11 NGS based metagenomics

This exposition has been developed by Knut Reinert.

In NGS based metagenomics we assume to be given an set of NGS sequencing reads which stem from an unknown set of taxa, that can contain more or less similar genomes. This naturally leads ot some basic problems in metagenomics.

- *Estimating the abundance* of the species contained in the sample, when the genomes are given
- *Binning the reads* when the genomes are not given, thereby given an estimate for the number and abundance of u nknown genomes.
- Producing *genome assemblies* of all the species contained in the sample when the genomes are not given.

In this lecture I present a method for the first problem.

The following exposition is mostly based on:

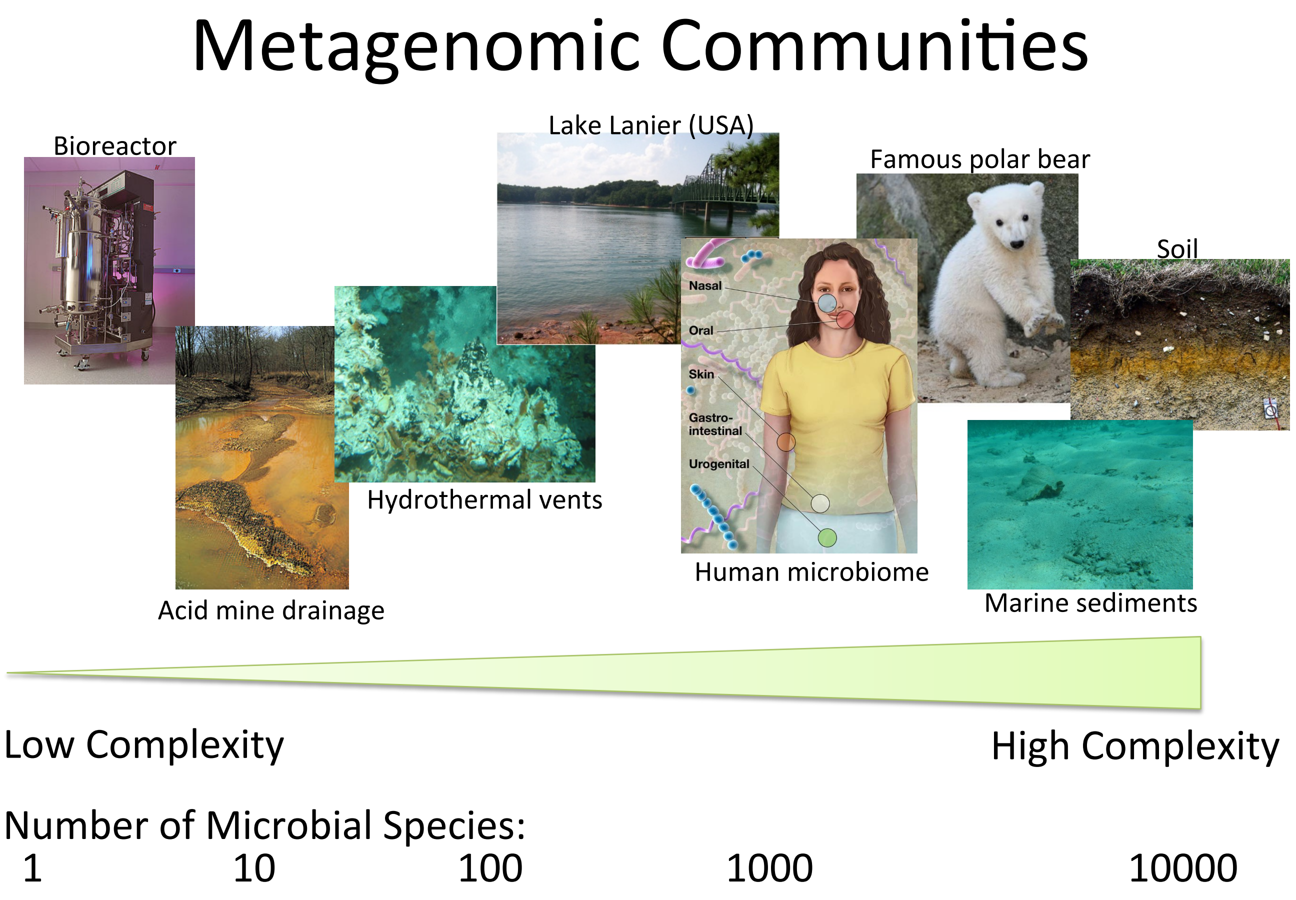
Lindner, M., Renard, B. : Metagenomic abundance estimation and diagnostic testing on species level (Nucleic Acids Research, 2012) and slides by Martin Lindner.

