Recruiting Environmental Genomes from Metagenomes

Naseer Sangwan · Pushp Lata · Amit Kumar Singh ·
Rup Lal

Published online: 9 February 2012
© Association of Microbiologists of India 2012

Next generation sequencing (NGS) has enabled us to understand the extraordinary plasticity of prokaryotic genomes. Metagenome (qualitative and quantitative analysis of eDNA) beholds the great potential of unveiling the variable regions of the individual genotypes. At the species level genomes are comprised mainly of two types of genetic content, the core genome (present in all strains) and variable genome (present only in few strains). These strain specific gene inventories are commonly known as Genomic islands (GIs) and when they are called upon using metagenomic reads they are called as Metagenomic Islands (MIs). MIs can be further explained as habitat independent gene acquisitions of a parent genotype that have been already reported from various environments. Till date, various studies have been published which can clearly explain the potential of metagenomics to not only reveal the diversity and functional gene content of previously unknown microbial diversity but also the genome recruitment of any of the enriched species or genotype [4]. Preliminary taxonomical analysis of EGTs (environmental gene tags) against any of the public sequence database can reveal the relative abundance of the actual microbial diversity (at various ranks) present at that particular environment. After quantifying the genomic fragments of dominant species, metagenomic recruitment plots can be generated using whole metagenomic data. Additionally, individual reads can be binned at various percentage identities using already sequenced neighbour as the reference genome. De novo assembly followed by assembly validation can further help to taxonomically identify the

assembled genotype. Here we are sharing (ahead of publication) a metagenomic recruitment plot of *Sphingobium chlorophenolicum* (Fig. 1) generated by aligning metagenomics reads of a HCH contaminated dumpsite soil (un-published data Sangwan and Lal et al.). Although sequencing depth is lower, one can easily observe the genomic fragments that are completely absent or misrepresented in the following metagenomic dataset. MIs were identified and checked for their functional status.

Metagenomic reads can be reference assembled directly or contigs/scaffolds from de novo assembly of binned EGTs can be reference mapped to the present reference genome sequences. Several strategies have been used to assemble, validate and taxonomically characterizing the contigs/scaffolds obtained from de novo or reference based metagenome assemblies [1]. In acid mine drainage (AMD) analysis Tyson et al. used a whole genome assembler algorithm to assemble the long, capillary sequencing reads. Later, scaffolds were assigned to microbial groups using their average GC% percentage [1]. Assembly was validated using mate pair constraints and taxonomy was decided using partial 16S rRNA sequence analysis. As the microbiota of the acid mine drainage was inhabited by only one or two naturally selected microbial groups extensive genome assembly was possible with their accurate taxonomical assignments. Where as in Permafrost analysis [5] a draft genome assembly of a previously unknown methanotroph was obtained using De-Bruijn based algorithm and single copy (total 19 genes) gene based validation. Total metagenomic data (pre-processed) was assembled on 41 K-mer length and contigs were taxonomically characterized (at various ranks) using MEGAN [2]. Coverage for individual contigs was calculated via remapping the metagenomic reads on contigs. It was observed that environmental methanotroph assembled was representing the dominant

N. Sangwan · P. Lata · A. K. Singh · R. Lal (✉)
Molecular Biology Laboratory, Department of Zoology,
University of Delhi, Delhi 110007, India
e-mail: ruplal@gmail.com

Fig. 1 *Sphingobium chlorophenolicum* genome and metagenomic islands. Metagenomic reads from a HCH (Hexachlorocyclohexane) contaminated dumpsite soil were aligned on the reference genome along with the coverage parameter. Each spot on the graph represents a metagenomic read aligned with respect to its percentage identity (y-axis). Numerically labelled regions are the metagenomic islands with minimum or completely absence of mapped reads

population responsible for oxidation of the methane produced after defrosting of the ice. It was the first successful attempt to assemble a draft genome from a complex matrix (soil).

The most discussed and still un-standardized parameter in the target metagenome assembly is the coverage cut-off. In acid mine drainage study scaffolds with $3\times$ to $10\times$ coverage were assembled and assigned to various microbial groups. Whereas in permafrost analysis average contig coverage was $>73\times$. Recently Luo et al. [3] have published some methodological standards for a individual genome assemblies from short metagenomic sequencing data. Metagenome assemblies were performed with simulated metagenomic reads spiked with sequence reads of a target genome. Reference genome coverage was plotted against several parameters (single base call error, N50 length and non target sequence) and it was observed that a minimum of $20\times$ coverage is required for a robust and valid metagenome assembly, targeting a specific genotype. Similar analysis was performed on non-simulated metagenomic datasets. Since the current studies were based upon the short sequence reads (illumina) their relevance with other NGS platforms are still to be evaluated. In summary, all of the current metagenomic assembly efforts are qualitatively in agreement with the fact that assembling and validating an environmental genome from metagenomic dataset is

still not a trivial task. Latest sequencing approaches like SIP (stable isotope probing) followed by whole genome sequencing and single cell genomics can really help to assemble a complete environmental genome (enriched) from a metagenomic dataset considering the coverage and sequencing artefacts.

References

1. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
2. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386
3. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. Individual genome assembly from complex community short-paired metagenomic datasets. *ISME J* doi:10.1038/ismej.2011.147
4. Lejla P, Rodriguez-Mueller B, Martin-Cuadrado A-B, Mira A, Rohwer F, Rodriguez-Valera F (2009) Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics* 10:570
5. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480:368–371