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.

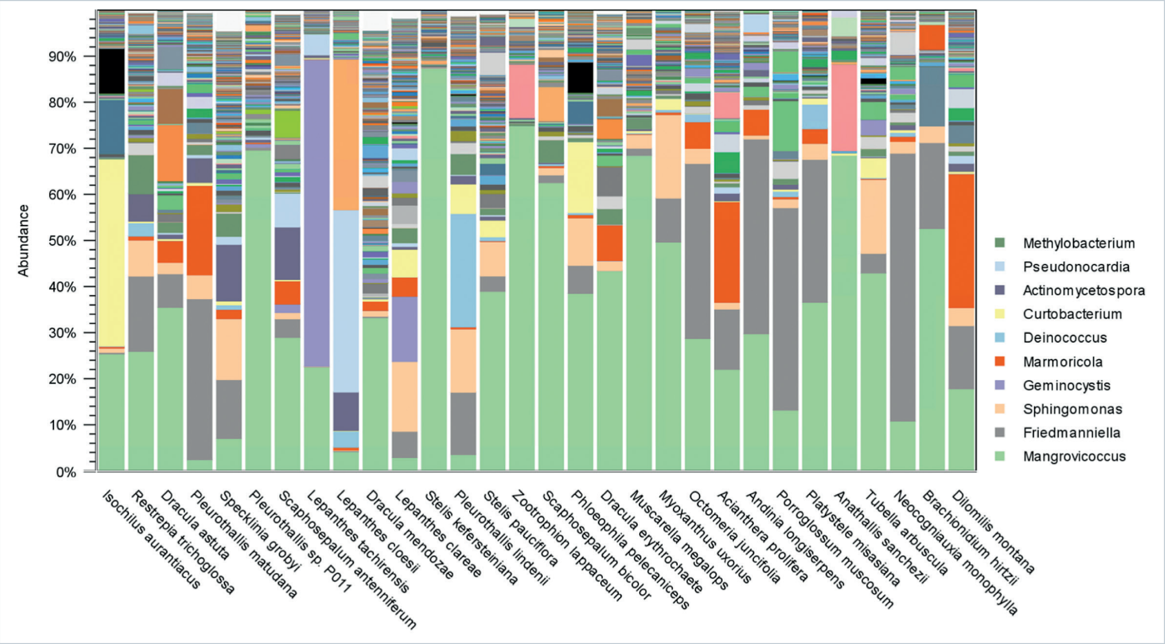
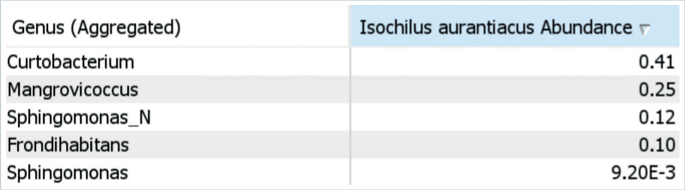


Figure 1.
Visualization of bacterial content in whole-genome sequencing files of 30 orchid species.



Genus (Aggregated)	Isochilus aurantiacus Abundance
Curtobacterium	0.41
Mangrovicoccus	0.25
Sphingomonas_N	0.12
Frondihabitans	0.10
Sphingomonas	9.20E-3

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

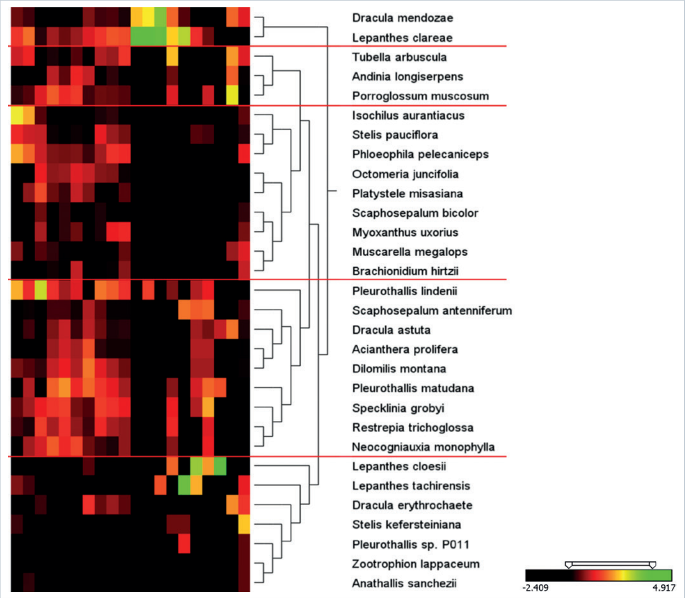
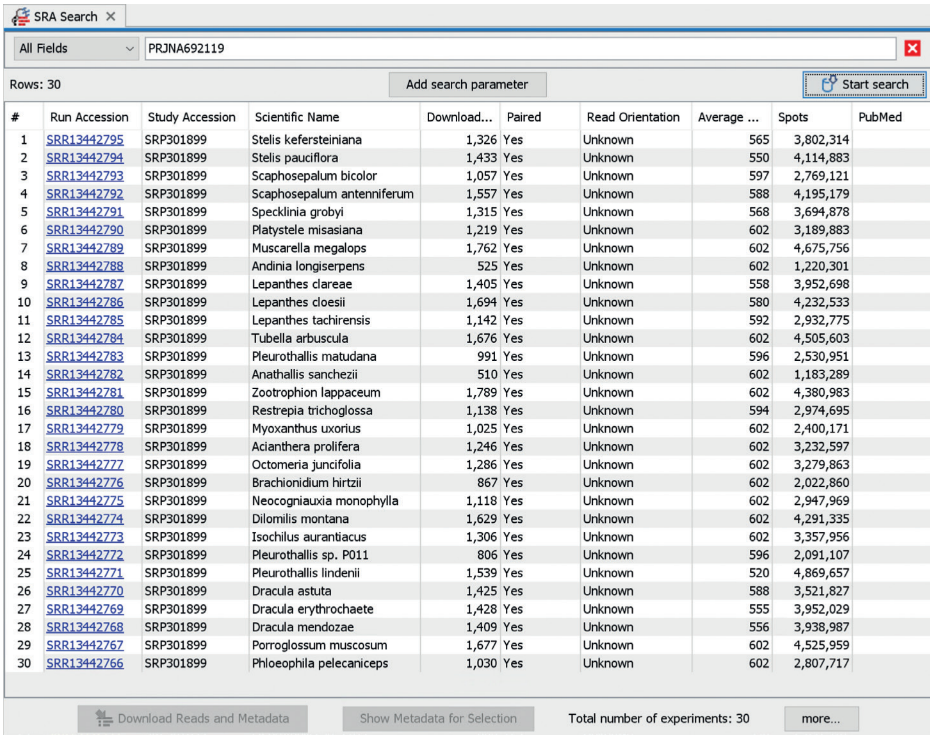


Figure 3.
Heat map with the 20 most represented bacterial species in the 30 orchid species analyzed; green indicates most abundant. Red lines separate the five major clusters based on microbiome composition.

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.

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).



