Chapter 15

Transkingdom Networks: A Systems Biology Approach to Identify Causal Members of Host–Microbiota Interactions

Richard R. Rodrigues, Natalia Shulzhenko, and Andrey Morgun

Abstract

Improvements in sequencing technologies and reduced experimental costs have resulted in a vast number of studies generating high-throughput data. Although the number of methods to analyze these "omics" data has also increased, computational complexity and lack of documentation hinder researchers from analyzing their high-throughput data to its true potential. In this chapter we detail our data-driven, transkingdom network (TransNet) analysis protocol to integrate and interrogate multi-omics data. This systems biology approach has allowed us to successfully identify important causal relationships between different taxonomic kingdoms (e.g., mammals and microbes) using diverse types of data.

Key words Omics, Transkingdom, Network analysis, Causal relationships

1 Introduction

Over the last decade assessing eukaryotic and prokaryotic genomes and transcriptomes have become extremely easy. With technologies like microarrays and next-generation sequencing, investigators now have faster and cheaper access to high-throughput "-omics" data [1]. This in turn has increased the number of analysis methods [2] and allows for the exploration of new and different biological questions to provide insights and better understanding of host, host–microbial systems, and diseases [3–5].

Studies usually focus on identifying differences between "groups" (e.g., healthy versus diseased or treatment versus control) or changes across a time course (e.g., development of an organism or progression of a disease). Depending on the biological questions, such studies generate one or more types of omics data [6–8], e.g., host gene expression and gut microbial abundance. Typically, studies analyze these omics data separately, comparing gene

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-4939-8728-3_15) contains supplementary material, which is available to authorized users.

Robert G. Beiko et al. (eds.), *Microbiome Analysis: Methods and Protocols*, Methods in Molecular Biology, vol. 1849, https://doi.org/10.1007/978-1-4939-8728-3_15, © Springer Science+Business Media, LLC, part of Springer Nature 2018

expression and microbial abundance between groups or across stages. Although such analysis methods have been very useful, they do not directly answer the most critical questions of host–microbiota interactions, i.e., which microbes affect specific pathways in the host and which host pathways/genes control specific members of the microbial community? Therefore, to answer those questions, these analyses are usually followed by literature searches to identify relationships between host genes and microbes.

Different algorithms and methods have been proposed to integrate multi-omics data [9–13]. More recently, a few published studies have not only integrated microbiome and host data, but have also been able to successfully test their computational predictions in the laboratory [14–19]. In this chapter we describe our data-driven, transkingdom network (TransNet) analysis pipeline (Fig. 1) that has allowed us to make validatable computational inferences. We construct networks using correlations between differentially expressed elements (e.g., genes, microbes) and integration of high-throughput data from different taxonomic kingdoms (e.g., human and bacteria). In fact, TransNet analysis can be applied to integrate any "Transomics" data, between as well as within taxonomic kingdoms, e.g., miRNA and gene expression, protein and metabolite, bacterial and host gene expression, or copy number, methylation, and gene expression, provided the different data are obtained from the same samples. Interrogation of this network allows us to pinpoint important causal relationships between data. For example, using this method we inferred and validated: (1) microbes and microbial genes controlling a specific mammalian pathway [15]; (2) a microbe that mediates effect of one host pathway on another [14]; (3) a host gene that mediates control of gut microbe through an upstream master regulator gene [14]. Below we show how TransNet analysis can be used to integrate host gene expression with microbial abundance to create transkingdom networks.

2 Materials

2.1 Program Availability

Our transkingdom network analysis pipeline is independent of programming language or software. However, for ease of access and usage simplicity, we have provided our pipeline as a convenient R package TransNetDemo (https://github.com/richrr/TransNetDemo) and supplementary document (File S1) in addition to the description provided. Although the user can choose to perform the following steps in a programming language or software of their choice, we suggest using our R package.

