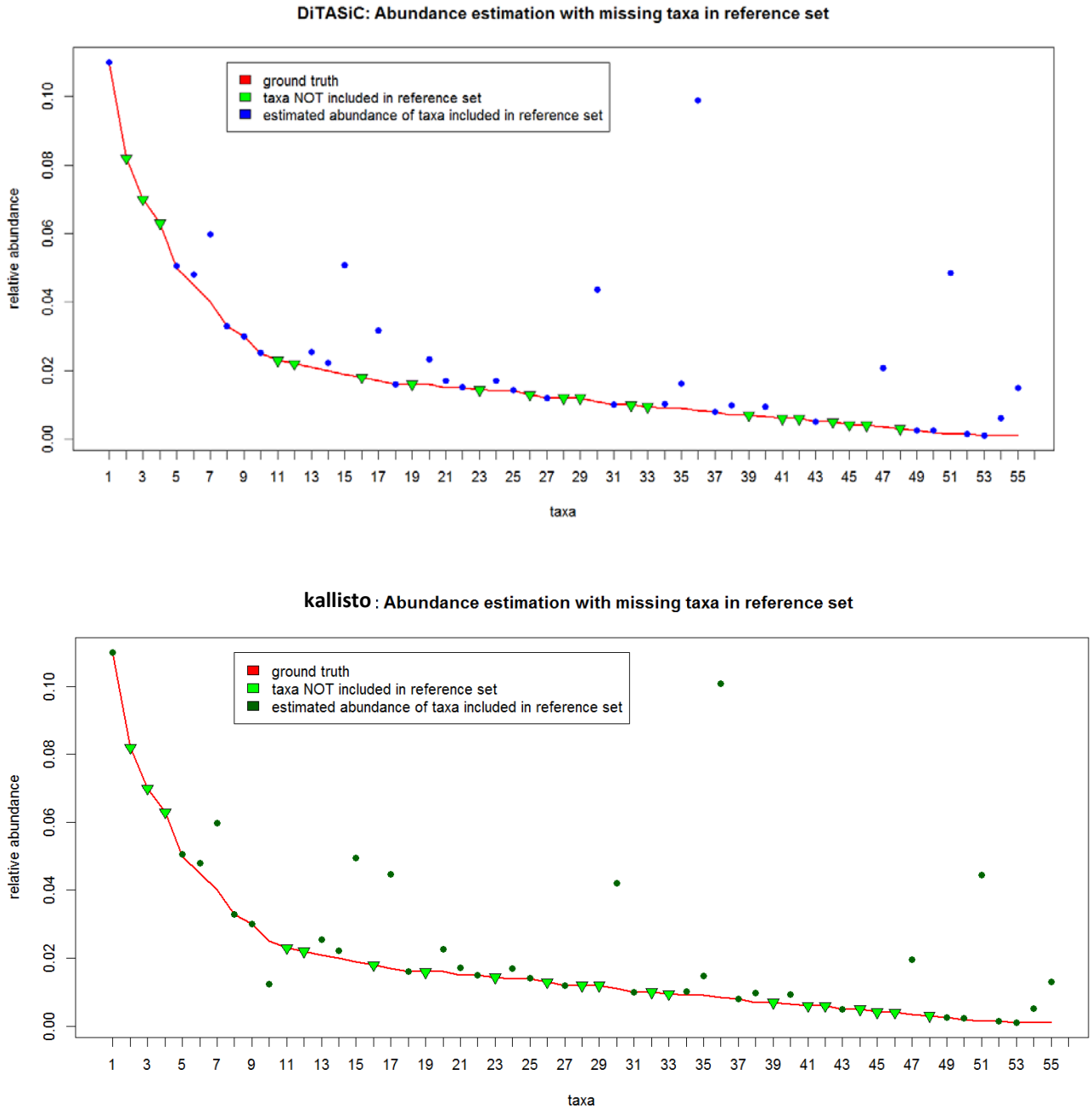
(c)



