Supplementary Materials

Here we provide additional details about AGOUTI.

Features

1. Scaffold hundreds to thousands of contigs, yielding more contiguous assemblies;
2. Reduce the number of gene models and update them simultaneously;
3. Record any inconsistencies with the original (input) scaffolding results;
4. Support break-and-continue feature such that some time-consuming steps can be skipped if the previous run is successful;
5. Generate a dot file ready for Graphviz to visualize the scaffolding path;
6. Satisfy the requirements of good bioinformatics software proposed here: http://www.acgt.me/blog/2015/10/18/we-asked-272-bioinformaticiansname-something-that-makes-you-angry-more-reflections-on-the-poor-state-of-software-documentation

Scaffolding

AGOUTI accepts assemblies as both contigs and scaffolds. In scaffold form, AGOUTI breaks assemblies at gaps of certain lengths, essentially reducing it to contig form (a “split” assembly). AGOUTI scaffolds on split assemblies, and will report inconsistencies between the RNA-based scaffolding it conducts and the original scaffolding.

AGOUTI starts by identifying “joining-pairs,” pairs of reads that are mapped to different contigs. It is through these pairs that many of the existing scaffolding algorithms are able to connect contigs into scaffolds (e.g. Boetzer et al., 2011; Hunt et al., 2014; Mortazavi et al., 2010). AGOUTI uses only those joining-pairs that are uniquely mapped, recording the mapping positions and orientations for all identified pairs. Besides mapping quality, AGOUTI further provides two additional parameters accessible from command line to filter out suspicious alignments: maximum percentage of mismatches per alignment allowed (-maxFracMM; 5% by default), and minimum percentage of alignment length allowed (i.e. the ratio of the alignment length to the read length; -minFracOvl; 70% by default). Each filter is applied to both ends of a pair. These two options can be disabled by specifying 100% mismatch rate and 0% alignment length. All of our AGOUTI evaluation were conducted with these two parameters disabled.

AGOUTI starts by building an edge-weighted adjacency graph using these joining-pairs. In the graph, each vertex represents a contig, and an edge connects two nodes if there are supporting joining-pairs between them. A weight is put on each edge as the number of supporting joining-pairs. The graph is simplified by keeping edges with a minimum weight (5 by default).

Prior to scaffolding, AGOUTI denoises the graph by identifying and removing erroneous edges. Such edges can result from many types of errors, for example from highly similar sequences on different chromosomes. The details of this module are as follows. Because each read-pair comes from a single cDNA fragment, AGOUTI requires it to not be separated by any number of genes in between. This can be done by first checking whether the joining-pairs are mapped to the gene models at the edges of the contigs, i.e. 5’ and 3’. Specifically, AGOUTI assigns each end of a joining-pair (i.e. left or right end) as 5 or 3 if it overlaps with the gene model at 5’ or 3’ of each contig (**Supplementary Figure 2A**). Each joining-pair is thus labeled with either 5-3, 5-5, 3-5, or 3-3. If contigs contain only a single gene, reads overlapping the gene can be either 5 or 3. It is worth noting that there are cases where the mapping positions of reads fail to overlap with gene models at both 5’ and 3’ ends. If these joining-pairs fall in between the terminal gene models, they are excluded, as they are probably the result of highly similar sequences of genes in different parts of the genome (**Supplementary Figure 2B**). Otherwise, AGOUTI will keep the links and create artificial gene models at correspondent locations (**Supplementary Figure 2C-D**). The artificial gene models not used in the scaffolding are discarded from the final updated gene annotation.

In addition to ensuring that joining-pairs map to the edges of contigs, AGOUTI checks the orientation of the reads in these pairs in order to denoise the graph to be traversed. As both ends of a read-pair are inwardly sequenced, orientation imposes another important constraint, and it must be considered in combination with the end assignments. For instance, a joining-pair with a label of 5-3 and mapped in a forward-reverse fashion could span multiple gene models in the middle, and would be removed (**Supplementary Figure 2E**). AGOUTI considers a pair of contigs for scaffolding as long as the joining-pairs supporting them follows either of the four valid combinations of the end assignments and the orientations, as demonstrated in **Supplementary Figure 3A-D**. AGOUTI also keeps track of the IDs of the pair of gene models used to connect each contig pair, and their correspondent orientations.

