# Taxonomic Assignment in Metagenomics with TANGO

**Daniel Alonso-Alemany[1], José C. Clemente[2], Jesper Jansson[3], Gabriel Valiente[4]**

[1]Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain
[2]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, USA
[3]Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
[4]Algorithms, Bioinformatics, Complexity and Formal Methods, Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain

*Corresponding author:* valiente@lsi.upc.edu

## Abstract

One of the main computational challenges facing metagenomic analysis is the taxonomic identification of short DNA fragments. The combination of sequence alignment methods with taxonomic assignment based on consensus can provide an accurate estimate of the microbial diversity in a sample. In this note, we show how recent improvements to these consensus methods, as implemented in the latest release of the TANGO tool, can provide an improved estimate of diversity in simulated datasets.

## Introduction

The diversity and richness of microbial populations can be characterised by several ecological indices, calculated by either grouping similar sequence reads into operational taxonomic units, or assigning them to the most similar taxa in a given taxonomy. While the former is useful for the study of unknown microbial communities, the latter is best suited when sequences and taxonomies of related species are already known.

The usual protocol for taxonomic assignment involves aligning the sequence reads to a set of reference sequences and, then, resolving any ambiguities (that is, a sequence being equally similar to more than one reference sequence) by assigning to a consensus sequence, such as the lowest common ancestor (LCA) of all the candidate sequences in a given taxonomy (Huson *et al*., 2007; Kunin *et al*., 2008; Liu *et al*., 2008). Sequence composition-based methods have also been used in taxonomic assignment (Diaz *et al*., 2009; McHardy *et al*., 2007; Wang *et al*., 2007).

Previous work on taxonomic assignment based on alignment has focused either on sequence reads of the 16S ribosomal RNA gene (Clemente *et al*., 2010, 2011; Ribeca and Valiente, 2011), or on whole metagenomic shotgun sequence reads (Gerlach *et al*., 2009; Krause et al., 2008). In this note, we show for the latter that recent improvements to consensus methods, as implemented in the latest release of the TANGO tool (Clemente *et al*., 2011), bring about an accurate estimate of the actual taxonomic diversity in a metagenomic data-set.

In the improved consensus method, ambiguous sequence reads are assigned to consensus sequences at a lower taxonomic rank than the LCA of the candidate reference sequences (increased specificity), at the expense of discarding some candidate reference sequences (reduced sensitivity). This is done by optimising the combined precision and recall (F-measure) of the taxonomic assignment (Clemente *et al*., 2010, 2011).

## Metagenomic data-set

The complexity of the signal obtained when sequencing metagenomic data makes it necessary to take a standardised data-set as the basis for analysis (Ribeca and Valiente, 2011). We have chosen the metagenomic data-set of Mavromatis *et al*. (2007), which was designed with the goal of simulating microbial communities of varying complexity: low-complexity communities, with one dominant population (simLC), as seen in bioreactor communities (García Martín *et al*., 2006; Strous *et al*., 2006); medium-complexity communities, with more than one dominant population flanked by low-abundance populations (simMC), as seen in acid mine drainage biofilm (Tyson *et al*., 2004)

**Table 1.** Phylogenetic distribution of the 113 microbial genomes.

| Domain | Phylum | Class | Genomes |
|---|---|---|---|
| Bacteria | Actinobacteria | Actinobacteria | 9 |
| | Bacteroidetes | Cytophagia | 1 |
| | Chlorobi | Chlorobia | 7 |
| | Chloroflexi | Chloroflexi | 1 |
| | Cyanobacteria | Cyanobacteria | 6 |
| | Deinococcus-Thermus | Deinococci | 1 |
| | Firmicutes | Bacilli | 13 |
| | | Clostridia | 8 |
| | Proteobacteria | Alphaproteobacteria | 17 |
| | | Betaproteobacteria | 13 |
| | | Gammaproteobacteria | 25 |
| | | Deltaproteobacteria | 6 |
| | | Epsilonproteobacteria | 1 |
| | | unclassified Proteobacteria | 1 |
| Archaea | Euryarchaeota | Methanomicrobia | 3 |
| | | Thermoplasmata | 1 |

and symbiotic microbes from eukaryotes (Woyke *et al.*, 2006); and high-complexity communities, with no dominant population (simHC), as seen in agricultural soil (Tringe *et al.*, 2005).

The Mavromatis *et al.* data-set was built by combining Sanger sequence reads selected at random from 113 microbial genomes. The phylogenetic composition of the metagenomic data-set, summarised in Table 1, shows a high abundance of Proteobacteria, Actinobacteria, and Firmicutes, as usual in most metagenomic samples (Gabor *et al.*, 2004; Manichanh *et al.*, 2008).

