Mol Genet Genomics (2014) 289:427–438
DOI 10.1007/s00438-014-0819-3

ORIGINAL PAPER

Towards an improved apple reference transcriptome using RNA-seq

Yang Bai · Laura Dougherty · Kenong Xu

Received: 16 October 2013 / Accepted: 25 January 2014 / Published online: 16 February 2014

Abstract The reference genome of apple (*Malus* × *domestica*) has been available since 2010. Despite being a milestone in apple genomics, the reference genome is difficult to be used as a reference in RNA-seq (RNA sequencing) analysis, a widespread technology in transcriptomic studies. One of the major limitations appears to be the low coverage of the reference transcriptome in RNA-seq mapping of reads. To improve the reference transcriptome, we obtained 14 sets of strand-specific RNA-seq data of 168.5 million reads in total from fruit of Golden Delicious (GD, the source of the reference genome) in varying growth and developmental stages. Using a combination of genome-guided assembly and de novo assembly, the apple reference transcriptome was improved to a collection of 71,178 genes or transcripts, which includes 53,654 genes predicted originally (with MDP prefixed in their IDs) and 17,524 novel transcripts. Of these novel transcripts, 8,144 were identified from reads directly mapped to the reference genome while the remaining 9,380 were extracted from de novo assemblies of reads that could not be initially mapped to the reference genome. Evaluating the improved apple reference transcriptome with reads from Golden Delicious and other genotypes used in this and other studies showed that it allowed 62.5 ± 9.3–82.3 ± 2.7 % of reads to be mapped, a marked increase from the low rates of 37.4 ± 7.7–46.6 ± 7.1 % offered by the original reference transcriptome. The improved reference transcriptome therefore represents a step forward towards a complete reference transcriptome in apple.

Keywords *Malus* × *domestica* · Transcriptome coverage · RNA sequencing · Transcript discovery

Introduction

The development of RNA sequencing (RNA-seq) technology (Mortazavi et al. 2008; Wilhelm and Landry 2009) has been a breakthrough in the characterization of complex eukaryotic transcriptomes. Compared with microarray-based global gene expression assays, which primarily employ molecular hybridization between a sample of unknown transcripts and an arrayed transcriptome, RNA-seq directly sequences the mRNA derived cDNA using a next generation sequencing (NGS) platform, such as Illumina HiSeq 2000/2500. The massive throughput of NGS machines allows RNA-seq to provide unprecedented resolution and depth of data, enabling simultaneous quantification of gene expression, discovery of novel transcripts and exons, detection of SNP and measurement of splicing variants (Chepelev et al. 2009; Wilhelm et al. 2010). Due to these advantages, RNA-seq has quickly become the choice in diverse transcriptomic studies attempting to investigate complex biological process in human, animal and plants as well as in microbes.

Early RNA-seq based transcriptomic studies in plants included *Arabidopsis* (Lister et al. 2008), grape (Zenoni et al. 2010), maize (Li et al. 2010) and rice (Zhang et al.

Communicated by S. Hohmann.

Y. Bai and L. Dougherty contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-014-0819-3) contains supplementary material, which is available to authorized users.

Y. Bai · L. Dougherty · K. Xu (✉)
Department of Horticulture, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456, USA
e-mail: kx27@cornell.edu

2010). It has now been used in plant species not only with a reference genome, but also without a reference genome (Ong et al. 2012; Ruttink et al. 2013). An essential step in RNA-seq data analysis is to map the short reads back to the reference genome or reference transcriptome so that the reads associated with a specific gene could be counted and then used to compare with other genes for differentiating their expression levels. When a reference genome is not available, an alternative reference transcriptome can be assembled by de novo assembly of RNA-seq reads directly (Ong et al. 2012; Ruttink et al. 2013). This alternative approach allows a much broader range of RNA-seq applications. However, chromosomal locality information and local and global genomic contexts would not be available in this alternative approach, and alternative splicing is also less readily detectable due to the nature of matured mRNA, which is the source of RNA-seq reads in most cases. In addition to the utility in identifying genes of splicing variants, nucleotide polymorphisms, and expression levels that are co-elevated or suppressed in certain pathways, RNA-seq has also become a tool of discovery for revealing novel dimensions hidden in plant transcriptomes, such as RNA editing in chloroplast and mitochondria (Sun et al. 2013; Suzuki et al. 2013) and long non-coding RNAs that function as endogenous microRNA (miRNA) target mimics preventing miRNAs from reaching their target genes (Wu et al. 2013).

Apple (*Malus* × *domestica*, $2n = 2x = 34$ usually) is one of the most important fruit crops in the world. Its genome sequences of 742.3 Mb (Velasco et al. 2010) are available from the Genome Databases for Rosaceae (GDR, http://www.rosaceae.org) and other sites. There are 63,541 predicted genes (or MDPs due to prefix MDP in gene IDs, e.g. MDP0000252114) in the consensus gene set in the genome. The source of the reference genome is Golden Delicious (GD), an apple variety grown widely throughout the major production areas in the US and abroad. RNA-seq based transcriptomics studies have also been reported in apple (Krost et al. 2012, 2013; Zhang et al. 2012; Gapper et al. 2013; Gusberti et al. 2013). However, using the predicted genes as a reference transcriptome has led to inconsistent RNA-seq reads mapping rates from 35.8 ± 3.7 % (unique reads, Gusberti et al. 2013) to 65 % (Gapper et al. 2013), leaving more than one-third of reads uncounted even if the non-specific reads were counted in Gusberti et al. (2013). In other RNA-seq based studies, the reference transcriptome was not used (Krost et al. 2012, 2013; Zhang et al. 2012). Although the gene prediction in the apple genome might not be perfect, the 63,541 predicted genes that represent the current version of apple reference transcriptome are invaluable resource and have been used in many studies since the genome sequences became available in 2010. Clearly, there is a need for improving the reference transcriptome by building on it in the apple research community. To address this need, we obtained 14 sets of strand-specific RNA-seq data from GD fruit in varying growth and developmental stages that were used in one of our previous studies (Wang and Xu 2012). Using a combination of genome-guided assembly and de novo assembly, the apple reference transcriptome was improved to a collection of 71,178 genes or transcripts, including 53,654 MDPs and 17,524 novel transcripts. Testing of RNA-seq mapping with reads from this and other studies indicated that the improved apple reference transcriptome increased the reads mapping rates to 62.5 ± 9.3–82.3 ± 2.7 % using high stringent mapping parameters, a considerable lift from the rates of 37.4 ± 7.7–46.6 ± 7.1 % offered by the original reference transcriptome.

Materials and methods

Plant materials and RNA isolation

Fruit of Golden Delicious (GD) was sampled from 14 time points from 1 week after full-bloom (WAF) through 20 WAF (at harvest) in 2010 as described previously (Wang and Xu 2012). The fruit samples were flash frozen in liquid nitrogen and stored at −80 °C before being used. For each sample, total RNA was isolated from 2 (young fruit)–3 g (mature or near mature fruit) of ground tissues pooled from at least five fruits according to Gasic et al. (2004) with modifications: Before tissue tearer homogenization, 1 ml Sarkosyl of 20 % (w/v) was added to 10 ml of the extraction buffer. The extracted total RNA was dissolved in EB buffer (Qiagen, Germantown, MD) supplemented by 1 × Ambion RNAsecure (Invitrogen/Life Technologies, Carlsbad, CA). To activate RNAsecure, the samples were incubated at 60 °C (in a water bath) for 10 min and then immediately put on ice. RNA quantity and quality were evaluated by Nanodrop 1,000 (Thermo Scientific, Waltham, MA) and Bioanalyzer 2100 with RNA 6000 Nano Chip (Agilent, Santa Clara, CA) as well as a 2 % agarose gel (using 1/10 RNA dilutions in EB buffer with 1 × Ambion RNA secure). Immediately prior to mRNA isolation, the RNA samples were treated with DNase I (amplification grade, Invitrogen) at 37 °C for 30 min followed by heat inactivation at 65 °C for 15 min.

