SUPPLEMENTARY INFORMATION for

Integration of short reads from multiple 16S rRNA regions generates high resolution microbial community reconstruction, by Amir et al.

SUPPLEMENTARY METHODS

SM.1 A short overview of theoretical results for COMPASS

We shortly describe the mathematical formulation of COMPASS and two theoretical results, that are described in detail and proven in (1).

Probabilistic setting: The probabilistic formulation of the problem is given by first normalizing the entries in each column of $\boldsymbol{A}$ by the corresponding sequence length. We denote the resulting matrix by $\boldsymbol{\hat{A}}$ hence $\boldsymbol{\hat{A}}$ij represents the probability of observing k-mer *i* in sequence *j*. We then provide a probabilistic generative model for the distribution of the measurement vector **y**. This read sampling distribution, which we note by $p_{\hat{A}}(\boldsymbol{x})$, is a function of $\boldsymbol{\hat{A}}$ and of the mixture's true vector **x**. The equivalence between the probabilistic setting and the one provided in this manuscript is straightforward.

Reconstruction Guarantees: A fundamental question is to what extent does the information provided by sampled short reads allows one to correctly reconstruct the species identities and frequencies. We first defined and studied species '*identifiability'*, namely the ability to correctly identify the species present in a mixture and estimate their frequencies when the number of reads and our computational power are unlimited. Understanding the conditions for identifiability is important, since when identifiability does not hold, profiling will always be ambiguous having some species' frequencies undetermined, regardless of the data analysis method used even for an infinite number of errorless reads. Mathematically, the condition for identifiability is that each mixture frequency vector **x** has a unique fingerprint, *i.e*. the distribution $p_{\hat{A}}(\mathbf{x})$ is uniquely determined by the vector **x**. This is achieved if the matrix **Â** is of full rank.

Identifiability is determined by both the similarity between sequences in the database and by the read length. The longer and more diverse are the sequenced regions, the more information they provide on the DNA sequence of different species in the mixture. We show that when the read length is long enough, the problem is identifiable. For realistic read lengths (e.g. 100nt), and for the Greengenes database used, it turns out that identifiability does not hold, *i.e*. there are certain species whose frequencies cannot be uniquely recovered. However, this has no severe consequences. We defined and analysed *'partial identifiability'* for the Greengenes database. Even if the entire frequency vector $\boldsymbol{x}$ cannot be uniquely determined from the read sampling distribution $p_{\hat{A}}(\boldsymbol{x})$, we can still recover correctly the frequencies x_j for specific species *j*, under specific algebraic condition which can be checked for each species (see (1)). It turns out that for the Greengenes database and realistic read lengths, the vast majority of species frequencies can be identified uniquely and profiling is ambiguous only for a small minority. For example, about 98.5% of the sequences in the Greengenes database used are uniquely identifiable for a read length of 100nt, and only the remaining 1.5% sequences would need longer read lengths in order to be distinguishable.

We next studied the reconstruction error as a function of the number of reads available. More specifically we proved upper-bounds on the difference between the correct vector $\boldsymbol{x}$ and the solution of Eq. 1 in the main text. The bounds depend on the similarity between sequences and the read length, manifested by the matrix $\boldsymbol{\hat{A}}$, and on the number of reads. Intuitively, we would expect reconstruction error to be lower as the number of reads increased, and as the similarity between the database sequences is decreased. Similar database sequences are expected to be harder to distinguish and lead to higher reconstruction error. We formulated this intuition mathematically, by proving an upper-bound on the mean squared error between the true and reconstructed vectors which is inversely proportional to the number of reads R, and to the smallest eigenvalue of the matrix $\mathbf{\hat{A}}^{\mathbf{T}}\mathbf{\hat{A}}$. For sequences showing high similarity, the eigenvalue will be small, hence more reads are needed for accurate reconstruction (at the extreme case, when the problem is not identifiable, the smallest eigenvalue is zero and the bound is meaningless).

More details, including additional results for other error measures, are available in (1).

SM.2 Performance exploration using extensive simulations

Simulating a mixture: Each simulation is characterized by the mixture parameters, the database specifications and by 'sequencing' parameters. A mixture is characterized by the number of bacteria *n*, and their frequency distribution, while the database *S* was either taken as the full 16S rRNA gene or the specific 750bp region that was actually amplified in our experiments. The 'sequencing' parameters were given by the number of reads *R* and their length *k*.

A simulation was performed by randomly selecting *n* bacteria from *S* and assigning their frequencies according to the chosen distribution. Reads were then simulated by selecting one bacterium from the mixture according to its frequency and then randomly selecting a read of length *k* from its sequence in *S* using a uniform distribution. This read is then subject to read errors according to a model described below. The COMPASS algorithm receives this set of reads together with *S* and outputs the reconstructed vector stating the estimated frequency of each bacteria in *S*. Unless mentioned otherwise, the algorithm used the read correction procedure described in the main text.

Error model for Illumina reads: To introduce read errors in simulated Illumina reads we used the error models in (2,3), which take into account position-dependent base substitutions errors with error probability exponentially increasing along the read. More specifically, the following confusion matrix is defined:

	A	C	G	T
A	1-p	0.3*p	0.22*p	0.18*p
C	0.5*p	1-p	0.22*p	0.6*p
G	0.35*p	0.15*p	1-p	0.22*p
T	0.15*p	0.55*p	0.56*p	1-p

where p depends on the location *l* along a read of length *k* using the following formula:

$$p(l) = E_0 e^{\alpha(l-1)}, \ \alpha = \log(E_k / E_1) / (k-1)$$

And E_1 and E_k correspond to the error probability in the first base and last positions, respectively. The values used were $E_1 = 0.5\%$ and $E_k = 3\%$.

SM.3 Phylogenetic tree building

Phylogenetic trees presented in Figures 5, S3 and S6 were calculated and plotted using standard MATLAB functions. Sequence pairwise distances were calculated using the Jukes-Cantor algorithm allowing pairwise alignment (seqpdist.m function). A tree was then constructed from these pairwise distances using single linkage method (seqlinkage.m function).

SUPPLEMENTARY RESULTS

SR.1 Profiling low frequency bacteria

We performed two sets of simulations where we replaced the 0.1% minimal frequency threshold value by 0.025% and used blocks of 4000 bacteria instead of 1000 bacteria in the divide and conquer step. The first set of simulations repeated the case of n=1000 bacteria (Figure 3 panel C). In the second scenario we simulated 500 bacteria having a power law distribution (1/x), in which case the frequency of more than 350 bacteria falls below 0.1% and the minimal frequency is about 0.03%. Results are significantly better than for using the former parameters. Changing the parameters increased simulation run times of each divide and conquer block by a factor of about 4 (see Materials and Methods). This presents a trade-off between run time and the minimal detection frequency, which can be 'tuned' depending on the specific experimental application.

	mixture of 500 bacteria frequency: power law		mixture of 1000 bacteria frequency: uniform	
	MM 0%	MM 2%	MM 0%	MM 2%
Weighted Precision	89%±3%	99%±1%	88%±2%	99%±1%
Weighted Recall	92%±2%	97%±1%	91%±1%	97%±1%

Table S1: Simulating cases of lower frequency bacteria using COMPASS. A lower threshold and a larger block size were applied.

SR.2 Experimental profiling of Fly samples: COMPASS vs. MG-RAST

Results of the Drosophila experiments are presented in Figure S3 in the same format described for 454-BLAST in the main text (Figure 5), hence its description is omitted.

Concordance between COMPASS and SRF at the MG-RAST finest resolution: The comparison shows a high agreement between the two methods, as evident from the matrix, detecting the same 4 main clusters of bacteria (*Wolbachia*, *Lactobacillus*, *Acetobacter* and 'Other' which comprised mostly *Pseudomonas*). A closer look into the matrix shows that all 23 sequences found by COMPASS had counterparts found by MG-RAST. 12 out of the 23 sequences found by COMPASS had identical counterparts by MG-RAST, while the other 11 sequences differed by up to 4nt. MG-RAST found 57 bacteria, many of which were later found to be false positives. Out of the 8 sequences found by MG-RAST, that differed from the COMPASS sequences by 7-16nt, 3 sequences from the Pseudomonas cluster had frequency of about 0.1-0.2% in MG-RAST and also appear in only one of the four samples. The other 5 cases occurred in the Wolbachia cluster – 2 of these were validated as false positives and the correctness of others could not be resolved (see next section).

COMPASS displays increased phylogenetic resolution: Each black entry in the matrix in Figure S3 which corresponds to a complete match between the methods, also manifests COMPASS's increased phylogenetic resolution compared to SRF. In all 12 complete matches between COMPASS and MG-RAST, COMPASS achieved higher resolution.

COMPASS displays less false positive detections than MG-RAST: See next section

SR.3 Validation via Sanger sequencing – comparison with MG-RAST results

Summary of results displays reduction in false-positively detected bacteria: Results for sample L2 are summarized in Figure S4 in the same format as Figure 6, thus we refer here only to MG-RAST. The COMPASS column is the same as in Figure 6. Results are also summarized in Table 3.

Wolbachia (Figure S4 left): The dominant high abundance group of sequences (W1) was selected by MG-RAST although it contained 11 bacteria, where only 3 bacteria which were the found by COMPASS were 'Sanger validated', displaying the improved phylogenetic resolution. MG-RAST predicted the existence of 4 additional bacteria, W3-W6 (total frequency of 5.4%), which were all false positives. Two more groups (AB025965.1 and EU137480.1) were not amplified by the Sanger primers, thus can not be deciphered.

Acetobacter (Figure S4 right): MG-RAST predicted the same 4 groups that were predicted by COMPASS and were all validated. However, group A1 contained 25 bacteria out of which the 18 bacteria shared by the COMPASS method were validated and other bacteria were found to be incorrect, which is another manifestation of increased resolution provided. The additional MG-RAST predicted group A5 having frequency of 4.9% contained 10 bacteria that were false positives.

