Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses

Xiaohuan Sun¹, Yue-Hua Hu^{2,*}, Jingjing Wang³, Chao Fang^{1,4}, Jiguang Li³, Mo Han¹, Xiaofang Wei³, Haotian Zheng^{1,5}, Xiaoqing Luo^{1,6}, Yangyang Jia¹, Meihua Gong³, Liang Xiao¹ and Zewei Song^{1,*}

¹ BGI-Shenzhen, Shenzhen 518083, China

² CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China

³ MGI, BGI-Shenzhen, Shenzhen 518083, China

⁴ Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Shenzhen 518083, China

⁵ BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China

⁶ State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China

ABSTRACT

Metabarcoding is a widely used method for fast characterization of microbial communities in complex environmental samples. However, the selection of sequencing platform can have a noticeable effect on the estimated community composition. Here, we evaluated the metabarcoding performance of a DNBSEQ-G400 sequencer developed by MGI Tech using 16S and internal transcribed spacer (ITS) markers to investigate bacterial and fungal mock communities, as well as the ITS2 marker to investigate the fungal community of 1144 soil samples, with additional technical replicates. We show that highly accurate sequencing of bacterial and fungal communities is achievable using DNBSEQ-G400. Measures of diversity and correlation from soil metabarcoding showed that the results correlated highly with those of different machines of the same model, as well as between different sequencing modes (single-end 400 bp and paired-end 200 bp). Moderate, but significant differences were observed between results produced with different sequencing platforms (DNBSEQ-G400 and MiSeq); however, the highest differences can be caused by selecting different primer pairs for PCR amplification of taxonomic markers. These differences suggested that care is needed while jointly analyzing metabarcoding data from different experiments. This study demonstrated the high performance and accuracy of DNBSEQ-G400 for short-read metabarcoding of microbial communities. Our study also produced datasets to allow further investigation of microbial diversity.

Submitted: 22 December 2020

Accepted: 18 March 2021

Published: 23 March 2021

* Corresponding authors. E-mail: huyuehua@xtbg.org.cn; songzewei@cngb.org

Published by GigaScience Press.

Preprint submitted at <https://doi.org/10.1101/2020.07.02.185710>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Gigabyte, 2021, 1–15

Subjects Genetics and Genomics, Metagenomics, Microbial Ecology

DATA DESCRIPTION

We examined the accuracy of DNBSEQ-G400 sequencing on bacterial and fungal mock samples, and on a large set of soil fungal communities, and produced a dataset suitable for

benchmarking and comparison studies. Independently sequenced PCR and sequencing run technical replicates were used to examine the performance of DNBSEQ-G400's single-end 400 bp (SE400) and paired-end 200 bp (PE200) sequencing modes. Moreover, to assess the performance consistency by a single sequencing mode within one platform, all soil samples were sequenced in SE400 mode three times. Finally, to validate these data, we compared the performance in measuring the alpha and beta diversity of the soil fungal communities between two sequencing platforms, using DNBSEQ-G400 and Illumina MiSeq.

CONTEXT

Culture independent studies have greatly expanded our knowledge of microbial diversity in recent decades [1–3]. Studies focused on characterizing microbial community composition often amplify the 16S or internal transcribed spacer (ITS) regions of rRNA [4, 5], which are the most commonly used barcoding regions for bacteria and fungi, respectively. High-throughput sequencing (HTS)-based metabarcoding allows the parallel analysis of large numbers of samples from different natural environments, especially those with complex taxonomic composition or low microbial biomass, such as soil [6], air [7], glaciers [8], and the deep sea [9].

Factors contributing to ecological biases in HTS-based metabarcoding have been reported in many studies, based on factors occurring both before (owing to marker gene length, AT/GC bias, the choice of PCR polymerase, primers and number of cycles [10–12]), and after sequencing (owing to choice of data-analyzing methods [13–15]). In addition, using mock communities, run-to-run variation in taxon presence and abundance was also observed among 16S HTS using Illumina HiSeq [16]. Similarly, bias among sequencing runs in detecting fungal taxa in soil samples was also reported [17], indicating the necessity of using positive controls in each sequencing study. Compared to the intensive focus on biases in pre- and post-HTS processes, relatively few studies have focused on the potential bias introduced by sequencers and sequencing modes.

HTS datasets of soil microbial communities can be produced by various sequencing platforms, including 454 (a platform no longer in use), IonTorrent, Pacific Biosciences (PacBio), Oxford Nanopore Technologies (ONT), and – the most frequently used – Illumina [6, 18]. In 2015, the DNBSEQ™ platform was released by MGI Tech., using combinatorial Probe-Anchor Synthesis (cPAS) and DNA Nanoball (DNB) techniques [19]. Since then, the DNBSEQ™ platform has been applied in various areas of genomics [19–21] and metagenomics [22, 23], producing results of highly comparable quality to other platforms [24, 25]. In 2017, an ultra-HTS sequencer, DNBSEQ-G400 (formerly known as MGISEQ-2000), which supported 2×200 paired-end and 400 single-end data, with a maximum throughput of 720 Gbp, was launched by MGI Tech. In comparison, Illumina's MiSeq and NovaSeq are capable of generating 2×300 and 2×250 paired-end data, have a 15 Gbp and 400 Gbp maximum throughput. With the comparable (and growing) read length and larger throughput, DNBSEQ-G400 can be seen as an attractive and highly capable alternative for microbial HTS analyses. Studies comparing both types of platforms are thus needed to reveal possible important differences between them and to support informed decision-making when selecting the sequencing strategy.

