Genome
An International Journal
for Genetics and Genomics

OPEN ACCESS | Article

Analysis and evaluation of different sequencing depths from 5 to 20 million reads in shotgun metagenomic sequencing, with optimal minimum depth being recommended

Jin Liu[a], Xiaokai Wang[b], Hailiang Xie[c], Qinghua Zhong[c], and Yan Xia[c]

[a]Department of Life Sciences, Yuncheng University, Yuncheng, Shanxi Province 044000, China; [b]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong 999077, China; [c]01Life Institute, Shenzhen 518000, China

Corresponding author: **Yan Xia** (email: xiayan@01lifetech.com)

Abstract

Our study was to analyze and evaluate the impact of different shotgun metagenomic sequencing depths from 5 to 20 million in metagenome-wide association studies (MWASs), and to determine the optimal minimum sequencing depth. We included a set of 200 previously published gut microbial shotgun metagenomic sequencing data on obesity (100 obese vs. 100 non-obese). The reads with original sequencing depths >20 million were downsized into seven experimental groups with depths from 5 to 20 million (interval 2.5 million). Using both integrated gene cluster (IGC) and metagenomic phylogenetic analysis 2 (MetaPhlAn2), we obtained and analyzed the read matching rates, gene count, species richness and abundance, diversity, and clinical biomarkers of the experimental groups with the original depth as the control group. An additional set of 100 published data from a colorectal cancer (CRC) study was included for validation (50 CRC vs. 50 CRC-free). Our results showed that more genes and species were identified following the increase in sequencing depths. When it reached 15 million or higher, the species richness became more stable with changing rate of 5% or lower, and the species composition more stable with ICC intraclass correlation coefficient (ICC) higher than 0.75. In terms of species abundance, 81% and 97% of species showed significant differences in IGC and MetaPhlAn2 among all groups with $p < 0.05$. Diversity showed significant differences across all groups, with decreasing differences of diversity between the experimental and the control groups following the increase in sequencing depth. The area under a receiver operating characteristic curve, AUC, of the obesity classifier for running the obesity testing samples showed an increasing trend following the increase in sequencing depth ($\tau = 0.29$). The validation results were consistent with the above results. Our study found that the higher the sequencing depth is, the more the microbial information in structure and composition it provides. We also found that when sequencing depth was 15 million or higher, we obtained more stable species compositions and disease classifiers with good performance. Therefore, we recommend 15 million as the optimal minimum sequencing depth for an MWAS.

Key words: sequencing depth, metagenome-wide association study, minimum sequencing depth, disease classifier

Résumé

La présente étude visait à analyser et évaluer l'impact de la profondeur de séquençage (de 5 à 20 millions de lectures) dans le cadre d'analyses d'association métagénomiques (MWAS) et de déterminer une profondeur minimale optimale. Les auteurs ont inclus 200 jeux de données métagénomiques publiées du microbiome intestinal et de leur lien avec l'obésité (100 obèses vs. 100 non-obèses). Les jeux de données qui comprenaient à l'origine une profondeur de >20 million de lectures ont été sous-échantillonnés pour constituer sept groupes expérimentaux dont le nombre de lectures retenues s'échelonnaient de 5 à 20 million de lectures (par intervalles de 2,5 million). À l'aide du clusters de gènes intégrés (IGC) et du logiciel MetaPhlAn2, les auteurs ont obtenu et analysé le taux d'alignement, le décompte génique, la richesse en espèces et leur abondance, la diversité et les biomarqueurs cliniques parmi les groupes expérimentaux en employant les données complètes originales comme groupe témoin. Un autre jeu de données publiées dans le cadre d'une étude sur le cancer colorectal (CRC) a été inclus pour fins de validation (50 cas de CRC vs. 50 sans CRC). Les résultats ont montré que plus de gènes et d'espèces ont pu être identifiés en augmentant la profondeur de lecture. À compter de 15 million de lectures, la richesse en espèces s'est stabilisée, avec des augmentations de 5 % ou moins, et la composition des espèces également, la corrélation entre classes (ICC) étant de plus de 0,75. En ce qui a trait à leur abondance, 81 % et 97 % des espèces ont présenté des différences significatives (p < 0,05) entre les groupes. Des différences significatives dans la diversité ont également été observées entre les groupes expérimentaux et le groupe témoin, ces différences s'amenuisant au gré de l'augmentation du nombre de lectures employées. L'aire sous la courbe (AUC) pour le classificateur d'obésité, employé pour analyser les échantillons de cette étude, a eu tendance à augmenter en parallèle avec l'augmentation de la profondeur de séquençage ($\tau = 0.29$). Les résultats de l'étude de validation étaient conformes à ceux décrits plus hauts. Les résultats de cette étude ont montré qu'une augmentation de la profondeur de séquençage entraîne une information plus riche sur la structure et la composition du microbiote. Les auteurs ont également trouvé qu'au-delà de 15 M de lectures, des résultats plus stables en matière de composition en espèces et une bonne performance dans la classification des échantillons étaient obtenus. Ainsi, les auteurs recommandent 15 million de lectures comme étant une profondeur minimale de séquençage optimale dans le contexte d'analyses MWAS. [Traduit par la Rédaction]

Mots-clés : profondeur de séquençage, analyse d'association métagénomique, profondeur minimale de séquençage, classificateur de maladie

Introduction

The past decade has witnessed unprecedented progress in microbiome research due to major advances in bioinformatics, the next-generation sequencing (NGS) approaches specifically. Using the NGS approaches, including the 16S RNA gene amplicon sequencing and the whole-genome shotgun sequencing, numerous metagenome-wide association studies (MWASs) have been conducted to help us understand better the role of the human microbiome in the human body physiologically and pathologically. It investigates the association between the human microbiome and diseases and provides potential solutions to conquer diseases from a very different perspective than before. For now, the associations between the microbiome and extensive diseases, including obesity, diabetes, autism, atherosclerosis, atopic dermatitis, irritable bowel syndrome, etc., are intensively investigated (Janda and Abbott 2007; Qin et al. 2012; Kelly et al. 2016; Wang and Jia 2016; Ghoshal and Gwee 2017; Gu et al. 2017; Liu et al. 2017; Lee et al. 2018; Sharon et al. 2019).

Most of the currently available MWASs applied the 16S rRNA sequencing, especially in large-scale studies for its lower cost. Thus, fewer data on shotgun metagenomic sequencing are available in the literature, despite the fact that shotgun metagenomic sequencing provides genetic information of higher taxonomic resolution and more intensive functional profiling (Rausch et al. 2019). However, the application of shotgun metagenomic sequencing is increasing in revealing more potential associations between the human microbiome and diseases.

Shotgun metagenomic sequencing produces raw sequencing reads with inevitably varied depths because of technological limitations. Researchers would then downsize the varied sequences to the same level to avoid statistical error either before or after being mapped to the reference database (Nielsen et al. 2014; Andersen et al. 2015; Fang et al. 2017; Ruppé et al. 2017). The targeted downsized depth was normally equal to or a bit higher than the minimum depth preset in an MWAS. However, there is not much guidance or discussion on how to preset a minimum sequencing depth besides the cost, nor on the impact of different depths on MWAS at sequencing depths between 5 and 20 million.

There are a few studies currently available in the literature illustrating the impact of various depths using shotgun metagenomic sequencing either with sequencing depth too shallow or sample size too small. Jovel et al. (2016) analyzed shotgun metagenomic sequencing depths below 100 000 reads and found that it offered increased resolution with taxonomy of more specific and functional classification of sequences as well as the discovery of new bacterial genes and genomes compared with 16S rRNA sequencing. Hillmann et al. (2018) evaluated the information content of shallow shotgun sequencing for depths lower than 1 million reads concluding that shotgun sequencing of 0.5 million reads is an alternative to 16S sequencing in large-scale studies. Zaheer et al. (2018) analyzed three various depths, 117, 59, and 26 million reads, for eight bovine fecal samples concluding that 59 million read count is suitable to describe the microbiome and resistome of bovine fecal samples. Gweon et al. (2019) investigated the impact of sequencing depth on antimicrobial resistance profiles and found that deep sequencing of 80 million reads per sample or higher was required depending on different environmental samples. However, the impact of different depths above 10 million using shotgun metagenomic sequencing with bigger sample sizes has not been properly studied in previous studies, nor has been the determination of the minimum depth.

