Reduced Metagenome Sequencing for strain-resolution taxonomic profiles

Lars Snipen*

Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences

P.O. Box 5003, NO-1432, Ås, NORWAY

Email: lars.snipen@nmbu.no

*Corresponding author

Inga-Leena Angell

Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences

P.O. Box 5003, NO-1432, Ås, NORWAY

Email: inga.angell@nmbu.no

Torbjørn Rognes

Department of Informatics, University of Oslo

P.O. Box 1080 Blindern, NO-0316, Oslo, NORWAY

Email: torognes@ifi.uio.no

Knut Rudi

Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences

P.O. Box 5003, NO-1432, Ås, NORWAY

Email: knut.rudi@nmbu.no

RESEARCH

Reduced Metagenome Sequencing for strain-resolution taxonomic profiles

Lars Snipen[1*], Inga-Leena Angell[1], Torbjørn Rognes[2,3] and Knut Rudi[1]

[*]Correspondence:
lars.snipen@nmbu.no
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432, Ås, NORWAY
Full list of author information is available at the end of the article

Abstract

Background: Studies of shifts in microbial community composition has many applications. For studies at species or subspecies levels, the 16S amplicon sequencing lacks resolution, and is often replaced by full shotgun sequencing. Due to higher costs, this restricts the number of samples sequenced. As an alternative to a full shotgun sequencing we have investigated the use of Reduced Metagenome Sequencing (RMS) to estimate the composition of a microbial community. This involves the use of double-digested restriction associated DNA sequencing, which means only a smaller fraction of the genomes are sequenced. The read sets obtained by this approach have properties different from both amplicon and shotgun data, and analysis pipelines for both can either not be used at all or do not explore the full potential of RMS data.

Results: We suggest a procedure for analyzing such data, based on fragment clustering and the use of a constrained ordinary least square de-convolution for estimating the relative abundance of all community members. Mock-community data sets shows the potential to clearly separate between strains even when the 16S is 100% identical and genome-wide differences is < 0.02, indicating RMS has a very high resolution. From a simulation study we compare RMS to shotgun sequencing and show that we get improved abundance estimates when the community has many very closely related genomes. From a real data set of infants guts we show that RMS is capable of detecting a strain-diversity gradient for *Escherichia coli* across time.

Conclusion: We find that RMS is a good alternative to either metabarcoding or shotgun sequencing when it comes to resolving microbial communities at the strain-level. Like shotgun metagenomics, it requires a good database of reference genomes, and is well suited for studies of the human gut or other communities where many reference genomes exist. A data analysis pipeline is offered, as an R package at https://github.com/larssnip/microRMS.

Keywords: metagenome; strains; ddRADseq

Background

The study of microbial communities relies on the sequencing of microbial DNA, and current practice can be divided into two main approaches: Metabarcoding, also known as amplicon or targeted sequencing, and shotgun sequencing of random fragments from the entire genomes [1]. The amplicon-approach is primarily used for revealing the taxonomic composition, but may also be used to study the distribution

of targeted functional genes [2]. Shotgun sequencing provides a potentially more detailed information about the community genomes, the microbiome, and is typically used for studies that digs beyond the composition and into the genomic function. Shotgun microbiome sequencing requires significantly more efforts in sequencing, data processing and analysis compared to meta-barcoding.

In most microbiome studies the composition is of interest, and in some cases it is all we require. Shifts in composition may be used as indicators in various ways, e.g. microbiota profiles in forensics [3], or in the surveillance of environments [4]. For communities like the human gut, extensive studies of the composition has given us the big picture, but recent investigations indicates that differences at the strain level may be crucial for phenotypic differences [5, 6]. Common to these problems is the need for high-resolution taxonomic profiles that can be collected with moderate efforts and in a reproducible way. Such studies often require many samples in order to capture the biological variation, and since sequencing and computational resources are always limited, the simpler amplicon approach is often preferred to a deep shotgun sequencing in order to get enough samples covered. However, the standard approach using the 16S rRNA gene marker has a limited resolution. If separation at the species or strain level is required, the 16S marker is in general too conserved, and a full shotgun sequencing seems necessary.

As an alternative to do a full shotgun metagenomic sequencing, the use of restriction enzymes to reduce the genomic sequence space has also been employed to investigate microbial communities [7, 8, 9]. The double-digested restriction associated DNA sequencing (ddRADseq) idea [10, 11, 12] has been along for some time, but its use for metagenome studies is quite new. The main advantage for this approach has been to reduce the sequencing efforts, and thereby costs per sample. In short, this means cutting DNA into fragments using two different restriction enzymes, followed by a PCR amplification and sequencing of the resulting amplicons. This procedure falls between the full shotgun sequencing and the classical use of a specified marker gene (16S). It has some resemblance to shotgun sequencing, since from each genome we sequence many different, in some sense random, fragments that varies in size and number between genomes. But, this also resembles metabarcoding, since multiple copies of a certain genome will produce the exact same fragments, and the reads are from these fragments each time. Thus, the approach has been termed Reduced Metagenome Sequencing (RMS). The wet-lab protocols for ddRADseq are well established and straightforward [8].

There are some difficulties that arise with the RMS approach. When target sequencing a pre-defined marker gene like the 16S, we may cluster reads into OTUs or sequence variants, where each cluster represents some taxon. RMS reads may also be clustered, but each taxon give rise to a variable number of distinct fragments, and it is difficult to infer the taxonomic composition from such read clusters without mapping to some references. Also, due to the variable lengths and compositions of the fragments, the PCR-amplification efficiency must be expected to vary and create biases. For purely predictive purposes, such reference-free approaches may still be useful, as suggested by [9].

In this paper we focus on the use of RMS data for estimating taxonomic profiles, as an alternative to a full shotgun sequencing. This means reads are in some way

mapped to a database of reference genomes, often referred to as closed-reference analysis. Our focus is on the estimation of high-resolution abundances, i.e. species or strain-level profiles, which is not attainable by conventional 16S sequencing. It is possible to use computational tools designed for shotgun data directly in the RMS setting. This may produce helpful results, but do not utilize all the information we have in this case. We propose an alternative analysis approach, using fragment clustering and a constrained least squares estimation. Based on mock community data and simulations, we demonstrate some important aspects of RMS data, and show the potential for RMS to improve composition estimates at the strain level. We also include an example of using RMS to estimate strain-diversiy for *Escherichia coli* in the infant gut microbiome. The analysis tools, along with some tutorial material, is freely available as an R package at the GitHub site https://github.com/larssnip/microRMS.

Results

We have explored the Reduced Metagenome Sequencing approach for studying the composition of microbial communities, with a focus on high-resolution profiles. The RMS idea is to cut genomes into fragments using restriction enzymes, then amplify and sequence the resulting fragments. In this study we have only considered the use of the restriction enzymes EcoRI and MseI. The data processing pipeline illustrated in Figure 1 will apply to any choice of enzymes, but some of the choices made along the way may change.

