GuaCAMOLE: GC-bias aware estimation improves the accuracy of metagenomic species abundances

Laurenz Holcik[1,2], Arndt von Haseseler[1,3], Florian G. Pflug[1,4*]

[1]Center for Integrative Bioinformatics Vienna (CIBIV), University of Vienna, Dr. Bohr Gasse 9, 1030 Vienna, Austria.
[2]Vienna BioCenter PhD Program, a Doctoral School of the University of Vienna and Medical University of Vienna, Dr. Bohr Gasse 9, 1030 Vienna, Austria.
[3]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, 1090 Vienna, Austria.
[4]Biological Complexity Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna, 904-0495 Okinawa, Japan.

*Corresponding author(s). E-mail(s): florian.pflug@univie.ac.at;

Abstract

GuaCAMOLE is a novel computational method which detects and removes GC bias from metagenomic sequencing data. Metagenomic sequencing measures the species composition of microbial communities, and has revealed the crucial role of microbiomes in the etiology of a range of diseases such as colorectal cancer. Quantitative comparisons of microbial communities are, however, affected by GC-content dependent biases. GuaCAMOLE works regardless of the specific amount or direction of GC-bias present in the data and requires only a single sample. The algorithm reports unbiased abundances and quantifies the amount of bias present in terms of GC-depdendent sequencing efficiencies. Experimental mock community data confirms both estimates to be accurate across a wide range of experimental protocols. In gut microbiomes of colorectal cancer patients we observe a clear bias against GC-poor species in the abundances reported by existing methods. GuaCAMOLE successfully removes this bias and corrects the abundance of clinically relevant GC-poor species such as *F. nucleatum* (28% GC) by up to a factor of two. GuaCAMOLE thus contributes to a better quantitative understanding of microbial communities by improving the accuracy and comparability of species abundances across experimental setups.

1 Background

Metagenomic sequencing has enabled the comprehensive and quantitative analysis of taxa abundances in a wide range of microbial communities (Morgan et al., 2010). It has uncovered the importance of microbiomes in health, disease, nutrition, ecology amongst others, and has revealed a complex interplay between microbial consortia and their hosts (Sunagawa et al., 2015; Integrative et al., 2019; Trivedi et al., 2020).

Metagenomic sequencing relies on comprehensive high-throughput sequencing of all DNA in a sample to quantify the abundance of all present taxa. To prepare a sample for sequencing, the DNA is extracted, purified, fragmented, amplified, and finally outfitted with sequencing adapters. Numerous protocols have been established for these library preparation steps, each differing in methods and materials used (Sato et al., 2019; Bowers et al., 2015; Jones et al., 2015). After sequencing, the reads are assigned to taxa and relative read counts are used as proxies for the taxa's abundances (Wood et al., 2019; Blanco-Míguez et al., 2023).

While metagenomic sequencing is in principle agnostic to the specific taxa in a sample, library preparation can introduce sequence-dependent biases. (Dohm et al., 2008). In particular, the *GC content* (i.e. fraction of G and C bases in the sequence) has been shown to strongly affect sequencing efficiency (Benjamini and Speed, 2012). Metagenomic sequencing is particularly affected because the genomic GC content often differs significantly between species (Sato et al., 2019). The magnitude and even direction of this bias, however, varies between different library preparation and sequencing protocols (Tourlousse et al., 2021). For example, a low GC content can either increase or decrease sequencing efficiency depending on the precise protocol used. As a result, computational correction for GC bias has been challenging (Browne et al., 2020).

The species on the extreme ends of the genomic GC content range are particularly prone to biases. Amongst these species we find pathogenic taxa such as *F. nucleatum* (28% GC content, associated with colorectal cancer) and *M. pneumoniae* (25% GC content, associated with pneumonia) (Yu et al., 2017; Han, 2015; Yang et al., 2017). With common sequencing protocols, the abundance of these taxa will be underestimated (Browne et al., 2020; Sato et al., 2019), and this can affect even comparisons between samples analyzed using the same protocol (McLaren et al., 2022).

Ideally, GC bias should therefore be removed on a per-sample level. Computational methods to remove GC bias have been developed for various sequencing-based methods, and have been shown to be crucial to avoid skewed results (Benjamini and Speed, 2012; Love et al., 2016). These methods, however, assume reads are aligned to a reference genome. For metagenomic samples possibly containing thousands of taxa, creating such an alignment is prohibitively computationally expensive. Instead, metagenomic reads are typically assigned to taxa using k-mer based methods (Wood and Salzberg, 2014; Dilthey et al., 2019). This makes existing methods inapplicable and requires a novel approach to GC bias correction.

2 Introduction

We present the GuaCAMOLE (Guanosine Cytosine Aware Metagenomic Opulence Least squares Estimation) algorithm for the efficient detection and removal of GC bias

from metagenomic samples. GuaCAMOLE is an alignment-free algorithms and instead assigns reads to taxa using Kraken2 (Wood and Salzberg, 2014). The algorithms also does not require calibration data or any a-priori assumptions about the quantitative relationship between GC content and bias (such as extremely GC-rich and GC-poor species being more prone to bias), and thus works equally well for all sequencing protocols.

Using both simulations and experimental data for mock communities (Tourlousse et al., 2021) we show that GuaCAMOLE uncovers GC bias present in samples and improves abundance estimates over Bracken (Lu et al., 2017) and MetaPhlAn4(Blanco-Míguez et al., 2023). We also demonstrate that GuaCAMOLE correctly infers protocol-specific GC-dependent sequencing efficiencies.

To show that GC bias correction can be relevant in a clinical setting, we apply GuaCAMOLE to meta-genomic stool samples of colorectal cancer patients (Yachida et al., 2019). Here, we observe that accounting for GC content significantly increases the estimated abundances of clinically relevant taxa on the low end of the GC spectrum such as *F. nucleatum*.

3 Results

The GuaCAMOLE algorithm processes the raw sequencing reads of a metagenomic sample and outputs bias-correct abundances for all detected taxa. GuaCAMOLE also infers and outputs GC-dependent sequencing efficiencies which reflect the probability (relative to the maximum) that a DNA fragment with a certain GC content successfully undergoes all library preparation steps and sequencing. These GC-dependent sequencing efficiencies thus measure the extent of the GC bias present in the raw data. Briefly, GuaCAMOLE works as follows (Fig. 1, see Methods for details): Reads are first assigned to individual taxa using Kraken2 (Wood et al., 2019), and within each taxa to discrete bins representing the read's GC content. Reads which cannot be assigned to a specific taxon unambiguously by Kraken2 are redistributed probabilistically to the likeliest taxon using the Bracken algorithm (Lu et al., 2017). Read counts in each taxon-GC-bin are then normalized based on expected read counts computed from the genome lengths and genomic GC content distributions of individual taxa. The resulting quotients only depend on the unknown abundances (one for each taxon) and unknown GC-dependent sequencing efficiency (one per GC-bin). From these quotients we then compute bias-corrected abundance estimates and the GC-dependent sequencing efficiencies. GuaCAMOLE reports the estimated abundances either as *sequence abundances* proportional to the total amount of DNA present, or *taxonomic abundance* proportional to the number of genomes (Sun et al., 2021).

