Application Note

Dynamics of microbial communities during coffee fermentation under different processing conditions

This application note shows how QIAGEN CLC Microbial Genomics Module, a part of QIAGEN CLC Genomics Workbench Premium, can help you analyze bacterial and fungal communities during the fermentation processes.

Introduction

Fermentation processes are common in various food preparation and preservation methods. Currently, fermentation research employs metagenomic information to assess the diversity and dynamics of microbial populations through the fermentation processes. The diversity and dynamics of microbial populations is often associated with the quality of the final products. Here, we demonstrate some of the tools in the QIAGEN CLC Microbial Genomics Module that assist in dissecting and visualizing amplicon-based metagenomics sequencing datasets. We analyze the dynamics of bacterial and fungal communities during the fermentation of coffee beans. The datasets used in this application note were published by Zhang et al. (1) and are available in GenBank (2, 3). The bacterial dataset contains 48 files with amplicon sequencing of the V4 region of the 16S rRNA gene. The fungal profile is based on amplicon sequencing of the ITS1 region of the 26S rRNA gene in the same 48 processing samples.

Summary

QIAGEN CLC Genomics Workbench Premium and QIAGEN CLC Microbial Genomics Module were used to analyze microbial communities during coffee fermentation procedures. Publicly available sequencing datasets were downloaded directly from GenBank to the software. Amplicon-based analysis workflows were used for the initial processing steps, such as sequencing quality control (QC), operational taxonomic unit (OTU) clustering and estimation of alpha and beta diversities. To understand the effects of different fermentation methods and processing durations on the microbial communities, the data were compared, visualized and analyzed using various tools and visualization options.

Data Analysis

Importing data from GenBank Short Read Archive (SRA) and modification of Metadata tables

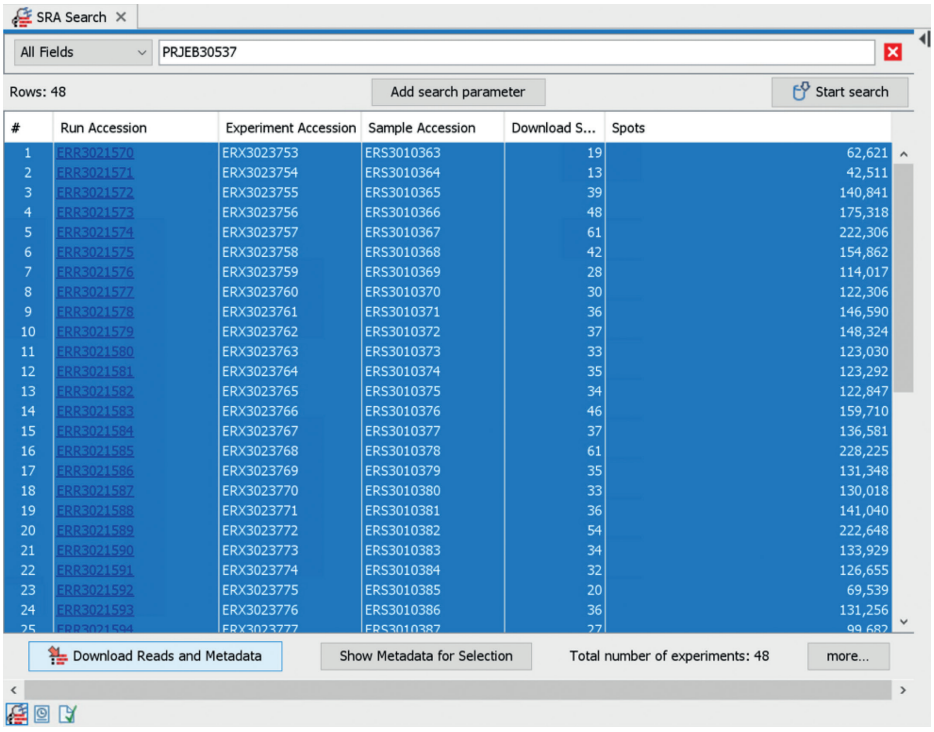
The bacterial metagenome data was imported directly to QIAGEN CLC Genomics Workbench from GenBank:

Download | Search for Reads in SRA | Search

criteria: BioProject ID PRJEB30537 | Select All| Download Reads and Metadata

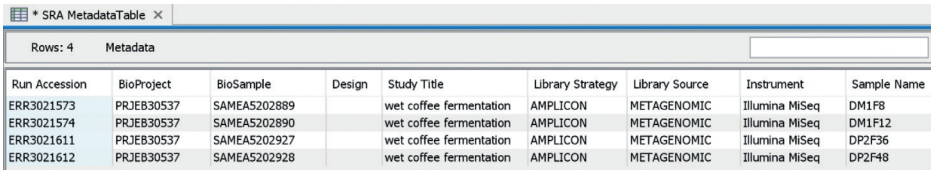
The 48 imported Illumina datasets contain reads from the V4 region of the 16S rRNA gene (Figure 1).

For these samples in GenBank, the experiment information is present only in the Sample Name column (Figure 2). In the paper (1), two fermentation conditions were tested: Depulping and demucilaging. They are abbreviated as DP and DM in each sample name. From each condition, the samples were collected at multiple time points. The fermentation time (in hours) is the number after "F" in the sample name. To set up various comparison experiments, we transferred the fermentation type to the "Design" column and created a new column "Processing time" with the fermentation time (Figure 3). QIAGEN CLC Genomics Workbench made it easy to modify the metadata tables. In the downstream analysis steps, these columns were used for the comparisons of fermentation conditions and duration.



#	Run Accession	Experiment Accession	Sample Accession	Download S...	Spots
1	ERR3021570	ERX3023753	ERS3010363	19	62,621
2	ERR3021571	ERX3023754	ERS3010364	13	42,511
3	ERR3021572	ERX3023755	ERS3010365	39	140,841
4	ERR3021573	ERX3023756	ERS3010366	48	175,318
5	ERR3021574	ERX3023757	ERS3010367	61	222,306
6	ERR3021575	ERX3023758	ERS3010368	42	154,862
7	ERR3021576	ERX3023759	ERS3010369	28	114,017
8	ERR3021577	ERX3023760	ERS3010370	30	122,306
9	ERR3021578	ERX3023761	ERS3010371	36	146,590
10	ERR3021579	ERX3023762	ERS3010372	37	148,324
11	ERR3021580	ERX3023763	ERS3010373	33	123,030
12	ERR3021581	ERX3023764	ERS3010374	35	123,292
13	ERR3021582	ERX3023765	ERS3010375	34	122,847
14	ERR3021583	ERX3023766	ERS3010376	46	159,710
15	ERR3021584	ERX3023767	ERS3010377	37	136,581
16	ERR3021585	ERX3023768	ERS3010378	61	228,225
17	ERR3021586	ERX3023769	ERS3010379	35	131,348
18	ERR3021587	ERX3023770	ERS3010380	33	130,018
19	ERR3021588	ERX3023771	ERS3010381	36	141,040
20	ERR3021589	ERX3023772	ERS3010382	54	222,648
21	ERR3021590	ERX3023773	ERS3010383	34	133,929
22	ERR3021591	ERX3023774	ERS3010384	32	126,655
23	ERR3021592	ERX3023775	ERS3010385	20	69,539
24	ERR3021593	ERX3023776	ERS3010386	36	131,256
25	ERR3021594	ERX3023777	ERS3010387	27	99,682

Figure 1
Download of sequencing datasets from GenBank SRA.



Run Accession	BioProject	BioSample	Design	Study Title	Library Strategy	Library Source	Instrument	Sample Name
ERR3021573	PRJEB30537	SAMEA5202889		wet coffee fermentation	AMPLICON	METAGENOMIC	Illumina MiSeq	DM1F8
ERR3021574	PRJEB30537	SAMEA5202890		wet coffee fermentation	AMPLICON	METAGENOMIC	Illumina MiSeq	DM1F12
ERR3021611	PRJEB30537	SAMEA5202927		wet coffee fermentation	AMPLICON	METAGENOMIC	Illumina MiSeq	DP2F36
ERR3021612	PRJEB30537	SAMEA5202928		wet coffee fermentation	AMPLICON	METAGENOMIC	Illumina MiSeq	DP2F48