In Figure 2 we see how RMS fragment number and lengths distribute for a selection of species typically found in the human gut. This will clearly change if other restriction enzymes are used. For each species we randomly selected 10 sequenced strains, and all results are based on retrieving the RMS fragments *in silico* from the genomes, using the cutting motifs `GAATTC` (EcoRI) and `TTAA` (MseI). In the upper panel we notice that the number of RMS fragments per megabasepairs varies a lot between species, but less within each species. The number of RMS fragments is typically limited by the occurrence of the longer cutting motif, in this case `GAATTC`. Since this is a GC-poor motif, there is an effect of GC-content, and the grey sector indicates where the numbers should have been had it been random DNA. In the lower panel the densities show how fragment lengths distribute. Here we selected three species only, having low (*F. nucleatum*), medium (*E.coli*) and high (*B. longum*) GC-content. The fragment lengths are typically governed by the occurrence of the shorter motif, in this case `TTAA`, and again there is a huge effect of GC-content. The GC-poor *F. nucleatum* has very short fragments, since the short motif occurs more frequently than in the more GC-rich species. There are fragments longer than 1000 bases, but these typically produce low signals after PCR amplification and are not shown here. For any choice of restriction enzymes, similar investigations should be made to see how many and how long fragments one should expect from the species most likely to be found in the targeted samples.

Figure 3 illustrates the potential for high-resolution using the RMS method. Here we have considered the 27 genomes used in our mock community below. In panel A we used the 16S sequence from each genome. These were aligned (MUSCLE, [13]) and the p-distance (1 minus identity, i.e. distance 0.01 corresponds to 99%

identity) between them computed using the ape-package in R [14]. We notice that strains within the same species are identical or close to identical, and even the two *Staphylococcus* species are difficult to separate, with p-distance < 0.01. OTUs based on 16S data are usually clustered at distance 0.03. In panel B we computed the p-distances based on whole genomes, using the fastANI software ([15]). This separates the *Staphylococci* better than 16S, but strains within the same species are again quite similar. The two strains of *Helicobacter pylori* have p-distance 0.04 between them, indicating 96% of their genomes are identical. In panel C is the correlation distance between genomes based on RMS fragment copy numbers. From the genomes we get the copy number matrix $\boldsymbol{X}$, with one row for each fragment cluster and one column for each genome. The two *L. gasseri* strains have a small correlation distance, making them as good as impossible to separate with RMS. But, the other species with multiple strains are far better off, and the correlation distance of 0.22 between the two *S. mutans* strains should be large enough for discrimination between them.

Previous use of ddRADseq have reported biases in the signals, most likely introduced in the PCR amplification of such fragments [16]. To investigate this we used mock data from [8], where some samples included a single genome only. In Figure 4 we have plotted the data from four such samples. From panel A we clearly see how fragment length affects the relative readcount signals by the banana-shaped cloud of strong signals. The cloud of very low signals (note log-transformed y-axis) is due to noise. In panel B there seems to be no bias due to GC-content. Based on repeated observations of similar patterns, we propose a length-normalization of the RMS signals. In panel C we show the effect of this procedure described in the Methods section. We also suggest, based on panels A and C, that only fragments within the length interval $30 - 500$ bases should be used in the downstream analyses, highlighted in panel C of Figure 4. In this interval the length bias is small, and the normalization procedure will not affect the data very much, which is always a good thing. From Figure 2 we also saw that most fragments are in this length-interval when using the current restriction enzymes. Clearly, these limits must be re-considered if other enzymes are used.

Next, we used the RMS approach on the mock community data. In Figure 5 we show a classical stacked barplot displaying the estimated composition of the 20 genome mock. The estimates are based on the constrained ordinary least squares (COLS, see Methods for details) procedure, with 0.1 trimming, described in the Methods section. The database contains the genomes from all the 27 genomes, but only the original 20 genomes were included in this sample. Proportions are well estimated, and of the seven extra genomes absent in the samples, six of them are correctly estimated to have no contributions. The only false positive (weak) signal is from *L. gasseri* ATCC 33323, as expected from the results in Figure 3. In Figure 6 we display the actual versus the predicted relative abundances as scatterplots for each of the other mock-combinations. Here the extra strains were spiked-in, one by one, and in one sample all seven were added. From this we note that proportions are in most cases well estimated, where strains who are absent are estimated with zero proportions, and when strains are spiked-in they are estimated with fairly accurate proportions. There are two exceptions. As seen also in Figure 5, the *L.*

gasseri ATCC 33323 strain is estimated as (weakly) present also in those cases it is absent, and the spiking-in of *H. pylori* NCTC 11637 seems to have failed in some way, coming out with much too low abundance in the two cases where it is present.

We also tested the RMS approach against a standard shotgun sequencing procedure using simulated data. We focused on human gut-like communities of three different resolutions, where community members have a minimum whole-genome p-distance of 0.05, 0.02 or 0.01. This resulted in 291, 601 and 1086 community genomes, respectively. In all cases the samples contained reads from 100 randomly sampled present genomes, but the databases (RMS and Kraken2) contained all community genomes, both those present and absent. Both from RMS and shotgun data we re-estimated the relative abundance of every single genome in the database, using the COLS method for RMS data and Kraken2 with a custom database for the shotgun data. To evaluate the results, we computed the Manhattan distance (or L_1 norm) between actual and predicted relative abundances, as suggested in [17]. Thus, a Manhattan distance of $D = 0$ means we estimate all relative abundances perfectly. In Figure 7 the actual versus the predicted abundances are plotted as scatterplots. We observe, as expected, that predictions are poorer for lower p-distances, i.e. it becomes more difficult to distinguish genomes as they become more similar. However, the difference between RMS (upper panels) and shotgun (lower panels) data is striking. With the RMS approach we can estimate the abundance of each genome quite well, while for shotgun data the variance becomes huge for the highest resolutions, with predicted abundances up to three times larger or smaller than the actual abundances.

Finally, we include some real data to illustrate the use of RMS for quantifying strain-diversity of *E. coli* in the infant gut. At the time of writing, there are 1066 complete *E. coli* genomes in the RefSeq database [18]. This is an example of a genome collection where we find some very similar genomes. We computed the whole-genome p-distance between all pairs of genomes, using the MASH software [19], as well as the RMS correlation distances described in the Methods section. In Figure 8 the left panel A scatterplot indicates how these distances relate to each other. We only plot the distance to the nearest neighbor for each genomes, and the grey dots are for all 1066 genomes. Note that some of distances are zero (both distance measures), indicating RefSeq contains multiple copies of identical genomes. The scatterplot also relates the correlation distance to the more familiar p-distance, and we observe that a correlation distance around 0.30 here corresponds to a p-distance of roughly 0.01, i.e. genomes of 99% identity.

Using all 1066 genomes turns out impossible, because the copy number matrix produces an infinite condition value. This is due to the identical genomes, which is (theoretically) impossible to separate. Hence, we employed the genome clustering described in the Methods section. Setting the maximum tolerated condition value at 100 produced 54 genome clusters, i.e. the *E. coli* population is divided into 54 subgroups with this resolution. The black dots in Figure 7 are the nearest neighbor distances for these 54 genomes. In the right panel B we indicate where these are located in a neighbor joining tree based on the p-distances between all strains.

Six of the samples from the infant guts were also subject to a conventional shotgun sequencing. We first made a comparison between shotgun and RMS, using Kraken2

and a custom database for assigning reads to the exact same 54 *E. coli* genomes that we used for the RMS analysis. Since we do not know the true composition of these samples, we only considered which strains were estimated to be present or absent in a sample. In the left panel A of Figure 9, we show a tree of the 54 strains, based on whole-genome p-distances, where we have colored the leaf nodes by how they were classified by either shotgun or RMS data in one of the samples. We can see that from the shotgun data, and the Kraken2 assignments, 51 of the 54 strains were assigned reads, and thereby being present (grey or black), while the RMS results only estimate 16 strains as present (black only), of which half is from the same clade at the top of the tree. The other five samples show a similar trend: The shotgun approach will assign reads to a majority of strains, while the RMS approach is more specific, stating fewer strains are present.

