

**Contiguity**

Construction and exploration of assembly graphs

Version 1.0

Licence: GPLv3

http://mjsull.github.io/Contiguity/

<Citation here>

mjsull@gmail.com

Contents

[Requirements 3](#_Toc396002438)

[Installation 3](#_Toc396002439)

[Overview 4](#_Toc396002440)

[General layout 4](#_Toc396002441)

[Menu 5](#_Toc396002442)

[File 5](#_Toc396002443)

[View 5](#_Toc396002444)

[Tools 5](#_Toc396002445)

[Canvas overview 6](#_Toc396002446)

[Context menus 6](#_Toc396002447)

[Contig Selection 7](#_Toc396002448)

[Colouring contigs 8](#_Toc396002449)

[Creating comparisons 8](#_Toc396002450)

[Viewing the Assembly 8](#_Toc396002451)

[Self comparisons 8](#_Toc396002452)

[Find Paths 8](#_Toc396002453)

[Constructing a Contig Adjacency Graph 8](#_Toc396002454)

[GUI 8](#_Toc396002455)

[Command-line 9](#_Toc396002456)

# Requirements

Contiguity comes precompiled for Windows, OSX and GNU/Linux. If you would like Contiguity to automatically generate comparison files for you, please have BLAST installed and in your path. Contiguity also accepts user generated comparison files in BLAST's tabbed output format with no headers (-outfmt 6).

**Running Contiguity as a python script**

**Python** – Contiguity has been tested using python 2.7, use earlier versions at your own risk.

**Tkinter (GUI)** – Included in most installations of python.

**Constructing contig adjacency graphs with Contiguity**

**Khmer** – While not required, it is strongly recommended you use Khmer to generate the CAG file unless you have a large amount of memory (>=16GB) a smallish sequencing run (5Mbp genome with < 200x coverage) and a lot of time.

**NCBI-BLAST** – BLAST is required for generating CAG files.

**Bowtie 2** – Bowtie 2 is required for generating CAG files with paired-end data.

# Installation

Executable versions of Contiguity for Windows, OSX and GNU/Linux are available at <https://mjsull.github.io/Contiguity> simply download unzip and run the Contiguity executable from within the unzipped folder.

