Rascaf: Improving Genome Assembly with RNA Sequencing Data

Li Song, Dhruv S. Shankar, and Liliana Florea*

Abstract

Abundant but short second-generation sequencing reads make assembly difficult, leading to fragmented genomes and gene annotations. Gene structure information from RNA sequences can be used to improve the completeness and contiguity of an assembly, but bioinformatics methods have been lacking. Rascaf is a highly efficient tool leveraging long-range continuity information from intron spanning RNA sequencing [RNA-seq] read pairs to detect new contig connections. It determines a heaviest path in an exon block graph that simultaneously represents a gene and the underlying contig relationships. Rascaf is more accurate than its competitors, highly precise, and finds thousands of new verifiable connections in several draft Rosaceae genomes. Lightweight and practical, it can be readily incorporated into sequencing pipelines to improve an assembly and its gene annotations.

Recent years have seen a tremendous increase in the number and diversity of sequenced genomes (Reddy et al., 2015). More than 13,000 eukaryotes have been sequenced or are in the process of sequencing, and more are planned including hundreds of plants and animals. Most model organisms have been sequenced under the umbrella of large genome projects undertaken by broad international consortia with the aim to create high-quality reference sequences (Arabidopsis Genome Initiative, 2000; Lander et al., 2001; Venter et al., 2001; Waterston et al., 2002, Rat Genome Sequencing Project Consortium, 2004; International Human Genome Sequencing Consortium, 2004; Tuskan et al., 2006). In recent years, second-generation sequencing technologies have dramatically accelerated the pace of generating new genomes,

Core Ideas

- Rascaf simultaneously improves an assembly and its gene annotations
- Rascaf finds thousands of new contig connections in draft Fragaria genomes
- Rascaf is a highly accurate, efficient, and practical tool for plant genome sequencing projects

L. Song and L. Florea, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, 1900 E Monument, Welch 113, Baltimore, MD 21205 and Dep. of Computer Science, Johns Hopkins Univ., 1900 E Monument, Welch 117, Baltimore, MD 21205; D.S. Shankar, Dep. of Biomedical Engineering, Univ. of North Carolina, 333 S. Columbia St., Chapel Hill, NC 27514. Received 11 Mar. 2016. Accepted 11 May 2016. *Corresponding author (florea@jhu.edu). Assigned to Associate Editor Scott Jackson.

Abbreviations: cDNA, complementary DNA; mRNA, messenger RNA; RNA-seq, RNA sequencing; STP, strong true positive; TP, true positive.
as reduced sequencing costs along with increased access to sequencing have made it possible for groups and even individual investigators to sequence the genome of the species they study. Virtually all of these projects will produce draft versions of the genomes, in which the chromosomes are assembled into a relatively large number of contigs separated by gaps. Annotation software will then use the contigs, typically within groups of contigs with known order and gap sizes (scaffolds) or full chromosomes, as the substrate on which to identify genes.

During a typical genome assembly process, overlapping reads are first used to build contigs, then contigs are connected into larger scaffolds using order and orientation information from mate–pair reads. Mates are sequenced from the two ends of DNA fragments in a size-selected library, and their relative distance (insert size), order, and orientation on the originating DNA sequence can be estimated with relatively high accuracy. Repetitive regions in the genome pose a significant challenge to assembly algorithms. To be able to reconstruct these sequences, insert sizes need to exceed the length of the repeat to allow anchoring the assembly onto the non-repetitive flanking regions. Therefore, a typical genome assembly project will require multiple insert-size libraries, spanning from 500 bp to 8 to 10 kb. There is a rich body of work in developing scaffolding algorithms based on mate pairs from whole-genome sequencing dating back to the assembly of the first sequenced eukaryotic genomes (reviewed in Hunt et al., 2014). However, building the critical long-insert libraries is expensive and labor intensive.

Once a draft genome sequence is produced, the first and most crucial step in its analysis is finding the genes, which then provide the basis for downstream studies of gene function and variation. Deep RNA-seq has become the primary means to characterize the genes of a species, and there are already a number of high-performance tools for RNA-seq read analysis, including alignment and transcript assembly tools (Wang et al., 2010; Trapnell et al., 2012; Dobin et al., 2013; Florea and Salzberg 2013; Kim et al., 2013; Mezlini et al., 2013; Song and Florea 2013; Kim et al., 2015). It also provides critical information about species-specific genes and alternatively spliced variants, including novel protein-coding genes and non-coding RNAs. Errors and gaps in the assembly can however interfere with correct gene and transcript annotation by fragmenting the genes, deleting or scrambling the exons, and by locally altering the gene’s sequence (Florea et al., 2011). Therefore, to aid investigators in their gene studies, every effort must be made to improve the quality of the assembly, particularly in the gene regions.

