

Supplementary Information and Figures for “*Consed*: A Graphical Editor for Next-Generation Sequencing” by David Gordon and Phil Green

S1 Input data requirements

Consed requires reads and alignments either in BAM format (Li *et al.*, 2009) or in *consed*'s editor-ready format (Gordon *et al.*, 1998). Several *de novo* assembly programs produce editor-ready format: Newbler (Margulies *et al.*, 2005), Velvet (Zerbino and Birney, 2008) (Jamison <https://github.com/dcurtisjamison/Velvet2Consed>), CAP3 (Huang and Madan, 1999), CAP4, PCAP (Huang *et al.*, 2003; Huang and Yang, 2005), and *phrap* (<http://www.phrap.org>).

For assemblers that do not provide read location information, reads must first be aligned to the consensus by an alignment program. *Consed*'s various features for detecting and correcting misassemblies have been implemented and iteratively refined in response to requests and feedback from users who have extensive experience performing these tasks. Since reads are the raw data of any assembly program, the prevailing view of such users is that detection of misassemblies is most reliable when read data is available and consequently read data is the basis for most of *consed*'s misassembly detection algorithms and other features.

As indicated in the main text, *bamScape* is used to identify regions of interest and to bring up the *consed* graphical editor on them. If the entire region to be edited is known in advance and contains just a few million reads, the *bamScape* step can be skipped by converting read alignments from BAM format into *consed*'s editor-ready input format using *consed*'s program *bam2Ace*. *Bam2Ace* can extract either all alignments or those from a user-specified list of regions and is run in batch (i.e. as running a separate process from the command line rather than through a graphical interface). Optionally *bam2Ace* can reduce read depth by choosing a sample of reads that preserves all variants and any read mates that map nearby. (Depth reduction can also be done starting with an editor-ready dataset rather than a BAM file.) *Consed* can convert output from the aligner *cross_match* (<http://www.phrap.org>), which requires fasta input, to editor-ready format.

Consed can directly read 454 sff files and display 454 mock traces in a manner similar to how it displays traces of Sanger read chromatograms, allowing inspection of the signal strength for mononucleotide runs. Most *consed* graphical editor features work best when quality values (Ewing and Green, 1998; Ewing *et al.*, 1998) are supplied, but a user-specified default value can be used when quality values are unavailable.

The input data for displaying tracks in *consed* is provided as WIG fixedStep files or BED files, downloaded from <http://www.genome.ucsc.edu> or created by the user.

S2 Other new features

There are several dozen new interactive lists, including discrepancies with a specified read, and exon boundaries in RNASeq alignments. A user-created file of locations can be displayed as an interactive list, either automatically at *consed* startup or under manual control. Keys can be configured to make particular edits and/or run external programs. Contigs can be sorted by number of reads or found by name. Additional batch mode functions include complementing contigs, exporting scaffolds, and making edits. Roughly 20 different types of report can be produced by *autoreport*.

