This work was funded by European Research Council (DOUBLE-TROUBLE 833522).

Inference of Ancient Polyploidy Using Transcriptome Data

Jia Li, Yves Van de Peer* and Zhen Li*

1. **Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium**
2. **VIB Center for Plant Systems Biology, VIB, Ghent, Belgium**

Correspondence to Yves Van de Peer or Zhen Li .

Abstract

Polyploidizations, or whole-genome duplications (WGDs), in plants have increased biological complexity, facilitated evolutionary innovation, and likely enabled adaptation under harsh conditions. Besides genomic data, transcriptome data have been widely employed to detect WGDs, due to their efficient accessibility to the gene space of a species. Age distributions based on synonymous substitutions (so-called K_S age distributions) for paralogs assembled from transcriptome data have identified numerous WGDs in plants, paving the way for further studies on the importance of WGDs for the evolution of seed and flowering plants. However, it is still unclear how transcriptome-based age distributions compare to those based on genomic data. In this chapter, we implemented three different de novo transcriptome assembly pipelines with two popular assemblers, i.e., Trinity and SOAPdenovo-Trans. We selected six plant species with published genomes and transcriptomes to evaluate how assembled transcripts from different pipelines

perform when using K_S distributions to detect previously documented WGDs in the six species. Further, using genes predicted in each genome as references, we evaluated the effects of missing genes, gene family clustering, and de novo assembled transcripts on the transcriptome-based K_S distributions. Our results show that, although the transcriptome-based K_S distributions differ from the genome-based ones with respect to their shapes and scales, they are still reasonably reliable for unveiling WGDs, except in species where most duplicates originated from a recent WGD. We also discuss how to overcome some possible pitfalls when using transcriptome data to identify WGDs.

1 Introduction

It is generally acknowledged that polyploidization, or whole-genome duplication (WGD), has played a significant role in the speciation, evolution, and adaptation of flowering plants, and WGDs have been identified in most angiosperm genomes [1,2,3,4]. Indeed, so far, all sequenced angiosperms, except for *Amborella trichopoda* [5] and *Aristolochia fimbriata* [6], seem to have experienced at least one WGD since the divergence of angiosperms [7]. Genome sequences, and especially well-assembled genomes, are great resources for unveiling (ancient) WGDs, because of structural information that can be used for finding synteny and/or collinearity (Chen et al. and Victor et al., in this volume). Synteny and/or collinearity analysis compares the location and order of homologous genes within a genome or between genomes. A WGD event can be identified through intragenome synteny or collinearity or by showing so-called double synteny with the genome of another species [8, 9].

It has been recognized that synteny or collinearity analysis is the most reliable approach for identifying ancient WGDs. However, the analysis depends on the continuity of genome assembly, the age of a WGD, and the rate of genome rearrangements after WGDs [10]. Although with the help of new sequencing technologies, a plant genome sequence becomes more accessible than ever before [11, 12], the assembly and annotation of a genome is still limited by computing resources and algorithm deficiency [13, 14]. Moreover, not all species of evolutionary and economic importance have their genomes sequenced (yet), as genome sequencing may be hindered by huge genome sizes and high levels of heterozygosity and ploidy [15,16,17,18]. For such

species, transcriptome sequencing provides an alternative solution to access gene space. Through de novo transcriptome assembly, the reconstructed gene content can be used to detect WGDs by applying approaches that are entirely independent of the genome sequence, such as standard approaches utilizing age distributions for all the paralogs in a species (or the whole paranome) based on the number of synonymous substitutions per synonymous site (K_S), or, alternatively, the number of transversions per four-fold degenerated site (4dTV) (Chen et al., in this volume). By plotting the number of duplicated genes against the age of the duplication event, the paranome age distribution shows a peak at a specific K_S value if a species has experienced a WGD [19, 20]. Thus, unlike identifying synteny or collinearity using genomes, the paranome age distribution only needs a well-represented gene space, which can be efficiently obtained by transcriptome sequencing these days [21].

Many studies have successfully employed transcriptomic data to build paranome age distributions for the detection of WGDs in various lineages of plants [22,23,24,25,26]. Even before the era of next-generation sequencing, Cui et al. [22] have applied the approach to Expressed Sequence Tag (EST) data and identified WGDs in species from Nymphaeales, Magnoliids, Gnetales, etc. Now, after more than a decade, analysis of fully sequenced genomes of the above species has confirmed most of these previously identified WGDs [10, 27,28,29]. In addition, paranome age distributions are often combined with phylogenomic approaches involving gene tree—species tree reconciliation to infer WGDs (see more details in Chen et al. in this volume). For instance, using 1124 plant transcriptomes, the one thousand plant (OneKP) transcriptome sequencing project has inferred 244 ancient WGDs across green plants based on paranome age distributions and gene tree—species tree reconciliations. Among the WGDs identified by OneKP, 65 (27%) could be verified by currently published genomes [30]. However, the OneKP project also seems to have missed several WGDs that could be identified through the use of entire genome sequences [31], suggesting that transcriptomes may still have less power than do complete genomes concerning the identification of WGDs.

Indeed, compared with all genes predicted in a genome, the gene space reconstructed by transcriptomes is neither complete nor nonredundant, so that it may affect the correct inference of WGDs. A characteristic of

transcriptomes is that many expressed genes are environment (condition) and developmental stage dependent [32] and, as a result, transcriptome sequencing cannot guarantee to cover the complete gene space of a species. For example, even a well-assembled plant transcriptome can just retrieve up to 75% of reference transcripts in a species [33].

Also, transcriptome assembly can be more complicated than genome assembly because sequencing reads from a transcriptome have different abundances at various gene loci, while usually, sequencing reads from a genome show a somewhat uniform coverage [34]. Due to the differences in transcriptome assembly algorithms, some assemblers are reasonably good at assembling transcriptomes of certain species in particular, but no assembler outperforms others in all species. By comparing 20 biological-based and reference-free metrics, five assemblers, namely Trinity, SPAdes, Trans-ABySS, Bridger, and SOAPdenovo-Trans, have been shown to be among the best tools for de novo transcriptome assembly [35].

Finally, transcriptome assembly can assemble different products transcribed from one gene locus, causing redundancy in the reconstructed gene space. On the one hand, a gene locus may produce different transcripts (isoforms) due to alternative splicing. On the other hand, a gene locus may have different alleles and may produce several allelic transcripts, especially in species with high heterozygosity. It has been well acknowledged that high heterozygosity is always an issue for both genome and transcriptome assembly [36, 37]. Thus, when isoforms meet allelic transcripts in transcriptome assembly, they may lead to redundant assembled transcripts that originated from the same gene locus [38]. Furthermore, if the gene locus has a highly similar duplicate, isoforms and allelic transcripts may match with the gene or its duplicate, and it is then difficult to distinguish between sequences from a gene locus and sequences from its duplicate [39], potentially leading to chimeric assembled transcripts [40, 41]. Therefore, selecting an assembled transcript to represent a gene locus is often not a trivial task. Failing to do so may artificially amplify gene family sizes when building paranome age distributions and complicate the identification of signature K_S peaks [42].

Although there are issues with assembling transcriptomes, building paranome age distributions based on transcriptomic datasets has become a widely adopted approach to detect WGDs. However, it is still unclear to what extent the paranome age distributions based on genes from transcriptomes are comparable to those based on genes from genomes. In this chapter, we selected six plant species with published genomes and transcriptomes to study how transcriptomic datasets perform when using paranome K_S distributions to detect the most recent well-documented WGDs in these six species. Using genes predicted in each genome as references, we evaluated the effects of missing reference genes, gene family clustering, and de novo assembled transcripts on the paranome K_S distributions. Our results show that although the transcriptome-based paranome K_S distributions differ from the genome-based ones, they are still reasonably reliable for identifying WGD when using cautiously.

2 Materials

2.1 Plant Genomes and Transcriptomes

We selected six plants with available genomes and RNA-seq datasets (Table 1): two monocot species, i.e., pineapple (*Ananas comosus*) and Phalaenopsis (*Phalaenopsis equestris*), and four eudicot species, i.e., Arabidopsis (*Arabidopsis thaliana*), papaya (*Carica papaya*), soybean (*Glycine max*), and grape (*Vitis vinifera*). The pineapple genome has experienced two ancient WGD events, namely σ and τ, and the most recent WGD has a peak at ~1.2 in the K_S distributions for the whole paranome [43]. The Phalaenopsis genome has one WGD identified with a peak at ~1.1 in the K_S distributions for the whole paranome [44]. In Arabidopsis, two WGDs have been uncovered (since the *γ* WGD shared by all core eudicots) in the K_S distributions for the whole paranome, and the most recent WGD has a signature K_S peak at ~0.8 [45]. Soybean has experienced a very recent WGD with a K_S peak at ~0.2 and retained more than 75% of the duplicated genes that originated from this WGD [46]. Grape and papaya have experienced no additional WGDs after the ancient hexaploidization (*γ*) that is shared by all the core eudicots [47]. The ancient hexaploidization (*γ*) signature K_S peak varies in different K_S distributions (grape ($K_S \approx 1.2$); papaya ($K_S \approx 1.8$)) for the simple reason that different species can have different synonymous substitution rates. RNA-Seq data from leaf,

root, and stem were collected for each species except for *C. papaya*, which only has transcriptomes from leaf and root. We also created a "mixed" sample in each species by merging RNA-Seq data from different tissues.

3 Methods

3.1 De Novo Assembly of Transcriptomes

For de novo transcriptome assembly, we implemented three standard pipelines that are widely employed in various evolutionary genomic studies [11, 31, 48,49,50]. Briefly, Trinity v2.12.0 [34] or SOAPdenovo-Trans v1.03 [51] was first used to de novo assemble cleaned RNA-Seq reads for each sample. After collecting the assembled transcripts, Transdecoder v5.0.2 (https://github.com/TransDecoder/TransDecoder/) was used to predict open reading frames (ORFs). Because redundant assembled transcripts resulted from alternative splicing, allelic transcripts, or highly similar duplicates still exist, the predicted ORFs were clustered by tools like CD-HIT v4.8.1 [52] aiming at selecting a representative sequence at each gene locus. Also, BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.0.0 [53] was integrated into each pipeline for a primary evaluation of the gene space.

3.1.1 Data Preprocessing

Trimmomatic v0.39 [54] with default parameters was used for removing Illumina sequencing adaptors, and low-quality bases in sequencing reads.

```
trimmomatic PE -summary trimmomatic.summary.log fq1.gz
fq2.gz clean.fq1.gz unpaired.fq1.gz clean.fq2.gz
unpaired.fq2.gz SLIDINGWINDOW:5:20 LEADING:3 TRAILING:3
MINLEN:50
```

- LEADING:3, TRAILING:3. Remove low-quality or N bases (below quality 3) at both ends.
- SLIDINGWINDOW:5:20. Scan the read with a 5-base wide sliding window, cutting when the average quality per base drops below 20.
- MINLEN:50. Discard reads shorter than 50 bases.

3.1.2 Pipeline 1

Pipeline 1 implements a straightforward approach based on Trinity for transcriptome assembly, followed by Transdecoder and CD-HIT for predicting ORFs and removing redundant assemblies, respectively. The pipeline or similar pipelines with different parameters, such as different identity thresholds in CD-HIT, have been widely used in studies like Ren et al. [55] and Cheon et al. [50].

```
Trinity --seqType fq --min_contig_length 150 –left
clean.fq1.gz --right clean.fq2.gz --output trinity_Mixed
```

- -min_contig_length: minimum assembled contig length to report.

```
TransDecoder.LongOrfs -t Mixed.trans.clean.fa
TransDecoder.Predict -t Mixed.trans.clean.fa cd-hit -i
Mixed.transdecoder.pep -c 0.99 -o
Mixed.transdecoder.cdhit.p1.pep
```

- -c: sequence identity threshold. 0.99 means that sequences with 99% identity are clustered.

3.1.3 Pipeline 2

Pipeline 2 is less commonly adopted than Pipeline 1, but it uses the transcript clustering information embedded within Trinity [49]. The clustering information is defined by shared sequence content when performing de novo transcriptome assembly [34]. Ideally, such a cluster can represent a gene locus, and transcripts in the cluster are considered isoforms derived from the gene locus. Therefore, Pipeline 2 selects the longest ORF from each trinity cluster as the representative ORF for a gene locus.

```
Trinity --seqType fq --min_contig_length 150 –left
clean.fq1.gz --right clean.fq2.gz --output trinity_Mixed
```

- -min_contig_length: minimum assembled contig length to report.

```
TransDecoder.LongOrfs -t Mixed.trans.clean.fa
TransDecoder.Predict -t Mixed.trans.clean.fa perl
Selecting_Trinity_transcript_based_on_Transdecoder_longest
_orfs.pl Mixed.trans.clean.fa Mixed.transdecoder.pep
Mixed.orf_info.xls Mixed.cluster.xls
Mixed.longest_orf.unigenes.fa Mixed.longest_orf.fa
```

- This is an in-house Perl script to extract the longest ORF in a Trinity cluster (see Code availability).

