







# Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids

Guanjing Hu , Corrinne E. Grover , Mark A. Arick II , Meiling Liu , Daniel G. Peterson  and Jonathan F. Wendel 

Corresponding authors: Corrinne E. Grover, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA. Tel: +1-515-294-7098. E-mail: corrinne@iastate.edu; Jonathan F. Wendel, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA. Tel: +1-515-294-7172. E-mail: jfw@iastate.edu

## Abstract

Polyploidy is a widespread phenomenon throughout eukaryotes. Due to the coexistence of duplicated genomes, polyploids offer unique challenges for estimating gene expression levels, which is essential for understanding the massive and various forms of transcriptomic responses accompanying polyploidy. Although previous studies have explored the bioinformatics of polyploid transcriptomic profiling, the causes and consequences of inaccurate quantification of transcripts from duplicated gene copies have not been addressed. Using transcriptomic data from the cotton genus (*Gossypium*) as an example, we present an analytical workflow to evaluate a variety of bioinformatic method choices at different stages of RNA-seq analysis, from homoeolog expression quantification to downstream analysis used to infer key phenomena of polyploid expression evolution. In general, EAGLE-RC and GSNAP-PolyCat outperform other quantification pipelines tested, and their derived expression dataset best represents the expected homoeolog expression and co-expression divergence. The performance of co-expression network analysis was less affected by homoeolog quantification than by network construction methods, where weighted networks outperformed binary networks. By examining the extent and consequences of homoeolog read ambiguity, we illuminate the potential artifacts that may affect our understanding of duplicate gene expression, including an overestimation of homoeolog co-regulation and the incorrect inference of subgenome asymmetry in network topology. Taken together, our work points to a set of reasonable practices that we hope are broadly applicable to the evolutionary exploration of polyploids.

**Key words:** allopolyploid; co-expression gene network; differential expression; homoeolog-specific read partitioning; RNA-seq

Guanjing Hu is a researcher in evolutionary genomics, functional genomics and whole-genome duplication (polyploidization).

Corrinne E. Grover is a researcher in evolutionary genomics, molecular evolution and plant systematics and evolution.

Mark A. Arick II is a researcher in genomics, bioinformatics and computational sciences.

Meiling Liu is a researcher in statistics and bioinformatics.

Daniel G. Peterson is a researcher in genomics, biocomputing and biotechnology.

Jonathan F. Wendel is a researcher in evolutionary genomics, molecular evolution and plant systematics and evolution.

Submitted: 17 December 2019; Received (in revised form): 6 February 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

## Introduction

Comparative transcriptomics has become a widely employed and powerful tool in plant evolutionary biology. Applications are many and diverse, including evolutionary rate estimation [1–3], reconstruction of species relationships [3–5] and the elucidation of co-expression and regulatory changes in gene networks [6, 7]. Next-generation sequencing has facilitated inexpensive and efficient transcriptomic profiling for species whose lack of existing genomic resources would have previously been an obstacle. A landmark example is the recent publication of transcriptomes from more than 1000 species of green plants, which substantially improved available resources and facilitated comparative transcriptomics and phylogenetics among previously underrepresented plants [8] ([www.onekp.com](http://www.onekp.com)). This success led to the 10KP project (<https://db.cngb.org/10kp/>), which aims to sequence 10 000 plant and protist genomes within the next 5 years to further advance our understanding of plant evolution and diversity.

In the context of comparative transcriptomics, polyploid genomes offer unique challenges due to the coexistence of highly similar duplicated genes (homoeologs). Polyploidy in plants is far more prevalent than once thought, acting historically and more recently to shape the genomes of all angiosperms and most other groups of plants [8–11]. One realization that has emerged in the last decade is that polyploidy is accompanied by massive transcriptomic responses, as reviewed [12–14]. These responses are many and varied, including biased homoeolog expression, condition-specific differential homoeolog usage, transgressive expression levels and expression level dominance. Duplicated gene expression patterns are coordinated in ways that are not fully understood and which depend on myriad factors, including dosage effects, gene balance, interactions among divergent *cis*- and *trans*-acting factors and various topological aspects of gene networks [6, 15–18].

Research on polyploid transcriptomes is divided into two broad categories with respect to the treatment of homoeologs: those that evaluate individual homoeolog expression separately and those that evaluate the aggregate expression of homoeologs. The ability to consider homoeologs separately depends largely upon the mode of origin (autopolyploid versus allopolyploid), the number of subgenomes and the extent of sequence divergence between homoeologs, as well as the genomic resources available. Distinguishing individual homoeolog expression levels is difficult when sequence divergence between homoeologs is too low, as often is the case with allopolyploids formed from recently diverged diploid parents or in evolutionarily young autopolyploids. When a reference genome or transcriptome is available for a polyploid, quantitation of individual homoeolog expression levels is possible if sequence divergence is sufficiently high, and aggregated expression can be derived from the summation of each homoeolog set. In many cases, reference genomes may only be available for one or more model diploids. These diploid genomes can be useful in analyses of duplicate gene expression in polyploids, but they require additional steps to characterize and partition homoeolog-specific reads. Regardless of the ploidy level of the reference genome, short RNA-seq reads may be difficult to explicitly map to individual homoeologs due to their near-duplicate nature (i.e. multi-mapped reads). That is, only a certain proportion of reads (related to divergence between homoeologous genomes) will contain homoeolog distinguishing variants (e.g. single-nucleotide polymorphisms [SNPs]). Only those reads that can be unambiguously assigned to specific homoeologs can be utilized for homoeolog transcript counting (Figure 1A).

As previously noted by Ilut et al. [19], the issue of ambiguous read mapping is prevalent in plants due to their natural genomic redundancy and is even more so for recent and/or higher-order polyploids. Many intrinsic and extrinsic factors affect the ability to partition homoeolog expression, including (1) the divergence between subgenomes, in terms of frequency and distribution of SNPs; (2) the number of subgenomes; (3) the sequencing strategy (e.g. read length and paired- versus single-ended reads) for generating RNA-seq data; (4) the quality of reference genome(s) and (5) the bioinformatic tools used for partitioning and/or quantifying homoeolog-specific reads, including methods for allocating ambiguous reads in general (such as RSEM [20] and Salmon [21]) and those specifically developed for polyploid systems (i.e. PolyCat [22], PolyDog [23], SniPloid [24], HyLiTE [25], HANDS/HAND2 [26, 27], HomoeoRoq [28] and EAGLE-RC [29, 30]).

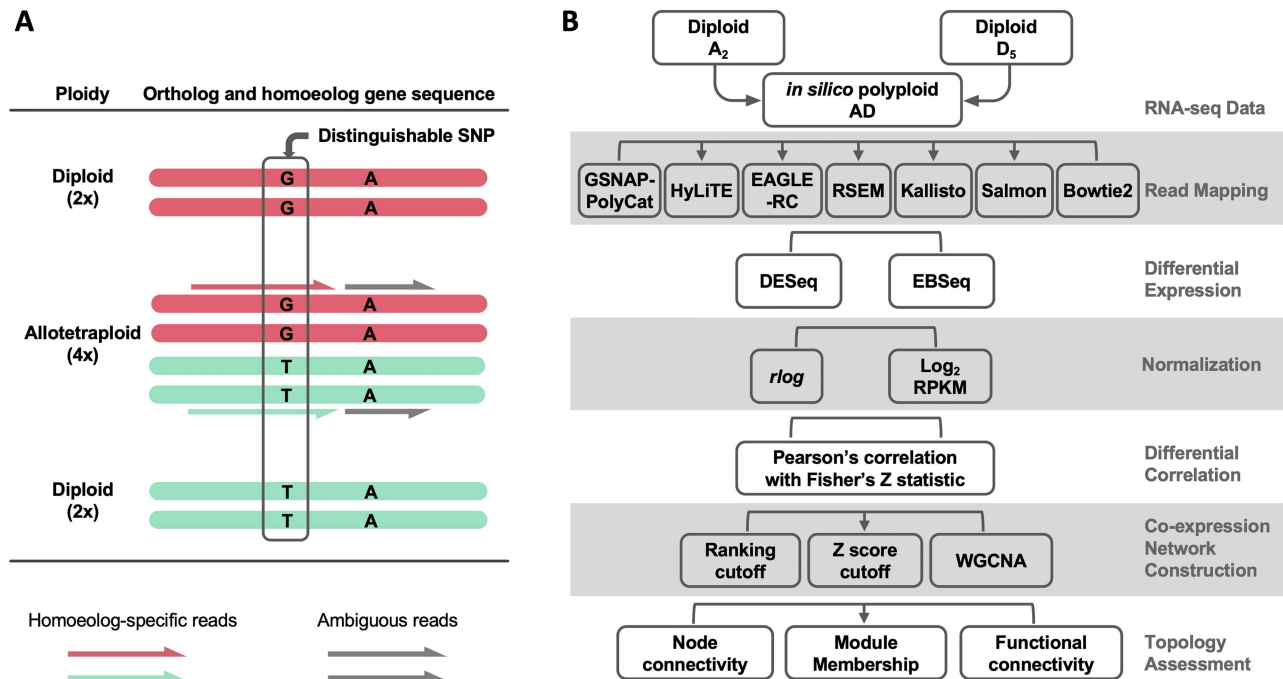
Given these complexities inherent in working with polyploid transcriptome data, the question arises as to how these factors affect our ability to derive accurate polyploid gene expression profiles. That is, how do the many issues noted above affect read assignment and our inferences of gene expression and co-expression characteristics? Here, we explore the causes and consequences of read ambiguity in homoeologous differential expression (DE) and co-expression networks using transcriptome data from the cotton genus (*Gossypium*) as an example (Figure 1B). Tetraploid cotton (represented here by *Gossypium hirsutum*; AD<sub>1</sub>) originated from an allopolyploidization event between an A-genome (*Gossypium herbaceum*- or *Gossypium arboreum*-like) and a D-genome (*Gossypium raimondii*-like) diploid species circa 1 to 2 million years ago (reviewed in [31]). Because there is no gold standard for true expression levels of At and Dt (t denotes subgenome) homoeologs in the polyploid AD<sub>1</sub> transcriptomes, we generated *in silico* allopolyploid datasets (AD) by combining reads from the A- and D-genome diploid transcriptomes (see Methods). This approach allowed us to evaluate the accuracy of 'homoeolog' expression against the actual diploid reads used for generating *in silico* dataset. Methodologically, we first evaluated a variety of bioinformatic method choices at different stages of RNA-seq data analysis, with the aim of generating insight into best practices that may be broadly applicable to other polyploid systems.

