ORIGINAL ARTICLE 10.1111/j.1469-0691.2012.03856.x

Evolution of genetic diversity using networks: the human gut microbiome as a case study

E. Bapteste, C. Bicep and P. Lopez

UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

Abstract

In order to study complex microbial communities and their associated mobile genetic elements, such as the human gut microbiome, evolutionists could explore their genetic diversity with shared sequence networks. In particular, the detection of remarkable structures in gene networks of the gut microbiome could serve to identify important functions within the community, and would ease comparison of data sets from microbiomes of various sources (human, ape, mouse etc.) in a single analysis.

Keywords: Evolution, exploratory science, genetic diversity, microbiome, network

Original Submission: 22 November 2011; **Revised Submission**: 28 November 2011; **Accepted:** 20 March 2012

Editors: A. Moya, Rafael Cantón, and D. Raoult

Clin Microbiol Infect 2012; **18** (Suppl. 4): 40–43

Corresponding author: E. Bapteste, UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

E-mail: eric.bapteste@snv.jussieu.fr

Introducing exploratory studies

The vast majority of genetic diversity is currently unknown. Most of it comprises mobile elements (phages and plasmids) and microbial communities that cannot be cultured under laboratory conditions. This situation also occurs for the human gut microbiome, a complex 'organ' (or ecosystem) with about 100 times as many genes as the human genome, about 70% of the protein coding genes without known homologues [1], a resident population of mobile genetic elements, and a high turnover of at least some of its members, of which 80% are uncultured. With so many unknowns in a biological system scientists can expect many original discoveries. For example, studies of the gut microbiome could unravel new gene forms, motifs, processes, functions, interactions and multi-level organizations affecting genetic diversity, and unravel links between the environment, diet, composition and function of the microbiome. How to enhance these discoveries is a particularly motivating issue.

One strategy to handle massive amounts of unknowns and an overwhelming wealth of data is called exploratory studies. Such studies go from (microbiome) data to hypotheses and rely heavily on the experimental design of most inclusive methods of genetic diversity, which seek for patterns in huge data sets with the fewest assumptions possible to ease the discovery of unrecognized regularities, phenomena and interactions. The assumed goal of exploratory studies is to foster the discovery of many unrecognized patterns and to actively generate novel hypotheses, in our case about genetic diversity [2]. In this approach, biologists do not know what types of results they will find, but can expect truly 'original' findings. As such, exploratory approaches differ from standard (or targeted) approaches that go from hypotheses to (microbiome) data and either support or reject pre-existing hypotheses.

In evolutionary biology, the standard approach is centred on the reconstruction of species and gene trees to organize the analysis of genetic diversity. The tree hypothesis *a priori* constrains the patterns and the processes to be identified and the discoveries to be made in a data set (e.g. genealogical relationships between taxa and genes in the microbiome). However, ecosystems like a microbiome with 10–100 trillion cells do not fit on a single branch on a tree. Moreover, transfer of genetic material between the mobile elements and the various lineages occupying the gut also creates an important evolutionary dynamic that is poorly captured by a

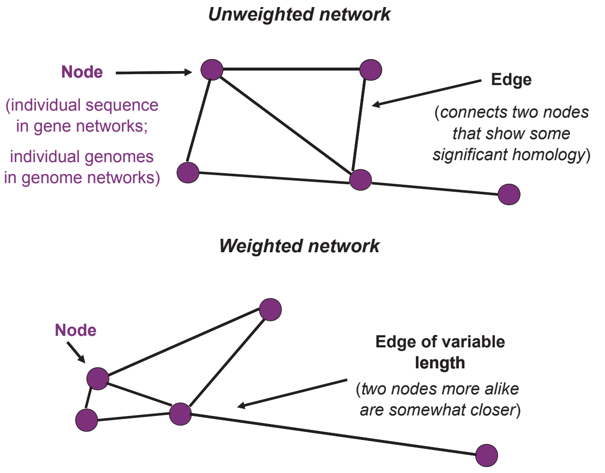
Unweighted network
Node
(individual sequence in gene networks;
individual genomes in genome networks)
Edge
(connects two nodes that show some significant homology)
Weighted network
Node
Edge of variable length
(two nodes more alike are somewhat closer)

FIG. 1. Scheme of shared sequence networks. Nodes are indicated by circles, edges by links between the circles.

tree-based model [3,4]. Consequently, we argue that exploratory evolutionary analyses, letting go of some phylogenetic assumptions, could be a desirable option to study the gut microbiome. We introduce a less constrained approach based on networks that could enhance (evolutionary) discoveries about gut microbiomes.

