



OPEN

Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota

Francesco Durazzi¹, Claudia Sala¹, Gastone Castellani², Gerardo Manfreda³, Daniel Remondini¹✉ & Alessandra De Cesare⁴

In this paper we compared taxonomic results obtained by metataxonomics (16S rRNA gene sequencing) and metagenomics (whole shotgun metagenomic sequencing) to investigate their reliability for bacteria profiling, studying the chicken gut as a model system. The experimental conditions included two compartments of gastrointestinal tracts and two sampling times. We compared the relative abundance distributions obtained with the two sequencing strategies and then tested their capability to distinguish the experimental conditions. The results showed that 16S rRNA gene sequencing detects only part of the gut microbiota community revealed by shotgun sequencing. Specifically, when a sufficient number of reads is available, Shotgun sequencing has more power to identify less abundant taxa than 16S sequencing. Finally, we showed that the less abundant genera detected only by shotgun sequencing are biologically meaningful, being able to discriminate between the experimental conditions as much as the more abundant genera detected by both sequencing strategies.

The study of the gut bacterial community composition has become a fast-developing field, both for the assessment of possible correlations with human diseases and pathologies^{1,2} and for the examination of the effects of diets and probiotics in animal productions^{3,4}. This developing field has been hugely affected by the emergence and optimization of high-throughput sequencing, that made metagenomics the key instrument to access complex ecosystems, such as the human and animal gut. The popularity of high-throughput sequencing is due to decreasing cost and augmented speed and scalability of experiments⁵, which are crucial aspects when researchers design a sequencing project.

The main objectives of gut metagenomic studies are: (i) the identification of the gut microbiota taxonomic composition, (ii) the characterization of the relative abundances of taxa, (iii) the description of the functional contribution of each taxon and (iv) the understanding of the intra-species and/or intra-population gene heterogeneity⁶. To this aim, metataxonomics and metagenomics strategies are used. Metataxonomics consists in the targeted sequencing of 16S rRNA gene hypervariable regions⁷, and allows representative bacterial taxonomic estimation⁸ even when a relatively small number of raw reads is obtained (i.e., as low as 18,000–20,000 reads per sample)^{9,10}. The overall sequencing output of metataxonomics is a set of clusters of nearly identical sequences, referred to as Operational Taxonomic Units (OTUs)¹¹ or Amplicon Sequence Variants (ASVs)¹². From the analysis of such clusters, information on the community diversity, richness and evenness can be derived¹³, while accounting for the degree of divergence between different ecosystems or sample types¹⁴. However, the choice of primers used to amplify 16S rRNA leads to potential biases in the representation of the taxonomic units^{15–18}.

Besides the mapping of the taxonomic composition of a sample, the most challenging task for metagenomic studies is the evaluation of the genic contribution of each member of the investigated community in terms of functional genes⁶. To address this issue, shotgun metagenomic sequencing is the most suitable strategy. Here, long DNA molecules, such as complete chromosomes, are randomly broken into fragments that are then sequenced¹⁹. Hence, metagenomic data deliver knowledge on the taxonomic composition of the ecosystem under study but

¹Department of Physics and Astronomy, University of Bologna, 40127 Bologna, Italy. ²Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, 40127 Bologna, Italy. ³Department of Agricultural and Food Sciences, University of Bologna, 40064 Ozzano dell'Emilia, Italy. ⁴Department of Veterinary Medical Sciences, University of Bologna, 40064 Ozzano dell'Emilia, Italy. ✉email: daniel.remondini@unibo.it