Stand-specific RNA-seq library construction and sequencing

For each sample, 5 µg total RNA was used to isolate mRNA to prepare a stand-specific RNA-seq library using NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) following the manufacturer's protocols with minor modifications. Briefly, mRNA was extracted with 15 µl of NEBNext Magnetic Oligo d(T)$_{25}$ and fragmented in NEBNext First Strand Synthesis Buffer by heating at 94 °C for 10 min. First strand cDNA was reverse transcribed form the fragmented mRNA and then used as template to synthesize double stranded cDNA with dUTP replacing dTTP. The resulting double-stand cDNA was end-repaired, dA-tailed and then ligated with NEBNext Adaptor. To remove unwanted large fragments, the adaptor-ligated cDNA was selected for size using Agencourt AMPure XP beads (Bechman Coulter, Pasadena, CA) in 0.6 volumes of the ligation reaction. For optimizing the size selection, another round of size selection was performed as described in Zhong et al. (2011), where the beads in 1.4 volumes of the cDNA solution were used. Next, the selected cDNA was digested with NEBNext USER enzyme and then enriched by PCR in the following conditions: 98 °C for 30 s; 14 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s; 72 °C for 5 min; and then held at 4 °C. The PCR-enriched cDNA libraries were purified by 1.4 volumes of Agencourt AMPure XP beads and eluted in 20 µl low TE buffer (10 mM Tris–HCl, pH 8.0, and 0.1 mM EDTA). To estimate whether or not the libraries were within the expected size range from 250 bp to 400 bp, 2 µl of the purified PCR products were analyzed by 2 % agarose gel electrophoresis and then visualized with ethidium bromide-staining. If the primer-dimer band (~80 bp) appeared, the libraries were purified again with 1.4 volumes of the beads. The purified libraries were then quantified by Qubit 2.0 Fluorometer using the dsDNA HS Assay Kit (Invitrogen/Life Technologies, Carlsbad, CA). The 14 multiplexed libraries with 60 ng each were pooled together for single-end sequencing of 101 bases without replication in one lane of Illumina HiSeq 2000 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY).

Reads processing and data analysis

The 14 sequence files of 180.8 million raw reads in total were generated by the Illumina pipeline in software CASAVA v1.8 in Sanger FASTQ format (available under NCBI SRA experiment number SRX392051). Only were the high quality reads (168.5 million) that passed the chastity filter (i.e. no more than one base call in the first 25 cycles has a chastity higher than 0.6) in the pipeline used, which accounted for 93.2 ± 1.3 % of the total raw reads of 180.8 million (Table S1). Data analyses were performed using CLC Genomics Workbench (CLC GW) v6.5 (CLCBio, Cambridge, Massachusetts). Three files of the apple reference genome (Velasco et al. 2010) *M. domestica* v1.0 (Md-v1.0 hereafter) were downloaded from the Genome Databases for Rosaseae (GDR, http://www.rosaceae.org). The first is the genome sequence file consisting of 122,107 contigs (MDCs hereafter); the second is the coding sequence (CDS) file for the consensus gene set containing 63,541 predicted genes (MDPs hereafter); and the third is the genome annotation file for the 63,541 genes in GFF format. The genome sequence file of 122,107 MDCs and the GFF file were combined by CLC GW to reconstruct the annotated apple reference genome (Md-v1.0) locally. To enable local BLAST search of the genome sequences, the sequences of 122,107 MDCs were converted into a BLAST database using CLC GW.

For RNA-seq mapping against reference genome Md-v1.0, we used only the gene regions defined by the MDPs (i.e. without including any bases up- or down-stream of MDPs). For convenience, the set of 63,541 MDPs was collectively called Md-v1.0-RT (apple reference transcriptome v1.0) in this study. The limit for read unspecific match to Md-v1.0-RT was set to 10. To map a read, the minimum length fraction is 0.8 and the minimum similarity is 0.98 (our empirical sequence identity threshold often effective in differentiating paralog sequences in the apple genome). These two high stringent parameters were also used in large gap read mapping of sequences against Md-v1.0. In de novo assembly, the word sizes 22 (for Step 1, Fig. 1) and 23 (for Steps 2 and 8, Fig. 1) and bubble size of 50 were chosen automatically by the CLC de novo assembler.

For transcript discovery, parameters were mostly set by default, but with the following changes: (1) the minimum length of ORF calling was raised from 100 bp to 150 bp; (2) for "gene" discovery, the maximum distance between events was set to 10 bp and the minimum length of "gene" was 150 bp.

Revision of the reference genome transcriptome (Md-v1.0-RT)

An approach of three rounds of de novo assembly, large gap read mapping and/or transcript discovery was taken for improving Md-v1.0-RT (Fig. 1). Briefly, Steps 1–5 in Round 1 were intended to reveal new transcripts directly from reference genome Md-v1.0 so that a revision of Md-v1.0-RT could be made straightforwardly. We began with a complexity reduction step by de novo assembling of the RNA-seq reads into contigs in each of the 14 samples, generating a set of contigs of 1,241,606. The unassembled reads were collected and pooled and were de novo assembled again, yielding another set of contigs of 193,860. The reads (12,472,992) that still remained unassembled were re-collected, and used along with the two sets of contig sequences (1,435,466 in sum) as input for large gap read mapping against reference genome Md-v1.0 using the CLC Large Gap Read Mapping tool. Revision of the reference transcriptome was conducted using the CLC Transcript Discovery tool, leading to the first revision of Md-v1.0-RT