Gene structures, in which introns may span thousands of bases, provide an effective way to increase the completeness and continuity of an assembly in silico. Several genome sequencing projects, starting with the human genome, have used gene information from independently generated expressed sequence tags and full-length messenger RNAs (mRNAs) to detect assembly errors or recruit additional contigs into the assembly (Venter et al., 2001; Dalloul et al., 2010; Shulaev et al., 2011; Wegryn et al., 2014). However, tools that could be systematically applied to any genome project and take advantage of the next-generation sequencing data being generated have been lacking. Traditional mate pair-based scaffolding methods (Pop et al., 2004; Dayarian et al., 2010; Boetzer et al., 2011; Gao et al., 2011; Donmez and Brudno 2013) rely on a uniform read coverage of the genome and a statistically well characterized insert-size distribution and cannot be directly applied to RNA-seq reads. Only two tools have been recently developed that take advantage of next-generation RNA sequencing: L_RNA_scaffolder (Xue et al., 2013) applies the gene-based approach using de novo transcript assemblies generated with tools such as Trinity (Haas et al., 2013), whereas AGOUTI (Zhang et al., 2015) employs the RNA-seq read alignments directly, in the context of known gene annotations, to detect new connections. However, de novo assembly of RNA-seq sequences as employed by L_RNA_scaffolder is challenging and error prone as well as time consuming. Chimeric transcript reconstructions can lead to incorrect scaffolds, and low-expression genes may be only partially reconstructed or missed entirely and therefore have limited impact.

We developed Rascaf (RnA-SCAFfolder), a novel tool that uses the alignments of RNA-seq reads to identify new contig connections in a fragmented genome and improves the completeness and accuracy of the genes and genome simultaneously. Rascaf uses an exon block graph to simultaneously represent a gene and the underlying contig relationships and to determine a heaviest contig path. Suggested contig connections can then be optionally validated by database searches for cross-species complementary DNA (cDNA) and protein evidence. When evaluated on both simulated and real sequence data, and against similar tools, Rascaf was both more accurate and highly efficient and therefore can be effectively used to increase the quality of new genome assemblies of plants and animals. More specifically:

1. Rascaf simultaneously improves an assembly and its gene annotations, resulting in longer scaffolds, more accurate scaffolds, and more complete gene models.
2. It has higher or comparable accuracy to the best of the other tools for each application tested.
3. Rascaf is highly precise with only a handful of misassemblies introduced, has a small memory footprint, and runs in minutes on a regular workstation for a typical RNA-seq data set and genome.
4. Rascaf identified 1000 to 10,000 new contig connections in the draft genomes of several Fragaria species and of the Rosaceae pear (Pyrus communis L.), thus increasing their utility.
5. The program can be used with a single or with multiple RNA-seq data sets simultaneously.

6. An optional in silico validation step searches the predicted contig joins against external cDNA or protein databases for independent evidence.

Rascaf is available free of charge under the GNU Public License from https://github.com/mourisl/Rascaf.

Materials and Methods
Rascaf builds an improved assembly in two stages. Stage 1, implemented in the program rascaf, determines a set of possible contig connections based on continuity information from alignments of paired-end RNA-seq reads. Once a set of possible connections is determined, Stage 2, implemented in the program rascaf-join, uses the connections to scaffold the new assembly and to generate the new genome sequence. The general framework is illustrated in Fig. 1, and the data structures and methods are described below.
Detecting Contig Connections

Step 1
The input to Rascaf is the draft genome in FASTA format and an alignment file of paired-end RNA-seq reads. If the assembled genome is in scaffolds, Rascaf first converts it to a contig-level (raw) assembly by splitting the sequences at runs of Ns.

Steps 2 and 3
The basic data structure employed by Rascaf is the exon block. An exon block denotes a maximal set of consecutive genomic coordinates covered by aligned RNA-seq reads corresponding to a block of overlapping exons. A gene block is an ordered set of exon blocks connected by spliced alignments corresponding to a portion of a gene located on the same contig.

Step 4
Rascaf builds a gene block graph as follows. Each gene block is represented by a pair of vertices \((L, R)\) connected by an edge (block edge). When the two reads in a pair span different gene blocks, mate edges are added to connect the \(L\) endpoint of one gene block to the \(R\) endpoint of the other (Supplemental Fig. S1). This data structure is similar to the contig graph in Huson et al. (2002). One important constraint on the gene block graph is that every path must alternate block and mate edges. Hence, setting the direction of one edge in a path will determine the directions of all remaining edges such that the concatenation of contig sequences along the path spells either the sequence of the genome or its reverse complement.

Each component of the gene block graph corresponds to a gene or a portion of a gene. Rascaf employs a greedy method to find the order of the gene blocks in each component, choosing the most supported mate edge and then reiterating the search to extend the path in both directions. The procedure is terminated when there is no possible extension, on encountering a previously visited node, or at a sudden significant drop in read support. The algorithm produces a path of gene blocks, or block scaffold. If there are any remaining edges where neither of the adjacent nodes was selected, the procedure is repeated to find additional block scaffolds.