SR.4 Computational resources

Table S2 A-C displays time and memory usage of COMPASS and EMIRGE. Experiments were performed using a virtual Linux (64 bit Ubuntu 12.04.2 LTS) machine with 6 cores and 37GByte of RAM. This virtual machine was hosted on a Dell PowerEdge R610 running VMware vSphere ESXi 5.0 with 48Gbytes of RAM and 2 Intel Xeon X5550 running at 2.659Ghz..The machine was solely dedicated to COMPASS and EMIRGE, running a single case at a time allowing usage of all 6 CPUs. We measured the time of each run and the peak memory usage (namely of all 6 cores). Results are presented when varying the number of reads, the read length and the number of bacteria in exactly the same settings as in Figure 3 and Figures S1 and S2. Each scenario was tested once, hence some variability in time estimates may occur.

EMIRGE results: Running times increase dramatically with increasing the number of reads and the number of bacteria, and also when decreasing the read length (memory exceeded the machine's memory for read lengths 35nt and 50nt). EMIRGE was applied while enabling parallel usage of bowtie in all 6 cores, which reduced its running times.

COMPASS results: Memory was independent of number of reads, read length and number of bacteria, either with or without correcting for read errors - each of the 6 CPUs used about 3.5 Gigabytes of RAM. In contrast, time-wise there were significant differences when applying read error correction in COMPASS.

Run time in COMPASS without read error correction seems to be independent of the number of reads, it slightly increases with the number of bacteria. In most cases time was in the order of 1-2 hours (in the case of a short read length of 35nt the running time was 3.5 hours). Time seems to be significantly shorter than of EMIRGE when number of reads is larger than a few hundreds of thousands and for read lengths shorter than 200nt and for any number of bacteria.

Running times of COMPASS with read error correction are significantly longer than EMIRGE in almost all scenarios. When the number of reads is 500,000 or 10^6 COMPASS takes around 11 hours. Also, a longer read length contains more errors that need to be corrected for, hence time increases with the read length, as opposed to EMIRGE and to COMPASS without read error correction for which time decreased with read length. When increasing the number of bacteria running times reach about 35 hours for 1000 bacteria.

Summary: Given the limited effect of read errors, read error correction may be ignored, without significant loss in performance (see e.g. Figure S8). This would dramatically reduce running times, to a regime well below the EMIRGE run times. However, in case sequencing results contain a large number of read errors and correction is required, one can simply employ a larger number of cores to reduce run time.

Remark regarding SRF using BLAST: Run times of 16S-V4 depend on the specific BLAST implementation, and are proportional to the number of unique reads. It is completely parallelizable since reads may be split among the cores. For example, processing a 1000 unique reads of 16S-V4 takes about 15 minutes using the MATLAB based BLAST over the 6 cores. Hence an experiment with 10,000 reads would take about 2.5 hours to analyse on our Linux machine.

Tables S2

	Time (Hours)			Memory (Gigabytes)	
Number of reads	COMPASS with correcting for read errors	COMPASS without correcting for read errors	EMIRGE	COMPASS	EMIRGE
10,000	0.85	1	0.03	19	0.24
20,000	1	1	0.08	19.7	0.24
50,000	1	0.7	0.16	19.1	0.4
100,000	1.3	0.7	0.35	19.4	0.7
500,000	7.3	1.23	3.5	20.1	4
1,000,000	11	0.83	8	20.5	9.6

Table S2-A, Run times and memory usage as a function of the number of reads. Simulations were performed using 200 bacteria, and read length of 100nt.

Read length	Time (Hours)			Memory (Gigabytes)	
	COMPASS with correcting for read errors	COMPASS without correcting for read errors	EMIRGE	COMPASS	EMIRGE
35	7	3.5	-	20.1	-
50	5.7	2.1	-	20.1	-
75	8.8	1.3	14	20.3	13.4
100	11	0.83	8	20.5	9.6
150	18	0.75	2.8	20.7	2.8
200	24	0.7	1.75	21.3	2.4

Table S2-B, Run times and memory usage as a function of the read length. Simulations were performed using 200 bacteria, and 10^6 reads. EMIRGE runs for read lengths 35nt and 50nt failed due to memory demands.

	Time (Hours)			Memory (Gigabytes)	
number of bacteria	COMPASS with correcting for read errors	COMPASS without correcting for read errors	EMIRGE	COMPASS	EMIRGE
10	8.7	0.75	0.8	20.6	4.6
100	22	1.23	6.2	20.5	6.3
200	25.5	1.3	10	20.3	12.4
400	32	1.6	12	20.3	11.6
600	29.5h	1.5	17.7	21	12.3
1000	34.7	1.8	19.5	20.5	15

Table S2-C, Run times and memory usage as a function of the number of bacteria. Simulations were performed using 10^6 reads of length of 100nt.

SUPPLEMENTARY FIGURES

Figure S1: Extensive simulations – EMIRGE performance as a function of several parameters

Panels (A)-(C) present weighted recall and precision as a function of different variables. Unless stated otherwise, parameters are as follows: each simulation contained 200 bacteria, randomly selected from the (full length) 16S database, with relative frequencies following a power law distribution (1/x). The read length was 100nt, and 10^6 reads were simulated. The simulated mixtures were exactly those simulated in Figure 3, although reads were not subject to errors. The blue and red lines denote weighted recall and precision respectively; a solid line refers to the case in which complete sequence identity is required, while a dashed line refers to the case where up to 2% differences in sequence are acceptable.

(A.) Effect of number of reads, with read number changed from 10^6 down to 10^4, and other parameters as above.

(B.) Effect of read length, with read length varying between 75 and 200, and other parameters as above. Due to memory limitations read lengths 35 and 50 were not tested.

(C.) Effect of number of bacteria in the original mixture. Here for each number of bacteria *n*, all bacterial frequencies are set to 1/*n*. Other parameters are as above

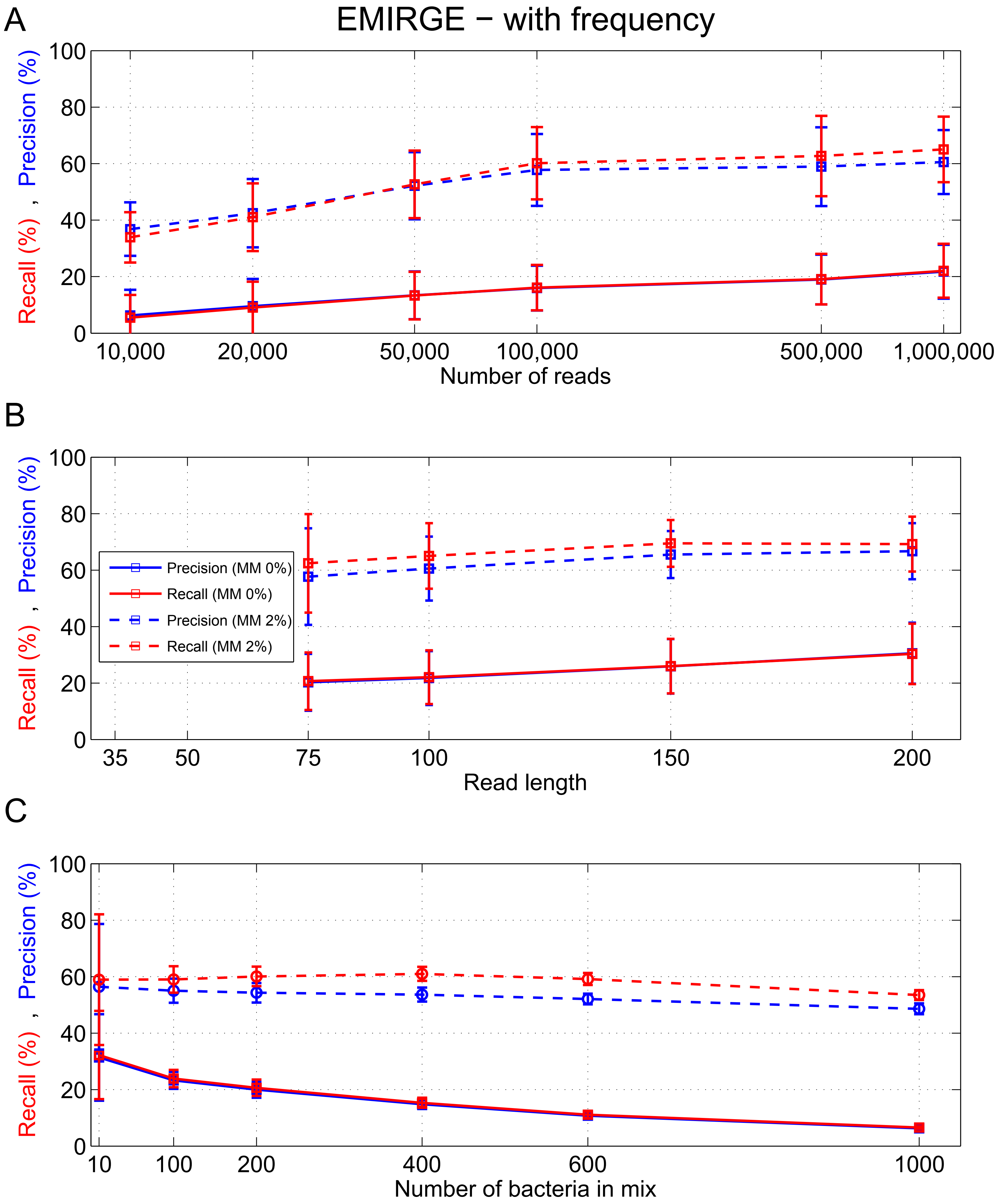

A
EMIRGE − with frequency
Recall (%) , Precision (%)
0
20
40
60
80
100
10,000
20,000
50,000
100,000
500,000
1,000,000
Number of reads
B
Recall (%) , Precision (%)
0
20
40
60
80
100
35
50
75
100
150
200
Read length
Precision (MM 0%)
Recall (MM 0%)
Precision (MM 2%)
Recall (MM 2%)
C
Recall (%) , Precision (%)
0
20
40
60
80
100
10
100
200
400
600
1000
Number of bacteria in mix

FIGURE S1

Figure S2: Extensive simulations – EMIRGE performance as a function of several parameters when frequency criterion is not enforced

The same as Figure S1, apart from the fact that the frequency criterion was not enforced.