METHODS

A protocol collection gathering together methods for amplification, library preparation and sequencing at DNBSEQ-G400 and Illumina MiSeq is available at protocols.io (Figure 1 [26]).

Figure 1. Protocol collection for efficient and stable metabarcoding using the DNBSEQ-G400 sequencer and validated by comprehensive community analyses. <https://www.protocols.io/widgets/doi?uri=dx.doi.org/10.17504/protocols.io.brtpm6mn>

Sampling and sequencing experimental design

Two mock communities containing bacterial and fungal strains, respectively, were used in our study. ZymoBIOMICS Microbial Community DNA Standard D6305 was purchased from ZymoBIOMICS™; the 16S V4 region was examined for Zymo mock. Strains for fungal mock community standard (equally mixed genomic DNA of *Aspergillus ustus*, *Trichoderma koningii*, *Penicillium expansum*, *Aspergillus nidulans*, *Penicillium chrysogenum*, *Trichoderma reesei* and *Trichoderma longibrachiatum*) were obtained from China General Microbiological Culture Collection Center; the ITS2 region was examined for this mock. Taxonomic identification of the strains was confirmed by Sanger sequencing of full-length rRNA [27]. The genomic DNA of *Lentinula edodes* was mixed with gDNA of the fungi *Flammulina velutipes*, *Pleurotus eryngii* and *Saccharomyces cerevisiae* (obtained as commercial products intended for human consumption), in ratios of 1:1, 1:10 and 1:100, respectively. Meanwhile, the ITS2 PCR product of *L. edodes* was mixed with the ITS2 PCR products of *F. velutipes*, *P. eryngii* and *S. cerevisiae*, in ratios of 1:1, 1:10 and 1:100, respectively, to determine sequencing sensitivity.

A total of 1276 topsoil samples (described in GigaDB [27]) were collected from Nabanhe (NBH) 20 ha (400 × 500 m) tropical rainforest plot, Ailaoshan (ALS) 20 ha (500 × 400 m) subtropical evergreen forest plot, and Lijiang (LJ) 25 ha (500 × 500 m) subalpine forest plot, Yunnan, southwest China. We collected topsoil cores at each site (0–10 cm in depth) by hammering a ring knife (10 cm in diameter) into the soil at a regular grid of points every 50 m. Each base sample was paired with three additional samples at three random selected distances out of 0.2 m, 0.5 m, 2.5 m, 5 m, 10 m, 15 m or 24.9 m, in random compass directions, to capture variation in soil fungal composition at finer scales. Before sampling, all litter and loose debris above the sample points was removed from the forest floor. As a result, 396, 396 and 484 soil cores were collected from NBH, ALS and LJ, respectively. Soil samples were immediately stored on ice during collection, then 0.5 g soil was weighed for DNA extraction. Soil samples were stored at −80 °C after collection, and DNA from these samples was extracted within 2 months.

DNA extraction

DNA extractions of soil samples and all fungal strains in this study were performed using the PowerSoil®DNA Isolation Kit (Mobio) according to the manufacturer's instructions. All soil samples were extracted once. DNA concentration was measured using Qubit 3.0 (Invitrogen).

Figure 2. Schematic representation of sequencing strategy at DNBSEQ-G400 platform and Illumina MiSeq platform. Filled squares indicate 1276 topsoil samples for three forest plots, and filled circles indicate the technical replicates in each sequencing run. PCR replicates were labeled by the number in each filled circle.

Amplification, library preparation and sequencing with DNBSEQ-G400 and Illumina MiSeq

Protocols are available at protocols.io [26] and outline the following. We amplified the fungal ITS1 region with primer pairs ITS1Fngs (GGTCATTTAGAGGAAGTAA) and ITS2ngs (TTYRCKRCGTTCTTCATCG); the ITS2 region with forward primer gITS7ngs (GTGARTCATCRARTYTTTG) and reverse primer ITS4ngsUni (CCTSCSCTTANTDATATGC). All amplifications were performed with Kapa Hifi DNA polymerase (Roche).