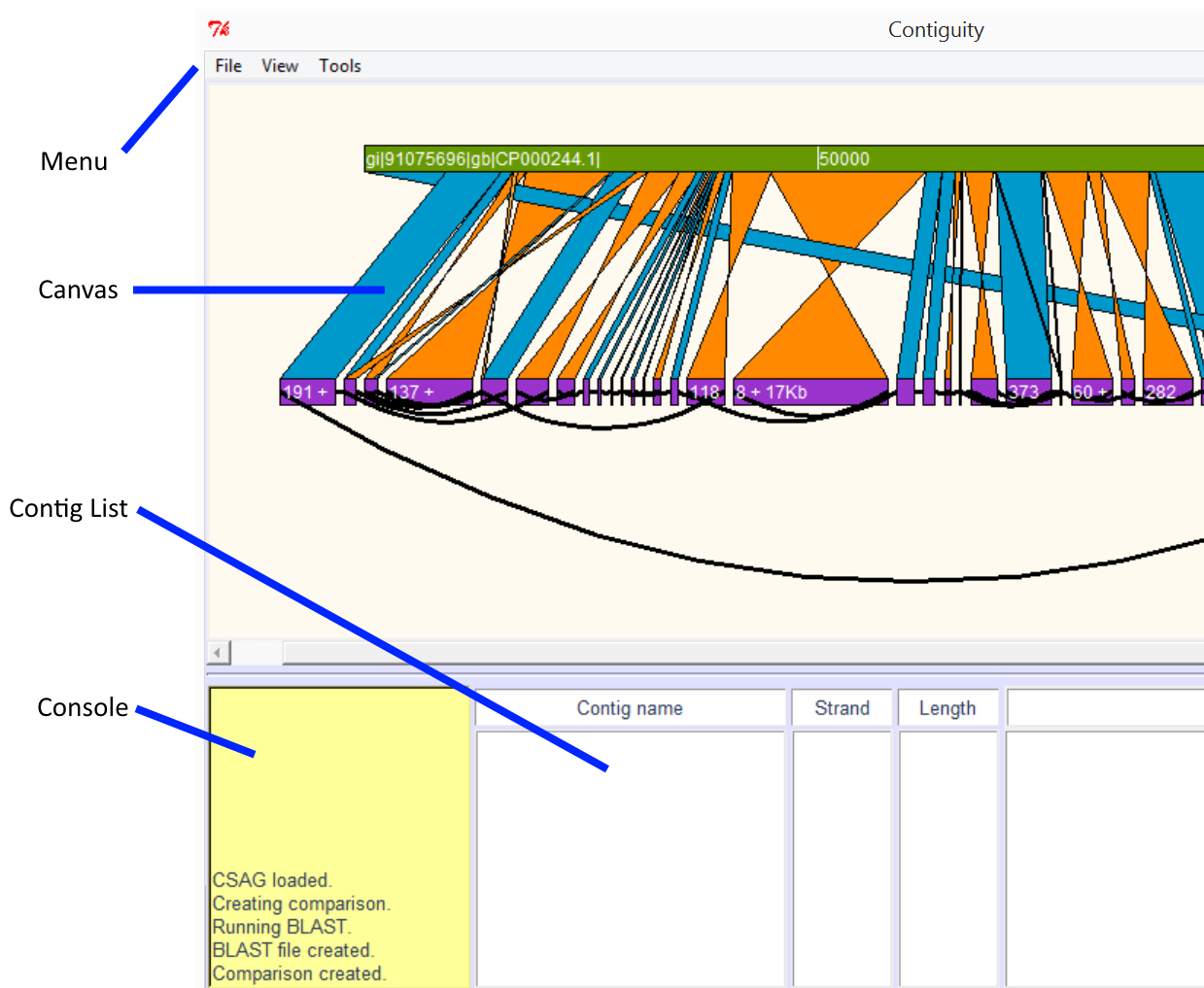
Alternatively an executable version of the source code is available at <https://github.com/mjsull/Contiguity>, please make sure you have satisfied all the requirements before executing.

# Citing Contiguity

<Paper will be available soon>

# Overview

## General layout



**Menu** – Dropdown menu

**Canvas** – Area where Contig, mapping and graph information is shown.

**Console** – Progress of processes running in the background is reported here.

**Contig** **List** – List of selected contigs

**ABOVE**: Comparison of an Illmunia assembly of E. coli to a virulence plasmid.

The Plasmid reference is shown in green. BLAST hits and inverted BLAST hits are shown in blue and orange respectively. Contigs are displayed in purple, a shortened version of the contig name on the contig when room allows. Its orientation and length is also shown where possible. If read information suggests that these contigs may be adjacent to each other, they are joined by a curved black line.

## Menu

### File

**Create CAG file –** Construct a CAG file (for more details see “Constructing a CAG” below)

**Create Comparison –** Create comparison (for more details see “Creating comparisons” below)

**Load Assembly –** Load FASTA/CAG/LastGraph/.dot

**Save Image –** Save canvas as a postscript image

**Change Working Directory –** Change the working directory from .contiguity\_wd

**Cancel Running Process –** Cancel any currently running processes

**Exit –** Cancel any currently running processes and quit Contiguity

### View

**View Assembly –** View loaded assembly (for more details see “view assembly” below)

**Self-Comparison –** Find and display sequence similarity between and optionally within contigs (for more details see “Self-comparisons” below)