PRJNA692119

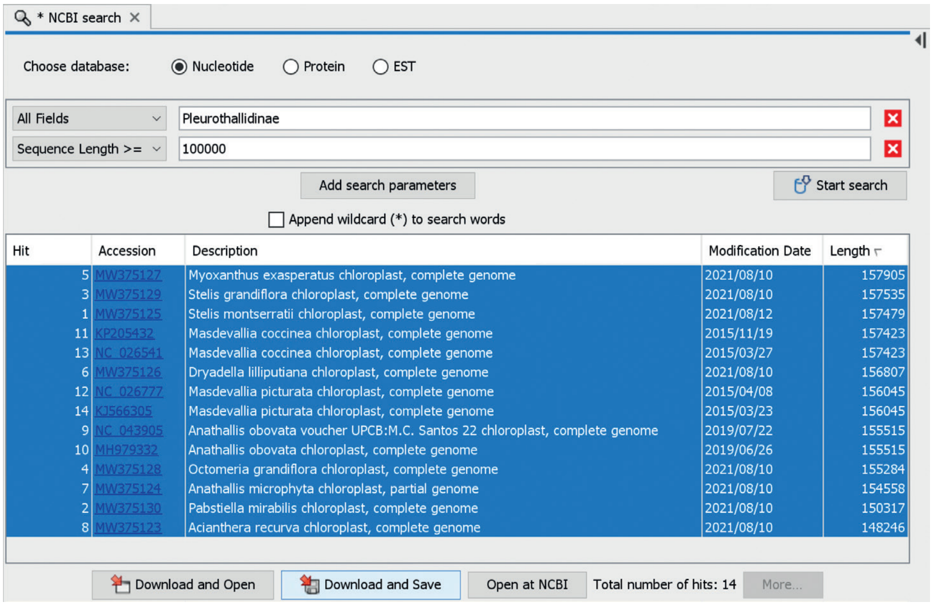
Rows: 30

#	Run Accession	Study Accession	Scientific Name	Download...	Paired	Read Orientation	Average ...	Spots	PubMed
1	SRR13442795	SRP301899	Stelis kefersteiniana	1,326	Yes	Unknown	565	3,802,314	
2	SRR13442794	SRP301899	Stelis pauciflora	1,433	Yes	Unknown	550	4,114,883	
3	SRR13442793	SRP301899	Scaphosepalum bicolor	1,057	Yes	Unknown	597	2,769,121	
4	SRR13442792	SRP301899	Scaphosepalum antenniferum	1,557	Yes	Unknown	588	4,195,179	
5	SRR13442791	SRP301899	Specklinia grobyi	1,315	Yes	Unknown	568	3,694,878	
6	SRR13442790	SRP301899	Platystele misasiana	1,219	Yes	Unknown	602	3,189,883	
7	SRR13442789	SRP301899	Muscarella megalops	1,762	Yes	Unknown	602	4,675,756	
8	SRR13442788	SRP301899	Andinia longiserpens	525	Yes	Unknown	602	1,220,301	
9	SRR13442787	SRP301899	Lepanthes clareae	1,405	Yes	Unknown	558	3,952,698	
10	SRR13442786	SRP301899	Lepanthes cloesii	1,694	Yes	Unknown	580	4,232,533	
11	SRR13442785	SRP301899	Lepanthes tachirensis	1,142	Yes	Unknown	592	2,932,775	
12	SRR13442784	SRP301899	Tubella arbuscula	1,676	Yes	Unknown	602	4,505,603	
13	SRR13442783	SRP301899	Pleurothallis matudana	991	Yes	Unknown	596	2,530,951	
14	SRR13442782	SRP301899	Anathallis sanchezii	510	Yes	Unknown	602	1,183,289	
15	SRR13442781	SRP301899	Zootrophion lappaceum	1,789	Yes	Unknown	602	4,380,983	
16	SRR13442780	SRP301899	Restrepia trichoglossa	1,138	Yes	Unknown	594	2,974,695	
17	SRR13442779	SRP301899	Myoxanthus uxorius	1,025	Yes	Unknown	602	2,400,171	
18	SRR13442778	SRP301899	Acianthera prolifera	1,246	Yes	Unknown	602	3,232,597	
19	SRR13442777	SRP301899	Octomeria juncifolia	1,286	Yes	Unknown	602	3,279,863	
20	SRR13442776	SRP301899	Brachionidium hirtzii	867	Yes	Unknown	602	2,022,860	
21	SRR13442775	SRP301899	Neocogniauxia monophylla	1,118	Yes	Unknown	602	2,947,969	
22	SRR13442774	SRP301899	Dilomilis montana	1,629	Yes	Unknown	602	4,291,335	
23	SRR13442773	SRP301899	Isochilus aurantiacus	1,306	Yes	Unknown	602	3,357,956	
24	SRR13442772	SRP301899	Pleurothallis sp. P011	806	Yes	Unknown	596	2,091,107	
25	SRR13442771	SRP301899	Pleurothallis lindenii	1,539	Yes	Unknown	520	4,869,657	
26	SRR13442770	SRP301899	Dracula astuta	1,425	Yes	Unknown	588	3,521,827	
27	SRR13442769	SRP301899	Dracula erythrochaete	1,428	Yes	Unknown	555	3,952,029	
28	SRR13442768	SRP301899	Dracula mendozae	1,409	Yes	Unknown	556	3,938,987	
29	SRR13442767	SRP301899	Porroglossum muscosum	1,677	Yes	Unknown	602	4,525,959	
30	SRR13442766	SRP301899	Phloeophila pelecaniceps	1,030	Yes	Unknown	602	2,807,717	

Total number of experiments: 30

Figure 4.
Search for files at NCBI SRA.

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



Choose database: ● Nucleotide ○ Protein ○ EST

All Fields: Pleurothallidinae

Sequence Length >=: 100000

□ Append wildcard (*) to search words

Hit	Accession	Description	Modification Date	Length
5	MW375127	Myoxanthus exasperatus chloroplast, complete genome	2021/08/10	157905
3	MW375129	Stelis grandiflora chloroplast, complete genome	2021/08/10	157535
1	MW375125	Stelis montserratii chloroplast, complete genome	2021/08/12	157479
11	KP205432	Masdevallia coccinea chloroplast, complete genome	2015/11/19	157423
13	NC_026541	Masdevallia coccinea chloroplast, complete genome	2015/03/27	157423
6	MW375126	Dryadella lilliputiana chloroplast, complete genome	2021/08/10	156807
12	NC_026777	Masdevallia picturata chloroplast, complete genome	2015/04/08	156045
14	KJ566305	Masdevallia picturata chloroplast, complete genome	2015/03/23	156045
9	NC_043905	Anathallis obovata voucher UPCB:M.C. Santos 22 chloroplast, complete genome	2019/07/22	155515
10	MH979332	Anathallis obovata chloroplast, complete genome	2019/06/26	155515
4	MW375128	Octomeria grandiflora chloroplast, complete genome	2021/08/10	155284
7	MW375124	Anathallis microphyta chloroplast, partial genome	2021/08/10	154558
2	MW375130	Pabstiella mirabilis chloroplast, complete genome	2021/08/10	150317
8	MW375123	Acianthera recurva chloroplast, complete genome	2021/08/10	148246

Total number of hits: 14

Figure 5.
Downloading chloroplast genomes from the subtribe Pleurothallidinae using the NCBI search tool.