In the right panel B of Figure 9 the boxplot shows the number of strains found present by RMS in all 94 infants at 4 different times after birth. The variance is large, partly due to biological variation. Still, the trend of a growing diversity by age is clear, and a simple ANOVA analysis confirmed a highly significant increase in diversity from birth (Meconium) to all later times, especially to 12 months.

Discussion

The upper panel of Figure 2 shows that, using the restriction enzymes of this study, the RMS fragment density varies by GC-content. Most genomes also have fewer fragments than expected in random DNA, indicating a negative selection of the cut sites. However, multiplying by genome size, we find that most genomes have in the range of hundred to thousand fragments. The lower panel of Figure 1 shows most fragments are rather short. This is a good thing, since longer fragments amplify poorly, and we found that by only focusing on fragments in the length-interval $30-500$ bases, we obtain strong signals without too much PCR bias. These results apply to our chosen restriction enzymes, and should always be investigated for any alternative choice of enzymes.

From Figure 3 we clearly see the potential for RMS to resolve strains at a level which is impossible with 16S, and even difficult with shotgun sequencing. In Panel A we used full-length 16S sequences, but still the separation is very poor between closely related genomes. The distances in panel B reflects how similar the genomes are in overall nucleotide identity. As expected, strains within a species have p-distance less than 0.05, i.e. more than 95% identical. Panel C demonstrates that even closely related strains have rather large correlation distance, indicating a good number of unique RMS fragments. This may seem strange, how can genomes be so similar in p-distance, but still have different RMS fragments? Mutations in restriction cut sites as well as re-arrangements of genomic regions will both create/destroy RMS fragments, but have little impact on the whole-genome distance. Only if two or more genomes share the vast majority of RMS fragments, we would see a small correlation-distances, and a shallow branch in the dendrogram. From panel C in Figure 2 we expect to be able to separate all genomes by RMS, except perhaps the two *L. gasseri* strains.

Since the RMS approach involves a PCR step, we must expect some biases. In Figure 4 (panel A) we observe a distinct effect of fragment length on the relative

signals strengths we get, based on single-genome samples from a previous study. However, the GC-content of the fragments does not seem to have any effect (panel B), unlike what was reported by [16]. The lower panels of Figure 4 is our proposed way of handling the length-bias. First, only use fragments in the length interval $30-500$ bases, highlighted in brown color in panel C. As we saw in Figure 2, short fragments account for the vast majority anyway, and we have found that $80-90\%$ of the reads will map to fragments in this interval. Next, we propose a simple normalization, illustrated in panel C. The cloud of strong signals are 'straightened'. Note that due to the log-transformed y-axis it looks like weak signals (noise) are heavily distorted by the normalization, while their values actually changes very little. If other restriction enzymes are used, the fragment lengths may be different, and the limits of $30-500$ should be reconsidered. However, the length bias corrected by the normalization will probably be of the same type, since this is a PCR effect which is independent of the restriction enzymes.

The mock data results in Figure 5 and 6 reveal that with the RMS approach and the COLS algorithm we can estimate relative abundances fairly well. It should be noted that the actual abundances are probably not exact, as they rarely are in experimental data. The mock composition was designed by 16S copies, and the transformation to genome copies is not without uncertainty, since most of these species are known to have variable 16S copy numbers. Most important is that strains who are absent from a sample are also estimated to zero abundance, i.e. they show up in the lower left corner of the panels in Figure 6. The exception is *L. gasseri* ATCC 33323. This is simply too similar to the other *L. gasseri* strain, as seen in Figure 3 as well. Only five fragment clusters are unique to the ATCC 33323 strain, and with some noise signals on some of these, it appears to be present even when it is not.

In virtually all cases the three replicates (marker types) of each sample show very similar results, indicating there is very little variance in the RMS procedure as such. Hence, any deviations between actual and estimated proportions are most likely due to some systematic effect. On closer inspection we found that the RMS data display what we denote as a fragment bias. Fragments unique to a genome should in theory all produce similar readcounts, with some variation due to the randomness of sequencing. This is not the case. Some fragments consistently produce strong signals and others weak. This is also remarkably stable across all samples where a particular genome is present. We accounted for this in our simulation study, adding a random scaling to all fragments, and supplementary Figure 2 shows the distribution of this fragment bias. So far we have failed to reveal the cause of this effect. If we could understand and compensate for it, this would improve the precision and thereby the resolution of the method even further.

We used simulated data to compare the RMS approach to the use of shotgun sequencing in combination with the Kraken2 tool for re-estimating the relative abundances of each genome. Kraken2 is only one out of several tools for estimating metagenome composition, but we chose this because it has a good reputation, will always try to classify reads to the genomes of its database, but most importantly, the genomes in the database can be easily customized. To make the comparison fair, the database must be identical for both the RMS and shotgun approach. Using

a generic database is bound to produce poorer results compared to one where the exact genomes under study are in the database. An alternative tool like MetaPhlAn2 assigns reads no lower than to the species level, but has the extension StrainPhlAn [20] for a strain level analysis. It is, however, difficult to compare StrainPhlAn output to the ones we get here, since we focus on relative abundances of *a priori* defined genomes, while StrainPhlAn identifies strains *a posteriori* by aligning reads to a set of marker genes and output a multiple sequence alignment. It seems to us these are two quite different approaches to a strain-level analysis.

In Figure 7 we show some results of our simulation study. The left panels are from a community where no members have a p-distance below 0.05 to another member. This is roughly a community with one genome from each species. Here both methods perform extremely well, and the shotgun+Kraken2 is the best, with Manhattan distance $D = 0.034$. However, as the communities (and databases) are filled up with more and more similar genomes, the picture changes (middle and right panels). The shotgun+Kraken2 results are getting dramatically poorer, with highly fluctuating estimates of relative abundances. The RMS+COLS approach is also poorer, but not nearly as bad. While the Kraken2 results seem to be fairly unbiased, but with a huge variance, the COLS results have a small variance, at the cost of some bias in giving weak abundance to some absent genomes, seen in the lower left corners of the 98% and 99% panels. Our explanation for these results is that with shotgun sequencing most reads will match multiple genomes in the database, and Kraken2 will then assign to the lowest common ancestor, i.e. the species. Thus, species abundances become extremely precise, but too few reads are left at the strain level to get reliable estimates. In our COLS algorithm for RMS data, we also have many fragment clusters who are present in multiple genomes, but since we have the copy number matrix with this exact information, the constrained least square solution spreads the signal across all genomes instead of assigning it to their common ancestor. It should be mentioned that this idea has some resemblance to what was proposed by [21], using methods from RNAseq data as an alternative approach for analyzing shotgun data. The Kraken2 software also has an extension in the Bracken software [22], re-estimating low-rank abundances based on the higher-rank assignments, but this does not currently re-estimate below the species rank.