Exploring genetic diversity with networks

Network-based methods based on sequence similarities have recently started providing fast and heuristic pictures of genes, genome evolution and the evolution of communities for various microbes, mobile elements and environments [5–11]. Such networks are graphs connecting nodes by edges, when the objects at the nodes share some similarity in their

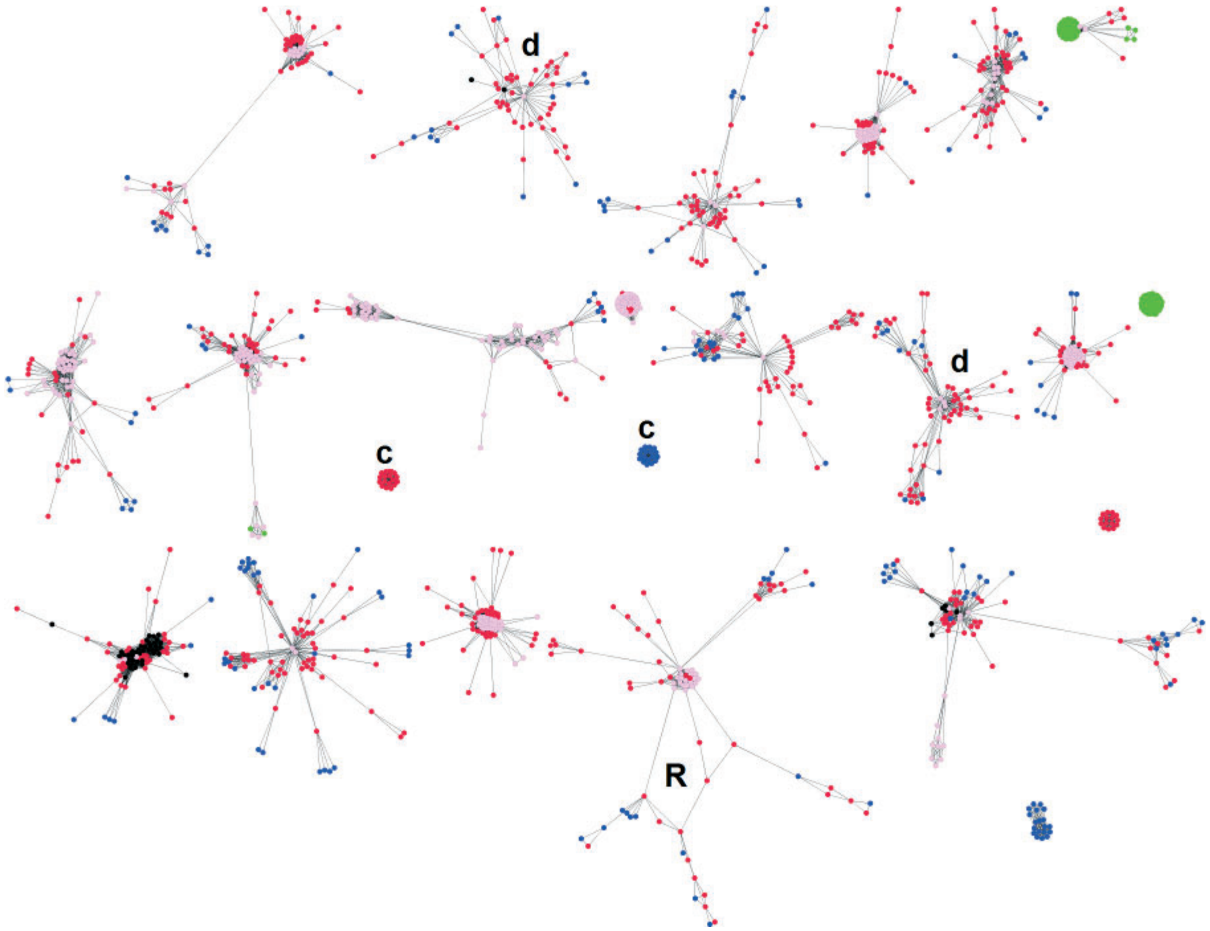

FIG. 2. Sample of the gene network for the human gut microbiome. Two sequences (nodes) are connected when they share significant homology (a BLAST threshold of <1e-5 and at least 20% identity in their aligned portions). Individual gene families correspond to separated subgraphs (connected components) with red/blue nodes for sequences from the Japanese/American gut microbiomes, respectively. Sequences of integrons, phages and plasmids are indicated in green, pink and black, respectively. The letter c illustrates potential conserved families, the letter d potential divergent families, and the letter R potential recombined families.

