

Contents

1	Introduction to QIIME 2	1
1.1	Overview of QIIME 2	1
1.2	Core Concepts in QIIME 2	3
1.2.1	Artifacts	3
1.2.2	Visualizations	4
1.2.3	Semantic Type	4
1.2.4	Plugins	4
1.3	Install QIIME 2	5
1.4	Store and Track Data	7
1.5	Extract Data from QIIME 2 Archives	8
1.6	Summary	8
	References	8
2	Introduction to R for Microbiome Data	11
2.1	Some Useful R Functions	11
2.1.1	Save Data into R Data Format	12
2.1.2	Read Back Data from R Data Format Files	14
2.1.3	Use Download.File() to Download File from Website	15
2.2	Some Useful R Packages for Microbiome Data	15
2.2.1	Readr	16
2.2.2	Ggpubr	17
2.2.3	Tidyverse	23
2.3	Specifically Designed R Packages for Microbiome Data	24
2.3.1	Phyloseq	24
2.3.2	Microbiome	34
2.3.3	ampvis2	47
2.3.4	curatedMetagenomicData	48
2.4	Some R Packages for Analysis of Phylogenetics	48
2.4.1	Ape	49

2.4.2	Phytools	50
2.4.3	Castor	52
2.5	BIOM Format and Biomformat Package	54
2.6	Creating Analysis Microbiome Dataset	57
2.7	Summary	61
	References	61
3	Basic Data Processing in QIIME 2	65
3.1	Importing Data into QIIME 2	65
3.1.1	Import FASTA Format Data	66
3.1.2	Import FASTQ Format Data	69
3.1.3	Import Feature Table	73
3.1.4	Import Phylogenetic Trees	75
3.2	Exporting Data from QIIME 2	76
3.2.1	Export Feature Table	76
3.2.2	Export Phylogenetic Trees	77
3.3	Extracting Data from QIIME 2 Archives	77
3.3.1	Extract Data Using the Qiime Tools Export Command	78
3.3.2	Extract Data Using Unzip Program on macOS	78
3.4	Filtering Data in QIIME 2	79
3.4.1	Filter Feature Table	80
3.4.2	Taxonomy-Based Tables and Sequences Filtering	83
3.4.3	Filter Distance Matrices	86
3.5	Introducing QIIME 2 View	87
3.6	Communicating Between QIIME 2 and R	87
3.6.1	Export QIIME 2 Artifacts into R Using qiime2R Package	87
3.6.2	Prepare Feature Table and Metadata in R and Import into QIIME 2	90
3.7	Summary	92
	References	92
4	Building Feature Table and Feature Representative Sequences from Raw Reads	95
4.1	Analyzing Demultiplexed Paired-End FASTQ Data	96
4.1.1	Prepare Sample Metadata	96
4.1.2	Prepare Raw Sequence Data	99
4.1.3	Import Data Files as Qiime Zipped Artifacts(.qza)	100
4.1.4	Examine and Visualize the Qualities of the Sequence Reads	101
4.2	Analyzing Demultiplexed Paired-End FASTQ Data Using DADA2 and q2-dada2 Plugin	103
4.2.1	Introduction to DADA2 and q2-dada2 Plugin	103

4.2.2	Denoise Sequences to Construct Feature Table and Feature Data with q2-dada2 Plugin	104
4.2.3	Summarize the Feature Table and Feature Data from q2-dada2 Plugin	108
4.3	Analyzing Multiplexed Paired-End FASTQ Data Using q2-dada2 Plugin	108
4.3.1	Prepare Sample Metadata	109
4.3.2	Prepare Raw Sequence Data	109
4.3.3	Import Data Files as Qiime Zipped Artifacts(.qza)	110
4.3.4	Demultiplexing Sequences	110
4.3.5	Summarize the Demultiplexing Results and Examine Quality of the Reads	112
4.4	Analyzing Demultiplexed Paired-End FASTQ Data Using Deblur and q2-deblur Plugin	113
4.4.1	Introduction to Deblur and q2-deblur Plugin	113
4.4.2	Process Initial Quality Filtering	114
4.4.3	Preliminary Works for Denoising with Deblur	114
4.4.4	Denoise Sequences with Deblur to Construct Feature Table and Feature Data	116
4.4.5	Summarize the Feature Table and Feature Data from Deblur	117
4.4.6	Remarks on DADA2 and Deblur	118
4.5	Summary	119
	References	120
5	Assigning Taxonomy, Building Phylogenetic Tree	123
5.1	Assigning Taxonomy	123
5.1.1	Bioinformatics Tools and Reference Databases	123
5.1.2	QIIME 2-Formatted and Maintained Taxonomic Reference Databases	125
5.1.3	DADA2-Formatted and Maintained Taxonomic Reference Databases	128
5.1.4	Introduction to q2-Feature-Classifer	128
5.1.5	Assign Taxonomy Using the q2-Feature-Classifer	132
5.1.6	Remarks on Taxonomic Classification	135
5.2	Building Phylogenetic Tree	135
5.2.1	Introduction to Phylogenetic Tree	135
5.2.2	Build a Phylogenetic Tree Using the Alignment and Phylogeny Commands	136
5.2.3	Remarks on the Taxonomic and Phylogenetic Trees	138
5.3	Summary	141
	References	141

