ZYMO RESEARCH

ZymoBIOMICS® Service Report: Shotgun Metagenomic Sequencing

Table of Contents

1. Workflow Checklist 1
2. Methods 2
3. References 3
4. Final Report Link 4
5. Raw Sequencing Data Link 4
6. Output File Structure: Root Folder 5
7. Output File Structure: Group Comparisons 6
8. Output File Structure: Taxonomy Analysis 7
9. Output File Structure: Functional Profiling 10

1. Workflow Checklist

Sample Received	✓
Sample Quality Evaluated	✓
Sample Prepared for Sequencing	✓
Next-Gen Sequencing	✓
Sequence Quality Check	✓
Bioinformatics Processing	✓
Data/Results	✓

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

2. Methods

The samples were processed and analyzed with the ZymoBIOMICS® Shotgun Metagenomic Sequencing Service (Zymo Research, Irvine, CA).

DNA Extraction: If DNA extraction was performed, one of three different DNA extraction kits was used depending on the sample type and sample volume and were used according to the manufacturer's instructions, unless otherwise stated. The kit used in this project is marked below.

☐ ZymoBIOMICS® DNA Miniprep Kit (Zymo Research, Irvine, CA)

☐ ZymoBIOMICS® DNA Microprep Kit (Zymo Research, Irvine, CA)

☒ ZymoBIOMICS®-96 MagBead DNA Kit (Zymo Research, Irvine, CA)

☐ N/A (DNA Extraction Not Performed)

Additional Notes: N/A

Library Preparation: Genomic DNA samples were profiled with shotgun metagenomic sequencing. Sequencing libraries were prepared with the option marked below.

☐ KAPA™ HyperPlus Library Preparation Kit (Kapa Biosystems, Wilmington, MA) with up to 100 ng DNA input following the manufacturer's protocol using internal single-index 8 bp barcodes with TruSeq® adapters (Illumina, San Diego, CA)

☒ Nextera® DNA Flex Library Prep Kit (Illumina, San Diego, CA) with up to 100 ng DNA input following the manufacturer's protocol using internal dual-index 8 bp barcodes with Nextera® adapters (Illumina, San Diego, CA)

All libraries were quantified with TapeStation® (Agilent Technologies, Santa Clara, CA) and then pooled in equal abundance. The final pool was quantified using qPCR.

Sequencing: The final library was sequenced on the platform marked below.

☐ HiSeq® (Illumina, San Diego, CA)

☒ NovaSeq® (Illumina, San Diego, CA)

Control Samples: The ZymoBIOMICS® Microbial Community Standard (Zymo Research, Irvine, CA) was used as a positive control for each DNA extraction, if performed. The ZymoBIOMICS® Microbial Community DNA Standard (Zymo Research, Irvine, CA) was used as a positive control for each targeted library preparation. Negative controls (i.e. blank extraction control, blank library preparation control) were included to assess the level of bioburden carried by the wet-lab process.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

2. Methods

Bioinformatics Analysis: Raw sequence reads were trimmed to remove low quality fractions and adapters with Trimmomatic-0.33 (Bolger et al., 2014): quality trimming by sliding window with 6 bp window size and a quality cutoff of 20, and reads with size lower than 70 bp were removed. Antimicrobial resistance and virulence factor gene identification was performed with the DIAMOND sequence aligner (Buchfink et al., 2015). Microbial composition was profiled with Centrifuge (Kim et al., 2016) using bacterial, viral, fungal, mouse, and human genome datasets. Strain-level abundance information was extracted from the Centrifuge outputs and further analyzed: (1) to perform alpha- and beta-diversity analyses; (2) to create microbial composition barplots with QIIME (Caporaso et al., 2012); (3) to create taxa abundance heatmaps with hierarchical clustering (based on Bray-Curtis dissimilarity); and (4) for biomarker discovery with LEfSe (Segata et al., 2011) with default settings ($p > 0.05$ and LDA effect size > 2).

3. References

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Buchfink, B., Xie, C., Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**:59-60.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335-336.

Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **12**:1721-1729.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

4. Final Report Link

The final report was zipped and can be accessed at the link below.

https://epiquest.s3.amazonaws.com/epiquest_in1000/GHBRUNXUJ5MK5XNQ4RKJQ6JYRBE7Z8YX/report/in1000.200719.report.zip

To view the report, please follow the steps below:

1. Download the .zip file from the report link above.
2. Extract all the contents of the downloaded .zip file to your desktop.
3. Open the extracted file and open the “Report” HTML to view the report.
4. Results for each group comparison are linked in Section 1. Results include taxonomy analysis, functional pathway profiling, antibiotic resistance profiling, and virulence factor profiling.
5. Results files and figures can be saved to your desktop by right-clicking and selecting “Save as.”