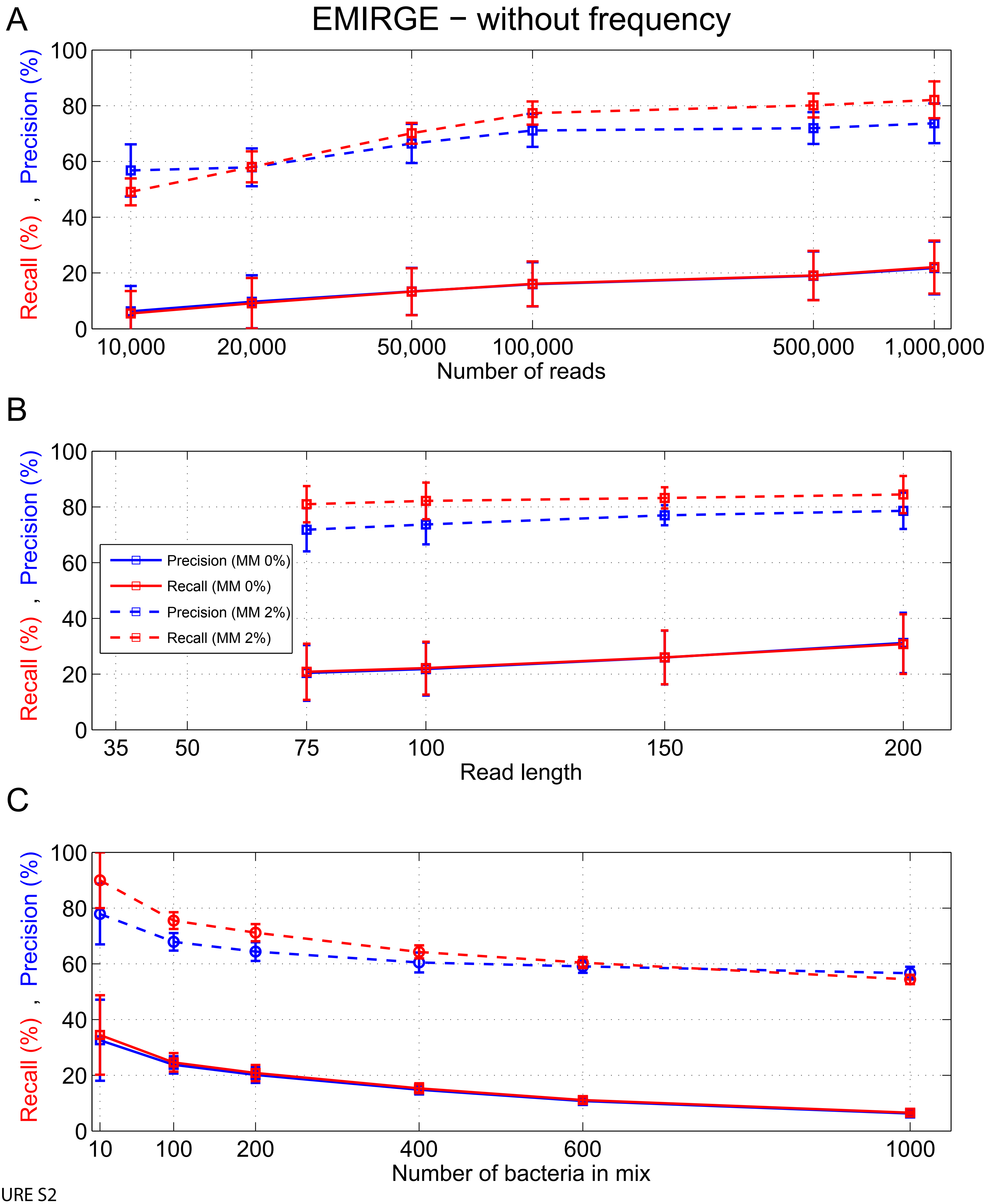

A
EMIRGE − without frequency
Recall (%) , Precision (%)
0
20
40
60
80
100
10,000
20,000
50,000
100,000
500,000
1,000,000
Number of reads
B
Recall (%) , Precision (%)
Precision (MM 0%)
Recall (MM 0%)
Precision (MM 2%)
Recall (MM 2%)
35
50
75
100
150
200
Read length
C
Recall (%) , Precision (%)
10
100
200
400
600
1000
Number of bacteria in mix

FIGURE S2

Figure S3: Reconstruction of Drosophila samples – COMPASS and MG-RAST

Results comparing MG-RAST SRF based on Roche 454 and Illumina-based COMPASS framework on four Drosophila samples, L1, L2, E1 and E2. Results are presented in the same format as in Figure 5 hence its description is omitted.

FIGURE S3

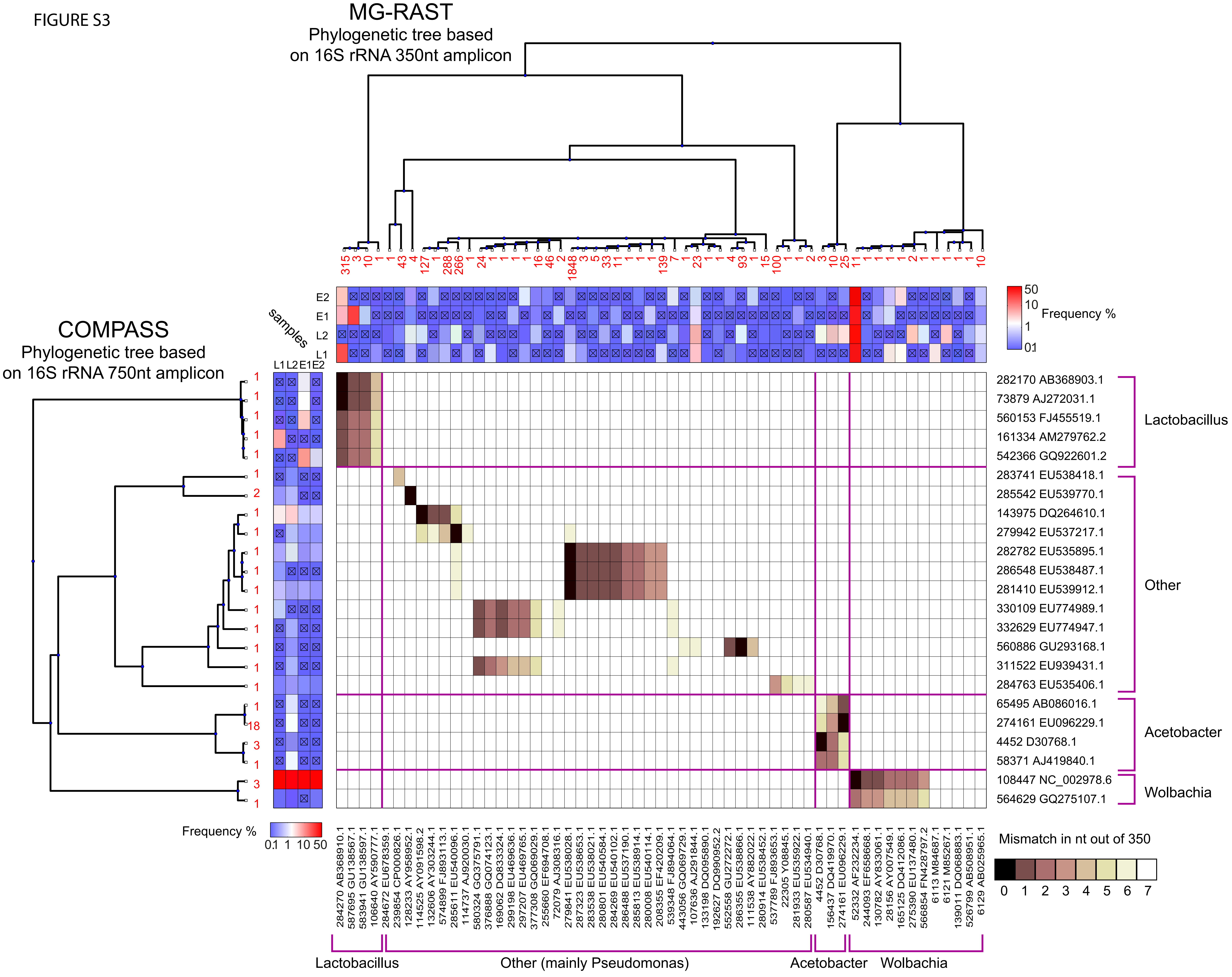
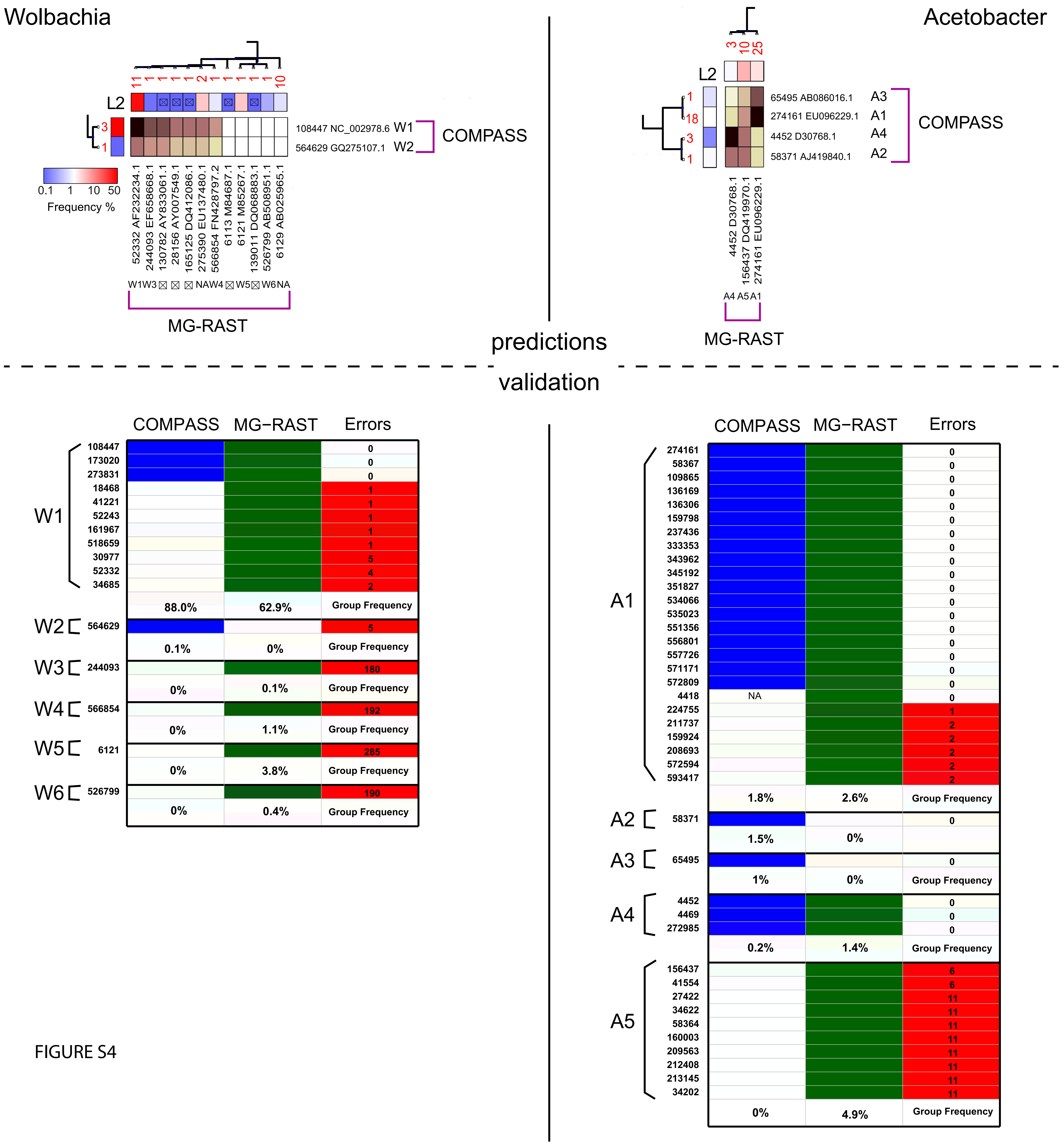

