The ISME Journal (2016) 10, 1020–1024
© 2016 International Society for Microbial Ecology All rights reserved 1751-7362/16
www.nature.com/ismej

SHORT COMMUNICATION

16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies

Matthew Perisin[1,2], Madlen Vetter[1], Jack A Gilbert[1,3,4,5] and Joy Bergelson[1,2]
[1]*Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA;* [2]*Committee on Microbiology, University of Chicago, Chicago, IL, USA;* [3]*Biosciences Department, Institute for Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL, USA;* [4]*Marine Biological Laboratory, Woods Hole, MA, USA and* [5]*College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China*

The 16S rRNA gene (16S) is an accepted marker of bacterial taxonomic diversity, even though differences in copy number obscure the relationship between amplicon and organismal abundances. Ancestral state reconstruction methods can predict 16S copy numbers through comparisons with closely related reference genomes; however, the database of closed genomes is limited. Here, we extend the reference database of 16S copy numbers to *de novo* assembled draft genomes by developing 16Stimator, a method to estimate 16S copy numbers when these repetitive regions collapse during assembly. Using a read depth approach, we estimate 16S copy numbers for 12 endophytic isolates from *Arabidopsis thaliana* and confirm estimates by qPCR. We further apply this approach to draft genomes deposited in NCBI and demonstrate accurate copy number estimation regardless of sequencing platform, with an overall median deviation of 14%. The expanded database of isolates with 16S copy number estimates increases the power of phylogenetic correction methods for determining organismal abundances from 16S amplicon surveys.

The ISME Journal (2016) **10,** 1020–1024; doi:10.1038/ismej.2015.161; published online 11 September 2015

Plummeting DNA sequencing costs have allowed insight into the remarkable diversity of the microbial world. The 16S rRNA gene (16S) dominates marker gene studies of bacterial and archaeal diversity, owing to its ubiquity, sequence conservation, and variable regions that allow taxonomic discrimination. However, given that 16S copy number can vary from 1–15 copies (Lee *et al.*, 2009), organismal abundances could be grossly misestimated, producing incorrect diversity estimates (Kembel *et al.*, 2012).

Two methods have emerged to normalize 16S amplicon counts based on copy numbers (Kembel *et al.*, 2012; Langille *et al.*, 2013). Both methods compare observed 16S sequences with a reference database of closed genomes to find related isolates with known copy numbers. An ancestral state is calculated and used to predict the 16S copy number of the organism in question. The accuracy of these methods depends on the relatedness between the observed and reference sequences, with prediction accuracy sharply declining with increasing genetic distance (Langille *et al.*, 2013). A larger and more diverse reference database would improve 16S copy number predictions, especially for samples originating from poorly characterized environments.

There are currently ~3200 closed and ~26 000 draft bacterial genomes deposited in NCBI. Copy numbers for closed genomes are accessible in the Ribosomal RNA Database (*rrn*DB; Stoddard *et al.*, 2014), but no such database exists for draft genomes because of difficulties in resolving copy numbers of highly conserved, repetitive genomic regions. Assemblies frequently contain the 16S in a single, overrepresented contiguous sequence fragment (contig). Although computational methods have emerged to identify and quantify copy number variation using next-generation sequencing data (Zhao *et al.*, 2013; Periwal and Scaria, 2014), these approaches are either designed for polyploids or do not provide the necessary sensitivity for 16S copy number quantification. Of these methods, read depth approaches have shown the most promise (Brynildsrud *et al.*, 2015; Greenblum *et al.*, 2015), but they have not been optimized for draft genomes comprising many contigs.

Building on read depth approaches, we demonstrate accurate and precise quantification of 16S copy number from draft genomes using the 16Stimator pipeline (Supplementary Figure 1). Accuracy was assessed by application to *de novo* assemblies of 12 endophytic bacterial isolates collected from *Arabidopsis thaliana* leaves, and copy number

Correspondence: J Bergelson, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St, Chicago, IL 60637, USA.
E-mail: jbergels@uchicago.edu

Received 7 March 2015; revised 29 June 2015; accepted 8 July 2015; published online 11 September 2015

confirmation by qPCR. We then apply our method to the raw sequence data of draft genomes deposited in NCBI. Thus, we have drastically increased the number of species with estimated 16S copy numbers, thereby improving phylogenetically based abundance estimation. All custom scripts are publicly available at https://bitbucket.org/perisin/16stimator.