**Add Contig –** Add contig (long or short name)

**Find Contig –** Go to specified contig on the canvas

**Zoom in –** Zoom in on the canvas

**Zoom out –** Zoom out on the canvas

**Stretch –** Stretch the canvas in the x-axis

**Shrink –** Shrink the canvas in the x-axis

### Tools

**Select All –** Select all contigs currently displayed on the canvas from left to right

**Clear Selected –** Clear all currently selected contigs

**Find Paths –** Find paths between selected contigs (For more information see “Find Paths” below)

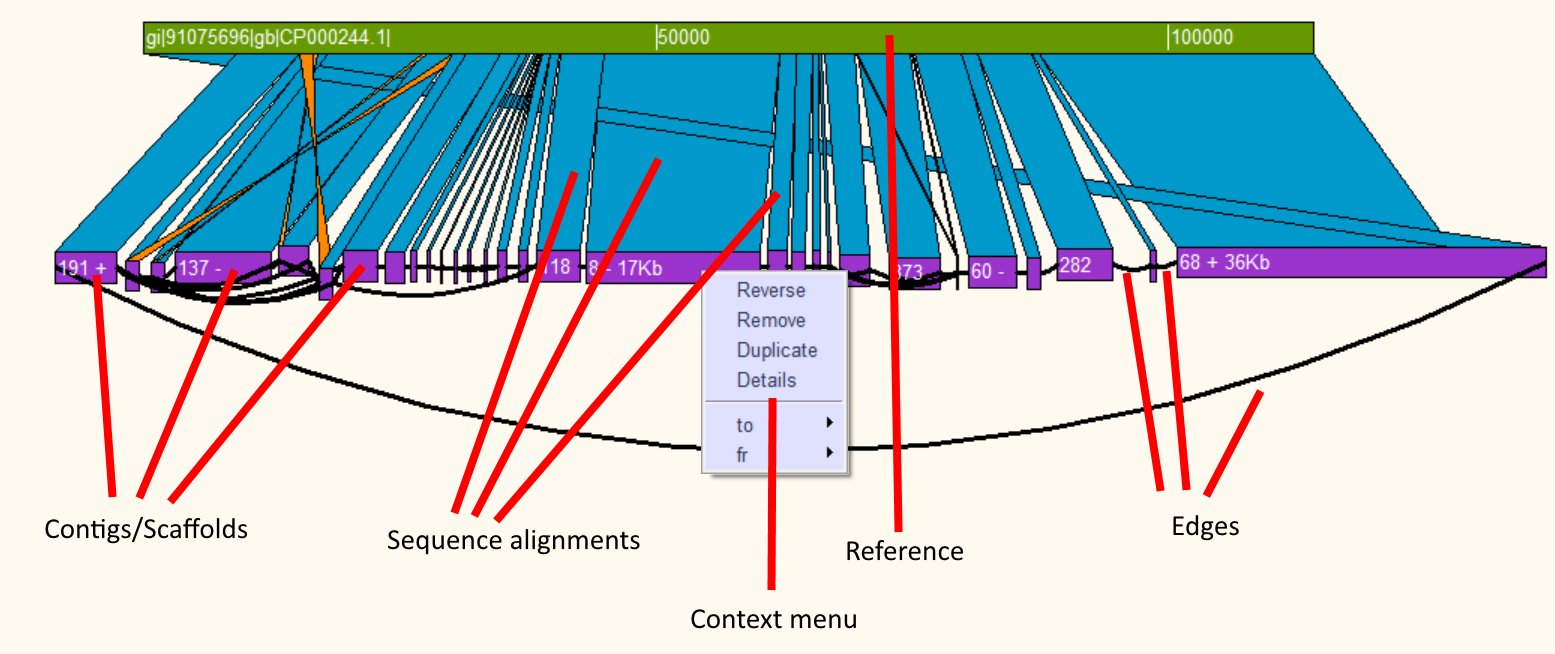
**Write FASTA –** Write a scaffold of currently selected contigs, if an edge is found between to sequential contigs, use the edge to fill in sequence or remove overlaps otherwise user defined scaffolding characters will be use

**Write multiFASTA –** Write currently selected contigs as a multiFASTA, each selected contig will be written as its own entry

**Colour options –** Change options relating to contig colours (for more information see “Contig Selection” below)

**Load Special –** Load custom colours from a text file (for more information see “Contig Selection” below)

## Canvas overview



**ABOVE**: Comparison of the Illumina assembly to a plasmid reference. Several contigs have been reversed and repeat region that was assembled into a single contig duplicated.