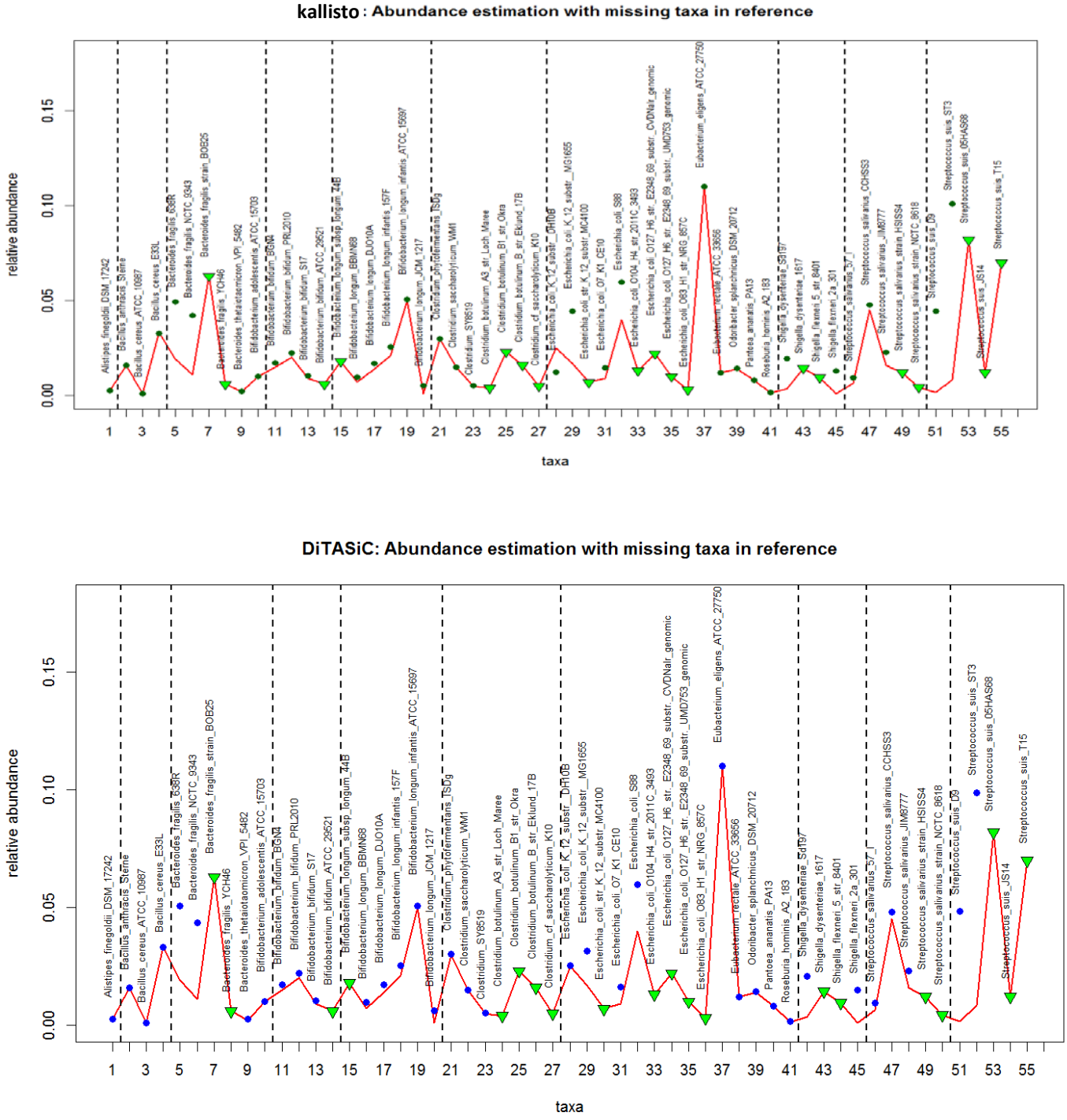
Supplementary Figure 11: **Robustness evaluation of (a) DiTASiC, (b) GASiC, and (c) kallisto, on two replicate samples** from the simulation data (data set 4 and data set 5, respectively; taxa are numbered according to the list given in Supplementary Figure 7). DiTASiC and kallisto show an overall robust performance in abundance estimation of all 35 taxa in the replicates, and a significant improvement compared to GASiC.

Supplementary Figure 12:

(A)



(B)



Supplementary Figure 12: Impact of missing taxa in a reference set on the abundance estimation. The red line refers to the ground truth values, points refer to the abundance estimates obtained by the corresponding tool, while triangles mark absent taxa. Vertical lines are drawn to define sections of strain clusters. In this study, reads derived from 55 taxa are contrasted to a reduced reference set of 35 taxa to investigate the impact of missing taxa in a selected reference set. One consequence is that 11% of reads are not aligned and therefore are eliminated from the subsequent model calculations; second, the abundances estimated for some taxa are overestimated by the tools. However, a closer look reveals that it always concerns closely related strains which show an increased abundance due to missing strains within their cluster. However, no overall abundance bias is observed. Noticeable, while DiTASiC only exhibits abundance overestimations, kallisto also shows underestimation and overestimations within one cluster to compensate for missing taxa.