sequences (Fig. 1). For instance, in genome networks, two nodes (genomes) are connected when they share at least a gene family (i.e. two genomes of *Escherichia coli*, each with a copy of a glucose dehydrogenase, will be connected). In gene networks [5], two nodes (individual sequences) are connected when they display more than a certain threshold of similarity (i.e. two glucose dehydrogenases will be connected when their reciprocal best BLAST score is <1e-5 and/or when they display >70% sequence identity). Importantly, within a gene network, many disconnected subnetworks are obtained, because many genes families are unrelated (i.e. glucose dehydrogenases have no homology with ribosomal proteins), therefore defining distinct connected components.

Using MetaGeneAnnotator [12], we predicted 311 265/195 521 genes in Japanese [1] and American [13] gut microbiomes, respectively. To study the evolution of their genetic diversity, they were included in a gene network with sequences of all the phages, plasmids and integrons publicly available (for a total of 748 688 sequences). The resulting network showed a huge genetic diversity, identifying connected components of various sizes and shapes (Fig. 2). Since networks are mathematical objects, the topology of these various connected components can be exploited to sort the connected components (hence the gene families) by describing the connectivity and relationships of their nodes, as well as their coefficient of clustering. Using such centralities, it is straightforward to identify various types of gene families: conserved ones, divergent ones, recombined ones etc. (Fig. 2; see also for an instance of a conserved family the translation initiation factor I, of a divergent family the type V secretory pathway proteins, of a recombined family the type I restriction endonuclease S subunit in Fig. 2 in [11]).

Connected components can also be sorted based on their composition (i.e. when they comprise only sequences from Japanese or American gut microbiomes, or from both of these microbiomes). This sorting, although still based on a limited data set, shows a very high genetic diversity in the human gut microbiome: only 39% of the 118 489 'American' genes and the 207 443 'Japanese' genes fell in shared gene families; 16 991 of the gene families that produced connected components (35% of the data) were only found in Japanese gut microbiomes; 12 644 (26% of the data) were only found in American gut microbiomes. The latter numbers are certainly too large to imagine that finding gene families in Japanese gut microbiomes simply reflects differences in diet. Not all such genes, if any, may be 'sushi genes' [14].