Our study was designed to evaluate the impact of different sequencing depths from 5 to 20 million using shotgun metagenomic sequencing on microbial information gained, including species richness and abundance, alpha and beta diversity, and obesity classifier based on biomarkers selected. We also aimed to determine the optimal minimum sequence depth between 5 and 20 million to preset in an MWAS on most cohorts.

Methods and materials

Materials

Metagenomic sequences of 200 fecal samples (100 obese vs. 100 non-obese) were selected from a case-control cohort study of 118 obese patients (95 obese patients and 23 obese patients with weight-loss treatment) and 105 normal-weight controls, picking those with more read counts and at baseline only. The dataset was downloaded from the European Bioinformatics Institute (EBI) database with BioProject ID (PRJID) of PRJEB12123.

All the above data were paired-end reads sequenced by Illumina HiSeq 4000 with sequences from 21 to 43 million reads (Table S1).

Another set of 100 published shotgun metagenomic sequencing data from a colorectal cancer (CRC)-associated study (EBI PRJID PRJEB10878) with 74 patients with CRC and 54 controls was adopted, picking those with more read counts (50 CRC vs. 50 control) (Yu et al. 2017).

Data of the CRC study were paired-end reads sequenced by Illumina HiSeq 2000, with sequences from 22 to 39 million reads (Table S1).

The supplementary script contains the R script code with comments and instructions.

Quality control

We used Trimmomatic (Bolger et al. 2014) with parameters LEADING:5, TRAILING:5, SLIDINGWINDOW:4:15, and MINLEN:70 for quality control of the sequencing reads and SOAP2 (Li et al. 2009) to remove host sequences (Liu et al. 2017).

Gene and taxonomic profiling

The qualified reads were aligned to the integrated gene cluster (IGC) (Li et al. 2014), a database for taxonomic profiling by SOAP2 (soap2.2.1: https://github.com/gigascience/bgi-soap2), with a threshold of more than 90% identity over 95% of the length. Sequence-based gene abundance profiling was calculated using methods applied by Qin et al. (2012). The taxonomic profiling was calculated according to the relative abundance of their respective genes.

The reads were also mapped using metagenomic phylogenetic analysis 2 (MetaPhlAn2) (v2.7.7), a tool to calculate taxonomic profiling (Truong et al. 2015) by bowtie2 with its defaulted parameters. It uses a library of clade-specific markers to provide pan-microbial (bacterial, archaeal, viral, and eukaryotic) quantification at the species level.

All the mapping results of IGC and MetaPhlAn2 were retained for the subsequent downsizing.

Downsizing reads

The original sequencing reads were downsized to 5, 7.5, 10, 12.5, 15, 17.5, and 20 million, respectively, as seven experimental groups, using R function sample without replacement. We calculated the profile of the downsized read tag with the mentioned mapping results of IGC and MetaPhlAn2. GUN parallel was used to speed up the process (Tange 2011).

Diversity

Alpha diversity (within-sample diversity) is calculated on the gene and taxonomy abundance profile of each sample according to the Shannon index and species richness (Hurlbert 1971). Bray–Curtis dissimilarity (Bray and Curtis 1957) is used to measure the between-sample distance and principal coordinates analysis (PCoA) is used to display the shifts of MetaPhlAn2 genus.

Impact of sequencing depth

The sequencing depth in this paper was treated as a categorical variable. We used the repeated-measures Analysis of variance (ANOVA) with different sequencing depths as the independent variables and sample subjects as the random effect to test whether the sequence depths influence matching rates and taxonomic profiling etc., as the response variables.

The changing rate of the response variables of each experimental group and the control group was compared by Kruskal–Wallis test. The false discovery rate was controlled using q value method for multiple-hypothesis tests (Storey 2002).

Stability of species richness between each experimental group and the control group was assessed using intraclass correlation coefficient (ICC) based on an absolute-agreement, single measurement, two-way mixed-effects model using R package "irr" (v0.84.1). The ICC values <0.5, 0.5–0.75, 0.75–0.9, and >0.90 are indicative of stability of poor, moderate, good, and excellent levels, respectively (Koo and Li 2016).

Identified gene markers

We randomly separated the 100 obese vs. 100 non-obese data into two sets, one with 75 obese vs. 75 non-obese as the training set, and another with 25 obese vs. 25 non-obese as the testing set (Table S1).

In the training set, we downsized the original reads into the mentioned seven depths. Gene markers as mapped in IGC were identified using Wilcoxon rank-sum test for each sequencing depths. The false discovery rate was controlled using q value method for multiple-hypothesis tests (Storey 2002).

Genes that showed significant differences in relative abundance between the obese and non-obese (Wilcoxon rank-sum test, $q < 1e^{-3}$) with median abundance $\geq 1e^{-7}$ in any group were identified as obesity marker genes. All gene markers were clustered into metagenomic species (MGS) using the approach with default parameters ($\geq$3 genes) described by Nielsen et al. (2014).

MGS-based classifier

MGS were sorted into confirmed and tentative groups using R package "Boruta" (v7.0.0) with default parameters (Kursa and Rudnicki 2010). Confirmed and tentative MGS were used to build the final random forest models as obesity classifiers using R package "randomForest" (v4.6-14) with default parameters (Breiman 2001).

The area under a receiver operating characteristic curve (AUC) of true-positive rates vs. false-positive rates of all the random forest models was calculated using respectively the training set (out-of-bag data) and the testing set at each depth, to depict the performance of the random forest models. The value for AUC ranges from 0 to 1. We adopted the standards recommended in "Applied Logistic Regression" and followed in our study (Hosmer et al. 2013):

0.5 = no discrimination
0.5–0.7 = poor discrimination
0.7–0.8 = acceptable discrimination
0.8–0.9 = excellent discrimination
$>$0.9 = outstanding discrimination

All the above analysis were run 100 times for each sequencing depth to ensure the stability, including the testing set, obtaining 100 AUC scores on models established for each depth. Dunn's all-pairs rank comparison test using R package "PMCMR" (v1.9.0) (Pohlert 2014) was applied to compare the AUC scores of the testing set on the models between each depth, and the p value adjusted by Benjamini and Hochberg (1995) method was calculated.

The trends of AUC scores and the sequencing depth as ordered categorical variables were assessed using Kendall rank correlation coefficient (τ) and its p value was computed using R package "Kendall" (v2.21).

Results

To evaluate the impact of different sequencing depths from 5 to 20 million, sequences in the experimental groups were analyzed in terms of read matching rates, genes, species, diversity, and clinical biomarker discovery, with the original depth as the control group. Models based on the biomarker genes identified on different depths were established as disease classifiers for obesity to evaluate the impact of sequencing depths.

Both IGC and MetaPhlAN2 were used for mapping to minimize deviation caused by different analysis methods and reference databases, in calculating gene richness and species and genus abundance specifically.

Read matching rates

Read matching rates, the percentage of reads aligned to the references database, of all samples at different depths were calculated, obtaining mean values of 2.5% in MetaPhlAn2 and 80.5% in IGC (Table S2). Matching rates of all sequences in IGC were then adopted for the subsequent comparative analysis by repeated-measures ANOVA, showing no significant difference among all groups ($p = 0.25$) (Fig. 1A), indicating that the sequencing depths from 5 to 20 million reads do not affect the read matching rate. Thus, all the depths we used in our study from 5 to 20 million were acceptable when read matching rate was considered only.

