Published online 31 August 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 1* ***e10***
doi:10.1093/nar/gks803

Metagenomic abundance estimation and diagnostic testing on species level

Martin S. Lindner and Bernhard Y. Renard*

Research Group Bioinformatics (NG4), Robert Koch-Institut, Nordufer 20, 13353 Berlin, Germany

Received April 27, 2012; Revised and Accepted July 31, 2012

ABSTRACT

One goal of sequencing-based metagenomic community analysis is the quantitative taxonomic assessment of microbial community compositions. In particular, relative quantification of taxons is of high relevance for metagenomic diagnostics or microbial community comparison. However, the majority of existing approaches quantify at low resolution (e.g. at phylum level), rely on the existence of special genes (e.g. 16S), or have severe problems discerning species with highly similar genome sequences. Yet, problems as metagenomic diagnostics require accurate quantification on species level. We developed Genome Abundance Similarity Correction (GASiC), a method to estimate true genome abundances via read alignment by considering reference genome similarities in a non-negative LASSO approach. We demonstrate GASiC's superior performance over existing methods on simulated benchmark data as well as on real data. In addition, we present applications to datasets of both bacterial DNA and viral RNA source. We further discuss our approach as an alternative to PCR-based DNA quantification.

Introduction

Metagenomic analysis of microbial communities using sequencing technologies increasingly draws attention (1) as the technical capabilities, both on the biological and computational side, evolve rapidly. Genome assembly is now even possible for low abundant species in complex metagenomic high coverage Next-Generation Sequencing (NGS) datasets (2) and the number of available reference sequences is increasing steadily.

Reference-based identification and quantification of the constituents is a key goal of metagenomic analysis and is a special case of 'taxonomic binning', i.e. finding the taxonomic affiliation of sequences in a dataset. Reads are typically assigned to nodes in a phylogenetic tree by either aligning them against the reference genomes or comparing statistical features of reads and references (3). However, abundance estimation is often not possible at species level (4) and is highly influenced by many factors such as genome length, genome similarity, reference set composition or phylogenetic structure.

One way is to align reads against a comprehensive reference sequence database using BLAST (5) and subsequently analyse the results with tools such as MEGAN (6). As reads—especially short NGS reads—often match to multiple genomes, MEGAN assigns these ambiguous reads to nodes in the pyhlogenetic tree by finding the 'Lowest Common Ancestor' node of all matching sequences. Assigning the reads to the Lowest Common Ancestor reduces the risk of a too optimistic assignment and thus of obtaining false positive matches; with the disadvantage that quantification may only be possible at a low resolution. Furthermore, MEGAN discards nodes with insufficient support, i.e. when the number of reads assigned to a node does not exceed a user-defined threshold. The graphical user interface makes MEGAN highly suitable for the visual inspection of metagenomic data. Yet, MEGANs read counts are influenced by several factors such as genome sizes or the presence of similar genomes in the phylogenetic tree, which makes MEGAN less suitable for quantitative metagenomic analyses.

Another tool based on read alignment, GAAS (7), uses an iterative procedure to estimate improved relative genome abundances and an average genome length. To this end, GAAS calculates genome length corrected alignment qualities (*E*-values) for all matching reads and uses this information to iteratively calculate weights for each reference genome. Yet, ambiguities of read matches are only considered indirectly via the corrected *E*-values, which is only suitable if the reference genomes have low similarity.

GRAMMy (8) successively improves on GAAS as it explicitly models read assignment ambiguities in a

*To whom correspondence should be addressed. Tel: +49 30 18754 2561; Fax: +49 30 18754 2328; Email: renardb@rki.de probability matrix. The problem is formulated as a finite mixture model which incorporates the read probability matrix and the genome lengths. The Expectation–Maximization algorithm is used to iteratively solve for the mixing parameters of the model: the relative genome abundances. In contrast to the previous methods, GRAMMy seeks to reflect the reference genome similarities in the mixture model. Yet, the similarity parameters are estimated from the alignment qualities of the reads to the reference genomes rather than from the reference genomes directly and are thus not accurate enough to allow robust abundance estimation in the case of highly similar reference genomes.

We observed that high similarity of reference sequences challenges all described methods. This can be problematic, for instance, in diagnostic settings, when the distinction between presence and absence of single species or relative abundance levels are of eminent importance. To overcome this limitation, we present Genome Abundance Similarity Correction (GASiC), a versatile algorithm to estimate corrected abundances on the species level by directly accounting for the reference genome similarities. We demonstrate that GASiC is able to provide accurate abundance estimates for reference genomes with high sequence similarity and for complex metagenomic communities. Its simulation-based approach makes GASiC more independent from biases introduced by the sequencing technology, differences in genome sizes, or composition and structure of the reference sequences. Furthermore, GASiC provides statistical tests for the presence of a species in the sample.