AGOUTI first starts to traverse the graph from leaf nodes, i.e. those that connect to only one other contig, and follows the highest-weighted edges until no further extension can be made (**Supplementary Figure 4A**). For an edge to be traversed, it is required to have a minimum number of supporting joining-pairs, but AGOUTI makes this parameter accessible from command line. Each walk gives a scaffolding path, where the shortest such path includes only two contigs. This is the basic scaffolding procedure designed in RNAPATH (Mortazavi et al. 2010). This scaffolding algorithm, however, ignores subgraphs made of only non-leaf vertices (**Supplementary Figure 4B**). Rather than randomly picking one, AGOUTI traverses such a subgraph from each of its nodes following highest-weighted edges. For the same group of vertices, AGOUTI records all possible orders. AGOUTI will then identify a best traversal order among them using the following steps. With all the scaffolding paths, AGOUTI next reconciles each one using constraints imposed by constituent gene models. Specifically, it examines each pair of vertices in a path using the gene model making the connection (**Supplementary Figure 5**). This process terminates at any vertex whose connection with the next would have intervening gene models between them (**Supplementary Figure 5**). An optimal path is the one recruiting all of its vertices. AGOUTI will give up checking other possible paths once an optimal path is achieved. Otherwise, it will pick a different node, re-walk the subgraph, and reconcile the new path. After trying every vertex, AGOUTI will choose a path with the largest number of nodes. If there are two paths having the same length, AGOUTI will pick the path of the highest total weight. AGOUTI marks all the vertices in the best path as visited and prevents them from being placed multiple times.

AGOUTI updates gene models according to the new assembly obtained by the scaffolding step. For each pair of contigs within a scaffold, AGOUTI merges the two gene models from which the connection was made, reverse-complementing contigs as needed.

Comparison of AGOUTI and RNAPATH

One major difference between AGOUTI and RNAPATH is the denoising step prior to scaffolding, which removing erroneous joining-pairs. We expected a noise-free graph to result in better scaffolding. We tested this by running RNAPATH on the same six assemblies analyzed in the main text. More specifically, we compared the performance of these algorithms on two datasets, one with all the joining-pairs (including noisy pairs), and the other using only the noise-free ones. Both sets of joining-pairs came from the same RNA-seq data. We also used the default settings of RNAPATH (i.e. a minimum of 2 supporting read-pairs) for both tests. Consistent with our expectation, RNAPATH, with the additional noisy edges, recovered fewer contigs across all six assemblies (**Supplementary Table 2**). This number was boosted when the noise-free data were used (compare the first two rows of each assembly in **Supplementary Table 2**).

Second, the scaffolding algorithm in AGOUTI is guided by evidence from gene models, in addition to weights. We expected this to result in more accurate scaffolding even when noise-free datasets were used. From the aforementioned runs on the noise-free datasets, we found that RNAPATH suffered from many more inter-chromosomal errors than AGOUTI (compare **Supplementary Table 3** to **Supplementary Table 4**). These errors occurred by joining contigs from different chromosomes. In addition, RNAPATH produced intra-chromosomal errors that placed contigs of the same chromosome in the wrong order. Interestingly, we observed that RNAPATH repetitively recruited the same contigs into different scaffolds with when given noisy data, but these errors disappeared with the denoised read-pairs (**Supplementary Table 4**). These differences in error rates could be due to the difference in the minimum number of joining-pairs required by AGOUTI and RNAPATH, rather than the scaffolding algorithms. We tested this by re-running AGOUTI on the six noise-free datasets, and decreased the minimum number of supporting joining-pairs to 2, the same setting as was used for RNAPATH runs. With this smaller number, AGOUTI committed only marginally more errors (**Supplementary Table 3**). This number is consistently smaller than the one obtained from RNAPATH across all six assemblies (**Supplementary Table 4**), supporting the advantage of using gene model-guided scaffolding in AGOUTI. Similarly, increasing the minimum number of joining-pairs to 5 when running RNAPATH still resulted in more error-prone results than AGOUTI (**Supplementary Table 4**). Most importantly, across all conditions AGOUTI kept low levels of errors while placing tens to hundreds more contigs.

Finally, there were paths scaffolded by AGOUTI but entirely missed by RNAPATH, e.g. a path consisting of only non-leaf vertices (**Supplementary Figure 4B**). Because RNAPATH initiates a graph walk only from leaf nodes--and these have out-degree 1--it ignores paths without leaves. Comparing results of AGOUTI and RNAPATH, the former always placed more contigs regardless of parameter settings.