MG-RAST
Phylogenetic tree based
on 16S rRNA 350nt amplicon
315 3 10 1 1 43 4 127 1 288 266 1 24 1 1 1 1 16 46 2 1848 3 5 33 11 1 1 1 139 7 1 23 1 1 4 93 1 15 100 1 1 2 3 10 25 11 1 1 1 1 2 1 1 1 1 1 10
E2
E1
L2
L1
samples
50
10
1
0.1
Frequency %
COMPASS
Phylogenetic tree based
on 16S rRNA 750nt amplicon
L1 L2 E1 E2
1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 18 3 1 3 1
282170 AB368903.1
73879 AJ272031.1
560153 FJ455519.1
161334 AM279762.2
542366 GQ922601.2
283741 EU538418.1
285542 EU539770.1
143975 DQ264610.1
279942 EU537217.1
282782 EU535895.1
286548 EU538487.1
281410 EU539912.1
330109 EU774989.1
332629 EU774947.1
560886 GU293168.1
311522 EU939431.1
284763 EU535406.1
65495 AB086016.1
274161 EU096229.1
4452 D30768.1
58371 AJ419840.1
108447 NC_002978.6
564629 GQ275107.1
Lactobacillus
Other
Acetobacter
Wolbachia
Frequency %
0.1 1 10 50
284270 AB368910.1
587695 GU138567.1
583941 GU138597.1
106640 AY590777.1
284672 EU678359.1
239854 CP000826.1
128233 AY958952.1
114525 AY091598.2
132606 AY303244.1
574899 FJ893113.1
285611 EU540096.1
114737 AJ920030.1
580324 GQ375791.1
376888 GQ074123.1
169062 DQ833324.1
299198 EU469636.1
297207 EU469765.1
377308 GQ069029.1
255660 EF694708.1
72079 AJ308316.1
279841 EU538028.1
287323 EU538653.1
283538 EU538021.1
280801 EU540584.1
284269 EU540102.1
286488 EU537190.1
285813 EU538914.1
280008 EU540114.1
208355 EF420209.1
539348 FJ894064.1
443056 GQ069729.1
107636 AJ291844.1
133198 DQ095890.1
192627 DQ990952.2
552558 GU272272.1
286355 EU538866.1
111538 AY882022.1
280914 EU538452.1
537789 FJ893653.1
22305 Y08845.1
281933 EU535922.1
280587 EU534940.1
4452 D30768.1
156437 DQ419970.1
274161 EU096229.1
52332 AF232234.1
244093 EF658668.1
130782 AY833061.1
28156 AY007549.1
165125 DQ412086.1
275390 EU137480.1
566854 FN428797.2
6113 M84687.1
6121 M85267.1
139011 DQ068883.1
526799 AB508951.1
6129 AB025965.1
Lactobacillus
Other (mainly Pseudomonas)
Acetobacter
Wolbachia
Mismatch in nt out of 350
0 1 2 3 4 5 6 7

Figure S4: Validation via Sanger sequencing – MG-RAST and COMPASS

The left and right sides in the upper part of the figure display a zoom in to the *Wolbachia* and *Acetobacter* regions of Figure S3 for samples L2, stating the predicted groups by MG-RAST and COMPASS. Results are presented in the same format as in Figure 6, hence its description is omitted. Bacteria are grouped according to Figure S3, for both COMPASS and MG-RAST predictions.



		COMPASS	MG−RAST	Errors
W1	108447	🟦	🟩	0
	173020	🟦	🟩	0
	273831	🟦	🟩	0
	18468		🟩	1
	41221		🟩	1
	52243		🟩	1
	161967		🟩	1
	518659		🟩	1
	30977		🟩	5
	52332		🟩	4
	34685		🟩	2
		88.0%	62.9%	Group Frequency
W2	564629	🟦		5
		0.1%	0%	Group Frequency
W3	244093		🟩	180
		0%	0.1%	Group Frequency
W4	566854		🟩	192
		0%	1.1%	Group Frequency
W5	6121		🟩	285
		0%	3.8%	Group Frequency
W6	526799		🟩	190
		0%	0.4%	Group Frequency

		COMPASS	MG−RAST	Errors
A1	274161	🟦	🟩	0
	58367	🟦	🟩	0
	109865	🟦	🟩	0
	136169	🟦	🟩	0
	136306	🟦	🟩	0
	159798	🟦	🟩	0
	237436	🟦	🟩	0
	333353	🟦	🟩	0
	343962	🟦	🟩	0
	345192	🟦	🟩	0
	351827	🟦	🟩	0
	534066	🟦	🟩	0
	535023	🟦	🟩	0
	551356	🟦	🟩	0
	556801	🟦	🟩	0
	557726	🟦	🟩	0
	571171	🟦	🟩	0
	572809	🟦	🟩	0
	4418	NA	🟩	0
	224755		🟩	1
	211737		🟩	2
	159924		🟩	2
	208693		🟩	2
	572594		🟩	2
	593417		🟩	2
		1.8%	2.6%	Group Frequency
A2	58371	🟦		0
		1.5%	0%	
A3	65495	🟦		0
		1%	0%	Group Frequency
A4	4452	🟦	🟩	0
	4469	🟦	🟩	0
	272985	🟦	🟩	0
		0.2%	1.4%	Group Frequency
A5	156437		🟩	6
	41554		🟩	6
	27422		🟩	11
	34622		🟩	11
	58364		🟩	11
	160003		🟩	11
	209563		🟩	11
	212408		🟩	11
	213145		🟩	11
	34202		🟩	11
		0%	4.9%	Group Frequency

FIGURE S4

Figure S5: Comparison Roche 454 and Illumina - human saliva

Experimental results comparing 454-BLAST SRF and Illumina based COMPASS on four human saliva samples. The format is identical to that of Figure 4. On the left a phylogenetic tree based on sequences inferred by COMPASS (with frequency higher than 1%), where a similar tree based on the 454-BLAST is shown on top (with frequency higher than 0.5%). Database accession names are shown on the left and below respectively (further details appear in Supplementary Datasets 4-5). The matrix in the center displays the similarity between sequences found by 454-BLAST and COMPASS, calculated over the mutual 350bp long sequence. Complete identity in shown in black, while 7 or more mismatches appear as white. The block structure displays the high similarity between the results of the two methods. 58 out of 82 COMPASS inferred bacteria are identical to 454-BLAST inferred bacteria. As opposed to the Drosophila samples, a rather large portion of 454-BLAST inferred bacteria were not found by COMPASS (see rightmost part of the matrix). The same phenomenon occurs, although to a smaller extent, for bacteria found by COMPASS. Both these effects are due to bacteria that were amplified by the relevant primer pair of one method but not with the primer pair of the other method, thus not allowing for correct comparison between methods.

The figure is better viewed on screen.