6	Clustering Sequences into OTUs	147
6.1	Introduction to Clustering Sequences into OTUs	147
6.1.1	Merge Reads	148
6.1.2	Remove Non-biological Sequences	148
6.1.3	Trim Reads Length	148
6.1.4	Discard Low-Quality Reads	149
6.1.5	Dereplicate Sequences	149
6.2	Introduction to VSEARCH and q2-vsearch	149
6.3	Closed-Reference Clustering	150
6.3.1	Introduction	150
6.3.2	Implement Cluster-Features-Closed-Reference	150
6.3.3	Remarks on Closed-Reference Clustering	152
6.4	<i>De Novo</i> Clustering	153
6.4.1	Introduction	153
6.4.2	Implement Cluster-Features-De-Novo	153
6.4.3	Remarks on <i>De Novo</i> Clustering	154
6.5	Open-Reference Clustering	155
6.5.1	Introduction	155
6.5.2	Implement Cluster-Features-Open-Reference	155
6.5.3	Remarks on Open-Reference Clustering	156
6.6	Summary	157
	References	157
7	OTU Methods in Numerical Taxonomy	161
7.1	Brief History of Numerical Taxonomy	162
7.2	Principles of Numerical Taxonomy	164
7.2.1	Definitions of Characters and Taxa	165
7.2.2	Sample Size Calculation: How Many Characters?	166
7.2.3	Equal Weighting Characters	169
7.2.4	Taxonomic Rank	169
7.3	Phenetic Taxonomy: Philosophy of Numerical Taxonomy	170
7.3.1	Phenetics: Numerical-Based Empiricism Versus Aristotle's Essentialism	170
7.3.2	Classification: Inductive Theory Versus Darwin's Theory of Evolution	173
7.3.3	Biological Classifications: Phenetic Approach Versus Cladistic Approach	174
7.4	Construction of Taxonomic Structure	176
7.4.1	Defining the Operational Taxonomic Units	177
7.4.2	Estimation of Taxonomic Resemblance	178
7.4.3	Commonly Clustering-Based OTU Methods	192
7.5	Statistical Hypothesis Testing of OTUs	208
7.5.1	Hypothesis Testing on Similarity Coefficients of OTUs	208
7.5.2	Hypothesis Testing on Clustering OTUs	209

7.6	Some Characteristics of Clustering-Based OTU Methods	210
7.6.1	Some Basic Questions in Numerical Taxonomy	211
7.6.2	Characteristics of Numerical Taxonomy	212
7.7	Summary	218
	References	219
8	Moving Beyond OTU Methods	227
8.1	Clustering-Based OTU Methods in Microbiome Study	228
8.1.1	Common Clustering-Based OTU Methods	228
8.1.2	Hierarchical Clustering OTU Methods	229
8.1.3	Heuristic Clustering OTU Methods	230
8.1.4	Limitations of Clustering-Based OTU Methods	231
8.1.5	Purposes of Using OTUs in Microbiome Study	240
8.1.6	Defining Species and Species-Level Analysis	241
8.2	Moving Toward Single-Nucleotide Resolution-Based OTU Methods	250
8.2.1	Concept Shifting in Bioinformatic Analysis	250
8.2.2	Single-Nucleotide Resolution Clustering-Based OTU Methods	252
8.3	Moving Beyond the OTU Methods	254
8.3.1	Entropy-Based Methods: Oligotyping	255
8.3.2	Denoising-Based Methods	256
8.4	Discussion on Moving Beyond OTU Methods	267
8.4.1	Necessity of Targeting Single-Base Resolution	267
8.4.2	Possibility of Moving Beyond Traditional OTU Methods	268
8.4.3	Issues of Sub-OTU Methods	270
8.4.4	Prediction of Sequence Similarity to Ecological Similarity	271
8.4.5	Functional Analysis and Multi-omics Integration	271
8.5	Summary	273
	References	273
9	Alpha Diversity	289
9.1	Abundance-Based Alpha Diversity Metrics	290
9.1.1	Chao 1 Richness and Abundance-Based Coverage Estimator (ACE)	290
9.1.2	Shannon Diversity	298
9.1.3	Simpson Diversity	300
9.1.4	Pielou's Evenness	304
9.2	Phylogenetic Alpha Diversity Metrics	305
9.2.1	Phylogenetic Diversity	306
9.2.2	Phylogenetic Entropy	306
9.2.3	Phylogenetic Quadratic Entropy (PQE)	307
9.3	Exploring Alpha Diversity and Abundance	307
9.3.1	Heatmap	308

