

An example of CAMBerVis usage
to identify gene families with
highly inconsistent annotations
of TISs among strains

We load example input data for *S. aureus*

The screenshot displays the CamberVis software interface. The main window is titled "CamberVis window" and has a light blue background. A dialog box titled "Load input data" is open in the center, featuring two tabs: "Examples" and "Load data". Under the "Examples" tab, there are two radio button options: "M. tuberculosis" and "S. aureus", with "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), a toolbar, and a task pane on the right with tabs for "Component...", "Tasks...", "MultigeneZ...", and "Component ...". The task pane shows a table with columns "ID", "JobID", "Time", "Tool", and "Status". At the bottom of the screen, an "Output - Data loading" window shows the message: "The file with info about strains been loaded." The system tray at the bottom right shows the "Load input data" window is active at 62% zoom.

We can personalize the panels localization (OPTIONAL)

The screenshot displays the CamberVis software interface. The main window, titled "CamberVis window", contains eight tracks of genomic data, each with a "main" label and a "Range: (0,100000)" indicator. The tracks are labeled: main: 04-02981, main: COL, main: ED133, main: ED98, main: JH1, main: JH9, main: JKD6008, and main: JKD6159. To the right, the "MultigeneZoom window" shows fields for "Multigene ID:", "Lengths:", and "Locus:", along with "Highlight alt." and "NCBI API blast" buttons. Below it, the "Component window" includes "Find" and "ClustalW2" buttons, and fields for "Connected component ID:", "Connected component type:", and "Number of multigenes:". At the bottom left, the "ComponentsStats window" displays a table with columns: cc id, # of multigenes, # of main multigenes, # of plasmid multigenes, # of strains, # of ann strains, max length, max # of TIS, and cc type. A tooltip over the table reads "This is a ComponentsStats window". At the bottom right, the "TasksTable window" has "Clean" and "Results" buttons and a table with columns: ID, JobID, Time, Tool, and Status.

| cc id | # of multigenes | # of main multigenes | # of plasmid multigenes | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|----------------------|-------------------------|--------------|------------------|------------|--------------|---------|
| 0 | 22 | 22 | 0 | 22 | 21 | 663 | 1 | ANCHOR |
| 1 | 22 | 22 | 0 | 22 | 21 | 636 | 1 | ANCHOR |
| 10 | 22 | 22 | 0 | 22 | 22 | 2007 | 1 | ANCHOR |
| 100 | 22 | 22 | 0 | 22 | 22 | 789 | 1 | ANCHOR |
| 1000 | 22 | 22 | 0 | 22 | 22 | 1986 | 2 | ANCHOR |
| 1001 | 22 | 22 | 0 | 22 | 22 | 744 | 1 | ANCHOR |
| 1002 | 22 | 22 | 0 | 22 | 22 | 600 | 1 | ANCHOR |

We sort the list of connected components (gene families) by the number of TISs. In the example we pick the gene family with ID:1425. It is an ANCHOR and has 5 different TISs annotated. Double click zooms in the genome browser.

The screenshot shows the CamberVis genome browser interface. The main window displays a list of connected components (gene families) sorted by the number of Transcription Start Sites (TISs). The selected component, ID:1425, is highlighted in blue. It is an ANCHOR and has 5 different TISs annotated. The interface includes a MultigeneZoom window, a Component window, and a ComponentsStats window.

ComponentsStats window

| cc id | # of multigenes | # of main multig... | # of plasmid mul... | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|---------------------|---------------------|--------------|------------------|------------|--------------|---------------|
| 552 | | | | | | 19 | 1647 | 13 NON_ANCHOR |
| 1409 | | | | | | 2 | 1317 | 6 NON_ANCHOR |
| 8 | | | | | | 22 | 642 | 6 NON_ANCHOR |
| 111 | | | | | | 15 | 426 | 5 NON_ANCHOR |
| 1191 | | | | | | 22 | 816 | 5 NON_ANCHOR |
| 1425 | | | | | | 22 | 888 | 5 ANCHOR |
| 1517 | | | | | | 22 | 1557 | 5 NON_ANCHOR |