The report data can also be accessed through the file explorer. The file structure of the data report through the file explorer is described on Pages 5-12 of this report.

5. Raw Sequencing Data Link

The raw sequencing data was zipped and can be accessed at:

Raw data for sample report is not available but would be linked here in your final report.

To view the raw sequencing data, download the Excel file from the link above. This file contains the list of all samples in the project and the links to download the raw data in fastq.gz format. Each sample has two files that represent Read 1 and Read 2 in paired-end sequencing. For example, “Sample 1” has two files, *zrxxxx_1_R1.fastq.gz* and *zrxxxx_2_R2.fastq.gz*.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

6. Output File Structure: Root Folder

The file structure of the root folder in the final data report is described below. Folder names are black bold font, file names are black regular font, and descriptions of the folder or file are green regular font.

Report.html	Link to HTML report with all data for all group comparisons.
#...<ComparisonName>.illumina.pe	Folder containing taxonomy analysis, functional pathway profiling, antibiotic resistance profiling, and virulence factor profiling for the comparison. See Page 7 for Output File Structure: Group Comparisons.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

7. Output File Structure: Group Comparisons

The file structure of the group comparison folders is described below. Folder names are black bold font, file names are black regular font, and descriptions of the folder or file are green regular font.

GroupOverview.html	Link to Group Overview results page for the group comparison.		
AbundanceTables *Read abundance for all taxa identified in samples.*	Abundance Table.csv	Read counts for every taxon identified in every sample at the highest resolution.	
	ReadDistribution Table.csv	Read distribution of host, microbial, and unclassified reads in each sample.	
All	Taxonomy analysis for all domains. See Page 8 for Output File Structure: Taxonomy Analysis.		
Antibiotic Resistance	**Summary**	<SampleName>. summary.csv	Raw for antibiotic resistance gene identification data.
	Summary.html	Link to table with links to antibiotic resistance results for each sample.	
Eukaryote	Taxonomy analysis for eukaryotes only. See Page 8 for Output File Structure: Taxonomy Analysis.		
FunctionalPathway	Analysis for functional pathways identified in samples. See Page 11 for Output File Structure: Functional Profiling.		
Prokaryote	Taxonomy analysis for prokaryotes only. See Page 8 for Output File Structure: Taxonomy Analysis.		
SampleInformation	**FastQC**	<SampleID>_R#_ fastqc.html	Read quality results for each read for each sample.
	ReadProcessing Summary.csv	Summary of raw reads, reads surviving, and reads dropped during quality trimming for each sample.	
	ReadProcessing Summary.html		
	Sample Metadata.csv	Summary of sample IDs, names, and group assignments.	
	Sample Metadata.html		
VirulenceFactor	**Summary**	<SampleName>. summary.csv	Raw virulence factor gene identification data.
	Summary.html	Link to table with links to virulence factor results for each sample.	
Virus	Taxonomy analysis for viruses only. See Page 8 for Output File Structure: Taxonomy Analysis.		

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

8. Output File Structure: Taxonomy Analysis

The file structure of the taxonomy analysis folders is described below. Taxonomy analysis folders included **All**, **Eukaryote**, **Prokaryote**, and **Virus**. The **Prokaryote** folder is used as an example to guide navigation through the results. The results for other taxonomy analysis folders can be navigated using the guidelines below with relevant folder/file name changes. Folders and files of greatest interest are highlighted in the table.

TaxonomyAnalysis.html	Link to Taxonomy Analysis results page for the domain and group comparison.	
AbundanceTables / <#.TaxonomicLevel> *The files in each taxonomic level folder contain the data and results at that level.*	abun_table.biom	Read abundance for every taxon in the Prokaryote domain identified in each sample.
	abun_table.tsv	
	abun_table_read_counts.tsv	Read counts for every taxon in the Prokaryote domain identified in each sample.
	ReadAbundance.tsv	Read counts for every taxon in the Prokaryote domain identified in each sample at the highest resolution.
AlphaDiversity / <#.TaxonomicLevel> *Alpha diversity measures species richness and evenness within each sample at different taxonomic levels.* *The files in each taxonomic level folder contain the data and results at that level.*	ObservedSp.csv	Raw data for each sample using observed species metric.
	ObservedSp_Barplot.png	Barplot using observed species metric.
	ObservedSp_Boxplot_<SubgroupName>.png	Boxplot using observed species metric plotted by subgroups, if indicated by customer.
	Shannon.csv	Raw data for each sample using Shannon metric.
	Shannon_Barplot.png	Barplot using Shannon metric.
	Shannon_Boxplot_<SubgroupName>.png	Boxplot using Shannon metric plotted by subgroups, if indicated by customer.

