CrusView Manual
CrusView is a tool for karyotype/genome visualization and comparison of crucifer species. It also provides functions to import new genomes.

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Requirement</td>
<td>3</td>
</tr>
<tr>
<td>Installation</td>
<td>3</td>
</tr>
<tr>
<td>General Installation Procedure</td>
<td>3</td>
</tr>
<tr>
<td>Further Installation Note for Linux Machine</td>
<td>3</td>
</tr>
<tr>
<td>Further Installation Note for Windows Machine</td>
<td>3</td>
</tr>
<tr>
<td>Memory Usage</td>
<td>4</td>
</tr>
<tr>
<td>Basic Data Format</td>
<td>4</td>
</tr>
<tr>
<td>Name Criteria</td>
<td>4</td>
</tr>
<tr>
<td>FASTA Format</td>
<td>4</td>
</tr>
<tr>
<td>GFF Format</td>
<td>4</td>
</tr>
<tr>
<td>BED format</td>
<td>4</td>
</tr>
<tr>
<td>Additional File Formats</td>
<td>5</td>
</tr>
<tr>
<td>Karyotype Format</td>
<td>6</td>
</tr>
<tr>
<td>List Format</td>
<td>6</td>
</tr>
<tr>
<td>Gmm Format</td>
<td>6</td>
</tr>
<tr>
<td>Track Format</td>
<td>6</td>
</tr>
<tr>
<td>Sample Data</td>
<td>6</td>
</tr>
<tr>
<td>Usage</td>
<td>6</td>
</tr>
<tr>
<td>Main Window</td>
<td>6</td>
</tr>
<tr>
<td>Switching between genomes and karyotypes</td>
<td>7</td>
</tr>
<tr>
<td>CNV and PAV detection</td>
<td>7</td>
</tr>
<tr>
<td>Mark Genes on the Two Genomes</td>
<td>8</td>
</tr>
<tr>
<td>Generating Customized Gene List</td>
<td>8</td>
</tr>
<tr>
<td>Comparing Window</td>
<td>8</td>
</tr>
<tr>
<td>Mark gene groups on the dotplot</td>
<td>9</td>
</tr>
<tr>
<td>Changing Gene Identity Cut-off</td>
<td>10</td>
</tr>
<tr>
<td>Analyzing Gene Subsets</td>
<td>10</td>
</tr>
<tr>
<td>Saving the selected gene Groups</td>
<td>10</td>
</tr>
<tr>
<td>Tandem duplicated gene prediction</td>
<td>10</td>
</tr>
<tr>
<td>Assembling Window</td>
<td>11</td>
</tr>
<tr>
<td>Basic operations</td>
<td>11</td>
</tr>
<tr>
<td>Functions</td>
<td>12</td>
</tr>
<tr>
<td>Exon-to-exon alignment plot</td>
<td>12</td>
</tr>
<tr>
<td>Single gene alignment</td>
<td>13</td>
</tr>
<tr>
<td>Multiple gene alignment</td>
<td>14</td>
</tr>
<tr>
<td>Import New Genome</td>
<td>15</td>
</tr>
<tr>
<td>Required Files</td>
<td>15</td>
</tr>
<tr>
<td>Assembled Genome</td>
<td>15</td>
</tr>
</tbody>
</table>
List of Figures

1. Main window .................................................. 7
2. CNV detection table ........................................... 8
3. Compare window ............................................... 9
4. Assembling window ............................................ 11
5. Single Gene Alignment Panel ................................. 13
7. Multiple Gene Alignment Panel ............................... 14
8. Multiple Gene Alignment Result .............................. 14
9. Diagram for Genome Importing Algorithm .................. 16
10. Information from Karyotype ................................ 27
11. Saving Current Scaffolds’ Arrangement into Main Screen 27
12. Information from the genes’ block IDs ....................... 28
13. Synteny Plot between pseudo-chr4 of Eutrema and chr2 of Arabidopsis. The purple lines indicate that the orthologs of the genes at the beginning of left chromosome are at around 11 Mb of the right chromosome ........................................... 28
14. Dotplot for pseudo-chr4 of Eutrema and chr2 of Arabidopsis. Gene groups OG5_212231, OG5_212230 and OG5_135168 are marked in red. Their members in Arabidopsis are at around 11 Mb .................................. 29
15. The input files and parameter settings for importing the Scaffolds of Eutrema into CrusView ................................................................. 30
16. The imported Eutrema’s “Genome” on the main window .... 31
17. Lyrata’s karyotype and shared syntenic region between scaffold 5 of Eutrema and chromosome 1 of Lyrata ....................................................... 31
18. The dotplot between chromosome 1, 3, 4 and 7 of Eutrema and Lyrata ................................................................. 32

List of Tables

1. System Requirement ............................................. 3
2. Description of BED Format ..................................... 5
3. 25 Predefined Block Colors ..................................... 5
4. Files Required when Importing Assembled Genome ............ 15
5. Homolog Pair File Description ................................ 15
6. The Location of three ortholog groups on chr4 of Eutrema and chr2 of Arabidopsis .............................. 29
### System Requirement

Currently, CrusView supports both Linux and Windows OS with a 1.6 or higher version Java virtual machine. We highly recommend user to run this application with at least 2Gb of memory. The minimum requirement provides the possibility of using CrusView in a relatively older machine. However, the user should avoid of using memory intensive functions such as importing genome sequences and generating multiple genes alignments. Under the minimum memory requirement, these functions may end without proper notice, which may result in unexpected result later.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU number</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Memory</td>
<td>512 Mb</td>
<td>2 Gb or more</td>
</tr>
<tr>
<td>Hard Disk</td>
<td>1Gb</td>
<td>50Gb</td>
</tr>
</tbody>
</table>

Tab. 1: System Requirement

### Installation

The installation procedures are similar in different operations systems. However, user should notice the differences between Java virtual machines in different systems.

