

RESEARCH

Open Access



# Understanding Horizontal Gene Transfer network in human gut microbiota

Chen Li<sup>†</sup>, Jiaying Chen<sup>†</sup> and Shuai Cheng Li<sup>\*†</sup>

## Abstract

**Background:** Horizontal Gene Transfer (HGT) is the process of transferring genetic materials between species. Through sharing genetic materials, microorganisms in the human microbiota form a network. The network can provide insights into understanding the microbiota. Here, we constructed the HGT networks from the gut microbiota sequencing data and performed network analysis to characterize the HGT networks of gut microbiota.

**Results:** We constructed the HGT network and perform the network analysis to two typical gut microbiota datasets, a 283-sample dataset of Mother-to-Child and a 148-sample dataset of longitudinal inflammatory bowel disease (IBD) metagenome. The results indicated that (1) the HGT networks are scale-free. (2) The networks expand their complexities, sizes, and edge numbers, accompanying the early stage of lives; and microbiota established in children shared high similarity as their mother (p-value = 0.0138), supporting the transmission of microbiota from mother to child. (3) Groups harbor group-specific network edges, and network communities, which can potentially serve as biomarkers. For instances, IBD patient group harbors highly abundant communities of *Proteobacteria* (p-value = 0.0194) and *Actinobacteria* (p-value = 0.0316); children host highly abundant communities of *Proteobacteria* (p-value =  $2.8785e^{-5}$ ) and *Actinobacteria* (p-value = 0.0015), and the mothers host highly abundant communities of *Firmicutes* (p-value =  $8.0091e^{-7}$ ). IBD patient networks contain more HGT edges in pathogenic genus, including *Mycobacterium*, *Sutterella*, and *Pseudomonas*. Children's networks contain more edges from *Bifidobacterium* and *Escherichia*.

**Conclusion:** Hence, we proposed the HGT network constructions from the gut microbiota sequencing data. The HGT networks capture the host state and the response of microbiota to the environmental and host changes, and they are essential to understand the human microbiota.

**Keywords:** HGT network, Scale free, Von Newman entropy, Network evolving, Community analysis

## Introduction

Human gut microbiota is a complex ecosystem consisting of a total of  $10^{14}$  bacteria [1]. Species richness makes gut microbiota harbor diverse metabolic functions and robust to disturbances, such as the invasion of pathogenic bacteria while maintaining host health. However, factors [2], such as age [3], environment [4], diet [5], can lead to significant shifts in the composition of the individual

microbiome over longitudinal periods. Microbes may share genetic materials with others to get beneficial traits through Horizontal Gene Transfer (HGT) to gain better adaptations. HGTs allow microbes to acquire genes from distant species that are not in a parent-offspring relationship [6]. It increases the genetic diversity of recipients and plays a vital role in the evolution of microbes. The complex interaction among species indicates that gut microbiota is a complex system.

Researchers have developed tools to detect HGTs. GIST [7] identified HGT events by searching for regions with different genomic signatures from the genome average, and it can detect HGT among distantly related

\*Correspondence: shuaicli@cityu.edu.hk

<sup>†</sup>Chen Li and Jiaying Chen contributed equally to this work  
Department of Computer Science, City University of Hong Kong, Hong Kong, Hong Kong SAR, China



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

genomes. While for some ancient HGT events, genomic signatures are similar to background subjected to the mutation process [8] and are hard to be recognized by GIST. Darkhorse [9] and HGTector [10] identified HGT events in genomes showing taxonomically discordant similarity to genes within a reference database. Compared to GIST, Darkhorse and HGTector are biased toward more ancient HGT events and not make use of phylogenetic trees to test for phylogenetic incongruence. MetaCHIP [11] identified HGT events based on best-match and phylogenetic incongruency approaches [12]. The main challenge of MetaCHIP is to detect recent HGT events [13]. Essentially, HGT is the insertion of foreign genes into recipient genomes. It is similar to inter-chromosomal translocation in an organism with multiple chromosomes [14]. Therefore, HGT can be treated as a kind of complicated structure variation (SV). No matter ancient or recent HGT events could be identified by mapping NGS reads against reference genomes. LEMON [15] applies split reads re-alignment [16] and DBSCAN [17] to detect HGT events.

Network science is a powerful tool to analyze and understand such complex systems. Various mathematical models are available to analyze and quantify such systems [18–22]. In order to use networks to model the human gut microbiota, one method calculates the correlation coefficients for the abundances of each pair microbes. The correlation matrix forms a network [23]. The positive and negative correlation indicates the two microbes may have cooperation or competition interact, respectively. However, this network model is incapable of capturing the potential mechanism for the interactions. Another method is based on the metabolic exchange between microbes to construct the Global metabolic interaction network of the human gut microbiota [24]. It requires annotations of enzymes and metabolic pathways. The HGT networks represent how human gut, oral, nasopharyngeal skin microbiomes share genetic material to adapt to the environment. They provide new insights into the dynamics of the microbiota. Through analyzing the HGT networks, we may gain new insights into the communities assemble, species interactions, and host-associated selection pressure on the microbiota. Kunin et al. [25] constructed the HGT network based on the reconstructed phylogenetic trees. It utilized protein sequences as evolutionary units, limiting its ability to detect HGT events in regions outside or across gene boundaries [26]. While these HGT events could be detected by LEMON, which takes whole metagenomic sequencing data as input, thus, LEMON could be used to create a complete HGT network.

We apply the HGT networks to study two typical human gut microbiota datasets.

The first is a 283-sample dataset of Mother-to-Child [3]. The development of infant gut microbiota plays an important role in establishing a healthy host–microbiome symbiosis, including the maturation of the immune system [27], nutrient utilization [28], and so on. The dynamic microbiota grows rapidly and is affected by factors such as delivery mode [29] and feeding [30]. As the importance of infant gut microbiota has been realized, how and where an infant acquires these microbes attract increasing attention. The maternal gut bacteria is considered as one important source, but the vertical inheritance remains largely unexplored. Ferretti et al. [31] utilized strain-level metagenomic profiling to track the mother-to-infant bacterial transmission; they find strains shared within the mother and infant pairs. However, their research is unable to capture the significant similarity of the gut microbiome in mother–child pairs by using Bray–Curtis dissimilarity.

The second is a 148-sample dataset of longitudinal Inflammatory bowel disease (IBD) metagenome [32]. The composition of gut microbiota is related to many diseases [33], such as IBD [34] and type 1 diabetes [35]. IBD is one most common groups of chronic inflammatory disorders affecting millions of people. The cause of IBD is associated with human genetic mutation and gut microbiota. The gut microbiota plays an important role in IBD [36]. The increasing of *Bacteroidetes* and decreasing of *Firmicutes* are observed in IBD patients [37]. Since gut microbiota is dynamic, the longitudinal gut microbiota is analyzed to capture the variation of gut microbiota composition over time in IBD patients [38]. The linkage between metagenomic functional potential and functional activity is built in recent research [32]. However, these work fail to model gut microbiota from a systematic perspective, which motivates us to apply HGT network analysis.

Our results consist of three aspects. First, we investigate the general characteristics of the HGT network. HGT network is a complex network, and we used power-law distribution and three heavy tail distributions to fit the network. The result demonstrates that the HGT networks are scale-free, which implies the HGT network holds important properties such as ultra-small world property and robust to random disruption. Second, we studied the dynamic change of HGT networks, especially for the Mother-to-Child data set. The increasing complexity (Von Newman Entropy), network size, and HGT event rate in child HGT networks accompany the growth of child gut microbiota during the first three months after birth. Furthermore, the high structural similarity ( $p$ -value = 0.0138) between the child and maternal HGT networks supports the transmission of microbiota from mother to child. Third, we analyze

phenotype-specific HGT communities and HGT edges. As for HGT communities, compared with control individuals, the phylum composition of IBD-specific HGT communities have significant increasing of *Proteobacteria* (p-value = 0.0194) and *Actinobacteria* (p-value = 0.0316). Compared with mother, the phylum composition of child-specific HGT communities have significant decreasing of *Firmicutes* (p-value =  $8.0091e^{-7}$ ) [39] and increasing of *Proteobacteria* (p-value =  $2.8785e^{-5}$ ) and *Actinobacteria* (p-value = 0.0015) [40, 41]. As for conserved HGT edges across multiple samples, IBD patients have increased HGT edges in pathogenic genus including *Mycobacterium*, *Sutterella*, and *Pseudomonas*, compared to non-IBD individuals. Child-specific HGT edges are mainly from *Bifidobacterium* and *Escherichia*. These differences imply the alteration of gut microbiota caused by the change of selection pressure.

As we can see, both in Mother-to-Child and longitudinal IBD data sets, by analyzing temporal HGT networks, we captured the significant changes of HGT networks. These changes reflect the alteration of gut microbiota under the change of host-associated selection pressure. Therefore, the HGT network is an effective model to describe the relationship between the gut microbiota and the host state. It provides a new perspective to observe the change of gut microbiota in the everchanging environment.

## Results

We applied HGT networks analysis to two human gut longitudinal metagenomic sequencing datasets: Mother-to-Child data set [3] and longitudinal IBD data set [32]. As described in [3], the Mother-to-Child data set contains 283 samples that are collected from 44 Finish families. Thirty-three families have children sampled at five-time points: birth, two weeks, and one, two, and three months and mother sampled at three-time points: gestational week 27, birth, and three months post-delivery. The remaining 11 families have children sampled at birth and mother sampled at gestational week 27 and birth. As described in [32], the longitudinal IBD data set contains 148 samples spanning 26 participants: 15 patients with Crohn's disease (CD), eight patients with ulcerative colitis (UC) and three non-IBD controls, here CD and UC are two main forms of IBD.

Following the method in, we constructed one HGT network per sample. Then we obtained 283 HGT networks from the Mother-to-Child data set and 148 HGT networks from the longitudinal IBD data set. Figure 1a illustrates the distribution of the size of HGT networks constructed from the two datasets. Network size is the number of nodes in the HGT network. The average child form individual network size of infant HGT networks

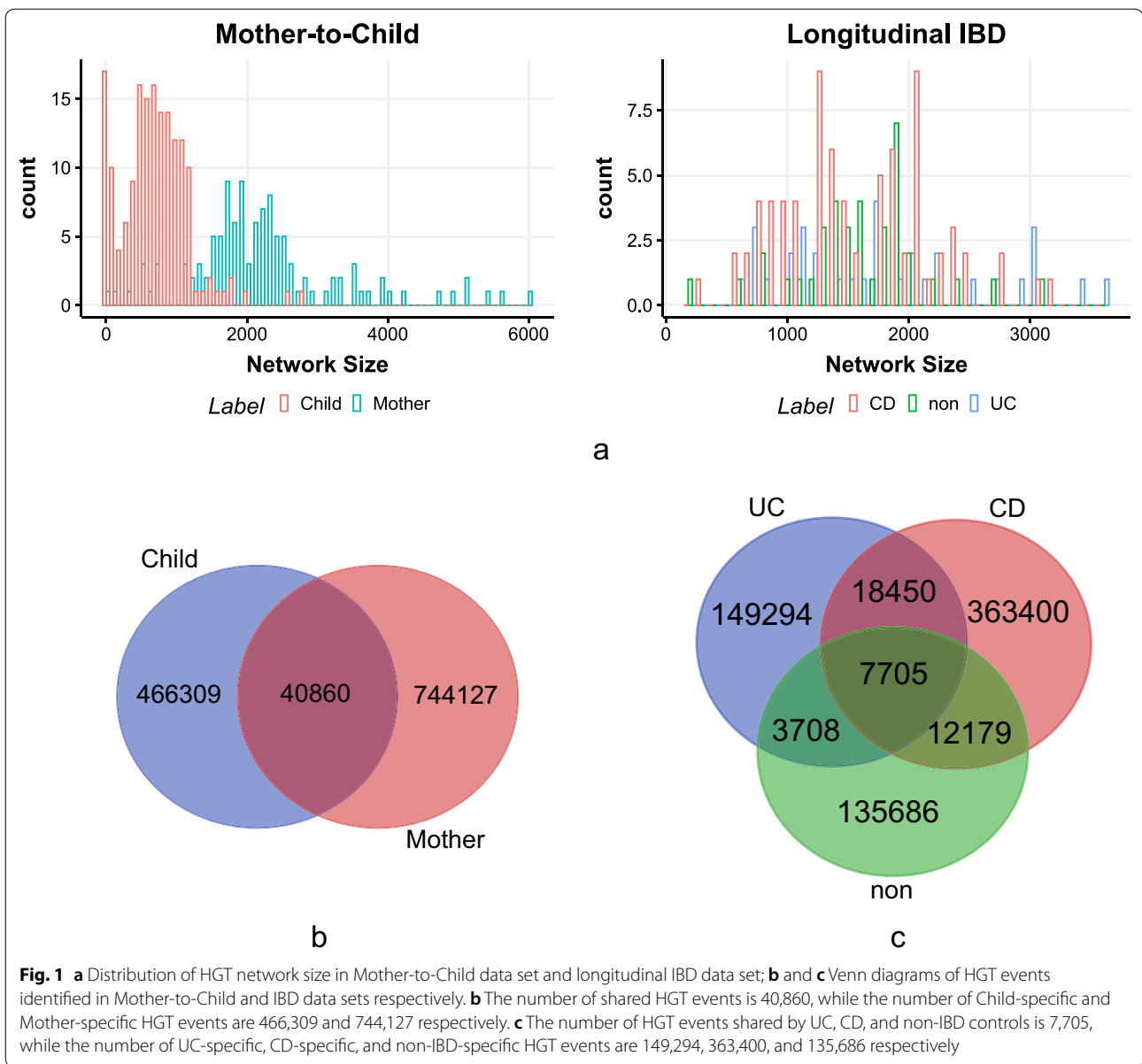
is 702.57, and the average network size of maternal HGT networks is 2228.12. There is an overlap between the distribution of network size for the HGT networks detected from CD, UC, and non-IBD controls individuals. Their average network sizes are 1604.71, 1795.03, and 1611.66, respectively. Figure 1b, c are Venn diagrams of HGT events identified in Mother-to-Child and IBD data sets, respectively. Common HGT events shared by different groups are much less than group-specific HGT events. In the Mother-to-Child data set, the number of shared HGT events is 40,860, while the number of Child-specific and Mother-specific HGT events are 466,309 and 744,127, respectively. In Longitudinal IBD, the number of HGT events shared by UC, CD, and non-IBD controls is 7705, while the numbers of UC-specific, CD-specific, and non-IBD-specific HGT events are 149,294, 363,400, and 135,686, respectively. So, HGT event overlap could effectively measure the overlap between HGT networks under different conditions.