S3 Documentation

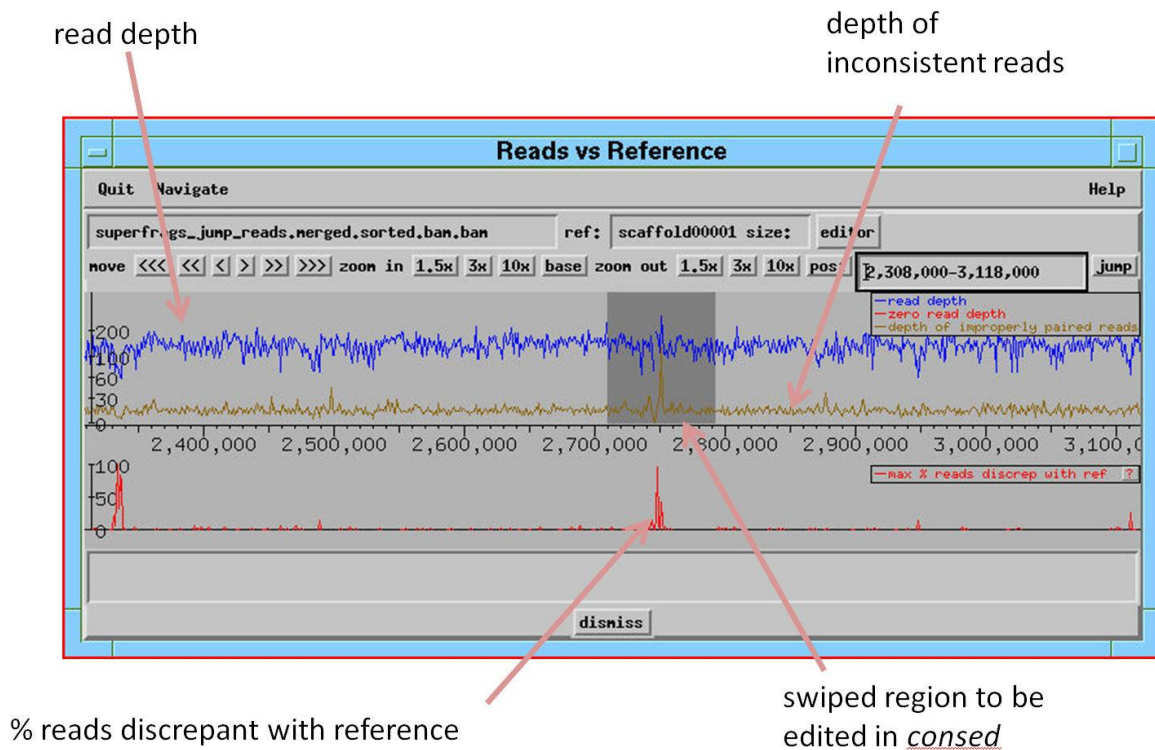
Full documentation for the current version and features introduced in each version (current and past) is found at <http://www.phrap.org/consed/consed.html> by clicking “documentation.”

REFERENCES

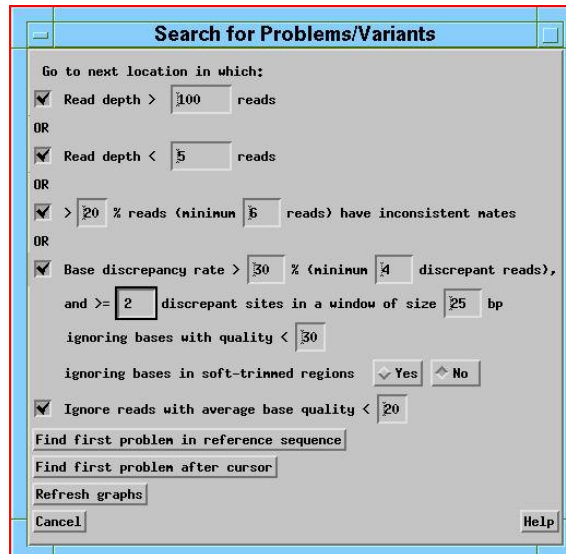
- Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res*, **8**, 186-194.
- Ewing, B., *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res*, **8**, 175-185.
- Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing, *Genome Res*, **8**, 195-202.
- Gordon, D., Desmarais, C. and Green, P. (2001) Automated finishing with autofinish, *Genome Res*, **11**, 614-625.
- Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program, *Genome Res*, **9**, 868-877.
- Huang, X., *et al.* (2003) PCAP: a whole-genome assembly program, *Genome Res*, **13**, 2164-2170.
- Huang, X. and Yang, S.P. (2005) Generating a genome assembly with PCAP, *Curr Protoc Bioinformatics*, **Chapter 11**, Unit11 13.
- Li, H., *et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078-2079.
- Margulies, M., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.
- Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res*, **18**, 821-829.