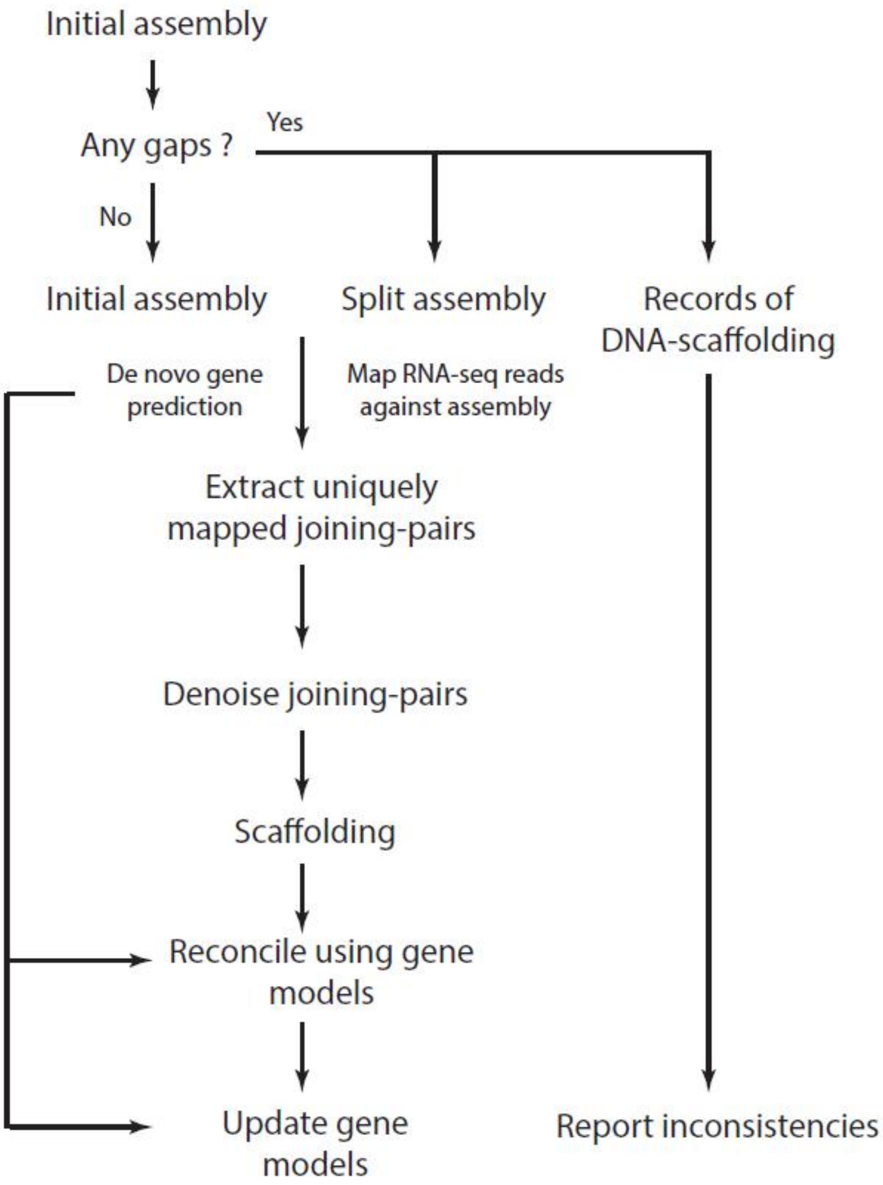
References

Boetzer, M. et al., 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4), pp.578–579.

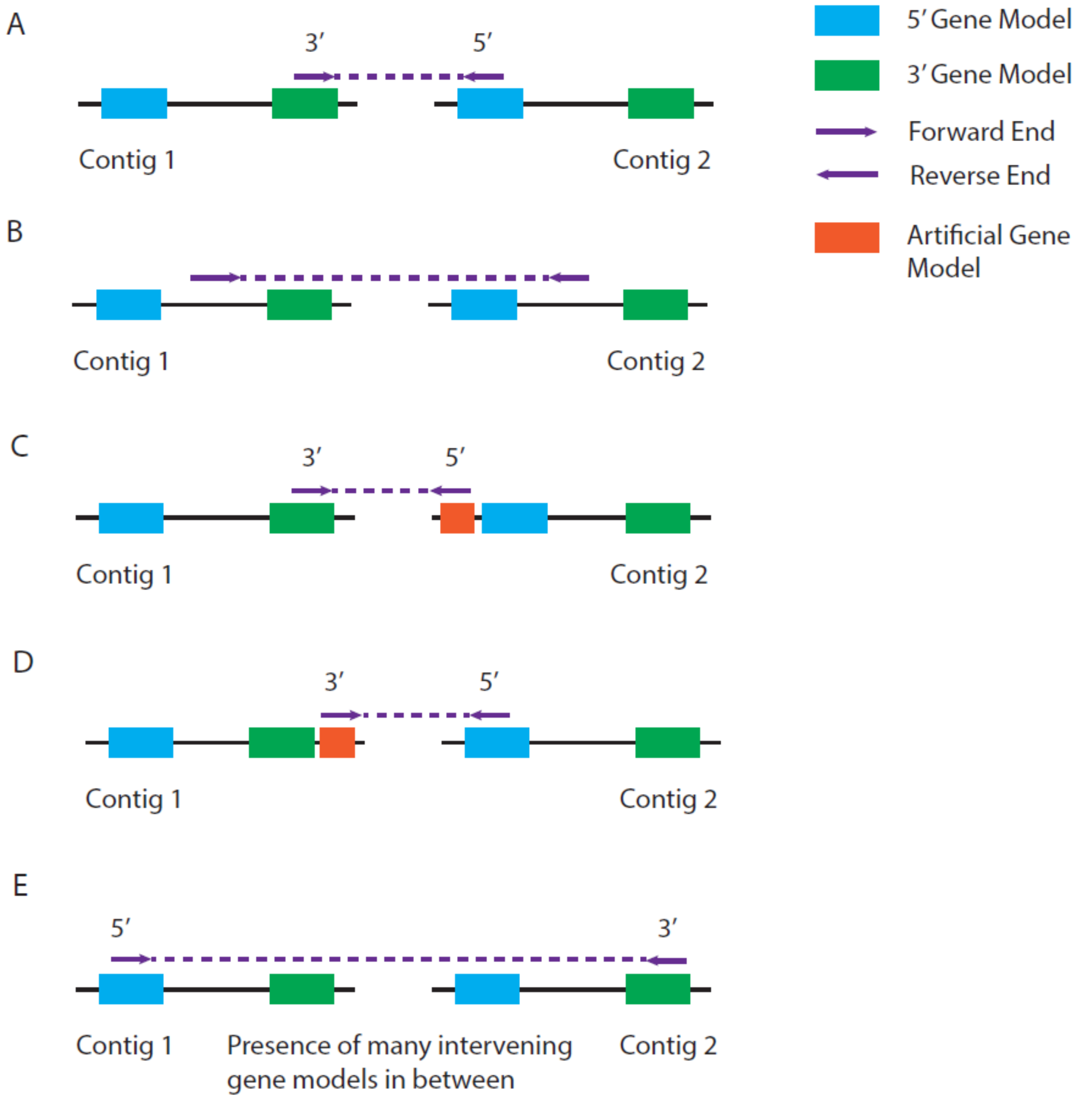
Hunt, M. et al., 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome biology*, 15(3), p.R42.