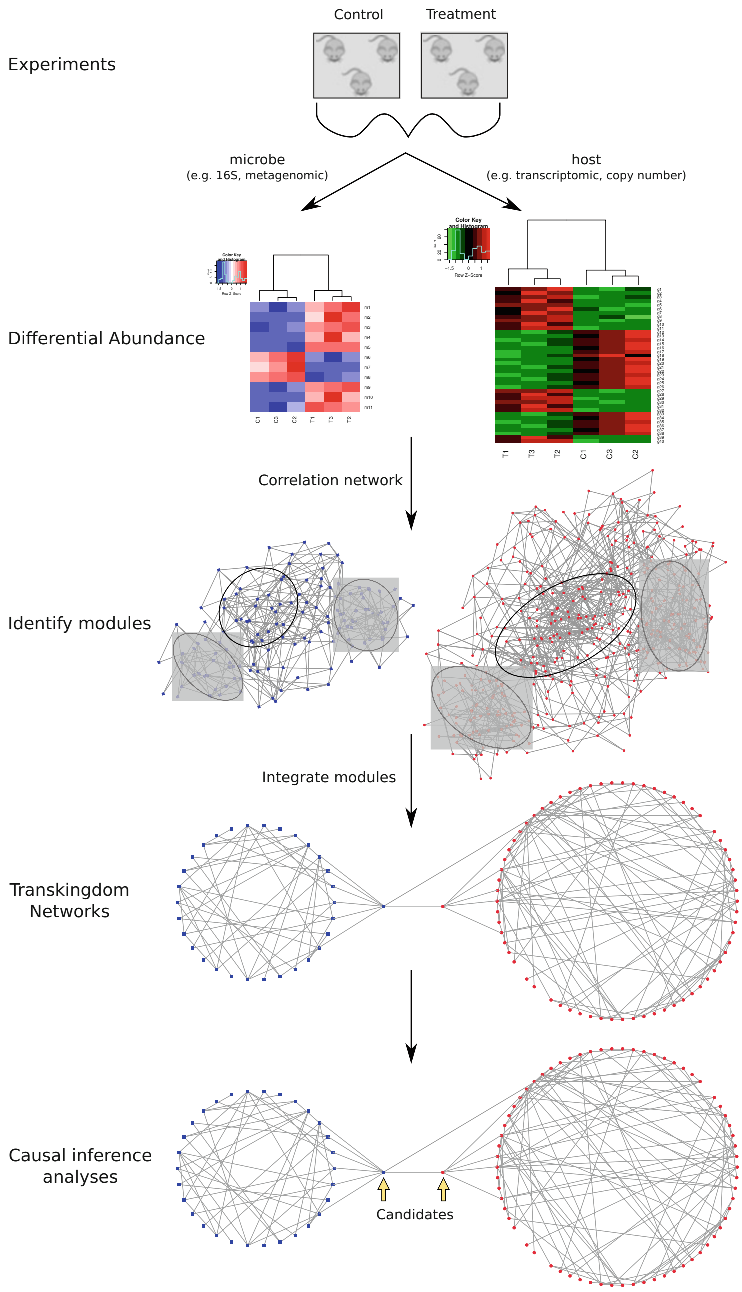

Fig. 1 Overview of transkingdom network analysis. Omics data for multiple data types (e.g., microbial, gene expression) are analyzed to identify differentially abundant elements (e.g., microbes, genes). For each group (e.g., treatment or control) co-expression networks are constructed for each data type followed by the identification of dense subnetworks (modules). Calculating correlations between module elements of the different data types creates the "transkingdom" network. Network interrogation of the transkingdom network allows identification of causal members and regulatory relationships

2.1.1 Required R Packages

Install the following packages along with their dependencies: stringr, ProNet, igraph, ggplot2, gplots from CRAN (https://cran.r-project.org/). The following commands will automatically install the required version of the packages (https://github.com/richrr/TransNetDemo/blob/master/DESCRIPTION).

2.1.2 Installing TransNetDemo

- library(devtools)
- install_github("richrr/TransNetDemo")
- library(TransNetDemo)

2.1.3 Code Referenced in the Chapter

The following scripts (available at https://github.com/richrr/TransNetDemo/tree/master/inst/demo and File S1) guide users in running TransNet:

- GeneDemo.R
- MicrobeDemo.R
- GeneMicrobeDemo.R
- Heatmaps.R

The following functions (available at https://github.com/richrr/TransNetDemo/tree/master/R and File S1) are used in the pipeline:

- Apply_sign_cutoffs.R
- Calc_bipartite_betweeness_centrality.R
- Calc_combined.R
- Calc_cor.R
- Calc_median_val.R
- Check_consistency.R
- Compare_groups.R
- Correlation_in_group.R
- Diff_abundance.R
- Get_shortest_paths.R
- Get_template_matrix.R
- Identify_subnetworks.R
- Puc_compatiable_network.R

2.2 Data Sources

Due to a variety of data generation technologies, biological questions, and software, description of every possible analysis is beyond the scope of this chapter. We expect that the user has access to tab-delimited file(s) containing the measurements of biological data type, e.g., gene expression, copy number, methylation, miRNA, or microbial abundances across samples. Depending on the data type the user can find reviews and protocol papers describing the analysis needed to produce “abundance” tables [20–24].