## HGT networks are scale free

Our study implies that the HGT network is scale-free. It is supported by the result that the degree distribution of the HGT network is better fitted by power-law than the other three heavy tail distribution [42]. We filtered out HGT networks with less than 100 nodes to ensure the remaining HGT networks have enough degree data to fit. Finally, we collected 256 HGT networks from the Mother-to-Child data set and 147 HGT networks from the longitudinal IBD data set. We applied *powerlaw* package [43] to estimate degree distribution of HGT networks.

The evaluation of the goodness of fit for power-law distribution is described in "Method" section. As illustrated in Fig. 2a, 100%, 94%, and 92% of HGT networks are better fitted by power-law than exponential, lognormal\_positive, and Weibull respectively in Mother-to-Child data set. Meanwhile, Fig. 2b demonstrates that in Longitudinal IBD data set 99%, 94%, and 88% of HGT networks are better fitted by power-law than exponential, lognormal\_positive, and Weibull. Therefore, the HGT network is scale-free since HGT networks have degree distribution better fitted by power-law than the other three heavy tail distributions. Besides, as we can see in Fig. 5, a vast number of nodes of the HGT network has a small degree. They are connected to a few hub nodes. These hub nodes would have a very large degree. Thus, the degree of a randomly selected node would be tiny or arbitrarily large, which means HGT networks do not have a meaningful internal scale [44]. It explains the scale-free property of the HGT network.

HGT networks in different groups harbor different distributions of fitted exponents  $\alpha$  which is the



parameter of power-law distribution. Figure 3a compares exponents  $\alpha$  fitted to maternal and child HGT networks respectively. The distribution of fitted exponents corresponding to Mother is significantly different to the one corresponding to child ( $p = 0.038$ , Student's t-test) and has larger mean value  $\mu_{mother} > \mu_{child}$ . Figure 3b compares exponents  $\alpha$  fitted to HGT networks of non-IBD, UC, and CD respectively. Three exponent distributions are significantly different to each other ( $p = 0.009$ , Student's t-test) and have  $\mu(UC) > \mu(CD) > \mu(Non-IBD)$ .

As shown in Fig. 3, the 95% Confidence Intervals  $CI_{95\%}$  of fitted exponent  $\alpha$  in all groups are in region (2, 3).

So HGT networks hold the ultra-small world property, which implies that HGT networks tend to form dense sub-graphs. Such a network structure implies that the HGT network is robust and could maintain a stable status [45].

#### Ultra-small world property of HGT networks

Since the degree exponent  $\alpha$  of the HGT network satisfies  $2 < \alpha < 3$ , HGT networks have the property of the ultra-small world [46]. The ultra-small world property means that the diameter  $d$  of HGT networks has a linear relationship with  $\ln \ln N$ , where  $N$  is the number of nodes in the network,  $\ln$  is the natural logarithm.

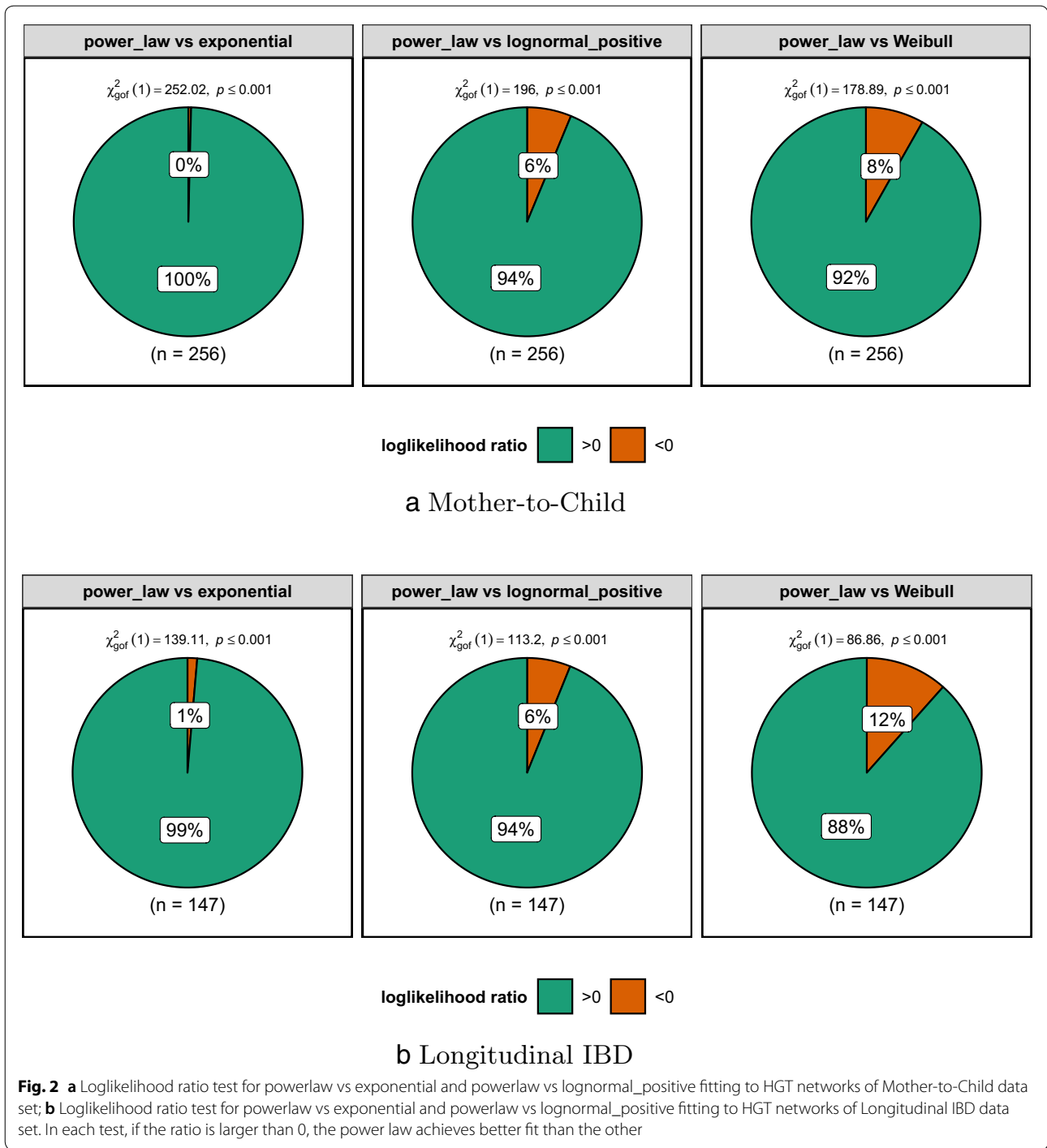
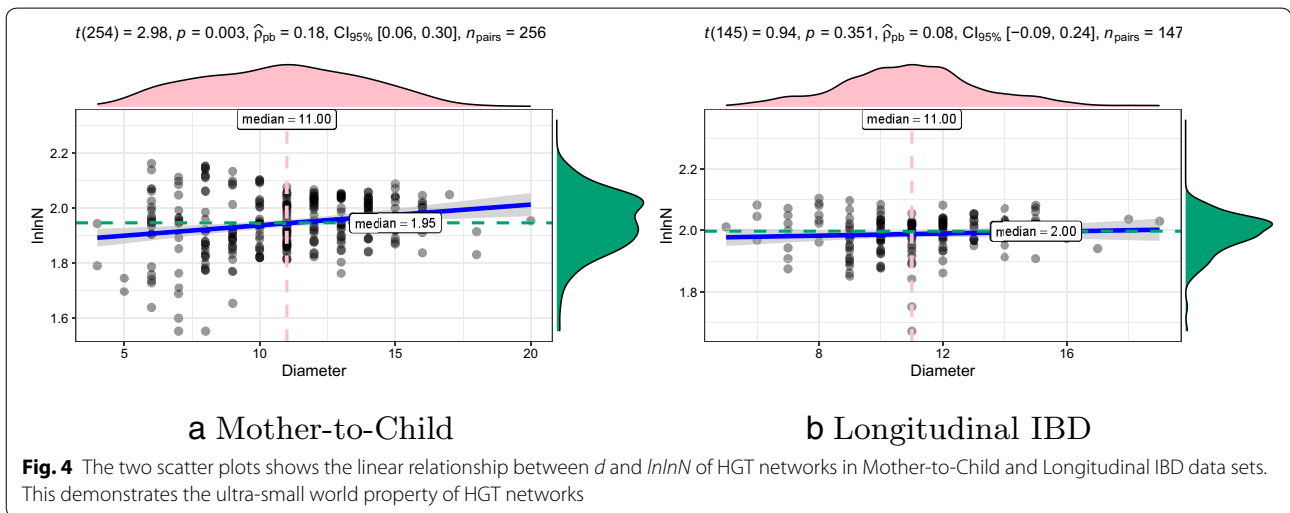
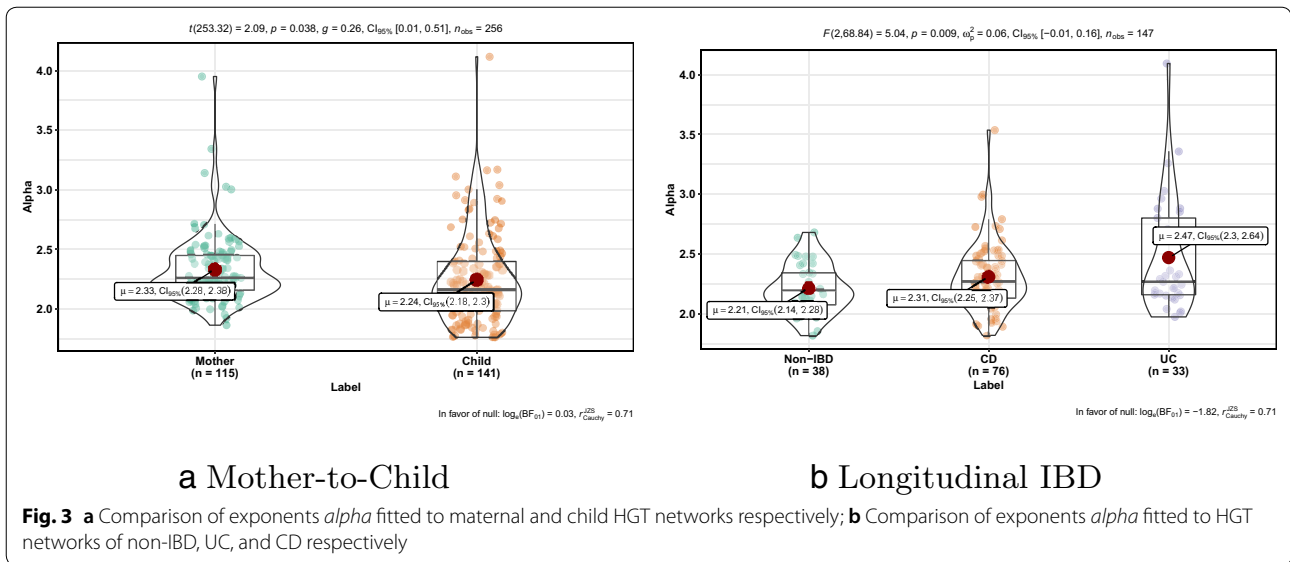


Figure 4 shows the linear relationship between  $d$  and  $\ln \ln N$  of HGT networks in Mother-to-Child and Longitudinal IBD data sets. The ultra-small diameter  $d$

increases as  $\ln \ln N$ , which is significantly slower growth than the  $\ln N$  derived for random networks. Therefore, the average distance in an HGT network is smaller than



