Pacific Symposium on Biocomputing 15:3-9(2010)

ACCURATE TAXONOMIC ASSIGNMENT OF SHORT PYROSEQUENCING READS

JOSÉ C. CLEMENTE

Center for Information Biology and DNA Databank of Japan
National Institute of Genetics
Yata 1111, Mishima, Japan
E-mail: jclement@lab.nig.ac.jp

JESPER JANSSON

Graduate School of Humanities and Sciences
Ochanomizu University
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
E-mail: jesper.jansson@ocha.ac.jp

GABRIEL VALIENTE

Algorithms, Bioinformatics, Complexity and Formal Methods Research Group
Technical University of Catalonia
E-08034 Barcelona, Spain
E-mail: valiente@lsi.upc.edu

Ambiguities in the taxonomy dependent assignment of pyrosequencing reads are usually resolved by mapping each read to the lowest common ancestor in a reference taxonomy of all those sequences that match the read. This conservative approach has the drawback of mapping a read to a possibly large clade that may also contain many sequences not matching the read. A more accurate taxonomic assignment of short reads can be made by mapping each read to the node in the reference taxonomy that provides the best precision and recall. We show that given a suffix array for the sequences in the reference taxonomy, a short read can be mapped to the node of the reference taxonomy with the best combined value of precision and recall in time linear in the size of the taxonomy subtree rooted at the lowest common ancestor of the matching sequences. An accurate taxonomic assignment of short reads can thus be made with about the same efficiency as when mapping each read to the lowest common ancestor of all matching sequences in a reference taxonomy. We demonstrate the effectiveness of our approach on several metagenomic datasets of marine and gut microbiota.

Background

The advent of next-generation sequencers has been accompanied by new computational challenges to deal with the ever increasing amounts of data produced.[1] In particular, metagenomic analysis of microbial communities[2,3] has resulted in a plethora of tools for comparative studies, such as determining the richness and diversity of communities,[4–7] or test their similarity.[8–11]

A more fundamental problem is how to determine the composition of a particular community given the set of pyrosequencing reads obtained from a sample, that is, what species (or strains) are present, and in what proportion. Two strategies have been proposed, based on whether a taxonomy is assumed or not. *Binning* approaches discard the use of bacterial taxonomies since they tend to be biased towards cultivable species, and apply instead some unsupervised classification method (clustering) on the reads to determine the structure of the population. Self-Organizing Maps,[12,13] Support Vector Machines,[14] z-score correlations,[15] or nearest neighbors[16] have been successfully utilized for this purpose. *Taxonomy-based* approaches, on the other hand, map the reads to known species in a given taxonomy, usually based on the 16S rRNA. Ambiguous fragments that cannot be unequivocally assigned to a specific taxon are mapped to an inner node of the taxonomy, usually the lowest common ancestor (LCA) of all sequences to which the read might be assigned.[17–20]

Both binning and taxonomy-based methods need to define a measure to compare sequences. *Similarity-based* methods use sequence identity to determine how alike sequences are: BLAST[18,19,21] and number of mismatches[22–25] are commonly used measures. *Sequence composition* methods use instead intrinsic features of the sequences to determine their similarity, such as their GC-content[26] or k-nucleotide frequencies.[14,15,20]

In this paper, we address the problem of how to assign ambiguous short reads with a taxonomy-based

Pacific Symposium on Biocomputing 15:3-9(2010)

approach and a measure of similarity based on the number of mismatches between sequences. The hidden assumption made by previous studies when assigning these fragments to the LCA is that a higher coverage should be preferred to a higher accuracy (see Fig. 1). Our work is a generalization aimed at maximizing the F-measure in order to assign ambiguous reads at inner nodes of the taxonomy that are not necessarily the LCA. Notice that the use of the F-measure in this context is just one of the several possible assignment strategies and it does not reflect the accuracy of the global assignment schema, which would also include unambiguously assigned reads and be affected by the chosen measure of similarity between fragments.