There is, as for any method, a limit to the resolution obtained by RMS. The example with 1066 *E. coli* genomes illustrates this. Many of these are more than 99% identical, some even 100%, as seen in Figure 8. We plotted the correlation distance between all genomes against the p-distance for the same pairs, to illustrate how they are related. An RMS correlation distance of around 0.30 corresponds roughly to a p-distance of 0.01 (99% identity) in this case. We employ a genome clustering, where we only keep a selection of the genomes, ensuring a minimum difference between them. This means each cluster centroide represents a subgroup of highly similar strains. When using the COLS algorithm the resolution is limited by the condition value of the fragment copy number matrix. A very large condition value indicates the estimated abundances will be unstable. Condition values of 10^2, 10^3 or even 10^4 may be used to obtain a gradually higher resolutions, but at the cost of more uncertain results. Even with the lowest threshold at 10^2 we get 54 subgroups in the analysis, and it is likely that in many cases such a resolution suffice. As seen

in Figure 8, these strains are typically $98-99\%$ identical (p-distance $0.02-0.01$) and represent the full tree of all strains quite well.

A shotgun sequencing should in theory be able to separate anything below 100% identical, but in reality not. Reads are not without errors, and read coverage is often poor for low-abundance taxa. The results in Figure 9 underlines this. The shotgun data indicate almost all genomes in the database are present in the sample. With the RMS approach, much fewer genomes are present. In one out of the six comparable samples the methods came out with the exact same genomes as present and absent. In the others the shotgun data always results in more detected strains. It is reasonable to suspect both methods are too sensitive, assigning too many subgroups as present, but RMS seems far better in this respect. The total fraction of *E. coli* is small in these samples (around 1%) but the absolute number of reads assigned to this species are in the same range for both methods (around 1000). It is in fact slightly larger for the RMS data, hence the increased prevalence from shotgun data is not due to increased coverage. For shotgun data reads can originate from all locations on the genomes, making it notoriously difficult to map a read correctly when genomes are as similar as here, and given that reads may contain errors. RMS reads are assigned to the *a priori* known fragments, and allows for some slack due to sequencing error. Also, if genome A and B share 50% of their fragments, but only the genome A fragments have signal, the COLS algorithm will assign abundance 0.0 to genome B even if 50% of its fragments have signal. This is possible because we know these fragments are shared with genome A, and since the unique genome A fragments have signal while the unique genome B have not, the shared fragment signals are all allocated to genome A, giving no abundance to genome B.

The boxplots in panel B of Figure 9 is an example of how we use RMS to detect a change in strain-diversity over time in the infant gut. The increasing diversity by age is as expected. This example also illustrates how patterns emerge because we were able to sequence many samples, rather than deep sequencing of a few, where the biological variation probably would obscure the results. Such a high-resolution analysis would not be possible by 16S analysis.

The RMS has been proposed as a low-cost alternative to a full shotgun sequencing [**?**], since we only sequence the amplified fragments accounting for a fraction of the entire genomes. This is true if you use a reference-free approach where you need to cluster the reads, and hence need to have sequenced the same region of a genome several times in order to say something about abundance. However, as long as reads are mapped to reference genomes, this difference in library complexity is no longer important. Instead, the potential gain in using RMS lies in precise estimates of strain-resolution profiles. As for shotgun data, there is no theoretical lower sequencing depth that is required, the more reads the better. For the mock data results in Figures 5 and 6, where strains separated nicely, each sample had between 1 and 2 million reads mapped to some fragments, resulting in mostly $10-100$ read per fragment. This we consider a very good coverage. As always, high coverage is needed for detecting low-abundance taxa, and the separation between closely related strains is less influenced by this. A bottleneck for RMS is the fragment bias previously mentioned. For some reason, fragments from the same genomes tend

to get quite different readcounts, in a reproducible way. If a genome has as very few fragments, the average readcount for these is not as stable as with many fragments.

We believe our results indicate the RMS approach for metagenome profiling is something to explore further. We have only used one pair of restriction enzymes in this study, but other enzymes are used for similar studies ([7, 9]. The choice of enzymes will affect the number and length of fragments, but apart from this the data analysis procedure we propose here may be used. In the supplied software (R package) there are options for using any pair of restriction enzymes. The RMS approach, like the shotgun metagenome approach, requires sequenced reference genomes to map against in order to produce taxonomic profiles. To obtain this at the strain level, we need good reference databases. The good news is that recent extensive efforts provide us with many new reference genomes, especially for the human gut [23, 24, 25, 26, 27, 28, 29]. We believe that with evolving sequencing technologies, the quality of metagenome aquired genomes (MAGs) will improve drastically, and the road lies open for more strain-level taxonomic profiling.

Conclusion

We have demonstrated that the RMS approach can be used for taxonomic profiling of microbial communities down to the strain level. Compared to the conventional 16S approach, we find that strains with identical 16S genes are clearly discriminated by RMS, and we can estimate abundances for such strains in the same sample. The reason for this is simply that even genomes with identical 16S sequences will in most cases differ in a fair number of RMS fragments, enough to obtain strain-specific signals for the COLS algorithm.

Compared to the shotgun metagenome approach, the RMS offers an advantage in only sequencing *a priori* known amplicons, and we may construct a copy number matrix revealing the relations between all reference genomes prior to any sequencing. From this information, and the suggested constrained ordinary least squares estimation algorithm, we can obtain strain-level abundance estimates at least as good as the much used metagenome tool Kraken2. A clustering of genomes into species subgroups is proposed, as a way of balancing high resolution against precision in estimated abundances.

Based on this, we conclude that the RMS approach is worth pursuing, as a tool for studies of composition in the human gut or other microbial communities of particular interest and where a comprehensive collection of reference genomes exists. An R-package with the data analysis methods suggested here, as well as tutorials, is available in GitHub at https://github.com/larssnip/microRMS.

Methods

Mock data

In order to test the RMS approach, and learn about how such data behave, a mock-community study was conducted. As a basis we used a mock community of 20 genomes obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project (Genomic DNA from Microbial Mock Community B (Even, Low Concentration), v5.1L, for 16S rRNA Gene Sequencing, HM-782D, [30]), see Table 1. This mock has been constructed to yield 100000 16S copies from each

included organism. We converted this into the number of genome copies by dividing 100000 by the 16S copy number for each organism, as listed in the Ribosomal RNA Database [31]. In addition to this mock itself, we spiked-in 7 additional DSMZ strains [32]. These strains were selected to be highly similar, and with identical 16S gene, to one of the existing strains in the mock, to see if we could separate signals from such closely related organisms. One strain was spiked-in at a time, producing 7 additional samples. The spiked-in genomes were also at 100000 16S copies, controlled by a droplet digital PCR procedure. Finally, a sample with all 27 strains was also used. All these 9 mock mixtures where done in triplicates, resulting in 27 samples. All samples were subject to the wet-lab procedures described in [8] to obtain paired-end Illumina HiSeq reads. The restriction enzymes EcoRI and MseI were used throughout this study. All the strains involved in all samples have whole genome sequence data publicly available, and these were downloaded from the NCBI Genome database [33].

Infant gut data

As an illustration of a high-resolution analysis, we used a set of RMS data from the gut of infants. The microbiome was sampled from feces of 94 infants at meconium (newborn) and 3, 6 and 12 months age. We used a genome collection consisting of all complete RefSeq genomes of *E. coli* (1066 genomes) in order to look at strain diversity in these samples. Six of the samples were also sequenced by conventional shotgun sequencing, for comparison. All RMS samples were subject to the wet-lab procedures described in [8]. Both RMS and shotguns samples were sequenced by Illumina HiSeq, resulting in 150 bp paired-end reads.