## Methods

All codes used in this study are available in GitHub <https://github.com/Wendellab/homoeologGeneExpression-Coexpression>. The R environment 3.5.0 was used for statistical analyses.

## Data availability and preparation

For generating *in silico* allopolyploid cotton (AD) datasets, matched RNA-seq data of the model diploid progenitors, i.e. *G. arboreum* (A<sub>2</sub>) and *G. raimondii* (D<sub>5</sub>), were obtained, each comprising 33 RNA-seq libraries under 12 sample conditions (see Supplementary Table S1 available online at <https://academic.oup.com/bib>). The seed dataset under the National Center for Biotechnology Information (NCBI) BioProject PRJNA179447 consists of 11 libraries (four seed developmental stages each with 2–3 biological replicates) for each diploid with 100 bp single-end reads and an average of 14.8 million reads per library. The flowering dataset under the NCBI BioProject PRJNA529417 consists of 22 libraries (eight tissues each with 2–3 biological replicates for each diploid) that were constructed with 300 bp insert size and sequenced with 150 bp paired-end reads and an average of 13.8 million read pairs per library.



**Figure 1.** Challenges of homoeolog gene expression analysis. (A) Using the allotetraploid cotton species as an example, only a small portion of RNA-seq reads contain diagnostic SNPs (i.e. homoeolog-specific reads) reflecting the parental origin of homoeologous genes. (B) An analytic workflow of RNA-seq analysis was applied to evaluate the use of homoeolog-specific reads to study duplicated gene expression and co-expression networks. A ground-truth, *in silico* dataset of allopolyploid cotton (AD) was generated from the parental diploid cotton A<sub>2</sub> and D<sub>5</sub> reads, which was analyzed using a variety of method choices.

Following adaptor and quality trimming via Sickle [v1.33] [32], the matched A<sub>2</sub> and D<sub>5</sub> libraries (at each condition and replicate) were adjusted to contain equivalent number of filtered reads and subsequently combined to generate the corresponding *in silico* allopolyploid (AD) datasets. For each pair of AD homoeologous genes, the gene regions that should be unambiguously assigned to subgenome (i.e. effective region), given the distinguishable variant distribution (in SNP index or variant candidate file [VCF] format) between homoeologs and the specific sequencing strategy involved, were detected using a custom script 'detectEffectiveRegion.r'. The proportion of each gene sequence that belongs to an effective region was calculated as %Eflen. We next introduced a metric of 'Ambiguity' for each pair of homoeologous genes as calculated by 1-%Eflen, because %Eflen is inversely correlated with the number of ambiguous reads that cannot be assigned via direct variant evidence.

### RNA-seq read mapping and homoeolog-specific read partitioning

The following seven pipelines (see [Supplementary Table S2](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>) were each independently applied to the diploid and AD polyploid datasets. Although newer versions of these programs are continuously being released, new versions whose fundamental mode of action is similar should yield comparable results, which can be verified through the testing presented here. Quantification of transcripts was conducted at the gene level based on the annotation of primary transcripts only.

#### GSNAP-PolyCat

This pipeline utilizes the SNP-tolerant capabilities of GSNAP [v2016-08-16] [33] to map polyploid reads to a single diploid

progenitor genome (here, *G. raimondii* [34]). The SNP-tolerant feature of GSNAP permits equivocal mapping of both A- and D-diploid-derived reads based on a priori SNP information, which may be extensive for some polyploid systems. Here, we used a previously generated genome-diagnostic SNP index [22] for mapping. The resulting alignments were sorted using SAMtools [v1.9] [35] and subsequently partitioned into homoeolog-specific reads using PolyCat [v1.3] [22]. Read counts were tabulated using HTSeq [v0.9.1] [36].

#### HyLiTE

This software automates the process of read mapping, SNP detection and read count partitioning in a single step [25]. Briefly, HyLiTE [v2.0.1] under Python 3 [v3.6.3] employs Bowtie2 [v2.3.4] [37] to map both diploid and polyploid reads to the reference gene models and sorts homoeologous reads based on the SNPs detected from mapping the diploid reads. This pipeline was separately tested using the D-genome *G. raimondii* [34] and the A-genome *G. arboreum* [38] references. Because HyLiTE performs 'on-the-fly' homoeoSNP identification using only the data contained within the analysis, read sorting is made possible for species with limited resources; however, the homoeoSNPs identified by HyLiTE are limited to those contained within the dataset under consideration even if additional information is available in other datasets. The final step of the pipeline automatically generates homoeolog-specific read count tables.

#### EAGLE-RC

In contrast to using a single subgenome or diploid progenitor genome as reference, this pipeline applies a subgenome-classification approach to assign reads to their origin based on

mapping statistics against each subgenome of the polyploid system. Reads were mapped against *G. raimondii* [34] and *G. arboreum* [38] reference genomes using STAR [v2.5.3a]. The variant candidate files with respect to each reference genome were generated from reciprocal LAST [v869] (<http://last.cbrc.jp/>) alignments as part of the EAGLE-RC pipeline. Based on these mapping results and subgenome-discriminating variants, the basic EAGLE model [29] was used to evaluate the likelihood of a read deriving from one subgenome (as the reference) as the null hypothesis versus deriving from the other subgenome as the alternative hypothesis. EAGLE-RC [v1.1.1] next calculated the probability for reads considering each subgenome as the reference and determined the winning hypothesis for read classification.

#### RSEM

While not specifically developed for polyploids, RSEM [v1.3.0] [20] and the following programs (i.e. Salmon [v0.9.1] [21] and Kallisto [v0.44.0] [39]) were developed to address the general issues of ambiguously mapped reads while also increasing mapping speed. RSEM automates read alignment to a set of reference transcripts using Bowtie2 [v2.3.4] and subsequently estimates feature counts using the EM algorithm, both at the gene and isoform levels. As the presence of homoeologs is bioinformatically similar to the presence of alleles of isoforms, RSEM may be suitable for disentangling homoeologous reads and estimating homoeolog abundance. We approximated the polyploid reference transcriptome by combining the *G. raimondii* transcriptome and the predicted *G. arboreum* ( $A_2$ ) transcripts based on the same SNP index used by GSNAP-PolyCat. That is, the *G. arboreum* transcripts here are simply the *G. raimondii* transcripts with homoeologous SNP sites replaced with *G. arboreum*-specific SNPs.

#### Kallisto

This method belongs to a class of read aligners known as ‘pseudoaligners,’ which leverage k-mer information to detect the transcripts that could have generated a given read without specifically aligning the read [39]. Kallisto, like other pseudoaligners, generates a De Bruijn graph of the k-mers present in a transcriptome to quickly assign reads based on intersecting read and transcriptome k-mer metrics. Kallisto was run under default parameters using the above-generated polyploid reference transcriptome.

