BIOINFORMATICS ORIGINAL PAPER
Vol. 29 no. 10 2013, pages 1260–1267
doi:10.1093/bioinformatics/btt147

Sequence analysis
Advance Access publication April 14, 2013

Analyzing genome coverage profiles with applications to quality control in metagenomics

Martin S. Lindner[1], Maximilian Kollock[1,2], Franziska Zickmann[1] and Bernhard Y. Renard[1,*]
[1]Research Group Bioinformatics (NG4), Robert Koch-Institut, 13353 Berlin, Germany and [2]Institute for Mathematics, Freie Universität Berlin, 14195 Berlin, Germany

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Genome coverage, the number of sequencing reads mapped to a position in a genome, is an insightful indicator of irregularities within sequencing experiments. While the average genome coverage is frequently used within algorithms in computational genomics, the complete information available in coverage profiles (i.e. histograms over all coverages) is currently not exploited to its full extent. Thus, biases such as fragmented or erroneous reference genomes often remain unaccounted for. Making this information accessible can improve the quality of sequencing experiments and quantitative analyses.

Results: We introduce a framework for fitting mixtures of probability distributions to genome coverage profiles. Besides commonly used distributions, we introduce distributions tailored to account for common artifacts. The mixture models are iteratively fitted based on the Expectation-Maximization algorithm. We introduce use cases with focus on metagenomics and develop new analysis strategies to assess the validity of a reference genome with respect to (meta-) genomic read data. The framework is evaluated on simulated data as well as applied to a large-scale metagenomic study, for which we compute the validity of 75 microbial genomes. The results indicate that the choice and quality of reference genomes is vital for metagenomic analyses and that validation of coverage profiles is crucial to avoid incorrect conclusions.

Availability: The code is freely available and can be downloaded from http://sourceforge.net/projects/fitgcp/.

Contact: RenardB@rki.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 31, 2013; revised on March 21, 2013; accepted on March 22, 2013

1 INTRODUCTION

Genome coverage, the number of sequencing reads mapped to a specific position within a reference genome, contains valuable information about reference genome and the mapping process and is easily accessible. Therefore, it is frequently consulted in bioinformatics analyses to improve decisions in algorithms or to provide meaningful information to the user. For instance, experimental design methods (Löwer *et al*., 2012) guide the experimentalist to achieve a specific average sequencing depth (i.e. genome coverage). After sequencing, the obtained reads can be mapped to a reference genome. Quality control tools (DeLuca *et al*., 2012; García-Alcalde *et al*., 2012) analyze the mapping data and report measures such as coverage information, mapping quality or error rate to the user. For example, Qualimap (García-Alcalde *et al*., 2012) visualizes the coverage profile and the coverage over the whole genome together with the GC content, which allows detecting biases in the sequencing process. If no reference genome is available, the reads can be assembled to complete genomes or at least longer contiguous sequences (contigs). The latter is nowadays possible for metagenomic data, i.e. datasets containing reads of many different species with different abundances. The assembler MetaVelvet (Namiki *et al*., 2012) uses the coverage information in the de Bruijn graph to connect contigs of similar coverage, as they are more likely to belong to the same organism. In addition to these examples, local coverage information is also used for detecting copy number alterations in genomes (e.g. Miller *et al*., 2011).

Despite these versatile applications of genome coverage, a vast amount of information commonly remains unused. Most current methods either use the average coverage over a certain sequence (DeLuca *et al*., 2012; Löwer *et al*., 2012) or describe the coverage profile using single probability distributions such as the negative binomial (Miller *et al*., 2011) or gamma (Hooper *et al*., 2010) distribution. Yet, to the best of our knowledge, more complex models such as mixtures of distributions are not used to fit genome coverage profiles (GCPs). Here, we suggest that more complex models can improve current methods and can open doors for new analysis strategies.

We see one application of complex coverage distribution models in metagenomics, where reference-based methods have become increasingly popular with the advent of high-throughput sequencing technologies (Mande *et al*., 2012). However, there are two major problems with reference genomes. First, the process of assembling and finishing reference genomes is time consuming and cumbersome and many reference genomes remain unfinished in the draft stage with varying qualities depending on the used sequencing technologies (Mavromatis *et al*., 2012). Draft genomes are typically a set of assembled contigs, where many contigs may be erroneous or, if assembled from metagenomic data, belong to different organisms. The second problem is of biological nature; evolution in the microbial world proceeds at high pace due to short replication times, and new subtypes or even species emerge perpetually. This causes different microbial species to have high genomic similarities. Therefore, the coverage is generally far from homogenous when mapping metagenomic reads to a reference genome; describing it with a single uni-modal

*To whom correspondence should be addressed.

