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.

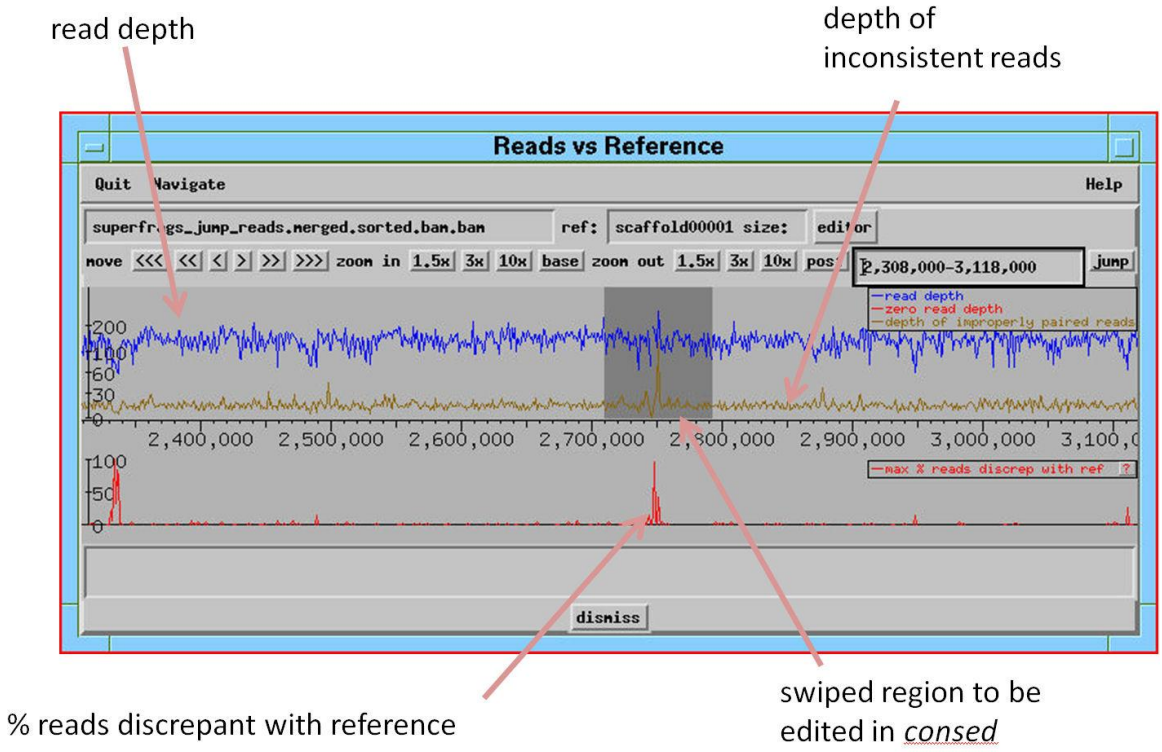
Li, H.*, et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078-2079.

Margulies, M.*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.

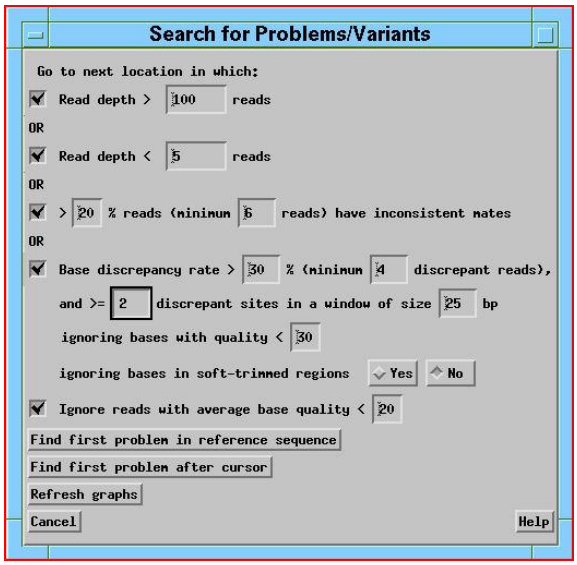
Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res*, **18**, 821-829.

