Gap5

James Bonfield, Wellcome Trust Sanger Institute

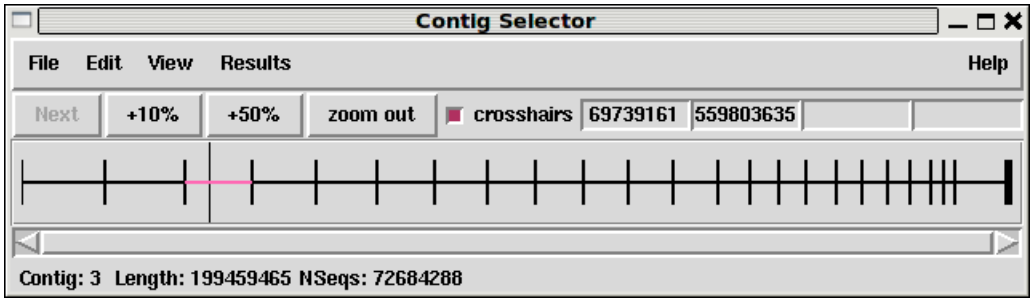
Short Contents

Table of Contents

1 Contig Selector

The gap5 Contig Selector is used to display, select and reorder contigs. It can be invoked from the gap5 View menu, but will automatically appear when a database is opened. In the Contig Selector all contigs are shown as colinear horizontal lines separated by short vertical lines. The length of the horizontal lines is proportional to the length of the contigs and their left to right order represents the current ordering of the contigs. This Contig Order is stored in the gap database and users can change it by dragging the lines representing the contigs in the display. The Contig Selector can also be used to select contigs for processing.

Unlike gap4, gap5 does not display annotations within the Contig Selector window.



The figure shows a typical display from the Contig Selector. At the top are the File, View and Results menus. Below that are buttons for zooming and for displaying the crosshair. The four boxes to the right are used to display the X and Y coordinates of the crosshair. The rightmost two display the Y coordinates when the contig selector is transformed into the contig comparator (see Chapter 2 [Contig Comparator], page 5). The two leftmost boxes display the X coordinates: the leftmost is the position in the contig and the other is the position in the overall consensus. The crosshair is the vertical line spanning the panel below.

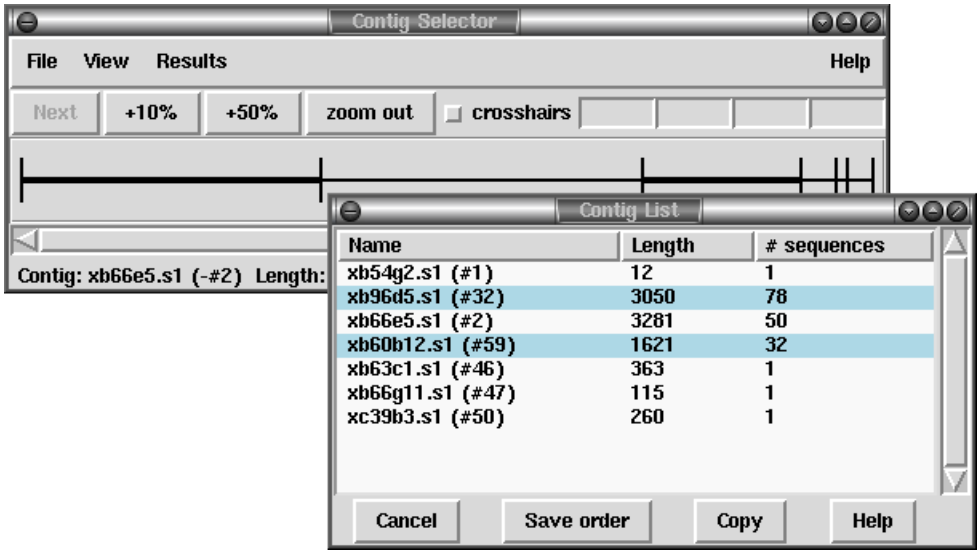
This panel shows the lines that represent the contigs and the currently active tags. Those tags shown above the contig lines are on readings and those below are on the consensus. Right clicking on a tag gives a menu containing “information” (to see the tag contents) and “Edit contig at tag” which invokes the contig editor centred on the selected tag.

The information line is showing data for the contig that is currently under the crosshair.

1.1 Selecting Contigs

Contigs can be selected by either clicking with the left mouse button on the line representing the required contig in the contig selector window or alternatively by choosing the "List contigs" option from the "View" menu. This option invokes a "Contig List" list box where

the contig names and numbers are listed in the same order as they appear in the contig selector window.



Within this list box the contig names can be sorted alphabetically on contig name or numerically on contig number. This is done by selecting the corresponding item from the sort menu at the top of the list box. Clicking on a name within the list box is equivalent to clicking on the corresponding contig in the contig selector. More than one contig can be selected by dragging out a region with the left mouse button. Dragging the mouse off the bottom of the list will scroll it to allow selection of a range larger than the displayed section of the list. When the left button is pressed any existing selection is cleared. To select several disjoint entries in the list press control and the left mouse button. The “Copy” button copies the current selection to the paste buffer.

Most commands require a contig identifier, which can be the contig name itself or the name/number of any reading within that contig. Gap5 always knows reading record numbers, but depending on the options used in tg_index when creating the assembly database the reading names may not be indexed. To specify a reading by record number, precede it by a # character, e.g. “#10000” means reading record number 10000, but “10000” means the contig or reading with name 10000.

Also any currently active dialogue boxes that require a contig to be selected can be updated simply by clicking on a contig in the contig selector or clicking on an entry in the "Contig Names" list box. For example, if the Edit contig command is selected from the Edit menu it will bring up a dialogue requesting the identity of the contig to edit. If the user clicks the left mouse button on a contig in the contig selector window, the contig editor dialogue will automatically change to contain the name of the selected contig. Some commands, such as the Contig Editor, can be selected from a popup menu that is activated by clicking the right mouse button on the contig line in the Contig Selector or clicking the right mouse button on the corresponding name within the "Contig List" list box. This simultaneously defines the contig to operate on and so the command starts up without dialogue.

Several contigs can be selected at once by either clicking on each contig with the left mouse button or dragging out a selection rectangle by holding the left mouse button down. Contigs which are entirely enclosed within the rectangle will be selected. Alternatively, selecting several contigs from the "Contig Names" list box will also result in each contig being selected. Selected contigs are highlighted in bold. Selecting the same contig again will unselect it.

The currently selected contigs are also kept in a 'list' named contigs.

1.2 Changing the Contig Order

The order of contigs is shown by the order of the lines representing them within the Contig Selector. The order of contigs can be changed by moving these lines using the middle mouse button, or Alt left mouse button. Several contigs may be moved at once by selecting several contigs using the above method. After selection, move the contigs with the middle mouse button, or Alt left mouse button, and position the mouse cursor where you want the selection to be moved to. Upon release of the mouse button the contigs will be shuffled to reflect their new order. The separator line at the point the contig was moved from increases in height.

The contig order is saved automatically whenever a contig is created or removed (eg auto assemble), including operations like disassemble which temporarily create contigs. The order can be saved manually using the Save Contig Order option on the File menu.

1.3 The Contig Selector Menus

The File menu contains only one command; "Exit". This simply quits the contig selector display.

The View menu gives access to the Results Manager (see ⟨undefined⟩ [Results Manager], page ⟨undefined⟩), allows contigs to be selected using a list box containing the contig names (See Section 1.1 [Selecting Contigs], page 1), and the list of selected contigs to be cleared.

The Results menu is updated on the fly to contain cascading menus for each of the plots shown when the contig selector is in its 2D Contig Comparator mode (see Chapter 2 [Contig Comparator], page 5). The contents of these cascading menus are identical to the pulldown menus available from within the Results Manager.

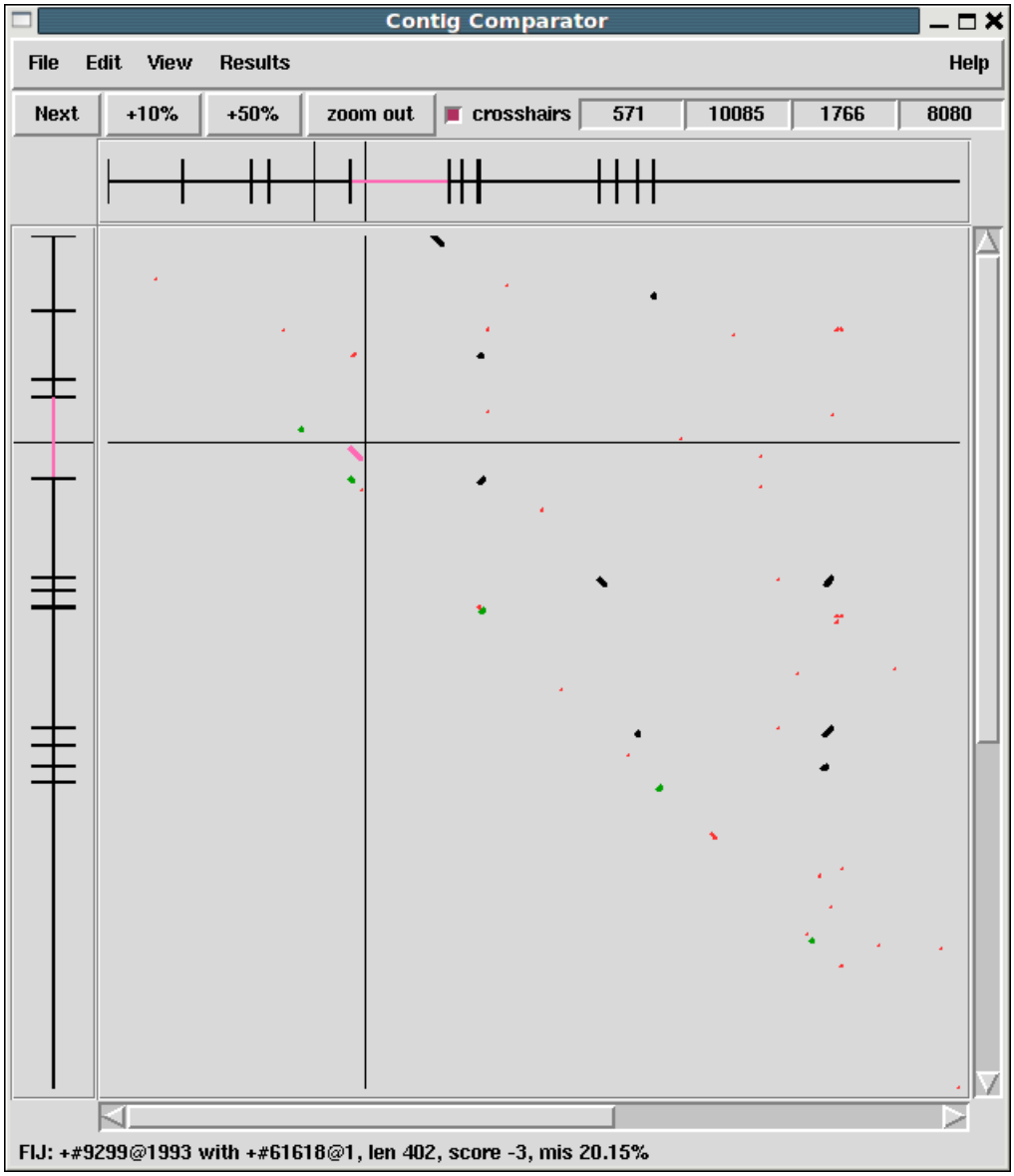
2 Contig Comparator

Gap5 commands such as Find Internal Joins (see Section 5.4 [Find Internal Joins], page 44) and Find Repeats (see Section 5.5 [Find Repeats], page 50) automatically transform the Contig Selector (see Chapter 1 [Contig Selector], page 1) to produce the Contig Comparator. To produce this transformation a copy of the Contig Selector is added at right angles to the original window to create a two dimensional rectangular surface on which to display the results of comparing or checking contigs. Each of the functions plots its results as diagonal lines of different colours. If the plotted points are close to the main diagonal they represent results from pairs of contigs that are in the correct relative order. Lines parallel to the main diagonal represent contigs that are in the correct relative orientation to one another. Those perpendicular to the main diagonal show results for which one contig would need to be reversed before the pair could be joined. The manual contig dragging procedure can be used to change the relative positions of contigs. See Section 1.2 [Changing the Contig Order], page 3. As the contigs are dragged the plotted results will be automatically moved to their corresponding new positions. This means that if users drag the contigs to move their plotted results close to the main diagonal they will be simultaneously putting their contigs into the correct relative positions.

By use of popup menus the plotted results can be used to invoke a subset of commands. For example if the user clicks the right mouse button over a result from Find Internal Joins a menu containing Invoke Join Editor (see Section 4.11 [The Join Editor], page 36) and Invoke Contig Editors (see Chapter 4 [Editing in gap4], page 17) will pop up. If the user selects Invoke Join Editor the Join Editor will be started with the two contigs aligned at the match position contained in the result. If required one of the contigs will be complemented to allow their alignment.

A typical display from the Contig Comparator is shown below. It includes results for Find Internal Joins in black, Find Repeats in red and Sequence Search in green. The currently highlighted item is shown in pink with a summary at the bottom of the screen. The orientation of this is from top-left to bottom-right indicating that the match is in the same orientation within both contigs (we can see some in the opposite orientation indicating that we need to reverse complement either of the two contigs before attempting any joins, although this will happen automatically). The crosshairs show the positions for a pair of

contigs. The vertical line continues into the Contig Selector part of the display, and the position represented by the horizontal line is also duplicated there.