#### Salmon

This method employs a lightweight, quasi-mapping strategy [40] similar to Kallisto and a two-phase estimation of expression. This two-phase estimation uses two forms of Bayesian inference [41, 42] to first estimate and then subsequently refine transcript-level abundances [21]. Using this method, Salmon is able to estimate abundance uncertainty due to ambiguously mapped reads, which are common with homoeologs. Salmon [v0.9.1] was also run with default parameters using the above-generated polyploid reference transcriptome and the option ‘keepDuplicates’ for indexing the transcriptome. Estimated transcript abundance is automatically returned by the program.

#### Bowtie2

To represent a standard transcriptome-based quantification approach, polyploid reads were mapped against the same polyploid reference transcriptome as above using Bowtie2 [v2.3.4] in—‘local’ mode to report one best alignment by

default. In order to retain only the uniquely mapped reads, alignments were next filtered with a mapping quality of 10 using SAMtools [v0.9.1] [35]. Read counts were generated using the alignment-based mode ‘Salmon quant’ from Salmon [v0.9.1].

#### Performance evaluation of estimating homoeolog expression

For each set of bioinformatically partitioned reads, multiple measures of performance were conducted. Because the true assignment of each *in silico* polyploid (AD) read is known and originates from only two sources ( $A_2$  and  $D_5$ ), assessing homoeolog assignment becomes a binary classification problem. When considering the A-subgenome only, correct assignment of  $A_2$ -derived reads to At is considered a TP (true positive), whereas incorrect assignment to Dt is a FN (false negative); correspondingly, correct assignment of  $D_5$ -derived reads to Dt is a TN (true negative), and incorrect assignment to At is a FP (false positive). Similarly, when considering the D-subgenome only, correct assignment of  $D_5$ -derived reads to Dt is a TP, whereas incorrect assignment to At is a FN; correct assignment of  $A_2$ -derived reads to At is a TN and incorrect assignment to Dt is a FP.

The prediction results of the binary classification can be arrayed as a  $2 \times 2$  confusion matrix, which summarizes the numbers of true/false positives/negatives (TP, FP, TN and FN) that can be evaluated using information retrieval statistics [43], such as Precision/Recall [44] and the Matthews correlation coefficient (MCC) [45]. The general formulas of these statistics are as follows:

$$\text{‘Precision’} = \frac{TP}{TP + FP},$$

$$\text{‘Recall’} = \frac{TP}{TP + FN},$$

$$\text{‘Accuracy’} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Here, we report both the  $F_1$  and MCC scores, which provide a generalized measure of accuracy; however, we note that MCC may be preferred because it accounts for more of the confusion matrix and is more balanced with respect to classes of very different sizes [46].

We also note that the results of binary classification measures for GSNAP-PolyCat and HyLiTE are somewhat misrepresentative of those pipelines. Because GSNAP-PolyCat and HyLiTE discard reads with no diagnostic SNPs, the number of TPs and FNs will be distorted for these pipelines, i.e. reduced and increased, respectively. In contrast, the remaining pipelines (i.e. RSEM, Salmon and Kallisto) use statistical inference to completely assign all reads to homoeologs. We therefore define two additional measures that reflect these differences, ‘Efficiency’ and ‘Discrepancy.’ Here, the measure of ‘Efficiency’ is simply the number of reads assigned to homoeologs



Table 1. Nine classes of DC changes

Category	Diploid correlation	Polyploid correlation	Description of DC pattern
P+/+	+	+	Both positive but different in <i>r</i> -value
P+/0	+	0	Loss of positive correlation
P+/-	+	-	Inversion from positive to negative correlation
P0/+	0	+	Gain of positive correlation
P0/0	0	0	Neither significant but different in <i>r</i> -value
P0/-	0	-	Gain of negative correlation
P-/+	-	+	Inversion from negative to positive correlation
P-/0	-	0	Loss of negative correlation
P-/-	-	-	Both negative but different in <i>r</i> -value

(regardless of accuracy) divided by the total number of reads. The overall difference between the obtained read count and expected true read count for each class was measured by their 'Discrepancy'

$$= \frac{\text{abs}(\text{obs} - \text{exp})}{\text{exp}}$$

### Gene expression analysis

The analysis of DE of genes was conducted using two methods, i.e. DESeq2 [v1.22.2] [47] and EBSeq [v1.22.1] [48]. DESeq2 takes a classical hypothesis testing approach to report nominal *P*-values, whereas EBSeq accommodates the uncertainty inherent in isoforms (here, homoeologs) using a Bayesian framework to return posterior probabilities for DE. A false discovery rate  $\alpha < 0.05$  was required to determine significant DE changes, which was applied to the adjusted *P*-values of DESeq2 [49] and the posterior probability ( $= 1 - \alpha$ ) of EBSeq. For each sample condition (tissue type or developmental stage), the pairwise comparison of At and Dt genes was conducted using both DESeq2 and EBSeq. Common to all sample conditions, the homoeolog-specific DE effect was analyzed with DESeq2 using a multifactor design (~ condition + homoeolog).

For evaluation, the DE analyses were conducted between (1) the parental diploid read counts and (2) the subgenomic read counts within the polyploid samples. Because these *in silico* polyploid data were derived from combining diploid libraries, the null hypothesis is that DE between inferred homoeologs should match the DE observed between the diploid libraries for those genes. We again treat this as a binary classification problem, marking each gene as DE or non-DE and comparing the observed number of DE genes in the polyploid libraries with the expected number derived from the diploid data. The same statistical measures of performance (i.e. 'Precision,' 'Recall,' 'Accuracy,'  $F_1$  and MCC) were calculated for each pipeline, as described above. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were calculated for each and visualized using the R package ROCR [v1.0-7] [50]. AUC scores reflect the probability that a random classification is correct, ranging from 0.0 to 1.0 [51, 52].

### Gene expression correlation analysis

Differential correlation (DC) analysis is commonly used to evaluate coordinated changes in gene expression, either independent of or in the context of co-expression network analyses. Both DC and network analyses require some form of variance-stabilizing transformation of the raw data. Several methods of normalization exist [53, 54], which have their own advantages

and nuances. Here two common methods were tested, i.e. RPKM followed by  $\log_2$  transformation and regularized logarithm (rlog) transformation as implemented in DESeq2.

Using the R package DGCA [v1.0.2] [55], Pearson correlation coefficients (*r*) and their corresponding *P*-values were calculated for each pair of genes across all 33 samples, which were subsequently classified as having a significant ( $P < 0.05$ ) positive correlation (+), a significant negative correlation (-) or not significantly different from zero (0). The gene pairwise matrices of *r* and corresponding correlation condition were each generated for the diploid (expected) and homoeologous (observed) gene expression datasets. Fisher's *z*-test [56] was used to identify significant changes between the expected and observed *r* values.

Given that the orthologous gene pair between diploids (expected) and the homoeologous gene pair in the polyploid (observed) each exhibit three possible within-pair correlation conditions (+, - or 0), together there are nine possible categories to describe the pattern of DC (Table 1). Among those, three categories (0/0, +/+ and -/-) indicate significant changes in within-pair *r* values with no corresponding change in the inference of correlation condition, whereas the other categories encompass changes in correlation condition that indicate misidentification of gene-to-gene correlations within the read partitioned dataset. We assessed enrichment of each class for each pipeline using a one-sided Fisher's exact test ( $P < 0.05$ ).

Finally, as previously described [15], we compiled a list of genes that are overrepresented with the gene-to-gene paired DC relationships (see above) to identify differentially co-expressed genes (DC genes). Briefly, the probability *P* of any pair of genes exhibiting a DC relationship is defined as the percent of DC pairs detected among all possible gene pairs. For a gene observed in *k* DC pairs among all possible pairs *n*, the probability of a 'differential co-expression gene' follows the binomial distribution model:

$$P_{DC} = \binom{n}{k} \cdot p^k q^{n-k}$$

$P_{DC}$  was corrected by the BH method [49] and a cutoff of 0.05 was used for identifying DC genes.

### Co-expression network construction

Co-expression networks are a multidimensional representation of the expression relationships among genes. Accordingly, construction of co-expression networks uses similarity scores from the pairwise gene expression profiles to generate an adjacency matrix which reflects connections between genes (as nodes) in the network [57]. Here, we used the Pearson correlation coefficients to calculate the matrix of similarity scores. Derived from

this correlation matrix, the adjacency matrix was used as the basis for a series of binary and weighted gene co-expression networks, which were generated for both the log<sub>2</sub>RPKM- and rlog-transformed read count tables from each expression estimation pipeline.