Sequencing and alignment errors may create false mate edges, leading to chimeric scaffolds. Rascaf uses alignment, read pair, and genomic context information to filter likely false positives. More specifically, Rascaf removes a mate edge if (i) it is supported by fewer than \(K\) reads (by default, \(K = 2\)); (ii) there are multiple possible connections with similar support, indicating an ambiguous and potentially error-prone connection; (iii) the concatenated sequence of exon blocks does not fit the mean and standard deviation of the insert size distribution for the RNA-seq reads; and (iv) the gene blocks appear to be duplicated in the assembly based on the overlap between their \(k\)-mer profiles, potentially indicating a paralogous connection.

Step 5
Once a set of block scaffolds is constructed, they are used as guides to find contig connections. Rascaf iteratively parses each block scaffold, starting from the one with the strongest read support, to create a list of contig connections and to decide the order and orientation of each contig within the scaffold. Connections that are incompatible with previously ordered contigs are ignored. In the end the procedure, implemented in the program rascaf, will determine a set of contig connections with known relative order and orientation.

Scaffolding Guided by Connections

Step 6
Once a set of contig connections is determined, rascaf-join incorporates them into a scaffolding algorithm to create a new assembled sequence. One ancillary benefit of separating the scaffolding from the detection of contig connections is that it allows combining multiple RNA-seq data sets, leveraging the variability in gene expression levels across the samples. For instance, the locus of a low-expression gene in one sample may be difficult to scaffold because connections here are hard to distinguish from noise, but this drawback can be mitigated when the gene is more richly covered in another data set.

Rascaf-join builds a contig graph that is similar in concept to the gene block graph described earlier. More specifically, each contig is represented by two vertices \((L_c, R_c)\) connected by a contig edge. Two contigs are connected by a scaffold edge if they are adjacent in a scaffold in the raw assembly or are part of a contig connection identified by Rascaf. With multiple RNA-seq data sets, scaffold edges from different data sets could potentially introduce cycles. Rascaf-join detects any cycles in the contig graph using a depth-first search algorithm and removes all scaffold edges previously identified by rascaf that are adjacent to the contigs in the cycle to create an acyclic graph. It then attempts to improve each scaffold in the original assembly, starting from the longest, as described below.

Given a scaffold \(S\) in the raw original assembly, rascaf-join attempts to fill gaps in \(S\) and to extend it from both ends. Suppose \(S\) contains \(n\) contigs, with the associated contig nodes \(L_1, R_1, \ldots, L_n, R_n\). Rascaf-join first finds the biconnected component in the contig graph containing all contigs from \(S\) (the biconnected component is a subgraph such that every node can be reached both from the path starting with \(L_1\overrightarrow{R_1}\) and from the path starting with \(R_n\overrightarrow{L_n}\)). Intuitively, a biconnected component contains the contigs from \(S\) as well as those contigs on a path that branches off and then returns to \(S\). It then converts the component into a directed acyclic graph by fixing the path starting with \(L_1\overrightarrow{R_1}\). Further, it uses a topological sort algorithm (Cormen et al., 1990) to order the contigs in the connected component and to produce a longer scaffold \(S'\) (Supplemental Fig. S2). In the end, Stage 2 generates a new assembly by recruiting...
additional contigs informed by the identified contig connections while adjusting the existing scaffolds as necessary to create more complete gene models.

While RNA-seq data present many advantages for genome scaffolding, it also has its drawbacks. For instance, the paired reads’ inner distance along the genome, which may include introns of unknown sizes, cannot be characterized statistically, which can introduce ambiguity in the contig order and may lead to local rearrangements within a scaffold. This is especially problematic for genomes with very long introns and short contigs, in particular, with contigs located entirely within introns of the genes. Therefore, while using RNA-seq or gene structure information provides a highly practical solution to filling in the scaffold structure and building more complete gene models, further validation using, for instance, optical or physical maps and other data types may be necessary to resolve the local contig order at high resolution.

**Results**

**Performance Evaluation on Simulated Control Data**

To obtain an accurate evaluation of performance on control data, we applied each program (Rascaf v.1.0.1; L_RNA_scaffolder (v. June 2013) and AGOUTI v.0.3.0-22) to a simulated data set. We generated an artificial genome by extracting the sequences of human chromosomes 1 and 12 and splitting them into contigs. For a more realistic model, we followed the contig size distribution of the *Prunus persica* v.1.0 genome (International Peach Genome Initiative, 2013; www.rosaceae.org). In parallel, we used FluxSimulator (Griebel et al., 2012) to generate 100 million 100-bp paired-end RNA-seq reads with average insert size 174 bp, respectively, using the GENCODE v.22 (www.gencode.org) gene annotations as reference. No chimeric reads were included. For Rascaf and AGOUTI, we mapped the ~15 million reads from chromosomes 1 and 12 to the contigs with a fast-spliced aligner, HISAT (Kim et al., 2015). Additionally, Rascaf has been adapted to incorporate partial alignments generated with BWA-mem (Li and Durbin, 2009), potentially obtained from spliced alignments spanning multiple contigs. For L_RNA_scaffolder, we first assembled the reads into transcripts with Trinity, and for AGOUTI we used as input annotation the gene models produced from the RNA-seq reads by Cufflinks (Trapnell et al., 2009).

