PRE-PRINT

Linking comparative genomics and environmental distribution patterns of microbial populations through metagenomics

Tom O. Delmont[1] and A. Murat Eren[1,2*]

*Correspondence:
meren@uchicago.edu

[1]The Department of Medicine, The University of Chicago, Chicago, 60637 IL, USA

[2]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, 02543 MA, USA

Full list of author information is available at the end of the article

Abstract

Combining well-established practices from comparative genomics and the emerging opportunities from assembly-based metagenomics can enhance the utility of increasing number of metagenome-assembled genomes (MAGs). Here we used protein clustering to characterize 48 MAGs and 10 cultivars based on their entire gene content, and linked this information to their environmental distribution patterns to better understand the microbial response to the 2010 Deepwater Horizon oil spill in the Gulf of Mexico coastline. Our results suggest that while most oil-associated bacterial populations originated from the ocean, a few actually emerged from the sand rare biosphere. These new findings suggest that there are considerable benefits to employ approaches from comparative genomics to study the whole content of newly identified genomes, and the investigation of emerging patterns in the environmental context can augment the efficacy of assembly-based metagenomic surveys.

Keywords: comparative genomics; metagenomics; meta-pangenomics; microbial ecology; oil spill; anvi'o

During the last two decades, the genomic content of more than 40,000 microbial isolates have been characterized and used to study the microbial gene pool, adaptation, and evolution [1, 2, 3, 4, 5]. Although cultivation-based methods have paved the way for the emergence of powerful comparative genomics approaches, comprehensive understandings of microbial distribution patterns and niche boundaries remained hard to achieve due to well-understood limitations of cultivation.

A complimentary solution emerges from assembly-based metagenomics, where both the genomic content, and the relative distribution of naturally occurring microbial populations can be recovered [6, 7]. To explore microbial systems using metagenome-assembled genomes (MAGs), researchers often rely on functional annotations, phylogenetically informative conserved gene families, or distribution patterns to identify metabolic potentials [8], evolutionary relatedness [9, 10], or co-occurrence of genomic collections [11]. Despite their efficacy, these approaches disregard the signal from genes that are not yet characterized, but may be critical for niche adaptation [12].

Characterizing often-novel MAGs by taking their entire gene content into consideration, in conjunction with their distribution recovered from metagenomic data could compliment current practices, and provide additional insights into the ecology of microbial ecosystems. Here we investigated MAGs and cultivar genomes associated with the 2010 Deepwater Horizon (DWH) oil spill to improve our understanding regarding the origin of oil-associated microbial populations using protein clusters (PCs), and their environmental distribution patterns.

The shotgun metagenomic dataset we studied contained 16 samples collected by Rodriguez-R et al. [13] from Pensacola Beach (Florida) (1) before the oil from DWH began to wash ashore, (2) during the oiling event, (3) and after the removal of oil from the beach. This dataset of 452 million reads was originally analyzed at the contig-level [13], and our re-analysis had resulted in the recovery of MAGs spanning various taxonomical groups [14]. Of those, 14 MAGs were mostly detected before or after the oiling event, while 34 MAGs and all ten cultivar genomes isolated from the same environment [15] were enriched only in oil-contaminated samples (Figure 1). The distribution patterns across environmental conditions revealed a strong link between the genomes enriched in oil-contaminated samples, and genes coding for urea metabolism [14]. Urea is a dissolved organic nitrogen compound used by marine microbes as a main source of nitrogen (Solomon et al., 2010), which prompted us to suggest that bacterial populations enriched in the Gulf of Mexico coastline during DWH originated from the ocean rather than emerging from the sand rare biosphere [14].

To revisit these findings, here we used 176,024 genes identified in the 48 MAGs and 10 cultivar genomes, and clustered them into 14,991 non-singleton protein clusters (PCs) (see Supplementary Methods). We then (1) clustered PCs based on their occurrence patterns across genomes, (2) organized genomes based on PCs they harbor, and (3) created a holistic display that conveys the distribution patterns of genomes in the environment (Figure 1). Five groups emerged from the organization of genomes based on the PCs they shared. Groups I-II, and IV were mostly enriched during oiling, and contained 32 MAGs and all 10 cultivar genomes. In contrast, most of the 16 MAGs in groups III and V were detected only before or after oiling.