(see figures starting next page)

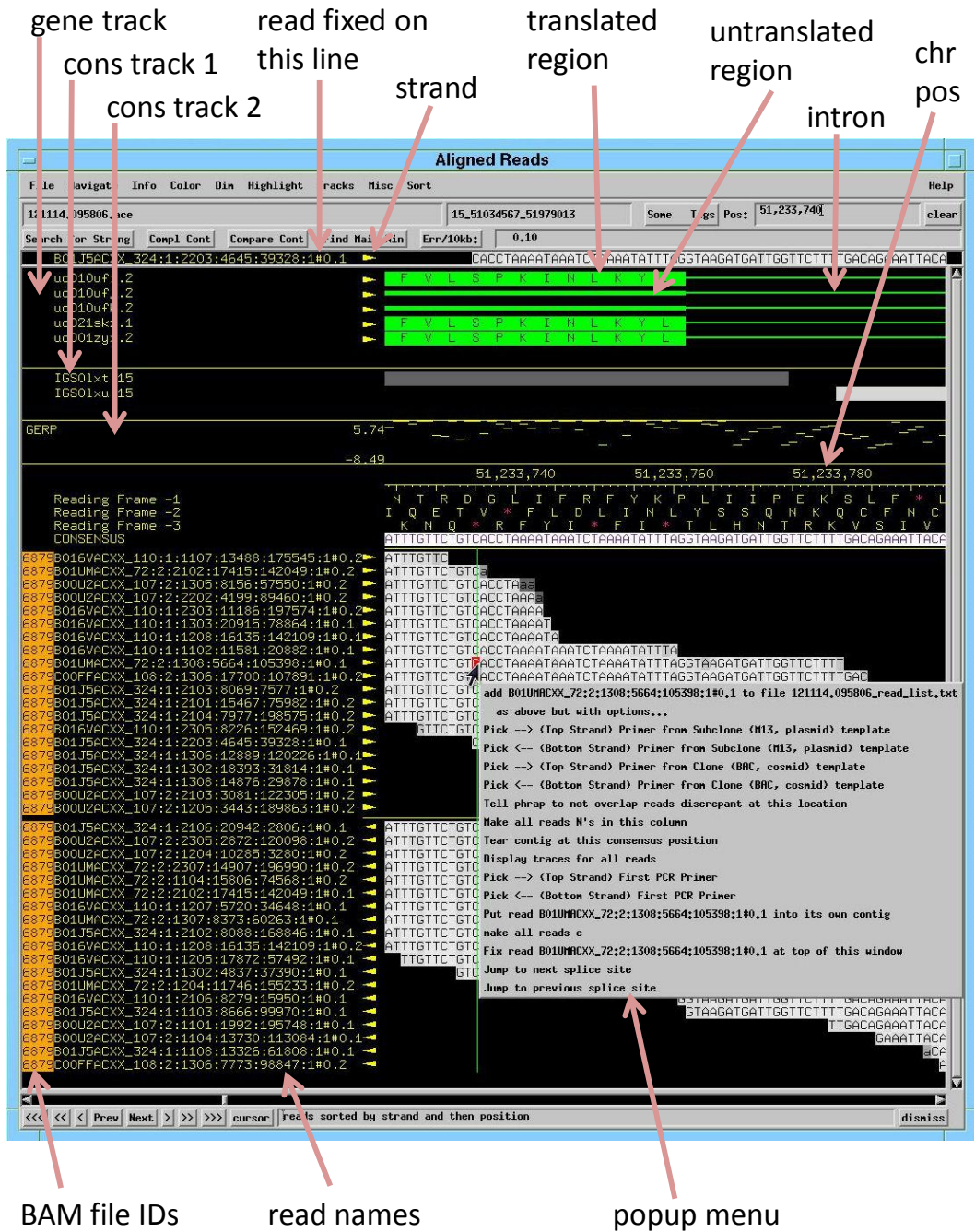
Acknowledgement We thank the Broad Institute, the Northwest Genome Center, and the Genome Institute at Washington University for data used in the figures.