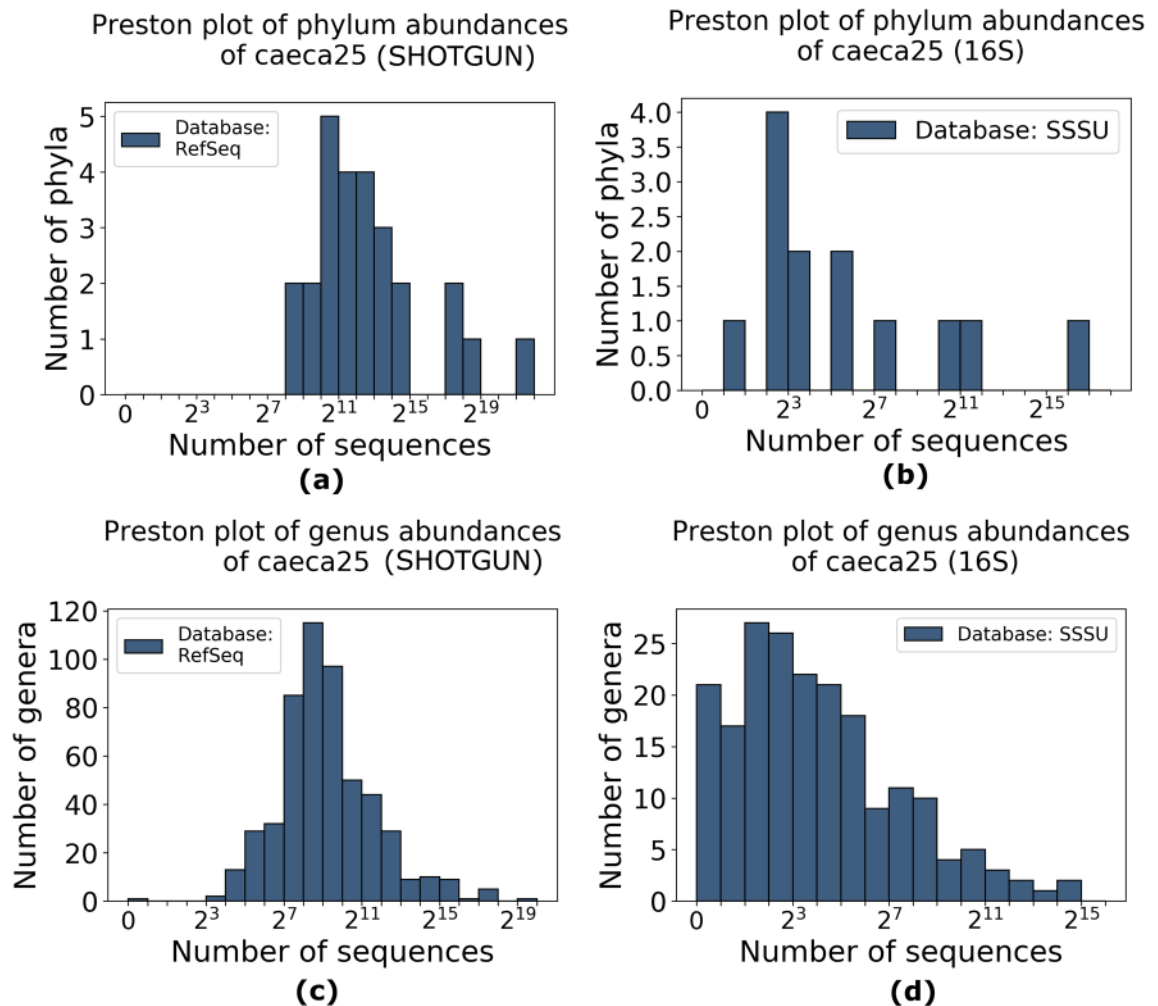


Figure 1. RSA histograms in logarithmic scale (Preston plots²¹) of bacterial abundances in one sample selected as an example (caeca25): (a) genera sampled by shotgun sequencing, (b) genera sampled by 16S rRNA sequencing, (c) phyla sampled by shotgun sequencing and (d) phyla sampled by 16S sequencing.

also on functional genes in the sample, an information that is not retrievable with 16S rRNA gene sequencing. On the other hand, shotgun metagenomics requires higher coverage than metataxonomics¹⁶.

In a previous study we gained insights into the effects of *Lactobacillus acidophilus* D2/CSL (CECT 4529) (LA) on the ecology of the bacteria colonising the chicken gastrointestinal (GI) tract²⁰. Specifically, we investigated the crop and caeca microbiomes in treated animals and in a control group at 1, 14 and 35 days of rearing, using shotgun metagenomic sequencing. In the present study, the same DNA samples investigated in the previous research were analysed using the targeted 16S rRNA gene sequencing (16S). Then, the results obtained with both sequencing strategies were compared to answer three broad questions: (1) what is the resolution of bacterial populations observed by shotgun sequencing as a function of the total number of reads; (2) how many bacterial genera are retrieved exclusively by one sequencing strategy and not by the other; (3) how much the two sequencing strategies retrieve information about the specific experimental conditions, namely the different compartments of gastrointestinal tracts and the sampling time. To address these questions, we studied the dependence between the capability of detecting bacterial populations and the total number of reads and we showed that, when a sufficient number of reads is available, shotgun sequencing finds a statistically significant higher number of taxa than 16S sequencing, corresponding to the less abundant. Finally, we analysed bacterial community profiles exclusive to each strategy, demonstrating that the genera detected only by shotgun sequencing are able to discriminate between the experimental conditions better than those detected only by 16S sequencing.

Results and discussion

Relative species abundance distribution. In order to evaluate sample quality, we analysed both the Relative Species Abundance distribution (RSA) and the rarefaction curves. For each sample, we compared the RSA derived by shotgun and 16S sequencing. RSA histograms in logarithmic scale show that the distributions obtained by shotgun and 16S have similar shape at phylum level (Fig. 1a, b). In Fig. 1b, the 16S sample is characterized by a more patchy distribution, having identified less phyla. At phylum level, both strategies produce

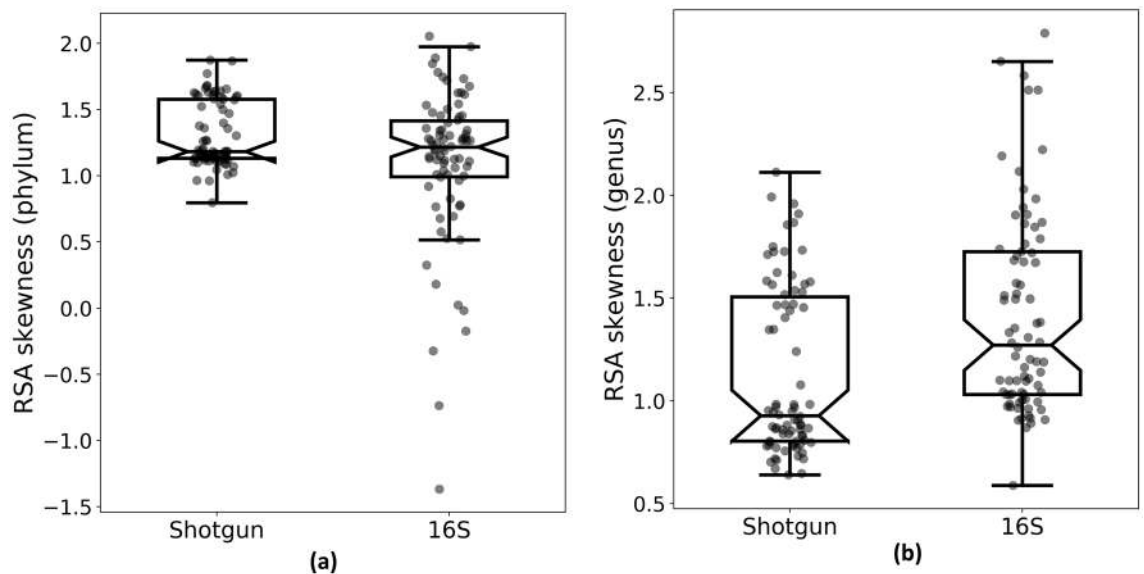


Figure 2. Box plot of the RSA skewness of bacterial communities at (a) phylum level and (b) genus level. Bacterial communities are sampled with (left) shotgun sequencing and (right) 16S sequencing.

positively skewed samples in the \log_2 -transformed distributions, except for 16S outliers, because none of the phyla is significantly rare (Fig. 2a).