For soil samples, ITS2 amplicons were randomly and equally distributed into three DNBSEQ runs for both 2×200 bp paired-end (PE200) and 400 bp single-end sequencing (SE400) (Figure 2). The left libraries of 1276 samples were repetitively sequenced two more times by SE400 using different DNBSEQ-G400 machines. Three randomly chosen soil samples from each forest plot (ALS268, LJ105 and NBH217) were all separately amplified three times as PCR replicates. The corresponding nine sequencing libraries were sequenced repetitively by three DNBSEQ runs. Sequencing was carried out using the entire lane of the Illumina MiSeq platform at the University of Minnesota Genomic Center (Figure 2). Amplicons of the 1276 forest soil samples were normalized and pooled randomly in equal molar ratios into eight parallel sequencing runs on the MiSeq platform. Again, soil samples (ALS268, LJ105 and NBH217) were all separately amplified three times as PCR replicates. The corresponding nine sequencing libraries were sequenced repetitively by eight MiSeq runs.

Metabarcoding data processing

On average, 688.9 million high-quality reads were produced per lane on the DNBSEQ platform using 2×200 paired-end mode; 88.2% of bases had a Phred score of at least Q30 for read1, and 86.6% bases had a Phred score of at least Q30 for read2. After barcode demultiplexing, an average of 558,000 reads per sample were obtained. For 400 bp

single-end mode in DNBSEQ, an average of 452.5 million high-quality reads per lane were produced, with 79.4% of bases scoring Q30. An average 736,000 reads per sample were obtained after barcode splitting.

All sequencing data were processed using a Python-based Snakemake pipeline [28]. For paired-end sequencing data (PE200) from DNBSEQ-G400, the sequencing adapter and the primer area were removed by cutadapt (cutadapt, RRID:SCR_011841) [29]. A 42-bp tail representing the large subunit (LSU) was removed from the end of the reverse primer, so the entire amplicon could be aligned globally to the reference. For the DNBSEQ library, about half of the amplicons were sequenced from the 5' end of the forward primer, and the rest from the 5' end of the reverse primer (similar to Bahram *et al.* [6]). These two types of amplicons were separated by recognizing the leading primers using cutadapt, then corrected to the same orientation using seqtk (Seqtk, RRID:SCR_018927) [30]. Low quality reads with a maximum expected error rate >1 and minimum lengths shorter than 100 bp were discarded using Vsearch -fastq_filter function [31]. R1 and R2 reads passing these quality filtering steps were aligned to SILVA (v132) [32] or UNITE (v7.2) databases [33] using BURST v0.99.8 [34], with a 97% sequence similarity threshold in BEST mode (only the first highest hit was reported for each query). Alignment results from R1 and R2 reads were compared to ensure this target continued to be met. Operational taxonomic unit (OTU) counts were calculated based on alignment results for the community profile.

Bioinformatic analysis was slightly modified for single-end sequencing data (SE400) from DNBSEQ-G400. Like PE200 data, reads were corrected to the same orientation by recognizing the leading primers using cutadapt. Reads were trimmed until they reached the maximum expected rate ≤ 1 (-trunclee 1). Reads shorter than 100 bp were then discarded. Community profiles were calculated using the same method as the PE200 data by aligning to the UNITE database.

Reads from MiSeq in PE300 mode were processed with a slightly modified method [28]. Briefly, 41 bp was cut from the 5' end of the reverse primer to remove the LSU region, owing to the 1-bp primer shift. The same quality control and alignment rules were conducted for MiSeq, similar to that described above. Community profiles were calculated using identical parameters as DNBSEQ-G400 PE200 data.

Sequencing data from both MiSeq and DNBSEQ-G400 was normalized to 5000 sequences per sample by picking the sampling result with median richness from 1000 independent rarefactions. Finally, samples with less than 5000 reads were filtered out. Ultimately, 1220 soil samples remained for MiSeq, 1182 for DNBSEQ-G400 PE200, and 1202, 1195, 1183 samples remained for the three sequencing replicates by SE400 modes. The remaining 1144 soil samples across all sequencing platforms were used in all further analyses.

Evaluation of the quality of DNBSEQ-G400 and MiSeq sequencing data

A heterogeneity G-test of goodness-of-fit and a pooled G-test of goodness-of-fit were performed using the RVAideMemoire package in R (RVAideMemoire, RRID:SCR_015657) [35]. Procrustes analysis was performed using the Procrustes and protest functions from the vegan package in R (vegan, RRID:SCR_011950) [36]. Correlation coefficients and statistics were calculated using the Procrustes permutation test in R [37]. Nonmetric multidimensional scaling (NMDS) analysis was performed using the metaMDS function from the vegan package in R [38]. Permutational multivariate analysis of variance

(PERMANOVA) was performed on the abundance of OTUs using Bray–Curtis distance with the *adonis* function from the *vegan* package in R [39]. Significant differences in abundance among major OTUs were examined using the Kruskal–Wallis test in R [37]. All data were represented by *ggplot2* in R (*ggplot2*, RRID:SCR_014601) [40]. All R scripts can be found in GitHub [28].