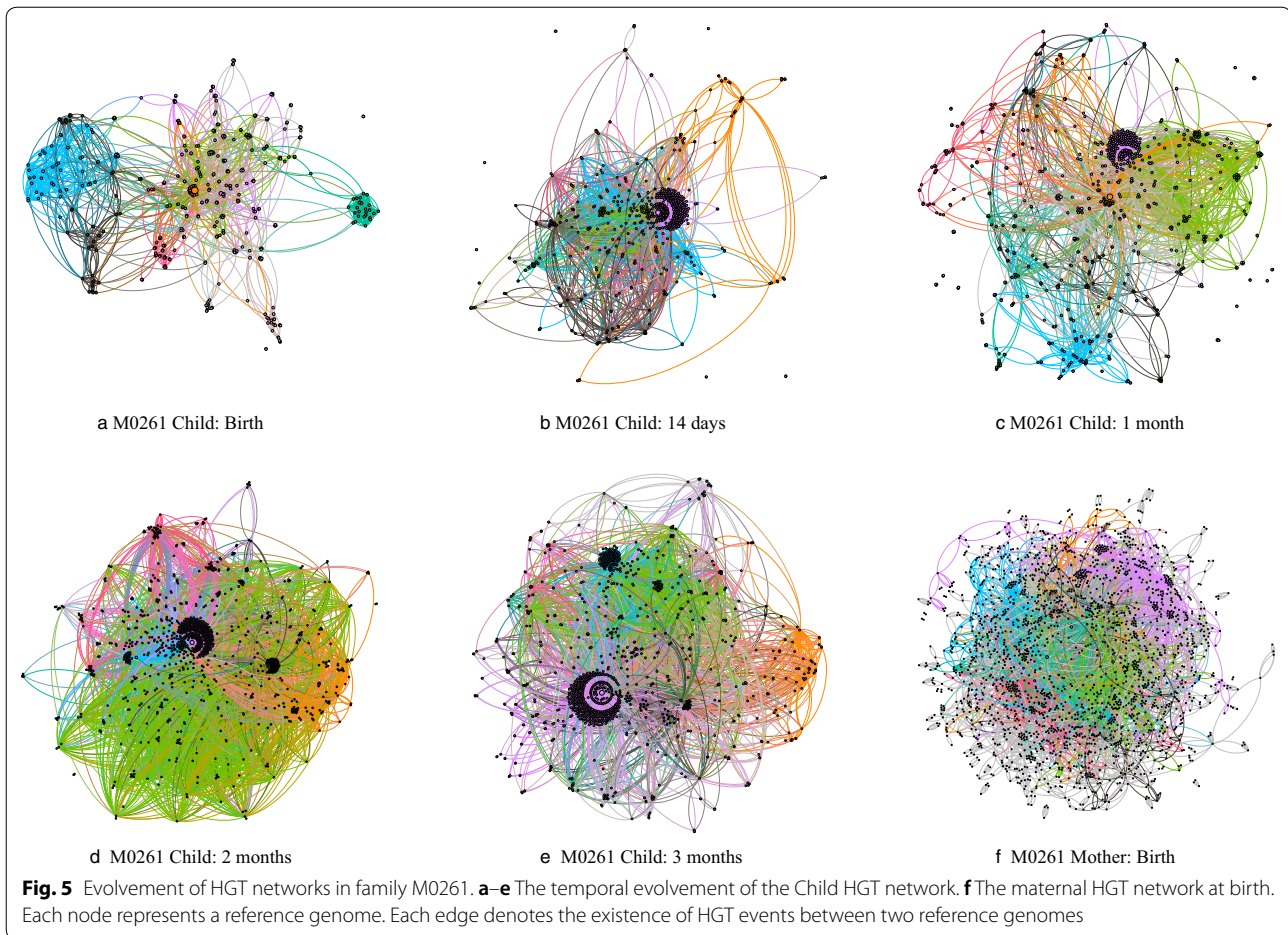
that in a random network, which imply the ultra-small world property of HGT networks.

**Analysis of Mother-to-Child HGT networks**  
**HGT network evolves**

We studied the evolvement of HGT networks across time from two aspects: network complexity and network similarity. Network complexity is measured by Von Newman entropy, network size, and HGT event rate. Network similarity is measured by the Jaccard index and degree correlation. We observed, first, child HGT networks have increasing network complexity in the first three months after birth. Second, maternal and child HGT networks share a family-specific significant similarity. Last, the individual-specific similarity of the HGT

network between different time points in IBD patients is significantly larger than that in non-IBD individuals. The evolvement of the infant HGT network describes the growth of gut microbiota. Figure 5 shows the temporal evolvement of child HGT networks across five-time points and the maternal HGT network at birth. As time goes by, the child HGT network becomes bigger and more complex due to the growth of child intestinal microbial strains harboring HGT events. Compared to child HGT networks, the maternal HGT network contains more nodes and edges, so it has a more complex internal structure.

We analyzed the evolvement of maternal HGT networks at three-time points: gestational week 27 (M\_Gest), birth (M\_Birth), and three months post-delivery

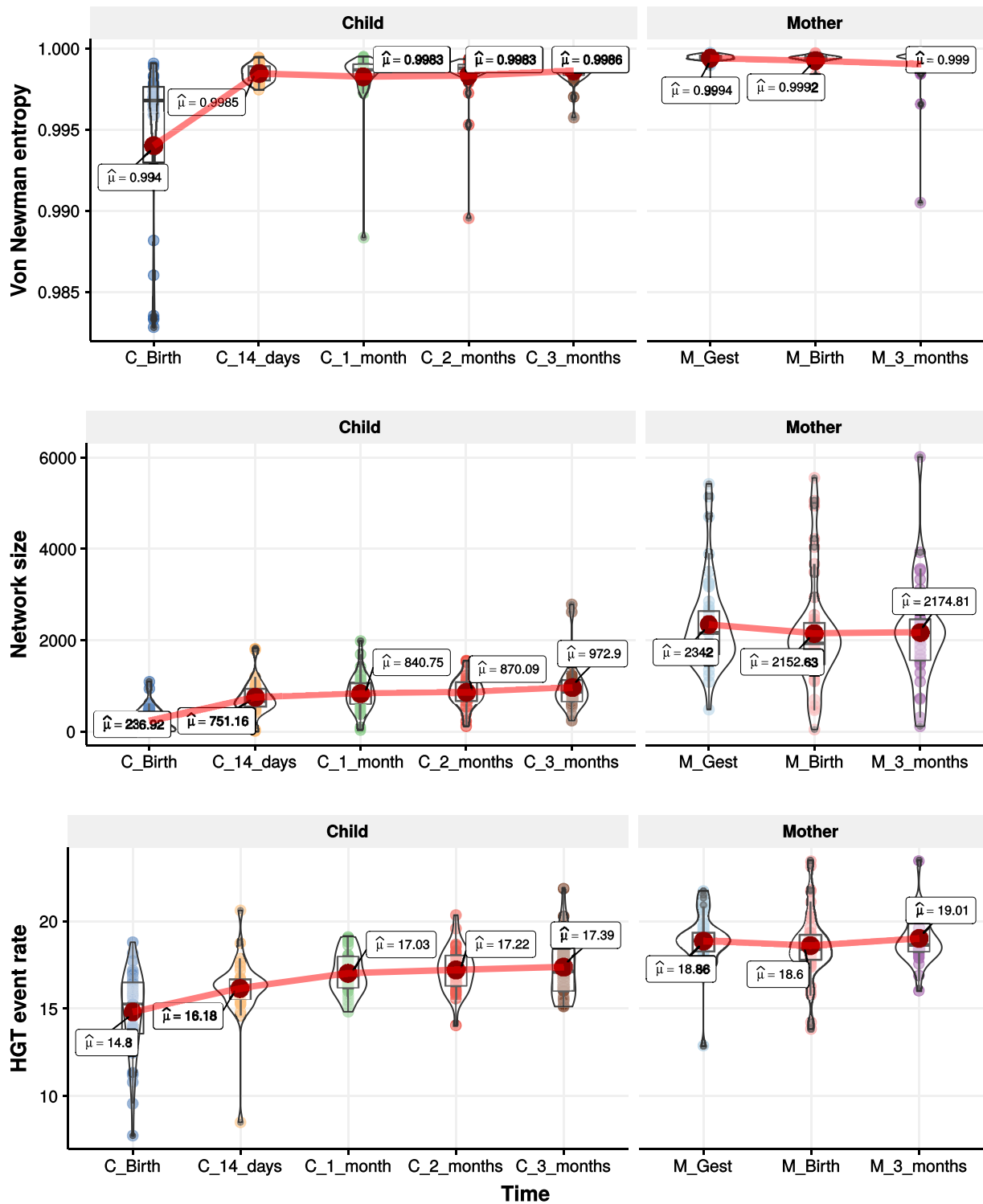


(M\_3\_months), and child HGT networks during the first three months. The evolution is measured from three aspects: Von Newman entropy, network size, and HGT event rate. The Von Newman Entropy measures the complexity of the network (see “Method” section) Network size is the number of nodes in the HGT network. HGT event rate is defined in “Method” section. It measures the number of HGT events detected in one sample. We calculate the three metrics for all HGT networks at each time point. As shown in Fig. 6, for child HGT networks, the three metric keep increasing during the first three months, which imply the increasing of the number of HGT events and the growing number of strains involved in HGT events. Compared to the evolution of child HGT networks, the three metrics of maternal HGT networks do not have increasing trends and maintain a relatively high level at the three-time points. This reflects the stability of mature gut microbiota. Furthermore, the increasing trends of the Von Newman Entropy, Network size, and HGT event rate of child HGT networks indicate the rapid growth of child gut microbiota during the first three months after birth. The average Von Newman

Entropy of child HGT networks rise from 0.994 to 0.9983. The average network size rises from 236.92 to 972.9, and the average HGT event rate rises from 14.8 to 17.39. This growth process could be described by the evolution of the HGT network. It usually takes 2 or 3 years to achieve the established microbiota among children [47]. To study the establishing of microbiota in child, it is worthwhile to analyze the evolution of the HGT network after three months, which could be in our future work (Fig. 6).

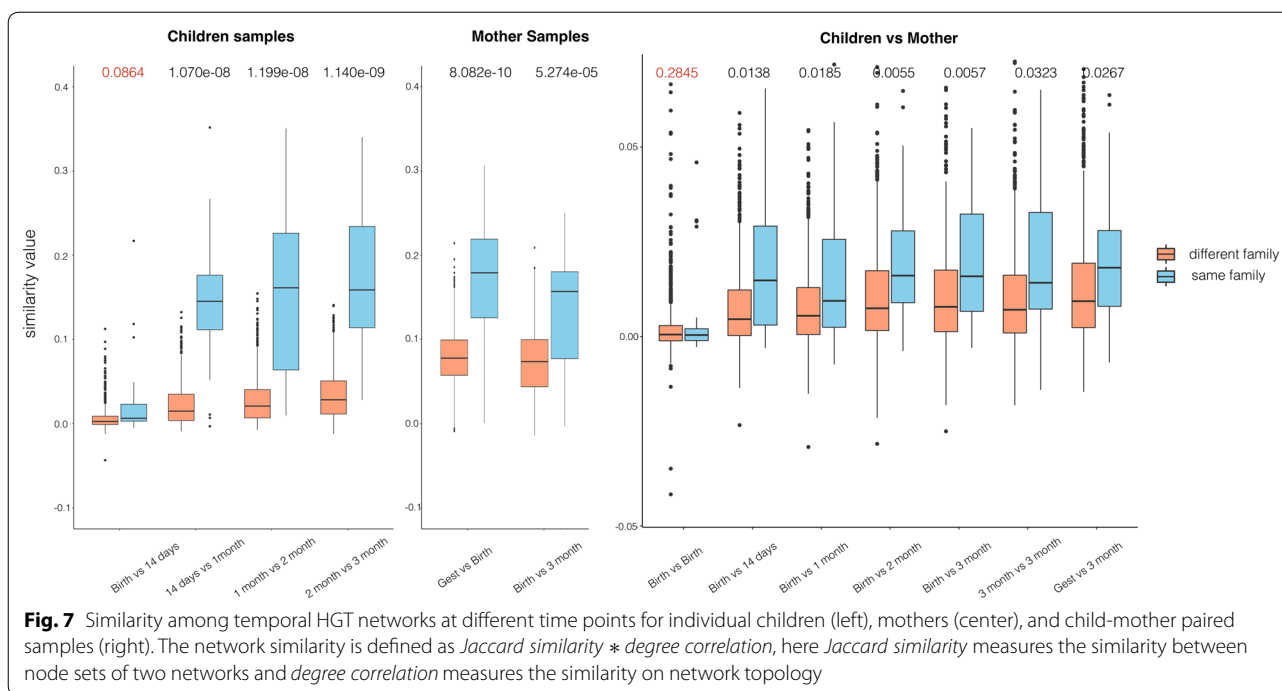
#### Family-specific similarity of maternal and child HGT networks

We measured the similarity of HGT networks to explore the evolution of dynamic gut microbiota (see “Method” section). First, child HGT networks show individual similarity. We compared HGT networks between adjacent time points chosen from the same individual and random two individuals. The network similarity is measured by *Jaccard similarity \* degree correlation*. As illustrated in Fig. 7, for child gut microbiota, HGT networks from the same child have significantly higher similarity over time than that from different children except for the first 2 weeks (Student’s t-test,



**Fig. 6** The complexity evolution of maternal and child HGT networks during the first three months are measured from three aspects: Von Newman Entropy, Network size, and HGT event rate





**Fig. 7** Similarity among temporal HGT networks at different time points for individual children (left), mothers (center), and child-mother paired samples (right). The network similarity is defined as *Jaccard similarity* \* *degree correlation*, here *Jaccard similarity* measures the similarity between node sets of two networks and *degree correlation* measures the similarity on network topology

$p = 0.0864, p = 1.07e^{-8}, p = 1.199e^{-8}, p = 1.14e^{-9}$ ).