**Using the canvas**

The contigs and the reference may be moved individually by dragging them with the left mouse button. The canvas can be zoomed in and out by using the mousewheel or using the “w” or “d” key. Contigs may be shrunk or stretched in the x dimension by using the “a” and “d” key. Context menus can be brought up by right clicking a contig or sequence alignment allowing the user to reverse, remove, duplicate or bring up details about that contig or alignment.

### Context menus

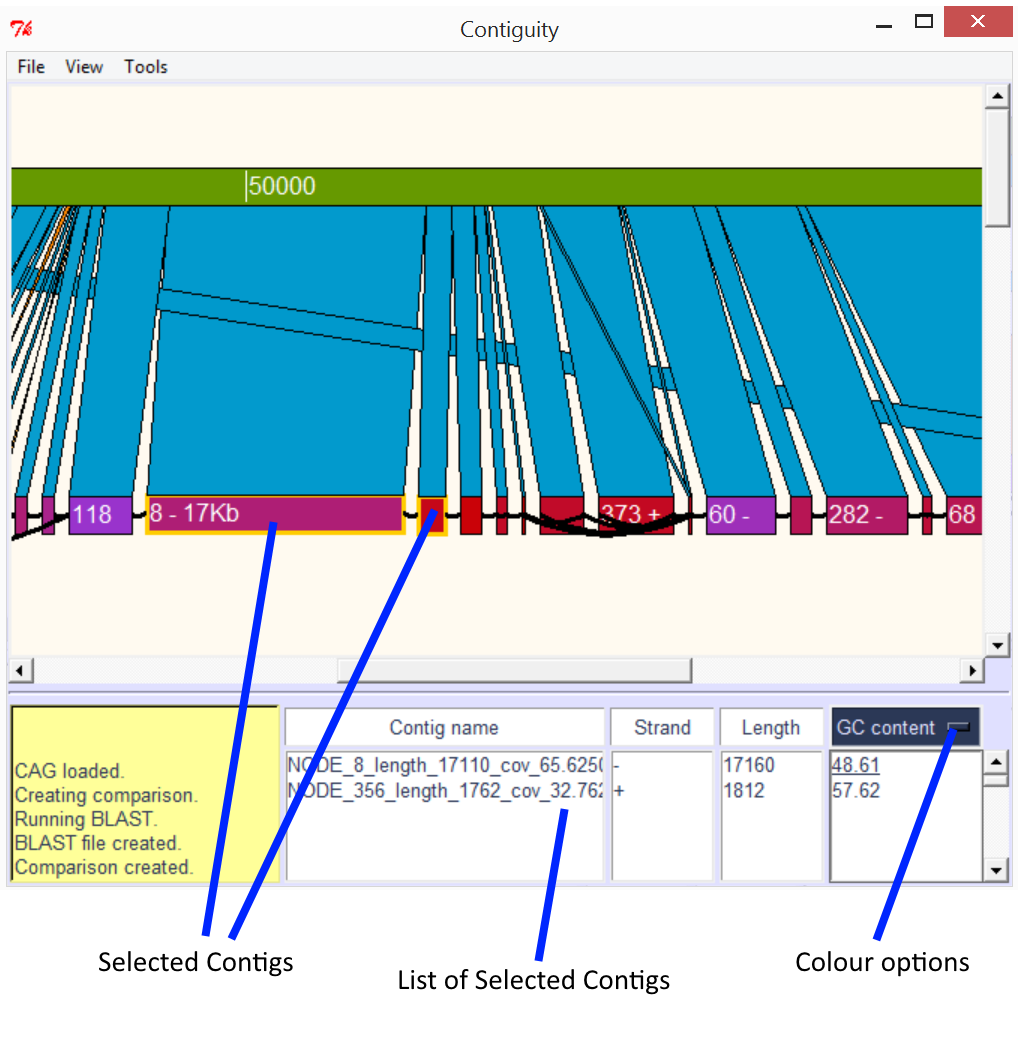
**Contig**

* **Reverse –** Reverse compliment the contig
* **Remove –** Remove the contig from the canvas
* **Duplicate –** Duplicate the contig
* **Details** – Bring up details about the contigs length, GC content, skew, coverage and edges
* **to/fr –** Brings up a menu that allows the user to highlight edges. Also allows the user to add, duplicate or move contigs attached to the 3’ end of the selected contig (to) or the 5’ end (fr)