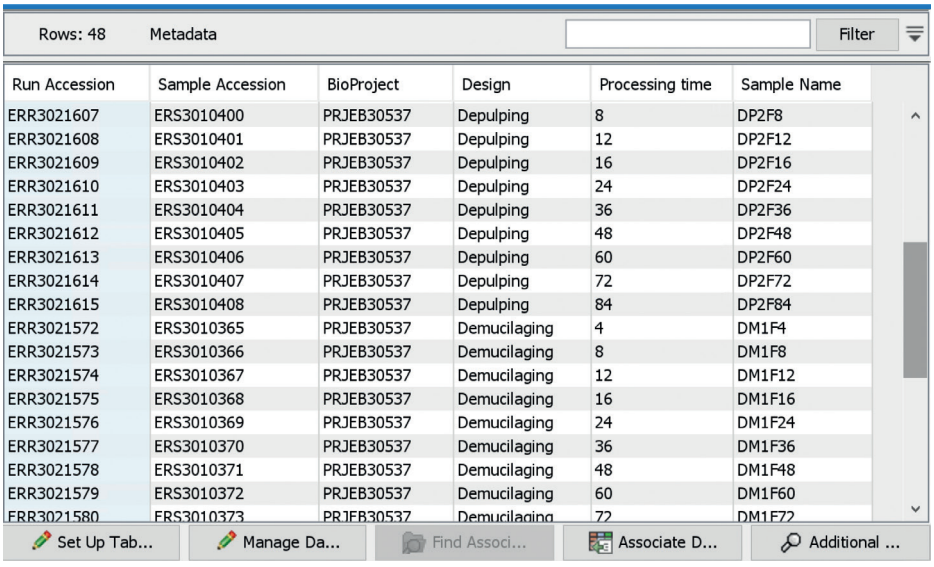
Figure 2
GenBank metadata information for four selected samples in the bacterial dataset.

Importing the microbial reference databases

Both the bacterial (Greengenes 97% v.13_8) and fungal (UNITE 97% v7.2) TaxPro reference databases were imported into the workbench using the Download Amplicon-Based Reference Database tool (Figure 4).

Sequencing data QC and OTU Clustering

All 96 sequencing data files (48 bacterial and 48 fungal) were submitted for QC and OTU clustering using the corresponding workflow in the Microbial Workflows folder (Figure 5). In this step, the low-quality reads were removed or trimmed, and the reads containing Illumina adapter sequences were discarded. Each set of reads was then mapped to the bacterial (Greengenes 97% v.13_8) or fungal (UNITE 97% v7.2) databases (reference based OTU clustering).



Rows: 48 Metadata Filter

Run Accession	Sample Accession	BioProject	Design	Processing time	Sample Name
ERR3021607	ERS3010400	PRJEB30537	Depulping	8	DP2F8
ERR3021608	ERS3010401	PRJEB30537	Depulping	12	DP2F12
ERR3021609	ERS3010402	PRJEB30537	Depulping	16	DP2F16
ERR3021610	ERS3010403	PRJEB30537	Depulping	24	DP2F24
ERR3021611	ERS3010404	PRJEB30537	Depulping	36	DP2F36
ERR3021612	ERS3010405	PRJEB30537	Depulping	48	DP2F48
ERR3021613	ERS3010406	PRJEB30537	Depulping	60	DP2F60
ERR3021614	ERS3010407	PRJEB30537	Depulping	72	DP2F72
ERR3021615	ERS3010408	PRJEB30537	Depulping	84	DP2F84
ERR3021572	ERS3010365	PRJEB30537	Demucilaging	4	DM1F4
ERR3021573	ERS3010366	PRJEB30537	Demucilaging	8	DM1F8
ERR3021574	ERS3010367	PRJEB30537	Demucilaging	12	DM1F12
ERR3021575	ERS3010368	PRJEB30537	Demucilaging	16	DM1F16
ERR3021576	ERS3010369	PRJEB30537	Demucilaging	24	DM1F24
ERR3021577	ERS3010370	PRJEB30537	Demucilaging	36	DM1F36
ERR3021578	ERS3010371	PRJEB30537	Demucilaging	48	DM1F48
ERR3021579	ERS3010372	PRJEB30537	Demucilaging	60	DM1F60
ERR3021580	ERS3010373	PRJEB30537	Demucilaging	72	DM1F72

Set Up Tab... Manage Da... Find Associ... Associate D... Additional ...