9.3.2	Boxplot	312
9.3.3	Violin Plot	316
9.4	Statistical Hypothesis Testing of Alpha Diversity	318
9.4.1	Summarize the Diversity Measures	319
9.4.2	Plot Histogram of the Diversity Distributions	319
9.4.3	Kruskal-Wallis Test	319
9.4.4	Perform Multiple Comparisons	321
9.5	Alpha Diversity Analysis in QIIME 2	322
9.5.1	Calculate Alpha Diversity Using Core-Metrics- Phylogenetic Method	322
9.5.2	Calculate Alpha Diversity Using Alpha Method	324
9.5.3	Calculate Alpha Diversity Using Alpha-Phylogenetic Method	326
9.5.4	Test for Differences of Alpha Diversity Between Groups	327
9.5.5	Alpha Rarefaction in QIIME 2	328
9.6	Summary	330
	References	331
10	Beta Diversity Metrics and Ordination	335
10.1	Abundance-Based Beta Diversity Metrics	335
10.1.1	Bray-Curtis Dissimilarity	337
10.1.2	Jaccard Dissimilarity	339
10.1.3	Sørensen Dissimilarity	340
10.2	Phylogenetic Beta Diversity Metrics	342
10.2.1	Unweighted UniFrac	342
10.2.2	Weighted UniFrac	343
10.2.3	GUniFrac	344
10.2.4	pldist	344
10.2.5	Calculate (Un)Weighted UniFrac and GUniFrac Distances Using the GUniFrac Package	345
10.2.6	Remarks on Rarefaction for Alpha and Beta Diversity Analysis	348
10.3	Ordination Methods	349
10.3.1	Introduction to Ordination	349
10.3.2	Ordination Plots in the ampvis2 Package	352
10.3.3	Principal Component Analysis (PCA)	353
10.3.4	Principal Coordinate Analysis (PCoA)	359
10.3.5	Nonmetric Multidimensional Scaling (NMDS)	362
10.3.6	Correspondence Analysis (CA)	366
10.3.7	Detrended Correspondence Analysis (DCA)	371
10.3.8	Redundancy Analysis (RDA)	374
10.3.9	Canonical Correspondence Analysis (CCA)	377
10.4	Beta Diversity Metrics and Ordination in QIIME 2	381
10.4.1	Calculate Beta Diversity Measures	381