**Table 2.** Distribution of sequence reads in the metagenomic data-set.

| | simLC | simMC | simHC |
|---|---|---|---|
| Most abundant | 28,861 | 22,956 | 2,384 |
| 2nd abundant | 9,277 | 16,577 | 2,248 |
| 3rd abundant | 5,168 | 10,484 | 2,191 |
| 4th abundant | 1,149 | 6,107 | 2,127 |
| 5th abundant | 1,109 | 4,868 | 2,083 |
| 6th abundant | 1,074 | 1,146 | 2,051 |
| Rest | 50,857 | 52,319 | 103,687 |

The distribution of sequence reads in the metagenomic data-set, summarised in Table 2, shows a low-complexity microbial community, with one dominant population (28,861 sequence reads from *Rhodopseudomonas palustris HaA2*); a mediumcomplexity microbial community, with three dominant populations (22,956 sequence reads from *Bradyrhizobium sp. BTAi1*, 16,577 sequence reads from *Rhodopseudomonas palustris BisB5*, and 10,484 sequence reads from *Xylella fastidiosa Dixon*) flanked by low-abundance populations; and a high-complexity microbial community, with no dominant population.

## Aligning sequence reads

The first step in the taxonomic analysis of a metagenomic data-set involves aligning the sequence reads to a database of known sequences from a large set of different organisms. Traditional alignment tools, such as BLAST (Altschul *et al.*, 1990) or BLAT (Kent, 2002), do not scale up to align millions or billions of sequence reads to a large reference genome (Horner *et al.*, 2010; Ribeca and Valiente, 2011; Trapnell and Salzberg, 2009). Microbial genomes are much shorter, though, making these tools appropriate for the alignment of sequence reads from envi-

Table 3: Ambiguous sequence reads in the metagenomic data-set.

| Data-set | No hit | One hit | Ambiguous | Total |
|---|---|---|---|---|
| simLC | 59 | 22,956 | 2,384 | 97,495 |
| simMC | 76 | 16,577 | 2,248 | 114,457 |
| simHC | 100 | 10,484 | 2,191 | 116,771 |

ronmental samples. Nevertheless, more efficient tools are available for the alignment of short and long sequence reads obtained using high-throughput sequencing technologies, including BWA (Li and Durbin, 2009), BWA/SW (Li and Durbin, 2010), and GEM (Ribeca, 2009).

We have used BLAST to align the 328,723 sequence reads to the 113 microbial genomes. Notice that a larger database is often used when the target sequences are not known beforehand. Ambiguities arise when a sequence read is aligned with more than one target sequence, and we have taken as candidate alignments all those sequences with the same E-value as the top BLAST hit. As shown in Table 3, ambiguous sequence reads represent about 20% of the metagenomic data-set. Sequence reads with no hit in the database of microbial genomes are the result of sequencing errors.

## Assigning sequence reads

Once the sequence reads have been aligned to reference sequences, the second step in the taxonomic analysis of a metagenomic data-set involves resolving ambiguities by mapping those reads with more than one possible assignment to species at the closest possible taxonomic rank. We have chosen as taxonomic reference the NCBI taxonomy (Sayers *et al*., 2009) for the 113 sampled microbial genomes. Again, notice that a larger taxonomy is often used when the target sequences are not known beforehand. Alternative taxonomies for microbial genomes include ARB-SILVA (Pruesse *et al*., 2007), Greengenes (DeSantis *et al*., 2006), RDP (Cole *et al*., 2009), and TOBA (Garrity *et al*., 2007).

We have used TANGO to assign the 328,723 sequence reads to the 113 microbial genomes at the closest possible taxonomic rank. As shown in Table 4, the optimal consensus method, F-measure-based assignment, resulted in assignments at a lower taxonomic rank than the classical consensus method, LCA-based assignment (Huson *et al*., 2007).

## Taxonomic diversity

Once the sequence reads have been assigned a taxonomy, the third and final step in the taxonomic analysis of a metagenomic data-set involves describing the diversity and richness of the sampled microbial population by means of ecological indices. Some widely accepted notions in ecology are those of $\alpha$-diversity (species diversity within an ecosystem), $\beta$-diversity (change in species diversity within an ecosystem), and $\omega$-diversity (phylogenetic difference between species in an ecosystem) (Faith, 1992; Whittaker, 1972). Among the latter, we have chosen the Clarke-Warwick taxonomic diversity index (Clarke and Warwick, 1998), which measures the

Table 4: Taxonomic distribution of the metagenomic data-set using consensus (LCA, top) and optimal (F-measure, bottom) taxonomic assignment.

| Data-set | Taxonomic rank | | | | | |
|---|---|---|---|---|---|---|
| | Domain | Phylum | Class | Order | Family | Genus |
| simLC | 126 | 104 | 134 | 56 | 2,785 | 5,295 |
| simMC | 194 | 176 | 174 | 101 | 2,784 | 5,219 |
| simHC | 272 | 219 | 230 | 111 | 822 | 11,164 |
| simLC | | 1 | 65 | 46 | 1,236 | 3,241 |
| simMC | | 10 | 90 | 104 | 1,179 | 3,191 |
| simHC | | 12 | 145 | 77 | 414 | 6,847 |

Table 5: Taxonomic diversity (Clarke-Warwick index) of the metagenomic data-set for consensus (LCA) and optimal (F-measure) taxonomic assignment, together with the actual taxonomic diversity.

| Data-set | Taxonomic diversity | |
|---|---|---|
| | LCA | F-measure |
| simLC | 3.8193 | 4.5798 |
| simMC | 4.1485 | 4.7993 |
| simHC | 4.9433 | 5.7422 |

average distance in the taxonomic reference between the sampled species.