For 16S copy number estimation from *de novo* assemblies, paired-end sequencing reads were generated for short and long-insert libraries using the Illumina HiSeq 2000 platform (San Diego, CA, USA). Quality-checked reads were assembled (Supplementary Methods) and sequence data submitted to NCBI (WGS and SRA databases, Supplementary Table S1). After annotation by RAST (Aziz *et al.*, 2008; Overbeek *et al.*, 2013), we extracted the positions of 16S and 73 single-copy, conserved genes (Sharon *et al.*, 2013). Alignment of sequencing reads back to the assembly allowed for calculation of read depth at each genomic position (Supplementary Methods). If there was partial resolution of 16S copies, as demonstrated by multiple 16S locations, we combined the read depths for these regions. We further divided each gene into non-overlapping windows, based on read length, and extracted read depth for each window. Read depth is commonly biased by GC% (Yoon *et al.*, 2009; Ross *et al.*, 2013), so we first tested for bias by fitting a linear model to read depths from single-copy gene windows. If there was a significant GC% effect, we used the model parameters to correct single-copy and 16S read depths (Supplementary Methods). As coverage can be quite variable, we estimated 16S copy number by dividing the median read depth of 16S windows by the median read depth of the single-copy conserved gene windows. We assessed the confidence of these estimates using permutations or asymptotic distribution-free confidence intervals (Price and Bonett, 2002) (PB), and independently confirmed copy numbers for endophytes by qPCR (Figure 1). Absolute quantification of 16S and single-copy genes was performed for genomic DNA and controls that contained a 1:1 ratio of 16S to single-copy gene (Lee *et al.*, 2008).

To test the computational pipeline, we processed similarly generated sequencing reads for *Escherichia coli* TY-2482 (GCA_000217695.2, SRR292678, SRR292862), *Bacteroides fragilis* HMW 615 (GCA_000297735.1, SRR488169, SRR488170), *Pseudomonas aeruginosa* PAO1 (GCA_000006765.1, SRR032420, SRR032832) and *Staphylococcus aureus* KPL1828 (GCA_000507725.1, SRR835799, SRR958927). These isolates have closed genomes or are closely related to isolates with closed genomes. As demonstrated in Figure 1, 16Stimator accurately estimates 16S copy numbers. The confidence of each estimate varies and depends on sequencing coverage variability within the 16S and single-copy gene regions. The permutation method for generating confidence intervals captures this variability to a greater extent than the PB method, which assumes the ratio of coverage medians is asymptotically normal and thereby lessens the influence of outliers.

To extend the database of genomes and draft genomes with known 16S copy numbers, we applied 16Stimator to all annotated genomes with corresponding SRA sequence reads that were published in NCBI by November 25, 2014. Of the 29 315 genome assemblies available, only 3291, representing 1519 unique bacterial species, were closed. After parsing our results for sequencing coverage, proportion of mapped reads and outliers of 16S copy estimates (Supplementary Methods), application of 16Stimator resulted in estimates for 26 353 different sequencing libraries representing 816 unique species, including 586 species without closed genomes (Supplementary Table S2). To test the accuracy of our method, we compared 16Stimator estimates with actual copy numbers for closed genomes and observed a strongly positive, linear correlation (Figure 2a, $r^2 = 0.82$ (0.75, 0.87), $P = < 2e–16$). The median absolute proportional deviation of our estimates to the actual copy numbers was 0.14. Furthermore, we found consistent reliability across sequencing technologies, except for estimates generated from the Illumina HiSeq 1000, Ion Torrent PGM and PacBio RS; note that there are few published draft genomes with corresponding sequencing reads for these platforms (Figure 2b). The large spread of points for the Illumina MiSeq category corresponds to estimates for *Listeria monocytogenes*. Though the genomes analyzed were annotated to have six 16S copies, *rrn*DB lists completed genomes with five or six copies (Stoddard *et al.*, 2014). This variation in true copy number increases the variation in the differences between our estimates and the actual copy numbers. Further, we observed effects on the accuracy of our estimates due to interactions between low sequencing coverage and GC correction of read depths (Supplementary Figure 4). For low-coverage sequencing libraries with GC correction, we underestimated copy numbers by 1–2 copies. For low-coverage sequencing libraries without GC correction, we overestimated copy numbers by 1–5 copies. These libraries did not meet the thresholds for GC correction and provide future grounds for improvement of our method. Though this bias seems specific to *L. monocytogenes* MiSeq sequencing, we have included options in our scripts to calculate copy numbers with or without corrections so each researcher can better judge GC effects on 16S copy number estimation.