MATERIALS AND METHODS

The GASiC workflow is depicted in Figure 1. As in most reference-based methods, the reads are first aligned to every genome in a set of references and the number of reads matching to each genome is counted. We call these counts the ‘observed abundances’, as opposed to the ‘abundance estimates’ which we want to obtain in the end. In the next step, GASiC constructs a similarity matrix encoding the alignment similarities between the reference sequences. The similarity matrix and the observed abundances are then used together in a linear system of equations, where GASiC solves for the corrected abundances using a constrained optimization routine to obtain the estimates. The whole procedure can be iterated using bootstrap (9) samples from the original dataset. This yields more stable abundance estimates and provides an intuitive non-parametric statistical test for the presence of a species.

We first introduce some notation that will be used in the following. Starting from the experiment side, the sequencing dataset is denoted as D, containing N reads in total. The reads may originate from a set of M Species $S = \{S_i, i = 1..M\}$ with known reference sequences or possibly from other sources (noise, contaminants) with no relation to any species in S. S_i is synonymously used for both the species itself as well as its reference sequence. For quantification of species we use the term ‘abundance’, which is the number of reads belonging to the species divided by the total number of reads N. Due to amplification biases, this abundance may not represent the true absolute abundance of the species in the data, but may be valuable when comparing abundances of multiple (in particular similar) species.

Alignment

The reads in D are aligned to all species S with an alignment method suitable for the characteristics of D. Then, we count the number of reads r_i from D that were successfully aligned to S_i, irrespective of the number of matching positions in S_i or matches to other species. In particular, we neither restrict ourselves to unique matches only, nor assume any phylogenetic structure within the S_i, as is done for example in MEGAN. If the dataset only contains very dissimilar species, the read counts r_i may already be suitable estimates for the true abundances. Otherwise, the r_i are in general highly disturbed and dominated by shared matches, such that the r_i cannot directly be used as abundance estimates.

Similarity estimation

A proper similarity estimation of the reference sequences is required to achieve accurate similarity correction of the r_i. The similarities between sequences are encoded in a similarity matrix $A = (a_{ij}), i,j = 1..M$, where a_{ij} denotes the probability that a read drawn from S_i can be aligned to S_j. In practice, we simulate a set of reads from every reference S_i with a read simulator which is able to imitate the sequencing technology and error characteristics of D. For example, Mason (10) and Grinder (11) simulate Illumina, 454 and Sanger reads; and dwgsim (sourceforge.net/projects/dnaa/) simulates Illumina, ABI SOLiD and IonTorrent reads. Then, we align the simulated reads of S_i to S_j using the very same settings as for aligning the reads in dataset D and count the number of matching reads $\tilde{r}_{ij}$. The matrix entries are then estimated as $a_{ij} = \frac{\tilde{r}_{ij}}{\tilde{r}_{ii}}$.

The key element of similarity estimation is a proper read simulation since we use the simulated reads to estimate the reference genome similarities, the source of ambiguous alignments. Thus, the simulated reads should have the read characteristics and the error characteristics of the instrument (read length, paired/single end, etc.) and should cover the reference genome at least once.

For very complex metagenomic communities with a high number of species M, the calculation of the complete similarity matrix may become infeasible because of its computational complexity $O(M^2)$. We recommend to first estimate similarities using, for example, fast k-mer-based methods (12) and refine the estimates via the simulation approach only for genomes with sufficiently high (e.g. $a_{ij} > 0.01$) similarity.

Similarity correction

We introduce a linear model to correct the r_i for the genome similarity using the similarity matrix A. Let c_i denote the true, but unknown, abundance of species S_i. We then assume that the observed abundance r_i is a