10.4.2	Explore Principal Coordinates (PCoA) Using Emperor Plots	382
10.5	Remarks on Ordination and Clustering	384
10.6	Summary	388
	References	388
11	Statistical Testing of Beta Diversity	397
11.1	Introduction to Nonparametric MANOVA Using Permutation Tests	397
11.2	Analysis of Similarity (ANOSIM)	398
11.2.1	Introduction of ANOSIM	399
11.2.2	Perform ANOSIM Using the Vegan Package	401
11.2.3	Remarks on ANOSIM	404
11.3	Permutational MANOVA (PERMANOVA)	405
11.3.1	Introduction to PERMANOVA	405
11.3.2	Perform PERMANOVA Using the Vegan Package	409
11.3.3	Remarks on PERMANOVA	417
11.4	Analysis of Multivariate Homogeneity of Group Dispersions	418
11.4.1	Introduction to the Function betadisper()	418
11.4.2	Implement the Function betadisper()	419
11.5	Pairwise PERMANOVA	422
11.5.1	Introduction to Pairwise PERMMANOVA	422
11.5.2	Implement Pairwise PERMMANOVA Using the RVAideMemoire Package	423
11.6	Identify Core Microbial Taxa Using the Microbiome Package	426
11.7	Statistical Testing of Beta Diversity in QIIME 2	428
11.7.1	Significant Testing of Bray-Curtis Distance	429
11.7.2	Significant Testing of Jaccard Distance	430
11.7.3	Significant Testing of Unweighted UniFrac Distance	430
11.7.4	Significant Testing of Weighted UniFrac Distance	431
11.8	Summary	431
	References	431
12	Differential Abundance Analysis of Microbiome Data	435
12.1	Zero-Inflated Gaussian (ZIG) and Zero-Inflated Log-Normal (ZILN) Mixture Models	436
12.1.1	Total Sum Scaling (TSS)	437
12.1.2	Cumulative Sum Scaling (CSS)	438
12.1.3	ZIG and ZILN Models	438
12.2	Implement ZILN via metagenomeSeq	440

12.3	Some Additional Statistical Tests in metagenomeSeq	453
12.3.1	Log-Normal Permutation Test of Abundance	453
12.3.2	Presence-Absence Testing of the Proportion/Odds	454
12.3.3	Discovery Odds Ratio Testing of the Proportion of Observed Counts	455
12.3.4	Perform Feature Correlations	455
12.4	Illustrate Some Useful Functions in metagenomeSeq	456
12.4.1	Access the MRexperiment Object	456
12.4.2	Subset the MRexperiment Object	457
12.4.3	Filter the MRexperiment Object or Count Matrix	458
12.4.4	Merge the MRexperiment Object	458
12.4.5	Call the Normalized Counts Using the cumNormMat() and MRcounts()	459
12.4.6	Calculate the Normalization Factors Using the calcNormFactors()	460
12.4.7	Access the Library Sizes Using the libSize()	460
12.4.8	Save the Normalized Counts Using the exportMat()	461
12.4.9	Save the Sample Statistics Using the exportStats()	461
12.4.10	Find Unique OTUs or Features	461
12.4.11	Aggregate Taxa	462
12.4.12	Aggregate Samples	463
12.5	Remarks on CSS Normalization, ZIG, and ZILN	465
12.6	Summary	466
	References	467
13	Zero-Inflated Beta Models for Microbiome Data	469
13.1	Zero-Inflated Beta Modeling Microbiome Data	469
13.2	Zero-Inflated Beta Regression (ZIBSeq)	471
13.2.1	Introduction to ZIBseq	471
13.2.2	Implement ZIBseq	473
13.2.3	Remarks on ZIBSeq	479
13.3	Zero-Inflated Beta-Binomial Model (ZIBB)	480
13.3.1	Introduction to ZIBB	480
13.3.2	Implement ZIBB	483
13.3.3	Remarks on ZIBB	488
13.4	Summary	489
	References	489
14	Compositional Analysis of Microbiome Data	491
14.1	Introduction to Compositional Data	492
14.1.1	What Are Compositional Data?	492
14.1.2	Microbiome Data Are Treated as Compositional	492
14.1.3	Aitchison Simplex	493
14.1.4	Challenges of Analyzing Compositional Data	494