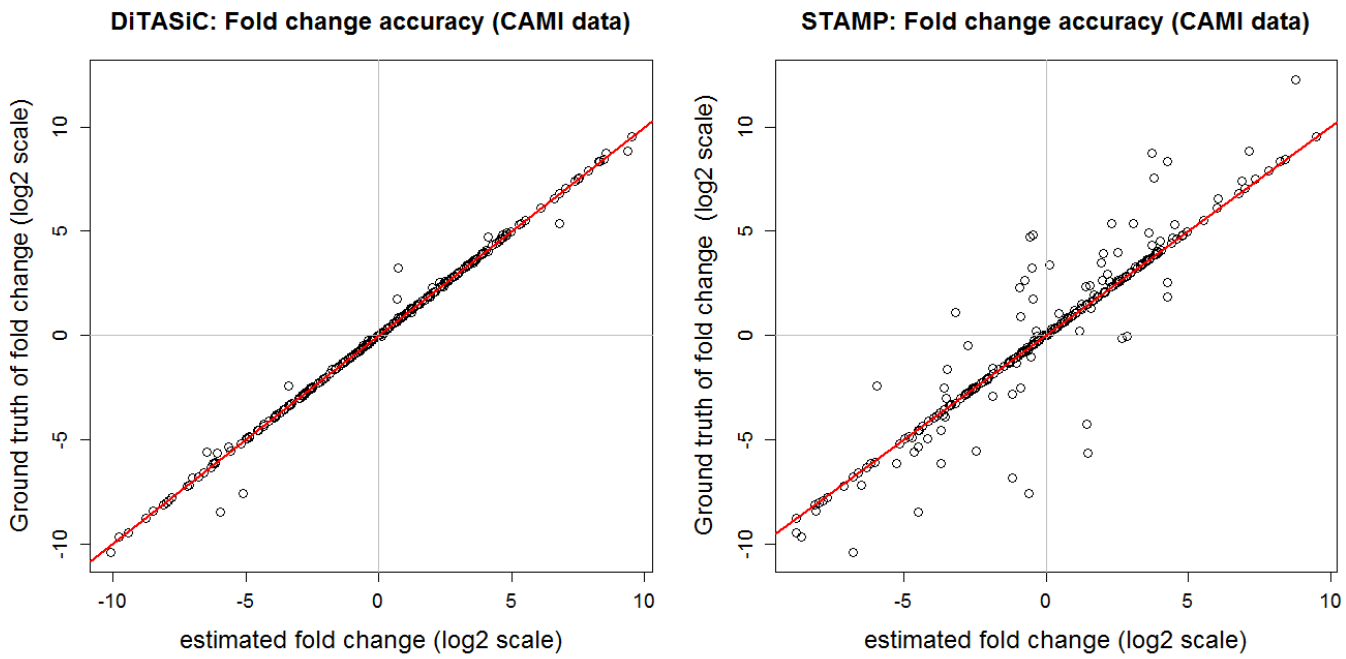
Supplementary Table S1

	Simulation Data			FAMeS Data (Pignatelli et al.)		
	Sim 1	Sim 2	Sim 3	LC	MC	HC
# absent taxa (TN)	25	21	19	10	12	10
# false-positives (FP)	0	0	0	0	0	0
# present taxa (TP)	10	14	16	112	110	112
# false-negatives (FN)	0	0	0	1	0	0
Sensitivity	1	1	1	0.991	1	1
Specificity	1	1	1	1	1	1

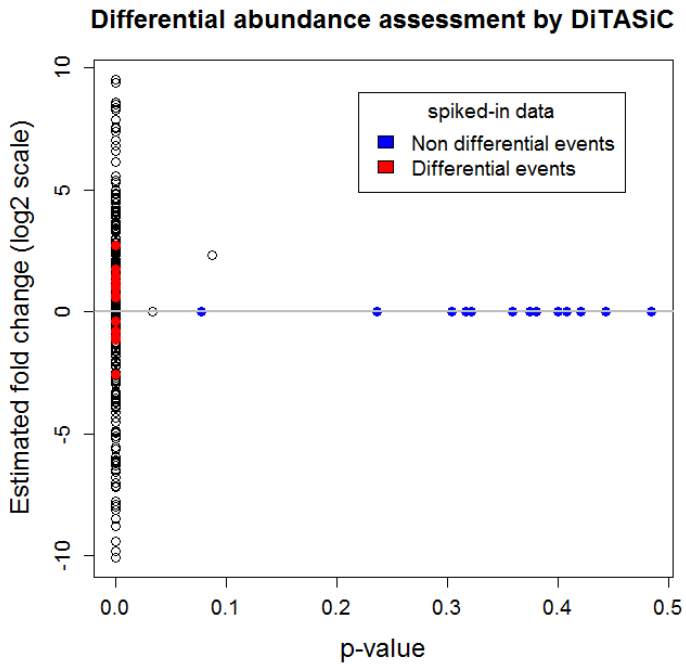
Supplementary Table 1: **Detection of absent taxa**. We tested the detection performance with different proportions of absent taxa included, namely in the simulation group (1) and the FAMeS data sets. In the simulation data of group (1) only 28%, 40% and 45% taxa out of the 35 provided references are abundant in the data. Absolute numbers of absent and present taxa of each data set are reported in this table as well as absolute numbers of false-positive or false-negative detections. DiTASiC achieves exact detections, resulting in a sensitivity and specificity of 100%. The proportion of absent taxa in the FAMeS data refers to 8%, 9% and 8% based on the reference set of 122 taxa overall. A sensitivity and specificity of 100% is again reported for DiTASiC for the MC and HC data. In the LC set a reduced sensitivity is caused by one missed abundant taxon.