3.1 GC bias is removed independent of its type

We first demonstrate that GuaCAMOLE infers the correct abundances and GC-dependent sequencing efficiencies independent of the specific type of GC bias present. We ran GuaCAMOLE on data simulated using three different models of GC bias (see Methods for details): peak efficiency at 50% GC, efficiency increasing with GC content, and efficiency decreasing with GC content (Fig. 2). For all three simulated

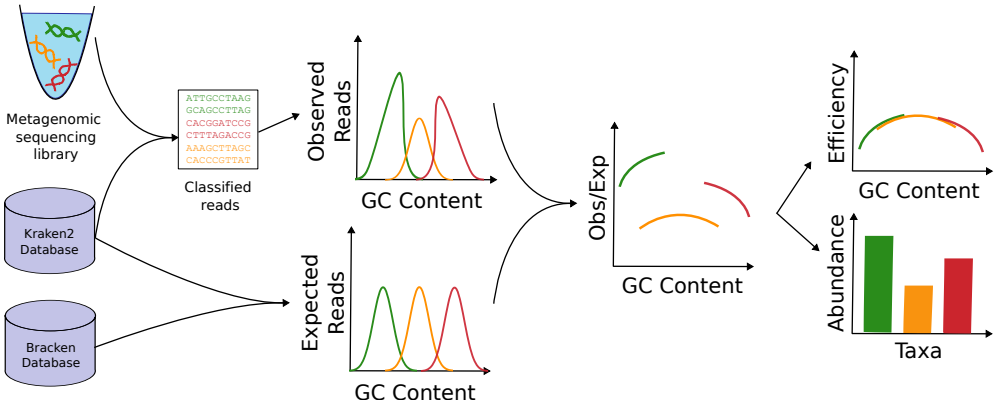


Fig. 1 The GuaCAMOLE algorithm. Reads are assigned to taxa using Kraken2/Bracken (Wood and Salzberg, 2014; Lu et al., 2017) and binned into discrete GC bins per taxon. Corresponding expected read counts are then computed for each taxon and GC bin from the reference genomes. The observed/expected quotients reflect the GC-dependent sequencing efficiencies scaled by each taxon's abundance. Abundances are estimated by finding the scaling factors for which the quotients form a continuous curve.

datasets, GuaCAMOLE produced virtually unbiased estimates (mean relative error less than 1%) and correctly recovered the GC-dependent sequencing efficiencies used for the simulation. The Bracken estimates showed considerable GC bias in comparison (relative errors 10% to 30% depending on the bias model).

3.2 Improved accuracy across a range of experimental protocols

We next demonstrate that GuaCAMOLE improves abundances estimates for experimental data produced using different protocols and can uncover the GC-dependent sequencing efficiencies of these protocols. We re-analzyed published data (Tourlousse et al., 2021) of a mock community sequenced using 28 different protocols (Table 1) with GuaCAMOLE, Bracken (Lu et al., 2017) and MetaPhlAn4 (Blanco-Míguez et al., 2023). The mock community comprises 19 bacterial species representative of human-associated microbiota and was sequenced using 11 different commercially available library preparation kits (labelled A-K below, see Table. 1). For each kit, Tourlousse et al. tested up to three PCR amplification regimes: 500 ng input DNA with no PCR amplification (suffix 0), 50 ng input DNA with 4-8 PCR cycles (suffix L), and 1 ng input DNA with 8-15 PCR cycles (suffix H).

We find that the GC-dependent sequencing efficiencies estimated by GuaCAMOLE differ strongly between different protocols (Fig. 3A). Some protocols show uniform efficiencies, while others show a strong dependence on the GC content. In accordance with the results of Tourlousse et al. (Tourlousse et al., 2021) we see that the protocols GH, DH, IL, IH and FH show the strongest dependency on GC content (Fig. 3A, colored lines). The nature of this dependence differs qualitatively between protocols. While protocols IL and IH show decreasing sequencing efficiency with increasing GC content, GH, DH and FH show an increase in efficiency for higher GC content.

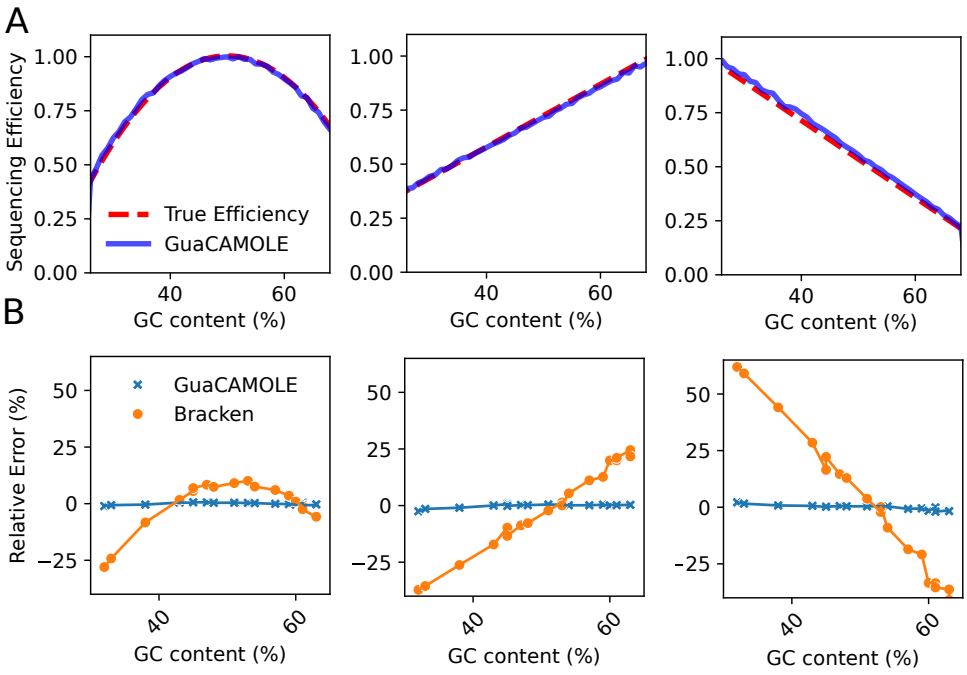


Fig. 2 Inference of unbiased abundances and GC-dependent sequencing efficiencies. Metagenomic sequencing reads for a community comprising 19 species (one represented by two strains) were simulated *in silico* (fragment length 400bp, read length 150bp) using three different models of GC bias (see Methods). **(A)** True GC-dependent sequencing efficiencies (green) and efficiencies estimated by GuaCAMOLE (blue). **(B)** Relative error $(a_j - A_j)/A_j$ of estimated abundances a_j vs. true abundances A_j for GuaCAMOLE (blue) and Bracken (yellow).