#### General Installation Procedure

CrusView is available as a Java web start application, which can be found in the following address.

http://www.cmbb.arizona.edu/crusview

The installation procedure will start automatically when running the program for the first time. If user promote the installation, a directory called CrusView should be created in either user’s home directory (Linux) or under C:\ (Windows).

#### Further Installation Note for Linux Machine

Some Linux operation systems use openJDK as default Java virtual machine, which limits the memory allocation. However, CrusView generally requires more memory than openJDK’s upper limit and CrusView may not run properly with less than 512 Mb of memory. We recommend users to use Sun Java 1.7 environment in Linux\(^1\). please read the Ubuntu help wiki for more information\(^2\).

#### Further Installation Note for Windows Machine

The amount of memory allowed to be allocated on Windows machine actually varies, depending on either the 32bit or the 64bit version Java virtual machine was installed. If it is the 32bit version, the maximum memory could be allocated is around 1200 Mb. Any attempts of launching CrusView with more than 1200 Mb of memory will eventually fail. Additionally, user should notice that some 64bit Windows Operation System may still have 32bit Java virtual machine installed.

Memory Usage

Most basic functions of CrusView runs fine under its minimum memory requirement, the only difference should just be how smooth it is. However, some memory intensive tasks are not in this scenario, such as loading a whole genome into CrusView’s internal database and generating large exon-to-exon alignment plot. For the former, the required memory depends on the length of the longest chromosome. For example, 512 Mb is sufficient for loading a 35 Mb chromosome. The latter varies on the matrix used in sequence alignment (longest exon length in first/up DNA segment) * (the length of second/down DNA segment).

Basic Data Format

The BED, GFF/GTF and FASTA format files are the tree basic file for a genome, which is important to CrusView’s functions. Additionally, to cope with the functions in CrusView, some customized file formats are used as well. User could find sample files in the CrusView Folder.

Name Criteria

To cope with CrusView’s internal database, a same base name is required for a genome’s BED, GFF/GTF and FASTA files. For instance, file Arabidopsis_Thaliana.bed, Arabidopsis_Thaliana.gff and Arabidopsis_Thaliana.fa can be used for species Arabidopsis Thaliana. Additionally, user should pay attention to the consistency of the IDs used in the three files, including both gene IDs and scaffold/contig IDs.

FASTA Format

Fasta format file is supported by CrusView. The whole string after > sign is used as the sequence ID.

GFF Format

GFF is supported by CrusView. However, duo to the various types of GFF/GTF files, user should provide the correct feature words for defining mRNA and exon regions when importing the file into CrusView’s internal database, which can be found at the third column of the GFF/GTF files.

BED format

CrusView uses a customized BED format file as its basic input, which can be loaded by function “Import bed/karyotype File” in the File menu. It is consist of ten columns, as described in table 1. Specifically, column 8 and 9 refers to the 24 synten blocks for Brassicaceae species\(^3\), which can be filled by 0 if not known. Given a block ID (alphabetically from A to X, and 0), the color for the genes will be interpreted by CrusView with the relationship shown in table 2. If the predefined 24 blocks and colors are not sufficient for a genome, customized gene color can be defined in column 9 (block ID) by filling a HEX color code directly, such as #FF0000 for red.

\(^3\)Detailed information of these blocks can be found in paper “The ABC’s of comparative genomics in the Brassicaceae: building blocks of crucifer genomes”. M. Eric Schranz, Martin A. Lysak, Thomas Mitchell-Olids, 2006.
Additional File Formats

Several other formats are used in CrusView, which are karyotype, list, gmm and track format. Sample files in these format can be found in CrusView’s sample data.

<table>
<thead>
<tr>
<th>Column Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chromosome ID</td>
</tr>
<tr>
<td>2</td>
<td>start position</td>
</tr>
<tr>
<td>3</td>
<td>end position</td>
</tr>
<tr>
<td>4</td>
<td>chromosome size</td>
</tr>
<tr>
<td>5</td>
<td>homolog/group ID</td>
</tr>
<tr>
<td>6</td>
<td>gene ID</td>
</tr>
<tr>
<td>7</td>
<td>identity cut-off</td>
</tr>
<tr>
<td>8</td>
<td>block number</td>
</tr>
<tr>
<td>9</td>
<td>block ID</td>
</tr>
<tr>
<td>10</td>
<td>gene description</td>
</tr>
</tbody>
</table>