Component window

Connected component ID: 1425
 Connected component type: ANCHOR
 Number of multigenes: 22 (in 22 out of 22 strains)

| | |
|-------------------------|--------------|
| SA2981_2478 - 2628474 | 04-02981 (1) |
| SACOL2555 - 2614414 | COL (1) |
| SAOV_2586c - 2641141 | ED133 (1) |
| SAAV_2608 - 2629864 | ED98 (1) |
| SaurJH1_2618 - 2713656 | JH1 (1) |
| SaurJH9_2565 - 2713780 | JH9 (1) |
| SAA6008_02579 - 2707982 | JKD6008 (1) |
| SAA6159_02438 - 2618241 | JKD6159 (1) |
| SAA6232_2706800 | MPSA323 (1) |

ComponentsStats window

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)

We run on-the-fly analysis by CLUSTALW to check how conserved is the gene family (we included promoters of length 60bp).

The screenshot displays the CamberVis software interface. The main window shows a multigene alignment with tracks for various strains. A 'Configure ClustalW task' dialog box is open, showing the following configuration:

- Connected component ID: 1425
- ATG 828: Mu50 (SAV2542)
- ATG 828: USA300-TCH1516 (USA300HOU_2534)
- ATG 828: JH1 (SaurJH1_2618)
- ATG 828: JKD6159 (SAA6159_02438)
- ATG 828: N315 (SA2330)
- ATG 828: MSSA476 (SAS2428)
- ATG 828: Newman (NWMN_2441)
- ATG 828: TW20 (SATW20_26630)
- ATG 828: ED133 (SAOV_2586c)
- ATG 828: JKD6008 (SAA6008_02579)
- ATG 828*: MW2 (MW2463)
- ATG 828*: NCTC8325 (SAOUH5C_02852)

Additional settings in the dialog:

- Your ClustalW: C:\Documents and Settings\Administrator\My Documents\Dropbox\camber-vis\ext\clustalw2.exe
- Nucleotide sequences
- Amino acid sequences
- Include promoters length: 60
- Buttons: Multiple alignment, Close, Run ClustalW

Other windows visible:

- MultigeneZoom window:** Multigene ID, Lengths, Locus, Highlight alt., NCBI API blast.
- ComponentsStats window:** Table with columns: cc id, # of multigenes, # of main multig..., # of plasmid mul..., # of strains, # of ann strains, max length, max # of TIS, cc type.
- TasksTable window:** Table with columns: ID, JobID, Time, Tool, Status.

| cc id | # of multigenes | # of main multig... | # of plasmid mul... | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|---------------------|---------------------|--------------|------------------|------------|--------------|------------|
| 552 | 177 | 177 | 0 | 22 | 19 | 1647 | 13 | NON_ANCHOR |
| 1409 | 13 | 13 | 0 | 3 | 2 | 1317 | 6 | NON_ANCHOR |
| 8 | 112 | 112 | 0 | 22 | 22 | 642 | 6 | NON_ANCHOR |
| 111 | 588 | 588 | 0 | 22 | 15 | 426 | 5 | NON_ANCHOR |
| 1191 | 35 | 35 | 0 | 22 | 22 | 816 | 5 | NON_ANCHOR |
| 1425 | 22 | 22 | 0 | 22 | 22 | 888 | 5 | ANCHOR |
| 1517 | 29 | 29 | 0 | 22 | 22 | 1557 | 5 | NON_ANCHOR |

| ID | JobID | Time | Tool | Status |
|----|-------------|-------------|----------|---------|
| 1 | e69d32cf... | 2011-07-... | ClustalW | WORKING |

We have computed basic statistics for the multiple alignment, there are only 2 mutation points. This suggests that the differences in annotations are caused by different annotation tools.

The screenshot displays the CamberVis software interface. The main window shows a multiple sequence alignment of DNA sequences. A ClustalW task window is open, displaying the alignment and a list of sequences. The ClustalW window shows the alignment length as 888 and identities as 827 (93%). The ClustalW task ID is e69d32cf-bfb2-4355-98ca-93abeb2d0741. The ClustalW window also shows a list of sequences with their IDs and lengths.