Fragment copy number matrix

There exists a number of computational tools for estimating the taxonomic composition of a community based on shotgun data, e.g. Kraken2, MetaPhlAn2, CLARK, Kaiju etc ([34, 35, 36, 37]). Common to all is that reads are somehow mapped to some database of reference genomes. This is also required for RMS data. Given the reference genomes, and the cutting patterns of the restriction enzymes used (EcoRI and MseI), all RMS fragments were collected *in silico* from each genome. The RMS fragments are simply all sub-sequences starting with an EcoRI motif `GAATTC` ($5'$ of fragment), and ending by the first downstream MseI motif `TTAA`, containing none of these motifs inside. Genomes will in general have many such RMS fragments of highly variable lengths.

Fragments from closely related genomes may be identical or very similar. Also, some fragments may occur multiple times within a single genome. To reflect this we did a clustering of the fragment sequences into what we denote *fragment clusters*, using the VSEARCH software [38] and some specified identity threshold, similar to OTU-clustering for 16S data. An identity threshold of 0.99 was used in this study, but other thresholds were tested without significant changes in results. Each fragment cluster was represented by its centroid sequence and the fragment cluster copy number was stored in a *copy number matrix*. This matrix $\boldsymbol{X}$ has one row for each fragment cluster, and one column for each genome, and the integer in cell $\boldsymbol{X}(i,j)$ is the number of fragments from genome j that belongs to cluster

i. If we have a large collection of genomes, this matrix becomes huge. However, most fragment clusters occur in only one, or a few, genomes and most cells in the matrix are zero. Thus, the copy number matrix was stored as a sparse matrix data type, allowing most matrix operations but using comparatively little memory. This copy number matrix is an essential ingredient in the estimation of community abundances, as described below.

Read processing

We used the software `VSEARCH` [38] for all processing of reads. All reads were subject to a quality filtering, keeping only reads with an expected error rate below 0.02. Read-pairs were then merged. Since RMS fragments vary in length, some longer fragments produce non-overlapping reads. Thus, non-merged reads were included as single reads, where the R2-reads were reverse-complemented. To maintain the correct per-fragment read count, all merged reads were given a count of 2, while the single reads counts as 1. All reads were then de-replicated to obtain fasta-files of unique reads for all samples. Proper use of the `--sizein` and `--sizeout` options in `VSEARCH` allows us to work with the smaller set of unique reads without losing any information about actual read abundances.

Next, the processed reads from each sample were mapped to the fragment cluster centroids, using `VSEARCH` and the identity threshold from the fragment clustering (0.99, see above). This produced a readcount matrix $\boldsymbol{Y}$, with one row for each fragment cluster and one column for each sample.

Length-normalization

We realized the need for correcting the readcount signals due to fragment length PCR bias. First, let $\boldsymbol{y}_k$ denote raw readcounts from sample k, i.e. column k in $\boldsymbol{Y}$. Thus, $\boldsymbol{y}_k(i)$ is the raw readcount for fragment cluster i, and L_i is the length of cluster centroid i. Setting aside all clusters with zero signal, the $c_i = \log_{10}(\boldsymbol{y}_k(i))$ is simply the log-readcount. Next, we fitted a locally weighted scatterplot smoother (loess) $S(c_i|L_i)$ to these data, thus $S(c_i|L_i)$ is a smooth curve describing how log-readcounts c_i vary by fragment length L_i. Then, a correction factor for fragment cluster i is given as

$$f_i = 10^{(S_{max} - S(c_i|L_i))}$$

where S_{max} is the maximum value on the loess-curve. The normalized readcount for any fragment cluster is then

$$\hat{\boldsymbol{y}}_k(i) = \boldsymbol{y}_k(i) \cdot f_i$$

This multiplicative adjustment means fragments with zero signal remain zero also after normalization. This normalization is done for each sample separately. If the database contains a huge number of fragment clusters (many genomes), only a random sub-sample of them may be used to fit the loess model, in order to save time and memory.

Constrained ordinary least squares (COLS)

If all fragment clusters were unique to a single genome, the abundance of each genome would naturally be estimated by averaging the read counts for their corresponding fragment clusters. However, many RMS fragment clusters may be found in several genomes, and more closely related genomes will share more fragment clusters.

Prior to sequencing we constructed the copy number matrix from the G genomes in the database. This results in C fragment clusters, thus the copy number matrix $\boldsymbol{X}$ has C rows and G columns. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_G)$ be the proportion of the various genomes in sample k, i.e. $\beta_j \geq 0$ and $\sum_j \beta_j = 1$. Then, it is reasonable to assume that

$$E(\boldsymbol{y}_k) = \alpha \boldsymbol{X}'\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{y}_k$ are the (normalized) readcounts for each of the C RMS fragment clusters given the data from sample k, $\boldsymbol{X}$ is the copy number matrix and α is some positive scaling factor relevant for sample k. Thus, the expected signal for fragment cluster i, $E(\boldsymbol{y}_k(i))$, is proportional to the linear combination of fragment cluster copy numbers and genome abundances.

Given this model, the scaled proportions can be estimated using the constrained ordinary least squares (COLS) approach: Find $\boldsymbol{s} = \alpha\boldsymbol{\beta}$ that minimizes

$$f(\boldsymbol{s}) = (\boldsymbol{y}_k - \boldsymbol{X}'\boldsymbol{s})'(\boldsymbol{y}_k - \boldsymbol{X}'\boldsymbol{s}) \quad \text{where} \quad s_j \geq 0 \tag{2}$$

From the requirement that the β's must sum to 1.0 we get that $\sum_j s_j = \alpha$ and the estimated relative abundance of each genome is

$$\hat{\boldsymbol{\beta}} = \frac{\hat{\boldsymbol{s}}}{\sum_j \hat{s_j}} \tag{3}$$

In the implementation of this de-convolution, we have added the possibility of a trimmed estimate. This means a two-stage estimation procedure: After the initial fitting of the model, as described above, the residuals $\boldsymbol{y}_k - \boldsymbol{X}'\hat{\boldsymbol{s}}$ for all fragment clusters are computed. Then, a user-selected fraction of the fragments with most extreme residuals are discarded, and the model is re-fitted on the trimmed fragment set. This makes the estimated abundances less sensitive to extreme signals from some fragments, but is also reduces the size of the data set.

Correlation distance and genome clustering

The COLS algorithm also indirectly suggests the maximum resolution possible to de-convolve. If two genomes are very similar, they will share most RMS fragment clusters, and their respective columns in the copy number matrix $\boldsymbol{X}$ become similar. The *correlation distance* between two genomes is simply 1 minus the correlation between their respective columns in $\boldsymbol{X}$. Thus, a correlation distance of 0.0 means the two columns are identical, and the genomes share all fragment clusters. A correlation distance can be maximum 2.0, meaning all fragments present in one genome is absent in the other, and vice versa.

When solving eq. (2) we need to invert the matrix $\boldsymbol{X}'\boldsymbol{X}$, and if two or more columns are too similar this matrix inversion becomes highly unstable resulting in poor abundance estimates. This instability is often quantified by the condition value of $\boldsymbol{X}'\boldsymbol{X}$. A perfect condition value of 1.0 means all columns in $\boldsymbol{X}$ are orthogonal, i.e. no shared fragments. As columns become more and more correlated, the condition value increases. By computing the condition value from $\boldsymbol{X}$ we get an idea of how solvable this is, prior to any experimental efforts.

