

## DATA NOTE

## Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing

Chao Fang<sup>1,2,3,†</sup>, Huanzi Zhong<sup>1,2,4,†</sup>, Yuxiang Lin<sup>1,2,3</sup>, Bing Chen<sup>1,2,3</sup>,  
Mo Han<sup>1,2,3</sup>, Huahui Ren<sup>1,2,3</sup>, Haorong Lu<sup>1,2</sup>, Jacob M. Luber<sup>5,6,7,8,9</sup>, Min Xia<sup>1,2</sup>,  
Wangsheng Li<sup>1,2</sup>, Shayna Stein<sup>6,10,11</sup>, Xun Xu<sup>1,2</sup>, Wenwei Zhang<sup>1</sup>,  
Radoje Drmanac<sup>1</sup>, Jian Wang<sup>1,12</sup>, Huanming Yang<sup>1,12</sup>,  
Lennart Hammarström<sup>13</sup>, Aleksandar D. Kostic<sup>7,8,9</sup>, Karsten Kristiansen<sup>1,2,4</sup>  
and Junhua Li<sup>1,14,2,3,\*</sup>

<sup>1</sup>BGI-Shenzhen, Building 11, Beishan Industrial Zone, Yantian, Shenzhen 518083, China, <sup>2</sup>China National GeneBank, BGI-Shenzhen, Dapeng New District, Shenzhen 518120, China, <sup>3</sup>Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Building 11, Beishan Industrial Zone, Yantian, Shenzhen 518083, China, <sup>4</sup>Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen Biocenter, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark, <sup>5</sup>Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, BIG Program Office, 10 Shattuck Street, Countway Library, 4th Floor, Boston, MA 02115, USA, <sup>6</sup>Graduate School of Arts and Sciences, Harvard University, Richard A. and Susan F. Smith Campus Center, 1350 Massachusetts Avenue, Suite 350, Cambridge, MA 02138-3654, USA, <sup>7</sup>Section on Pathophysiology and Molecular Pharmacology, Joslin Diabetes Center, One Joslin Place, Boston, MA 02215, USA, <sup>8</sup>Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center, One Joslin Place, Boston, MA 02215, USA, <sup>9</sup>Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, 4th Floor, Boston, MA 02115, USA, <sup>10</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, MA 02215-5450, USA, <sup>11</sup>Department of Biostatistics, Harvard TH Chan School of Public Health, 655 Huntington Avenue, Building 2, 4th Floor, Boston, MA 02115, USA, <sup>12</sup>James D. Watson Institute of Genome Sciences, No.51, Zhijiang Road, Xihu District, Hangzhou, Zhejiang Province, 310058, P. R. China, <sup>13</sup>Division of Clinical Immunology and Transfusion Medicine, Department of Laboratory Medicine, Clinical University Hospital, Huddinge, SE-14186 Stockholm, Sweden and <sup>14</sup>School of Bioscience and Biotechnology, South China University of Technology, No.381 Wushan Road, Tianhe District, Guangzhou, Guangdong 510630, China

\*Corresponding author: Junhua Li, BGI-Shenzhen, Building 11, Beishan Industrial Zone, Yantian, Shenzhen 518083, China. Tel: +86 13929566296; E-mail: [lijunhua@genomics.cn](mailto:lijunhua@genomics.cn)

†Equal contribution

Received: 26 August 2017; Revised: 21 November 2017; Accepted: 19 December 2017

© The Author(s) 2018. Published by Oxford University Press. 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.

## Abstract

**Background:** More extensive use of metagenomic shotgun sequencing in microbiome research relies on the development of high-throughput, cost-effective sequencing. Here we present a comprehensive evaluation of the performance of the new high-throughput sequencing platform BGISEQ-500 for metagenomic shotgun sequencing and compare its performance with that of 2 Illumina platforms. **Findings:** Using fecal samples from 20 healthy individuals, we evaluated the intra-platform reproducibility for metagenomic sequencing on the BGISEQ-500 platform in a setup comprising 8 library replicates and 8 sequencing replicates. Cross-platform consistency was evaluated by comparing 20 pairwise replicates on the BGISEQ-500 platform vs the Illumina HiSeq 2000 platform and the Illumina HiSeq 4000 platform. In addition, we compared the performance of the 2 Illumina platforms against each other. By a newly developed overall accuracy quality control method, an average of 82.45 million high-quality reads (96.06% of raw reads) per sample, with 90.56% of bases scoring Q30 and above, was obtained using the BGISEQ-500 platform. Quantitative analyses revealed extremely high reproducibility between BGISEQ-500 intra-platform replicates. Cross-platform replicates differed slightly more than intra-platform replicates, yet a high consistency was observed. Only a low percentage (2.02%–3.25%) of genes exhibited significant differences in relative abundance comparing the BGISEQ-500 and HiSeq platforms, with a bias toward genes with higher GC content being enriched on the HiSeq platforms. **Conclusions:** Our study provides the first set of performance metrics for human gut metagenomic sequencing data using BGISEQ-500. The high accuracy and technical reproducibility confirm the applicability of the new platform for metagenomic studies, though caution is still warranted when combining metagenomic data from different platforms.

**Keywords:** BGISEQ-500; quantitative metagenomic analyses; next-generation sequencing

## Data Description

To evaluate the performance of the BGISEQ-500 platform for metagenomic sequencing, stool samples were collected from 20 healthy adults in the Stockholm regional area. Fecal DNA was extracted and sequenced on the BGISEQ-500 sequencer. The quality of raw data was evaluated and filtered by an in-house developed quality control (QC) pipeline to obtain high-quality data (see the Methods and Additional files 2 and 3 for details). Qualitative and quantitative analyses were conducted to evaluate the intra-platform reproducibility. In addition, data obtained by sequencing of the same fecal DNA samples on the HiSeq 2000 platform and the HiSeq 4000 platform were included for cross-platform comparison (see Fig. 1 and the Methods for details).