ClustalW task ID: e69d32cf-bfb2-4355-98ca-93abeb2d0741

Caret position: 48
Caret on sequence: SATW20_26630 - TW20 888

Alignment length: 888
Identities: 827 (93%)

ClustalW window sequence list:

- SA2300 - M315 888
- SAB2416c - RF122 888
- HMPREF0772_10649 + TCH60 888
- SA0V_2586c - ED100 888
- SAW_2608 - ED98 888
- SAUSA300_2480 - USA300-IPR0757 888
- SAA6008_02579 - JKD6008 888
- SAS2428 - M3SA476 888
- SAW_2526 - M303 888
- MW2463 - MW2 888
- SA2981_2478 - 04-02981 888
- SACOL2555 - COL 888
- SAU2542 - M350 888
- USA300H0V_2534 - USA300-TCH1516 888
- Saur_JH1_2618 - JH1 888
- NUMB_2441 - Newman 888
- SAP1G2592 - ST398 888
- SAR2622 - M3SA252 888
- SAA6159_02438 - JKD6159 888
- SA0UHSC_02852 - NCTC8325 888
- Saur_JH9_2565 - JH9 888
- SATW20_26630 - TW20 888

ComponentsStats window:

| cc id | # of multigenes | # of main multig... | # of plasmid mul... | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|---------------------|---------------------|--------------|------------------|------------|--------------|------------|
| 552 | 177 | 177 | 0 | 22 | 19 | 1647 | 13 | NON_ANCHOR |
| 1409 | 13 | 13 | 0 | 3 | 2 | 1317 | 6 | NON_ANCHOR |
| 8 | 112 | 112 | 0 | 22 | 22 | 642 | 6 | NON_ANCHOR |
| 111 | 588 | 588 | 0 | 22 | 15 | 426 | 5 | NON_ANCHOR |
| 1191 | 35 | 35 | 0 | 22 | 22 | 816 | 5 | NON_ANCHOR |
| 1425 | 22 | 22 | 0 | 22 | 22 | 888 | 5 | ANCHOR |
| 1517 | 29 | 29 | 0 | 22 | 22 | 1557 | 5 | NON_ANCHOR |

TasksTable window:

| ID | JobID | Time | Tool | Status |
|----|-------------|-------------|----------|--------|
| 1 | e69d32cf... | 2011-07-... | ClustalW | OK |

By queries to NCBI database (nr) we can check which TIS is most often annotated.
 We run NCBI queries for different TISs.

The screenshot displays the CamberVis software interface. A central dialog box titled "Configure NCBI BLAST task" is open, showing the following configuration:

- Multigene ID:** SA2981_2478
- Multigene:**

```

MLVDIKHMKYFIEVVKQGGMTNASKSLYIAQPTISKAIKDIENEMGTPLFDRSKRHLIL
TDACQIFYERSKEIVALDYLPSEMERLNGLETGHINMGMSAVMMNKILINILGAFHQQY
PNVTYNIENGCKTIEQQIINDEVDICVTTLPVDHHIFDYTTLDKEDLRLIIVSREHRLAK
YETVKLEDLAGEDFILFNKDLYLNDKIIENAKNVGFPVNTVAQISQWHVIEDLVNLEGI
SILPTSISEQLNQDVKLLRIEDAHVHWELGVVWFKDKQLSHATTKWIEFLKDRLG*
                
```
- Task configuration:**
 - Amino acid sequences
 - Nucleotide sequences
- Database:** nr
- E-value:** 1e-6
- TIS:** ATG 888*
- Your email:** m.wozniak@mimuw.edu.pl
- Tool-name:** CamberVis

Below the configuration, there is a section for "Important policy restrictions" with the following text:

Before you send any task, please, be informed about NCBI's policy:
 * do not send requests less than once every 3 seconds.
 * limit queries to the off peak hours of 9 PM to 5 AM Eastern Standard Time (USA).
 * use the `email=` field (and the `stool=` for distributed software), in BLAST URL API, so that we can track your project and contact you if there is a problem
 For more details, visit: <http://www.ncbi.nlm.nih.gov/BLAST/Doc/urllapi.html>

At the bottom of the dialog, the status is "Sending the request." and there are "Close" and "Run NCBI BLAST API" buttons.