Continued next page.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

8. Output File Structure: Taxonomy Analysis

BetaDiversity / <#.TaxonomicLevel> *Beta diversity measures the dissimilarity in taxonomic composition between two or more samples.* *The files in each taxonomic level folder contain the data and results at that level.*	**biplot**	**emperor_ required_ resources**	Folder containing data and figures to generate the plots. The index.html link will not function if this folder is edited.
		index.html	Three-dimensional beta diversity plot. Color labeling based on unique sample names. Options can be changed to show different colors, different groupings, etc.
	coordinates	pcoa_binary_ jaccard_ abun_table.txt	Principal coordinates calculated using the Jaccard index.
		pcoa_bray_curtis_ abun_table.txt	Principal coordinates calculated using the Bray-Curtis matrix.
	dist	binary_jaccard_ abun_table.txt	Distance matrix calculated using the Jaccard index.
		bray_curtis_ abun_table.txt	Distance matrix calculated using the Bray-Curtis matrix.
CompositionBarplots *Taxa abundance plots at different taxonomic levels.*	**charts**	Folders containing data and figures to generate the bar charts. The bar_charts.html link will not function if these folders are edited.	
	css		
	js		
	raw_data		
	bar_charts.html	Bar charts and relative abundance tables at every taxonomic level. Figures and tables can be exported.	
Heatmaps_ <SubGroupName> / <#.TaxonomicLevel> *Taxa abundance heatmaps at different taxonomic levels.* *The files in each taxonomic level folder contain the data and results at that level.* *The results in each folder were plotted by subgroup.*	heatmap_with_ sample_clustering. pdf	Hierarchically-clustered heatmap based on Bray-Curtis dissimilarity using specified category/group information. Color labeling based on category/group.	
	heatmap_without_ sample_clustering. pdf	Heatmap using specified category/group information. Color labeling based on category/group.	
	new_abun_table.tsv	Raw data used to plot the heatmaps.	

Continued next page.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

8. Output File Structure: Taxonomy Analysis

LEfSe_ <SubGroupName> *LEfSe biomarker discovery folder.* *The results in each folder were analyzed and plotted by subgroup.*	**Figures**	Folder which contains abundance distribution plots for each taxon that is significantly different between groups. Referenced by Biomarkers.html file. Associated taxon found by matching file name to file name in Column F of LEfSe_Results.csv.
	Biomarkers.html	Interactive plot of the distribution of identified biomarkers among all samples. Click on the bars of biomarkers to access the abundance distribution profile among groups.
	Biomarkers.pdf	Identified biomarkers listed by group definition and effect size.
	Cladogram.pdf	Identified biomarkers (colored based on groups) in a context of phylogenetic tree.
	LEfSe_Input.txt	LEfSe input file.
	lefse_legend.png	Legend used for the LEfSe biomarkers HTML plot. Legend will be missing if this file is modified or deleted.
	LEfSe.Results.csv	Raw data of effect size (column D) and P-values (column E) from statistical analysis. Column A = taxon. Column C = group with the highest abundance. Column F = name of associated abundance distribution plot for taxon, found in the Figures folder.
SampleInformation	SampleMetadata.csv	File with sample ID, sample name, and group assignment information used in the group comparison.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

9. Output File Structure: Functional Profiling

The file structure of the **FunctionalPathway** folder is described below.