FIGURE S5

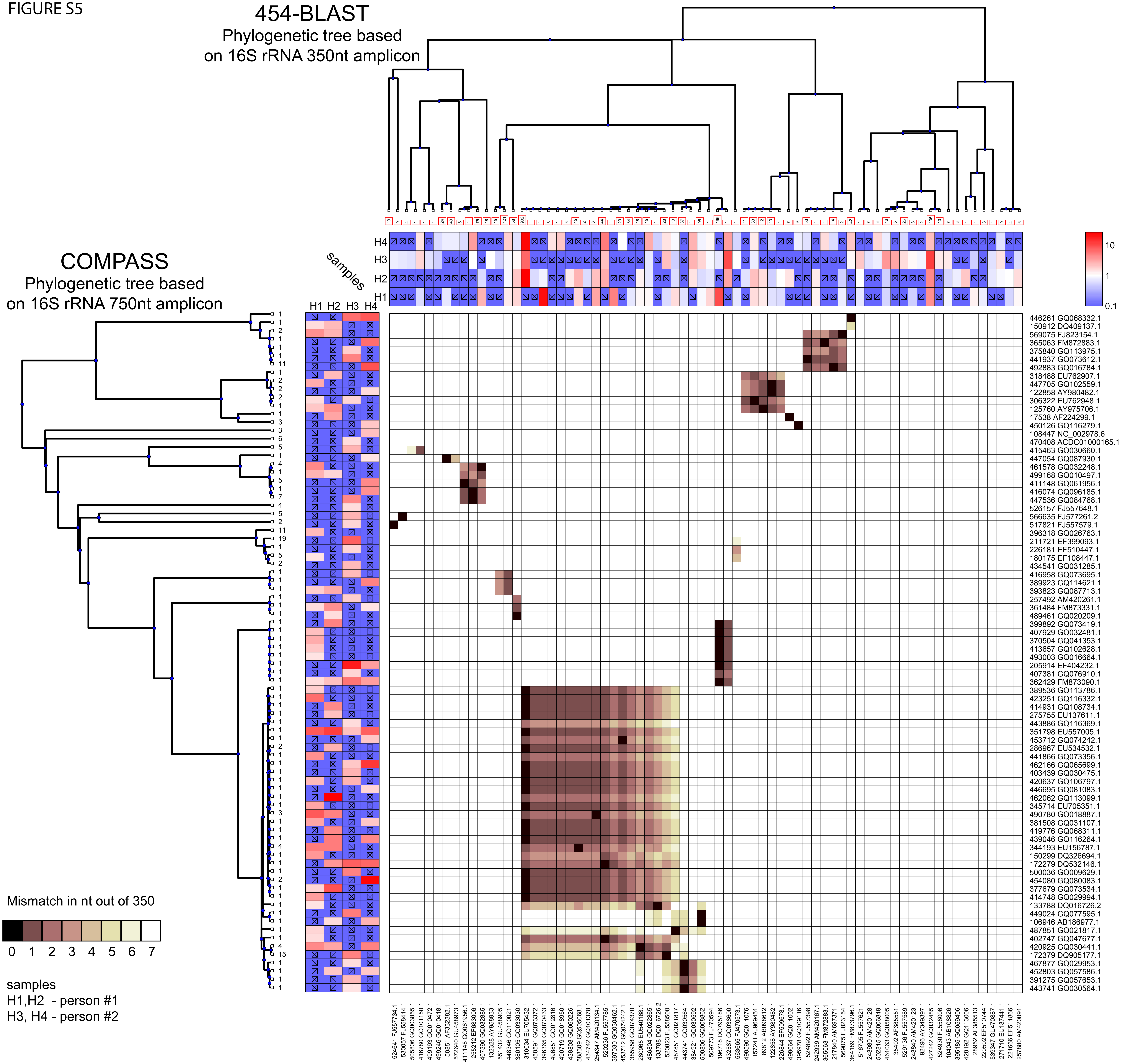

454-BLAST
Phylogenetic tree based
on 16S rRNA 350nt amplicon
COMPASS
Phylogenetic tree based
on 16S rRNA 750nt amplicon
samples
H1 H2 H3 H4
10
1
0.1
Mismatch in nt out of 350
0 1 2 3 4 5 6 7
samples
H1,H2 - person #1
H3, H4 - person #2
446261 GQ068332.1
150912 DQ409137.1
569075 FJ823154.1
365063 FM872883.1
375840 GQ113975.1
441937 GQ073612.1
492883 GQ016784.1
318488 EU762907.1
447705 GQ102559.1
122858 AY980482.1
306322 EU762948.1
125760 AY975706.1
17538 AF224299.1
450126 GQ116279.1
108447 NC_002978.6
470408 ACDC01000165.1
415463 GQ030660.1
447054 GQ087930.1
461578 GQ032248.1
499168 GQ010497.1
411148 GQ061956.1
416074 GQ096185.1
447536 GQ084768.1
526157 FJ557648.1
566635 FJ577261.2
517821 FJ557579.1
396318 GQ026763.1
211721 EF399093.1
226181 EF510447.1
180175 EF108447.1
434541 GQ031285.1
416958 GQ073695.1
389923 GQ114621.1
393823 GQ087713.1
257492 AM420261.1
361484 FM873331.1
489461 GQ020209.1
399892 GQ073419.1
407929 GQ032481.1
370504 GQ041353.1
413657 GQ102628.1
493003 GQ016664.1
205914 EF404232.1
407381 GQ076910.1
362429 FM873090.1
389536 GQ113786.1
423251 GQ116332.1
414931 GQ108734.1
275755 EU137611.1
443886 GQ116369.1
351798 EU557005.1
453712 GQ074242.1
286967 EU534532.1
441866 GQ073356.1
462166 GQ065699.1
403439 GQ030475.1
420637 GQ106797.1
446695 GQ081083.1
462062 GQ113099.1
345714 EU705351.1
490780 GQ018887.1
381508 GQ031107.1
419776 GQ068311.1
439046 GQ116264.1
344193 EU156787.1
150299 DQ326694.1
172279 DQ532146.1
500036 GQ009629.1
454080 GQ080083.1
377679 GQ073534.1
414748 GQ029994.1
133788 DQ016726.2
449024 GQ077595.1
106946 AB186977.1
487851 GQ021817.1
402747 GQ047677.1
420925 GQ030441.1
172379 DQ905177.1
467877 GQ029953.1
452803 GQ057586.1
391275 GQ057653.1
443741 GQ030564.1

Figure S6: Validation via Sanger sequencing - cartoon

Cartoon of validation process and chromatogram analysis. Family specific primer pair is applied to amplify the sample, resulting in a multi-peak chromatogram at specific locations. All nucleotides whose amplitude was at least 3% of the maximum were called present at each location. Example of two Sanger-validated sequences (1 and 2) and a sequence that was ruled out (3) is shown.

FIGURE S6

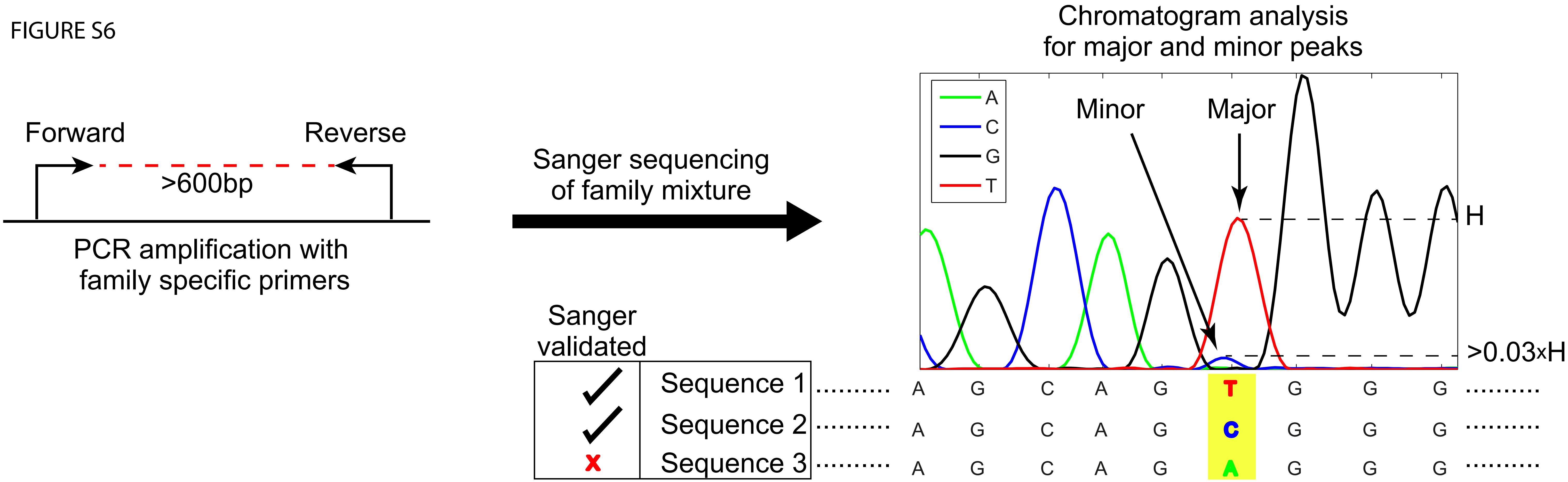

Forward
Reverse
>600bp
PCR amplification with family specific primers
Sanger sequencing of family mixture
Chromatogram analysis for major and minor peaks
A
C
G
T
Minor
Major
H
>0.03xH
Sanger validated
✓ Sequence 1 .... A G C A G T G G G ....
✓ Sequence 2 .... A G C A G C G G G ....
✗ Sequence 3 .... A G C A G A G G G ....

Figure S7: Sanger sequencing results - Drosophila samples

(A). Sanger sequencing of sample L2 using *Wolbachia*-related primer pair. Results for forward (panels A,C) and reverse (panels B,D) reactions. Each row corresponds to a bacterium potentially amplified by the Sanger primer pair (see Supplementary figure S5), presenting potential mismatches with the analyzed chromatogram. Gray corresponds to a match between the major chromatogram peak and the relevant bacteria, while white corresponds to the case where a match exists with a minor peak. We mark by black all positions where both major and minor chromatogram peaks differ from the bacteria's nucleotide. The rows are ordered in ascending number of errors. For example, the first 22 bacteria in panel A share the same sequence which matched the major peaks of the Sanger sequence. The 23rd bacterium ('4421') had two mismatches with the major peaks but matched the minor peaks in those locations. The 29th bacterium (`224775`), on the other hand, did not match the major or minor peaks at one location and was not 'Sanger-validated'.