Tab. 2: Description of BED Format

<table>
<thead>
<tr>
<th>Block ID</th>
<th>Hex Color Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>#dedede</td>
</tr>
<tr>
<td>A</td>
<td>#f4ea00</td>
</tr>
<tr>
<td>B</td>
<td>#f4ea00</td>
</tr>
<tr>
<td>C</td>
<td>#f4ea00</td>
</tr>
<tr>
<td>D</td>
<td>#e4291f</td>
</tr>
<tr>
<td>E</td>
<td>#e4291f</td>
</tr>
<tr>
<td>F</td>
<td>#32489b</td>
</tr>
<tr>
<td>G</td>
<td>#32489b</td>
</tr>
<tr>
<td>H</td>
<td>#32489b</td>
</tr>
<tr>
<td>I</td>
<td>#980d7a</td>
</tr>
<tr>
<td>J</td>
<td>#980d7a</td>
</tr>
<tr>
<td>K</td>
<td>#f6a800</td>
</tr>
<tr>
<td>L</td>
<td>#f6a800</td>
</tr>
<tr>
<td>M</td>
<td>#f6a800</td>
</tr>
<tr>
<td>N</td>
<td>#f6a800</td>
</tr>
<tr>
<td>O</td>
<td>#008040</td>
</tr>
<tr>
<td>P</td>
<td>#008040</td>
</tr>
<tr>
<td>Q</td>
<td>#008040</td>
</tr>
<tr>
<td>R</td>
<td>#008040</td>
</tr>
<tr>
<td>S</td>
<td>#72c9e9</td>
</tr>
<tr>
<td>T</td>
<td>#72c9e9</td>
</tr>
<tr>
<td>U</td>
<td>#72c9e9</td>
</tr>
<tr>
<td>V</td>
<td>#f6c2d1</td>
</tr>
<tr>
<td>W</td>
<td>#f6c2d1</td>
</tr>
<tr>
<td>X</td>
<td>#f6c2d1</td>
</tr>
</tbody>
</table>

Tab. 3: 25 Predefined Block Colors
**Karyotype Format**

Karyotype format files are used for visualization and as an input for KGBassembler. Each line of this file starts with a chromosome ID, followed by a set of tab delimited block IDs and their relevant orientation.

**List Format**

It is simply a list of gene IDs (one ID in each line) for the filtering function in comparing window. It can be generated by function “Add Current Table to CompareMode Filter” in Edit menu or manually prepared by the user.

**Gmm Format**

GMM file contains a set of gene IDs and their annotations, which are used as gene markers when modifying the assembling result.

**Track Format**

Track format file is similar with BED format files, which is consist of three columns, in terms of segment, start and end. It is used for adding an additional track when modifying the assembling result. For example, it can denote the position of centromeres or telomeres.

**Sample Data**

Sample data can be downloaded through CrusView’s internal function in File menu. Which includes Assembly Data (For running through the Assembly procedure), Additional Assembly Data (For generating BED/GFF/FASTA file according to the assembly result), Genome Data (GFF/GTF file and FASTA file for two sample species), External Genome (Raw data from one genome, which could be used for generating new BED file used in CrusView).

**Usage**

Generally speaking, CrusView provides functions for visualizing and comparing genomes/karyotypes in different scales and functions to importing new genomes into CrusView. In this section, we introduce the functions in the main window and comparing window, followed by functions for generating exon-to-exon alignment plots, confirming to the scale from genome to gene level. Then, we introduce the way to import new genome into CrusView, regardless it is assembled or partially assembled.

**Main Window**

The main window pops up when CrusView starts. As shown in figure 1, it is consisted of five parts: 1) the trees of genomes/karyotypes at the top left; 2) the currently focus query genome in the middle; 3) the detected blocks from the query genome at the top right; 4) the reference genome under these blocks’ legends; 5) The information table at the bottom.
Fig. 1: Main window

Besides the supporting functions in CrusView’s menu, the main window is typically designed for fast browsing a set of genomes/karyotypes, detecting copy number variation (CNV) and present/absent variation (PAV) between genomes and investigating a subset of genes. These functions are further described below.

**Switching between genomes and karyotypes**

- The trees of genomes and karyotypes
  - Trees of genome/karyotypes provides the way to switch the focused query genomes/karyotypes. The two trees can be switched by clicking the button at the very top. Left clicking on one genome changes the currently focused query genome and the relative information table. Left clicking on one karyotype open a new popup window showing the karyotype of the selected species.

- The reference genome
  - The reference genome can be changed to the current focused query genome by clicking the “clip” icon on the tool bar or selecting function “Update Right Screen” in Edit menu.

**CNV and PAV detection**

These two functions can be found in Analysis menu. The functions will scan through all genes in the currently focused two genomes and the result will be shown as a table. This table can be sorted by
clicking its header. And a right click on a row would trigger an additional popup menu, from which user can search the groupID from external links. Please notice that this external link can be changed in CrusView’s settings.

![image](image.png)

**Fig. 2: CNV detection table**

Mark Genes on the Two Genomes

A gene or a group of genes can be marked on the two genomes. Firstly, user need to find the interesting gene/group at the information table. Then, user can right-click on the corresponding row, a popup window should appear afterwards. If function “Mark Gene” is selected, this gene’s position on the two currently focused genome will be pointed with a black arrow. If function “Mark Gene Group” is selected, all genes within this particular ortholog group will be marked on the two genomes.

Generating Customized Gene List

A subset of gene IDs in the information table can be further investigated in both main window and comparing window. User need to type in one string (transporter, e.g.) in the text box on the tool bar. After clicking the search button, only rows in the information table containing the exact string in the text box remain. User can then investigate this subset of genes in the main window, or save the genes’ IDs via function “Add Current Table to CompareMode Filter” in the Edit menu. This list of IDs should be saved in “list” folder and also appear in Compare Mode’s filter afterwards.

Comparing Window

The comparing window opens if user clicks “Compare Selected Genomes” icon on the tool bar or select the same function in Analysis Menu. The currently focused two genomes will be shown in the window. As shown in figure 3, it generates synteny plot and dotplot for a pair of chromosomes.
The first operation required is to select one chromosome from each species. The synteny plot and dotplot should appear promptly. Homologs between the two chromosomes can be selected on the plots by dragging the mouse with left-button. A group ID table should appear afterwards. Then, a set of further analysis can be performed on the selected gene group, including marking them on the dotplot, searching external websites, retrieving their sample sequences, generating exon-to-exon alignment plot and detecting all homologs of this gene group. Four additional functions are also provided in this window, which are changing gene identity cut-off for the plots, working with gene filter set (only those genes in the list could be selected), saving currently selected gene list and predicting tandem duplicated genes.