On the other hand, at genus level, the two strategies display different shapes (Fig. 1c, d, Supplementary Fig. S1, S2, S3, S4). Indeed, the \log_2 -transformed distributions derived by shotgun sequencing generally have a skewness closer to zero compared to those obtained by 16S, i.e. are more symmetrical (Figure 2b): a paired Student's *t*-test on the skewness shows a significant difference between them ($P = 8 \cdot 10^{-6}$). This indicates that shotgun samples are characterized by a higher sampling size. According to Preston, left-skewed shapes of the RSA can be explained as artefacts of small sample size^{21,22}, since insufficient sampling of the original space produces a truncation of the left tail of the RSA, increasing its skewness.

In shotgun samples, the RSA skewness at genus level is related to the total number of reads (Supplementary Fig. S5): the shotgun samples with the lowest total number of reads have the largest skewness. Specifically, Supplementary Figure S5 shows that shotgun samples cluster in two groups, one characterized by a low number of reads (# reads < 500,000, 28/78 samples) and a highly skewed RSA (greater than the 16S median), and one with a high total number of reads (# reads > 500,000, 50/78 samples) and a less skewed RSA.

Noticeably, the high-skewness group includes all 9 samples from 1st day, all 15 crop samples from 14th day and 4 out of 18 crop samples from 35th day. The samples collected at day 1 were very poor in terms of biomass and the crop samples contained more feed residues than caecal samples, making the DNA extraction less efficient both in terms of DNA quantity and quality. For the comparative analysis we removed samples with less than 500,000 reads being characterized by a low quality. This choice was corroborated by the analysis of the rarefaction curves, showing that shotgun samples with less than 500,000 reads do not reach a plateau in terms of identified genera (Supplementary Fig. S6). All the 50 samples included in the comparative analysis have a total number of reads > 500,000 and a skewness lower than the median of 16S samples, indicating a good sampling depth. Since included samples were characterized by a high microbial load, we are confident to extend the results of the following analyses only to samples with few contaminant DNA and low cross-contaminations. Nonetheless, we have shown that shotgun samples have a RSA similar to 16S samples when a low number of total reads is available, thus hypothesizing that in differential analyses carried on samples with a low microbial load regime, shotgun sequencing could perform similarly to 16S sequencing or even worse.

For a balanced comparison, also 16S samples corresponding to the discarded shotgun samples were removed.

Differential analysis for the experimental conditions. Since in many situations a metagenomic analysis is used to discriminate between different experimental conditions, we compared the results of differential analysis performed on reads obtained by the two strategies. To this aim, we analysed the fold changes of genera abundances between compartments of the GI tract and between sampling times (Fig. 3 for caeca vs crop, Supplementary Fig. S7 for 14th vs. 35th day) common to both sequencing strategies (288 genera for caeca vs crop, and 246 for 14th vs. 45th day). Comparing the genera abundances between caeca and crop, 16S identified 108 statistically significant differences (adjusted $P < 0.05$ with DESeq2), while shotgun identified 256; 28 genera, corresponding to 9.7% of total common genera, were not identified as significantly different by either strategy. Among the 104 genera identified to be different between caeca and crop by both sequencing strategies, 93.3% (97/104) showed a concordant fold change. Concerning the genera with different abundance between sampling times, 16S detected 58 statistically significant changes (adjusted $P < 0.05$ with DESeq2), while 75 were detected by shotgun. The 80% (16/20) of the 20 genera with different abundance between the 14th and 35th day for both sequencing strategies showed a concordant fold change (see Table). The discrepancies seem to be related