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).

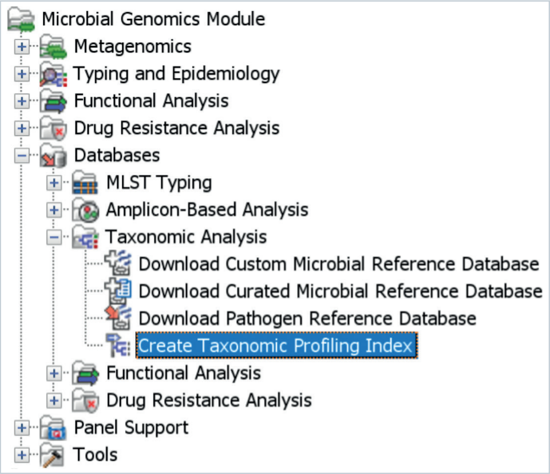


Figure 6.
Taxonomic analysis tools.

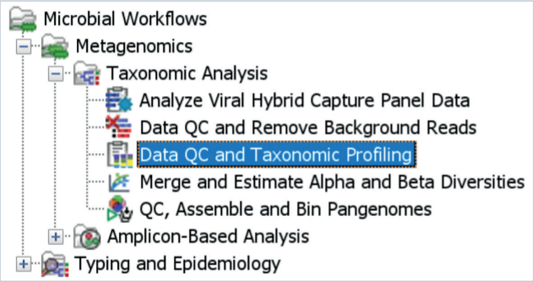
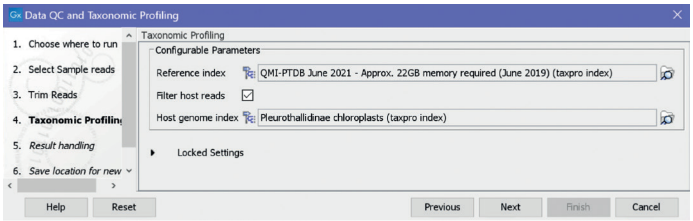

Microbial Workflows
Metagenomics
Taxonomic Analysis
Analyze Viral Hybrid Capture Panel Data
Data QC and Remove Background Reads
Data QC and Taxonomic Profiling
Merge and Estimate Alpha and Beta Diversities
QC, Assemble and Bin Pangenomes
Amplicon-Based Analysis
Typing and Epidemiology

Figure 7.
"Data QC and Taxonomic Profiling" workflow.



Taxonomic Profiling

Configurable Parameters

Reference index: QMI-PTDB June 2021 - Approx. 22GB memory required (June 2019) (taxpro index)

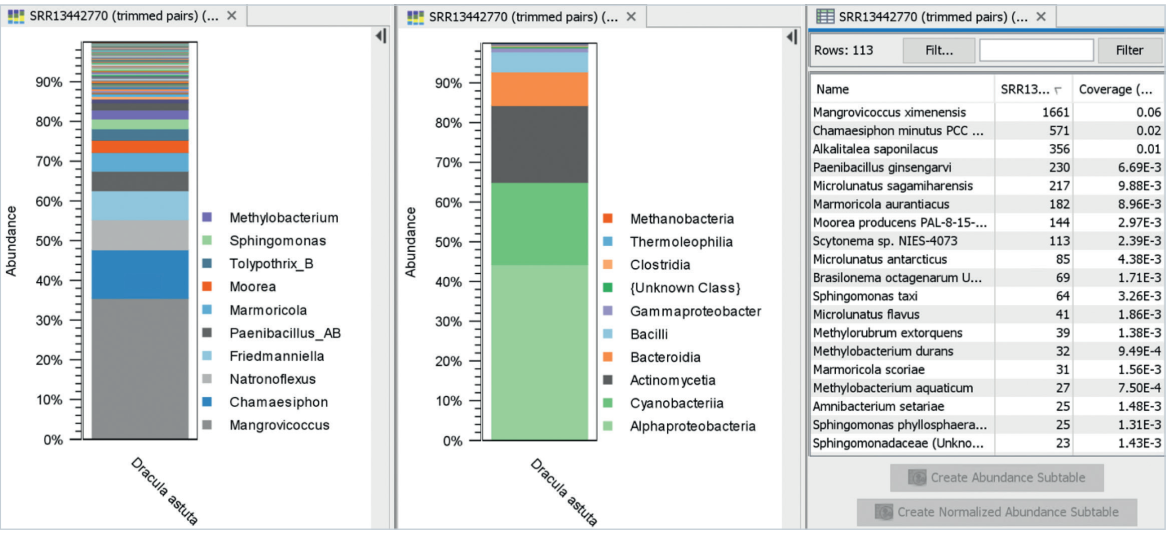
Filter host reads: ☑

Host genome index: Pleurothallidinae chloroplasts (taxpro index)

Locked Settings

Figure 8.
Setting "reference index" and "host genome index" parameters for taxonomic profiling.

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).



Name	SRR13...	Coverage (...
Mangrovicoccus ximenensis	1661	0.06
Chamaesiphon minutus PCC ...	571	0.02
Alkalitalea saponilacus	356	0.01
Paenibacillus ginsengarvi	230	6.69E-3
Microlunatus sagamiharensis	217	9.88E-3
Marmoricola aurantiacus	182	8.96E-3
Moorea producens PAL-8-15-...	144	2.97E-3
Scytonema sp. NIES-4073	113	2.39E-3
Microlunatus antarcticus	85	4.38E-3
Brasilonema octagenarum U...	69	1.71E-3
Sphingomonas taxi	64	3.26E-3
Microlunatus flavus	41	1.86E-3
Methylorubrum extorquens	39	1.38E-3
Methylobacterium durans	32	9.49E-4
Marmoricola scoriae	31	1.56E-3
Methylobacterium aquaticum	27	7.50E-4
Amnibacterium setariae	25	1.48E-3
Sphingomonas phyllosphaera...	25	1.31E-3
Sphingomonadaceae (Unkno...	23	1.43E-3

Figure 9.
Bacterial taxonomic profiling table (right) for the microbiome of *Dracula astuta*. The left chart displays the aggregation by genus, and the second chart the aggregation by taxonomic class.

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.

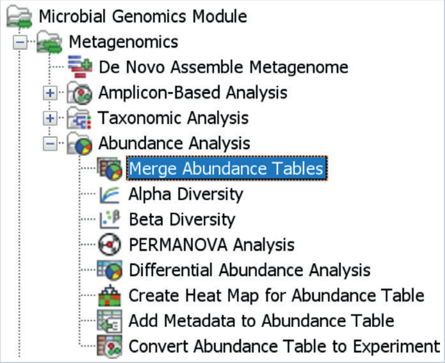

Microbial Genomics Module
Metagenomics
De Novo Assemble Metagenome
Amplicon-Based Analysis
Taxonomic Analysis
Abundance Analysis
Merge Abundance Tables
Alpha Diversity
Beta Diversity
PERMANOVA Analysis
Differential Abundance Analysis
Create Heat Map for Abundance Table
Add Metadata to Abundance Table
Convert Abundance Table to Experiment

Figure 10.
The abundance analysis tools, with the “Merge Abundance Tables” tool selected.

