

ABySS-Explorer v1.3.0: User Manual

prepared by:

Ka Ming Nip, Cydney Nielsen, Shaun Jackman, Inanc Birol

Canada's Michael Smith Genome Sciences Centre

November 2011

ABySS-Explorer is an interactive Java application that employs a novel graph-based representation to display a sequence assembly and associated meta data. The tool was designed with the ABySS sequence assembler in mind and was motivated by the need to examine assembly structure, in particular contig connectivity and supporting paired-end read relationships.

Table of Contents

1. [Requirements](#)
2. [Running ABySS-Explorer](#)
3. [User Interface and Usage](#)
 - A. [Importing Assemblies](#)
 - B. [Assemblies Panel](#)
 - C. [Settings Panel](#)
 - D. [Statistics View](#)
 - E. [Plots View](#)
 - F. [Navigator View](#)
 - G. [Search Box](#)
 - H. [Miscellaneous Features](#)
4. [Terminology](#)
5. [External Packages](#)
6. [Technical Support](#)

1. Requirements

System Requirements:

- Java SE Run Time Environment 6 or above
- Smallest screen resolution supported: 640 x 480
- optional: internet connection and default web browser for BLAT

Dataset Requirements:

- Assembly must be generated by ABySS 1.2.0 or above
- All files from the same assembly must be in the same directory
- Files used: *-contigs.fa, *-contigs.dot, *.adj, *.dot, *-3.hist, coverage.hist, *-3.dist, *-5.path

Preferred structure of assembly directories:

```
<name>/
  k<k-value>/
    coverage.hist
    <name>-3.dist
    <name>-3.hist
    <name>-5.adj
    <name>-5.path
    <name>-contigs.fa
    <name>-contigs.dot
    ...
  k<another k-value>/
  ...
```

2. Running ABySS-Explorer

The command for launching ABySS-Explorer is:

```
java -jar <PATH_TO_JAR> [argument]
```

where:

<PATH_TO_JAR> is the path to the ABySS-Explorer JAR file.

Make sure the JAR file and the "lib" directory are in the same directory.

[argument] is optional and can be any one of the following:

i) paths to one or more ABySS assembly directories or their parent directory

ii) path to either an ADJ or a DOT file

If no paths were specified, ABySS-Explorer would attempt to check whether the current directory is an ABySS assembly directory or a parent directory to one or more ABySS assembly directories.

iii) **--help** display information for help and exit

iv) **--version** output version information and exit

It is recommended that sufficient amount of heap space is requested when launching ABySS-Explorer. For example, to request 1024 MB of maximum heap space:

```
java -Xmx1024m -jar AbyssExplorer.jar
```

Please consult Java's documentation for more information on java commands:

<http://java.sun.com/javase/6/docs/technotes/tools/windows/java.html>

Alternatively, ABySS-Explorer can be initially launched from the [Java Web Start link](#) on our website. If prompted about unverified digital signature, simply click "Run" to continue. ABySS-Explorer can then be launched from the icon created on the desktop.

3. User Interface and Usage

The user interface is composed of seven components shown in Figure 1.