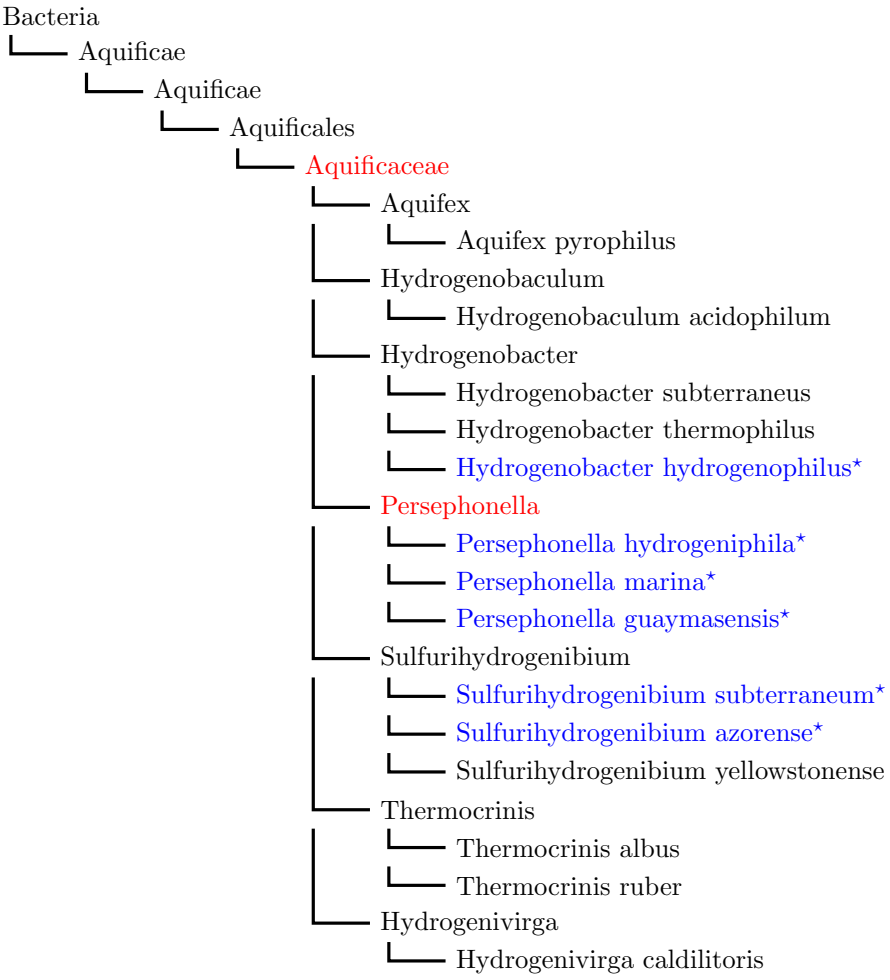

Fig. 1. Coverage and accuracy of assigning ambiguous reads to the LCA. The assignment of an ambiguous read to the family *Aquificaceae*, the LCA of the six matching taxa *H. hydrogenophilus*, *P. hydrogeniphila*, *P. marina*, *P. guaymasensis*, *S. subterraneum*, and *S. azorense*, marked up with a star, has a 100% coverage (recall) but a $6/14 = 43\%$ accuracy (precision). The assignment to the genus *Persephonella*, instead, has a coverage of $3/6 = 50\%$ and an accuracy of 100%.

Methods

Given a reference taxonomy T, a set R of short reads, and a threshold value k of sequence similarity, let R_i be the ith read, let M_i be the leaves of T matching R_i with up to k mismatches, let T_i be the subtree of T rooted at the lowest common ancestor of M_i, and let N_i be the leaves of T_i not matching R_i with up to k mismatches. Let also $L_i = M_i \cup N_i$.

