A large-scale comparative metagenomic analysis of short-read sequencing platforms indicates high taxonomic concordance and functional analysis challenges

Kinga Zielińska[1*], Kateryna Pantiukh[2*], Paweł P Łabaj[1], Tomasz Kosciolek[3#] and Elin Org[2#]

[1] Małopolska Centre of Biotechnology, Jagielonian University, Kraków, Poland
[2] Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia;
[3] Sano Centre for Computational Medicine, Kraków, Poland
* These authors contributed equally
[#] correspondence: Elin Org elin.org@ut.ee, Tomasz Kosciolek t.kosciolek@sanoscience.org

Abstract

Driven by the increasing scale of microbiome studies and the rise of large, continuously expanding population cohorts, the volume of sequencing data is growing rapidly. As such, ensuring the comparability of data generated across different sequencing platforms has become a pressing concern in efforts to uncover robust links between the microbiome and human health. In this study, we conducted a comprehensive comparison of taxonomic and functional profiles from 1,351 matched human gut microbiome sample pairs, sequenced using both the MGISEQ-2000 (MGI) and NovaSeq 6000 (Illumina NovaSeq) platforms. Taxonomic profiles showed high concordance within and between platforms: 96.44 ± 5.96% of species were shared between MGI–MGI pairs, and 92.07 ± 5.20% were shared between MGI and NovaSeq pairs. The proportion of platform-specific species was low, at 3.42% for MGI–MGI comparisons and 5.89% for MGI–NovaSeq comparisons. No significant differences in Shannon diversity were observed for either within-platform or between-platform comparisons. However, functional profiles revealed notable discrepancies between platforms, which were attributed to differences in pre-sequencing protocols.

Importance

Our findings demonstrate robust taxonomic comparability between MGI and NovaSeq platforms, while revealing systematic functional differences that should be carefully considered in cross-platform and even cross-cohort metagenomic studies.

Introduction

Metagenomics has emerged as a powerful tool for uncovering key roles that microorganisms play in human health and environmental ecosystems (1),(2). Recent advancements in Next Generation Sequencing (NGS) technologies have enabled increasingly fine-scale taxonomic resolution, while also significantly expanding our understanding of microbial functions and

their dynamic interactions with hosts and other microbes. However, simultaneously with technological advancements, the expansion of NGS has been followed by new protocols and sequencing platforms. This introduced biases, complicating interpretation and reducing the reproducibility of signals identified through metagenomic data analyses.

A recent study by Forry et al. compared the Firmicutes:Bacteroides ratios defined from the same biological samples independently by 44 participating labs (3). The results revealed a substantial impact of a protocol choice, introducing both biases in measurements tied to specific methodological decisions, as well as influencing measurement robustness, as evidenced by the variability in results across labs employing similar methods. Beyond these protocols, metagenomic results - specifically microbial load and composition - are highly susceptible to biases stemming from sample storage at the earliest stages. This is particularly true for low-biomass samples, underscoring the necessity of an efficient, universal storage method to ensure robust cross-study comparisons (4).

The major providers of massive parallel sequencing technologies currently used in metagenomics are Illumina (sequencing by synthesis) and MGI/BGI (DNBSEQ) technology. In Illumina's sequencing-by-synthesis approach, fluorescently labeled nucleotides are incorporated one by one during the sequencing process and a camera captures the fluorescent signals to determine the base sequence (5). Illumina offers high sequencing accuracy and is well adopted globally, which makes it easier to integrate into established workflows. On the other hand, rapidly growing in popularity DNBSEQ technology amplifies DNA into DNA nanoballs (DNBs) and sequences them on patterned arrays (5),(6). The method is claimed to reach similar accuracy levels to Illumina while offering a more cost-effective solution.

Existing comparative analyses of sequencing by synthesis and DNBSEQ technologies conclude that the two could be used interchangeably. The comparisons cover a range of fields, starting with whole-genome DNA sequencing (7), through paleogenomics (8) and bacterial genome assembly. A range of sequencing-specific biases has been discovered, however, including a higher number of insertions detected in HiSeq 2000 and more deletions in BGISEQ-500 genome assemblies, as well as a bias toward genes with higher GC content being enriched on the HiSeq platforms (9). These could have profound impacts on the downstream metagenomic analyses, but the scopes and sizes of the available studies are limited and have not been further explored.

Here, we take advantage of a large population-based metagenomic dataset (1,351 individuals) that was simultaneously sequenced using the Illumina NovaSeq 6000 (hereafter NovaSeq) and MGISEQ-2000 (hereafter MGI) platforms, allowing us to directly compare two distinct sequencing technologies for metagenomic applications. We begin by evaluating the robustness of MGI sequencing through an analysis of samples sequenced twice (N=53) sequenced with MGI, assessing the degree of intra-sequencing variability. Building on this, we conduct a comprehensive comparative analysis of the full dataset, examining matched samples from both MGI and NovaSeq platforms to identify sequencing-specific biases (N=1,351).

Results

Study design and cohort overview

Our aim was to evaluate the consistency and comparability of microbiome sequencing outputs across two different platforms, with a specific focus on understanding how technological variations impact both taxonomic and functional profiles. To address this, we utilized a subset of the Estonian microbiome cohort (EstMB), which underwent sequencing on both NovaSeq and MGI platforms.

The original EstMB cohort is a population-based cohort consisting of 2,509 gut metagenomic samples, sequenced using paired-end metagenomic shotgun sequencing on the Illumina NovaSeq 6000 platform (hereafter referred to as NovaSeq), generating an average of 30.63 ± 3.12 million paired reads per sample. A comprehensive description of the cohort is available in our previous publication (**Figure 1a)** (10)**.**

A subset of the EstMB cohort was sequenced on the MGISEQ-2000 platform developed by MGI Tech Co, Ltd (hereafter referred to as MGI) with deeper coverage for metagenomic reconstruction purposes. This subset, referred to as the Estonian Microbiome deep cohort (EstMB-deep), includes 1,729 gut metagenomic samples, all of which are also represented in the main EstMB cohort. Additionally, 53 of the 1,729 samples were sequenced twice as technical replicates. Sequencing on the MGI platform yielded an average of 108.71 ± 42.08 million paired reads per sample. While the read length for MGI sequencing was consistent with that of the NovaSeq platform, the coverage was over three times deeper **(Figure 1a, 1b).**

To evaluate both within-platform and cross-platform variability, we first analyzed a subset of 53 samples that were randomly selected and sequenced twice on the MGI platform (MGI–MGI pairs). We then expanded the analysis to include 1,729 samples that were sequenced on both the MGI and NovaSeq platforms, forming MGI–NovaSeq pairs. To ensure a fair comparison between platforms, MGI reads were subsampled first to match the read depth of their NovaSeq counterparts. As a result of the initial quality assessment, sample pairs in which either sample failed quality control were excluded from further analysis. The resulting high-quality dataset comprised 1,351 MGI–NovaSeq sample pairs and consisted of 71.87% females and 28.12% males, ranging in age from 23 to 89 years with a mean age of 49.7 years **(Figure 1c).** The samples then underwent downstream taxonomic and functional profiling and subsequent cross-platform comparisons. (**Figure 1a**).