FunctionalResults.html	Link to Functional Profiling results page for the group comparison.	
Heatmap_GeneFamily_CPM_ <SubGroupName> *Gene family abundances with species information.* *The results in each folder were plotted by subgroup.*	Heatmap_with_ SampleClustering.pdf	Hierarchically-clustered heatmap based on Bray-Curtis dissimilarity using specified category/group information. Color labeling based on category/group.
	Heatmap_without_ SampleClustering.pdf	Heatmap using specified category/group information. Color labeling based on category/group.
	Raw_Data.tsv	Gene family abundance in counts per million* for most abundant gene families identified in each sample.
Heatmap_PathwayAbundance_ <SubGroupName> *Functional pathway abundances.* *The results in each folder were plotted by subgroup.*	Heatmap_with_ SampleClustering.pdf	Hierarchically-clustered heatmap based on Bray-Curtis dissimilarity using specified category/group information. Color labeling based on category/group.
	Heatmap_without_ SampleClustering.pdf	Heatmap using specified category/group information. Color labeling based on category/group.
	Raw_Data.tsv	Pathway abundance in counts per million* for most abundant pathways identified in each sample.
Heatmap_SpeciesPathwayAbundance_ <SubGroupName> *Functional pathway abundances with species information.* *The results in each folder were plotted by subgroup.*	Heatmap_with_ SampleClustering.pdf	Hierarchically-clustered heatmap based on Bray-Curtis dissimilarity using specified category/group information. Color labeling based on category/group.
	Heatmap_without_ SampleClustering.pdf	Heatmap using specified category/group information. Color labeling based on category/group.
	Raw_Data.tsv	Pathway abundance in counts per million* with species identification for most abundant pathways identified in each sample.

*Counts per million (CPM) are counts for the gene family or pathway of interest divided by the total read count and multiplied by one million. $CPM = \frac{Read\ Counts\ of\ Interest}{Total\ Read\ Counts} \times 10^6$

Continued next page.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

9. Output File Structure: Functional Profiling

LEfSe_SpeciesPathwayAbundance_<SubGroupName> *LEfSe biomarker discovery folder.* *The results in each folder were analyzed and plotted by subgroup.*	**Figures**	Folder which contains abundance distribution plots for each taxon that is significantly different between groups. Referenced by Biomarkers.html file. Associated taxon found by matching file name to file name in Column F of LEfSe_Results.csv.
	Biomarkers.html	Interactive plot of the distribution of identified biomarkers among all samples. Click on the bars of biomarkers to access the abundance distribution profile among groups.
	Biomarkers.pdf	Identified biomarkers listed by group definition and effect size.
	Cladogram.pdf	Identified biomarkers (colored based on groups) in a context of phylogenetic tree.
	LEfSe_Input.txt	LEfSe input file.
	pathlefse_legend.png	Legend used for the LEfSe biomarkers HTML plot. Legend will be missing if this file is modified or deleted.
	LEfSe.Results.csv	Raw data of effect size (column D) and P-values (column E) from statistical analysis. Column A = taxon. Column C = group with the highest abundance. Column F = name of associated abundance distribution plot for taxon, found in the Figures folder.

Continued next page.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com

ZYMO RESEARCH

9. Output File Structure: Functional Profiling

RawData	combined_pathway_abun_filt.tsv	Concatenated pathway abundance data and pathway abundance data with species identification for each pathway identified in each sample.
	gene_fam_cpm.tsv	Gene family abundance in counts per million* for gene families identified in each sample.
	gene_fam_cpm_filt.tsv	Gene family abundance in counts per million* for most abundant gene families identified in each sample.
	pathway_abun_cpm.tsv	Pathway abundance in counts per million* for pathways identified in each sample.
	pathway_abun_filt.tsv	Pathway abundance in counts per million* for most abundant pathways identified in each sample.
	pathway_cov.tsv	Pathway coverage** data for pathways identified in each sample.
	pathway_cov_filt.tsv	Pathway coverage** data for most abundant pathways identified in each sample.
	species_gene_fam_cpm_filt.tsv	Gene family abundance data with species identification for the most abundant gene families identified in each sample.
	species_pathway_abun_filt.tsv	Pathway abundance with species information in counts per million* for most abundant pathways identified in each sample.
	species_pathway_cov_filt.tsv	Pathway coverage** data for most abundant pathways identified in each sample with species information.
SampleInformation	SampleMetadata.csv	File with sample ID, sample name, and group assignment information used in the group comparison.

*Counts per million (CPM) are counts for the gene family or pathway of interest divided by the total read count and multiplied by one million. $CPM = \frac{Read\ Counts\ of\ Interest}{Total\ Read\ Counts} \times 10^6$

**Coverage represents the confidence score for the pathway with a range of 0 to 1 where 0 is no confidence and 1 is 100% confidence.

ZYMO RESEARCH CORP.
Phone: (949) 679-1190 ▪ Toll Free: (888) 882-9682 ▪ Fax: (949) 266-9452
services@zymoresearch.com ▪ info@zymoresearch.com ▪ www.zymoresearch.com