(see figures starting next page)

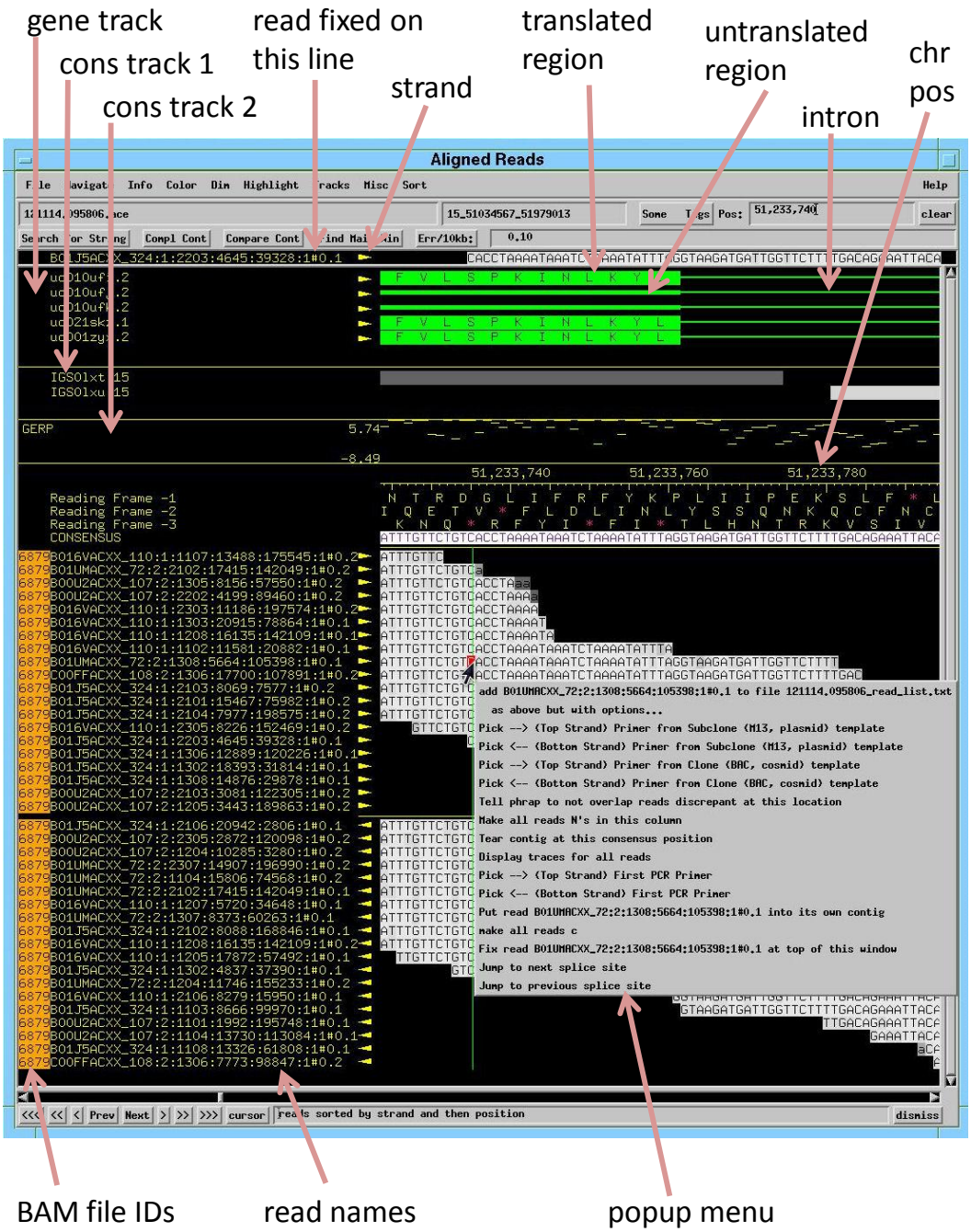
Acknowledgement We thank the Broad Institute, the Northwest Genome Center, and the Genome Institute at Washington University for data used in the figures.