Figure 3
Modified Metadata table with the "Design" and "Processing time" columns.

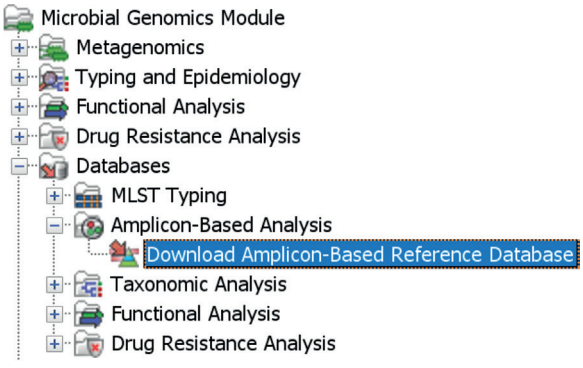


Figure 4
The Download Amplicon-Based Reference Database tool is in the Databases folder.

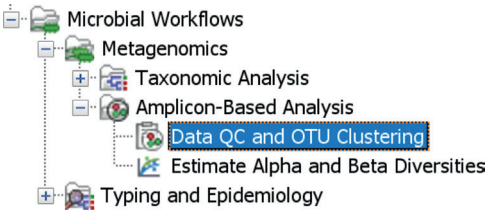


Figure 5
The Data QC and OTU Clustering workflow.

This workflow created trimming and clustering reports as well as OTU tables for each sample. The OTU tables contain abundance counts for each detected OTU. It is possible to visualize OTU counts individually for each sample, but visualization after merging multiple samples provides more insights. In the next step, to compare the microbial composition across multiple samples, we merge individual OTU tables into one combined table.

Bacterial community and dynamics during fermentation

The bacterial OTU tables were merged using the Merge Abundance Tables tool (Figure 6). *Leuconostoc* and *Lactococcus* were detected as the most abundant genera present by the end of the fermentation processes. Figure 7 visualizes the prokaryotic composition of 36 fermentation samples. The other 12 samples are various controls and extended depulping fermentation samples and are not shown here. The two left blocks of nine samples are two biological replicates of the demucilaging fermentation samples, and the two right blocks of nine samples are two biological replicates of depulping fermentation samples. The samples were sorted by fermentation time, starting with 4 hours and ending with 72 hours. The features were aggregated and color-coded by genus abundance. From this display of data, we observed that demucilaging fermentation results in a more complex microbial population in the early stages of the process compared to depulping fermentation. However, the depulping process results in more bacterial diversity in the later stages of fermentation compared to demucilaging.

The visualization options allow the OTU counts to be displayed across multiple samples using different metadata information and various feature aggregation units.

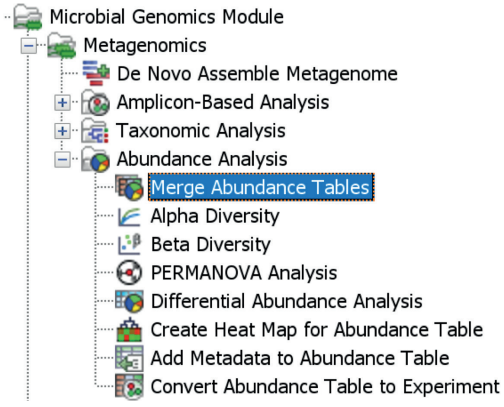


Figure 6
The Merge Abundance Tables tool in the Abundance Analysis folder.

Figure 7
Visualization of prokaryotic diversity at various time points of demucilaging and depulping fermentation of coffee for demucilaging and depulping (two replicates each). Sample names are give on the X-axis.

Fungal diversity and dynamics during fermentation

The fungal dynamics were found to be somewhat less reproducible in the biological replicates compared to the bacterial dynamics (Figure 8). *Pichia* and *Papiliotrema* were the most represented genera on average; however, they did not seem to be associated with the fermentation duration. *Kazachstania* (blue bars in Figure 8 appeared to be more indicative of the fermentation duration as it was better represented in the later stages of fermentation in both types of fermentation processes. *Pyrenochaeta* was more represented in the demucilaging samples.

