

An example of CAMBerVis usage to identify connected components (gene families) that are highly conserved but annotated in very few strains.

We load example data for *S. aureus*

The screenshot displays the CamberVis software interface. The main window is titled 'CamberVis window' and is currently blank. A 'Load input data' dialog box is open in the center, featuring two tabs: 'Examples' and 'Load data'. Under the 'Examples' tab, there are two radio button options: 'M. tuberculosis' (unselected) and 'S. aureus' (selected). At the bottom of the dialog, there is a progress bar labeled 'Loading unified annotations', a 'Load example' button, and a 'Close' button. The background interface includes a menu bar (File, Display, Navigate, Window, Tools, Help) and several other windows: 'MultigeneZoom window' with fields for Multigene ID, Lengths, and Locus; 'Component window' with fields for component ID and type; 'ComponentsStats window' with a table of statistics; and 'TasksTable window' with a table of task results.

07:19 18:51 SOCCF-CBRI-001 Workshop ScreenHunter

File Display Navigate Window Tools Help

CamberVis window

MultigeneZoom window

Multigene ID:
Lengths:
Locus:

Highlight alt. NCBI API blast

Load input data

Examples Load data

M. tuberculosis

S. aureus

Loading unified annotations

Load example Close

Component window

Find ClustalW2

Component ID:
Component type:
Number of multigenes:

ComponentsStats window

cc id	# of multigenes	# of main multig...	# of plasmid mul...	# of strains	# of ann strains	max length	max # of TIS	cc type
-------	-----------------	---------------------	---------------------	--------------	------------------	------------	--------------	---------

TasksTable window

ID	JobID	Time	Tool	Status
----	-------	------	------	--------

We sort gene families by the column „# of strains” in order to identify highly conserved gene families (with genes present in all strains)

ComponentsStats window

cc id	# of multigenes	# of main multig...	# of plasmid mul...	# of strains	# of ann strains	max length	max # of TIS	cc type
0	22	22	0	22	21	64		
1	22	22	0	22	21	63		
10	22	22	0	22	22	200		
100	22	22	0	22	22	78		
1000	22	22	0	22	22	198		
1001	22	22	0	22	22	74		
1002	22	22	0	22	22	60		

Output - Data loading

Click on header to sort by column. Double click to zoom the selected component.

- cc_id:** connected component identifier
- # of multigenes:** number of multigenes in the connected component
- # of main multigenes:** number of multigenes in the connected component located on the main genome
- # of plasmid multigenes:** number of multigenes in the connected component on plasmids
- # of strains:** number of strains with at least one multigene in the connected component
- # of ann strains:** number of strains with at least one originally annotated multigene in the connected component
- max length:** length of the longest multigene in the connected component
- max # of TIS:** maximal number of TISs among multigenes in the connected component
- cc_type:** type of the connected component (ANCHOR / NON-ANCHOR)

Then we sort the same table by the column „# of ann strains” means the number of strains with at least one annotated element of ta given gene family. There are examples of gene families with elements predicted in all strains, but annotated in only one strain.

The screenshot displays the CamberVis software interface. The main window shows genomic tracks for eight strains: 04-02981, COL, ED133, ED98, JH1, JH9, JKD6008, and JKD6159. Each track shows gene annotations with colored bars and arrows. The interface includes a menu bar (File, Display, Navigate, Window, Tools, Help), a toolbar, and several floating windows:

- MultigeneZoom window:** Displays Multigene ID, Lengths, and Locus. It has buttons for 'Highlight alt.' and 'NCBI API blast'.
- Component window:** Includes 'Find' and 'ClustalW2' buttons. It shows fields for 'Connected component ID:', 'Connected component type:', and 'Number of multigenes:'.
- ComponentsStats window:** A table with columns: cc id, # of multigenes, # of main multig..., # of plasmid mul..., # of strains, # of ann str..., max length, max # of TIS, and cc type.
- Output - Data loading window:** Contains a table with the same columns as ComponentsStats, showing data for cc IDs 1024, 1190, 126, 129, 134, 1434, and 146.
- TasksTable window:** Has 'Clean' and 'Results' buttons.

At the bottom, a help text box provides instructions: 'Click on header to sort by column. Double click to zoom the selected component.' It also defines the following terms:

- cc_id:** connected component identifier
- # of multigenes:** number of multigenes in the connected component
- # of main multigenes:** number of multigenes in the connected component located on the main genome
- # of plasmid multigenes:** number of multigenes in the connected component on plasmids
- # of strains:** number of strains with at least one multigene in the connected component
- # of ann strains:** number of strains with at least one originally annotated multigene in the connected component
- max length:** length of the longest multigene in the connected component
- max # of TIS:** maximal number of TISs among multigenes in the connected component
- cc_type:** type of the connected component (ANCHOR / NON-ANCHOR)

families are usually short. The gene family number 1-24 is a typical

camber-vis 201007282301 07:10 18:00 SDC/ CBB/ 001 Workshop Screenshots

File Display Navigate Window Tools Help

CamberVis window

MultigeneZoom window

Multigene ID: SAAV_0591
Lengths: 303*,
Locus: (636183,636485)