2.1 Examining Results and Using Them to Select Commands

Moving the cursor over plotted results highlights them, and the information line gives a brief description of the currently highlighted match. This is in the form:

match name: *contig1_number*@*position_in_contig1*, with *contig2_number*@*position_in_contig2*, *length_of_the_match*

For Find Internal Joins the percentage mismatch is also displayed.

Several operations can be performed on each match. Pressing the right mouse button over a match invokes a popup menu. This menu will contain a set of options which depends on the type of result to which the match corresponds. The following is a complete list, but not all will appear for each type of result.

Information
: Sends a textual description of the match to the Output Window.

Hide
: Removes the match from the Contig Comparator. The match can be revealed again by using "Reveal all" within the Results Manager.

Invoke contig editors
Invoke join editors
: When invoked these options bring up their respective displays to show the match in greater detail.

Remove
: Removes the match from the Contig Comparator. The match cannot be revealed again by using "Reveal all" within the Results Manager.

One of the items in the popup menu may have an asterisk next to it. This is the default operation which can also be performed by double clicking the left mouse button on the match. For Repeat or Find Internal Joins matches this will normally be the Join Editor, or two Contig Editors when the match is between two points in the same contig.

The crosshairs can be toggled on and off and a diagonal line going from top left to bottom right of the plot can also be displayed if required. This is useful as a guide for moving the contigs such that their matches lie upon the diagonal line.

The "Results" menu on the contig selector window provides a similar mechanism of accessing results, but at the level of all matches in a particular search. This is simply a menu driven interface to the Results Manager window (see ⟨undefined⟩ [Results Manager], page ⟨undefined⟩), but containing only the results relevant to the contig comparator window.

2.2 Automatic Match Navigation

The "Next" button of the contig comparator window automatically invokes the default operation on the next match from the current active result. This provides a mechanism to step through each match in turn ensuring that no matches have been missed.

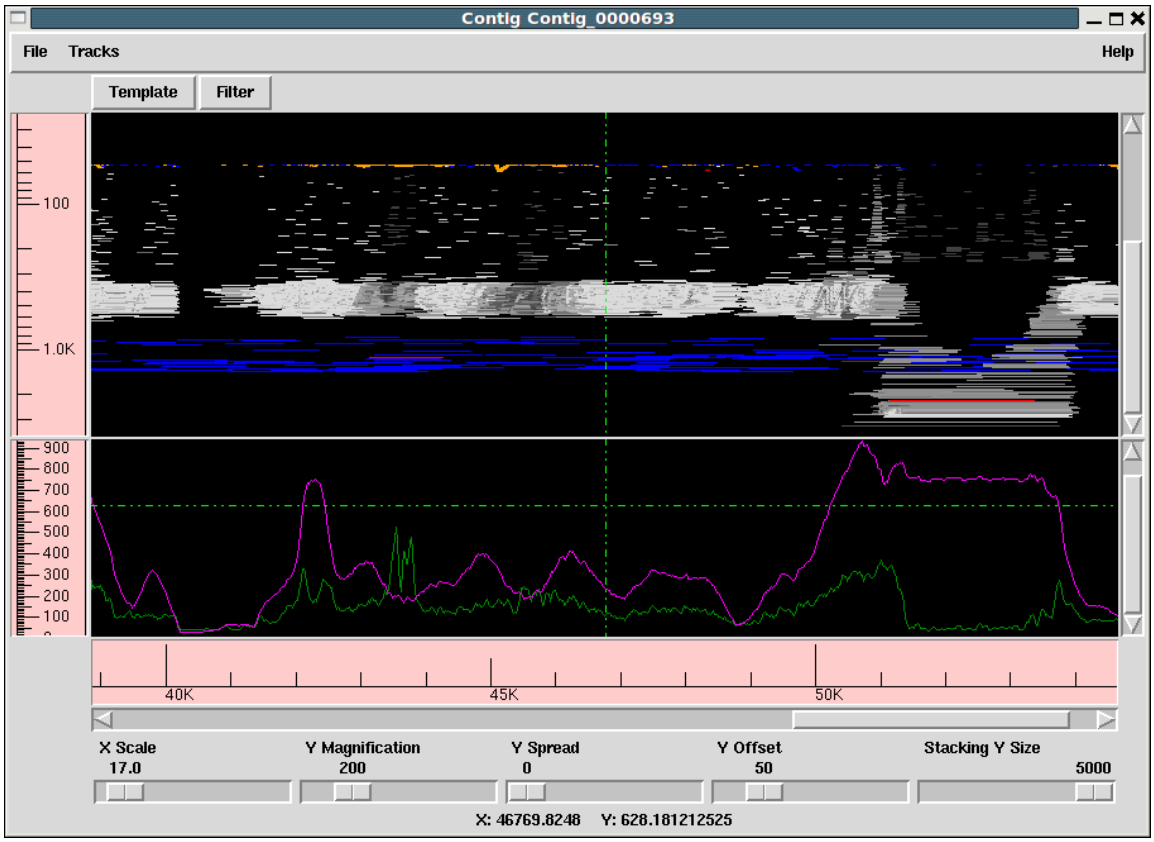
With a single result (set of matches) plotted, the "Next" button simply steps through each match in turn until all have been seen. Moving the mouse above the "Next" button, without pressing it, highlights the next match and displays brief information about it in the status line at the bottom of the window. To step through the matches in "best first" order, select the "Sort Matches" option from the relevant name in the Results menu. The exact order is dependent on the result in question, but is generally arranged to be the most interesting ones first.

Bringing up another result now directs "Next" to step through each of the new matches. To change the result that "Next" operates on, use the Result menu to select the "Use for 'Next'" option in the desired result. Alternatively, double clicking on a match also causes "Next" to process the list starting from the selected result.

The "Next" scheme remembers any matches that have been previously examined either by itself or by manually double clicking, and will skip these. To clear this 'visited' information select "Reset 'Next'" in the Results Manager.

3 Template Display

The template display is a graphical overview of a single contig. It allows us to see how much data we have, how long the fragments are and how they relate to each other (whether they are forming valid pairs).



The window consists of one or more tracks, by default showing the reading template layout at the top and a sequence / read-pair coverage plot at the bottom. The Tracks menu allows us to turn these on and off.

Below the main menu bar is a series of buttons that bring up new dialogues for controlling how the data is to be display and what is to be displayed.

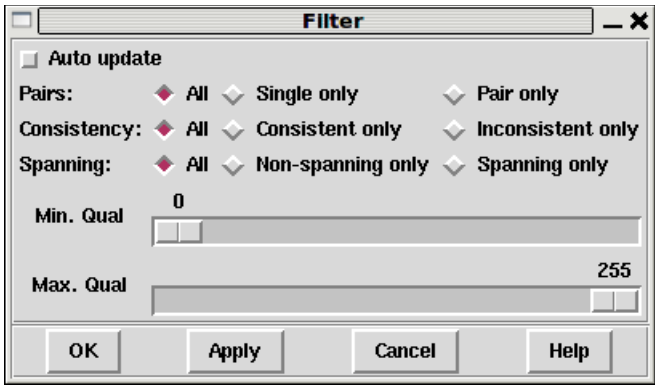
Then come a graphic plot per track. A cross-hair automatically tracks the cursor, indicating the X and Y coordinates (in appropriate units) in the status line at the bottom of the window. The track displays can be moved by either using the horizontal and vertical scrollbars at the bottom and right hand edges of the window, or by clicking and dragging the contents of the window. While dragging the display will not update to show newly visible regions of a contig until the left mouse button is released.

Finally the bottom contains a scrollbar and ruler for positioning and a series of controls. The X scale simply controls how many base-pairs of the contig are covered by he window.

The X scale number is arbitrary, but is interpreted in an exponential manner so it is easy to rapidly zoom in or zoom out. All other controls in the bottom panel do not affect the reading coverage track, so they are covered in the template track section below.

3.1 Filtering data

By default all templates are used for drawing the tracks, but there are times when we may wish to focus on specific problem data or to exclude it from our graphics.



The Filter button at the top of the Template Display brings up the dialogue shown above. Making changes to this dialogue either have an instant impact on the display (when "Auto update" is enabled) or instead only when we hit Apply or OK to dismiss the dialogue.

The Pairs: section allows us to select either reads on all templates, reads that are the sole read for that template, or reads that are paired on a template. Note that the definition of a pair here is strictly dependant on how many reads for a template are in the gap5 database rather than the library preparation strategy. So a paired-end template for which only one read is in the gap5 database (perhaps due to failure to map) is classified as "single".

The Consistency section can be used to select all, consistent only or inconsistent only data. This requires read-paired data (single reads cannot be inconsistent as so are considered as consistent). The interpretation of inconsistent currently is that the two reads of a pair do not point towards one another, but in future releases this is planned to check the correct orientation for that library type as for some constructions it is normal to have reads pointing in the same orientation.

The Spanning section governs whether to display read pairs with one read in this contig and the other read in another contig. Handling templates with more than two reads is still on-going work, but when finished a spanning read-pair will be one with any read not in this contig.

Underneath these are two sliders applied in addition to the above filters. They allow removal of any read or read-pair (depending on the type of data being plotted) with a mapping quality outside the selected range.

3.2 Template plot

This is the main body of the template display window. The default plot will be showing read-pairs, mainly coloured by mapping quality with the insert size governing the Y coordinate. Larger inserts are at the bottom of the track while shorter ones are at the top.

The colours used are as follows:

blue This is a template with only one reading present. It could be either a pair with one end not in this assembly, or a true single-ended sequencing experiment. The horizontal size of the line is now the length of the individual sequence rather than the computed length of the insert.

orange This is a template with one reading present in another contig. The size of the line is derived from the size of the data in this contig (typically a single reading).

red This template is considered as inconsistent in some manner, typically due to the relative position and orientation of the forward and reverse sequences being incorrect.

grey (variety of)
Any consistent read-pair is coloured by the mapping quality, by default using the average of the individual sequence mapping qualities. Lighter shades represent higher mapping qualities.

The row of scale bars at the bottom of the window control how data is to be plotted. They are:

X Scale Controls how many base-pairs in the contig to plot. Higher values indicate more base pairs, but with an exponentially growing scale.

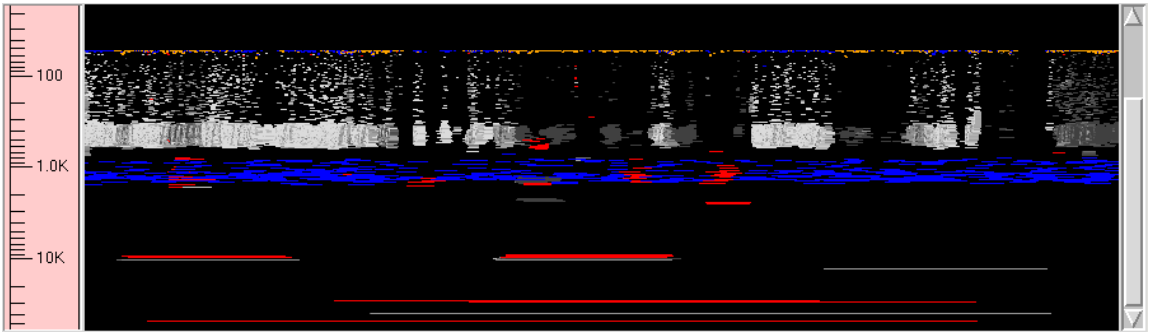
Y Magnification
Governs the amount of vertical space consumed by the template track. This has no impact on the depth track.

Y Offset Adds a small shift to the Y position of data prior to plotting. This is of little use unless Separate Strands has also been selected, where upon this allows the two halves of the plot to be brought closer together. (Effectively meaning the a plot can go from -1000 to -100 and +100 to +1000 instead of -1000 to +1000 with a blank area in the middle if our sequences are a minimum of 100 bases long.)

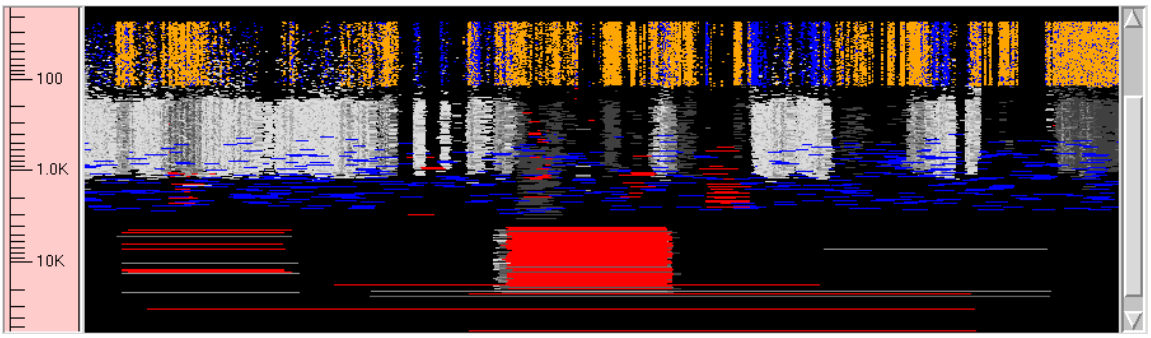
Stacking Y Size
Only of use in Stacking Y-Position mode. This vertically groups together data of similar length, allowing a basic approach of separating short-read and long-read technologies. The Y layout is performed in steps of "Stacking Y Size". To pack reads tightly together regardless of length, set this to the maximum value possible.

Y Spread This adds a small perturbation to the computed Y coordinates of lines in the template track. When the Y coordinate is derived based on the insert size of the read-pair it is not always clear whether a line represents a single item or

many items stacked perfectly on top of one another. The Y spread control compensates for this.