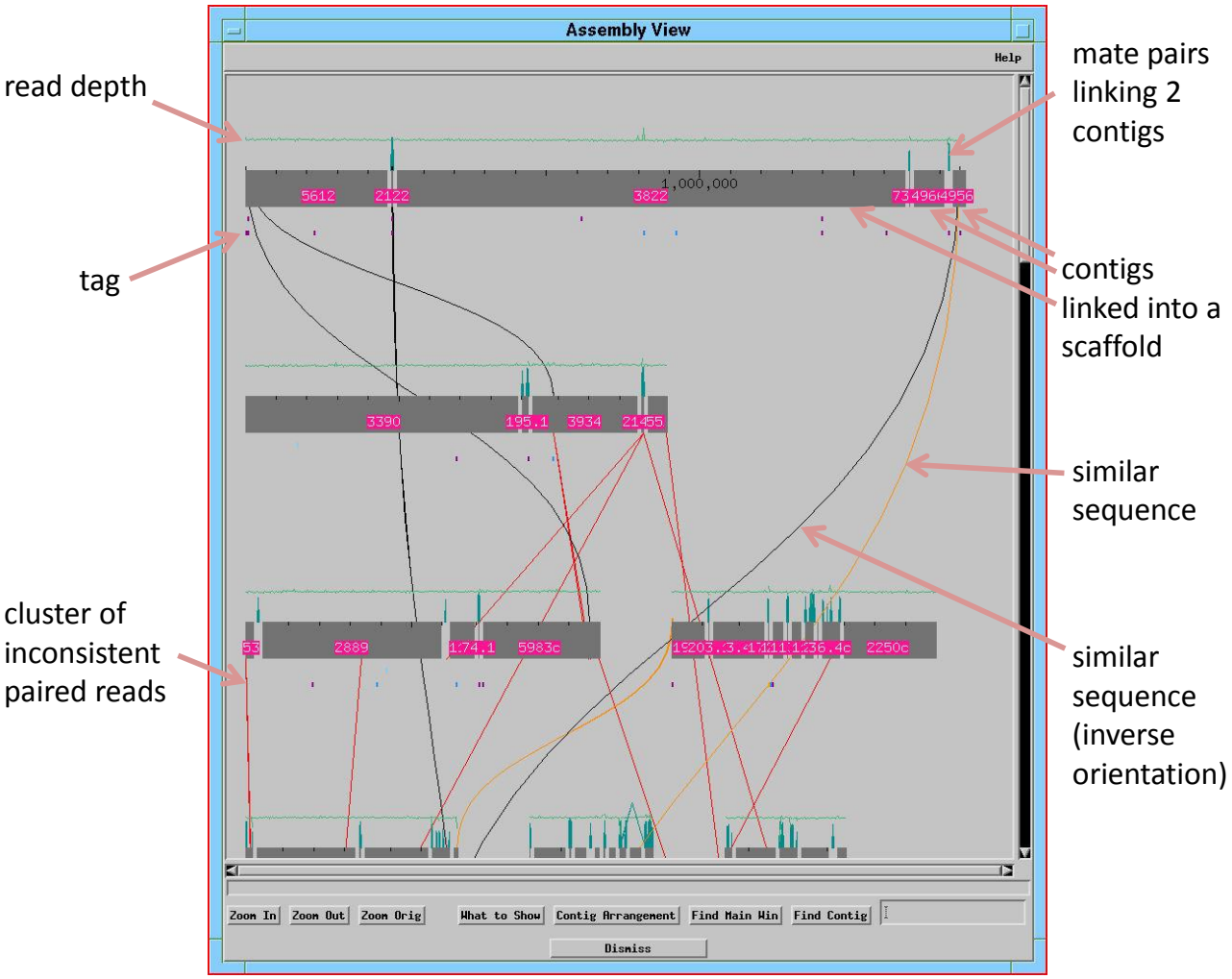
Supplementary Fig. S1. *BamScape's* Reads vs Reference Window. This example shows summary information for paired Illumina reads aligned to a reference sequence. Swiping a region pops up a window that shows where any inconsistently mapped mates cluster, and allows the user to bring up *consed's* graphical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



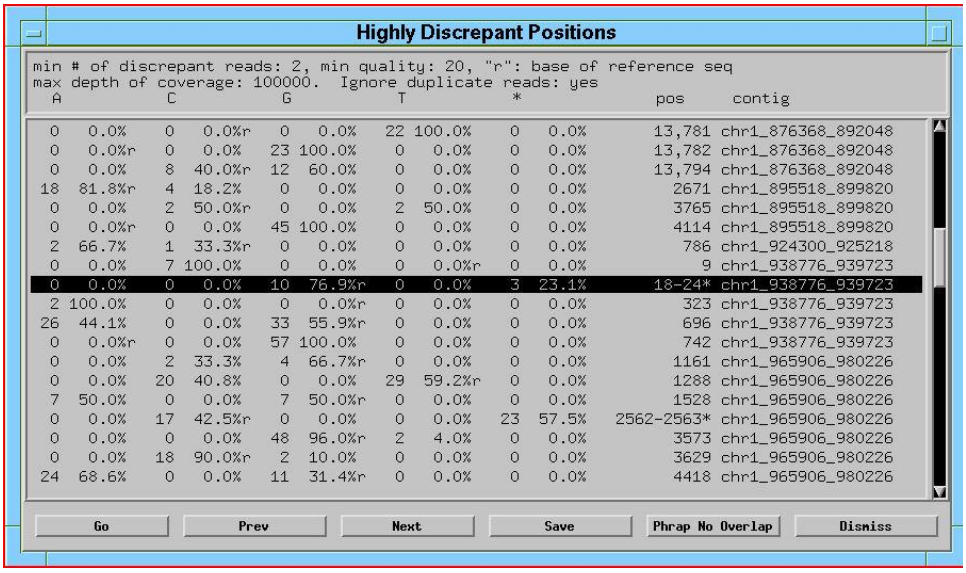
Supplementary Fig. S2. *BamScape's* Search for Problems/Variants Window. A problem can be defined as excessively high or low depth of coverage, too many reads with inconsistently placed mates or too many discrepant reads. Each filter can be turned off by unchecking it. For read base discrepancies, each 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