3.1.4 Pipeline 3

Pipeline 3 is similar to Pipeline 1, but SOAPdenovo-Trans replaces Trinity as the transcriptome assembler in the pipeline. It is the pipeline that has been used in the OneKP project [30].

```
SOAPdenovo-Trans-31mer all -s Mixed.soap.conf -K 25 -F -o
Mixed.soap_trans; GapCloser -a Mixed.soap_trans.scafSeq -b
Mixed.soap.conf -o Mixed.soap_trans.GapCloser.fa
```

- "Mixed.soap.conf": the configuration file for SOAPdenovo-Trans-31mer and GapCloser

```
max_rd_len=150[LIB]rd_len_cutoff=150avg_ins=200revers
e_seq=0asm_flags=3map_len=32q1=clean.fq1.gzq2=clean.f
q2.gz
```

- -K: kmer size
- -F: fill gaps in scaffolds

```
TransDecoder.LongOrfs -t Mixed.trans.clean.fa
TransDecoder.Predict -t Mixed.trans.clean.fa cd-hit -i
Mixed.transdecoder.pep -c 0.99 -o
Mixed.transdecoder.cdhit.p1.pep
```

- -c: sequence identity threshold. 0.99 means that sequences with 99% identity are clustered.

3.1.5 BUSCO Evaluation

For BUSCO evaluation, the following exemplar command-line was used to infer the completeness of gene space for each sample based on 1614 BUSCOs in the database of embryophyta_odb10.

```
busco -I Mixed.transdecoder.cdhit.pep -l embryophyta_odb10
-m prot
```

The numbers of predicted ORFs from the three pipelines for different samples are shown in Table 2. In general, Pipeline 1 predicted the most ORFs in all the examined transcriptomes, whereas Pipeline 2 and Pipeline 3 generated different but similar numbers of ORFs (Table 2). Similarly, for a specific sample, Pipeline 1 resulted in the most Complete BUSCOs (including both nonduplicated and duplicated), followed by Pipeline 2 and then Pipeline 3. The fractions of Complete BUSCOs are comparable among different pipelines, except for Pipeline 3, which showed worse performance in assembling the RNA-Seq reads from *G. max* (Fig. 1). Additionally, in all the species and samples, ORFs from Pipeline 1 always have the highest fractions of Duplicated BUSCOs, suggesting that Pipeline 1 produced a certain level of gene space redundancy in different samples.

Further, we directly compared the numbers of predicted ORFs and the numbers of genes in the reference genomes (Fig. 2). Compared to the number of reference genes in the genomes, Pipeline 1 tends to predict many more ORFs. It should be noted that the differences in numbers between the predicted ORFs from Pipeline 1 and those from Pipelines 2 and 3 differ to various extents in the six species, which is likely correlated with the heterozygous level of sequences in the RNA-Seq reads. For example, *A. thaliana* is a self-pollinated plant with heterozygosity as low as 0.5% [56]. The numbers of predicted ORFs in all three pipelines are close to the number of reference genes. However, compared with other species, Pipeline 1 in *A.*

comosus and *G. max* predicted more ORFs than the reference genes (Fig. 2). Both species must have more heterozygous sequences in their RNA-Seq reads because the *A. comosus* genome has a high heterozygosity of 2% [43], while the *G. max* genome still contains many duplicated genes resulted from a very recent WGD [46]. In addition, the BUSCO results for the predicted ORFs from Pipeline 1 in *A. comosus* and *G. max* also show higher fractions of Duplicated BUSCOs than that in other species (Fig. 1), indicating that they still have redundant ORFs resulting from allelic transcripts or duplicated genes. Because both high heterozygosity and recent duplicates with highly similar sequences can cause similar issues in de novo transcriptome assembly, our results suggest that Pipeline 1 may behave suboptimal in dealing with highly heterozygous sequencing reads.

In contrast, compared to the number of genes in the completely sequenced genomes, Pipelines 2 and 3 predicted fewer or sometimes relatively comparable numbers of ORFs. Both pipelines resulted in lower fractions of Duplicated BUSCOs than Pipeline 1 (Fig. 1), suggesting that they removed more ORFs with similar sequences. However, compared with other species, both pipelines predicted much fewer ORFs than the reference genes in *G. max*, but Pipelines 2 and 3 have different performances in the BUSCO evaluations. ORFs from Pipeline 2 have nearly equivalent fractions of Complete BUSCOs but almost no Duplicated BUSCOs. This may be an issue for a species still retaining duplicates from a recent WGD, as we expect more duplicated genes in the genome. Because the duplicated genes retained from the very recent WGD in *G. max* are still quite similar, Trinity could falsely cluster true paralogs with similar sequence content during the assembly process. For the ORFs from Pipeline 3, they have lower fractions of Complete BUSCOs, but higher fractions of Fragmented BUSCOs than do Pipelines 1 and 2 in the samples of *G. max* (Fig. 1), suggesting that Pipeline 3 with SOAPdenovo-Trans handles duplicated genes with highly similar sequences differently from Pipeline 2. In either case, both Pipelines 2 and 3 performed less well in species with a recent WGD than in species without.

3.2 Building K_S Distributions for the Whole Paranomes

Finally, we used the wgd v1.2 program [57] to build K_S distributions for the whole paranomes based on the predicted ORFs from de novo transcriptome assemblies and the reference genes in genomes (Chen et al., in this volume). The wgd suite integrates commonly used K_S and collinearity analysis workflows with Gaussian mixture modeling and result visualization tools, providing researchers with a convenient way to detect WGD events based on genomic or transcriptomic data. Below, we take the predicted ORFs from the mixed sample of *V. vinifera* as an example for using wgd:

```
wgd dmd -I 3 Mixed.selected.transdecoder.p1.cds -o
01.wgd_dmd –nostrictcds
```

- -I: --inflation FLOAT inflation factor for MCL.
- -nostrictcds: do not enforce proper CDS sequences, which means all the cds, including the complete cds with start codon and stop codon and the incomplete cds, will be used for clustering.

```
wgd ksd 01.wgd_dmd/Mixed.selected.transdecoder.p1.cds.mcl
Mixed.selected.transdecoder.p1.cds -o 02.wgd_ksd
```

Using the K_S distributions for the whole paranome of *V. vinifera* as an example (Fig. 3), our results show that K_S distributions based on transcriptomes are different from those based on genes predicted in complete genomes. Also, the transcriptome-based K_S distributions are different from each other depending on the different pipelines used. The peak representing the hexaploidization event in the *V. vinifera* genome at ~1.2 is evident in the K_S distributions based on ORFs from Pipelines 2 and 3, but less so in the K_S distribution based on ORFs from Pipeline 1, the reason being that the transcriptome-based K_S distribution based on ORFs from Pipeline 1 exhibits an abnormally high number of duplicates at low K_S values (0–0.1). Such an abnormally high peak overshadows the WGD signature K_S peak in *V. vinifera*, leading to potential failures in detecting WGDs.

For the rest of the chapter, we further compared the transcriptome-based and genome-based K_S distributions in the six plant species by considering the effects of missing reference genes, gene family clustering, and de novo assembled transcripts. Because the mixed sample in each species combines all the expressed genes from different tissues, it always contains considerably more ORFs than individual tissues, no matter which pipeline was used (Table 2). In addition, the mixed samples from the six species also have the most Complete BUSCOs in all pipelines (Fig. 1). Therefore, for building K_S distributions and further analyses in the chapter, we only focus on the de novo assemblies based on the mixed samples.

4 Missing Reference Genes and K_S Distributions

The differences in numbers of ORFs and genes indicate that the transcriptome assemblies missed some reference genes in the genome and produced some unknown ORFs. Here, we define the missing reference genes as those that exist in the reference genomes but do not appear in the predicted ORFs in the transcriptome assemblies. Unknown ORFs are defined as the ORFs that only exist in the transcriptomes but are not found in the reference genome. Because missing reference genes must affect K_S distributions, we first determine the gene space that could be reconstructed by the three de novo assembly pipelines, using the predicted genes in the genomes as references.

4.1 Gene Space Reconstructed by Transcriptome Assembly

To obtain an upper bound of the gene space that can be reconstructed by transcriptome sequencing in a species, we first mapped all the RNA-Seq reads of each species to their corresponding genome by Hisat2 v2.1.0 with a parameter "--dta" to only report alignments tailored for transcriptome assemblers [58]. The upper bound of gene space for a sample was then defined as the number of reference genes mapped by at least two RNA-Seq reads. Here, we used a loose cut-off of two RNA-Seq reads for estimating the upper bounds in different species not only because it is the minimum number of RNA-Seq reads used in Trinity [34], but it is also the minimum requirement for doing any assembly. The results show that about 48–69% of the reference genes have the potential to be assembled, and the upper bounds of gene space vary in different species (Fig. 4).

Then, we mapped the ORFs predicted by the three pipelines to their corresponding genomes by BLAT v3.5 [59] to determine how many reference genes could be retrieved from the assembled transcripts. Apparently, not all the reference genes supported by two or more RNA-Seq reads could be assembled. The fractions of reconstructed genes vary from species to species, with the highest in *A. comosus* and the lowest in *C. papaya* and *P. equestris* (Fig. 4). For different pipelines, in general, Pipeline 1 assembled more reference genes than Pipelines 2 and 3, but the fractions are close in each species except for *G. max*. Also, for the predicted ORFs, most are complete ORFs with 100% coverage or nearly complete ORFs with a coverage ≥ 90% of the corresponding reference genes. The only exception is that Pipeline 3 assembled most ORFs with a coverage less than 90% of the reference genes in *G. max*. As discussed above, this may be related to the poor performance of SOAPdenovo-Trans on species that have undergone a recent WGD.

Because each species has a significant fraction of missing reference genes, we wonder if this would affect identifying WGD signals in K_S distributions. To this end, we marked the reference genes in the genome-based K_S distributions according to their assembly status in the transcriptomes (Fig. 5). We noticed that the potential K_S signals could be recovered by paralogous pairs that are completely or partially assembled in the predicted ORFs. Although the missing reference genes make up a certain proportion of genes at each K_S interval, especially at intervals with small K_S values, they do not really affect the appearance of WGD peaks in the K_S distributions, reassuring the application of transcriptomic dataset in detecting WGDs.

4.2 Redundant ORFs

By mapping the predicted ORFs to the reference genes in each genome, we also evaluated the transcriptome assembly redundancy of the three pipelines. The mapping results demonstrate that Pipeline 1 could produce redundant ORFs, excessing to half of the gene loci having more than one mapped ORFs, especially for *G. max* (Fig. 6). On the contrary, the proportions of gene loci to which only one ORF was mapped increased dramatically for ORFs from Pipelines 2 and 3. Pipeline 2 demonstrates a better capability to remove redundancies because only one representative ORF was kept in each Trinity

cluster, but at a risk of falsely clustering paralogous genes into one Trinity cluster. Pipeline 3 seems to be more moderate than the other two pipelines, balancing the number of ORFs and redundancy.

5 Gene Family Clustering and K_S Distributions

Gene families containing paralogous gene pairs are another basis for building K_S distributions. However, in transcriptome assemblies, isoforms and allelic assemblies together with missing genes can confound gene family identification, so we investigated the effects of gene family clustering on the transcriptome-based K_S distributions. We identified 4242, 4683, 3120, 10,979, 3136, and 3680 multigene families in genomes with reference genes in *Ananas comosus Arabidopsis thaliana, Carica papaya, Glycine max, Phalaenopsis equestris,* and *Vitis vinifera*, respectively. For the predicted ORFs, those from Pipeline 1 always have more gene families identified than the reference genes, while the ones from Pipelines 2 and 3 have a comparable number of gene families as the reference genes (Table 3).