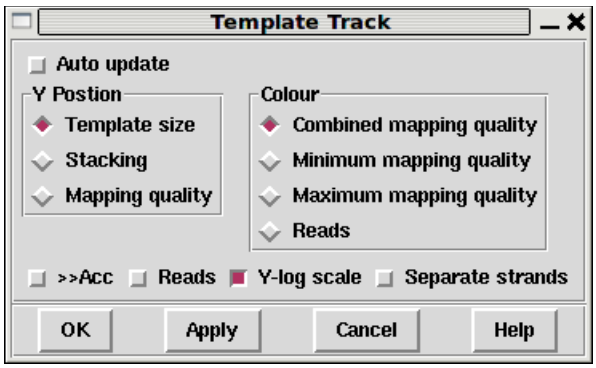
Template track with Y spread of 0.



Template track with Y spread of 50.

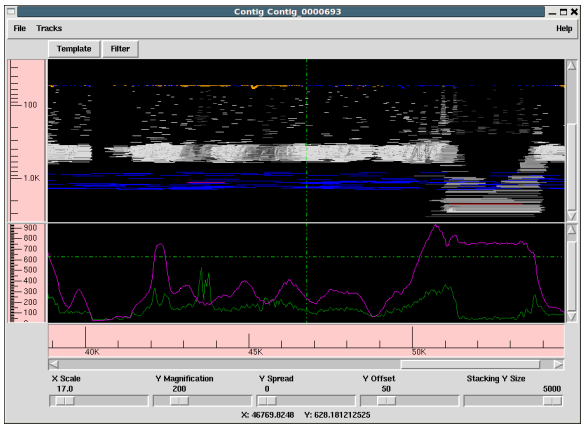
3.2.1 Controlling The Y Layout.

The layout and type of data in the template track can be controlled using the Template button at the top of the main template display window.



The Y Position section controls how the Y coordinates are computed when plotting data (with X being tied to the position in the assembly or reference). It can be one of three settings.

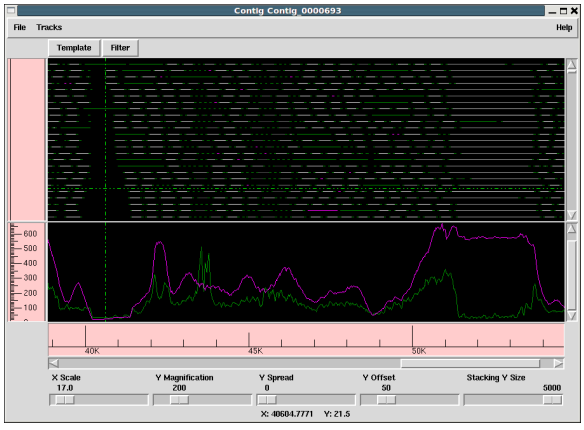
Template size



The default mode. The size of an object is defined to be the number of bases it spans. This is normally the size of a read-pair, or if the pair spans contigs or if only readings are shown it is the size of a single reading instead. Larger objects are at the bottom of the window. This Y method very clearly reveals indels in a mapped assembly. It sometimes also sometimes reveals misassemblies.

Given that items of identical size will stack on top of one another, of particular use to this display mode is the Y Spread control in the main window.

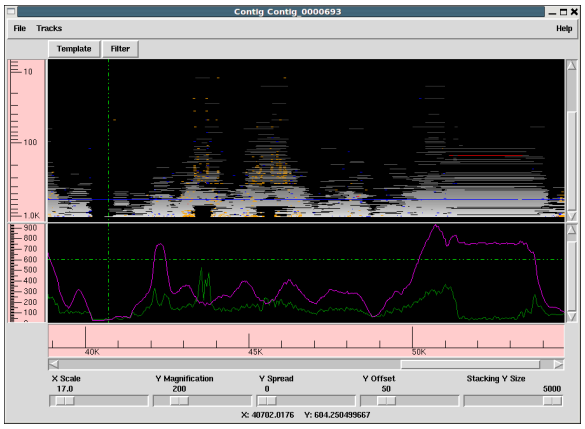
Stacking



A more traditional view - each and every item is allocated its own non-overlapping Y coordinate (although low Y magnifications may imply these are drawn at the same Y pixel).

It is still possible to partially group items by their insert size using the “Stacking Y Size” control in the main window.

Mapping Quality



Finally we can display data collated by the mapping score. This is typically only available for mapped assemblies. This plot sometimes helps to reveal regions where all the data present is of poor mapping quality, indicating a likely repeat.

Adjacent to the Y Position frame is the Colour frame. This controls the colour of the lines drawn in the template display rather than their location.

Combined mapping quality
Minimum mapping quality
Maximum mapping quality

For templates with multiple reads visible, we have a variety of mapping qualities. Often these individual sequence mapping qualities will differ, but we wish to draw a single line for the template with a single colour. These three methods control whether we take the average, minimum or maximum values from the individual sequences on this template.

Reads The line typically represents the entire span of the insert, but we may not have sequence data for all of the template. This colour mode will also draw the portions of the template that we have known sequence for, in green for forward strand sequences and magenta for reverse strand sequences. Any remaining portion of template between the reads is drawn using the combined mapping quality.

At the bottom of this dialogue is a row of check buttons.

“>>Acc” enables accurate mode, but be warned this can be very slow. When the template display is drawn it fetches all data within the visible portion plus a little bit ether side. From this reads from the same template are paired up. However when a template spans a substantially larger range than is shown we may only have fetched one read for this template. We do know that such a template forms a pair, but we do not know the exact location of the other end or even whether it is in this contig. The assumption is that it is not, and the template is drawn in orange. Enabling accurate mode will work out the precise location of the other end and if it is present elsewhere within this contig then the insert size will be correctly determined and the plot adjusted accordingly.

The “Reads” checkbutton (not to be confused with the Reads colour selector) disables all drawing of read-pairing and template lines, instead drawing lines to represent the known DNA sequence instead.

“Y-log scale” controls whether we plot our Y values using log or linear scales.

“Separate strands” attempts to classify all templates as coming from the top or bottom strand of DNA (based on the orientation of the sequences on that template, although sometimes these are conflicting). It then splits the plot in two, forming an approximate mirror image. This may be of use in some transcriptome sequencing experiments.

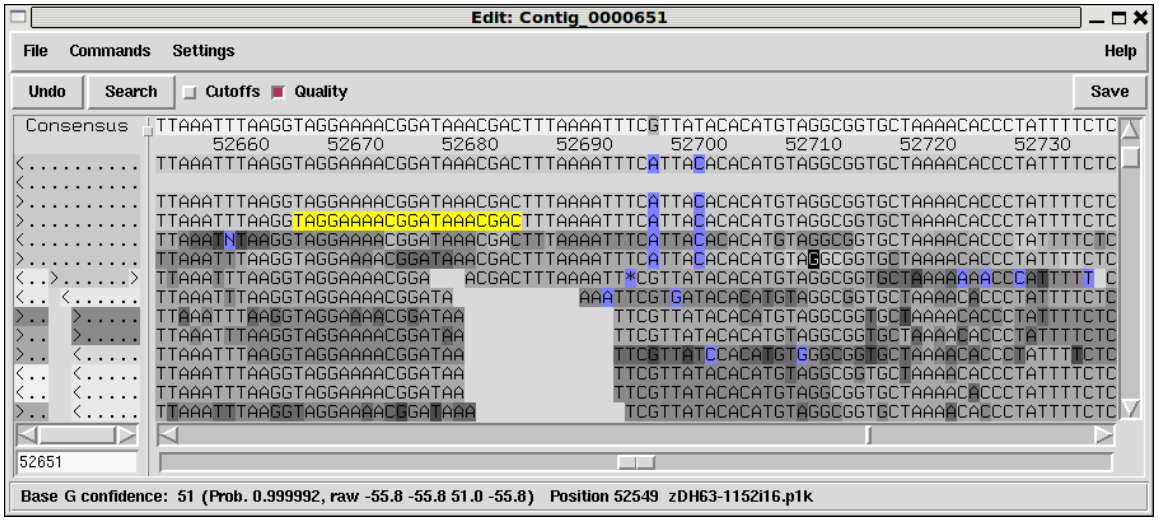
3.3 Depth / Coverage Plot

The depth track shows coverage of both individual readings and read-pairs, where a read-pair counts as +1 coverage over the entire length it spans rather than just the portion directly sequenced.

The filter options for (in)consistent read pairs also apply here, giving the option to only show depth of consistent pairs.

4 Editing in Gap5

The gap5 Contig Editor is designed to allow rapid checking and editing of characters in assembled readings. Very large savings in time can be achieved by its sophisticated problem finding procedures which automatically direct the user only to the bases that require attention. The following is a selection of screenshots to give an overview of its use.

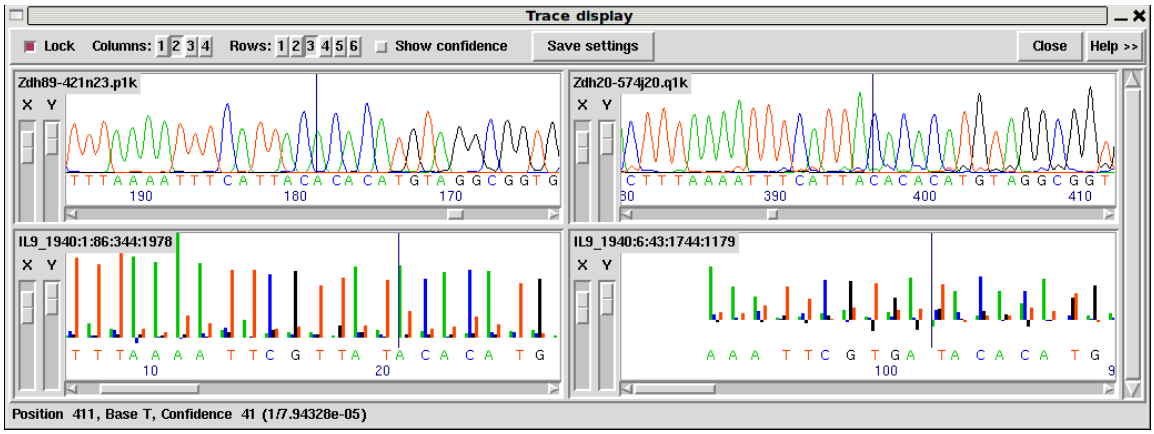


The figure above shows a screendump from the Contig Editor showing the consensus for a small region of a contig and the aligned reads. The main components are, top-most menu bar; common buttons and controls beneath this; the main name and sequence panels to the left and right; scrollbars and jog-control; a status text line at the bottom.

The names panel on the left can show either reading names or a small ASCII diagram representing their position, orientation and mapping quality as a grey-scale. The sequences to the right in the screenshot has base quality shown in grey (dark being poor, light being good) with disagreements to the consensus at the top shown in blue. The consensus line also shows base qualities. You may notice we have a mixture of long and short sequences, with the longer ones being at the top. This screenshot is from a mixed assembly of Illumina short-read data and ABI Sanger-method capillary sequences.

One base is drawn in inverse video (a “G”). This is the current location of the editing cursor. We can move this we arrow keys or clicking with the left mouse button. It behaves much like the editing cursor in a word processor and need not be visible in the portion of the contig we are viewing.

Also visible is a set of bases coloured yellow. These are an OLIGO annotation. Gap5 supports a wide variety of annotation types (often also referred to as "tags"). These are covered later in more detail.



This figure is an example of the Trace Display showing three capillary traces and an Illumina trace from readings in the previous Contig Editor screendumps. Note that this demonstrates the possibility of showing the raw trace data for new short-read sequencing technologies, but typically this is not available due to the high storage size.

4.1 Moving the visible segment of the contig

The contig editor displays only one segment of the entire contig, although several contig editors can be in use at once. Below the sequence is a scrollbar and below that a "jog" control. The scrollbar behaves as expected, allowing rapid positioning anywhere within the contig using the middle mouse button or left-clicking and dragging the slider. However with extremely long contigs (for example 100Mb) it can become tricky to move by the desired amount. Each pixel on the scrollbar may represent 100Kb worth of data, so dragging the scrollbar is only approximate positioning. Equally so clicking in the trough to move a screen-full at a time can be too small. This is where the jog-control can be of use.

By default this is always centred. Clicking and dragging this left or right starts to scroll the editor, at a speed proportional to how far away from the centre the jog is dragged. Releasing the mouse button stops automatically scrolling and recentres the jog control.

The final, more precise, manner of positioning the editor view is with the text entry box in the bottom left corner. Type in any coordinate here and press return to jump straight to that location. Note however that Gap5's coordinates are currently always in padded form; that is to say that a gap in the consensus caused by an insertion in one of the aligned sequences is still counted as a base position.

For particularly deep displays the vertical scrollbar on the right edge of the window will also be useful. While scrolling in X, the editor attempts to keep the same sequences visible on screen. To do this it may automatically adjust the Y scrollbar for you due to changing layout of sequences. (By default the top-most sequence is always the sequence that starts furthest left and the bottom most is the sequence starting furthest right.)

If you have a mouse wheel, this may also be used for small scrolling. By itself it scrolls in Y one sequence at a time. With the Control key held down it scrolls in larger increments. Using the Shift key in conjunction with the mouse wheel scrolls in X instead, with Shift+Control to scroll in larger increments.

The displayed portion of the contig is separate from the current location of the editing cursor. This is displayed as a black rectangle with typically a light coloured letter inside it. Any editing keys operate on the base underneath this or to the base immediately preceding it for Delete. We cover the topic of editing later (see Section 4.3 [Editing], page 21), however moving the editing cursor is also another way of scrolling the editor.