kit name	frag. method	PCR cycles (0/L/H)	abbr.
Accel NGS 2S Plus DNA Library Kit	physical	0 / 4 / 9	A
QIAseq FX DNA Library Kit	enzymatic	0 / 8 / 12	B
TruSeq (Nano) DNA (PCR-Free) Library Prep Kit	physical	0 / 8 / 8	C
KAPA HTP Library Preparation Kit	physical	0 / 5 / 15	D
KAPA HyperPrep (PCR-free) Kit	physical	0 / 4 / 14	E
KAPA HyperPrep + KAPA Frag	enzymatic	0 / 4 / 14	F
KAPA HTP + KAPA Frag	enzymatic	- / 5 / 15	G
NEBNext Ultra II DNA Library Prep Kit	physical	- / 4 / 9	H
NEBNext Ultra II FS DNA Library Prep Kit	enzymatic	- / 4 / 9	I
Nextera DNA Flex Library Prep Kit	enzymatic	- / 4 / 12	J
SMARTer ThruPLEX DNA-Seq Kit	physical	- / 6 / 11	K

Table 1 DNA Library Preparation Kits used for the DNA mock community (Tourlousse et al., 2021)

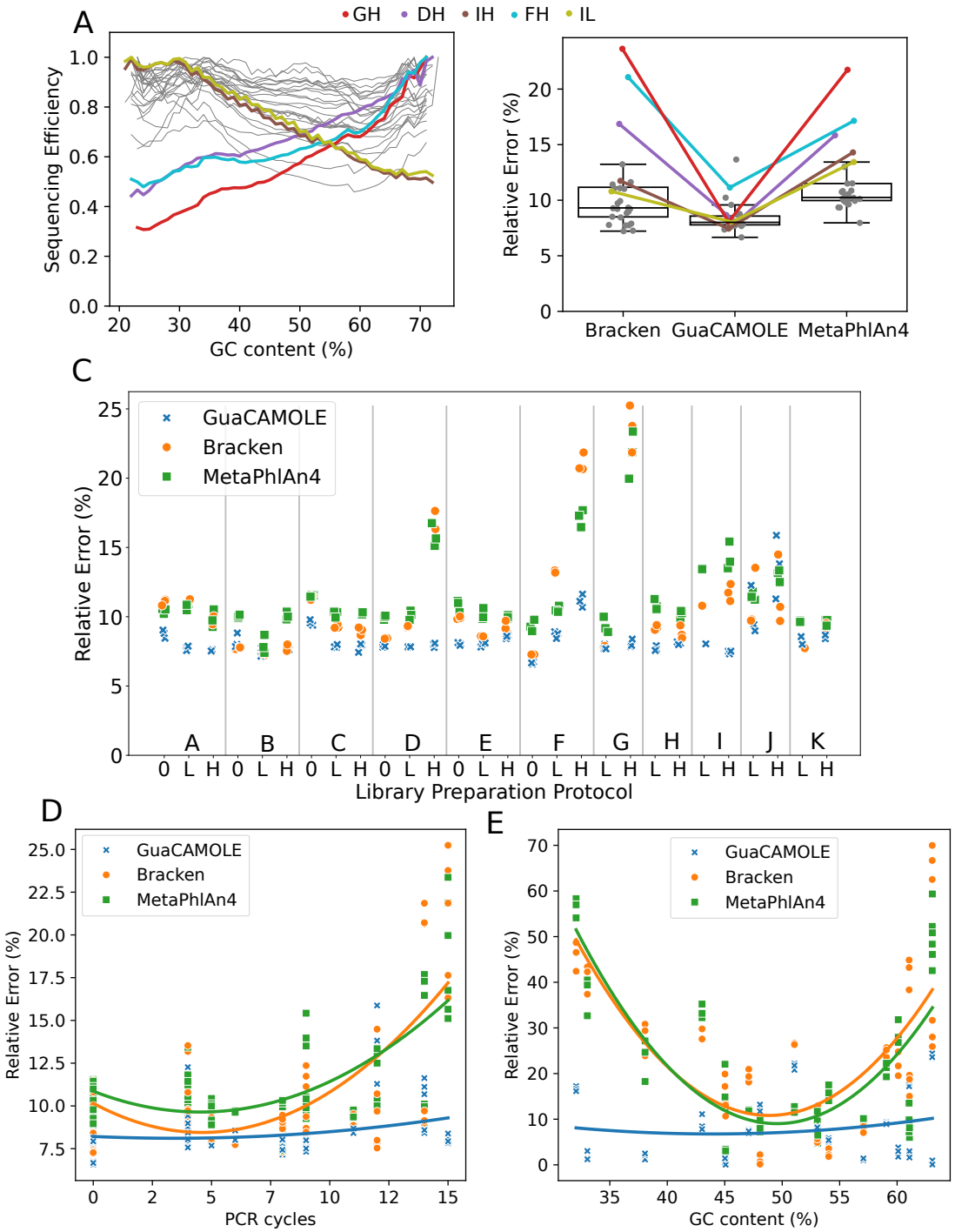


Fig. 3 Performance of GuaCAMOLE for experimental mock community data. The mock community (Tourlousse et al., 2021) contains 19 species and was sequenced in triplicates using 28 different library preparation protocols (Table 1). We re-analzyed the reads with GuaCAMOLE, Bracken and MetaPhlAn4. GuaCAMOLE was set to report taxonomic abundances, Bracken results were manually adjusted for genome length. Relative errors are $|a_j - A_j|/A_j$ with A_j the truth and a_j the estimate. **(A)** Estimated GC-dependent sequencing efficiencies of the 28 protocols by GuaCAMOLE. Highlighted protocols GH, DH, IH, FH, IL were found by Tourlousse et al. to exhibit the strongest dependency of efficiency on GC content. **(B)** Relative abundance estimation errors per protocol for GuaCAMOle, Bracken and MetaPhlAn4 (averaged over all 19 species). **(C)** Relative abundance estimation error of GuaCAMOlE, Bracken and MetaPhlAn4 for the three replicates of each protocol. **(D)** Relative abundance estimation error vs. number of PCR cycles used in each protocol. Lines show quadratic best fit. **(E)** Relative abundance estimation error of each taxon averaged across protocols vs. genomic GC content. Lines show quadratic best fit.