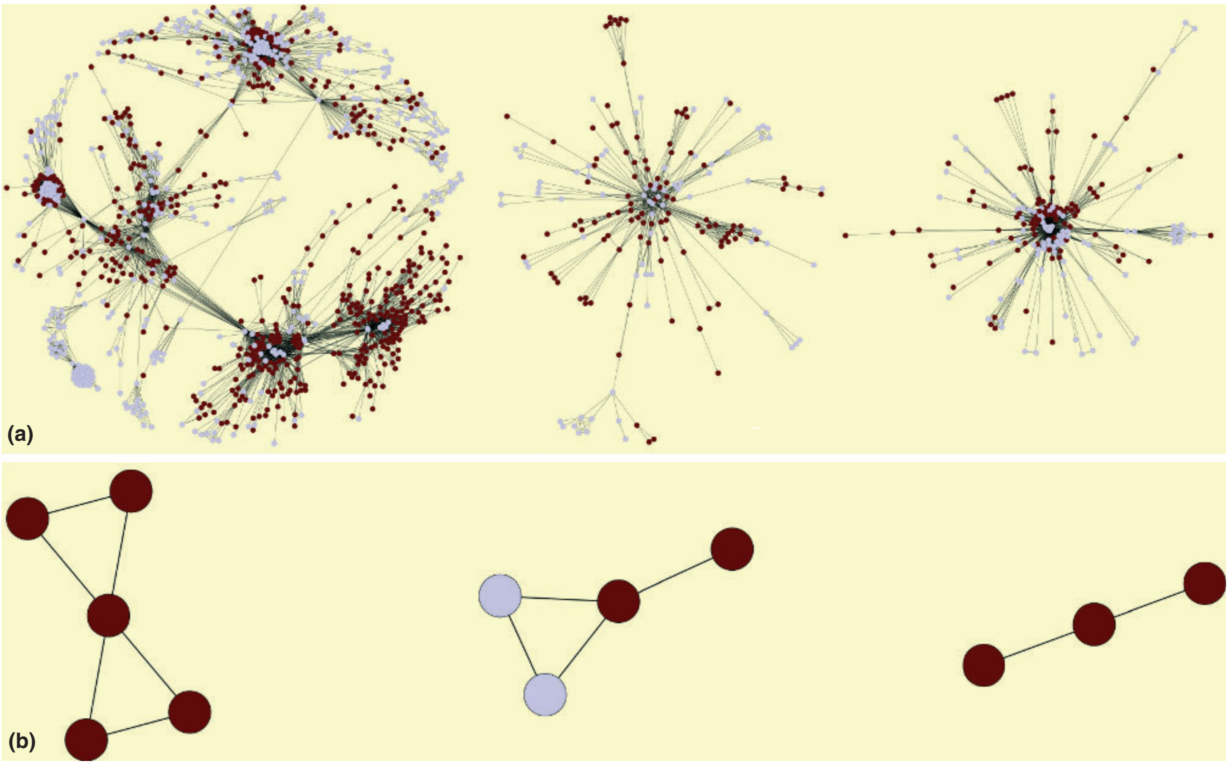

FIG. 3. Typical connected components for genes involved in different functions. (a) Three connected components of genes involved in the metabolism of carbohydrates, from left to right. (b) Three connected components of genes involved in the metabolism of cell motility and chemotaxis, from left to right. The nodes in brown were annotated to fulfil these functions using MG-RAST; nodes in grey had no known functions.

Processes and functions structure the evolution of genetic diversity

Interestingly, in our network, 8.8% of the connected components (4296 gene families) mixed sequences from the human gut microbiome with sequences from mobile elements, suggesting that these gene families may be mobile. If so, plasmids and associations of mobile elements appear to play a prevalent role in the mobilization of genes in human gut microbiomes. Remarkably about 10% of the mobile genes are carried not only by one type of DNA vehicle (phage, plasmid or integrons) but by many. These results confirm that lateral gene transfer plays an important role in the convergence and expansion of the gene sets of gut microbiomes. Also, the presence of some particular genes and functions may be more important than the presence of some particular species in the gut microbiome. Thanks to the mobility of these gene families the song (the function) sometimes matters more than the singer (the species that fulfil the function), as the singer can be replaced [15]. If true, function and gene transfer structure genetic diversity in gut microbial communities.

The latter claim seems supported when connected components are sorted by function. Some gene families with different functions evolve differently in the gut microbiome: they present different topologies. For instance, gene families involved in the metabolism of carbohydrate are much more diversified (e.g. large diameter, higher number of nodes) than gene families involved in cell motility and chemotaxis, which present smaller connected components with a very reduced diversity (Fig. 3). This result reflects that carbohydrate metabolism is much more important than cell motility in the gut (since the uptake of carbohydrate imposes a regular selective pressure on gut microbial communities, while bacteria are naturally stirred up in the guts). What is exciting then is to further explore gene networks of the microbiomes to determine what other functions are associated with the other most diverse connected components, and to include always more microbiome data in the analyses.

Transparency Declaration

Nothing to declare.

References

1. Kurokawa K, Itoh T, Kuwahara T et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 2007; 14: 169–181.
2. Burian R. *Experimentation, exploratory*. In: Springer Encyclopedia of Systems Biology, Springer, Berlin, 2011.
3. Bapteste E, O'Malley MA, Beiko RG et al. Prokaryotic evolution and the tree of life are two different things. *Biol Direct* 2009; 4: 34.
4. Doolittle WF, Bapteste E. Pattern pluralism and the tree of life hypothesis. *Proc Natl Acad Sci U S A* 2007; 104: 2043–2049.
5. Bittner L, Halary S, Payri C et al. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biol Direct* 2010; 5: 47.
6. Dagan T, Martin W. Getting a better picture of microbial evolution en route to a network of genomes. *Phil Trans R Soc Lond B Biol Sci* 2009; 364: 2187–2196.
7. Fondi M, Fani R. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ Microbiol* 2010; 12: 3228–3242.
8. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A* 2010; 107: 127–132.
9. Kloesges T, Popa O, Martin W, Dagan T. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol Biol Evol* 2010; 2: 1057–74.
10. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* 2008; 25: 762–777.
11. Beauregard-Racine J, Bicep C, Schliep K, Lopez P, Lapointe FJ, Bapteste E. Of woods and webs: possible alternatives to the tree of life for studying genomic fluidity in *E. coli*. *Biol Direct* 2011; 6: 39.
12. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 2008; 15: 387–396.
13. Turnbaugh PJ, Hamady M, Yatsunenko T et al. A core gut microbiome in obese and lean twins. *Nature* 2009; 457: 480–484.
14. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* 2010; 464: 908–912.
15. Doolittle WF, Zhaxybayeva O. Metagenomics and the units of biological organization. *Bioscience* 2010; 60: 102–112.