DATA VALIDATION AND QUALITY CONTROL

The accuracy of amplicon sequencing on DNBSEQ-G400

The performance of the DNBSEQ-G400 for HTS was first examined by 16S V4 amplicons of the broadly used ZymoBIOMICS Microbial Community Standard. The proportions of each taxon in the Zymo mock community were well reproduced (Figure 3a), with no significant difference in the theoretical mock community relative abundance detected by heterogeneity G-test of goodness-of-fit ($P = 0.98$) and pooled G-test of goodness-of-fit ($P = 1$).

An even mock community with seven common fungal species belonging to the same genus was used to assess the classification resolution at species level by DNBSEQ-G400 (Figure 3b). The ITS2 region of the seven species was identified. Some species were assigned to the correct genus but others, for example *A. nidulans*, was identified as *A. quadrilineatus*, and *P. expansum* was identified as *P. marinum*. This was a consequence of identical ITS2 sequences between these species in the UNITE database. Thus, the resolution of the ITS2 marker allowed taxonomic assignments at genus level to be as accurate as possible (Figure 3b). The abundances of five fungi were slightly underestimated, and the overestimation of *T. longibrachiatum* was about 2-fold (Figure 3b).

Three paired combinations in geometric series (1:1, 1:10, and 1:100) of fungal taxa from ITS1 amplicons were also used to evaluate the performance of DNBSEQ-G400 (Figure 3c). Genomic DNA usually contains an undefined number of rRNA copies, so to quantify the accuracy between the amount of input and the results obtained by sequencing, we compared mixtures with the same geometric series, using both raw genomic DNA and PCR products. Generally, PCR product mixtures reproduced the original proportions better ($P = 0.18$ by heterogeneity G-test of goodness-of-fit, and $P = 0.15$ by pooled G-test of goodness-of-fit) than raw genomic DNA ($P = 0.00$ by heterogeneity G-test of goodness-of-fit) (Figure 3c). This was logical since preferential PCR amplification of certain templates skewed the representation of taxa with amplification with prior-mixed DNA. Two pairs of combinations, *L. edodes* with *F. velutipes* and *L. edodes* with *P. eryngii*, were generally restored better than with the *L. edodes* and *S. cerevisiae* combination (Figure 3c). More specifically, *L. edodes* was overestimated (Figure 3c).

Reproducibility and precision of two DNBSEQ-G400 sequencing modes

ITS2 amplicons from forest soil samples were examined using PE200 and SE400 sequencing modes of DNBSEQ-G400 (Figure 4). Thirty-three classes were identified by both modes from the ALS forest plot. Forty-two and forty-three fungal classes were detected by PE200 and SE400, respectively, from the LJ forest plot, of which 40 were shared by both modes. Finally, from the NBH plot, 42 overlapped classes were found by both modes, and 43 and 45 fungal genera were identified by PE200 and SE400, respectively. Heatmap analysis showed the frequency of occurrence of classes remained similar across the sequencing platforms,

Figure 3. Histograms showed compositions of mock communities detected by DNBSEQ-G400 PE200 mode, (a) ZymoBIOMICS Microbial Community Standards, (b) equal genomic DNA mixtures of seven fungi mock, (c) Fungus mixed with differential ratios. The expected ratios were shown with dotted lines for b and c. Error bars indicate the standard deviation.

especially for the top 20 most abundant classes (Figure 4a). SE400 has higher sensitivity since it recovered a few more fungal classes (GS18, Kickxellomycetes, Paraglomeromycetes and Pucciniomycetes); however, PE200 performed better on the ALS and LJ samples, with more recovered classes of Zoopagomycetes and Archaeorhizomycetes. More variable classes, such as the Kickxellomycetes, Entorrhizomycetes, and Entomophthoromycetes, were found in low abundance (Figure 4a).

We next compared the two sequencing modes at the genus level; most detected genera overlapped within PE200 and SE400 (Figure 4b). Most of the fungal genera that differed between the two modes were the low abundance ones (genera abundance <0.1%), with a few exceptions. *Archaeorhizomyces* was the only genus highly detected by PE200 and barely detected by SE400. A few more genera were more abundantly detected by SE400 than

Figure 4. (a) All classes from the three forest plots detected by DNBSEQ-G400 are represented as log₁₀ (frequency) by heatmap. (b) Venn diagrams show the overlap of identified fungal genera by DNBSEQ-G400 PE200 and SE400 modes. ALS = AiLaoShan, LJ = LiJiang, NBH = NaBanHe.

PE200, including *Alpinaria*, *Purpureocillium*, *Cladorrhinum*, *Didymella*, *Talaromyces*, *Phoma*, *Helicodendron*, *Volucrispora*, *Stagonosporopsis*.

Sample-to-sample comparisons allowed us to further quantify the proportion difference of detected OTUs by the two sequencing modes on a global scale. We estimated the similarity of soil taxa beta-diversity between PE200 and SE400 using Procrustes analysis [41]. The two sequencing modes resulted in a high correlation coefficient among microbial taxa using the Procrustes permutation test (ALS = 0.948, LJ = 0.911, NBH = 0.867). These high similarities support the view that our findings are real and independent of the possible methodological biases introduced by the different sequencing modes.