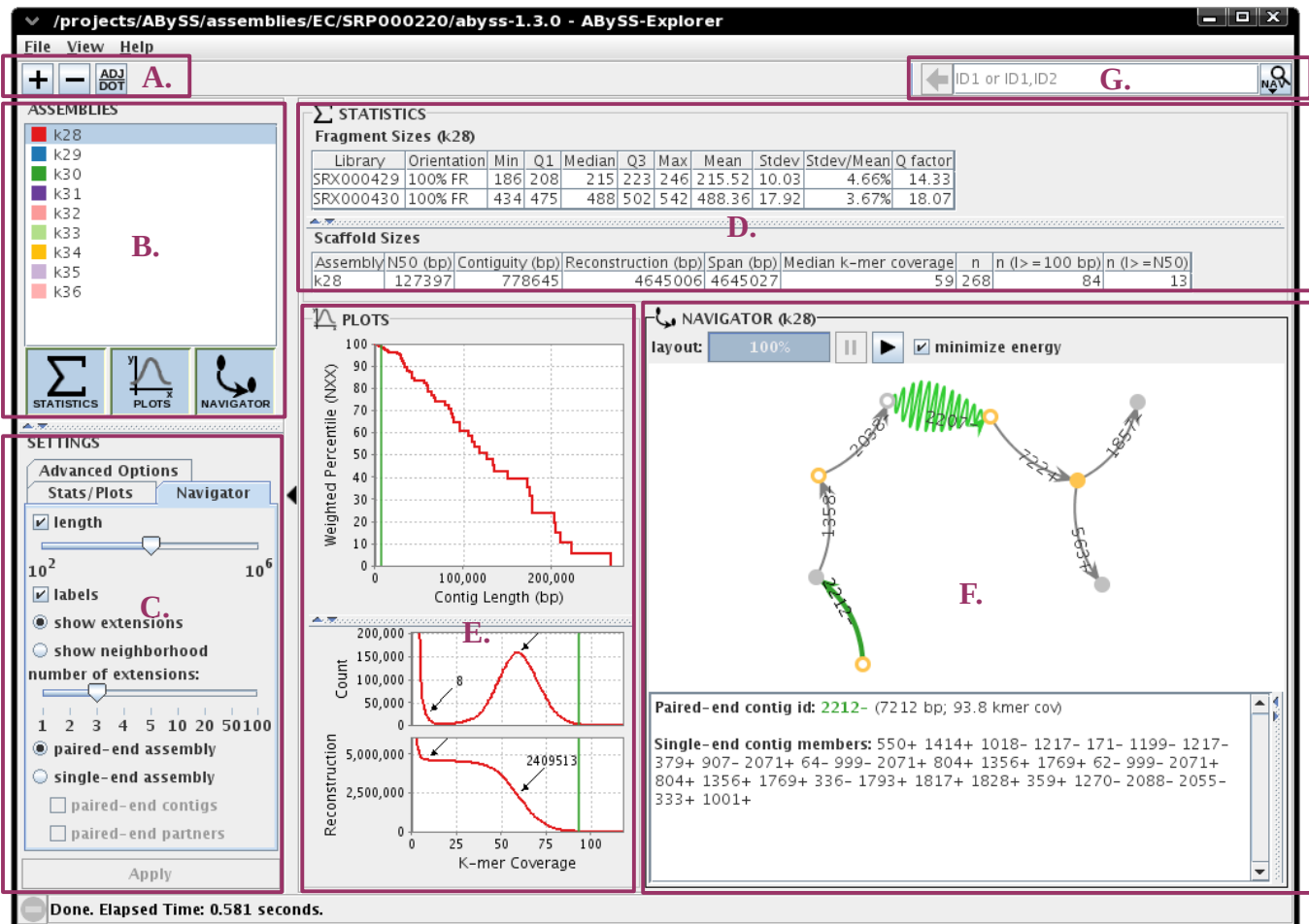


Figure 1. The user interface of ABySS-Explorer v1.3.0.

- A. [Importing Assemblies](#)
- B. [Assemblies Panel](#)
- C. [Settings Panel](#)
- D. [Statistics View](#)
- E. [Plots View](#)
- F. [Navigator View](#)
- G. [Search Box](#)

A. Importing Assemblies

Assemblies can be imported in two ways using the buttons shown in Figure 2:

- i. importing an ABySS assembly directory
- ii. importing an ADJ or DOT file



Figure 2. The buttons for importing assemblies. These commands are also available in the *File* menu.

The '+' button adds one or more ABySS assemblies to the *Assemblies* panel

When the file chooser is launched, one or more ABySS assembly directories or their parent directories can be selected.

The '-' button removes one or more selected assemblies from the *Assemblies* panel

The 'ADJ/DOT' button loads one file, either an ADJ file or a DOT file.

The ADJ files must reside in a directory whose name starts with 'k' followed by the value of k.

When an ABySS assembly directory is loaded, all three visualization views (*Statistics*, *Plots*, *Navigator*) can be launched. When either an ADJ file or DOT file is loaded, only the *Navigator* view is available.

B. Assemblies Panel



Figure 3. The *Assemblies* panel.

The *Assemblies* panel (Figure 3) lists the ABySS assemblies loaded.

The colored squares denote the colors that will be used in the *Plots* view.

Multiple assemblies can be selected for *Statistics* view and *Plots* view, but *Navigator* view can only view one assembly at a time.

The views can be launched from the three toggle-buttons. When there are no changes in the selection of assemblies or in the settings (described in the next section), pressing these buttons would only toggle the visibility of the views.

Otherwise, pressing the green buttons would apply any new settings and regenerate the views.

C. Settings Panel

The *Settings* panel consists of the settings for the three views. Changes in settings are only applied when the *Apply* button is pressed. The *Apply* button is available (green) when there are changes in settings.

i) Stats/Plots:

The *Stats/Plots* tab (Figure 4) has all the basic settings both the *Statistics* view and the *Plots* view.

Figure 4. The *Stats/Plots* tab of the *Settings* panel.

min. contig length (bp) field

sets the length (in base-pairs) of the smallest contig used to generate the statistics and plots

N50-plot & Coverage plot check box

shows the N50-plot and the k-mer coverage plot when the *Plots* button is pressed

N50-plot Y-axis unit drop-down box

sets the Y-axis unit for N50-plot as either *percentile* or *reconstruction*

N50-plot X-axis scale drop-down box

sets the X-axis scale for N50-plot as either *linear* or *logarithmic*

Fragment Size Distribution check box

shows the *Fragment Size Distribution* plot when the *Plots* button is pressed

select libraries button

brings up a menu to select libraries for *Fragment Sizes* statistics and the *Fragment Size Distribution* plot. In general, this button is not used if the assembly was based on one library.

ii) Navigator:

The *Navigator* tab (Figure 5) has all the settings for *Navigator* view.

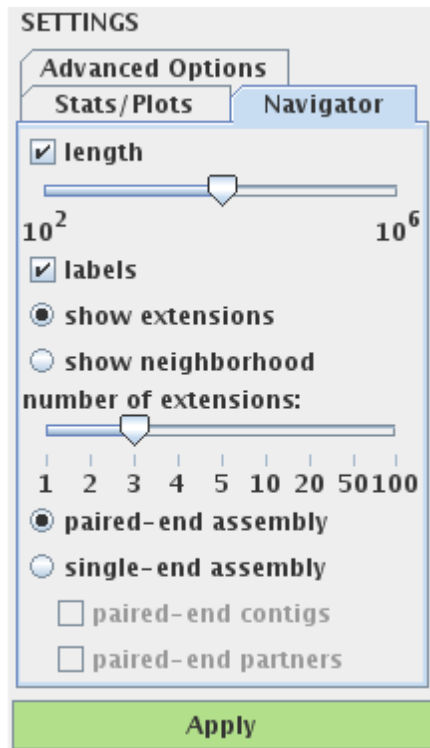


Figure 5. The *Navigator* tab of the *Settings* panel.

***length* check-box**

enables rendering of length of contigs as a wave packet

***length* slider**

sets the length represented as half-period of the wave packet; available when *length* check-box is checked

***show extensions* radio button**

constructs a neighborhood such that only extensions to the seed contig are shown; edges shown are limited to the number of extensions

***show neighborhood* radio button**

constructs a neighborhood such that all neighbor edges are shown; edges shown are limited to the size of neighborhood

***size of neighborhood/number of extensions* slider**

sets the size of the neighborhood of the seed contig

***labels* check box**

shows the identifier labels of the contigs

***paired-end assembly* radio button**

displays the graph of the paired-end assembly

***single-end assembly* radio button**

displays the graph of the single-end assembly

***paired-end contigs* check-box**

highlights the single-end contigs that make up a paired-end contig; available when the single-end assembly radio button is selected

***paired-end partners* check-box**

highlights the paired-end partners; available when the single-end assembly radio button is selected

iii) Advanced Options:

The *Advanced Options* tab (Figure 6) has the advanced settings for the *Statistics* view and the *Plots* view.