**Mark gene groups on the dotplot**

- Marking gene groups
  - Dots (gene pairs) are automatically marked if their group IDs are selected in the table. Also, a gene group can be marked via a function in the popup menu of the table (right click).

- Marking multiple gene groups with different colors
  - After marking a gene group on the dotplot, user could fix the current gene markers and then change the color used for marking genes by “Settings” in File menu. The second gene
Comparing Window

• Cleaning all marks
  – All marks, including those generated by tandem duplication prediction, can be removed by function “Clean Marker” in the dotplot’s pop up menu (right click on the dotplot).

Changing Gene Identity Cut-off

The gene identity cut-off used for generating plots can be changed by clicking the “cut” icon on the tool bar. The default value is 60%.

Analyzing Gene Subsets

Gene subsets can be selected on the tool bar, if list format files are in “list” directory before this compare window launched. By default, all dots (gene pairs) can be selected on the dotplot. After defining a gene subset, only gene pairs within this subset can be selected.

Saving the selected gene Groups

The currently selected gene groups can be saved by clicking “(Save Selected Gene Groups)” icon on the tool bar.

Tandem duplicated gene prediction

A simple function for tandem duplication prediction is provided in dotplot’s popup menu (right click on the dotplot). It assumes that gene homologs within a certain distance are tandem duplicated (5 Kb by default). If predicted as tandem duplicated, the color of a dot (gene pair) will be changed to a predefined color, which is green by default.
Assembling Window

The assembling window opens after successfully running through function “Run KGBassembler” or “Load KGBassembler’s result” in the Analysis menu. As shown in figure 4, it shows the reference karyotype used for assembling and KGBassemblers’ result. Auto-scaled gene density is illustrated along the scaffolds/contigs. Also, additional information can be added to guide the curating procedure, such as gene markers and position of centromere on the scaffolds/contigs.

![Assembling window diagram](image)

**Fig. 4: Assembling window**

**Basic operations**

- Changing scaffold’s position
  - Firstly put your mouse on top of a scaffold and hold mouse’s left button. Then, move the pointer to one of the empty rectangles between scaffolds, its color should be changed to green automatically. Then, release the mouse’s left button to put the scaffold at this new position.

- Changing scaffold’s orientation

---

4 the guide for preparing files and choosing parameters involved in this step can be found at KGBassembler’s website: [http://www.cmbb.arizona.edu/KGBassembler/](http://www.cmbb.arizona.edu/KGBassembler/)
- This operation is similar with the previous one. But before releasing the mouse’s left button, user should press “shift” key at first, after which the rectangle’s color should be changed to red, indicating that the orientation of the scaffold/contig has been changed.

Functions

• min Block Length
  - This function defines the minimum length for scaffolds/contigs to be shown on the plot. It is typically useful for editing an assembly with a lot of small scaffolds/contigs. The default value is 200 kb.

• Push to Main Screen
  - This function save the current arrangement of the scaffolds into a BED file, and then add this BED file to the genome tree in the main window. There are two options available, one is to save a simple BED file, which need no additional information. However, this BED file doesn’t contain any ortholog groups or alignment identity to the genes on reference genome, which limits its usage in CrusView. The other option provides a full BED file with all information needed for CrusView, but it asks for additional input files such as the result of orthoMCL, identity for protein pair, the gene feature file (GFF) for the original scaffolds and etc. A set of sample files can be downloaded as additional assembly data in the File menu.

• Add Marker
  - This function adds a set of markers (GMM format file) to the plot, which can be genetic markers. It asks for the lines of markers that user would like to use, which actually depends on the marker’s density. These markers will then be shown as gray/red sticks below the scaffolds, together with annotations under the sticks. If the color for one marker is gray, its annotation should be found directly under it. If the color is red, the annotation’s place has been changed due to space limitation on the plot.

• Add Track
  - This function adds one track file to the assembly window, which can be position of centromeres. If added, this track is plotted between the gene density plot and the scaffolds/contigs.

• Save assembly
  - Provided the scaffolds’ original GFF file and FASTA file, the new files confirming to the arrangement of the scaffolds can be save under CrusView’s root folder with this function.

Exon-to-exon alignment plot

After loading correct GFF and FASTA files into CrusView’s internal database, these plots can be generated through functions “Single gene alignment plot” and “Multiple genes alignment plot” in the Analysis Menu, or the similar functions integrated into the compare window. Both functions will extract exon sequences from the upper segment, and then align them to the second sequence, regardless its annotated exon location.
Single gene alignment

As shown in figure 5, given a gene’s ID, this function automatically detects its orthologs in the other species and generates exon-to-exon alignment plot upon user’s selection.

(a) Search with a gene ID

(b) Select one of the detected homologs

Fig. 5: Single Gene Alignment Panel

The result is shown in a new window, and the detail alignment information can be found in the text window below. The plot can be saved with the function provided in File menu. However, the detailed alignment information could only be saved manually by copy and paste them to another word editor.

Fig. 6: Single Gene Alignment Result
**Multiple gene alignment**

This function firstly extracts chromosome level information from CrusView’s inner database, and then asks the user to define two regions for alignment. Additional features such as whether adding gene ID to the plot, the size of the plot and the length of the scale bar are also provided.