To further examine the sequencing reproducibility of DNBSEQ-G400, the same library of all soil samples was sequenced in triplicate in SE400 mode (replicates SE400_1, SE400_2, SE400_3), with the second and third replicate sequenced on a different DNBSEQ-G400

Figure 5. (a) Non-metric multidimensional scaling (NMDS) analysis and (b) Permutational multivariate analysis of variance (PERMANOVA) results showed no significant sequencing mode, run and PCR on the composition of fungal communities in three forest plots generated by DNBSEQ-G400 PE200 and SE400. (c) bar chart shows the percentage of major OTUs in three forest plots at genus level in DNBSEQ-G400 sequencer. ALS = AiLaoShan, LJ = LiJiang, NBH = NaBanHe.

machine. Consistent performance of DNBSEQ-G400 was observed. The average correlation coefficients of fungal composition from the three sequencing repeats were 0.96, 0.99 and 0.99 for ALS, LJ and NBH, respectively. Furthermore, the three replicates of all soil samples were significantly correlated (the average correlation coefficient was 0.98 for SE400_1 with SE400_2, 0.99 for SE400_1 with SE400_3, and 0.99 for SE400_2 with SE400_3; $P < 0.001$).

Consistency of DNBSEQ-G400 evaluated by technical replicates

To evaluate the sequencing stability of the DNBSEQ-G400 platform, we performed NMDS ordination on the technical replicates of samples ALS268, LJ105 and NBH217, which were produced by three independent sequencing runs. Both PCR and run replicate analysis results were tightly clustered, and all technical replicates were separated according to their geological locations (Figure 5a). The observations were supported by higher R^2 value of samples organized by plot ($R^2 = 0.956$, $P = 0.001$) compared with samples organized by platform ($R^2 = 0.016$), PCR ($R^2 = 0.002$) and sequencing runs ($R^2 = 0.0007$) according to PERMANOVA (Figure 5b).

Technical replicates sequenced using two modes of DNBSEQ-G400 shared major fungal genera, but had varied relative abundance (Figure 5c). The differences in major fungal groups recovered by PE200 and SE400 were all less than 3%, except for *Dothideomycetes* in LJ, and *Psathyrellaceae* in NBH. Sequencing mode did not contribute significantly to the

Figure 6. (a) Alpha-diversity was displayed by species richness (Chao1), Shannon and Simpson diversity index in all soil samples at different forest plots sequenced by MiSeq and DNBSEQ-G400. Data are visualized as box-and-whisker plots showing the median and the interquartile (midspread) range (boxes containing 50% of all values), the whiskers (representing the 25 and 75 percentiles) and the extreme data points. Numbers above boxes indicate significant differences (p value) by Kruskal-Wallis test. (b) Beta-diversity was displayed by NMDS of all soil species from 3 forest plots sequenced by MiSeq and DNBSEQ-G400.

observed β -diversity comparing to site ($P = 0.77$ for ALS, $P = 0.28$ for LJ, $P = 0.56$ for NBH by PERMANOVA) (Figure 5a, b).

Taxonomic composition comparison between MiSeq and DNBSEQ-G400 with the same soil DNA samples

The ITS2 region of the same soil DNA samples were previously amplified with primer pairs 5.8SR_Nextera and ITS4_Nextera (dual-index [DI] approach) and were sequenced by Illumina MiSeq (unpublished). To evaluate the comparability and reuse availability of metabarcoding data from different sequencing platforms, we compared the two sequencing results in this study. Owing to a wider coverage of universal primers [42] in DNBSEQ, more species were recovered by DNBSEQ than MiSeq based on several alpha-diversity indexes (Figure 6a). Microbial profiles of samples from the same forest plots were clustered together by beta-diversity analysis (Figure 6b). A small but significant difference ($R^2 = 0.0377$) among sequencing platforms was observed by PERMANOVA ($P = 0.001$). Taxonomic characterization was affected by the selection of PCR amplification primers; this is consistent with previous reports [43, 44]. Since the amount of metabarcoding performed with DNBSEQ is likely to increase in the future, a comprehensive comparison of different sequencing platforms, such as the comparison described here, will provide useful information for design and standardization of experimental conditions.

DISCUSSION

We comprehensively assessed the amplicon sequencing results of positive mock communities and a large number of soil samples using the DNBSEQ-G400 sequencing