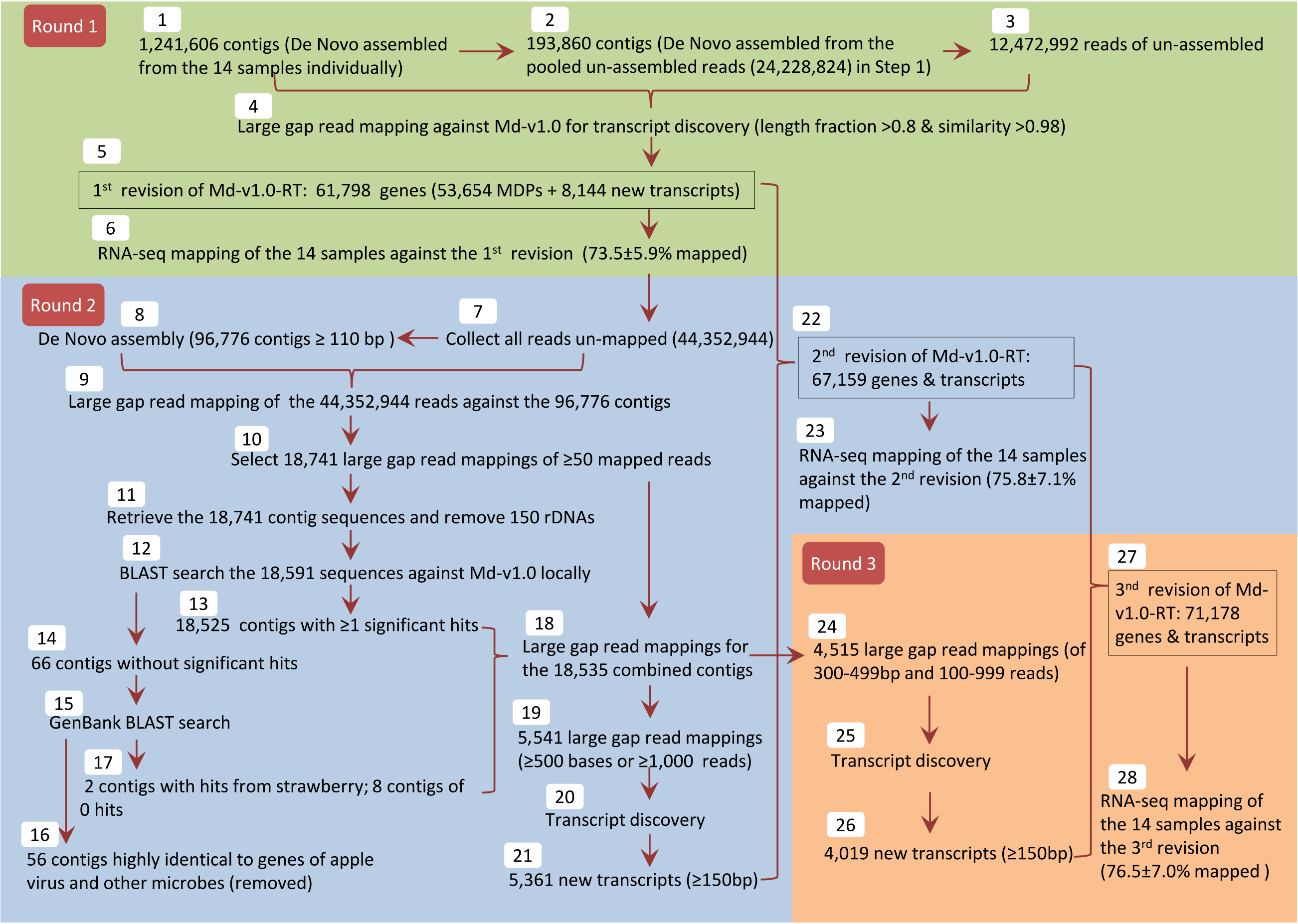
Round 1
1
1,241,606 contigs (De Novo assembled from the 14 samples individually)
2
193,860 contigs (De Novo assembled from the pooled un-assembled reads (24,228,824) in Step 1)
3
12,472,992 reads of un-assembled
4
Large gap read mapping against Md-v1.0 for transcript discovery (length fraction >0.8 & similarity >0.98)
5
1st revision of Md-v1.0-RT: 61,798 genes (53,654 MDPs + 8,144 new transcripts)
6
RNA-seq mapping of the 14 samples against the 1st revision (73.5±5.9% mapped)
Round 2
8
De Novo assembly (96,776 contigs ≥ 110 bp)
7
Collect all reads un-mapped (44,352,944)
9
Large gap read mapping of the 44,352,944 reads against the 96,776 contigs
10
Select 18,741 large gap read mappings of ≥50 mapped reads
11
Retrieve the 18,741 contig sequences and remove 150 rDNAs
12
BLAST search the 18,591 sequences against Md-v1.0 locally
13
18,525 contigs with ≥1 significant hits
14
66 contigs without significant hits
15
GenBank BLAST search
17
2 contigs with hits from strawberry; 8 contigs of 0 hits
16
56 contigs highly identical to genes of apple virus and other microbes (removed)
18
Large gap read mappings for the 18,535 combined contigs
19
5,541 large gap read mappings (≥500 bases or ≥1,000 reads)
20
Transcript discovery
21
5,361 new transcripts (≥150bp)
22
2nd revision of Md-v1.0-RT: 67,159 genes & transcripts
23
RNA-seq mapping of the 14 samples against the 2nd revision (75.8±7.1% mapped)
Round 3
24
4,515 large gap read mappings (of 300-499bp and 100-999 reads)
25
Transcript discovery
26
4,019 new transcripts (≥150bp)
27
3rd revision of Md-v1.0-RT: 71,178 genes & transcripts
28
RNA-seq mapping of the 14 samples against the 3rd revision (76.5±7.0% mapped)

Fig. 1 Flow chart of sequence analyses conducted to improve the apple reference transcriptome. Steps are indicated by the numbers in white boxes. Round 1 includes Steps 1–6, Round 2 Steps 7–23, and Round 3 Steps 24–28. Md-1.0: apple reference genome *Malus* × *domestica* v1.0. Md-1.0-RT: apple reference transcriptome associated with Md-1.0. MDPs: genes predicted in Md-1.0 and Md-1.0-RT

(Step 5). RNA-seq mapping evaluation of this initial revision of Md-v1.0-RT was completed using the reads from the 14 RNA-seq samples (Step 6). Steps 7 through 20 in Round 2 were attempted to discover new transcripts from the reads that could not be mapped in Step 6 so that more novel transcripts could be identified to achieve the second revision of Md-v1.0-RT. RNA-seq mapping evaluation of the second revision of Md-v1.0-RT was performed in Step 23 where the reads from the 14 samples were used again. The third round was largely a supplementary step to Round 2 to identify new transcripts that were not selected early. Augmenting the second revision of MD-1.0-RT with the new transcripts in Round 3 resulted in the improved reference transcriptome Md-v1.0-RT, i.e. the third revision of Md-v1.0-RT.

MapMan gene ontology of new transcripts

The sequences of the new transcripts were retrieved to BLAST search multiple databases using the web-based search tool Mercator (http://mapman.gabipd.org/web/guest/mercator). The databases searched include: TAIR-Arabidopsis TAIR proteins (release 10), PPAP-SwissProt/UniProt Plant Proteins, CHLAMY-JGI Chlamy release 4 Augustus models, ORYZA-TIGR5 rice proteins, KOG: Clusters of orthologous eucaryotic gene database (KOG), CDD—conserved domain database, and IPR—interpro scan. The output file of Mercator not only contains the best hits in databases, but also assigns MapMan's gene ontology (Thimm et al. 2004) for each input sequence if possible.

Chromosomal locality and apple genome origin of new transcripts

Chromosomal locality or the apple genome origin of the new transcripts was either deduced from their associated home MDCs if they were identified in Round 1 (Fig. 1), or evidenced from significant ($E < 10^{-10}$) hits in BLAST

Table 1 RNA-seq mapping of reads against the current version of apple reference transcriptome (Md-v1.0-RT) and its revisions

Reference transcriptome	Read mapping	Overall no. of reads	Mean no. of reads per sample	SD of mean no. of reads- per sample	% of total reads	% of total reads-SD
MD-v1.0-RT	Mapped	72,398,775	5,171,341	792,594	42.8	4.5
	Uniquely	55,134,893	3,938,207	596,381	32.6	3.3
	Non-specifically	17,263,882	1,233,134	198,280	10.2	1.2
	Unmapped	96,128,386	6,866,313	623,257	57.2	4.5
	Total	168,527,161	12,037,654	986,658	100.0	0.0
1st revision	Mapped	124,174,217	8,869,587	1,175,967	73.5	5.9
	Uniquely	100,207,386	7,157,670	1,060,606	59.3	6.1
	Non-specifically	23,966,831	1,711,917	229,393	14.2	1.2
	Unmapped	44,352,944	3,168,067	652,320	26.5	5.9
	Total	168,527,161	12,037,654	986,658	100.0	0.0
2nd revision	Mapped	127,996,932	9,142,638	1,298,164	75.8	7.1
	Uniquely	105,370,507	7,526,465	1,068,633	62.4	5.9
	Non-specifically	22,626,425	1,616,173	234,296	13.4	1.3
	Unmapped	40,530,229	2,895,016	797,807	24.2	7.1
	Total	168,527,161	12,037,654	986,658	100.0	0.0
3rd revision	Mapped	129,277,684	9,234,120	1,298,218	76.5	7.0
	Uniquely	105,409,556	7,529,254	1,054,598	62.4	5.7
	Non-specifically	23,868,128	1,704,866	248,086	14.1	1.4
	Unmapped	39,249,477	2,803,534	783,221	23.5	7.0
	Total	168,527,161	12,037,654	986,658	100.0	0.0

searches against reference genome Md-v1.0 or GenBank that were conducted in Rounds 2 or 3 (Fig. 1).