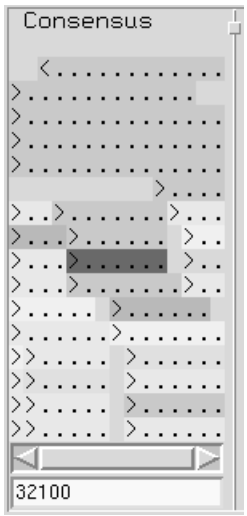
Finally the Page Up and Page Down keys scroll the editor left or right by 1Kb. Used with Shift the moves in increments of 10Kb, with Control in increments of 100Kb and with both Shift and Control in increments of 1Mb.

FIXME: Add Home and END too for start/end of contig?

4.2 Names

At the left side of the editor window is the "names panel". This either displays an ASCII pictorial summary of the sequence layout or the actual sequence names themselves depending on the settings in use. Between the names panel and the sequences panel is a vertical line, visible at the right edge of the above image. This can be dragged left and right to adjust the proportion of display dedicated to the names and sequence panels.

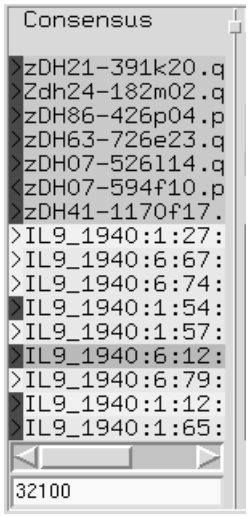
The default name display looks like this:



This plot is a mini diagram of the way the sequences overlap. Here the > and < symbols represent the start of sequences, assembled on either the forward or reverse strand, with the ... sections reflecting their relative lengths. The background shading indicates the mapping quality of the sequence (which may not be available in many cases, depending on how the assembly was derived). This should indicate the likelihood that the sequence has been assembled to the correct point. Sequence that appears to map elsewhere (eg due to

a repeat) will be dark grey while unique sequence will be light grey or white. Moving the mouse cursor over a sequence will tell you the precise mapping quality along with additional information such as the sequence name, the technology used (Sanger, Illumina, 454, etc), and whether it is part of a pair of sequences.

In the editor Settings menu is a checkbox labelled "Pack Sequences". When checked we permit multiple sequences to be drawn in the same row. Unchecking this reverts to the Gap4 style of display where each sequence has its own dedicated row. This also has an affect on the names panel, which switches to showing the sequence names, as below.



This still uses the > and < symbols to reflect strand and grey scales for representing the mapping quality. The > and < are now also coloured independently. A dark grey > or < indicates that the read is not paired, while light means it forms a pair. (In future this may be expanded to indicate read-pair consistency and pairs spanning contigs.)

At the bottom of the names panel is an editable text field containing the current "padded" display position. This is updated automatically as we scroll through the editor, or it can be used to jump the editor to specific points by typing in a new location and pressing the enter key.

In both display modes, pressing the right mouse button brings up a context sensitive menu containing operations relevant to that specific sequence. This may contain the following commands.

Copy to clipboard
: This copies the sequence name to the clipboard for use in a subsequent paste operation. Note that there is no visual cue that this has happened. The same function may also be achieved by left-clicking and dragging the mouse horiztonally, as if attempting to highlight a region of text.

Goto...
: This lists other sequences sharing the same template, such as the other end of a read-pair. Selecting this command will jump the editor to the left-most base in that sequence. If the sequence is in another contig then a new editor will be

created, unless one already exists for that contig in which case that other editor will be moved accordingly.

4.3 Editing

Editing can take up a significant portion of the time taken to finish a sequencing project. Gap5 has a selection of searches (see Section 4.6 [Searching], page 27) designed to speed up this process. The problems that require most attention are conflicts between good bases. Where base confidence values are present it should be unnecessary to edit all conflicting bases as, generally, this will amount to adjusting poor quality data to agree with good quality data in which case the consensus sequence should be correct anyway.

Pads in the consensus should not be considered a problem requiring edits because it is possible to output the consensus sequence (from the main gap5 File menu) with pads stripped out. Obviously poorly defined pads (a mixture of several alignment padding characters and real bases) require checking in the same manner as other poorly defined consensus bases.

To change a base simply overtype with a new base call, one of a,c,g or t in lowercase. Alternatively a base can be changed to an alignment padding character by pressing "*". These new bases and pads automatically get given a quality value of 100, but see below for how to adjust this. The consensus cannot be edited in this manner.

To insert a gap into sequence press "i". At present only alignment pads can be inserted, not bases, although the pads can subsequently be edited to turn them into bases. The "i" key also permits insertions of gaps into the consensus, which it achieves by inserting into every sequence aligned at that position.

Bases may be deleted by pressing the Delete or Backspace key. This deletes the base immediately to the left of the current editing cursor. Note that if Delete or Backspace is pressed with the editing cursor on the consensus this removes an entire column of data. Deleting anything other than alignment padding characters (either in sequences or the consensus) is a dangerous operation needing careful thought. To prevent accidental removal of data therefore, to delete anything other than "*" you must press Control in conjunction with Delete or Backspace.

4.3.1 Moving the editing cursor

Nearly all editing operations happen at the location of the editing cursor. This cursor appears as a black block containing the base in a light colour, instead of the usual black base on a light background.

The simplest mechanism of moving the cursor is using the left mouse button. Alternatively the following keys can be used.

```
Left arrow or Control b        Move left one base
Right arrow or Control f       Move right one base
Up arrow or Control p          Move up one base
Down arrow or Control n        Move down one base
Control a                      Move editing cursor to start of sequence
Control e                      Move editing cursor to end of sequence
Meta or Alt <                  Move editing cursor to start of contig
Meta or Alt >                  Move editing cursor to end of contig
```

If any of these move the editing cursor outside of the visible region, the editor will scroll to accommodate. Control-a and Control-e with the editor on the consensus line will also jump to the start and end of the contig.

If "Cutoffs" are shown (see Section 4.3.4 [Adjust the Cutoff Data], page 22) the cursor may be placed in the cutoff data too. Note that turning off displaying cutoff data would then leave the editor on an invisible base, so it is moved to the consensus line instead.

4.3.2 Adjusting the Quality Values

Each base has its own quality value. Assembly will allow only values between 1 and 99 inclusive. A quality value of 0 means that this base should be ignored. A quality value of 100 means that this base is definitely correct and the consensus will be forced to be the same base type and will be given a consensus confidence of 100. If two conflicting bases both have a quality of 100 the consensus will be a dash with a confidence of 0.

Newly added bases or replaced bases are assigned a quality of 100.

Several keyboard commands are available to edit the quality value of an individual base.

```
[                          Set quality to 0 and move cursor right
]                          Set quality to 100 and move cursor right
Shift   Up-Arrow           Increment quality by 1
Control Up-Arrow           Increment quality by 10
Shift   Down-Arrow         Decrement quality by 1
Control Down-Arrow         Decrement quality by 10
```

Finally note that quality values can also be made visible by clicking on the "Quality" checkbutton at the top of the editor. This shows the quality by use of a grey scale.

4.3.3 Adjusting the alignment coordinates

On rare occasions we may need to move an entire sequence a small amount to achieve an optimal alignment, rather than simply inserting or deleting pads.

This is achieved by using Control plus the left and right arrow keys while the editing cursor is anywhere on the sequence.

4.3.4 Adjusting the Cutoff Data

Sequences typically consist of a good quality "used" portion and poor quality "clipped" or "cutoff" portions at the 5' and 3' ends of the sequence. Although for short sequencing technologies it's quite likely we have no cutoff data at all. The reason for this is that the low quality ends of sequences may have a sufficient number of errors that the sequence

alignment algorithms are no longer confident they have the correct bases aligned, or event that the sequence simply disagrees too much.

By default these are not shown, although you may see blank lines in the display as room is left for this sequence even when it is not visible. The cutoff data may be displayed by pressing the "Cutoffs" check-button at the top of the editor. The cutoff sequence will then be displayed in grey. We call the boundary between the cutoff data and the used data the cutoff position. These positions can be adjusted by pressing the "<" (left cutoff) or ">" (right cutoff) keys. In both cases the cutoff point is between the base with the editing cursor and the base to the left of the editing cursor.

4.3.5 Summary of Editing Commands

A brief summary of these editing operations can be seen below:

Key	Location	Action
a,c,g,t,*	Reading	Change base
i	Reading	Insert pad
delete	Reading	Delete * to left
Ctrl delete	Reading	Delete any base to left
Control Left	Reading	Move reading left
Control Right	Reading	Move reading right
[	Reading	Set quality to 0
]	Reading	Set quality to 100
Shift Up	Reading	Incr. quality by 1
Shift Down	Reading	Decr. quality by 1
Ctrl Up	Reading	Incr. quality by 10
Ctrl Down	Reading	Decr. quality by 10
<	Reading	Set left cutoff
>	Reading	Set right cutoff
i	Consensus	Insert column of pads
delete	Consensus	Delete * to left
Ctrl delete	Consensus	Delete any base to left

4.4 Selections

It is possible to highlight an area of a reading or the consensus sequence in preparation for performing some further action upon it. Such examples of actions are: creating annotations and pasting into a new window. We call these highlighted areas "selections". They are displayed as an underlined region.

The simplest way to make a selection is using the left mouse button. Pressing the mouse button marks the base beneath the cursor as the start of the selection. Then, without releasing the button, moving the mouse cursor adjusts the end of the selection. Finally releasing the button will allow normal use of the mouse again. If while marking a selection we reach the edge of the window then the editor will automatically start scrolling for us.

Sometimes we may wish to make a particularly long selection, or just extend an existing selection after we've already released the mouse button. This can be done by using shift left mouse button to adjust the end of the selection. Hence we can mark the start of the selection using the left button, scroll along the contig to the desired position, and set the end using the shift left button.

The selection is stored in the "clipboard". This allows for the usual "cut and paste" operations between applications, although the contig editor only supports this in one direction (as it is not possible to "paste" into the window). The mechanism employed for this follows the usual X Windows standard of using the middle mouse button.

A quick summary of the mouse selection commands follows.

```
Left button                          Position editing cursor to mouse cursor
Left button (drag)                   Mark start and end of selection
Shift left button                    Adjust end of selection
Middle button (in another window)    Copy selected sequence
```

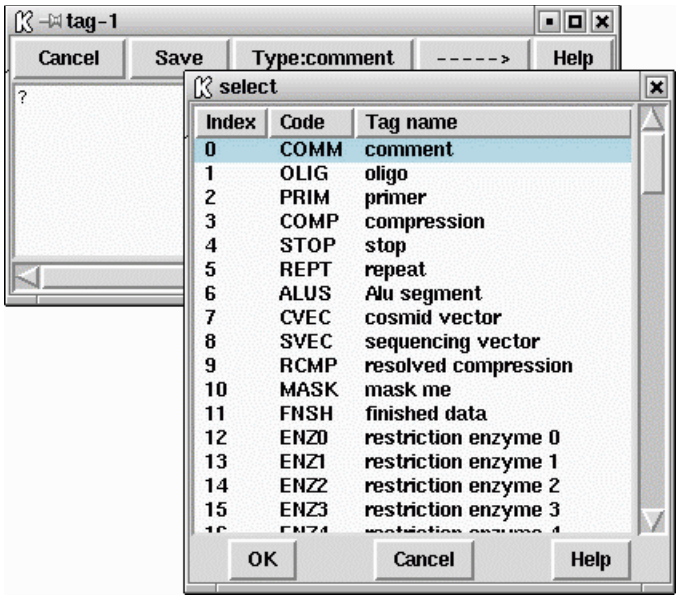
4.5 Annotations

Annotations (or tags) can be placed at any position on readings or on the consensus. They are usually used to record positions of primers for walking, or to mark sites, such as repeats or compressions, that have caused problems during sequencing. Each annotation has a type such as "primer", a position, a length, a strand (forward, reverse or both) and an optional comment. Each type and strand has an associated colour that will be shown on the display. For information on searching for annotations see Section 4.6.2 [Searching by Tag Type], page 27, and Section 4.6.1 [Searching by Annotation Comments], page 27.

FIXME: not all of the tag editor features are supported yet; specifically the Move/Copy functionality and storing strand information.

To create an annotation, make a selection and then select "Create Tag" from the contig editor commands menu. See ⟨undefined⟩ [The Commands Menu], page ⟨undefined⟩. This

will bring up a further window; the "tag editor" (shown above). The "Type:" button at the top of the editor invokes a selectable list from which tag types can be chosen. See below.



Use this to select the desired type of annotation.

[FIXME: To implement. Next the strand of the annotation can be selected. This will be displayed as one of "<——>", "<——" and "——>".] The comment (the box beneath the buttons) can be edited using the usual combination of keyboard input and arrow keys. The "Save" button will exit the tag editor and create the annotation. To abandon editing without creating the annotation use the "Cancel" button.

To edit an existing annotation, position the editing cursor within a annotation and select "Edit Tag" from the commands menu. This will be a cascading menu, typically showing one tag. If multiple tags coincide at the same sequence position you will be able to chose which tag to edit. Once again the tag editor will be invoked and operates as before. The **F11** key is also a shortcut for editing the top-most tag underneath the editor cursor. When editing, the "Save" will save the edited changes and "Cancel" will abandon changes.

Removing a annotation involves positioning the editing cursor within an annotation and selecting "Delete Tag" from the commands menu. As with "Edit Tag" this is a cascading menu to allow you to chose which tag at a specific point to delete. The **F12** key is a shortcut to remove the top-most tag underneath the editor cursor.

As usual, "undo" can be used to undo any of these annotation creations, edits and removals.