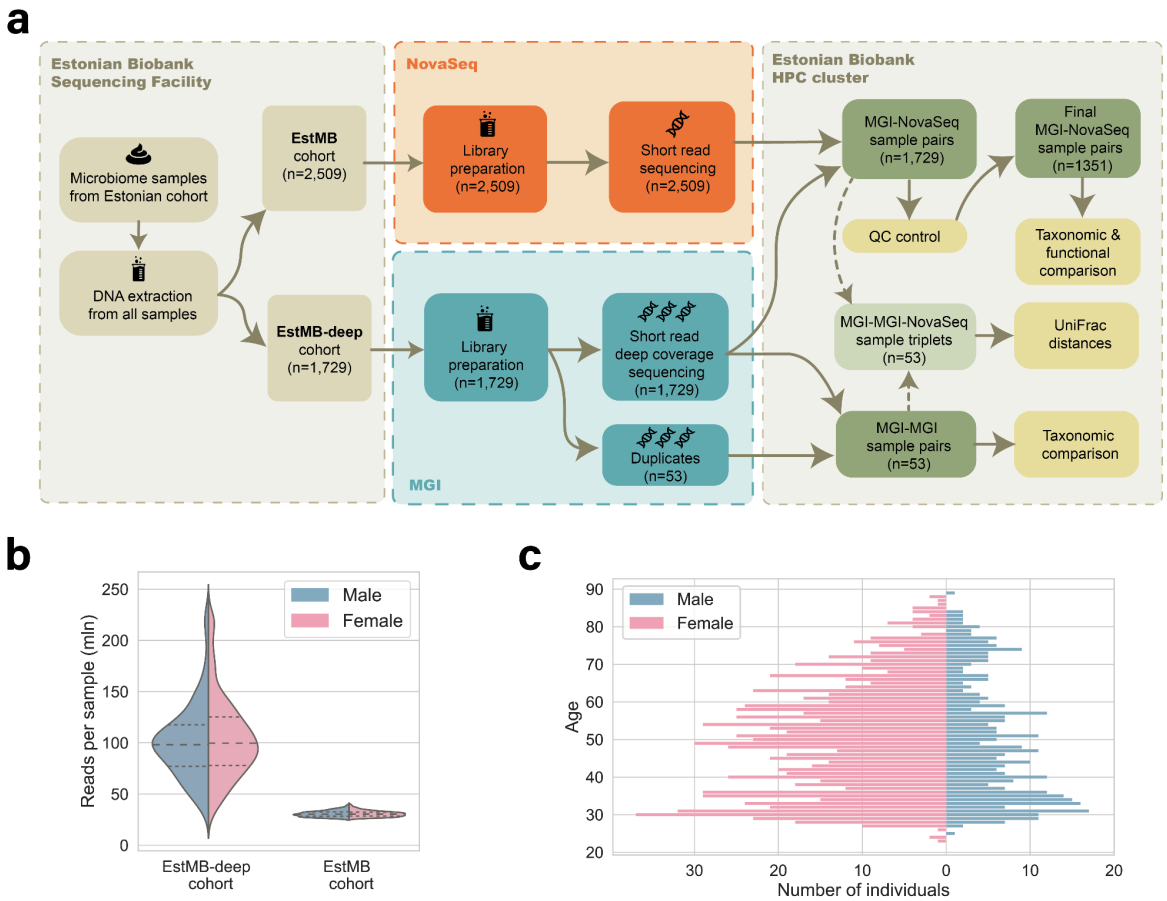

Figure 1. Study Overview. a. Study design and cohort characteristics **b.** Distribution of the number of reads per sample for MGI (EstMB-deep cohort) and NovaSeq (EstMB cohort) sequenced samples from the final sample set (N=1,351). **c.** Gender and age distributions of individuals in the final sample set (N=1,351).

Technical reproducibility: intra-platform comparison of MGI replicates

To evaluate intra-platform variability, we first examined whether replicate sequencing runs on the same platform introduce any systematic variation. This step established a baseline for assessing consistency within a single platform prior to cross-platform comparisons. Specifically, we analyzed a subset of 53 biological samples from our deeply sequenced cohort, sequenced twice on the MGISEQ-2000 platform. These replicate datasets - referred to as MGI–MGI sample pairs (**Figure 1a**) - originated from two sets (MGI-1 and MGI-2) and had been rarefied to equal read counts for a controlled comparison. To evaluate taxonomic consistency between the sample sets, we defined several comparison metrics: (1) the percentage of set-specific species, calculated as the number of species detected exclusively in one of the two sample sets divided by the total number of detected species; (2) the difference in the number of detected species; (3) the percentage of shared species; (4) the difference in species prevalence; (5) the difference in species relative abundance; and (6) the within-pair difference in Shannon diversity index. The summary of these results is presented in **Table 1**.

Across both MGI sequencing runs, a total of 1,083 species belonging to 380 genera were identified. Of these, 3.42% (37 species) were detected exclusively in one of the two runs - 17

species uniquely to the MGI-1 and 20 to the MGI-2 set (**Figure 2a, Supplementary Table S1**). These set-specific species exhibited a low prevalence and were typically of low relative abundance. Notably, 36 of the 37 species were observed in only a single sample, while one species (*Lactococcus cremoris*) was detected in two samples. The mean relative abundance of these set-specific species was approximately 2.25×10^{-5} %, in contrast to 7.67×10^{-4} % for species shared between both runs (**Figure 2b, Supplementary Figure 1**).

To evaluate the consistency of species detection between the two sequencing runs, we calculated a difference in the species counts for each MGI–MGI sample pair and visualized the distribution of these differences (**Figure 2c**). The distribution was centered around a mean difference of 0.85 species with a standard deviation of 2.88. Although the distribution may seem approximately normal, the Shapiro-Wilk test rejected normality ($p \ll 0.05$). We therefore used the Wilcoxon signed-rank test, which showed a significant difference between runs ($p = 0.032$) with a small effect size (Cohen's $d = 0.29$, where the coefficient of +/- 0.80 represents a strong effect). Taken together with the visual inspection, these results suggest that the observed difference was minor and likely attributable to random variation.

To extend our comparison, we assessed the overlap of species detected within MGI-MGI sample pairs. As illustrated in **Figures 2d** and **2e**, most sample pairs exhibited a high level of concordance, with more than 90% of species shared between the replicates (mean = 96.44% ± 5.96%). In addition, species prevalence and relative abundance exhibited strong agreement across sequencing runs, demonstrating high correlation with R^2 values approaching 1.0. (**Figures 2f** and **2g**). Finally, we compared Shannon diversity index values, a commonly used measure of community richness often associated with biological variables such as disease status. A paired t-test revealed no significant differences in Shannon index values between the two sequencing runs ($p > 0.05$), indicating consistent diversity estimates across replicates (**Figure 2h**).

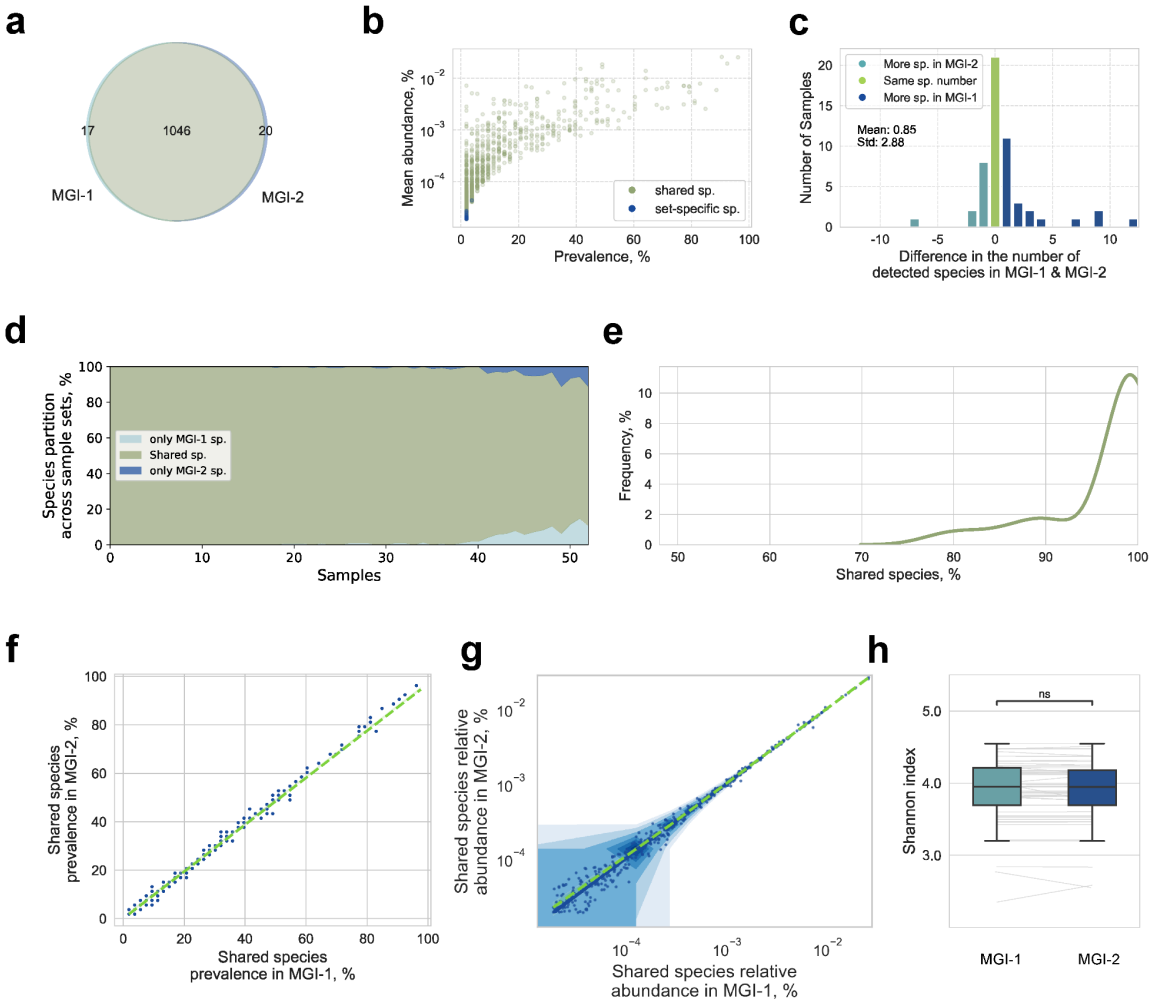

Figure 2. Intra-platform baseline comparison (N=53). **a.** Venn diagram of the number of species detected in MGI-1 and MGI-2. **b.** Prevalence versus mean relative abundance of detected species, with species found only in one run highlighted in blue. **c.** Distribution of differences in the number of detected species within sample pairs. Pairs with identical species counts are represented by green bars, while sample pairs with more species in MGI-1 or MGI-2 are represented by light or dark blue bars, respectively. **d.** The percentage of shared and unique run-specific species within each sample pair. **e.** Distribution of the percentage of shared species across sample pairs. **f.** Comparison of the prevalence of shared species between matched samples. **g.** Comparison of the relative abundance of shared species in matched samples. **h.** Comparison of the Shannon diversity Index within sample pairs.