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.

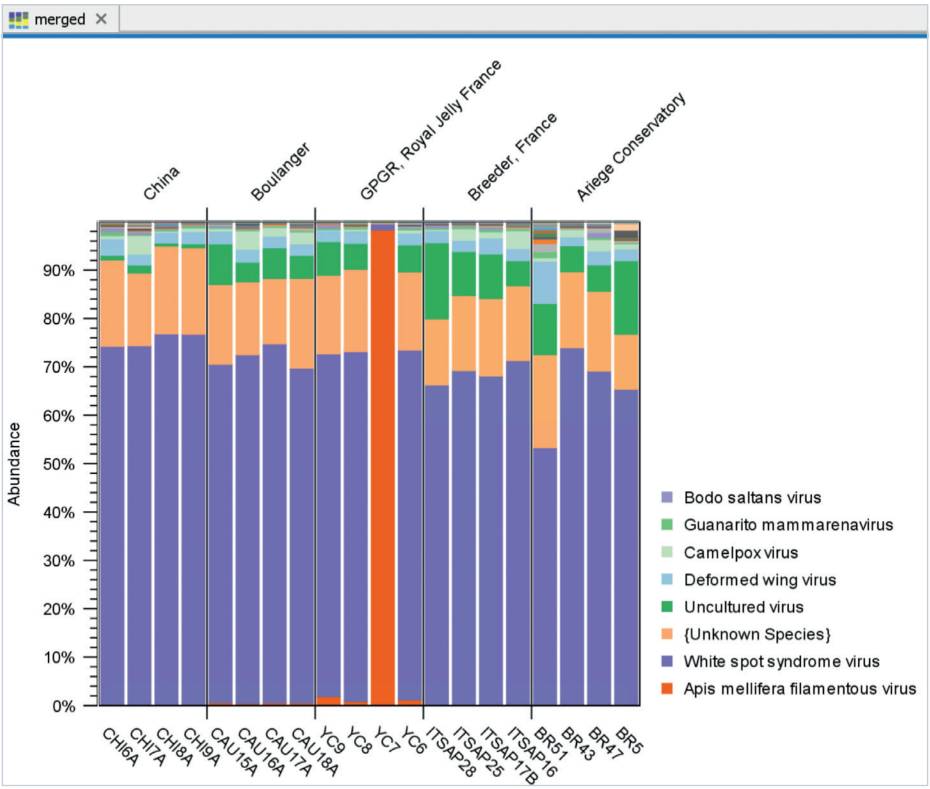
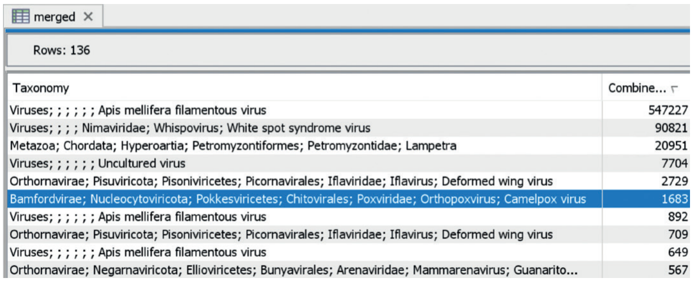


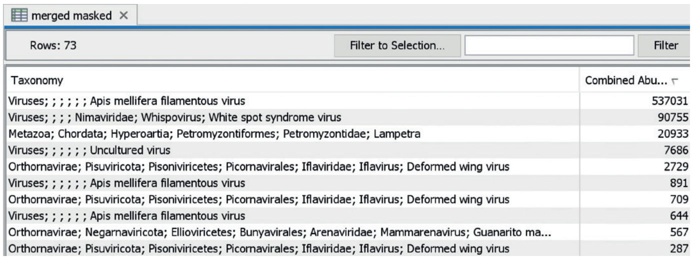
Figure 11.
Visualization of viral content in whole-genome sequencing files of 20 honeybee drones from various locations.



Rows: 136

Taxonomy	Combine...
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	547227
Viruses; ; ; ; Nimaviridae; Whispovirus; White spot syndrome virus	90821
Metazoa; Chordata; Hyperoartia; Petromyzontiformes; Petromyzontidae; Lampetra	20951
Viruses; ; ; ; ; ; Uncultured virus	7704
Orthornavirae; Pisuviricota; Pisoniviricetes; Picornavirales; Iflaviridae; Iflavirus; Deformed wing virus	2729
Bamfordvirae; Nucleocytoviricota; Pokkesviricetes; Chitovirales; Poxviridae; Orthopoxvirus; Camelpox virus	1683
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	892
Orthornavirae; Pisuviricota; Pisoniviricetes; Picornavirales; Iflaviridae; Iflavirus; Deformed wing virus	709
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	649
Orthornavirae; Negarnaviricota; Ellioviricetes; Bunyavirales; Arenaviridae; Mammarenavirus; Guanarito...	567

Figure 12.
Combined counts for the 10 most abundant viruses in 20 honeybee samples.



Rows: 73

Taxonomy	Combined Abu...
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	537031
Viruses; ; ; ; Nimaviridae; Whispovirus; White spot syndrome virus	90755
Metazoa; Chordata; Hyperoartia; Petromyzontiformes; Petromyzontidae; Lampetra	20933
Viruses; ; ; ; ; ; Uncultured virus	7686
Orthornavirae; Pisuviricota; Pisoniviricetes; Picornavirales; Iflaviridae; Iflavirus; Deformed wing virus	2729
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	891
Orthornavirae; Pisuviricota; Pisoniviricetes; Picornavirales; Iflaviridae; Iflavirus; Deformed wing virus	709
Viruses; ; ; ; ; ; Apis mellifera filamentous virus	644
Orthornavirae; Negarnaviricota; Ellioviricetes; Bunyavirales; Arenaviridae; Mammarenavirus; Guanarito ma...	567
Orthornavirae; Pisuviricota; Pisoniviricetes; Picornavirales; Iflaviridae; Iflavirus; Deformed wing virus	287

Figure 13.
Combined counts for the 10 most abundant viruses in 20 honeybee samples after matching against the repeat masked viral reference database.