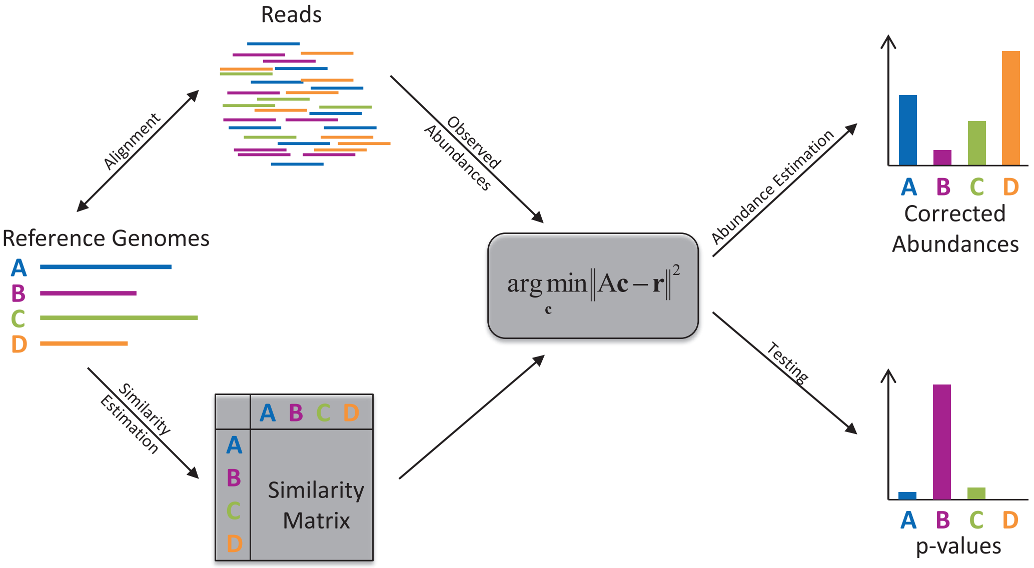

Figure 1. GASiC workflow. Metagenomic reads are first aligned to the reference genomes and matching reads are counted for each genome (observed abundances). GASiC then uses the reference genomes to construct a similarity matrix encoding the genome similarities while considering influences of the applied sequencing technology. The similarity matrix and the observed abundances are used in a linear system of equations to model the influence of reference genome similarities on read alignment. GASiC solves the system of equations using a constrained optimization routine to calculate the estimated true abundances of the reference genomes in the dataset. Bootstrapping from the reads delivers stable abundance estimates and allows GASiC to test for the presence of each species in the dataset.

mixture of the true abundances c_j of all species S_j, weighted with the estimated probability a_{ij} that a read from j can be aligned to i:

$$\sum_j a_{ij} c_j = r_i.$$

To simplify notation, we use a matrix representation of the true and the observed abundances, i.e. $\mathbf{c} = (c_1, c_2, ..., c_M)^{\mathrm{T}}$ and $\mathbf{r} = (r_1, r_2, ..., r_M)^{\mathrm{T}}$. In matrix notation, this can be written as

$$A\mathbf{c} = \mathbf{r}.$$

Since direct inversion of the matrix A may result in instable abundance estimates, we formulate the solution for $\mathbf{c}$ as a non-negative LASSO (13,14) problem:

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmin}} \, ||A\mathbf{c} - \mathbf{r}||_2$$

$$\text{s.t. } \hat{c}_i \geq 0 \; \forall i \text{ s.t. } \sum_i |\hat{c}_i| \leq 1.$$

The constraints enforce the result to be meaningful, i.e. each estimated relative abundance $\hat{c}_i$ must be equal to or greater than zero and the sum of all relative abundances must be less than or equal to one. The first conditions also ensure that the correction produces abundances lower than or equal to the measured abundances. The last condition allows the presence of reads from a totally unrelated species, since the abundances are allowed to sum up to less than or equal to one. It also enforces the sparsity of results such that only meaningful contributions have abundances larger than 0. We solve the constraint optimization problem with the COBYLA method implemented in SciPy (www.scipy.org/).

Error estimation and testing

We apply a bootstrapping procedure on the steps described before, first, to estimate how errors in the input data propagate through the correction algorithm and, second, to calculate P-values to test for the presence of a species in the sample. To this end, we generate B bootstrap samples from the dataset D and perform similarity correction for each sample separately, yielding a distribution $\hat{c}_{i,b}$ $b = 1..B$ of abundances for each species i. We calculate the average abundance $\overline{c_i}$ and estimate the standard error $\sigma_i = \sqrt{\mathrm{VAR}(\hat{c}_{i,b})}$. To test whether a species is present in the sample, we count how many bootstrap samples yielded a higher abundance than an a priori defined detection threshold t:

$$p(c_i > t) = \frac{\#(c_{i,b} > t, b = 1..B)}{B}.$$

Quality check

As the composition of the reference genome set is critical for the complete method, GASiC offers an additional quality check after the alignment to reference genomes. The quality check step analyses the outputted SAM files of the read alignment tool and provides helpful statistics to the user to judge the appropriateness of the results. Besides reporting statistical measures, such as the number of mapped reads or the average genome coverage, GASiC generates a coverage histogram which often allows the user to exclude certain genomes from the reference set or to detect possibly important missing reference genomes. For example, a high number of uncovered bases in combination with a typical Poisson distribution at higher coverage may indicate that the considered species is not contained in the dataset, but a closely related species. In addition to the statistics and the histogram, GASiC produces warning messages in critical setups, e.g. when the dataset may be too small for abundance estimation or large parts of the genome are not covered although there is evidence for the genome in the dataset.