## Methods

### Healthy subject enrollment and sampling

Twenty Swedish healthy adults living in the Stockholm regional area were enrolled as part of a large study cohort: “Characterization of the intestinal microbiome in patients with IgA deficiency.” The detailed inclusion and exclusion criteria were as follows: (1) no diagnosed gastrointestinal problems (inflammatory bowel disease, celiac disease, or lactose intolerance); (2) no antibiotic treatment for at least 60 days; (3) no intake of yoghurt products for at least 5 days prior to sampling. Feces specimens were collected at home by each participant, immediately frozen in the home freezer, and transferred to the laboratory on dry ice and kept frozen at  $-80^{\circ}\text{C}$  until processed.

### DNA extraction

The stool DNA was extracted in accordance with the MetaHIT protocol as described previously [1]. The DNA concentration was estimated by Qubit (Invitrogen).

### Library preparation and sequencing

#### For sequencing using the BGISEQ-500 platform

Five hundred ng of input DNA was used for library formation and fragmented ultrasonically with Covaris E220 (Covaris, Brighton, UK), yielding 300 to 700 bp of fragments. Sheared DNA with

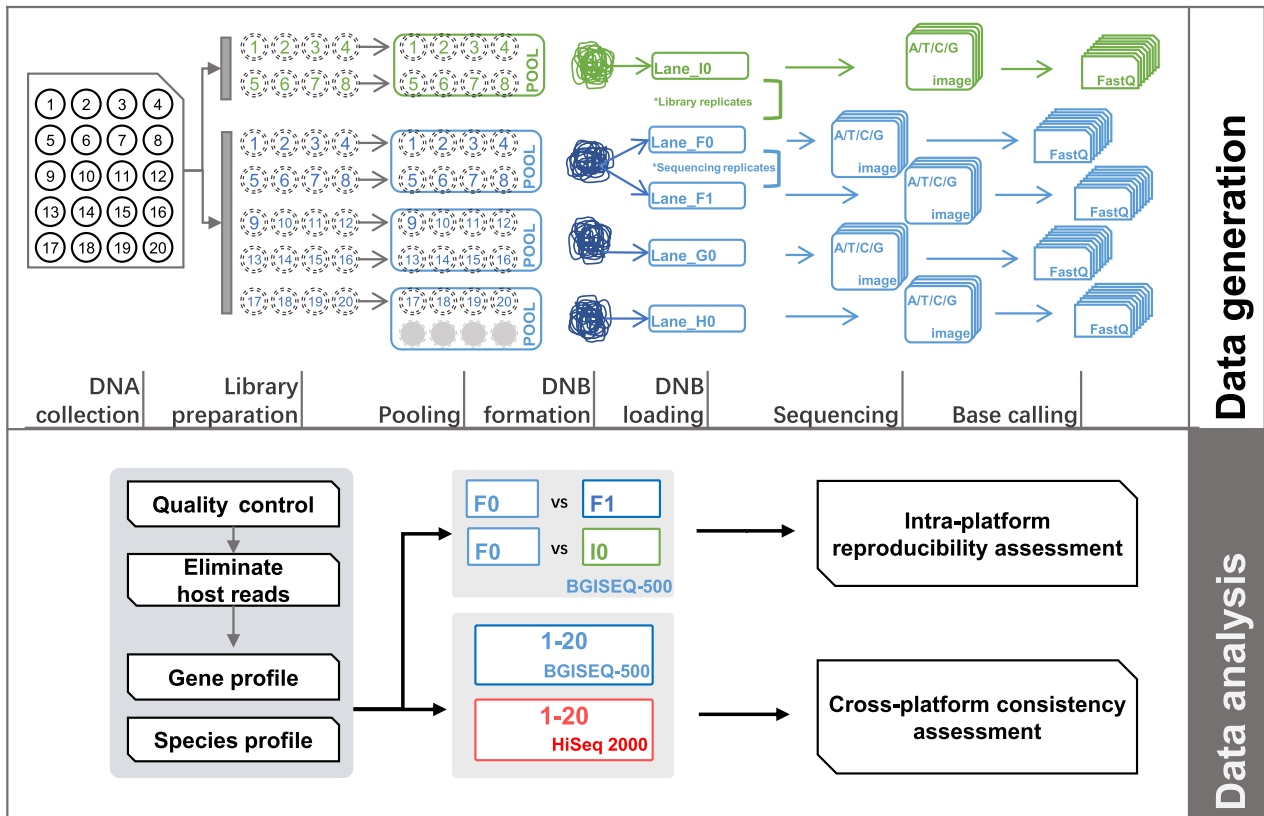
out size selection was purified with an Axygen™ AxyPrep™ Mag PCR Clean-Up Kit. An equal volume of beads was added to each sample, and DNA was eluted with 45  $\mu\text{L}$  TE buffer. We performed end-repairing and A-tailing with a 2:2:1 mixture of T4 DNA polymerase (ENZYMATICS™ P708–1500), T4 polynucleotide kinase (ENZYMATICS™ Y904–1500), and rTaq DNA polymerase (TAKARA™ R500Z). Twenty ng of purified DNA was used, and enzymes were heat-inactivated at  $75^{\circ}\text{C}$ . Adaptors with specific barcodes (Ad153 2B) were ligated to the DNA fragment by T4 DNA ligase (ENZYMATICS™ L603-HC-1500) at  $23^{\circ}\text{C}$ . After the ligation, PCR amplification was carried out. Fifty-five ng of purified PCR products was denatured at  $95^{\circ}\text{C}$  and ligated by T4 DNA ligase (ENZYMATICS™ L603-HC-1500) at  $37^{\circ}\text{C}$  to generate a single-strand circular DNA library. Eight barcoded libraries were pooled in equal amounts to make DNA Nanoballs (DNB). Each DNB was loaded into 1 lane for sequencing.

Sequencing was performed according to the BGISEQ-500 protocol (SOP AO) employing the SE100 mode as described previously [2]. For reproducibility analyses, DNA from the same 8 subjects (S01–S08) were processed twice following the same protocol as described above to serve as library replicates, and 1 of the DNBs from the same 8 subjects was sequenced twice as sequencing replicates. As shown in Fig. 1, a total of 36 datasets were generated using the BGISEQ-500 platform.