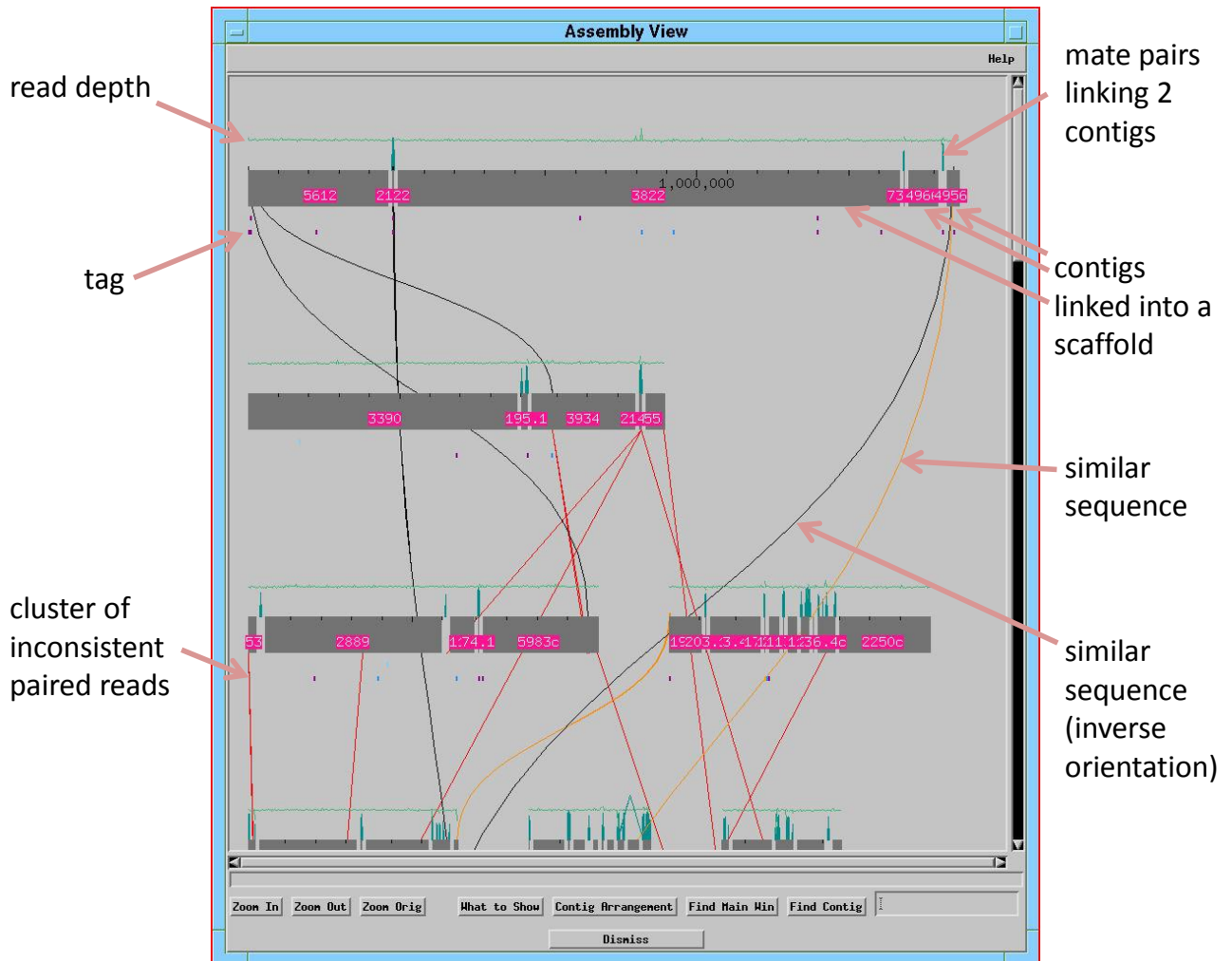
Supplementary Fig. S1. *BamScape's* Reads vs Reference Window. This example shows summary information for paired Illumina reads aligned to a reference sequence. Swiping a region pops up a window that shows where any inconsistently mapped mates cluster, and allows the user to bring up *consed's* graphical editor (Supplementary Fig. S3) on the swiped region. The navigate menu (top line) brings up the Search for Problems/Variants Window (Supplementary Fig. S2).