Supplementary Figure 13:

(a)



(b)



Supplementary Figure 13: **(a) Fold change accuracy achieved by DiTASiC in comparison to fold change accuracy obtained by STAMP in the CAMI data, and (b) differential abundance assessment using p-values by DiTASiC.** (a) Fold change estimates are proven to be highly accurate for DiTASiC with an *SSE* 19 times smaller compared to the STAMP output. This is depicted in the plots by fold change estimates found on the diagonal for DiTASiC, while many estimates are divergent from the diagonal in the plot by STAMP. (b) Computed p-values by the statistical framework in DiTASiC prove to clearly separate the spiked-in non-differential and differential taxa. Other taxa of the data set, holding fold change values greater than zero, also receive very small p-values stating differential abundance, but cannot be further confirmed here.

Supplementary Table S2

Sample comparison	DiTASiC	STAMP
FAMeS: LC vs. MC	0.0047	0.5089
FAMeS: LC vs. HC	0.0013	0.4992
FAMeS: MC vs. HC	0.0051	0.0986
CAMI : S1 vs S2	25.07	476.91

Supplementary Table 2: ***SSE* values of fold change accuracy obtained by DiTASiC in comparison to STAMP in different sample comparisons.** *SSE* values of DiTASiC are significantly smaller compared to the ones computed by STAMP, indicating the importance of read ambiguity resolution and integration of abundance estimate uncertainties for differential abundance analysis.

Data sets & Model descriptions

Data Set Description

1) Simulation Data:

Nine data sets comprise 35 reference genomes from bacterial strains downloaded from NCBI, two additional data sets (set 10,11) were extended by further strain and sub-strain sequences (total 55 reference genomes) to create a high strain cluster density.

Each data set consists of 750,000 reads of 100bp length simulated by Mason (Holtgrewe, 2010), following Illumina read characteristics with default parameter settings. Reads are simulated according to the following abundance profiles.

Mason parameters:

- Total number of simulated reads: 750000
- Read length: 100 bp
- Replicate study Sim 4/5: Default.seed = 2048 (Sim 4), seed = 22 (Sim 5)

Taxa abundance list (1):