5.1 Presence and Absence of Gene Families

Next, we assessed how well gene families identified on the basis of the reference genes could be predicted based on the ORFs by the three de novo assembly pipelines. To determine the correspondence between gene families of the predicted ORFs and the reference genes, the predicted ORFs were aligned to the reference genes according to Chen et al. [49] using BLAT v3.5 [59]. For subsequent analysis, poor hits with a match length shorter than 100 bp and identity lower than 95% were discarded. The hit with the highest bit-score in BLAT was kept when an ORF had multiple hits to reference genes. If a transcriptome-based gene family only contains ORFs mapped to a single genome-based gene family and vice versa, their correspondence could be precisely determined. However, in many cases, ORFs in transcriptome-based gene families have hits to multiple genome-based gene families and vice versa, so we defined the correspondence of a transcriptome-based gene family and a genome-based gene family if they reciprocally had the most hits to each other. In addition, transcriptome-based gene families that could not match the criteria on correspondence but had hits to genome-based gene families were

considered problematic families. Gene families with no hits to genome-based gene families were considered unknown families.

It turns out that the ORFs from Pipeline 1 could identify 70–80% of the genome-based gene families. The identification ratio of genome-based gene families decreased to 60–70% for the ORFs from Pipelines 2 and 3 (Table 3). The only exception is *G. max*, which missed nearly half of genome-based gene families in Pipelines 2 and 3, possibly related to the aforementioned assembly issues. The ratios of identified gene families are slightly higher than those of reconstructed gene spaces. As more than half of the gene families of a species exist in the transcriptome assemblies, the transcriptome-based K_S distributions should uncover signature peaks in the K_S distributions for the whole paranome. However, besides the presence and absence of gene families, the shape of K_S distributions also depends on gene family sizes.

5.2 Size Differences of Gene Families

To compare sizes between the transcriptome-based and genome-based gene families, we only selected the corresponding gene families and plotted cumulative distributions for the size differences between a pair of transcriptome- and genome-based gene families. Further, to measure if the size differences are significantly larger or smaller, we sampled the exact number of genes from all the corresponding genome-based gene families 100 times in each species. Finally, in each resampled set, we calculated z-scores for the size differences between the resampled gene families and the genome-based gene families. We hence could define that a transcriptome-based gene family with a size difference larger and smaller than a *z*-score of 2 and −2, i.e., outside twice standard deviations, is a gene family with significantly different sizes (Fig. 7).

Our results show that, for the transcriptome-based gene families, the ones reconstructed from the ORFs from Pipelines 2 and 3 do vary from the sizes of their corresponding genome-based gene families. However, most of them still have similar sizes as the genome-based gene families. In contrast, the ORFs from Pipeline 1 have more gene families that are significantly larger than their corresponding genome-based gene families, in line with the results that there

are more redundant ORFs at the same gene locus in Pipeline 1 (Fig. 6). Heterozygosity also affects gene family clustering of ORFs in transcriptome assembly. Comparing *A. thaliana* with low heterozygosity and *A. comosus* with high heterozygosity, we found that the latter species has more transcriptome-based gene families significantly larger than their corresponding genome-based ones. Interestingly, ORFs from different pipelines produced similar fractions of gene families smaller than the corresponding genome-based gene families, except for *G. max*. As a species with a recent WGD, ORFs in *G. max* not only have much fewer genome-based gene families, but the ORFs from Pipeline 1 formed too many significantly larger gene families, and the ORFs from Pipelines 2 and 3 formed too many significantly smaller gene families, suggesting that none of the implemented assembly pipelines are ideally suited for this species.

5.3 Gene Family Sizes and K_S Distributions

To illustrate the effects of gene family identification and size changes on the transcriptome-based K_S distributions, we depicted different kinds of gene families classified above in the transcriptome-based K_S distributions and compared them with the genome-based K_S distributions (Fig. 8). Gene families that are significantly larger than their corresponding genome-based ones mainly appear in the ORFs from Pipeline 1. They contribute a certain fraction to ORF pairs with $K_S < 0.1$ in the histograms. In the K_S distributions based on the ORFs from Pipelines 2 and 3, such large gene families tend to stand out in gene families of ORF pairs with large K_S values, such as those in *A. comosus* and *V. vinifera*. To certain extent, the problematic gene families also contribute to gene families of ORF pairs with $K_S < 0.1$, but their fractions seem to have no preference toward K_S values. In addition, the fractions of unknown gene families are also different from species to species, but in many species, they tend to be present in gene families of ORF pairs with low K_S values ($K_S < 0.5$).

Our results would suggest that gene families that are larger than their corresponding genome-based gene families and the unknown gene families that do not exist in the “true” genomes may inflate the number of ORF pairs at low K_S values in the K_S distributions, which would affect K_S peaks for WGDs,

especially for ORFs from Pipelines 1 and 2. For the K_S distributions based on ORFs from Pipeline 1, the WGD K_S peaks, except the one for the WGD in *A. thaliana* expected at $K_S \approx 0.8$, are hidden in the tail of the histograms, simply because there are too many ORF pairs with small K_S values resulting from large and unknown gene families.

For the K_S distributions based on ORFs from Pipeline 2, these have in general fewer ORF pairs than the K_S distributions based on the reference genes, especially for the ORF pairs with small K_S values (Fig. 8), suggesting that Pipeline 2 may collapse many recent duplicates that would usually compromise the number of duplicates with small K_S values. However, despite somehow obscure, the K_S peaks representing WGD events can still be seen in the K_S distributions for *A. comosus*, *A. thaliana*, *P. equestris*, and *V. vinifera*. Although removing the unknown gene families in the K_S distributions may help increase the visibility of the signature K_S peaks for WGDs, it is impossible to do so when the investigated species has no information on its actual gene space, a typical situation when utilizing transcriptomic data for WGD detection.

Although having fewer duplicate gene pairs, the K_S distributions based on ORFs from Pipeline 3 seem comparable with the K_S distributions based on the reference genes. All the K_S peaks for the acknowledged WGDs, except the most recent one in *G. max*, can be identified. The inflated effects of the unknown gene families for ORF pairs with small K_S values are also mild in these K_S distributions. However, the K_S value for the peak in *P. equestris* seems to shift toward a smaller K_S value because of the inflated effects.

Specifically, none of the K_S distributions based on ORFs from the three de novo transcriptome assembly pipelines shows the K_S peak at ~0.2 in *G. max* for its most recent WGD, which has produced a highly duplicated genome with around 75% of the genes still present in a multicopy status [46]. All three de novo assembly pipelines seem to have issues with a genome with a high proportion of recently duplicated genes. They either generate many more significantly larger gene families (for Pipeline 1) or significantly smaller gene families (for Pipelines 2 and 3), suggesting a fine(r)-tuned de novo transcriptome assembly pipeline is required for species that still retained most duplicates after a recent WGD event.

6 De Novo Assemblies and K_S Distributions

Although the three de novo assembly pipelines for transcriptomes show different performances on gene space completeness (Fig. 4) and redundancy (Fig. 6), as well as gene family sizes (Fig. 7), it is still not clear how assembled transcripts or ORFs affect K_S distributions. To this end, we classified the predicted ORFs into five distinct groups based on the mapping results of ORFs to gene loci, i.e., "Correct," "Isoform," "Isoform-like," "Fragmented," and "Unknown" ORFs (Fig. 9). Specifically, a "Correct" ORF is the only best match to a reference gene locus, where the ORF should cover 95% of coding sequences of the reference gene. For the redundant ORFs, although sometimes recent gene duplications may confound redundancy, such ORFs are mainly from different alleles or various isoforms at a gene locus. Because it is difficult to clearly distinguish whether a predicted ORF is from a different allele or an isoform (or both), we here used "Isoform" ORFs to represent ORFs that best match the same reference gene. The "Isoform" ORFs have both start and stop codons, and they only contain sequences from the exons predicted in the reference genome. In contrast, some ORFs with start and stop codons have the best match to the same reference gene, but they may contain extra exons or partial sequences of exons in the reference genome. Because reference gene predictions may also be problematic, we defined such ORFs as "Isoform-like" ORFs. For ORFs that have no start codon and/or stop codons but could be mapped to a part of a reference gene, we classified them as "Fragmented" ORFs. In the end, the rest are the "Unknown" ORFs that could not be mapped to the reference genomes or any genes thereof.

Among the six plant species investigated, *A. thaliana* has the highest proportion of Correct ORFs, ranging from 23.4% to 50.0% for the three pipelines (Fig. 10). The other species show massive reductions in the proportions of the correct ORFs. *A. thaliana* and *G. max* also have higher proportions of Fragmented ORFs than other investigated species. In addition, *A. thaliana* and *G. max* have the most diminutive proportions of the Unknown ORFs, in line with the gene family analyses where both species have much fewer unknown gene families than other species (Fig. 8). The unknown ORFs may have different sources. They could be de novo assembly artifacts or assembly of contaminated reads in RNA-Seq samples. For example, we found

that many Unknown ORFs in *A. comosus* are from microorganisms by annotating these ORFs with the NCBI Nonredundant database, indicating potential contamination when preparing the samples for transcriptome sequencing.

Concerning the pipelines, Pipeline 1 contains the most Isoform ORFs, while Pipeline 2 contains the least of such ORFs, whereas the proportions of Isoform-like ORFs are relatively similar among the three pipelines. In the transcriptome-based K_S distributions based on ORFs from Pipeline 1 (Fig. 11), ORF pairs with $K_S < 0.1$ include many Isoform or Isoform-like ORFs. If an assembly pipeline could not remove the Isoform or Isoform-like ORFs, they would be considered extra members for those gene families. Hence, they result in many gene families with significantly larger sizes than their corresponding genome-based gene families. There is no such pattern in the transcriptome-based K_S distributions based on ORFs from Pipeline 2, while the first bars in the transcriptome-based K_S distributions based on ORFs from Pipeline 3 have more Isoform-like ORFs than Isoform ORFs.

The Fragmented ORFs exist in ORF pairs with all sorts of K_S values. Again, in the transcriptome-based K_S distributions based on ORFs from Pipeline 1, the Fragmented ORFs are mainly present in ORF pairs with $K_S < 0.1$. However, they are also found in ORF pairs with different values, likely forming the problematic gene families in Fig. 8. Because the Fragmented ORFs are relatively short and contain one or a few conserved domains, they might disturb gene family identifications by incorrectly joining different gene families or falsely forming independent gene families.

7 Discussion

Because of the ease of transcriptome sequencing, transcriptomic data have been widely adopted to infer WGDs, and it has become standard practice to examine transcriptomes from dozens, if not hundreds of species, to detect WGDs in a large-scale phylogeny [30, 55, 60, 61]. Such studies systematically allow to investigate the importance of WGDs for the evolution of green plants. However, there are some methodological concerns using K_S distributions with transcriptomic datasets because they may lead to fallacious conclusions drawn from falsely detected WGD events [42, 62]. Here, we compared genome-

based K_S distributions with transcriptome-based K_S distributions resulting from three different de novo assembly pipelines. Our results show that with proper transcriptome assembly, although the transcriptome-based K_S distributions have different shapes than the genome-based ones, they have the power to identify WGDs but may fail with species that still retained most duplicates from a recent WGD.

Transcriptome assembly pipelines, especially the steps that remove redundant assemblies, have significant impacts on inferring WGDs based on K_S distributions. Despite missing some reference genes and gene families, all the implemented assembly pipelines here could reconstruct enough genes and gene families for WGD detection, if transcriptomes are relatively well sequenced. However, gene space redundancy is a more severe issue than (lack of) completeness. We show that in a pipeline (Pipeline 1) that has been widely applied to various studies (maybe with different parameters, though), Isoform or Isoform-like ORFs are treated as genuine duplicated genes, and most of them have small K_S values less than 0.1. As a result, the extraordinary large numbers of duplicates with small K_S values overshadow the signature WGD peaks in the examined K_S distributions. Some efforts could alleviate the effects of redundant ORFs, for example, through clustering redundant ORFs with a decreasing identity in CD-HIT or by removing ORF pairs with minimal K_S values [42, 62]. However, if the cut-offs used to remove redundancy are too stringent, they have limited effects on the number of redundant ORFs in the transcriptome assembly. On the other hand, if they are too loose, they may collapse genuine paralogous genes with small K_S values and leave an artificial K_S peak slightly larger than 0.1. The peak could then be falsely identified as evidence for a recent WGD event or considered an artifact after removing too many genuine paralogous genes with small(er) K_S values [55]. Although some of the recent WGDs seem to be supported by analyses using genomic data, the K_S peak values from the transcriptome-based K_S distributions are sometimes different from the ones in the genome-based K_S distributions [42], suggesting the inference of WGDs may be still arbitrary and requires further corroboration. Moreover, determining the cut-offs for removing redundant ORFs in transcriptome assembly may require prior knowledge about sequenced species, such as the heterozygosity. It is, of course, helpful to have such information before sequencing a species, but it is

also not an easy task for studies including hundreds of species, especially if the cut-offs are required to be fine-tuned from species to species.