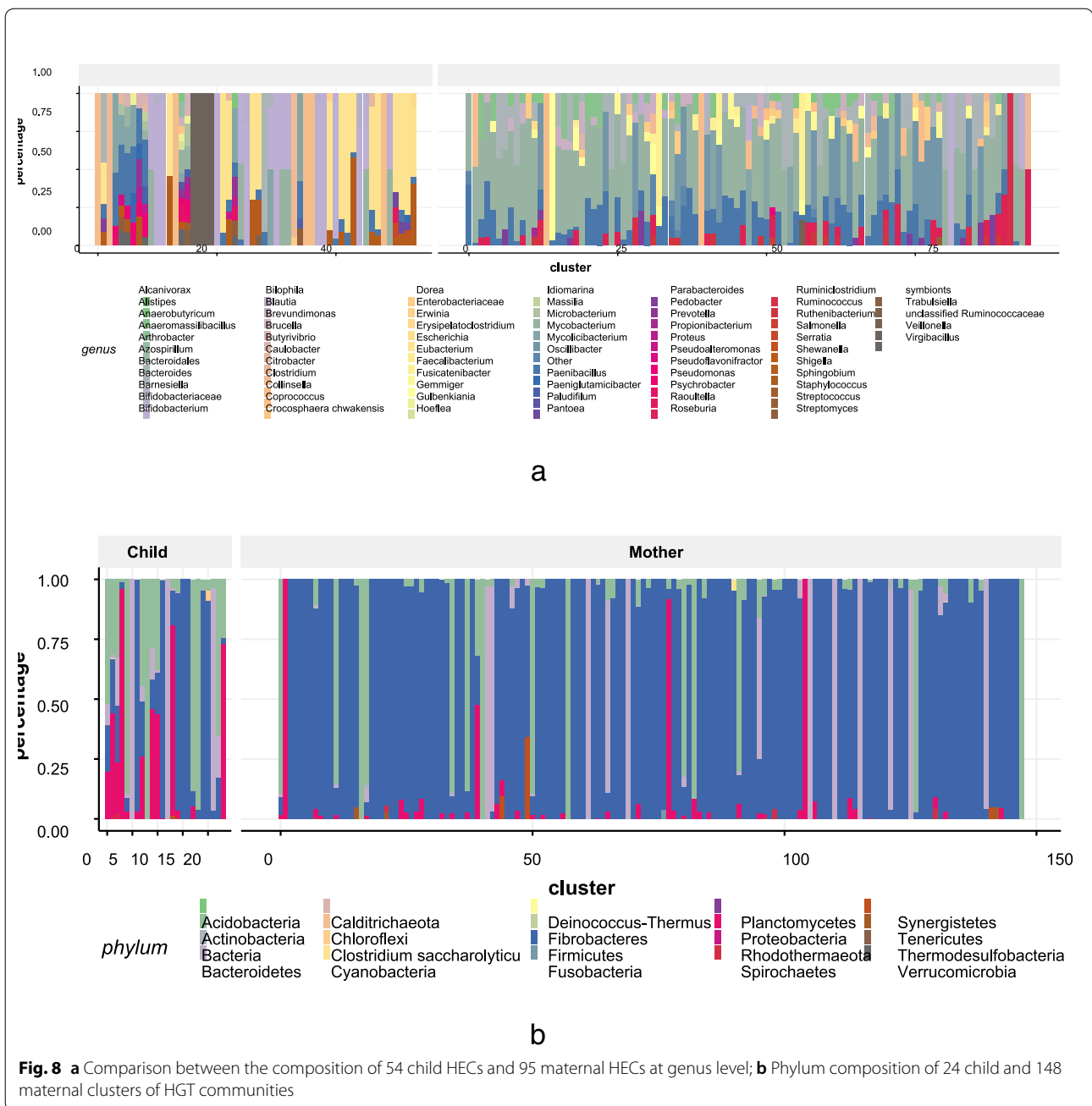
This demonstrates the child form individual-specific gut microbiota after one month since birth. Second, for maternal gut microbiota, the HGT network maintains significant similarity between adjacent time points in the same individual compared to random paired HGT networks ( $p = 8.082e^{-10}, p = 5.274e^{-5}$ ).

Third, to explore the transmission of the HGT network from mother to child, we compared maternal and child HGT networks within- and across-families. As illustrated in Fig. 7, we compared maternal HGT networks at birth to child HGT networks at birth, 14 days, 1 month, 2 months ( $p = 0.2845, p = 0.0138, p = 0.0185, p = 0.0055$ ). HGT networks of a mother and her child have significant similarities than those from random two families except for maternal and child HGT networks at birth ( $p = 0.2845$ ). Besides, we also compared the child HGT network at three months to maternal HGT networks at gestational week 27, birth, and three months post-delivery ( $p = 0.0267, p = 0.0057, p = 0.0323$ ). These results show that there is significant similarity between HGT networks of a mother and her child. Therefore, the mother does pass along microbes harboring family-specific HGT events to her child, which leads to the similarity of gut microbiota between a mother and her child. This is a family-specific gut microbial similarity captured by their HGT networks.

#### Conserved edges in Mother-to-Child HGT network

We first focus on edges in the HGT network and identify multiple conserved edges corresponding to host states. In HGT network, the edge is constructed according to HGT events (see “Method” section). The existence of conserved edges across multiple samples implies that HGT events occurred in multiple samples. We analyze conserved edges by clustering HGT events. By applying cluster analysis to HGT events (see “Method” section), we found HGT events that occurred in similar samples form HGT event clusters (HECs). For each HEC, we determine its label according to the host state of samples to which the majority of HGT events belong (see “Method” section). Further analysis on HECs helps explain how different host states affect the genus composition of conserved edges.

We get 54 child HECs and 95 maternal HECs. Child and maternal HECs have different composition at genus level, see Fig. 8a. The genus composition of HECs shows that the *Bifidobacterium* is the genus that significantly different in child and mother (21.59% in Child vs. 1.82% in Mother,  $p\text{-value} = 2.1604e^{-6}$ ), which is consistent with the finding in [48] that *Bifidobacterium* is the predominant bacteria in the child’s gut. Besides, we also found that *Escherichia* (20.24% in Child vs 0% in Mother,  $p\text{-value} = 1.6749e^{-8}$ ), *Microbacterium* (6.48% in Child vs 33.7% in Mother,  $p\text{-value} = 1.1742e^{-10}$ ),



*Mycolicibacterium* (0% in Child vs 11.78% in Mother,  $p\text{-value} = 1.5939e^{-4}$ ) are different in HECs of child and mother. *Escherichia* includes a number of pathogenic species such as *Escherichia coli* [49]. The high percentage of *Escherichia* contained in child-specific HECs is one common health risk to infants [50]. We have observed 13 other species of *Escherichia* in HECs of child including *Escherichia* sp. KTE172, *Escherichia* sp. 1\_1\_43, *Escherichia* sp. 4\_1\_40B, *Escherichia* sp. B1147, *Escherichia* sp. TW14182, *Escherichia*

*albertii*, *Escherichia* sp. TW15838, *Escherichia* sp. KTE159, *Escherichia fergusonii* ATCC 35469, *Escherichia* sp. KTE52, *Escherichia marmotae*, *Escherichia* sp. TW10509, and *Escherichia* sp. TW09231.

**HGT communities differ at different age state**

To find out preserved local network structure in multiple HGT networks, we detected communities in each HGT network. These communities are defined as HGT communities. Then we identified and determined the

label of HGT community clusters (HCCs) (see “Method” section).

We got 24 child and 148 maternal HCCs in the Mother-to-Child data set, and get the phylum composition of these HCCs as shown in Fig. 8b. The average phylum composition of child HCC is *Firmicutes*: 35.3%, *Actinobacteria*: 29.8%, *Proteobacteria*: 19.4%, *Bacteroidetes*: 15.1%, Others: 0.4%. The average phylum composition of maternal HCC is *Firmicutes*: 78.2%, *Actinobacteria*: 10.1%, *Bacteroidetes*: 7.9%, *Proteobacteria*: 3.2%, Others: 0.6%. Compared with child HCCs, the increasing of *Firmicutes* (p-value =  $8.0091e^{-7}$ ), the decreasing of *Proteobacteria* (p-value =  $2.8785e^{-5}$ ) and *Actinobacteria* (p-value = 0.0015) are significant. Genomes belonging to *Firmicutes* which plays an important role in maintaining the gut health [39] are the main maternal HCCs members. Communities in child HCCs are more diverse at the phylum level since child HGT network contains more nodes belonging to genus from *Proteobacteria* and *Actinobacteria*. As shown in the genus composition of child HCCs, these genus include *Bifidobacterium*, *Escherichia*, etc.

### Analysis of longitudinal IBD HGT networks

#### Disease-specific similarity of individual HGT networks

We also explored the HGT network similarity in different disease states using the longitudinal IBD data set. We utilize four similarity metrics, including Jaccard index, degree correlation, Pagerank correlation, and clustering coefficient, to measure network similarity. For each individual, we calculate the four metrics among his HGT networks at different time points (tables and heatmaps are

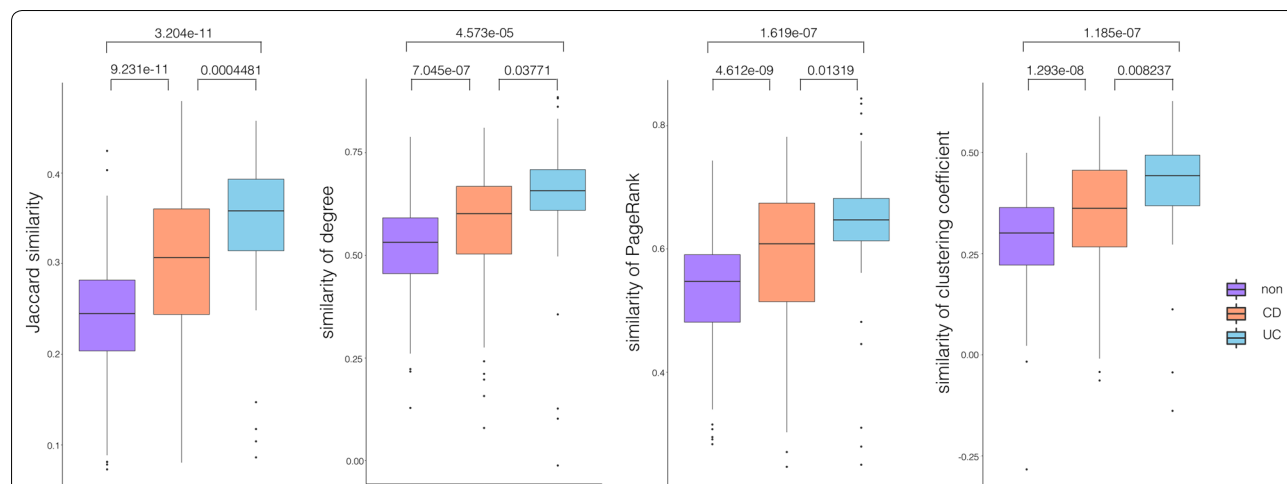
in Additional files 1 to 5). Finally we get disease-specific similarity metrics in CD, UC, and non-IBD controls as shown in Fig. 9. The four similarity metrics in Non-IBD samples are significantly lower than those in IBD samples. The lower similarity among Non-IBD HGT networks at different time points indicates that healthy gut microbiota is more flexible than IBD gut microbiota.