Genes

Genes identified in the training samples by IGC at each depth were calculated, obtaining gene count from 4 514 745 to 5 346 170 at depths from 5 to 20 million, and 5 499 409 in the control group, indicating that more genes would be detected in deeper sequencing groups (Table 1). Genes identified in one sample by IGC at each depth were calculated, obtaining a median gene count of 400 000–700 000 at depths from 5 to 20 million reads, revealing a positive correlation between the gene count identified and sequencing depth (Fig. 1B, Table S3).

The changing rate of all-sample gene count between each experimental group and the control group was compared by Kruskal–Wallis test showing significant difference ($p < 2.2e^{-16}$), descending from 42.33% to 7.89% following the increase in sequencing depth (Fig. 1C).

Fig. 1. Data of all samples at different sequencing depths matching to IGC. (A) Matching rates of each group at different depths; (B) gene count identified by IGC of each group at different depths; (C) the change rate of gene count difference between each experimental group and the control group. For all box plots, the central lines, boxes, and whiskers represent the median, interquartile range (IQR), and 1.5 times the IQR, respectively. For panels A and B, the samples are in the order of value in the control group.

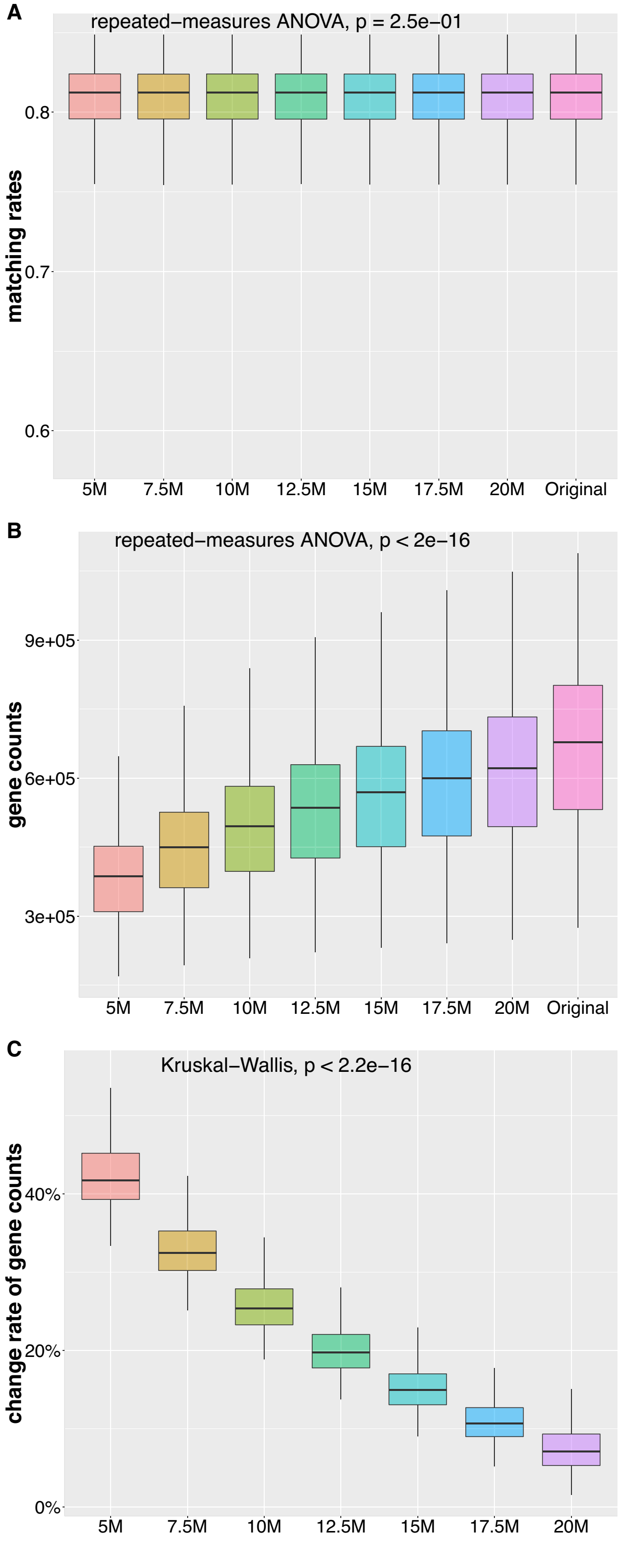


Table 1. Overview of clinical biomarkers.

Disease associated		5M	7.5M	10M	12.5M	15M	17.5M	20M	Original
Obese	Gene number (found in at least 1 sample)	4 514 745	4 780 130	4 956 647	5 087 213	5 191 526	5 273 953	5 346 170	5 499 409
	Gene number (median abundance in any groups $\geq$ 1e^{-7})	193 367	227 106	245 589	257 921	267 053	273 492	278 816	289 406
	Gene number (Wilcox q < 1e^{-3})	23 986	29 030	32 234	34 650	36 336	37 637	38 931	40 741
	MGS number (gene number $\geq$ 3)	128	167	188	193	219	217	221	222
	Gene number of the top MGS	2178	2359	2469	2560	2606	2654	2692	2792
	Percentage of shared genes of the top MGS	96.01%	88.64%	84.69%	81.68%	80.24%	78.79%	77.67%	74.89%
CRC	Gene number (found in at least 1 sample)	4 294 116	4 573 468	4 760 439	4 897 850	5 006 464	5 094 728	5 168 971	5 381 711
	Gene number (median abundance in any groups $\geq$ 1e^{-7})	172 517	204 795	223 613	235 915	245 041	252 053	256 706	269 656
	Gene number (Wilcox q < 1e^{-3})	14 772	18 869	21 649	23 871	25 398	27 190	28 011	31 022
	MGS number (gene number $\geq$ 3)	115	149	176	175	188	215	224	239
	Gene number of the top MGS	1832	2188	2345	2544	2644	2724	2758	2960
	Percentage of shared genes of the top MGS	79.20%	66.32%	61.88%	57.04%	54.88%	53.27%	52.61%	49.02%

Note: MGS, clustered into metagenomic species; CRC, colorectal cancer.

Species

Species richness of each group identified in IGC and MetaPhlAn2 was calculated, showing increasing richness from 488 to 510 in IGC and 370 to 492 in MetaPhlAn2, following the increase in depths (Table S4). Species richness among each groups including the control group was compared using Kruskal–Wallis test showing significant difference ($p < 2e^{-16}$), with higher species richness being detected following the increase in sequence depths (Figs. 2A and 2B). Difference in species richness between each experimental group and the control group was compared using Kruskal–Wallis test showing significant difference ($p < 2.2e^{-16}$), obtaining changing rate descending from 16.23% to 2.8% in IGC and 32.1% to 6.44% in MetaPhlAn2, following the increase in sequence depths (Fig. 2C). Lower changing rate of species richness represents higher stability. It decreased to 5% and lower in IGC only when the sequencing depth was 15 million or deeper, and to 10% and lower in MetaPhlAn2 only when the sequencing depth was 15 million or deeper (Fig. 2C).

Species compositions of the samples at different sequencing depths were obtained by calculating the ICC of different groups based on the species richness in IGC and metaPhlan2. When ICC is larger than 0.75, it represents a relatively stable species composition. The ICCs of the 5 and 7.5 million groups were below 0.5 with poor stability of the species composition. The ICCs of the 10 and 12.5 million groups were between 0.5 and 0.75, representing moderate stability of the species composition. The ICCs of the 15 and 17.5 million groups were 0.75–0.9 with good stability, and that of the 20 million group was larger than 0.9, doing excellent (Fig. 2D).

Species abundance of a certain individual species in all groups was obtained based on IGC and MetaPhlAn2 and analyzed using repeated-measures ANOVA analysis. Species abundance of 81% species from IGC and 97% species from MetaPhlAn2 presented significant difference at different depths with q value < 0.05. The abundance of three species, *Bacteroides stercoris*, *Klebsiella pneumoniae*, and *Veillonella atypica*, was significantly different at different depths in both reference databases (Table S4) (Fig. 3).