To create a gold reference, we consider all pairs of ordered and oriented contigs in the assembly that are supported by read pairs. Let $M$ be the size of this set and let $N$ be the number of scaffold edges in the set of contig paths predicted by the program being evaluated. A contig pair $(c_1,c_2)$ in the reference data set is said to be satisfied, or is a true positive (TP), if $c_1$ and $c_2$ appear in the same order and orientation in a contig path ($c_1$ and $c_2$ need not be adjacent). Conversely, a scaffold edge in a contig path is said to match the reference, or is a strong true positive (STP), if its two contig nodes are connected by a read pair in the gold reference in the same order and orientation. With these concepts in place, we define sensitivity $S_n = TP/M$, and precision $P_r = STP/N$.

For this test case, Rascaf using both full and partial alignments had the best overall performance, at 0.763 sensitivity and precision 0.995, with L_RNA_scaffolder a very close second, at 0.741 sensitivity and precision 1.0 (Fig. 2; Supplemental Table S3). This simple simulated example also illustrates one limitation when using RNA-seq reads directly without prior assembly, as implemented in Rascaf and AGOUTI, particularly for RNA-seq libraries with short insert sizes. Intuitively, a short insert size produces more read pairs sampled from the same exon and where one or both reads could span the boundary of the intron. Such cross-contig spliced alignments are missed by current alignment software. Indeed, both Rascaf’s and AGOUTI’s performance is improved.

**Fig. 2.** Performance evaluation of programs on simulated data. Sensitivity ($S_n$) and precision ($P_r$) are defined in the text.
with longer fragments (Supplemental Table S4), and their relative performance vis-à-vis the assembly-based L_RNA_scaffold is fully recovered when incorporating partial (clipped) alignments produced with BWA-mem.

In Silico Validation of Pyrus communis Genome Improvement

To assess the usefulness and accuracy of programs on a real assembly project, we applied them to improve the completeness and contiguity of the P. communis (‘Bartlett’) genome (Chagne et al., 2014). The pear genome has recently been sequenced using second-generation (Roche 454) single-end reads from 2- and 7-kb insert libraries, resulting in a 577 Mb assembly in 142,083 scaffolds (184,520 contigs). Since no gold reference is available in this case, we performed an in silico validation by searching the concatenated gene sequences spanning each contig connection against the RefSeq gene database (BLASTn, dc-megablast option [Altschul et al., 1997]). Note that since AGOUTI does not report the underlying gene structure, we could not include it in the evaluation. A match to a database homolog that spans the connection then greatly increases the confidence of the prediction. We queried each connection sequence and inspected the BLAST alignments for evidence of consistent coverage spanning the junction point. We deemed an alignment to be positive, and therefore provide proof for the connection, if all gene block sequences were contained in the alignment in the same order or orientation. We then distinguish between uncertain and potentially novel connections, in which alignments are compatible in order and orientation but cover only a subset of the gene blocks, and likely erroneous connections, which either show rearrangements between the gene block segments or portions of the segments (rearranged) or in which portions of the query sequence match different sequences in the database (chimeric). Note that both chimeric and negative connections may in fact reflect errors in other species’ genomes or gene annotations rather than decision errors made by the tools. Lastly, since the programs may also report connections between the contigs from the same scaffold that are already known, we excluded any such connections from the performance measurements below.

When run with the SRR1609135 RNA-seq data set, comprised of 24.2 million 101-bp paired-end reads sampled from pear leaves, Rascaf produced 1286 and L_RNA_scaffold generated 707 new putative connections (Table 1). Of these, 1218 (94.7%) of Rascaf connections were classified as positive, and an additional 55 (4.3%) were either uncertain or with no homolog in the database and could potentially represent connections from novel genes or extensions of annotated genes. For L_RNA_scaffold, 474 (67%) were validated, and an additional 152 (21.5%) were uncertain or unaligned. Only 13 (1.0%) Rascaf connections were chimeric or showed evidence of rearrangement compared with 81 (11.5%) for L_RNA_scaffold. Therefore, Rascaf detected >2.5× as many positive (validated) connections than L_RNA_scaffold and twice as many likely connections when the unaligned and uncertain cases were included, and reported four times fewer chimeric and rearranged (negative) cases. Hence, it was both more sensitive and more precise than L_RNA_scaffold by a wide margin.