For constructing binary networks, a hard threshold was applied to similarity scores to determine whether a pair of genes should be connected in the network, resulting an adjacency matrix containing only 0 and 1 values. Two types of hard thresholds were tested, specifically rank-based and Fisher's Z-statistics [56]-based thresholds. A set of rank-based cutoffs (5, 1, 0.5 and 0.1%) were applied to these similarity scores in order to select the top-ranked connections as possible edges. Following Fisher's z-transformations to convert each Pearson correlation coefficient to a Z-score, a set of cutoffs (i.e. 1.5, 2.0, 2.5 and 3.0) were used to retain correlations with Z-score above the cutoff value as edges. The performance of network construction was evaluated as a binary classification problem; that is, because we expected to see the edges inferred from the expected expression (diploid) retained in the polyploid network, we were able to create a confusion matrix from the presence or absence of edges compared to what was expected. The edge classification was again evaluated with a ROC curve using the R package ROC [50]. Due to the large gene number in the network (>60 000 genes), a 10% random sampling of genes was used for computation with 10 iterations.

While binary networks have their own utility, weighted co-expression networks are frequently used for reasons enumerated elsewhere [58], including the ability to quantify network connections. Weighted networks use soft thresholding to assign connection strengths to gene pairs, thereby allowing the adjacency matrix to present network connections quantitatively. Using the R package weighted gene co-expression network analysis (WGCNA) [v1.68] [59, 60], a set of soft thresholds (1, 12, 24) were applied for automatic network construction with the *blockwiseModules* function and the following parameters: *corType* = 'Pearson,' *networkType* = 'signed,' *TOMType* = 'signed' and *minModuleSize* = 100. The performance of each polyploid network construction was evaluated against the reference network generated using the diploid data. Preservation of the reference network modules by AD dataset was calculated using the WGCNA function *modulePreservation* with 200 permutations. In general, modules with the derived preservation score *Z*<sub>summary</sub> > 10 are interpreted as strong preservation.

### Network topology measures and functional connectivity assessment

Node connectivity and functional connectivity (FC) are two metrics that may provide insight into the importance and/or function of a given gene in a network. Node connectivity (*k*) measures the connectivity of any given node in the network, either by counting the number of connected edges (for a binary network) or summing the connected edge weights (for a weighted network). FC uses the 'guilt-by-association' principle to measure network quality under the assumption that genes with similar functions should be connected in a well-constructed network. A neighbor voting algorithm from the R package EGAD [v1.10.0] [61] was used to classify genes into functional groups based on the functionality of their connected genes (i.e. their neighborhood). This package uses the known functional labels of genes (e.g. Gene Ontology [GO] and Kyoto Encyclopedia of Genes and Genomes [KEGG] annotations) and the voting algorithm as a

binary classifier to return true or false predictions for those functional labels; the performance of the neighbor voting functional assignment can then be assessed by an ROC curve. The derived AUC characterizes the degree to which an input network topology can predict the gene membership of a functional category, which intuitively corresponds to the assessment of functional connectivity. GO and KEGG terms were extracted from the v2.1 annotation of *G. raimondii* reference genome downloaded from Phytozome ([www.phytozome.net](http://www.phytozome.net)).

## Results

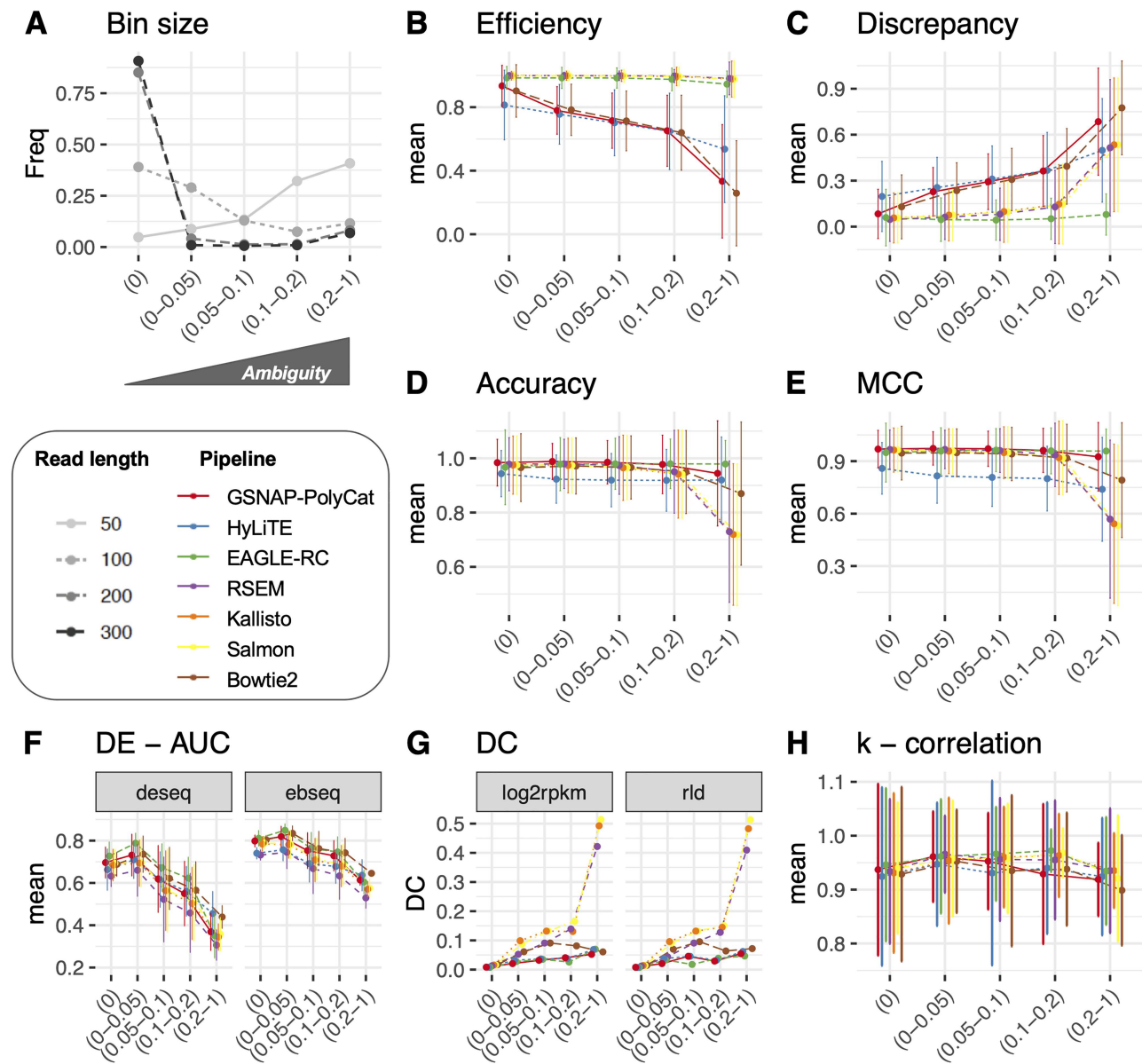
### Subgenome divergence and homoeolog read ambiguity: the problem

As mentioned in the introduction, the issue of ambiguous read mapping is prevalent in polyploids and in plants in general because of means other than polyploidy that generate paralogs. Accurately partitioning polyploid reads is bioinformatically challenging (Figure 1A), and the consequences of inaccurate partitioning are unknown. The proportion of ambiguous reads is dependent both on subgenome divergence and the sequencing strategy, and the subsequent treatment (i.e. removal or statistically based assignment) can affect the outcome of downstream analyses. Here, we evaluated the performance of seven different pipelines (see Supplementary Table S2 available online at <https://academic.oup.com/bib>) in assigning reads to polyploid genomes and the effects of their treatment of ambiguous reads on downstream analyses of duplicated gene expression (Figure 1B). Accordingly, we introduced the metric of 'Ambiguity' for each pair of homoeologous genes, which corresponds to the percentage of a gene region that cannot be distinguished between homoeologs (see Methods). Ideally, if the homoeologous sequences were sufficiently divergent and the sequencing reads were long enough to consistently contain homoeolog distinguishable variants (e.g. SNPs), all reads could be assigned with zero 'Ambiguity'; however, these conditions are rarely met by existing data.

In tetraploid *Gossypium*, where the average sequence divergence (in coding regions) between homoeologs is approximately 1.5% [22], only 5% of homoeologous gene pairs can be unambiguously mapped ('Ambiguity' = 0) by 50 bp RNA-seq reads, whereas over 90% can be unambiguously mapped by 300 bp reads (Figure 2A). In the following analysis, we binned homoeologous gene pairs into five increasing levels of 'Ambiguity,' i.e. 0, 0–0.05, 0.05–0.1, 0.1–0.2 and 0.2–1.0. These bins were next used to relate the performance of read assignment and other duplicated gene expression patterns to the level of read ambiguity (Figure 2).