To the left of each figure appears a summary of results for each sequence. We present 3 types of information - regarding the whole Sanger sequence, the overlap between the Sanger sequences and the COMPASS Illumina amplicon (750) and regarding the overlap with the 454-BLAST ampilcon (350). For each of these we present (from left to right) the bacteria accession number, the length of the overlapping region with the Sanger sequence, the number of mismatches between the bacterium and the major and minor peaks in the chromatogram. The accession numbers for the 750 and 350 cases are given by the relevant representative bacterium in Figure 3.

The lengths of the forward and reverse reactions were 456nt and 656nt, respectively, and 705nt in total.

(B) The same for *Acetobacter*. The lengths of the forward and reverse reactions were 411nt and 511nt, respectively, and 511nt in total.

The figure is better viewed on screen.

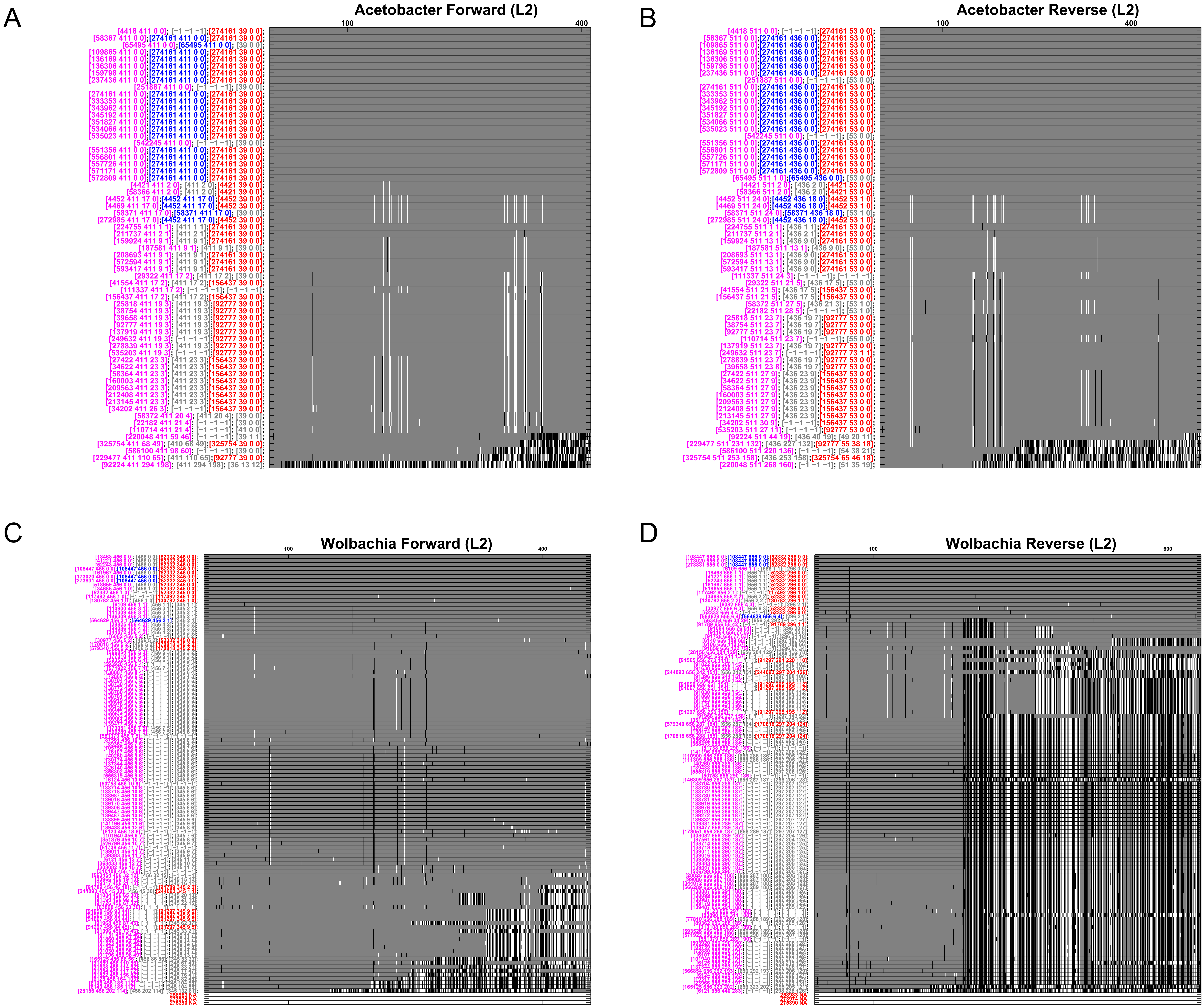

A
Acetobacter Forward (L2)
100
400
[4418 411 0 0]; [−1 −1 −1]; [274161 39 0 0];
[58367 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[65495 411 0 0]; [65495 411 0 0]; [39 0 0];
[109865 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[136169 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[136306 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[159798 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[237436 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[251887 411 0 0]; [−1 −1 −1]; [39 0 0];
[274161 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[333353 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[343962 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[345192 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[351827 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[534066 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[535023 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[542245 411 0 0]; [−1 −1 −1]; [39 0 0];
[551356 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[556801 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[557726 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[571171 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[572809 411 0 0]; [274161 411 0 0]; [274161 39 0 0];
[4421 411 2 0]; [411 2 0]; [4421 39 0 0];
[58366 411 2 0]; [411 2 0]; [4421 39 0 0];
[4452 411 17 0]; [4452 411 17 0]; [4452 39 0 0];
[4469 411 17 0]; [4452 411 17 0]; [4452 39 0 0];
[58371 411 17 0]; [58371 411 17 0]; [39 0 0];
[272985 411 17 0]; [4452 411 17 0]; [4452 39 0 0];
[224755 411 1 1]; [411 1 1]; [274161 39 0 0];
[211737 411 2 1]; [411 2 1]; [274161 39 0 0];
[159924 411 9 1]; [411 9 1]; [274161 39 0 0];
[187581 411 9 1]; [411 9 1]; [39 0 0];
[208693 411 9 1]; [411 9 1]; [274161 39 0 0];
[572594 411 9 1]; [411 9 1]; [274161 39 0 0];
[593417 411 9 1]; [411 9 1]; [274161 39 0 0];
[29322 411 17 2]; [411 17 2]; [39 0 0];
[41554 411 17 2]; [411 17 2]; [156437 39 0 0];
[111337 411 17 2]; [−1 −1 −1]; [−1 −1 −1];
[156437 411 17 2]; [411 17 2]; [156437 39 0 0];
[25818 411 19 3]; [411 19 3]; [92777 39 0 0];
[38754 411 19 3]; [411 19 3]; [92777 39 0 0];
[39658 411 19 3]; [411 19 3]; [92777 39 0 0];
[92777 411 19 3]; [411 19 3]; [92777 39 0 0];
[137919 411 19 3]; [411 19 3]; [92777 39 0 0];
[249632 411 19 3]; [−1 −1 −1]; [92777 39 0 0];
[278839 411 19 3]; [411 19 3]; [92777 39 0 0];
[535203 411 19 3]; [−1 −1 −1]; [92777 39 0 0];
[27422 411 23 3]; [411 23 3]; [156437 39 0 0];
[34622 411 23 3]; [411 23 3]; [156437 39 0 0];
[58364 411 23 3]; [411 23 3]; [156437 39 0 0];
[160003 411 23 3]; [411 23 3]; [156437 39 0 0];
[209563 411 23 3]; [411 23 3]; [156437 39 0 0];
[212408 411 23 3]; [411 23 3]; [156437 39 0 0];
[213145 411 23 3]; [411 23 3]; [156437 39 0 0];
[34202 411 26 3]; [−1 −1 −1]; [156437 39 0 0];
[58372 411 20 4]; [411 20 4]; [39 0 0];
[22182 411 21 4]; [−1 −1 −1]; [39 0 0];
[110714 411 21 4]; [−1 −1 −1]; [41 0 0];
[220048 411 59 46]; [−1 −1 −1]; [39 1 1];
[325754 411 68 49]; [410 68 49]; [325754 39 0 0];
[586100 411 98 60]; [−1 −1 −1]; [39 0 0];
[229477 411 110 65]; [411 110 65]; [92777 39 0 0];
[92224 411 294 198]; [411 294 198]; [36 13 12];
B
Acetobacter Reverse (L2)
100
400
[4418 511 0 0]; [−1 −1 −1]; [274161 53 0 0];
[58367 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[109865 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[136169 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[136306 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[159798 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[237436 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[251887 511 0 0]; [−1 −1 −1]; [53 0 0];
[274161 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[333353 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[343962 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[345192 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[351827 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[534066 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[535023 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[542245 511 0 0]; [−1 −1 −1]; [53 0 0];
[551356 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[556801 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[557726 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[571171 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[572809 511 0 0]; [274161 436 0 0]; [274161 53 0 0];
[65495 511 1 0]; [65495 436 0 0]; [53 0 0];
[4421 511 2 0]; [436 2 0]; [4421 53 0 0];
[58366 511 2 0]; [436 2 0]; [4421 53 0 0];
[4452 511 24 0]; [4452 436 18 0]; [4452 53 1 0];
[4469 511 24 0]; [4452 436 18 0]; [4452 53 1 0];
[58371 511 24 0]; [58371 436 18 0]; [53 1 0];
[272985 511 24 0]; [4452 436 18 0]; [4452 53 1 0];
[224755 511 1 1]; [436 1 1]; [274161 53 0 0];
[211737 511 2 1]; [436 2 1]; [274161 53 0 0];
[159924 511 13 1]; [436 9 0]; [274161 53 0 0];
[187581 511 13 1]; [436 9 0]; [53 0 0];
[208693 511 13 1]; [436 9 0]; [274161 53 0 0];
[572594 511 13 1]; [436 9 0]; [274161 53 0 0];
[593417 511 13 1]; [436 9 0]; [274161 53 0 0];
[111337 511 24 3]; [−1 −1 −1]; [−1 −1 −1];
[29322 511 21 5]; [436 17 5]; [53 0 0];
[41554 511 21 5]; [436 17 5]; [156437 53 0 0];
[156437 511 21 5]; [436 17 5]; [156437 53 0 0];
[58372 511 27 5]; [436 21 3]; [53 1 0];
[22182 511 28 5]; [−1 −1 −1]; [53 1 0];
[25818 511 23 7]; [436 19 7]; [92777 53 0 0];
[38754 511 23 7]; [436 19 7]; [92777 53 0 0];
[92777 511 23 7]; [436 19 7]; [92777 53 0 0];
[110714 511 23 7]; [−1 −1 −1]; [55 0 0];
[137919 511 23 7]; [436 19 7]; [92777 53 0 0];
[249632 511 23 7]; [−1 −1 −1]; [92777 73 1 1];
[278839 511 23 7]; [436 19 7]; [92777 53 0 0];
[39658 511 23 8]; [436 19 7]; [92777 53 0 0];
[27422 511 27 9]; [436 23 9]; [156437 53 0 0];
[34622 511 27 9]; [436 23 9]; [156437 53 0 0];
[58364 511 27 9]; [436 23 9]; [156437 53 0 0];
[160003 511 27 9]; [436 23 9]; [156437 53 0 0];
[209563 511 27 9]; [436 23 9]; [156437 53 0 0];
[212408 511 27 9]; [436 23 9]; [156437 53 0 0];
[213145 511 27 9]; [436 23 9]; [156437 53 0 0];
[34202 511 30 9]; [−1 −1 −1]; [156437 53 0 0];
[535203 511 27 11]; [−1 −1 −1]; [92777 53 0 0];
[92224 511 44 19]; [436 40 19]; [49 20 11];
[229477 511 231 132]; [436 227 132]; [92777 55 38 18];
[586100 511 220 136]; [−1 −1 −1]; [54 38 21];
[325754 511 253 158]; [436 253 158]; [325754 65 46 18];
[220048 511 268 160]; [−1 −1 −1]; [51 35 19];
C
Wolbachia Forward (L2)
100
400
288993 NA
273474 NA
275390 NA
D
Wolbachia Reverse (L2)
100
600
288993 NA
273474 NA
275390 NA