Cross-platform reproducibility of taxonomic composition

After evaluating intra-platform variability, we addressed the central question of our study: whether MGI–NovaSeq sequencing platforms produce biologically comparable results. To address this, we performed a comprehensive analysis of paired samples generated on both sequencing platforms. Our evaluation encompassed not only taxonomic composition but also raw sequencing read quality and functional profiling; all of which were subsequently compared.

Assessment of intra-and inter-platform variability using MGI-MGI-Novaseq triplets

Before proceeding with the large-scale analysis of MGI–NovaSeq sample pairs, we first examined intra- and inter-platform variability using Shannon entropy and UniFrac distances on the subset

of the 53 samples. This time, however, we added the matched NovaSeq sample, which resulted in a set of samples sequenced three times: twice with the MGI and once with the NovaSeq platforms (MGI-MGI-NovaSeq triads). To enable biologically meaningful comparison, we used the UniFrac metric, which quantifies differences in microbial community composition based on bacterial phylogenetic relationships. We found no statistically significant difference between intra-platform and inter-platform UniFrac distances (PERMANOVA p-value = 1; **Supplementary Figure 2a**) and matched samples showed substantial overlap in the three-dimensional PCoA space (**Supplementary Figure 2b**). However, in 79% of the triads, the two MGI replicates were more similar to each other than to the corresponding NovaSeq replicate. Moreover, the mean distance between the MGI and NovaSeq samples was approximately 2.5 times greater than the distance between the MGI replicates, suggesting that intra-platform variability was consistently lower than cross-platform variability.

Raw reads sequencing quality assessment

To assess raw sequencing quality, we ran FastQC on all samples from the MGI–NovaSeq sample pairs (1,729 samples per platform). All samples passed key quality metrics, including per-base sequence quality and adapter content. One sample from each platform was flagged for "Overrepresented sequences" (those samples were from different pairs). GC content distributions were overall similar between the platforms, with a mean of 45.27 ± 1.82% for MGI and 45.60 ± 1.61% for NovaSeq. Despite the similarity, a paired Wilcoxon test indicated a statistically significant difference in the GC content ($p = 4.25 \times 10^{-25}$), although the effect size was small (Cohen's $d \approx 0.24$) (**Figure 3a**). The slightly higher GC content observed in NovaSeq samples aligns with the platform's known bias against low-GC regions, likely attributable to PCR amplification and cluster generation artifacts inherent to Illumina's bridge amplification (11). In contrast, DNBSEQ technology, utilizing rolling circle amplification to generate DNA nanoballs before loading onto a patterned array, aims to minimize such amplification biases.

Based on sequence duplication levels, 12.2% of all samples did not meet quality thresholds and were flagged as "warn" or "fail", indicating potential issues such as PCR amplification bias, low library complexity, or contamination. The failure rate was substantially higher for MGI (15.9%) compared to NovaSeq (8.4%) (**Figure 3b**). Notably, the failed samples did not consistently overlap across platforms, suggesting that the observed quality issues likely reflected platform-specific biases, differences in library preparation, or other technical artefacts. If this had been due to early-stage sample processing errors, both platforms would have been affected in a similar manner.

To ensure the robustness of the comparison, sample pairs containing at least one quality warning ("warn" or "fail", N=378) were excluded from downstream analysis. This yielded a final dataset of 1,351 high-quality sample pairs, which were subsequently used for taxonomic and functional assessments.

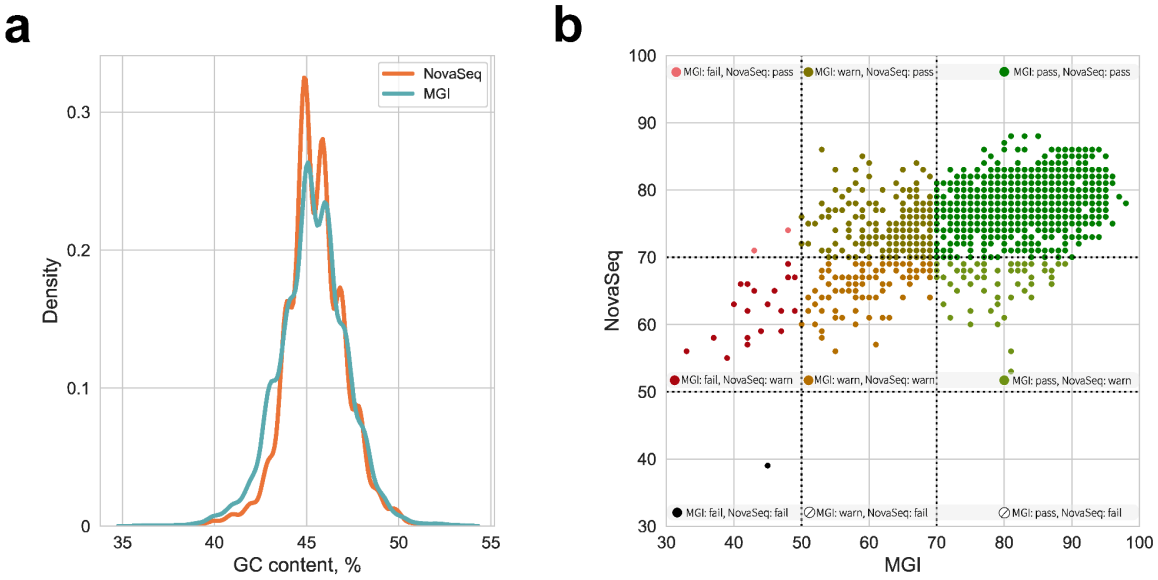

Figure 3. Inter-platform raw sequencing quality comparison (n=1,729) a. Distribution of the GC content across MGI and NovaSeq samples **b.** Sample pairs with their corresponding quality tags, assigned based on sequence duplication levels.

Cross-platform concordance in taxonomic profiles

We performed taxonomic comparisons for all sample pairs from the final dataset (N=1,351). Across both sequencing platforms, we identified a total of 2,953 species spanning 808 genera. Among these, 5.89% (174 species) were detected exclusively by one platform - 88 species were unique to MGI and 86 to NovaSeq (**Figure 4a, Supplementary Table S2**). Most of the platform-specific taxa were rare: 154 out of 174 appeared in only one sample, and their mean relative abundance (1.30×10^{-6} %) was markedly lower than that of the shared species (2.85×10^{-4} %) (**Figure 4b**).

To assess the consistency in species detection, we compared the difference in species counts for each paired sample and visualized the distribution of these differences (**Figure 4c**). The distribution was centered around a mean difference of -0.50, with a standard deviation of 4.13. The distribution failed the Shapiro–Wilk test ($p \ll 0.05$), prompting rejection of its normality and the use of the Wilcoxon signed-rank test, which showed a significant difference in species counts between platforms ($p = 1.03 \times 10^{-7}$). However, the effect size was small (Cohen's d = -0.12), and visual inspection suggested the difference was minimal and likely due to random variation.

To extend our comparison, we assessed the overlap of species detected within the matched samples. As shown in **Figures 4d** and **4e**, most pairs exhibited a high level of concordance, with more than 90% of species shared between the replicates (mean = 92.07 ± 5.20%). In addition, species prevalence and relative abundance exhibited strong agreement across sequence runs, demonstrating high correlation with R^2 values approaching 1.0 (**Figures 4f** and **4g, Supplementary Figure 3**).

No significant difference in Shannon diversity was observed between the platforms (paired t-test, $p > 0.05$) (**Figure 4h**). Additionally, Procrustes analysis demonstrated excellent concordance between sample pairs across the two platforms ($M^2 = 0.02$, $p = 0.0001$; **Supplementary Figure 4**). Differential abundance analysis with MaAsLin2 initially revealed three species enriched in either platform (two from the *ER4* genus family *Oscillospiraceae* and one *Paraprevotella*), but they all failed to pass the correction for multiple testing (**Supplementary Table S3**).