14.1.5	Fundamental Principles of Compositional Data Analysis	495
14.1.6	The Family of Log-Ratio Transformations	496
14.1.7	Remarks on Log-Ratio Transformations	498
14.2	ANOVA-Like Compositional Differential Abundance Analysis	500
14.2.1	Introduction to ALDEx2	500
14.2.2	Implement ALDEx2 Using R	505
14.2.3	Remarks on ALDEx2	517
14.3	Analysis of Composition of Microbiomes (ANCOM)	518
14.3.1	Introduction to ANCOM	519
14.3.2	Implement ANCOM Using QIIME 2	520
14.3.3	Remarks on ANCOM	526
14.4	Analysis of Composition of Microbiomes-Bias Correction (ANCOM-BC)	528
14.4.1	Introduction to ANCOM-BC	528
14.4.2	Implement ANCOM-BC Using the ANCOMBC Package	532
14.4.3	Remarks on ANCOM-BC	547
14.5	Remarks on Compositional Data Analysis Approach	549
14.6	Summary	551
	References	551
15	Linear Mixed-Effects Models for Longitudinal Microbiome Data	557
15.1	Introduction to Linear Mixed-Effects Models (LMMs)	558
15.1.1	Advantages and Disadvantages of LMMs	558
15.1.2	Fixed and Random Effects	559
15.1.3	Definition of LMMs	559
15.1.4	Statistical Hypothesis Tests	560
15.1.5	How to Fit LMMs	561
15.2	Identifying the Significant Taxa Using the nlme Package	563
15.2.1	Introduction to LMMs in Microbiome Research	563
15.2.2	Longitudinal Microbiome Data Structure	564
15.2.3	Fit LMMs Using the Read Counts as the Outcome	565
15.3	Modeling the Diversity Indices Using the lme4 and LmerTest Packages	570
15.3.1	Introduction to the lme4 and lmerTest Packages	571
15.3.2	Fit LMMs Using the Diversity Index as the Outcome	572
15.4	Implement LMMs Using QIIME 2	576
15.4.1	Introduction to the QIIME Longitudinal Linear-Mixed-Effects Command	576
15.4.2	Fit LMMs in QIIME 2	577
15.4.3	Perform Volatility Analysis	579

15.5	Remarks on LMMs	581
15.6	Summary	582
	References	582
16	Introduction to Generalized Linear Mixed Models	587
16.1	Generalized Linear Models (GLMs) and Generalized Nonlinear Models (GNLMs)	588
16.2	Generalized Linear Mixed Models (GLMMs)	588
16.3	Model Estimation in GLMMs	590
16.4	Algorithms for Parameter Estimation in GLMMs	591
16.4.1	Penalized Quasi-Likelihood-Based Methods Using Taylor-Series Linearization	592
16.4.2	Likelihood-Based Methods Using Numerical Integration	593
16.4.3	Markov Chain Monte Carlo-Based Integration	595
16.4.4	IWLS and EM-IWLS Algorithms	596
16.5	Statistical Hypothesis Testing and Modeling in GLMMs	599
16.5.1	Model Selection in Statistics	599
16.5.2	Model Selection in Machine Learning	599
16.5.3	Information Criteria for Model Selection	600
16.5.4	Likelihood-Ratio Test	605
16.5.5	Vuong Test	607
16.6	Summary	608
	References	608
17	Generalized Linear Mixed Models for Longitudinal Microbiome Data	615
17.1	Generalized Linear Mixed Models (GLMMs) in Microbiome Research	616
17.1.1	Data Transformation Versus Using GLMMs	617
17.1.2	Model Selection in Microbiome Data	618
17.1.3	Statistical Hypothesis Testing in Microbiome Data	619
17.2	Generalized Linear Mixed Modeling Using the glmmTMB Package	621
17.2.1	Introduction to glmmTMB	621
17.2.2	Implement GLMMs via glmmTMB	624
17.2.3	Remarks on glmmTMB	646
17.3	Generalized Linear Mixed Modeling Using the GLMMadaptive Package	647
17.3.1	Introduction to GLMMadaptive	647
17.3.2	Implement GLMMs via GLMMadaptive	647
17.3.3	Remarks on GLMMadaptive	659
17.4	Fast Zero-Inflated Negative Binomial Mixed Modeling (FZINBMM)	660

17.4.1	Introduction to FZINBMM	660
17.4.2	Implement FZINBMM Using the NBZIMM Package	664
17.4.3	Remarks on FZINBMM	668
17.5	Remarks on Fitting GLMMs	669
17.6	Summary	670
	References	671
18	Multivariate Longitudinal Microbiome Models	675
18.1	Overview of Multivariate Longitudinal Microbiome Analysis	675
18.1.1	Multivariate Distance/Kernel-Based Longitudinal Models	676
18.1.2	Multivariate Integration of Multi-omics Methods	677
18.1.3	Univariate Analysis Versus Multivariate Analysis	678
18.2	Nonparametric Microbial Interdependence Test (NMIT)	678
18.2.1	Introduction to NMIT	679
18.2.2	Implement NMIT Using R	680
18.2.3	Implement NMIT Using QIIME 2	687
18.2.4	Remarks on NMIT	690
18.3	The Large P Small N Problem	692
18.4	Summary	693
	References	693
	Correction to: Bioinformatic and Statistical Analysis of Microbiome Data	C1
	Index	697