Alternatively, relying on the algorithms in de novo transcriptome assemblers is another solution to eliminate redundant ORFs. Trinity and SOAPdenovo-Trans are two de novo transcriptome assemblers based on the *de Bruijn* graph. Both assemblers have been used for detecting WGDs with K_S distributions [63,64,65]. In Trinity, sequencing reads in the same assembly graph or Trinity cluster are separately assembled, and the final assembled transcripts retain the cluster information. Because a Trinity cluster is often considered to be corresponding to a gene, our Pipeline 2 uses the clustering information to select one representative sequence for each Trinity cluster. As shown in the results of the BUSCO evaluation and K_S distributions, Pipeline 2 removed many duplicated ORFs and scaled-down the overall number of duplicates in the K_S distributions. Nevertheless, the pipeline is inclined to remove more ORF pairs with K_S less than 0.5, leading to failures in detecting very recent WGD events.

Like Trinity, SOAPdenovo-Trans also provides cluster or gene information after transcriptome assembly, but it shows a more reasonable number of gene loci than does Trinity. As shown by us (this chapter) and others [35, 66], Trinity does assemble more Complete BUSCOs than SOAPdenovo-Trans, but it also produces a higher proportion of Duplicated BUSCOs. For instance, in *A. thaliana*, SOAPdenovo-Trans reported 8242 gene loci, one-fourth of which have multiple isoforms with a maximum number of isoforms up to five. On the contrary, Trinity reported 46,364 gene loci, in which 7412 gene loci have multiple isoforms with a maximum number of up to 98 isoforms. Therefore, the assembled results of SOAPdenovo-Trans have much lower redundancy than the results of Trinity, reducing the efforts to further remove redundant ORFs. In addition, our results show that the K_S distributions using ORFs assembled by SOAPdenovo-Trans (Pipeline 3) are more comparable to the genome-based K_S distributions with respect to their shapes and scales. Meanwhile, the running time for SOAPdenovo-Trans is shorter than that of Trinity, which is significantly meaningful when conducting studies with hundreds of species, likely explaining why SOAPdenovo-Trans has been chosen in the OneKP to some extent.

Besides transcriptome assembly, there are other concerns related to RNA-Seq sequencing that may affect using K_S distributions to infer WGDs, such as RNA extraction, library preparation, sequencing depth, and other unexpected events like sample contamination. Compared to genome sequencing and assembly, transcriptome sequencing and assembly are more unstable, and hence challenging to measure their quality. However, if the data volume of the transcriptome is not high enough, the resulting K_S distributions would be less informative and suitable for WGD inference, due to the insufficient gene space. For instance, compared to other samples, the leaf sample of *P. equestris* only produced one-third of total ORFs (Table 1 and Fig. 1). Another issue that should be avoided is sample contamination, as found in the root sample of *A. comosus*, because alien ORFs from other species may have unexpected effects on K_S distributions.

As newly developed sequencing technologies have already been applied to transcriptome sequencing [67], single-molecule and long-read sequencing may help solve some issues related to short-read sequencing. For example, Yue et al. [68] have used PacBio Single Molecule, Real-Time (SMRT) sequencing technology to generate full-length transcriptome data for a sterile triploid *Crocus sativus,* and have successfully identified a recent WGD event in its evolutionary history. In any case, transcriptomic data, increasingly generated as supplements to genomic data, are a great asset for the identification and delineation of ancient polyploid events.

Codes Availability

Commands and scripts used in this study are available at https://github.com/li081766/transcriptome_WGD_project.

References

1. Guo J, Xu W, Hu Y et al (2020) Phylotranscriptomics in cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. Mol Plant 13:1117–1133

CrossRef CAS Google Scholar

2. Sheehan H, Feng T, Walker-Hale N et al (2020) Evolution of l-DOPA 4, 5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales. New Phytol 227:914–929

CrossRef CAS Google Scholar

3. Xiang Y, Huang C-H, Hu Y et al (2017) Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Mol Biol Evol 34:262–281

CAS Google Scholar

4. Van de Peer Y, Ashman T-L, Soltis PS et al (2020) Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell 33:11–26

CrossRef Google Scholar

5. Albert VA, Barbazuk WB, Depamphilis CW et al (2013) The Amborella genome and the evolution of flowering plants. Science 342:1241089

CrossRef Google Scholar

6. Qin L, Hu Y, Wang J et al (2021) Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome. Nat Plants 7(9):1239–1253

CrossRef CAS Google Scholar

7. Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat Rev Genet 18:411–424

CrossRef Google Scholar

8. Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. Nat Rev Genet 5:752–763

CrossRef Google Scholar

9. Tang H, Bowers JE, Wang X et al (2008) Synteny and collinearity in plant genomes. Science 320:486–488

CrossRef CAS Google Scholar

10. Wan T, Liu Z, Leitch IJ et al (2021) The Welwitschia genome reveals a unique biology underpinning extreme longevity in deserts. Nat Commun 12:4247

CrossRef CAS Google Scholar

11. Belser C, Istace B, Denis E et al (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants 4:879–887

CrossRef CAS Google Scholar

12. Michael TP, Jupe F, Bemm F et al (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun 9:1–8

CrossRef Google Scholar

13. Kersey PJ (2019) Plant genome sequences: past, present, future. Curr Opin Plant Biol 48:1–8

CrossRef CAS Google Scholar

14. Salzberg SL (2019) Next-generation genome annotation: we still struggle to get it right. Genome Biol 20:92

CrossRef Google Scholar

15. Unamba CIN, Nag A, Sharma RK (2015) Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. Front Plant Sci 6:1074

CrossRef Google Scholar

16. Kyriakidou M, Tai HH, Anglin NL et al (2018) Current strategies of polyploid plant genome sequence assembly. Front Plant Sci 9:1660

CrossRef Google Scholar

17. Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr Opin Plant Biol 54:26–33

CrossRef CAS Google Scholar

18. Voshall A, Moriyama EN (2020) Next-generation transcriptome assembly and analysis: impact of ploidy. Methods 176:14–24

CrossRef CAS Google Scholar

19. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

CrossRef CAS Google Scholar

20. Tuskan GA, DiFazio S, Jansson S et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604

CrossRef CAS Google Scholar

21. Li Z, Barker MS (2020) Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. GigaScience 9:giaa004

CrossRef CAS Google Scholar

22. Cui L, Wall PK, Leebens-Mack JH et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749

CrossRef CAS Google Scholar

23. Cai L, Xi Z, Amorim AM et al (2019) Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. New Phytol 221:565–576

CrossRef CAS Google Scholar

24. Godden GT, Kinser TJ, Soltis PS et al (2019) Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints. Genome Biol Evol 11:3393–3408

CAS Google Scholar

25. Wang Y, Nie F, Shahid MQ et al (2020) Molecular footprints of selection effects and whole genome duplication (WGD) events in three blueberry species: detected by transcriptome dataset. BMC Plant Biol 20:1–14

Google Scholar

26. Vanneste K, Sterck L, Myburg AA et al (2015) Horsetails are ancient polyploids: evidence from Equisetum giganteum. Plant Cell 27:1567–1578

CrossRef CAS Google Scholar

27. Zhang L, Chen F, Zhang X et al (2020) The water lily genome and the early evolution of flowering plants. Nature 577:79–84

CrossRef CAS Google Scholar

28. Rendón-Anaya M, Ibarra-Laclette E, Méndez-Bravo A et al (2019) The avocado genome informs deep angiosperm phylogeny, highlights

introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proc Natl Acad Sci U S A 116:17081–17089

CrossRef Google Scholar

29. Chen J, Hao Z, Guang X et al (2019) Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. Nat Plants 5:18–25

CrossRef CAS Google Scholar

30. Leebens-Mack JH, Barker MS, Carpenter EJ et al (2019) One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685

CrossRef Google Scholar

31. Wong GK-S, Soltis DE, Leebens-Mack J et al (2020) Sequencing and analyzing the transcriptomes of a thousand species across the tree of life for green plants. Annu Rev Plant Biol 71:741–765

CrossRef CAS Google Scholar

32. Strickler SR, Bombarely A, Mueller LA (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species. Am J Bot 99:257–266

CrossRef CAS Google Scholar

33. Honaas LA, Wafula EK, Wickett NJ et al (2016) Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome. PLoS One 11:e0146062

CrossRef Google Scholar

34. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512

CrossRef CAS Google Scholar

35. Hölzer M, Marz M (2019) De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 8:giz039

CrossRef Google Scholar

36. Tigano A, Sackton TB, Friesen VL (2018) Assembly and RNA-free annotation of highly heterozygous genomes: the case of the thick-billed murre (Uria lomvia). Mol Ecol Resour 18:79–90

CrossRef CAS Google Scholar

37. Freedman AH, Clamp M, Sackton TB (2021) Error, noise and bias in de novo transcriptome assemblies. Mol Ecol Resour 21:18–29

CrossRef CAS Google Scholar

38. Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14:328

CrossRef CAS Google Scholar

39. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65

CrossRef CAS Google Scholar

40. Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

CrossRef CAS Google Scholar

41. Hahn MW, Zhang SV, Moyle LC (2014) Sequencing, assembling, and correcting draft genomes using recombinant populations. G3 4:669–679

CrossRef Google Scholar

42. Wang H, Guo C, Ma H et al (2019) Reply to Zwaenepoel et al.: Meeting the challenges of detecting polyploidy events from transcriptomic data. Mol Plant 12:137–140

CrossRef CAS Google Scholar

43. Ming R, VanBuren R, Wai CM et al (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47:1435–1442

CrossRef CAS Google Scholar

44. Cai J, Liu X, Vanneste K et al (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Genet 47:65–72

CrossRef CAS Google Scholar

45. Vanneste K, Baele G, Maere S et al (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. Genome Res 24:1334–1347

CrossRef CAS Google Scholar

46. Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

CrossRef CAS Google Scholar

47. Jaillon O, Aury J-M, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

CrossRef CAS Google Scholar

48. Zhang G-Q, Liu K-W, Li Z et al (2017) The Apostasia genome and the evolution of orchids. Nature 549:379–383

CrossRef CAS Google Scholar

49. Chen L-Y, Morales-Briones DF, Passow CN et al (2019) Performance of gene expression analyses using de novo assembled transcripts in polyploid species. Bioinformatics 35:4314–4320

CrossRef Google Scholar

50. Cheon S, Zhang J, Park C (2020) Is phylotranscriptomics as reliable as phylogenomics? Mol Biol Evol 37:3672–3683

CrossRef CAS Google Scholar

51. Xie Y, Wu G, Tang J et al (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666

CrossRef CAS Google Scholar

52. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

CrossRef CAS Google Scholar

53. Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

CrossRef Google Scholar

54. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

CrossRef CAS Google Scholar

55. Ren R, Wang H, Guo C et al (2018) Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant 11:414–428

CrossRef CAS Google Scholar

56. Chin C-S, Peluso P, Sedlazeck FJ et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054

CrossRef CAS Google Scholar

57. Zwaenepoel A, Van de Peer Y (2019) wgd—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35:2153–2155

CrossRef CAS Google Scholar

58. Kim D, Paggi JM, Park C et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915

CrossRef CAS Google Scholar

59. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

CAS Google Scholar

60. Li Z, Baniaga AE, Sessa EB et al (2015) Early genome duplications in conifers and other seed plants. Sci Adv 1:e1501084

CrossRef Google Scholar

61. Stull GW, Qu X-J, Parins-Fukuchi C et al (2021) Gene duplications and genomic conflict underlie major pulses of phenotypic evolution in gymnosperms. Nat Plants 7(8):1015–1025

CrossRef Google Scholar

62. Zwaenepoel A, Li Z, Lohaus R et al (2019) Finding evidence for whole genome duplications: a reappraisal. Mol Plant 12:133–136

CrossRef CAS Google Scholar

63. Johnson MG, Malley C, Goffinet B et al (2016) A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). Mol Phylogenet Evol 98:29–40

CrossRef Google Scholar

64. Devos N, Szövényi P, Weston DJ et al (2016) Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta). New Phytol 211:300–318

CrossRef CAS Google Scholar

65. Clark JW, Puttick MN, Donoghue PCJ (2019) Origin of horsetails and the role of whole-genome duplication in plant macroevolution. Proc R Soc B Biol Sci 286:20191662

CrossRef Google Scholar

66. Chopra R, Burow G, Farmer A et al (2014) Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. PLoS Data 9:e115055

CrossRef Google Scholar

67. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

CrossRef CAS Google Scholar

68. Yue J, Wang R, Ma X et al (2020) Full-length transcriptome sequencing provides insights into the evolution of apocarotenoid biosynthesis in Crocus sativus. Comput Struct Biotechnol J 18:774–783