Figure 6. The *Advanced Options* tab of the *Settings* panel.

file to compare drop-down box

sets the file for *Scaffold/Contig Sizes* statistics and *N50-plot*

This drop-down box lists the names of all *.fa and *.adj files in the assembly directories added.

However, there are special cases where the file selected is ignored for N50-plot. When the *single-end assembly* radio-button from *Navigator* settings is selected, either *-5.adj or *-5.fa (instead of *-contigs.fa) would be used to generate the N50-plot. When the *paired-end assembly* radio-button is selected, *-contigs.fa would be used to generate the N50-plot. Note that *-contigs.fa contains the paired-end assembly whereas *-5.fa and *-5.adj contain the single-end assembly.

unit of length drop-down box

sets the unit of length for contigs in *Scaffold/Contig Sizes* statistics and N50-plot as one of:

bp:	number of base-pairs
k-mer:	number of k-mers (available to *.fa files only)
nol-bp:	number of non-overlapping base-pairs (available to *.adj files only)

When the *Navigator* view is active, the unit of length will always be *bp* for N50-plot.

D. Statistics View

Σ STATISTICS										
Fragment Sizes (k28)										
Library	Orientation	Min	Q1	Median	Q3	Max	Mean	Stdev	Stdev/Mean	Q factor
SRX000429	100% FR	186	208	215	223	246	215.52	10.03	4.66%	14.33
SRX000430	100% FR	434	475	488	502	542	488.36	17.92	3.67%	18.07

Scaffold Sizes									
Assembly	N50 (bp)	Contiguity (bp)	Reconstruction (bp)	Span (bp)	Median k-mer coverage	n	n (l >= 100 bp)	n (l >= N50)	
k28	127397	778645	4645006	4645027	59	268	84	13	
k29	119060	774386	4647979	4648002	52	251	82	14	
k30	119293	773560	4640097	4640103	45	243	83	13	

Figure 7. The *Statistics* view.

The tables in *Statistics* view (Figure 7) can be sorted by column. Clicking the column header would sort the table. In addition, a table can be copied to the clipboard via the right-click menu.

i) Fragments Sizes

file required: *-3.hist

If multiple assemblies are selected in the *Assemblies* panel, only the *Fragment Sizes* statistics from the assembly with the highest contiguity is displayed. The statistics is calculated based on outliers-trimmed data. Outliers are outside the range:

$$[Q1 - 1.5*(Q3 - Q1), Q3 + 1.5*(Q3 - Q1)]$$

where Q1 and Q3 are the 1st and 3rd quartiles in the raw data, respectively. Note that the Q1 and Q3 displayed are calculated after removing outliers.

If over 1% of the fragment sizes are negative, separate statistics are calculated for the positive and the negative fragment sizes. Outliers are also trimmed separately. The fragment size is calculated as the position of the reverse read minus the position of the forward read, which is positive for pairs oriented forward-reverse and negative for pairs oriented reverse-forward.

Double-clicking a row displays the fragment size distribution of the corresponding library in a tab within the *Plots* view.

ii) Scaffold/Contig Sizes Statistics

file required: *.fa or *.adj

The file used can be selected from the *file to compare* drop-down box from the *Advanced Options* tab in the *Settings* panel.

The largest values in the *Reconstruction*, *N50*, *Contiguity* columns have a pinkish red colour.

The *Span* column is hidden when the assembly contains no gaps of Ns. In this case, the table is called *Contig Sizes* instead of *Scaffold Sizes*.

E. Plots View

N50-plot and k-mer coverage plot are displayed together under the same tab. They are vertically stacked when the *Navigator* view is shown, but they are placed side by side when the *Navigator* view is hidden. Fragment size distribution are displayed in separate tabs. Tabs for N50-plot/coverage plot and fragment size distribution can be shown/hidden by checking/unchecking the check-boxes in Stats/Plot Settings. Please see the "Settings" section for more details.

In any plot, the user can drag to pan and roll the mouse-wheel to zoom. The user can right-click to bring up the context menu for advanced settings from JFreeChart.

Plot-specific hot keys are available when the cursor is directly above the plots:

- A: (for "auto-range") auto-range on both axes to show all data points on the plot
- P: (for "prvious") toggles between the current and previous zoom
- L: (for "legend") toggles between showing and hiding the legend of the plot
- S: (for "special"; available to k-mer coverage plots only) zoom to the region of the coverage threshold and median k-mer coverage

i) N50-plot (see Figure 8)

file required: *.fa or *.adj

The file used can be selected from the *file to compare* drop-down list from the *Advanced Options* tab in the *Settings* panel. When the *Navigator* view is active, *-contigs.fa is used for paired-end assemblies and *-5.adj is used for single-end assemblies.

This plot provides a quick way to visualize all N_{xx} (eg. N50, N80, etc.) of a single assemblies. It also allows comparison of one or more assemblies throughout the different N_{xx}.

The contigs are arranged in descending order, such that the largest contig (N0) is at the lower right corner while the smallest contig (N100) is at the top left corner.

ii) K-mer Coverage Plot (see Figure 8)

file required: coverage.hist

The k-mer coverage plot consists of two subplots sharing a common *K-mer Coverage* axis.

The upper plot is a plot of the histogram in coverage.hist (ie. Count vs. K-mer Coverage). The annotations in the upper plot display ABySS's threshold coverage (near the local minimum) and median coverage (near the local maximum).

In the lower plot, the value for *Reconstruction* is the integral of the upper plot to the right of each *K-mer Coverage* value. The annotations in the lower plot display the reconstruction at ABySS's threshold coverage and median coverage.

iii) Fragment Sizes Distribution (see Figure 9)

file required: *-3.hist

The fragment size distribution is a plot of the histogram in *-3.hist (ie. Count vs. Fragment Size). The median is marked with an orange vertical line and the inter-quartile range is shaded in blue.