### Artificial allopolyploid datasets permit assessment of fidelity in homoeologous read assignment

We generated *in silico* allotetraploid (AD) datasets for multiple sample conditions (tissues, developmental timepoints, etc.; see Supplementary Table S1 available online at <https://academic.oup.com/bib>) as a ground-truth reference. For these, we combined equal amounts of reads from two diploids, *G. arboreum* (A<sub>2</sub>) and *G. raimondii* (D<sub>5</sub>), which represent the model diploid progenitors for a clade of naturally occurring polyploids in *Gossypium*. As these datasets are diploid-derived, the amount of A- and D-derived reads in the AD datasets is known, and the ability of each pipeline to accurately reconstruct this becomes testable.



**Figure 2.** Homoeologous read ambiguity and consequences. (A) Given the specific sequencing read length (i.e. 50, 100, 200 and 300 bp), the homoeologous gene pairs from *Gossypium* were binned by 'Ambiguity' into five groups: 0, 0-0.05, 0.05-0.1, 0.1-0.2 and 0.2-1.0, the first of which indicates complete read assignment via SNP differentiation. The y-axis refers to the bin size of each gene group. These 'Ambiguity' bins were used to relate the performance of read assignment (B-E), DE (F), DC (G) and the analysis of node connectivity  $k$  (H). Error bars represent the standard deviation.

Because the seven pipelines differ in how they treat ambiguous reads, either discarding them (GSNAP-PolyCat, HyLiTE and Bowtie2) or statistically partitioning them (EAGLE-RC, RSEM, Salmon and Kallisto), we first evaluated the 'Efficiency' and 'Discrepancy' of read assignment. 'Efficiency' simply measures the percentage of reads assigned, considering all the reads versus those partitioned into each subgenome. As shown in Table 2, RSEM, Salmon and Kallisto all achieved 100% read assignment due to their underlying statistical inference of origin for ambiguous reads; however, they tend to slightly overestimate the number of At reads. EAGLE-RC assigned slightly fewer reads (96.3% of the total) through competitive mapping to the A- and D-subgenomes. Since Bowtie2, GSNAP-PolyCat and HyLiTE discard ambiguous reads, their 'Efficiency' negatively correlates with 'Ambiguity,' as expected (Figure 2B), with less than 90% of total

reads partitioned into subgenome (Table 2). In contrast to RSEM, Salmon and Kallisto, there appears to be a classification bias in both GSNAP-PolyCat and HyLiTE that leads to more reads being characterized as derived from the reference genome; this bias is most significant for HyLiTE (Table 2; At 78.5% versus Dt 85.8% based on  $D_5$  reference and At 82.7% versus Dt 80.3% based on  $A_2$  reference; Student's  $t$ -test  $P < 0.05$ ). It is worth noting that an average of 45% more reads per library, regardless of the diploid origin, were mapped to the  $D_5$  than the  $A_2$  reference by HyLiTE, indicating the higher quality of the  $D_5$  genome reference and its gene models; hence, only the  $D_5$ -based HyLiTE datasets were included in the following analyses. Although EAGLE-RC uses both references for mapping and quantification was restricted to homoeologous gene pairs (excluding subgenome-unique genes), a higher 'Efficiency' was obtained for Dt than At reads

Table 2. Overall and subgenome assessment of homoeolog expression estimation

	GSNAP-PolyCat		HyLiTE		EAGLE-RC	RSEM	Salmon	Kallisto	Bowtie2
Reference	D <sub>5</sub>	D <sub>5</sub>	A <sub>2</sub>	A <sub>2</sub> and D <sub>5</sub>	Polyploid AD				
Efficiency	87.7 ± 1.5%	82.2 ± 0.7%	81.5 ± 0.7%	96.3 ± 0.8%	<b>100.0%</b>	<b>100.0%</b>	<b>100.0%</b>	84.9 ± 1.5%	
At	86.7 ± 1.6%	78.5 ± 1.0%	82.7 ± 0.7%	95.5 ± 0.9%	101.0 ± 0.6%	101.6 ± 0.5%	101.5 ± 0.6%	84.6 ± 1.5%	
Dt	88.6 ± 1.6%	85.8 ± 0.6%	80.3 ± 0.9%	97.1 ± 1.0%	99.1 ± 0.6%	98.5 ± 0.5%	98.6 ± 0.6%	85.1 ± 1.6%	
Discrepancy	12.7 ± 1.5%	18.1 ± 0.8%	18.7 ± 0.8%	<b>5.1 ± 0.8%</b>	<b>5.1 ± 0.6%</b>	<b>5.1 ± 0.6%</b>	<b>5.1 ± 0.6%</b>	<b>5.1 ± 0.6%</b>	15.9 ± 1.5%
At	13.4 ± 1.6%	21.0 ± 1.1%	17.5 ± 0.8%	5.4 ± 0.9%	5.4 ± 0.6%	5.2 ± 0.6%	5.2 ± 0.6%	16.0 ± 1.5%	
Dt	12.1 ± 1.5%	14.8 ± 0.6%	20.0 ± 0.9%	4.8 ± 0.8%	4.8 ± 0.6%	5.1 ± 0.6%	5.0 ± 0.6%	15.7 ± 1.5%	
Precision	–	–	–	–	–	–	–	–	–
At	<b>98.4 ± 0.1%</b>	92.4 ± 0.3%	91.9 ± 0.4%	<b>98.1 ± 0.1%</b>	95.1 ± 0.2%	94.3 ± 0.2%	94.4 ± 0.2%	95.5 ± 0.1%	
Dt	<b>97.7 ± 0.5%</b>	91.2 ± 0.6%	91.6 ± 0.5%	<b>97.2 ± 0.5%</b>	97.0 ± 0.5%	96.5 ± 0.4%	96.4 ± 0.4%	96.1 ± 0.5%	
Recall	–	–	–	–	–	–	–	–	–
At	<b>97.7 ± 0.5%</b>	90.6 ± 0.6%	93.1 ± 0.4%	<b>97.2 ± 0.4%</b>	96.5 ± 0.5%	96.4 ± 0.4%	96.4 ± 0.4%	96.4 ± 0.5%	
Dt	<b>98.1 ± 0.1%</b>	94.0 ± 0.3%	91.3 ± 0.4%	<b>97.9 ± 0.1%</b>	96.2 ± 0.2%	95.4 ± 0.2%	95.4 ± 0.2%	95.0 ± 0.1%	
F <sub>1</sub> score	–	–	–	–	–	–	–	–	–
At	<b>98.7 ± 0.2%</b>	92.2 ± 0.4%	92.9 ± 0.3%	<b>98.1 ± 0.2%</b>	97.2 ± 0.3%	96.8 ± 0.2%	96.9 ± 0.2%	97.0 ± 0.3%	
Dt	<b>98.7 ± 0.2%</b>	93.1 ± 0.4%	92.1 ± 0.3%	<b>98.0 ± 0.2%</b>	97.2 ± 0.3%	96.7 ± 0.2%	96.7 ± 0.2%	96.9 ± 0.3%	
Accuracy*	<b>98.2 ± 0.2%</b>	93.7 ± 0.7%	93.6 ± 0.3%	<b>97.7 ± 0.2%</b>	96.3 ± 0.3%	95.7 ± 0.2%	95.8 ± 0.2%	96.2 ± 0.2%	
MCC*	<b>96.8 ± 0.5%</b>	84.5 ± 0.7%	84.2 ± 0.7%	<b>95.8 ± 0.4%</b>	95.5 ± 0.5%	94.7 ± 0.4%	94.8 ± 0.4%	94.2 ± 0.5%	

Note: The best performance for each metric is marked in bold text.

\*Same values for At and Dt reads.

(82.7% versus 80.3%, respectively). This result may also reflect differences in assembly and annotation quality between the reference genomes. We also evaluated the ‘Discrepancy’ for each pipeline, which measures the absolute difference between the obtained homoeolog read counts and the expected counts; this measure is affected by both the assignment ‘Efficiency’ and binary classification measures (see Methods). Due to the high ‘Efficiency’ guaranteed by the algorithms of EAGLE-RC, RSEM, Salmon and Kallisto, these pipelines exhibit the lowest ‘Discrepancy’ (5.1%; Table 2), while the highest ‘Discrepancy’ was found in HyLiTE (18.1% and 18.7%), followed by 15.9% in Bowtie2 and 12.7% in GSNAP-PolyCat. In general, ‘Discrepancy’ from the actual read numbers increases as the level of ‘Ambiguity’ increases (Figure 2C), as expected; however, EAGLE-RC performs robustly across ‘Ambiguity’ bins.