11.1 Metagenomic abundance estimation

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.



Metagenomic Communities

Bioreactor

Acid mine drainage

Hydrothermal vents

Lake Lanier (USA)

Human microbiome

Famous polar bear

Soil

Marine sediments

Low Complexity

High Complexity

Number of Microbial Species:

1 10 100 1000 10000

11.2 Megan

One way is to align reads against a comprehensive reference sequence database using BLAST or another local alignment tool and subsequently analyse the results with tools such as MEGAN (Huson et al.).

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.

11.3 GAAS

Another tool based on read alignment, GAAS (Angly et al.), 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.

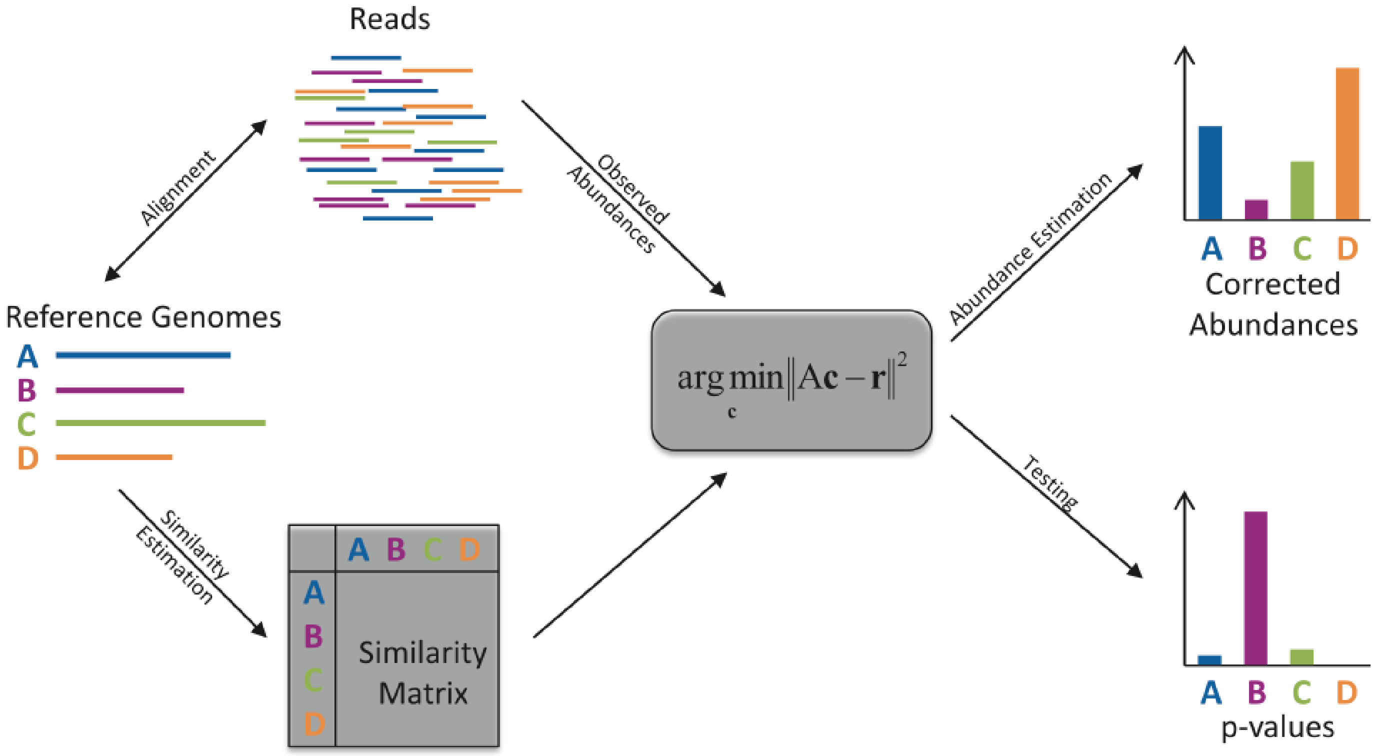
11.4 GAAS

GRAMMy (Xia et al.) successively improves on GAAS as it explicitly models read assignment ambiguities in a 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.