An example of positively validated connection at the P. communis locus homologous to the Malus × domestica GDSL esterase/lipase At3g48460-like gene region (accession: XM_008395632.1) is shown in Fig. 3A, and an uncertain connection at the P. × bretschneideri peptidyl-prolyl cis-trans isomerase CYP71-like gene homolog (accession: XR_668135.1) missing support for its terminal 256 bp is shown in Fig. 3B. For some of the connections classified as negative, the rearrangement resided in an existing contig rather than being introduced by our procedure (e.g., P. × bretschneideri cryptochrome-1-like gene homolog, accession: XM_009380716.1; Fig. 3C). Lastly, Fig. 3D illustrates a chimeric connection as a result of a duplication within the M. × domestica UDP-glycosyltransferase 74F1-like gene homolog (accession: XM_008396047). Therefore, Rascaf achieved >95% precision and can be highly trusted to improve the sequence of a draft genome assembly while keeping the number of errors to a minimum.

Assembly Improvement using Multiple RNA Sequencing Data Sets

We further assessed the programs’ performance on a high-quality genome system, which can serve as a more realistic control experiment. The 121-Mb genome of the model plant Arabidopsis thaliana Col-0 (Arabidopsis Genome Initiative, 2000) has been reported and is one of the most extensively studied systems to date and therefore is the closest genome model yet to serving as a gold reference. As an additional goal, we sought to determine the feasibility and benefits of improving the assembly using multiple RNA-seq data sets simultaneously.

We used whole-genome DNA sequence data (SRA accession: SRR1810274; 60 million 100-bp reads) and assembled it with SOAPdenovo2 (Luo et al., 2012). After filtering small scaffolds shorter than 500 bp, the draft genome assembly consisted of 37,948 contigs organized in 8082 scaffolds. We also downloaded 11 RNA-seq data sets sampled from plant leaves, root, and shoot apex (SRR2187604, SRR2080045, SRR971148, SRR1106559, SRR1187932, SRR1781769, SRR2060632, SRR2061405, SRR2895388, SRR2895627, and SRR2895761; HISAT alignment statistics are included in Supplemental Table

<table>
<thead>
<tr>
<th>Program</th>
<th>Validated</th>
<th>Uncertain</th>
<th>Unaligned</th>
<th>Rearranged</th>
<th>Chimeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rascaf</td>
<td>1218</td>
<td>42</td>
<td>13</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>L_RNA_scaffold</td>
<td>474</td>
<td>128</td>
<td>24</td>
<td>38</td>
<td>43</td>
</tr>
</tbody>
</table>

Table 1. In silico validation of programs on the Pyrus communis genome by BLAST searches against the National Center for Biotechnology Information RefSeq mRNA database.
S5). For the multisample analysis, we added each data set to a growing pool to incrementally evaluate their impact on assembly improvement. Among the programs, only Rascaf can seamlessly integrate multiple RNA-seq data sets for analysis, potentially identifying and resolving internal conflicts. Nevertheless, we also ran L_RNA_scaffolder on the combined sets of transcripts assembled from the RNA-seq data. For AGOUTI, which was not designed to handle multiple data sets simultaneously, we ran the process iteratively (Supplemental Table S6).

Comparative Evaluation of Programs

We first analyzed the performance of all programs by measuring the improvement on the assembly for a single RNA-seq data set (SRR2187604) using the Quast (Gurevich et al., 2013) evaluation software to compare against the reference A. thaliana genome. To calculate its statistics, Quast analyzes all scaffolds 500 bp or longer by aligning them to the reference genome with nucmer (Kurtz et al., 2004). Rascaf found the most new connections, reducing the number of scaffolds by ~400, with or without incorporating partial BWA-mem alignments. All three tools improved the NGA50, a measure of assembly continuity, by 2 to 4 kb (5–9%) as shown in Table 2 (top) (NGA50 is defined as the minimum length of a scaffold alignment such that 50% of the aligned portion of the assembly to the reference genome is in scaffold alignments this size or longer). At the same time, Rascaf and AGOUTI introduced a comparable number of misassemblies, while L_RNA_scaffolder was more imprecise. More in-depth analyses of the putative misassemblies introduced by the programs revealed that, in fact, most
of these involved contigs that were misassembled in the original SOAPdenovo2 assembly. In several other cases, the reported misassembly was due to the intercontig gap length (1 kb) default parameter setting in Quast, which was too short to accommodate gaps potentially introduced by introns. After correcting for measurement errors and errors propagated from problematic contigs within the original assembly, the number of effective misassemblies introduced by each program was significantly reduced to 10 to 114 (0.8–8.8% of the total). Using Rascaf with additional partial BWA-mem matches did not bring significant improvement, likely because of the longer insert size in the RNA-seq library (245 bp) coupled with shorter exon sizes in *A. thaliana*. Overall, Rascaf demonstrated better performance than the other tools, and both read alignment-based tools were considerably more precise than the transcript-based L_RNA_scaffolder.