Estimation of alpha and beta diversities

The alpha and beta diversities (Figure 5) were generated using the Estimate Alpha and Beta Diversities tool in the Microbial Workflows folder. Figure 9 shows the alpha diversities by the phylogeny in the bacterial samples. This plot demonstrates the same finding observed earlier in Figure 7: depulping samples have more microbial diversity at the late fermentation stages than late-stage demucilaging samples. Here, we show only the processing data points between 12 and 72 hours as they were the most indicative.

Figure 8
Bar chart of abundance of fungal genera in the fermentation samples for demucilaging and depulping (two replicates each). Processing time in hours is given on the X-axis.

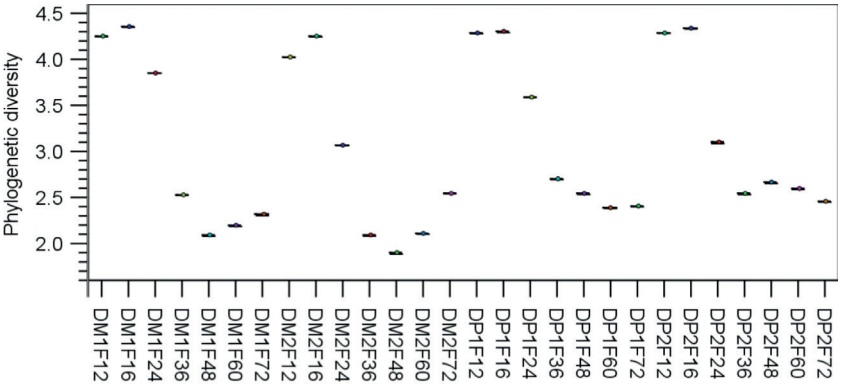


Figure 9
Bacterial alpha diversity in fermentation samples. DM: demucilaging process; DP: depulping process. The numbers following "F" in the sample name indicate the processing time in hours.

The Bray-Curtis beta diversity plot demonstrates that the changes in the microbial dynamics converge in the later stages of both fermentation processes, depulping and demucilaging (Figure 10).

Differential abundance analysis

Differential abundance analysis was performed using the corresponding tool in the Abundance Analysis folder (Figure 11). For this analysis, all bacterial datasets were used and the analysis was set to compare each fermentation time point against control samples, i.e., samples taken at 0 hours. The data in the output table can then be visualized as Venn diagrams. Any of the Venn diagram sections can be selected and extracted to a new table (Figure 12). We selected the intersection of differentially abundant genera in three comparisons, 36 hours vs 0 hours, 48 hours vs 0 hours, and 60 hours vs 0 hours. There were 33 differentially abundant genera in this intersection. Only five of genera became relatively more abundant in the fermentation process, with *Leuconostoc* and *Lactococcus* the most abundant at these late stages of fermentation.

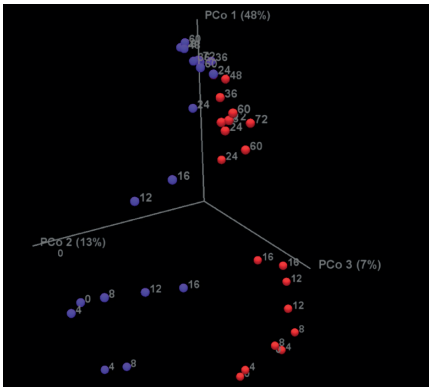


Figure 10
Beta diversity plot for the bacterial datasets. Red dots are depulped samples, and purple dots are demucilaged samples. The numbers indicate the processing time in hours.