For the protocols most strongly affected by GC content (GH, DH, IH, FH, IL) GuaCAMOLE reduces the mean relative abundance error drastically compared to Bracken and MetaPhlAn4 (Fig. 3B, colored lines). For other protocols GuaCAMOLE overall also shows the smallest and most consistent mean error but the effect is less pronounced (Fig. 3, boxplots). Looking at individual protocols, GuaCAMOLE is amongst the algorithms with smallest mean error in every case except protocol JH (Fig. 3C). A common source of GC bias is PCR amplification (Aird et al., 2011), and accordingly the advantage of GuaCAMOLE over Bracken and MetaPhlAn4 increases with the number of PCR cycles (Fig. 3D). However, GuaCAMOLE also offers a clear advantage over the other algorithms for PCR-free protocols A0, C0, E0, H0 (Fig. 3C).

The quantification error per bacterial species shows for GuaCAMOLE only a weak residual dependence on genomic GC content (Fig. 3E). In comparison, the quantification error of both Braken and MetaPhlAn4 increase significantly for taxa on the extremal ends of the GC content range.

3.3 Correct abundance estimation of GC-poor species

In the microbiomes of 80 colorectal cancer patients published by Yachida et al. (Yachida et al., 2019) we observe the abundances estimated by Bracken to gradually decline starting at a genomic GC content of about 40% and to drop to half the maximum for a GC content of about 30% (Fig. 4A). This suggests the presence of considerable bias against GC-poor species in the data, which affects about a third of species (below 40% GC) to some degree, and 12% of species (below 30% GC) strongly (Fig. 4B).

The GC-dependent sequencing efficiencies estimated by GuaCAMOLE confirm this bias and indeed estimates the sequencing efficiency to drop to around 50% at a genomic GC content about 30% (Fig. 4C). The estimates efficiencies also show some bias against GC-rich species, although less pronounced. These findings are consistent with previous reports about the Nextera XT kit used by Yachida et al. (Browne et al., 2020; Sato et al., 2019). The bias-corrected abundance reported by GuaCAMOLE are much more uniform across different genomic GC contents, indicating little or no residual GC bias (Fig. 4A). For GC-poor species, the abundances reported by GuaCAMOLE are up to two-fold higher than those reported by Bracken (Fig. 4D).

Amongst the GC-poor species whose abundance is severely underestimated by Bracken, we find *F. nucleatum* which has been associated with colorectal cancer (Yu et al., 2017; Yang et al., 2017; Han, 2015). *F. nucleatum* was observed in 7 out of 80 samples, and its abundance was consistently underestimated by Bracken (on average 1.6-fold).

3.4 Outlier removal

Due to sequence similarity between genomes, a certain fraction of reads typically gets assigned to taxa not actually present in the sample. Filtering detected taxa based on read counts is a common way to remove some of these false-positives, but this is not always effective. GuaCAMOLE additionally filters taxa based on the relative deviations of observed from expected read counts (see Methods). Briefly, if these *residuals*

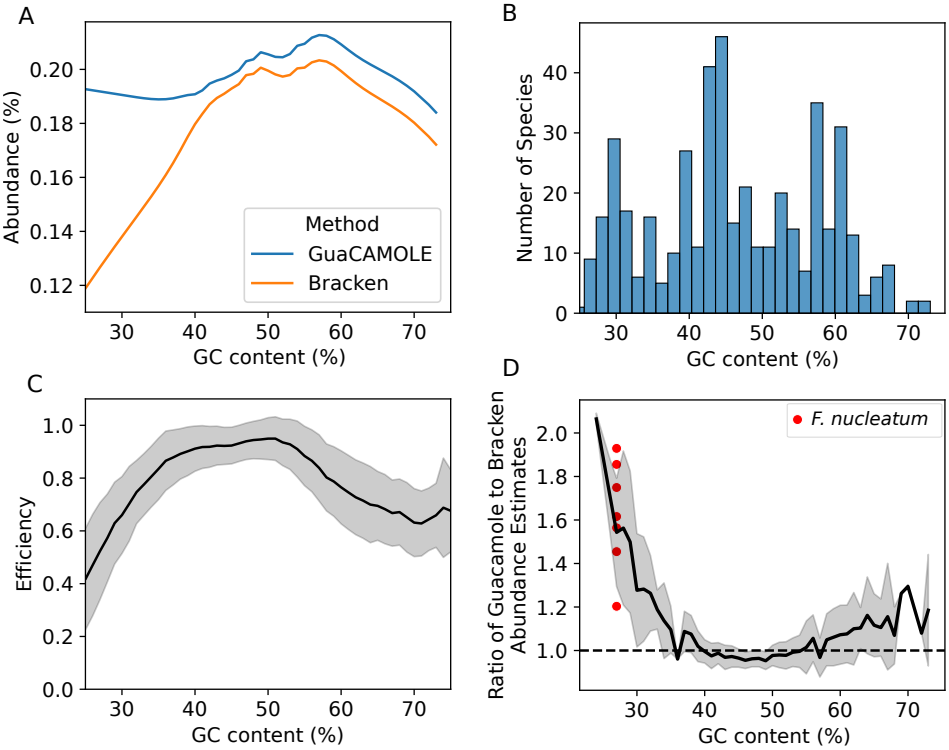


Fig. 4 Performance of GuaCAMOLE for human gut microbiomes. The data published by Yachida et al. (Yachida et al., 2019) comprising gut microbiomes of 80 colorectal cancer (CRC) patients was re-analyzed with GuaCAMOLE (in sequence abundance mode) and Bracken. GuaCAMOLE detected 447 distinct taxa, taxa flagged as outliers by GuaCAMOLE were removed from the Bracken output as well. **(A)** Estimated abundances vs. genomic GC content for GuaCAMOLE and Bracken (smoothed with Loess fraction 0.7). **(B)** Distribution of genomic GC content amgonst 447 detected taxa. **(C)** Mean and std. dev. of GC-dependent sequencing efficiencies across all 80 samples. **(D)** Ratio of GuaCAMOLE and Bracken abundances estimates vs. species GC content. Shaded area shows st. dev., red dots show *F. nucleatum* ratio in each sample.

vary more across the GC bins of a taxon than a pre-defined threshold, the taxon is flagged as an outlier and removed. After removing a taxon, all abundances are recomputed and another round of outlier detection is initiated. This process is designed to remove false-positive taxa and to prevent outliers from skewing the efficiency and abundance estimates (Figs. A2 and A3 in Appendix A Supplemental Information).

