Workflows for extracting and analyzing microbiomes from whole-genome data of plant and animal species

Learn how to extract and analyze microbial sequences associated with host organisms using the CLC Microbial Genomics Module of QIAGEN® CLC Genomics Workbench Premium

Most plant and animal sequencing data also contain reads from microbiomes associated with the sequenced organism. We describe how to extract and analyze the microbiome data from publicly available datasets of host species. The first dataset used in this application note contains whole-genome sequencing data for 30 orchid species from the Prague Botanical Garden (1). The second dataset includes multiple samples of honeybees collected from various hives in different geographic locations (2).

Prokaryotic metagenomes of 30 orchid species

Results
In our analysis, we extracted and analyzed the microbial portion of the orchid-leaf sequencing data submitted to the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI). Twenty nine orchid species in this dataset are from the Pleurothallidinae subtribe and one species, *Isochilus aurantiacus*, is from the Ponerinae subtribe (this species was used as an outgroup in the orchid taxonomic research reported in reference 1). After prokaryotic taxonomic analysis with the QIAGEN CLC Genomics Workbench Premium, we created a visualization plot of the bacterial content in these 30 files (Figure 1). The outgroup species *Isochilus aurantiacus* (the first column in Figure 1) shows a somewhat distinct bacterial metagenomic profile, with the largest representation of *Curtobacterium* (41%), *Sphingomonas* (12%) and *Frondihabitans* (10%) (Figure 2). Notably, *Mangrovicoccus* (the light green bars in Figure 1) is present in all species, but to different extents in different species.
The heat map in Figure 3 shows the relative abundance of each bacterial genus in each sample and demonstrates that the orchid samples are clustered into five main branches according to microbial composition. The microbiome counts may reflect the specificity of some bacterial species to certain orchids. The geographical or ecological origins of these orchid specimens might also influence microbiome compositions. Because all orchid specimens used in the study were greenhouse-grown, the varying bacterial compositions suggest that orchid specimens may preserve the associated microbiome after a change in cultural conditions.

**Figure 1.**
Visualization of bacterial content in whole-genome sequencing files of 30 orchid species.

**Figure 2.**
The most abundant bacterial genera in the sequencing data of *Isochilus aurantiacus*.

<table>
<thead>
<tr>
<th>Genus</th>
<th><em>Isochilus aurantiacus</em> Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curtobacterium</td>
<td>0.41</td>
</tr>
<tr>
<td>Mangrovococcus</td>
<td>0.25</td>
</tr>
<tr>
<td>Sphingomonas_N</td>
<td>0.12</td>
</tr>
<tr>
<td>Frondithabitans</td>
<td>0.10</td>
</tr>
<tr>
<td>Sphingomonas</td>
<td>9.20E-3</td>
</tr>
</tbody>
</table>

**Figure 3.**
Heat map with the 20 most represented bacterial species in the 30 orchid species analyzed; green indicates most abundant. Red lines separate the five major clusters based on microbiome composition.
Workflow description

Sequencing reads and metadata used in this study were downloaded directly from NCBI using a search for the project ID “PRJNA692119” in the “SRA Search” dialog (Figure 4).

Before bacterial taxonomic profiling, we would ideally remove all orchid reads. Unfortunately, full genome information was not available for any of the 30 orchid species. However, chloroplast genomes from the subtribe Pleurothallidinae (Figure 5) were available. The chloroplast genome files...
were combined into one file, which was then used to create the “Host Genome Index” with the corresponding tool under the “Databases” folder in the CLC Microbial Genomics Module (Figure 6).

To identify the microbial reads in the orchid samples, it was necessary to download a taxonomic database workbench. For bacterial taxonomic profiling, the QMI-PTDB TaxPro index was downloaded using the “Download Curated Microbial Reference Database” tool (Figure 6).