**Blast**

* **Query/Subject**
  + **Move –** Move to query/subject so it is aligns with its match
  + **Goto –** Move to the query or subject
  + **Write –** Write the aligned query or subject sequence to a FASTA file
* **Remove –** Remove the hit
* **Details –** Show details about the hit such as identity, mismatches etc.
* **To back/front –** Move the blast hit to the back or front of of the canvas
* **Highlight –** Outline the hit with a specific colour

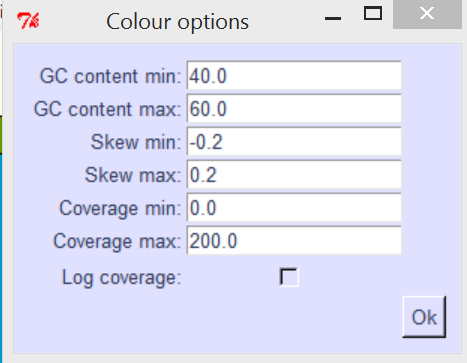
## Contig Selection



**ABOVE:** Assembled contigs aligned to a plasmid. Contigs 8 and 356 have been selected, and are displayed along with their orientation and length in the contig list section. GC content has been selected from a drop down menu (colour options) and the GC content of both contigs is displayed. Contigs on the canvas are now also coloured according to their GC content. Contigs with 60% or greater GC content are coloured red, contigs with 40% or lower GC content are coloured purple, contigs in-between are coloured on a gradient between red and purple. Selected contigs are outlined with a yellow box.

### Colouring contigs

Contigs can be coloured according to GC content, GC Skew, AT Skew, Coverage or user provided colours. Contigs are coloured purple if the selected value is equal to or under the minimum value or red if the colour is equal to or more than the maximum value, contigs with a value in between the minimum and maximum value are coloured on a gradient from purple to red. Minimum and maximum values can be changed from the colour options menu under tools.



**Custom colours and values**

Custom colours and values may be loaded. To load a file select “Load Custom” from the “Tools” menu and select a file with the custom values. The file should be in a tab delimited format with the first column being full contig names, the second column containing the custom value for the corresponding column and the third column containing hex colour codes.

e.g.

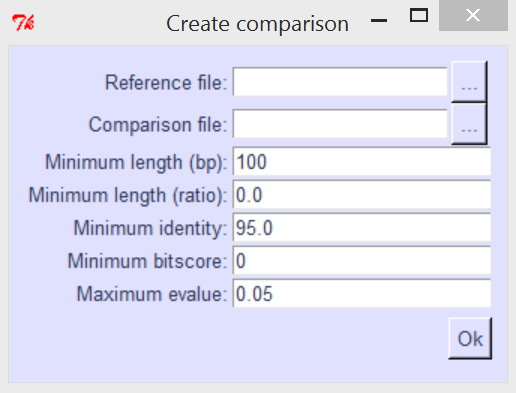
node\_8\_length\_10123\_cov\_52.121 20 #FF1C28

node\_9\_length\_32191\_cov\_24.121 30 #12FF28

node\_10\_length\_1021\_cov\_100.111 40 #12311C

## Creating comparisons

Comparison can be created between a loaded assembly and a reference by selecting “Create comparison” from the “File” menu. If you have NCBI-BLAST+ installed and in your path comparisons can be generated on the fly by providing Contiguity with a reference file. Alternative an alignment file in BLAST’s tab-delimited (no header) format may be provided. Alignment results will displayed if they satisfy the following parameters.