The transkingdom network analysis method can be applied to any experimental design (e.g., treated/untreated, control/disease). As an example we will use simulated data from a simple experimental design, where 25 mice each are fed either high fat high sucrose (HFHS) or normal chow diet (NCD) for 8 weeks, to investigate the effects of diet on host–microbial interactions. At the end of the experiment, among other phenotypic measurements (e.g., body weight, enzyme levels, hormone levels), the gene expression levels and microbial abundance in the gut (e.g., ileum) of the samples were measured. Depending on resource availability, high confidence and consistent results can be achieved by increasing the number of samples per group and/or repeating the above experiment multiple times. In this example data, we have two such experiments. A brief description of how to generate the abundance tables is mentioned below. Information about how the network analysis protocol can be adapted to answer some other biological questions have been mentioned in Subheading 4 of this chapter.

2.3 Gene Expression Analysis

Several different technologies, each with their own pros and cons, allow for the measuring of transcriptome levels in an organism. Although microarrays were extensively used over the last two decades, the availability of cheap and efficient library preparation kits and sequencing methods allow for the expression measurements of known and novel genes using RNA-Seq technologies [25].

In case of RNA-Seq data, the sequencing facilities usually provide fastq files that contain raw reads per sample (demultiplexed) (*see* **Note 1**). Here, the number of reads corresponding to a particular gene is proportional to that gene's expression level. Software like FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), PRINSEQ [26], or cutadapt [27] can be used for adapter removal and quality control. Depending on the availability of a gold-standard reference host genome sequence, gene expression abundance can be measured using the Tuxedo [21] or Trinity [28] pipeline. Both of these pipelines permit the analysis of single or paired reads and different read lengths (*see* **Note 2**) while outputting a file containing the expression levels (number of reads) of genes (rows) present in each sample (columns). The obtained read counts can be normalized by simple (e.g., quantile normalization, counts per million (CPM), reads per kilobase per million mapped reads (RPKM) [29]) or more sophisticated methods (e.g., DESEQ [30], edgeR [31]) (*see* **Notes 3** and **4**).

2.4 Microbial Abundance Analysis

The advent of next-generation technologies has helped in the study of microbial richness and diversity. Scientists no longer need to rely on cultivation methods and can directly sequence the microbiome, helping explore previously unknown microbes. The amplicon-based sequencing technologies rely on using a gene marker (16S

[32, 33] ribosomal RNA gene, Internal Transcribed Spacer [34, 35], etc.) to identify microbial presence and abundance. Although relatively cheaper than shotgun metagenomics, they rely on databases of known genomic markers to identify microbes and rarely provide taxonomy at the species or strain levels. The shotgun metagenomics sequencing approach does a better job at surveying the entire genome of microbes since it does not focus on amplifying specific genes. Consequently, it provides fine-grained taxonomic information along with a more accurate representation of the microbial structure and function, including the previously unknown "dark matter" microbes [36].

Software like QIIME [37], MOTHUR [38], etc. provide all-in-one toolkits that can demultiplex, perform quality control, and analyze the amplicon-based sequences. Similar to RNA-Seq data, the fastq files obtained from the sequencing facility need to be processed for the removal of barcodes, adapter, and primers followed by filtering to retain high quality sequences. The reads are grouped (binned) per sequence similarity (usually at 97% threshold) into operational taxonomic units (OTUs). The taxonomy of a known microbe (or the ancestor taxonomy of the top matches) closest to the representative sequence of the OTU is assigned to all the reads in that OTU. The tools output a file containing the abundance (number of reads) of OTUs (rows) present in each sample (columns). The obtained read counts can be relativized or cumulative sum scaling (CSS) [39] normalized.

Shotgun metagenomic data can be analyzed [36] using tools such as MG-RAST [40], MEGAN [41], MetaPhlAn [42], and HUMAnN [43]. Although most of these software packages provide taxonomic and functional analyses, they are not standalone. Demultiplexing and quality control need to be done before the reads are imported in the software. Especially in case of host–microbe systems, PuMA (http://blogs.oregonstate.edu/morgunshulzhenkolabs/softwares/puma) provides an all-inclusive software pipeline that can be more user-friendly. PuMA uses cutadapt for quality control and Bowtie [44] to identify reads that match the host genome and discards these "contaminating" reads from downstream analysis. The remaining microbial reads are aligned to a database of known protein sequences using DIAMOND [45], followed by taxonomic and functional (e.g., SEED, COG, KEGG) assignments using MEGAN. PuMA outputs a file containing the abundance of microbes and pathways (rows) in each sample (columns). The appropriate normalization techniques from the RNA-Seq or amplicon sequencing methods can be performed on the abundance table.