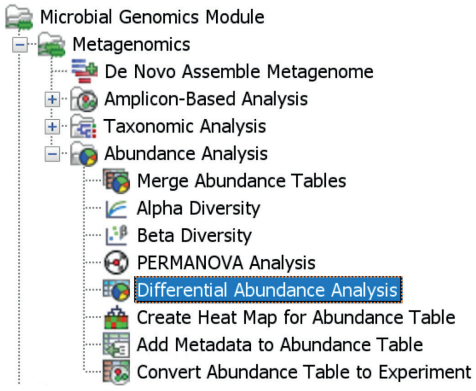
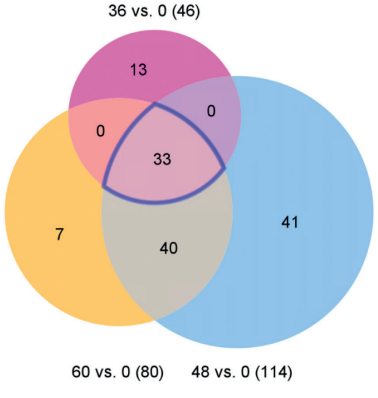


Figure 11
The Abundance Analysis folder of the QIAGEN CLC Microbial Genomics Module contains various visualization and statistical tools. Differential Abundance Analysis tool is selected.



Rows: 33 Venn diagram table — Filter to Selection... — Filter

Name	36 vs. 0			48 vs. 0		
	Max group mean	Fold change	FDR p-value	Max group mean	Fold change	FDR p-value
Salmonella; 4384058	303.75	113.22	9.77E-4	339.50	43.78	9.85E-3
Leuconostoc; 565537	2,270.75	104.27	6.58E-4	2,354.75	36.42	5.43E-3
Oryzae; 4423978	133.00	100.15	0.03	247.00	82.51	0.02
Leuconostoc; 4482944	10,107.25	50.88	1.52E-3	11,226.75	18.61	0.03
Lactococcus; 716006	29,530.00	47.97	9.77E-4	39,374.00	19.07	0.02
Enterobacteriaceae; ...	293.00	40.03	4.70E-3	391.25	18.28	0.03
Veronii; 276867	378.25	-16.17	0.05	378.25	-50.75	1.93E-4
R4-41B; 3701020	197.75	-25.71	0.01	197.75	-311.20	2.61E-7
Pseudomonas; 961783	255.50	-25.72	0.05	255.50	-3,105.61	1.43E-4
Agrobacterium; 561885	97.75	-26.61	0.04	97.75	-38.43	2.16E-3
Mesophilicum; 542475	243.00	-28.83	0.03	243.00	-65.73	4.34E-4
Diaphorobacter; 946104	235.00	-31.21	0.02	235.00	-2,861.15	9.41E-5
Rhizobiales; 750263	95.00	-31.78	0.03	95.00	-1,085.47	9.25E-4
Pseudomonas; 541859	120.25	-33.47	0.02	120.25	-246.29	3.31E-5
ACK-M1; 545460	49.25	-38.62	0.04	49.25	-609.44	2.89E-3
Streptomyces; 39463	644.50	-44.56	0.03	644.50	-727.78	8.14E-6
Prosthecobacter; 516...	76.25	-45.80	0.04	76.25	-1,198.30	1.37E-3
Methylosinus; 361082	127.50	-59.91	0.02	127.50	-1,239.89	9.22E-4
Alcaligenaceae; 661229	2,049.75	-68.72	9.77E-4	2,049.75	-286.07	7.96E-8
Comamonadaceae; 3...	393.75	-76.08	0.02	393.75	-497.27	2.24E-5
Comamonadaceae; 5...	122.75	-88.94	2.21E-3	122.75	-1,513.24	2.59E-4
Legionellaceae; 1108...	121.25	-105.37	4.05E-3	121.25	-1,873.91	2.25E-4
Comamonadaceae; 7...	289.25	-190.06	7.94E-3	289.25	-4,416.64	2.25E-4
Comamonadaceae; 3...	83.75	-300.02	0.03	83.75	-1,042.40	7.78E-4
Polynucleobacter; 56...	2,622.25	-314.59	9.77E-4	2,622.25	-1,165.57	4.10E-6
Isosphaeraceae; 685...	52.50	-315.65	0.04	52.50	-1,222.39	1.02E-3
Polynucleobacter; 33...	310.25	-378.45	9.77E-4	310.25	-8,197.20	4.17E-5
Microbacteriaceae; 7...	1,147.75	-441.78	4.24E-3	1,147.75	-2,466.24	2.42E-5
Gyiorum; 839270	71.25	-473.68	0.03	71.25	-1,770.97	5.86E-4
Trebouxiophyceae; 4...	143.75	-475.67	0.03	143.75	-1,850.72	6.30E-4
Sediminibacterium; 5...	108.50	-523.51	0.03	108.50	-1,661.49	6.30E-4
Comamonadaceae; 1...	158.50	-622.79	0.01	158.50	-1,934.10	2.25E-4
Alcaligenaceae; 25051	97.50	-807.41	0.02	97.50	-592.41	9.15E-5