**Reference file –** Reference file in FASTA or multiFASTA format

**Comparison file –** Alignment file in BLAST’s tab-delimited format

**Minimum length –** Only alignments of length (in base pairs) equal to or greater than this value will be shown

**Minimum length (ratio) –** Fraction of the contig sequence aligned to show alignment i.e. 0.5 means that an alignment needs to cover at least half the query to be shown

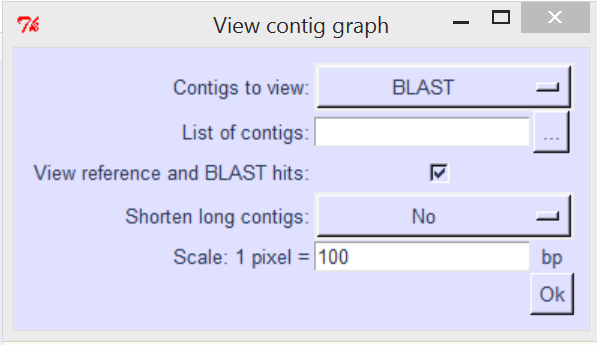
**Minimum identity –** Only alignments with an identity equal to or greater than this value (in percent) will be shown

**Minimum bitscore –** Only alignments with a bitscore equal to or greater than this value will be shown

**Maximum evalue –** Only alignments with an expect score lower than or equal to this value will be shown

## Viewing the Assembly

Once the assembly has been loaded by selecting “View” and then “View assembly”. If you also wish to view a comparison between the assembly and a reference ensure it has been created before viewing the assembly.



**Contigs to view**

* **BLAST –** view only contigs that align to the reference
* **ALL -**  view all contigs
* **List –** view a user defined set of contigs
* **Filter –** view only contigs that do not align to the reference

**List of contigs –** If “List” is selected from “Contigs to view:” please provide a text file with a list of contig names with each name in its own row

e.g.

node\_8\_length\_10123\_cov\_52.121

node\_9\_length\_32191\_cov\_24.121

node\_10\_length\_1021\_cov\_100.111

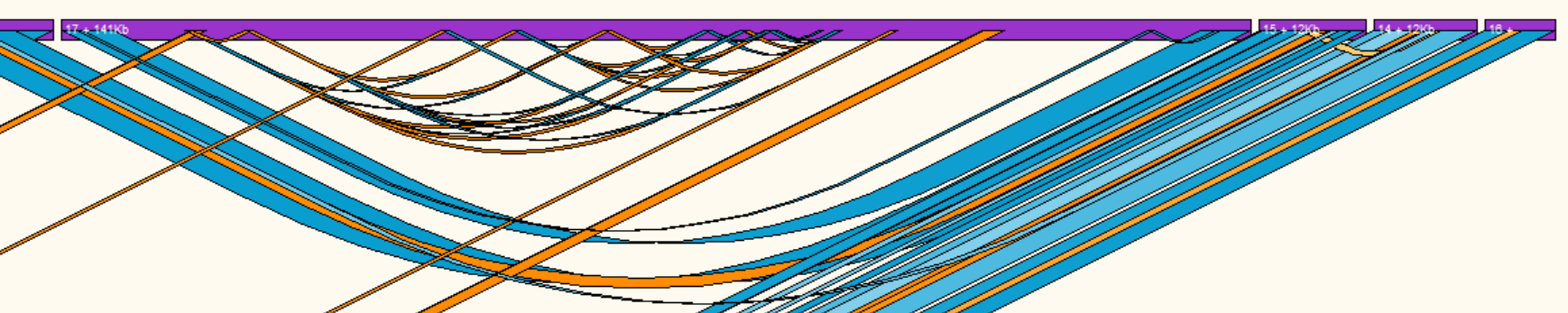
**View reference and BLAST hits –** When graph is drawn on the canvas also draw the reference and alignments

**Shorten long contigs**

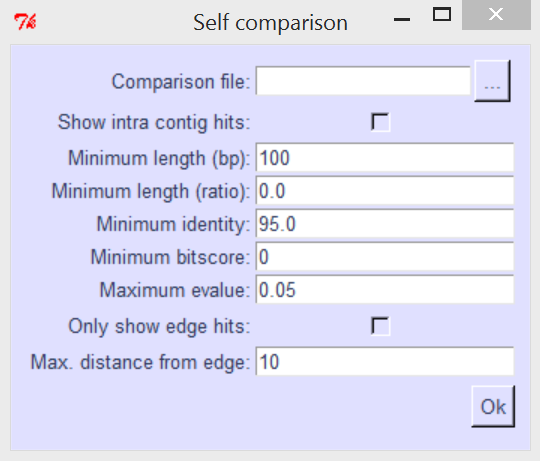
* **No –** Contig length on canvas proportional to length in base pairs
* **Constant –** All contigs 200px wide
* **Min –** All contigs at least 200px wide
* **Log –** Contig length on canvas proportional to the log of its length