Evaluation of the improved reference genome with RNA-seq data from other sources

Two sets of published RNA-seq paired-end data (Table S2) generated from the Illumina platform were downloaded from NCBI Sequence Read Archive (SRA). The first set contained two samples ERR033805 and ERR033806 (Krost et al. 2012, 2013). Both samples were collected from meristem tissues of apple breeding selections, but differed in their growth habit: ERR033805 was obtained from a standard selection while ERR033806 a columnar selection. The second set was a six-sample (ERR313216, ERR313217, ERR313224, ERR313225, ERR313226 and ERR313239) subset representing the 24 samples used in investigating apple scab ontogenic resistance (Gusberti et al. 2013). These samples were derived from mRNA of Golden Delicious leaves challenged by pathogen *Venturia inaequalis* or mock-inoculated by water. The two datasets were trimmed by three parameters (a quality limit of 0.05-a probability of error for a base called, an ambiguous limit of 2 and a minimum number of nucleotides of 15) to remove low quality reads or low quality bases in the reads prior to RNA-seq mapping (Table S2).

Results

Mapping of RNA-seq reads to the apple reference transcriptome (Md-v1.0-RT)

The current version of apple reference genome (Md-v1.0) was released in 2010 (Velasco et al. 2010). It comprises 122,107 MDCs (genomic contigs) that were annotated with 63,541 MDPs (predicted genes). To evaluate the coverage of Md-v1.0-RT, the 14 sets of RNA-seq data from fruit of Golden Delicious (GD) were mapped against the reference transcriptome. The mean input was 12,037,654 ± 986,658 reads per sample, with 168,527,161 reads in total (Table 1). The mean reads mapping rate was 42.8 ± 4.5 % (34.4–49.6 %), including 32.6 ± 3.4 % mapped uniquely and 10.2 ± 1.2 % non-specifically. Of the mapped reads, 10.9 ± 1.4 % was mapped to the introns (data not shown). These data suggested that a majority (57.2 ± 4.5 %) of reads were not counted in the RNA-seq read mapping process, and that alternative splicing variants were likely common within the MDPs as nearly 11 % of mapped reads were mapped to the predicted intron regions.

Improvement of the apple reference transcriptome (Md-v1.0-RT)

To improve reference transcriptome Md-v1.0-RT, the 14 sets of RNA-seq data from GD fruit were used to uncover novel transcripts. An approach of three rounds of de novo assembling, large gap read mapping, transcript discovery and/or RNA-seq mapping evaluation was carried out for this purpose (Fig. 1; Table 1).

In Round 1, the two sets of de novo assembled contigs (1,435,466) and the unassembled reads (12,472,992) from the 14 samples were mapped to 49,553 of the 122,107 MDCs in Md-v1.0 using the CLC Large Gap Read Mapping tool. The remaining 72,554 MDCs received zero reads. The CLC Transcript Discovery tool was used for the first revision of reference transcriptome Md-v1.0-RT that contained 61,798 genes or transcripts, including 53,654 MDPs and 8,144 new transcripts (Step 5, Fig. 1) which collectively covered 176.6 Mb. The 53,654 MDPs accounted for 160.7 Mb (2,995 ± 2,784 bp/gene) of the transcriptome and comprised two gene sets. The first set of 53,579 genes corresponded to MDPs defined in the original Md-v1.0-RT, but with many enhanced by alternative splicing variants and/or new 5′ or 3′ sequences (Fig. 2a, c). The second set of 74 genes that represented 154 single MDPs in the original Md-v1.0-RT as each of the 74 were combined from two or more MDPs due to the presence of bridging reads between them (Fig. 2b, e) on the same strand. The 8,144 new transcripts had an accumulated exon length of 6.5 Mb (796 ± 536 bp/transcript) and represented transcribed regions that were not included in Md-v1.0-RT (Fig. 2a, d). Alternative splicing variants were detected in the new transcripts as well (Fig. 2d). There were 9,887 MDPs in the original Md-v1.0-RT that were excluded from the first revision because the genomic contigs (MDCs) harboring these MDPs were among the 72,554 MDCs that received zero reads in the large gap read mapping. Repeating the mapping of the original RNA-seq data for the 14 samples using the first revision of Md-v1.0-RT resulted in the mapping 73.5 ± 5.9 % of reads (Fig. 1; Table 1), an increase of 30.7 percentiles from the coverage of 42.8 ± 4.5 % when using the original Md-v1.0-RT (Table 1).

The aim of the second round of revision was to uncover more novel transcripts. To do so, the reads that could not be mapped to the first revision of Md-v1.0-RT in RNA-seq mapping (Step 6, Fig. 1) were collected and de novo assembled into 96,776 contigs of minimal 110 bases (Steps 7–8, Fig. 1). When the reads were mapped back to these contigs using the CLC Large Gap Read Mapping tool, 18,741 mappings of at least 50 reads were selected (Steps 9–10, Fig. 1). By retrieving their corresponding contig sequences and then BLAST search against local databases for rDNAs and the apple genome (Md-v1.0), 150 contigs highly similar to rDNAs were removed and 18,525 contigs of at least one significant hit ($E < 10^{-10}$) in Md-v1.0 were identified (Steps 11–13, Fig. 1). The 66 contigs without a significant hit in Mdv1.0 were BLAST searched against GenBank, leading to removing 56 of them similar to apple virus and other microbes' sequences (Steps 14–16, Fig. 1). The remaining ten, including two with hits from strawberry and eight with zero hits, were pooled with the 18,525 contigs of hits in Md-v1.0, which allowed retrieving 18,535 corresponding large gap read mappings (Steps 13, 15, 17 and 18, Fig. 1). Applying the CLC Transcript Discovery tool to 5,541 of the 18,535 large gap read mappings of at least 500 bases in length or at least 1,000 mapped reads revealed 5,361 new transcripts (Steps 19–21, Figs. 1, 3a, b). Combining these new transcripts with the first revision of Md-v1.0-RT led to the second revision of Md-v1.0-RT (Step 22, Fig. 1). The 5,361 new transcripts accounted for transcribed regions of 3,481 kb (649 ± 330 bp/transcript) of the transcriptome and included many splice variants (Fig. 3a, b). RNA-seq mapping of the 14 samples against the second revision of Md-v1.0-RT resulted in a read mapping rate of 75.8 ± 7.1 %, a 2.3-percentile increase over that obtained using the first revision.

The intent of the third round of revision was to identify transcripts that could account for the reads represented by the 4,515 large gap read mappings that were not selected in the second round, but had 100–999 mapped reads with contig length of 300–499 bp (Step 24, Fig. 1). Using again the CLC Transcript Discovery tool, we obtained 4,019 new transcripts (Steps 25–26, Figs. 1, 3c, d) that are equivalent to an accumulated transcribed region of 1,503 kb (374 ± 58 bp/transcript). Alternative splicing variants were detectable in this set of transcripts as well (Fig. 3d). Supplementing the second revision of the reference transcriptome with these 4,019 new transcripts (Step 27, Fig. 1) denoted the third revision of Md-v1.0-RT. Mapping of RNA-seq reads from the 14 samples to the third revision of Md-v1.0-RT resulted in a mean read mapping rate of 76.5 ± 7.0 % (Table 1; Step 28, Fig. 1). Overall, this final revision of reference transcriptome Md-v1.0-RT, which is available at the Genome Database for Rosaceae (http://www.rosaceae.org/), contained 71,178 genes or transcripts covering 172.2 Mb, of which 53,654 are MDPs and 17,524 novel transcripts.