FIGURE S7

Figure S8: Correcting for read errors

A. Distribution of the difference between *weighted precision* when incorporating read error correction and without correcting for read errors. *Weighted precision* was calculated for the MM 0% case. The histogram is positively skewed displaying an advantage of applying read error correction in COMPASS. The red dashed line corresponds to the median difference.

B. The same for *weighed recall*. C-D the same as A-B using MM 2%. The effect of read error correction is less pronounced in these cases.

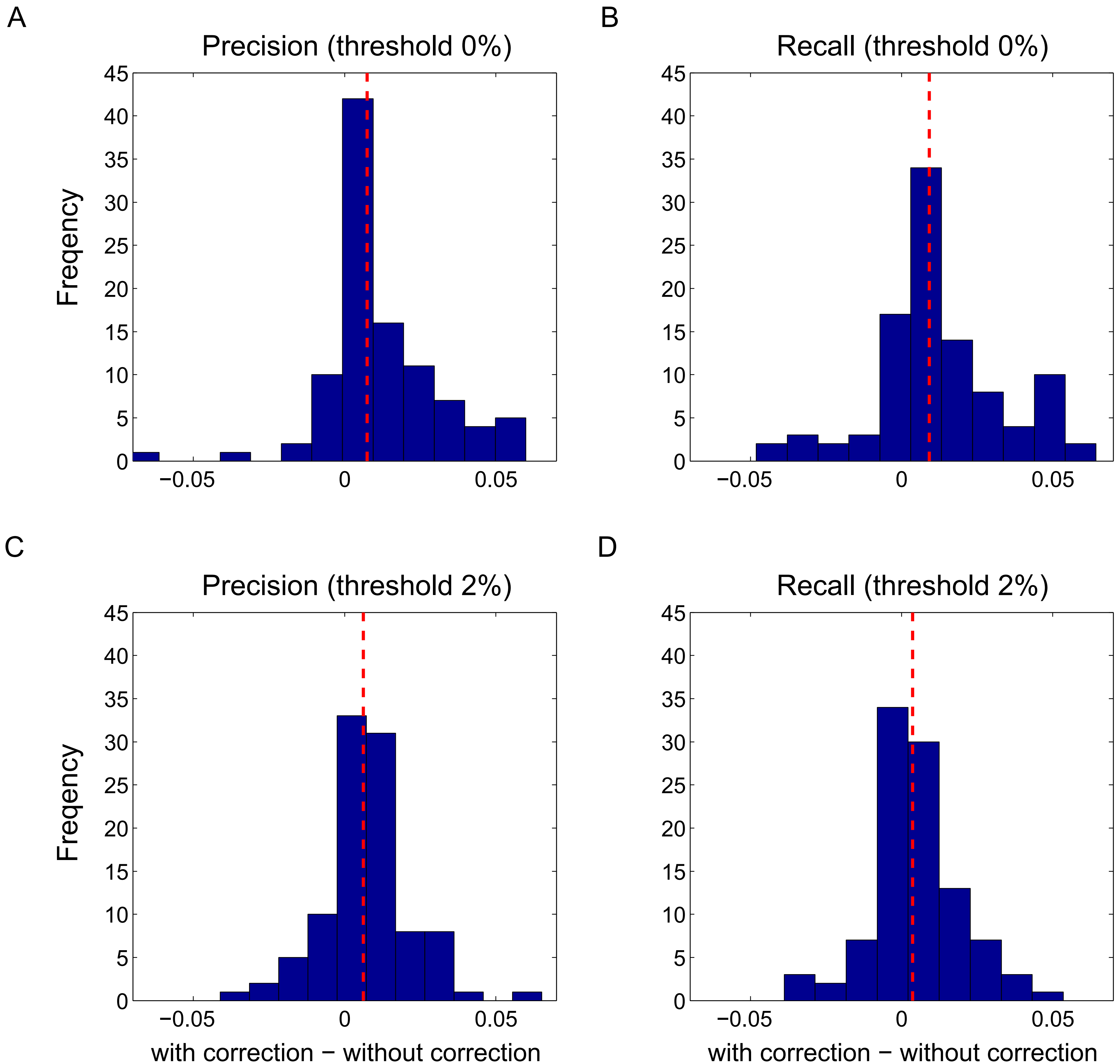

A
Precision (threshold 0%)
B
Recall (threshold 0%)
C
Precision (threshold 2%)
D
Recall (threshold 2%)
Freqency
Freqency
45
40
35
30
25
20
15
10
5
0
−0.05
0
0.05
with correction − without correction
with correction − without correction

FIGURE S8

Figure S9: Illumina and Roche 454 Primers

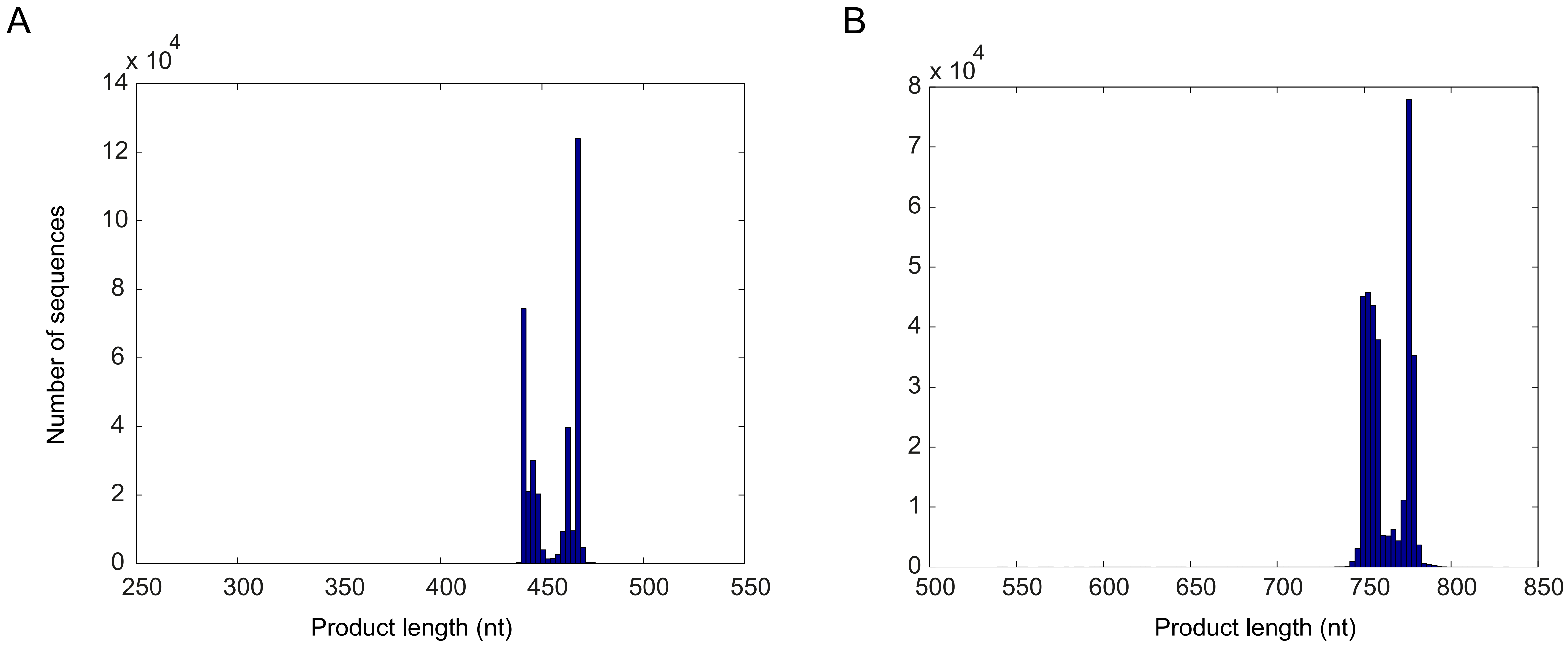
Figure S5: Primers used for Illumina and Roche 454 sequencing. Histograms show the distribution of product length for all potentially amplified 16S sequences out of the 455,055 sequence database.