Taxa Name	GenBank accession number	Ground Truth: relative taxa abundance							
		Group (1)			Group (2)				
		Sim 1	Sim 2	Sim 3	Sim 4 ~5	Sim 6	Sim 7	Sim 8	Sim 9
Alistipes finegoldii DSM 17242	GCF_000265365.1	0.01	0.01	0	0.01	0.01	0.005	0.001	0.01
Bacillus anthracis str. Sterne	GCF_000008165.1	0.02	0.04	0.02	0.013	0.025	0.025	0.002	0.013
Bacillus cereus ATCC 10987	GCA_000008005.1	0.3	0.2	0.15	0.005	0.005	0.009	0.003	0.005
Bacillus cereus E33L	GCA_000011625.1	0.2	0.25	0	0.015	0.01	0.01	0.008	0.019
Bacteroides fragilis 638R	GCA_000210835.1	0	0	0	0.01	0.01	0.15	0.004	0.005
Bacteroides fragilis NCTC 9343	GCA_000025985.1	0	0.01	0.05	0.008	0.015	0.015	0.13	0.008
Bacteroides thetaiotaomicron VPI-5482	GCA_000011065.1	0	0	0	0.02	0.02	0.01	0.17	0.02
Bifidobacterium adolescentis ATCC 15703	GCA_000010425.1	0	0.02	0.02	0.003	0.009	0.009	0.009	0.003
Bifidobacterium bifidum BGN4	GCA_000265095.1	0	0	0	0.006	0.006	0.002	0.002	0.006
Bifidobacterium bifidum PRL2010	GCA_000165905.1	0.21	0.18	0.1	0.014	0.02	0.011	0.011	0.014
Bifidobacterium bifidum S17	GCA_000164965.1	0	0.01	0	0.007	0.007	0.03	0.025	0.014
Bifidobacterium longum BBMN68	GCA_000166315.1	0.14	0.14	0.05	0.15	0.008	0.008	0.008	0.15
Bifidobacterium longum DJO10A	GCA_000008945.1	0	0	0	0.02	0.02	0.003	0.02	0.02
Bifidobacteriumlongum infantis 157F	GCA_000196575.1	0	0	0.03	0.1	0.05	0.05	0.005	0.1
Bifidobacterium longum infantis ATCC 15697	GCA_000020425.1	0	0	0	0.01	0.01	0.014	0.006	0.02
Bifidobacterium longum JCM 1217	GCA_000196555.1	0	0.01	0.01	0.007	0.03	0.03	0.007	0.007
Clostridium phytofermentans ISDg	GCA_000018685.1	0	0	0	0.015	0.015	0.015	0.015	0.015
Clostridium saccharolyticum WM1	GCA_000144625.1	0.08	0	0.04	0.005	0.005	0.08	0.08	0.005
Clostridium SY8519	GCA_000270305.1	0.02	0.02	0.21	0.01	0.004	0.004	0.008	0.01
Escherichia coli K-12 substr. DH10B	GCA_000019425.1	0	0	0	0.008	0.008	0.035	0.035	0.008
Escherichia coli K-12 substr. MG1655	GCA_000005845.1	0	0	0	0.02	0.011	0.011	0.011	0.02
Escherichia coli O7:K1 str. CE10	GCA_000227625.1	0	0	0.12	0.013	0.009	0.013	0.001	0.013
Escherichia coli S88	GCA_000026285.1	0	0	0	0.006	0.006	0.006	0.11	0.006
Eubacterium eligens ATCC 27750	GCA_000146185.1	0	0	0	0.02	0.015	0.015	0.01	0.02
Eubacterium rectale ATCC 33656	GCA_000020605.1	0.01	0	0.04	0.007	0.007	0.012	0.012	0.017
Odoribacter splanchnicus DSM 20712	GCA_000190535.1	0.01	0.03	0.05	0.13	0.007	0.007	0.007	0.13
Pantoea ananatis PA13	GCA_000233595.1	0	0	0	0.04	0.04	0.04	0.016	0.04
Roseburia hominis A2-183	GCA_000225345.1	0	0	0	0.015	0.015	0.026	0.003	0.015
Shigella dysenteriae Sd197	GCA_000012005.1	0	0	0.02	0.003	0.022	0.022	0.13	0.003
Shigella flexneri 2a str. 301	GCA_000006925.2	0	0	0	0.018	0.018	0.002	0.017	0.032
Streptococcus salivarius 57.I	GCA_000305335.1	0	0.03	0.08	0.03	0.1	0.005	0.018	0.03
Streptococcus salivarius CCHSS3	GCA_000253335.1	0	0	0	0.009	0.18	0.18	0.009	0.009
Streptococcus salivarius JIM8777	GCA_000253315.1	0	0	0	0.12	0.12	0.12	0.002	0.08
Streptococcus suis D9	GCA_000231885.1	0	0.05	0.01	0.033	0.033	0.022	0.005	0.033
Streptococcus suis ST3	GCA_000204625.1	0	0	0	0.1	0.13	0.004	0.1	0.1

Taxa abundance list (2):