Pacific Symposium on Biocomputing 15:3-9(2010)

Further, consider some arbitrary, but fixed, ordering of the nodes of T, say in postorder, let $T_{i,j}$ be the subtree of T rooted at the jth node of T_i in postorder, let $M_{i,j}$ be the leaves of $T_{i,j}$ matching R_i with up to k mismatches, and let $N_{i,j}$ be the leaves of $T_{i,j}$ not matching R_i with up to k mismatches.

For the ith read and the jth node of T_i in postorder, the leaves of T_i can be partitioned in the following four subsets (see Fig. 2):

- $TP_{i,j} = M_{i,j}$ (true positives)
- $FP_{i,j} = N_{i,j}$ (false positives)
- $TN_{i,j} = N_i \setminus N_{i,j}$ (true negatives)
- $FN_{i,j} = M_i \setminus M_{i,j}$ (false negatives)

Then, the precision of classifying R_i as T_j is $P_{i,j} = |TP_{i,j}|/(|TP_{i,j}| + |FP_{i,j}|)$, and the recall is $R_{i,j} = |TP_{i,j}|/(|TP_{i,j}| + |FN_{i,j}|)$. The combined F-measure of precision and recall is $F_{i,j} = 2P_{i,j}R_{i,j}/(P_{i,j} + R_{i,j})$.

It is easy to see that $F_{i,j} = 2P_{i,j}R_{i,j}/(P_{i,j} + R_{i,j}) = 2|TP_{i,j}|/(2|TP_{i,j}| + |FP_{i,j}| + |FN_{i,j}|) = 2|TP_{i,j}|/(|TP_{i,j}| + |FP_{i,j}| + |M_i|) = 2|M_{i,j}|/(|L_{i,j}| + |M_i|)$. This gives a simple algorithm for computing the best possible taxonomic rank to which each read can be assigned, in time linear in the size of T_i. Given the set M_i of matching sequences for a read R_i, it suffices to compute the sets $L_{i,j}$ and $M_{i,j}$ for each node j in T_i during a bottom-up traversal of T_i.[27,28] Notice that it takes time linear in the size of M_i to find the root of T_i, because T has constant height, and no additional preprocessing of T is required.[29]

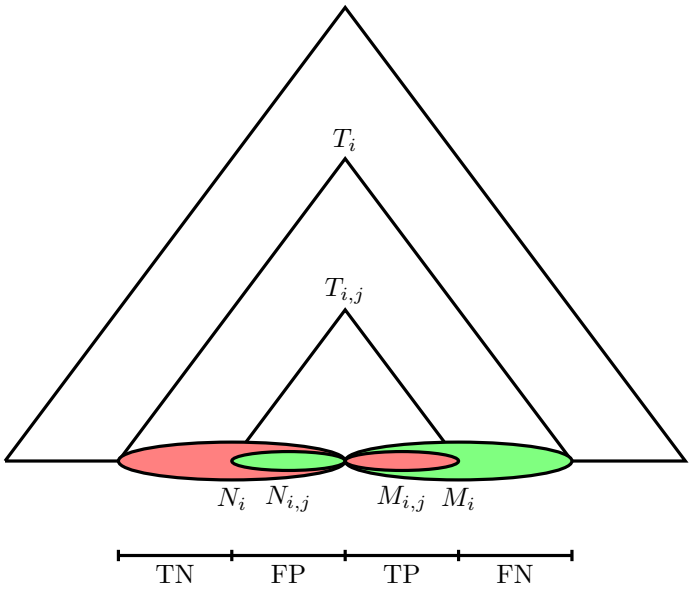

Fig. 2. Precision and recall of assigning the ith read to the jth node of T_i in postorder.

Results and Discussion

It is not completely clear yet what microbial community structure different environments possess, but results so far seem to indicate a high degree of variability both in environmental[30] and gut samples,[31] with significant differences between gut and other microbiomes.[32] The distribution of functions seems to be more conserved though,[33] indicating that a core functionality can be achieved through different species distributions. Understanding this correlation requires accurate measurements of both variables, and our work aims at reducing the amount of error introduced by the assignment of ambiguous fragments to the LCA of a group of species.