Supplementary Fig. S2. *BamScape*'s Search for Problems/Variants Window. A problem can be defined as excessively high or low depth of coverage, too many reads with inconsistently placed mates or too many discrepant reads. Each filter can be turned off by unchecking it. For read base discrepancies, each type (A,C, G, T, deletion, insertion) is tabulated separately and the threshold applied to each. (This reduces the "noise" due to base-calling errors.) Discrepant sites are only reported if there are at least 2 such sites within a 25 bp window; to report putative SNPs as well, change the "2" to "1". Clicking "Find first problem in reference sequence" or "Find first problem after cursor" causes a blinking cursor to move in the Reads vs Reference Window



Supplementary Fig. S3. Aligned Reads Window. This example shows Illumina reads, a gene track, an indel-purified conservation track, a GERP conservation track, and bottom-strand amino acid translations. (Track files were created using the table browser at <http://www.genome.ucsc.edu>.) One read has been placed just above the gene track to facilitate viewing during scrolling. Reads are shown sorted by strand and then left end position, but this can be switched instantly to sorts by base or base quality at the cursor position, alphabetically by read name, or by an arbitrary order specified in a file. Either chromosome positions or contig positions can be displayed. The popup menu appears upon pushing the right mouse button.



Supplementary Fig. S4. Assembly View Window. Contigs (dark bars) are shown arranged in scaffolds, one or several scaffolds per line. Scaffolds are either determined by linking mate pairs or specified by the user (the contigs can be manually rearranged). The set of tags to be displayed is configurable. The sequence similarity curves (orange and black) can be filtered by various characteristics of the match: e.g. length, % similarity, whether between or within contigs, whether or not at contig ends, or requiring the match to include a specified location. The user can click a similarity curve to view the alignment of the similar regions in a Compare Contigs Window (Supplementary Fig. S6 bottom). Clicking on the red lines displays information about the inconsistent read pairs with the option to remove those reads from the contig and possibly reassemble them. To reduce “noise” due to chimeras, only inconsistent read pairs confirmed by at least N other read pairs linking the same two locations are shown, where N is set such that (by default) no more than one cluster would occur by chance.