26% of the 12,983 functions we identified distributed differentially between the five groups (p ¡ 0.05; Figure 1). We incorporated a subset of these functions into our display to highlight the biological relevance of groups (Figure 1). Among them, genes coding for urea metabolism were widespread in groups I-II, and IV but missing in groups III and V, providing a stronger partitioning of urea metabolism by PCs compared to distribution patterns alone (Figure 1). Groups I-II-IV were also enriched with genes coding for TRAP transporters (specialized in the uptake of organic acids), glutathione reductase and glutathione S-transpherase (oxidative stress and detoxification), revealing distinct nutrient acquisition and stress response strategies compared to groups III and V (Figure 1). Furthermore, genes coding for flagellar biosynthesis and twitching mobility were widespread in all groups except group III. Instead, group III was enriched in genes involved in gliding motility, providing a means to travel in environments with low water content ([16]).

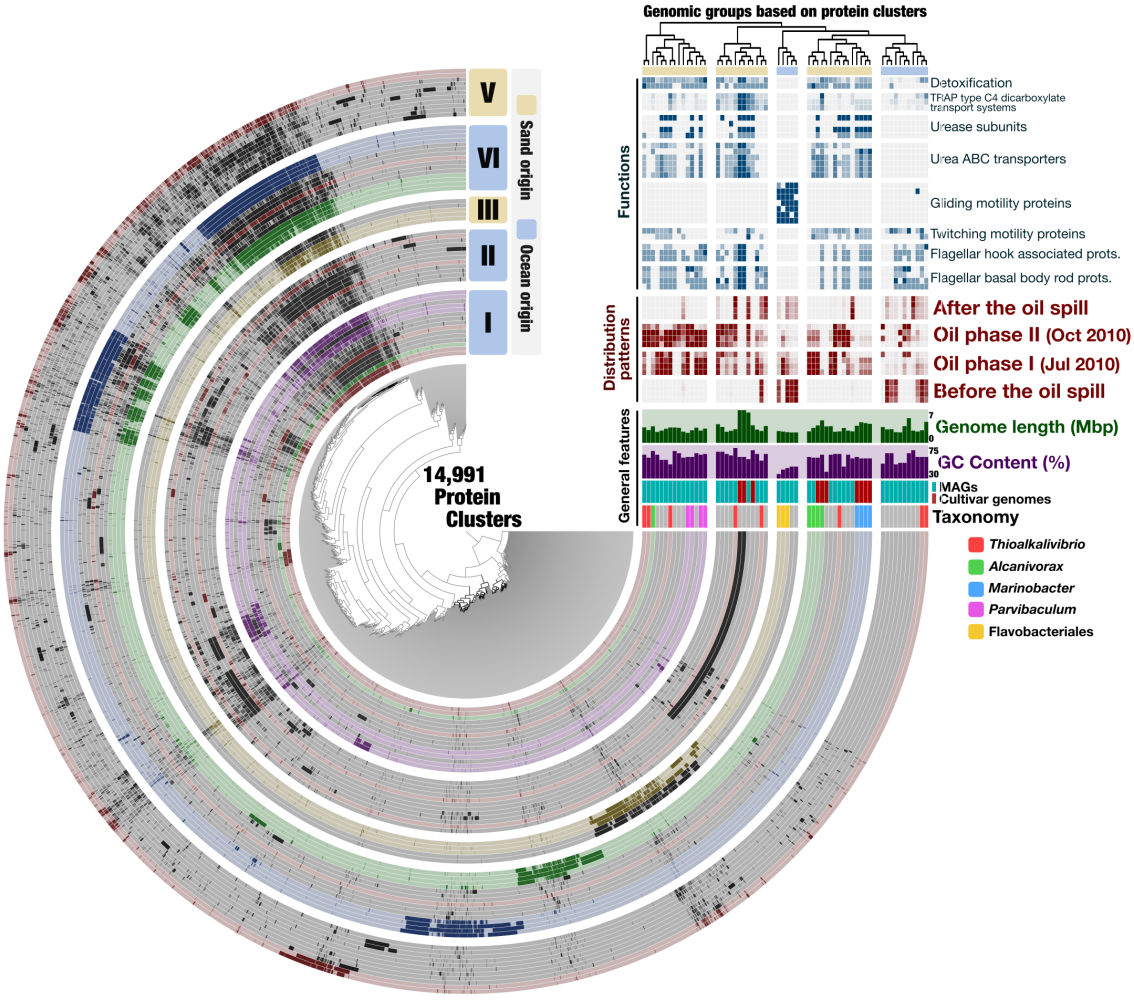

Figure 1 Clustering of 58 microbial genomes identified from the Gulf of Mexico coastline based on 14,991 protein clusters (PCs). Each radial layer represents a genome, and each bar in a layer represents the occurrence of a PC. The organization of genomes is defined by the shared PCs (the second tree on the right-top). In addition, general features, environmental distribution patterns, and a selection of differentially occurring functions are displayed for each genome. A high-resolution copy of this image is temporarily hosted here.