**Scale: 1 pixel = … bp –** Ratio for determining size of contigs in pixels.

## Self comparisons



**ABOVE:** Self comparison of a pacbio assembly, direct repeats are shown with blue ribbons, inverted repeats are shown with orange ribbons. Hits with a low identity are lighter than hits with a high identity.



**Comparison file –** Alignment file in BLAST’s tab-delimited format

**Show intra contig hits –** Show hits mapping from and to the same contig

**Minimum length –** Only alignments of length (in base pairs) equal to or greater than this value will be shown

**Minimum length (ratio) –** Fraction of the contig sequence aligned to show alignment i.e. 0.5 means that an alignment needs to cover at least half the query to be shown

**Minimum identity –** Only alignments with an identity equal to or greater than this value (in percent) will be shown

**Minimum bitscore –** Only alignments with a bitscore equal to or greater than this value will be shown

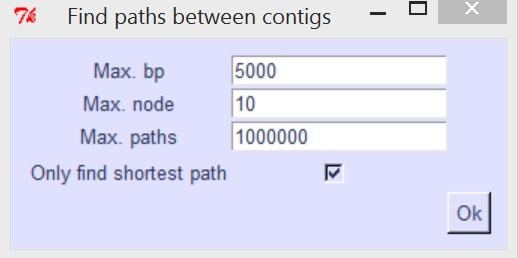
**Maximum evalue –** Only alignments with an expect score lower than or equal to this value will be shown

**Only show edge hits –** Only show hits that map to the edge of the contig

**Max. distance from edge –** Distance from the edge of a to consider hit an “edge hit”

## Find Paths

Find paths allows the user to find the shortest or all paths between all selected contigs. Simply select the contigs you would like to find paths between and then select “Find paths” from the “tools” menu. “Find paths” will search through the contig adjacency graph until it finds another selected contig or reaches the user defined node or base pair limit.



**Max bp –** Depth of search in base pairs

**Max node –** Depth of search in contigs

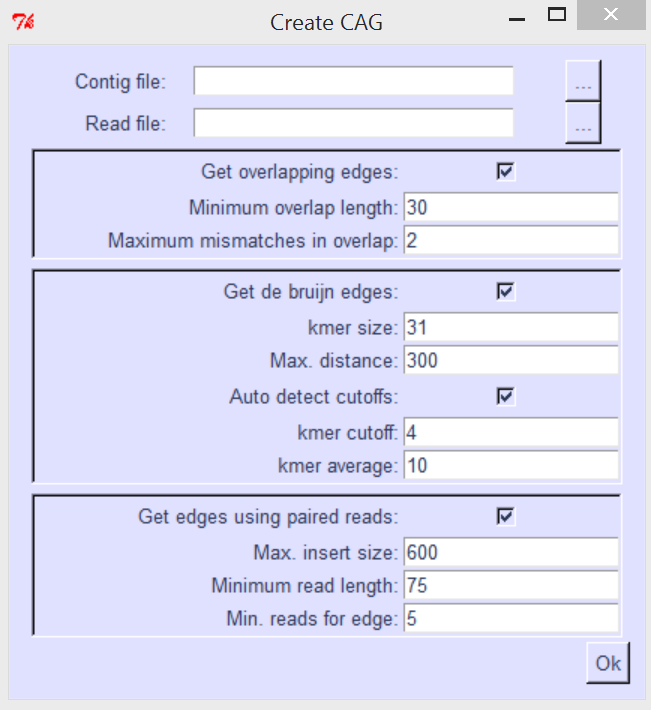
**Max paths –** If this value is reached Contiguity will lower the amount of nodes being searched