#### Conserved edges in longitudinal IBD HGT network

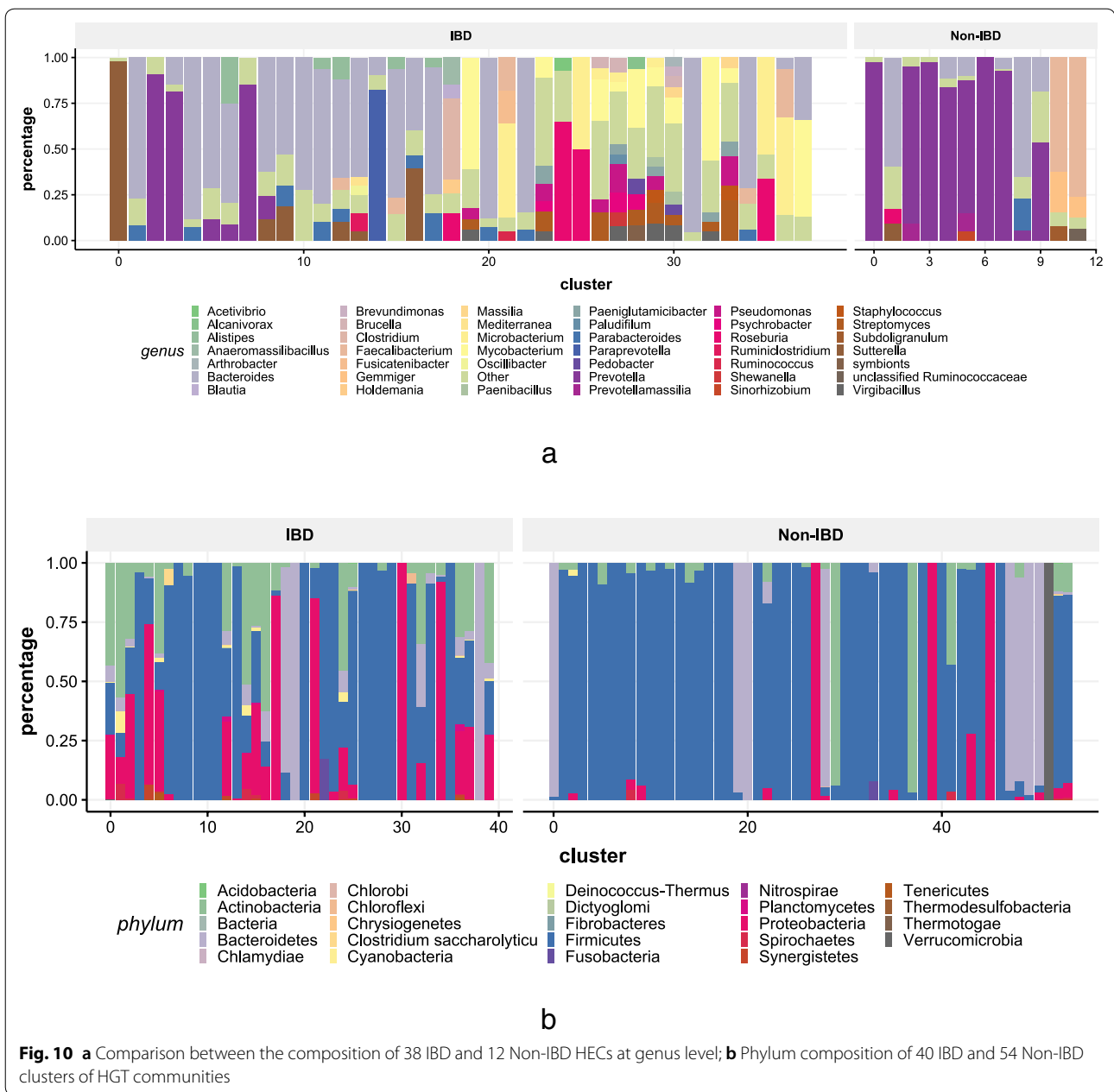
We get 38 IBD HECs and 12 Non-IBD HECs. Figure 10a compares the composition of these HECs at genus level. Compared with IBD HECs, more HGT events in Non-IBD HECs occur in *Prevotella* (57.4% in Non-IBD vs. 7.7% in IBD, p-value = 0.4527), which is a critical bacterium for healthy microbiota [51]. In contrast, more HGT events in detected IBD HECs are contained in pathogenic genus such as *Mycobacterium* (0% in Non-IBD vs 7.18% in IBD, p-value = 0.2937), *Sutterella* (0.78% in Non-IBD vs 4.82% in IBD, p-value = 0.3894) [52], *Pseudomonas* (0% in Non-IBD vs 1.63% in IBD, p-value = 0.0006). These genera could be treated as potential biomarkers. *Pseudomonas* plays an important role in IBD [53]. These findings demonstrate that groups harbor group-specific edges (Additional files 2, 3, 4, 5). Host states affect the conservation of HGT edges. Identifying these conserved edges helps to locate the key HGT events under specific conditions.

#### HGT communities differ between the IBD and non-IBD individuals

We found 40 IBD and 54 Non-IBD HCCs in the longitudinal IBD datasets. Figure 10b shows the distribution of phylum composition of them. The figure implies that



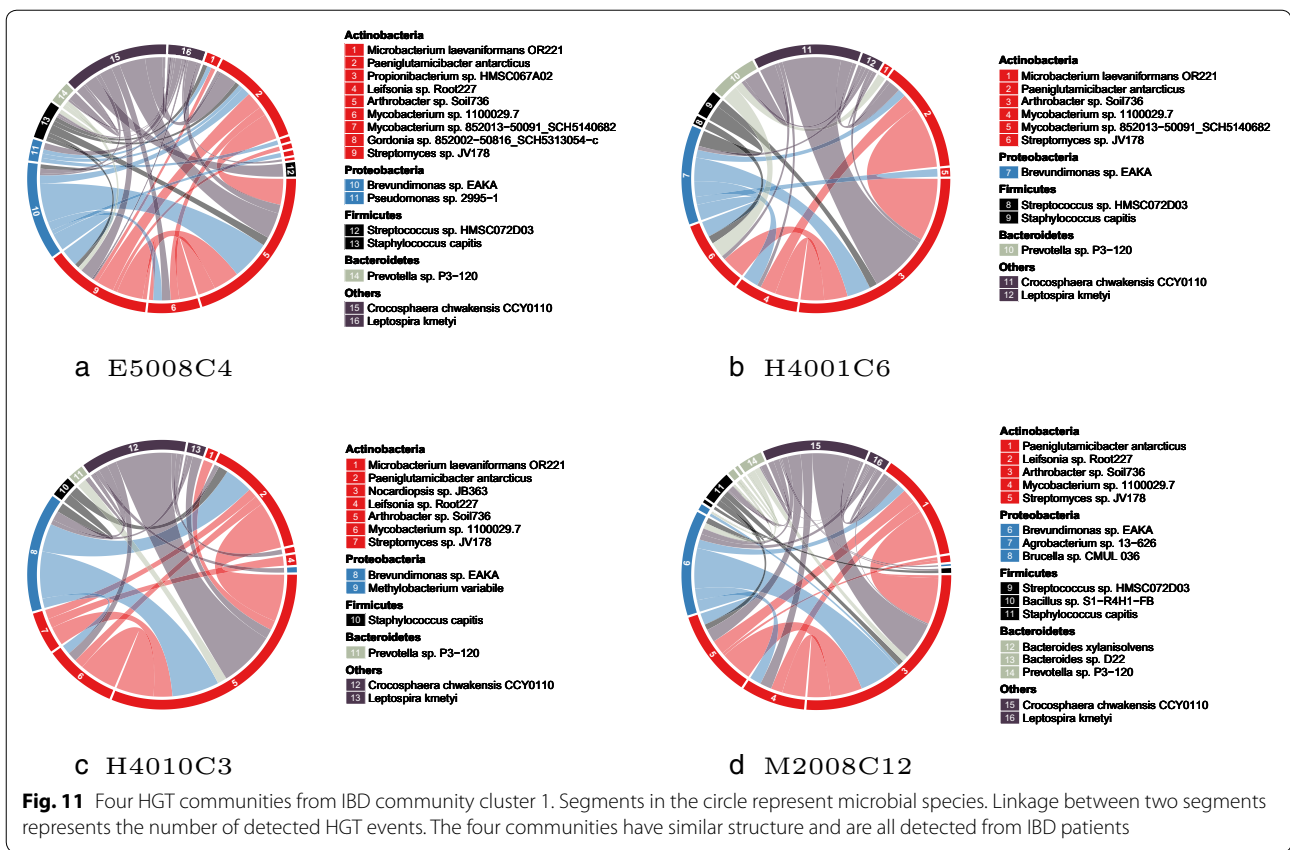
**Fig. 9** Compares the similarity among individual HGT networks at different time points in CD, UC, and non-IBD controls. The four similarity metrics are Jaccard index, degree correlation, pagerank correlation, and clustering coefficient



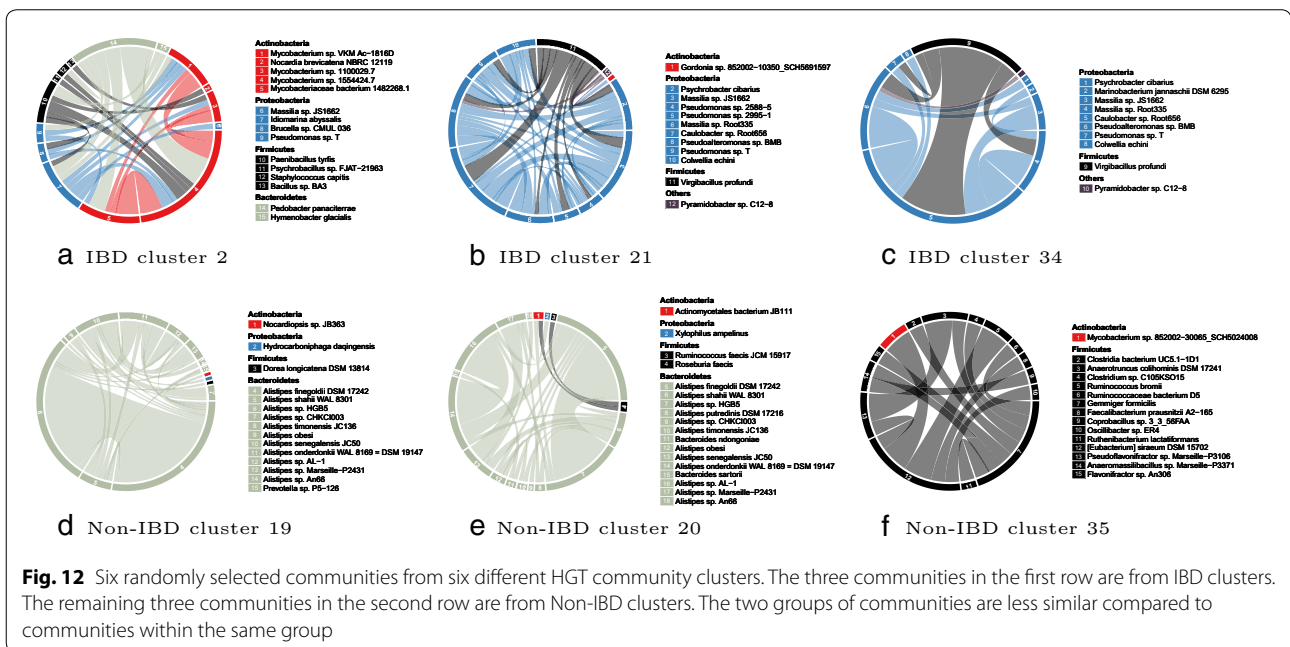
IBD HGT communities have different phylum composition with Non-IBD HGT communities. The average phylum composition of Non-IBD HCC is *Firmicutes*: 70.7%, *Bacteroidetes*: 14.4%, *Proteobacteria*: 6.8%, *Actinobacteria*: 5.9%, *Verrucomicrobia*: 1.9%, Others: 3%. The average phylum composition of IBD HCC is *Firmicutes*: 53.6%, *Proteobacteria*: 19.6%, *Actinobacteria*: 14.5%, *Bacteroidetes*: 9.9%, Others: 2.4%. Compared with Non-IBD HGT communities, IBD HCCs have less genome nodes from *Firmicutes* and *Bacteroidetes*

which are the two most dominant phyla in the large intestine of healthy adults [39]. While the phylum composition of IBD-specific HGT communities have significant increasing of *Proteobacteria* (p-value = 0.0194) and *Actinobacteria* (p-value = 0.0316). Many species belonging to *Proteobacteria* and *Actinobacteria* have strong association with the pathogenesis of IBD [40, 41].

Figure 11 illustrates the four similar communities from one IBD HCC. The four communities are detected



**Fig. 11** Four HGT communities from IBD community cluster 1. Segments in the circle represent microbial species. Linkage between two segments represents the number of detected HGT events. The four communities have similar structure and are all detected from IBD patients



**Fig. 12** Six randomly selected communities from six different HGT community clusters. The three communities in the first row are from IBD clusters. The remaining three communities in the second row are from Non-IBD clusters. The two groups of communities are less similar compared to communities within the same group

from four different IBD patients: E5008C4, H4001C6, H4010C3, and M2008C12. They have a high percentage of overlapping species and share similar linking patterns

between species. Since HGT communities in this cluster are mainly from the IBD group, their common structure can be used as a biomarker of IBD.

As shown in Fig. 12, HCCs detected from IBD and Non-IBD possess communities with different composition. Figure 12a–c are three randomly selected communities from three IBD HCCs. Figure 12d–f are three randomly selected communities from three non-IBD HCCs. Compared to HGT communities from clusters with the same label, HGT communities from clusters with different labels share less common members and therefore are much more dissimilar.

Moreover, the species in the community of IBD HCC are found to be associated with IBD. Figure 12a contains species (*Mycobacterium* sp. VKM Ac-1816D, *Mycobacterium* sp. 1100029.7, and *Mycobacterium* sp. 1554424.7) from genus *Mycobacterium*, which includes pathogens known to cause IBD [41, 54, 55]. Figure 12b, c contains species (*Pseudomonas* sp. 2588-5, *Pseudomonas* sp. 2995-1, and *Pseudomonas* sp. T) belong to genus *Pseudomonas*, which is also associated with IBD [40].

