SUPPLEMENTARY MATERIAL FOR THE PAPER: "RASCAF: IMPROVING GENOME ASSEMBLY WITH RNA-SEQ DATA"

Authors: Li Song, Dhruv S. Shankar, Liliana Florea

Table of contents:

Additionally, scripts used in the evaluation of the software can be obtained from: https://github.com/mourisl/rascaf_paper_scripts

Figure S1. Methods – finding contig connections (*rascaf*)

There are four types of possible connections between two contigs (L_1,R_1) and (L_2,R_2) as dictated by the paired-end reads, represented by the mate edges below (thick lines). (1) Both contigs are in the forward orientation (1,2). (2) Contig 2 needs to be reversed (1,-2). (3) Contigs 1 and 2 must be swapped (2,1). (4) Contig 1 is reversed, and contigs 1 and 2 are swapped (2,-1).

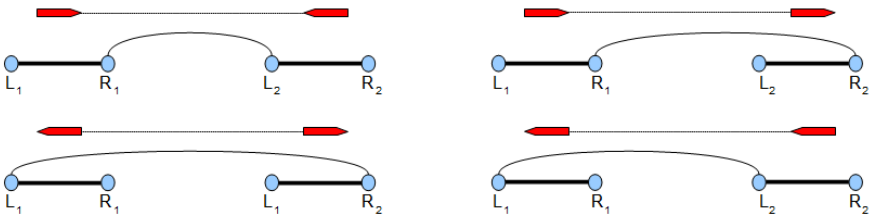

Figure S2. Methods – scaffolding *(rascaf-join)*

This example illustrates scaffolding starting from a raw assembly with 5 contigs (blue boxes): contigs 1-3 are connected into scaffold S, and contigs 4 and 5 are singletons. In stage 1, *rascaf* detects connections between contigs (1,4), (4,2) and (3,-5). The resulting contig graph has 5 (contig) nodes and 10 edges: 5 *contig edges* (thick lines), which connect the two nodes representing a contig, and 5 *mate edges*, representing either adjacency relationships in the original scaffold ((1,2) and (2,3)) or connections detected by *rascaf* based on RNA-seq paired-end reads ((1,4), (4,2) and (3,-5)). Starting from scaffold S, *rascaf-join* then traverses the contig graph to determine a bi-connected component and then uses a topological sort algorithm to determine a component path (scaffold) containing contigs 1,4,2 and 3 that satisfies all precedence relationships. Note that contig 5 is not part of the bi-connected component, as L_5 and R_5 can be reached from $L_1 \rightarrow R_1$ but not from $R_3 \rightarrow L_3$ in scaffold S. We call R_3, L_5, R_5 a *dangling* path. To add dangling paths to the scaffold, *rascaf-join* traverses the component path starting from the last contig and moving backwards. In iteration i, it greedily chooses a maximal path starting with $L_i \rightarrow R_i$ and trims it if it reaches a contig that has already been added to the scaffold. This path is then inserted in the component path (scaffold) following node R_i. Reverse dangling paths, which can be reached from $L_3 \rightarrow R_3$, are analyzed similarly.