Diversity

For alpha diversity, Shannon index in all groups was computed with species abundance obtained in both IGC and MetaPhlAn2, showing significant difference ($p = 5.58e^{-04}$) using repeated-measures ANOVA, with changing rate less than 0.1% between the experimental groups and the control group in IGC, descending following the increase in sequencing depth; with species abundance obtained in MetaPhlAn2, the changing rate is less than 2.5% (Figs. 4A and 4B). The difference in species diversity between experimental groups and the control group becomes smaller following the increase in the sequencing depth with very small changing rate of less than 0.1% in IGC and less than 2.5% in MetaPhlAn2. Thus, the provided depths from 5 to 20 million were considered acceptable in term of species diversity in our study.

For beta diversity, Bray–Curtis dissimilarity in all groups was calculated for changing rate between samples with genus abundance obtained in both IGC and MetaPhlAn2, but adopting results of MetaPhlAn2 only, showing significant difference ($p < 2.2e^{-16}$) using repeated-measures ANOVA (Bray and Curtis 1957). The maximum compositional dissimilarity of one sample among the experimental groups based on Bray–Crurtis dissimilarity goes downward following the increase in sequencing depths, presenting significant difference (>0.05) in 38 samples out of 200 samples (Fig. 4C). Less than 5% samples at 15 million sequence depth and above presented larger than 0.05 beta diversity between the experimental groups and the control group (Fig. 4D). The smaller the proportion of samples having low similarity between the experimental groups and the control group is, the more comprehensive information a sequence data provides, thus 15 million was taken here as the recommended minimum depth in terms of beta diversity in our study.

Clinical biomarker discovery

The number of obesity marker genes being identified increased following the increase in sequencing depths, from 23 986 to 40 741 (Table 1). A total of 13 662 obesity marker genes intersected in all groups were identified, with 3650 intersecting in depths excluding the 5 million (Fig. 5A) (Table 1).

Obesity marker genes identified were clustered into obesity MGS. MGS < 3 genes were then excluded. As few as 128 qualified MGS in the 5 million depth group and as much as 222 qualified MGS in the 20 million depth group were obtained, with the largest MGS composing over 2000 genes (Table 1).

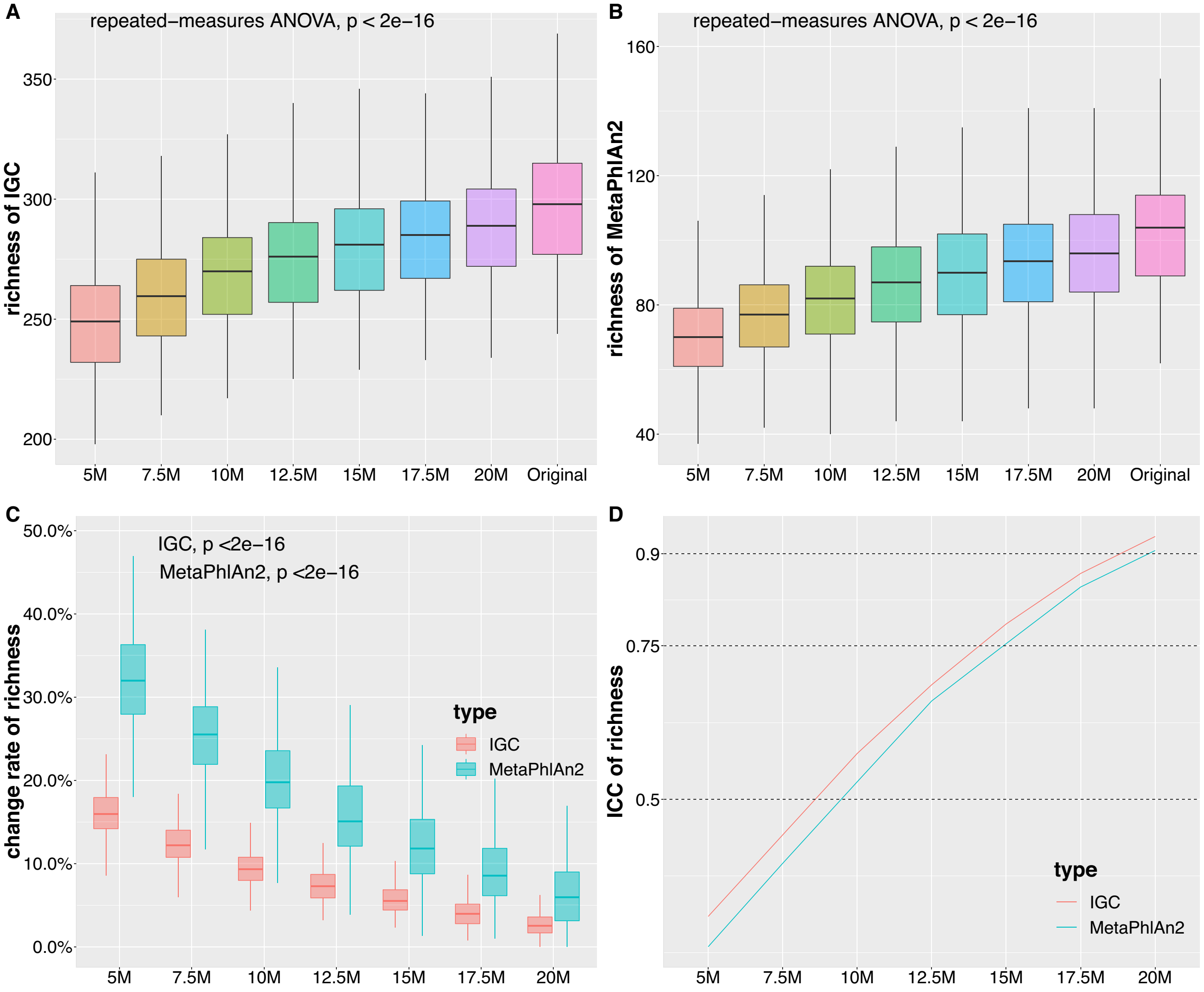
Based on the qualified obesity MGS, 100 obesity models were established for each sequencing depth using random forest method. The AUC scores of the training set are all approx. 0.90, representing excellent or outstanding performance of the models. We then tried to classify the samples of the testing set with the obesity models and obtained AUC scores between 0.82 and 0.85 (excellent) with an increasing trend following the increase in sequencing depths ($\tau = 0.29$) (Fig. 5B) (Table 2). We calculated the p value adjusted by BH method of the testing set AUC scores and came up with p values mostly less than 0.05 (Table S6).

Validation

We validated the results with a set of 100 metagenomic sequencing data selected from an independent CRC-associated study, splitting into a training set of 38 CRC vs. 38 control and a testing set of 12 CRC vs. 12 control data (Table S1).

Read matching rates of all samples at different depths were calculated, obtaining a mean value of 74.2% in IGC, showing no significant difference among all groups by repeated-measures ANOVA analysis ($p = 0.996$) (Fig. 6A, Table S2). The count of genes identified by IGC was from 4 294 116 to 5 168 971 at sequencing depths from 5 to 20 million, indicating that more genes are detected at deeper sequencing depths (Table 1). Species richness identified in IGC increased from 480 to 495 following the increase in depths. Species abundance of 72% species from IGC and 91% species from MetaPhlAn2 presented significant difference with q

Fig. 2. Comparison of the species richness obtained in IGC and MetaPhlAn2. (A, B) IGC and MetaPhlAn2 species richness of each group; (C) the change rate of IGC and MetaPhlAn2 species richness difference between each experimental group and the control group; (D) the ICC of IGC and MetaPhlAn2 species richness between each experimental group and the control group. For all box plots, the central line, boxes, and whiskers represent the median, interquartile range (IQR), and 1.5 times the IQR, respectively. For panels A and B, the samples are in the order of value in the control group. ICC, intraclass correlation coefficient.



value < 0.05 (Table S4). The abundance of three species *Bacteroides stercoris*, *Klebsiella pneumoniae*, and *Veillonella atypica* was significantly different at different sequencing depths in both reference databases (Fig. S1).