CrossRef CAS Google Scholar

Download references

Acknowledgments

YVdP acknowledges funding from the FWO (G090919N), from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522), and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

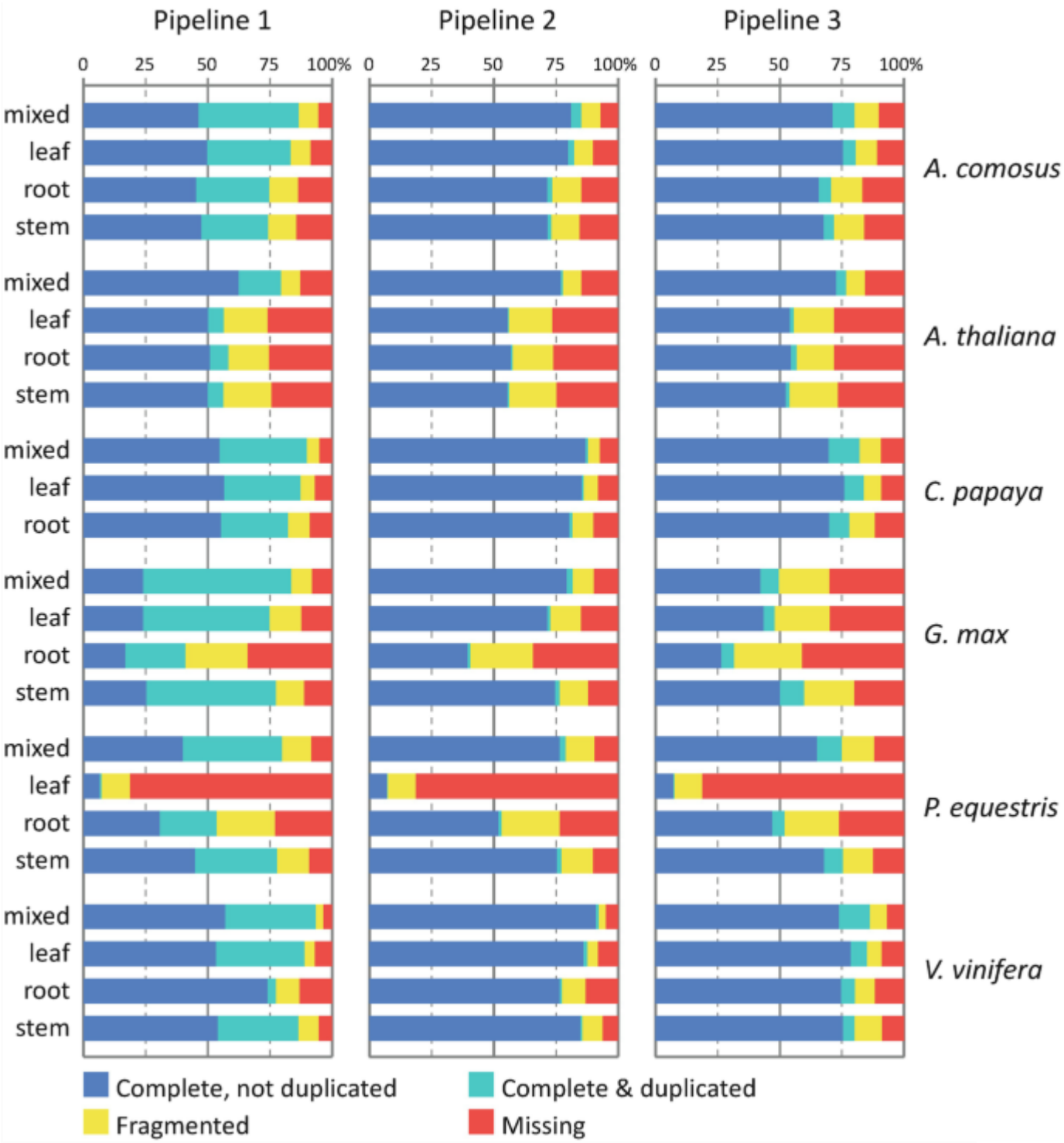

Fig. 1 BUSCO evaluations of predicted ORFs by the three different pipelines for various tissues.

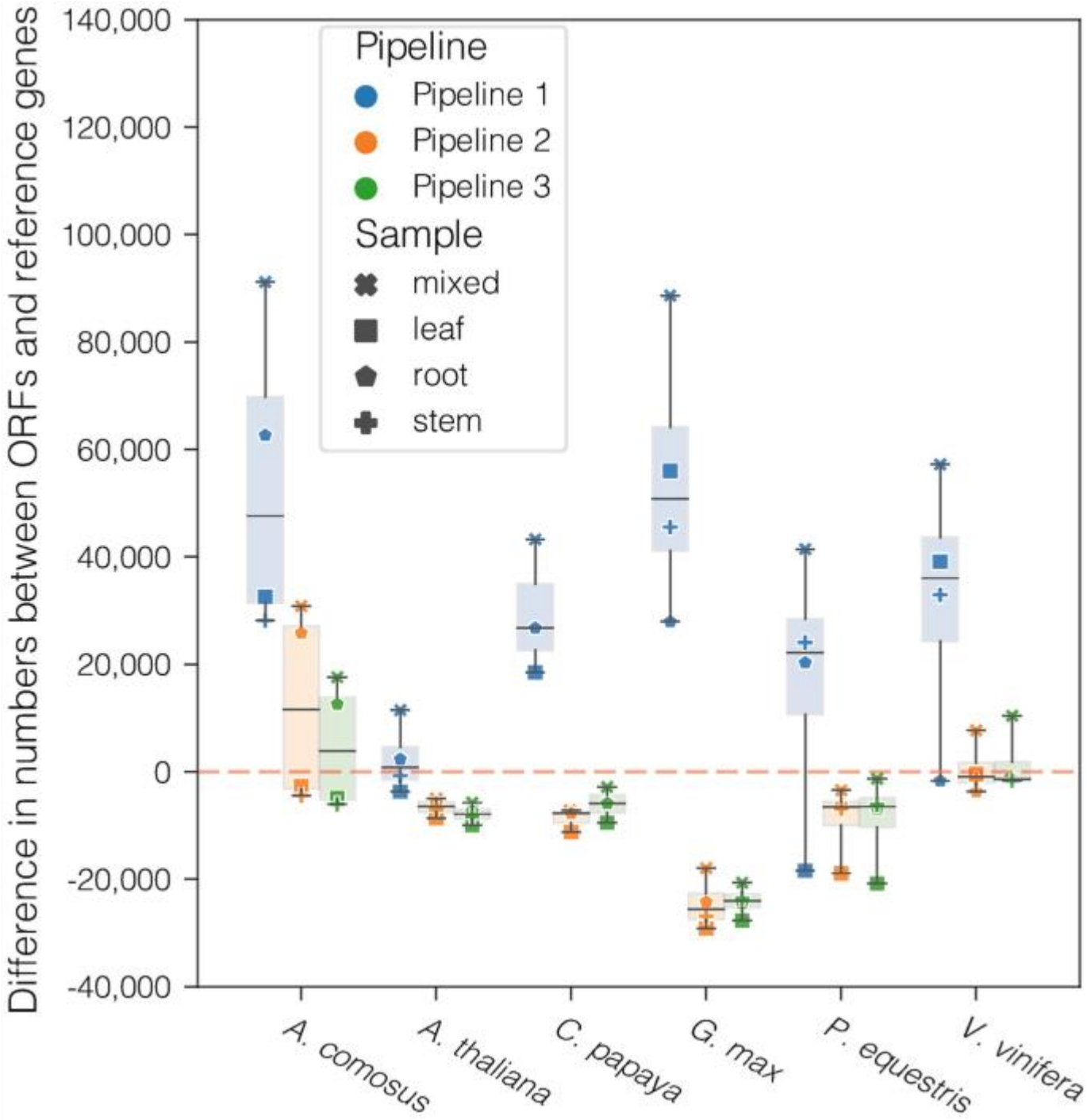

Fig. 2 Differences in numbers between the predicted ORFs and reference genes. The red dashed line at 0 shows no difference in numbers. A dot above the red dashed line means that the number of predicted ORFs exceeds the number of reference genes. A dot below the red dashed line means the number of predicted ORFs is less than the number of reference genes.

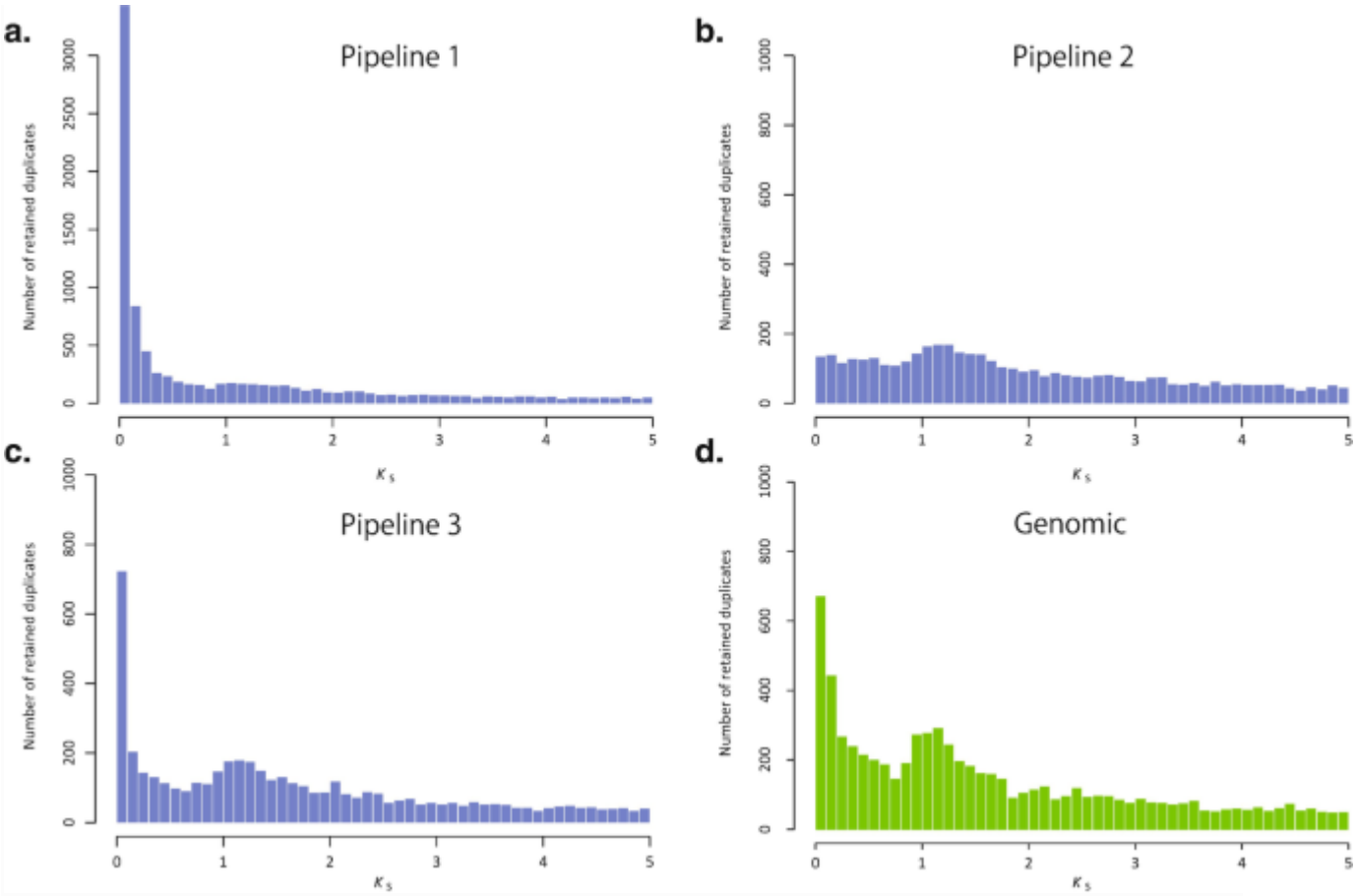

Fig. 3 K_S distributions for the whole paranome of *Vitis vinifera*. (**a**–**c**) The K_S distributions are based on the ORFs predicted by the three transcriptome assembly pipelines with the mixed sample in *V. vinifera*. (**d**) The K_S distribution based on the reference genes from the *V. vinifera* genome.