Instead of trying to estimate the abundance of all genomes, we *cluster* them into groups, and replace them by the group centroide genomes, as a representative of each group. The centroide is the one with smallest sum of distances to all the others in the same group. This basically means we get fewer columns and rows in $\boldsymbol{X}$. We employed a clustering procedure as follows:

1. Compute the correlation distance between all pairs of genomes from the columns of $\boldsymbol{X}$.
2. Compute a single linkage hierarchical clustering of the genomes based on this. This results in a dendrogram.
3. Each height in the dendrogram corresponds to an alternative clustering. Choose the largest dendrogram height resulting in a copy number matrix with condition value below a user-specified tolerance.

In this way, the user specifies a tolerated upper condition value (e.g. 100 or 1000), and genomes will be clustered to the finest resolution not violating this. A larger tolerance value leads to a finer resolution, but also more unstable estimates.

Simulation study

We also included a simulation study, where we compared the RMS approach to shotgun metagenome sequencing at various resolutions. Genome similarity was computed as whole-genome p-distance, i.e. 1.0 minus the Average Nucleotide Identity (ANI). A whole-genome p-distance of 0.0 means identical genomes and above 0.3 means very different genomes. Strains from the same species usually have p-distance below 0.05. In most real communities, like the human gut, we must expect some closely related strains, having a p-distances well below 0.05.

In [27] 1520 genomes from the human gut were isolated and sequenced. The whole-genome p-distances between all pairs of these genomes were computed using the MASH software [19], and then used to form clusters at three different resolutions: p-distance 0.05, 0.02 and 0.01. The cluster centroids were used as community members. The following procedure was applied to all communities, separately: From a community of G genomes, a sample contained reads from 100 randomly selected genomes, i.e. 100 of the G genomes are present, the resmaining $G-100$ are absent. Their abundances were exponentially distributed such that the largest abundance was 100 times the lowest abundance (dynamic range of 100), see supplementary figure 1. Let $f_1, f_2, ..., f_G$ be the relative abundance for each of the G genomes in the community, i.e. 100 of them are positive and the rest are zero, and they all sum to 1.0. These values form the actual relative abundances that we later tried to estimate.

This was repeated 25 times for each community, forming 25 different samples. Note that for each sample new 100 present genomes were randomly selected from the sub-population, thus different genomes were present/absent in each sample.

Reads were simulated using the ART software [39], using Illumina HiSeq 2500 error profiles, resulting in paired-end reads of 150 bases. For each sample we simulated 1 million read-pairs, either as a shotgun sample or as an RMS amplicon sample.

Shotgun data

The ART software requires the user to supply the reference sequences to simulate from as well as the number of read-pairs to generate. In shotgun metagenome sequencing, the probability of a read-pair to originate from genome g is proportional to the abundance of the genome multiplied by its size. After fragmentation of the genomic DNA, the reads are sampled from this fragment pool, and larger and more abundant genomes will contribute with more fragments. Thus, if z_g is the size of genome g, we form a weight for genome g as

$$w_g = f_g \cdot z_g$$

Given that we sequenced a million read-pairs, these were spread out among the genomes by random sampling using the probabilities

$$p_g = w_g / \sum_{j=1}^{G} w_j$$

resulting in read-counts $r_1, r_2, ..., r_G$ for each genome. Note that genomes with zero abundance get zero reads. Finally, read-pairs were simulated from each genome, given these read-counts, and assembled into a pair of fastq-files. This was then repeated for each sample, producing new sets of fastq-files.

RMS data

Instead of random fragmentation of the genomic DNA, the RMS protocol results in amplicons based on the fragments we get from restriction enzyme cutting. For each genome sequence we collected the RMS fragments *in silico*, again using the EcoRI and MseI restriction enzyme cutting motifs. Next, we have observed two main biases in how the RMS fragments from a given genome contributes to the pool of sequenced amplicons:

First, there is a length bias, especially very long fragments are poorly amplified. Let l_{gk} be the factor that scales the amplification of fragment k in genome g. This is a function of fragment length only, and in supplementary figure 2 we show the function we used for simulating this.

Second, we have also observed that some fragments are consistently more or less represented in the reads from a given genome. We denote this the fragment bias. Let v_{gk} be this fragment bias factor for fragment k in genome g, i.e. it may scale the amplification of fragment k up ($v_{gk} > 1$) or down ($v_{gk} < 1$). These factors were sampled at random from the distribution in supplementary figure 2, once for each genome, and then used forever after. Both this distribution, as well as the length bias function, were estimated from real RMS data, using the restriction enzymes described above.

Together, this means that the fragments from genome g get the weights

$$w_k = f_g \cdot l_{gk} \cdot v_{gk}$$

where $k = 1, 2, ...F_g$ and F_g is the number of fragments in genome g. All fragments, together with their weights, were assembled for all abundant genomes, and the read-count for each fragment/amplicon was sampled at random, again using probabilities $p_g = w_g / \sum_{j=1}^{G} w_j$.

Note that for shotgun data the weights are only affected by genome abundance and size, while RMS data is affected by genome abundance, number of fragments, length distribution and fragment-bias distribution for the present genomes.

Databases

The databases contained all G genomes of the community, both the 100 present at various levels and the $G - 100$ absent.

For each community, all RMS fragments were found in all G genomes, and a copy number matrix was constructed using a 0.99 identity threshold, as described above.

For the shotgun data we used the Kraken2 software [34] to obtain relative abundance estimates. This tool has shown good results in several benchmarking studies [40, 41, 42], but more importantly, is equipped with excellent facilities for building a custom database. In order to make a fair comparison to the RMS approach, the database of reference genomes must be the same as in the RMS case. Thus, custom Kraken2 databases were constructed, containing all G genomes of the respective communities. Also, the taxonomy was extended correspondingly, to have a taxonomy-id for every single genome, making it possible for Kraken2 to list hits to each genome.

Analysis

The analysis of the RMS data was carried out as described above, but without any genome clustering, resulting in an estimate of the relative abundance of every genome in the database.

For the shotgun data, Kraken2 and its custom database was used to assign reads to the genomes, using the default confidence level of 0.0. Only reads assigned to the genome level were counted, since this is our focus. The read count for a genome was divided by the genome size (basepairs), to produce the genome signal. Finally, these signals were divided by the total sum of signals, to produce relative abundances for all genomes in the communities.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The mock datasets generated and analysed during the current study are available in the Sequence Read Archive repository, under the accession PRJNA574678, see https://www.ncbi.nlm.nih.gov/bioproject/PRJNA574678/. The computational methods described in this paper are available as an R-package. It is currently available at GitHub, together with some tutorials describing the analysis steps, see https://github.com/larssnip/microRMS.

Competing interests
The authors declare that they have no competing interests.

Funding
This project has been financed by the Norwegian University of Life Sciences and the project DigiSal NFR 248792.

Author's contributions
LS proposed the data analysis methods, did all R programming and drafted the manuscript. IA did all lab work related to experiments. TR did all C++ programming related to VSEARCH. KR conceived the idea. All were involved in discussing methods and editing the manuscript.

Acknowledgements
Not applicable.

Author details
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432, Ås, NORWAY. [2]Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316, Oslo, NORWAY. [3]Department of Microbiology, Oslo University Hospital, Rikshospitalet, P.O. Box 4950 Nydalen, NO-0424, Oslo, NORWAY.