Technical details

We implemented GASiC in the Python programming language (www.python.org), making extensive use of the high performance scientific computing libraries SciPy and NumPy (www.scipy.org). Since GASiC is independent from the choice of the alignment algorithm and read simulator, we already integrated interfaces to a set of tools. The user can add custom interfaces easily, a brief manual is provided within the code. We set value on comprehensible and well-documented code, such that GASiC can easily be adapted to the users needs without deeper knowledge of Python.

GASiC requires the widespread SAM alignment format (15) as output from the alignment tool to analyse the results, since most alignment tools either directly support SAM output or alignment results can be readily converted into SAM files.

RESULTS

We sought to corroborate the key features of GASiC with corresponding experiments. First, we compared GASiC with previous methods on a common reference dataset. Second, we demonstrate GASiC's power to disambiguate abundances of highly similar bacteria and to test for the presence of species. Third, we present a potential application besides metagenomics: we analysed a published viral dataset and compared GASiC's results with abundance levels obtained by a quantitative PCR method. The experimental settings are described in detail in Supplementary Methods.

FAMeS dataset

The established metagenomic FAMeS (16) reference datasets contain shotgun sequencing reads of 113 microbial species mixed into three datasets with low, medium and high complexity. The low complexity dataset `simLC` simulates a bioreactor community with one dominant and many low abundant genomes. The `simMC` dataset mimics a moderately complex community, as for example found in acid mine drainage biofilms, with few dominating species flanked by low abundant ones. A typical metagenomic dataset with high complexity and no dominant species is simulated in `simHC`. Ground truth is available, making these datasets an excellent choice to compare metagenomic algorithms.

Xia *et al*. compared the performance of the tools MEGAN, GAAS and GRAMMy on the FAMeS dataset, see (8) for details. We extended this comparison and measured GASiC's Relative Root Mean Squared Error and Average Relative Error (RRMSE and AVGRE) on all datasets. Given the true abundances t_i and the corrected abundances c_i, $i = 1..M$, the error measures are defined as follows:

$$\mathrm{RRMSE} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\frac{|c_j - t_j|}{t_j}\right)^2}$$

$$\mathrm{AVGRE} = \frac{1}{M}\sum_{j=1}^{M}\frac{|c_j - t_j|}{t_j}.$$

RRMSE measures the sum of 'squared' relative errors, whereas AVGRE is the sum of 'absolute' relative errors. Thus, RRMSE is more sensitive to outliers. The error measures of MEGAN, GAAS and GRAMMy, as reported in (8), and GASiC are compared in Table 1. Detailed results are reported in Supplementary Table S1. GASiC strongly reduces the estimated errors on all three datasets compared with the competing methods; the strongest error reduction is achieved on the high complexity simHC dataset, where the error rates are reduced by 51.9% and 60.5% for RRMSE and AVGRE, respectively. In particular, this high increase of accuracy demonstrates GASiC's ability to quantify low abundances correctly, even when a large number of reference genomes are used. Also the differing genome lengths (ranging from 1.0 Mbp to 9.7 Mbp) did not pose an obstacle for GASiC.

Mixed *Escherichia coli*/EHEC dataset

In the second experiment, we combined two real datasets, *E. coli* DH10B and *E. coli* TY-2482, in selected fractions. Both datasets were acquired with a IonTorrent PGM device. *E. coli* TY-2482 is highly similar to *E. coli* DH10B and received attention in the so called 'German 2011 EHEC outbreak' and we therefore, respectively, term the datasets *E. coli* and *EHEC* for a better differentiation. All combined datasets were analysed with GASiC and GRAMMy, the two best performing tools from the previous experiment. In addition to the *E. coli* and the EHEC references, we included *Shigella flexneri* as phantom reference. Herewith we challenged the tools, first, to distinguish highly similar reference genomes over a wide range of abundances and, second, to exclude reference genomes not present in the data. Figure 2 shows the estimated relative abundances of both tools for *E. coli*, EHEC and *Shigella*. Detailed results are reported in Supplementary Table S2. In contrast to GRAMMy, GASiC provides stable abundance estimates, especially in the case of low abundances. It persistently rules out the presence of all phantom references correctly, where the diagnostic detection threshold *t* in GASiC was set to disregard abundances below 1%. The statistical test for the presence of a genome assigns high *P*-values to *Shigella* in all datasets, to EHEC and *E. coli* only at concentrations of 1% or below, proving GASiC suitable for detecting the presence of low abundant genomes.