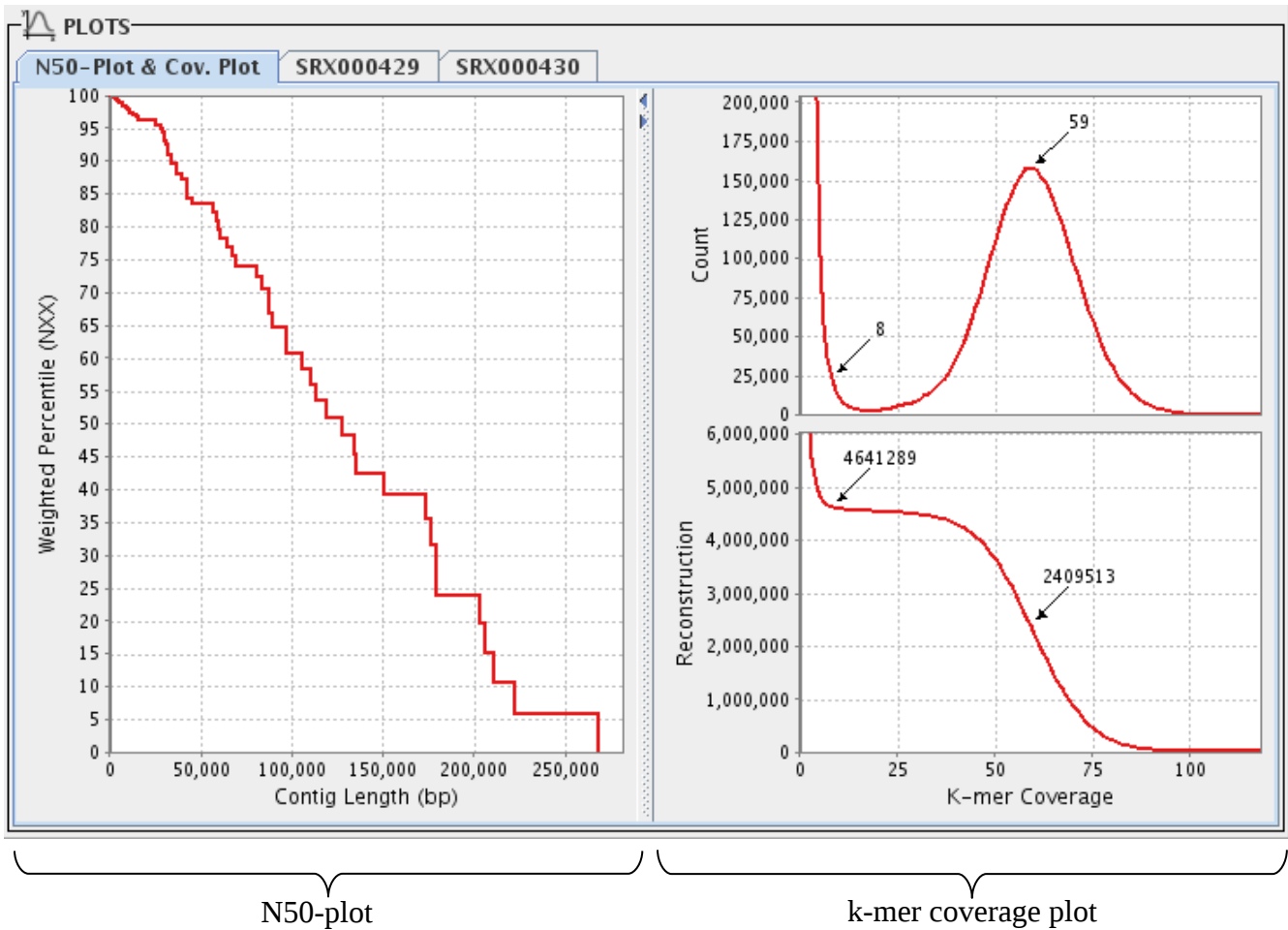


Figure 8. An N50-plot and a k-mer coverage plot in the Plots view.

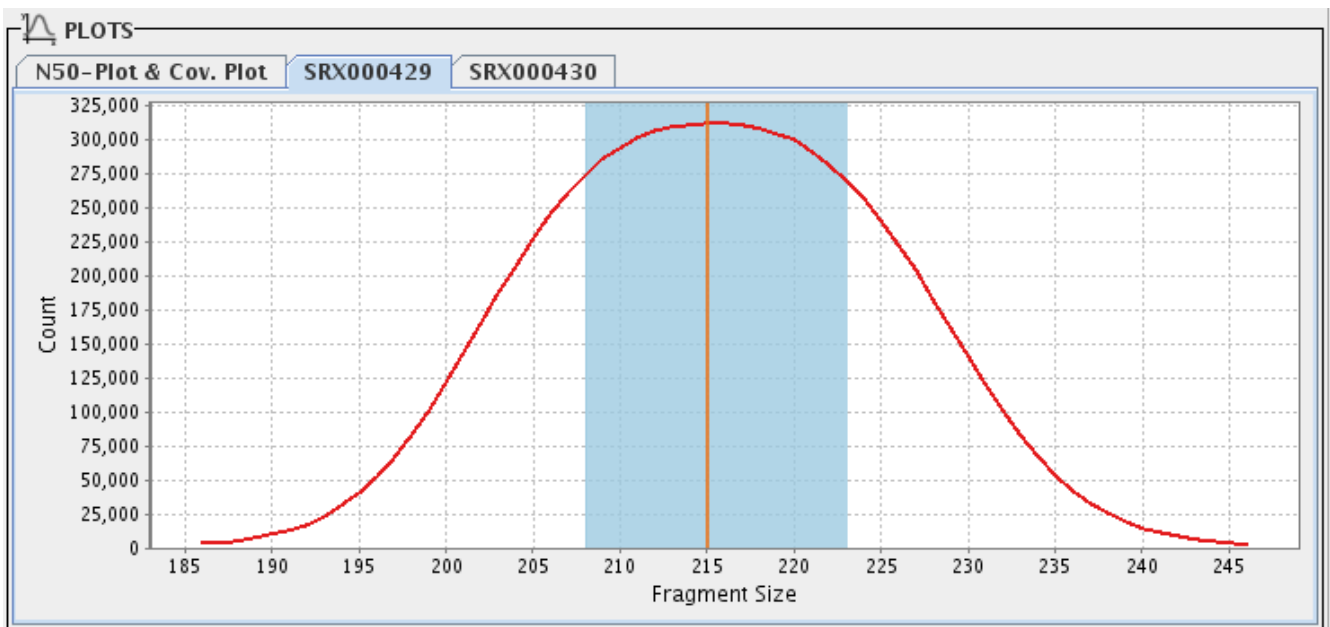


Figure 9. A fragment size distribution in the Plots view.