#### Gene fusions associated with HGT events in Longitudinal IBD data set

HGT event can cause the combination of parts of two genes belonging to two different genome sequences to fusion a gene (see “Method” section). From the Longitudinal IBD dataset, we found multiple HGT-caused gene fusion events that are related to the HGT mechanism, as listed following. We detected 2186 gene fusion events, which involve 1280 genes and 800 different gene functions. Gene fusion events containing genes associated with the mechanism of HGT include *recombinase family protein* (51/2186) [56], *plasmid mobilization relaxosome protein Mob* (39/2186) [57], *IS110 family transposase* (23/2186) [58], *site-specific integrase* (23/2186) [59], *conjugal transfer protein Tra* (20/2186) [60] and *transposase* (2/2186) [58]. Specifically, as shown in Table 1 in Appendix section, we detected 7 and 1 fusions of multidrug transporter genes in IBD and Non-IBD samples, respectively. The HGT event column denotes the two genome references for each HGT event. The number in the bracket denotes the HGT breakpoint position on the reference. Each HGT event consists of two HGT breakpoints. Fusion Gene A and B columns describe the information of the two fusion genes. The label column denotes the label of each sample. The multidrug transporter gene could encode multidrug transporters, which play an important role in multidrug resistance. By identifying and ejecting xenobiotic substances, multidrug transporters protect bacteria against antibacterial agents [61]. Products of the detected multidrug transporter genes involved in HGT events include *multidrug SMR transporter* [62], *multidrug transporter AcrB* [63], *multidrug transporter MatE* [64], and *multidrug efflux RND transporter permease subunit* [65].

## Method

### Reference genomes construction

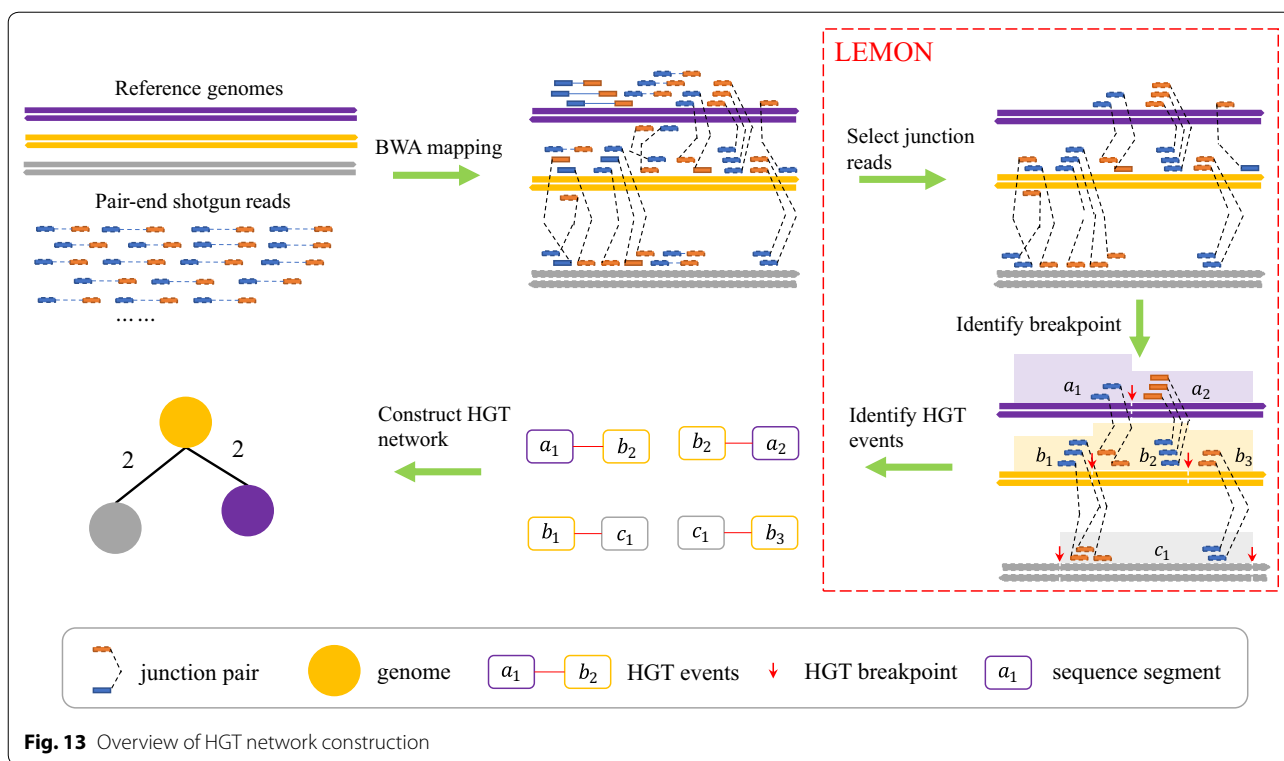
To construct HGT networks, we downloaded all bacterial genomes from the National Center for Biotechnology Information (NCBI). For each taxonomy, we choose the genome with the minimal scaffold number and highest completeness, whose contamination is less than 10% in Genome Taxonomy Database (GTDB) taxonomy evaluation results [66]. Finally, we obtain 109,419 bacterial genomes, which consist of 16,093 species with 1,246,881 scaffolds.

### HGT network construction

Figure 13 illustrates the construction procedure of the HGT network from raw NGS data. Firstly, we utilized Burrows–Wheeler Aligner (BWA) to map paired-end reads against reference genomes. Then we took the aligned reads as the input of LEMON [15] to detect HGT breakpoints on reference genomes. According to the HGT breakpoints, the genome sequences are split to segments, such as  $a_1$ ,  $a_2$ ,  $b_1$ , and so on. One HGT event is defined as the linkage of two segments belonging to two different genome sequences due to HGT, such as  $(a_1, b_2)$ ,  $(a_2, b_2)$ ,  $(b_2, c_2)$  in Fig. 13. In order to construct an HGT network, for any two reference genomes, if there exist HGT events between them, the two reference genomes are treated as nodes and linked by an HGT event. The weight of an edge is the number of HGT events between the two reference genomes. The three steps in the red box are key parts of LEMON, which are described in detail in [15].

### Evaluate the goodness of fit for power-law distribution

The scale-free network has its degree distribution  $P_k$  follows a power law  $P_k \sim k^{-\alpha}$ , where  $k$  is degree and  $\alpha$  is degree exponent whose value typically satisfies  $2 < \alpha < 3$ . As suggested in [43], in order to evaluate the goodness of fit for the HGT network to power-law distribution, we compared it with the other three heavy-tail distributions: exponential distribution, Weibull distribution, and lognormal distribution with positive parameter  $\mu$ . Let  $D = d_1, \dots, d_n$  denote the node degree set of HGT network,  $\forall d_v \in D$ , we have  $d_v \geq 1$  since an HGT event links two reference genomes, which means a genome node in HGT network should link at least another node. So the median of  $D$  has  $\text{median}(D) \geq 1$ . As for distribution  $\text{Lognormal}(\mu, \sigma^2)$ , its median is  $\exp(\mu)$ . Therefore, if we use  $\text{Lognormal}(\mu, \sigma^2)$  to fit  $D$ , we must have  $\exp(\mu) \geq 1$ , which means  $\mu \geq 0$ . We use the log-likelihood ratio test [42] to compare the goodness of two fits. In our experiments, for each HGT network, we performed three tests: power-law vs. exponential,



power-law vs. lognormal\_positive, and power-law vs. Weibull. In each test, if the ratio is larger than 0, the power-law achieves a better fit than the other.

**Von Newman entropy**

Structural complexity is an important characteristic of complex networks. It greatly determines the function and status of complex networks. The growth of temporal networks often leads to structural change, which also means a change of complexity. Ye et al. [67] proposed an approximation of von Neumann entropy to measure the complexity of dynamic networks. Since the simplified von Neumann entropy can be interpreted as the thermodynamic entropy of the network, we can describe the complex dynamic system from the perspective of statistical thermodynamics. Therefore, we applied the simplified von Neumann entropy to measure the complexity of temporal HGT networks. Let  $H(V, E)$  denote the HGT network,  $V$  denotes the vertex set, which represents genomes linked by HGT events,  $E$  is an edge set, which indicates whether there exist HGT events between two genome nodes. Then the von Neumann entropy  $H_{VN}$  is defined as follows,

$$H_{VN} = 1 - \frac{1}{|V|} - \frac{1}{|V|^2} \sum_{(u,v) \in E} \frac{1}{d_u d_v} \tag{1}$$

Here  $d_u$  is the degree of node  $u$ .  $|V|$  is the number of nodes,  $(u,v)$  is an edge in  $E$ .

**HGT event rate**

We take next-generation sequencing (NGS) short reads as input of LEMON to detect HGT events on reference genomes. The amount of raw data and the length of reference genomes can affect the number of detected HGT events in one sample. We normalized the number of HGT events and define the normalized number as *HGT event rate*. Let  $H$  denotes the number of detected HGT events in sample  $S$ ,  $G = \{g_1, g_2, \dots, g_N\}$  is the set of genomes linked by HGT events. Denote  $r_i$  as the number of reads uniquely mapped onto genome  $g_i$ ,  $g_i \in G$ , and denote  $l_i$  is the length of  $g_i$ . So we get two sets  $R = \{r_1, r_2, \dots, r_N\}$  and  $L = \{l_1, l_2, \dots, l_N\}$ . The HGT event rate  $\bar{H}$  is calculated as follows

$$\bar{H} = \ln \frac{H}{\sum_{i=1}^N r_i / \sum_{i=1}^N l_i} \tag{2}$$

here  $\sum_{i=1}^N r_i / \sum_{i=1}^N l_i$  is approximate to the average read depth.  $H$  represents the logarithm of the rate between the number of HGT events and the read depth.

**Similarity metric for HGT networks**

After obtaining the HGT networks, we calculated the similarity between networks according to the following measurements. First, we measured the similarity between networks by Jaccard similarity of species that present in the

network. If two HGT networks share more species, then they obtain higher Jaccard similarity. Besides species present and absent, we also measured network similarity based on the topology property for species in networks, e.g., degree, PageRank, clustering coefficient. For two networks, we find species shared by both networks, record their degree in two networks, and calculate the Spearman correlation on the two-degree lists. The generated correlation implies the consistency of the importance rank for every species in two networks at the aspect of different topology properties. Finally, we assigned these comparison pair into multiple groups and studied the similarity distribution. To test whether the similarity between networks in the same family is significantly higher than it between different families, we applied a T-test to the similarity values between these two groups in different sample times.

### Detecting and clustering HGT communities

We detected community in each HGT network in each sample and to study the community evolution by comparing those detected HGT communities across time/individual. First, for community detection, we applied the Leiden algorithm [68], which is based on the Louvain algorithm. Louvain algorithm is a popular community detection methods which optimize the modularity in the network by local move and aggregate network iteratively. However, it can generate badly connect communities sometimes. Leiden overcomes it by adding a smart local move to refine the partition of nodes in each iteration. Therefore Leiden generates more robust and well-connected communities, which is a better solution in our situation. After community detection, we further clustered HGT communities detected in all different individuals and different times. By applying clustering analysis on communities, we can find common communities across samples and calculate their composition within a group. We call the common community group as HGT community clusters (HCCs). For clustering, the distance between communities is calculated from their Jaccard similarity; then, we apply a hierarchical cluster on all identified communities to find HCCs.

### HGT event cluster

For each HGT event, we recorded the number of samples it present. HGT events that show in less than the minimum sample threshold are filtered out. The minimum samples threshold is set as 5 and 4 in Mother-to-Child and Longitudinal IBD experiments, respectively. Then we measured the relationship between HGT events by their Jaccard similarity considering the samples they show up. After convert similarity to distance, we applied hierarchical clusters and determined classes by dynamic tree cut. We set distance to 0.6 and minimum class sizes as 10

and 20 for Mother-to-Child and Longitudinal IBD data sets, respectively. The outcome of the hierarchical result is HGT event clusters (HECs). Each HEC refers to the group of HGT events which occur together.

### Label clusters of HGT communities and events

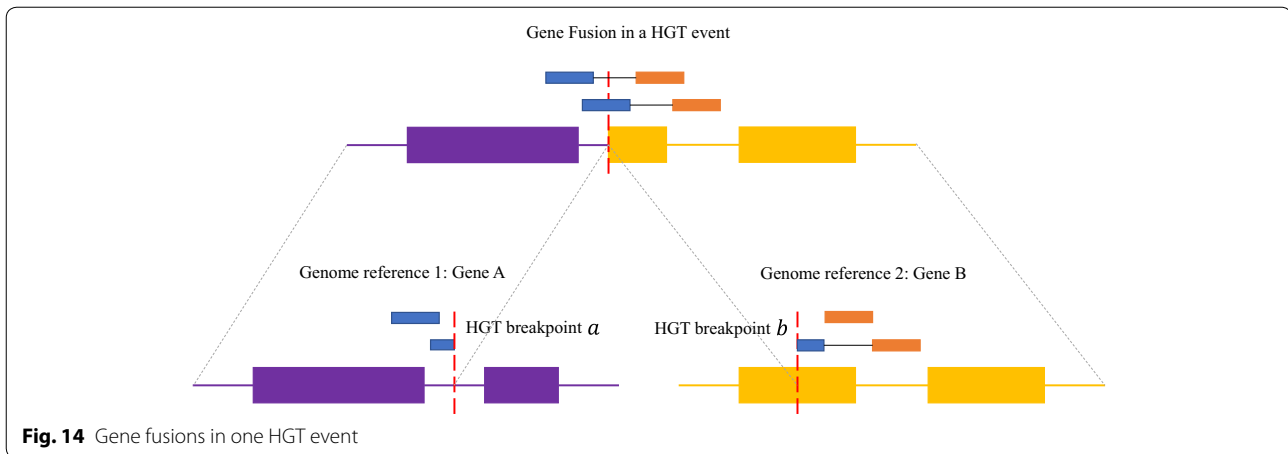
Since each cluster of HGT communities/events (HCC/HEC) consists of HGT communities/events from different groups, it may contain multiple labels with an unequal number of communities/events. The label of the cluster should be determined by the predominant label in the cluster. Therefore we determined the label of cluster as following. Without loss of generality, we take the determination of labels for HCCs in longitudinal IBD data set as an example. From longitudinal IBD data set, we got 94 HCCs  $\{cluster_i = [c_{i1}, \dots, c_{in}], i = 1, \dots, 94\}$ , here  $cluster_i$  represents the  $i$ th HCC,  $c_i$  denotes the communities in  $cluster_i$ . According to the group of the sample from which the community is detected, each community is labeled as Non-IBD or IBD. For community  $c_i$ , we denote  $l(c_i)$  as its label. We counted the number of communities belonging to different labels and let  $COUNT_i(Non-IBD)$  and  $COUNT_i(IBD)$  represent the number of communities having label  $Non-IBD$  and  $IBD$  respectively in  $cluster_i$ . If  $COUNT_i(Non-IBD) > COUNT_i(IBD)$ , then  $l(cluster_i) = Non-IBD$ . Otherwise  $l(cluster_i) = IBD$ . However, the number of Non-IBD samples  $num(Non-IBD)$  is 38 and the number of IBD samples  $num(IBD)$  is 109, the larger value of  $num(IBD)$  make  $cluster_i$  tend to contain more communities with IBD labels. To correct the bias, we set  $l(cluster_i)$  as follows,

$$l(cluster_i) = \begin{cases} \text{Non-IBD} & \frac{COUNT_i(Non-IBD)}{num(Non-IBD)} > \frac{COUNT_i(IBD)}{num(IBD)} \\ \text{IBD} & \frac{COUNT_i(Non-IBD)}{num(Non-IBD)} < \frac{COUNT_i(IBD)}{num(IBD)} \end{cases} \quad (3)$$