### Performance Improvement with Multiple RNA Sequencing Data Sets

In a second experiment, we assessed the impact of using multiple RNA-seq data sets on the completeness and accuracy of the resulting genome and its gene annotations. Since AGOUTI does not provide support for simultaneous analysis using multiple RNA-seq data sets, we ran the program iteratively, generating a new temporary assembly and improving it with the cumulative data set at the next step. As before, we used Quast to assess the impact on the assembled genome sequence when using 0 (raw assembly), 1, 2, 6, and 11 RNA-seq data sets. As shown in Table 2 (bottom) for Rascaf, the number of scaffolds and NGA50 values gradually improve as more data sets are added (by 5–10 and 9–19%, respectively). The number of effective misassemblies remains small (10–77), however, representing a growing fraction (0.8–5.9%) of the total assembly errors. AGOUTI was more aggressive in identifying new connections and produced slightly longer scaffolds at the expense of introducing more errors (Supplemental Table S6), whereas L_RNA_scaffolder led to shorter scaffolds and a significantly larger (seven- to eight-fold) number of errors. We attribute Rascaf’s high precision to its ability to identify and correct inconsistencies that arise from combining multiple RNA-seq data sets as implemented in rascaf-join.

Further, we assessed the impact of Rascaf’s assembly improvement procedure on the accuracy and completeness of the gene repertoire. For each of the original and intermediate assemblies, we aligned the 35,215 *A. thaliana* RefSeq mRNA transcript sequences to the scaffolds using the spliced alignment programs ESTmapper and sim4db (Istrail et al., 2004; Walenz and Florea 2011). For each set, we plotted the percentage of transcript sequences that have more than f% of their bases in the primary alignment, for varying coverage levels f (f = 5, 10, …, 100). The coverage curves are plotted in Fig. 4. Adding RNA-seq data improves the coverage and the number of transcript sequences aligned at all coverage levels, especially at the higher coverage cutoffs. In particular, adding the first RNA-seq data set brings the number of transcripts with 90% coverage from 33,529 to 33,183 in the original assembly and that number further increases to 33,847 when all 11 RNA-seq data sets are included. Most importantly, Rascaf increases the coverage for 506 to 991 genes across the 11 assemblies (Supplemental Table S7).

Therefore, using multiple RNA-seq sources improves both the assembled sequence and the gene annotations on a broader scale, albeit the benefits reach diminishing returns as more data sets recapitulating similar sets of genes are included.

### Improving the Assembly and Gene Annotations of Sequenced Fragaria

Next-generation sequencing has dramatically accelerated the pace of genome sequencing projects, with dozens of new draft genomes being reported every few months. To illustrate the benefits of using RNA-seq to improve the quality of a genome assembly and its annotations, as implemented in Rascaf, we applied the method to several

---

**Table 2. Evaluation of Arabidopsis thaliana** assemblies for single RNA sequencing (RNA-seq) data (top) and with 0, 1, …, 11 RNA-seq data sets (bottom), using Quast (Gurevich et al., 2013).**

<table>
<thead>
<tr>
<th>Programs</th>
<th>Raw assembly</th>
<th>Rascaf</th>
<th>Rascaf+BWA-mem</th>
<th>L_RNA_scaffolder</th>
<th>AGOUTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single RNA-seq set (SRR2187604)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffolds</td>
<td>8082</td>
<td>7686</td>
<td>7674</td>
<td>7759</td>
<td>7771</td>
</tr>
<tr>
<td>NGAS0†</td>
<td>42,479</td>
<td>46,331</td>
<td>46,828</td>
<td>44,441</td>
<td>45,667</td>
</tr>
<tr>
<td>Misassemblies</td>
<td>1153</td>
<td>1180</td>
<td>1188</td>
<td>1296</td>
<td>1177</td>
</tr>
<tr>
<td>Problematic scaffolds</td>
<td>1412</td>
<td>1434</td>
<td>1434</td>
<td>1536</td>
<td>1433</td>
</tr>
<tr>
<td>Effective misassemblies</td>
<td>na</td>
<td>10</td>
<td>14</td>
<td>114</td>
<td>10</td>
</tr>
</tbody>
</table>

| Multiple RNA-seq sets     |              |            |                |                 |        |
| Data sets                 | 0 (raw)      | 1          | 2              | 6               | 11     |
| Scaffolds                 | 8082         | 7686       | 7626           | 7283            | 7222   |
| NGAS0                     | 42,479       | 46,331     | 46,898         | 49,673          | 50,571 |
| Misassemblies             | 1153         | 1180       | 1190           | 1281            | 1302   |
| Problematic scaffolds     | 1412         | 1434       | 1437           | 1473            | 1478   |
| Effective misassemblies   | na           | 10         | 14             | 66              | 77     |

† NGA50 is defined as the minimum length of a scaffold alignment such that 50% of the aligned portion of the assembly to the reference genome is in scaffold alignments this size or longer.
sequenced and in-progress Rosaceae. The octoploid genome of the cultivated strawberry (*Fragaria × ananassa* Duchesne ex Rozier) has been sequenced and draft assembled, along with those of four wild diploid relatives, *F. inumae* Makino, *F. nipponica* Makino, *F. nubicola* (Hook. f.) Lindl. ex Lacaita, and *F. orientalis* Losinsk (Hirakawa et al., 2014). All genomes were sequenced with a combination of Illumina and Roche 454 technologies and assembled de novo. For our analyses, we downloaded up-to-date draft assemblies of these five *Fragaria* species from the Genome Database for Rosaceae (www.rosaceae.org) with quality indicators listed in Table 3.

