The ISME Journal (2014) 8, 2349–2351
© 2014 International Society for Microbial Ecology All rights reserved 1751-7362/14
www.nature.com/ismej

COMMENTARY

Estimating coverage in metagenomic data sets and why it matters

Luis M Rodriguez-R and Konstantinos T Konstantinidis

The ISME Journal (2014) **8,** 2349–2351; doi:10.1038/ismej.2014.76; published online 13 May 2014

A ‘metagenome’ is the theoretical collection of genomes from all members of a given microbial community, and a ‘metagenomic data set’ is the subset captured in a given sequencing event. Although these terms are often used interchangeably and metagenomic data sets are regularly called metagenomes by synecdoche, their relationship is analogous to sample and population in statistics. The fraction of the metagenome represented in the metagenomic data set, termed coverage (not to be confused with the repetition of features, termed sequencing depth), is of key importance in assessing statistical significance of features sampled (taxa, genes and so on). However, quantitative computational methods to assess the level of coverage are limited, a problem we have recently attempted to solve. In extreme cases, where small data sets are used to characterize complex communities, misleading inferences can arise. For instance, random variation can be frequently mistaken for real differences in comparisons of metagenomic data sets with extreme differences in coverage. Further, insufficient coverage also reduces the detection limits and statistical power of the comparisons, hiding real, ecologically relevant trends and differences (Figure 1). We demonstrate here how available solutions can determine the level of sequencing coverage obtained by metagenomic data sets and thus, guide their robust analysis and comparison.

One widely used qualitative method to estimate coverage is a rarefaction curve, sometimes also called a collector or complexity curve. This method relies on the observation that the curve of rarefied counts of any feature (for example, operational taxonomic units, named species, predicted genes, functional categories or even short motifs) should plateau if the sample is close to saturation. Use of rarefaction curves in microbial community studies was popularized by tools such as mothur (Schloss *et al.*, 2009) and recently extended to include accurate projections at higher sequencing efforts by preseq (Daley and Smith, 2013), which allows the estimation of coverage across features (arithmetic mean). However, this technique and others like it typically rely on a high-quality assembly, comprehensive reference data sets or both, which are often unavailable for complex or poorly characterized communities (with the probable exception of ribosomal RNA (rRNA) genes). Moreover, the preseq projection is optimized for single-species data sets and, therefore, does not scale for mixtures of species, making it insufficient for accurate estimations with complex metagenomic data sets. Without accurate projections, rarefaction curves can only be used to determine whether a data set is close to saturation, a useful but insufficient assessment of coverage. Performing this task with rRNA genes is also problematic; largely because their high sequence conservation frequently masks important levels of genetic and ecological differentiation among closely related organisms (Caro-Quintero and Konstantinidis, 2012).

Another approach is to estimate the coverage of one or a few target species in the metagenomic data set using simple statistical approaches such as the Lander–Waterman expressions (Lander and Waterman, 1988), while ignoring the remaining genomes of the community. Such methods are useful in studies targeting specific species in a community in order, for instance, to recover complete genomes. The main drawbacks of this approach include a lack of implemented software and the requirement of reliable estimates for genome size and abundance of the target species, which often poorly represent the community as a whole. No matter how limiting this approach may appear, it can be applied to many available metagenomic data sets, is based on robust statistical frameworks (Wendl *et al.*, 2012) and the interpretation of coverage is straightforward: breadth of the genome covered by sequencing reads.

Finally, genome-wide approaches that capitalize on community modeling and/or modeling of contig sequencing depth have been proposed (for example, Hooper *et al.*, 2010; Stanhope, 2010). Such approaches are independent of comprehensive reference databases, which broadens their applicability, but depend on assumed abundance (and genome size) distributions and high-quality assemblies. Moreover, no software has been available to facilitate their application to real metagenomes.

We recently presented Nonpareil (Rodriguez-R and Konstantinidis, 2013) as an alternative approach.