11.5 General overview

Here is the general overview of the method



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.

11.6 Similarity estimation

The reads in D are aligned to all species S with the *same* read mapper and *identical settings*. 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.

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 can not directly be used as abundance estimates.

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.

For this the authors simulate a set of reads from every reference S_i with a read simulator. Then, they map the simulated reads of S_i to S_j using the very same settings as for mapping 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 read mapper used was BowTie2.

The authors 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 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:

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

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

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

Obviously one can try to compute c now simply by computing $c = A^{-1}r$. But as it can be seen below direct inversion of the matrix A may result in instable abundance estimates.

Hence the authors formulate the solution for $\mathbf{c}$ as a non-negative LASSO (least absolute shrinkage and selection operator) problem:

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmin}} ||A\mathbf{c} - \mathbf{r}||_2$$
$$\text{s.t. } \hat{c}_i \geq 0\ \forall i \text{ and } \sum_i |\hat{c}_i| \leq 1$$

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.

For example given the measured abundances (note that through reads that match on multiple genomes the sum can be greater than one):

$$\begin{pmatrix} 0.3241 & 0.2756 & 0.6799 & 0.5296 \end{pmatrix}$$

and a similarity matrix

$$\begin{pmatrix} 1 & 0.0034 & 0.2258 & 0.2758 \\ 0.0058 & 1 & 0.4420 & 0.4561 \\ 0.3566 & 0.4354 & 1 & 0.7009 \\ 0.3901 & 0.4601 & 0.7000 & 1 \end{pmatrix}$$

The naive matrix inversion yields:

$$\begin{pmatrix} 0.1771 & -0.0082 & 0.5791 & 0.0589 \end{pmatrix}$$

While the constrained, nonnegative LASSO yields:

$$\begin{pmatrix} 0.1801 & 1.292e-27 & 0.5766 & 0.0540 \end{pmatrix}$$

This example is from a real data of three bee viruses (follows later).

11.7 P-Value computation

The authors apply a bootstrapping procedure 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, they 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. Then they 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, they first count how many bootstrap samples yielded a higher abundance than an a priori defined detection threshold t:

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

Using those probabilities, they test, whether they can reject the Null-Hypotheses ”The concentration of a species is smaller than t”. For small probabilities they have to reject it, for high probabilities, they do not reject it and the P-value for the test is then 1 indicating that the concentration of a species under the threshold.

11.8 Evaluation - Experiment 1

Lets see how the method works.

There is an established metagenomic reference dataset (FAMeS) that contains 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 using the following measures:

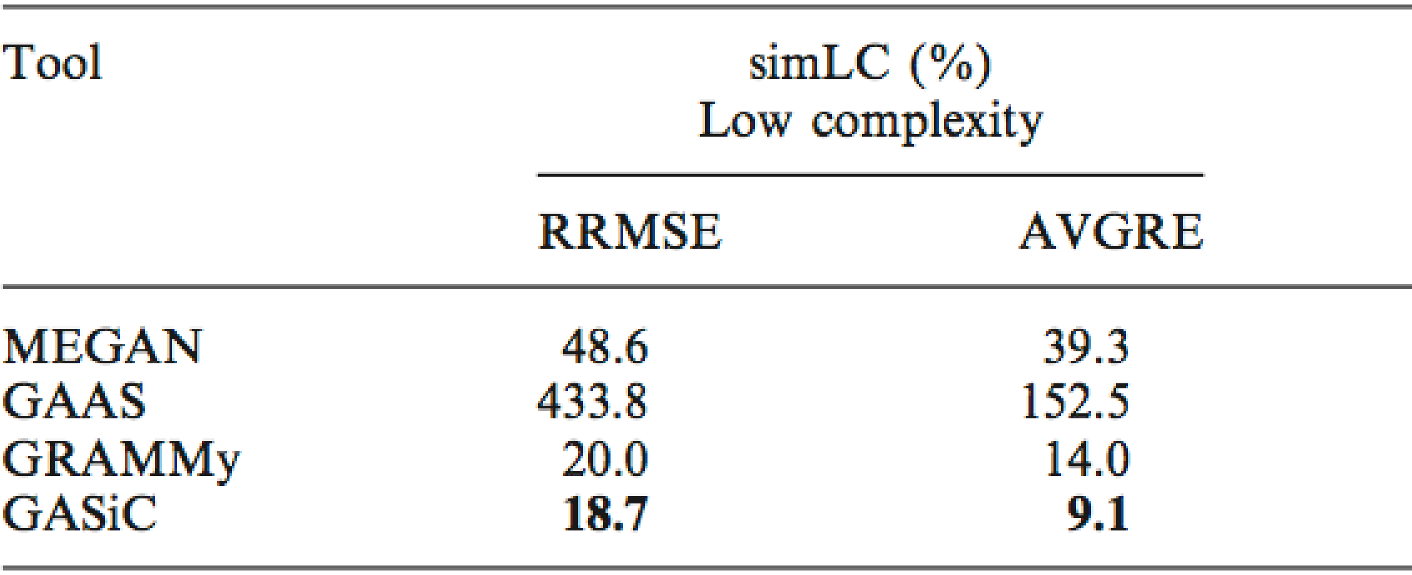
$$RRMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\frac{\mid c_j - t_j \mid}{t_j}\right)^2}$$