F. Navigator View

files required for paired-end assembly mode: *-contigs.dot, *-5.path

files required for single-end assembly mode: *-5.adj, *-5.path, *-3.dist

The Navigator view (Figure 10) displays the graph of an assembly. In general, edges represent contigs and vertices represent the overlap between contigs.

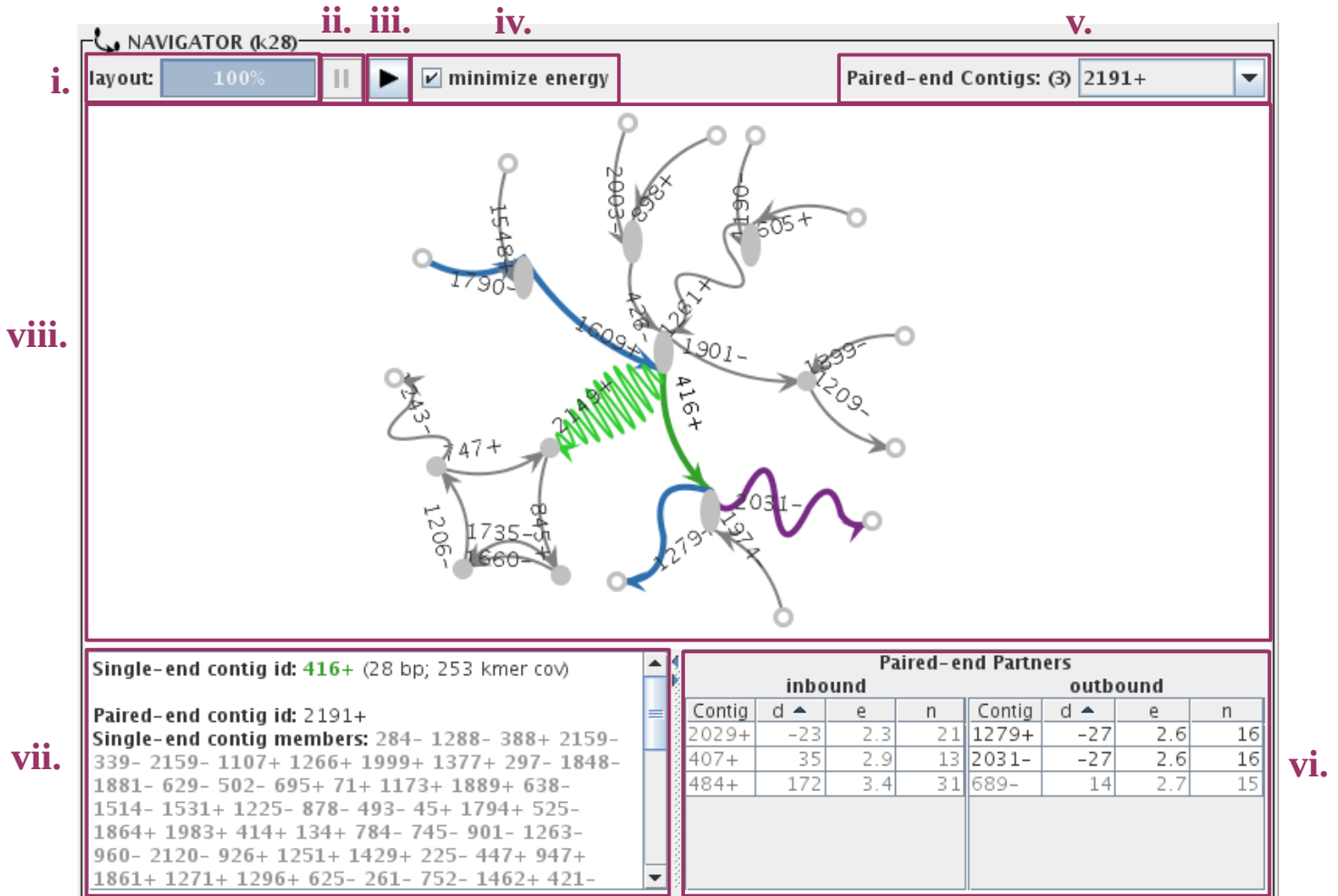


Figure 10. The Navigator view in single-end assembly mode.

- i. *layout progress bar*: shows the progress of the layout of the graph
- ii. *pause button*: pauses the layout and displays the current state of the graph
- iii. *resume button*: resumes the layout
- iv. *minimize energy check-box*: layout stops when the energy of the graph is minimized or 2000 iterations has run; otherwise, the edges and vertices are not spaced out
- v. *paired-end contigs drop-down box*: shows the selected contig's corresponding paired-end contig(s) (for single-end assembly mode only)
- vi. *paired-end partners table*: lists the inbound and outbound paired-end partners of the selected contig and their distance estimates (for single-end assembly mode only)
- vii. *information pane*: displays information about the selected contig or vertex
- viii. *visualization viewer*: displays the graph and allows user interactions

There are three basic transformation operations on the graph within the *visualization viewer*:

- pan: drag the background with mouse
- zoom: roll the mouse-wheel
- rotate: drag the background with mouse while holding the “Ctrl” key

The right-click pop-up menu within the *visualization viewer* has several options:

- *Refocus on this contig* (available if clicked on an edge): performs a Navigator search on the contig
- *Show nucleotide sequence* (available if clicked on an edge): performs a sequence search on the contig
- *Screen capture*: saves a snapshot of the graph
- *Center graph*: moves the graph to the center to the *visualization viewer*

Selected vertices can be moved by dragging with mouse. Multiple vertices can be selected by creating a selection rectangle while holding the “Shift” key.