Some tags may contain graphical controls instead of the usual text panel. These are encoded with the master gap4 tag database (*GTAGDB*) by specifying the default tag text to be a piece of "ACD" code. A full description of the (modified for gap4/5) ACD syntax is not available currently, but it is strongly modelled on the the EMBOSS ACD syntax which has documentation at

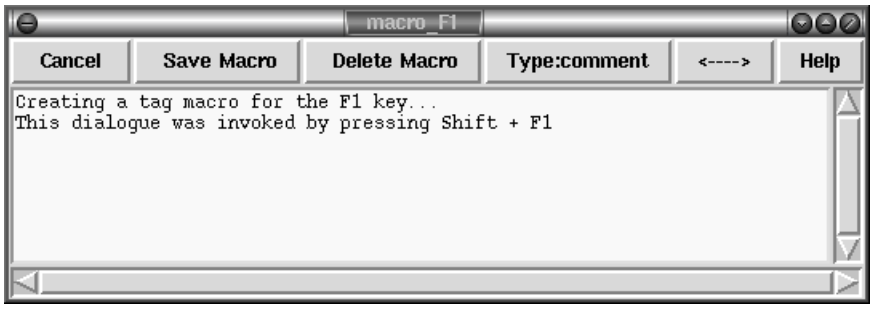
http://www.emboss.org/Acd/index.html .

It is possible to add your own tag types by modifying either the system *GTAGDB* file or creating your own *GTAGDB* file in your home directory (for all your databases) or the current directory (for just those in that directory).

For rapid editing and deleting the F11 and F12 keys may be used. These edit and delete the top-most tag underneath the editing cursor. If you wish to edit or delete the tag underneath the mouse cursor instead (and hence save a mouse click) use Shift F11 and Shift F12 for edit and delete.

The Control-Q key sequence may be used to toggle the displaying of tags. Pressing it once will prevent all tags from being displayed in the editor. This is sometimes useful to see any colouring information underneath the tag. Pressing Control-Q once more will redisplay them.

4.5.1 Annotation Macros



For rapid annotating a series of 10 macros may be programmed. Press Shift and a function key between F1 and F10 to bring up the macro editor. This look much like the normal tag editor except that **Save** is replaced with **Save Macro** and saving does not actually create a tag on the sequence. To use the macro, highlight the bases you wish and press the function key corresponding to that macro - F1 to F10. For a single base pair tag you do not need to underline a region as the tag will automatically cover the base underneath the editing cursor. To remember these permanently use the "Save Tag Macros" option in the "Settings" menu.

If you have an existing tag you wish to rapidly duplicate to many places, use Control plus a function key to copy the tag underneath the editing cursor to that numbered tag macro. This is simply a short cut for Shift and the function key, but without needing to manually replicate the tag type and textual comment.

You may find that some function keys are already programmed to do other things (such as raise or lower windows), depending on the windowing environment in use. If this is the case either modify the configuration of your windowing system or simply use another macro key.

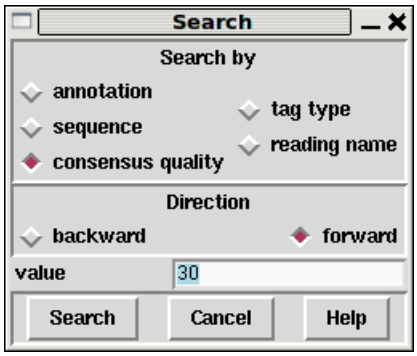
```
Shift   F1-F10          Create a tag macro via a dialogue window
Control F1-F10          Create a tag macro from tag at editor cursor
F1-F10                  Apply a tag macro (create a real tag)
```

4.6 Searching

The contig editor's searching ability and its links to the consensus calculation algorithm are crucial in determining the efficiency with which contigs can be checked and corrected. The consensus is calculated "on the fly" and changes in response to edits. For editing, the most important search functions are those which reveal problems in the consensus whilst ignoring all bases that are adequately well determined. The standard search type is therefore by consensus quality. By default this is done in the forward direction and for a quality value of 30, although this is configurable by changing the collowing lines in the gap5rc file.

```
set_def CONTIG_EDITOR.SEARCH.DEFAULT_TYPE       consquality
set_def CONTIG_EDITOR.SEARCH.DEFAULT_DIRECTION  forward
set_def CONTIG_EDITOR.SEARCH.CONSQUALITY_DEF    30
```

Pressing the "Search" button brings up a separate search window. This allows the user to select the direction of search, the type of search, and a value to search on. The value is entered into a value text box, then pressing the "search" button performs the search. If successful, the cursor is positioned accordingly.



The Control-s and Control-r key bindings in the editor are equivalent to searching for the next or previous match. Both key bindings will bring up the search window if it is not currently displayed (and not search), otherwise they perform the search currently selected in that window.

As is described below, there are several search modes.

4.6.1 Search by Annotation Comments

This positions the cursor at the start of the next tag which has a comment containing the string specified in the value box. The search performed is a regular expression search, and certain characters have special meaning. Be careful when your string contains ".", "*", "[", "]", "\", "^" or "$". The search can be performed either forwards or backwards from the current cursor position. Searching with an empty value will find all tags.

4.6.2 Search by Tag Type

This positions the cursor at the start of the next tag of the specified type. To change the type, click on the currently listed tag type, which displays a tag type selection dialogue.

The search can be performed either forwards or backwards of the current cursor position. To find all tags, use "Search by Annotation Comments", with an empty text box.

4.6.3 Search by Sequence

This positions the cursor at the start of the next segment of sequence that matches the value specified in the text box. The search is case insensitive, ignores pads, and can allow a specified number of mismatches. Unlike Gap4, Gap5's sequence search only looks in the consensus sequence. It also operates either forwards or backwards from the current editing cursor position.

4.6.4 Search by Consensus Quality

This positions the cursor on the consensus at the next position where the quality of the consensus is below a given threshold. The quality threshold should be entered into the value box and should be within the range of 0 to 100 inclusive.

4.6.5 Search by Reading Name

This positions the cursor at the left end of the reading specified in the value text box. Note that not all reading names may be indexed by Gap5 and that the search will not find unindexed names. See `tg_index -t` for information on creating Gap5 databases with reading name indices.

The reading name has to be an exact match and so currently does not find prefix strings. If multiple sequences exist with the same name (which should be strongly discouraged) then it is undefined which will be found first.

4.7 The Settings Menu

The purpose of this menu is to configure the operation of the contig editor. Settings can be saved using the "Save settings" button, but this does not save any tag macros. These may be saved separately using the "Save Macros" option. Settings for the following options can be changed.

- Highlight Disagreements
  - By dots
  - By foreground colour
  - By background colour
  - Case sensitive
- Set quality threshold
- Pack sequences
- Hide annoations
- Save tag macros
- Save settings

4.7.1 Highlight Disagreements

This toggles between the normal sequence display (showing the current base assignments) and one in which those assignments that differ from the consensus are highlighted. It makes scanning for problems by eye much easier.

Several modes of highlighting are available: "By dots" will only display the bases that differ from the consensus, displaying all other bases as full stops if they match or colons if they mismatch but are poor quality. The definition of poor quality here can be adjusted using the "Set quality threshold" option of the Settings menu. The base colours are as normal (ie reflecting tags and quality).

Highlight disagreements "By foreground colour" and "By background colour" displays all base characters, but colours those that differ from the consensus. Bases which differ by are below the difference quality threshold are shaded in light blue while high quality differences are dark blue. This allows easier visual scanning of the context that a difference occurs in, but it may be wise to disable the displaying of tags (hint: control-Q toggles tags on and off).

Finally the "Case sensitive" toggle controls whether upper and lower case bases of the same base type should be considered as differences.

4.7.2 Pack Sequences

This controls whether the editor allocates one row per sequence or whether it is permitted to pack multiple sequences onto a single row, assuming they do not overlap.

The latter allows for a more compact plot which is desirable when dealing with short sequences, however it has the side effect that the reading names can no longer be listed in the names panel to the left.

4.7.3 Hide Annotations

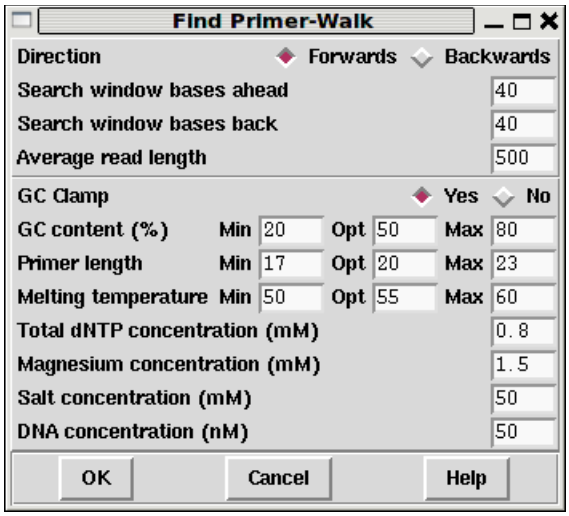
Sometimes we need to see the background shading underneath an annotation, for example to see the base quality or if we have Highlight Disagreements turned on using the *by background colour* mode. This option simply hides all annotations from display until it is selected again to reveal them once more.

The Control-Q keyboard shortcut has the same effect.

4.8 Primer Selection

The "Find Primer Walk" function from the Commands menu is an interface to the Primer3 program (builtin to Gap5 so it does not need an external installation). Currently it only allows for selection of a single internal oligo suitable for "walking" along a template. It is designed for manual finishing work and is not appropriate for automatic finishing. Future plans are to add PCR support.

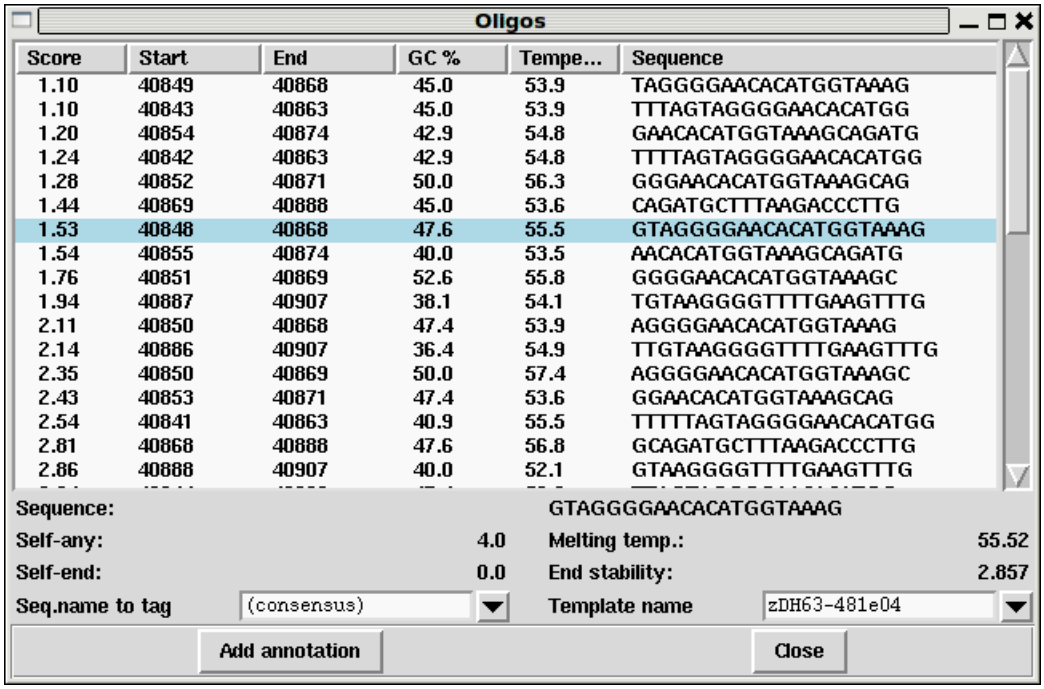
The command brings up its own dialogue window.



The top portion of this window controls where to look for primers. By default it will be either side of the editing cursor location. We also specify here what strand we wish to run our experiment on.

Below this are a series of Primer3 parameters. Please see the Primer3 documentation for a full description of these.

Upon hitting OK, and assuming that some primers can be found, a new window showing the available choices is presented.



The primers show are sorted by Primer3 score, with lower being better. Clicking on any of the other headings in the table allows the data to be re-sorted by that column. Clicking the left mouse button on any line will show the location of this primer in the main editor window as an underlined region. It also updates the bottom half of the Oligos window with further details.

At the bottom of the window are two editable selections. The left most labelled "Seq. name to tag" allows us to pick a sequence we wish to place an oligo (`OLIG`) annotation on, which defaults to the consensus sequence. The right selection box labelled "Template name" is an list of identified templates at this region, however this is not necessarily exhaustive as it only includes the sequences at this position and may miss some read-pairs that span this region. If you have a specific template in mind you can also type in the name of it to here.