![Multiple Gene Alignment Panel](image)

**Fig. 7: Multiple Gene Alignment Panel**

The result is shown in the figure below. The gene ID is left justified with its location on the sequence segment. If the gene IDs’ locations are overlapping, the function will try to plot the second ID on the top of the first one. The maximum IDs allowed for overlapping is four. If user would like to align two large pieces of segments with a bunch of IDs, the user may like to use a larger plot length.

![Multiple Gene Alignment Result](image)

**Fig. 8: Multiple Gene Alignment Result**
Import New Genome

CrusView provides the functions to import a new genome, no matter if the chromosomes are assembled or not. However, one additional karyotype file is needed for the partially assembled genomes (scaffold/contig level).

Required Files

Several files are needed for generating the proper Bed file for CrusView, which is shown in the table below.

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFF/GTF</td>
<td>The gene feature file</td>
</tr>
<tr>
<td>OrthoMCL Folder</td>
<td>The output from OrthoMCL website</td>
</tr>
<tr>
<td>ID mapping File (txt format)</td>
<td>file for mapping gene IDs (reference genome related)</td>
</tr>
<tr>
<td>Homolog Pair File (txt format)</td>
<td>summary from BLAST result (reference genome related)</td>
</tr>
<tr>
<td>Reference Block File (txt format)</td>
<td>the assigned block ID for each gene in reference genome</td>
</tr>
<tr>
<td>Reference Group File (txt format)</td>
<td>the group ID for each gene in reference genome</td>
</tr>
</tbody>
</table>

Tab. 4: Files Required when Importing Assembled Genome

The GFF/GTF file could usually be found in the release of a new genome and the OrthoMCL result can be obtained by submitting its protein sequences to OrthoMCL’s website\(^5\). As Arabidopsis Thaliana is used for the reference genome in CrusView, the required ID mapping file and reference block/group files are prepared and ready to download as sample data. The only file required to be generated by user is the homolog pair file, which can be extracted from BLAST result. We are happy to provide the Perl script for extracting the result from BLAST’s alignment result upon request. The description of this file is provided in the table below.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>query gene ID</td>
</tr>
<tr>
<td>2</td>
<td>subject gene ID</td>
</tr>
<tr>
<td>3</td>
<td>query gene identity</td>
</tr>
<tr>
<td>4</td>
<td>subject gene identity</td>
</tr>
<tr>
<td>5</td>
<td>query gene length</td>
</tr>
<tr>
<td>6</td>
<td>subject gene length</td>
</tr>
</tbody>
</table>

Tab. 5: Homolog Pair File Description

Assembled Genome

With the files mentioned, we could generate the BED file for CrusView directly. However, several parameters need to be defined for this procedure, which are minimum sequences length, block inferring ratio, block masking ratio, window size, step size and the keyword for genes in the GFF/GTF file. The minimum length is for removing short scaffolds, which is set to 1 Mb by default.

\(^5\) [http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=proteomeUploadForm](http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=proteomeUploadForm)
The block inferring ratio is the first cut-off for assigning block IDs to genes on a sequence segment. For a segment with 10 genes (so the window size is 10), a 0.5 ratio would pass these genes to the second cut-off, if 5 or more of them contain the same block ID. The block masking ratio is the second cut-off for assigning block IDs. Given the information that each gene on the segment contribute one block ID, the cut-off is defined as the ratio between the largest number of block IDs and the second largest. If both cut-offs are satisfied, the genes are assigned with the new block ID. The window size defines the length of the segment by number of genes on it. And the step size defines how the window slides on the scaffold/chromosome. Step size 1 would cause the window slides over one gene at a time.
Partially Assembled Genome

The partially assembled genome can be imported into CrusView. However, one additional karyotype file regarding the block arrangement on the chromosomes is required. The input files for running KGBassembler are need to be prepared as well. Please read KGBassembler’s manual for instructions\(^6\). CrusView calls KGBassembler internally and provides the interface for manually modification of the predicted arrangement afterwards. In the assembly window, one function is provided to generate the BED file accordingly, and one additional function will help user to convert the genome sequence and GFF/GTF file accordingly.

---

Function List

In this section, we will go through all functions provided by CrusView.

Main Window

The functions in File, Edit and Analysis Menu, and the functions in tool bar and information table are described below.

File Menu

- Update Trees
  - Scan the files in “bed” and “karyotype” folder, and update the trees accordingly.

- Delete Currently Focused Genome
  - Delete the currently focus query genome and update the tree.

- Download Sample Data
  - Download the sample data from CrusView’s website.

- Import Gene Feature File
  - import either GFF or GTF file into CrusView’s internal database, the keywords for gene and exon in the files need to be provided.

- Import Genome Sequence
  - import genome sequence into CrusView’s internal database.

- Import Bed/Karyotype File
  - Copy Bed/Karyotype file into their relative folder and update the tree.

- Import Gene Description
  - import gene annotation/description to the currently focused query genome. The temporarily imported data won’t be written to the Bed file, while the permanently imported data will be.

- Import None Gene Features
  - This Function would import a set of non-gene features, such as genome breakpoint into CrusView. The input file should be consist of four columns, which are the scaffold/chromosome ID, start position, end position and corresponding description. Again, the temporarily imported data won’t be written to the Bed file, while the permanently imported data will be.

- Generate New Bed File
  - This function would generate the Bed file for CrusView, with the information from GFF/GTF file, reference block/group file, homolog pairs file and etc. Please read the Import New Genome section for details.
• Save
  – Save the plots in PNG/SVG/PDF format.