#### For sequencing using the HiSeq 2000 and 4000 platforms

One  $\mu\text{g}$  DNA was sheared to 350 bp using the Covaris LE220 (Covaris, Inc., Woburn, MA, USA), size selected using AMPure XP beads (Beckman Coulter, Brea, CA, USA). Adapters were then ligated. Twenty libraries were prepared following BGI's protocol [3]. Five libraries were pooled for each lane, and sequencing was performed on an Illumina HiSeq 2000 using V3 reagents for 100 bp of paired-end reads. The base-calling was performed using Illumina pipeline Real Time Analysis (RTA; version 1.13.48) to process the raw fluorescent images and call sequences.

For both platforms, raw data containing multiple subjects were first split into separate files based on subject-specific barcodes. The samples of the 20 subjects sequenced by both BGISEQ-500 and HiSeq 2000 were used to assess the compatibility of metagenomic data across these platforms. For comparison, only the forward reads from HiSeq 2000 were used.



**Figure 1:** Schematic model summarizing the study design and analysis strategy. Schematic diagram depicting the process of data generation, including collection of fecal samples and extraction of DNA from 20 healthy subjects, library preparation, and sequencing strategy for BGISEQ-500 and HiSeq 2000. Each circle indicates 1 independent subject, with subject ID shown in the circle. For BGISEQ-500, each sample was sheared and tagged with a unique barcode to prepare libraries, then equal amounts of DNA fragments from 8 samples were pooled together for DNB formation, loading, and sequencing. In total, 20 samples were sequenced in 3 lanes (F0, G0, and H0). Of them, DNA from 8 subjects (S01-S08) was utilized to perform library construction and sequencing twice; the corresponding 8 paired datasets from lane I0 (green) and lane F0 (blue) were considered library replicates. DNBs from the same 8 subjects were loaded and sequenced twice to generate 8 paired sequencing replicates (lane F0 and lane F1). Twenty datasets from HiSeq 2000 were also generated in this study. The detailed assessment and comparison analyses of metagenomic datasets between intra- and inter-platforms are shown below.

### Quality control of sequencing data

To evaluate the data quality from the 2 different sequencing platforms, raw FASTQ reads from BGISEQ-500 and HiSeq 2000 were converted into Sanger Phred+33 quality score format and Phred+64 quality score format, respectively [4]. Quality assessment by base position revealed lower-quality scores in the beginning of raw reads from the HiSeq 2000 platform compared with BGISEQ-500 and a gradually decreasing trend of quality toward the 3'-end of reads on both platforms (Additional file 1). Considering that a routinely tail-trimming QC pipeline would not be sensitive to detection and filtering of reads with randomly distributed low-quality bases, we developed an overall accuracy (OA) control strategy for quality adjustment (Additional file 2). By using this approach, 96.06% of the raw reads remained as high-quality reads, which attained an average length of 85 bp, with 90.56% of bases scoring Q30 and above. The parameters of sequencing performance both before and after the QC process are presented in Additional file 3.

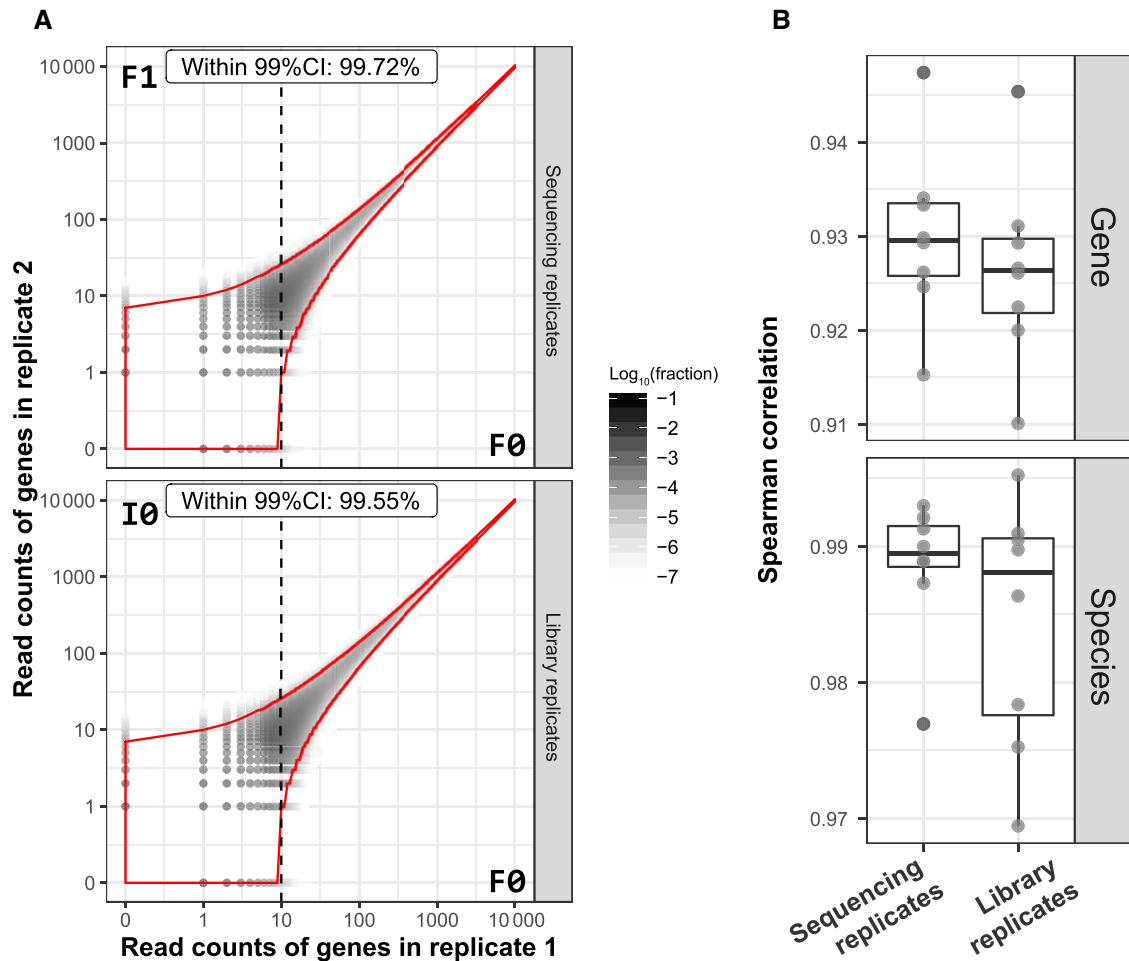
### Alignment and quantification of metagenome content