Ancestral reconstruction methods currently rely on a reference database of closed genomes, primarily human-associated microbes, to generate accurate metagenome predictions and abundance corrections from 16S amplicon surveys. The accuracy of these methods decreases with sequence divergence to reference, described by Langille *et al.* (2013) as nearest sequenced taxon index (NSTI). 16Stimator can fill holes in the reference database by generating

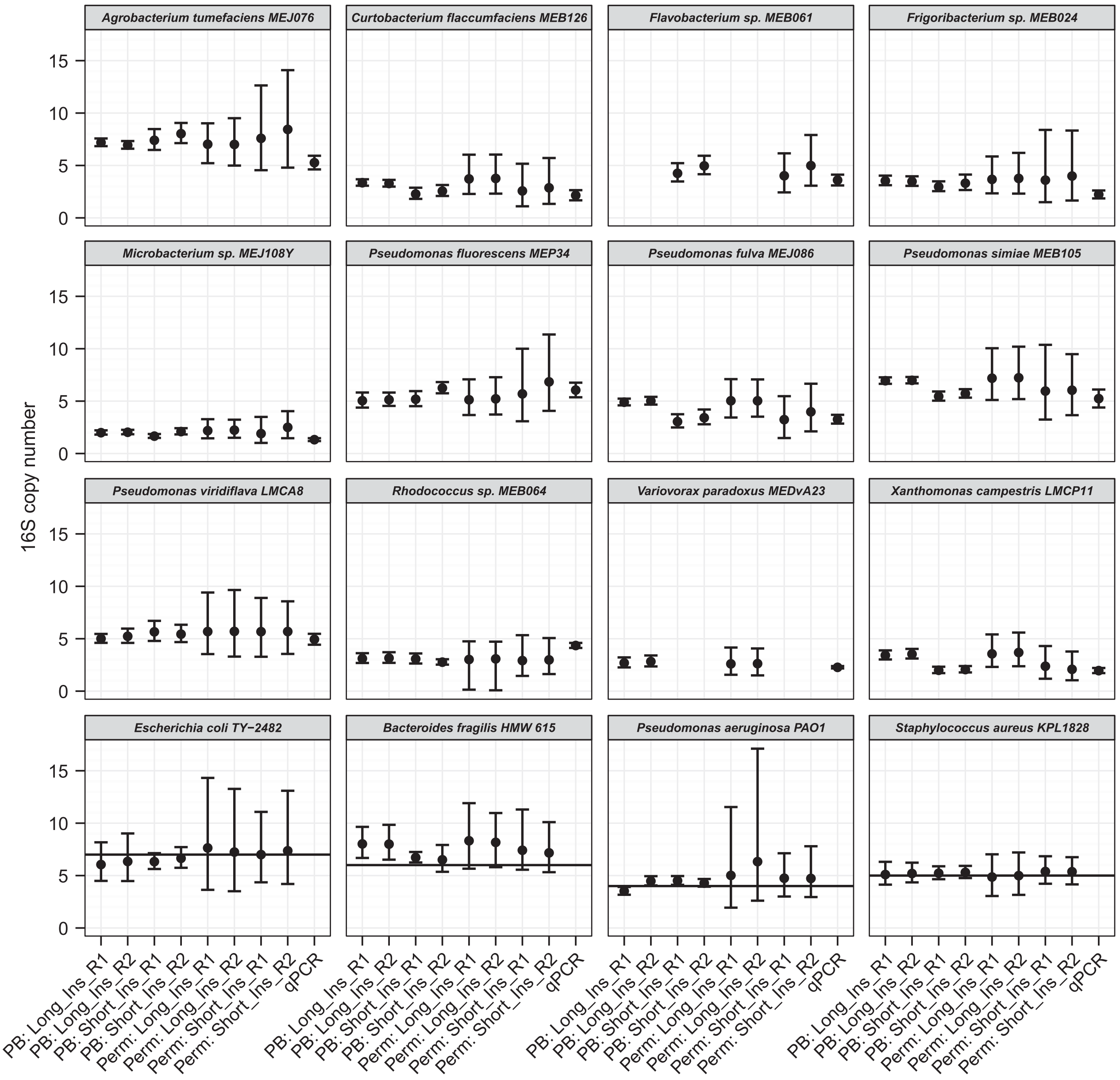

Figure 1 16S copy number estimates from *de novo* assemblies. For each endophytic isolate, paired-end sequencing reads (R1, R2) were generated on the Illumina HiSeq 2000 from short (~250 bp) and long (~2500) insert libraries (Short_Ins and Long_Ins, respectively). For closed-genome controls, similarly generated sequencing reads were downloaded from SRA: *Escherichia coli* TY-2482 (GCA_000217695.2, SRR292678, SRR292862), *Bacteroides fragilis* HMW 615 (GCA_000297735.1, SRR488169, SRR488170), *Pseudomonas aeruginosa* PAO1 (GCA_000006765.1, SRR032420, SRR032832) and *Staphylococcus aureus* KPL1828 (GCA_000507725.1, SRR835799, SRR958927). The 16Stimator pipeline was used to estimate 16S copy number as the ratio of median coverage for 16S and single-copy genes. Confidence intervals (95%) were either calculated as in Price and Bonett (2002) (PB), or via permutations (Perm). For endophytic isolates, 16S copy numbers were independently verified by absolute quantification via qPCR with the mean and standard deviation of technical replicates shown. For closed-genome controls, each horizontal line marks the *rrn*DB (Stoddard *et al*., 2014) consensus 16S copy number for each species. Note: the short-insert library for MEDvA23 and the long-insert library for MEB061 did not meet quality thresholds. 16S copy number was not experimentally determined by qPCR for *E. coli* TY-2482, *B. fragilis* HMW 615, *P. aeruginosa* PAO1 and *S. aureus* KPL1828.

16S copy number estimates from draft assemblies, obviating the need to fully close each genome. Alternatively, ancestral reconstruction methods could predict 16S copy number for a single genome based on its 16S sequence relatedness to reference, closed genomes. From cross-validation of 16S copy number estimation, Langille *et al.* (2013) showed PICRUSt accuracy decreased with increasing NSTI. As our estimates are solely based on sequencing reads from a single genome, without comparison with references, they do not suffer from decreased performance with increasing NSTI. When we compared our method with PICRUSt for Greengenes 99% OTU copy number predictions (McDonald *et al.*, 2012),