• Settings
  – Modify a bunch of settings in CrusView, such as the identity cut-off used for generating exon-to-exon alignment plots, the external links defined in CrusView and the colors for plots in CrusView.

• Exit
  – Exit CrusView.

**Edit Menu**

• Show Table in Database
  – show all the tables in CrusView’s internal database.

• Sort Current Table by Position
  – Sort current table by their position. The scaffold/chromosome ID and the gene starting position will be sorted with ascending order.

• Update Right Screen
  – Update the reference genome with the currently focused query genome.

• Clear Markers
  – Clean all markers painted on the query and reference genomes

• Main Screen [+]
  – Zoom in for the currently focused query genome

• Main Screen [-]
  – Zoom out for the currently focused query genome

• Right Screen [+]
  – Zoom in for the reference genome

• Right Screen [-]
  – Zoom out for the reference genome
Analysis Menu

- Add Current Table to CompareMode filter
  - Save the list of gene IDs currently in the table to a file in “list” directory.

- Compare Current Two Genomes
  - launch the compare window with the currently focused query genome and reference genome

- Calculate Copy Number Variation
  - Provided group ID for each gene in the currently focused genomes, this simple function calculate the copy number changes of a gene group. The result table can be sorted by clicking on its header.

- Calculate Present/Absent Variation
  - Provided group ID for each gene in the currently focused genomes, this simple function search for genes that do not show up in the other genome by sharing a same group ID.

- Single Gene Alignment Plot
  - Generate exon-to-exon gene alignment for a pair of genes. The detail alignment information can be displayed in a text window.

- Multiple Gene Alignment Plot
  - Generate exon-to-exon gene alignment for two user defined segments. User should be careful when setting the length of the second segment, as the length of the longest exon on the first segment and the length of the second segment determines the size of the matrix. Larger plots generally requires more memory.

- Run KGBassembler
  - Run KGBassembler with required files. The way to prepare the files can be found in KGBassembler’s manual. The assembly window will show up afterwards.

- Load KGBassembler’s result
  - The function loads KGBassembler’s result directly and the assembly window will show up afterwards.

Tool Bar

- Update Right Screen
  - Update the reference genome with the currently focused query genome.

- Zoom in Main Screen
  - Zoom in for the currently focused query genome
Function List

• Zoom out Main Screen
  – Zoom out for the currently focused query genome

• Clear Markers
  – Clean all markers painted on the query and reference genomes

• Compare Current Two Genomes
  – Launch the compare window with the currently focused query genome and reference genome

• Down Right Screen Zoom in
  – Zoom in for the reference genome

• Down Right Screen Zoom out
  – Zoom out for the reference genome

• Keyword Search
  – Filtering the current table with one keyword. Only genes with this keyword remain.

Functions in Information Table

• Mark Gene
  – Mark selected gene on the query genome, and one of its homologs in the reference genome.

• Mark Gene Group
  – Mark all genes belong to this gene group on the both genomes.

• External Link For Gene
  – Query the gene ID to one external website. One Internet browser would be launched automatically.

• External Link For Group
  – Query the group ID to one external website. One Internet browser would be launched automatically.

• Reset Table
  – Clean all filtering made on the table, including the effect of keyword search and marking gene group.
**Compare Window**

- Zoom in Segment Plot
  - Zoom in Segment Plot

- Zoom out Segment Plot
  - Zoom out Segment Plot

- Dotplot Zoom In
  - Zoom In Dotplot

- Dotplot Zoom Out
  - Zoom Out Dotplot

- Save Selected Gene
  - Save currently selected genes into a txt file

- Change Identity Cut-off
  - Change the identity cut-off used for generating plots. The default value is 60%

- Filter Gene Set
  - Select filter gene set. Only genes in the selected list can be selected on the plots. These list can be generated through the function in the main window, or directly copy the list format files in the ‘list’ folder.

- Keep Markers
  - Keep currently painted markers. The markers remain painted even if they are not selected on the dotplot.

**Right Click on Dotplot**

- Clean Marker
  - Clean all markers on the dotplot.

- Predict Tandem Duplicated Genes
  - Given a distance cut-off between two homologs, the homologs locate within this cut-off will be marked as tandem duplicated genes. The default color for painting is green.

- Multiple Gene Alignment
  - Given a selected area, the Multiple Gene Alignment Panel (Figure 7.) would appear accordingly, which can generate gene alignment plot. This function requires the support of the internal database, where the genome sequence is imported.
Function List

Function in Gene Table

- Right Click
  - Mark in Dotplot
    - Mark the gene group in Dotplot
  - Link for Gene
    - If the gene group ID is also an gene ID, open external link for this gene ID.
  - Link for Group
    - Open external link for this group ID.
  - Show Fasta File
    - Show two sample sequences for this gene group. One from each species. This function requires the support of the internal database, where the genome sequence is imported.
  - Gene Comparison Plot
    - It is a similar to the function used for single gene alignment plot in main window. The difference is that user could select the two genes from the provided list.
  - Homologs
    - List all genes in one gene group.

- Single Left Click
  - Mark the gene group in Dotplot

- Drag and Select Multiple gene Groups
  - Mark the gene group in Dotplot

- Double Left click on One Gene Group
  - Quick sample single gene alignment plot between two genes from the two species.

Assembly Window

- Zoom [+]
  - Zoom In the Curation Window

- Zoom [-]
  - Zoom out the Curation Window

- minBlockLength
  - Set the minimum length for a scaffold/contig to show up in the window.