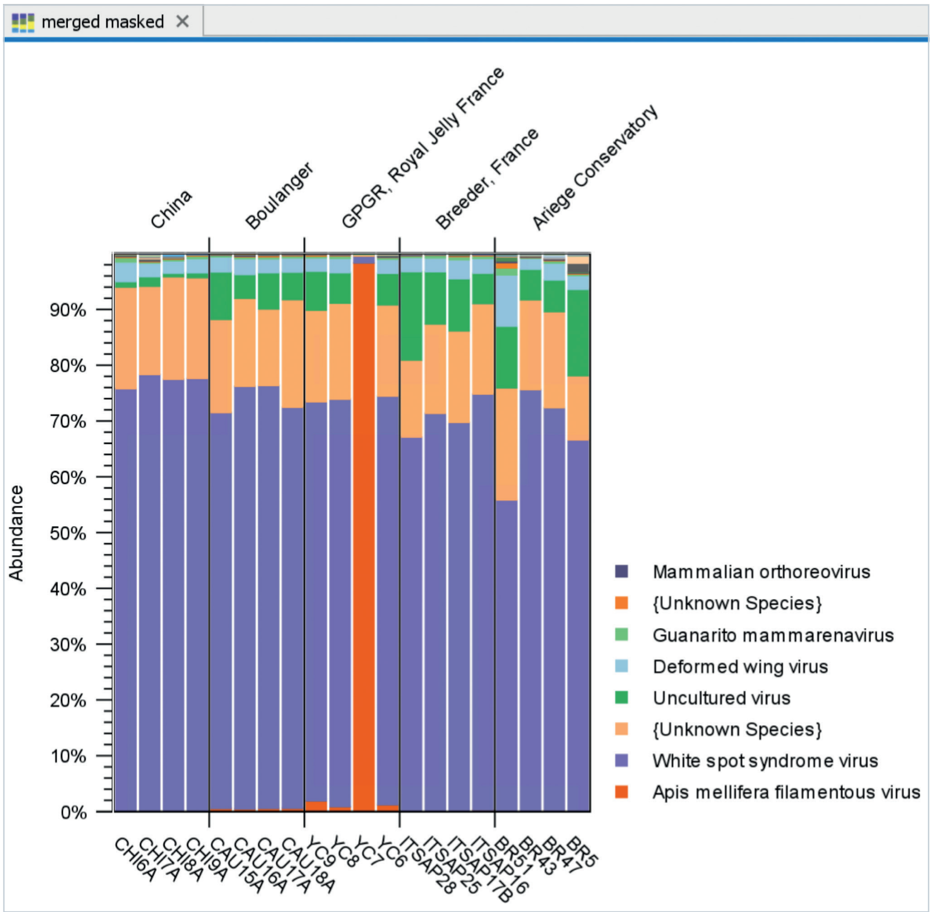


Figure 14.
Visualization of viral content in whole-genome sequencing files of 20 honeybee drones from various locations, using a repeat masked viral reference database.

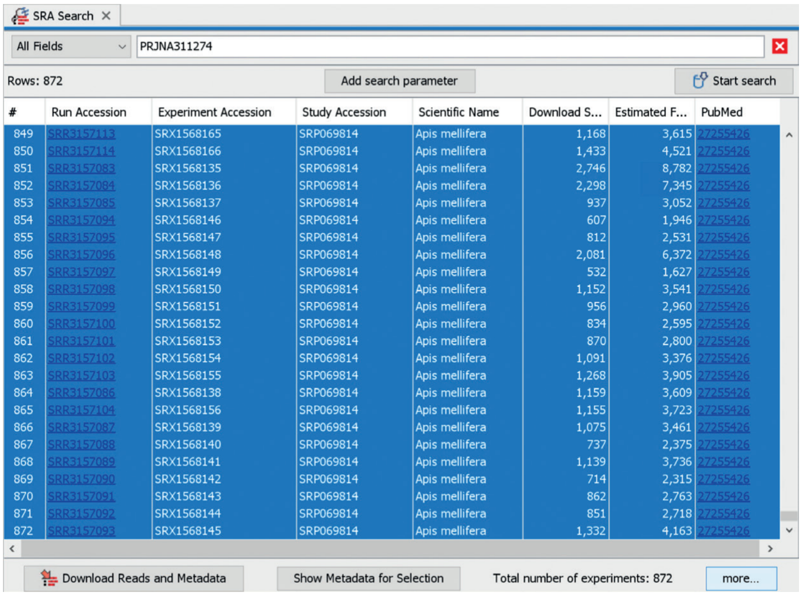
Workflow description

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.



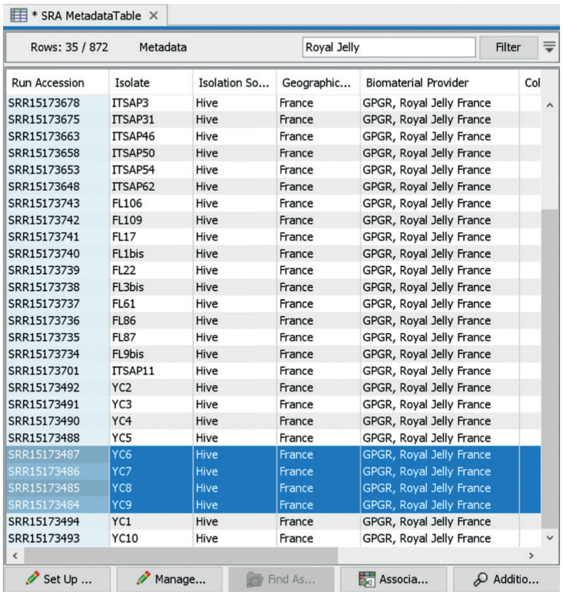
PRJNA311274

Rows: 872

#	Run Accession	Experiment Accession	Study Accession	Scientific Name	Download S...	Estimated F...	PubMed
849	SRR3157113	SRX1568165	SRP069814	Apis mellifera	1,168	3,615	27255426
850	SRR3157114	SRX1568166	SRP069814	Apis mellifera	1,433	4,521	27255426
851	SRR3157083	SRX1568135	SRP069814	Apis mellifera	2,746	8,782	27255426
852	SRR3157084	SRX1568136	SRP069814	Apis mellifera	2,298	7,345	27255426
853	SRR3157085	SRX1568137	SRP069814	Apis mellifera	937	3,052	27255426
854	SRR3157094	SRX1568146	SRP069814	Apis mellifera	607	1,946	27255426
855	SRR3157095	SRX1568147	SRP069814	Apis mellifera	812	2,531	27255426
856	SRR3157096	SRX1568148	SRP069814	Apis mellifera	2,081	6,372	27255426
857	SRR3157097	SRX1568149	SRP069814	Apis mellifera	532	1,627	27255426
858	SRR3157098	SRX1568150	SRP069814	Apis mellifera	1,152	3,541	27255426
859	SRR3157099	SRX1568151	SRP069814	Apis mellifera	956	2,960	27255426
860	SRR3157100	SRX1568152	SRP069814	Apis mellifera	834	2,595	27255426
861	SRR3157101	SRX1568153	SRP069814	Apis mellifera	870	2,800	27255426
862	SRR3157102	SRX1568154	SRP069814	Apis mellifera	1,091	3,376	27255426
863	SRR3157103	SRX1568155	SRP069814	Apis mellifera	1,268	3,905	27255426
864	SRR3157086	SRX1568138	SRP069814	Apis mellifera	1,159	3,609	27255426
865	SRR3157104	SRX1568156	SRP069814	Apis mellifera	1,155	3,723	27255426
866	SRR3157087	SRX1568139	SRP069814	Apis mellifera	1,075	3,461	27255426
867	SRR3157088	SRX1568140	SRP069814	Apis mellifera	737	2,375	27255426
868	SRR3157089	SRX1568141	SRP069814	Apis mellifera	1,139	3,736	27255426
869	SRR3157090	SRX1568142	SRP069814	Apis mellifera	714	2,315	27255426
870	SRR3157091	SRX1568143	SRP069814	Apis mellifera	862	2,763	27255426
871	SRR3157092	SRX1568144	SRP069814	Apis mellifera	851	2,718	27255426
872	SRR3157093	SRX1568145	SRP069814	Apis mellifera	1,332	4,163	27255426