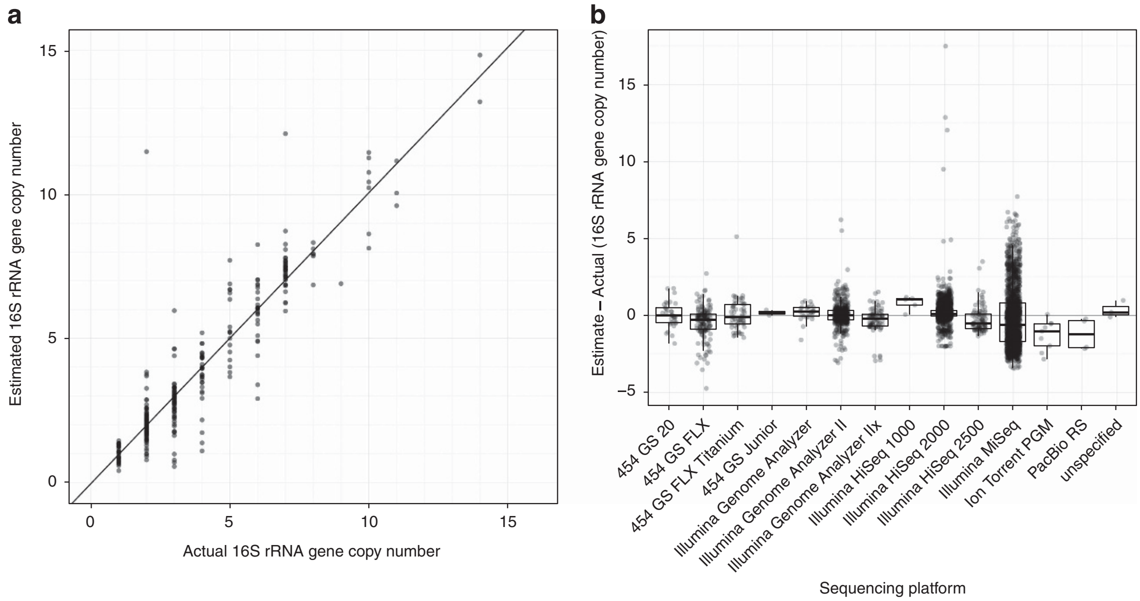

Figure 2 (**a**) Accuracy of 16Stimator pipeline. 16Stimator was applied to sequencing libraries generated from isolates with closed genomes. Plotting estimate versus actual 16S copy numbers assessed accuracy. Each point corresponds to the point estimate for 16S copy number generated from a separate sequencing library. As the number of sequencing libraries per isolate varies, we randomly chose one library for each Isolate:Bioproject combination (283 total) and then fit a linear model (lm function in R). This process was repeated 1000 times to obtain 95% confidence intervals for the mean intercept (–0.05 (–0.13, 0.04) $P = 0.65$ (0.24, 0.98)), slope (1.01 (0.98, 1.03) $P < 2e–16$), and adjusted r^2 values (0.82 (0.75, 0.87)). The above plot represents one such iteration. (**b**) Bias of 16Stimator by sequencing technology. Each point corresponds to results from different sequencing libraries mapped back to their respective closed genomes. Actual 16S copy number values were subtracted from estimates and grouped by sequencing platform. (454 GS 20, $n = 33$; 454 GS FLX, $n = 145$; 454 GS FLX Titanium, $n = 57$; 454 GS Junior, $n = 2$; Illumina Genome Analyzer, $n = 26$; Illumina Genome Analyzer II, $n = 521$; Illumina Genome Analyzer IIx, $n = 78$; Illumina HiSeq 1000, $n = 5$; Illumina HiSeq 2000, $n = 2002$; Illumina HiSeq 2500, $n = 92$; Illumina MiSeq, $n = 1671$; Ion Torrent PGM, $n = 9$; PacBio RS, $n = 4$; Unspecified, $n = 3$).

there was an overall positive correlation between estimates (Pearson correlation coefficient = 0.50, $P < 0.001$), but this correlation deteriorates with NSTI (Supplementary Figure 2). We did not observe an effect of NSTI or phylogeny on confidence interval size for our estimates for the Greengenes OTUs (Supplementary Figure 3 and 5, respectively). We therefore conclude that 16Stimator provides more accurate estimates than PICRUSt for 16S copy number prediction from single genomes.

Nevertheless, ancestral reconstruction methods provide remarkable insight into the metagenomic potential of a sample based on the observed 16S amplicons (Kembel *et al.*, 2012; Langille *et al.*, 2013). Our 16Stimator pipeline has drastically increased the number of bacterial species and strains with estimated ribosomal copy numbers. This expanded database will increase the power of ancestral state reconstruction methods for metagenome predictions and organismal abundance corrections for 16S amplicon surveys. We recommend that this method be applied to new genome submissions in NCBI to continually expand the database. As sequencing costs decline and read lengths increase, we envision higher quality genome assemblies with fully resolved repetitive regions. Nevertheless, our estimates greatly enhance the database now and provide an interim means to continue improving it.