Mortazavi, A. et al., 2010. Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Research*, 20(12), pp.1740–1747.

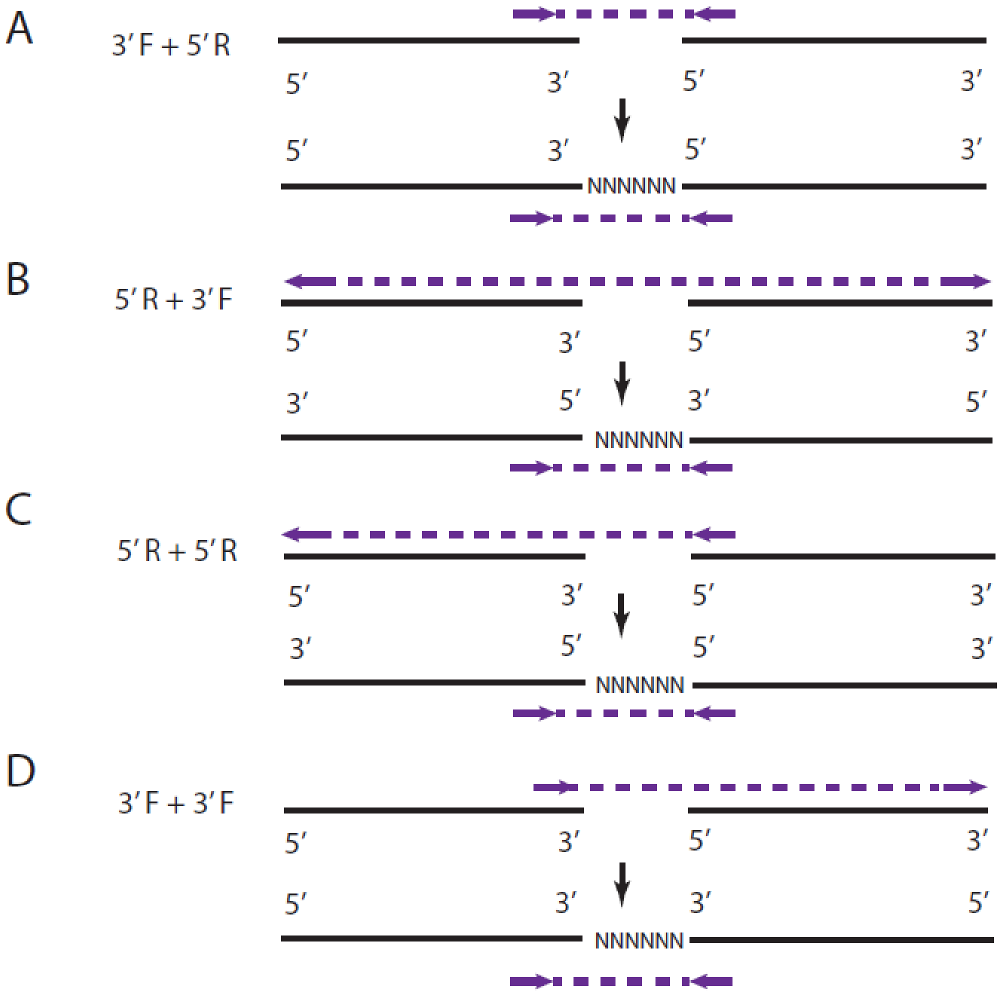
Supplementary Figure 1. AGOUTI workflow



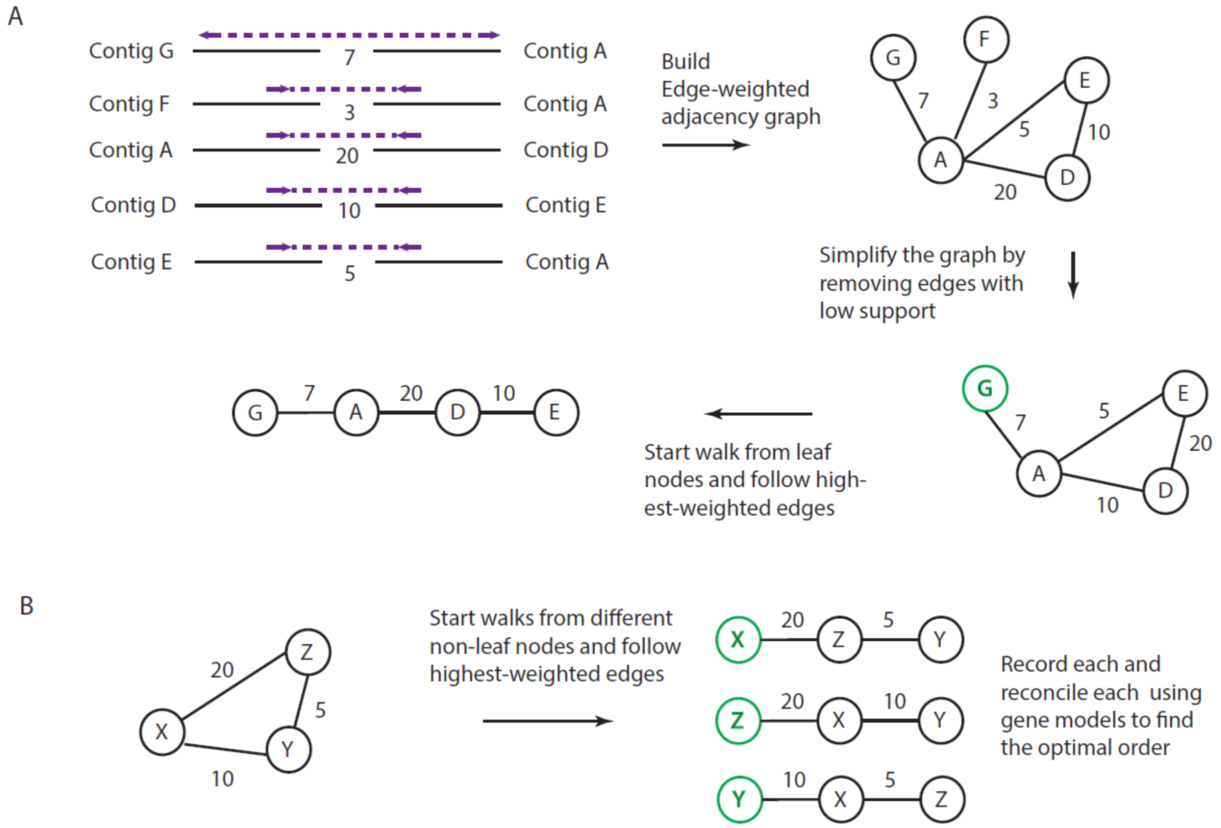
Supplementary Figure 2. Denoise joining-pairs by first making sure they are mapped to 5’-most and 3’-most gene models. (A) For each joining-pair connecting two contigs, AGOUTI assigns each end (i.e. forward and reverse) to 5’-most and 3’-most gene models on the two contigs. In this case, we have labeled the ends of the joined contigs 3’ and 5’, respectively. Doing so ensures that each joining-pair does not go across any gene models (i.e. there are no intervening gene models). (B) A joining-pair fails to map to any gene model at the edges of the two contigs. AGOUTI does not use such joining-pairs in scaffolding. (C) The reverse end of the joining-pair is mapped 5’ of the 5’-most gene model on Contig 2. AGOUTI will create an artificial gene model accordingly, and give an end label of 5’. (D) Similar to (C), the forward end is mapped 3’ of the 3’-most gene model on Contig 1. AGOUTI will create an artificial gene model and give an end label of 3’. (E) Orientation imposes an important constraint. In this case, joining the contigs in the correct orientation shows that there are multiple intervening gene models between them, and this pair is therefore ignored. Here we only show the gene models at the edges of the contigs. There can be many genes in between them.