Evaluation of the revised apple reference transcriptome with RNA-seq data from other sources

To evaluate the revised reference transcriptome, two sets of RNA-seq data that were reported in previous studies (Krost

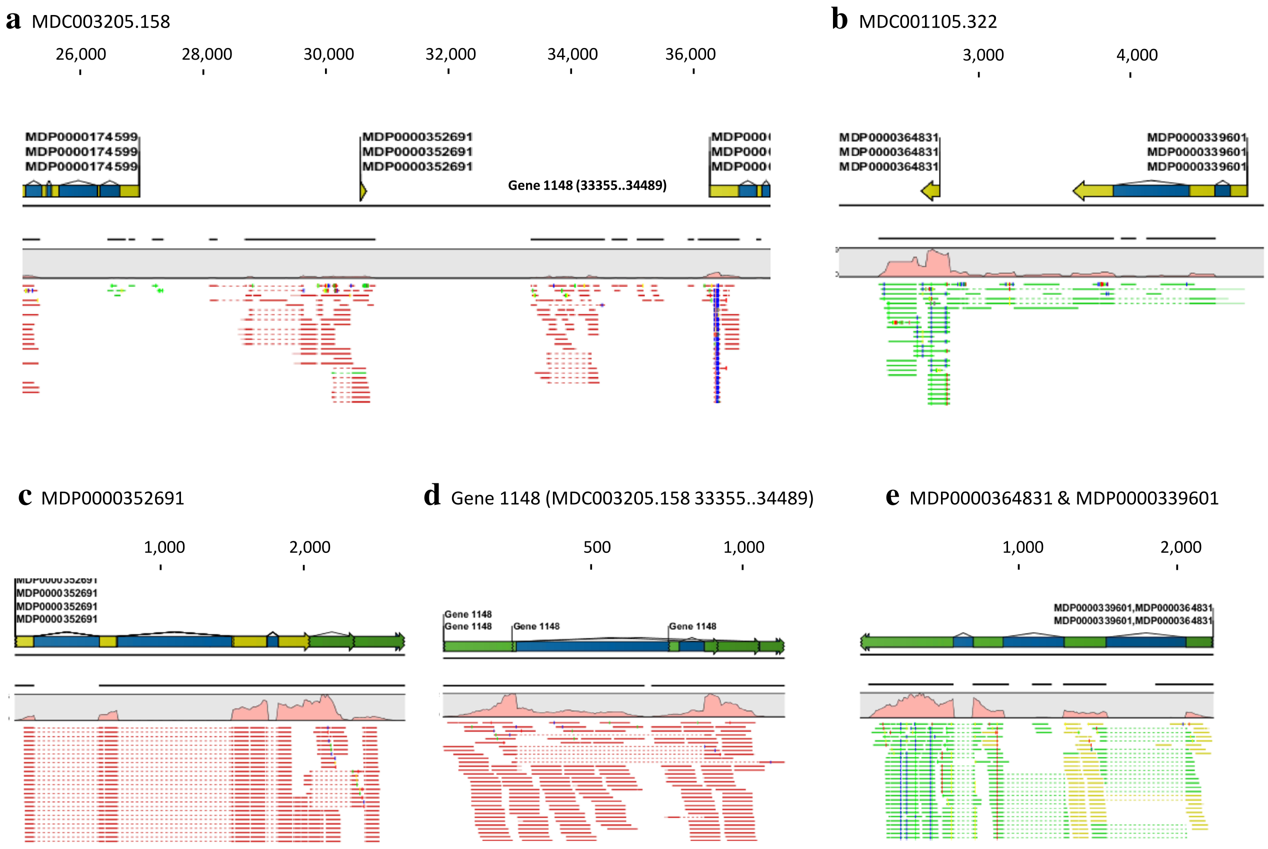

Fig. 2 Large gap read mapping and transcript discovery and verification. The length (bp) of MDCs and MDPs or transcripts is shown by the numbers on *top*. The *blue*, *green* and *yellow colors* in the graphic annotation of genes represent gene, mRNA and CDS (coding sequences), respectively. The mapped strand-specific reads are shown in *red line* (for one strand) or *green line* (for the other stand). Reads that are not specifically mapped are indicated with *yellow lines*. The *dotted* region in reads corresponds to an intron. **a**. Large gap read mapping of reads and read contigs onto MD C003205.158 (shown a section). Note the reads and read contigs under MDP0000352691 and Gene 1148. **b**. Large gap read mapping of reads and read contigs onto MDC001105.322 (shown a section). Note the reads and read contigs that bridge genes between MDP0000364831 and MDP0000339601 on the same strand. **c–e** RNA-seq mappings of reads for genes MDP0000352691 (**c** note the new sequences expanded beyond the original coverage of MDP0000352691 in **a**), gene 1148 (**d** this proves gene 1148 to be a novel transcript between MDP0000352691 and MDP0000277388 in MDC003205.158-see **a**), and MDP0000364831 and MDP0000339601 that were merged into one gene (**e**) (color figure online)

et al. 2012, 2013; Gusberti et al. 2013) were used (Table S2). After trimming/removing low quality bases and reads, the first dataset (ERR033805 and ERR033806) in Krost et al. (2012, 2013) contained 80.9 ± 3.3 million reads per sample and the second dataset (ERR313216, ERR313217, ERR313224, ERR313225, ERR313226 and ERR313239) in Gusberti et al. (2013) contained 73.6 ± 10.2 million reads per sample. RNA-seq mapping against the revised Md-v1.0-RT showed that reads of 82.3 ± 2.7 % in the first dataset and 62.6 ± 7.7 % in second dataset were mapped. In comparison, only 46.6 ± 7.1 % of reads in the first dataset and 37.4 ± 7.7 % in the second dataset were mapped using the original Md-v1.0-RT (Table S3). These results indicated that when reads from other tissues or genotypes were mapped with the revised reference, much higher rates of coverage were obtained.

MapMan gene ontology of the new transcripts

Over the process of three rounds of transcript discovery, a total of 17,524 new transcripts were identified, including 8,144 identified from reads directly mapped to reference genome Md-v1.0, and 9,380 identified from the unmapped reads. To understand their putative functions as well as their possible MapMan's gene ontology (Thimm et al. 2004), multiple databases were searched with the web-based search tool Mercator (http://mapman.gabipd.org/web/guest/mercator) using the 17,524 new transcripts as BLAST queries. Of these, 7,724 (44.1 %) had significant returns ($E < 10^{-10}$), but only 6,978 (39.8 %) were assigned with a MapMan ontology in one of the 34 bins while 746 (4.3 %) were placed to 'not assigned and no ontology' (Fig. 4). The remaining 9,800 (55.9 %) were returned without significant