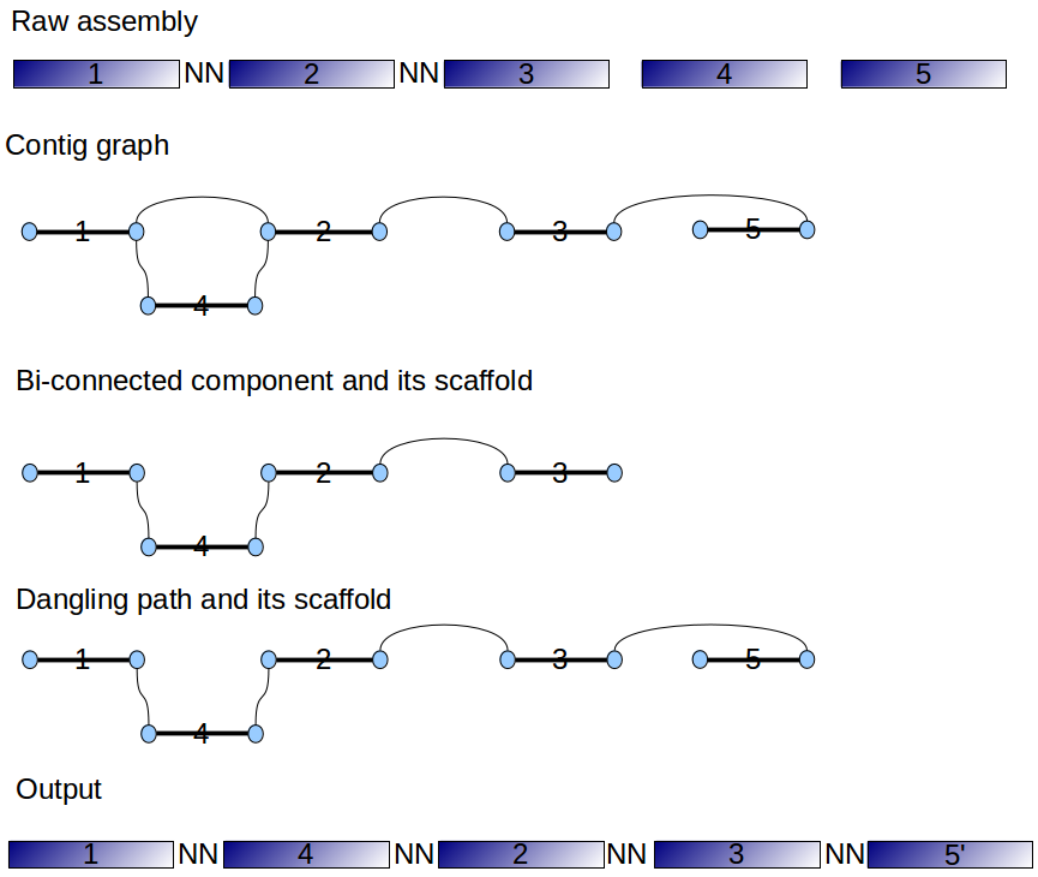

Table S3. Performance of programs on the simulated data set. Sn = TP/M, Pr = STP/N, M = 590.

Program	N	TP	STP	Sn	Pr
Rascaf	247	273	247	0.463	1.00
Rascaf+BWA-mem	395	450	393	0.763	0.995
L_RNA_scaffolder	375	437	375	0.741	1.00
AGOUTI	197	208	192	0.352	0.974
AGOUTI+BWA-mem	374	416	364	0.705	0.973

Table S4. Effects of library insert size on program performance, on the simulated data. A) Sn and Pr are defined as in Table S3, and m = average insert size for the library.

Program	m=174 bp		m=225 bp		m=275 bp	
	Sn	Pr	Sn	Pr	Sn	Pr
Rascaf	0.463	1	0.596	1	0.656	1
Rascaf+BWA-mem	0.763	0.995	0.774	0.997	0.740	0.993
L_RNA_scaffolder	0.741	1	0.739	0.995	0.725	0.984
AGOUTI	0.352	0.974	0.498	0.989	0.573	0.987

Table S5. Alignment statistics for the ERR430941 and SRR1930097 RNA-seq data sets mapped with HISAT to the Fragaria genomes.

Species	Mapped	%	>1 matches	%	Non-concordant	%
ERR430941 (68,982,904 reads)						
F. iinumae	47,855,465	69.4	452,802	0.9	9,074,580	19.0
F. nipponica	40,638,897	58.9	539,286	1.3	11,633,842	28.6
F. nubicola	42,135,327	61.1	562,084	1.3	11,543,566	27.4
F. orientalis	34,778,425	50.4	673,390	1.9	12,035,649	34.6
F. x ananassa	46,308,597	67.1	6,003,079	13.0	13,675,282	29.5
SRR1930097 (76,157,614 reads)						
F. iinumae	47,549,112	62.4	578,913	1.2	10,370,142	21.8
F. nipponica	46,033,521	60.4	805,434	1.7	10,940,696	23.8
F. nubicola	51,722,939	67.9	555,747	1.1	9,644,349	18.6
F. orientalis	43,478,759	57.1	525,460	1.2	10,262,802	23.6
F. x ananassa	49,274,091	64.7	6,367,284	12.9	12,587,844	25.5

Table S6. Evaluation of *A. thaliana* assemblies improved with AGOUTI and L_RNA_scaffolder. AGOUTI was applied iteratively with 1,2,…11 RNA-seq data sets, whereas L_RNA_scaffolder was run with the full set of transcripts.

AGOUTI (iterative)					
Data sets	0 (raw)	1	2	6	11
Scaffolds	8,082	7,771	7,679	7,174	7,109
NGA50	42,479	45,667	47,021	50,027	51,316
Misassemblies	1,153	1,177	1,209	1,401	1,417
Problematic scaffolds	1,412	1,433	1,439	1,450	1,454
Effective misassemblies	-	10	22	81	84
L_RNA_scaffolder (batch)					
Scaffolds	8,082	7,761	7,502	7,154	6,961
NGA50	42,479	44,399	45,491	46,328	46,627
Misassemblies	1,153	1,296	1,459	1,731	1,896
Problematic scaffolds	1,412	1,536	1,618	1,772	1,870
Effective misassemblies	-	114	216	416	566

Table S7. Number of genes with increased coverage in the *A. thaliana* assemblies after applying Rascaf with 0, 1, …, 11 RNA-seq data sets.

Assembly	Genes with gain > 0	Genes with gain >5%
A. thaliana 0 (raw)	na	na
A thaliana 1	506	479
A. thaliana 2	580	549
A. thaliana 3	718	668
A thaliana 4	869	811
A. thaliana 5	889	830
A. thaliana 6	937	872
A. thaliana 7	945	880
A. thaliana 8	985	919
A. thaliana 9	985	919
A. thaliana 10	985	917
A. thaliana 11	991	922

Figure S8. Transcript coverage plots for the five Fragaria species, before and after application of Rascaf.

For each assembly, the number of transcripts with a fraction x or more of bases contained in the primary alignment (y-axis) was plotted for coverage cutoffs (x-axis) x = 0, 0.05, …, 0.95, 1.00. Assemblies modified with Rascaf have higher rates of transcript coverage (higher curves), for any coverage cutoff.