The overall coherence between the partitioning of genomes based on shared PCs, and their functional potential and distribution patterns, supports the hypothesis of a distinct origin for oil-associated microbes (Figure 1). This holistic strategy elucidates which bacterial populations likely originated from the ocean (groups I-II, and IV), and separates them from the native sand bacterial populations (groups III and V). Interestingly, groups associated with the sand ecosystem include genomes undetected before the oiling event, revealing a previously overlooked finding that some oil-associated populations emerged from the sand rare biosphere.

PCs also show improvement over taxonomy. For instance, eight *Thioalkalivibrio* genomes in the dataset occurred in four of the five genomic groups, reminding that co-existing microbial populations sharing the same taxonomical affiliation can strongly diverge functionally and may not necessarily originate from the same

ecosystem. In another example, all three Flavobacteriales MAGs clustered in group III, yet one of them was oil-associated while the two other occurred mostly before the perturbation, suggesting an oil-triggered shift in the relative distribution of Flavobacteriales MAGs, possibly within the same niche.

In summary, combining comparative genomics, environmental distribution patterns, and functional potential of genomes in a single context allowed us to achieve a more detailed depiction of the ecology of microbes in an oil-challenged environment by enhancing the utility of MAGs. This approach revealed co-occurring microbial populations that originate from distinct ecosystems, and differentially occurring microbial populations that share the same one. Our findings also showed that although most of the oil-degrading microorganisms in Pensacola Beach originated from the marine environment, some others likely emerged from the sand rare biosphere. These new findings demonstrate the benefits of studying the whole content of newly identified MAGs, and investigating emerging patterns in the environmental context.

Acknowledgements

We thank Julie Reveillaud for her comments on our manuscript. As two research parasites, we have the utmost respect and gratitude towards Rodriguez-R et al., Overholt et al., and others who studied the 2010 Deepwater Horizon oil spill and made their data publicly available.

Methods

Genomes and metagenomes. The sand metagenomes were generated by Rodriguez-R et al. [13] and includes four samples collected before the oil began to wash ashore (May 2010), eight samples collected during the oiling event (four in July 2010, and four in October 2010), and four samples collected after removal of oil from the beach (June 2011). Raw metagenomic sequencing data for these 16 samples are publicly available under NCBI BioProject ID PRJNA260285. Data for ten cultivar genomes from Overholt et al. [15] are publicly available under NCBI BioProject ID PRJNA217943. Cultures were grown using oil as the sole source of carbon from samples collected from Pensacola Beach and other Florida beaches affected by the oil spill. FASTA files for 48 MAGs [14], ten cultivar genomes, and anvi'o files to reproduce Figure 1 are publicly available via https://dx.doi.org/10.6084/m9.figshare.3199459.

Recovering the distribution patterns of genomes and generating protein clusters. After noise filtering (see [14] for details), short reads from each metagenomic sample were mapped back to cultivar genomes and MAGs using CLC Genomics Workbench (version 6) (http://www.clcbio.com) by requiring a minimum of 97% sequence identity over 100% of the read length. Anvi'o reported the mean coverage and portion coverage of each genome in all metagenomic datasets. To generate protein clusters we used ITEP [17] (with inflation and maxbit parameters set to 2.0 and 0.1) to profile GenBank files for the 10 cultivar genomes and 48 MAGs using a server running Linux CentOS version 6.4. Figure 1 reports the occurrence of protein clusters across genomes.

Taxonomic and functional annotation. We used the RAST platform [18] to infer the taxonomy and functionality of genomes, and to acquire GenBank files that contain the location and the translated protein sequence for open reading frames identified in each genome. We used STAMP [19] to perform ANOVA test to determine differentially occurring functions between groups of genomes as identified by protein clusters.

Generating the anvi'o display and network visualization. We used anvi'o v1.2.2 (available from http://github.com/meren/anvio) to analyze protein clusters and integrate the environmental data. We transformed the ITEP tab-delimited output for protein clusters into an anvi'o compatible format using the script `anvi-script-itep-to-data-txt`. We then created a hierarchical clustering of protein clusters based on their distribution across the 58 genomes using the script `anvi-matrix-to-newick`, which employed Euclidean distance as distance measure. We used `anvi-gen-samples-info-database` to create an additional anvi'o database to display basic features of genomes (i.e., GC-content, or genome length), as well as presence of select functions, and environmental distribution patterns across 16 metagenomic samples. Finally, we visualized all this information using the program `anvi-interactive`. User tutorials for the meta-pangenomic workflow is available online at the URL http://merenlab.org.