The high-quality reads of the BGISEQ and HiSeq platforms were then aligned to hg19 using SOAP2.22 (identity  $\geq 0.9$ ) to remove human reads (SOAPaligner/soap2, [RRID:SCR\\_005503](https://doi.org/10.1093/bioinformatics/btt053)) [5]. The retained clean reads were aligned to the integrated gene catalog

(IGC) by using SOAP2.22 (identity  $\geq 0.95$ ) [5]. As shown in Additional file 4, the clean reads from BGISEQ-500 reached an average IGC mapping rate of 77.77% and an average unique mapping rate of 63.27%, which are comparable to the mapping rates of reads from the HiSeq 2000 platform. The IGC mapping ratio of subject S01 (54.58%) was significantly lower in the HiSeq 2000 dataset than in the BGISEQ-500 dataset (Additional file 4). Therefore, we eliminated subject S01 for subsequent analysis. To eliminate the influence of different numbers of reads per sample in intra- or cross-platform analyses, uniquely mapped reads were down-sized to 20 million for each subject. Gene relative abundance (RA) was calculated based on the down-sized mapped reads, as previously described [1]. Relative species abundance in each sample was assessed using MetaPhlan2 [6].

### Intra-platform reproducibility

To estimate the probability distribution of gene occurrence in duplicate experiments, we assessed the expected read count fluctuations based on 20 million IGC uniquely mapped reads (Supplementary Methods). As shown in Fig. 2A, more than 99.5% of genes in replicate 1 (F0) exhibited the expected read count fluctuations in the corresponding sequence replicate 2 (F1) and library replicate 2 (I0; 99% confidence interval [CI]). This indi-



**Figure 2:** Evaluation of intra-platform reproducibility. **A**, Detecting mapped read count fluctuations of genes between intra-platform replicates. Unique IGC mapped reads were downsampled to 20 million for each subject, and the read count fluctuations were estimated (Supplementary Methods). The x-axis represents mapped read counts of a gene in replicate 1 (F0), and the y-axis represents mapped read counts of that gene in replicate 2 (F1 as sequencing replicate and I0 as library replicate). The area bordered by the red line represents the 99% confidence interval (CI) of genes showing the expected read count fluctuations in their replicates. The dashed line indicates that, at 99% CI, genes with greater than or equal to 10 reads in replicate 1 (x-axis) could be detected (with mapped reads greater than or equal to 1) in replicate 2 (y-axis). **B**, Spearman's correlation coefficient. Genes with greater than or equal to 10 mapped reads per sample were retained as highly reproducible genes and used for Spearman correlation analysis. Both library and sequence replicates showed very high correlations at the gene levels (0.930 and 0.926) and species levels (0.984 and 0.989).

icates a high reproducibility of gut microbial gene detection using the BGISEQ-500 platform.

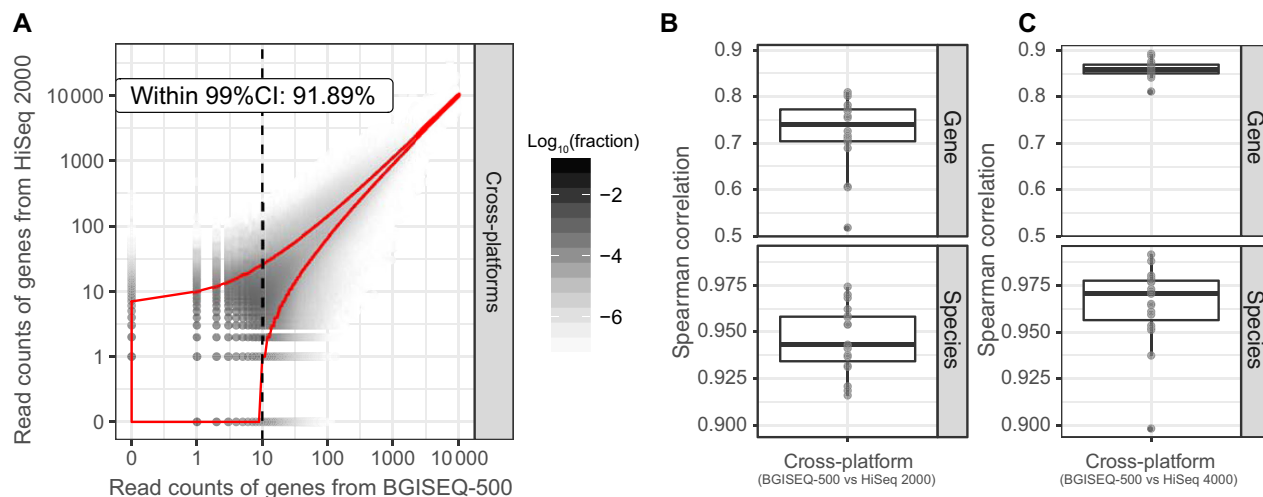
To assess the consistency of relative abundance identification of gut microbial genes, we performed Spearman correlation analysis based on highly reproducible (HR) genes and species profiles (Supplementary Methods). Both sequence replicates and library replicates showed high consistency at the gene level (Spearman's  $\rho > 0.91$ ) and species level (Spearman's  $\rho > 0.97$ ) (Fig. 2B). We further quantified the mean difference between replicates by using area left of the cumulative curve (ALC) (Supplementary Methods) [7]. The cumulative distributions of replicate differences were plotted (Additional file 5A). The mean gene relative abundance differences between sequence replicates ranged from 1.008- to 1.323-fold change (Additional file 5B). Similarly, the differences between library replicates ranged from 1.011- to 1.340-fold change (Additional file 5B). Together, these results suggest that very little variation was introduced by library preparation and sequencing processes.