Pressing the "Add annotation" button then creates an oligo annotation. The text associated with the annotation will depend on the primer chosen, but an example follows.

```
Sequence        AACACATGGTAAAGCAGATG
Template        zDH64-714h06
GC              40.0
Temperature     53.45
Score           1.54377204143
Date_picked     Thu Aug 12 17:31:18 BST 2010
Oligoname       ??
```

4.9 Traces

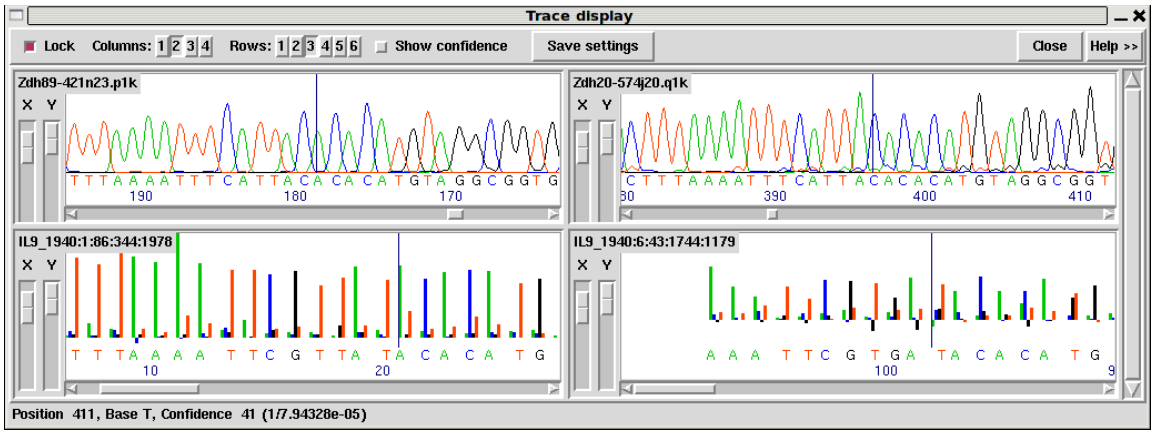
The original trace data from which the readings where derived can be displayed by double clicking (two quick clicks) with the left or middle mouse button on the area of interest. Control-t has the same effect. The trace will be displayed centred around the base clicked upon and the name of the reading in the contig editor will be highlighted. Double clicking on the consensus displays traces for all the readings covering that position.

Moving the mouse pointer over a trace base causes the display of an information line at the bottom of the window. This gives the base type, its position in the sequence, and its confidence value.

There are two forms of trace display which are selected using the "Compact" button at the top of the Trace display. The compact form differs by not showing the Info, Diff, Comp. and Cancel buttons at the left of each trace.

Note that gap5 does not store the trace files in the project database: it stores only their names and reads them when required. By default it will attempt to look for them in the current working directory (likely the same directory as the gap database). However this

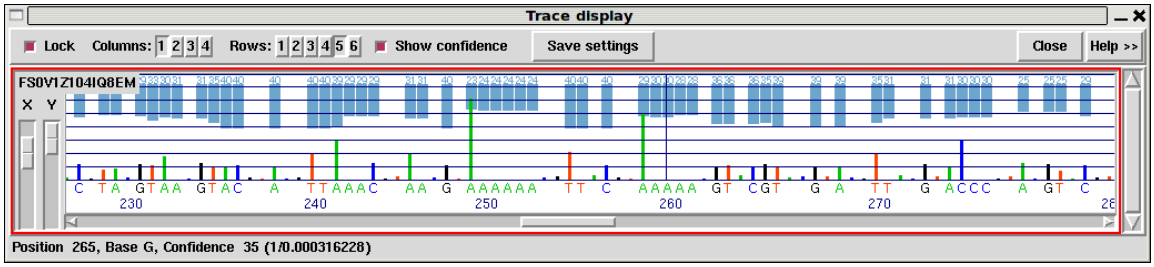
can be adjusted to look in other directories or via URLs using "Trace file location" in the main gap5 configure menu (see ⟨undefined⟩ [Trace File Location], page ⟨undefined⟩).



This figure is an example of the Trace Display showing three capillary traces and an Illumina trace. On the top line, the Lock checkbutton keeps the trace data in sync with the editor cursor position. The layout is controlled by the Columns and Rows selectors at the top of the window; 2 column by up to 3 rows in the above screenshot. Show confidence draws coloured bars and a numerical value representing the quality of each individual base-call.

The main trace panels each have the sequence name displayed in the top left corner. Below this are X and Y zoom controls on the left and the actual trace data on the right. The style of this will depend on the type of trace. Sanger chromatograms take multiple samples per base and are subsequently analysed (base-called) to identify the peaks and the number/type of bases represented by that peak. These are drawn using smooth lines, examples of which can be seen in the top row of the image above. Illumina GA instruments are "clocked" in that each and every measurement corresponds to one base. These are drawn using a stick plot, as seen in the bottom row of the screen-shot. Note that it is quite likely you will not have the processed trace data available for Illumina GA sequences due to size constraints, so the above is simply an example of what *could* be viewed rather than a typical example.

454 instruments use pyro-sequencing and so produce a variable number of bases per measurement, with each measurement being clocked to a specific cycle (flow) on the sequencing instrument. Hence 454 data is also drawn using a stick plot, although with potentially multiple bases per measurement. An example is visible below.



The horizontal rulers in this plot correspond to normalised peak intensities for 1.0, 2.0 and so on to indicate 1, 2, 3... bases per flow. Clearly visible are flows of approximate height 1 (C T A G T on the left), 2 (the following AA) and 0 (the G between the left most C and T). Above these the confidence bars are visible.

Right clicking on a trace will bring up a popup menu containing the following options.

Information
: Displays some basic textual information about the trace. The information available will vary by trace type, but it may include details such as the length, instrument and run-date.

Save
: Saves the trace in ZTR format to a local file on disk. This can be useful for when you are using a remote service for fetching traces or extracting them from an archive such as .sff or .srf file.

Complement
: Reverse complements the trace display. This does not modify data in any way, but simply adjusts how it is drawn.

Quit
: Removes this trace from the trace window. If it is the last displayed trace then the window will be removed too.

4.10 The Editor Information Line

The very bottom line of the editor display is text line used by the editor to display pieces of useful information. Currently this gives information on individual bases, readings, the contig, and tags, as the mouse is moved over the appropriate object. Each type of object we move the mouse pointer over (sequence base, consensus base, sequence name panel, annotation) has its own list of information to display which can be configured using a format string stored in your *$HOME/.gap5rc* file.

Typically you will not need to modify these, but if you choose to do so the default values to start from are shown below.

```
# Mouse-over a sequence the reading name panel
set_def READ_BRIEF_FORMAT \
Reading:%n(#%Rn)  Tech:%V  Length:%l(%L)  MappingQ:%m%**/%*m  Pos:%S%p / %*S%*p

# Mouse-over the "Consensus" label in the name panel
set_def CONTIG_BRIEF_FORMAT  \
Contig:%n(#%Rn)   Length:%l  Start:%s  End:%e

# Mouse-over a base in a sequence
set_def BASE_BRIEF_FORMAT1  \
 Base %b confidence:%4.1c (Prob. %Rc, raw %4.1A %4.1C %4.1G %4.1T)   Position %Rp  %n

# Mouse-over a base in the consensus
set_def BASE_BRIEF_FORMAT2  \
Base confidence:%4.1c (Prob. %Rc)  A=%4.1A C=%4.1C G=%4.1G T=%4.1T *=%4.1*  Position %p

# Mouse-over an annotation
set_def TAG_BRIEF_FORMAT  \
Tag type:%t  Comment:"%.100c"
```

The text output is as listed above, but replacing percent-code strings with a relevant piece of text. In many cases a capital R indicates raw mode to display a numerical value instead of a string. For example `%n` in READ_BRIEF_FORMAT will be replaced by the sequence name while `%Rn` will be replaced by the sequence record number. The full syntax of percent expansion is as follows:

- A percent sign.
- An optional minus sign to request left alignment of the information. When displaying information in a specific field with where that data does not fill the entire space allowed the information will, by default, be right justified. Adding a minus character here requests left justification.
- An optional minimum field width. This is a decimal number indicating how much space to leave for this information.
- An optional precision for numbers or maximum field width for strings. This is given as a fullstop followed by a decimal number.
- An optional 'R' to specify Raw mode. This changes the meaning of many (but not all) of the expansion requests to give a numercial representation of the data. For example %n is a reading name and %Rn is a reading number.
- Th expansion type itself. This is either one or two letters. See below for full details of their meanings.

To programmers this syntax may seem very similar to `printf`. This is intentional, but do not assume it is the same. Specifically the print syntax of `%#`, `%+` and `%0` will not work.

4.10.1 Reading Information

Used when we move the mouse over a sequence name in the names panel or a sequence base-call. Example output is **Reading:xc04a1.s1(#74) Tech:Sanger Length:295(474) MappingQ:50**. Note that not all expansions make sense when used in the names panel as no cursor X position is available.

%%	A single % sign
%n	Reading name. Raw mode: record number
%#	Reading record number
%p	Position in sequence. Raw mode: position in contig.
%l	Clipped sequence length
%L	Unclipped sequence length
%s	Start of clip
%e	End of clip
%S	Sense (whether complemented) - "<<" or ">>". Raw mode: 0/1
%d	Strand - "+" or "-". Raw mode: 0/1
%b	Base call
%c	Confidence value of called base (phred style). Raw mode: probability

%A
%C
%G
%T Individual confidence (phred style) of A,C,G,T component in log-odds form. Raw mode: probability value.

%m Mapping Quality. Raw mode: probability of correctly mapped.

%V Instrument type - Sanger, Illumina, SOLiD, 454 or Unknown.

4.10.2 Contig Information

For the CONTIG_BRIEF_FORMAT and BASE_BRIEF_FORMAT2 the following expansions apply. These operate on contigs and the consensus sequence.

%% Single % sign

%n Contig name. Raw mode: contig record number.

%# Contig record number

%p Position in contig

%l Length of contig

%s Contig start coordinate

%e Contig end coordinate

%b Called consensus base

%c Score for called consensus base. Raw mode: probability value

%A
%C
%G
%T
%* Individual confidence for A,C,G,T,* base types in log-odds form. Raw mode: as a probability value.

4.10.3 Tag Information

The TAG_BRIEF_FORMAT string is used to display annotation summaries. The possible percent encodings are as follows.

%% Single % sign

%p Tag position

%t Tag type (always 4 characters)

%l Tag length

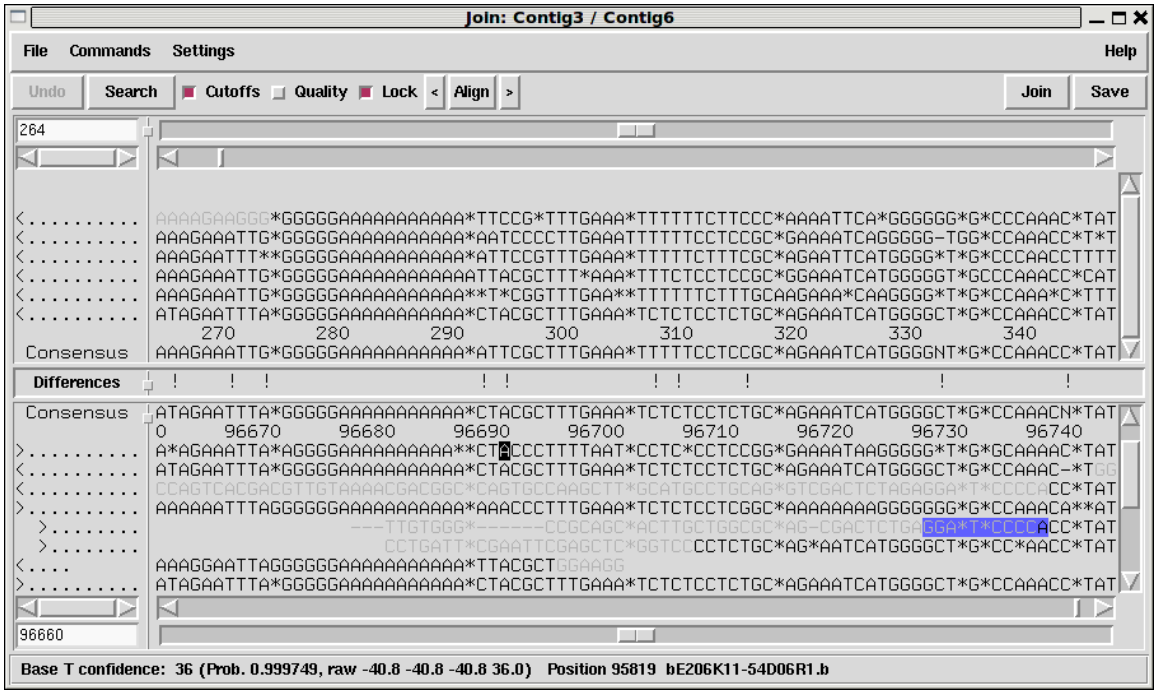
%# Tag number (0 if unknown)

%c Tag comment

4.11 The Join Editor

Contigs are joined interactively using the Join Editor. This is simply a pair of contig editor displays stacked above one another. The top editor is flipped in Y so that the consensus appears at the bottom. This allows the two consensus sequences to be adjacent to one another, separated only by a "differences" line. Note that it is essential to align the contigs over the full length of their overlap. It is much more difficult to achieve this after a join has been made, and until the alignment is correct, the consensus sequence will be nonsense.

The few differences between the Join Editor and the Contig Editor can be seen in the figure below. Otherwise all the commands and operations are the same as those for the Contig Editor



One difference is the Lock button. When set (as it is in the illustration) scrolling either contig will also scroll the other contig.

The Align button aligns the overlapping consensus sequences and adds pads as necessary. The alignment routine assumes that the two contigs are already in approximately the right relative position (as they are immediately after the Join Editor has been invoked from Find Internal Joins, or Find Repeats). If they are not you may get better results by manually positioning then before hand.

The "<" and ">" buttons either side of the "Align" button perform the alignment from the editing cursor to the start of the contig and and from the cursor to the end of the contig only. Alignment end-gaps are penalised at the curosr position but not for the alignment end at the contig start/end position. These buttons are useful for when multiple alignment positions may be valid, such as is the case with an overlap consisting entirely of a short tandem repeat.

It should be noted that each of the pair of editors comprising the Contig Editor maintains its own undo history, and using Align is likely to add to both undo histories. There is only one Undo button, but it applies to the editor last clicked within. A hint is given as to which of the two editors this is by highlighting the editor in a red border when the mouse is moved over the Undo button.