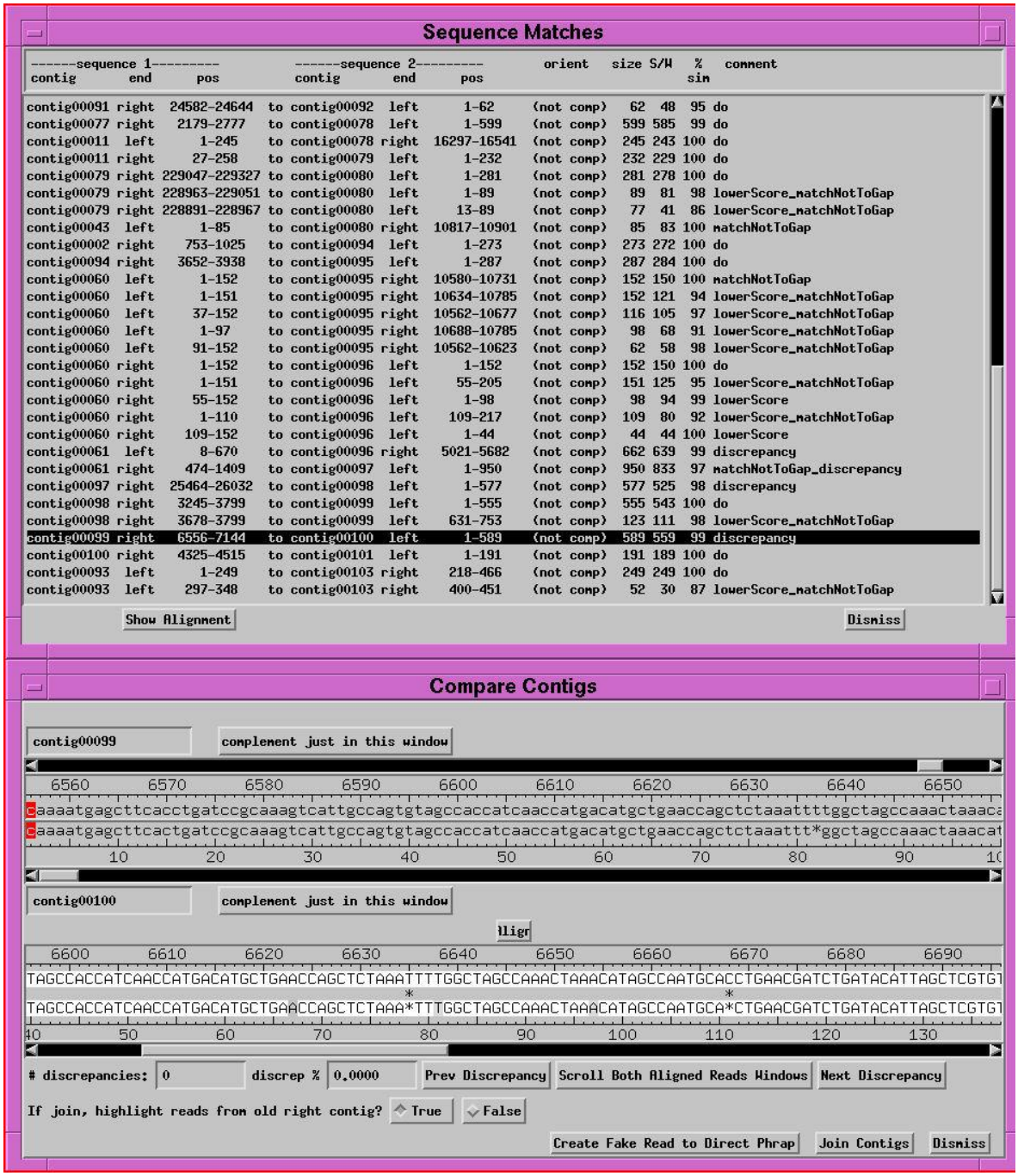
Supplementary Fig. S3. Aligned Reads Window. This example shows Illumina reads, a gene track, an indel-purified conservation track, a GERP conservation track, and bottom-strand amino acid translations. (Track files were created using the table browser at http://www.genome.ucsc.edu.) One read has been 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.



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.



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.