Highly Discrepant Positions												
min # of discrepant reads: 2, min quality: 20, "r": base of reference seq												
max depth of coverage: 100000. Ignore duplicate reads: yes												
A	C	G	T	*	pos	contig						
0	0.0%	0	0.0%	r	0	0.0%	22	100.0%	0	0.0%	13,781	chr1_876368_892048
0	0.0%	r	0	0.0%	23	100.0%	0	0.0%	0	0.0%	13,782	chr1_876368_892048
0	0.0%	8	40.0%	r	12	60.0%	0	0.0%	0	0.0%	13,794	chr1_876368_892048
18	81.8%	r	4	18.2%	0	0.0%	0	0.0%	0	0.0%	2671	chr1_895518_899820
0	0.0%	2	50.0%	r	0	0.0%	2	50.0%	0	0.0%	3765	chr1_895518_899820
0	0.0%	r	0	0.0%	45	100.0%	0	0.0%	0	0.0%	4114	chr1_895518_899820
2	66.7%	1	33.3%	r	0	0.0%	0	0.0%	0	0.0%	786	chr1_924300_925218
0	0.0%	7	100.0%	r	0	0.0%	0	0.0%	0	0.0%	9	chr1_938776_939723
0	0.0%	0	0.0%	r	10	76.9%	0	0.0%	3	23.1%	18-24*	chr1_938776_939723
2	100.0%	0	0.0%	r	0	0.0%	0	0.0%	0	0.0%	323	chr1_938776_939723
26	44.1%	0	0.0%	r	33	55.9%	0	0.0%	0	0.0%	696	chr1_938776_939723
0	0.0%	r	0	0.0%	57	100.0%	0	0.0%	0	0.0%	742	chr1_938776_939723
0	0.0%	2	33.3%	r	4	66.7%	0	0.0%	0	0.0%	1161	chr1_965906_980226
0	0.0%	20	40.8%	r	0	0.0%	29	59.2%	0	0.0%	1288	chr1_965906_980226
7	50.0%	0	0.0%	r	7	50.0%	0	0.0%	0	0.0%	1528	chr1_965906_980226
0	0.0%	17	42.5%	r	0	0.0%	0	0.0%	23	57.5%	2562-2563*	chr1_965906_980226
0	0.0%	0	0.0%	r	48	96.0%	2	4.0%	0	0.0%	3573	chr1_965906_980226
0	0.0%	18	90.0%	r	2	10.0%	0	0.0%	0	0.0%	3629	chr1_965906_980226
24	68.6%	0	0.0%	r	11	31.4%	0	0.0%	0	0.0%	4418	chr1_965906_980226

Supplementary Fig. S5. Highly Discrepant Positions Window. Each line indicates the number and percentage of reads having a given base at the specified location. The reference base is indicated by an ‘r’ after the corresponding percentage. Clicking on a line or ‘next’ causes the Aligned Reads Window to appear showing read data at that location. A range of positions in the “pos” column indicates a multi-base deletion. Options, which can be changed in another popup window and saved for use in other viewing sessions, include: minimum # of discrepant reads, quality below which discrepancies should be ignored, maximum depth of coverage, whether to count all or just the first of multiple reads starting at the same location (to disallow counting potentially duplicate reads), whether only indels should be shown or both indels and substitutions, and whether or not to ignore locations at which the consensus base is from a user-defined list of characters (e.g. N). Clicking “Phrap No Overlap” increases the qualities of discrepant bases of the highlighted line in order to avoid, during subsequent re-assembly, overlaps of reads discrepant at that location.