Pressing the Join button will display a small dialogue box informing you of the length and percentage match of the overlap between the two contigs. At this point you can decide to make the join, to not make the join (both of which remove the editors from the screen) or to cancel which leaves the join editor visible still to permit further editing.

4.12 Using Several Editors at Once

Several editors can be used simultaneously, even on the same contig. In the latter case, it is useful to understand the difference between the data and the view of the data.

Each operating Contig Editor is a view of the data for a particular contig. With two editors viewing the same contig, making changes in either will modify the data that both are viewing, hence the change will be visible in both editors. Similarly, using Undo in either will undo the changes to both.

Interaction between Contig Editors and Join Editors is more complicated and generally isn't advised. However such interactions work consistently with the notion of views of contigs. For example, suppose there are two Contig Editors open on two separate contigs, and in addition to these a Join Editor displaying both contigs. Making the join in the Join Editor will update the two stand-alone Contig Editors so that they are each viewing the correct positions in the new contig, even though they're both now viewing the same contig.

4.13 Quitting the Editor

The Exit operation in the File menu quits the editor. If changes have been made since the last save you will be asked whether you wish to save these changes. Answering "Cancel" abandons the exit process and provides control of the editor again, otherwise the appropriate action will be taken and the editor quitted.

4.14 Summary

4.14.1 Keyboard summary for editing window

("Left", "Right", "Up", "Down" refer to the appropriate arrow keys.)

```
Page Up                         Scroll left by 1Kb
Shift-Page Up                   Scroll left by 10Kb
Control-Page Up                 Scroll left by 100Kb
Shift-Control-Page Up           Scroll left by 1Mb

Page Down                       Scroll right by 1Kb
Shift-Page Down                 Scroll right by 10Kb
Control-Page Down               Scroll right by 100Kb
Shift-Control-Page Down         Scroll right by 1Mb
```

Left arrow or Control-b	Move editing cursor left one base
Right arrow or Control-f	Move editing cursor right one base
Up arrow or Control-p	Move editing cursor up one base
Down arrow or Control-n	Move editing cursor down one base
Control-a	Move editing cursor to start of sequence
Control-e	Move editing cursor to end of sequence
Alt-comma	Move editing cursor to start of contig
Alt-fullstop	Move editing cursor to end of contig
Control-t	Display trace
Control-s	Search forward
Control-r	Search backwards
Control-q	Toggle tag display
<	Set left cutoff clip point
>	Set right cutoff clip point
[	Set confidence to 0
]	Set confidence to 100
Shift Up	Increase confidence of base by 1
Shift Down	Decrease confidence of base by 1
Control Up	Increase confidence of base by 10
Control Down	Decrease confidence of base by 10
a, c, g, t or *	Overwrite base with a new call.
i	Insert pad (or column if in consensus)
Backspace or Delete	Delete padding character
Ctrl-Backspace or Ctrl-Delete	Delete base (any base type)
Control-right arrow	Move sequence right 1 base-pair
Control-left arrow	Move sequence left 1 base-pair

4.14.2 Mouse summary for editing window

Left button	Position editing cursor to mouse cursor
Left button (drag)	Mark start and end of selection
Shift left button	Adjust end of selection
Left button (double click)	Display trace
Right button	Display commands menu
Mouse-wheel	Vertically scroll the editor
Control mouse-wheel	Vertically scroll the editor, fast
Shift mouse-wheel	Horizontally scroll the editor
Shift Control mouse-wheel	Horizontally scroll the editor, fast

4.14.3 Mouse summary for names window

Left button + drag	Copy sequence name to clip-board
Right button	Display popup menu
Mouse-wheel	Vertically scroll the editor
Control mouse-wheel	Vertically scroll the editor, fast

5 Assembling and Adding Readings to a Database

There are two main types of assembly - denovo and mapped - with the latter not really being a true assembly at all.

Denovo assembly consists of an assembly of DNA fragments without typically knowing any of the goal target sequence. Hence it compares sequence fragments against each other in order to form contigs. Mapped assembly makes uses of a known reference sequence and compares all sequence fragments against the reference, which is a far simpler and faster process than denovo assembly.

Gap5 however has neither denovo or mapped assembly built-in. Instead it relies on externally running standard command-line tools. At present this consists purely of using bwa for a mapped assembly, but in future this will be expanded upon.

This means that the Assembly menu currently only contains a "Map Reads" sub-menu, which is turn has multiple choices for bwa usage. You will not be directly able to join contigs using these facilities or to fill holes in the contig, although this is possible by manually following some of the steps outlined below and using an alternate step for generating the SAM file.

5.1 Importing with tg_index

To enable efficient editing of data, Gap5 needs its own database format for storing sequence assemblies. Formats such as BAM are good at random access for read-only viewing, but are not at all amenable to actions such as reverse complementing a contig and joining it to another.

Hence we need a tool that can take existing assembly formats and convert them to a form suitable for Gap5. The `tg_index` program performs this task. It is strictly a command line tool, although in some specific cases Gap5 has basic GUI dialogues to wrap it up.

One or more input files may be specified. The general form is:

`tg_index` *[options]* `-o` *gap5_db_name input_file_name ...*

An example usage is:

```
tg_index -z 16384 -o test_data.g5 test_data.bam
gap5 test_data.g5 &
```

File formats supported are SAM, BAM, ACE, MAQ (both short and long variants), CAF and BAF. Tg_index typically automatically detects the type of file, but in rare cases you may need to explicitly state the input file type.

Tg_index options:

-o *filename*
: Creates a gap5 database named *filename* and *filename*`.aux` If not specified the default is "g_db".

-a
: Append to an existing database, instead of creating a new one (which is the default action).

-n When appending, the default behaviour is to add reads to existing contigs if contigs with the appropriate names already exist. This option always forces creation of new contigs instead.

-g When appending to an existing database, assume that the alignment has been performed against an ungapped copy of the consensus exported from this database. (This is internally used when performing mapped assemblies as they consist of exporting the consensus, running the external mapped alignment tool, and then importing the newly generated alignments.)

-m
-M Forces the input to be treated as MAQ, both short (-m) and long (-M) formats are supported. By default the file format is automatically detected.

-A Forces the input to be treads as ACE format.

-B Forces the input to be treads as BAF format.

-C Forces the input to be treads as CAF format.

-b
-s Forces the input to be treads as BAM (-b) or SAM (-s) format. SAM must have @SQ headers present. Both need to be sorted by position.

-z *bin_size* Modifies the size of the smallest allowable contig bin. Large contigs will contain child bins, each of which will contain smaller bins, recursing down to a minimum bin size. Sequences are then placed in the smallest bin they entirely fit within. The default minimum bin size is 4096 bytes. For very shallow assemblies increasing this will improve performance and the decrease disk space used. Ideally 5,000 to 10,000 sequences per bin is an approximate figure to aim for.

-u Store unmapped reads only (from SAM/BAM only)

-x Store SAM/BAM auxillary key:value records too.

-p
-P Enable (-p) or disable (-P) read-pairing. By default this is enabled. The purpose of this is to link sequences from the same template to each other such that gap5 knows the insert size and read-pairings. Generally this is desirable, but it adds extra time and memory to identify the pairs. Hence for single-ended runs the option exists to disable attempts at read-pairing.

-f Attempt a faster form of read-pairing. In this mode we link the second occurrence of a template to the first occurrence, but not vice versa. This is sufficient for the template display graphical views to work, but will cause other parts of the program to behave inconsistently. For example the contig editor “goto...” popup menu will sometimes be missing.

-t
-T Controls whether to index (-t) or not (-T) the sequence names. By default this is disabled. Adding a sequence name index permits us to search by sequence name or to use a sequence name in any dialogue that requires a contig identifier. However it consumes more disc space to store this index and it can be time consuming to construct it.

-r *nseq* Reserves space for at least *nseq* sequences. This generally isn't necessary, but if the total number of records extends above 2 million (equivalent to 2 billion sequences, or less if we have lots of contigs, bins and annotation records to write) then we run out of suitable sequence record numbers. This option preallocates the lower record numbers and reserves them solely for sequence records.

-c *compression_method*
Specifies an alternate compression method. This defaults to *zlib*, but can be set to either *none* for fastest speed or *lzma* for best compression.

5.2 Mapped assembly by bwa aln

This function runs the bwa program using the "aln" method for aligning sequences. It is appropriate for matching most types of short-read data.

The GUI is little more than a wrapper around command line tools, which can essentially be repeatedly manually as follows.

1. Calculate and save the consensus for all contigs in the database in fastq format.
2. Index the consensus sequence using "bwa index".
3. Map our input data against the bwa index using "bwa aln". Repeat for reverse matches too.
4. Generate SAM format from the alignments using "bwa samse" or "bwa sampe".
5. Convert to BAM and sort by position.
6. Import the BAM file, appending to the existing gap5 database (equivalent to `tg_index -a`).

5.3 Mapped assembly by bwa dbwtsw

This function runs the bwa program using the "dbwtsw" method for aligning sequences. This should be used when attempting to align longer sequences or data with lots of indels.

The GUI is little more than a wrapper around command line tools, which can essentially be repeatedly manually as follows.

1. Calculate and save the consensus for all contigs in the database in fastq format.
2. Index the consensus sequence using "bwa index".
3. Map our input data against the bwa index using "bwa dbwtsw".
4. Convert to BAM and sort by position.
5. Import the BAM file, appending to the existing gap5 database (equivalent to `tg_index -a`).

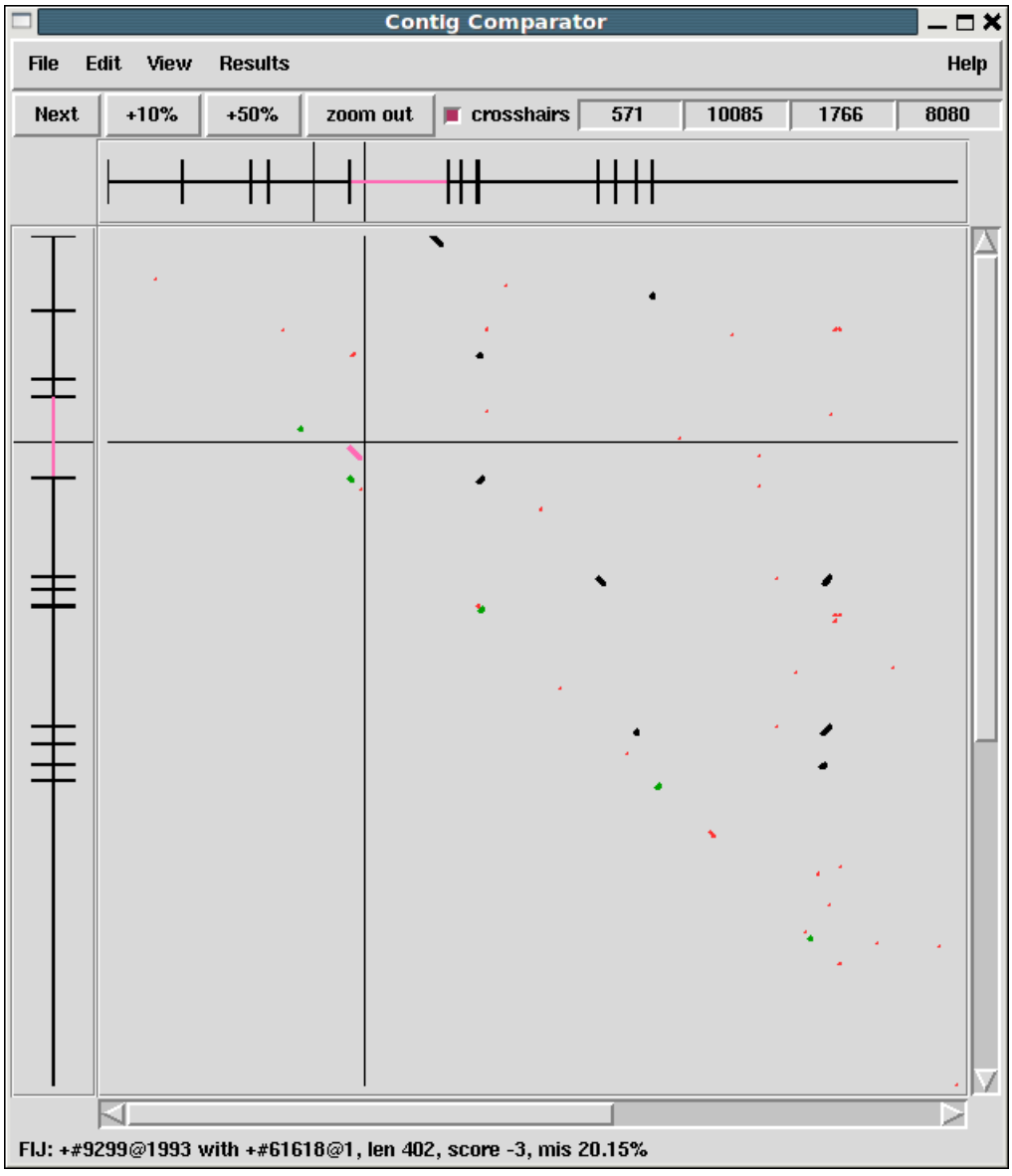
--split()

5.4 Find Internal Joins

The purpose of this function (which is invoked from the gap5 View menu) is to use sequences already in the database to find possible joins between contigs. Generally these will be joins that were missed or judged to be unsafe during assembly and this function allows users to examine the overlaps and decide if they should be made. During assembly joins may have been missed because of poor data, or not been made because the sequence was repetitive. Also it may be possible to find potential joins by extending the consensus sequences with the data from the 3' ends of readings which was considered to be too unreliable to align during assembly i.e. we can search in the "hidden data".