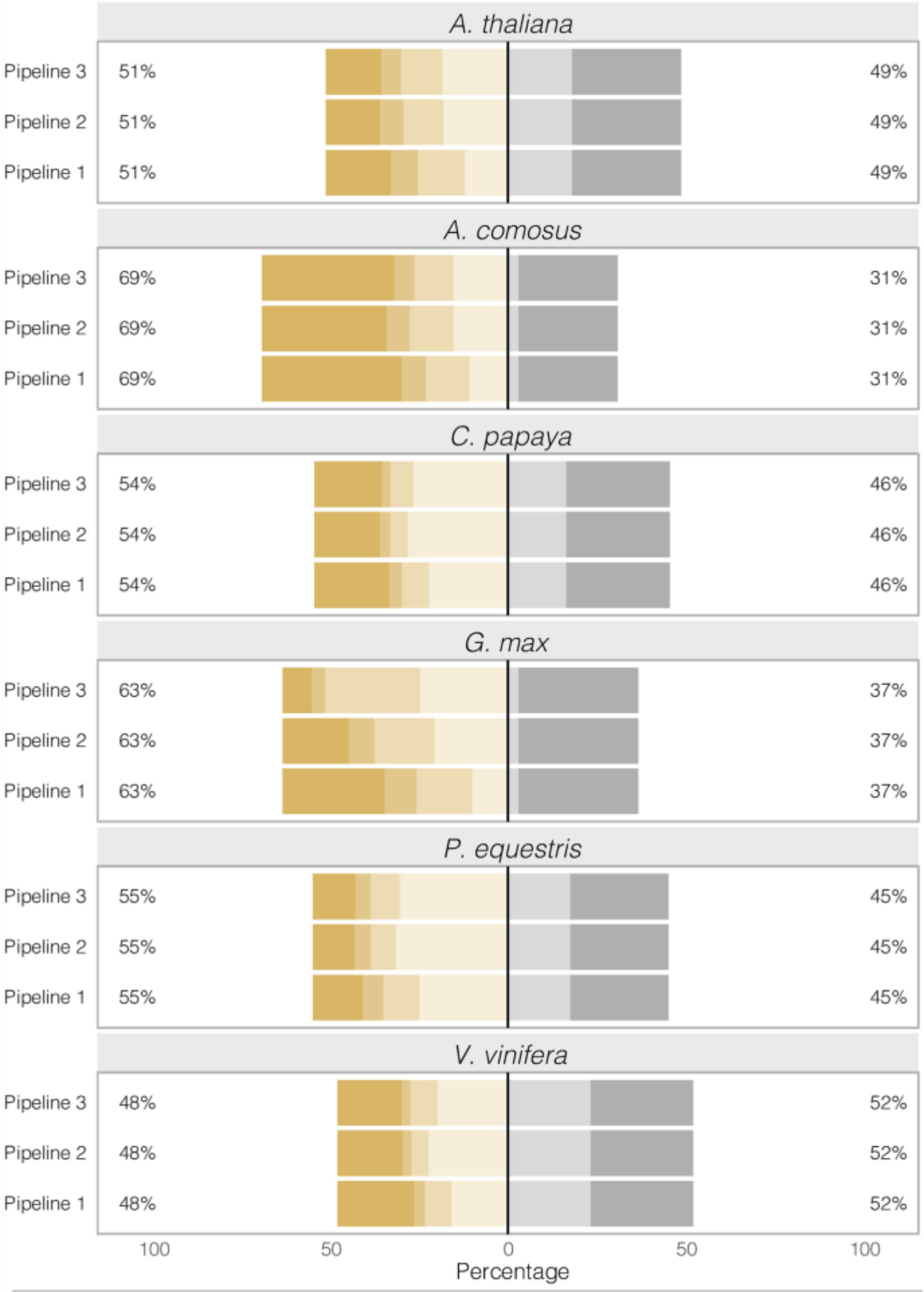
A. thaliana
Pipeline 3
51%
49%
Pipeline 2
51%
49%
Pipeline 1
51%
49%
A. comosus
Pipeline 3
69%
31%
Pipeline 2
69%
31%
Pipeline 1
69%
31%
C. papaya
Pipeline 3
54%
46%
Pipeline 2
54%
46%
Pipeline 1
54%
46%
G. max
Pipeline 3
63%
37%
Pipeline 2
63%
37%
Pipeline 1
63%
37%
P. equestris
Pipeline 3
55%
45%
Pipeline 2
55%
45%
Pipeline 1
55%
45%
V. vinifera
Pipeline 3
48%
52%
Pipeline 2
48%
52%
Pipeline 1
48%
52%
100
50
0
50
100
Percentage

Fig. 4 Reference genes supported by RNA-Seq reads and assembled ORFs, based on the mixed sample in each species. The left part shows the percentages of reference genes supported by at least two RNA-Seq reads (see details in the main text). Assembled ORFs have different coverages for reference genes: “Complete (100%),” “Nearly complete (≥90%),” and “Fragmented (<90%).” The reference genes supported by RNA-Seq reads but without any assembled ORFs are shown as “Unassembled.” The right part shows the percentages of reference genes that are not supported “No read,” or only supported by “One read” in the mixed samples.

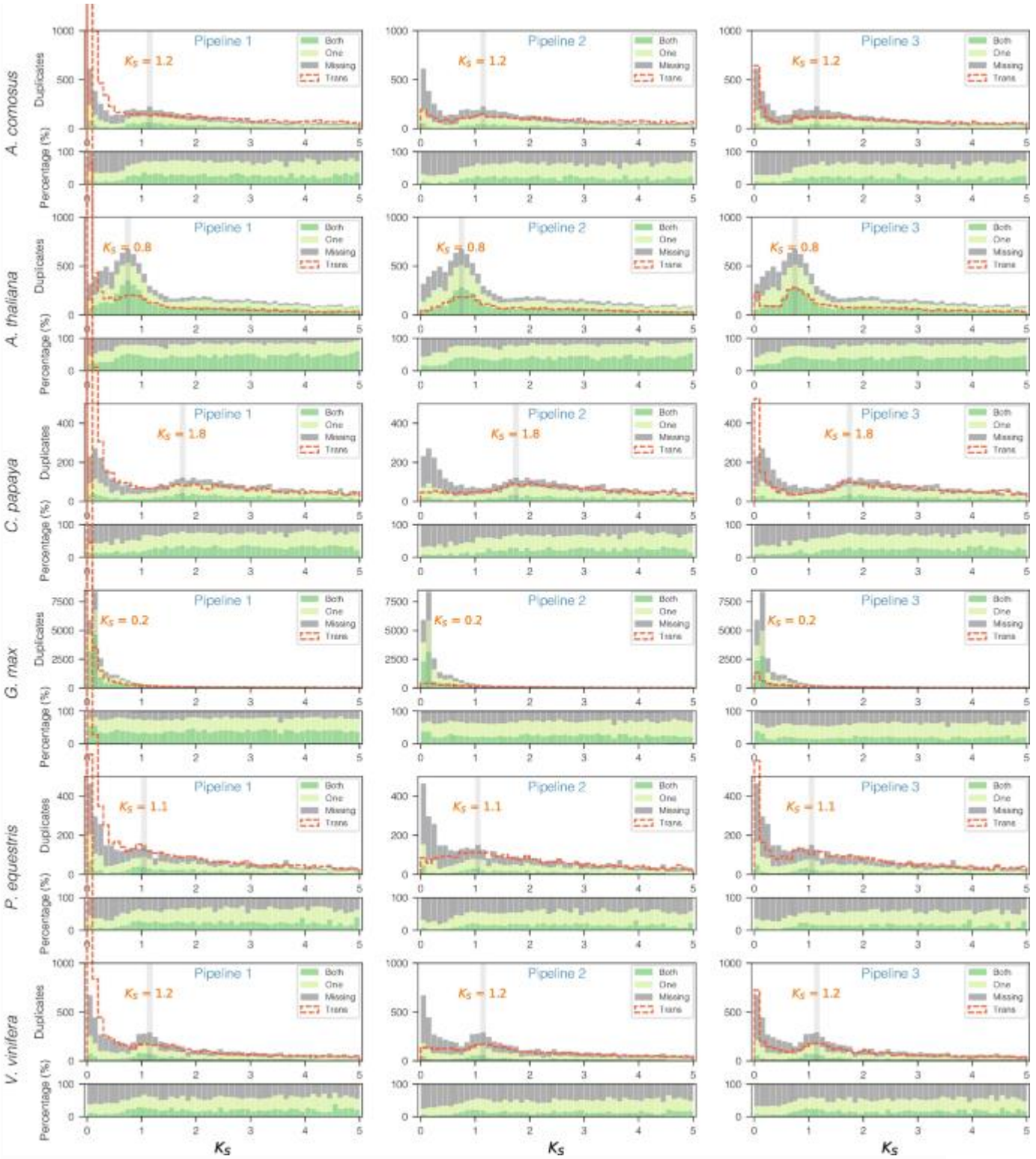

Fig. 5 The impact of the reconstructed gene space on genome-based K_S distributions. "Both" means that the two reference genes of one paralogous pair are present in the predicted ORFs. "One" means that only one of the two reference genes of one paralogous pair is present in the predicted

ORFs. “Missing” means neither of the two genes of one paralogous pair is present in the predicted ORFs. The upper part of each subplot shows a genome-based K_s distribution, and the lower part shows the percentages of the three groups at each K_s interval. The red dashed line denotes the transcriptome-based K_s distribution. The gray rectangle denotes the K_s peak of each species.

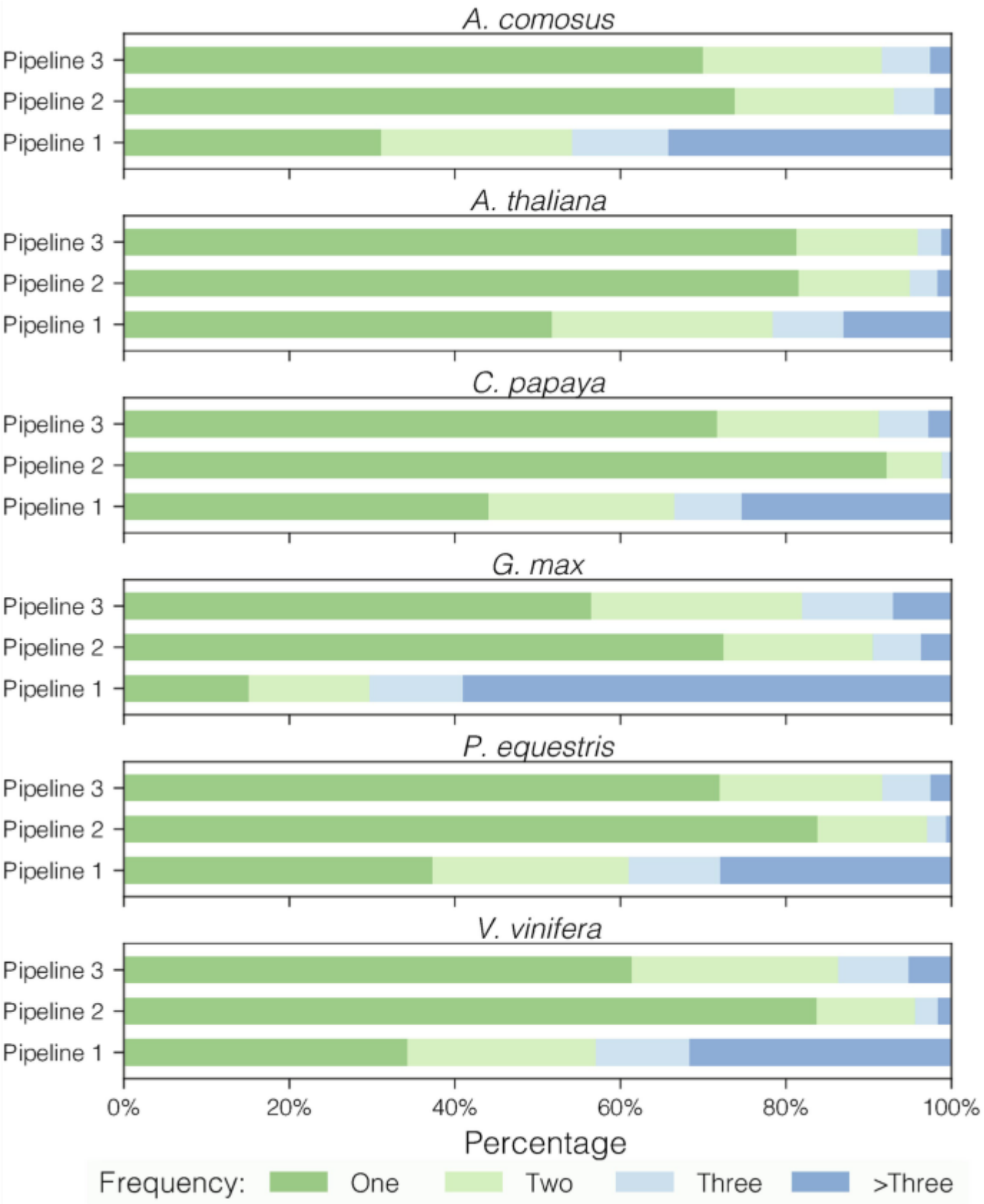
A. comosus
A. thaliana
C. papaya
G. max
P. equestris
V. vinifera
Pipeline 3
Pipeline 2
Pipeline 1
0%
20%
40%
60%
80%
100%
Percentage
Frequency:
One
Two
Three
>Three

Fig. 6 The redundancy of ORFs at each gene locus. “Frequency” refers to the number of ORFs that could be mapped to each gene locus. “One” means that a gene locus only has one ORF mapped. “Two” means that a gene locus has two ORFs mapped. “Three” means that a gene locus has three ORFs mapped. “>Three” means that a gene locus has more than three ORFs mapped.