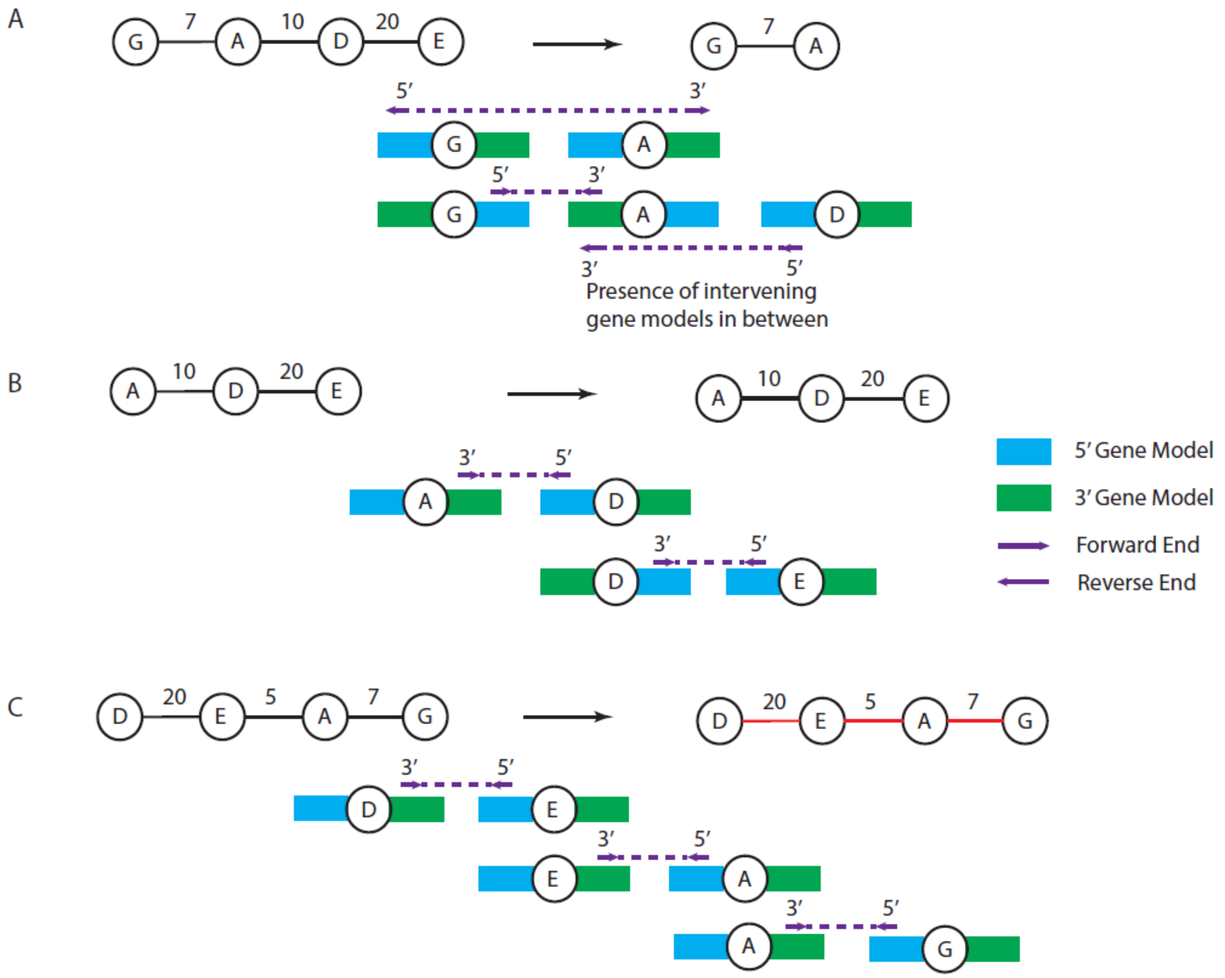
Supplementary Figure 3. Denoise joining-pairs by further considering end-assignments with orientation constraints. The top row of each case shows the combination of the end-labels and orientation of a joining-pair. The bottom row demonstrates the orientation of the two contigs with the joining-pair after scaffolding. Because of the way each read-pair is sequenced (i.e. facing each other), we need to make sure the two contigs are scaffolded in a way such that this expected orientation is not violated. There are four combinations (A-D) of the end-assignments and the orientation satisfying these requirements. For example, 5'R + 3'F means that one end of the joining-pair is mapped to the 5'-most gene model in the reverse orientation, while the other end is mapped to the 3'-most gene model in the forward orientation. If we reverse both sequences, we can make a valid scaffold between the two contigs using the joining-pair.