platform. The DNBSEQ-G400 platform could restore the Zymo mock communities with negligible bias in our study (Figure 3a). High reproducibility of mock communities was also observed on MiSeq based on a previous study [45], showing an equivalent performance of the two sequencers. The substantial differences in quantification detected in an even mock community with seven fungi (Figure 3b) suggested sequencing bias against certain taxa. However, since this mock community sample was mixed based on genomic DNA content, rRNA copy number in genomic DNA might affect the detected taxonomic composition. In communities with different ratios of four fungi (Figure 3c), the abundance of *L. edodes* was overestimated in a combination of *L. edodes* and *S. cerevisiae*. The bias of relative proportion is best explained by the ITS1 length differences of the two species (*L. edodes* ITS1 is 243 bp, *S. cerevisiae* is 366 bp). Since DNBSEQ-G400 uses rolling-cycle amplification of sequencing signals, short fragments might be amplified more than long sequences in a given time, giving the impression of a higher abundance of *L. edodes*. This corresponds to much smaller biases in the abundance estimates of pairs of species with comparable ITS1 lengths: *F. velutipes* (247 bp) and *P. eryngii* (252 bp) relative to *L. edodes* (Figure 3c). This overestimation of taxa with short amplicons seems to be a general bias in metabarcoding; it was also observed as the main source of bias in observed community composition in other sequencers of Illumina and PacBio RSII and Sequel I [12].

Comprehensive comparisons of ITS2 amplicons from >1000 soil samples revealed highly comparable performance between the PE200 and SE400 modes of DNBSEQ-G400. Different fungal classes were confined to classes with very low abundance (Figure 4). A few profound differences between platforms at class-level abundances were detected; for example, the Archaeorhizomycetes class was highly detected by DNBSEQ PE200 and MiSeq in ALS and LJ, but not by SE400. Archaeorhizomycetes often have highly divergent sequences [42]; the missed detection in SE400 mode suggests the importance of 3' sequences for assignment of certain taxa. Genera that were highly detected by SE400 modes but barely detected by PE200 were also poorly detected by PE200. Because we aligned R1 and R2 reads separately in paired sequencing modes, the read length for alignments is longest in SE400, and shorter in MiSeq, followed by PE200. Thus, differentially detected abundances in certain genera is probably because the read length affects the coverage of the recognition area during alignment. Taken together, these results suggest that care should be taken while jointly analyzing metabarcoding data from different sequencers or different sequencing modes.

A higher R^2 value was observed while evaluating run effects on technical replicates, suggesting that a large effect on taxonomic profile is caused by sequencing run (Figure 5b). As previous reports have showed Illumina sequencer run bias against certain fungal and archaeal taxa [17], technical controls must be included when executing large-scale sequencing efforts.

To date, comparable sequencing performance between DNBSEQ sequencers and Illumina sequencers has been reported in many studies, including palaeogenomic sequencing [21], metagenomics [24], and single-cell transcriptomics [25]. In contrast, significant differences in taxonomic composition based on short reads produced by MiSeq and DNBSEQ again indicated the potential problem of repeatability of metabarcoding caused by different primers and library preparation procedures. This has also been observed in some other studies [15, 18].

The rapid advances in sequencing technologies allows discovery of unknown microorganisms in natural complex samples, using techniques such as shotgun

metagenomics and long-read DNA sequencing methods [46–48]. However, it remains a difficult task for scientists to reflect the true taxonomic composition of a soil sample, owing to the largely undiscovered diversity; many undescribed species are genomically complex and have no reference sequences in the databases [49]. Compared with metagenomics, metabarcoding can provide sufficient taxonomic information in a more cost-effective manner, which makes it possible for researchers to expand their surveys of microbial communities to both large spatial and long temporal scales.

REUSE POTENTIAL

Here, we evaluated the performance of the DNBSEQ-G400 platform for high-throughput metabarcoding in investigating the composition of soil microbial communities. The platform produced largely similar results as Illumina MiSeq (the currently dominant platform used for this purpose), but with a higher throughput and at potentially lower cost. The sequencing results of bacterial and fungal mock communities in this study could provide a reference for future evaluation studies. Samples from three forest plots were from a long-term ongoing project. The fungal communities data can support further studies of the forest ecosystem in southwest China. The dataset in this study represents a substantial benchmark for opportunities and challenges presented by combining datasets produced by different sequencing platforms.

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

Sequencing data for all samples have been deposited to the China National GeneBank DataBase (CNGBdb) [50], with the accession number CNP0000069, and the European Bioinformatics Institute (EBI) with the accession number PRJEB28018. All supporting data are deposited in the *GigaScience* GigaDB repository [27].

DECLARATIONS

LIST OF ABBREVIATIONS

ALS: AiLaoShan; cPAS: combinatorial Probe-Anchor Synthesis; DNB: DNA Nanoball; HTS: high-throughput sequencing; ITS, internal transcribed spacer; LJ: LiJiang; LSU, large subunit; NBH: NaBanHe; NMDS: Non-metric multidimensional scaling; OTU: operational taxonomic unit; PE200: paired-end 200bp; SE400: single-end 400bp.

FUNDING

This research was supported by the National Natural Science Foundation of China (grant numbers 31570380, 31300358, 31100312), the West Light Foundation of the Chinese Academy of Sciences to Yue-Hua Hu, the CAS 135 program (grant number 2017XTBG-T01), the Natural Science Foundation of Yunnan Province (grant number 2015FB185), the Southeast Asia Biodiversity Research Institute, Chinese Academy of Sciences (grant number 2016CASSEABRIQG002), the National Key Basic Research Program of China (grant number 2014CB954100), and the Applied Fundamental Research Foundation of Yunnan Province (grant number 2014GA003).