Pacific Symposium on Biocomputing 15:3-9(2010)

While feature-based and binning approaches[13,14] require long fragments (more than 1K bp), taxonomical methods can work with shorter reads, which can be as effective as longer sequences for taxonomic assignment provided that the region of the 16S rRNA is adequately chosen.[34,35] The algorithm we introduce here is also very efficient and can process large number of fragments in time at most linear in the number of reference sequences for each fragment, providing a useful tool to quickly test hypotheses about microbial communities.

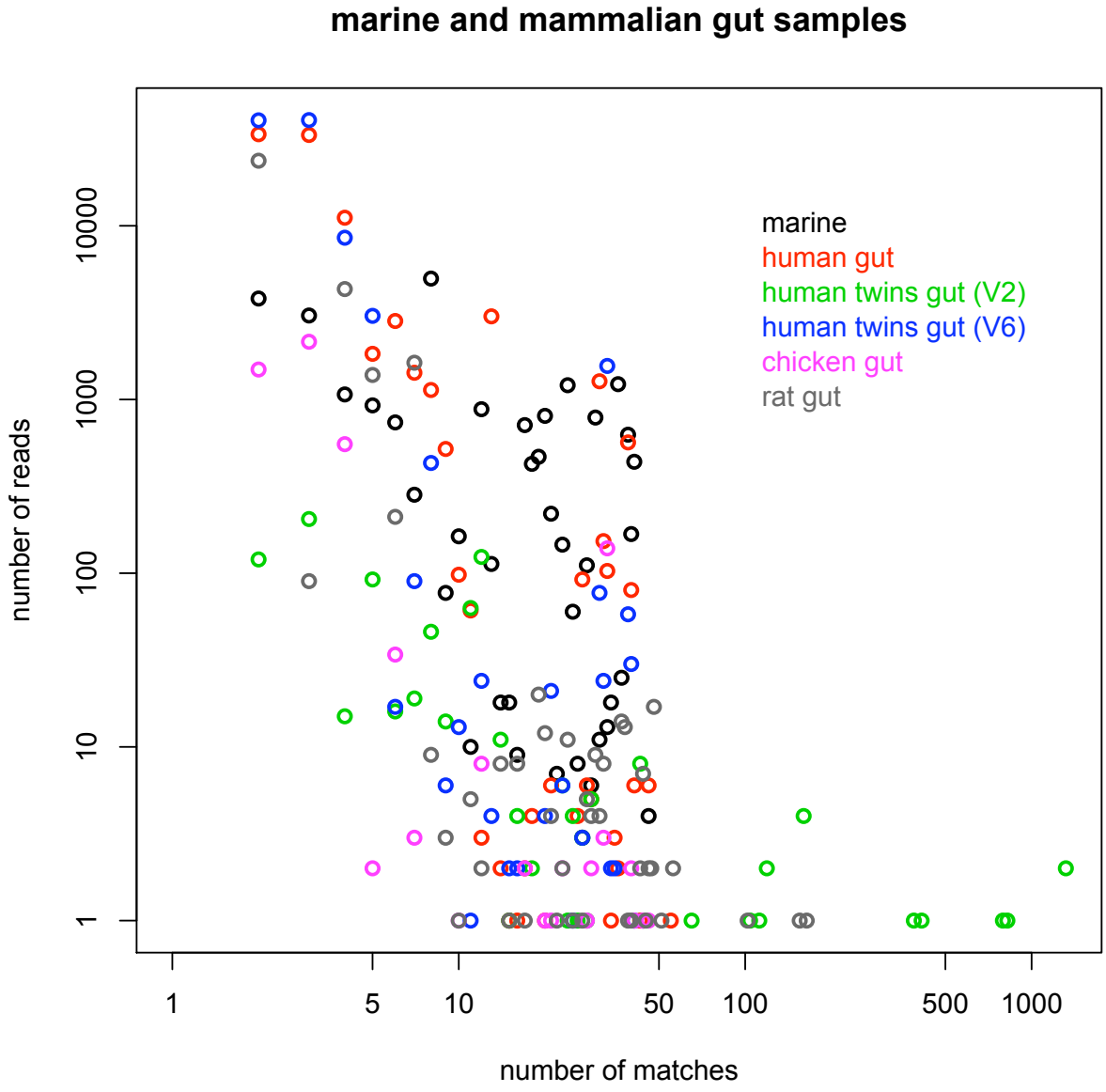

Fig. 3. Distribution of the 23,621 marine, 91,335 human gut, 776 human twins gut (V2 region), 94,999 human twins gut (V6 region), 4,395 chicken gut, and 31,509 rat gut pyrosequencing reads ambiguously matched with up to 2 mismatches to two or more of the 5,165 sequences in the reference bacterial taxonomy.

In order to demonstrate the effectiveness of our approach, we have studied taxonomic assignment in several microbial communities: marine,[30] human gut,[36] lean and obese human twins gut,[31] chicken gut,[37]

Pacific Symposium on Biocomputing 15:3-9(2010)

and rat gut[38] samples. The samples themselves contain 454 pyrosequencing tags for a variable region of 16S rRNA, between 50 and 329 bp in length, and for each of these bacterial communities, we have used both the LCA approach and our approach to assign each of the pyrosequencing reads at the best possible taxonomic rank, using a reference bacterial taxonomy of 5,165 near-full-length type cultures of high quality[17] with a uniform scheme of seven taxonomic ranks (domain, phylum, class, order, family, genus, species).

The taxonomy covers the whole spectrum of known bacteria, and the dominant phyla are Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes, and Tenericutes, with 1,925, 1,285, 1,178, 355, and 160 species, respectively. The near-full-length 16S reference sequences range from 1,202 to 1,780 bp.