and

$$AVGRE = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{\mid c_j - t_j \mid}{t_j}\right)$$

Given the true abundances t_i RRMSE measures the sum of squared relative errors, whereas AVGRE is the sum of absolute relative errors. Thus, RRMSE is more sensitive to outliers.

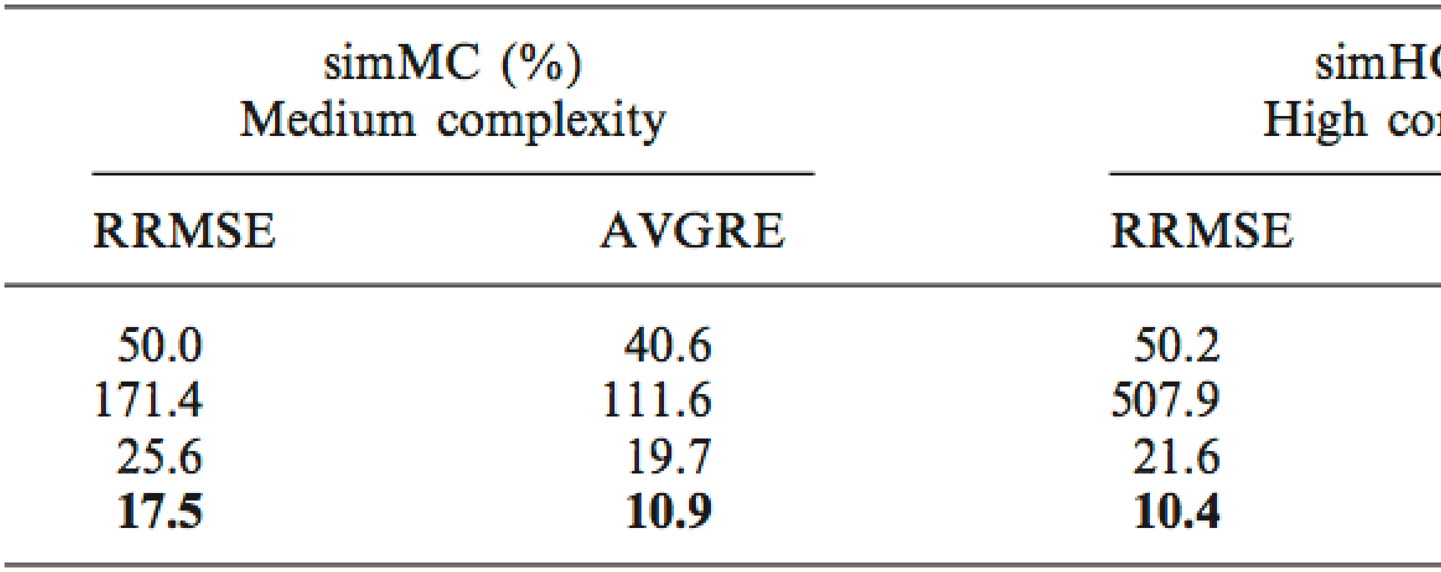
Lets first have a look at the low complexity data set.



Tool	simLC (%) Low complexity	
	RRMSE	AVGRE
MEGAN	48.6	39.3
GAAS	433.8	152.5
GRAMMy	20.0	14.0
GASiC	**18.7**	**9.1**

We can see that GASiC reduces the both error measures in estimating the abindances of the genomes.