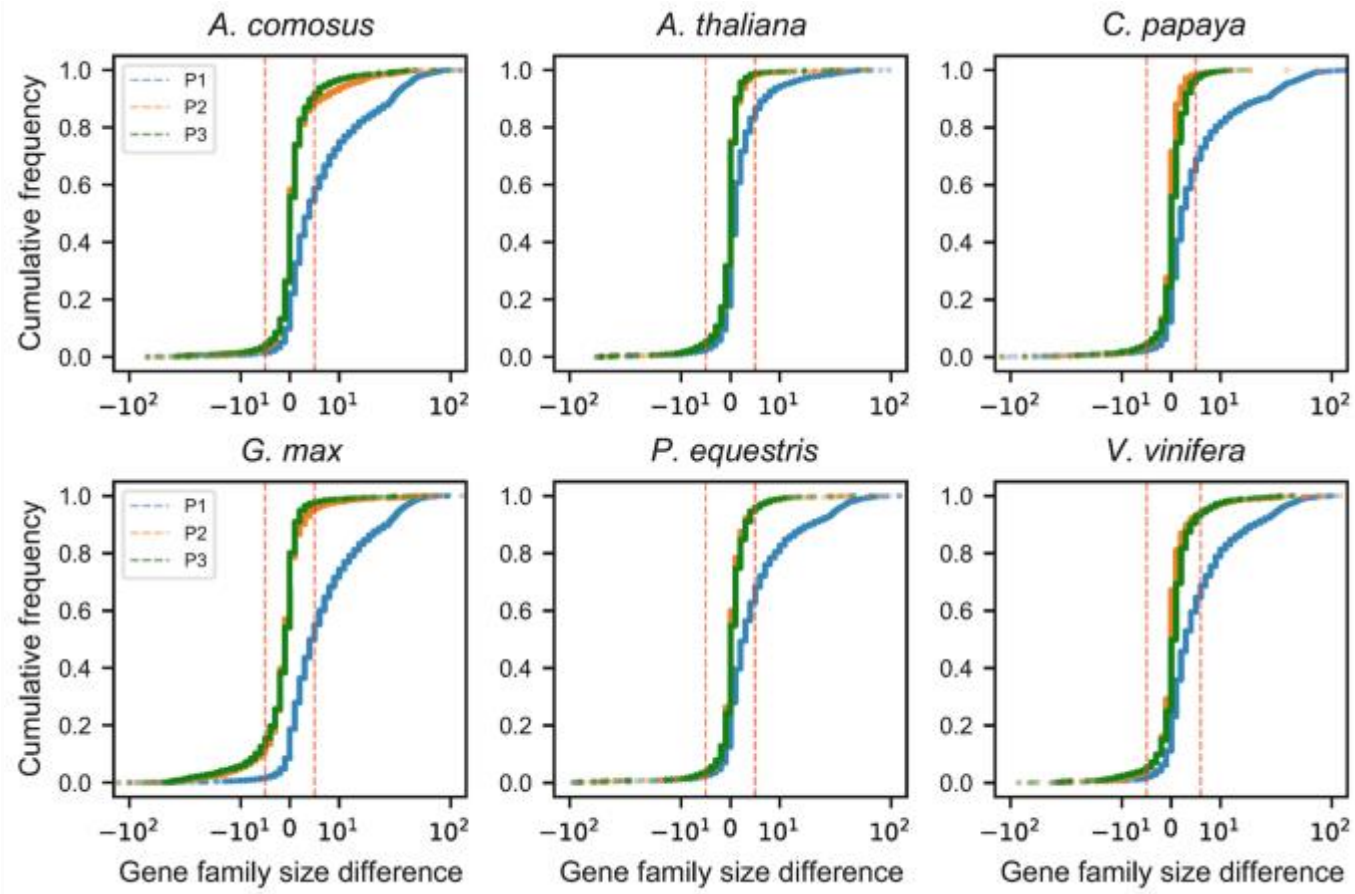

Fig. 7 Cumulative distributions for size differences of gene families between the corresponding transcriptome-based and genome-based gene families. The vertical dashed lines in each subplot correspond to the sizes differences with *z*-scores of −2 and 2, respectively (see details in the main text).

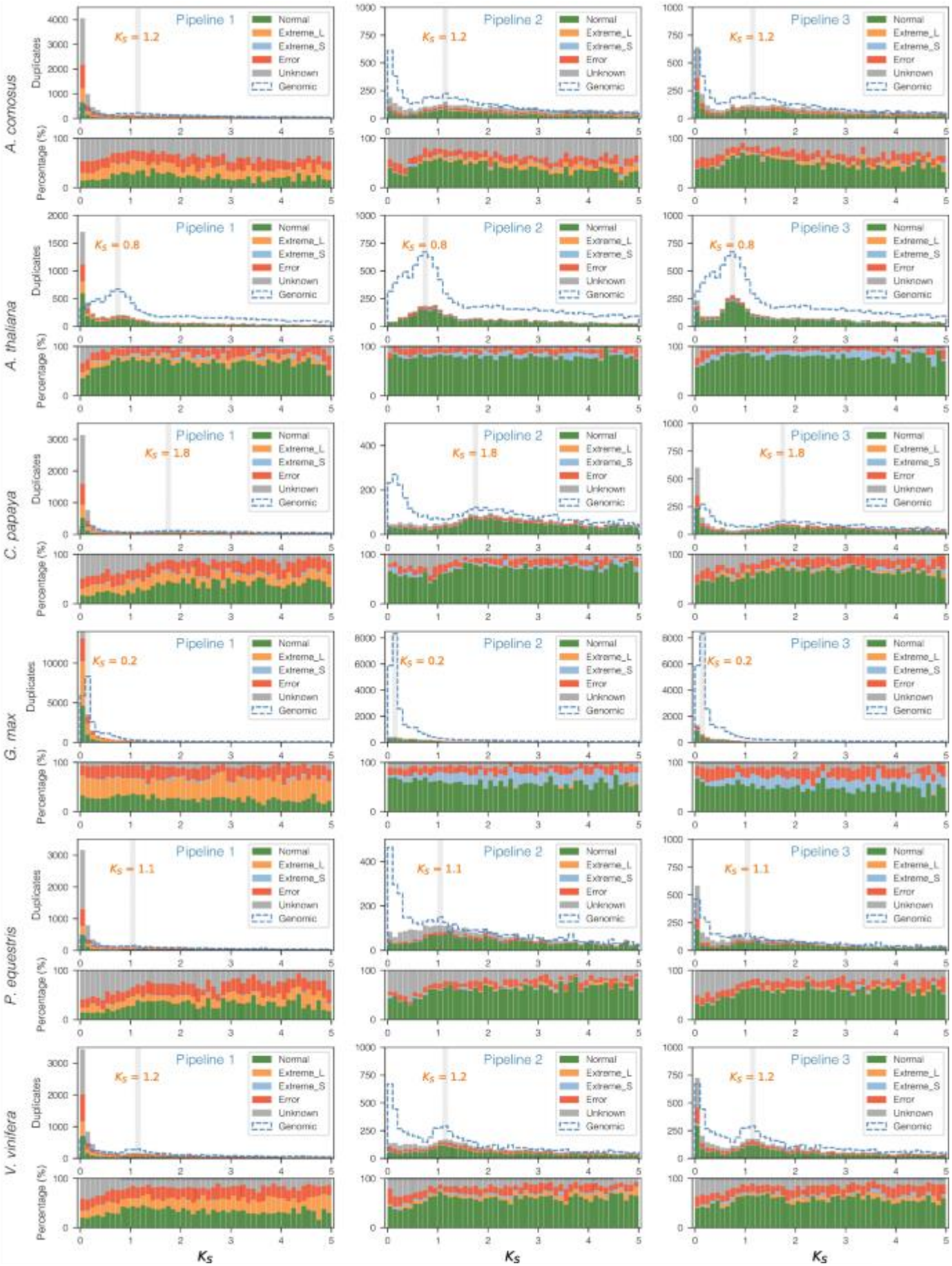
A. comosus
A. thaliana
C. papaya
G. max
P. equestris
V. vinifera
Pipeline 1
Pipeline 2
Pipeline 3
Ks = 1.2
Ks = 0.8
Ks = 1.8
Ks = 0.2
Ks = 1.1
Normal
Extreme_L
Extreme_S
Error
Unknown
Genomic
Duplicates
Percentage (%)
Ks

Fig. 8 The impacts of transcriptome-based gene families with extreme sizes on the transcriptome-based K_s distributions. Gene families have extreme sizes if they are significantly larger or smaller than their corresponding genome-based gene families (see details in the main text). The upper part of each subplot shows a transcriptome-based K_s distribution. The lower part shows the percentages of the different kinds of gene families at each K_s interval. The blue dashed line shows the genome-based K_s distribution. The gray rectangle denotes the K_s peak of each species.

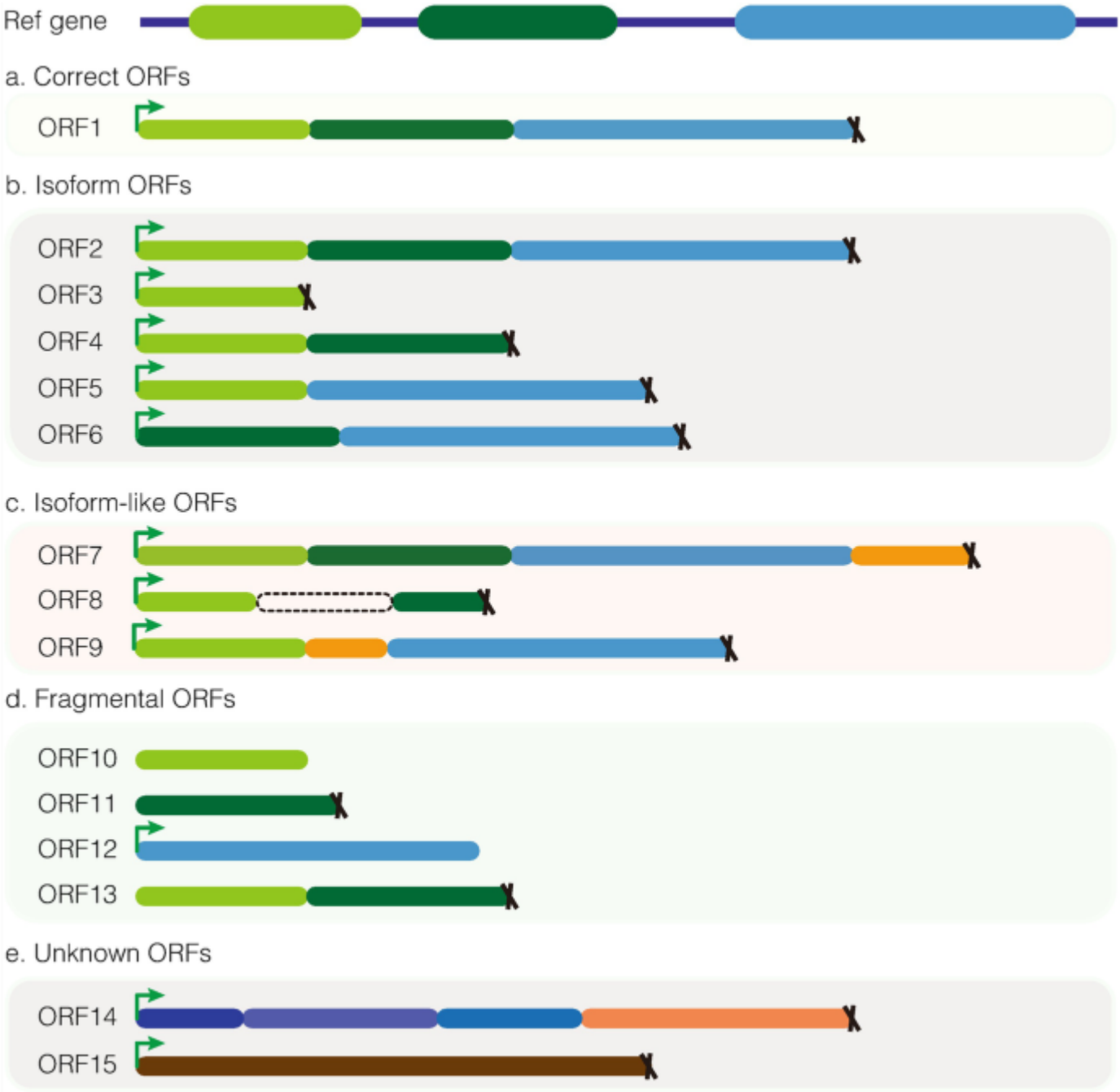

Fig. 9 The classification of assembled ORFs. "ORF1–15" are examples of predicted ORFs mapped to their corresponding reference gene. Five distinct groups of ORFs, including "Correct," "Isoform," "Isoform-like," "Fragmented," and "Unknown" have been depicted. The green arrow denotes a start codon, and the black cross denotes a stop codon.