We ran 100 times for each sequencing depth on the CRC training set to establish models, verified the models with the CRC testing set, and calculated the AUC scores (Table 2). The AUC scores of the CRC training set were between 0.95 and 0.97 (outstanding). The AUC scores of the CRC testing set were between 0.66 and 0.73, with poor to acceptable discrimination and increasing trend following the increase in sequencing depth ($\tau = 0.42$) (Fig. 6B). We calculated the p value adjusted by BH method of the testing set AUC scores and came up with p values mostly less than 0.05 (Table S6).

Discussion

With the improvement of sequencing technology and experience we have obtained, more and more MWASs of large-scale population have been carried out using shotgun metagenomic sequencing, inevitably producing varied depths of sequences (Li et al. 2014; Dai et al. 2018; Thingholm et al. 2019; Visconti et al. 2019; Asnicar et al. 2021). For future studies with validity, high efficiency, and cost effectiveness, a relatively optimal and unified minimum sequencing depth for an MWAS is expected to be recommended to obtain sequencing data of the same level. Hence, an evaluation of different sequencing depths within a certain range is required.

To our best knowledge, our study is the first to have evaluated the impact of different sequencing depths from 5 to

Fig. 3. Comparison of the species abundance obtained from IGC and MetaPhlAn2. (A) *Bacteroides stercoris* abundance of IGC in each group; (B) *Bacteroides stercoris* abundance of MetaPhlAn2 in each group; (C) *Klebsiella pneumoniae* abundance of IGC in each group; (D) *Klebsiella pneumoniae* abundance of MetaPhlAn2 in each group; (E) *Veillonella atypica* abundance of IGC in each group; (F) *Veillonella atypica* abundance of MetaPhlAn2 in each group.

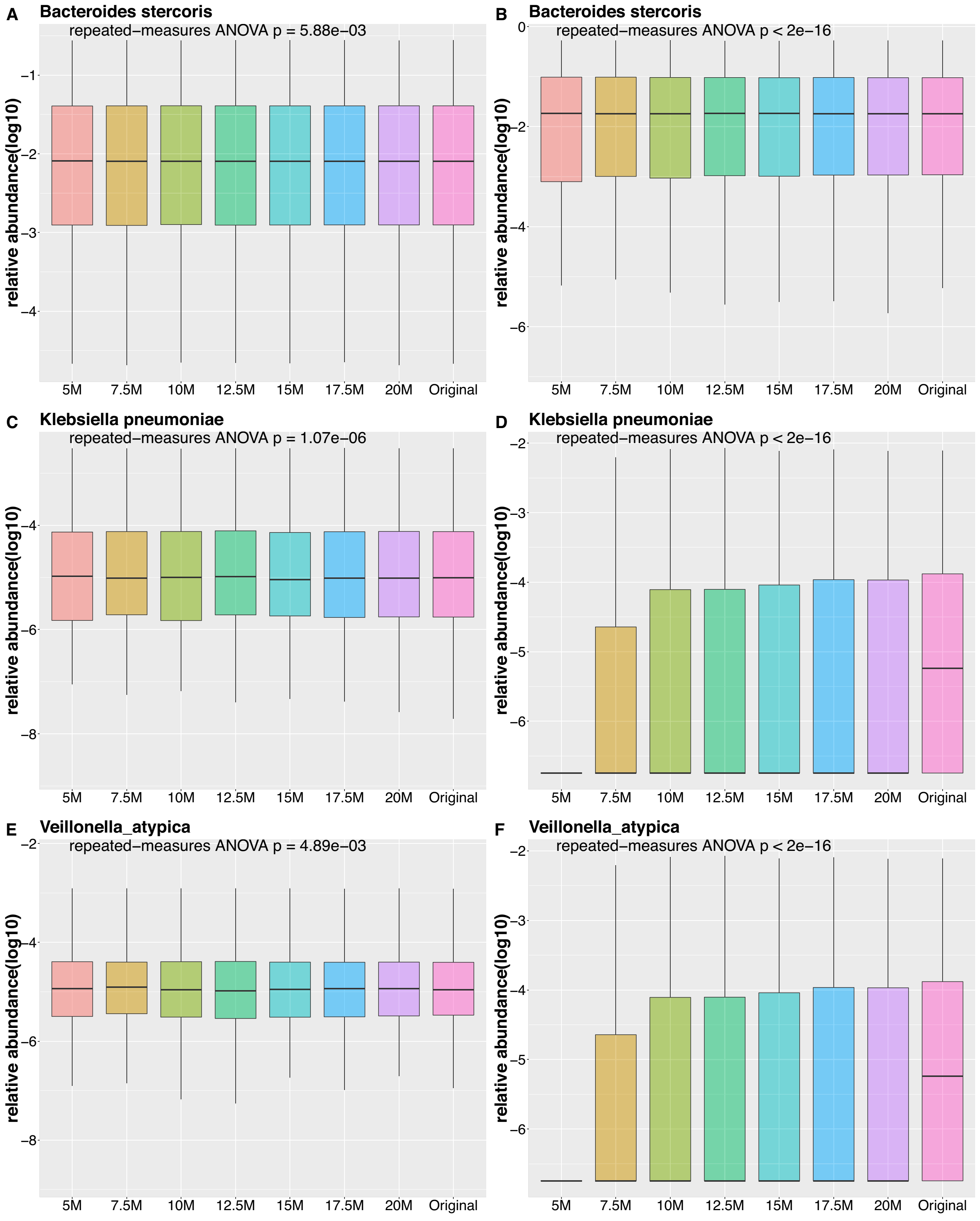
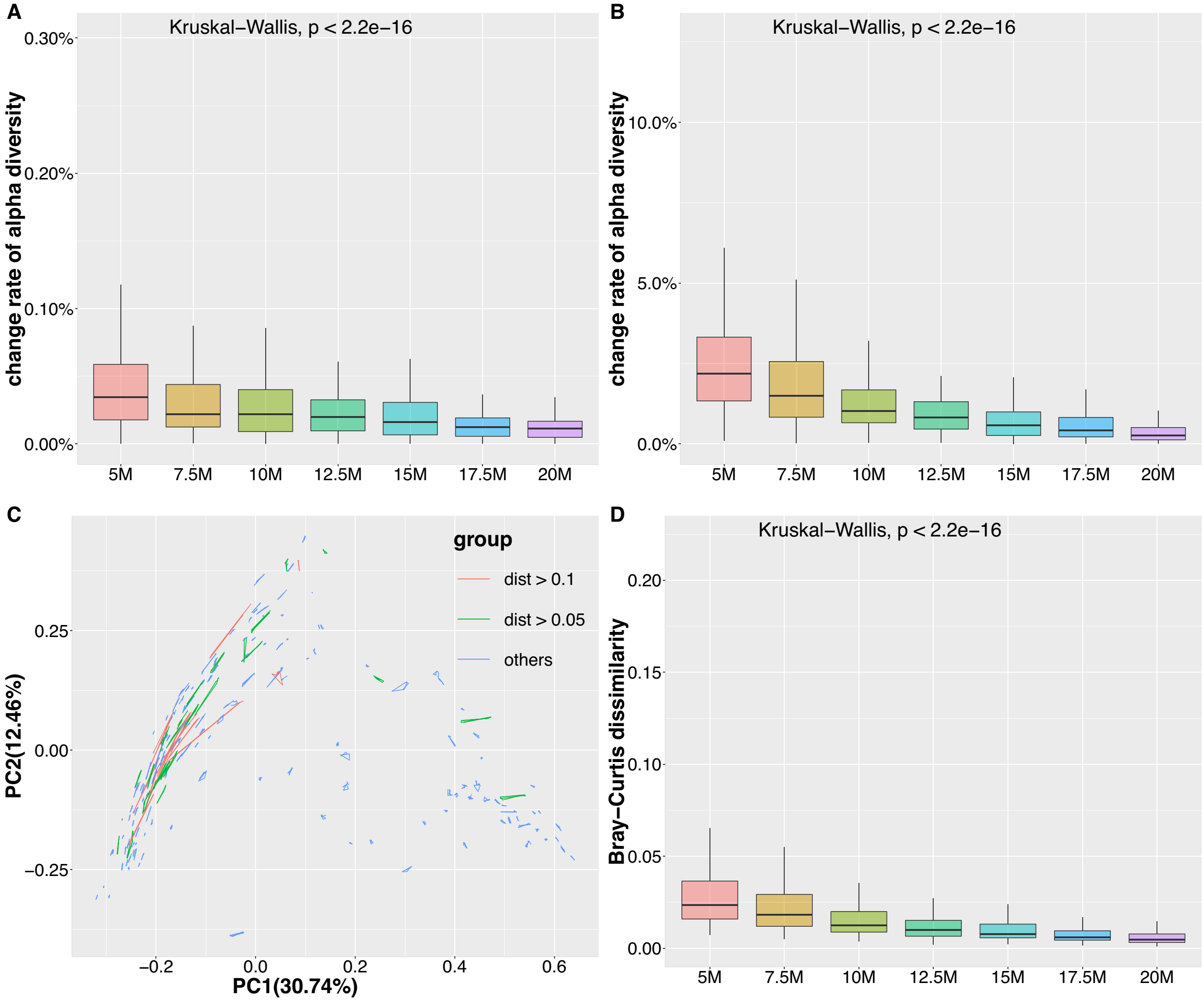