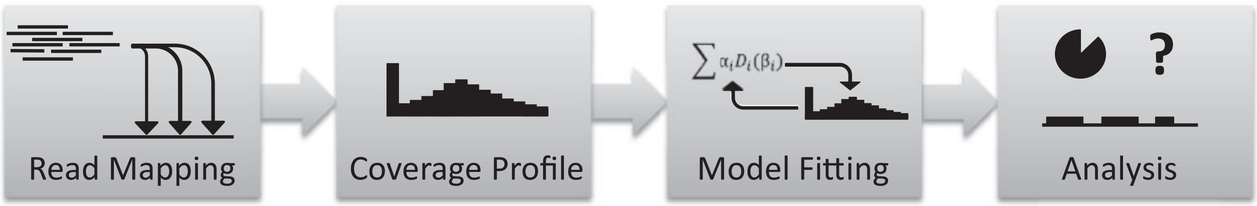

Fig. 1. Method overview. Starting with a set of reads mapped to a reference genome, we construct a GCP and fit a mixture of probability distributions to the profile. This procedure is the basis for further analysis steps concerning the reference genome, the mapping process and the read dataset

distribution would not be appropriate. Here, more complex models can have the power to disentangle and quantify different contributors to the genome coverage.

In this article, we present a framework for fitting complex mixtures of probability distributions to GCPs. We demonstrate in simulated experiments that the proposed framework can produce reliable and robust results and present a real data experiment, where we use our framework to reanalyze the data presented in a large-scale metagenomic study.

2 METHODS

The proposed method has four steps (Fig. 1). After mapping a set of sequence reads to a reference genome, the mapped reads are analyzed and a GCP is constructed. Then, a mixture model of customized probability distributions is fitted to the profile using an iterative procedure that is similar to the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977), seeking to identify and distinguish different contributors in the profile. Further analysis on the fit parameters can then be used to answer questions about the reference genome and the mapping process, such as the validity of a reference genome or the occurrence of multiple related organisms in one dataset. The presented method is not a new invention in itself, it is rather a combination of established statistical methods, which we demonstrate to be useful for analyzing GCPs. The novelty of this contribution is the composition of the mixture models and the subsequent analysis steps.

2.1 Genome coverage profiles

When reads are mapped to a reference genome, the per-base coverage for each position in the reference genome is given as the number of reads covering that position. We term the *histogram* over all per-base coverages the GCP. A GCP encapsulates valuable information about the relation between the reference genome and the genome(s) contained in the dataset. In the following, we solely operate on GCPs, as they provide a condensed view on the mapping of reads to a reference genome.

A GCP can take shape in various ways: First, if the reference genome matches perfectly to the reads contained in the dataset, the genome is homogeneously covered and the GCP consists of a uni-modal distribution, as depicted in Figure 2a. In reality, the reads and the reference genome will differ owing to mutations and errors in the reads. Therefore, the differing parts of the reference genome will not be covered by reads and lead to an excess of zero-coverage counts in the GCP, as shown in Figure 2b. Note that the distribution has a tail at low coverages, which we discuss in Section 2.5. As a third type (shown in Fig. 2c), a reference genome may have an overall low coverage as well as positions differing from the reads. Then, positions with zero coverage may be caused either by a locally differing sequence or the position was not covered by chance owing to the statistical fluctuations in the coverage. In addition to the three simple types, a GCP can also be a more complex combination of coverage distributions, as shown in Figure 2d. In this example, the dataset contained two genomes, A and B, with differing coverages. Both genomes share parts with the reference genome and also have similarities among each other.

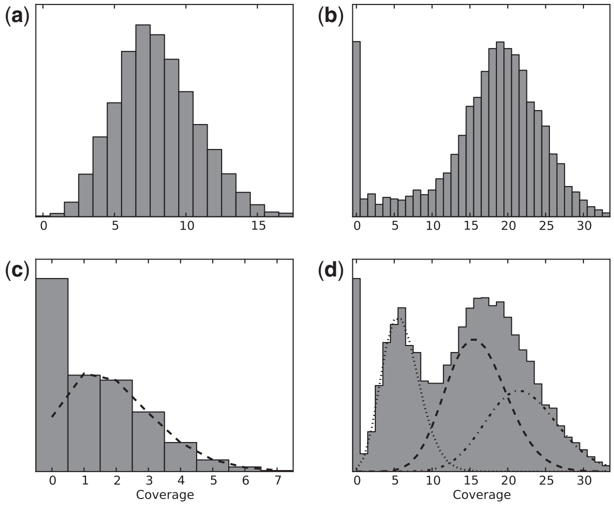

Fig. 2. Exemplary GCPs after mapping a set of reads to a reference genome. (**a**) Reads matching perfectly to reference genome. (**b**) Reference genome partially covered by reads; the covered areas have high coverage. (**c**) Reference genome partially covered by reads with low coverage. The dashed curve has a non-zero value at zero coverage and thus adds to the number of positions that were not covered due to disagreement with the reads. (**d**) Reference genome partially covered by reads from two organisms A and B with different abundances, yielding a mixture of four distributions: (i) a zero-distribution for the parts of the reference not covered, (ii) the coverage caused by organism A (mean = 6×, dotted curve), (iii) the coverage caused by organism B (mean = 16×, dashed curve) and (iv) the coverage where A and B map to the same position (22×, dash-dotted curve)

2.2 Genome coverage distributions

In this section, we give a short overview of probability distributions, which we consider relevant for describing GCPs. The simplest assumption we can make is the *random sampling* property of shotgun sequencing devices, meaning that we assume a uniform distribution of the reads over the genome. When reads are mapped to a genome, the coverage of each position follows a *Poisson* distribution $P(x|\lambda)$. The Poisson distribution is well studied and has one parameter λ, which simplifies fitting observed distributions. However, the Poisson distribution is often too narrow for fitting real genome coverage distributions, in particular for metagenomic data. This effect is called over-dispersion and occurs frequently in biological data. A common way (Bliss and Fisher, 1953) to account for over-dispersion is assuming that the Poisson parameter λ is distributed according to a second distribution. When λ is assumed to be gamma distributed, we obtain a *negative binomial* distribution $NB(x|a, b)$, which has two shape parameters and is thus harder to fit than the Poisson distribution. However, the negative binomial turned out to model GCPs well for both high and low coverages and has been used, for example, in differential expression analysis (Anders and Huber, 2010).