Here  $\frac{COUNT_i(Non-IBD)}{num(Non-IBD)}$  and  $\frac{COUNT_i(IBD)}{num(IBD)}$  represent the relative amount of  $Non-IBD$  and  $IBD$  communities that are contained in  $cluster_i$  respectively. If  $\frac{COUNT_i(Non-IBD)}{num(Non-IBD)} < \frac{COUNT_i(IBD)}{num(IBD)}$ , it means that HGT communities sharing common structure in  $cluster_i$  are mainly from IBD samples and their common structure could act as a candidate biomarker of IBD.

Let  $[g_j^1, \dots, g_j^m]$  denotes the genomes set of community  $c_j$ . For each cluster of communities  $cluster_i = [c_{i1}, \dots, c_{in}]$ , we collect its communities' genomes together and get the genome set  $G_i = [g_{i1}^1, \dots, g_{i1}^{m_{i1}}, \dots, g_{ij}^{m_{ij}}, \dots, g_{in}^{m_{in}}]$  for  $cluster_i$ , here  $m_{ij}$  is the number of genomes in community  $c_{ij}$ . The phylum/genus to which each genome in  $G_i$  belongs could be found on NCBI. Then, for  $cluster_i$ , we get the phylum set  $P_i = [phylum(g_{i1}^1), \dots, phylum(g_{i1}^{m_{i1}}), \dots, phylum(g_{ij}^{m_{ij}}), \dots, phylum(g_{in}^{m_{in}})]$ . Finally, we figured out the composition of  $cluster_i$  under





different conditions at the phylum level by calculating the percentage of each phylum in  $P_i$ . For each cluster of HGT events, we collect its genus set and figure out the percentage of each genus.

#### Gene fusions in HGT events

We further extended the HGT network study to gene function.

Figure 14 denotes one fusion gene in an HGT event. The fusion gene contains two gene parts belonging to Gene A on genome reference 1 and Gene B on genome reference 2 respectively and  $a$  and  $b$  are the two HGT breakpoints. Since the two HGT breakpoints are in two gene regions, the combination of parts of two genes caused by HGT forms a fusion gene. We detected HGT breakpoints by LEMON and identified the gene fusion. So we can link the HGT event to gene function. The gene annotations (including Gene ID, start and end positions, function description) are collected from NCBI.

#### Discussion and conclusion

HGT is the process of sharing genetic material among different microbial species. It links different species by transferring genetic information. By analyzing HGT networks constructed from two longitudinal metagenomic sequencing data sets: Mother-to-Child and longitudinal IBD data sets, we found the HGT network is scale-free, whose degree distribution follows a power law. Most nodes have a small degree and are connected by hub nodes. Moreover, statistics support the ultra-small world property of the HGT network. The distance between two randomly nodes in the HGT network could maintain small despite the growth of the network. Such a stable internal structure demonstrates the robustness of the HGT network in the ever-changing environment. The HGT network also provides us an efficient way to model human gut microbiota. The

development of the child gut microbiota during the first three months after birth could be captured by the evolution of temporal HGT networks. The increasing of HGT network complexity and size is led by the growth of strains harboring HGT events. Furthermore, we have found a significant similarity between the family-specific child and maternal HGT networks. Therefore, HGT network could characterize bacterial transmission patterns from mother to child. It also demonstrates that maternal gut bacteria may be an essential source of child gut bacteria.

Analysis of the HGT community and HGT event demonstrates that age and disease change the internal structure of HGT networks. Compared with healthy individuals, the inflammatory conditions of the gastrointestinal tract in IBD patients are suitable for the growth of bacteria from *Proteobacteria* and *Actinobacteria* phylum. So HGT communities contain more bacteria from *Proteobacteria* and *Actinobacteria* and fewer bacteria from *Firmicutes*. Similar change is also observed in newborn children. By clustering HGT communities, we could find similar communities across multiple HGT networks that are from the same group. These similar communities reflect the influence of a specific host state on the structure of the HGT network and could be treated as potential biomarkers. The formation of the HGT community cluster is led by similar HGT events across samples. Through clustering HGT events, we realized that in IBD patients, the composition change of HGT communities is led by the increasing of HGT events mainly contained in the pathogenic genus *Mycobacterium*, *Sutterella*, and *Pseudomonas*. While in newborn children, we observed the increase of HGT events contained in *Bifidobacterium* and *Escherichia*. These differences reflect the alteration of gut microbiota in different conditions. The summarization of fusion genes in HGT events helps us better realize genes associated with HGTs. Many fusion genes

in HGT events encode proteins, including recombinase family protein, plasmid mobilization relaxosome protein Mob, conjugal transfer protein Tra, and so on, that facilitate the horizontal transfer of genetic material. Furthermore, in IBD patients, more fusion genes caused by HGT events encode multidrug transporter proteins. This reveals that beneficial HGT events contribute to the survival of associated microbial strains in the gut under the specific selection pressure.

In summary, through the HGT network, we research human gut microbiota from a systematic perspective. The network analysis of Mother-to-Child and longitudinal IBD data sets demonstrate the characteristic of HGT networks differs under different conditions. In the future, we will apply our pipeline to analyze more HGT networks. It helps us get a deeper understanding of the relationship between host states and microbial interactions.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s13099-020-00370-9>.

## Acknowledgements

Not applicable.

## Authors' contributions

SCL conceived the project. CL and JXC performed the analyses and evaluated the results, CL and JXC wrote the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

## Funding

The work is supported by City University of Hong Kong (Project 7004886). The funding body did not play any role in the design of the study and collection, analysis, interpretation of data, and manuscript writing.

## Availability of data and materials

283 Metagenomic samples were deposited to Sequence Read Archive (BioProject: PRJNA475246). 148 Metagenomic samples were deposited to Sequence Read Archive (BioProject: PRJNA389280). HGT detection package LEMON is available at <https://github.com/lichen2018/LEMON>. All analysis codes are freely available at <https://github.com/lichen2018/HGT-network>.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Appendix

### Multidrug transporter gene fusions in HGT events

**Additional file 1.** Spearman correlation between degrees of nodes in HGT networks in IBD dataset.

**Additional file 2.** Jaccard similarity of nodes in HGT networks in IBD dataset.

**Additional file 3.** Spearman correlation between PageRank of all nodes in HGT networks in IBD dataset.

**Additional file 4.** Spearman correlation between the clustering coefficient of all nodes in HGT networks in IBD dataset.

**Additional file 5.** The heatmaps of similarity matrix for IBD HGT networks measured using Jaccard similarity (Fig. 1), Spearman correlation between degrees (Fig. 2), Spearman correlation between PageRank (Fig. 3), and Spearman correlation between the clustering coefficient (Fig. 4). The heatmap of similarity matrix for Infant HGT networks measured using Jaccard similarity\*degree correlation (Fig. 5).

**Table 1** Multidrug transporter gene fusions in HGT events

| HGT event                                      | Gene A             |  | Gene B             |  | Label |
|--|--------------------|--|--------------------|--|-------|
|  | Gene symbol        | Description  | Gene symbol        | Description  |       |
| NZ_DS362246.1(124896)<br>NZ_DS499674.1(613633) | BACUNI RS1296      | Multidrug transporter                                | BACSTE RS1143      | Hypothetical protein   | CD    |
| NZ_GG697149.2(303360)<br>NZ_GG697156.2(32286)  | FAEPRAA2165 RS0137 | Multidrug SMR transporter                            | FAEPRAA2165 RS1330 | Stage II sporulation protein   | CD    |
| NZ_DS499665.1(25120) NZ_<br>DS499672.1(423580) | BACSTE RS0155      | Hypothetical protein                                 | BACSTE RS0626      | MexE family multidrug efflux<br>RND transporter periplasmic<br>adaptor subunit | Non   |
| NZ_DS499676.1(402917)<br>NZ_DS499675.1(12487)  | BACSTE RS1390      | Hybrid sensor histidine<br>kinase/response regulator | BACSTE RS1181      | Multidrug transporter AcrB   | CD    |
| NZ_DS499672.1(447434)<br>NZ_DS499662.1(199650) | BACSTE RS0636      | ATP-dependent Clp protease<br>Clp                    | BACSTE RS0090      | Multidrug transporter MatE   | UC    |
| NZ_DS499671.1(180331)<br>NZ_DS499673.1(122374) | BACSTE RS0406      | Glycoside hydrolase family                           | BACSTE RS0707      | Multidrug transporter AcrB   | UC    |
| NZ_DS499671.1(61003) NZ_<br>DS499673.1(330329) | BACSTE RS0350      | Hypothetical protein                                 | BACSTE RS0784      | Multidrug efflux RND trans-<br>porter permease subunit                         | UC    |
| NZ_DS499673.1(123680)<br>NZ_DS499677.1(203460) | BACSTE RS0707      | Multidrug transporter AcrB                           | BACSTE RS1579      | UDP-3-O-[3-hydroxymyristoyl]<br>N-acetylglucosamine<br>deacetylase             | UC    |