(A) Histogram for the primers used for Roche 454 sequencing. These primers potentially amplify a product of approximately 450bp.

(B) same as (A) for primers used for Illumina sequencing, resulting in a product of approximately 750bp.

The table below summarizes the properties of these primers, presenting the median product length, the number of amplified sequences and the number of unique sequences. Although the number of amplified bacteria is quite similar, the number of unique sequences is about 30% higher in the 750bp primers. This indicates potential improvement in resolution achieved using the larger region.

FIGURE S9



	Forward	Reverse	Median length	# of amplified sequences	# of unique sequences
Illumina	CTCCTACGGGAGGCAGCAG	GGGTTGCGCTCGTTGCG	759	327,716	231,299
Roche-454	CTCCTACGGGAGGCAGCAG	GGACTACCAGGGTATCTAATCC	463	343,455	176,647

Figure S10: Gel of sonicated samples

Agarose gel of DNA samples following sonication. This figure presents length distribution of DNA prior to library preparation for Illumina sequencing. Following PCR amplification using primers producing an amplicon of 750bp (See Figure S5), samples were sonicated (Bioruptor Diagenode) between 80 and 100 cycles (30/30 second on/off), such that length distribution is on the range of 100-300bp. Panel A shows Drosophila larvae samples L1,L2, and eggs samples E1 and E2, and panel B shows human saliva samples H3, H4, H1 and H2.

FIGURE S10

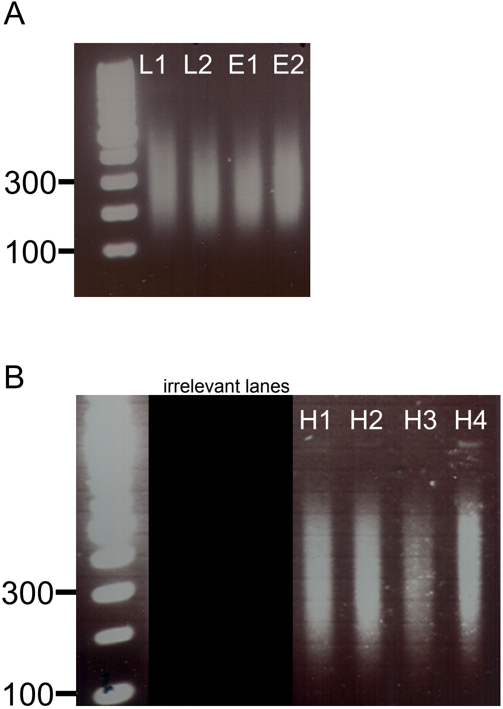


Figure S11: Experimental non uniform coverage and normalization

(A) Read distribution for sample L1 shows short and long-range bias. The number of 100-mers starting at each position is shown. In case a 100-mer appears in more than one bacterium in the 16S database, its average position was used. For clarity, data is shown for forward reads only. Unequal coverage is manifested both in short range bias (*i.e.* variability between neighboring nucleotides) and in long range bias (*i.e.* different average coverage for different regions).

(B) Global bias profile following 90-mer averaging. The coverage based on a running average over 20 bases is shown for forward (red) and reverse (cyan) reads of sample L1 following short-range bias normalization via representing each 100-mer as 11 sliding 90-mers. The running average ignored the first 10 nucleotides due to the large discontinuity in case of very short sonication. Pattern similarity between forward and reverse reads indicates that this bias originates from sonication effects.

(C) Read distribution for sample L1 following short and long-range bias corrections (as described in Methods section). The number of reads of each 90-mer was normalized according to the global bias profile value at the 90-mer position (averaged over all database sequences containing the 90-mer). The process was independently performed for forward and reverse reads, which were then combined.

FIGURE S11

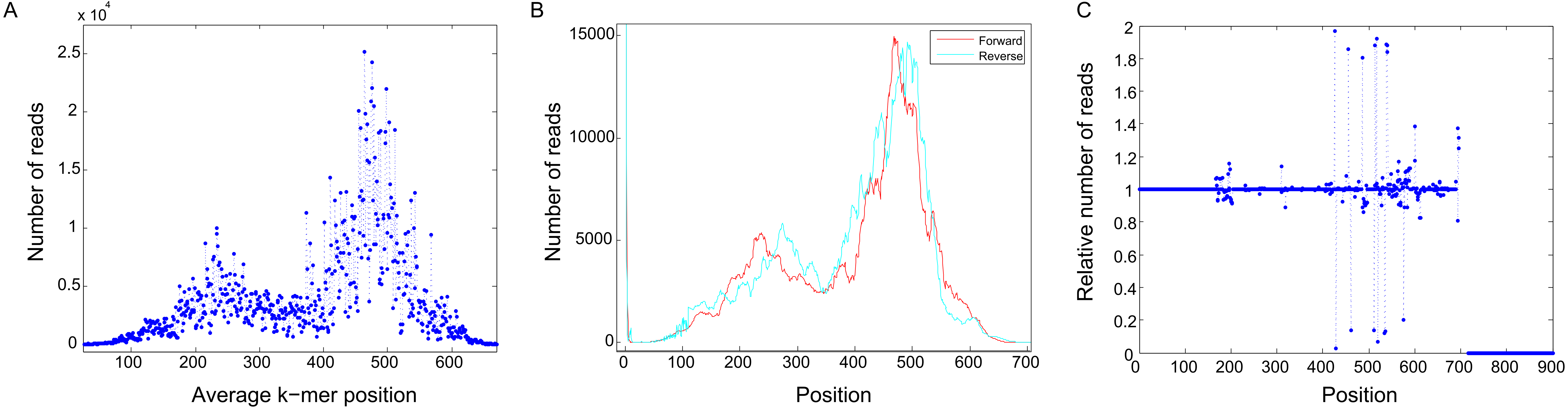

A
x 10^4
2.5
2
1.5
1
0.5
0
100
200
300
400
500
600
Number of reads
Average k−mer position
B
15000
10000
5000
0
0
100
200
300
400
500
600
700
Forward
Reverse
Number of reads
Position
C
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
700
800
900
Relative number of reads
Position

Table S3: Bacteria in simulated toy example

Sequence	Frequency
Eubacterium rectale ATCC 33656 [Eubacterium rectale;]	14.17%
Collinsella aerofaciens ATCC 25986 [470145 NZ_AAVN02000007.1]	11.71%
Blautia hydrogenotrophica DSM 10507 [469888 NZ_ACBZ01000217.1]	11.58%
Desulfovibrio piger GOR1 [51535 AF192152.1]	10.21%
Clostridium symbiosum ATCC 14940 [Clostridium symbiosum]	9.44%
Escherichia coli str. K-12 substr. MG1655 [9659 U18997.1]	8.14%
Marvinbryantia formatexigens DSM 14469 [94718 AJ318527.2]	7.88%
Bacteroides ovatus ATCC 8483 [469818 NZ_AAXF02000050.1]	7.88%
Bacteroides ovatus ATCC 8483 [469818 NZ_AAXF02000050.1]	7.88%
Bacteroides caccae ATCC 43185 [248140 AAVM02000008.1]	6.35%
Bacteroides sp. str. D1 [469729 ACAB01000173.1]	5.90%

Sequences also appear in Dataset 10 (see supplementary datasets)

SUPPLEMENTARY DATASETS

Supplementary Dataset 1: 454-BLAST bacteria of Drosophila samples:
454BLAST_results_L1_L2_E1_E2.fa

Supplementary Dataset 2: COMPASS bacteria of Drosophila samples:
COMPASS_results_L1_L2_E1_E2.fa

Supplementary Dataset 3: MG-RAST bacteria of Drosophila samples:
MGRAST_results_L1_L2_E1_E2.fa

Supplementary Dataset 4: 454-BLAST bacteria of human saliva samples:
454BLAST_results_H1_H2_H3_H4.fa

Supplementary Dataset 5: COMPASS bacteria of human saliva samples:
COMPASS_results_H1_H2_H3_H4.fa

The above FASTA files include the header and sequence of bacteria that appear in Figures 6, S3 and S5. Note that in cases that several bacteria share the same sequence over the amplicon (a 'group' in our notation), the FASTA file contains the representative that appears in the figure.

Supplementary Datasets 6-9:

Supplementary Datasets 6: MG-RAST read classification file of sample L1:
4526879.3_MGRAST_results_sample_L1_sequences_annotated_by_Greengenes.fna

Supplementary Datasets 7: MG-RAST read classification file of sample L2:

4526880.3_MGRAST_results_sample_L2_sequences_annotated_by_Greengenes.fna

Supplementary Datasets 8: MG-RAST read classification file of sample E1:

4526878.3_MGRAST_results_sample_E1_sequences_annotated_by_Greengenes.fna

Supplementary Datasets 9: MG-RAST read classification file of sample E2:

4526877.3_MGRAST_results_sample_E2_sequences_annotated_by_Greengenes.fna

An example for the header's format in these MG-RAST files is the following:

>4526879.3|102|Greengenes|550951 16S ribosomal RNA [Lactobacillus plantarum]

Where:

4526879.3 is the project ID

102 is the original read number that appears in the reads' fasta file

550951 is the Greengenes number of the classified bacterium

[Lactobacillus plantarum] is the MG-RAST classification

Supplementary Dataset 10: Sequences for bacteria in the toy mixture as a dataset

bacteria_toy_mixture.fa

REFERENCES

1. Zuk, O., Amir, A., Zeisel, A., Shamir, O. and Shental, N. (2013) Accurate Profiling of Microbial Communities from Massively Parallel Sequencing Using Convex Optimization. *SPIRE 2013, LNCS 8214, O. Kurland, M. Lewenstein, and E. Porat (Eds.)*, 279-297.
2. Kao, W.-C., Stevens, K. and Song, Y.S. (2009) BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res*, **19**, 1884-1895.
3. Kao, W.-C. and Song, Y.S. (2011) naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *J Comput Biol*, **18**, 365-377.