In Figure 2b, we observed a tail on the low-coverage end of the main distribution. The magnitude of the tail depends on the fragmentation of the reference genome with mutations, as we discuss in detail in Section 2.5. The shape of the tail is determined by the original parent distribution; therefore, the tail distribution does not have its own shape parameters and can rather be considered as an extensional distribution. We implemented the *Poisson tail* and the *negative binomial tail* in our framework and give a detailed mathematical description in the Supplementary Text.

Finally, we introduce the *zero* distribution $z(x)$, which is useful to describe the excess of zero coverage positions. The zero distribution has probability 1 at zero, and 0 everywhere else. It is static as it has no shape parameters, but proves its usability in combination with other distributions. The zero-inflated Poisson and zero-inflated negative binomial are defined as mixed distributions of zero and Poisson or zero and negative binomial. These zero-inflated (ZI) distributions were used, for example, to model the number of defects in manufactured items (Lambert, 1992), but can also be applied for GCPs: the areas where the reference genome agrees with the reads in the dataset yields a coverage distribution according to Poisson or negative binomial, the areas with disagreement are modeled by the zero distribution.

We want to regard mixtures consisting of more than one probability distribution and write the joint distribution function as follows:

$$f(x, \alpha|\beta) = \alpha_0 \cdot z(x) + \sum_{i=1}^{k} \alpha_i \cdot D_i(x|\beta_i) \qquad (1)$$

where k is the number of non-zero distributions, α_i are the non-negative mixing coefficients that sum to 1 and give a weight to each distribution, and D_i is a Poisson, negative binomial or tail distribution with the corresponding set of parameters β_i. Fitting these mixture models to data cannot be done directly, but requires an iterative method, as described in the following section.

2.3 Iterative algorithm for fitting the mixture model

The algorithm starts with a set of initial parameters for the distributions D_i that can either be user defined or estimated from the GCP. With every iteration, the algorithm adjusts the parameters to increase the likelihood of the data given the mixture model in Equation (1). The iteration is stopped when an accuracy threshold is reached, e.g. the change of the likelihood drops below a predefined value. The algorithm repeatedly computes the so-called expectation step followed by an adjustment of the parameter set to improve the accuracy with which the corresponding model describes the data, i.e. the likelihood of the data given the model with parameters after iteration $t+1$ should be greater or equal to the likelihood after iteration t. This assumption is guaranteed if maximum likelihood estimation is used when adjusting the parameter set. In this case, the procedure is known as the EM algorithm (Dempster *et al.*, 1977).

2.3.1 Expectation step Following the initialization, the expectation step estimates conditional probabilities identical to the EM algorithm. Using the current set of parameters $\beta^{(t)}$, we compute for each observed coverage value the probability of belonging to distribution D_i. With these coverage-wise probabilities, we re-estimate the mixing coefficients α_i for each distribution D_i. See Supplementary Text for more details.

2.3.2 Parameter estimation step In the second step, we optimize the parameter set $\beta^{(t)}$ by fitting the mixture model with respect to the previously calculated mixing coefficients $\alpha^{(t)}$. When fitting the distributions D_i, we have to decide whether to use moment-based or maximum likelihood estimates. Using the method of moments, for 1-(2-) parametric distributions, we take the sample mean (the sample mean and the variance, respectively) and calculate the distribution parameters from these moments. Maximum likelihood estimation directly selects a set of parameters $\beta^{(t)}$ that maximizes the likelihood. Owing to the nature of our data there is no ultimate solution. Either method might be more suitable depending on the situation. Yet, for the negative binomial distribution, there is no closed form of the maximum likelihood estimator and requires application of, for example, Newton's method. The method of moments proves to yield similar results for the negative binomial distribution and is numerically more robust. The zero and the tail distributions do not require parameter estimation as the former has no shape parameters and the latter inherits the parameters from the parent distribution.

2.4 Genome–Dataset Validity score

The standard scenario is that we have one reference genome and a set of genomic reads originating from one or more unknown genome(s). Then, we define the Genome–Dataset Validity (GDV) score as the fraction of the reference genome that has a counterpart in the genome(s) underlying the reads. The true similarity of the two sequences is not directly observable, as the unknown genome is realized as a set of short reads.

The naïve way to estimate the similarity of both sequences is by mapping the reads to the genome and measuring the fraction of the genome that was covered by reads. This estimate can be sufficiently good for high genome coverages, such as coverages above 10×. Here, the likelihood that a location shared between both genomes remains uncovered is negligibly small. Almost all sites on the known genome not covered by reads can be considered to be different from the unknown genome. In contrast, for low abundances, the probability that a base is not covered by reads although it is shared by both genomes can not be neglected anymore. Assume, for example, a simple model where the base coverages over the genome follow a Poisson distribution. While the probability of not covering a base at 10× coverage is 0.0045%, it rises to 13.5% for 2× coverage and 36.8% for 1× coverage.