Taxa Name	GenBank accession number	Ground Truth: relative taxa abundance Group (3): Sim 10	Ground Truth: relative taxa abundance Group (3): Sim 11
Alistipes_finegoldii_DSM_17242	GCF_000265365.1	0.01	0.0025
Bacillus_anthracis_Sterne	GCF_000008165.1	0.013	0.016
Bacillus_cereus_ATCC_10987	GCA_000008005.1	0.005	0.001
Bacillus_cereus_E33L	GCA_000011625.1	0.015	0.033
Bacteroides_fragilis_638R	GCA_000210835.1	0.01	0.019
Bacteroides_fragilis_NCTC_9343	GCA_000025985.1	0.008	0.011
Bacteroides_fragilis_strain_BOB25	GCA_000965785.1	0	0.063
Bacteroides_fragilis_YCH46	GCA_000009925.1	0	0.006
Bacteroides_thetaiotaomicron_VPI_5482	GCA_000011065.1	0.008	0.002
Bifidobacterium_adolescentis_ATCC_15703	GCA_000010425.1	0.02	0.01
Bifidobacterium_bifidum_BGN4	GCA_000265095.1	0.003	0.015
Bifidobacterium_bifidum_PRL2010	GCA_000165905.1	0.006	0.02
Bifidobacterium_bifidum_S17	GCA_000164965.1	0.014	0.009
Bifidobacterium_bifidum_ATCC_29521	GCA_001025135.1	0	0.006
Bifidobacterium_longum_subsp_longum_44B	GCA_000261265.1	0	0.018
Bifidobacterium_longum_BBMN68	GCA_000166315.1	0.15	0.007
Bifidobacterium_longum_DJO10A	GCA_000008945.1	0.02	0.014
Bifidobacterium_longum_infantis_157F	GCA_000196575.1	0.1	0.021
Bifidobacterium_longum_infantis_ATCC_15697	GCA_000020425.1	0.01	0.05
Bifidobacterium_longum_JCM_1217	GCA_000196555.1	0.007	0.001
Clostridium_phytofermentans_ISDg	GCA_000018685.1	0.015	0.03
Clostridium_saccharolyticum_WM1	GCA_000144625.1	0.005	0.015
Clostridium_SY8519	GCA_000270305.1	0.01	0.005
Clostridium_botulinum_A3_str_Loch_Maree	GCA_000019545.1	0	0.004
Clostridium_botulinum_B1_str_Okra	GCA_000019305.1	0	0.023
Clostridium_botulinum_B_str_Eklund_17B	GCA_000307125.1	0	0.016
Clostridium_cf_saccharolyticum_K10	GCA_000210535.1	0	0.005
Escherichia_coli_K_12_substr__DH10B	GCA_000019425.1	0.008	0.025
Escherichia_coli_K_12_substr__MG1655	GCA_000005845.1	0.02	0.017
Escherichia_coli_str_K_12_substr_MC4100	GCA_000499485.1	0	0.007
Escherichia_coli_O7_K1_CE10	GCA_000227625.1	0.013	0.009
Escherichia_coli_S88	GCA_000026285.1	0.006	0.04
Escherichia_coli_O104_H4_str_2011C_3493	GCA_000299455.1	0	0.013
Escherichia_coli_O127_H6_str._E2348_69_substr._CVDNalr_genomic	GCA_000442065.2	0	0.022
Escherichia_coli_O127_H6_str._E2348_69_substr._UMD753_genomic	GCA_000442085.2	0	0.01
Escherichia_coli_O83_H1_str_NRG_857C	GCA_000183345.1	0	0.003
Eubacterium_eligens_ATCC_27750	GCA_000146185.1	0.02	0.11
Eubacterium_rectale_ATCC_33656	GCA_000020605.1	0.007	0.012
Odoribacter_splanchnicus_DSM_20712	GCA_000190535.1	0.13	0.014
Pantoea_ananatis_PA13	GCA_000233595.1	0.04	0.008
Roseburia_hominis_A2_183	GCA_000225345.1	0.015	0.0014
Shigella_dysenteriae_Sd197	GCA_000012005.1	0.003	0.0035
Shigella_dysenteriae_1617	GCA_000497505.1	0	0.0144
Shigella_flexneri_5_str_8401	GCA_000013585.1	0	0.0095
Shigella_flexneri_2a_301	GCA_000006925.2	0.018	0.001
Streptococcus_salivarius_57_I	GCA_000305335.1	0.03	0.0065
Streptococcus_salivarius_CCHSS3	GCA_000253335.1	0.009	0.045
Streptococcus_salivarius_JIM8777	GCA_000253315.1	0.12	0.016
Streptococcus_salivarius_strain_HSISS4	GCA_000448685.2	0	0.012
Streptococcus_salivarius_strain_NCTC_8618	GCA_000785515.1	0	0.0042
Streptococcus_suis_D9	GCA_000231885.1	0.033	0.0015
Streptococcus_suis_ST3	GCA_000204625.1	0.1	0.0085
Streptococcus_suis_05HAS68	GCA_000168355.3	0	0.082
Streptococcus_suis_JS14	GCA_000186405.1	0	0.012
Streptococcus_suis_T15	GCA_000494895.1	0	0.07

2) CAMI Data set:

Within the CAMI challenge (https://data.cami-challenge.org) (Sczyrba *et al.*, 2017), we selected a benchmark data set of medium complexity, which is provided for testing tools, with a ground truth of taxa proportions being available (‘*2. Toy Test Dataset Medium_Complexity’*). It comprises two 15 gb samples each holding about 150 million paired-end reads of 100 bp length based on HiSeq sequencing. A total of 225 bacterial and archaea genomes are present in both samples. Different clusters of strains with high sequence similarities are present within the 128 genera and 199 species. The relative abundances of the taxa range from 0.00009% to 8% in a medium complexity environment with median values of 0.1% and 0.08% for the samples, respectively. Comparison of the two samples yields taxa fold changes with a large span from 0.0009 to 1024. However, no ground truth is given for differential abundance classification and only fold change accuracy can be evaluated.