Highlight alt. NCBI API blast

AAAAGAT ACTTT GTT AAGAACT ATT AAGCATT ATGAAGACTT AAT ATT GGAGGT GTCGCC
AT GAT ACAACAAAT AACACAT ATT ATGATT ATT AGTT CACT CATT ATTTTT GGAATT GCA
TT AAT CAT CT GTT ATTT AGATT AAT CAAGGGACCT ACAACAGCAGAT CGT GT CGTT ACA
TTT GAT ACAACAAAGT GCT GT CGT AAT GT CAATT GT GGGT GT GTT AAGT GT ACTT ATGGCC
ACCGTT CTTT CTT AGATT CAAT CAT GCT CATT GCCATT AT ATCTTT GT AAGT CT GTT
TCAAT AT CACGCTTT ATTGCT GGGGGCAT GT GTTT AAT GGAAT AACAAAAGAAAT CTT
TAG

Caret position:

Component window

Find ClustalW2

Connected component ID: 1024
Connected component type: ANCHOR
Number of multigenes: 22 (in 22 out of 22 strains)

x - 1743454 04-02981 (1)
x - 1754596 COL (1)
x - 1737867 ED133 (1)
x - 1717949 ED98 (1)
x - 1833533 JH1 (1)
x - 1833659 JH9 (1)
x - 1756127 JKD6008 (1)
x - 1736682 JKD6159 (1)
x - 1810542 MDS252 (1)

ComponentsStats window

cc id	# of multigenes	# of main multig...	# of plasmid mul...	# of strains	# of ann strains	max length	max # of TIS	cc type
3365	22	22	0	22	1	123	1	ANCHOR
818	22	22	0	22	1	123	1	ANCHOR
1922	22	22	0	22	1	117	1	ANCHOR
1024	22	22	0	22	1	111	1	ANCHOR
134	22	22	0	22	1	111	1	ANCHOR
3460	22	22	0	22	1	111	1	ANCHOR
2093	22	22	0	22	1	108	1	ANCHOR

Output - Data loading

Gene id: SAOUHSC_01780
Strand: -
Locus: (1680746,1680636)
TISs: 111*,
Conn comp id: 1024
The conn. comp. contains 22 multigenes in 22 strains,
1 strains contain at least one annotated multigene.

TasksTable window

Clean Results

ID	JobID	Time	Tool	Status
----	-------	------	------	--------

main: Mu3 Range: (1785696,1787806)
main: Mu50 Range: (1784296,1786406)
main: N315 Range: (1707903,1710013)
main: NCTC8325 Range: (1679636,1681746)
main: Newman Range: (1655526,1657636)
main: RF122 Range: (1814924,1817034)
main: ST398 Range: (1655526,1657636)

The problem also can be found in gene families with pretty long genes like. Connected component number 2791 have elements with length 327bp. We zoom the genome browser on in by double click. Here the spurious annotation might be caused by another highly overlapping gene family: 944, which is annotated in all strains except NCTC8325. It may be a case of overlapping genes or wrong annotations.

Gene id: x
Strand: +
Locus: (620050,620352)
TISs: 303,
Conn comp id: 944
 The conn. comp. contains 22 multigenes in 22 strains,
 21 strains contain at least one annotated multigene.

MultigeneZoom window
 Multigene ID: x
 Lengths: 303,
 Locus: (620050,620352)
 Highlight alt. NCBI API blast

```

AAAAGAT AGTTT GTT AAGAACT ATT AAGCATT AT GAAGACTT AAT ATT GGAGGT GT CGCC
AT GAT AC AAAC AAT AACACAT ATT AT GATT ATT AGTT CACT CATT ATTTT GGAAAT GCA
TT AAT CAT CT GTT ATTT AGATT AAT CAAGGGACCT ACAACAGCAGAT CGT GT CGT AC A
TTT GAT ACAAC AAGT GCT GT C GT AAT GT CAATT GT GGGT GT GTT AAGT GT ACTT AT GGGC
ACCGTT CTTT CTT AGATT CAAT CAT GCT CATT GCCATT AT AT CTTT GT AAGT CT GTT
T CAAT AT CACGCTTT ATT GGT GGGGGCCAT GT GTT AAT GGAAT AAC AAAAGAAAT CTT
TAG
  
```

Caret position: 172

Component window
 Find ClustalW2
 Connected component ID: 2791
 Connected component type: ANCHOR
 Number of multigenes: 22 (in 22 out of 22 strains)

x - 672700	04-02981 (1)
x - 708688	COL (1)
x - 687840	ED133 (1)
x - 636146	ED98 (1)
x - 715387	JH1 (1)
x - 715511	JH9 (1)
x - 701859	JKD6008 (1)
x - 660108	JKD6159 (1)
x - 681782	MP54252 (1)

ComponentsStats window

cc id	# of multigenes	# of main multig...	# of plasmid mul...	# of strains	# of ann strains	max length	max # of TIS	cc type
2353	22	22	0	22	1	393	1	ANCHOR
2791	22	22	0	22	1	327	1	ANCHOR
2810	22	22	0	22	1	288	1	ANCHOR
1520	22	22	0	22	1	225	1	ANCHOR
3505	22	22	0	22	1	225	1	ANCHOR
1702	22	22	0	22	1	219	1	ANCHOR
2310	22	22	0	22	1	216	1	ANCHOR

TasksTable window

ID	JobID	Time	Tool	Status
1	055aa0d...	2011-07-...	ClustalW	OK
2	94bcd5e...	2011-07-...	ClustalW	OK
3	05d3039...	2011-07-...	ClustalW	OK
4	7a9e865...	2011-07-...	ClustalW	OK
5	11f6d85d...	2011-07-...	NCBI API	OK