The marine (V6 region) samples themselves range from 50 to 100 bp, with an average length of 62 bp; the human gut (V6 and V3 regions) samples range from 50 to 165 bp, average 101 bp; the human twins gut (V2 region) samples range from 50 to 317 bp, average 231 bp; the human twins gut (V6 region) samples range from 50 to 129 bp, average 60 bp; the chicken gut (V6 region) samples range from 55 to 75 bp, average 60 bp; and the rat gut (V4 region) samples range from 50 to 329 bp, with an average length of 231 bp.

We have built a suffix array for the 5,165 sequences in the reference bacterial taxonomy using the GEM-do-index tool,[39] and have matched each of the pyrosequencing reads to the 5,165 reference sequences using the GEM-mapper tool[39] with appropriate parameter settings for finding all matching sequences with up to 2 mismatches, which is about 99% identity for reads of 200 bp. The distribution of those pyrosequencing reads that could not be unambiguously matched to a single sequence in the reference bacterial taxonomy is given in Fig. 3.

The pyrosequencing reads that matched two or more sequences in the reference bacterial taxonomy were assigned to the LCA of the matching sequences in the taxonomy, and they were also assigned at the best possible taxonomic rank using our method. The distribution of reads assigned at the taxonomic rank of domain, phylum, class, order, family, and genus using the LCA of the matching sequences in the taxonomy is shown in Table 1, and the distribution of reads assigned at the taxonomic rank of class, order, family, genus, and species using the new method is shown in Table 2.

Table 1. Number of ambiguous pyrosequencing reads assigned at various taxonomic ranks using the LCA of the matching sequences in the reference bacterial taxonomy of 5,165 sequences.

taxonomic rank	number of reads					
	marine V6	human V6, V3	twins V2	twins V6	chicken V6	rat V4
domain			40			1
phylum	29	5,498	3	13,133	130	49
class	12,099	2,354		1,854	154	3
order	976	5	13	8	8	35
family	3,428	49,647	371	2,343	1,441	3,582
genus	7,089	33,831	349	77,661	2,662	27,839
	23,621	91,335	776	94,999	4,395	31,509