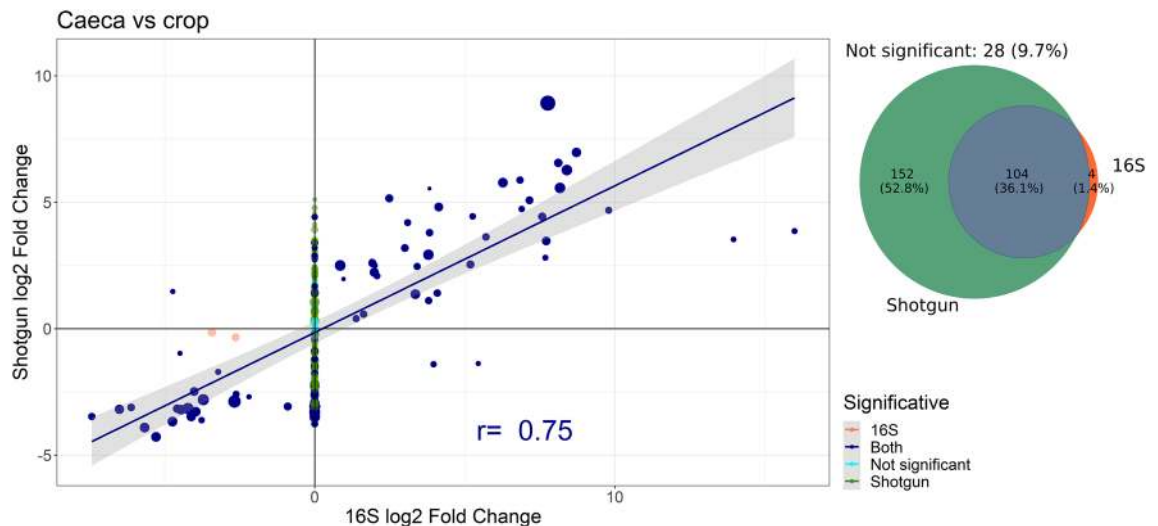


Figure 3. Fold changes between caeca and crop in genera identified by both strategies. Some fold changes are shrunk toward zero by the DESeq2 algorithm (see “Methods” section). Points with a statistically significant change for both strategies are represented in blue, for shotgun only in green, for 16S only in orange and without a significant change in cyan (adjusted $P > 0.05$ with DESeq2). Point size is the \log_{10} of average number of reads from shotgun strategy mapping to each genus. Pearson’s correlation coefficient r and regression line are computed only on points with statistically significant fold changes according to both strategies (“Both” group in figure legend and in Table 1).

	16S & shotgun	16S only	Shotgun only	Not significant
Caeca versus crop	93.3 (97/104)	75.0 (3/4)	72.4 (110/152)	64.3 (18/28)
14th versus 35th day	80.0 (16/20)	55.3 (21/38)	58.2 (32/55)	57.1 (76/133)

Table 1. Percentage of genera having a concordant change in abundance in shotgun and 16S samples with a statistically significant change for: both sequencing strategies (1st column), 16S only (2nd column), Shotgun only (3rd column), neither (4th column).

to detection issues in 16S samples: indeed, all the seven discordant changes in caeca vs crop are caused by the partial or total absence of a genus in 16S samples (as shown in Supplementary Fig. S8). This is possibly due to the fact that these genera are close to the detection limit of 16S strategy, for which we have provided an estimate in the next section. On the other hand, three of the four discordant genera in the 14th vs 35th day showed actual discrepancies between shotgun and 16S changes, not necessarily caused by detection issues (see Supplementary Fig. S9).

Noticeably, shotgun sequencing found 152 statistically significant changes in genera abundance between caeca and crop of chickens that 16S sequencing failed to detect, while 16S found only 4 changes that shotgun sequencing did not identify (Fig. 3).

Genera detection and abundance quantification. The agreement between the taxonomic profiles estimated with the two strategies was further evaluated computing for each sample the Pearson’s correlation coefficient (r) between the taxonomic abundances of genera common to 16S and shotgun sequencing. Overall, we observed a good agreement between the taxonomic abundances found by the two strategies (Fig. 4, Supplementary Fig. S10), with an average correlation of 0.69 ± 0.03 in caeca (all p -values $< 10^{-11}$ for correlation significance between 16S-shotgun sample pairs) and 0.75 ± 0.05 in the crop (all p -values $< 5 \cdot 10^{-5}$).

A larger difference is observed between the number of identified taxa by the two strategies. The histogram on the left of Fig. 4 shows that, in the selected sample, a great majority of genera detected by shotgun sequencing are not found by 16S sequencing (green portion of the bars). This phenomenon is mostly observed on the leftmost bins, that represent low abundance genera (Supplementary Fig. S11, S12, S13, S14 for single samples). These results are confirmed in Fig. 5 and Supplementary Table S1, showing that (a) all the phyla detected in 16S samples were identified in shotgun samples and (b) only a small number of genera (about 23% for caeca samples and 11% for crop) is recovered by both strategies. The percentage of reads mapping to the genera identified by both sequencing strategies is large (on average 89% for caeca and 99% for crop). This means that genera identified by both strategies are the most abundant ones, e.g. those mapped by most of the reads. In summary, shotgun sequencing always identifies more taxa than 16S sequencing, as also reflected on the rarefaction curves

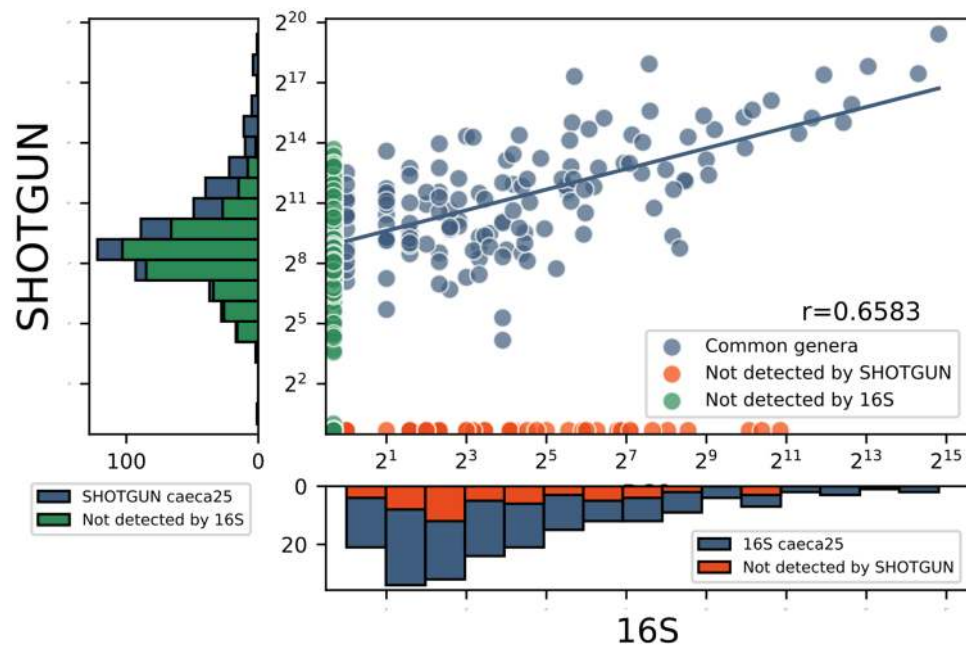


Figure 4. Scatter plot of 16S and shotgun genera abundances of one sample selected as example (caeca25). Histograms display stacked bars, where every column is divided in a part corresponding to the abundance of genera detected by both sequencing strategies (blue) and the other part is relative to genera detected exclusively by only one strategy (red for 16S and green for shotgun). Pearson's correlation coefficient is computed only for the common genera. Logarithmic (\log_2) scale helps to recognize that less abundant genera identified by shotgun sequencing are almost not detected by 16S sequencing.