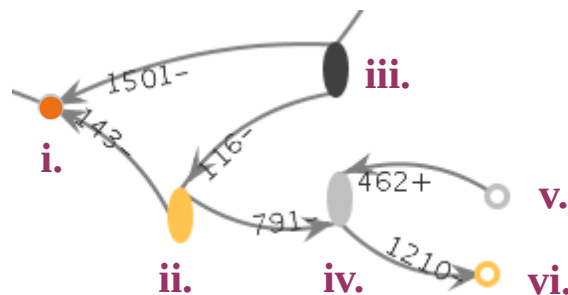


Figure 11. Vertex shapes and colours.

Default colour: iv, v

Default shape: ii, iii, iv

Selected: iii

Collapsed: i, v, vi

All incident edges shown: i, ii, iii, iv

One or more incident edges not shown: v, vi

Inferred overlap: ii, vi

Large overlap: i

All vertices have two poles and are displayed as an oval by default. If all incident edges displayed connect to the same pole, then the two poles of the vertex would be collapsed together, in which the vertex becomes a circle.

Vertices are gray and filled by default. They are dark gray when selected. They are unfilled when one or more incident edges are not shown. Double-clicking an unfilled vertex would center the graph on one of the edges that are not shown. The *Navigator* options *show extensions*, *size of neighbourhood*, and *number of extensions* can limit the number of edges displayed.

A vertex is said to have an "inferred overlap" if the overlap between some incident contigs is not found. In this case, the vertex is yellow.

A vertex is said to have an "large overlap" when:

- the overlap is larger (more negative) than the threshold distance and
- the overlap is larger than half the length of the longer contig of the overlap.

The threshold overlap is calculated as: $Q1 - 3 * (Q3 - Q1)$, where Q1 and Q3 are the 1st and 3rd quartile of the overlap in the assembly. In this case, the vertex is orange.

Edges are gray by default. The seed edge of a neighbourhood is light green. A selected edge is green and is thicker than a regular edge. See Figure 12.

The inbound and outbound contigs with paired-end read support to the currently selected contig are coloured light and dark purple, respectively. See Figure 13.

Single-end contigs that are members of a paired-end contig are coloured blue. A blue gradient indicates the directionality of the paired-end contig, with light blue at the 5' end and dark blue at the 3' end. See Figure 14.

A click on an edge selects the edge. Double-clicking an edge displays its reverse-complement (ie. from '+' to '-' and vice versa). If the Plots view is also active, a cursor (green vertical line) would be drawn for contig length and k-mer coverage. See Figure 15.

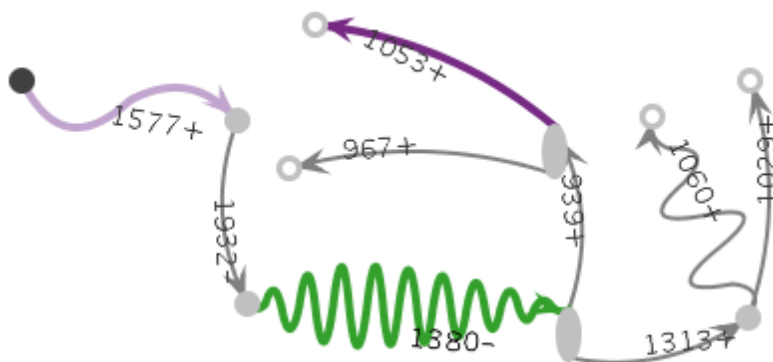


Figure 12. Edge colours.

Default: 1577+

Selected: 1932-

Seed: 1380-



Paired-end Partners							
inbound				outbound			
Contig	d ▲	e	n	Contig	d ▲	e	n
1577+	-2	0.7	235	1053+	-31	3.4	10
				1695+	-18	0.6	271

Figure 13. Graph showing the inbound and outbound partners of a selected contig.