While ‘Efficiency’ and ‘Discrepancy’ provide generic measures of read partitioning based on the numbers expected, they do not account for whether each read is accurately assigned. Therefore, the results of each pipeline were arrayed in a 2 × 2 confusion matrix (i.e. TP, FN, etc.; [43]), and the performance of the pipeline was evaluated using the information retrieval metrics of ‘Precision,’ ‘Recall,’ ‘Accuracy,’ F<sub>1</sub> score and MCC. In the context of information retrieval (as implemented here), ‘Precision’ measures how many of the reads assigned to a given subgenome (A or D) were correctly identified, ‘Recall’ measures how many of each subgenome were retrieved from the mixed population (relative to expectations), and ‘Accuracy’ measures how well each pipeline correctly identifies one subgenome while excluding the other; the measures F<sub>1</sub> and MCC account for more of the confusion matrix and attempt to generalize the results into a single score of performance (see Methods for details). The results in Table 2 show that that GSNAP-PolyCat and EAGLE-RC generally performed better in all information retrieval metrics, meaning that they recovered more relevant reads for each subgenome while excluding reads from the other subgenome. The four generic alignment-based approaches (i.e. Bowtie2, RSEM, Salmon and Kallisto) showed comparable performance to each other, with only a slight reduction in all scores relative to GSNAP-PolyCat. Only HyLiTE stands out as performing relatively

poor compared to the other pipelines; however, it is noteworthy that the other pipelines (except EAGLE-RC) all utilized the same SNP information (1 251 736 coding region SNPs) derived from rich genomic resequencing data [23], whereas HyLiTE conducted on-the-fly SNP calling from the input parental diploid RNA-seq datasets (754 670 and 871 224 coding region SNPs based on A<sub>2</sub> and D<sub>5</sub> references, respectively). This most likely explains the relatively poor performance of HyLiTE, as tested here. The high performance of EAGLE-RC is likely due to the use of both mapping statistics and subgenomic variants (630 309 670 and 630 313 coding region SNPs based on A<sub>2</sub> and D<sub>5</sub> references, respectively). Interestingly, as shown in Figure 2D and E, EAGLE-RC, GSNAP-PolyCat and HyLiTE exhibit relatively consistent performance across ‘Ambiguity’ bins, indicating that their accuracy (as measured by ‘Accuracy’ and MCC) is largely static, irrespective of homoeolog divergence. Bowtie2, RSEM, Salmon and Kallisto, however, perform nearly as well as GSNAP-PolyCat when the expected amount of homoeologous ambiguity is low, but quickly descend when ‘Ambiguity’ goes above 20% (Figure 2D and E).

### The inference of homoeolog expression divergence is affected by the choice of expression estimating pipeline

Expression divergence of homoeologs, both relative to one another and to their progenitor genomes, is a major component of polyploid research. Allopolyploidy reunites formerly diverged genes (and their regulatory context) into a common nucleus while simultaneously generating massive redundancy. Consequently, observed transcriptomic changes are myriad (reviewed in [13]) and include homoeolog expression bias (reviewed by [12, 13]) and functional divergence [62–65]. Since our ability to accurately describe expression changes depends upon our ability to accurately represent expression, we evaluated the extent to which each pipeline accurately represented DE between homoeologs. That is, the homoeolog DE results derived from each pipeline inferred AD dataset were compared to the expected DE results between the diploid orthologs from which the AD datasets were derived. While many methods



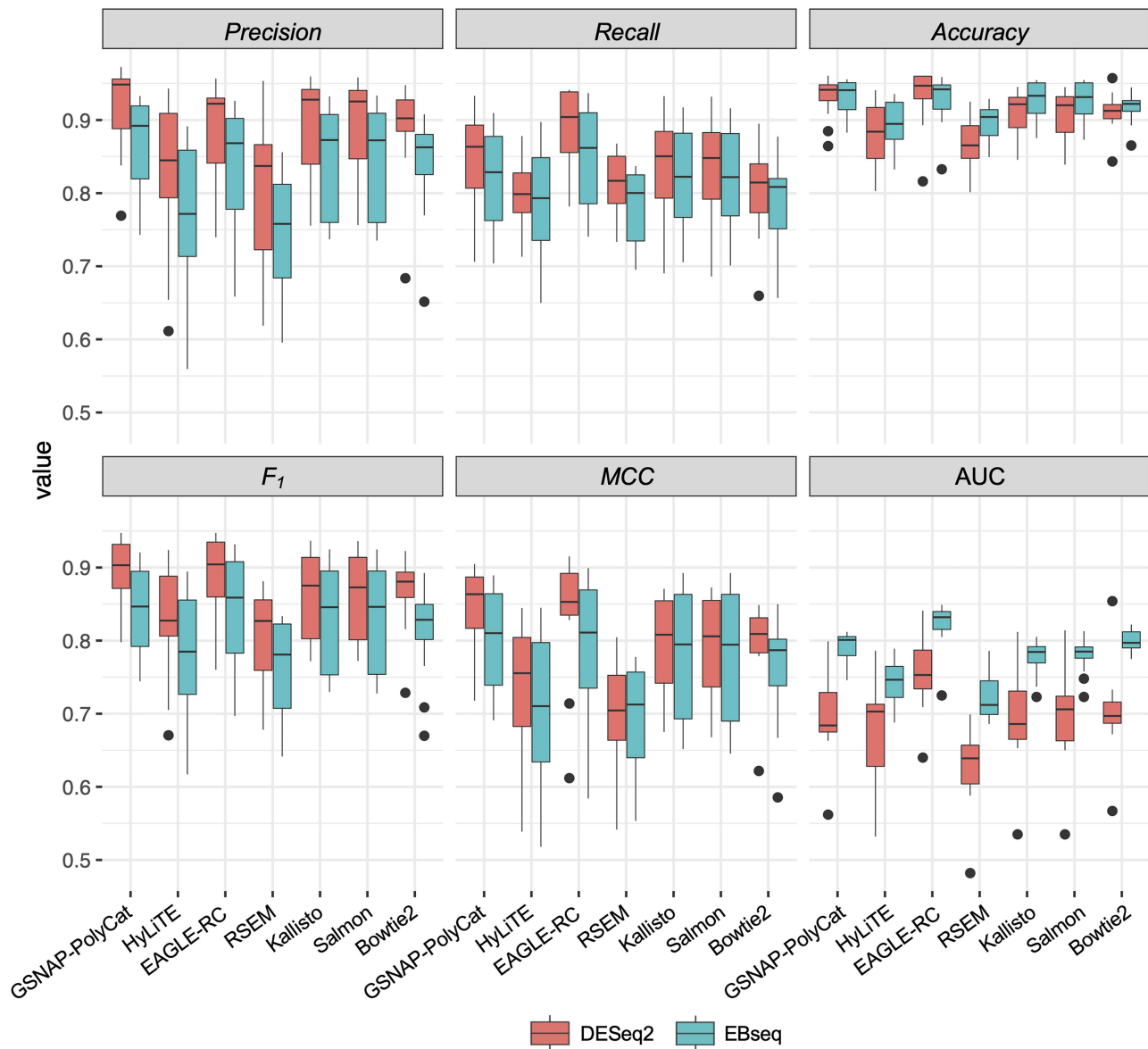
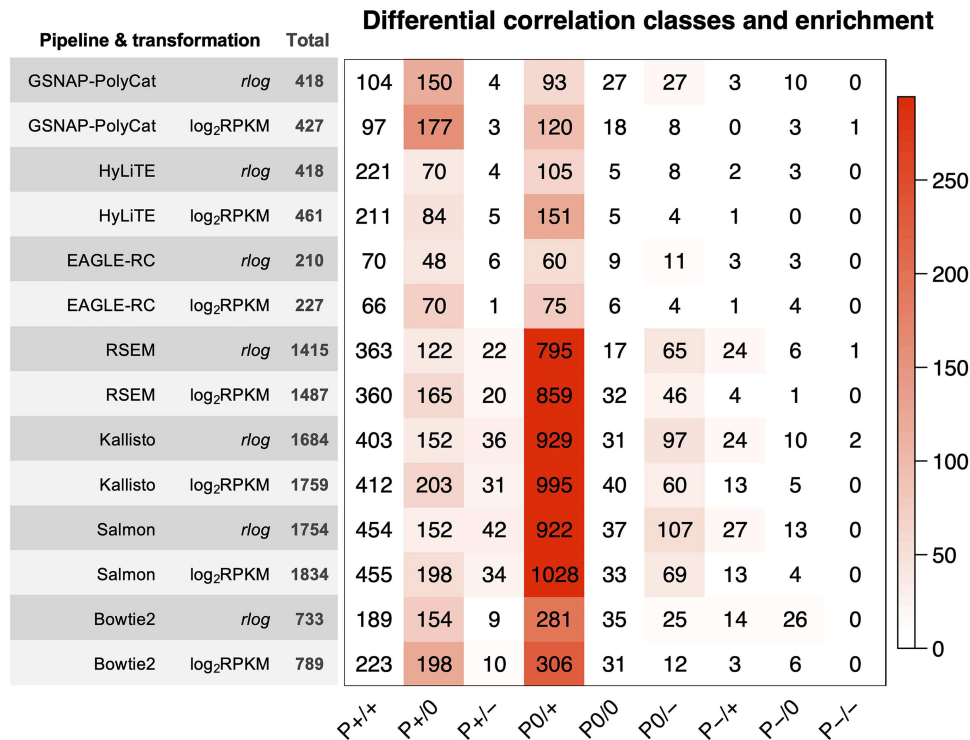