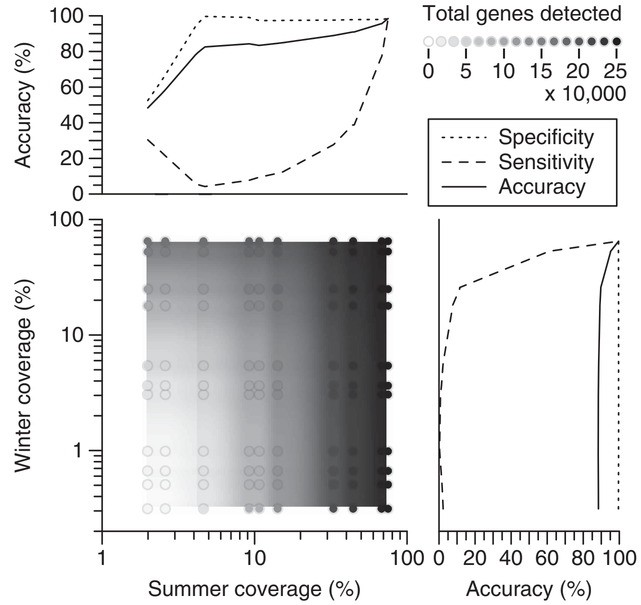

Figure 1 Effect of average coverage on detection of differentially abundant features. The abundance of nonredundant genes (assembled and clustered at 98% amino-acid identity) detected in the metagenomes of Lake Lanier (Atlanta, GA, USA; Sequence Read Archive Projects SRP028408, SRP005437-9; abundance estimates were based on read-mapping at 95% nucleotide identity) was compared between three summer and two winter samples, at different levels of subsampling (0.01–50% of the total data set) and the coverage was computed using Nonpareil (coverage axes). The main panel (bottom-left) shows the number of detected genes, represented by the color of the circles (see legend). The values between subsamples were estimated using bicubic interpolation. Note that the detection of genes is more strongly affected by the coverage in summer data sets owing to lower gene richness in the winter data sets. The additional panels correspond to the comparisons of the subsamples against the complete (not subsampled) data sets, which showed 64% and 75% coverage for winter and summer, respectively. The comparison between complete winter and summer data sets (top-right circle in main panel) was used as a reference for the definition of true/false positives (TP/FP) and true/false negatives (TN/FN). Sensitivity was defined as TP/(TP + FN), specificity as TN/(FP + TN) and accuracy of the test as (TP + TN)/(TP + TN + FN + FN). Sensitivity, specificity and accuracy were interpolated using cubic splines with smoothing parameter 0.6. Differential abundance was defined as adjusted P-value $\leqslant 0.1$ in the negative binomial test implemented in DESeq (Anders and Huber, 2010). Note that sensitivity drops rapidly when coverage of any (or both) of the collections of data sets decreases, while specificity is typically high, except at extreme differences in coverage. In general, the accuracy was compromised (<90%) in data sets with >twofold difference of coverage.

Nonpareil examines redundancy among the individual reads of a whole-genome shotgun metagenomic data set to quantitatively assess the abundance-weighted average coverage of the data set (for example, Figure 1). Therefore, Nonpareil is independent of assembly, reference databases or abundance distribution models, and allows for direct comparisons between data sets and with other quantitative metrics. Furthermore, it projects the average coverage at larger sequencing efforts, providing an estimate of the amount of sequencing required to reach any given coverage and means to quickly rank diversity in metagenomic data sets before assembly or taxonomic classification (Figure 2). Finally, Nonpareil uses empirical cutoffs to determine redundant reads, which represent well the area of genetic discontinuity frequently observed among the sequence-discrete populations that typify natural microbial communities based on previous metagenomic surveys (Caro-Quintero and Konstantinidis, 2012). Accordingly, Nonpareil does not distinguish between subpopulations. It is important to note, however, that Nonpareil estimates are based on the organisms recovered in a metagenomic data set, that is, they represent abundance-weighted values, analogous to how metagenomic data sets preferentially represent the abundant organisms in a sample. Thus, in cases where the goal is to characterize all members of the community, or rare members preferentially, and most of these members are not represented in the metagenomic data set due to very high species richness and/or relatively low sequencing effort, Nonpareil estimates may be limited, and should be complemented with genome- or marker-based estimations.

Using Nonpareil, we were able to directly compare the abundance-weighted average coverage of subsampled data sets with frequent analyses in microbial ecology studies. Fewer genes were identified as differentially abundant between data sets with lower coverage at a nearly log-linear rate (Figure 1, main panel), and both the significance and power of the statistical test decreased in these cases. The sensitivity of the tests rapidly declined as coverage decreased, while the specificity experienced a dramatic drop when comparing data sets with extremely different coverage, indicating a high rate of false positives (Figure 1, smaller panels). In general, we have observed that data sets with average coverage above 60% perform better in terms of assembly and detection of differentially abundant genes (see also Rodriguez-R and Konstantinidis, 2013), and comparisons of data sets with extreme differences in coverage (for example, >twofold) should be avoided.