Figure 12
The output from the Differential Abundance Analysis. Left: the Venn diagram. Right: table showing the genera extracted from the intersection of the Venn diagram.

Concluding Remarks

The data extractions, filtering parameters, and statistical cutoffs can be changed after processing. The same data presented in Figure 7 are shown in Figure 13 grouped by the fermentation duration and aggregated by species.

Another example showing the flexibility of QIAGEN CLC Microbial Genomics Module in postprocessing data visualization is presented in Figure 14. These additional Venn diagrams were created from the same Differential Abundance Analysis file used to produce Figure 12.

The Venn diagram on the left of Figure 14 has the intersection selected of the same three comparison data points, 36 hours vs 0 hours, 48 hours vs 0 hours, and 60 hours vs 0 hours. However, the cutoff parameters are now set to at least 100x-fold change and FDR to <0.01. Under these criteria, we find just five genera that are differentially abundant in all three sets. On the right side of Figure 14, the comparison of 60 hour vs 0 hour was replaced with 24 hour vs 0 hour. With the default parameters of 1.5-fold change and FDR of <0.05, six genera are selected as differentially abundant in these three comparisons. When looking for intersections of more than three comparisons, they can be filtered from the table as shown in Figure 15. Here we selected the genera that are differentially abundant in five comparisons using FDR values <0.05.

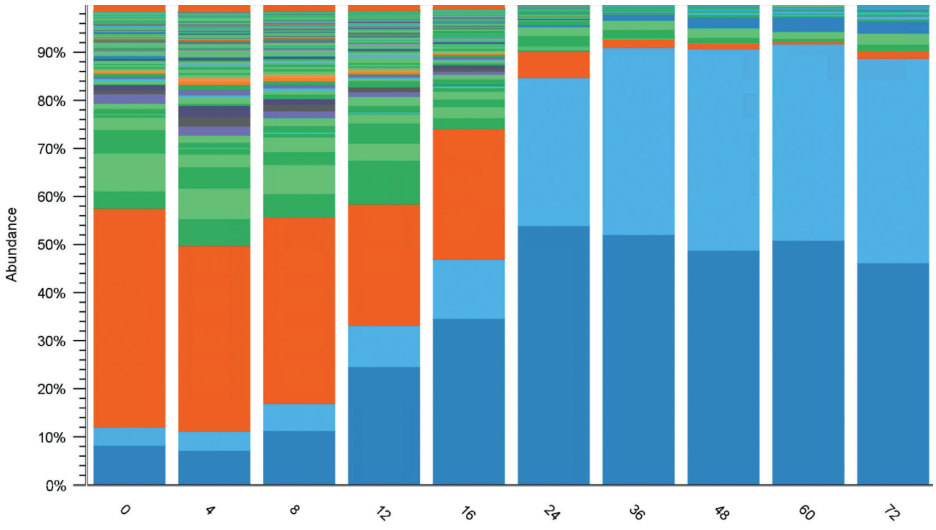


Figure 13
Microbial dynamics in coffee fermentation. Forty datasets are grouped by fermentation duration into ten groups. The numbers indicate the fermentation time in hours.