Total number of experiments: 872

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



Rows: 35 / 872 Metadata

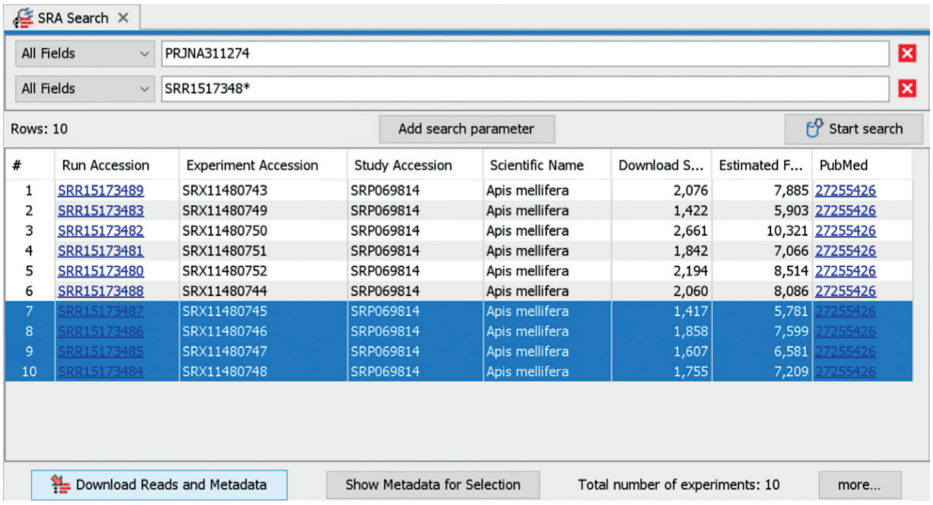
Royal Jelly

Run Accession	Isolate	Isolation So...	Geographic...	Biomaterial Provider	Col
SRR15173678	ITSAP3	Hive	France	GPGR, Royal Jelly France	
SRR15173675	ITSAP31	Hive	France	GPGR, Royal Jelly France	
SRR15173663	ITSAP46	Hive	France	GPGR, Royal Jelly France	
SRR15173658	ITSAP50	Hive	France	GPGR, Royal Jelly France	
SRR15173653	ITSAP54	Hive	France	GPGR, Royal Jelly France	
SRR15173648	ITSAP62	Hive	France	GPGR, Royal Jelly France	
SRR15173743	FL106	Hive	France	GPGR, Royal Jelly France	
SRR15173742	FL109	Hive	France	GPGR, Royal Jelly France	
SRR15173741	FL17	Hive	France	GPGR, Royal Jelly France	
SRR15173740	FL1bis	Hive	France	GPGR, Royal Jelly France	
SRR15173739	FL22	Hive	France	GPGR, Royal Jelly France	
SRR15173738	FL3bis	Hive	France	GPGR, Royal Jelly France	
SRR15173737	FL61	Hive	France	GPGR, Royal Jelly France	
SRR15173736	FL86	Hive	France	GPGR, Royal Jelly France	
SRR15173735	FL87	Hive	France	GPGR, Royal Jelly France	
SRR15173734	FL9bis	Hive	France	GPGR, Royal Jelly France	
SRR15173701	ITSAP11	Hive	France	GPGR, Royal Jelly France	
SRR15173492	YC2	Hive	France	GPGR, Royal Jelly France	
SRR15173491	YC3	Hive	France	GPGR, Royal Jelly France	
SRR15173490	YC4	Hive	France	GPGR, Royal Jelly France	
SRR15173488	YC5	Hive	France	GPGR, Royal Jelly France	
SRR15173487	YC6	Hive	France	GPGR, Royal Jelly France	
SRR15173486	YC7	Hive	France	GPGR, Royal Jelly France	
SRR15173485	YC8	Hive	France	GPGR, Royal Jelly France	
SRR15173484	YC9	Hive	France	GPGR, Royal Jelly France	
SRR15173494	YC1	Hive	France	GPGR, Royal Jelly France	
SRR15173493	YC10	Hive	France	GPGR, Royal Jelly France	

Figure 16.
Searching the metadata table for Royal Jelly samples and identifying run accession IDs for subsequent import using the “SRA Search” table.

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.

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).



All Fields: PRJNA311274

All Fields: SRR1517348*

Rows: 10

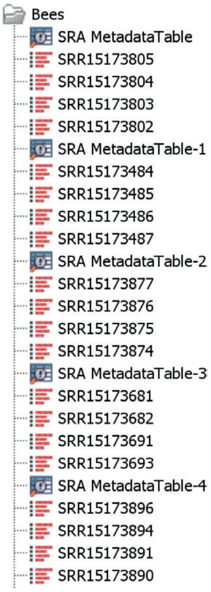
#	Run Accession	Experiment Accession	Study Accession	Scientific Name	Download S...	Estimated F...	PubMed
1	SRR15173489	SRX11480743	SRP069814	Apis mellifera	2,076	7,885	27255426
2	SRR15173483	SRX11480749	SRP069814	Apis mellifera	1,422	5,903	27255426
3	SRR15173482	SRX11480750	SRP069814	Apis mellifera	2,661	10,321	27255426
4	SRR15173481	SRX11480751	SRP069814	Apis mellifera	1,842	7,066	27255426
5	SRR15173480	SRX11480752	SRP069814	Apis mellifera	2,194	8,514	27255426
6	SRR15173488	SRX11480744	SRP069814	Apis mellifera	2,060	8,086	27255426
7	SRR15173487	SRX11480745	SRP069814	Apis mellifera	1,417	5,781	27255426
8	SRR15173486	SRX11480746	SRP069814	Apis mellifera	1,858	7,599	27255426
9	SRR15173485	SRX11480747	SRP069814	Apis mellifera	1,607	6,581	27255426
10	SRR15173484	SRX11480748	SRP069814	Apis mellifera	1,755	7,209	27255426