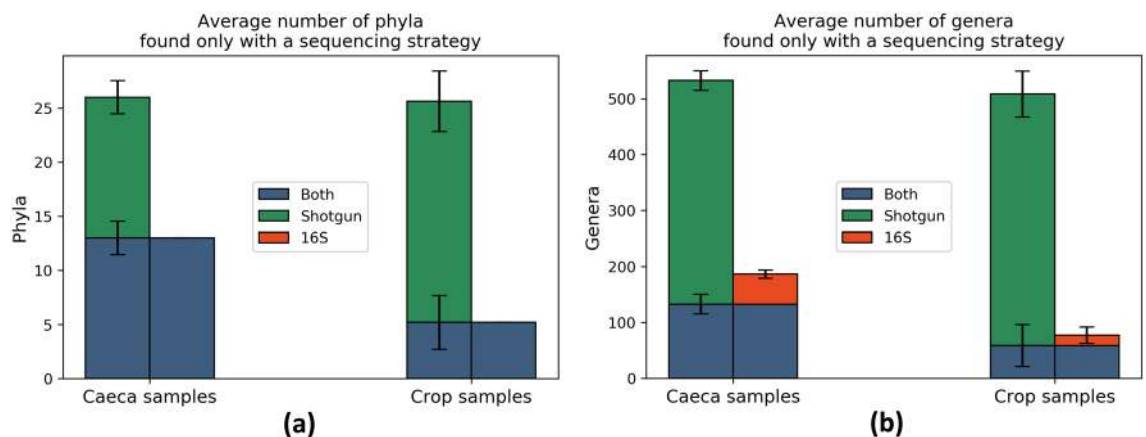


Figure 5. Average number of (a) phyla and (b) genera found within caeca and crop samples. The length of the error bars is equal to the standard deviation computed on all the samples.

(Supplementary Fig. S6). Moreover, 16S sequencing predominantly detects taxa that are also identified by shotgun sequencing (133 of 187 genera and 14 of 14 phyla on average) (Fig. 5).

The relationship between abundances detected by 16S and shotgun metagenomics was further investigated fitting a linear regression model on the abundance of genera common to both strategies in each sample. We considered, for each shotgun-16S pair of samples, the logarithmic abundances obtained with 16S as independent variable and those obtained with shotgun as dependent variable, so that the intercept in this model represents the number of shotgun reads corresponding to genera that are mapped to one single read by 16S sequencing, that we consider as a detection limit. Here, samples with low number of reads ($< 500,000$) were included for completeness. Results show that the model intercept increases as a function of the total number of reads available in shotgun samples (Fig. 6). The regression intercept and the total number of shotgun reads are positively correlated (Pearson's coefficient = 0.93, $P < 10^{-16}$).

Hence, given the total number of reads of a shotgun sequencing sample, the genera mapped by a number of reads lower than the model intercept are the most likely to be undetected by 16S strategy. For example,

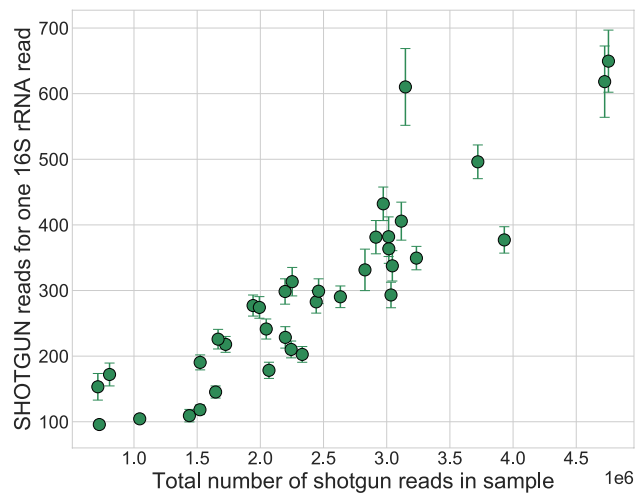


Figure 6. Intercepts of shotgun vs 16S abundance linear regressions of caeca samples against the total number of reads in each set, representing the number of shotgun sequences corresponding to one 16S sequence. Error bars correspond to the confidence interval for the parameter provided by the fit.

considering the sample depicted in Fig. 4 with a total number of reads equal to 2,972,671 with shotgun strategy, genera with a number of reads approximately < 350 would probably not be detected by 16S sequencing, corresponding to 88% of total detected genera in that shotgun sample.

Sample stratification by experimental condition. Since in many studies an unsupervised analysis (i.e. clustering) based on taxonomic profiling is performed, in order to identify possible sample stratification due to experimental conditions or to other unknown factors, we calculated the Bray–Curtis beta diversity and performed a Principal Coordinate Analysis (PCoA). We considered for every sample a n -dimensional ($n = 678$) vector of abundances, considering the genera that are common to all samples (a graphical representation is showed in Supplementary Figure S15). A quantitative evaluation of the separation in the PCoA space of samples labelled by experimental conditions was obtained through the mean Silhouette Score (SS) of the samples at genus level. We compared the silhouette scores either on the totality of genera identified by shotgun (“SHOTGUN”) or 16S (“16S”) strategy, or on the subset of genera detected exclusively by shotgun (“SHOTGUNex”) or by 16S (“16Sex”), to evaluate their ability to discriminate between known experimental conditions (compartment of the GI tract and sampling time). Results show that when samples are labelled according to the compartment of the GI tract, the Silhouette Score is high (i.e., close to 1) for both strategies (see Supplementary Tab. S2 and Supplementary Fig. S16). However, the separation between groups is stronger in shotgun samples than in 16S samples. This result remains true even when the PCoA of shotgun samples is computed considering only shotgun-specific genera.