We used available RNA-seq reads from *F. × ananassa* (SRA accession: ERR430941, 70 million 100-bp paired-end reads) and separately from the close relative *F. vesca* (SRA accession: SRR1930097, 76 million 101-bp paired-end reads). *Fragaria vesca* was previously sequenced and assembled with a combination of Illumina and 454 reads (Shulaev et al., 2011) and has undergone multiple rounds of curation to improve both the assembly and its gene models.

Rascaf found thousands of new putative connections for each genome, both when using the *F. × ananassa* and the cross-species *F. vesca* RNA-seq reads. We first assessed the likelihood and confidence of the connections using BLAST searches against the RefSeq mRNA database, as described above (Table 4). More than 96% of the connections found in each case could be validated, increasing to >99% likely connections when adding the uncertain and unaligned sequences, except for *F. × ananassa* improved with same-species RNA-seq reads, which had 89% positive validation rate and 98% likely connections rate.

We further assessed the impact on the completeness and accuracy of the gene annotations as an overall measure of assembly accuracy in gene regions. For this purpose, sequences of *F. vesca* RefSeq mRNA transcripts were mapped to each of the assemblies using ESTmapper and sim4db, and the portion of the bases contained in the primary alignment of each sequence was measured. The cumulative coverage plots were then computed as described in the previous section. Figure 5 shows the coverage plots for two of the species (*F. inumae* and *F. nipponica*) before and after applying the procedure, and plots for all five species are included in the Supplemental

### Table 3. Summary of quality indicators for the original (raw) assemblies of the sequenced *Fragaria* species (Hirakawa et al. 2014).

<table>
<thead>
<tr>
<th>Assembly</th>
<th>No. of scaffolds</th>
<th>Total length (bp)</th>
<th>N50†</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>F. inumae</em> (v1.0)</td>
<td>117,822</td>
<td>199,627,645</td>
<td>3309</td>
</tr>
<tr>
<td><em>F. nipponica</em> (v1.0)</td>
<td>215,024</td>
<td>206,414,979</td>
<td>1275</td>
</tr>
<tr>
<td><em>F. nubicola</em> (v1.0)</td>
<td>210,780</td>
<td>203,686,576</td>
<td>1291</td>
</tr>
<tr>
<td><em>F. orientalis</em> (v1.0)</td>
<td>323,163</td>
<td>214,184,046</td>
<td>722</td>
</tr>
<tr>
<td><em>F. × ananassa</em> (v1.0)</td>
<td>625,966</td>
<td>697,765,214</td>
<td>2201</td>
</tr>
</tbody>
</table>

† N50 is the minimum length of a scaffold such that half of the assembled genome sequence is in scaffolds of this size or longer.
Table 4. In silico evaluation of predicted Rascaf connections in the Fragaria genomes by BLAST searches against the NCBI RefSeq mRNA database.

<table>
<thead>
<tr>
<th>Species</th>
<th>Validated</th>
<th>Uncertain</th>
<th>Unaligned</th>
<th>Rearranged</th>
<th>Chimeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rascaf with ERR430941</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. iinumae</td>
<td>5267</td>
<td>119</td>
<td>19</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>F. nipponica</td>
<td>7496</td>
<td>155</td>
<td>32</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>F. nubicola</td>
<td>8447</td>
<td>198</td>
<td>59</td>
<td>26</td>
<td>59</td>
</tr>
<tr>
<td>F. orientalis</td>
<td>10,147</td>
<td>279</td>
<td>112</td>
<td>16</td>
<td>72</td>
</tr>
<tr>
<td>F. × ananassa</td>
<td>5866</td>
<td>365</td>
<td>201</td>
<td>44</td>
<td>109</td>
</tr>
<tr>
<td>Rascaf with SRR1930097</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. iinumae</td>
<td>1613</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F. nipponica</td>
<td>2710</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>F. nubicola</td>
<td>3644</td>
<td>14</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>F. orientalis</td>
<td>3880</td>
<td>25</td>
<td>9</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>F. × ananassa</td>
<td>1876</td>
<td>36</td>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Fig. S8. More specifically, using ERR430941 leads to a significant increase (>5%) in the coverage of 6684 RefSeq transcripts in F. iinumae and 8967 in F. nipponica, with 3192 to 10,302 transcripts showing gains in coverage in the remaining species (Table 5). It also increases the numbers of complete or near-complete transcripts, defined as having 90% or more bases in a single alignment, from 10,800 to 14,562 (34.8% increase) for F. iinumae and from 5997 to 8918 (48.7% increase) for F. nipponica. These results are consistent with those in Table 4 and demonstrate that Rascaf significantly improves the completeness of genes in all five of these species.