Competing interests

None to declare.

Author details

[1]The Department of Medicine, The University of Chicago, Chicago, 60637 IL, USA. [2]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, 02543 MA, USA.

References

1. Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., DeJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F., Trust, T.J.: Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature **397**(6715), 176–80 (1999). doi:10.1038/16495

2. Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Reeve, J.N.: Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. Journal of bacteriology **179**(22), 7135–55 (1997)

3. Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Acinas, S.G., González, J.M., Pedrós-Alió, C.: Ecology of marine Bacteroidetes: a comparative genomics approach. The ISME journal **7**(5), 1026–37 (2013). doi:10.1038/ismej.2012.169

4. Kumar, V., Sun, P., Vamathevan, J., Li, Y., Ingraham, K., Palmer, L., Huang, J., Brown, J.R.: Comparative genomics of Klebsiella pneumoniae strains with different antibiotic resistance profiles. Antimicrobial agents and chemotherapy **55**(9), 4267–76 (2011). doi:10.1128/AAC.00052-11

5. Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Lou, Y., Rohksar, D., Lucas, S., Huang, K., Goodstein, D.M., Hawkins, T., Plengvidhya, V., Welker, D., Hughes, J., Goh, Y., Benson, A., Baldwin, K., Lee, J.-H., Díaz-Muñiz, I., Dosti, B., Smeianov, V., Wechter, W., Barabote, R., Lorca, G., Altermann, E., Barrangou, R., Ganesan, B., Xie, Y., Rawsthorne, H., Tamir, D., Parker, C., Breidt, F., Broadbent, J., Hutkins, R., O'Sullivan, D., Steele, J., Unlu, G., Saier, M., Klaenhammer, T., Richardson, P., Kozyavkin, S., Weimer, B., Mills, D.: Comparative genomics of the lactic acid bacteria. Proceedings of the National Academy of Sciences of the United States of America **103**(42), 15611–6 (2006). doi:10.1073/pnas.0607117103

6. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**(6978), 37–43 (2004). doi:10.1038/nature02340

7. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O.: Environmental genome shotgun sequencing of the Sargasso Sea. Science (New York, NY) **304**(5667), 66–74 (2004)

8. Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M.: Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science (New York, N.Y.) **331**(6016), 463–7 (2011). doi:10.1126/science.1200387

9. Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., Armbrust, E.V.: Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science (New York, N.Y.) **335**(6068), 587–90 (2012). doi:10.1126/science.1212665

10. Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., Banfield, J.F.: Unusual biology across a group comprising more than 15% of domain Bacteria. Nature **523**(7559), 208–211 (2015). doi:10.1038/nature14486

11. Delmont, T.O., Eren, A.M., Maccario, L., Prestat, E., Esen, Ö.C., Pelletier, E., Le Paslier, D., Simonet, P., Vogel, T.M.: Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Frontiers in microbiology **6**, 358 (2015). doi:10.3389/fmicb.2015.00358

12. Zhu, C., Delmont, T.O., Vogel, T.M., Bromberg, Y.: Functional Basis of Microorganism Classification. PLoS computational biology **11**(8), 1004472 (2015). doi:10.1371/journal.pcbi.1004472

13. Rodriguez-R, L.M., Overholt, W.A., Hagan, C., Huettel, M., Kostka, J.E., Konstantinidis, K.T.: Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. The ISME journal (2015). doi:10.1038/ismej.2015.5

14. Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., Delmont, T.O.: Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ **3**, 1319 (2015). doi:10.7717/peerj.1319

15. Overholt, W.A., Green, S.J., Marks, K.P., Venkatraman, R., Prakash, O., Kostka, J.E.: Draft genome sequences for oil-degrading bacterial strains from beach sands impacted by the deepwater horizon oil spill. Genome announcements **1**(6) (2013). doi:10.1128/genomeA.01015-13

16. Spormann, A.M.: Gliding motility in bacteria: insights from studies of Myxococcus xanthus. Microbiology and molecular biology reviews : MMBR **63**(3), 621–41 (1999)

17. Benedict, M.N., Henriksen, J.R., Metcalf, W.W., Whitaker, R.J., Price, N.D.: ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC genomics **15**(1), 8 (2014). doi:10.1186/1471-2164-15-8

18. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O.: The RAST Server: rapid annotations using subsystems technology. BMC genomics **9**, 75 (2008). doi:10.1186/1471-2164-9-75

19. Parks, D.H., Beiko, R.G.: Identifying biologically relevant differences between metagenomic communities. Bioinformatics (Oxford, England) **26**(6), 715–21 (2010). doi:10.1093/bioinformatics/btq041