References
1. Breitwieser, F., Lu, J., Salzberg, S.: A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics, 1–15 (2017)
2. Liu, B., Zhang, X., Bakken, L., Snipen, L., Frostegård, .: Rapid succession of actively transcribing denitrifier populations in agricultural soil during an anoxic spell. Frontiers in Microbiology **9**(3208) (2019)
3. Nataas Hanssen, E., Hovde Liland, K., Gill, P., Snipen, L.: Optimizing body fluid recognition from microbial taxonomic profiles. Forensic Science International: Genetics **37**, 13–20 (2018)
4. Triadò-Margarit, X., Veillette, M., Duchaine, C., Talbot, M., Amato, F., Minguillòn, M.C., Martins, V., de Miguel, E., Casamayor, E.O., Moreno, T.: Bioaerosols in the barcelona subway system. Indoor Air **27**(3), 564–575 (2017)
5. Segata, N.: On the road to strain-resolved comparative metagenomics. mSystems **3**(2) (2018)
6. Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., Segal, E.: Structural variation in the gut microbiome associates with host health. Nature **568**(7750) (2019)
7. Liu, M., Worden, P., Monahan, L., DeMaere, M., Burke, C., Djordjevic, S., Chalres, I., Darling, A.: Evaluation of ddradseq for reduced representation metagenome sequencing. PeerJ **19**(5:e3837) (2017)
8. Ravi, A., Avershina, E., Angell, I., Ludvigsen, J., Manohar, P., Padmanaban, S., Nachimuthu, R., Snipen, L., Rudi, K.: Comparison of reduced metagenome and 16s rrna gene sequencing for determination of genetic diversity and mother-child overlap of the gut associated microbiota. Journal of Microbial Methods **149**, 44–52 (2018)
9. Hess, M., Rowe, S., Van Stijn, T., Henry, H., Hickey, S., Brauning, R., McCulloch, A., Hess, A., Kirk, M., Kumar, S., Pinares-Patiño, C., Kittelmann, S., Wood, G., Janssen, a., McEwan, J.: A restriction enzyme reduced representation sequencing approach for low-cost, highthroughput metagenome profiling. PLOS ONE **15**(4) (2020)
10. Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Homes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., Zabeau, M.: Aflp: a new technique for dna fingerprinting. Nucl. Acid Res. **23**(21), 4407–4414 (1995)
11. Lowry, D., Hoban, S., Kelley, J., Lotterhos, K., Reed, L., ntolin, M., Storfer, A.: Breaking rad: an evaluation of the utility of restriction site-associated dna sequencing for genome scans of adaptation. Molecular Ecology Resources **17**, 142–152 (2016)
12. Vendrami, D.L.J., Forcada, J., Hoffman, J.I.: Experimental validation of in silico predicted rad locus frequencies using genomic resources and short read data from a model marine mammal. BMC Genomics **20**(72) (2019)
13. Edgar, R.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5) (2004)
14. Popescu, A., Huber, K., Paradis, E.: ape 3.0: new tools for distance based phylogenetics and evolutionary analysis in r. Bioinformatics **28**, 1536–1537 (2012)
15. Jain, C., Rodriguez, L., Phillippy, A., Konstantinidis, K., Aluru, S.: High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. Nature Communications **9**(5114) (2018)
16. DaCosta, J., Sorenson, M.: Amplification biases and consistent recovery of loci in a double-digest rad-seq protocol. PLoS ONE **9**(9) (2014)
17. Meyer, F., Bremges, A., Belmann, P., Janssen, S., McHardy, A., David Koslicki, D.: Assessing taxonomic metagenome profilers with opal. Genome Biology **20**(51) (2019)
18. NCBI Genome/RefSeq. https://ftp.ncbi.nlm.nih.gov/genomes/refseq/
19. Ondov, B., Treangen, T., Melsted, P., Mallonee, A., Bergman, N., Koren, S., Phillippy, A.: Mash: fast genome and metagenome distance estimation using minhash. Genome Biology **17**(132) (2016)
20. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., Segata, N.: Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research **27**, 626–638 (2017)
21. Schaeffer, L., Pimentel, H., Bray, N., Melsted, P., Pachter, L.: Pseudoalignment for metagenomic read assignment. Bioinformatics **33**(14), 2082–2088 (2017)
22. Lu1, J., Breitwieser, F., Thielen, P., Salzberg, S.: Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. **DOI 10.7717/peerj-cs.104** (2017)
23. Almeida, A., Mitchell, A., Boland, M., Forster, S., Gloor, G., Tarkowska, A., Lawley, T., Finn, R.: A new genomic blueprint of the human gut microbiota. Nature **568**(7753) (2019)

24. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., Pike, L.J., Louie, T., Browne, H.P., Mitchell, A.L., Neville, B.A., Finn, R.D., Lawley, T.D.: A human gut bacterial genome and culture collection for improved metagenomic analyses. Nature biotechnology **37**, 186–192 (2019)
25. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., Kyrpides, N.C.: New insights from uncultivated genomes of the global human gut microbiome. Nature **568**(7753) (2019)
26. Pasolli, E., Asnicar, F., Manara, S., Quince, C., Huttenhower, C., Segata, N.: Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell **176**, 649–662 (2019)
27. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R., Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., Lin, Y., Tang, S., Huo, G., Xu, X., Hou, Y., Liu, X., Wang, J., Yang, H., Kristiansen, K., Li, J., Jia, H., Xiao, L.: 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature Biotechnology **34**, 179–185 (2019)
28. Almeida1, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z., Pollard, K., Parks, P., Hugenholtz, P., Segata, N., Kyrpides, N., Finn, R.: A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. bioRxiv preprint (doi: https://doi.org/10.1101/762682) (2019)
29. Hiseni, P., Rudi, K., Wilson, R., Hegge, F., Snipen, L.: Humgut: A comprehensive human gut prokaryotic genomes collection filtered by metagenome data. bioRxiv preprint (doi: https://doi.org/10.1101/2020.03.25.007666) (2020)
30. Pearce, M.M., Hilt, E.E., Rosenfeld, A.B., Zilliox, M.J., Thomas-White, K., Fok, C., Kliethermes, S., Schreckenberger, P.C., Brubaker, L., Gai, X., Wolfe, A.J.: The female urinary microbiome: A comparison of women with and without urgency urinary incontinence. mBio **5**(4) (2014)
31. Stoddard, S., Smith, B., Hein, R., Roller, B., Schmidt, T.: rrndb: improved tools for interpreting rrna gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research **43**, 593–598 (2015)
32. Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures. https://www.dsmz.de/
33. NCBI Genome Database. https://www.ncbi.nlm.nih.gov/genome
34. Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**(R46) (2014)
35. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N.: Metaphlan2 for enhanced metagenomic taxonomic profiling. Nature methods **12**, 902–903 (2015)
36. Ounit, R., Wanamaker, S., Close, T., Lonardi, S.: Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics **16**(236) (2015)
37. Menzel, P., Ng, K.L., Krogh, A.: Fast and sensitive taxonomic classification for metagenomics with kaiju. Nature communications **7**(11257) (2016)
38. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F.: Vsearch: a versatile open source tool for metagenomics. PeerJ (4:e2584) (2016)
39. Huang, W., Li, L., Myers, J., Marth, G.: Art: a next-generation sequencing read simulator. Bioinformatics **28**(4) (2012)
40. Lindgreen, S., Adair, K., Gardner, P.: An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports **6**(19233) (2015)
41. McIntyre, A., Ounit, R., Afshinnekoo, E., Prill, R., Hénaff, E., Alexander, N., Minot, S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R., Rosen, G., Mason, C.: Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology **18**(182) (2017)
42. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Sparholt Jørgensen, T., Shapiro, N., Blood, P.D., Gurevich, A., Bai1, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hestbjerg-Hansen, L., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Kang, D.D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H., Liao, Y., Silva, G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., McHardy, A.C.: Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nature Methods **14**(11) (2017)

Figures

Figure 1 An illustration of the suggested RMS profiling procedure. The left branch is executed once for a collection of reference genomes, except that the clustering of genomes may be done at various resolutions depending on later use. The right branch is done for each set of samples. The supplied R package has a tutorial illustrating all steps.