In the background, the "ComponentsStats window" shows a table of components:

| cc id | # of multigenes | # of main multig... | # of plasmid mul... | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|---------------------|---------------------|--------------|------------------|------------|--------------|------------|
| 552 | 177 | 177 | 0 | 22 | 19 | 1647 | 13 | NON_ANCHOR |
| 1409 | 13 | 13 | 0 | 3 | 2 | 1317 | 6 | NON_ANCHOR |
| 8 | 112 | 112 | 0 | 22 | 22 | 642 | 6 | NON_ANCHOR |
| 111 | 588 | 588 | 0 | 22 | 15 | 426 | 5 | NON_ANCHOR |
| 1191 | 35 | 35 | 0 | 22 | 22 | 816 | 5 | NON_ANCHOR |
| 1425 | 22 | 22 | 0 | 22 | 22 | 888 | 5 | ANCHOR |
| 1517 | 29 | 29 | 0 | 22 | 22 | 1557 | 5 | NON_ANCHOR |

At the bottom right, a "Results" window shows a table of query results:

| ID | JobID | Time | Tool | Status |
|----|-------------|-------------|----------|---------|
| 1 | e69d32cf... | 2011-07-... | ClustalW | OK |
| 2 | 927e2c4f... | 2011-07-... | NCBI API | OK |
| 3 | ee8f7de8... | 2011-07-... | NCBI API | WORKING |

The results show that the TIS which correspond to the longest gene is most often annotated in other strains. It suggest that this is the correct one.

The screenshot displays the CamberVis software interface. The main window shows a genomic map with several tracks for different strains (main: 04-02981, COL, ED133, ED98, JH1, JH9, JKD6008, JKD6159). A MultigeneZoom window is open, showing the Multigene ID: SA2981_2478 and its lengths: 828,864,879,888*,885. The Locus is (2629361,2628474). The NCBI API task results window is also open, showing the task ID: ee8f7de8-0283-443e-b95b-c16d056f2f2b. The results table is as follows:

| ID | Target name | E-value | Aligned TISs |
|----|-------------------------------------|---------|--------------|
| 1 | >ref NP_373066.1 transcription ... | 0.0 | YES |
| 2 | >ref ZP_06334661.1 LysR family... | 0.0 | YES |
| 3 | >ref ZP_05687795.1 LysR family... | 0.0 | YES |
| 4 | >ref ZP_05610994.1 transcriptio... | 0.0 | YES |
| 5 | >ref YP_187348.1 LysR family tr... | 0.0 | NO |
| 6 | >ref YP_041967.1 LysR family r... | 0.0 | NO |
| 7 | >gb ADL66578.1 LysR family tra... | 0.0 | NO |
| 8 | >ref NP_647280.1 hypothetical ... | 0.0 | NO |
| 9 | >ref ZP_03612494.1 transcriptio... | 2e-157 | NO |
| 10 | >gb EFV88474.1 bacterial regula... | 2e-157 | NO |
| 11 | >ref ZP_04798125.1 LysR family... | 4e-157 | NO |
| 12 | >ref ZP_04817760.1 LysR family... | 6e-157 | NO |
| 13 | >ref ZP_07841950.1 transcriptio... | 1e-156 | NO |
| 14 | >ref NP_765661.1 transcription ... | 2e-155 | NO |
| 15 | >ref ZP_04676835.1 transcriptio... | 5e-149 | NO |

The ComponentsStats window shows the following data:

| cc id | # of multigenes | # of main multigenes | # of plasmid multigenes | # of strains | # of ann strains | max length | max # of TIS | cc type |
|-------|-----------------|----------------------|-------------------------|--------------|------------------|------------|--------------|------------|
| 552 | 177 | 177 | 0 | 22 | 19 | 1647 | 13 | NON_ANCHOR |
| 1409 | 13 | 13 | 0 | 3 | 2 | 1317 | 6 | NON_ANCHOR |
| 8 | 112 | 112 | 0 | 22 | 22 | 642 | 6 | NON_ANCHOR |
| 111 | 588 | 588 | 0 | 22 | 15 | 426 | 5 | NON_ANCHOR |
| 1191 | 35 | 35 | 0 | 22 | 22 | 816 | 5 | NON_ANCHOR |
| 1425 | 22 | 22 | 0 | 22 | 22 | 888 | 5 | ANCHOR |
| 1517 | 29 | 29 | 0 | 22 | 22 | 1557 | 5 | NON_ANCHOR |

The TasksTable window shows the following data:

| ID | JobID | Time | Tool | Status |
|----|-------------|-------------|----------|--------|
| 1 | e69d32cf... | 2011-07-... | ClustalW | OK |
| 2 | 927e2c4f... | 2011-07-... | NCBI API | OK |
| 3 | ee8f7de8... | 2011-07-... | NCBI API | OK |
| 4 | 2a11668... | 2011-07-... | NCBI API | OK |