For the mock community data of Tourlousse et al., GuaCAMOLE's outlier removal reduces the number of false-positive taxa from 18 ± 9 to 8 ± 6 on average. The removed taxa together account for about 0.4% of all reads. Since the deviation between observed and expected read counts differs considerably between taxa and protocols (Fig. A1 in Appendix A Supplemental Information), the appropriate number of outlier removal rounds differs as well. For most protocols 4 rounds are appropriate and yield the most accurate efficiency estimates (Fig. A2 in Appendix A Supplemental Information). The

exceptions are protocols JH and B0 where 4 rounds occasionally remove some taxa known to be present; for these protocols we have thus restricted GuaCAMOLE to 3 rounds of outlier removal.

In the gut microbiome data of Yachida et al. 4 rounds of outlier removal reduce the number of taxa from 231 ± 61 to 56 ± 21 which account for $76\% \pm 15\%$ of the sequencing reads. After removing these taxa, the GC-dependent sequencing efficiencies agree much more closely between taxa (Fig. A3 in Appendix A Supplemental Information). It is possible, however, that some of the outliers we remove are not false-positives, but rather represent taxa present in the sample whose reference genome is inaccurate.

4 Discussion

GuaCAMOLE infers both bias-corrected abundances and GC-dependent sequencing efficiencies from a single sample without prior information about the amount or direction of GC-bias present in the data. The algorithm is agnostic about the specific sequencing protocol used and can correctly detect and correct for GC-bias without calibration data or prior knowledge about the expected type of bias. For most sequencing protocols, the bias-corrected abundances reported by GuaCAMOLE are more accurate than those reported by both Bracken and MetaPhlAn4. The advantage provided by GuaCAMOLE increases with the amount of GC bias present, and thus in particular with the amount of PCR amplification done prior to sequencing. Interestingly, we don't observe an advantage of MetaPhlAn4 over Bracken even though we might expect the marker gene based approach of MetaPhlAn4 to be less susceptible to bias. In fact, Bracken and MetaPhlan4 often show a relatively similar quantification error. This further corroborates that the improvement offered by GuaCAMOLE does indeed stem from successful correction of GC bias and not from other algorithmic differences.

In addition to bias-corrected abundances, GuaCAMOLE reports accurate GC-dependent sequencing efficiencies. This is useful as a quality control to check that library preparation and sequencing perform as expected. It also provides a way to estimate the amount of bias that affects taxa which remained unobserved. Finally, it allows different library preparation and sequencing protocols to be compared without the need for mock communities with known abundances.

GC bias can affect the abundance estimates of clinically relevant pathogens such as *F. nulceatum* which has been associated with a range of diseases (Han, 2015; Yu et al., 2017; Yang et al., 2017). We observe this in published microbiomes of colorectal cancer patients (Yachida et al., 2019) where the abundance of *F. nulceatum* is underestimated about 1.6 fold before GC bias correction. Since the bias in this study is caused by the widely used Nextera XT library preparation kit (Browne et al., 2020; Sato et al., 2019) other studies are likely affected by a similar bias.

GuaCAMOLE detecs false-positive taxa by checking for outliers within the deviations of observed from expected read counts. This offers more power than read-count thresholding, and ensures that such outliers do the skew the estimated sequencing efficiencies and abundances. However, this outlier detection assumes reasonably accurate reference genomes. Therefore, taxa which are present but whose reference genomes are inaccurate are at risk of being flagged an outlier and removed. If this is a concern, the

GC-dependent efficiencies reported by GuaCAMOLE can be used to correct the bias present in the abundances estimated by other tools such as Bracken or MetaPhlAn4. This effectively "borrows" information about GC bias between species with similar GC content. For Bracken, GuaCAMOLE already implements this mode of operation as an option.

5 Conclusion

GuaCAMOLE provivdes a computational method to correct for GC bias in sequencing protocols. For a wide range of sequencing protocols, GuaCAMOLE substationally improves abundance estimates over alternative methods. Species whose abundance estimates are improved include clinically relevant taxa. GuaCAMOLE is in principle applicable to all types of meta-genomic samples, but performs best when reasonably complete reference genomes are available for all species. To facilitate its use by the community and its integration into standard pipelines, GuaCAMOLE is available as an easy-to-use and fast Python package under https://github.com/Cibiv/GuaCAMOLE.

6 Methods

6.1 The GuaCAMOLE algorithm

GuaCAMOLE operates on a pre-defined taxonomy comprising nodes $K_1, \ldots, K_n$ which are arranged in a tree. We often refer to these nodes simply as *taxa*. Leaf nodes represent individual species (or strains) whereas internal nodes represent higher taxonomic groups such as genera, families etc. Leaf nodes always have an associated genome, for internal nodes this is optional. The taxonomy together with all associated genomes is refered to as a *database*.

GuaCAMOLE estimates the abundances of these taxa from a meta-genomic sequencing library containing a number N of sequencing reads. Each read is assumed to stem from one of the taxa in the taxonomy. The GC content of a read is the fraction of bases which are either guanin (G) or cytosin (C). We assign reads into one of b equally spaced bins according to their GC content, and denote the GC content by the index g of the respective bin.

GuaCAMOLE assumes that the composition of the sequencing library depends on (i) the abundances $a_1, \ldots, a_n$ of all taxa (zero for all taxa not present in the sample), (ii) the GC-dependent sequencing efficiency η_g, (iii) the genomic GC content distributions $f(j, g)$ of the taxa (normalized such that $\sum_g f(j, g) = 1$), and (iv) the lengths $l_1, \ldots, l_n$ of the taxa's genomes. In terms of these quantities, GuaCAMOLE assumed that the number $O(j, g)$ of fragments stemming from taxon j with GC content g in the library is

$$O(j, g) = N \cdot \frac{a_j \cdot \eta_g \cdot l_j \cdot f(j, g)}{\sum_{g=1}^{b} \sum_{i=1}^{n} a_i \cdot \eta_g \cdot l_i \cdot f(i, g)}. \tag{1}$$

Note that abundances a_i and efficiencies η_g are defined only up to a factor by Eq. (1), we normalize these quantities by demanding that $\sum_j a_j = \max_g \eta_g = 1$.

GuaCAMOLE estimates abundances and GC-dependent efficiencies by plugging observed fragment counts $O(j, g)$, GC distributions $f(j, g)$ and genome lengths l_g into Eq. (1) and solving for $a_1, \ldots, a_n$ and $\eta_1, \ldots, \eta_b$. Note that this system is typically over-determined: it contains on the order of nb equations for $n + b$ unknowns.