Furthermore, 80 453 and 80 184 HR genes detected in at least 6 pairs of replicates in sequence and library replicates were

used for statistical tests, respectively (Supplementary Methods). Paired tests of gene abundances revealed no significant difference between BGISEQ-500 technical replicates (false discovery rate [FDR]  $< 0.05$ , Benjamini-Hochberg adjustment). Collectively, these findings demonstrate that the BGISEQ-500 platform, across the entire process of library preparation and sequencing of metagenomic DNA samples, provides highly reproducible and well-controllable results.

### Cross-platform consistency

Previously, shotgun metagenomic DNA sequence reads have mostly been generated using Illumina platforms, warranting evaluation of data consistency between the BGISEQ-500 and Illumina platforms; 91.89% of the genes in the BGISEQ-500 datasets showed expected read count fluctuations in HiSeq 2000 (99% CI) that were less than intra-platform replicates (Fig. 3A). Spearman correlation of HR gene and species profile of cross-platform samples reached 0.724 and 0.948 (Fig. 3B). Compared with intra-platform variations, cross-platform comparison showed a



**Figure 3:** Evaluation of inter-platform consistency. For 19 cross-platform replicates at 99% CI, 91.89% genes in the BGISEQ-500 datasets showed the expected mapped read count fluctuations using HiSeq 2000 (A). The Spearman correlation analyses revealed high agreement within 19 pair of platform replicates between BGISEQ-500 and HiSeq 2000 (B) (an average Spearman's  $\rho$  of 0.724 at gene level [top] and 0.948 at species level [bottom]) and between BGISEQ-500 and HiSeq 4000 (C) (an average Spearman's  $\rho$  of 0.859 at gene level [top] and 0.965 at species level [bottom]).

slightly greater difference. The differences in relative abundance between cross-platform groups ranged from 1.409- to 2.015-fold change (Additional file 5B).

Among 349 479 HR genes detected in at least 6 pairs of cross-platform replicates, the relative abundance of 11 350 (3.25%) genes differed significantly between these 2 platforms (FDR < 0.05, Benjamini-Hochberg adjustment). Among them, 2051 were detected by paired t tests, and 9299 were detected by paired sign tests (Supplementary Methods). Additionally, these 11 350 genes showed a bimodal distribution in GC content (Fig. 4A). AT-rich genes were enriched in the BGISEQ-500 dataset. Conversely, the relative abundances of GC-rich genes were higher in the HiSeq 2000 dataset (Fig. 4B). In accordance with the taxonomic annotation of IGC, 25.37% of the genes that differed in relative abundance were assigned to known species (Additional file 6).

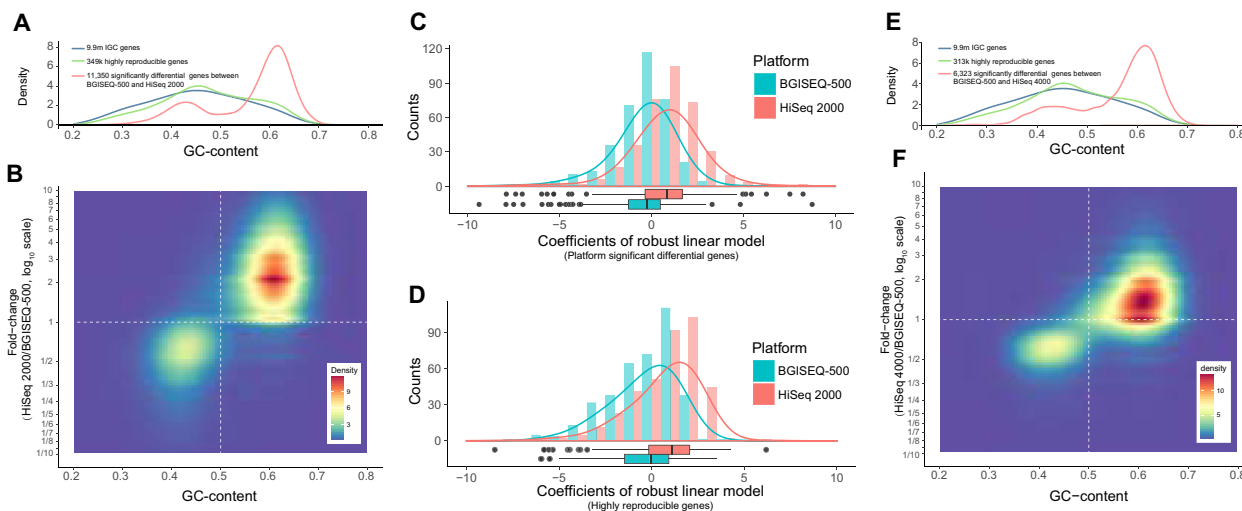
Assuming that the abundance of most genes from a species should be even and independent of their GC content, we conducted robust linear regression analysis of the correlation between the abundance of genes and their GC content for each species (Supplementary Methods). Based on the genes in the top 20 species exhibiting the most significant differences in abundance, the median of regression coefficient of the BGISEQ-500 dataset was close to 0, namely  $-0.095$  (Fig. 4C; Additional file 7), whereas, the regression coefficient of the HiSeq 2000 dataset was 0.925, indicating a slightly positive correlation between gene abundance and their GC content. The regression coefficient between all tested genes from the 20 species and their GC contents exhibited a similar tendency (Fig. 4D; Additional file 7). Additionally, generalized linear model (GLM) regression analysis was conducted to investigate the associations between approximate relative species abundance and GC content across the 2 platforms. MetaPhlan2 [6] was utilized to generate estimates of relative abundance for each species in each sample. The GC content of each species was retrieved from NCBI. Samples were classified as either high/low abundance (above/below median = 0.2844), either high/low GC content (above/below median = 43.8%) with respect to sequencing platform (BGISEQ-500 or Illumina) (Fig. 5). A log-linear model was used to model the total number of species in each of the 8 categories (abundance high/low, GC content high/low, BGI/Illumina), and a likelihood

ratio test then suggested that the association between relative species abundances and their GC content did not vary across the BGISEQ-500 and HiSeq 2000 sequencing platforms ( $P = 0.323$ , chi-square test) (Supplementary Methods).