If it has not already occurred, use of this function will automatically transform the Contig Selector into the Contig Comparator. Each match found is plotted as a diagonal line in the Contig Comparator, and is written as an alignment in the Output Window. The length of the diagonal line is proportional to the length of the aligned region. If the match is for two contigs in the same orientation the diagonal will be parallel to the main diagonal, if they are not in the same orientation the line will be perpendicular to the main diagonal. The matches displayed in the Contig Comparator can be used to invoke the Join Editor (see Section 4.11 [The Join Editor], page 36) or Contig Editor. See Chapter 4 [Editing in gap4], page 17. Alternatively, the "Next" button at the top left of the Contig Comparator can be used to select each result in turn, starting with the best, and ending with the worst. When this is in use, users can find the match in the Contig Comparator which corresponds

to the next result by placing the cursor over the Next button. The plotted match and the contigs involved will turn white.



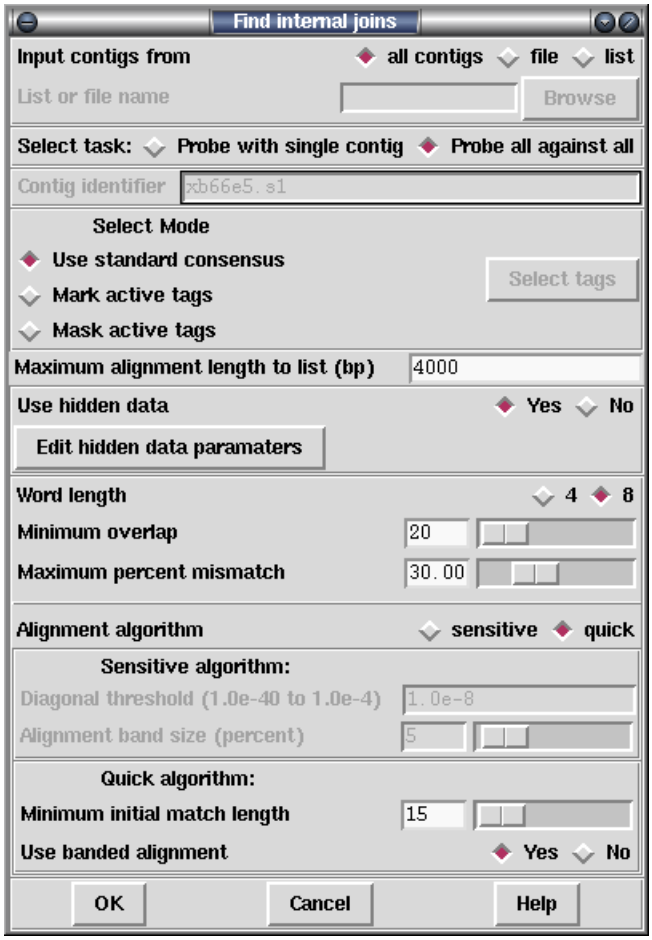
A typical display from the Contig Comparator is shown in the figure above.

To define the match all numbering is relative to base number one in the contig: matches to the left (i.e. in the hidden data) have negative positions, matches off the right end of the contig (i.e. in the hidden data) have positions greater than that of the contig length. The convention for reporting the positions of overlaps is as follows: if neither contig needs to be complemented the positions are as shown. If the program says "contig x in the -sense" then the positions shown assume contig x has been complemented. For example, in

the results given below the positions for the first overlap are as reported, but those for the second assume that the contig in the minus sense (i.e. 443) has been complemented.

```
Possible join between contig   445 in the + sense and contig   405
Percentage mismatch after alignment =  4.9
       412        422        432        442        452        462
    405  TTTCCCGACT GGAAAGCGGG CAGTGAGCGC AACGCAATTA ATGTGAG,TT AGCTCACTCA
          ::::::::: : ::::::::  ::::: ::: :::::::::: :::::::::: ::::::::::
    445  *TTCCCGACT G,AAAGCGGG TAGTGA,CGC AACGCAATTA ATGTGAG*TT AGCTCACTCA
      -127       -117       -107        -97        -87        -77
       472        482        492        502        512
    405  TTAGGCACCC CAGGCTTTAC ACTTTATGCT TCCGGCTCGT AT
         :::::::::: :::::::::: :::::::::: :::::::::: ::
    445  TTAGGCACCC CAGGCTTTAC ACTTTATGCT TCCGGCTCGT AT
       -67        -57        -47        -37        -27
Possible join between contig   443 in the - sense and contig   423
Percentage mismatch after alignment = 10.4
        64         74         84         94        104        114
    423  ATCGAAGAAA GAAAAGGAGG AGAAGATGAT TTTAAAAATG AAACG*CGAT GTCAGATGGG
         :::: ::::: :::::::::: :::::::::: ::::::  :: ::::: :::: :::::::::
    443  ATCG,AGAAA GAAAAGGAGG AGAAGATGAT TTTAAA,,TG AAACGACGAT GTCAGATGG,
      3610       3620       3630       3640       3650       3660
       124        134        144        154        164
    423  TTG*ATGAAG TAGAAGTAGG AG*AGGTGGA AGAGAAGAGA GTGGGA
         ::: :::::: :::::::::: :: ::::::::  ::: ::::: :: ::
    443  TTGGATGAAG TAGAAGTAGG AGGAGGTGGA ,GAG,AGAGA GTTGG*
      3670       3680       3690       3700       3710
```

5.4.1 Find Internal Joins Dialogue



The contigs to use in the search can be defined as "all contigs", a list of contigs in a file "file", or a list of contigs in a list "list". If "file" or "list" is selected the browse button is activated and gives access to file or list browsers. Two types of search can be selected: one, "Probe all against all" compares all the contigs defined against one another; the other "Probe with single contig", compares one contig against all the contigs in the list. If this option is selected the Contig identifier panel in the dialogue box is ungreyed. Both sense of the sequences are compared.

If users elect not to "Use standard consensus" they can either "Mark active tags" or "Mask active tags", in which cases the "Select tags" button will be activated. Clicking on this button will bring up a check box dialogue to enable the user to select the tags types they wish to activate. Masking the active tags means that all segments covered by tags that are "active" will not be used by the matching algorithms. A typical use of this mode is to avoid finding matches in segments covered by tags of type ALUS (ie segments thought to be Alu sequence) or REPT (ie segment that are known to be repeated elsewhere in the data (see ⟨undefined⟩ [Tag types], page ⟨undefined⟩). "Marking" is of less use: matches will

be found in marked segments during searching, but in the alignment shown in the Output Window, marked segments will be shown in lower case.

Some alignments may be very large. For speed and ease of scrolling Gap5 does not display the textual form of the longest alignments, although they are still visible within the contig comparator window. The maximum length of the alignment to print up is controlled by the “Maximum alignment length to list (bp)” control.

The default setting for the consensus is to "Use hidden data" which means that where possible the contigs are extended using the poor quality data from the readings near their ends. To ensure that this additional data is not so poor that matches will be missed, the program uses algorithms which can be configured from the "Edit hidden data parameters" dialogue. Two algorithms are available. Both slide a window along the reading until a set criteria is met. By default an algorithm which sums confidence values within the window is used. It stops when a window with < "Minimum average confidence" is found. The other algorithm counts the number of uncalled bases in the window and stops when the total reaches "Max number of uncalled bases in window". The selected algorithm is applied to all the readings near the ends of contigs and the data that extends the contig the furthest is added to its consensus sequence.

If your total consensus sequence length (including a 20 character header for each contig that is used internally by the program) plus any hidden data at the ends of contigs is greater than the current value of a parameter called maxseq, Find Internal Joins may produce an error message advising you to increase maxseq. Maxseq can be set on the command line (see ⟨undefined⟩ [Command line arguments], page ⟨undefined⟩) or by using the options menu (see ⟨undefined⟩ [Set Maxseq], page ⟨undefined⟩).

The search algorithms first finds matching words of length "Word length", and only considers overlaps of length at least "Minimum overlap". Only alignments better than "Maximum percent mismatches" will be reported.

There are two search algorithms: "Sensitive" or "Quick". The quick algorithm should be applied first, and then the sensitive one employed to find any less obvious overlaps.

The sensitive algorithm sums the lengths of the matching words of length "Word length" on each diagonal. It then finds the centre of gravity of the most significant diagonals. Significant diagonals are those whose probability of occurence is < "Diagonal threshold". It then uses a dynamic programming algorithm to align around the centre of gravity, using a band size of "Alignment band size (percent)". For example: if the overlap was 1000 bases long and the percentage set at 5, the aligner would only consider alignments within 50 bases either side of the centre of gravity. Obviously the larger the percentage and the overlap, the slower the aligment.

The quick algorithm can find overlaps and align 100,000 base sequences in a few seconds by considering, in its initial phase only matching segments of length "Minimum initial match length". However it does a dynamic programming alignment of all the chunks between the matching segments, and so produces an optimal alignment. Again a banded dynamic algorithm can be selected, but as this only applies to the chunks between matching segments, which for good alignments will be very short, it should make little difference to the speed.

After the search the results will be sorted so that the best matches are at the top of a list where best is defined as a combination of alignment length and alignment percent identity. This list can be stepped through, one result at a time using the Contig Joining Editor, by clicking on the "Next" button at the top left of the Contig Comparator.

5.5 Find Repeats

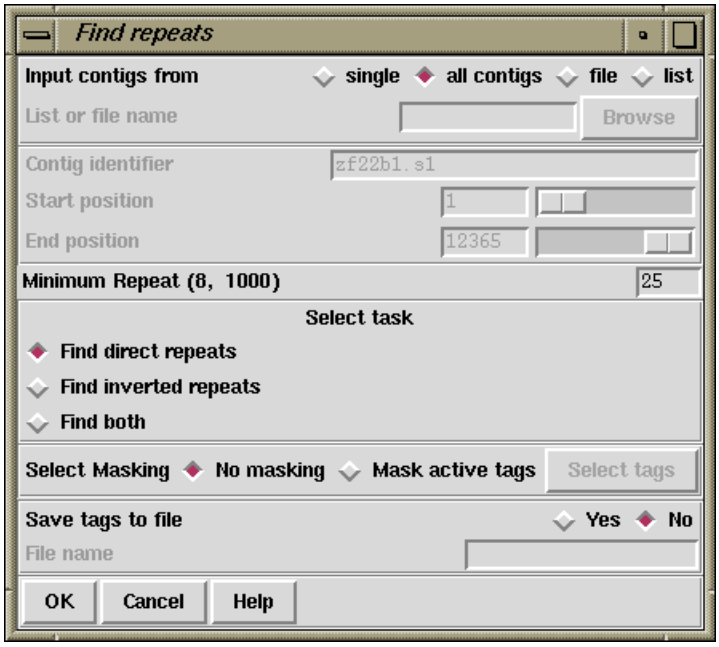
The purpose of this function (which is invoked from the gap5 View menu) is to find exact repeats in contig consensus sequences. An exact repeat is defined as a run of consecutive identical ACGT characters; no mismatches or gaps are permitted.

If it has not already occurred, selection of this function will automatically transform the Contig Selector into the Contig Comparator. See Chapter 2 [Contig Comparator], page 5. Each match found is plotted as a diagonal line in the Contig Comparator. The length of the diagonal line is proportional to the length of the match.

If the match is for two contigs in the same orientation the diagonal will be parallel to the main diagonal, if they are not the line will be perpendicular to the main diagonal. The matches displayed in the Contig Comparator can be used to invoke the Join Editor (see Section 4.11 [The Join Editor], page 36) or Contig Editors (see Chapter 4 [Editing in gap5], page 17), and an Information button will display data about the match in the Output window. e.g.

```
Repeat match
    From contig xb54a3.s1(#26) at 78
    With contig xb62h3.s1(#3) at 1
    Length 37
```

This means that position 78 in the contig with xb54a3.s1 (reading number 26) at its left end matches 37 bases at position 1 in the contig with xb62h3.s1 (number 3) at its left end.



Users can elect to search a "single" contig, or compare "all contigs", or a subset of contigs defined in a list or a file. If "file" or "list" is selected the browse button is activated and gives access to file or list browsers. If they choose to analyse a single contig the

dialogue concerned with selecting the contig and the region to search becomes activated. The "Minimum Repeat" defines the smallest match that the algorithm will report. The algorithm will search only for repeats in the forward direction "Find direct repeats", or only those in the reverse direction "Find inverted repeats", or both "Find both".

If "Mask active tags" is selected the "Select tags" button is activated. Clicking on this button will bring up a check box dialogue to enable the user to select the tags types they wish to activate. Masking the active tags means that all segments covered by tags that are "active" will not be used in the matching algorithm. A typical use of this mode is to avoid finding matches in segments covered by tags of type ALUS (ie segments thought to be Alu sequence) or that already covered by REPT tags. See ⟨undefined⟩ [Tag types], page ⟨undefined⟩.

After the search is complete clicking on "Yes" in the "Save tags to file" panel will activate the "File name" box and all repeats on the list will be written to a file. This file can be used with "Enter tags" (see ⟨undefined⟩ [Enter Tags], page ⟨undefined⟩) to create REPT tags for all the repeats found. Note that "Enter tags" will remove all the results plotted in the contig comparator.

Note that the current version of Find Repeats has a limit to the number of repeats it can store. The limit depends on the current maximum consensus length, so if you want to increase the limit, reset the maximum consensus length. This can be done using the "Set maxseq" item in the "Options" menu.

Index

K

M

N

O

P

Q

R

S

T

U

V