In summary, the user needs at least one of each of the following files before starting network analysis:

- Mapping file: tab-delimited file containing the group (e.g., treated/untreated, control/disease) affiliation for each sample with "Factor" and "SampleID" as column headers, respectively.
- Data files: tab-delimited files containing the abundance of elements (host genes and microbes) per sample, where the elements and samples are rows and columns, respectively. Importantly, each sample must have both types of data available.
  - Normalized gene expression file: the column "IdSymbol" contains the unique genes while the remaining columns contain their expression levels across different samples.
  - Normalized otu abundance file: the column "IdSymbol" contains the unique microbes while the remaining columns contain their abundance across different samples.

3 Methods

The following steps will help to identify key elements of a system from high-confidence modules of a multi-omics network. We show the first few steps with the gene abundance file(s) using the code from the GeneDemo.R (GD) file available in our package. It is straightforward to run similar steps on the microbe abundance file (s); however, we have also provided the code in MicrobeDemo.R (MD) file for ease of use.

- Start by setting defaults for variables that you will use in the analysis, such as significance thresholds (GD: lines 7–9), groups to be compared (GD: lines 11–13), and headers of relevant columns from the mapping (GD: lines 14–15) and abundance files (GD: line 16).
- Next you want to identify the differentially expressed elements (GD: line 29). The network analysis can be performed using all (differentially and non-differentially expressed) elements (genes, microbes, etc.). However, we suggest identifying the elements that show differential abundance between groups (*see* **Notes 5** and **6**), using code from Compare_groups.R (Cg) file, to focus on the important elements and make the analyses computationally efficient.
  - Read from the mapping file to extract the samples from each group (Cg: lines 11–20).
  - Read from the gene abundance file (Cg: lines 22–25).
  - Then perform test for differential abundance using code from Diff_abundance.R (Da) file. This function returns the mean and median for each group along with the fold change and *p*-value (Da: lines 8–28).

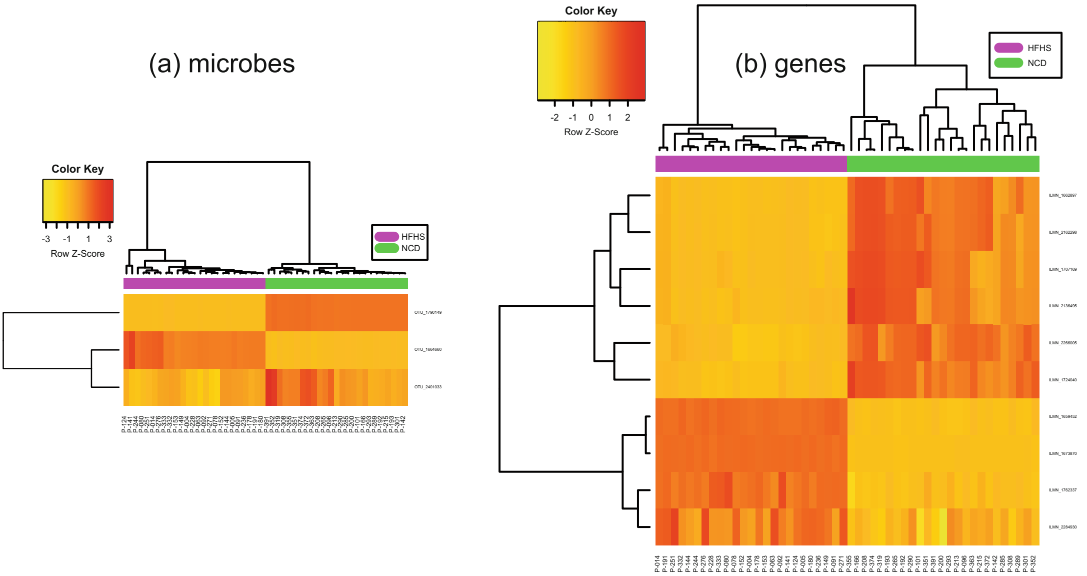

Fig. 2 Heat map from hierarchical clustering of differentially abundant elements. Rows indicate (**a**) microbes and (**b**) genes, while columns indicate samples. The purple and green colors indicate samples belonging to the groups A (HFHS) and B (NCD), respectively. The yellow and orange colors indicate decrease and increase, respectively, in expression or abundance, and color intensity corresponds to the level of fold change

  - Next, account for multiple testing using Benjamini-Hochberg's FDR calculation (Cg: lines 38–41).
  - Finally, select the differentially expressed genes using appropriate FDR cutoff (GD: line 33) (<0.05) (Fig. 2b).