In follow-up experiments, we challenged GASiC under complicated conditions (Supplementary Methods and Supplementary Table S3). First, we added more highly similar phantom genomes to the reference set and observed that GASiC still provided accurate estimates

Table 1. Benchmark comparison. In addition to MEGAN based, GAAS and GRAMMy abundance estimates (8), we calculated abundance estimates with GASiC for all reference genomes in the FAMeS datasets simLC, simMC and simHC

Tool	simLC (%) Low complexity		simMC (%) Medium complexity		simHC (%) High complexity	
	RRMSE	AVGRE	RRMSE	AVGRE	RRMSE	AVGRE
MEGAN	48.6	39.3	50.0	40.6	50.2	40.8
GAAS	433.8	152.5	171.4	111.6	507.9	165.8
GRAMMy	20.0	14.0	25.6	19.7	21.6	14.7
GASiC	**18.7**	**9.1**	**17.5**	**10.9**	**10.4**	**5.8**

The four tools are compared by their relative error (RRMSE and AVGRE, see Methods section). The lowest error rates are shown in bold font. GASiC reduces the relative error on all datasets and improves on GRAMMy, the best existing tool, by up to 60%. Best results are achieved on the high complexity dataset simHC, indicating that GASiC provides a particularly large benefit for complex mixtures where more corrections are necessary and low concentrations exist which are more difficult to estimate.

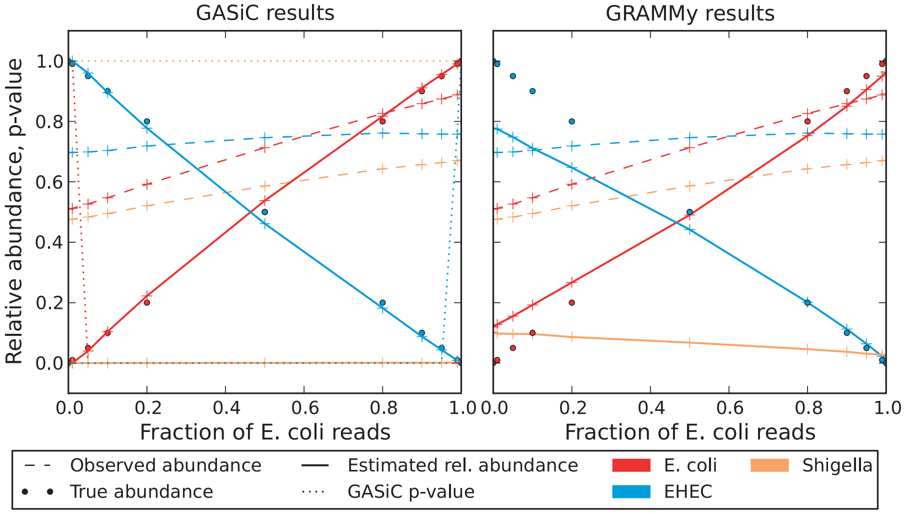

Figure 2. Comparison of GASiC and GRAMMy on synthetic datasets with varying concentrations of real *E. coli* and EHEC reads. Both algorithms estimated the relative abundances of the highly similar bacteria *E. coli*, EHEC, and *Shigella* in all datasets and GASiC tested (*P*-value) for the absence of each bacterium. GRAMMy was challenged by the similarity of the bacteria and deviated strongly from the expected relative concentrations. For *Shigella*, which was not present in the sample, GRAMMy incorrectly estimates abundances up to 10%. GASiC provided more stable abundance estimates at all concentrations and also correctly identified *Shigella* as not present in the dataset and accordingly assigned high *P*-values.

for all reference sequences. In a second experiment, we added noise reads to the dataset simulating a very distant unknown species in the metagenome. As the reads did not match to any of the reference sequences, GASiC's estimates were not affected by the noise reads. In a third experiment, we removed the EHEC genome from the reference set to simulate the effect of having a novel species in the dataset with high similarity to existing ones. Both GASiC and GRAMMy respond to the EHEC reads by overestimating the abundances of species with high similarity to EHEC, where GASiC produced overall better estimates than GRAMMy. Yet, more distant species are not affected. In this case, GASiC's quality check provides useful information to the experimentator, as it suggests that *Shigella* may not be present in the dataset. This contradicts GASiC's estimates and should encourage the experimentator to check manually whether a reference genome is missing. Lastly, we replaced the *E. coli* genome with contigs assembled from the original *E. coli* reads. Although the assembly consisted of 154 contigs and only accounted for 95% of the *E. coli* genome, GASiC was able to provide robust estimates for all involved species.

Viral RNA quantification

To demonstrate a possible application of GASiC beyond metagenomics, we analysed RNA data from a study on viral recombination in *Apis mellifera*, the honey bee. Moore *et al.* (17) analysed viral RNA of 40 honeybee pupae, many of them infested by Varroa destructor mites. They identified novel recombinations of the two *Picornavirales*, Deformed Wing Virus (DWV) and Varroa Destructor Virus-1 (VDV-1). The reference genomes of the recombinants, VDV-1_{DVD} and VDV-1_{VVD}, were published such that both the original and the recombinant sequences were available.