In the next step, we used the prebuilt “Data QC and Taxonomic Profiling” workflow to map and count bacterial reads in the data files (Figure 7). All 30 sequencing-read files were submitted at once by selecting the “Batch” option. In the “Taxonomic Profiling” step, we selected the previously downloaded bacterial reference index, along with the constructed Pleurothallidinae chloroplast index (Figure 8).

![Figure 6. Taxonomic analysis tools.](image)

![Figure 7. “Data QC and Taxonomic Profiling” workflow.](image)

![Figure 8. Setting “reference index” and “host genome index” parameters for taxonomic profiling.](image)
The workflow generated 30 taxonomic profile tables, one for each orchid sample. The tables contain the counts and the coverage of each detected bacterial species in the sample (Figure 9). As shown in Figure 9, the counts can be visualized as stacked charts using various aggregation options, such as genus (left chart) and taxonomic class (right chart).

For the comparative analyses and visualizations of all 30 orchid samples, we combined the taxonomic profile tables using the “Merge Abundance Tables” tool (Figure 10). The resulting merged table contains counts for all detected species in all samples. The merged data, shown in Figure 1, is analyzed using various tools under the “Abundance Analysis” folder (Figure 10). Here we clustered the samples using the “Create Heat Map for Abundance Table” tool, generating the heat map in Figure 3.

The heat maps can be constructed using various distance measures and cluster linkages. The map shown in Figure 3 was constructed by selecting the Euclidian distance and complete linkage options.

The choice of analytical and visualization options is limited by the availability of metadata associated with the dataset. For the orchid dataset, the only differentiating metadata information available is the species name. We explore the analysis and visualization options using additional metadata fields from the second dataset.
Viral metagenomes of honeybee drones

Results
In this project, various honeybee next-generation (NGS) samples were analyzed for the presence of virus reads. Twenty samples from five different origins (biomaterial providers) were imported into QIAGEN CLC Genomics Workbench Premium and run through the taxonomic analysis. The viral content in these 20 files is shown in Figure 10: samples are sorted by origin. Whereas all samples from Royal Jelly France contained detectable amounts of *Apis mellifera* filamentous virus (red bars), sample YC7 contained extremely high counts of this virus. All other samples contained similar counts for white spot syndrome virus (purple) and an unknown species (orange). The deformed wing virus counts (light blue) were elevated in the BR51 sample from the Ariège Conservatory. The samples from China contained fewer counts for uncultured virus (green). The unexpectedly high counts for camelpox virus (light green in Figure 10) called for additional analysis using a repeat masked viral reference database (Figure 11). Camelpox virus is mostly specific to camelids and is not expected to be present in bee samples. To eliminated the erroneous counts for camelpox, a repeat masked viral reference database was created using the tools provided in the CLC Microbial Genomics Module (Figure 12). Using the masked reference database also reduced the number of viral species detected from 136 (Figure 11) to 73 (Figure 12). However, the counts for the five most abundant viral species in the repeat masked data were similar to the results with the unmasked database (Figure 11 and Figure 12). The visualization plot produced with the repeat masked viral reference database is cleaner (Figure 13), but is still very similar to the unmasked plot in Figure 10: most samples have the same four abundant species of viruses, except YC7, for which the filamentous virus represents most of the viral load.

![Figure 11. Visualization of viral content in whole-genome sequencing files of 20 honeybee drones from various locations.](image-url)
**Workflow description**

**Import of selected SRA files**

The sequencing reads and metadata were downloaded directly from NCBI using project ID “PRJNA311274” in the “SRA Search” dialog (Figure 14). There are 872 runs in this project, and we loaded them all by repeatedly clicking the “more...” button. After selecting all runs, we opened the metadata table by clicking the “Show Metadata for Selection” button.
In the work described by Wragg et al, 2016 (2), sequencing involved honeybee drones from multiple locations in Europe and China, with the focus on the honeybee populations managed by the Royal Jelly company in France. Here we analyze only 20 samples from this large dataset for the presence of viral reads. Four samples from five different locations (biomaterial providers) were selected and imported into the CLC Genomics Workbench Premium.