On the other hand, when grouping samples according to sampling time within the same compartment of the GI tract, Silhouette Scores are generally lower for both strategies, while still achieving a good separation (see Fig. 7). Figure 7 plots refer to the Silhouette Scores of Supplementary Table S3, in which full samples (Fig. 7a,b) reach an almost perfect separation, though the low compactness leads to worse scores with respect to GI tract labelling. An interesting result is that abundance profiles of genera found only in shotgun sequencing (SHOTGUNex) have a positive silhouette score, while 16Sex samples have a smaller silhouette score, close to 0 (a value representing low-quality clustering, very close to random). This result highlights that shotgun sequencing detects low-abundance genera that carry significant information about the experimental biological factors, while 16S-specific genera fail to correctly cluster samples based on one of the experimental factors considered (namely sampling time), and a good separation is obtained only when the most abundant genera common to both strategies are considered.

Conclusion

This comparative analysis, based on 78 chicken gut metagenomes, showed that shotgun sequencing recovered more information about low-abundance genera than 16S sequencing, when a sufficient number of reads were available for taxonomic profiling ($> 500,000$ reads). Most of the bacterial genera were identified by shotgun sequencing, while, on average, 16S recovered less than 31% of the genera and less than 50% of the phyla.

In agreement with our study, Campanaro et al. (2018)¹⁷ showed that several phyla were strongly under-represented in the 16S amplicon analysis in comparison to random shotgun DNA sequencing. Moreover, Laudadio et al. (2018)¹⁶ highlighted the higher resolution of taxonomic analyses performed by shotgun metagenomics as compared to 16S sequencing at different taxonomic levels, using the number of taxa identified in each sample as a metric to evaluate performance. Tessler et al. (2017)²³ and Shah et al. (2011)²⁴ showed instead that 16S rRNA sequencing identified more diverse bacterial phyla and families than shotgun sequencing, but we remark that in both studies the number of reads per sample used for taxonomic profiling was about 100 times smaller than in our