Sequence Matches

sequence 1			sequence 2			orient	size	S/W	% sim	comment
contig	end	pos	contig	end	pos					
contig00091	right	24582-24644	to contig00092	left	1-62	(not comp)	62	48	95	do
contig00077	right	2179-2777	to contig00078	left	1-599	(not comp)	599	585	99	do
contig00011	left	1-245	to contig00078	right	16297-16541	(not comp)	245	243	100	do
contig00011	right	27-258	to contig00079	left	1-232	(not comp)	232	229	100	do
contig00079	right	229047-229327	to contig00080	left	1-281	(not comp)	281	278	100	do
contig00079	right	228963-229051	to contig00080	left	1-89	(not comp)	89	81	98	lowerScore_matchNotToGap
contig00079	right	228891-228967	to contig00080	left	13-89	(not comp)	77	41	86	lowerScore_matchNotToGap
contig00043	left	1-85	to contig00080	right	10817-10901	(not comp)	85	83	100	matchNotToGap
contig00002	right	753-1025	to contig00094	left	1-273	(not comp)	273	272	100	do
contig00094	right	3652-3938	to contig00095	left	1-287	(not comp)	287	284	100	do
contig00060	left	1-152	to contig00095	right	10580-10731	(not comp)	152	150	100	matchNotToGap
contig00060	left	1-151	to contig00095	right	10634-10785	(not comp)	152	121	94	lowerScore_matchNotToGap
contig00060	left	37-152	to contig00095	right	10562-10677	(not comp)	116	105	97	lowerScore_matchNotToGap
contig00060	left	1-97	to contig00095	right	10688-10785	(not comp)	98	68	91	lowerScore_matchNotToGap
contig00060	left	91-152	to contig00095	right	10562-10623	(not comp)	62	58	98	lowerScore_matchNotToGap
contig00060	right	1-152	to contig00096	left	1-152	(not comp)	152	150	100	do
contig00060	right	1-151	to contig00096	left	55-205	(not comp)	151	125	95	lowerScore_matchNotToGap
contig00060	right	55-152	to contig00096	left	1-98	(not comp)	98	94	99	lowerScore
contig00060	right	1-110	to contig00096	left	109-217	(not comp)	109	80	92	lowerScore_matchNotToGap
contig00060	right	109-152	to contig00096	left	1-44	(not comp)	44	44	100	lowerScore
contig00061	left	8-670	to contig00096	right	5021-5682	(not comp)	662	639	99	discrepancy
contig00061	right	474-1409	to contig00097	left	1-950	(not comp)	950	833	97	matchNotToGap_discrepancy
contig00097	right	25464-26032	to contig00098	left	1-577	(not comp)	577	525	98	discrepancy
contig00098	right	3245-3799	to contig00099	left	1-555	(not comp)	555	543	100	do
contig00098	right	3678-3799	to contig00099	left	631-753	(not comp)	123	111	98	lowerScore_matchNotToGap
contig00099	right	6556-7144	to contig00100	left	1-589	(not comp)	589	559	99	discrepancy
contig00100	right	4325-4515	to contig00101	left	1-191	(not comp)	191	189	100	do
contig00093	left	1-249	to contig00103	right	218-466	(not comp)	249	249	100	do
contig00093	left	297-348	to contig00103	right	400-451	(not comp)	52	30	87	lowerScore_matchNotToGap

Compare Contigs

contig00099 complement just in this window

6560 6570 6580 6590 6600 6610 6620 6630 6640 6650

caaaaatgagcttcactgatccgcaaaagtccattgcccagtgtagccaccatcaaccatgacatgctgaaccagctcctaaattttggctagccaaactaaacat

caaaaatgagcttcactgatccgcaaaagtccattgcccagtgtagccaccatcaaccatgacatgctgaaccagctcctaaattttggctagccaaactaaacat

10 20 30 40 50 60 70 80 90 100

contig00100 complement just in this window

align

6600 6610 6620 6630 6640 6650 6660 6670 6680 6690

TAGCCACCATCAACCATGACATGCTGAACAGCTCTAAATTTTGCTAGCCAAACTAAACATAGCCAAATGCACCTGAACGATCTGATACATTAGCTCGTGT

TAGCCACCATCAACCATGACATGCTGAACAGCTCTAAATTTTGCTAGCCAAACTAAACATAGCCAAATGCACCTGAACGATCTGATACATTAGCTCGTGT

40 50 60 70 80 90 100 110 120 130

discrepancies: 0 discrep % 0.0000

If join, highlight reads from old right contig? True False

Supplementary Fig. S6. Exploring potential joins. The Sequence Matches Window (top) indicates pairs of similar regions. End: the contig end that matches; not comp: not in complemented orientation; S/W: Smith-Waterman score; % sim: % similarity; do: making the join is recommended; discrepancy: there is a high quality discrepancy between the regions; matchNotToGap: the similar region doesn't extend to the end of one of the sequences, so making the join would create discrepancies; lowerScore: there is a better match somewhere else. The user can click "Show Alignment" to examine the alignment in the Compare Contigs Window (bottom) and then click "Join Contigs" to make the join (Gordon *et al.*, 2001). Discrepancies in the Compare Contigs Window can be examined using the "Prev Discrepancy" and "Next Discrepancy" buttons. Grayscale indicates base quality.