The effect becomes more pronounced when going to the medium and high complexity samples.



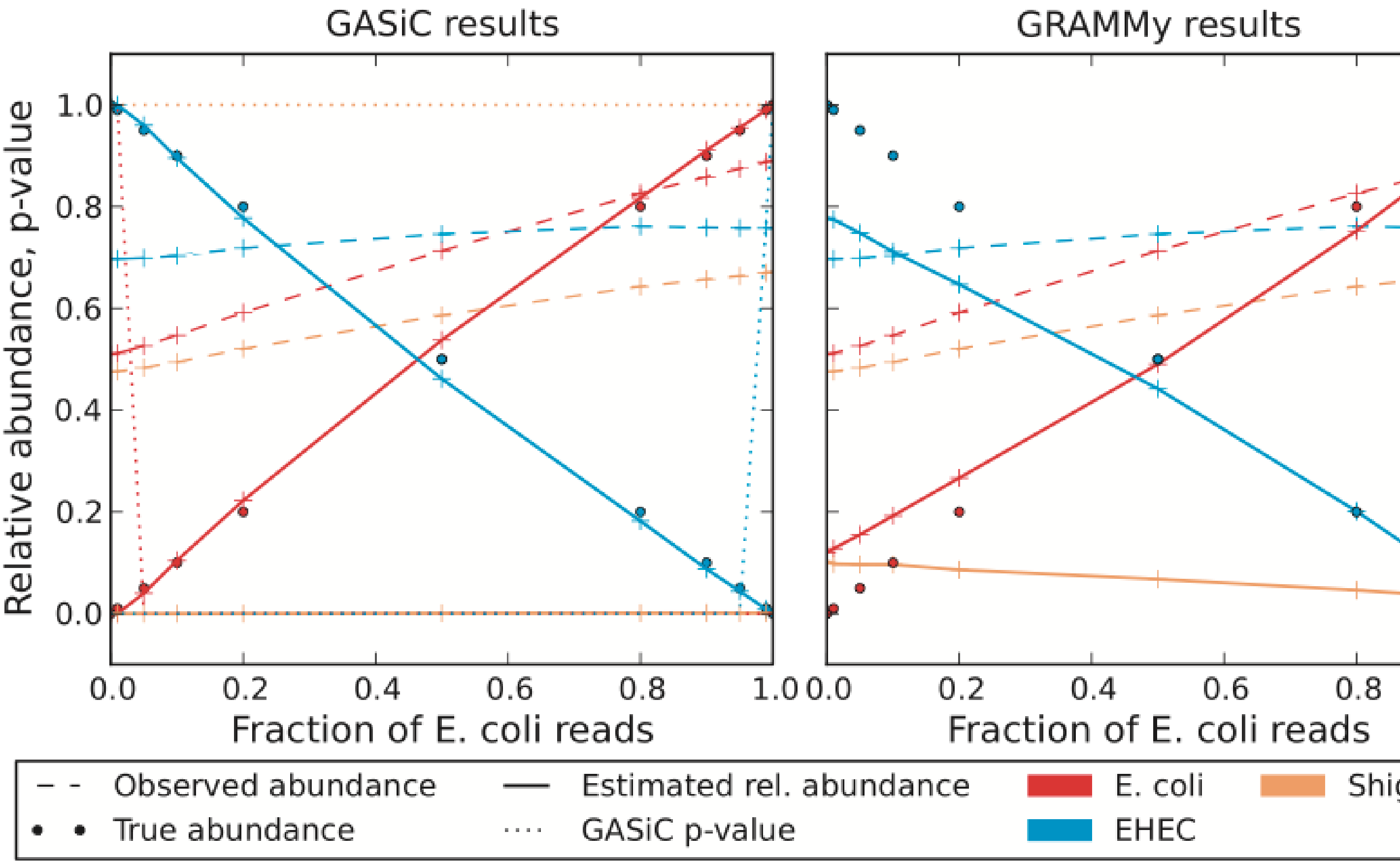
simMC (%) Medium complexity		simH High co
RRMSE	AVGRE	RRMSE
50.0	40.6	50.2
171.4	111.6	507.9
25.6	19.7	21.6
17.5	**10.9**	**10.4**

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.

11.9 Evaluation - Experiment 2

Here two *E. coli* genomes (E. coli DH10B and E. coli TY-2482) data sets 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*. Hence they 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, they included *Shigella flexneri* as phantom reference. Hence they 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.