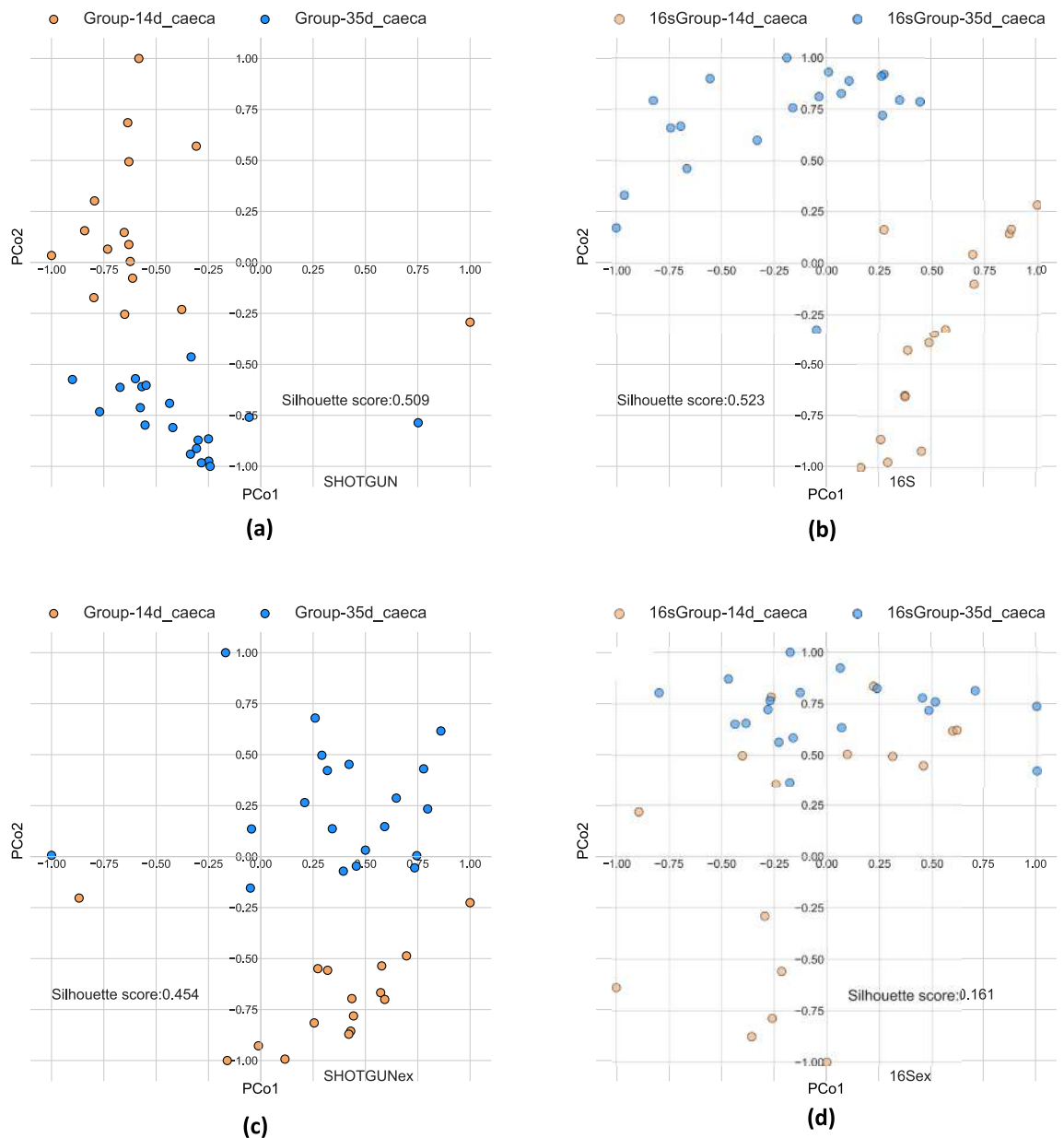


Figure 7. PCoA based on the beta-diversities between samples (Bray–Curtis metric), computed on genera abundances of caeca samples normalized by DESeq2, labelled by sampling time: 14th day (gold), 35th day (cyan). **(a, b)** with all genera detected respectively by shotgun **(a)** and 16S **(b)**; with genera found exclusively by shotgun **(c)** or 16S **(d)**.

study and in the two previously mentioned papers, thus this undersampling might be the reason of this apparent inconsistency (see Supplementary Table S4). In our work, we demonstrated that the total number of sequenced reads is a critical factor to perform a robust analysis with shotgun sequencing at low taxonomic (genus) level. Furthermore, analysing the changes in genera abundance due to different experimental conditions (i.e., different compartments of the GI tract and sampling time), we showed that shotgun sequencing is more sensitive than 16S, being able to identify a larger number of genera that are significantly affected by these factors (Fig. 3).

The RSA distributions obtained by the two sequencing strategies showed quantitative and qualitative differences at genus level, in particular for the left tail. When the total number of reads was as low as $\sim 200,000$ in shotgun samples, the RSA shapes became strongly positively skewed, similarly to those obtained with 16S. On the other hand, increasing the sampling intensity enabled to detect less abundant genera²¹. The genera detected only by shotgun sequencing are not very abundant, because even if they constitute about 75% of the identified genera, the associated read count is less than 9% of the total reads. Nonetheless, these less abundant genera provided reliable information and showed significant correlation with the different experimental conditions. The same conclusion could not be drawn for low-abundance genera exclusively identified by 16S sequencing. The latter were able to discriminate different compartments of the GI tract but failed to stratify samples according

to the sampling time. Finally, our results allowed to provide an estimation of the number of genera that would likely be detected by shotgun sequencing and not by 16S sequencing.

Methods

Animals and treatments. The experimental trial from which the samples were collected has been previously described in De Cesare et al. (2020)²⁰. In our study, we compared bacterial abundance profiles in 78 samples undergoing both 16S and shotgun metagenomic sequencing. Overall, 40 samples were from caeca (4 from 1st day, 16 from 14th day and 20 from 35th day) and 38 from crop (5 from 1st day, 15 from 14th day and 18 from 35th day) were collected.

Shotgun sequencing was performed as previously described²⁰, whereas for amplicon sequencing, the libraries were prepared following the Illumina 16S Library preparation protocol¹⁰, amplifying the variable V3 and V4 regions of the 16S rRNA. Sequencing was performed in paired-end at 150 bp in the Illumina MiSeq^{25,26}. The maximum output of the v2 kit is 15 million reads per run, meaning approximately 187,500 reads per sample.

Since biomasses extracted from GI tracts (caecum in particular) have typically a high microbial load²⁷ and host depletion methods can be not effective for DNA libraries²⁸, the samples were not depleted from host DNA. The reads not assigned to Bacteria domain were hence removed after taxonomic profiling.

Overall, 78 metagenomes were analysed: 40 metagenomes from caeca (i.e., 4 from 1st day, 16 from 14th day and 20 from 35th day) and 38 from crop (i.e., 5 from 1st day, 15 from 14th day and 18 from 35th day). The 16S and shotgun metagenomes analysed as well as their number of sequences mapping to Bacteria are detailed in Supplementary Table S5, with genera rarefaction curves in Supplementary Figure S6.

Processing and taxonomic profiling. Both 16S and shotgun raw reads were pre-processed and assigned to a genus using MG-RAST with default parameters²⁹. Specifically, 16S reads were taxonomically classified using the Silva SSU reference database³⁰, while the RefSeq database³¹ was used for shotgun reads. Singleton reads were discarded.

MG-RAST advices against the reliability of taxonomic profiling at species level²⁹, in particular for samples obtained by shotgun sequencing. For this reason, we choose to perform the analysis of the bacterial communities up to genus level.

Statistical analysis methods. Data were processed and visualized in Python 3.6 and R 3.6.0 using custom scripts.

For each sample, the Relative Species Abundance distribution (RSA) was computed counting the number of genera that have a certain abundance. RSAs were visualized as Preston plots^{21,32}. Rarefaction curves were computed with an extension³³ of the R package phyloseq³⁴ implementing the package vegan³⁵.

Counts normalization and differential genera abundance analysis were performed by DESeq2 package³⁶, considering as significant those changes with an adjusted p-value lower than 0.05. High variability in the dispersion of the counts was adjusted with a shrinkage procedure by DESeq2, that led to smaller estimates of the fold changes in samples with a low number of total reads. For each sample, we calculated and visualized the Pearson's correlation coefficient between the taxonomic abundances obtained using 16S and shotgun, along with a p-value for correlation significance. For this analysis, we considered only genera that were present in both profiles.

Beta diversity was computed using the Bray–Curtis distance^{37,38} and considering normalized counts retrieved by DESeq2, that keeps into account the differences in sample size. Principal Coordinate Analysis (PCoA)³⁹ was performed to visualize the samples based on the beta diversity. Silhouette Scores were calculated to assess the correspondence between sample displacement in the PCoA space and experimental factors (i.e. compartments of gastrointestinal tract and sampling time). Sampling time corresponds to the days of rearing of the chickens (1st, 14th and 35th).