The samples from each selected location were imported as individual batches along with the associated metadata tables. Figure 15 shows the metadata table with four files from Royal Jelly France.

![Figure 15](image1)

**Figure 15.** Searching for files at NCBI SRA and selecting all entries for metadata retrieval.

All the names of the selected files in the metadata table started with “SRR1517348”. A search for “SRR1517348*” in the “SRA Search” table returned 10 files (Figure 16), from which we downloaded the four desired files by clicking the “Download Reads and Metadata” button.

![Figure 16](image2)

**Figure 16.** Searching the metadata table for Royal Jelly samples and identifying run accession IDs for subsequent import using the “SRA Search” table.
In the same manner, we downloaded the remaining 16 samples and metadata (Figure 17) from four other biomaterial providers (China, Boulanger, “Breeder, France” and Ariège Conservatory).

Figure 17.
Search for run accession IDs in the SRA table.

Bee genome
It is preferable to filter out the host reads before mapping and counting the metagenomic reads. The bee genome is available and was downloaded by searching for the bee reference genome ID in the NCBI search tool, as shown in Figure 18. Sixteen chromosome files and the mitochondrion file were selected and imported into the CLC Genomics Workbench Premium as a single file. The genome was converted to the genome index using the “Create Taxonomic Profiling Index” tool under the “Databases” folder in the CLC Microbial Genomics Module (Figure 6).

Figure 18.
Whole-genome sequence files with metadata, as they appear after importing into QIAGEN CLC Genomic Workbench Premium.

Figure 19.
Downloading the bee reference genome using the NCBI search tool.
Viral reference database

The viral reference database was imported using the “Download Curated Microbial Reference Database” tool under the “Databases” folder in the CLC Microbial Genomics Module. The Clustered Reference Viral Database was downloaded as a taxonomic profiling index by checking the corresponding box in the dialog window.

To create the repeat masked reference, we downloaded the database again as a sequence list by checking “as Sequence List” in the dialog, as shown in Figure 19. We then masked the repeats in the sequences using the “Mask Low-Complexity Regions” tool under the “Databases” folder (Figure 20). The masked sequences were converted to taxonomic profiling index using the corresponding tool under the “Taxonomic Analysis” folder (Figure 20).

![Figure 20. Download of the viral reference database.](image)

![Figure 21. The tools for creating the masked reference database.](image)
**Taxonomic profiling**

To map and count the viral reads in the data files, we used the prebuilt “Data QC and Taxonomic Profiling” workflow (Figure 22). The sequencing reads were submitted in batches of four files with the corresponding metadata files. The viral reference indexes (standard or masked) along with the bee genome index were selected in the “Taxonomic Profiling” step of the workflow (Figure 21). This creates taxonomic profiling tables, which we combined using the “Merge Abundance Tables” tool (Figure 10). The merged tables contain the counts for all detected viral species in all samples, along with combined counts across all samples (Figure 11 and Figure 12). The counts can be visualized in multiple ways using various phylogenetic, metadata and aggregation criteria. Figure 22 shows how the results in Figure 10 and Figure 13 can be aggregated according to biomaterial provider, applying settings to show just the five most abundant viruses and displaying three taxonomic levels for each virus.

**Figure 22.**
Selecting the reference viral database index and the host bee genome index.

**Figure 23.**
Visualization of taxonomic data, aggregated by source (biomaterial provider).
Conclusions

The tools available in QIAGEN CLC Genomics Workbench Premium and CLC Microbial Genomics Module allow all-in-one analysis of NGS datasets. This application note demonstrates how additional insights can be extracted using publicly available sequencing data. The whole-genome files explored here were originally created to study the genomes of orchids and bees. However, samples from most host organisms also contain a microbial species that leave their signatures in NGS files. The workflows presented here can be used to identify microsymbionts and pathogens and demonstrates how this information can be used applications, such as identification of ecological footprints, sanitary analysis and forensics.

References