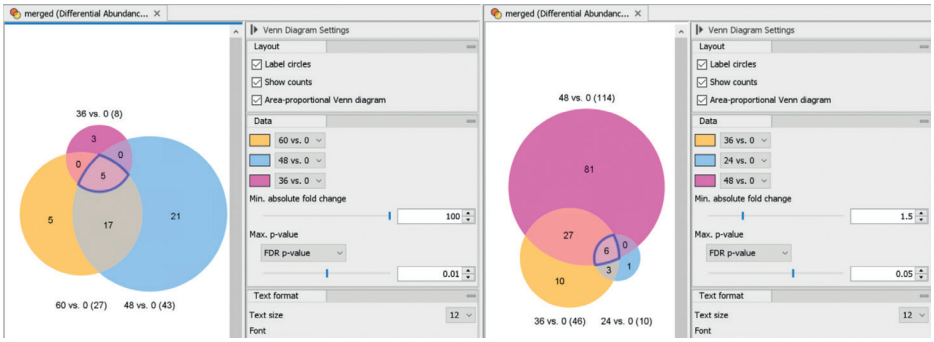
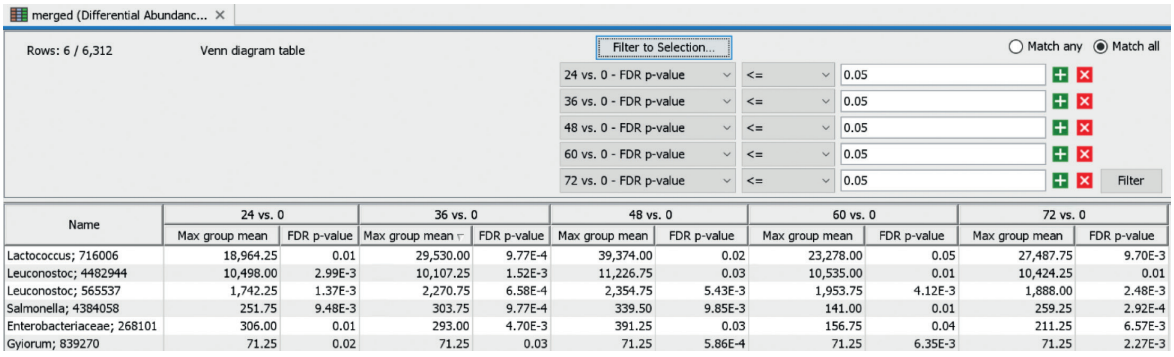


Figure 14
Visualization and data extraction from Venn diagrams.

The tools available in the QIAGEN CLC Genomics Workbench Premium, which includes the QIAGEN CLC Microbial Genomics Module, allow all-in-one analysis of publicly available fermentation datasets. Multiple comparisons using different statistical and visualization options can be applied to the same datasets to gain new insights.



Name	24 vs. 0		36 vs. 0		48 vs. 0		60 vs. 0		72 vs. 0	
	Max group mean	FDR p-value	Max group mean	FDR p-value	Max group mean	FDR p-value	Max group mean	FDR p-value	Max group mean	FDR p-value
Lactococcus; 716006	18,964.25	0.01	29,530.00	9.77E-4	39,374.00	0.02	23,278.00	0.05	27,487.75	9.70E-3
Leuconostoc; 4482944	10,498.00	2.99E-3	10,107.25	1.52E-3	11,226.75	0.03	10,535.00	0.01	10,424.25	0.01
Leuconostoc; 565537	1,742.25	1.37E-3	2,270.75	6.58E-4	2,354.75	5.43E-3	1,953.75	4.12E-3	1,888.00	2.48E-3
Salmonella; 4384058	251.75	9.48E-3	303.75	9.77E-4	339.50	9.85E-3	141.00	0.01	259.25	2.92E-4
Enterobacteriaceae; 268101	306.00	0.01	293.00	4.70E-3	391.25	0.03	156.75	0.04	211.25	6.57E-3
Gyiorum; 839270	71.25	0.02	71.25	0.03	71.25	5.86E-4	71.25	6.35E-3	71.25	2.27E-3

Figure 15
Data extraction from the Differential Analysis table using FDR cutoffs.

References

1. Zhang SJ, et al. (2019) Influence of various processing parameters on the microbial community dynamics, metabolomic profiles and cup quality during wet coffee processing. Frontiers in Microbiology 10:2621.
2. National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/bioproject/590915 (accessed August 23, 2023)
3. National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/bioproject/590916 (accessed August 23, 2023)

Visit **digitalinsights.qiagen.com/MGM** to learn more about QIAGEN CLC Microbial Genomics Module and go to **digitalinsights.qiagen.com/GXWBP** to request your free trial.