The described iterative algorithm can improve on the naïve approach and provide reliable estimates for much lower coverages. Depending on the coverage distribution, we can fit a mixture of a Poisson or negative binomial distribution and a zero distribution to the GCP. The contribution of the zero distribution should then roughly correspond to the fraction of the reference genome that has no counterpart in the unknown genome(s). Therefore, we calculate the GDV score as follows:

$$GDV = 1 - \alpha_0.$$

This calculation has a clear advantage over the naïve approach: at low coverages, the probability that a position is not covered by chance (and not due to dissimilarity) is high, and the naïve approach is at risk of overestimating the fraction of the genome with no counterpart. In contrast, the mixture model approach makes use of the positions with higher coverage to estimate the probability of obtaining zero coverage by chance and thus provides more realistic and more reliable estimates.

2.5 Genome fragmentation estimation

In addition to the pure similarity of the sequencing reads and the reference genome, we can also question how fragmented the reference genome is with respect to the reads. With fragmentation, we mean the number of contiguous sequence fragments in the genome that conform with the reads. For example, single-nucleotide polymorphisms, insertions or deletions in the reads with respect to the reference genome can be the cause for genome fragmentation. As discussed in detail in the Supplementary Text, the effects at the fragment borders manifest in a tail structure in the GCP. The tail distributions can be included in the mixture model in

Equation (1) and their mixing coefficients can be estimated with the described iterative algorithm. The size of the tail (i.e. α_{Tail}) allows estimating the genome fragmentation directly, as the size of the tail depends on the number of fragment borders. With the genome length L and the read length RL, the number of fragments can be estimated as follows:

$$Frag \approx \frac{\alpha_{Tail} \cdot L}{2 \cdot RL}. \quad (2)$$

A high fragmentation of the genome gives rise to additional correction terms for models with zero inflation. Due to the excess of zero coverage positions at fragment borders, which is most pronounced for genomes with partially non-zero low ($1\times - 10\times$) coverage and zero coverage elsewhere, we introduce a correction term z_{corr} for the mixing coefficient of the zero component of the model. The correction term is a function of the number of fragments in the genome $Frag$, the genome length L and the average delay $\bar{d}$ of the first read mapped behind a fragment border:

$$z_{corr} = \frac{2 \cdot Frag \cdot \bar{d}}{L}.$$

We provide a detailed derivation of the number of fragments and the correction term in the Supplementary Text.

3 EXPERIMENTS AND RESULTS

The framework described in the previous section provides a powerful tool for solving problems related to reference genomes and genome coverage distributions. Here we present three experiments that demonstrate the applicability of the framework in a metagenomic context. In the first experiment, we demonstrate that the proposed algorithm can fit complex mixtures of distributions to GCPs and evaluate the influence of the choice of model on the fit quality. In the second experiment, we demonstrate the robustness of the framework: quantities calculated from the GCP fits (here: GDV scores) are stable over a wide range of genome coverages. This is crucial in metagenomics, where the number of mapped reads per genome is typically small, but can be high for single abundant species. In the third experiment, we apply the framework on real data and thereby illustrate a further application: we reanalyze data from a large-scale human gut metagenomic study and compute the GDV scores for a selected set of reference genomes.

3.1 Fitting complex mixture models

In this experiment, we evaluate the performance of the algorithm for fitting complex mixture models to multi-modal GCPs. Thus, we created a dataset with reads of two organisms sharing large genomic regions, *Escherichia coli* and *Shigella boydii*. We simulated 100 000 reads for *E.coli* and 600 000 reads for *S.boydii* with 75 bp length and Illumina sequencing characteristics using the Mason read simulator (Holtgrewe, 2010). These reads were then mapped to the *E.coli* reference genome with Bowtie (Langmead *et al*., 2009). We expected the genome to be homogeneously covered by the *E.coli* reads and locally by additional *S.boydii* reads. Yet, the number of *E.coli* reads could only account for 1.5× coverage. This challenged the algorithm in two ways. First, the low *E.coli* coverage caused a fraction of genome positions to have zero coverage; yet, they should not be explained by a zero distribution. Second, the *S.boydii* fragments with high coverage produced a tail in the GCP, which overlapped with the *E.coli* distribution. For fitting, we used models consisting of three components: (i) a zero distribution (abbreviated by z), (ii) a Poisson (p) or negative binomial (n) distribution for the *E.coli* reads and (iii) a Poisson, negative binomial, Poisson with tail (pt) or negative binomial with tail (nt) distribution for the *S.boydii* reads. The initialization was chosen such that component (ii) fitted the *E.coli* peak and (iii) fitted the *S.boydii* peak.

All models were fitted to the GCP using an accuracy threshold of 0.1% and the zero-correction was calculated for the models with tail distribution. To compare the models by numbers, we calculated the Kolmogorov–Smirnov test statistic, the maximum absolute difference $d_{\max}$ between the observed and the estimated cumulative mass function.