Figure 3. Performance evaluation of DE analysis between homoeologous genes.

exist for comparing DE among samples, we selected two of the most popular methods, namely, DESeq2 and EBSeq, to compare both stringency and accuracy in general and in the context of the different pipelines. That is, following each quantification pipeline, the same expected and observed read count tables were supplied to DESeq2 and EBSeq analyses.

Overall, DESeq2 detected an average of 16% more significant changes in expression than EBSeq (see [Supplementary Table S3](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>; paired Student's t-test  $P < 0.05$ ), suggesting that by default, the latter is more stringent. Across the 12 sample conditions, pairwise homoeolog expression divergence was detected between 8% and 44% of the homoeologous gene pairs (28 401 pairs by EAGLE-RC and 37 223 pairs by other pipelines), without significant differences between the observed and expected datasets (see [Supplementary Table S3](https://academic.oup.com/bib) available online at <https://academic.oup.com/bib>; analysis of variance (ANOVA) formula: DE = pipeline + DE method + dataset; pipeline  $P < 0.001$ , DE method

$P < 0.001$  and dataset  $P = 0.23$ ); however, the homoeolog expression divergence common to all sample conditions was significantly underestimated from the observed than expected datasets (51% versus 53%; paired Student's t-test  $P < 0.05$ ). As shown in [Figure 3](#), a relatively high level of DE 'Accuracy' (above 80%) was consistently inferred. Regardless of which DE method was used, the expression datasets generated by GSNAP-PolyCat and EAGLE-RC outperformed those by other pipelines (Salmon/Kallisto/Bowtie2 > HyLiTE/RSEM) in identifying the true expression divergence between homoeologs. While DESeq2 appeared to perform better than EBSeq according to the measures of 'Precision,' 'Recall,'  $F_1$  and MCC, the AUC scores suggested that EBSeq is more robust than DESeq2 to separate binary classes ([Figure 3](#)), particularly for genes exhibiting high 'Ambiguity' ([Figure 2F](#)). For both methods, their AUCs were negatively correlated with 'Ambiguity,' reflecting the strong dependence of DE analysis on the extent of homoeolog sequence divergence ([Figure 2F](#)).



**Figure 4.** Inference changes in co-expression relationships between homoeologs. For each of the 14 combinations of homoeolog expression estimation pipelines and data transformation methods (row), the number of DC changes between observed and expected datasets is shown for each DC category (column). Cell color indicates the magnitude of significant overrepresentation based on  $-\log_{10}(P\text{-value})$  of Fisher's exact test (i.e.  $P=0.05$  is converted to 1.3). For example, the number in category  $P_{0/+}$  of the bottom row indicates that 306 homoeolog pairs showed DC changes from no significant correlation (0) to significantly positive correlation (+) due to the estimation error from the Salmon pipeline followed by  $\log_2$ RPKM transformation.

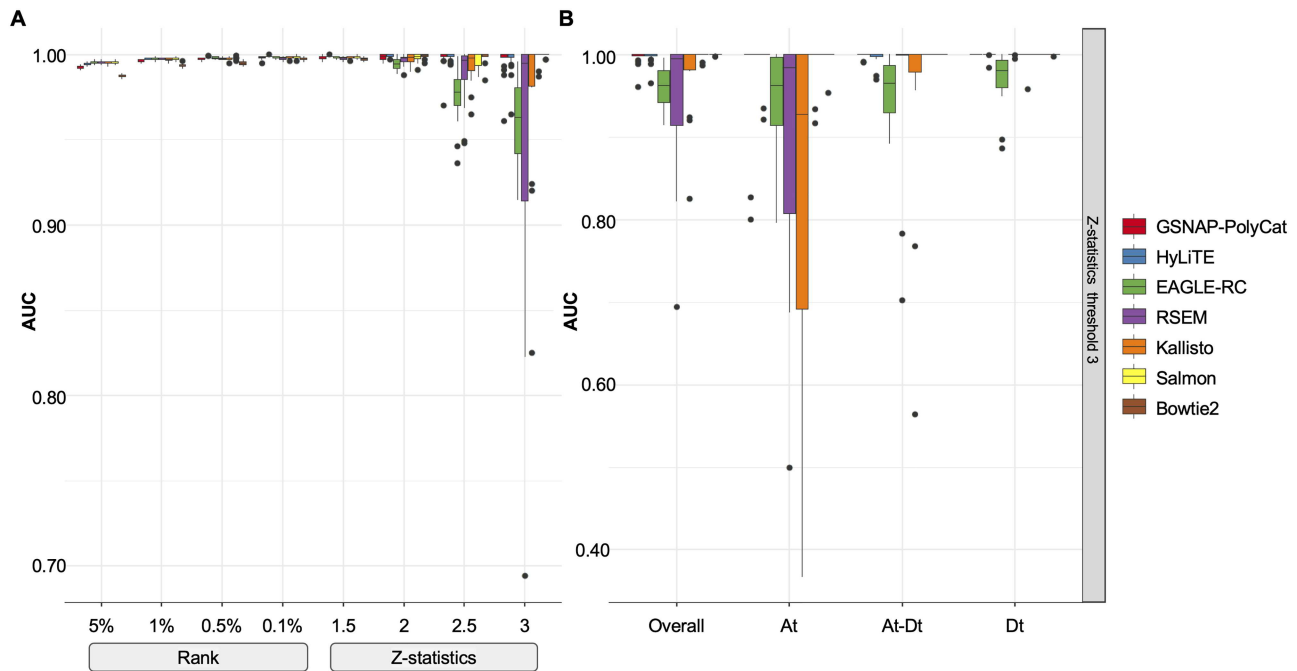
Co-expression relationships between homoeologs were measured using Pearson correlation coefficient across multiple sample conditions. Approximately 1–5% of homoeolog pairs (210–227 out of 28 401 pairs by EAGLE-RC and 418–1834 out of 37 223 pairs by other pipelines) exhibited significant changes due to incorrect read assignment. As shown in Figure 4, EAGLE-RC introduced the smallest numbers of DC changes, followed by GSNAP-PolyCat and HyLiTE, which outperformed Bowtie2, RSEM, Salmon and Kallisto. Artificially induced DC was most prominent in those homoeologous gene pairs exhibiting higher ‘Ambiguity’ (Figure 2G), with the highest bin (i.e. 0.2–1) exhibiting a nearly 4-fold increase in DC than other bins for RSEM, Salmon and Kallisto. Among the nine categories of DC changes (Figure 4, columns), the class of  $0/+$  was most significantly enriched for RSEM, Kallisto, Salmon and Bowtie2. This suggests that the majority of DC changes due to read partitioning errors lead to gains in correlation, generally changing our inferences from no significant correlations (0) to significantly positive correlations (+). Together with the observation of generally fewer DE genes common to all sample conditions, these results indicate that read partitioning methods could lead to an overestimation of co-regulation between homoeologous genes due to incorrect homoeolog expression estimation, consequently restricting our ability to infer expression divergence and/or possible functional divergence of duplicated genes. Notably, these patterns were consistent for both the *rlog* and  $\log_2$ RPKM data transformation methods. In addition to DC between homoeologous gene pairs, we also conducted identification and classification of DC patterns for all possible gene pairs (see Supplementary Table S4 available online at

<https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bba035/5811916>), resulting in 0.3–1.5% global pairwise DC changes, which affected 3.5–15.2% of total genes (i.e. DC genes enriched with DC pairs) in their co-expression relationships.