**Only find shortest path –** Only report the shortest path (in base pairs) found

# Constructing a Contig Adjacency Graph

## GUI

Contig adjacency graphs can be created by selecting “Create CAG” from the “File” menu



**Contig file –** FASTA file of contigs or scaffolds

**Read file –** Interleaved fastq file - read1\_left, read1\_right, read2\_left... orientated as such --> <--

**Get overlapping edges –** Find edges by looking for overlapping edges (Requires BLAST)

**Minimum overlap length –** Minimum overlap to consider two contigs adjacent

**Maximum mismatches in overlap –** Maximum number of mistmatches allowed in that overlap

**Get de bruijn edges –** Find edges using a De Bruijn graph

**Kmer size –** kmer size used to construct De Bruijn graph

**Max. distance:** Distance to search in the De Bruijn grah for an adjacent contig (this value + kmer size)

**Auto detect cutoffs –** Automatically detect coverage cutoff and median unique kmer frequency

**kmer cutoff –** kmers with a frequency less than this value will not be traversed

**kmer average –** median unique kmer frequency: all kmers with a frequency greater than this value will be traversed

**Get edges using paired reads:** Find edges by mapping paired-end reads (Bowtie-2 required)

**Max. insert size:** Only reads with an insert size smaller than this will be counted

**Minimum read length:** Only reads that have at least this value of bases mapping will be counted

**Min. reads for edge:** Minimum mapping reads to create an edge between two contigs

## Command-line

Contig adjacency graphs can also be created from the command line

USAGE: Contiguity.py -cl -c <contig\_file.fa> -fq <read\_file.fq> -o <output\_folder>

contig file: FASTA file of contigs or scaffolds

read file: Interleaved fastq file - read1\_left, read1\_right, read2\_left etc... orientated as such --> <--

output folder: folder to put output files in, can and will overwrite files in this folder, will create folder if folder

doesn't exist

Only other option to keep in mind is -rl if the read length is not 101bp

optional arguments:

-h, --help show this help message and exit

-co CONTIG\_FILE, --contig\_file CONTIG\_FILE

fasta file of assembled contigs or scaffolds

-rf READ\_FILE, --read\_file READ\_FILE

read file

-o OUTPUT\_FOLDER, --output\_folder OUTPUT\_FOLDER

output folder

-k KMER\_SIZE, --kmer\_size KMER\_SIZE

k-mer size for finding adjacent contigs [31]

-max\_d MAX\_DISTANCE, --max\_distance MAX\_DISTANCE

maximum distance apart in the de bruijn graph for

contigs to count as adjacent [300]

-kmer\_a KMER\_AVERAGE, --kmer\_average KMER\_AVERAGE

All k-mers above half this value will be traversed

[auto]

-kmer\_c KMER\_CUTOFF, --kmer\_cutoff KMER\_CUTOFF

cutoff for k-mer values [auto]

-ov OVERLAP, --overlap OVERLAP

minimum overlap to create edge [kmer\_size-1]

-rl MIN\_READ\_LENGTH, --min\_read\_length MIN\_READ\_LENGTH

Minimum read length [75]

-max\_mm MAX\_MISMATCH, --max\_mismatch MAX\_MISMATCH

maximum number of mismatches to count overlap [2]

-lo LONG\_OVERLAP\_IDENT, --long\_overlap\_ident LONG\_OVERLAP\_IDENT

minimum percent identity to create an edge where there

is a long overlap [85]

-mp MINIMUM\_PAIRS\_EDGE, --minimum\_pairs\_edge MINIMUM\_PAIRS\_EDGE

Minimum pairs to create edge [2]

-is MAX\_INSERT\_SIZE, --max\_insert\_size MAX\_INSERT\_SIZE

Upper bound on insert size [600]

-cl, --command\_line Run contiguity in command line mode

-no, --no\_overlap\_edges

Don't get overlap edges

-nd, --no\_db\_edges Don't get De Bruijn edges

-np, --no\_paired\_edges

Don't get paired-end edges

-km, --khmer Don't use khmer for De Bruijn graph contruction (not

recommended)

Thanks for using Contiguity