Fig. 4. Comparison of alpha diversity and beta diversity in each group. (A) The change rate of IGC alpha diversity difference between each experimental group and the control group; (B) the change rate of MetaPhlAn2 alpha diversity difference between each experimental group and the control group; (C) PCoA displaying the shifts in MetaPhlAn2 genus Bray–Curtis dissimilarity, with samples of one individual lined together and colored by the maximum distance; (D) MetaPhlAn2 genus Bray–Curtis dissimilarity of the same individual of each experimental group and the control group. For all box plots, the central line, boxes, and whiskers represent the median, interquartile range (IQR), and 1.5 times the IQR, respectively.

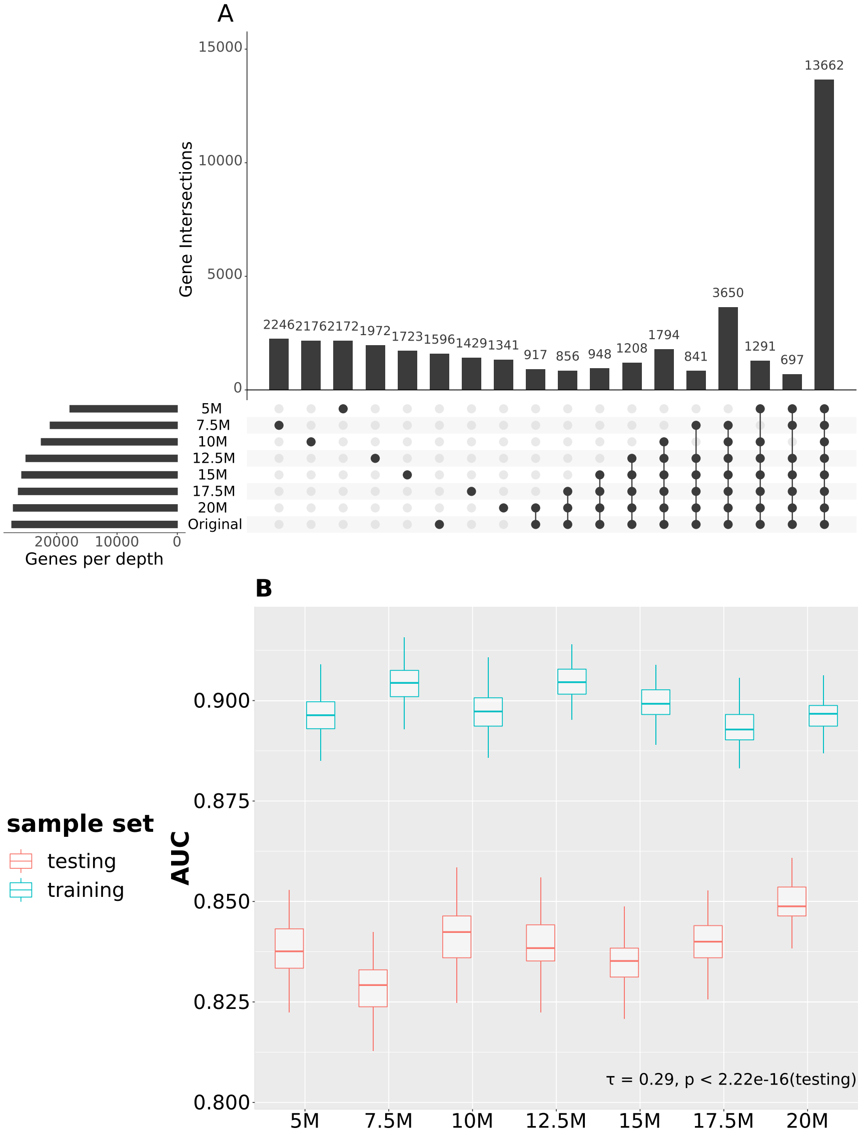


20 million reads of human gut microbiome in MWASs with reference databases, where an optimal sequencing depth between is recommended. To eliminate deviation caused by analysis methods and reference databases, we used IGC and MetaPhlAn2 at the same time while evaluating different aspects of the data, specifically read matching rate, gene richness, species richness and abundance, diversity, and clinical biomarkers.

As for read matching rate, our study showed clear results that it was ideal enough at 5 million reads depth, which was more than 75%. We observed that the MetaPhlAn2-based rate fluctuated more than the IGC-based rate. The reason might be that there are fewer reference genes in MetaPhlAn2 than in IGC.

We found a significant positive correlation between different sequence depths and the identification of genes, as well as the species richness. It adds value of another dimension to Li's finding of the rapid increasing of genes identified following the increase in sample size (Li et al. 2014). The count of genes identified in our study with 200 sample at sequence depths from 5 to 20 million echoes to that in Li's study at sequence depth of 11 million with 200 samples. Zaheer et al. (2018) indicated similar results that richness was higher at the genus and species level with the increase in sequence depth by analyzing three different sequence depths of 117, 59, and 26 million reads. However, Shao et al. (2019) mentioned in their study investigating the effects of delivery ways towards gut microbiome that they observed no correlation

Fig. 5. Biomarker genes identified and the AUC of each of seven disease classifiers established with training set samples. (A) Count of marker genes identified in each group (the bars); count of genes intersection between groups identified (the columns); groups selected for comparison (the black dots); count of genes intersection between groups smaller than 1% of the total marker genes were excluded; (B) AUCs of the models at each depth using the training set and the testing set calculated 100 times. AUC, the area under a receiver operating characteristic curve.



between the sequencing depth (5.47–66.5 million reads) and species richness.

We also found that the species composition becomes more stable following the increasing of sequencing depth. Species composition obtained from sequencing depths of 15 million and deeper was in good and excellent stability, while that from sequencing depths of lower than 10 million was poor. Thus, 15 million is the optimal sequencing depth we recommended when species composition is considered.