We used GASiC to estimate viral abundances for both the original and the recombinant genomes in the published NGS dataset used for identifying the recombinant genomes. This data posed a particularly difficult problem, since the reference sequences showed up to 96% sequence identity (Supplementary Table S4). Furthermore, since the considered species are RNA viruses, the reference sequences are only representatives for 'quasispecies clouds' of highly similar sequences (18). As the divergence of a quasispecies cloud is lower than the distance between the considered reference sequences (<4%), GASiC should be able to correct for the given similarities, although we expect the results to be not as precise as in other experiments.

GASiC's estimates are shown in Figure 3 and Supplementary Table S5, demonstrating that the high sequence similarities caused strong corrections to the number of matching reads. After correction, VDV-1_{DVD} was estimated as the most abundant virus while very low abundances were estimated for VDV-1. The high *P*-value ($P = 0.53$) suggests that VDV-1 is not present in the dataset. Furthermore, we see that recruiting only unique matches to estimate abundances would be misleading in this case, suggesting DWV as most abundant virus. We compared our estimates with the qRT-PCR results reported by Moore *et al.*, although they used different bee pupae for qRT-PCR than for sequencing. Yet, the results should be comparable since all pupae were collected from the same apiary. Moore *et al.* also found no evidence for VDV-1 and measured significant levels of VDV-1_{DVD} in all examined 25 bee pupae. DWV was found in 23 of 25 pupae, but at lower levels than VDV-1_{DVD}, and VDV-1_{VVD} was found in 15 of 25 pupae. A direct quantitative comparison with our estimates is not possible due to the differing biological samples and due to our estimates possibly being distorted by the quasispecies cloud nature of the viral RNA. Nevertheless, the virus levels obtained by Moore *et al.* coincide with the abundance estimates calculated by GASiC.

Furthermore, we estimated the viral abundances with GRAMMy to compare both tools on data with highly similar reference genomes. The experiment is described in the Supplementary Methods and GRAMMy's estimates are reported in Supplementary Table S5. We observe substantial differences between the GRAMMy estimates and the qRT-PCR/GASiC estimates. VDV-1 could not be found in the PCR experiment and was estimated to insignificantly low abundances by GASiC, yet, GRAMMy estimates (10.6 ± 0.3)% abundance for VDV-1. GRAMMy estimates DWV as most abundant virus, whereas the other methods identify VDV-1_{DVD} as most abundant and only observe relatively low abundances for DWV. The both recombinants, having very high similarity, were estimated by GRAMMy to about equal abundances of 27%.

DISCUSSION

Our experiments demonstrate GASiC's wide range of applicability in species quantification tasks. The FAMeS benchmark dataset consists of very few but long reads, thus only a very small number of reads is available for each reference genome. Although the long reads are ideal for metagenomic assembly and are thus frequently used for metagenomic analyses, the low number of reads encumbers quantification and thus challenges the algorithms. We demonstrated that GASiC greatly outperforms