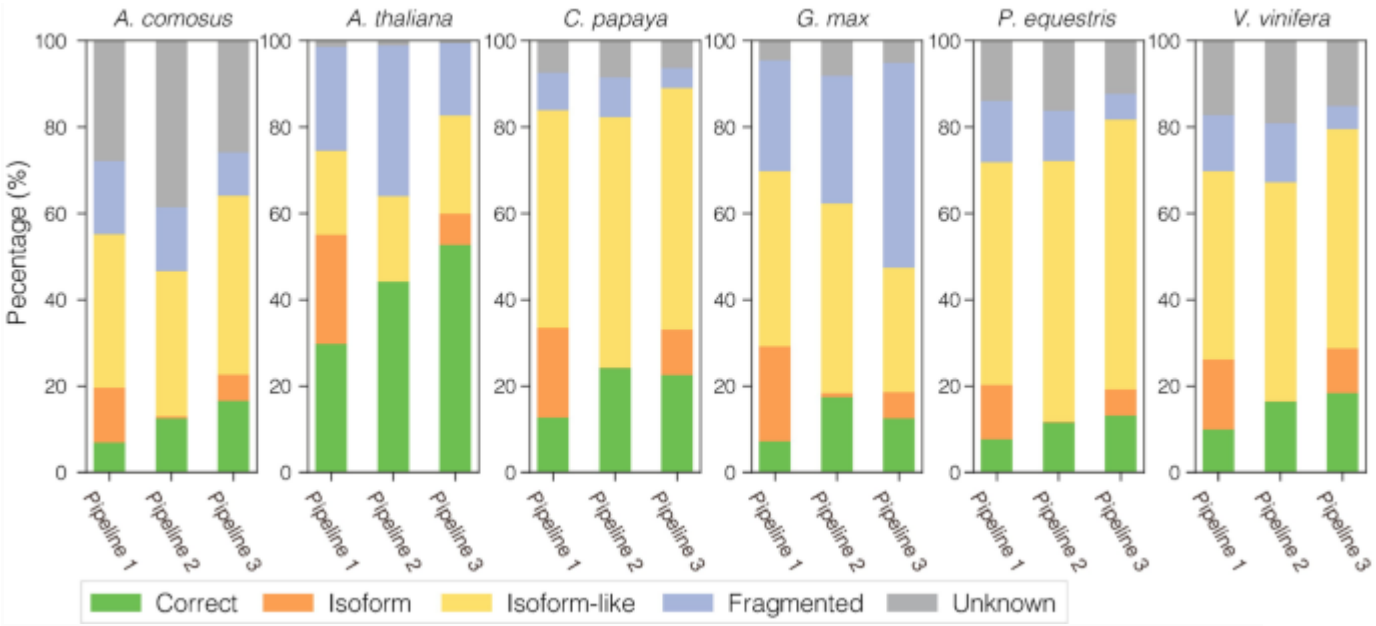

Fig. 10 The percentage of different groups of ORFs in each pipeline.

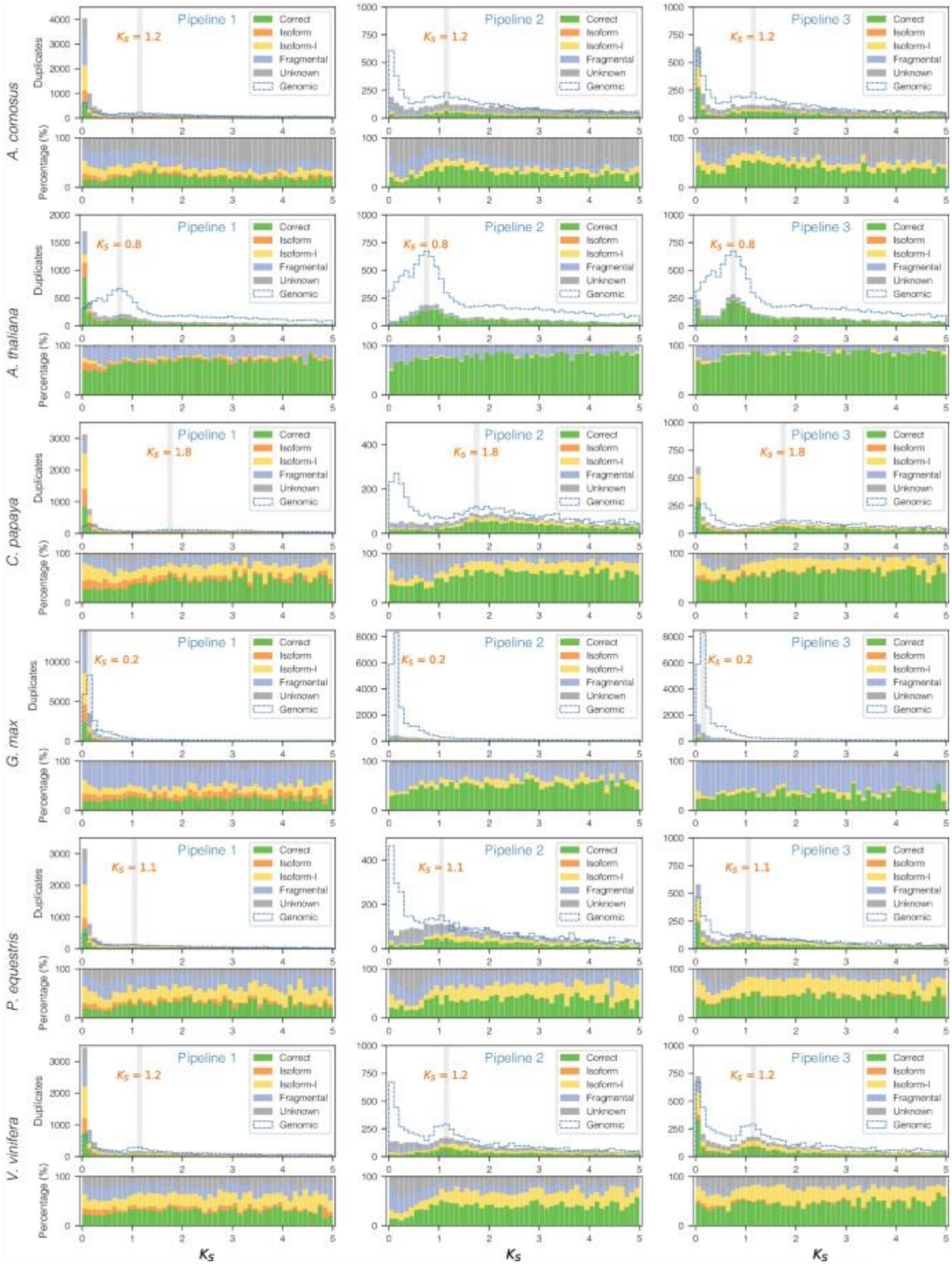
A. comosus
A. thaliana
C. papaya
G. max
P. equestris
V. vinifera
Pipeline 1
Pipeline 2
Pipeline 3
Ks = 1.2
Ks = 0.8
Ks = 1.8
Ks = 0.2
Ks = 1.1
Correct
Isoform
Isoform-I
Fragmental
Unknown
Genomic
Duplicates
Percentage (%)
Ks

Fig. 11 The impacts of various categories of ORFs on transcriptome-based K_s distributions. Five different groups of ORFs are assigned to a transcriptome-based K_s distribution (the upper part of each subplot). The lower part of each subplot shows the percentages of the five groups of ORFs at each K_s interval. The dark blue dashed line depicts the genome-based K_s distribution. The gray rectangle denotes the K_s peak of each species. “Isoform-l” means the “Isoform-like ORFs” group.

Table 1 Summary of the examined plants in this study

Species[a]	Genome size (Mb)	Gene number	RNA-Seq data		
			Tissue	SRA ID	Size (Gb)
Ananas comosus	316	25,440	Leaf	SRR7663722	4.45
			Root	SRR7663721	4.40
			Stem	SRR7663702	4.23
Phalaenopsis equestris	1086	29,431	Leaf	SRR2080202	1.20
			Root	SRR2080194	4.49
			Stem	SRR2080200	5.95
Arabidopsis thaliana	120	27,655	Leaf	SRR3993754	1.15
			Root	SRR3993762	1.28
			Stem	SRR3993761	1.07
Carica papaya	343	27,768	Leaf	SRR7145703	6.38
			Root	SRR7145705	7.15
Glycine max	978	56,044	Leaf	SRR12744739	6.76
			Root	SRR12744729	7.02
			Stem	SRR12744731	6.75
Vitis vinifera	486	26,346	Leaf	SRR9970452	8.53
			Root	ERR3814249	5.83
			Stem	SRR9970442	8.47

1. [a]Genome sequences and annotations were downloaded from PLAZA 4.5, except the one of *Ananas comosus*, which was retrieved from NCBI with PRJEB33121

Table 2 The number of predicted ORFs

Species	Samples	Pipeline 1	Pipeline 2	Pipeline 3
Ananas comosus				
	Mixed	116,585	56,267	43,024
	Leaf	58,009	22,781	20,507
	Root	88,084	51,251	38,035
	Stem	53,561	20,986	19,400
Arabidopsis thaliana				
	Mixed	39,126	22,651	21,887
	Leaf	23,951	18,995	17,699
	Root	29,981	21,855	20,280
	Stem	26,922	20,601	19,262
Carica papaya				
	Mixed	71,012	20,606	24,916
	Leaf	46,262	16,515	18,302
	Root	54,506	20,002	21,900
Glycine max				
	Mixed	144,687	38,102	35,400
	Leaf	112,034	26,867	28,361
	Root	83,968	31,817	31,592
	Stem	101,569	29,133	32,485

Species	Samples	Pipeline 1	Pipeline 2	Pipeline 3
Phalaenopsis equestris				
	Mixed	70,865	25,982	28,129
	Leaf	11,025	10,510	8624
	Root	49,741	23,205	22,562
	Stem	53,474	22,504	23,313
Vitis vinifera				
	Mixed	83,542	34,044	36,757
	Leaf	65,442	25,969	24,692
	Root	24,634	22,645	25,289
	Stem	59,287	24,866	24,691

Table 3 The number of gene families in different species obtained by different pipelines

		Pipeline 1		Pipeline 2		Pipeline 3	
Species	**Reference**	**Number of gene families**	**Ratio (%) of recovered reference gene families**	**Number of gene families**	**Ratio (%) of recovered reference gene families**	**Number of gene families**	**Ratio (%) of recovered reference gene families**
Ananas comosus	4242	12,127	72.63	6399	59.05	6188	61.83
Arabidopsis thaliana	4683	5755	71.34	3358	63.25	3652	64.89
Carica papaya	3120	7512	76.41	2802	66.73	4645	71.38
Glycine max	10,979	13,285	74.46	5048	36.78	5989	37.96
Phalaenopsis equestris	3136	8740	74.14	3211	60.40	4763	66.14
Vitis vinifera	3680	9309	80.54	3880	67.12	5822	73.34