Figure 3 depicts the fits of selected mixture models, and detailed results about the mixing coefficients and fit errors are provided in the Supplementary Table. The results show a prominent difference between models with and without tail: models with tail fit the observed GCP much better (average $d_{max} = 0.0022$) than the models without tail (average $d_{max} = 0.0073$). The simplest model, zpp (see Fig. 3a), yields the highest fit error of all models ($d_{max} = 0.0141$). For the models without tail, the fit error decreases as the model complexity (parameters to fit) increases. The difference of the fit error between the models with tail is overall lower than between the models without tail: the lowest fit error is achieved by zpnt ($d_{max} = 0.0018$), the highest by znnt ($d_{max} = 0.0026$). In particular, the fit error does not decrease with increasing model complexity. Furthermore, the model fits with tail are highly similar: besides the similar fit error, they also have almost identical mean values μ for the two non-zero distributions [distribution (ii): $1.57 < \mu < 1.60$; distribution (iii): $10.54 < \mu < 10.57$].

The relative sizes of the tail distributions are on par with the other distributions, indicating a high degree of fragmentation of the *E.coli* genome compared with *S.boydii.* The number of *S.boydii* fragments in *E.coli* can be estimated via Equation (2); depending on the model, there are between 8032 and 8289 *S.boydii* fragments in the *E.coli* genome. The contribution of the zero distribution is estimated to exactly zero in all models except zpp and zpn.

Further experiments using more complex models (e.g. znnnt) do not reduce the fit error. The spare distributions either take the same shape as one of the two original distributions or their mixing coefficients are reduced to zero, depending on the start parameters.

This experiment shows that our algorithm can fit complex mixture models to GCPs accurately. Best results are obtained when the complexity and the selected distributions in the model match the data, but more complex models do not decrease accuracy and should thus be chosen in doubt. The low fit errors of the models with tail distribution support the usefulness of the tail distribution concept. Although our iterative algorithm is not guaranteed to converge to an optimal solution as EM does, we see that the fit results are highly similar, in particular for the models with tail.

3.2 Influence of average coverage

In this experiment, we demonstrate the robustness of our framework over a wide range of genome coverages. Information about the genomes—both the source of the reads and the

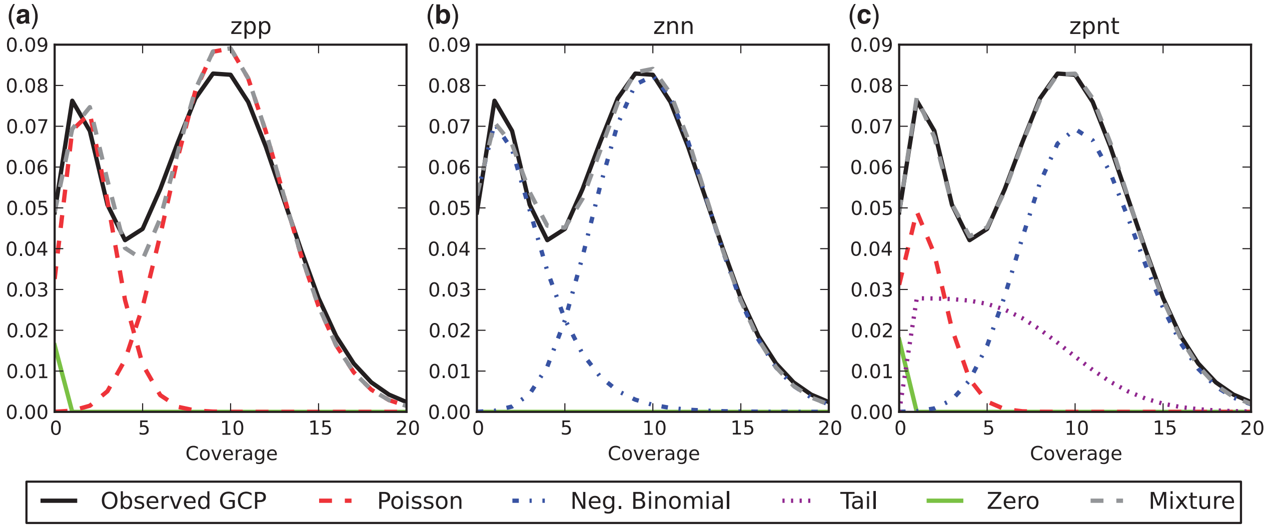

Fig. 3. Influence of the choice of mixture model for fitting GCPs. Three exemplary models are shown: (**a**) zero and two Poisson distributions (zpp), (**b**) zero and two negative binomial distributions (znn), (**c**) zero, Poisson and negative binomial distribution with tail (zpnt). Model (c) yields the lowest fit error ($d_{max} = 0.0018$), but is more complex than models (a) and (b). Model (a) has the lowest complexity, but yields the highest fit error ($d_{max} = 0.0141$)

reference—derived from GCPs should generally not be affected by the overall number of reads mapped to the reference genome.

We used the *Shigella flexneri* genome as reference and simulated datasets of short (75 bp) Illumina reads from the *E.coli* genome with Mason. The smallest dataset contained 1000, the largest 10 Million *E.coli* reads. We used Bowtie to map these datasets to the *S.flexneri* reference genome and fitted ZI Poisson and ZI negative binomial models, both with and without zero correction, to the GCPs. The GDV score was calculated for each model based on the fit parameters as described in Section 2.4. The true GDV score was estimated from the dataset with 10 Million simulated *E.coli* reads to be 0.826. Due to the high coverage, at least one read starts at each position in the *E.coli* genome, and a mapping should therefore cover all fragments in the *S.flexneri* genome that are identical with *E.coli* and at least 75 bp long.