To further document the quantitative consistency and performance regarding GC content observed for the BGISEQ-500 and HiSeq 2000 platforms, the same 20 DNA samples were processed to construct libraries and sequenced on an Illumina HiSeq 4000 platform using the HiSeq 3000/4000SBS Kit (300 cycles) for 100 bp of paired-end reads.

The raw sequencing reads were filtered as described above. After QC and removing host reads, an average of 26.35 million clean reads was generated for each sample. The HiSeq 4000 dataset showed comparable high-quality and IGC mapping ratios close to that of the BGISEQ-500 and HiSeq2000 datasets (Additional file 3). Because of the low number of sequencing reads from the HiSeq 4000 platform, all IGC uniquely mapped reads for each subject (ranging from 13.18 to 21.54 million) were used for validation analyses without downsizing. The forward reads of 19 subjects (subject S01 removed) from HiSeq 4000 were used for further analyses to be consistent with the cross-platform comparison described above.

Interestingly, the Spearman correlation coefficients of HR genes and species profiles between BGISEQ-500 and HiSeq 4000 samples were 0.859 and 0.965, respectively (Fig. 3C). These correlation coefficients were higher than those observed comparing the BGISEQ-500 and HiSeq 2000 datasets (Fig. 3B) and comparing the HiSeq 2000 and HiSeq 4000 datasets (Additional file 8). Statistical analysis revealed smaller quantitative differences between the BGISEQ-500 and HiSeq 4000 platforms than that between the BGISEQ-500 and HiSeq 2000 platforms, with only 6323 (2.02% of 313 020) HR genes showing significantly different relative abundances (FDR < 0.05, Benjamini-Hochberg adjustment). The same bimodal distribution patterns in GC content were observed among these 6323 HR genes, with an enrichment of GC-rich genes in the HiSeq 4000 dataset, as well as a slight enrichment of AT-rich genes in BGISEQ-500 dataset. Additionally, the abundance fold changes of these genes were smaller than those observed comparing the BGISEQ-500 and HiSeq 2000 datasets (Fig. 4E and F).



**Figure 4:** A, GC content distributions of genes that differed significantly in abundance between platforms. Density curves showing a comparison of GC content distributions of the total 9.9 million IGC genes (blue), all 349 479 highly reproducible (HR) genes (green), and all 11 350 genes that differed significantly in abundance between the 2 platforms (red line). B, Two-dimensional plot showing the GC content distribution of genes that differed significantly in abundance between the BGISEQ-500 and HiSeq 2000 platforms. The x-axis indicates the GC content of genes, the y-axis indicates fold-changes of gene relative abundance (RA), which is calculated by  $\log_{10}$  transformed mean RA in the HiSeq 2000 datasets/mean RA in the BGISEQ-500 datasets. C, D, Density histograms showing the coefficients of a robust linear model for relative abundance of genes from the top 20 species and their GC content for genes that differed significantly in abundance between the 2 platforms (C) and for all HR genes (D). D, E, Density curves (E) and 2-dimensional plot (F) showing the GC content distributions of HR genes that differed significantly in abundance between the BGISEQ-500 and HiSeq 4000 platforms.

In summary, despite that the HiSeq 2000 platform showed a slight enrichment of reads on a relatively small number of high-GC content genes, metagenomic datasets from BGISEQ-500 and HiSeq 2000 exhibited comparable cross-platform consistency regarding gene detection and quantification. The high cross-platform quantitative consistency was further documented using a HiSeq 4000 dataset, with a lower number of genes exhibiting differences in abundance between the platforms, and exhibiting the same bimodal distribution pattern of GC content of these genes.

## Discussion

The BGISEQ-500 platform has lately proven its robust performance in connection with several sequencing applications including whole-genome sequencing, RNA-seq, and small RNA-seq [2, 8–10]. Unlike these applications, which focus on a single organism, metagenomics deals with a complex mixture of DNA from multiple organisms. One of the key challenges for metagenomics studies is the accurate identification and quantification of taxa and functions in metagenomic samples, which could be affected by both the complexity of environmental samples and the bioinformatic tools.

Though rapid updates of Illumina sequencing techniques have provided up to 300 bp long and highly accurate reads, the read lengths available for the BGISEQ-500 were limited to single-end 50 bp, single-end 100 bp, and paired-end  $2 \times 50$  bp when this study was launched. Hence, we chose the longer mode of single-end 100 bp sequencing for the BGISEQ-500 platform and evaluated its performance.