To test the biological interpretability of the results, we examined associations between microbial taxa and health-related variables using linear regression, separately for each sequencing platform. Taxonomic associations were overall consistent between the MGI and NovaSeq datasets, with only two taxa showing platform-specific significance but with negligible effect sizes (**Supplementary Table S4**).

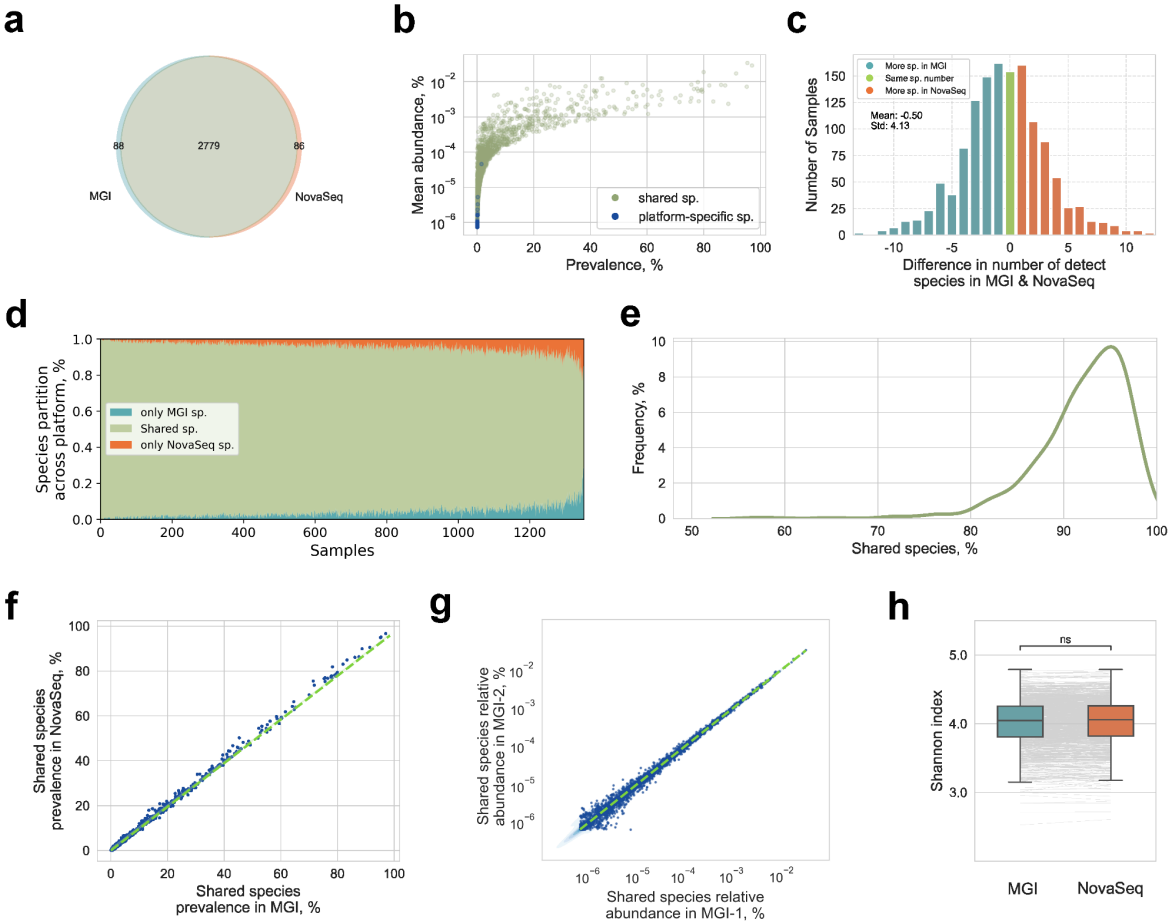

Figure 4. Taxonomic composition comparison between sequencing platforms (n=1351) **a.** Proportion of species shared between platforms and those detected exclusively in MGI or NovaSeq. **b.** Prevalence and relative abundance of all detected species, with platform-specific species (found only in one platform) highlighted in blue. **c.** Distribution of differences in the number of species detected per sample pairs. Pairs with identical species counts are shown in green, those with more species in MGI are shown in light blue and with more species in NovaSeq are shown in orange. **d.** Percentage of shared and platform-specific species across sample pairs **e.** Distribution of the percentage of shared species within each sample pair **f.** Species prevalence for taxa shared between platforms within sample pairs. **g.** Relative abundance of shared species within sample pairs **h.** Comparison of Shannon diversity índex within each sample pair.

The results of the intra- and inter-platform taxonomic comparisons are summarised in **Table 1**.

Table 1. Comparison summary: intra-platform (MGI-MGI) and inter-platform (NovaSeq-MGI) taxonomy metrics.

Metric	Intra-platform comparison (MGI-MGI sample pairs)	Inter-platform comparison (NovaSeq-MGI sample pairs)
Number of sample pairs	53	1,351
Percentage of set-specific species	3.42% (37 out of 1,083 total species)	5.89% (174 out of 2953 total species)
Difference in the number of species per sample pair (Wilcoxon signed-rank test)	significant difference ($p = 0.038$) Effect size: small (Cohen's $d = 0.29$)	significant difference ($p = 1.03 \times 10\text{-}7$) Effect size: very small (Cohen's d = -0.12)
Mean percentage of shared species within sample pairs	96.44% ± 5.96%	92.07% ± 5.20%
Correlation of species' prevalence and relative abundance between platforms within sample pairs	R^2=0.999	R^2=0.999
Difference in Shannon diversity index between platforms within sample pairs (paired t-test)	$p > 0.05$	$p > 0.05$

Platform-specific functional differences reflect sequencing depth and complexity

To systematically assess differences in functional profiling between the sequencing platforms, we randomly selected 200 matched sample pairs from our final dataset of 1,351 metagenomic samples. Functional annotation was performed with mifaser, which resulted in the identification of 1,192 unique enzymes (EC numbers) across all samples. Of these, 3.52% (42 enzymes) were detected exclusively by one platform - 38 enzymes uniquely to MGI and 4 uniquely to NovaSeq (**Figure 5a** and **b**).

The degree of functional similarity within each matched pair, based on enzyme presence/absence, ranged from 80% to 97% - mirroring the pattern observed for taxonomic profiles (**Figure 5c**). However, the MGI consistently yielded a higher number of detected enzymes per sample: in every pair, the MGI sample either matched or exceeded the enzyme count of its NovaSeq counterpart. To further examine this discrepancy, we conducted a differential abundance analysis using MaAsLin2, which revealed 131 enzymes significantly enriched in one platform over the other (FDR-adjusted $p < 0.05$, **Supplementary Table 5**). These differences thus extended beyond enzyme counts, highlighting systematic variation in the functional detection.

We hypothesized that differences in sequence diversity might underlie the observed functional discrepancies between the platforms. To investigate this, we performed a k-mer analysis to evaluate the complexity of sequence data captured by each platform, as variation in k-mer diversity can indicate platform-specific biases affecting functional annotation independently of taxonomic composition. Using the Jellyfish tool, we quantified unique k-mers and found that MGI samples consistently contained 5–10 times more unique k-mers than their NovaSeq counterparts (**Figure 5d**). Despite the pronounced difference in k-mers diversity, the number of reads assigned taxonomic labels was similar across platforms. However, MGI samples yielded several-fold more functionally annotated reads (**Figure 5e**), suggesting that the increased sequence diversity in MGI data has limited impact on taxonomic profiling, but substantially improves functional annotation.

To explore whether functional difference was linked to taxonomic differences, we ordered sample pairs by decreasing taxonomic overlap and compared their functional similarity. No correlation was observed between the two (Spearman's $\rho = 0.005$, $p > 0.05$, **Figure 5c** and **g**), indicating that functional differences are not simply a consequence of taxonomic variation. Further analysis of genome coverage across a representative genomic fragment revealed platform-specific biases: MGI and Illumina samples exhibited non-overlapping and uneven coverage across genomic regions (**Supplementary Figure 5**, see Methods: genome coverage). Despite these functional differences, we observed no strong clustering by sequencing platform. In a plot representing each sample as three components - shared, MGI-unique, and Illumina-unique enzyme fractions - the values broadly overlapped across platforms with MGI points slightly but consistently higher (**Figure 5f**).