Total number of experiments: 10

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

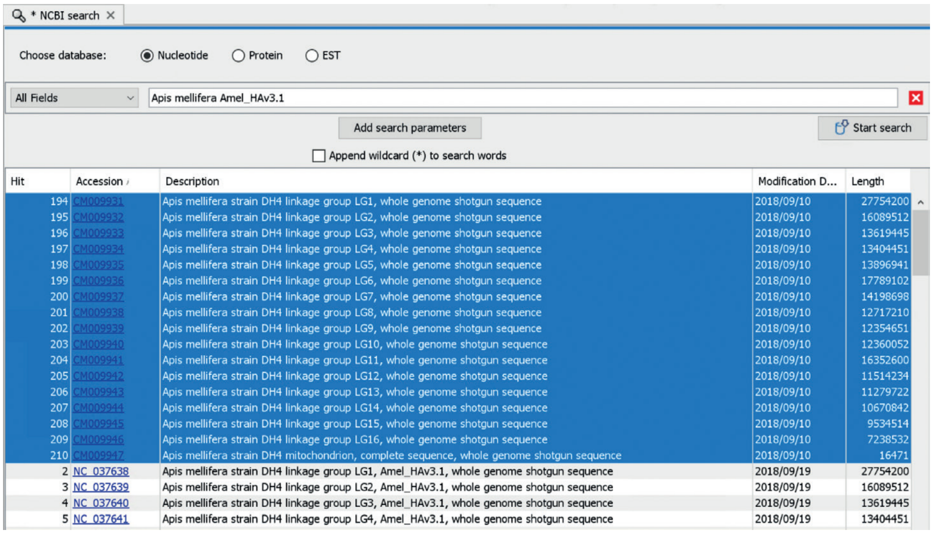
Bee genome

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).



- Bees
  - SRA MetadataTable
  - SRR15173805
  - SRR15173804
  - SRR15173803
  - SRR15173802
  - SRA MetadataTable-1
  - SRR15173484
  - SRR15173485
  - SRR15173486
  - SRR15173487
  - SRA MetadataTable-2
  - SRR15173877
  - SRR15173876
  - SRR15173875
  - SRR15173874
  - SRA MetadataTable-3
  - SRR15173681
  - SRR15173682
  - SRR15173691
  - SRR15173693
  - SRA MetadataTable-4
  - SRR15173896
  - SRR15173894
  - SRR15173891
  - SRR15173890

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



Choose database: ● Nucleotide ○ Protein ○ EST

All Fields: Apis mellifera Amel_HAv3.1

□ Append wildcard (*) to search words

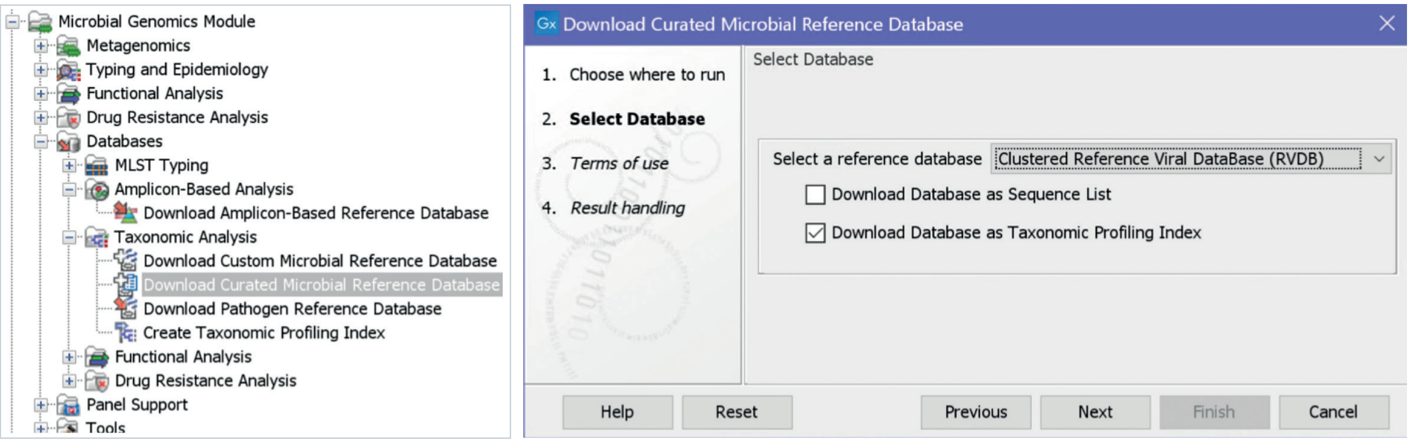
Hit	Accession	Description	Modification D...	Length
194	CM009931	Apis mellifera strain DH4 linkage group LG1, whole genome shotgun sequence	2018/09/10	27754200
195	CM009932	Apis mellifera strain DH4 linkage group LG2, whole genome shotgun sequence	2018/09/10	16089512
196	CM009933	Apis mellifera strain DH4 linkage group LG3, whole genome shotgun sequence	2018/09/10	13619445
197	CM009934	Apis mellifera strain DH4 linkage group LG4, whole genome shotgun sequence	2018/09/10	13404451
198	CM009935	Apis mellifera strain DH4 linkage group LG5, whole genome shotgun sequence	2018/09/10	13896941
199	CM009936	Apis mellifera strain DH4 linkage group LG6, whole genome shotgun sequence	2018/09/10	17789102
200	CM009937	Apis mellifera strain DH4 linkage group LG7, whole genome shotgun sequence	2018/09/10	14198698
201	CM009938	Apis mellifera strain DH4 linkage group LG8, whole genome shotgun sequence	2018/09/10	12717210
202	CM009939	Apis mellifera strain DH4 linkage group LG9, whole genome shotgun sequence	2018/09/10	12354651
203	CM009940	Apis mellifera strain DH4 linkage group LG10, whole genome shotgun sequence	2018/09/10	12360052
204	CM009941	Apis mellifera strain DH4 linkage group LG11, whole genome shotgun sequence	2018/09/10	16352600
205	CM009942	Apis mellifera strain DH4 linkage group LG12, whole genome shotgun sequence	2018/09/10	11514234
206	CM009943	Apis mellifera strain DH4 linkage group LG13, whole genome shotgun sequence	2018/09/10	11279722
207	CM009944	Apis mellifera strain DH4 linkage group LG14, whole genome shotgun sequence	2018/09/10	10670842
208	CM009945	Apis mellifera strain DH4 linkage group LG15, whole genome shotgun sequence	2018/09/10	9534514
209	CM009946	Apis mellifera strain DH4 linkage group LG16, whole genome shotgun sequence	2018/09/10	7238532
210	CM009947	Apis mellifera strain DH4 mitochondrion, complete sequence, whole genome shotgun sequence	2018/09/10	16471
2	NC_037638	Apis mellifera strain DH4 linkage group LG1, Amel_HAv3.1, whole genome shotgun sequence	2018/09/19	27754200
3	NC_037639	Apis mellifera strain DH4 linkage group LG2, Amel_HAv3.1, whole genome shotgun sequence	2018/09/19	16089512
4	NC_037640	Apis mellifera strain DH4 linkage group LG3, Amel_HAv3.1, whole genome shotgun sequence	2018/09/19	13619445
5	NC_037641	Apis mellifera strain DH4 linkage group LG4, Amel_HAv3.1, whole genome shotgun sequence	2018/09/19	13404451

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).