The estimated GDV score of *S.flexneri* for the *E.coli* reads is summarized in Figure 4 and Supplementary Figure S3. Curve (a) shows the estimated GDV score for the ZI Poisson mixture models (without and with zero correction), (b) for the corresponding ZI negative binomial models. The estimated GDV score of the *S.flexneri* genome is close to the estimated true GDV score (gray dotted line) for all mixture models when the number of reads is above 1 Million. In the range from 100 000 to 1 Million reads, the ZI Poisson with correction yields the best estimates; it keeps the estimates on an almost constant level. For the ZI negative binomial model, the correction has a much smaller influence and the estimates are slightly worse than the corrected Poisson model. In the low-coverage regime (below 1.5× coverage or 100 000 reads), the tail effect is not observed anymore and all models reduce to the ZI Poisson and ZI negative binomial, respectively. The Poisson model yields lower estimates as the number of reads decreases and therefore becomes increasingly unreliable. On the other hand, the negative binomial model yields relatively good estimates down to low numbers of reads (~10 000), which corresponds to only 0.16× coverage in the covered fragments.

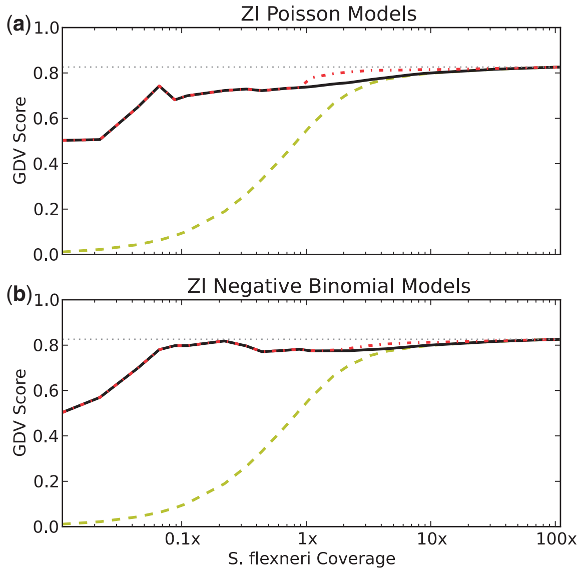

Fig. 4. Estimating the GDV score. The charts show the estimated fraction of the *S. flexneri* genome that is similar to the *E.coli* genome, i.e. the GDV score, depending on the average coverage of the *S.flexneri* genome with *E.coli* reads. We used ZI Poisson (**a**) and negative binomial (**b**) mixture models (see text). Each model was fitted without (black solid line) and with (red dash-dotted line) zero correction. The yellow dashed line is the fraction of the genome that was covered by reads. The gray dotted line is the estimated true GDV score

The results of this experiment suggest two different strategies for the selection of the mixture models: for coverages below 1×, the plain ZI negative binomial distribution yields the best results and allows determining the GDV score with acceptable accuracy. For local coverages of 2× and above, the ZI Poisson model with zero-correction produces highly accurate estimates and outperforms all other models. There, the advantage of the Poisson model is 2-fold: in addition to the better estimates, the Poisson model has one parameter less than the negative binomial model, and parameter estimation is faster. Conclusively, we see that the estimated GDV score is largely independent of genome coverage and estimation is possible even below 1× coverage.

3.3 Application: metagenomics

The validity of reference genomes is crucial for a sound interpretation of the data, in particular in metagenomics. Thus, we estimate GDV scores on real metagenomic data. The work by Qin *et al.* (2010) serves as a test case; they sequenced the metagenomic communities in faecal samples of 124 European individuals on the Illumina platform (75 bp paired end reads) and conducted exhaustive analysis to provide insight into the composition of genes and bacterial species in the human gut. As one result, they report a list of 75 prevalent bacterial species, the *common core*, which were present (genome coverage >1%) in a large number of individuals. We obtained the original reference genomes of the common core and selected 17 genomes that were originally found in all 124 individuals with at least 1% coverage. The metagenomic reads of individual MH0012 were downloaded from the corresponding EBI database (accession numbers: ERX004076–ERX004082).

In this experiment, we estimated the GDV score of the selected reference genomes with respect to the metagenome of individual MH0012. The 93 Million paired-end Illumina reads were mapped to the selected reference genomes using Bowtie 2 (Langmead and Salzberg, 2012), and the coverage profile was calculated subsequently for each genome. We used Bowtie 2 here, as the metagenomic data and the quality of the reference genomes requires a higher tolerance toward mismatches between reads and reference, which cannot be accomplished with Bowtie. In the next step, we fitted a mixture model of a zero distribution, two negative binomial distributions (with maximum likelihood estimation), and a negative binomial tail distribution to the GCPs. We preferred the negative binomial over the Poisson distribution here, as we expected over-dispersion due to a high biological variability in the metagenomic data. Two negative binomial distributions were chosen with genomic similarities in mind, where one distribution should fit the matches from the correct species and the other distribution should account for the noisy matches obtained by organisms with partial sequence similarity. The fit error was calculated as in the first experiment.