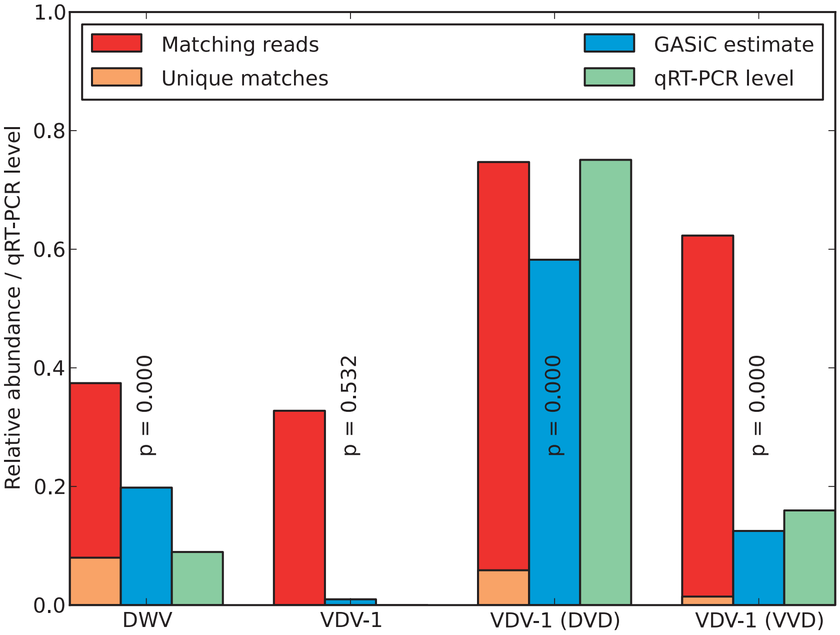

Figure 3. Estimation of viral abundances based on NGS and qRT-PCR. GASiC estimated the abundances of the highly similar bee viruses DWV, VDV-1, VDV-1_{DVD} and VDV-1_{VVD} in the viral RNA dataset acquired by Moore *et al.* (17). The abundances are displayed in relation to the total number of reads. GASiC's estimates coincide with the qRT-PCR quantification in the original paper: VDV-1_{DVD} was estimated as the most abundant virus and VDV-1 was correctly identified as not present in the dataset. The displayed relative qRT-PCR levels were calculated as described in Supplementary Methods. Interestingly, only considering the unique reads would have yielded misleading estimates (DWV as most abundant) in this experiment due to the high reference similarities.

all current competing algorithms on the FAMeS benchmark dataset. On the other hand, we demonstrated in the *E. coli*/EHEC experiment that GASiC handles mixtures of short read (80 bp) datasets of highly similar species better than GRAMMy, the best competing algorithm, and provides reliable tests for the presence of a species in the dataset. Also the different data sources did not challenge GASiC: whereas the aforementioned two datasets are bacterial DNA sequences, the bees dataset from the last experiment contains viral RNA reads. Also the extremely high sequence similarity (up to 96% nucleotide identity) of the viral reference sequences did not challenge GASiC.

This generality is mainly due to the fact that GASiC is independent from the underlying alignment algorithm: genomic similarities are estimated by aligning simulated reads to the reference genomes using the very same alignment tool and settings as for aligning the metagenomic reads. Thereby, tool characteristics are automatically canceled out.

Furthermore, GASiC is independent from any phylogenetic information, genome annotation or marker genes. Thus, GASiC is not restricted to the bacterial or viral domain only, but can be applied to sequences of any source, as long as reference sequences are available. This makes GASiC particularly appealing for metagenomic analyses, where large fractions of the analysed community may be uncategorized or a mixture of viral and bacterial sequences may be present.

We demonstrated that the common practice to only consider uniquely matching reads for abundance estimation can be heavily misleading. The high genomic similarity of the two bee viruses VDV-1_{VVD} and VDV-1_{DVD} yields relatively low numbers of unique reads for both of them, although VDV-1_{DVD} was the most abundant genome in the dataset.

One obvious drawback of GASiC is its need for reference sequences. Especially in complex metagenomic datasets, typically not all constituents are sequenced or even known. We identified four typical scenarios when GASiC can be applied: (i) when the metagenomic community is well-known from previous studies and comprehensive reference databases are available. This can be the case in metagenomic time series experiments, where the same community is sequenced repeatedly to observe temporal changes in the relative abundances of species. (ii) GASiC can be used to identify genomes present in a metagenomic dataset, when the community structure is not precisely known, but exhaustive databases of reference sequences are available. We demonstrated that GASiC still provides reliable estimates when more genome sequences are added to the reference set; this is particularly interesting for diagnostic settings of well-specified organisms and also for future applications, since the number of available reference genomes increases rapidly. (iii) GASiC can be applied when the scope of the study is to estimate abundances for a well-known-closed subset of sequences, i.e. a set of sequences which has a sufficiently high genomic distance to all other genomes, such that the probability of falsely aligning reads to sequences of the closed subset is very low. We observed (Supplementary Methods) that unknown sequence reads with low similarity do not diminish GASiC's accuracy. These closed subsets can be obtained, for example, by clustering sequences by similarity or using tools such as MEGAN to carefully pick references by hand. (iv) GASiC is applicable in experiments with high sequencing coverage or low community complexity, such that a preceding assembly step could directly deliver the references for quantification (2). We demonstrated this (Supplementary Methods and Supplementary Table S3) by replacing the *E. coli* genome in the Mixed *E. coli*/EHEC dataset experiment by contigs readily assembled from *E. coli* reads and obtained GASiC estimates similar to using the *E. coli* reference.

We see difficulties for the application of GASiC when the reference set composition is insufficient; e.g. when the dataset contains reads of a novel species which is highly similar to an existing species or a known species obtained novel genomic fragments via gene transfer (EHEC) or recombination (DWV/VDV-1). We also expect problems in precisely estimating abundances in small datasets containing high numbers of species, which is often the case for traditional Sanger sequencing experiments. However, the quality check step in GASiC outputs warnings when the risk of misinterpretation of results arises and thus serves as an automated indicator of these situations.