AUTHORS' CONTRIBUTIONS

X.S., L.X., Y.H. and Z.S. designed the study. Y.H. collected all soil samples from the three forests. X.S., H.M., H.Z., X.L. performed the amplification experiments. J.W., J.L., X.W. and

M.G. performed the DNBSEQ-G400 sequencing events. X.S., C.F., Y.J., analyzed the sequencing data. X.S., Y.H. and Z.S. wrote the manuscript.

ACKNOWLEDGEMENTS

The authors are very grateful to colleagues at BGI-Shenzhen and China National Genebank (CNCB), Shenzhen for library constructions and discussions. We thank Professor Cene Gostinčar for helpful comments.

REFERENCES

- 1 Venter JC et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 2004; 304(5667): 66–74, doi:10.1126/science.1093857.
- 2 Land M et al. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, 2015; 14: 141–161, doi:10.1007/s10142-015-0433-4.
- 3 Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF, Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature*, 2018; 558: 440–444, doi:10.1038/s41586-018-0207-y.
- 4 Schoch CL et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl Acad. Sci. USA*, 2012; 109(16): 6241–6246, doi:10.1073/pnas.1117018109.
- 5 Hamady M, Lozupone C, Knight R, Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.*, 2010; 4: 17–27, doi:10.1038/ismej.2009.97.
- 6 Bahram M et al. Structure and function of the global topsoil microbiome. *Nature*, 2018; 560: 233–237, doi:10.1038/s41586-018-0386-6.
- 7 Leung MHY, Tong X, Tong JCK, Lee PKH, Airborne bacterial assemblage in a zero carbon building: A case study. *Indoor Air*, 2018; 28(1): 40–50, doi:10.1111/ina.12410.
- 8 Garcia-Lopez E, Rodriguez-Lorente I, Alcazar P, Cid C, Microbial communities in coastal glaciers and tidewater tongues of Svalbard archipelago, Norway. *Front. Mar. Sci.*, 2019; 5: 512, doi:10.3389/fmars.2018.00512.
- 9 Varliero G, Bienhold C, Schmid F, Boetius A, Molari M, Microbial diversity and connectivity in deep-sea sediments of the South Atlantic Polar Front. *Front. Microbiol.*, 2019; 10: 665, doi:10.3389/fmicb.2019.00665.
- 10 Klindworth A et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.*, 2013; 41(1): e1, doi:10.1093/nar/gks808.
- 11 Gohl DM et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.*, 2016; 34: 942–949, doi:10.1038/nbt.3601.
- 12 Castaño C et al. Optimized metabarcoding with Pacific Biosciences enables semi-quantitative analysis of fungal communities. *New Phytol.*, 2020; 228(3): 1149–1158, doi:10.1111/nph.16731.
- 13 Lindahl BD et al. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytol.*, 2013; 199(1): 288–299, doi:10.1111/nph.12243.
- 14 Anslan S, Nilsson RH, Wurzbacher C, Baldrian P, Tedersoo L, Bahram M, Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycKeys*, 2018; 39: 29–40, doi:10.3897/mycokeys.39.28109.
- 15 Bolyen E et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, 2019; 37: 852–857, doi:10.1038/s41587-019-0209-9.
- 16 Yeh Y-C, Needham DM, Sieradzki ET, Fuhrman JA, Taxon disappearance from microbiome analysis reinforces the value of mock communities as a standard in every sequencing run. *mSystems*, 2018; 3(3): 00023-18, doi:10.1128/msystems.00023-18.
- 17 Song Z, Schlatter D, Gohl DM, Kinkel LL, Run-to-run sequencing variation can introduce taxon-specific bias in the evaluation of fungal microbiomes. *Phytobiomes J.*, 2018; 2: 165–170, doi:10.1094/PBIOMES-09-17-0041-R.
- 18 Tedersoo L, Drenkhan R, Anslan S, Morales-Rodriguez C, Cleary M, High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. *Mol. Ecol. Resour.*, 2018; 19(1): 47–76, doi:10.1111/1755-0998.12959.