- We highly recommend that if you have datasets obtained from replicate experiments or in different sample cohorts that you perform the above steps for each experiment and do meta-analysis [16, 18, 46, 47] (GD: lines 39–47).
  - First, the meta-analysis selects for genes that show fold change direction consistency across datasets (Check_consistency.R), e.g., upregulated (or downregulated) across all experiments.
  - Second, for the genes showing consistent fold change direction use Fisher's method to calculate a combined *p*-value (Calc_combined.R) from the individual *p*-values (from comparison test) across multiple experiments.
  - Then apply appropriate significance thresholds (Apply_sign_cutoffs.R) based on individual *p*-value (<0.3) in each dataset, combined (Fisher's) *p*-value across datasets (<0.05), and FDR (<0.1) across the combined *p*-values to identify consistently differentially abundant elements.
  - Ensuring the same direction of regulation in all datasets and restricting individual *p*-values at each individual dataset allows

controlling of heterogeneity between datasets. Note that mere calculation of Fisher *p*-value for meta-analysis followed by application of FDR is not sufficient for accurate identification of differential abundance/expression.

- Determining associations between elements (e.g., genes and/or microbes) is central for network reconstruction. Defining strength and sign of correlation (GD: line 56) can help to determine whether two elements (i.e., biological entities represented by nodes in a network) have a positive or negative interaction. Such information about potential relationships (*see* **Note 7**), using code from Correlation_in_group.R (Cig) file, is crucial for interrogating and understanding the regulatory mechanisms between elements. Note, correlations are calculated using data from samples representing one group (phenotypic class), never pooling samples from all groups for estimation of correlation. Therefore, the following steps should be performed for each group separately.
  - Read from the mapping file to extract the samples from a group (Cig: lines 11–18).
  - Read from the gene abundance file (Cig: lines 20–23).
  - Then create pairs (Cig: lines 25–32) from the consistent genes obtained in the previous step.
  - Next perform test for correlation on gene pairs using code from Calc_cor.R (Ccr) file. This function returns the correlation and *p*-value (Ccr: lines 8–17).
  - Next, account for multiple testing using Benjamini-Hochberg's FDR calculation (Cig: lines 48–50).
  - Finally, select the significantly correlated gene pairs using appropriate FDR cutoff (GD: line 60) < 0.1.
- We highly recommend that if you have datasets obtained from replicate experiments or different sample cohorts that you perform the above steps for each experiment and do meta-analysis (GD: lines 65–72).
  - First, the meta-analysis selects for gene pairs that show correlation direction consistency across datasets (Check_consistency.R), e.g., positive (or negative) across all experiments.
  - The next steps of combining the individual *p*-values (from correlation test) and applying multiple significance cutoffs are similar to those in the meta-analysis of genes.
- At this point you have a network for a single group where nodes are genes and edges indicate significant correlation. Next, we identify the proportion of unexpected correlations (PUC) [48] (GD: line 83). Edges in a network where the sign of correlations does not correspond to the direction of change are unexpected (*see* **Note 8**), are not likely to contribute to the process under

investigation, and hence discarded using code from Puc_compatiable_network.R (Pcn) file.

- First, for each gene pair identify the sign of correlation (Pcn: lines 47–53).
- Second, calculate if each gene in the pair has the same direction of regulation (i.e., fold change) (Pcn: lines 56–65).
- Pairs are expected and kept (Pcn: lines 70–80) if they satisfy either of these conditions:

  Positively correlated genes have the same fold change direction.

  Negatively correlated genes have different fold change direction.

- At this point you have a network consisting of regulatory relationships. Next, the obtained network can be systematically studied to answer different biological questions (*see* **Note 9**). Most often, network interrogation relies on identifying highly inter-connected sets of nodes. Such a subnetwork is called a module (or cluster). Identify clusters (GD: line 89) using the MCODE method (*see* **Note 10**) from the Identify_subnetworks.R file (Fig. 3b).
- Repeat the above steps for the microbial (or any other data type) abundance file(s) to obtain heat map (Fig. 2a) and clusters (Fig. 3a) per biological data type (e.g., genes, microbes). Refer to the code in MicrobeDemo.R file.