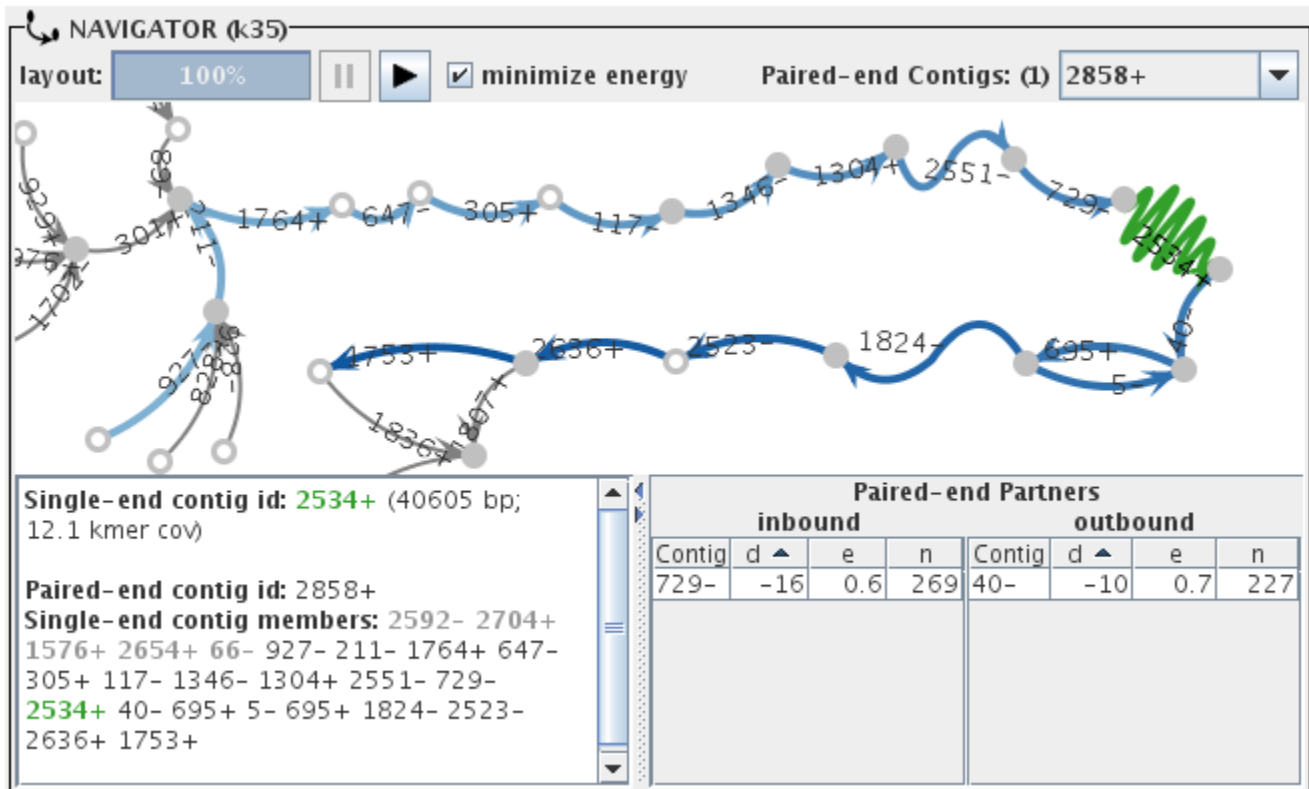


Figure 14. Graph showing single-end members of a paired-end contig 2858+. Contig 2534+ is selected and is a member of the paired-end contig. The light blue path 927-,729- are members of the paired-end contig and are 5' to 2534+. The dark blue path 40-, 753+ are also members of the paired-end contig but 3' to 2534+. Other members (grey contig ids in the *information pane*) are not shown.

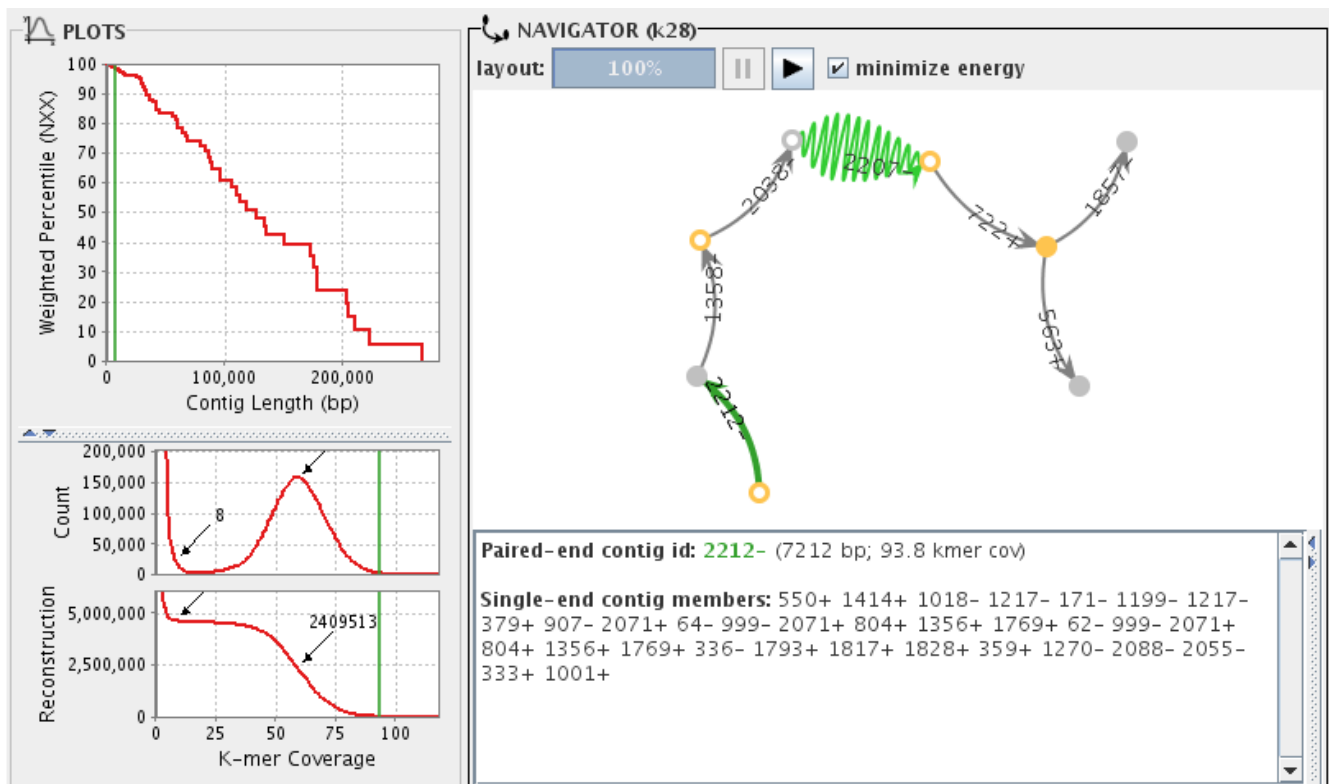


Figure 15. Plot view and Navigator view.

G. Search Box

There are two modes of contig search (Figure 15). Modes of search can be switched by clicking and holding over the search mode button. A click on the button performs a search.