Received: 14 March 2020 Accepted: 23 June 2020  
Published online: 09 July 2020

## References

- Putignani L, Chierico FD, Petrucca A, Vernocchi P, Dallapiccola B. The human gut microbiota: a dynamic interplay with the host from birth to senescence settled during childhood. *Pediatric Res*. 2014;76(1):2–10. <https://doi.org/10.1038/pr.2014.49>.
- Lozupone CA, Stombaugh JJ, Gordon JL, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature*. 2012;489(7415):220–30. <https://doi.org/10.1038/nature11550>.
- Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, Ilonen J, Ferretti P, Pasolli E, Tett A, Asnicar F, Segata N, Vlamakis H, Lander ES, Huttenhower C, Knip M, Xavier RJ. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe*. 2018;24(1):146–1544. <https://doi.org/10.1016/j.chom.2018.06.007>.
- Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JL. Human gut microbiome viewed across age and geography. *Nature*. 2012;486(7402):222–7. <https://doi.org/10.1038/nature11053>.
- Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, Kim AD, Shmigel AK, Syed AN, Walter J, Menon R, Koecher K, Knights D. Daily sampling reveals personalized diet-microbiome associations in humans. *Cell Host Microbe*. 2019;25(6):789–8025. <https://doi.org/10.1016/j.chom.2019.05.005>.
- Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genetics*. 2008;9(8):605–18. <https://doi.org/10.1038/nrg2386>.
- Hasan MS, Liu Q, Wang H, Fazekas J, Chen B, Che D. GIST: Genomic island suite of tools for predicting genomic islands. *Bioinformatics*. 2012;28(4):203–5. <https://doi.org/10.1093/bioinformatics/btq082>.
- Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*. 1997;44(4):383–97. <https://doi.org/10.1007/pl00006158>.
- Podell S, Gaasterland T. *Genome Biology*. 2007;8(2):16. <https://doi.org/10.1186/gb-2007-8-2-r16>.
- Zhu Q, Kosoy M, Dittmar K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics*. 2014;15(1):717. <https://doi.org/10.1186/1471-2164-15-717>.
- Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome*. 2019;7(1). <https://doi.org/10.1186/s40168-019-0649-y>.
- Douglas GM, Langille MGI. Current and promising approaches to identify horizontal gene transfer events in metagenomes. *Genome Biol Evol*. 2019;11(10):2750–66. <https://doi.org/10.1093/gbe/evz184>.
- McInnes RS, McCallum GE, Lamberte LE, van Schaik W. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Curr Opin Microbiol*. 2020;53:35–43. <https://doi.org/10.1016/j.mib.2020.02.002>.
- Trappe K, Marschall T, Renard BY. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics*. 2016;32(17):595–604. <https://doi.org/10.1093/bioinformatics/btw423>.
- Li C, Jiang Y, Li S. LEMON: a method to construct the local strains at horizontal gene transfer sites in gut metagenomics. *BMC Bioinform*. 2019;20:S23. <https://doi.org/10.1186/s12859-019-3301-8>.
- Karakoc E, Alkan C, ORoak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE. Detection of structural variants and indels within exome data. *Nat Methods*. 2011;9(2):176–8. <https://doi.org/10.1038/nmeth.1810>.
- Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E. (ed) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96, pp. 226–231. AAAI Press, Portland, Oregon (1996). <http://dl.acm.org/citation.cfm?id=3001460.3001507>
- Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998;393(6684):440–2. <https://doi.org/10.1038/30918>.
- Albert R, Barabási A-L. Statistical mechanics of complex networks. *Rev Modern Phys*. 2002;74(1):47–97. <https://doi.org/10.1103/revmodphys.74.47>.
- Hassan MK, Islam L, Haque SA. Degree distribution, rank-size distribution, and leadership persistence in mediation-driven attachment networks. *Physica A*. 2017;469:23–30. <https://doi.org/10.1016/j.physa.2016.11.001>.
- Bianconi G, Barabási A-L. Competition and multiscaling in evolving networks. *Europhys Lett (EPL)*. 2001;54(4):436–42. <https://doi.org/10.1209/epl/i2001-00260-6>.
- Caldarelli G, Capocci A, Rios PDL, Muñoz MA. Scale-free networks from varying vertex intrinsic fitness. *Phys Rev Lett*. 2002;89:25. <https://doi.org/10.1103/physrevlett.89.258702>.
- Bauer E, Thiele I. From network analysis to functional metabolic modeling of the human gut microbiota. *mSystems*. 2018; <https://doi.org/10.1128/mSystems.00209-17>.
- Sung J, Kim S, Cabatbat JJT, Jang S, Jin Y-S, Jung GY, Chia N, Kim P-J. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat Commun*. 2017; <https://doi.org/10.1038/ncomms15393>.
- Kunin V. The net of life: reconstructing the microbial phylogenetic network. *Genome Res*. 2005;15(7):954–9. <https://doi.org/10.1101/gr.3666505>.
- Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. *PLOS Comput Biol*. 2015;11(5):1004095. <https://doi.org/10.1371/journal.pcbi.1004095>.
- Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-Y M, Glickman JN, Garrett WS. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science*. 2013;341(6145):569–73. <https://doi.org/10.1126/science.1241165>.
- Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S. Host-gut microbiota metabolic interactions. *Science*. 2012;336(6086):1262–7. <https://doi.org/10.1126/science.1223813>.
- Wampach L, Heintz-Buschart A, Fritz JV, Ramiro-García J, Habier J, Herold M, Narayanasamy S, Kaysen A, Hogan AH, Bindl L, Bottu J, Halder R, Sjöqvist C, May P, Andersson AF, de Beaufort C, Wilmes P. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat Commun*. 2018; <https://doi.org/10.1038/s41467-018-07631-x>.
- Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT, Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Wang J. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*. 2015;17(5):690–703. <https://doi.org/10.1016/j.chom.2015.04.004>.
- Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, Beghini F, Bertorelli R, Sanctis VD, Bariletti I, Canto R, Clementi R, Cologna M, Crifò T, Cusumano G, Gottardi S, Innamorati C, Masè C, Postai D, Savoì D, Duranti S, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Milani C, Mangifesta M, Anzalone R, Viappiani A, Yassour M, Vlamakis H, Xavier R, Collado CM, Koren O, Tateo S, Soffiati M, Pedrotti A, Ventura M, Huttenhower C, Bork P, Segata N. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe*. 2018;24(1):133–1455. <https://doi.org/10.1016/j.chom.2018.06.005>.
- Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, Ananthakrishnan AN, Andrews E, Barron G, Lake K, Prasad M, Sauk J, Stevens B, Wilson RG, Braun J, Denson LA, Kugathasan S, McGovern DPB, Vlamakis H, Xavier RJ, Huttenhower C. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol*. 2018;3(3):337–46. <https://doi.org/10.1038/s41564-017-0089-z>.
- Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, Kumar N, Stares MD, Rodger A, Brocklehurst P, Field N, Lawley TD. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature*. 2019;574(7776):117–21. <https://doi.org/10.1038/s41586-019-1560-1>.
- Parekh PJ, Balart LA, Johnson DA. The influence of the gut microbiome on obesity, metabolic syndrome and gastrointestinal disease. *Clin Transl Gastroenterol*. 2015;6(6):91. <https://doi.org/10.1038/ctg.2015.16>.
- Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, Ross MC, Lloyd RE, Doddapaneni H, Metcalf GA, Muzny D, Gibbs RA, Vatanen T, Huttenhower C, Xavier RJ, Rewers M, Hagopian W, Toppari J, Ziegler A-G,

- She J-X, Akolkar B, Lernmark A, Hyoty H, Vehik K, Krischer JP, Petrosino JF. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*. 2018;562(7728):583–8. <https://doi.org/10.1038/s41586-018-0617-x>.
36. Schaubek M, Clavel T, Calasan J, Lagkouvardos I, Haange SB, Jehmlich N, Basic M, Dupont A, Hornef M, von Bergen M, Bleich A, Haller D. Dysbiotic gut microbiota causes transmissible crohns disease-like ileitis independent of failure in antimicrobial defence. *Gut*. 2015;65(2):225–37. <https://doi.org/10.1136/gutjnl-2015-309333>.
  37. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology*. 2014;146(6):1489–99. <https://doi.org/10.1053/j.gastro.2014.02.009>.
  38. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, McClure EE, Dunklebarger MF, Knight R, Jansson JK. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*; 2017. <https://doi.org/10.1038/nmicrobiol.2017.4>.
  39. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, Thomas LV, Zoetendal EG, Hart A. The gut microbiota and host health: a new clinical frontier. *Gut*. 2015;65(2):330–9. <https://doi.org/10.1136/gutjnl-2015-309990>.
  40. Mukhopadhyay I, Hansen R, El-Omar EM, Hold GL. Ibd what role do proteobacteria play? *Nat Rev Gastroenterol Hepatol*. 2012;9(4):219–30. <https://doi.org/10.1038/nrgastro.2012.14>.
  41. Graham DY, Markesich DC, Yoshimura HH. Mycobacteria and inflammatory bowel disease. *Gastroenterology*. 1987;92(2):436–42. [https://doi.org/10.1016/0016-5085\(87\)90139-9](https://doi.org/10.1016/0016-5085(87)90139-9).
  42. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. *SIAM Rev*. 2009;51(4):661–703. <https://doi.org/10.1137/070710111>.
  43. Alstott J, Bullmore E, Plenz D. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*. 2014;9(1):85777. <https://doi.org/10.1371/journal.pone.0085777>.
  44. Barabási A-L. *Network Science*. London: Cambridge University Press; 2016. <https://www.xarg.org/ref/a/1107076269/>
  45. Cohen R, Erez K, ben-Avraham D, Havlin S. Resilience of the internet to random breakdowns. *Phys Rev Lett*. 2000;85(21):4626–8. <https://doi.org/10.1103/physrevlett.85.4626>.
  46. Cohen R, Havlin S. Scale-free networks are ultrasmall. *Phys Rev Lett*. 2003; <https://doi.org/10.1103/physrevlett.90.058701>.
  47. Robertson RC, Manges AR, Finlay BB, Prendergast AJ. The human microbiome and child growth first 1000 days and beyond. *Trends Microbiol*. 2019;27(2):131–47. <https://doi.org/10.1016/j.tim.2018.09.008>.
  48. Lau A.S.-Y, Xiao J.-Z, Liang M.-T. *Bifidobacterium* for infants: essence and efficacy. In: *Microbiology monographs*. Springer, Cham. 2015. pp. 39–72. [https://doi.org/10.1007/978-3-319-23213-3\\_3](https://doi.org/10.1007/978-3-319-23213-3_3)
  49. Bettelheim KA, Goldwater PN. *Escherichia coli* and sudden infant death syndrome. *Front Immunol*. 2015; <https://doi.org/10.3389/fimmu.2015.00343>.
  50. Taddei CR, Oliveira FF, Duarte RTD, Talarico ST, Takagi EH, Carvalho IIR, Gomes FMS, Brandt K, Martinez MB. High abundance of *Escherichia* during the establishment of fecal microbiota in Brazilian children. *Microbial Ecol*. 2014;67(3):624–34. <https://doi.org/10.1007/s00248-014-0381-x>.
  51. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, Vadder FD, Arora T, Hallen A, Martens E, Björck I, Bäckhed F. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell Metabolism*. 2015;22(6):971–82. <https://doi.org/10.1016/j.cmet.2015.10.001>.
  52. Hiippala K, Kainulainen V, Kalliomäki M, Arkkila P, Satokari R. Mucosal prevalence and interactions with the epithelium indicate commensalism of *Sutterella* spp. *Front Microbiol*. 2016; <https://doi.org/10.3389/fmicb.2016.01706>.
  53. Wagner J, Short K, Catto-Smith AG, Cameron DJS, Bishop RF, Kirkwood CD. Identification and characterisation of *Pseudomonas* 16s ribosomal DNA from ileal biopsies of children with Crohn's disease. *PLoS ONE*. 2008;3(10):3578. <https://doi.org/10.1371/journal.pone.0003578>.
  54. Mpofu CM, Campbell BJ, Subramanian S, Clarke SM, Hart CA, Cross A, Roberts CL, McGoldrick A, Edwards SW, Rhodes JM. Microbial mannan inhibits bacterial killing by macrophages: A possible pathogenic mechanism for Crohn's disease. *Gastroenterology*. 2007;133(5):1487–98. <https://doi.org/10.1053/j.gastro.2007.08.004>.
  55. Subramanian S, Roberts CL, Hart CA, Martin HM, Edwards SW, Rhodes JM, Campbell BJ. Replication of colonic Crohn's disease mucosal *Escherichia coli* isolates within macrophages and their susceptibility to antibiotics. *Antimicrob Agents Chemother*. 2007;52(2):427–34. <https://doi.org/10.1128/aac.00375-07>.
  56. Rajeev L, Malanowska K, Gardner JF. Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions. *Microbiol Mol Biol Rev*. 2009;73(2):300–9. <https://doi.org/10.1128/mmb.00038-08>.
  57. Zrímeč J, Lapanje A. DNA structure at the plasmid origin-of-transfer indicates its potential transfer range. *Sci Rep*. 2018; <https://doi.org/10.1038/s41598-018-20157-y>.
  58. Cuevas A, Kanoksilapatham W, Gonzalez JM. Evidence of horizontal gene transfer by transposase gene analyses in *Ferrovibacterium* species. *PLoS ONE*. 2017;12(4):0173961. <https://doi.org/10.1371/journal.pone.0173961>.
  59. Wang J, Liu Y, Liu Y, Du K, Xu S, Wang Y, Krupovic M, Chen X. A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic Acids Res*. 2018;46(5):2521–36. <https://doi.org/10.1093/nar/gky005>.
  60. Park W, Jeon CO, Hohnstock-Ashe AM, Winans SC, Zylstra GJ, Madsen EL. Identification and characterization of the conjugal transfer region of the pCg1 plasmid from naphthalene-degrading *Pseudomonas putida* CG1. *Appl Environ Microbiol*. 2003;69(6):3263–71. <https://doi.org/10.1128/aem.69.6.3263-3271.2003>.
  61. Neyfakh AA. Natural functions of bacterial multidrug transporters. *Trends Microbiol*. 1997;5(8):309–13. [https://doi.org/10.1016/s0966-842x\(97\)01064-0](https://doi.org/10.1016/s0966-842x(97)01064-0).
  62. Bay DC, Rommens KL, Turner RJ. Small multidrug resistance proteins: a multidrug transporter family that continues to grow. *Biochimica Biophysica Acta (BBA)*. 2008;1778(9):1814–38. <https://doi.org/10.1016/j.bbamem.2007.08.015>.
  63. Eicher T, Cha H-J, Seeger MA, Brandstatter L, El-Delik J, Bohnert JA, Kern WV, Verrey F, Grutter MG, Diederichs K, Pos KM. Transport of drugs by the multidrug transporter AcrB involves an access and a deep binding pocket that are separated by a switch-loop. *Proc Natl Acad Sci*. 2012;109(15):5687–92. <https://doi.org/10.1073/pnas.1114944109>.
  64. Kuroda T, Tsuchiya T. Multidrug efflux transporters in the MATE family. *Biochimica Biophysica Acta (BBA)*. 2009;1794(5):763–8. <https://doi.org/10.1016/j.bbapap.2008.11.012>.
  65. Nikaido H, Takatsuka Y. Mechanisms of RND multidrug efflux pumps. *Biochimica Biophysica Acta (BBA)*. 2009;1794(5):769–81. <https://doi.org/10.1016/j.bbapap.2008.10.004>.
  66. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. 2018;36(10):996–1004. <https://doi.org/10.1038/nbt.4229>.
  67. Ye C, Torsello A, Wilson RC, Hancock ER. Thermodynamics of time evolving networks. In: *Graph-Based Representations in Pattern Recognition*. Chamm: Springer. 2015. p. 315–24. [https://doi.org/10.1007/978-3-319-18224-7\\_31](https://doi.org/10.1007/978-3-319-18224-7_31).
  68. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*. 2019. p. 9.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.