Scenario (iv) is particularly interesting as it is applicable when a metagenomic community is barely known, which is the case in many metagenomic studies. Yet, a complete assembly of all constituents of the sample is unrealistic, even in the case of a community with low complexity. Yet, GASiC is able to estimate abundances of single-assembled contigs or groups of contigs when algorithmically treated as a discrete 'species'. For example, rough estimates of species (groups of contigs) abundances or abundances of single genes (encoded on the contigs) can be obtained in this way. This concept can also be applied to fragments of genomes, as for example to fragmented RNA viruses or functional units in the genome. As observed in the viral RNA quantification experiment, quantifying complete genomes may be prone to errors when recombination occurred. Quantification of fragments may lead to more meaningful results if the recombinant genomes are not known. Nevertheless, it is not directly possible to detect recombination events with GASiC, although highly differing abundance estimates of fragments may be a sign for recombination.

CONCLUSION

We conclude that GASiC is a highly accurate and robust tool for genome abundance estimation and detection on the species level in metagenomic datasets. The similarities of reference genomes, being the main source of ambiguities in most metagenomic methods, are used directly to correct observed abundances. No prior information is needed for the analysis apart from the reference, making GASiC suitable for a broad range of applications. GASiC reduces quantitative error by as much as 60% over the best existing approaches for complex mixtures and quantitatively distinguishes even highly related organisms with more than 95% sequence similarity. We obtained accurate estimates on both viral and bacterial datasets from different sequencing platforms. Furthermore, we observed that GASiC's abundance estimates conform with virus levels obtained with qRT-PCR. This indicates that additional PCR-based quantification may become unnecessary if NGS data are available.

AVAILABILITY

The GASiC tool and source code are available for download at http://sourceforge.net/projects/gasic/. All data used in the experiments are available online. See Supplementary Table S6 for URLs and accession numbers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6, Supplementary Methods and Supplementary References [19–21].

ACKNOWLEDGEMENTS

We thank Franziska Zickmann, Martina Fischer and Anke Penzlin (Robert Koch-Institute, Germany) for critical reading of the manuscript and useful comments.

FUNDING

Generous start up funds from the Robert Koch-Institute. Funding for open access charge: Institutional funding (Robert Koch-Institute, Berlin, Germany).

Conflict of interest statement. None declared.

REFERENCES

1. Allen,E. and Banfield,J. (2005) Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.*, **3**, 489–498.
2. Iverson,V., Morris,R.M., Frazar,C.D., Berthiaume,C.T., Morales,R.L. and Armbrust,E.V. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, **335**, 587–590.
3. Wooley,J., Godzik,A. and Friedberg,I. (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.
4. Liu,B., Gibbons,T., Ghodsi,M., Treangen,T. and Pop,M. (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genom.*, **12**, S4.
5. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
6. Huson,D., Auch,A., Qi,J. and Schuster,S. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
7. Angly,F.E., Willner,D., Prieto-Davó,A., Edwards,R.A., Schmieder,R., Vega-Thurber,R., Antonopoulos,D.A., Barott,K., Cottrell,M.T., Desnues,C. *et al.* (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.*, **5**, e1000593.
8. Xia,L., Cram,J., Chen,T., Fuhrman,J. and Sun,F. (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, **6**, e27992.
9. Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.
10. Holtgrewe,M. (2010) Mason – a read simulator for second generation sequencing data. *Technical report TR-B-10-06*. Institut für Mathematik und Informatik, Freie Universität Berlin.
11. Angly,F.E., Willner,D., Rohwer,F., Hugenholtz,P. and Tyson,G.W. (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94.
12. Reinert,G., Chew,D., Sun,F. and Waterman,M. (2009) Alignment-free sequence comparison (I): statistics and power. *J. Comp. Biol.*, **16**, 1615–1634.
13. Renard,B.Y., Kirchner,M., Steen,H., Steen,J.A.J. and Hamprecht,F.A. (2008) NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, **9**, 355.
14. Efron,B., Hastie,T., Johnstone,I. and Tibshirani,R. (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.
15. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
16. Mavromatis,K., Ivanova,N., Barry,K., Shapiro,H., Goltsman,E., McHardy,A.C., Rigoutsos,I., Salamov,A., Korzeniewski,F., Land,M. *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.
17. Moore,J., Jironkin,A., Chandler,D., Burroughs,N., Evans,D.J. and Ryabov,E.V. (2011) Recombinants between Deformed wing virus and Varroa destructor virus-1 may prevail in Varroa destructor-infested honeybee colonies. *J. Gen. Virol.*, **92**, 156–161.
18. Fishman,S.L. and Branch,A.D. (2009) The quasispecies nature and biological implications of the hepatitis C virus. *Infect. Genet. Evol.*, **9**, 1158–1167.
19. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
20. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
21. Chevreux,B., Wetter,T. and Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. *Bioinformation Systems e.V.: Proceedings of the German Conference on Bioinformatics (GCB)*, Vol. 99, pp. 45–56.