The run time of the framework was measured for the 17 selected reference genomes. Fitting the model to the GCP was accomplished on average within 2.9 s (minimum: 2.1 s, maximum: 4.5 s). Another 1–3 min need to be added for calculating the GCP from the SAM file before fitting. In general, we observe that the run time is strongly correlated to the number of matching reads, i.e. the genome coverage, and the complexity of the model.

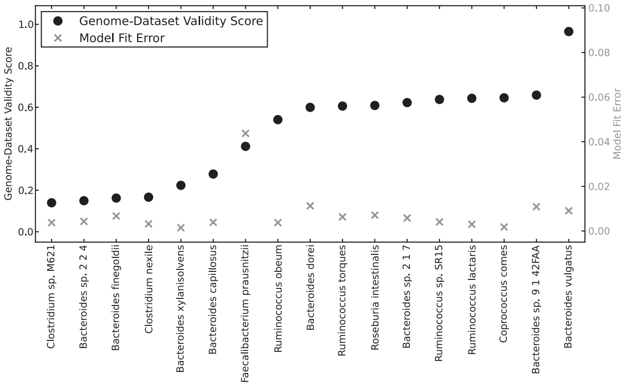

Fig. 5. GDV scores for human gut metagenome. The validity of 17 prevalent bacterial species with respect to one metagenomic human gut sample was estimated by fitting the GCPs as described in Section 2.4. The GDV score (dots) ranges from 0.140 for *Clostridium sp. M621* to 0.965 for *B.vulgatus*. A lower fit error (crosses, right axis) indicates a more trustworthy GDV score estimate

The GDV score of the 17 selected reference genomes is shown in Figure 5 and ranges from 0.140 (*Clostridium sp. M621*) to 0.965 (*Bacteroides vulgatus*). All genomes have a moderate average coverage (minimum 8×, maximum 47×), and the coverage has only low correlation to the GDV score (Pearson correlation coefficient: $r = 0.34$). The fit error is below 0.02 for all genomes except *Faecalibacterium prausnitzii*. Despite the high-error level compared with the first experiment, manual inspection of the fits indicates that the assumed model is sufficient for the complexity of data. Manual inspection of *F.prausnitzii* showed that the GCP was too complex for the assumed model and required an additional negative binomial component.

Considering that *E.coli* reads mapped to a *S.flexneri* genome yield a GDV score of >0.8, as demonstrated in the previous experiment, the numbers observed in this experiment are rather low. One reason may be that the selected reference genomes originally served as representatives for clusters of similar genomes. Furthermore, most reference genomes were sets of separate contigs, indicating that the reference genomes could be incomplete or have low quality. This is prototypic for metagenomics, as the majority of bacteria are still not or only poorly sequenced, such that a reference genome with low GDV score may be not a good, but the best possible choice. The only high-quality reference genome (no contigs) is *B.vulgatus*, which achieves by far the highest GDV score (GDV = 0.96).

A similar picture can be observed on the full set of 75 genomes (see Supplementary Figure); the GDV scores are in the range from 0.013 (*Enterococcus faecalis*) up to 0.998 (*Clostridium leptum*). Interestingly, four out of the seven best scored genomes are high-quality genomes and the other three have <100 contigs. Manual validation confirmed the high validity and showed homogenous coverage over the genomes, only interrupted by small gaps and single high-coverage parts. On the other hand, the genome with the worst score—the gut bacterium *E.faecalis*—is also a high-quality genome. Manual validation showed that *E.faecalis* was not covered homogeneously. This underlines the features of the GDV score: to achieve a high validity, reference genomes must be homogeneously covered by reads and must have high quality. A high genome quality but a low and inhomogeneous coverage, indicating that the species itself is not present, is correctly penalized by a low score.

4 DISCUSSION

We introduced GCPs as a means to extract quantitative information from mapping data. By fitting mixtures of probability distributions to the GCP, we obtain valuable information about the reference genomes and the mapping process, such as the fraction of the genome that could not be covered by reads or if there is more than one organism contributing to the coverage. This makes the proposed framework a powerful tool for the analysis of mapping data without restriction to the application.

The introduced GDV score is a simple, yet powerful, measure for how well a reference genome fits to the mapped reads. Especially in metagenomics, reference genomes are typically not required to fit perfectly to the data; nevertheless, the degree of divergence should not become too large. As one example, we observed a GDV score of 0.82 in the experiment in Section 3.2, where we mapped *E.coli* reads to a *S.flexneri* genome. This illustrates a relatively high biological divergence between data and reference despite a high GDV score. We assessed GDV scores in a real metagenomic experiment conducted by Qin *et al.* (2010) and observed surprisingly low scores for genomes that were originally considered to be present in the dataset; only 9 of 75 reference genomes achieved scores >0.8. This is an imposing example for high discrepancy between metagenomic data and reference genomes, which we presume to be a common challenge of metagenomic experiments. One of the major reasons might be the quality of the reference genomes: as microbes from metagenomic experiments are typically not cultivable, their genomes must be assembled from environmental samples, which is significantly more complicated and error prone than assembly from pure samples. In the experiment at hand, 37 of 75 reference genomes consisted of >100 (up to 1700) separate contigs, only six genomes were one contiguous sequence. The framework proposed and applied in this work makes these flaws quantifiable.