Scaffold sequences and structure organization for the five genomes after application of Rascaf can be obtained from our web site at http://ccb.jhu.edu/people/florea/research/Rascaf/

Availability and Run-Time Considerations

The Rascaf package was developed using a combination of C++ and Perl modules. More specifically, source code for the two executables rascaf and rascaf-join was written in C++. A Perl module implementing the in silico BLAST evaluation against an external National Center for Biotechnology Information database, as an optional postprocessing step, is included in the package. The user can optionally invoke the search and modify it to point to a local or external database and to the BLAST search program of choice. Run times for analyses illustrated in this manuscript were roughly 15 min when using a single RNA-seq data set and <2 h for 11 RNA-seq samples when run sequentially on a machine with 512 GB RAM and AMD Opteron 2.1 GHz processor not including the HISAT alignment step. Memory used was <1 GB RAM in all cases, as alignment files are being read and processed sequentially. Running times for AGOUTI and L_RNA_scaffolder alone were comparable; however, application of these tools in practice incurred additional run times for preassembling the RNA-seq reads into transcripts with Trinity (up to 1 d) and for generating the Cufflinks gene annotations. Therefore, Rascaf is highly memory and time efficient and can be readily used in assembly projects large and small and on a wide variety of platforms.

Rascaf is available free of charge for all distributed under a GNU General Public License and can be obtained from https://github.com/mouriisl/Rascaf.

Conclusions

Fast and affordable next-generation sequencing has brought genome sequencing to the fingertips of groups and even individual researchers. However, assembling the short reads into a finished genome represents a significant challenge, demanding highly specialized expertise and sophisticated bioinformatics methods. Most of the genomes thus produced will be in draft form, consisting largely of ordered and oriented contigs grouped in scaffolds, but also many orphan contigs for which there may not be sufficient information to allow placement into scaffolds or chromosomes. We present a software tool, called Rascaf, that takes advantage of the continuity information from paired-end Illumina RNA-seq reads to identify new connections among contigs and use them to recruit additional contigs into an assembly and to improve the organization of scaffolds. As most genome sequencing projects also deep sequence the transcriptome, often of multiple tissues or organs, as part of gene annotation efforts, these RNA-seq sequences represent an abundant and readily available resource that can be effectively used to improve the genome sequence. Several genome sequencing projects have already used gene structure information to further help orient contigs and recruit additional contigs. However, these efforts largely used locally developed methods and relied on long cDNA and mRNA sequences and more recently transcripts assembled de novo from Illumina reads, which are prone to introducing chimeric connections. These programs often rely on parameters, for instance to specify the criteria for alignments to be deemed significant, which need to be adjusted by the expert user. We present a practical and easy-to-use software tool, Rascaf, that can be universally used for any genome with no or very little need for calibration.

Rascaf improves the assembly and its gene annotations simultaneously, finding thousands of contig connections and contributing additional sequence to thousands of genes in the tested Rosaceae draft assemblies. It is also very precise, and almost every contig connection could be verified by independent evidence. Importantly, by separating the connection detection and scaffolding processes, Rascaf can incorporate multiple RNA-seq data sets, thus compensating for low-expression genes in any one sample while also eliminating likely errors that are revealed as incompatibilities. As an ancillary benefit, the split design reduces the end-to-end running time, allowing for parallelization of RNA-seq alignments across data sets as well as eliminating the overhead with building intermediate genome indices. Lastly, it provides a built-in mechanism for the user to intervene in the curation process; the user can manually inspect and edit the connections file.
filter or add connections before processing them through the rascaf-join scaffold. For convenience, scripts are provided to search and filter the connections against databases of proteins and cDNA sequences before scaffolding.

As another unique feature, Rascaf can incorporate partial alignment information, which is particularly beneficial for RNA-seq libraries with short insert sizes. Future work will assess incorporating alignments spliced across contigs, which cannot be generated by current alignment programs. Also, while Rascaf appears to work reasonably well with RNA-seq sequences from a very close relative, the performance is likely limited, since current alignment tools are not equipped to handle sequence differences that arise from evolutionary changes, including block insertion–deletion mutations and evolutionary mutation patterns.

In comparisons with the only two other programs, Rascaf had higher or comparable accuracy in all tests and the lowest overall processing times. Fast and accurate, Rascaf is a highly practical and much needed addition to the current genome assembly and assembly curation compendium of tools.

Supplemental Information Available

Supplemental information is available with the online version of this manuscript.

Acknowledgments

This work was supported in part by National Science Foundation award IOS-1339134 to LF.

References