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank John Wilmes and Stefano Allesina for helpful discussions, and the Center for Research Informatics at the University of Chicago for computational resources. MP was supported by a Department of Education GAANN fellowship and a NIH Genetics & Regulation training grant. This work was supported by grant DOE DE-AC02-06CH11357 to JAG and grants NSF MCB0603515 and James S. McDonnell Foundation 220020237 to JB.

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA *et al.* (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

Bryníldsrud O, Snipen L-G, Bohlin J. (2015). CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. *Bioinformatics* **31**: 1708–1715.

Greenblum S, Carr R, Borenstein E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**: 583–594.

Kembel SW, Wu M, Eisen JA, Green JL. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* **8**: e1002743.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–821.

Lee C, Lee S, Shin SG, Hwang S. (2008). Real-time PCR determination of rRNA gene copy number: absolute and relative quantification assays with *Escherichia coli*. *Appl Microbiol Biotechnol* **78**: 371–376.

Lee ZM-P, Bussema C, Schmidt TM. (2009). *rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res* **37**: D489–D493.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T *et al.* (2013). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res* **42**: D206–D214.

Periwal V, Scaria V. (2014). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics* **31**: 1–9.

Price RM, Bonett DG. (2002). Distribution-free confidence intervals for difference and ratio of medians. *J Stat Comput Simul* **72**: 119–124.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R *et al.* (2013). Characterizing and measuring bias in sequence data. *Genome Biol* **14**: R51.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. (2014). *rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res* **43**: D593–D598.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**: 1586–1592.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**: S1.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)