The first experiment showed that the iterative algorithm is able to fit complex mixtures of highly specialized probability distributions to GCPs. The impact of the tail distributions became apparent, as they significantly reduced the fit error. The second experiment showed that quantities calculated on fitted GCPs are robust toward influences of the average genome coverage. There, we observed stable estimates of the GDV score over a wide range of coverages, starting at average coverages below 0.2×. Although our method is robust toward the average coverage, it can be sensitive to the mapping parameters: more restrictive mapper settings typically yield a lower GDV score and a higher influence of the tail distributions. This has to be considered when comparing GDV scores over different experiments. The iterative algorithm encounters limitations in extreme cases, for example, when the average coverage is very low, but locally extremely high. This can be the case when a genome is not present in the data, but shares a gene with other highly abundant genomes. Then, the algorithm may fail to fit the low-coverage distribution, as intended by the user, but tries to fit the extremely high noise contributions. In other cases, the standard start parameters are inappropriate, such that the algorithm ends up in a local probability maximum instead of fitting the distribution as intended. These problems demonstrate that visual inspection of the fit is necessary and this is supported by the framework. Common strategies used for the EM algorithm are also possible, such as the initialization with different or manually determined starting parameters.

Here, we focused on applications in metagenomics; however, the information obtained by fitting the coverage distribution is by no means limited to metagenomics but can be used for other purposes, such as experimental design and sequencing depth estimation (Hooper *et al.*, 2010), the detection of copy number variations (Miller *et al.*, 2011) or metagenome assembly (Namiki *et al.*, 2012). As an example, metagenomic sequencing experiments can be designed in a way, such that the GDV score can be calculated robustly for reference genomes with a certain minimum abundance in the sample. The minimum amount of sequencing required can be found by finding the minimum required coverage for a robust GDV score calculation in a simulation-based experiment, as presented in Section 3.2. Tools for estimating species abundances in metagenomic data, such as GRAMMy (Xia *et al.*, 2011), GASiC (Lindner and Renard, 2012) or READSCAN (Naeem *et al.*, 2013), can make use of the GDV score to more precisely estimate the abundance of the organism truly contained in the dataset, if the used reference genomes have a low validity. Observations give rise to the assumption that our method is applicable to single contigs in unfinished reference genomes. There, the GDV score may support the identification of chimeric or erroneous contigs and thus contribute to the improvement of reference genomes. New applications arise, for example, in metagenomics where the information from the GCPs can be used to estimate the evolutionary distance of unknown organisms in the data to known organisms by mapping the reads to the known genomes and calculating GDV scores. In conjecture with phylogenetic information, the GDV score can be used to narrow the truly contained organism down to a certain area of a phylogenetic tree by excluding reference genomes yielding a lower GDV score.

5 CONCLUSION

Genome coverage information is commonly consulted for assessing the quality of mapping data. Yet, we argue that the coverage information is not sufficiently exploited and does allow a deeper evaluation of the mapping process and the suitability of reference genomes. Thus, we introduced GCPs as a powerful tool to extract valuable information about the interplay of read data and reference genomes and described an algorithm for fitting specialized probability distribution functions to GCPs. In simulated experiments, we showed that the proposed algorithm performs well on complex GCPs and over a wide range of genome coverages, including coverages as low as 0.2×. We demonstrated a use case of our framework in metagenomics, where we could apply our approach to quantify the discrepancy of metagenomic data and reference genomes. The observations suggest that the selection of reference genomes for metagenomic experiments should be done carefully and the validity of the reference genomes should be integrated in the further analysis.

ACKNOWLEDGEMENTS

We would like to thank Junjie Qin (BGI-Shenzhen) for help with metagenomic reference genomes and Martina Fischer and Robert Rentzsch (NG4, Robert Koch-Institut) for critical reading of the manuscript and useful comments.

Funding: This work was supported by the German Federal Ministry of Health [IIA5-2512-FSB-725 to B.Y.R.].

Conflict of Interest: none declared

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bliss,C. and Fisher,R. (1953) Fitting the negative binomial distribution to biological data. *Biometrics*, **9**, 176–200.

DeLuca,D. *et al.* (2012) RRNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.

Dempster,A. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–39.

García-Alcalde,F. *et al.* (2012) Qualimap: evaluating next generation sequencing alignment data. *Bioinformatics*, **28**, 2678–2679.

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.

Hooper,S. *et al.* (2010) Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics*, **26**, 295–301.

Lambert,D. (1992) Zero–inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Langmead,B. and Salzberg,S. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lindner,M. and Renard,B. (2012) Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.*, **41**, e10.

Löwer,M. *et al.* (2012) Confidence-based somatic mutation evaluation and prioritization. *PLoS Comput. Biol.*, **8**, e1002714.

Mande,S. *et al.* (2012) Classification of metagenomic sequences: methods and challenges. *Brief. Bioinformatics*, **13**, 669–681.

Mavromatis,K. *et al.* (2012) The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLoS One*, **7**, e48837.

Miller,C. *et al.* (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.

Naeem,R. *et al.* (2013) READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics*, **29**, 391–392.

Namiki,T. *et al.* (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Xia,L. *et al.* (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, **6**, e27992.