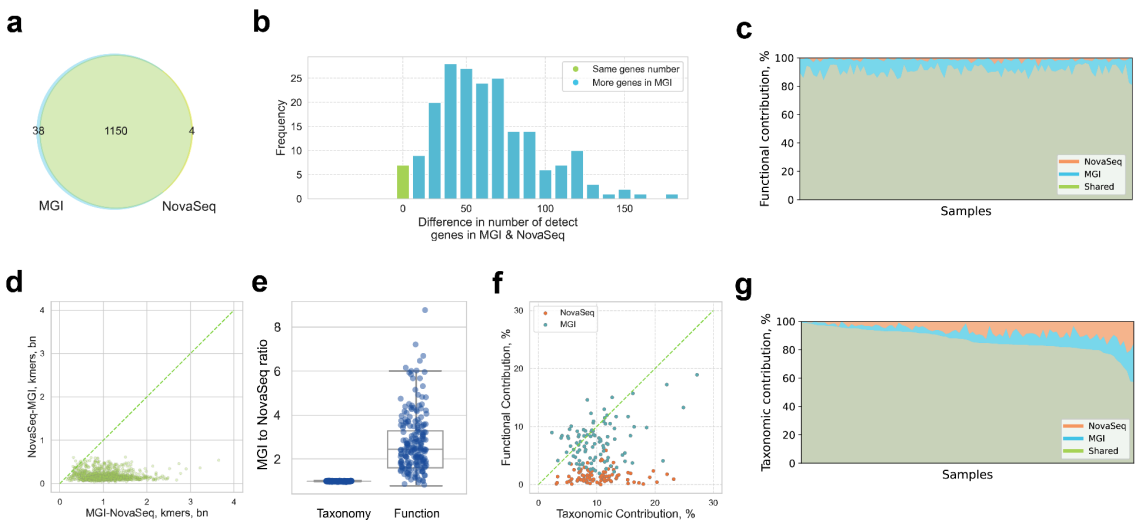

Figure 5. Function-based comparison of the MGI and NovaSeq batches (n=200). (a) Proportion of genes shared between platforms and platform-specific genes, detected exclusively in MGI or NovaSeq platforms. (b) Overlap of functions in matched MGI and NovaSeq samples. c) A functional overlap in 200 exemplary samples with different levels of taxonomic overlap, sorted by descending functional overlap. (d) Unique k-mers per sample: in NovaSeq versus MGI and vice versa. (e) Contributions of taxonomic and functional microbiome components found uniquely in MGI or NovaSeq. (f) Percentage of reads with taxonomic and functional assignments per sample, presented as MGI to NovaSeq ratio. (g) A taxonomic overlap of matched samples, sorted by descending functional overlap in (c).

Discussion

Previous comparative analyses of MGI and Illumina sequencing technologies have largely suggested their potential interchangeability for applications in metagenomics and clinical diagnostics (12),(13). However, these existing investigations often suffer from limited sample sizes, which can hinder the generalizability and statistical power of the findings. Furthermore, many of these studies have not rigorously isolated the impact of the sequencing technology itself, instead conflating it with other upstream methodological variables, such as library preparation protocols or bioinformatics pipelines. This makes it challenging to definitively attribute observed differences or similarities solely to the sequencing platform. Our study addresses these critical gaps by presenting one of the first large-scale comparisons of MGI and Illumina NovaSeq sequencing technologies. We provide a comprehensive and objective evaluation of MGI's inherent robustness and consistency, examining its performance not only through extensive intra-platform comparisons but also through a direct and detailed assessment against the widely adopted NovaSeq platform. This work aims to offer clearer insights into the reliability and comparability of these leading sequencing systems, which is crucial for advancing the field of metagenomics.

A deeper investigation into sequencing quality parameters revealed differences in GC content distribution and sequence duplication levels between platforms, suggesting potential sequencing biases. A published study focusing specifically on the GC content across numerous sequencing platforms in the context of metagenomic analysis highlighted the challenge of identifying GC-poor organisms (11). Specifically, the authors found that workflows using MiSeq and NextSeq were hindered by major GC biases, with problems becoming increasingly severe outside the 45–65% GC range. This led to a falsely low coverage in GC-rich and especially GC-poor sequences, where genomic windows with 30% GC content had >10-fold less coverage than windows close to 50% GC content. In addition, they showed a tight correlation of the GC content with coverage biases. On the other hand, observed discrepancies, mostly within the 40-50% GC coverage range, were not systematic enough to influence major biological interpretations. We found strong agreement in taxonomic profiles, with an over 94% overlap in species, genera, and families. Key diversity metrics, including Shannon entropy, UniFrac distances and disease-microbiome associations did not differ significantly between sequencing technologies. However, MGI samples tended to cluster more closely together than with their matched NovaSeq counterparts, indicating that while intra-platform variation is minimal, inter-platform differences did exist.

Our observations revealed a substantial discrepancy in the functional overlap, which was considerably lower and did not align with the taxonomic findings. This suggests that factors beyond mere taxonomic presence are influencing the functional profiles we observed. A recent study by the ABRF consortium sheds light on this issue, demonstrating how the initial amount of the starting material and subsequent PCR duplication rates can introduce significant biases (14). Their research, though focused on RNA-seq, found that low input material leads to a reduced number of detected genes and increased noise in expression counts. These conclusions are highly relevant to metagenomics, especially with the rise of higher multiplexing strategies.

Higher multiplexing, while enabling larger study sizes at a more affordable price, inherently reduces the amount of input material per sample. This, in turn, can lead to higher duplication rates and uneven coverage of the genomes being studied. Taxonomic annotation tools typically consider multiple genomic positions and rely on the copy number for abundance determination, making them less susceptible to uneven coverage. In contrast, functional profiling depends on smaller genomic fragments and requires reasonably equal coverage across the genome, a condition much more challenging to achieve when input material is limited and duplication rates are high. This phenomenon aligns well with our findings, where the MGI samples were sequenced in smaller sets and at higher depth, which maintained higher and more consistent original coverage. Although rarefaction was applied to normalize read depth, this correction was insufficient to eliminate disparity in the functional detection.

Conclusions

Our study confirms that MGI sequencing provides a reliable and robust alternative to NovaSeq for taxonomic microbiome profiling. While both platforms yield comparable taxonomic results, we observed differences at the functional level. Our results highlight the need for methodological consistency when comparing results across different sequencing technologies, especially in metagenomics studies using a gene catalog.

For accurate and comparable outcomes, standardized sample preparation is a critical first step. Furthermore, the experimental design should prioritize sequencing depth to ensure comprehensive results. We strongly recommend limiting the degree of sample multiplexing, as this directly increases the number of reads allocated to each sample. The resulting higher coverage is crucial for capturing a wider dynamic range of genome coverage and achieving more reliable functional annotations.

Methods

Cohort overview

The Estonian Microbiome Cohort (EstMB) was established in 2017, when stool, oral, and blood samples were collected from 2,509 Estonian Biobank (EstBB) participants. The EstBB is a volunteer-based population cohort initiated in 1999 that currently includes over 202,282 genotyped adults (≥ 18 years old) across Estonia. A detailed overview of the Estonian Biobank (EstBB) is described in Milani et al. 2025 (15) and an overview of the Estonian Microbiome Cohort (EstMB), including omics and phenotypic data availability, is described in Aasmets & Krigul et al. (10). In brief, EstMB cohort includes 1,308 women (69.65 %) and 570 men (30.35 %), representing ages 23 to 89. Extensive information was collected on the EstMB participants, including electronic health records (EHRs) and the participants' anthropometric measurements (age, gender BMI, stool sample etc.).

The Estonian Microbiome Deep cohort (EstMB-deep) includes a subset of samples from the EstMB that have been resequenced with deep coverage with MGISEQ-2000 platform (N = 1,729). Samples for EstMB-deep were randomly selected from EstMB, with the number of samples based on DNA availability.

Microbiome sample collection and DNA extraction

The participants collected a fresh stool sample immediately after defecation with a sterile Pasteur pipette and placed it inside a polypropylene conical 15 mL tube. The participants were instructed to time their sample collection as close as possible to the visiting time in the study centre The samples were stored at −80 °C until DNA extraction. The median time between sampling and arrival at the freezer in the core facility was 3 h 25 min (mean 4 h 34 min) and the transport time wasn't significantly associated with alpha (Spearman correlation, p-value 0.949 for observed richness and 0.464 for Shannon index) nor beta diversity (p-value 0.061, R-squared 0.0005). Microbial DNA extraction was performed after all samples were collected using a QIAamp DNA Stool Mini Kit (Qiagen, Germany). For the extraction, approximately 200 mg of stool was used as a starting material for the DNA extraction kit, according to the manufacturer's instructions. DNA was quantified from all samples using a Qubit 2.0 Fluorometer with a dsDNA Assay Kit (Thermo Fisher Scientific).

Metagenomic data processing

Illumina NovaSeq 6000 platform

The shotgun metagenomic paired-end sequencing was performed for whole EstMB cohort (N=2,509) by Novogene Bioinformatics Technology Co., Ltd. using the Illumina NovaSeq6000 platform, resulting in 4.62 ± 0.44 Gb of data per sample (insert size, 350 bp; read length, 2 x 250 bp). A NEBNext® Ultra™ DNA Library Prep Kit for NovaSeq (NEB,USA) was used for generating sequencing libraries, according to the manufacturer's recommendations. The reads were trimmed for quality and adapter sequences. The host reads that aligned to the human genome were removed using SOAP2.21 (parameters: -s 135 -l 30 -v 7 -m 200 -x 400).