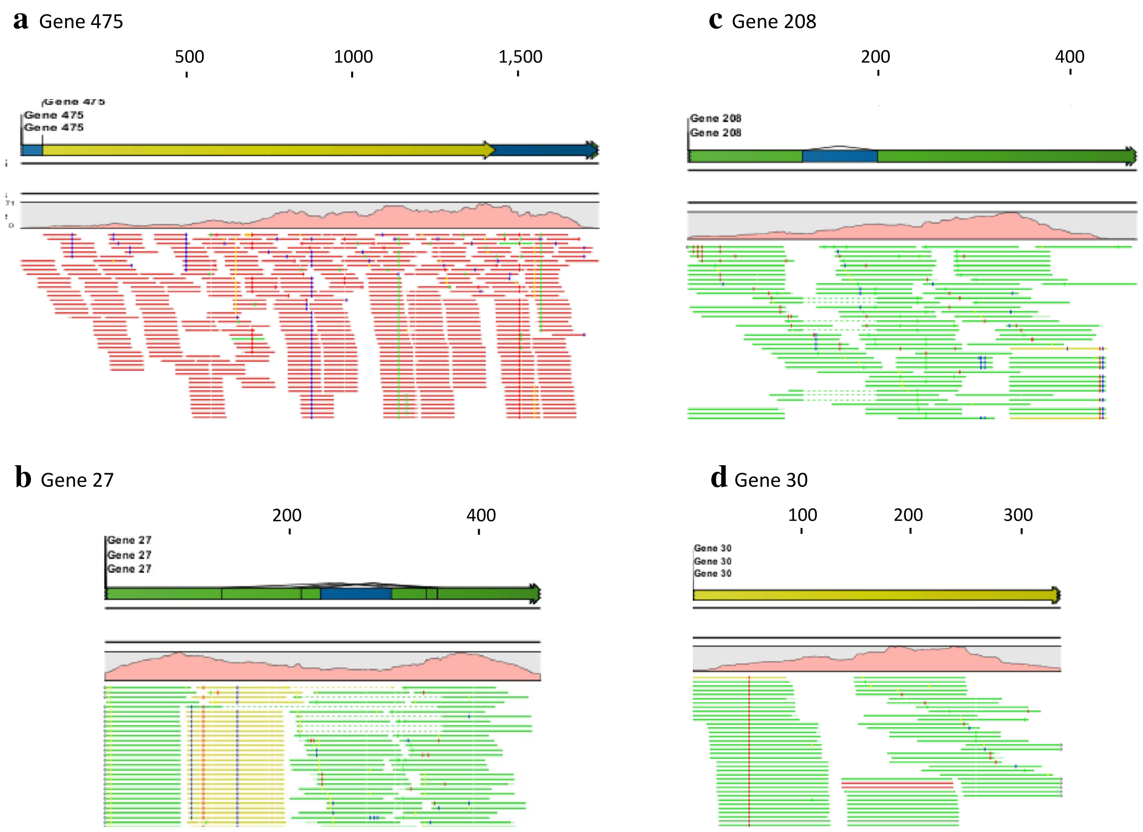

Fig. 3 Novel transcripts identified from reads unmapped initially. The transcripts are annotated or presented with the same elements and color schemes as described in Fig. 2. **a**, **b** Examples of new transcripts revealed in Round 2 (Fig. 1). **c**, **d** Examples of new transcripts uncovered in Round 3 (Fig. 1) (color figure online)

hits and assigned to category 'not assigned and unknown' (9,664 or 55.2 %) or failed to be processed (136 or 0.8 %). Among the assigned bins, 'Protein' (1,581 or 9.0 %) and 'RNA' (1,002 or 5.7 %) were the most abundant while bins 'gluconeogenesis/glyoxylate cycle' (3 or 0.02 %) and 'micro RNA, natural antisense etc' (1 or 0.01 %) the least (Fig. 4).

Chromosomal locality or apple genome origin of the new transcripts

Chromosomal locality of the 8,144 transcripts identified in the first round of revision was inferred straightforwardly from their harboring MDCs in the apple genome Md-v1.0. It showed that 7,582 (93.1 %) transcripts were located in MDCs anchored to one of the 17 chromosomes while 562 (6.9 %) were found in unanchored MDCs (Fig. 5). The apple genome origin of the 9,380 new transcripts that were identified from the unmapped reads in the second and third rounds of revisions (Fig. 1) was evaluated by BLAST searches against Md-v1.0 and GenBank. The BLAST searches found that 9,375 of them had one or more significant hits ($E < 10^{-10}$) in the apple genome Md-v1.0 (Table 2). Of the 9,375 of significant hits in the apple genome, 7,594 (81.0 %) had the highest sequence identity greater than 98.0 %, and the rest 1,781 (19.0 %) ranged from 70.0 through 98.0 %. Among the remaining five transcripts, two had significant hits of strawberry sequences in GenBank and three did not show any significant similarities with any sequences in GenBank (Table 2). These data suggested that 9,377 of the 9,380 new transcripts are of the apple genome origin. The three without any significant hits in GenBank are also likely of the apple genome origin (see "Discussions").

Discussions

Identification of novel transcripts

The massive throughput of RNA-seq has been effective in identifying novel transcripts in annotated genomes (Trapnell et al. 2010; Zhang et al. 2010; Roberts et al. 2011). Through direct mapping of RNA-seq reads to the apple reference genome Md-v1.0, we identified 8,144 novel transcripts. These transcripts represent a fraction of the apple

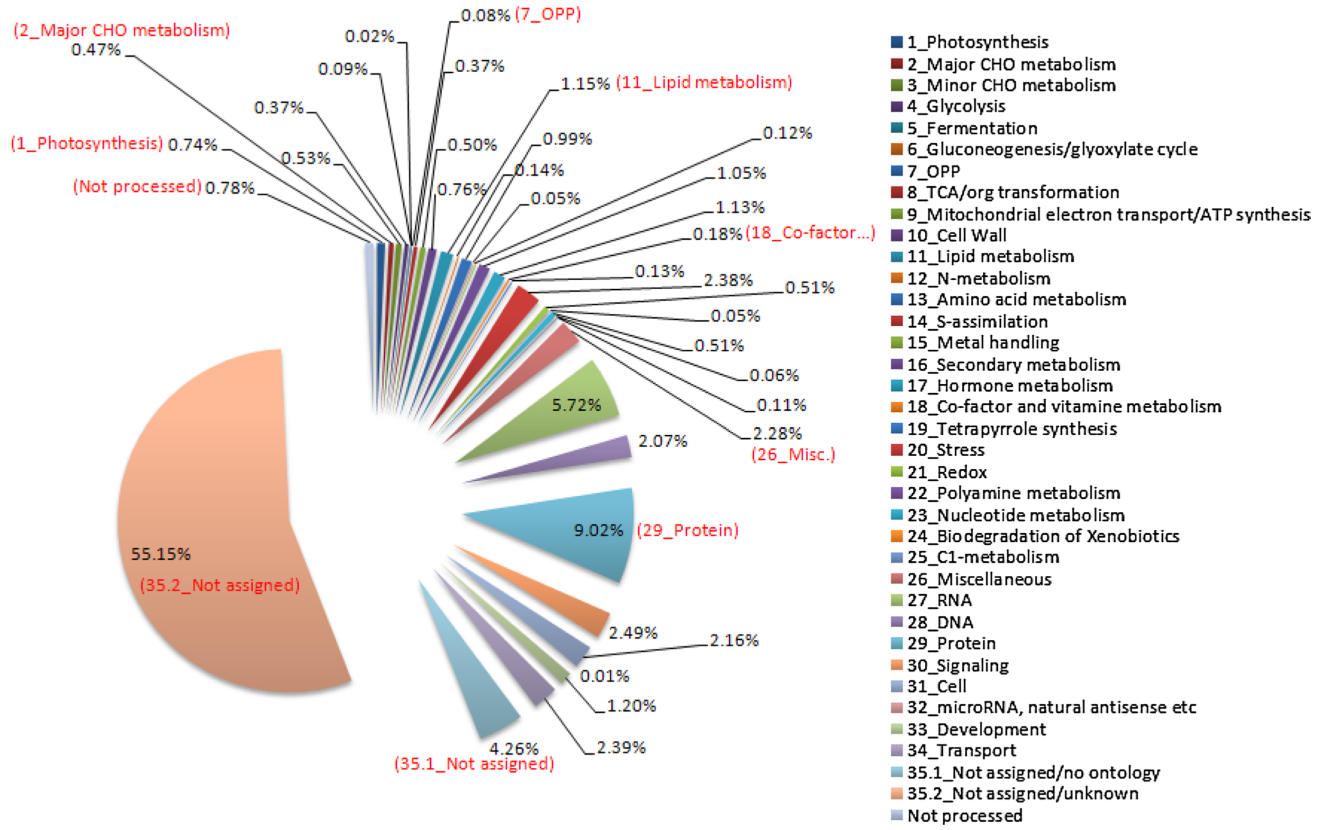

Fig. 4 Distribution of the 17,524 novel transcripts in MapMan bins. The percentage was calculated from the number of transcripts in a given bin over the total transcripts of 17,524. Each bin is shown by a piece of the pie and presented clock-wide with the first and several other bins labeled in *red*. The key for each of 35 MapMan bins is listed on *right* (color figure online)