Figure 2 The distribution of RMS fragment number and lengths in some selected genomes, using the restriction cut sites `GAATTC` (EcoRI) and `TTAA` (MseI). There are 10 genomes for each of the species. In the upper panel each dot corresponds to a genome, and the grey sector indicates where the dots should be if it was purely random DNA. In the lower panel each density is based on data from 10 genomes.

Figure 3 The dendrograms display hierarchical clustering of the 27 genomes in the mock study. In panel A the distances are p-distances computed from a multiple alignment of the 16S sequences from each genomes. In panel B the p-distances are based on whole-genome comparisons, and in panel C we used correlation distances based on RMS fragment copy numbers.

Figure 4 In panel A fragment signal (relative readcounts) is plotted against fragment length, and in panel B against fragment GC-content. Each dot corresponds to a fragment cluster, and data from the four single-genome samples are displayed together. In panel C is shown the effect of the simple length-normalization on a single sample. Raw readcounts are normalized as described in the text. The red-brown color highlight the fragments within the length-interval $30-500$ bases. Note the log-transformed y-axes in all panels.

Figure 5 Actual and estimated composition of the mock with 20 genomes and no extra genomes spiked-in. The extra strains, absent from the samples, are indicated by the tan colors. The left bar is the actual compositions, and the three additional bars (A, B and C) show estimates from the three replicates of this sample.

Figure 6 Each panel show actual versus estimated proportions for all 27 genomes as markers. Each marker color corresponds to a genome, and the three marker types are the three replicates. Species with a single genome is colored in shades of grey, while those where two or three genomes are similar have more distinct coloring. There are eight different mocks, all consisting of the 20 genomes in the original mock, but with various additional genomes spiked-in. The header above each panel indicate which genome has been spiked-in. The gray line in the background is where marker should be if the estimates were perfect. Note that in the lower left corner (those who are absent and predicted to be absent) many markers of different colors overlap each other.

Figure 7 The scatter plots show actual versus predicted relative abundances for the simulated data. Each dot is a relative abundance of a genome, and each panel contain results from 25 samples. The upper panels are RMS data estimated by COLS, and the lower panels shotgun data estimated by Kraken2. The resolution of the communities increase from left to right, as indicated by the upper panel headers. The average Manhattan distance D is displayed within each panel.

Figure 8 In panel A each dot indicates distance to the nearest neighbor from an *E. coli* genome, measured either as RMS correlation distance (x-axis) or whole-genomes p-distance (y-axis). The grey dots are the results for all 1066 genomes, and the black dots are for the 54 genomes left after clustering with maximum condition value 100. In panel B the neighbor joining tree is based on the p-distances between all strains, and the black tips indicate the cluster centroide genomes.

Figure 9 The left panel A shows a tree for the 54 *E. coli* strains, based on the whole-genome p-distances, with a scale marker in the lower right corner. The leaf nodes are marked by how they were classified as present in a single sample. No leaf marker means classified as absent by both shotgun+Kraken2 and RMS+COLS. Grey markers indicate classified as present by shotgun+Kraken2 only. Black dots indicate classified as present by both methods. No genomes were classified as present by RMS+COLS only. In the right panel B the boxplot shows the number of genomes, out of the 54, estimated to be present by RMS+COLS over time in all infant gut samples (Meconium is newborn feces). There are 94 samples behind each box.

Tables

Table 1 Table 1. For each genome is listed its size (megabasepairs), GC-content and the number of RMS fragments in the $30-500$ length interval. Genomes with an asterisk after its name were spiked-in, and not part of the original mock.

Genome	Size	GC	RMS-fragments
Acinetobacter baumannii strain 5377	3.98	0.39	961
Schaalia odontolytica strain 1A.21	2.39	0.65	92
Bacillus cereus strain NRS 248	5.22	0.36	2025
Bacteroides vulgatus strain ATCC 8482	5.16	0.42	2047
Clostridium beijerinckii strain NCIMB 8052	6.00	0.30	2463
Cutibacterium acnes strain KPA171202	2.56	0.60	300
Deinococcus radiodurans strain R1	3.06	0.67	115
Enterococcus faecalis ATCC 19433*	2.87	0.38	902
Enterococcus faecalis ATCC 29212*	3.01	0.37	922
Enterococcus faecalis strain OG1RF	2.74	0.38	822
Escherichia coli strain K12 substrain MG1655	4.64	0.51	920
Helicobacter pylori NCTC 11637*	1.60	0.39	233
Helicobacter pylori strain 26695	1.67	0.39	255
Lactobacillus gasseri ATCC 33323*	1.82	0.35	662
Lactobacillus gasseri strain 63 AM	1.89	0.35	675
Listeria monocytogenes strain EGDe	2.94	0.38	1504
Neisseria meningitidis strain MC58	2.27	0.52	332
Pseudomonas aeruginosa strain PAO1-LAC	6.26	0.66	143
Rhodobacter sphaeroides strain ATH 2 4 1	4.13	0.69	92
Staphylococcus aureus strain TCH1516	2.88	0.33	861
Staphylococcus epidermidis FDA strain PCI 1200	2.50	0.32	854
Streptococcus agalactiae ATCC 13813*	2.11	0.35	661
Streptococcus agalactiae strain 2603 V R	2.16	0.36	692
Streptococcus mutans ATCC 25175*	1.99	0.37	671
Streptococcus mutans strain UA159	2.03	0.37	680
Streptococcus pneumoniae ATCC 6305*	2.02	0.40	709
Streptococcus pneumoniae strain TIGR4	2.16	0.40	771

Additional Files

Additional file 1 — Supplementary figure 1
All simulated samples contained reads from 100 randomly selected genomes, and their relative abundances in the sample were according to this barplot. The largest abundance is 100 times the smallest. Different genomes were selected as the most/least abundant and absent ones in each sample, but this abundance distribution was used every time.

Additional file 2 — Supplementary figure 2
In order to simulate RMS data, some known biases were introduced to the signals. The upper panel shows the fragment-length bias used. All signals were scaled by this function, i.e. fragments of length around 200 bases remained close to unchanged (scale 1.0) while signals from shorter or longer fragments were scaled down. The lower panel shows the fragment-bias distribution. For each fragment within a genome, a factor was sampled from this distribution, and the signals from the fragments were scaled accordingly. The mean value of this distribution is 1.0, but some fragments may have signals up to six times as large, or down to almost nothing. Both the length-bias function and the fragment-bias distribution were estimated from real RMS data.