- 19 Fehlmann T et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenetics*, 2016; 8: 123. doi:10.1186/s13148-016-0287-1.
- 20 Huang J et al. A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience*, 2017; 6(5): gix024. doi:10.1093/gigascience/gix024.
- 21 Mak SST et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience*, 2017; 6(8): gix049. doi:10.1093/gigascience/gix049.
- 22 Zou Y et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.*, 2019; 37: 179–185, doi:10.1038/s41587-018-0008-8.
- 23 Chen C et al. The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. *Nat. Commun.*, 2017; 8: 875. doi:10.1038/s41467-017-00901-0.
- 24 Fang C et al. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience*, 2018; 7(3): gix133. doi:10.1093/gigascience/gix133.
- 25 Natarajan KN et al. Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol.*, 2019; 20: 70. doi:10.1186/s13059-019-1676-5.
- 26 Sun X, Hu Y, Song Z, Protocols for “Efficient and stable metabarcoding sequencing data using DNBSEQ-G400 sequencer validated by comprehensive community analyses”. protocols.io. 2021; <http://dx.doi.org/10.17504/protocols.io.brtpm6mn>.
- 27 Sun X et al. Efficient and stable metabarcoding sequencing data using DNBSEQ-G400 sequencer validated by comprehensive community analyses. GigaScience Database. 2021; <http://dx.doi.org/10.5524/100824>.
- 28 Sun X, Metabarcoding data processing. 2020; <https://github.com/XiaohuanSun/metabarcoding-data-processing>.
- 29 Martin M, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, 2011; 17(1): 10–12, doi:10.14806/ej.17.1.200.
- 30 Li H, Seqtk (Seqtk-1.3 (r106)). 2018; <https://github.com/lh3/seqtk>.
- 31 Rognes T, Flouri T, Nichols B, Quince C, Mahé F, VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 2016; 4: e2584. doi:10.7717/peerj.2584.
- 32 Quast C et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.*, 2013; 41(D1): D590–D596, doi:10.1093/nar/gks1219.
- 33 UNITE Community. UNITE QIIME release. Version 01.12.2017. PlutoF. 2017; <https://doi.org/10.15156/BIO/587481>.
- 34 Al-Ghalith G, Knights D, BURST enables optimal exhaustive DNA alignment for big data. Zenodo. 2020; <https://doi.org/10.5281/zenodo.806850>.
- 35 Mangiafico S, An R Companion for the Handbook of Biological Statistics. Version 1.3.3. New Brunswick, NJ: Rutgers Cooperative Extension 2015.; pp. 32–34.
- 36 Peres-Neto PR, Jackson DA, How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the Mantel test. *Oecologia*, 2001; 129: 169–178, doi:10.1007/s004420100720.
- 37 Xiong R, Blot K, Meulenet JF, Dessirier JM, Permutation tests for generalized procrustes analysis. *Food Qual. Preference*, 2008; 19(2): 146–155, doi:10.1016/j.foodqual.2007.03.003.
- 38 Oksanen J et al. vegan: Community Ecology Package. R package version 2.5-7. 2020; <https://CRAN.R-project.org/package=vegan>.
- 39 Warton DI, Wright ST, Wang Y, Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol. Evol.*, 2012; 3(1): 89–101, doi:10.1111/j.2041-210X.2011.00127.x.
- 40 Gómez-Rubio V, ggplot2 – Elegant graphics for data analysis (2nd Edition). *J. Stat. Softw.*, 2017; 77: doi:10.18637/jss.v077.b02.
- 41 Oksanen J, Multivariate analysis of ecological communities in R: vegan tutorial. 2015; <https://john-quensen.com/wp-content/uploads/2018/10/Oksanen-Jari-vegantutor.pdf>.
- 42 Tedersoo L, Tooming-Klunderud A, Anslan S, PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *New Phytol.*, 2018; 217(3): 1370–1385, doi:10.1111/nph.14776.
- 43 Gohl DM et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.*, 2016; 34: 942–949, doi:10.1038/nbt.3601.

- 44 **Castle SC et al.** DNA template dilution impacts amplicon sequencing-based estimates of soil fungal diversity. *Phytobiomes J.*, 2018; 2: 100–107, doi:[10.1094/PBIOMES-09-17-0037-R](https://doi.org/10.1094/PBIOMES-09-17-0037-R).
- 45 **McGovern E, Waters SM, Blackshields G, McCabe MS**, Evaluating established methods for Rumen 16S rRNA amplicon sequencing with mock microbial populations. *Front. Microbiol.*, 2018; 9: 1365. doi:[10.3389/fmicb.2018.01365](https://doi.org/10.3389/fmicb.2018.01365).
- 46 **Deamer D, Akeson M, Branton D**, Three decades of nanopore sequencing. *Nat. Biotechnol.*, 2016; 34: 518–524, doi:[10.1038/nbt.3423](https://doi.org/10.1038/nbt.3423).
- 47 **Eid J et al.** Real-time DNA sequencing from single polymerase molecules. *Science*, 2009; 323(5910): 133–138, doi:[10.1126/science.1162986](https://doi.org/10.1126/science.1162986).
- 48 **Wang O et al.** Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.*, 2019; doi:[10.1101/gr.245126.118](https://doi.org/10.1101/gr.245126.118).
- 49 **Martiny AC**, The ‘1% culturability paradigm’ needs to be carefully defined. *ISME J.*, 2020; 14: 10–11, doi:[10.1038/s41396-019-0507-8](https://doi.org/10.1038/s41396-019-0507-8).
- 50 **Chen FZ et al.** CNGBdb: China National GeneBank DataBase, Hereditas. 2020; <http://dx.doi.org/10.16288/j.ycz.20-080>.