MGISEQ-2000 platform

Metagenomic paired-end sequencing was performed for EstMB-deep cohort (N=1,729) using DNBSEQI technology resulting in 22.47 ± 4.2 Gb of data per sample (insert size, 350 bp; read length, 2 x 250 bp). Library preparation and circularization of equimolarly pooled libraries were done with MGIEasy FS DNA Library Prep Set (MGI Tech Co, Ltd, Shenzhen, China) according to the standard protocol. The shotgun metagenomic paired-end sequencing was performed by MGISEQ-2000 High-throughput Sequencing Set (FCL PE150) according to the manufacturer's instructions (MGI Tech, Shenzhen, China). Sequencing reads' quality control (QC) was performed using Fastqc (v 0.12.1) (16) and human reads were filtered using Bowtie2 (v 0.6.5) against the GRCH build 38, patch release 14 (17),(18).

Bioinformatics and statistical analysis

Sample selection

For downstream analysis, we first selected samples that were sequenced on both sequencing platforms and were present in both the EstMB and EstMB-deep cohorts (N=1729). Since the number of NovaSeq reads was consistently lower than the number of

MGI reads for each sample pair, we excluded pairs with fewer than 13 million or more than 20 million NovaSeq reads to avoid potential biases in comparative analyses. Next, we adjusted the number of MGI reads to match the number of NovaSeq reads within each sample pair with seqtk v1.3 (19). The final dataset consisted of $1{,}351$ samples, with an average of 4 GB of sequencing data or approximately 20 million reads per sample. For the intra-sequencing comparison, we used samples which had been sequenced with MGI twice (n=53).

Read quality comparison

The FastQC report files were aggregated using custom scripts (parse_fastqc_NovaSeq.py, parse_fastqc_mgi.py) and subsequently analyzed using standard Python packages, including pandas, matplotlib, and numpy (fastQC.ipynb). The quality control was conducted to assess base sequence quality, adapter content, sequence length distribution, GC content, sequence duplication levels, and overrepresented sequences. Differences in GC content between sequencing platforms were evaluated using the Wilcoxon signed-rank test, and correlation was assessed using Pearson's correlation coefficient. Samples exceeding sequence duplication thresholds - indicative of PCR amplification bias, low library complexity, or potential contamination - were flagged. Overrepresented sequences were identified based on FastQC's criteria for sequence enrichment beyond expected distributions. For all downstream analyses, we included only those sample pairs in which both samples passed all quality criteria (N = 1,351).

Taxonomic profiling

Metagenomic sequencing data were assigned taxonomic labels using Kraken2 v2.1.1 and Bracken v2.8 with a custom database (GTDB release 214) (20), (21). Kraken report files were processed using custom scripts (parse_kraken_NovaSeq.py, parse_kraken_mgi.py) to aggregate the relative abundance values across all samples, retaining only values greater than 0.01% in the summary file. Species relative abundance was defined based on the Kraken's 'fraction_total_reads' values and prevalence was calculated as the number of samples in which the species was detected.

Statistical analysis

Differences in species counts between sample pairs were tested for normality using the Shapiro–Wilk test. As distributions deviated significantly from normality ($p \ll 0.05$), comparisons between runs and platforms were performed using the Wilcoxon signed-rank test. Effect sizes were calculated using Cohen's d.

Further analyses

Alpha and beta diversity measures (UniFrac and procrustes analyses) were performed using QIIME 2 with default parameters (22). Functional predictions were made on a random subset of 200 samples with mifaser with standard parameters. K-mers were calculated with jellyfish v.2.3.0 (23). Differential enrichment analysis was done with MaAsLin2 in its default mode (24). We examined the association between centered log-ratio (CLR) transformed species abundances and participants' health status for common diseases in the Estonian population. We selected 33 common diseases based on ICD10 codes from the electronic health

records, each with over 100 prevalent cases within the EstMB cohort. The remaining samples were considered as controls for each studied disease. Linear regression models, adjusted for BMI, gender, and age, were constructed to evaluate the association between the selected diseases and CLR-transformed species abundance. A stringent Bonferroni correction was applied to the significance level, adjusting for the number of analyzed species.

Genome coverage

To evaluate sequencing coverage of key microbial species across samples, we selected four representative gut bacteria frequently used in metagenomic benchmarking: *Akkermansia muciniphila*, *Bacteroides uniformis*, *Faecalibacterium prausnitzii*, and *Prevotella copri*. For each species, reference genomes were downloaded from the NCBI genome database (accessions: GCA_000020225.1, GCA_025147485.1, GCA_003324185.1, and GCA_025151535.1, respectively). Genome sequences in FASTA format were concatenated into a single reference file and indexed using Bowtie2 to create a searchable index (17).
To determine coverage, post-QC paired-end sequencing reads from each sample were individually aligned to the combined reference using Bowtie2. The alignment was converted to BAM format and sorted using SAMtools (25). Per-base genome coverage was then calculated with BEDTools' genomecov function, using the reference genome file to define genome sizes (26). This process was repeated for each sample to quantify species-level coverage separately across all conditions.
Genomic coverage comparisons between the two sequencing methodologies were performed by visual inspection of mapped reads in IGV (Integrative Genomics Viewer) (27). For illustrative purposes, a representative genomic fragment was randomly selected. This particular region clearly demonstrated the observed differences in coverage uniformity between the evaluated technologies.

Code availability

The source code for the analyses is available at GitHub:
https://github.com/Chartiza/2024_Illumina_vs_BGI

Acknowledgements

We express our gratitude to all individuals who made valuable contributions to the Estonian Microbiome cohort - Oliver Aasmets, Kertu Liis Krigul, Kreete Lüll and Steven Smith. We thank the Estonian Biobank research team: Andres Metspalu, Lili Milani, Tõnu Esko and Mait Metspalu for providing health data from Estonian Biobank participants. Data analysis was carried out in part in the High-Performance Computing Centre of University of Tartu, and we thank HPC Support Team of the Institute of Computer Science at the University of Tartu for delivering exceptional service and assistance in installing the necessary programs on the cluster. Special thanks to the Writing Retreat organized by the Institute of Genomics, University of Tartu for providing a conducive atmosphere for writing this paper. We also thank MMHP (Million Metagenomes of Human Project) for providing the sequencing for EstMB-deep cohort. Last but not least, we thank Łukasz Szydłowski for useful discussions during the early stages of this study and Rafał Mostowy for inviting Elin Org as a part of his EMBO Installation Grant, which established this project.

Funding

This work was funded by Estonian Research Council grant (PUT 1371, PRG1414 to E.O) and an EMBO Installation grant (No. 3573 to E.O.). EstMB sample collection was supported by Estonian Center of Genomics/Roadmap II project No 16-0125. This research was also conducted as a part of the NCN Sonata BIS grant number 2020/38/E/NZ2/00598. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grants no. PLG/2023/016234 and no. PLG/2024/017180. TK is funded by the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

Bibliography

1. Pinto Y, Bhatt AS. 2024. Sequencing-based analysis of microbiomes. Nat Rev Genet 25:829–845.
2. Gilbert JA, Azad MB, Bäckhed F, Blaser MJ, Byndloss M, Chiu CY, Chu H, Dugas LR, Elinav E, Gibbons SM, Gilbert KE, Henn MR, Ishaq SL, Ley RE, Lynch SV, Segal E, Spector TD, Strandwitz P, Suez J, Tropini C, Whiteson K, Knight R. 2025. Clinical translation of microbiome research. Nat Med 31:1099–1113.
3. Forry SP, Servetas SL, Kralj JG, Soh K, Hadjithomas M, Cano R, Carlin M, Amorim MG de, Auch B, Bakker MG, Bartelli TF, Bustamante JP, Cassol I, Chalita M, Dias-Neto E, Duca AD, Gohl DM, Kazantseva J, Haruna MT, Menzel P, Moda BS, Neuberger-Castillo L, Nunes DN, Patel IR, Peralta RD, Saliou A, Schwarzer R, Sevilla S, Takenaka IKTM, Wang JR, Knight R, Gevers D, Jackson SA. 2024. Variability and bias in microbiome metagenomic sequencing: an interlaboratory study comparing experimental protocols. Sci Rep 14:9785.
4. Marotz C, Cavagnero KJ, Song SJ, McDonald D, Wandro S, Humphrey G, Bryant M, Ackermann G, Diaz E, Knight R. 2021. Evaluation of the Effect of Storage Methods on Fecal, Saliva, and Skin Microbiome Composition. mSystems 6:10.1128/msystems.01329-20.
5. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E. Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang G-D, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA,

Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, vandeVondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

6. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T, Sun N, Rao J, Wang J, Zhang W, Chen Y, Liao S, Jiang H, Liu X, Yang Z, Mu F, Gao S. 2017. A reference human genome dataset of the BGISEQ-500 sequencer. GigaScience 6:gix024.
7. Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A. 2020. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. PLOS ONE 15:e0230301.
8. Zhu K, Du P, Xiong J, Ren X, Sun C, Tao Y, Ding Y, Xu Y, Meng H, Wang C-C, Wen S-Q. 2021. Comparative Performance of the MGISEQ-2000 and Illumina X-Ten Sequencing Platforms for Paleogenomics. Front Genet 12.
9. Hu T, Chen J, Lin X, He W, Liang H, Wang M, Li W, Wu Z, Han M, Jin X, Kristiansen K, Xiao L, Zou Y. 2024. Comparison of the DNBSEQ platform and Illumina HiSeq 2000 for bacterial genome assembly. Sci Rep 14:1292.
10. Aasmets O, Krigul KL, Lüll K, Metspalu A, Org E. 2022. Gut metagenome associations with extensive digital health data in a volunteer-based Estonian microbiome cohort. Nat Commun 13:869.
11. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A, Hansen LH. 2020. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience 9:giaa008.
12. Kumar A, Robertsen EM, Willassen NP, Fu J, Hjerde E. 2023. Comparative analysis of HiSeq3000 and BGISEQ-500 sequencing platform with shotgun metagenomic sequencing data. Genomics Inform 21:e49.
13. Han S, Zhao Z, Yang L, Huang J, Wang Y, Feng J. 2023. The performance of metagenomic next-generation sequencing in diagnosing pulmonary infectious diseases using authentic clinical specimens: The Illumina platform versus the Beijing Genomics Institute platform. Front Pharmacol 14.
14. Zajac N, Vlachos IS, Sajibu S, Opitz L, Wang S, Chittur SV, Mason CE, Knudtson KL, Ashton JM, Rehrauer H, Aquino C. 2025. The impact of PCR duplication on RNAseq data generated using NovaSeq 6000, NovaSeq X, AVITI, and G4 sequencers. Genome Biol 26:145.
15. Milani L, Alver M, Laur S, Reisberg S, Haller T, Aasmets O, Abner E, Alavere H, Allik A, Annilo T, Fischer K, Hofmeister R, Hudjashov G, Jõeloo M, Kals M, Karo-Astover L, Kasela S, Kolde A, Krebs K, Krigul KL, Kronberg J, Kruusmaa K, Kukuškina V, Kõiv K, Lehto K, Leitsalu L, Lind S, Luitva LB, Läll K, Lüll K, Metsalu K, Metspalu M, Mõttus R, Nelis M, Nikopensius T, Nurm M, Nõukas M, Oja M, Org E, Palover M, Palta P, Pankratov V, Pantiukh K, Pervjakova N, Pujol-Gualdo N, Reigo A, Reimann E, Smit S, Rogozina D, Särg D, Taba N, Talvik H-A, Teder-Laving M, Tõnisson N, Vaht M, Vainik U, Võsa U, Yelmen B, Esko T, Kolde R, Mägi R, Vilo J, Laisk T, Metspalu A. 2025. The Estonian Biobank's journey from biobanking to personalized medicine. Nat Commun 16:3270.
16. LaMar D. 2015. FastQC https://qubeshub.org/resources/fastqc.
17. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
18. The 1000 Genomes Project Consortium, Corresponding authors, Auton A, Abecasis GR, Steering committee, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A,

Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Production group, Baylor College of Medicine, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, BGI-Shenzhen, Wang J, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Broad Institute of MIT and Harvard, Lander ES, Altshuler DM, Gabriel SB, Gupta N, Coriell Institute for Medical Research, Gharani N, Toji LH, Gerry NP, Resch AM, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Illumina, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Max Planck Institute for Molecular Genetics, Lehrach H, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, McDonnell Genome Institute at Washington University, Mardis ER, Wilson RK, Fulton L, Fulton R, US National Institutes of Health, Sherry ST, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, University of Oxford, McVean GA, Wellcome Trust Sanger Institute, Durbin RM, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Analysis group, Affymetrix, Schmidt JP, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Albert Einstein College of Medicine, Auton A, Campbell CL, Kong Y, Marcketta A, Baylor College of Medicine, Gibbs RA, Yu F, Antunes L, Bainbridge M, Muzny D, Sabo A, Huang Z, BGI-Shenzhen, Wang J, Coin LJM, Fang L, Guo X, Jin X, Li G, Li Q, Li Y, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Bilkent University, Alkan C, Dal E, Kahveci F, Boston College, Marth GT, Garrison EP, Kural D, Lee W-P, Fung Leong W, Stromberg M, Ward AN, Wu J, Zhang M, Broad Institute of MIT and Harvard, Daly MJ, DePristo MA, Handsaker RE, Altshuler DM, Banks E, Bhatia G, del Angel G, Gabriel SB, Genovese G, Gupta N, Li H, Kashin S, Lander ES, McCarroll SA, Nemesh JC, Poplin RE, Cold Spring Harbor Laboratory, Yoon SC, Lihm J, Makarov V, Cornell University, Clark AG, Gottipati S, Keinan A, Rodriguez-Flores JL, European Molecular Biology Laboratory, Korbel JO, Rausch T, Fritz MH, Stütz AM, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Beal K, Clarke L, Datta A, Herrero J, McLaren WM, Ritchie GRS, Smith RE, Zerbino D, Zheng-Bradley X, Harvard University, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Human Gene Mutation Database, Cooper DN, Ball EV, Stenson PD, Illumina, Bentley DR, Barnes B, Bauer M, Keira Cheetham R, Cox A, Eberle M, Humphray S, Kahn S, Murray L, Peden J, Shaw R, Icahn School of Medicine at Mount Sinai, Kenny EE, Louisiana State University, Batzer MA, Konkel MK, Walker JA, Massachusetts General Hospital, MacArthur DG, Lek M, Max Planck Institute for Molecular Genetics, Sudbrak R, Amstislavskiy VS, Herwig R, McDonnell Genome Institute at Washington University, Mardis ER, Ding L, Koboldt DC, Larson D, Ye K, McGill University, Gravel S, National Eye Institute, NIH, Swaroop A, Chew E, New York Genome Center, Lappalainen T, Erlich Y, Gymrek M, Frederick Willems T, Ontario Institute for Cancer Research, Simpson JT, Pennsylvania State University, Shriver MD, Rutgers Cancer Institute of New Jersey, Rosenfeld JA, Stanford University, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Tel-Aviv University, Halperin E, Baran Y, The Jackson Laboratory for Genomic Medicine, Lee C, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Thermo Fisher Scientific, Hyland FCL, Translational Genomics Research Institute, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, US National Institutes of Health, Sherry ST, Xiao C, University of California, San Diego, Sebat J,