Above you see results 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 concentra- tions.

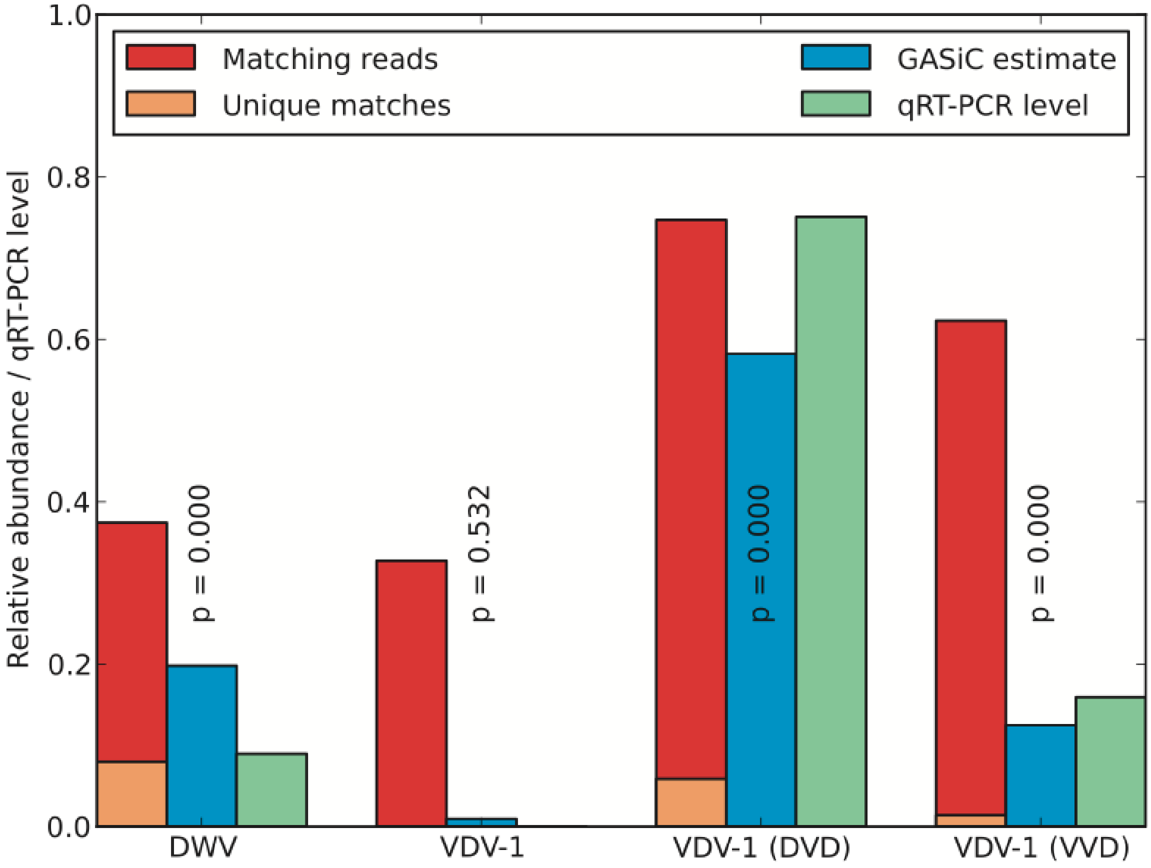
For Shigella, GRAMMy incorrectly estimates abundances up to 10%. GASiC provided more stable abundance estimates at all concentrations and also correctly identified Shigella as not present in the dataset and accordingly assigned high P-values

11.10 Evaluation - Experiment 3

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. 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-1DVD and VDV-1VVD, were published such that both the original and the recombinant sequences were available.

They 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.



GASiC's estimates are shown above demonstrating that the high sequence similarities caused strong corrections to the number of matching reads. After correction, VDV-1DVD 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. Moore et al. also found no evidence for VDV-1 and measured significant levels of VDV-1DVD in all examined 25 bee pupae. DWV was found in 23 of 25 pupae, but at lower levels than VDV-1DVD, and VDV-1VVD was found in 15 of 25 pupae.

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% abundance for VDV-1. GRAMMy estimates DWV as most abundant virus, whereas the other methods identify VDV-1DVD 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%.

11.11 Conclusion

The proposed method is a 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.