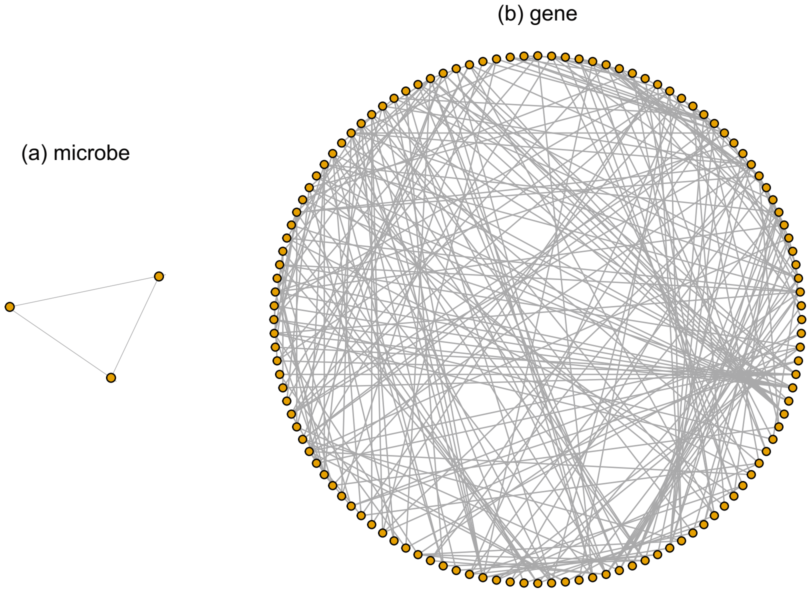

Fig. 3 Clusters obtained from the correlation networks. The PUC compatible (**a**) microbe and (**b**) gene networks for an individual group (HFHS) are mined to identify densely connected subnetworks. Edges indicate significant correlation between elements

- The next step is to integrate subnetworks to create transkingdom networks using code from the GeneMicrobeDemo.R (GMD) file. Note that at this point you have already identified modules from the gene and the microbe networks. Similar to the above steps, create pairs between nodes from the different modules (GMD: line 29) (*see* **Note 11**), calculate correlations within a group (GMD: line 32), and identify significant pairs based on single (GMD: line 36) or meta (GMD: lines 41–49) analysis. Next, apply PUC analysis and remove unexpected edges from this transkingdom (gene–microbe) network as it is done for regular gene expression (and microbial abundance) network (GMD: line 58).
- Combining the gene–gene correlations (edges from the gene subnetworks) (GMD: line 74), microbe–microbe correlations (edges from the microbe subnetworks) (GMD: line 77), and the gene–microbe correlations (GMD: line 80) creates the full transkingdom network (GMD: line 83) (Fig. 4).

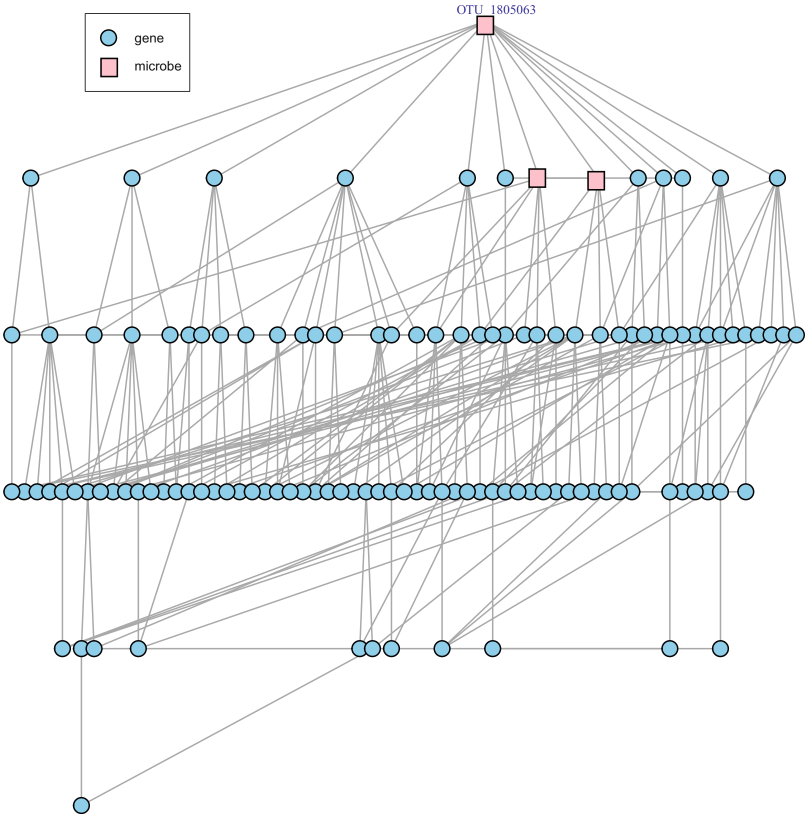

Fig. 4 Transkingdom network. A full network, for the HFHS group, contains gene–gene, microbe–microbe, and gene–microbe edges. Edges indicate significant correlation between elements. The blue circle and pink square indicate gene and microbe nodes, respectively. The labeled node has the highest BiBC measurement among microbes and is therefore considered to be important and a potential causal player in the experiment