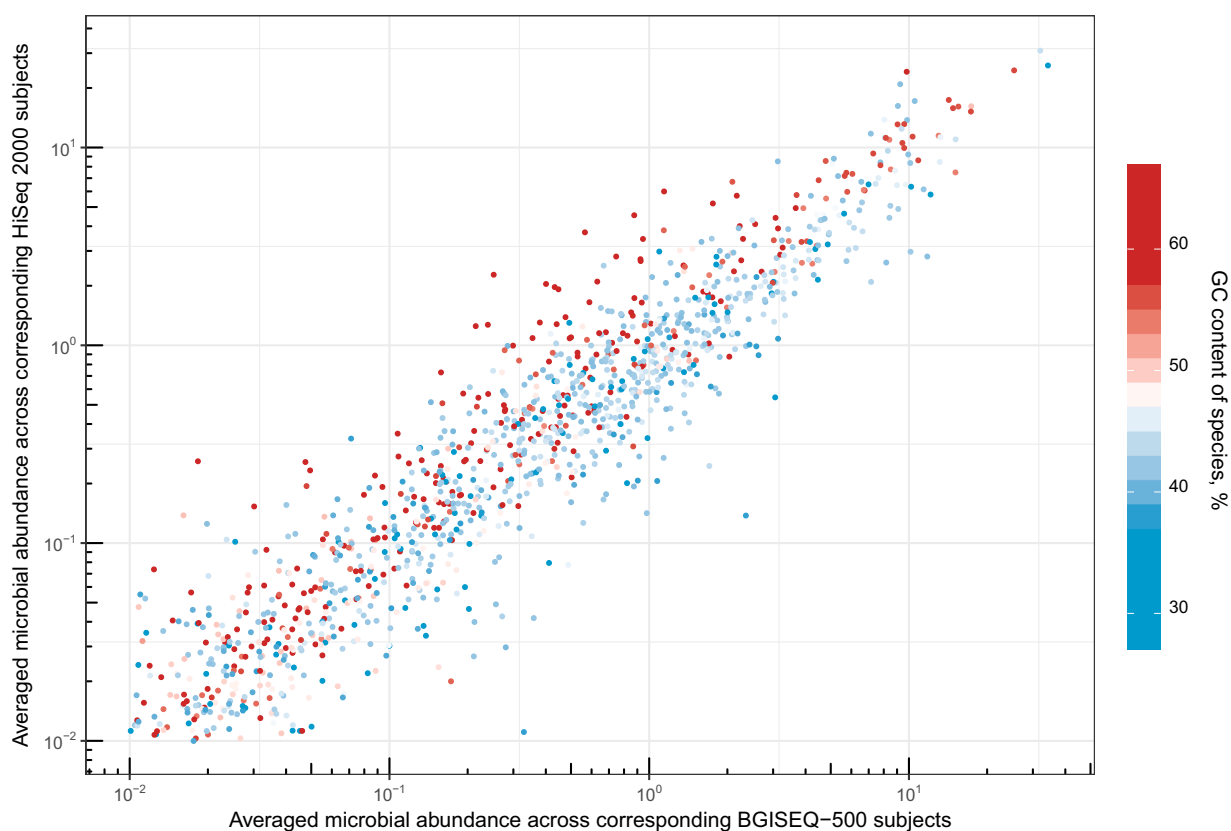
For the most well-studied microbiomes, such as the human gut microbiome, read lengths of 100 bp using Illumina HiSeq technology have been extensively applied, and the performance has been well demonstrated by numerous high-powered studies, including both reference gene catalog-based metagenome-wide association studies and Metage-

nomic Phylogenetic Analysis-based large-scale quantitative studies [11, 12]. On the other hand, for complex samples from environments such as soil, in which colonizing bacteria remain largely unknown, no reference gene catalogs have been established due to the inherent difficulty of assembly. In these cases, microbial genes were thereby identified using read-based homology searches, and thus, longer read lengths would enable more accurate gene identification.

In this work, we have evaluated the performance and validated the performance and robustness of the BGISEQ-500 platform for reference gene catalog-based human gut metagenome studies. We have developed an overall accuracy control-based QC method, which can detect random quality drops within reads and provide high-quality reads with minimal compromise of length. As the most widely used and acknowledged platform in the metagenomics field, datasets generated from Illumina platforms (HiSeq 2000/4000) were used throughout the cross-platform comparison.

By comparing metagenomic sequencing datasets from the BGISEQ-500 platform, we demonstrated excellent stability and reproducibility in intra-technical replications, providing evidence for the robustness and applicability of this new sequencing platform for metagenomics studies. We further demonstrated high consistency between the BGISEQ-500 and HiSeq 2000 platforms, with only a very small fraction of high-GC content genes showing a slight enrichment using the HiSeq 2000 platform. We furthermore compared the datasets from the HiSeq 4000 platform and BGISEQ-500 platform, and corroborated the high cross-platform consistency and platform-dependent GC distribution patterns.

As reported previously [13, 14], DNA extraction and library preparation methodology may affect both qualitative analysis and quantitative results in human microbiome research. In addition to that, even though we observed only minor differences in relative gene abundances comparing the BGISEQ-500 and the Illumina platform using single-end 100-bp read length, our



**Figure 5:** Comparison of relative species abundance between BGISEQ-500 and HiSeq 2000. Averaged microbial abundance calculated with Metaphlan2 across BGI replicates plotted against microbial abundance for the corresponding Illumina replicates for all samples. Species are colored by GC content.

results clearly point to the importance of using the same platform and technology for metagenomics studies in order to avoid the possible introduction of platform-dependent differences. In cases where comparison of data generated on different platforms is desirable, the possible platform-dependent confounding effects should be evaluated by well-designed analyses detecting possible confounding factors and biases before conclusions are drawn.

Finally, the results described in this paper emphasize the need for future use of benchmarking controls, including sequencing of defined microbial communities to elucidate the nature of possible biases associated with different preparation methodologies and sequencing platforms.

### Availability of supporting data

Metagenomic sequencing data for all samples have been deposited in the European Bioinformatics Institute (EBI) database under accession code PRJEB35961. Supporting data are also available from the *GigaScience* repository, *GigaDB* [15].

### Additional files

Additional file 1: Base quality assessment. Quality score heatmap showing the distribution of base Phred scores of all raw SE100 reads from BGISEQ-500 (A) and all forward reads from the HiSeq 2000 platform.

Additional file 2: Quality control with overall accuracy (OA) control strategy. A, The distribution of per base Phred score (top, orange), per base accuracy (middle, olive), and the overall accuracy curve of a randomly selected read from the BGISEQ-

500 platform (bottom, green) (see details on the OA-based QC pipeline in the Supplementary Methods). B, Identify QC parameters based on overall accuracy and high-quality reads ratio. Box plots showing the average OA of high-quality reads (green) and high-quality reads ratio (grey) based on different OA fragment thresholds. To balance the data accuracy and retention rate of high-quality reads after filtering, we chose  $OA_{\text{fragment}}$  as 0.8. Based on this cutoff, 96.06% of the raw reads were retained as high-quality reads with an average OA value greater than 90%.

Additional file 3: Summary of data production and quality control.