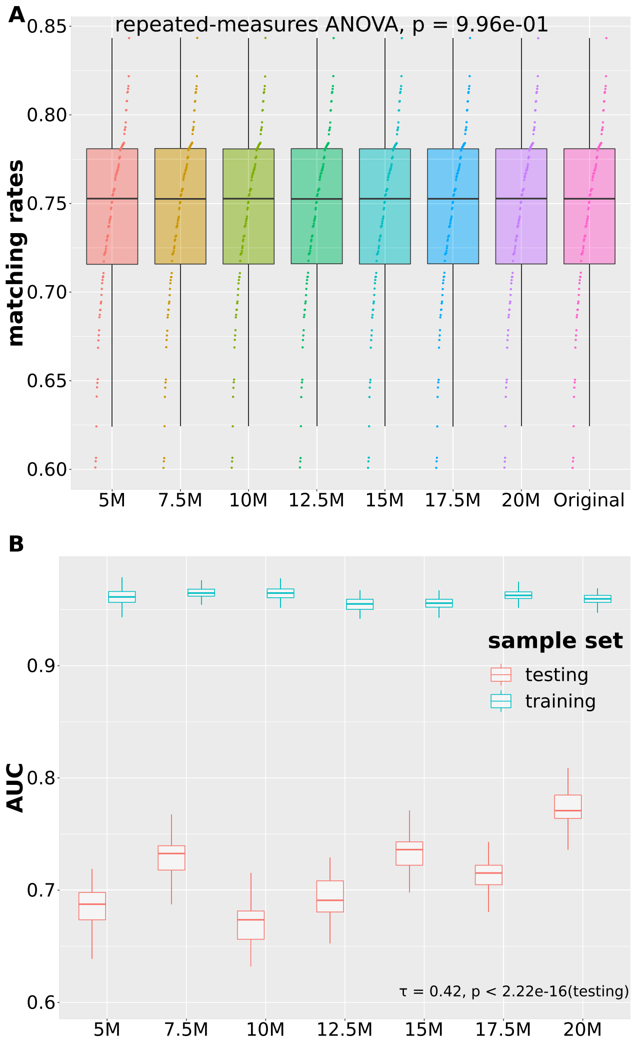
Our study also indicated that the species abundance was significantly influenced by sequencing depths, with a total of 493 species out of 603 being identified whose abundance was significantly influenced, which conflicts with some previous studies. Specifically, *Bacteroides stercoris*, *Klebsiella pneumoniae*, and *Veillonella atypica* were the species whose abundance was significantly found to be differing by both reference databases (Fig. 3, Fig. S1). Jovel et al. (2016) also indicated the same correlation in sequence depths from 500 to 100 000 reads, indicating that increasing the number of sequences results in more consistent estimations of bacteria relative abundance. On the other hand, the study of Hillmann et al. (2018) presented the results that the species profiles at 0.5 million and 2.5 billion sequences had high correlation of 0.990, implying little influence of sequence depth towards species richness and abundance. However, they only analyzed two samples for this purpose and even failed to analyze that at a fixed increment depth, which compromises the significance of the mentioned correlation (Hillmann et al. 2018). However, they

Table 2. Overview of disease classifiers using random forest method.

Disease associated		5M	7.5M	10M	12.5M	15M	17.5M	20M
Obese	Feature number	59.44 (2.43)	62.72 (3.15)	58.52 (4.64)	60.06 (3.21)	60.9 (3.55)	60.37 (3.14)	56.86 (3.21)
	AUC score of the training set	0.8963 (0.0045)	0.9042 (0.0049)	0.8974 (0.0051)	0.9046 (0.0047)	0.8991 (0.0045)	0.8933 (0.0047)	0.8964 (0.0041)
	AUC score of the testing set	0.838 (0.007)	0.8284 (0.0074)	0.8418 (0.0079)	0.8391 (0.0076)	0.8347 (0.0063)	0.84 (0.0061)	0.8496 (0.0056)
CRC	Feature number	49.69 (1.88)	54.16 (2.72)	56.53 (2.57)	56.78 (2.97)	56.53 (2.76)	53.82 (2.72)	54.64 (2.65)
	AUC score of the training set	0.961 (0.0069)	0.965 (0.0048)	0.9645 (0.0057)	0.9544 (0.0059)	0.9553 (0.006)	0.9622 (0.0051)	0.9593 (0.0051)
	AUC score of the testing set	0.6843 (0.0199)	0.7295 (0.0168)	0.6697 (0.0204)	0.6916 (0.0182)	0.7339 (0.0161)	0.7137 (0.014)	0.7727 (0.0166)

Note: Data are mean value (standard deviation) being run for 100 times. AUC, the area under a receiver operating characteristic curve; CRC, colorectal cancer.

Fig. 6. Validation analysis by CRC set. (A) Matching rates of each group at different depths; (B) AUCs of the models at each depth using the training set and the testing set calculated 100 times. AUC, the area under a receiver operating characteristic curve.



simply described a result in the supplementary note without providing any more evidence or data to support.

As for diversity, our study proved that the sequencing depth will significantly influence alpha diversity, which is the diversity of one sample, but with small changing rate less than 0.1%. In terms of beta diversity, 38 samples significantly differed in diversity at different sequence depths (>0.05) (Fig. 4, Fig. S2). The enterotypes proposed by Arumugam et al. (2011) group gut microbiome into three types according to beta diversity and they found that the sequencing depths would compromise the consistency of the enterotypes of one sample. However, we propose to adopt the probability of each of three enterotypes for the sample, which are more consistent at different depths (Table S4). The small the proportion of samples having low beta diversity (lower than 0.05) is, the more the comprehensive information a sequence data provides. We set less than 5% a standard for the ranking and found that groups with depths at 15 million and above had less than 5% samples presenting larger than 0.05 beta diversity (Fig. 4B, Fig. S2B). Therefore, 15 million depth was taken here as the recommended minimum depth in terms of beta diversity in our study. Our study also indicated that besides the samples, different sequence depths would also make the results of enterotype analysis unstable.

Regarding biomarker discovery, we found that although more marker genes were identified following the increase in sequence depth, the count of marker genes identified as intersected was not influenced significantly. We adopted MGS instead of marker genes as the biomarker in our study for the following considerations. First, researchers focus more on species in an MWAS to investigate the association between diseases and gut microbiome (Qin et al. 2012; Pedersen et al. 2016; Liu et al. 2017; Zhao et al. 2018; Hildebrand et al. 2021). Second, marker genes are too large in quantity and data dimension reduction is required to reduce the number of genes sharing the same features and cluster them into the MGS for further study. Lastly, we can validate on species level by qPCR but not on gene level (Yu et al. 2017).

We built random forest models based on the MGS as the disease classifiers and found that the reliability of the disease classifiers at different sequencing depths followed a certain pattern. The obesity models classified the obesity testing samples into groups as excellent with AUC scores from 0.82 to 0.85, increasing following the increase in the sequencing depths. In our validation set, although the CRC models classified the testing CRC samples as poor/acceptable with AUC scores from 0.67 to 0.77, it followed the same patterns with increased AUC scores following the increase in sequencing depths. It indicated that the increase in sequencing depth would improve the reliability of obesity classifiers.

The performance of the obesity classifiers and the CRC classifiers in our study varied significantly with obesity classifiers performing excellently and CRC classifiers poorly/acceptably, which is seen in other studies as well. The obesity classifiers established by Liu et al. (2017) grouped the obesity testing set outstandingly with ACU scores at 0.9214. The CRC classifier established in Feng's study got AUC scores of the testing set at 0.59 (Feng et al. 2015). The CRC classifier established by Yu's team got AUC scores of the testing set at 0.72 and 0.77 (Yu et al. 2017). We assumed that one reason is that the sample size is small. In our study, we got 24 samples out of 100 as the testing set to validate the performance of the model established by the other 76 datasets. More importantly, we assumed that the association between different diseases and the gut microbiota is different and that more phenotypic data shall be included when establishing CRC classifiers (Dai et al. 2018).

Conclusion

We concluded that the deeper the sequence, the more the microbial information in structure and composition it provides, specifically in gene identification, species richness and abundance, alpha- and beta diversity, and disease-associated biomarkers. More specifically, we found that when sequencing depth is 15 million or higher, we obtain more stable species compositions and disease classifiers with good performance. Therefore, we recommended 15 million as the optimal minimum sequencing depth for reference-based MWASs on either IGC or Metaphlan2.