Microbial Genomics Module
Metagenomics
Typing and Epidemiology
Functional Analysis
Drug Resistance Analysis
Databases
MLST Typing
Amplicon-Based Analysis
Download Amplicon-Based Reference Database
Taxonomic Analysis
Download Custom Microbial Reference Database
Download Curated Microbial Reference Database
Download Pathogen Reference Database
Create Taxonomic Profiling Index
Functional Analysis
Drug Resistance Analysis
Panel Support
Tools
Gx Download Curated Microbial Reference Database
Help
Reset
Previous
Next
Finish
Cancel

1. Choose where to run
2. **Select Database**
3. *Terms of use*
4. *Result handling*

Select Database

Select a reference database: Clustered Reference Viral DataBase (RVDB)

☐ Download Database as Sequence List

☑ Download Database as Taxonomic Profiling Index

Figure 20.
Download of the viral reference database.

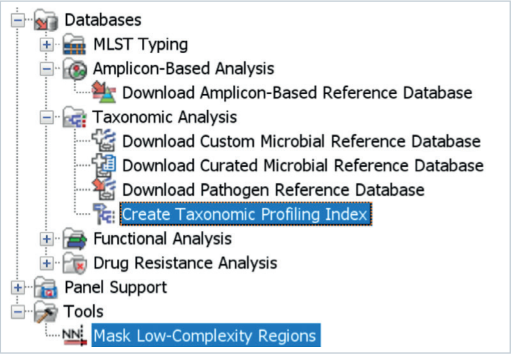
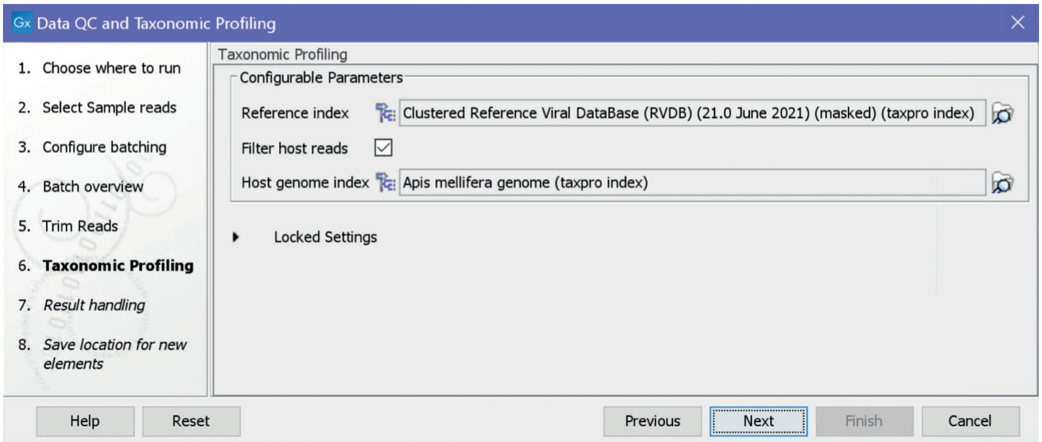

Databases
MLST Typing
Amplicon-Based Analysis
Download Amplicon-Based Reference Database
Taxonomic Analysis
Download Custom Microbial Reference Database
Download Curated Microbial Reference Database
Download Pathogen Reference Database
Create Taxonomic Profiling Index
Functional Analysis
Drug Resistance Analysis
Panel Support
Tools
Mask Low-Complexity Regions

Figure 21.
The tools for creating the masked reference database.

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.



1. Choose where to run
2. Select Sample reads
3. Configure batching
4. Batch overview
5. Trim Reads
6. **Taxonomic Profiling**
7. *Result handling*
8. *Save location for new elements*

Taxonomic Profiling

Configurable Parameters

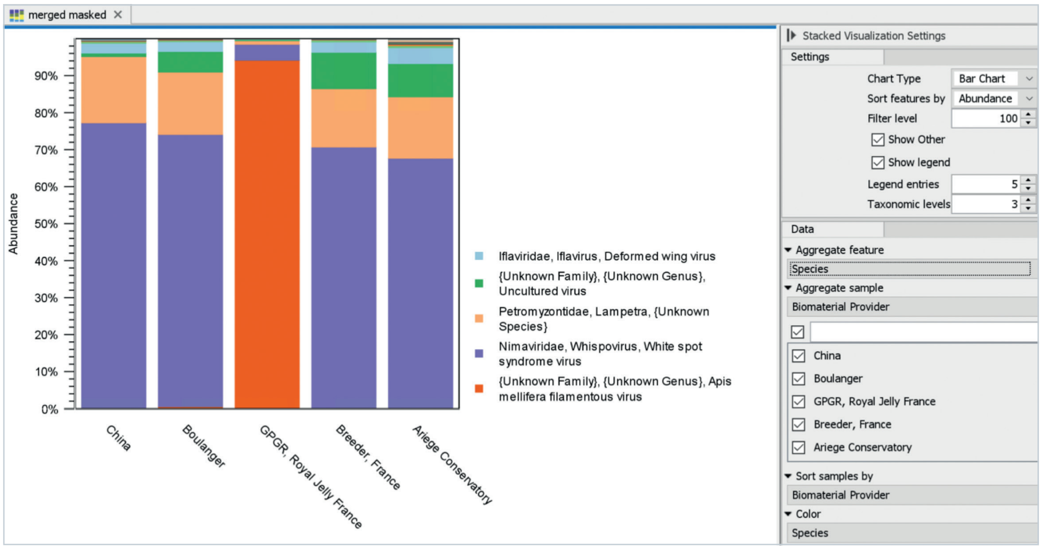
Reference index: Clustered Reference Viral DataBase (RVDB) (21.0 June 2021) (masked) (taxpro index)

Filter host reads: ☑

Host genome index: Apis mellifera genome (taxpro index)

Locked Settings

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


merged masked
Abundance
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
China
Boulanger
GPGR, Royal Jelly France
Breeder, France
Ariege Conservatory
Iflaviridae, Iflavirus, Deformed wing virus
{Unknown Family}, {Unknown Genus}, Uncultured virus
Petromyzontidae, Lampetra, {Unknown Species}
Nimaviridae, Whispovirus, White spot syndrome virus
{Unknown Family}, {Unknown Genus}, Apis mellifera filamentous virus

Stacked Visualization Settings

Settings

Chart Type: Bar Chart

Sort features by: Abundance

Filter level: 100

☑ Show Other

☑ Show legend

Legend entries: 5

Taxonomic levels: 3

Data

Aggregate feature: Species

Aggregate sample: Biomaterial Provider

☑

☑ China

☑ Boulanger

☑ GPGR, Royal Jelly France

☑ Breeder, France

☑ Ariege Conservatory

Sort samples by: Biomaterial Provider

Color: Species

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

1. Chumová Z, et al. (2021). Repeat proliferation and partial endoreplication jointly shape the patterns of genome size evolution in orchids. Plant J.;**107**(2):511–524. doi: 10.1111/tpj.15306. Epub 2021 May 25. PMID: 33960537.

2. Wragg D, et al. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. Sci Rep.;**6**:27168. doi: 10.1038/srep27168. PMID: 27255426; PMCID: PMC4891733.

Learn more and request a free trial at **digitalinsights.qiagen.com/GXWBP**.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.