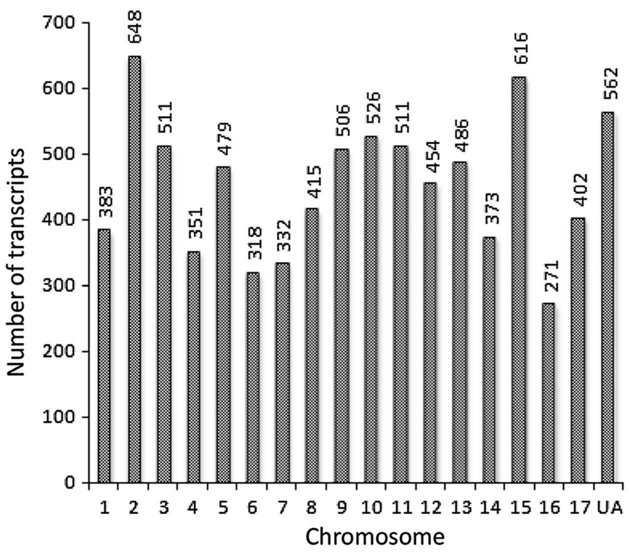

Fig. 5 Chromosomal distribution of the 8,144 novel transcripts. *UA* unanchored in the apple genome

reference transcriptome missing in Md-v1.0-RT. Furthermore, by de novo assembling of reads that could not be mapped to Md-v1.0 in RNA-seq mapping, an additional 9,380 novel transcripts were uncovered. Taken together, these revisions represent another important section of the apple reference transcriptome. Collectively, we revealed 17,524 new transcripts in this study. Similar results were reported in other plant species. For example, mapping of RNA-seq reads had identified 5, 285 genes that were never annotated previously in the cucumber genome (Li et al. 2011). In barley, RNA-seq based gene annotations led to a report of 34,276 novel transcribed genomic regions (Olson et al. 2013).

The 9,380 new transcripts were identified from the reads that could not be mapped initially. To provide evidence that they were of apple origin rather than from other sources or contamination, the 9,380 new transcripts were BLAST searched against Md-v1.0 and GenBank databases. The searches found that 9,377 of them were returned with a significant hit(s) in apple (9,375) or strawberry (2) (Table 2), strongly suggesting that these transcripts are of the apple genome origin. However, the chromosomal locality of the 9,380 new transcripts could not be determined in this study. To locate them on chromosomes, a dedicated effort is necessary. It is highly likely that some of these new transcripts would come from genomic regions currently missing in

Table 2 The number of novel transcripts returned with one or more significant hits in BLAST searches

Databases searched	Highest sequence identity (%)	*E* value	No. of transcripts	Percent
Md-v1.0	98.01–100.00	9.044E−18 to 0	7,594	80.96
Md-v1.0	90.01–98.00	6.732E−24 to 0	1,665	17.75
Md-v1.0	85.01–90.00	8.307E−22 to 0	75	0.80
Md-v1.0	70.01–85.00	8.893E−29 to 0	41	0.44
GenBank	74.53–78.10	1.66E−29 to 2.63E−71	2	0.02
Md-v1.0 and GenBank	NA	NA	3	0.03
Total			9,380	100

The cutoff is $E < 10^{-10}$

Md-v1.0. Determining their chromosomal locality would likely aid gap filling in the apple reference genome.

The three transcripts that had no significant hits in BLAST searches were also likely of the apple genome origin as they are not from contaminations of any known sources. Contamination by non-apple transcripts was indeed detected in the RNA-seq reads. In the BLAST searches, there were 56 transcripts assembled from the unmapped reads that were identified nearly identical to genes of apple virus and other microbes, including apple chlorotic leaf spot virus, apple stem pitting virus, and apple green crinkle associated virus. These non-apple transcripts were removed in the process of finding this set of new transcripts (Steps 14–16, Fig. 1). Since the 56 contaminant transcripts were found along with 9,377 apple transcripts, the likelihood for the three transcripts being of non-apple genome origin would be low. The chromosomal locality of these three transcripts will be determined together with the majority (9,377) of this group of novel transcripts.

A large number (9,627 or 55.2 %) of the 17,524 new transcripts are unknown in MapMan gene ontology (code 35.2) (Fig. 4). Several factors might have contributed to this observation. First, they were indeed new genes and there were no characterized orthologs in the databases. Second, they were un-translated regions (UTRs) disassociated with their coding regions of certain genes. Third, they were non-coding RNAs, which are common in plant genomes (Qi et al. 2013; Wu et al. 2013). Examining the 17,524 new transcripts showed that a CDS longer than 200 bp could only be found in 8,629 of them. This strongly suggested that the majority of the remaining 8,895 represent either UTRs or non-coding RNAs, largely explaining the observation.

There were 74 genes derived from merging two or more MDPs (154 single MDPs in total) because of the bridging reads between them on the same strand. These merges appeared to be similar to gene fusions that generate hybrid genes when chromosomal rearrangements bring two separate genes together. However, based on studies in plant genomes, such as Arabidopsis and rice, gene fusion is a rare and slow process (Nakamura et al. 2007). The observations of bridging reads between adjacent MDPs on the same stand were therefore unlikely caused by gene fusion, but by the imperfect gene prediction in Md-v1.0.

Improvement of the apple reference transcriptome

The apple reference genome Md-v1.0 has been available since 2010 (Velasco et al. 2010). Although the efforts are underway, Md-v2.0 has not been released thus far. We presented here an approach of repeated read mapping and transcript discovery using CLC Genomics Workbench to improve the original version of reference transcriptome Md-v1.0-RT. The major improvement is the identification of the 17,524 new transcripts. The improved reference transcriptome allowed 76.5 ± 7.0 % of the reads in the 14 samples mapped in RNA-seq mapping, representing a marked increase from 42.8 ± 4.5 %, and the reads mapping rate associated with the original Md-v1.0-RT. Testing of the improved reference with two published datasets from other genotypes and/or tissue types also showed a notable improvement in coverage of read mapping. In the first dataset (Krost et al. 2012, 2013), the coverage was increased from 46.6 ± 7.1 to 82.3 ± 2.7 %. In the second dataset (Gusberti et al. 2013), it was improved from 37.4 ± 7.7 to 62.5 ± 9.3 %. These results suggested that the improved apple reference transcriptome may be used in RNA-seq based studies involving tissues beyond fruit and genotypes beyond Golden Delicious.