6.2 The number of reads per taxon and GC bin

To compute observed read counts $O(j, g)$, reads are first assigned to taxonomic nodes using Kraken2 (Wood et al., 2019). This yields the number of assigned reads $M(j)$ for every node j in the taxonomy. The reads assigned to each node are then further subdivided according to their GC content to obtain $M(j, g)$, the number of reads assigned to taxon j with GC content g.

The counts $M(j, g)$, however, are biased by identical regions in the genomes of different taxa. When Kraken2 is unable to unambiguously assigned a read to a taxon due identical sequences within multiple genomes, Kraken2 assigns those reads to the lowest common ancestor (LCA) of all matching taxa. To correct for this systematic bias, we use the same approach as the Bracken algorithm introduced by Lu et al. (Lu et al., 2017). Bracken computes the conditional probabilities $P(r \in G_j | K_i)$ that a read which was assigned to taxon i actually stems from a descendant j of i, and redistributes reads assigned to i accordingly. GuaCAMOLE does the same, but keeps track of the GC content when redistributing reads. To estimate the number of reads in GC bin g that stem from taxon j, we thus compute

$$\tilde{O}(j, g) = \sum_{i=1}^{n} M(i, g) \cdot P(r \in G_j | K_i). \tag{2}$$

Note that we have assumed here for simplicity that $P(r \in G_j | K_i)$ does not depend on the read's GC content.

To make it possible to compare abundances of higher taxonomic levels such as genera, we then sum the corrected read counts over descendants. The per-taxon and per-GC-bin read counts plugged into Eq. (1) are thus

$$O(j, g) = \tilde{O}(j, g) + \sum_{i \in D_j} \tilde{O}(i, g) \tag{3}$$

where D_j denotes the descendants of node j.

6.3 The genomic GC content distributions

For Eq. (1) to hold, the observed counts $O(j, g)$ must, in theory, arise by sampling taken from the the genomic GC content distributions $f(j, g)$. These distribution must thus take the redistribution of reads into account. To find $f(j, g)$, we first compute the individual GC distributions $q(j, g)$ of the genomes in the taxonomy. Given the fragment length ℓ_f and read length ℓ_r of the experimental data, $q(j, g)$ reflects the fraction of windows of length ℓ_f whose GC content within the parts covered by reads (i.e. the first and last ℓ_r bases for paired-end reads) is g. Here, we use the correct

experimental fragment- and read length to avoid systematic errors. We then find the expected GC content distribution of the reads assigned to a specific taxon

$$h(j,g) = \sum_{i=1}^{n} q(i,g) \cdot P(r \in G_i | K_j). \tag{4}$$

Here, we have taken into account that Kraken2 will assign some reads to taxa at higher taxonomic levels. To find the expected GC distribution after fragment redistribution, we mimick Eq. (2) and compute

$$\tilde{f}(j,g) = \frac{\sum_{i=1}^{n} h(i,g) \cdot M(i) \cdot P(r \in G_i | K_j)}{\sum_{i=1}^{n} M(i) \cdot P(r \in G_i | K_j)}. \tag{5}$$

Finally, we proceed similarly to Eq. (3) and average the GC distributions of all descendants, weighted by their fragment counts,

$$f(j,g) = \frac{\tilde{f}(j,g) + \sum_{i \in D_j} \tilde{f}(i,g) \cdot \sum_{\gamma} \tilde{O}(i,g)}{\sum_{\gamma=1}^{b} \left(\tilde{f}(j,\gamma) + \sum_{i \in D_j} \tilde{f}(i,\gamma) \cdot \sum_{\gamma} \tilde{O}(i,\gamma) \right)}. \tag{6}$$

Since the tails of these distributions are typically noisy, we restrict these distributions to the range between the 2.5% and 97.5% quantile for every taxon j.

6.4 Genome lengths

We assign a single genome length l_j to every taxon j independent of its taxonomic level or number of associated genomes. To do so, we average over the lengths of all assigned genomes of the taxon and all of its descendants. To account for the observed read distribution, we weight each genome length with the prior probability $P(K_j)$ that a random read stems from taxon j as computed by Bracken (Lu et al., 2017).

6.5 Estimating abundances

To estimate abundances $a_1, \ldots, a_n$, we re-arrange Eq. (1) into the following expression for the GC-dependent efficiencies η_g in bin g,

$$\eta_g = \frac{C}{N} \cdot \underbrace{\frac{O(j,g)}{l_j f(j,g)}}_{\text{Obs/Exp reads}} \cdot \frac{1}{a_j}. \tag{7}$$

where $C = \sum_{g=1}^{b} \sum_{i=1}^{h} a_i \cdot \eta_g \cdot l_i \cdot f(i,g)$ is a normalization factor. Note the correspondence to Fig. 1: For a fixed taxon j, η_g is proportional to the obs/exp ratio $O(j,g)/l_j f(j,g)$. After scaling with inverse abundances a_j^{-1} these ratios become comparable across taxa.

Given abundances $a_1, \ldots, a_n$, Eq. (7) provides a separate estimate of η_g for every taxon whose genomic GC distribution overlaps g. This allows us to estimate the abundances by maximising the agreement between these separate estimates of η_g. In terms of the inverse abundances $a_1^{-1}, \ldots, a_n^{-1}$ this can be expressed as minimization of the quadratic form

$$G(a_1^{-1}, \ldots, a_n^{-1}) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{g=1}^{b} \left(\frac{O(i,g)}{l_i f(i,g)} \cdot a_i^{-1} - \frac{O(j,g)}{l_j f(j,g)} \cdot a_j^{-1} \right)^2 . \tag{8}$$

Here, we have dropped the pre-factor C/N from the GC-dependent efficiencies η_g. In practice, we drop all terms from the sum in Eq. (8) that are either undefined or unreliable. A term is undefined if one of the two GC distribution does not overlap bin g (i.e. $f(i,g)$ or $f(j,g)$ is undefined). A term is considered to be unreliable if the total number of reads assigned to one of the taxa including descendants (i.e. $\sum_{D_j} M(j)$ for taxon j, similarly for i) lies below some user-defined threshold (minimum read threshold, default 500).