Ethical approval and informed consent. The experiments were conducted after obtaining the approval of Ethical Committee of the University of Bologna on 17/3/2014 (ID: 10/79/2014). All experiments were performed in accordance with relevant guidelines and regulations.

Data availability

All 16S and shotgun metagenomic sequences tested as part of this comparative study were deposited in MG-RAST (<http://metagenomics.anl.gov/>) and are public available under the projects named newlacto16S (<http://www.mg-rast.org/mgmain.html?mgpage=project&project=mgp91466>) and newlacto (<http://www.mg-rast.org/mgmain.html?mgpage=project&project=mgp13081>), respectively.

Received: 13 February 2020; Accepted: 29 December 2020

Published online: 04 February 2021

References

1. Maccaferri, S., Biagi, E. & Brigidi, P. Metagenomics: key to human gut microbiota. *Dig. Dis.* **29**, 525–530 (2011).
2. Cornejo-Pareja, I., Muñoz-Garach, A., Clemente-Postigo, M. & Tinahones, F. J. Importance of gut microbiota in obesity. *Eur. J. Clin. Nutr.* **72**, (2018).
3. Stanley, D., Hughes, R. J. & Moore, R. J. Microbiota of the chicken gastrointestinal tract: influence on health, productivity and disease. *Appl. Microbiol. Biotechnol.* **98**, 4301–4310 (2014).
4. De Cesare, A. et al. Effect of dietary supplementation with *Lactobacillus acidophilus* D2/CSL (CECT 4529) on caecum microbiota and productive performance in broiler chickens. *PLoS One* **12**, (2017).
5. ten Hoopen, P. et al. The metagenomic data life-cycle: Standards and best practices. *Gigascience* **6**, (2017).

6. Scholz, M. B., Lo, C.-C. & Chain, P. S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* **23**, (2012).
7. Chistoserdova, L. Functional metagenomics: recent advances and future challenges. *Biotechnol. Genet. Eng. Rev.* **26**, 335–352 (2009).
8. Choi, K. Y., Lee, T. K. & Sul, W. J. Metagenomic analysis of chicken gut microbiota for improving metabolism and health of chickens: a review. *Asian-Australasian J. Anim. Sci.* **28**, 1217–1225 (2015).
9. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
10. 16S Metagenomic sequencing library preparation. in 1–28 (Illumina MiSeq System, 2013).
11. Cole, J. R. *et al.* Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, (2014).
12. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
13. Colwell, R. K. Biodiversity: concepts, patterns, and measurement. in *The Princeton Guide to Ecology* 257–263 (2009).
14. Sala, C. *et al.* Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. *BMC Bioinformatics* **17**, (2016).
15. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, (2017).
16. Laudadio, I. *et al.* Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *Omi. A J. Integr. Biol.* **22**, 248–254 (2018).
17. Campanaro, S., Treu, L., Kougias, P. G., Zhu, X. & Angelidaki, I. Taxonomy of anaerobic digestion microbiome reveals biases associated with the applied high throughput sequencing strategies. *Sci. Rep.* **8**, (2018).
18. Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. & Cotter, P. D. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* **16**, 123 (2016).
19. Weinstock, G. M. Genomic approaches to studying the human microbiota. *Nature* **489**, 250–256 (2012).
20. De Cesare, A. *et al.* Effect of Lactobacillus acidophilus D2/CSL (CECT 4529) Supplementation in Drinking Water on Chicken Crop and Caeca Microbiome. *PLoS One* 1–18 (2020) <https://doi.org/10.1371/journal.pone.0228338>.
21. Preston, F. W. The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948).
22. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton University Press, 2001).
23. Tessler, M. *et al.* Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* **7**, (2017).
24. Shah, N., Tang, H., Doak, T. G. & Ye, Y. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac. Symp. Biocomput.* 165–76 (2011) https://doi.org/10.1142/9789814335058_0018.
25. De Cesare, A. *et al.* Impact of a probiotic-based cleaning product on the microbiological profile of broiler litters and chicken caeca microbiota. *Poult. Sci.* **98**, 3602–3610 (2019).
26. MiSeq Reagent Kits v2. <https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/miseq-reagent-kit-v2.html>.
27. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies : Issues and Recommendations. *Trends Microbiol.* **27**, 105–117 (2018).
28. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
29. Wilke, A. *et al.* *MG-RAST Manual for version 4, revision 3*. (2017).
30. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, (2013).
31. O'leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, 733–745 (2015).
32. Magurran, A. E. *Measuring Biological Diversity*. (Wiley-Blackwell, 2003).
33. Mariadassou, M. Phyloseq-extended. https://github.com/mahendra-mariadassou/phyloseq-extended/blob/master/R/graphical_methods.R.
34. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).
35. Oksanen, J. *et al.* *vegan: Community Ecology Package*. (2019).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
37. Faith, D. P., Minchin, P. R. & Belbin, L. *Compositional dissimilarity as a robust measure of ecological distance*. **69**, 57–68 (1987).
38. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).
39. Wickelmaier, F. An introduction to MDS. *Ind. Commer. Train.* **10**, 11–18 (2003).

Acknowledgements

A.D.C., D.R., G.C. and G.M. were supported by H2020 EU COMPARE project n. 643476. A.D.C. and G.M. acknowledge Dr. Antonio Parisi of the Istituto Zooprofilattico Sperimentale della Puglia e della Basilicata for support in 16S sequencing.

Author contributions

A.D.C. and D.R. designed and supervised the analysis. F.D. and C.S. analyzed the data. G.C. and G.M. supervised the analysis and interpreted the results. All authors wrote the manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-82726-y>.

Correspondence and requests for materials should be addressed to D.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021