The coverage of a reference transcriptome in mapping RNA-seq reads is affected by many factors, such as the purity of source mRNA, reads length, and mapping parameters, especially the minimum length fraction and the minimum similarity. In this study, we set the minimum length fraction and the minimum similarity to 0.80 and 0.98, respectively. Testing of RNA-seq mapping using the six samples from Gusbeti et al. (2013) (Table S2) against the original reference transcriptome Md-v1.0-RT showed that the 0.80/0.98 combination of parameters allowed mapping 28.8 ± 5.9 % of the reads uniquely (Table S3). This is equivalent to an 8.2-percentile reduction from the unique mapping rate (37.0 ± 3.7 %) for the same six samples reported in Gusberti et al. (2013), where the CLC default mapping parameters (0.90 in the minimum length fraction and 0.80 in the minimum similarity) were used. The mapping parameters (0.80/0.98) are therefore probably more stringent than the CLC default settings in RNA-seq mapping, suggesting that the reference coverage was likely under-estimated in this study. Using Md-v1.0-RT, a read mapping rate of 65 % was reported (Gapper et al. 2013). It was possible that the shorter read length (40 bases) and lower minimum similarity (0.95) used in the study might have contributed to the observation of a relatively high read coverage for the original reference transcriptome Md-v1.0-RT.

In the three-round read mapping and transcript discovery approach, we restricted RNA-seq reads to Golden Delicious, the source of the reference genome Md-v1.0 to avoid potential uncertainties from other genotypes. This might not be necessary given the high reads mapping rate (82.3 ± 2.7 %) from the two other genotypes (Krost et al. 2012, 2013). Nevertheless, the approach appeared to be effective as it had led to an improved reference transcriptome that gave high coverage in mapping RNA-seq reads. The important difference of this approach from the de novo transcriptome assembly approach (Krost et al. 2012, 2013; Zhang et al. 2012) is that the improved reference transcriptome was mostly built on the existing reference genome and transcriptome. Since much of the improvement (e.g. the 8,144 novel transcripts) could be localized on chromosomes, the resulting transcriptome, if used, would readily allow studies to put findings under the context of genome.

The high read mapping rates were obtained from samples under normal growth conditions. It remains to be seen whether or not the improved reference transcriptome would also provide a high coverage for read mapping when samples were treated with biotic and abiotic stresses. In the six samples (Table S2) from Gusberti e al. (2013), samples (ERR313216, ERR313224, ERR313239) challenged by *V. inaequalis* appeared to have a lower read mapping rate (58.2 ± 3.3 %) than their non-challenged controls (ERR313217, ERR313225 and ERR313226) did (66.9 ± 12.2 %), but the difference was insignificant in *t* test ($P = 0.2322$). A noteworthy point is that there were 9,887 MDPs not included in the improved reference transcriptome due to zero reads mapped to their home MDCs. Although excluding these MDPs had no or little effect on mapping RNA-seq reads in this study, they might become relevant if their expression was highly specific to certain tissues, conditions, or growth and developmental stages. It might be necessary to include them in the reference transcriptome when the improved reference transcriptome is not satisfactory in mapping RNA-seq reads. Another important point is that the improved reference genome is just one step forward towards the complete apple reference transcriptome. Much more work remains to be continued, which includes, but not limited to, mRNA sequence backed precise annotation of all protein coding genes that include UTRs and alternative splicing variants, and non-coding RNAs.

Conclusion

We improved the current apple reference transcriptome Md-v1.0-RT using three rounds of read mapping and transcript discovery based on the RNA-seq data from fruit of Golden Delicious at 14 stages of growth and development. The major improvement is the identification of 17,524 novel transcripts that either not annotated or missing in the current reference genome. The improved reference transcriptome considerably increased the RNA-seq mapping rates in the samples studied, including those from genotypes rather than Golden Delicious. The improvement represents a step forward towards a complete reference transcriptome in apple.

References

Chepelev I, Wei G, Tang QS, Zhao KJ (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res 37:1 (e106)–8 (e106)

Gapper NE, Rudell DR, Giovannoni JJ, Watkins CB (2013) Biomarker development for external CO_2 injury prediction in apples through exploration of both transcriptome and DNA methylation changes. AoB Plants 5:plt021. doi:10.1093/aobpla/plt021

Gasic K, Hernandez A, Korban SS (2004) RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. Plant Mol Biol Rep 22: 437–438

Gusberti M, Gessler C, Broggini GAL (2013) RNA-Seq analysis reveals candidate genes for ontogenic resistance in *Malus-Venturia* pathosystem. PLoS ONE 8(11):e78457. doi:10.1371/journal.pone.0078457

Krost C, Petersen R, Schmidt ER (2012) The transcriptomes of columnar and standard type apple trees (*Malus* × *domestica*)—a comparative study. Gene 498:223–230

Krost C, Petersen R, Lokan S, Brauksiepe B, Braun P, Schmidt E (2013) Evaluation of the hormonal state of columnar apple trees (*Malus* × *domestica*) based on high throughput gene expression studies. Plant Mol Biol 81:211–220

Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, Reidel EJ, Turgeon R, Liu P, Sun Q, Nelson T, Brutnell TP (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet 42:1060–1067

Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K (2011) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genom 12:540

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Nakamura Y, Itoh T, Martin W (2007) Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. Mol Biol Evol 24:110–121

Olson A, Klein RR, Dugas DV, Lu Z, Regulski M, Klein PE, Ware D (2013) Expanding and vetting sorghum bicolor gene annotations through transcriptome and methylome sequencing. Plant Genome. doi:10.3835/plantgenome2013.08.0025 (Posted online 13 Sept. 2013)

Ong WD, Voo L-YC, Kumar VS (2012) De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing. PLoS ONE 7:e46937

Qi X, Xie S, Liu Y, Yi F, Yu J (2013) Genome-wide annotation of genes and noncoding RNAs of foxtail millet in response to simulated drought stress by deep sequencing. Plant Mol Biol 83:459–473

Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325–2329

Ruttink T, Sterck L, Rohde A, Bendixen C, Rouzé P, Asp T, Van de Peer Y, Roldan-Ruiz I (2013) Orthology guided assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in *Lolium perenne*. Plant Biotechnol J 11:605–617

Sun T, Germain A, Giloteaux L, Hammani K, Barkan A, Hanson MR, Bentolila S (2013) An RNA recognition motif-containing protein is required for plastid RNA editing in Arabidopsis and maize. Proc Natl Acad Sci 110:E1169–E1178

Suzuki H, Yu J, Ness S, O'Connell M, Zhang J (2013) RNA editing events in mitochondrial genes by ultra-deep sequencing methods: a comparison of cytoplasmic male sterile, fertile and restored genotypes in cotton. Mol Genet Genomics 288:445–457

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-U174

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel C-E, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42:833–839

Wang A, Xu K (2012) Characterization of two orthologs of REVERSION-TO-ETHYLENE SENSITIVITY1 in Apple. J Mol Biol Res 2:24–41

Wilhelm BT, Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249–257

Wilhelm BT, Marguerat S, Goodhead I, Bahler J (2010) Defining transcribed regions using RNA-seq. Nat Protoc 5:255–266

Wu H-J, Wang Z-M, Wang M, Wang X-J (2013) Widespread long noncoding RNAs as endogenous target mimics for MicroRNAs in plants. Plant Physiol 161:1875–1884

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. Plant Physiol 152:1787–1795

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20:646–654

Zhang Y, Zhu J, Dai H (2012) Characterization of transcriptional differences between columnar and standard apple trees using RNA-Seq. Plant Mol Biol Rep 30:957–965

Zhong S, Joung J-G, Zheng Y, Chen Y-R, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ (2011) High-throughput Illumina strand-specific RNA sequencing library preparation. Cold Spring Harbor Protoc. doi:10.1101/pdb.prot5652