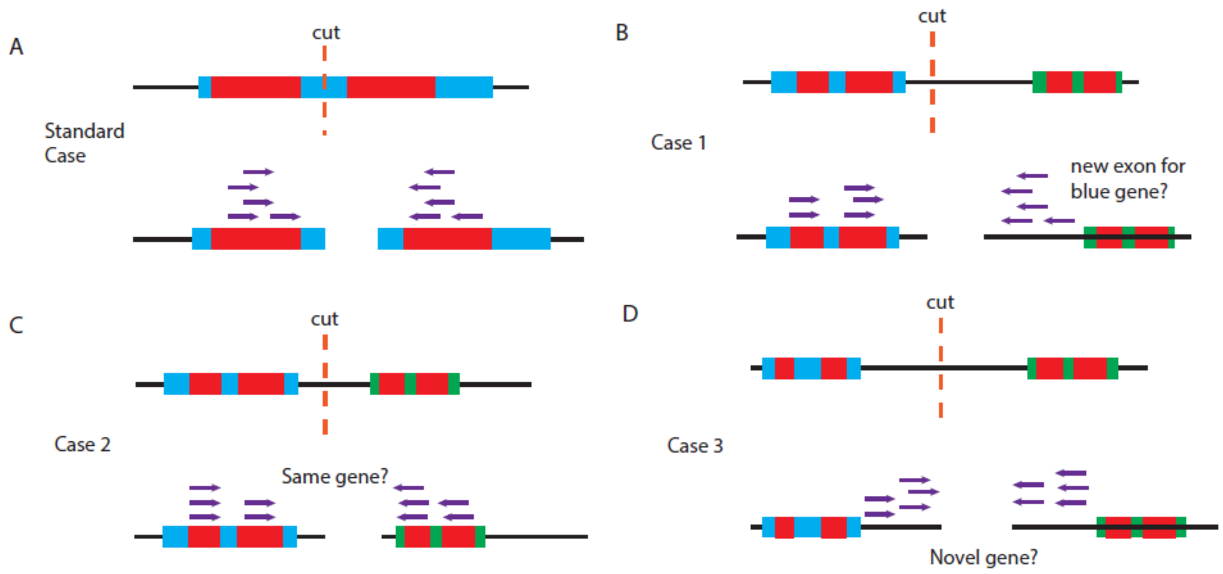
Supplementary Figure 4. Scaffolding. (A) AGOUTI first builds an edge-weighted adjacency graph made up of contigs (vertices; black lines) and the joining-pairs between them (edges; purple arrows). Edges are weighted by the number of supporting joining-pairs. The graph is further simplified by removing edges with weight less than a user-specified value, and denoised using constraints described in the text and shown in **Supplementary Figures 2 and 3**. AGOUTI starts from leaf nodes (green vertices) and follows the highest-weighted edges. Each walk gives a scaffolding path, where the shortest such path only two contigs. (B) Subgraphs with only non-leaf nodes that are ignored by RNAPATH. AGOUTI tries to traverse the subgraph starting from different vertices (green vertices). It records all the possible orders, each of which will be reconciled using constituent gene models to find the optimal one.



Supplementary Figure 5. Scaffolding reconciliation using constituent gene models. Here shows how to use gene models to reconcile scaffolding paths . Each contig is denoted by the letter in the circle. The blue and green boxes represent the gene models at the 5’ and 3’ ends of the contig. A joining-pair connecting two contigs is shown in purple, and the orientation is indicated by arrows. Contigs are reverse-complemented as needed. (A) The scaffolding path obtained by following highest-weighted edges. Examining the gene model between each pair of the contigs in the path tells us that the extension from A to D violates the requirement of no intervening gene models between two contigs. Therefore, the reconciled path contains only two contigs, rather than four. (B) The current best path is not the optimal one because it recruits only a subset of all vertices. AGOUTI therefore picks another vertex and re-traverses the subgraph. After reconciliation, the new path becomes the best path as it has more vertices than the last one. (C) Similarly, AGOUTI next starts from node D and gets a new path. The reconciled path contains all four vertices in the subgraph, and therefore AGOUTI uses it as the optimal (edges in red) one and stops checking other possible ones.



Supplementary Figure 6. Evaluation of whether each pair of contigs was connected because of the existence of an underlying gene. The top row of each panel mimics one sequence being assembled into two contigs. “Cut” is the site where the split occurs. The bottom row shows, for the two contigs, whether they are brought together because of exons of the same gene. Both “blue” and “green” boxes are genes, and arrows in purple represent joining-pairs. (A) Standard case. Two contigs are connected because they carry two exons of the same gene. 95% of the contig pairs scaffolded by AGOUTI fell in this category. (B) Case 1. Only one end of a joining-pair overlaps a predicted gene on either contig. This suggests the existence of another exon yet to be added to the same gene. (C) Case 2. The joining-pairs are mapped to two different annotated genes in the *C. elegans* genome, suggesting that the two genes should be merged into one. (D) Case 3. The joining-pairs are not mapped to any predicted genes, which may indicate the existence of a novel gene.



Supplementary Table 1. Performance of AGOUTI scaffolding.