Additional file 4: Assessment of reference coverage for metagenomic sequencing data. After QC and removing potential human-related reads, an average of 95.98% and 97.91% of raw reads were obtained from BGISEQ-500 (blue) and HiSeq 2000 (red), respectively, and were defined as clean reads (left panel). For the BGISEQ-500 platform, an average of 77.77% of total clean reads could be mapped to IGC; the averaged mapping rate on HiSeq 2000 was 75.45% (middle panel). Additionally, the HiSeq 2000 dataset from subject S01 showed a low IGC mapping rate of 54.58% and was subsequently marked as an outlier and removed before cross-platform comparison. For both platforms, more than 62% of total clean reads were uniquely mapped (right panel) and used for further quantification analysis.

Additional file 5: ALC value for sequencing, library, and cross-platform replicates. The ALC value is the area left of the cumulative distribution curve: Thus, a low ALC value denotes high reproducibility. The ALC value is determined by the relative abundance differences of highly reproducible genes between replicates. The x-axis represents  $\log_2$ -transformed difference-folds,

and the y-axis represents the cumulative proportion of gene relative abundance difference for intra- and cross-platform (A). The box plot shows the estimated fold-change differences in relative gene abundance between replicates calculated by ALC values (B).

Additional file 6: Summary of species annotation for the genes that differed significantly in abundance between the BGISEQ500 and the HiSeq 2000 platforms.

Additional file 7: Robust linear regression analysis between gene relative abundance and GC content.

Additional file 8: Spearman's correlation between datasets from HiSeq 2000 and HiSeq 4000.

## Abbreviations

ALC: area left of the cumulative curve; cPAS: combinatorial Probe-Anchor Synthesis; DNB: DNA Nanoball; FDR: false discovery rate; GLM: generalized linear model; HR: highly-reproducible; IGC: integrated gene catalog; OA: overall accuracy; QC: quality control; RA: relative abundance.

## Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Karolinska University Hospital (2016/2502-31/2) and the radiation protection committee of the Karolinska Hospital (K2016-4511).

## Competing interests

The authors declare that Chao Fang, Huanzi Zhong, Yuxiang Lin, Bin Chen, Mo Han, Huahui Ren, Haorong Lu, Min Xia, Wangsheng Li, Xun Xu, Wenwei Zhang, Radoje Drmanac, Jian Wang, Huanming Yang, Karsten Kristiansen, and Junhua Li are employees of BGI.

## Funding

This study was supported by the National Key Research and Development Program of China (No. 2017YFC0909700), the National Natural Science Foundation of China (No. 31601073), the Shenzhen Municipal Government of China (No. JSGG20160229172752028), the Shenzhen Municipal Government of China Peacock Plan (No. KQTD20150330171505310), the American Diabetes Association Pathway Award No. 1-17-INI-13 (A.D.K.), Smith Family Foundation Award (A.D.K.), NIH/NHGRI T32 HG002295, PI: Peter J Park (J.M.L.), and NIH/NIGMS T32 GM074897, PI: Xihong Lin and Curtis Huttenhower (S.S.).

## Authors' contributions

J.L. and K.K. conceived and directed the project. J.L. routinely managed the project at BGI-Shenzhen. L.H. was responsible for collection of fecal samples. M.H., W.Z., and R.D. contributed to metagenomic library construction for the BGISEQ-500 platform. H.Z., B.C., H.L., M.X., and W.L. designed the technical replicates and sequencing experiments. C.F. developed the quality control method for BGISEQ-500 sequencing data. J.L., H.Z., and C.F. designed the analyses. C.F., H.Z., H.R., and Y.L. performed the bioinformatic analyses. H.R. contributed to the statistics methods. J.M.L. and S.S. conducted the MetaPhlan2 results and GLM analysis. H.Z., C.F., J.L., K.K., J.M.L., and A.D.K. interpreted the data. L.H., X.X., J.W., and H.Y. participated in text revision and discussions. C.F., H.Z., and J.M.L. wrote the first version of the manuscript. J.L., K.K., and A.D.K. revised the manuscript.

## Acknowledgements

We thank all the volunteers who participated in this study. We thank Dr Yin Shan for providing useful statistical suggestions. We thank Dr Hui Jiang, Yanyan Zhang, and Xia Zhao for their useful suggestions and discussion on metagenomic library construction. We thank Dr Huijie Jia and Dr Lise Madsen for helpful discussion of results. We also thank Dr Zhe Zhang for helpful discussion and suggestions on the revised manuscript. We gratefully acknowledge colleagues at BGI for DNA extraction, library preparation and sequencing experiments, and helpful discussions.

## References

1. Qin J, Li Y, Cai Z et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60.
2. Huang J, Liang X, Xuan Y et al. A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* 2017;**6**(5):1–9.
3. Taylor P, Liu L, Hu N et al. A brief utilization report on the Illumina HiSeq 2000 sequencer. *Mycology* 2011;**2**:169–91.
4. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998;**8**(3):186–94.
5. Li J, Jia H, Cai X et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;**32**(8):834–41.
6. Truong DT, Franzosa EA, Tickle TL et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**(10):902–3.
7. Mestdagh P, Hartmann N, Baeriswyl L et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods* 2014;**11**(8):809–15.
8. Mak SST, Gopalakrishnan S, Carøe C et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* 2017;**6**(8):1–13.
9. Fehlmann T, Reinheimer S, Geng C et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenet* 2016;**8**:123.
10. Zhang B, Zhang W, Nie RE et al. Comparative transcriptome analysis of chemosensory genes in two sister leaf beetles provides insights into chemosensory speciation. *Insect Biochem Mol Biol* 2016;**79**:108–18.
11. Liu R, Hong J, Xu X et al. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med* 2017;**23**:859–68.
12. Zhernakova A, Kurilshikov A, Bonder MJ et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 2016;**352**:565–9.
13. Jones MB, Highlander SK, Anderson EL et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A* 2015;**112**:14024–9.
14. Hart ML, Meyer A, Johnson PJ et al. Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. *PLoS One* 2015;**10**.
15. Fang C, Zhong H, Lin Y et al. Supporting data for “Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing.” *GigaScience Database* 2017. <http://dx.doi.org/10.5524/100395>.