Therefore we extend the data set by simulating spike-in data: we selected 30 new strains from genera already present in the original set. A total of 20 million reads per sample are simulated from the new references based on a defined abundance given for each sample. Simulations are conducted using Mason (Holtgrewe, 2010), with error profiles matching the original reads, and are subsequently merged with the original set. Abundances of the added taxa are defined such that a ground truth of 15 differential and 15 non-differential events is created for additional differential assessment.

Mason parameters:

- Total number of simulated reads: N.sim = 20,000,000
- Read length: 100 bp
- Seed: 22
- Mismatch probability (begin): 0.005
- Mismatch probability (avrg): 0.01
- Mismatch probability (end): 0.03

* Mismatch probabilities are assessed by a pre-processing script which conducts a quick read-subset mapping for an approximate mismatch inference (refer to DiTASiC manual)

Merge ‘simulated set’ with ‘original set’:

Total number of reads (original CAMI set): N.org = 149,136,946

Factor = N.org / (N.org ~ N.sim) = 0.882

→ Relative abundance values (ground truth) of original CAMI reads are normalized by *Factor*

→ Relative abundance values (ground truth) for simulated reads created to sum up to (1-*Factor*) = 0.118

Taxa abundance list of the simulated 30 taxa (spiked into original CAMI set):

GenBank accession number	**Ground Truth: relative taxa abundance for sample 1 ~ 2**			
	Set 1	Set 1 – normalized values for Mason Simulation	Set 2	Set 2 – normalized values for Mason Simulation
GCA_900094705.1	0.005	0.04237288	0.005	0.04237288
GCF_000020965.1	0.01	0.08474576	0.005	0.04237288
GCF_000222305.1	0.0072	0.061016947	0.0072	0.061016947
GCF_000333455.1	0.003	0.025423728	0.0045	0.038135592
GCF_000385945.1	0.0015	0.012711864	0.0015	0.012711864
GCF_000428765.1	0.004	0.033898304	0.0023	0.019491525
GCF_000429685.1	0.013	0.110169488	0.013	0.110169488
GCF_000463735.1	0.003	0.025423728	0.0017	0.014406779
GCF_000470655.1	0.0055	0.046610168	0.0055	0.046610168
GCF_000471625.1	0.002	0.016949152	0.0048	0.040677965
GCF_000585495.1	0.0082	0.069491523	0.0082	0.069491523
GCF_000716525.1	0.001	0.008474576	0.0028	0.023728813
GCF_000817975.1	0.004	0.033898304	0.004	0.033898304
GCF_001298525.1	0.0033	0.027966101	0.0063	0.053389829
GCF_001402715.1	0.002	0.016949152	0.002	0.016949152
GCF_001418395.1	0.0066	0.055932202	0.003	0.025423728

GCF_001418715.1	0.004	0.033898304	0.004	0.033898304
GCF_001484195.1	0.0024	0.020338982	0.005	0.04237288
GCF_001485005.1	0.0015	0.012711864	0.0015	0.012711864
GCF_001514055.1	0.012	0.101694912	0.002	0.016949152
GCF_001514495.1	0.0044	0.037288134	0.0044	0.037288134
GCF_001544695.1	0.001	0.008474576	0.0027	0.022881355
GCF_001546055.1	0.003	0.025423728	0.003	0.025423728
GCF_001591345.1	0.0025	0.02118644	0.004	0.033898304
GCF_001591385.1	0.0013	0.011016949	0.0013	0.011016949
GCF_001592205.1	0.002	0.016949152	0.0015	0.012711864
GCF_001606025.1	0.0017	0.014406779	0.0017	0.014406779
GCF_001636425.1	0.0007	0.005932203	0.0023	0.019491525
GCF_001720585.1	0.001	0.008474576	0.0066	0.055932202
GCF_900044055.2	0.0012	0.010169523	0.0012	0.010169523
Sum:	0.118	1	0.118	1

3) Illumina 100 (*i100*) data set by Mende et al. (2012):

We applied the *i100* benchmark data set provided in the publication by Mende *et al.*, consisting of a total of 53.33 million single reads (~26.6 million paired reads) of 75 bp length following Illumina read characteristics. The reads are derived from 100 unique bacterial genomes and were originally simulated by the iMESSi metagenomics simulator.