Assembly	# contigs	# predicted gene models	Minimum supporting joining-pairs	# contigs scaffolded	# scaffolds in the final assembly	Scaffold N50	# gene models in the final assembly
1	12,196	23,822	2	5,342	8,527	36,052	21,780
			5	4,450	9,200	32,611	22,019
2	8,636	22,372	2	3,877	5,976	73,770	20,953
			5	3,235	6,452	66,927	21,071
3	7,336	21,768	2	3,091	5,243	99,384	20,657
			5	2,541	5,637	97,667	20,770
4	6,066	21,348	2	2,674	4,243	125,549	20,325
			5	2,243	4,576	119,046	20,430
5	4,586	20,719	2	1,966	3,284	258,507	19,978
			5	1,621	3,531	231,117	20,062
6	2,126	19,791	2	941	1,501	642,283	19,411
			5	766	1,625	566,481	19,455

Supplementary Table 2. Performance of RNAPATH scaffolding.

Assembly	Use of denoised joining-pairs	Minimum supporting joining-pairs	# contigs scaffolded	# scaffolds in the final assembly	Scaffold N50
1	No	2	3,421	9,841	28,769
	Yes	2	5,323	8,528	36,349
	Yes	5	4,416	9,220	32,608
2	No	2	2,430	6,933	58,959
	Yes	2	3,861	5,982	73,499
	Yes	5	3,205	6,469	66,927
3	No	2	1,980	5,968	85,802
	Yes	2	3,084	5,242	100,639
	Yes	5	2,531	5,640	97,667
4	No	2	1,618	4,937	103,844
	Yes	2	2,671	4,239	128,243
	Yes	5	2,240	4,573	119,046
5	No	2	1,225	3,760	202,360
	Yes	2	1,961	3,285	258,507
	Yes	5	1,610	3,537	231,117
6	No	2	511	1,774	492,192
	Yes	2	934	1,504	642,283
	Yes	5	763	1,625	566,481

Supplementary Table 3. Evaluation of AGOUTI scaffolding order and orientation.

Assembly	Minimum supporting joining-pairs	# contig pairs scaffolded [1]	# inter-chromosomal error	# intra-chromosomal error	# contigs placed repeatedly
1	2	3,669	2	2	0
	5	2,996	1	0	0
2	2	2,660	2	0	0
	5	2,184	0	0	0
3	2	2,093	0	1	0
	5	1,699	0	0	0
4	2	1,823	1	0	0
	5	1,490	1	0	0
5	2	1,302	0	0	0
	5	1,055	0	0	0
6	2	625	1	0	0
	5	501	0	0	0

[1] This number is calculated from the number of contigs scaffolded (Supplementary Table 1). For example, a scaffold made up of three contigs has two pairs of contigs.

Supplementary Table 4. Evaluation of RNAPATH scaffolding order and orientation.

Assembly	Use of denoised joining-pairs	Minimum supporting joining-pairs	# contig pairs scaffolded	# inter-chromosomal error	# intra-chromosomal error	# contigs placed repeatedly
1	No	2	2,366	6	7	12
	Yes	2	3,667	7	11	0
	Yes	5	2,975	2	4	0
2	No	2	1,703	8	14	1
	Yes	2	2,653	3	12	0
	Yes	5	2,166	1	4	0
3	No	2	1,378	3	10	11
	Yes	2	2,093	4	10	0
	Yes	5	1,695	0	5	0
4	No	2	1,138	7	5	10
	Yes	2	1,826	2	9	0
	Yes	5	1,492	1	10	0
5	No	2	825	1	2	0
	Yes	2	1,300	1	4	0
	Yes	5	1,048	0	1	0
6	No	2	351	6	4	0
	Yes	2	621	1	3	0
	Yes	5	500	0	2	0

Supplementary Table 5. Evaluation of AGOUTI scaffolding in terms of gene models.

Assembly	Minimum supporting joining-pairs	# contig pairs scaffolded	# contigs pairs correctly reflecting existing gene models	# Case 1	# Case 2	# Case 3
1	2	3,667	3,421	92 (19)[1]	77 (19)	77 (10)
	5	2,995	2,859	56 (16)	56 (16)	24 (3)
2	2	2,658	2,460	72 (17)	47 (13)	79 (10)
	5	2,184	2,072	39 (10)	39 (13)	34 (7)
3	2	2,093	1,928	59 (15)	42 (9)	64 (4)
	5	1,699	1,610	37 (13)	25 (5)	27 (2)
4	2	1,822	1,696	51 (11)	32 (6)	43 (5)
	5	1,489	1,427	24 (8)	26 (5)	12 (2)
5	2	1,302	1,215	41 (7)	20 (2)	26 (5)
	5	1,055	1,012	18 (5)	13 (0)	12 (4)
6	2	624	582	22 (3)	6 (1)	14 (4)
	5	501	483	10 (3)	4 (0)	4 (3)

[1] The number in the parenthesis shows the number of non-consecutive contig pairs in each case.