As shown in Table 5, the closer the measured taxonomic diversity in the metagenomic data-set is to the actual taxonomic diversity in the sampled population, the more accurate the assignment is: that is, when classical consensus (LCA) is replaced by the optimal consensus (F-measure) method.

## Conclusion

The combination of sequence alignment methods with taxonomic assignment based on consensus provides an accurate estimate for the composition of a sample of sequence reads of the 16S ribosomal RNA gene. We have shown that for sequence reads of whole microbial genomes, recent improvements to consensus methods also bring about an accurate estimate of the microbial diversity in a metagenomic sample.

## Acknowledgements

DA was supported by the Ministry of Economy and Knowledge of the Government of Catalonia and the European Social Fund. JJ was supported by the Special Coordination Funds for Promoting Science and Technology, Japan.

## References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-410. doi:10.1016/S0022-2836(05)80360-2
2. Clarke KR, Warwick RM (1998) A taxonomic distinctness index and its statistical properties. *J Appl Ecol* **35**, 523-531. doi:10.1046/j.1365-2664.1998.3540523.x
3. Clemente JC, Jansson J, Valiente G (2010) Accurate taxonomic assignment of short pyrosequencing reads. *Pacific Symp Biocomput* **15**, 3-9. doi:10.1142/9789814295291 _ 0002
4. Clemente JC, Jansson J, Valiente G (2011) Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics* **12**, 8. doi:10.1186/1471-2105-12-8
5. Cole JR, Wang Q, Cardenas E, Fish J, Chai B et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141-D145. doi:10.1093/nar/gkn879
6. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Briodie EL *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069-5072. doi:10.1128/AEM.03006-05
7. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* **10**, 56.
8. Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**, 1-10. doi:10.1016/0006-3207(92)91201-3
9. Gabor EM, Alkema WBL, Janssen DB (2004) Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**, 879-886. doi:10.1111/j.1462-2920.2004.00640.x
10. García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**, 1263-1269. doi:10.1038/nbt1247
11. Garrity GM, Lilburn TG, Cole JR, Harrison SH, Euzeby J *et al.* (2007) The taxonomic outline of bacteria and archaea. TOBA release 7.7. Michigan State University Board of Trustees, http://www.taxonomicoutline.org/.
12. Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J. (2009) WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* **10**, 430. doi:10.1186/1471-2105-10-430
13. Horner DS, Pavesi G, Castrignanò T, De Meo PD, Liuni S *et al.* (2010) Bioinformatics approaches for genomics and post genomics applications of nextgeneration sequencing.

*Brief Bioinform* **11**, 181-197. doi:10.1093/bib/bbp046

14. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386. doi:10.1101/gr.5969107
15. Kent WJ (2002) BLAT—The BLAST-like alignment tool. *Genome Res* **12**, 656-664. doi:10.1101/gr.229202
16. Krause L, Diaz NN, Goesmann A Kelley S, Nattkemper TW *et al.* (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**, 2230-2239. doi:10.1093/nar/gkn038
17. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72**, 557-578. doi:10.1128/MMBR.00009-08
18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760. doi:10.1093/bioinformatics/btp324
19. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595. doi:10.1093/bioinformatics/btp698
20. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36**, e120. doi:10.1093/nar/gkn491
21. Manichanh C, Chapple CE, Frangeul L, Gloux K, Guigo R *et al.* (2008) A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res* **36**, 5180-5188. doi:10.1093/nar/gkn496
22. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**, 495-500. doi:10.1038/nmeth1043
23. McHardy AC, García Martín H, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**, 63-72. doi:10.1038/nmeth976
24. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W *et al.* (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**, 7188–7196. doi:10.1093/nar/gkm864
25. Ribeca P (2009) GEM: Genomic multi-tool. http://gemlibrary.sourceforge.net/.
26. Ribeca P, Valiente G (2011) Computational challenges of sequence classification in microbiomic data. *Brief Bioinform*. In press. doi:10.1093/bib/bbr019
27. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH *et al.* (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **39**, D38-D51. doi:10.1093/nar/gkq1172
28. Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A *et al.* (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**, 790-794. doi:10.1038/nature04647
29. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* **27**, 455-458. doi:10.1038/nbt0509-455
30. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308**, 554-557. doi:10.1126/science.1107851
31. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **424**, 37-43.
32. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267. doi:10.1128/AEM.00062-07
33. Whittaker RH (1972) Evolution and measurement of species diversity. *Taxon* **21**, 213-251.
34. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950-955. doi:10.1038/nature05192