Table 2. Number of ambiguous pyrosequencing reads assigned at various taxonomic ranks using our method in the reference bacterial taxonomy of 5,165 sequences.

taxonomic rank	number of reads					
	marine V6	human V6, V3	twins V2	twins V6	chicken V6	rat V4
class			2			
order			4			2
family	860	2,150	16	195	3	57
genus	17,705	8,441	411	2,353	210	3,622
species	5,056	80,744	343	92,451	4,182	27,828
	23,621	91,335	776	94,999	4,395	31,509

Pacific Symposium on Biocomputing 15:3-9(2010)

These results show that only 3,213 of the 23,621 marine ambiguous reads (13.60%), 4,231 of the 91,335 human gut ambiguous reads (4.63%), 35 of the 776 human twins gut (V2 region) ambiguous reads (4.51%), 635 of the 94,999 human twins gut (V6 region) ambiguous reads (0.67%), 45 of the 4,395 chicken gut ambiguous reads (1.02%), and 48 of the 31,509 rat gut ambiguous reads (0.15%) were actually assigned to the LCA of the matching sequences using our method.

The remaining 96.67% of the ambiguous reads were assigned at a deeper taxonomic rank than the LCA of the matching sequences using the new method. While assigning a read to the LCA of the matching sequences in the taxonomy tends to produce assignments at the ranks of class, order, family, and genus, the new method produces more accurate assignments at the ranks of genus and species.

Conclusions

We have shown in this paper that ambiguities in the taxonomy dependent assignment of pyrosequencing reads can be resolved in an accurate way by mapping each read to the node of a reference taxonomy with the best combined value of precision and recall, in time linear in the size of the taxonomy subtree rooted at the lowest common ancestor of the matching sequences, given a suffix array for the sequences in the reference taxonomy. We have demonstrated the effectiveness of this approach on several metagenomic datasets of marine and gut microbiota, by showing that most reads are actually assigned at a deeper taxonomic rank than the LCA of the matching sequences in the reference taxonomy.

The experimental results were obtained using a reference bacterial taxonomy of 5,165 near-full-length type cultures of high quality.[17] The incompleteness and bias towards cultivable species of the taxonomy might affect these results. Most species in the gut of an individual are rare,[40] and the microbiome has a small number of deep-branching taxa with large diversity at the leaves, with different humans showing different patterns of abundance of microbial species.[41] As our knowledge of the human microbiome expands, we expect the number of unclassified species to diminish and the effectivenes of taxonomical methods to improve consequently.

Acknowledgements

JC was supported by Grant-in-Aid for JSPS Fellows from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, No. 2008086. JJ was supported by the Special Coordination Funds for Promoting Science and Technology. GV was supported by the Spanish government and the EU FEDER program under projects MTM2006-07773 COMGRIO and PCI2006-A7-0603. We want to thank Chaysavanh Manichanh for several discussions on the topic of this paper, and Paolo Ribeca for developing GEM.[39]

References

1. J. Shendure and H. Ji, *Nature Biotechnology* **26**, 1135 (2008).
2. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight and J. I. Gordon, *Nature* **449**, 804 (2007).
3. J. Venter, K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelson, W. Nelson, D. Fouts, S. Levy, A. Knap, M. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers and H. Smith, *Science* **304**, 66 (2004).
4. P. D. Schloss and J. Handelsman, *Applied and Environmental Microbiology* **71**, 1501 (2005).
5. P. D. Schloss and J. Handelsman, *Applied and Environmental Microbiology* **72**, 6773 (2006).
6. P. D. Schloss and J. Handelsman, *BMC Bioinformatics* **9**, p. 34 (2008).
7. V. Seguritan and F. Rohwer, *BMC Bioinformatics* **2**, p. 9 (2001).
8. C. Lozupone and R. Knight, *Applied and Environmental Microbiology* **71**, 8228 (2005).
9. A. Martin, *Applied and Environmental Microbiology* **68**, 3673 (2002).
10. P. D. Schloss and J. Handelsman, *Applied and Environmental Microbiology* **72**, 2379 (2005).
11. D. Singleton, M. Furlong, S. Rathbun and W. Whitman, *Applied and Environmental Microbiology* **67**, 4374 (2001).
12. T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya and T. Ikemura, *DNA Research* **12**, 281 (2005).