- Push To Main Screen
  - Save current arrangement of scaffolds as one Bed file. Two options are available.

- Add Marker
- Add makers for genes on the scaffolds/contigs.

- Add Track
  - Add one additional track under the scaffolds/contigs, in the form of little black bars.

- Save Assembly
  - Provided the original GFF and FASTA format files for the scaffolds/contigs, save the new GFF and FASTA format files according to the current scaffold/contig arrangement.
Supporting Scripts

Parser for Tandem Repeat Finder’s Result

The parser for converting the result TRF (tandem repeat finder) is provided in CrusView. Please visit our website for downloading the script and the sample case. The input files including a .dat file from TRF and a .ini file prepared by the user. A sample .ini file and its description is provided below.

```
[genome]
chr1 = scaffold_1+
chr2 = scaffold_2+
chr3 = scaffold_3+
chr4 = scaffold_4+
chr5 = scaffold_5+
chr6 = scaffold_6+
chr7 = scaffold_7+
chr8 = scaffold_8+
[scaf]
scaffold_5 = 21221946
scaffold_6 = 25113588
scaffold_3 = 24464547
scaffold_4 = 23328337
scaffold_1 = 33132539
scaffold_7 = 24649197
scaffold_2 = 19320864
scaffold_8 = 22951293
[parameter]
gap = 100
base = LyrataTRF
[trf]
path = Alyrata_107.fa.2.7.7.80.10.50.500.dat
```

- The “genome” section contains the scaffolds composition for each chromosome. The “+/-” sign at the end of each scaffold indicate the scaffold’s orientation. Multiple scaffolds can be tied together by “,” sign.
- The scaf section contains the length of each scaffold.
- The parameter section defines the gap length defined between scaffolds, if one chromosome contains multiple scaffolds. It also defines the base name for “group ID” and “geneID” columns in the result BED file. User should avoid from using simple base names, as same “groupID” among multiple genomes may be a problem.
- The trf section defines the path to the TRF’s output file.

To run the script, user can type the following command in a terminal.

```perl
crusView_trf_paser.pl crusView_trf_paser.ini >result.bed
```
Question & Answers

1. The external links are wrong and misleading. Is there something wrong with CrusView?

There are two links for a certain ID, which is for searching this ID from two websites. Typically one for gene, and the other for the ortholog group which this gene belongs to. It is user’s responsibility to use the correct function. And these external websites can be changed in CrusView. If user would like to use two gene’s websites, he or she could modify the settings. The function name “External link for group ID” would be weirded thereafter, though.

2. I would like to change the external websites defined in CrusView. What should I do?

The external links are customizable via “Settings” function in File menu. For example, if user would like to change the external link for gene ID to TAIR website, he or she could change the settings with the following steps.

(a) search a gene ID via the website directly. After seeing the result, a link can be fetched from the browser, such as the following one.

```
http://www.arabidopsis.org/servlets/Search?type=general
&search_action=detail&method=1&show_obsolete=F&name=AT5G01010
&sub_type=gene&SEARCH_EXACT=4&SEARCH_CONTAINS=1
```

(b) Find the gene ID that was used, which is AT5G01010 in this example. We separate this link into two parts, which are called head and tail in CrusView.

**Head:**
```
http://www.arabidopsis.org/servlets/Search?type=general
&search_action=detail&method=1&show_obsolete=F&name=
```

**Tail:**
```
&sub_type=gene&SEARCH_EXACT=4&SEARCH_CONTAINS=1
```

(c) Fill them into corresponding text boxes in “General” Tab of “Settings” and then click “Save”.

3. I tried to save a plot, but it is not found at the place I save it. Is it a bug in CrusView?

No. CrusView could only save a plot that is currently visible. When a plot window is closed by the user, its information is wiped out from memory at once, although the save plot option may be enabled.

4. Some scaffolds I am trying to assemble are small or containing repetitive sequences, what should I do? (Manual placement of unassembled scaffolds containing certain extent of gene synteny information.)

CrusView provides the possibility to aid the difficulties during genome assembly, including the cases for assembling small scaffolds or scaffolds with high density of repetitive sequences. Assuming most of the other scaffolds’ arrangement has been determined, the issue can be interpreted as finding the scaffold’s position and orientation on a pseudo-genome.

For a scaffold with at least a dozen of genes, the user can firstly check its major block color by zoom in, or more specifically, view its genes’ block ID on the main screen. Confirming to the karyotype provided and further information such as gene density plot, synteny plot and dotplot, the user is likely to figure out the scaffold’s position among other scaffolds. A sample case for finding the location of scaffold 28 of *Eutrema Salsugineum* is provided below.
Scaffold 28 of Eutrema is a scaffold less than 500 kb in length, with only a dozen of genes. As shown in figure 10, the major block color for this scaffold is purple, telling us its block ID is either I or J.

Thus in this case, we have to further inspect this scaffold with functions on the main screen. After placing this scaffold at the beginning of Chr4, the current arrangement of the scaffolds is saved as assembly “checking_scaf_28” via function “Push to Main Screen”.