Regularisation

If the taxa in a sample can be partitioned into two sets A and B such that Eq. (8) contains no term containing an abundance from A and a abundance from B, the relative abundances between sets A and B are undefined. In Fig. 1 this would be represented as two groups of taxa whose GC efficiency curves do not mutually overlap. In this situation, Eq. (8) as stated does not have a unique minimum. To avoid this, we regularize the quadratic G by penalizing large differences in efficiency between neighbouring GC bins. Using Eq. (7) we express η_g sans the prefactor C/N as a weighted average of taxon-specified efficiencies,

$$\lambda_g = \frac{1}{n_g} \sum_{i=1}^{n} \log\big(O(i,g) + 1\big) \cdot \frac{O(i,g)}{l_i f(i,g)} \cdot a_i^{-1}. \tag{9}$$

We now define the regularized objective function

$$\tilde{G}(a_1^{-1}, \ldots, a_n^{-1}) = \frac{1-r}{n^2} G(a_1^{-1}, \ldots, a_n^{-1}) + r \sum_{k=1}^{b} \sum_{l=1}^{b} e^{-|k-l|} (\lambda_k - \lambda_l)^2 \tag{10}$$

which is still quadratic since λ_g is linear in $a_1^{-1}, \ldots, a_n^{-1}$. The regularized program is thus still efficiently solvable. Here, r is a hyperparameter that controls the amount of regularization to apply. Smaller values of r allow more extreme and small-scale variations in sequencing efficiency to be corrected, but increase the risk of incorrect estimates in the case of taxa partitions with non-overlapping GC distributions.

To find abundances $a_1, \ldots, a_n$ we first minimizing the regularised objective function $\tilde{G}$ in terms of $a_1^{-1}, \ldots, a_n^{-1}$ subject to $\sum_i a_1^{-1} = 1$ using the Python package *cvxopt*.

We then compute $a_1, \ldots, a_n$, normalized such that $\sum_j a_j = 1$, and compute the GC-dependent sequencing efficiencies $\eta_g = \lambda_g / \max_\gamma \eta_\gamma$.

6.6 Outlier detection and removal

The set of taxa with a non-zero number $O(j, g)$ of assigned reads often contains taxa which are not actually present in the sample. The reads witnessing such a false-positive taxon are consequently not uniformly random draws from the taxon's genome, and we hence expect to see some deviation from Eq. (1). To detect false-positives we therefore look for outliers amongst the relative residuals $\xi(j, g) = \big(O(j, g) - \bar{O}(j, g)\big)/\bar{O}(j, g)$, where $\bar{O}(j, g)$ is the expected number of reads computed using Eq. (1). Taxa are removed if the variation $\phi(j) = \max_g \xi(j, g) - \min_g \xi(j, g)$ of $\xi(j, g)$ of their residuals $\xi(j, g)$ exceeds a predefined threshold T. After removing taxa, all abundances are re-computed, the threshold is halved, and another round of false-positive removals is done. We stop after a specified number of rounds (per default 4).

6.7 Simulated mock community dada

To simulate metagenomic sequencing libraries of a mock community under different models of GC bias, we first simulated an unbiased library by randomly by drawing 1 million 400bp fragments from the reference genomes of the 20 taxa (19 distinct species) in the mock community of Tourlousse et al. (Tourlousse et al., 2021). When then sub-sampled the combined library (in which each taxon has a sequence abundance of 0.05) to artificially introduce GC bias, retaining each fragment with probability $\bar{\eta}(g)$. Here, $\bar{\eta}(g)$ is the GC-dependent sequencing efficiency which we set to: $\bar{\eta}(g) = 1 - 10(g - 0.5)^2$ for the efficiency peaking at 50% GC, $\bar{\eta}(g) = 0.125 + 1.25g$ for increasing efficiency with GC content, and $\bar{\eta}(g) = 1 - 1.25g$ for decreasing efficiency. Finally, we generated a pair of 150bp reads from the two ends of each fragment. For the analysis of the simulated libraries we used a taxonomy containing only the 20 genomes present in the library and ran GuaCAMOLE (in sequence abundance mode on the species level and with fragment length set to 400bp) and Bracken, both with read threshold 500.

6.8 Experimental mock community data

For the analysis of the mock community data of Tourlousse et al. (Tourlousse et al., 2021) (SRA accession SRS7661134), we used the RefSeq release 220 database containing human, archaeal, viral, plasmid and bacterial DNA. We ran GuaCAMOLE (in taxonomic abundance mode), Bracken and MetaPhlAn4. For Bracken and GuaCAMOLE we set the read threshold to 500. For GuacAMOLE we set the read length to 150bp and the fragment length to the value observed for each protocol: 200bp for D0, IL and IH, 250bp for DH, BL, BH, EH, EL, 325bp for F0, FL, FH, 350bp for JH, AH, C0, CL, 400bp for AL, A0, JL, and 300bp for all other protocols. GuaCAMOLE was run with 3 false positive removal iterations. For MetaPhlAn4 we the CHOCOPhlAn database v202103 with default parameters, and manually corrected the classification of *F. prausnitzii* to *F. duncaniae* since this recent reclassification is

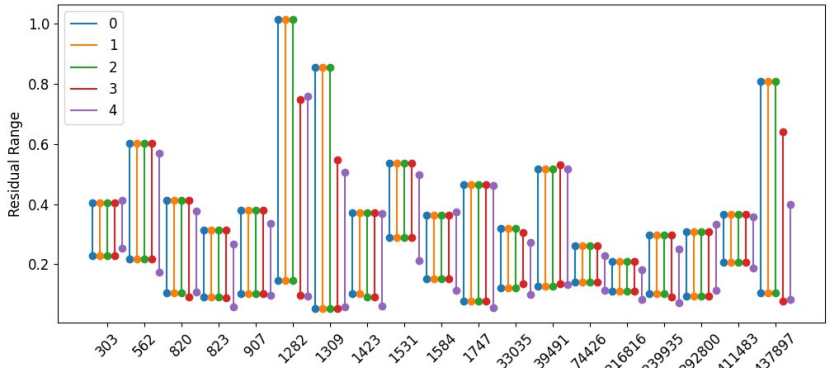
not reflected by CHOCOPhlAn v202103. Since Bracken always reports sequence abundances, we adjusted the Bracken-estimated abundances using the same genome length estimates as GuaCAMOLE uses to make them comparable.

6.9 Analyzing Colorectal Cancer Microbiome Data

We downloaded human gut microbiomes of 80 individuals (DDBJ Sequence Read Archive accession DRA006684) published by Yachida et al. (Yachida et al., 2019). The microbiomes were obtained from fecal samples of 40 colorectal cancer patients and 40 healthy individuals. We ran GuaCAMOLE in sequence abundance mode using the RefSeq 220 taxonomy with parameters read length 150bp, read threshold 1000, regularisation 0.1 and 4 outlier removal rounds. For comparison we ran Bracken with the same database and read threshold 1000. Since abundances ranged over multiple orders of magnitude, we computed the average abundance of taxon j as $a_j = \exp\left((1/n)\sum_k \log a_{j,k}\right)$ where $a_{j,k}$ denotes the abundance in sample k.