- Finally, identify elements that are crucial for crosstalk between the different modules in a network using bipartite betweenness centrality (BiBC) (GMD: lines 92–119) (*see* **Note 12**). This approach involves calculating the shortest paths between nodes from different modules using code from Get_shortest_paths.R file. The elements with the highest BiBC measurement (GMD: lines 123–128) are more likely to be critical in mediating the transfer of signals between the different modules of a network and candidates for further experimentation.

4 Notes

The above protocol was written for a step-by-step introduction to transkingdom network analysis. Although the above experimental setup and analyses should suffice in most cases please see the following suggestions for other alternatives to the analysis.

1. Attaching unique barcodes to samples in the amplification step of sequencing library preparation allows multiple samples to be pooled in a single sequencing run. This process is termed as multiplexing. After sequencing the barcodes can be used to separate reads per sample, a process termed as demultiplexing.
2. The DNA fragment (template) can be sequenced in single or both directions and is referred to as single-end or paired-end sequencing. Read length refers to the number of bases sequenced per DNA fragment. For example, 250-bp paired-end sequencing means that 250 bases were sequenced from each end of the DNA giving one fastq file per sample for each end of the read.
3. Data normalization is a crucial step in analysis and network reconstruction [49]; hence, choose the appropriate normalization method for your biological data [50, 51]. Normalization methods differ in how they account for the sequencing depth (in next-generation sequencing data), gene or transcript length, estimation of data variability; however, no normalization method universally outperforms other methods. However, if unsure about which normalization to use we recommend quantile normalization since, in our experience, it works well for most biological data.
4. In the case of microarray data, hybridization facilities usually provide scan files (Affymetrix CEL, Illumina IDAT, or GenePix GPR) that contain the intensity of probes per sample. Here, the probe intensity is proportional to the corresponding gene expression level. Software like Affymetrix® Expression Console™, Illumina's GenomeStudio, and GenePix® Pro, as well as packages like affy [52] and limma [53], allow for background

correction, normalization, and summarized probe intensities while outputting a file containing the expression levels of genes (rows) present in each sample (columns).

5. Depending on the experimental design and biological question apply appropriate parametric (paired or unpaired *t*-test, analysis of variance (ANOVA), multivariate ANOVA (MANOVA), etc.) and nonparametric (Man-Whitney, Wilcoxon rank sum test, Multi-response Permutation Procedures (MRPP), etc.) tests to identify differential abundance.

6. It is common practice to visualize the levels of differentially abundant elements. The code from Heatmaps.R file can help to visualize the significant genes and microbes from our example.

7. Pearson or Spearman correlation analyses between two elements from the same samples should suffice. However, use partial correlation [54] or other methods [55] to detect correlations and reduce indirect interactions.

8. If two elements have a regulatory relationship we expect them to behave in certain ways. For example, consider two groups. Two positively correlated genes in a group should have the same direction of fold change between two groups. On the other hand, two negatively correlated genes should have the opposite direction of fold change [48].

9. The network analysis can be extended to identify differentially correlated genes in co-expression networks obtained for the different groups and uncover regulatory mechanisms in phenotypic transitions [56, 57].

10. Cfinder and graph clustering (MCL) [19] are some other tools to help identify modules in networks.

11. In our example, gene expression was correlated with taxon abundance to identify genes and microbes with similar or opposite variation across samples within a group. Such pairs indicate potential associations between the nodes. Correlations between other data types are possible, provided that the measurements are available from the same set of samples.

12. In our example, we used the bipartite betweenness centrality measure to identify elements that are important for crosstalk between the different modules of the network. This is because the nodes with high BiBC scores lie on the largest number of shortest paths taken by nodes between the different modules to communicate with each other and therefore have more control over information passing in the network. BiBC assumes that every pair of node pass equally important messages and that nodes always communicate using the shortest paths which may not be true in each case. Therefore, depending on the biological question, the user can inspect multiple network

topology properties such as the degree, eccentricity, and centrality measures using *NetworkAnalyzer* in Cytoscape to identify important elements in the full transkingdom network.

Acknowledgments

The authors thank Karen N. D'Souza, Khiem Lam, and Dr. Xiaoxi Dong for their help in writing the book chapter. This work was supported by the NIH U01 AI109695 (AM) and R01 DK103761 (NS).