Here, we advocated for the estimation of the average coverage obtained in metagenomic studies, and briefly presented the advantages of different approaches. Figure 2 shows how coverage is not simply a function of data set size (often the only indication to coverage in metagenomic studies), but largely depends on the complexity of the communities sampled. Figure 1 shows that quantitative estimations of coverage can serve as a basis for the adjustment of statistical tests, applicable to most, if not all, metagenomics studies. We recommend using at least one of the above-mentioned tools to estimate coverage (directly or indirectly) when analyzing metagenomic data sets, taking into consideration the objectives of the study and the nature of the data sets.

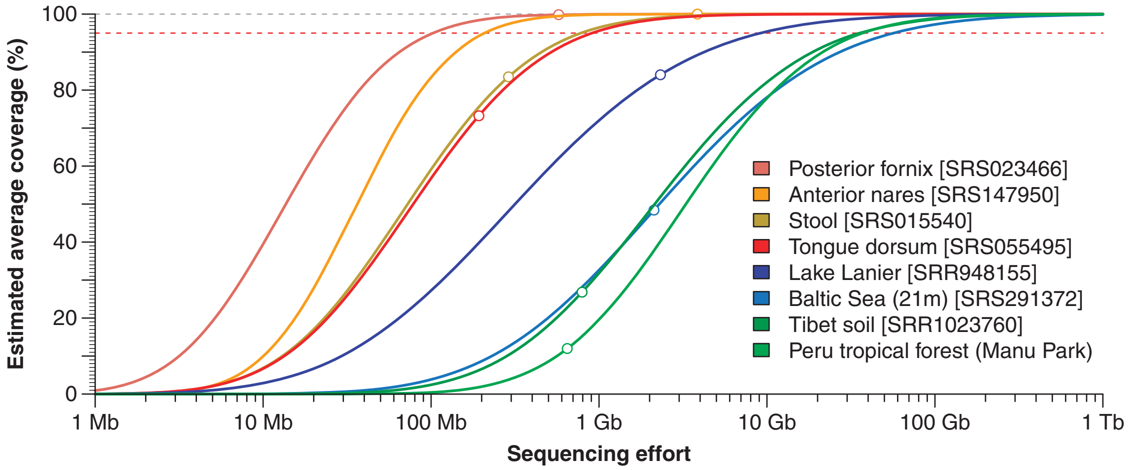

Figure 2 Comparison of diversity and coverage in available metagenomic data sets using Nonpareil curves. The abundance-weighted average coverage is presented as a function of sequencing effort in the form of Nonpareil curves (Rodriguez-R and Konstantinidis, 2013) for selected available metagenomic data sets. Note that more diverse communities require larger sequencing efforts to achieve the same level of coverage, hence located rightward in the plot. Four samples of the Human Microbiome Project are shown that represent communities in the human microbiome of varying diversity, all of which are less diverse than selected environmental samples. Soil (Tibet soil and Peru tropical forest) and marine (Baltic sea, 21 m depth) samples are the most diverse among those selected. The Sequence Read Archive identifier of each sample is provided within squared brackets, except for the Peru tropical forest sample obtained from Fierer *et al.* (2012).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Michael R Weigand for helpful suggestions regarding the manuscript. Our work is supported in part by the US DOE Office of Science, Biological and Environmental Research Division (BER), Genomic Science Program, Award Nos. DE-SC0006662 and DE-SC0004601, and by the US National Science Foundation Award No. 1241046.

LM Rodriguez-R and KT Konstantinidis are at the Center for Bioinformatics and Computational Genomics, and School of Biology, Georgia Institute of Technology, Atlanta, GA, USA.
KT Konstantinidis is also at the School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.
E-mail: kostas@ce.gatech.edu

References

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

Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**: 347–355.

Daley T, Smith AD. (2013). Predicting the molecular complexity of sequencing libraries. *Nat Methods* **10**: 325–327.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL *et al.* (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci* **109**: 21390–21395.

Hooper SD, Dalevi D, Pati A, Mavromatis K, Ivanova NN, Kyrpides NC. (2010). Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics* **26**: 295–301.

Lander ES, Waterman MS. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231–239.

Rodriguez-R LM, Konstantinidis KT. (2013). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**: 629–635.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Stanhope SA. (2010). Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. *PLoS One* **5**: e11652.

Wendl MC, Kota K, Weinstock GM, Mitreva M. (2012). Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J Math Biol* **67**: 1141–1161.