Antaki D, Gujral M, Noor A, Ye K, University of California, San Francisco, Burchard EG, Hernandez RD, Gignoux CR, University of California, Santa Cruz, Haussler D, Katzman SJ, James Kent W, University of Chicago, Howie B, University College London, Ruiz-Linares A, University of Geneva, Dermitzakis ET, University of Maryland School of Medicine, Devine SE, University of Michigan, Abecasis GR, Min Kang H, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Kate Wing M, Zhan X, University of Montréal, Awadalla P, Hodgkinson A, University of North Carolina at Chapel Hill, Li Y, University of North Carolina at Charlotte, Shi X, Quitadamo A, University of Oxford, Lunter G, McVean GA, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, University of Puerto Rico, Oleksyk TK, University of Texas Health Sciences Center at Houston, Fu Y, Liu X, Xiong M, University of Utah, Jorde L, Witherspoon D, Xing J, University of Washington, Eichler EE, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Weill Cornell Medical College, Khurana E, Wellcome Trust Sanger Institute, Durbin RM, Hurles ME, Tyler-Smith C, Albers CA, Ayub Q, Balasubramaniam S, Chen Y, Colonna V, Danecek P, Jostins L, Keane TM, McCarthy S, Walter K, Xue Y, Yale University, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Y, Harmanci AO, Jin M, Lee D, Liu J, Jasmine Mu X, Zhang J, Zhang Y, Structural variation group, BGI-Shenzhen, Li Y, Luo R, Zhu H, Bilkent University, Alkan C, Dal E, Kahveci F, Boston College, Marth GT, Garrison EP, Kural D, Lee W-P, Ward AN, Wu J, Zhang M, Broad Institute of MIT and Harvard, McCarroll SA, Handsaker RE, Altshuler DM, Banks E, del Angel G, Genovese G, Hartl C, Li H, Kashin S, Nemesh JC, Shakir K, Cold Spring Harbor Laboratory, Yoon SC, Lihm J, Makarov V, Cornell University, Degenhardt J, European Molecular Biology Laboratory, Korbel JO, Fritz MH, Meiers S, Raeder B, Rausch T, Stütz AM, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Paolo Casale F, Clarke L, Smith RE, Stegle O, Zheng-Bradley X, Illumina, Bentley DR, Barnes B, Keira Cheetham R, Eberle M, Humphray S, Kahn S, Murray L, Shaw R, Leiden University Medical Center, Lameijer E-W, Louisiana State University, Batzer MA, Konkel MK, Walker JA, McDonnell Genome Institute at Washington University, Ding L, Hall I, Ye K, Stanford University, Lacroute P, The Jackson Laboratory for Genomic Medicine, Lee C, Cerveira E, Malhotra A, Hwang J, Plewczynski D, Radew K, Romanovitch M, Zhang C, Translational Genomics Research Institute, Craig DW, Homer N, US National Institutes of Health, Church D, Xiao C, University of California, San Diego, Sebat J, Antaki D, Bafna V, Michaelson J, Ye K, University of Maryland School of Medicine, Devine SE, Gardner EJ, University of Michigan, Abecasis GR, Kidd JM, Mills RE, Dayama G, Emery S, Jun G, University of North Carolina at Charlotte, Shi X, Quitadamo A, University of Oxford, Lunter G, McVean GA, University of Texas MD Anderson Cancer Center, Chen K, Fan X, Chong Z, Chen T, University of Utah, Witherspoon D, Xing J, University of Washington, Eichler EE, Chaisson MJ, Hormozdiari F, Huddleston J, Malig M, Nelson BJ, Sudmant PH, Vanderbilt University School of Medicine, Parrish NF, Weill Cornell Medical College, Khurana E, Wellcome Trust Sanger Institute, Hurles ME, Blackburne B, Lindsay SJ, Ning Z, Walter K, Zhang Y, Yale University, Gerstein MB, Abyzov A, Chen J, Clarke D, Lam H, Jasmine Mu X, Sisu C, Zhang J, Zhang Y, Exome group, Baylor College of Medicine, Gibbs RA, Yu F, Bainbridge M, Challis D, Evani US, Kovar C, Lu J, Muzny D, Nagaswamy U, Reid JG, Sabo A, Yu J, BGI-Shenzhen, Guo X, Li W, Li Y, Wu R, Boston College, Marth GT, Garrison EP, Fung Leong W, Ward AN, Broad Institute of MIT and Harvard, del Angel G, DePristo MA, Gabriel SB, Gupta N, Hartl C, Poplin RE, Cornell University, Clark AG, Rodriguez-Flores JL, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Clarke L, Smith RE, Zheng-Bradley X, Massachusetts General Hospital, MacArthur DG, McDonnell Genome Institute at Washington University, Mardis ER, Fulton R, Koboldt DC, McGill University, Gravel S, Stanford University, Bustamante CD, Translational Genomics Research Institute, Craig DW, Christoforides A, Homer N, Izatt T, US National Institutes of Health, Sherry ST, Xiao C, University of Geneva,

Dermitzakis ET, University of Michigan, Abecasis GR, Min Kang H, University of Oxford, McVean GA, Yale University, Gerstein MB, Balasubramanian S, Habegger L, Functional interpretation group, Cornell University, Yu H, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Clarke L, Cunningham F, Dunham I, Zerbino D, Zheng-Bradley X, Harvard University, Lage K, Berg Jespersen J, Horn H, Stanford University, Montgomery SB, DeGorter MK, Weill Cornell Medical College, Khurana E, Wellcome Trust Sanger Institute, Tyler-Smith C, Chen Y, Colonna V, Xue Y, Yale University, Gerstein MB, Balasubramanian S, Fu Y, Kim D, Chromosome Y group, Albert Einstein College of Medicine, Auton A, Marcketta A, American Museum of Natural History, Desalle R, Narechania A, Arizona State University, Wilson Sayres MA, Boston College, Garrison EP, Broad Institute of MIT and Harvard, Handsaker RE, Kashin S, McCarroll SA, Cornell University, Rodriguez-Flores JL, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Clarke L, Zheng-Bradley X, New York Genome Center, Erlich Y, Gymrek M, Frederick Willems T, Stanford University, Bustamante CD, Mendez FL, David Poznik G, Underhill PA, The Jackson Laboratory for Genomic Medicine, Lee C, Cerveira E, Malhotra A, Romanovitch M, Zhang C, University of Michigan, Abecasis GR, University of Queensland, Coin L, Shao H, Virginia Bioinformatics Institute, Mittelman D, Wellcome Trust Sanger Institute, Tyler-Smith C, Ayub Q, Banerjee R, Cerezo M, Chen Y, Fitzgerald TW, Louzada S, Massaia A, McCarthy S, Ritchie GR, Xue Y, Yang F, Data coordination center group, Baylor College of Medicine, Gibbs RA, Kovar C, Kalra D, Hale W, Muzny D, Reid JG, BGI-Shenzhen, Wang J, Dan X, Guo X, Li G, Li Y, Ye C, Zheng X, Broad Institute of MIT and Harvard, Altshuler DM, European Molecular Biology Laboratory, European Bioinformatics Institute, Flicek P, Clarke L, Zheng-Bradley X, Illumina, Bentley DR, Cox A, Humphray S, Kahn S, Max Planck Institute for Molecular Genetics, Sudbrak R, Albrecht MW, Lienhard M, McDonnell Genome Institute at Washington University, Larson D, Translational Genomics Research Institute, Craig DW, Izatt T, Kurdoglu AA, US National Institutes of Health, Sherry ST, Xiao C, University of California, Santa Cruz, Haussler D, University of Michigan, Abecasis GR, University of Oxford, McVean GA, Wellcome Trust Sanger Institute, Durbin RM, Balasubramaniam S, Keane TM, McCarthy S, Stalker J, Samples and ELSI group, Chakravarti A, Knoppers BM, Abecasis GR, Barnes KC, Beiswanger C, Burchard EG, Bustamante CD, Cai H, Cao H, Durbin RM, Gerry NP, Gharani N, Gibbs RA, Gignoux CR, Gravel S, Henn B, Jones D, Jorde L, Kaye JS, Keinan A, Kent A, Kerasidou A, Li Y, Mathias R, McVean GA, Moreno-Estrada A, Ossorio PN, Parker M, Resch AM, Rotimi CN, Royal CD, Sandoval K, Su Y, Sudbrak R, Tian Z, Tishkoff S, Toji LH, Tyler-Smith C, Via M, Wang Y, Yang H, Yang L, Zhu J, Sample collection, British from England and Scotland (GBR), Bodmer W, Colombians in Medellín, Colombia (CLM), Bedoya G, Ruiz-Linares A, Han Chinese South (CHS), Cai Z, Gao Y, Chu J, Finnish in Finland (FIN), Peltonen L, Iberian Populations in Spain (IBS), Garcia-Montero A, Orfao A, Puerto Ricans in Puerto Rico (PUR), Dutil J, Martinez-Cruzado JC, Oleksyk TK, African Caribbean in Barbados (ACB), Barnes KC, Mathias RA, Hennis A, Watson H, McKenzie C, Bengali in Bangladesh (BEB), Qadri F, LaRocque R, Sabeti PC, Chinese Dai in Xishuangbanna, China (CDX), Zhu J, Deng X, Esan in Nigeria (ESN), Sabeti PC, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Gambian in Western Division – Mandinka (GWD), Jallow M, Sisay Joof F, Corrah T, Rockett K, Kwiatkowski D, Indian Telugu in the UK (ITU) and Sri Lankan Tamil in the UK (STU), Kooner J, Kinh in Ho Chi Minh City, Vietnam (KHV), Tịnh Hiê`n T, Dunstan SJ, Thuy Hang N, Mende in Sierra Leone (MSL), Fonnie R, Garry R, Kanneh L, Moses L, Sabeti PC, Schieffelin J, Grant DS, Peruvian in Lima, Peru (PEL), Gallo C, Poletti G, Punjabi in Lahore, Pakistan (PJL), Saleheen D, Rasheed A, Scientific management, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Green ED, Duncanson A, Dunn M, Schloss JA, Wang J, Yang H, Writing group, Auton A, Brooks LD, Durbin RM, Garrison EP, Min Kang H, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68–74.

19. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11:e0163962.
20. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.
21. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci 3:e104.
22. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857.
23. Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770.
24. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C. 2021. Multivariable association discovery in population-scale meta-omics studies. PLOS Comput Biol 17:e1009442.
25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
26. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
27. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.