References

1. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5 (1):16–18
2. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11 (1):31–46
3. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351
4. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141
5. Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92 (5):255–264
6. Erickson AR et al (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 7(11):e49138
7. Moreno-Risueno MA, Busch W, Benfey PN (2010) Omics meet networks—using systems approaches to infer regulatory networks in plants. Curr Opin Plant Biol 13(2):126–131
8. Imhann F et al (2016) Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. In: Gut
9. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7(3):198–210
10. Gehlenborg N et al (2010) Visualization of omics data for systems biology. Nat Methods 7(3 Suppl):S56–S68
11. Poirel CL et al (2013) Reconciling differential gene expression data with molecular interaction networks. Bioinformatics 29(5):622–629
12. Zhang W, Li F, Nie L (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology 156(Pt 2):287–301
13. Greer R et al (2016) Investigating a holobiont: Microbiota perturbations and transkingdom networks. Gut Microbes 7(2):126–135
14. Greer RL et al (2016) Akkermansia muciniphila mediates negative effects of IFNgamma on glucose metabolism. Nat Commun 7:13329
15. Morgun A et al (2015) Uncovering effects of antibiotics on the host and microbiota using transkingdom gene networks. Gut 64 (11):1732–1743
16. Mine KL et al (2013) Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. Nat Commun 4:1806
17. Schirmer M et al (2016) Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell 167(4):1125–1136 e8
18. Shulzhenko N et al (2011) Crosstalk between B lymphocytes, microbiota and the intestinal epithelium governs immunity versus metabolism in the gut. Nat Med 17 (12):1585–1593
19. Dong X et al (2015) Reverse enGENEering of Regulatory Networks from Big Data: A Roadmap for Biologists. Bioinform Biol Insights 9:61–74

20. Caporaso JG et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336
21. Trapnell C et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578
22. Laird PW (2010) Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 11(3):191–203
23. Krumm N et al (2012) Copy number variation detection and genotyping from exome sequence data. Genome Res 22(8):1525–1532
24. Perez-Diez A, Morgun A, Shulzhenko N (2007) Microarrays for cancer diagnosis and classification. Adv Exp Med Biol 593:74–85
25. Zhao S et al (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9(1):e78644
26. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863–864
27. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJ 17(1):10
28. Haas BJ et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8):1494–1512
29. Mortazavi A et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):621–628
30. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106
31. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40 (10):4288–4297
32. Stackebrandt E, Goebel BM (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Evol Microbiol 44(4):846–849
33. Lane DJ et al (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82 (20):6955–6959
34. Brookman JL et al (2000) Identification and characterization of anaerobic gut fungi using molecular methodologies based on ribosomal ITS1 and 185 rRNA. Microbiology 146 (Pt 2):393–403
35. Schoch CL et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 109(16):6241–6246
36. Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5:209
37. Kuczynski J et al (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics 10:7 Chapter 10. Unit
38. Schloss PD et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537–7541
39. Paulson JN et al (2013) Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10(12):1200–1202
40. Meyer F et al (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 9:386
41. Huson DH, Weber N (2013) Microbial community analysis using MEGAN. Methods Enzymol 531:465–485
42. Segata N et al (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9 (8):811–814
43. Lindgreen S, Adair KL, Gardner PP (2016) An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep 6:19233
44. Langmead B et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
45. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60
46. Rodrigues RR, Barry CT (2011) Gene pathway analysis of hepatocellular carcinoma genomic expression datasets. J Surg Res 170(1): e85–e92
47. Morgun A et al (2006) Molecular profiling improves diagnoses of rejection and infection in transplanted organs. Circ Res 98(12): e74–e83
48. Yambartsev A et al (2016) Unexpected links reflect the noise in networks. Biol Direct 11 (1):52
49. Saccenti E (2017) Correlation patterns in experimental data are affected by normalization procedures: consequences for data analysis and

network inference. J Proteome Res 16 (2):619–634

50. Hua YJ et al (2008) Comparison of normalization methods with microRNA microarray. Genomics 92(2):122–128
51. Li P et al (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinform 16:347
52. Gautier (2004) L., et al., affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20(3):307–315
53. Ritchie (2015) M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47
54. de la Fuente A et al (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20 (18):3565–3574
55. Weiss S et al (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J 10 (7):1669–1681
56. Thomas LD et al (2016) Differentially correlated genes in co-expression networks control phenotype transitions. F1000Res 5:2740
57. Skinner J et al (2011) Construct and Compare Gene Coexpression Networks with DAPfinder and DAPview. BMC Bioinform 12:286