Pacific Symposium on Biocomputing 15:3-9(2010)

13. C. Martin, N. N. Diaz, J. Ontrup and T. W. Nattkemper, *Bioinformatics* **24**, 1568 (2008).
14. A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz and I. Rigoutsos, *Nature Methods* **4**, 63 (2007).
15. H. Teeling, A. Meyerdierks, M. Bauer, R. Amann and F. O. Glöckner, *Environmental Microbiology* **6**, 938 (2004).
16. N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus and T. W. Nattkemper, *BMC Bioinformatics* **10**, p. 56 (2009).
17. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity and J. M. Tiedje, *Nucleic Acids Research* **37**, 141 (2009).
18. D. H. Huson, A. F. Auch, J. Qi and S. C. Schuster, *Genome Research* **17**, 377 (2007).
19. Z. Liu, T. Z. DeSantis, G. L. Andersen and R. Knight, *Nucleic Acids Research* **36**, p. e120 (2008).
20. Q. Wang, G. M. Garrity, J. M. Tiedje and J. R. Cole, *Applied and Environmental Microbiology* **73**, 5261 (2007).
21. L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards and J. Stoye, *Nucleic Acids Research* **36**, 2230 (2008).
22. B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, *Genome Biology* **10**, p. R25 (2009).
23. H. Li and R. Durbin, *Bioinformatics* **25**, 1754 (2009).
24. R. Li, Y. Li, K. Kristiansen and J. Wang, *Bioinformatics* **24**, 713 (2008).
25. N. Malhis, Y. Butterfield, M. Ester and S. J. M. Jones, *Bioinformatics* **25**, 6 (2009).
26. H. G. Martín, N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon and P. Hugenholtz, *Nature Biotechnology* **24**, 1263 (2006).
27. G. Valiente, *Algorithms on Trees and Graphs* (Springer, 2002).
28. G. Valiente, *Combinatorial Pattern Matching Algorithms in Computational Biology using Perl and R* (Taylor & Francis/CRC Press, 2009).
29. M. A. Bender, M. Farach-Colton, G. Pemmasani, S. Skiena and P. Sumazin, *Journal of Algorithms* **57**, 75 (2005).
30. M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl, *Proc. Natl. Acad. Sci. USA* **103**, 12115 (2006).
31. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon, *Nature* **457**, 480 (2009).
32. R. E. Ley, C. Lozupone, M. Hamady, R. Knight and J. I. Gordon, *Nature Reviews Microbiology* **6**, 776 (2008).
33. E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White and F. Rohwe, *Nature* **452**, 629 (2008).
34. Z. Liu, C. Lozupone, M. Hamady, F. D. Bushman and R. Knight, *Nucleic Acids Research* **35**, p. e120 (2007).
35. C. Manichanh, C. E. Chapple, L. Frangeul, K. Gloux, R. Guigó and J. Dore, *Nucleic Acids Research* **36**, 5180 (2008).
36. L. Dethlefsen, S. Huse, M. L. Sogin and D. A. Relman, *PLoS Biology* **6**, p. e280 (2008).
37. VAMPS, Visualization and analysis of microbial population structure project, AGT_CKN_Bv6—Chicken intestinal microbiota, (2009).
38. C. Manichanh, Rat intestinal microbiota, Private communication, (2009).
39. P. Ribeca, GEM—GEnomic Multi-tool, http://gemlibrary.sourceforge.net/, (2009).
40. L. Dethlefsen, M. McFall-Ngai and D. A. Relman, *Nature* **449**, 811 (2007).
41. R. E. Ley, D. Peterson and J. I. Gordon, *Cell* **124**, 837 (2006).