Then, on the main screen, we use the filter function with keyword “chr4” to select all genes on this pseudo chromosome 4. And then, we use function “Sort Current Table By Position” on the Edit Menu to sort the genes. The information for genes on scaffold 28 shows up at the beginning of the table, as shown in Figure 12.
The genes’ block IDs turns out to be either 1 or 0. Known block ID 0 is for those genes can’t get one block ID, we know that scaffold 28’s block ID is 1. Thus we know there are only 4 possible positions to place scaffold 28 on the pseudo chromosome, which are both sides of scaffold 15 and the right sides of scaffold 21 and scaffold 12. To further determine it, we need the information from the syntenic plot and dotplot. So we’ll compare this pseudo chromosome with chromosome 2 of Arabidopsis, as the two share similar syntenic blocks. Shown in the syntenic plot below, scaffold 28’s syntenic region on Arabidopsis is at around 11Mb. The dotplot on the next page further confirmed this with several orthologs. We could also infer from the dotplot that the real location of scaffold 28 is at around 20Mb of pseudo chromosome 4, which is just at the right side of scaffold 12.
Fig. 14: Dotplot for pseudo-chr4 of Eutrema and chr2 of Arabidopsis. Gene groups OG5_212231, OG5_212230 and OG5_135168 are marked in red. Their members in Arabidopsis are at around 11 Mb.

To determine the orientation of a scaffold without further information such as pair-end reads’ mapping result or genetic markers, we have to assume there’s no structural variations in the syntenic region. However, for this small scaffold, the orientation is still hard to determine from the visual, since it contains just a dozen of genes which are almost flat on the dotplot. After generating a new pseudo genome with scaffold 28 in the right place, we could summarize the location of the mentioned ortholog groups in the following table. The order of the orthologs indication current orientation “+” follows our assumption.

<table>
<thead>
<tr>
<th>Group</th>
<th>Start Position on Eutrema</th>
<th>Start Position on Arabidopsis</th>
</tr>
</thead>
<tbody>
<tr>
<td>OG5_135168</td>
<td>20752628</td>
<td>11388620</td>
</tr>
<tr>
<td>OG5_212230</td>
<td>20857306</td>
<td>11393150</td>
</tr>
<tr>
<td>OG5_212231</td>
<td>20952242</td>
<td>11401550</td>
</tr>
</tbody>
</table>

Tab. 6: The Location of three ortholog groups on chr4 of Eutrema and chr2 of Arabidopsis

In the sample case above, we find out that the location of scaffold 28 is on the right side of scaffold 12 with orientation “+”. A similar procedure can be done for another scaffold with repetitive sequences, as long as it contains a dozen of genes. For those scaffolds contains fewer genes, the chance to find its location via CrusView is not high. However, CrusView provides the function to integrate further information in the curation step, such as genetic markers and position of centromere/telemere, which may increase the chance to find a scaffold’s location on the genome.

5. I have scaffolds for a genome without karyotype information. Could CrusView help on inferring the pseudo-genome? (Compare a plant genome without karyotype to the reference
Yes, but not always. CCP-derived karyotype is important for CrusView and it’s hard to infer the whole genome without it. However, for certain Brassicaceae species, we know that they may share 24 large syntenic blocks. Also, with the orthologs from *Arabidopsis Thaliana*, each gene on these scaffolds can be assigned with one predicted block ID. If most genes on a scaffold can be assigned with correct block IDs, the block ID composition of this scaffold may be inferred. Additionally, we also know the ancestral karyotype for all Brassicaceae species, which is *Arabidopsis Lyrata*’s karyotype. By taking advantage of these general information and the help from dotplots, we may be able to assemble some scaffolds into pseudo-chromosomes, especially for those chromosomes undergo less structural variations. Again, we provide a sample case below.

In this sample case, we assume that we don’t have the karyotype data for Eutrema but annotated scaffolds, and we could assemble some of these scaffolds into pseudo-chromosomes. Firstly, we need to predict the block IDs for each gene in order to see the block compositions of each scaffold. To do that, we could trickily use CrusView’s function “Generate New Bed file” on the File menu, although it is implemented for importing assembled genomes. The description of the required input file and parameter settings can be found in the “Import New Genome” section of this manual. The screenshots in Figure 15 shows the input files, the parameter settings used in this sample case.

Fig. 15: The input files and parameter settings for importing the Scaffolds of Eutrema into CrusView

After a while, a new “Genome” called Eutrema_Scaffolds shows up on CrusView’s main window. Some of its scaffolds are shown in the figure 16 below. We could see that some scaffolds, such as scaffold 1, 5 and 7, are almost occupied by only one block color.
In particular, we could see scaffold 5 may share the same block color with chromosome 1 of Lyrata. An additional dotplot in CrusView could further confirm it. Figure 17 below shows the karyotype of *Arabidopsis Lyrata* and the further dotplot.

Fig. 16: The imported Eutrema’s “Genome” on the main window

Fig. 17: Lyrata’s karyotype and shared syntenic region between scaffold 5 of Eutrema and chromosome 1 of Lyrata
By repeating the same strategy, we can find that scaffolds for chromosome 1, 3, 4 and 7 of Eutrema could possibly be assembled, as we could have the following information from CrusView: 1) the predicted block colors of the genes on these scaffolds are mostly yellow, blue, purple and light blue, for the 4 chromosomes respectively; 2) these scaffolds share large syntenic regions with chromosome 1, 3, 4, 7 of Lyrata that nearly cover the whole chromosomes (shown in the figure 18); 3) the karyotype of Lyrata shows chromosome 1, 3, 4, 7 in Lyrata’s block color are also yellow, blue, purple and light blue (shown in Figure 17a), respectively.

Fig. 18: The dotplot between chromosome 1, 3, 4 and 7 of Eutrema and Lyrata

For other scaffolds with multiple block colors (scaffold 2 of Eutrema, for example), it is not quite feasible to assemble them into chromosomes with Lyrata’s karyotype, as chromosomal level structural variations may have taken place.