Reads:
According to the publication, we retrieved the paired read sample 'illumina_100species.1.fq' and ''illumina_100species.2.fq' from the link: http://www.bork.embl.de/~mende/simulated_data/

Reference sequences:
We refer to Table2 *(Genomes Used in the Medium Complexity Metagenome and Estimated Coverage (100 genomes))* of the Supplementary Material of Mende *et al*. As stated in their description, the dataset includes all chromosomes of the genomes as well as all plasmids. Chromosome and additional plasmids sequences were retrieved according to the provided accessions for the i100 data available from http://www.bork.embl.de/~mende/simulated_data/bacterial_data.txt.

Ground Truth of Abundance Proportions:
We refer to a slightly corrected version of the i100 ground truth table provided by Schaeffer et al. (2017), named 'i100_truth.csv' available from https://github.com/pachterlab/metakallisto. The table follows the format s*pecies*, *abundance*, *counts*, and *genome size*. Thereby, '*counts*' corresponds to the column '*Est_proportion of total sequence'* of the table by Mende et al. with minor corrections. The given '*counts*' are used as ground truth (named *GT.counts*).

DiTASiC calculation
parameters used for the matrix calculation (default settings), defined parameters:
- Read length: 75 bp
- Mismatch probability (begin): 0.007
- Mismatch probability (avrg): 0.013
- Mismatch probability (end): 0.036

* Mismatch probabilities are assessed by a pre-processing script which conducts a quick read-subset mapping for an approximate mismatch inference (refer to DiTASiC manual)
Note: DiTASiC uses the reads as single end reads

kallisto calculation
kallisto quant command, only parameter: - l 75 (length)
Note: kallisto is run in paired end read mode

Evaluation

Parameter outputs	**kallisto** (*paired mode*)	**DiTASiC** (*single mode*)
n (number of taxa (exact genome level))	100	100
T (number of reads processed; see also in .json output file of kallisto)	26667004	53334008
A (number of aligned reads)	26202326	46516552
μ (true absolute counts, ground truth GT)	'*GT.counts*' (see description above, sum(GT.counts) = T)	[*GT.counts* / sum (*GT.counts*)] * T (scaled for the number of single reads)
t (absolute count estimate)	kallisto count estimates	DiTASiC count estimates

$$AVGRE = \frac{1}{n}\sum_{i}^{n}\frac{\left|t_i * \frac{T}{A} - \mu_i\right|}{\mu_i} \qquad RRMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}\left(\frac{\left|t_i * \frac{T}{A} - \mu_i\right|}{\mu_i}\right)^2}$$

$$SSE = \sum_{i=1}^{n}\left(\frac{t_i}{A} - \frac{\mu_i}{T}\right)^2$$

Evaluation values computed:

	Exact Genome level		
	AVGRE	RRMSE	SSE
DiTASiC	**0.86**	**2.19**	**8.23 e-06**
kallisto (reproduced)	1.09 *	5.38 *	5.62 e-05

Compare to Table 1 provided in the publication by Schaeffer *et al.:*

	Exact Genome level	
	AVGRE	RRMSE
kallisto	**0.97 ***	**5.42 ***
Bracken	-	-
CLARK	-	-
GASiC	7.21	19.31
eXpress	2.57	11.92

- CLARK and Bracken results are reported by Schaeffer *et al.* to be missing as "they do not output strain level counts."

* Evaluation values of kallisto reproduced in our computed i100 study and evaluation values given by Schaeffer *et al.* are shown to be very similar. The minor value differences observed might be explained by minor changes in reference sequences of new or older versions available in NCBI. (NCBI download of this study: 03/15/2017)

Evaluation of overdispersion in the GLM models:

In this manuscript, we focus on the bias of abundance estimates of individual taxa within one sample occurring due to highly similar genome sequences present. The generalized linear model (GLM) model aims to reduce this bias by resolving the shared read counts. Generalized linear models provide a flexible system which allows for different distribution models. Our proposed GLM is based on an identity link function, and the corresponding error in the model is assumed to follow a poisson distribution: one rationale is the discrete count setting and second is the assumption that after the correction for read ambiguities the bias should not exceed the variance of a poisson distribution.
The latter assumption is constantly reviewed by applying a 'Dispersion test' to the computed GLM model. The test is part of the R-package 'AER' and "tests the null hypothesis of equidispersion in Poisson GLMs against the alternative of overdispersion and/or underdispersion" (*Cameron, A.C. and Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. Journal of Econometrics*).
We applied the test to all 17 models computed for the simulation data sets, the Pignatelli data sets, the *i100* data and the CAMI sets:
In all cases, the null hypothesis holds using an alpha level of 0.05 and showing p-values of 1, no overdispersion is reported for the data sets considered.
The dispersion test is integrated in the code. If the test indicates overdispersion, the GLM model will be re-calculated using an identity link function and a 'quasipoisson' error term which enables modelling of overdispersion.