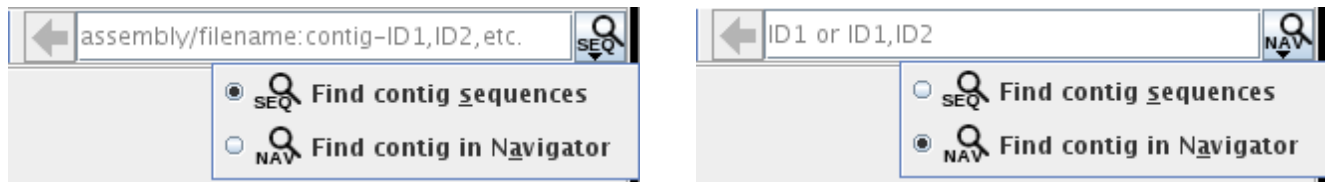


Figure 15. Modes of search and the corresponding format of query.

i) Sequence Search:

This search finds the nucleotide sequences of the query contigs in the *-contigs.fa
format of query: <A>/<F>:<C>

<A> = name of the assembly as displayed in the Assemblies panel

This field is not necessary if an assembly is selected or a DOT/ADJ file is loaded.

<F> = file name

This field is not necessary unless the file of interest is not *-contigs.fa.

<C> = one contig id or a comma/space-separated list of contig ids

This field is always required.

The sequences found are displayed in another window. The user can perform a web-BLAT search on the sequences found. Please note that there are restrictions on the usage of web-BLAT.

Example queries:

123

k32:123

k32/ecoli-4.fa:123

ii) Navigator Search:

Navigator search applies the settings when the *Apply* button is enabled (green).

query: one contig id (ie. 100+)

This search displays the neighbourhood of the query contig in the *Navigator* view.

query: two contig ids separated by a comma (ie. 100+,105+)

This search displays contigs that form the shortest path between the two contigs of interest in the Navigator. The path found ignores the direction of edges and so may not be a valid sequence path.

If the query has more than two contig ids (either separated by spaces or commas), a sequence search is performed instead.

H. Miscellaneous Features

- i) Double-clicking a view title maximizes the corresponding view by hiding other views shown. The hidden views can be shown by pressing the toggle-buttons in Assemblies panel.
- ii) Assemblies panel and Settings panel can be hidden or shown by pressing the triangle button on the divider.
- iii) The status bar at the bottom of the interface displays the current status. Pressing the stop button to the left of the status bar can kill the current running process (ie. loading file, rendering the layout of a graph in Navigator view).

4. Terminology

This is a list of terms used in this user manual and ABySS-Explorer.

ABySS assembly directory

This directory contains a single ABySS assembly and it must have either/both:

- i. a value of *k* in the DOT files
- ii. a name that starts with 'k' followed by the value of *k* used to generate the assembly (ie. k50) and has one or more files with extensions, *.fa or *.adj

Contig length units (bp, k-mer, nol-bp)

bp: number of base-pairs
k-mer: number of k-mers
= length in base-pairs – *k* + 1
nol-bp: number of non-overlapping base-pairs
= length in base-pairs; if both out-degree and in-degree is not 1
= length in base-pairs – *k* + 1; if either out-degree or in-degree is 1
= length in base-pairs – 2*k* + 2; if both out-degree or in-degree is 1

Contiguity

The square root of the sum of the squares of contig lengths of an assembly.

$$\sqrt{\sum_i x_i^2}$$

Extensions

Two edges are extensions of each other when the head of one edge connect to the tail of another edge at the same pole of the shared vertex. Also, any edges within a path are extensions of each other.

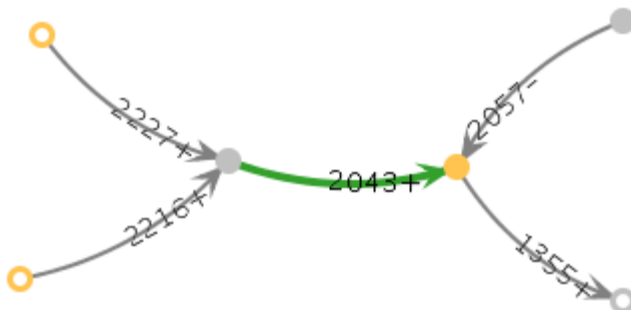
Consider the example in Figure 12, where edges 1577+, 1932-, and 1380- form a path:

Edge 1577+ is an extension of edges 1932- and 1380-.

Edge 1932- is an extension of edges 1577+ and 1380-.

Edge 1380- is an extension of edges 1577+ and 1932-.

Also, consider this example below:



Edges 2227+, 2216+, 1355+ are extensions of 2043+.

Edge 2057- is not an extension of 2043+, but it is an extension of 1355+.

However, this graph is the size-one neighbourhood of 2043+.

Neighbourhood

A graph that shows the neighbours (and next-neighbours) of a seed contig.

For example:

A neighbourhood of size one consists of all immediate neighbours of the seed contig.

A neighbourhood of size two consists of all immediate neighbours and next-neighbours of the seed contig.

A neighbourhood of size three consists of all immediate neighbours, next-neighbours, and next-next-neighbours of the seed contig.

N_{xx} (eg. N₅₀)

N_{xx} is the contig length, where contigs of same length or longer make up xx% of the reconstruction of an assembly.

Paired-end Partner

A contig that has paired-end support (distance estimate) with another contig.

Quality factor (Q factor)

The median divided by the interquartile range.

Reconstruction

The sum of all contig lengths of an assembly.

Scaffold span

The length of a scaffold; the sum of contig length and the number of N's in the sequence.

Span

The sum of all scaffold spans of an assembly.

5. External Packages

ABySS-Explorer uses the following external Java packages:

1. JFreeChart
<http://www.jfree.org/jfreechart/>
2. Java Universal Network Graph Framework (JUNG)
<http://jung.sourceforge.net/>
3. Batik SVG Toolkit
<http://xmlgraphics.apache.org/batik/>

All external packages required by ABySS-Explorer are included in the download.

6. Technical Support

Please direct your bug reports, questions, and suggestions to the ABySS Google Group:
abyss-users@googlegroups.com

You can also read and search existing discussions on the Google Group at:
<http://groups.google.com/group/abyss-users>

- End of User Manual -