Article information

History dates

Received: 14 December 2021
Accepted: 25 July 2022
Accepted manuscript online: 8 August 2022
Version of record online: 6 September 2022

Copyright

Data availability

The datasets analyzed are available from European Bioinformatics Institute (PRJID ERP013562 and ERP012177) (URL: https://www.ebi.ac.uk), open accessed.

Author information

Author ORCIDs

Hailiang Xie https://orcid.org/0000-0002-3626-6316

Competing interests

The authors declare that there are no competing interests.

Funding information

This research was supported by the horizontal cooperation project with Shenzhen Aimigene Institute (#201910D968) and the doctoral fund of Yuncheng University (#YQ-2019030).

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/gen-2021-0120.

References

Andersen, L.O., Bonde, I., Nielsen, H.B., and Stensvold, C.R. 2015. A retrospective metagenomics approach to studying Blastocystis. FEMS Microbiol. Ecol. **91**: fiv072. doi:10.1093/femsec/fiv072.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., et al. 2011. Enterotypes of the human gut microbiome. Nature, **473**: 174–180. doi:10.1038/nature09944.

Asnicar, F., Berry, S.E., Valdes, A.M., Nguyen, L.H., Piccinno, G., Drew, D.A., et al. 2021. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat. Med. **27**: 321–332. doi:10.1038/s41591-020-01183-8.

Benjamini, Y., and Hochberg, Y 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, **30**: 2114–2120. doi:10.1093/bioinformatics/btu170.

Bray, J.R., and Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. **27**: 325–349. doi:10.2307/1942268.

Breiman, L. 2001. Random forests. Mach. Learn. **45**: 5–32. doi:10.1023/A:1010933404324.

Dai, Z., Coker, O.O., Nakatsu, G., Wu, W.K.K., Zhao, L., Chen, Z., et al. 2018. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome, **6**: 70. doi:10.1186/s40168-018-0451-2.

Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., et al. 2017. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. GigaScience, **7**: gix133.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. 2015. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat. Commun. **6**: 6528. doi:10.1038/ncomms7528.

Ghoshal, U.C., and Gwee, K.A. 2017. Post-infectious IBS, tropical sprue and small intestinal bacterial overgrowth: the missing link. Nat. Rev. Gastroenterol. Hepatol. **14**: 435–441. doi:10.1038/nrgastro.2017.37.

Gu, Y., Wang, X., Li, J., Zhang, Y., Zhong, H., Liu, R., et al. 2017. Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. Nat. Commun. **8**: 1785. doi:10.1038/s41467-017-01682-2.

Gweon, H.S., Shaw, L.P., Swann, J., De Maio, N., AbuOun, M., Niehus, R., et al. 2019. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. Environ. Microbiome, **14**: 7. doi:10.1186/s40793-019-0347-1.

Hildebrand, F., Gossmann, T.I., Frioux, C., Özkurt, E., Myers, P.N., Ferretti, P., et al. 2021. Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host Microbe, **29**: 1167.e9–1176.e9.

Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., et al. 2018. Evaluating the information content of shallow shotgun metagenomics. mSystems, **3**: e00069–18. doi:10.1128/mSystems.00069-18.

Hosmer, D.W.J., Lemeshow, S., and Sturdivant, R.X. 2013. Applied logistic regression. John Wiley & Sons, Inc., Hoboken, NJ.

Hurlbert, S.H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology, **52**: 577–586. doi:10.2307/1934145.

Janda, J.M., and Abbott, S.L. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J. Clin. Microbiol. **45**: 2761–2764. doi:10.1128/JCM.01228-07.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S. Mitchel, T., et al. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front. Microbiol. **7**: 459. doi:10.3389/fmicb.2016.00459.

Kelly, T.N., Bazzano, L.A., Ajami, N.J., He, H., Zhao, J., Petrosino, J.F., et al. 2016. Gut microbiome associates with lifetime cardiovascular disease risk profile among Bogalusa Heart Study participants. Circ. Res. **119**: 956–964. doi:10.1161/CIRCRESAHA.116.309219.

Koo, T.K., and Li, M.Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. **15**: 155–163. doi:10.1016/j.jcm.2016.02.012.

Kursa, M.B., and Rudnicki, W.R. 2010. Feature selection with the Boruta package. J. Stat. Softw. **36**: 1–13. doi:10.18637/jss.v036.i11.

Lee, M.J., Kang, M.J., Lee, S.Y., Lee, E., Kim, K., Won, S., et al. 2018. Perturbations of gut microbiome genes in infants with atopic dermatitis according to feeding type. J. Allergy Clin. Immunol. **141**: 1310–1319. doi:10.1016/j.jaci.2017.11.045.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. **32**: 834–841. doi:10.1038/nbt.2942.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., et al. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, **25**: 1966–1967. doi:10.1093/bioinformatics/btp336.

Liu, R., Hong, J., Xu, X., Feng, Q., Zhang, D., Gu, Y., et al. 2017. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. Nat. Med. **23**: 859–868. doi:10.1038/nm.4358.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. **32**: 822–828. doi:10.1038/nbt.2939.

Pedersen, H.K., Gudmundsdottir, V., Nielsen, H.B., Hyotylainen, T., Nielsen, T., Jensen, B.A.H., et al. 2016. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature, **535**: 376–381. doi:10.1038/nature18646.

Pohlert, T. 2014. The pairwise multiple comparison of mean ranks package (PMCMR). R package. Availabe from https://CRAN.R-project.org/package=PMCMR.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, **490**: 55–60. doi:10.1038/nature11450.

Rausch, P., Ruhlemann, M., Hermes, B.M., Doms, S., Dagan, T., Dierking, K., et al. 2019. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome, **7**: 133. doi:10.1186/s40168-019-0743-1.

Ruppé, E., Lazarevic, V., Girard, M., Mouton, W., Ferry, T., Laurent, F., et al. 2017. Clinical metagenomics of bone and joint infections: a proof of concept study. Sci. Rep. **7**: 7718. doi:10.1038/s41598-017-07546-5.

Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A. Simpson, N., et al. 2019. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature, **574**: 117–121. doi:10.1038/s41586-019-1560-1.

Sharon, G., Cruz, N.J., Kang, D.W., Gandal, M.J., Wang, B., Kim, Y.M., et al. 2019. Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. Cell, **177**: 1600.e17–1618.e17. doi:10.1016/j.cell.2019.05.004.

Storey, J. D. 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Series B Stat. Methodol. **64**: 479–498. doi:10.1111/1467-9868.00346.

Tange, O. 2011. GNU parallel: the command-line power tool. The USENIX Magazine, February 2011: 42-47.

Thingholm, L.B., Rühlemann, M.C., Koch, M., Fuqua, B., Laucke, G., Boehm, R., et al. 2019. Obese individuals with and without type 2 diabetes show different gut microbial functional capacity and composition. Cell Host Microbe, **26**: 252.e10–264.e10.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods, **12**: 902–903. doi:10.1038/nmeth.3589.

Visconti, A., Le Roy, C.I., Rosa, F., Rossi, N., Martin, T.C., Mohney, R.P., et al. 2019. Interplay between the human gut microbiome and host metabolism. Nat. Commun. **10**: 4505. doi:10.1038/s41467-019-12476-z.

Wang, J., and Jia, H. 2016. Metagenome-wide association studies: fine-mining the microbiome. Nat. Rev. Microbiol. **14**: 508–522. doi:10.1038/nrmicro.2016.83.

Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., et al. 2017. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut, **66**: 70–78.

Zaheer, R., Noyes, N., Polo, R.O., Cook, S.R., Marinier, E., Domselaar, G.V., et al. 2018. Impact of sequencing depth on the characterization of the microbiome and resistome. Sci. Rep. **8**: 5890. doi: 10.1038/s41598-018-24280-8.

Zhao, L., Zhang, F., Ding, X., Wu, G., Lam, Y.Y., Wang, X., et al. 2018. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science, **359**: 1151–1156. doi:10.1126/science.aao5774.