Appendix A Supplementary Information

A



B

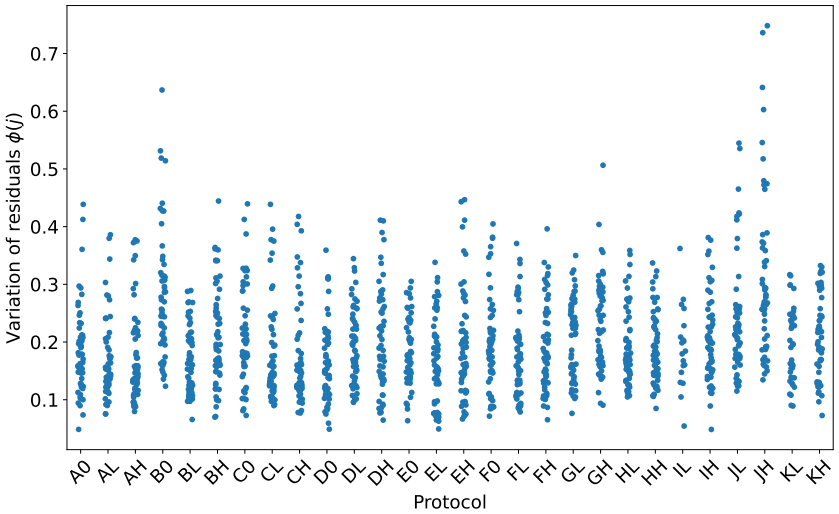


Fig. A1 Variation of residuals for mock community data. Variations $\phi(j) = \max_g \xi(j,g) - \min_g \xi(j,g)$ for the 19 species in the mock community data of Tourlousse et al. (Tourlousse et al., 2021). **(A)**. Minimal and maximal value of variation $\phi(j)$ of residuals for each taxon j observed across all samples 28 sequencing protocols after 0, 1, 2, 3 and 4 rounds of outlier removals. **(B)**. Variation $\phi(j)$ of residuals of individual samples per protocol after iteration 3.

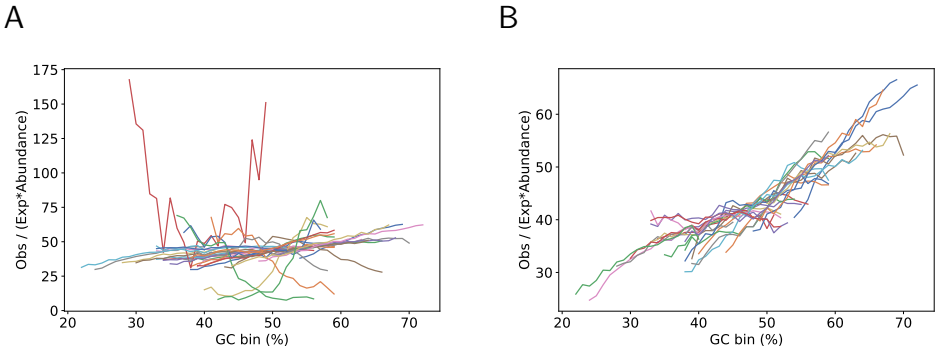


Fig. A2 Inferred taxon-specific efficiencies for mock community data. Estimated taxon-specific GC-dependent efficiencies for protocol DH replicate 1. Y-axis shows η_g as defined in Eq. (7) without the pre-factor C/N. **(A)** Before outlier removal. **(B)** After 4 rounds of outlier removal.

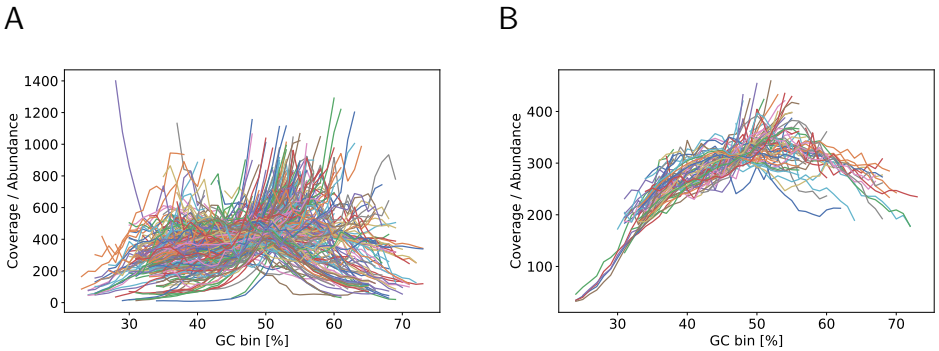


Fig. A3 Inferred taxon-specific efficiencies for a human gut microbiome sample. Estimated taxon-specific GC-dependent efficiencies for sample DRR127476. Y-axis shows η_g as defined in Eq. (7) without the pre-factor C/N. **(A)** Before outlier removal. **(B)** After 4 rounds of outlier removal.

References

Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. PloS one. 2010;5(4):e10209.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359.

Integrative H, Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, et al. The integrative human microbiome project. Natur. 2019;569(7758):641–648.

Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. Plant–microbiome interactions: from community assembly to plant health. Nature reviews microbiology. 2020;18(11):607–621.

Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Research. 2019;26(5):391–398.

Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. Bmc Genomics. 2015;16:1–12.

Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proceedings of the National Academy of Sciences. 2015;112(45):14024–14029.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome biology. 2019;20(1):1–13.

Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology. 2023;p. 1–12.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research. 2008;36(16):e105.

Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research. 2012;40(10):e72–e72.

Tourlousse DM, Narita K, Miura T, Sakamoto M, Ohashi A, Shiina K, et al. Validation and standardization of DNA extraction and library construction methods for metagenomics-based human fecal microbiome measurements. Microbiome. 2021;9(1):1–19.

Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience. 2020;9(2):giaa008.

Yu T, Guo F, Yu Y, Sun T, Ma D, Han J, et al. Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy. Cell. 2017;170(3):548–563.

Han YW. *Fusobacterium nucleatum*: a commensal-turned pathogen. Current Opinion in Microbiology. 2015;23:141–147.

Yang Y, Weng W, Peng J, Hong L, Yang L, Toiyama Y, et al. *Fusobacterium nucleatum* Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor-κB, and Up-regulating Expression of MicroRNA-21. Gastroenterology. 2017;152(4):851–866.e24.

McLaren MR, Nearing JT, Willis AD, Lloyd KG, Callahan BJ. Implications of taxonomic bias for microbial differential-abundance analysis. bioRxiv. 2022;p. 2022–08.

Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nature biotechnology. 2016;34(12):1287–1291.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):1–12.

Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nature communications. 2019;10(1):3066.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017;3:e104.

Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nature medicine. 2019;25(6):968–976.

Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, et al. Challenges in benchmarking metagenomic profilers. Nature methods. 2021;18(6):618–626.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology. 2011;12:1–14.