### Robust construction of gene co-expression networks by the rank-based binary method and WGCNA

Gene co-expression networks are commonly used to summarize the multidimensionality of gene expression data into clusters of genes with putatively related functions (i.e. modules). In the context of polyploidy, co-expression networks can be used to assess the functional relatedness among genes and homoeologs, reveal changes in homoeolog usage and characterize the genetic interplay between subgenomes [6]. We use both weighted and unweighted networks to assess the influence of variation in read partitioning on our inferences of co-expression.

Constructing unweighted co-expression networks requires a binary classifier (or hard threshold) to decide whether there exists a connection (i.e. an edge) between each pair of genes. As shown in Figure 5A, different rank-based thresholds (5, 1, 0.5 or 0.1% of top-ranked correlations become edges) yielded robust classification of the expected edges (based on diploid expression) with AUC scores close to 1. In contrast, the performance of Z-statistics-based thresholds (i.e. significant correlations with Z-score above 1.5, 2.0, 2.5 or 3.0 become edges) was more variable (AUC of 0.8–1) depending on the stringency of the Z-thresholds. These results indicated that the rank-based method is more robust here than Z-statistics to infer binary gene co-expression network.



**Figure 5.** Performance of binary co-expression network construction. (A) Boxplot of AUC scores were shown using different homoeolog estimation pipelines (color) and binary thresholds (x-axis). (B) Taking the Z-score threshold of 3, for example, AUC scores were compared among subnetworks: Overall, all edges considered; At, edges within the A-genome subnetwork; Dt, edges within the D-genome subnetwork; At-Dt, edges connecting genes across A- and D-subnetworks.

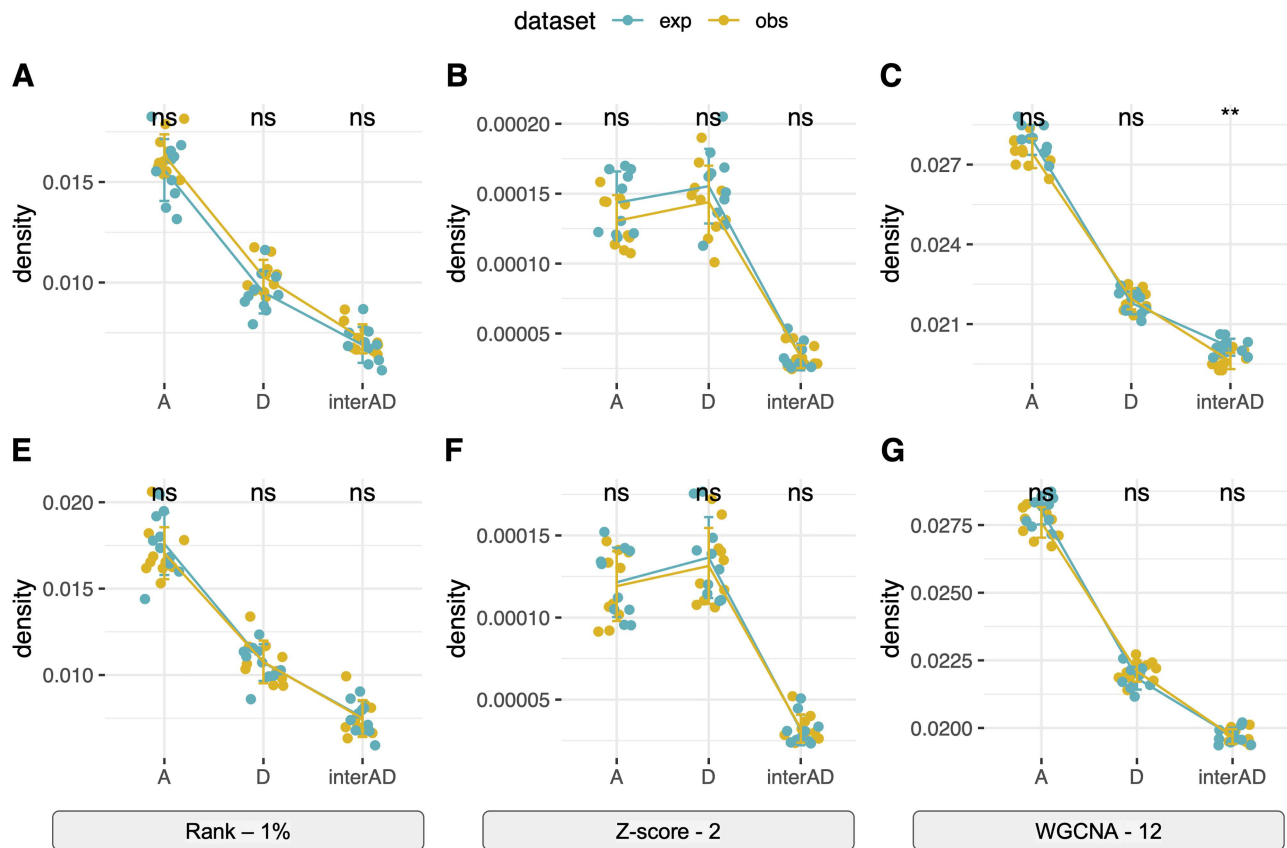
In addition to the network construction methods (ANOVA formula:  $AUC = \text{construction} + \text{pipeline} + \text{transformation}$ ; construction  $P < 2e-16$ ), the choice of read estimation pipeline also matters ( $P = 1.58e-08$ ) with the performance of RSEM and EAGLE-RC significantly falling behind others (Tukey's HSD test  $P < 0.05$ ); no significant difference was found between the  $r\log$  and  $\log_2\text{RPKM}$  transformation (ANOVA  $P = 0.615$ ). The lower performance of EAGLE-RC networks seems counterintuitive given its superiority in the previous evaluation; however, transcript quantification for EAGLE-RC was limited to only those homoeologous gene pairs that were annotated in both genomes, resulting in circa 9000 fewer genes assayed due to annotation differences between the reference genomes. Thus, the derived homoeologous gene networks could be susceptible to scaling errors between subgenomes due to data exclusion. Interestingly, while not unexpected, edge inference within the D-genome subnetwork is significantly more robust than edges within the A-genome subnetwork or those across subnetworks (Figure 5B; ANOVA and Tukey's HSD test  $P < 0.05$ ). This observation likely reflects quality differences in the mapping reference, i.e. the high-quality D-genome reference versus the inferred or actual A-genome sequences (see Methods).

In WGCNA networks, the quantitative strength of network connections is considered to maximize information captured in the network. The topological preservation tests of expected modules (based on diploid expression) exhibited high preservation ( $Z_{\text{summary}} > 10$ ) for almost all modules (see Supplemental Figure S1A available online at <https://academic.oup.com/bib>), regardless of soft threshold (i.e. 1, 12 or 24; see Methods), homoeolog read estimation pipeline and method of normalization. This result suggests that WGCNA-based inference of gene modular structure is rather robust.

In addition to the separate topological evaluation above (binary networks by edge inference AUC and WGCNA networks by module preservation), node connectivity ( $k$ ) and network

functional connectivity (FC) were calculated for each binary and weighted co-expression network constructed. Each AD network constructed was evaluated against the expected (diploid-based) network. Pearson correlation coefficients between the expected and observed networks suggest that both  $k$  and FC were rather consistent across different homoeolog expression estimation pipelines (ANOVA formula:  $\text{correlation} = \text{construction} + \text{pipeline} + \text{transformation}$ ; construction  $P > 0.05$ ), whereas the method of network construction could strongly influence topology ( $P < 2e-16$ ; see Supplemental Figure S1B–D available online at <https://academic.oup.com/bib>). Notably, normalization method affected  $k$  ( $P < 2e-16$ ;  $\log_2\text{RPKM}$  outperforms  $r\log$ ) but not FC ( $P = 0.08$ ). As shown in Supplemental Figure S1B–D, available online at <https://academic.oup.com/bib>, both rank-based binary construction and weighted gene network construction methods equally outperformed all but the least strict Z-statistics methods. The accurate inference of  $k$  (measured by correlation between observed and expected data; Figure 2H) is negatively correlated with 'Ambiguity,' albeit weakly. This diminished relationship is expected as the network property of each gene is intrinsically determined by all the other genes, thereby obscuring the impact of ambiguity per gene.

In addition, the measure of FC can be used to statistically evaluate the functional significance of network topology [61]. According to the 'guilt-by-association' principle [66], genes with similar functional properties tend to interact or be clustered together in biological networks. Thus, higher FC indicates more reasonable network topology. As shown in Supplemental Figure S2, available online at <https://academic.oup.com/bib>, the highest FC scores were observed for WGCNA networks ( $AUC = 0.64\text{--}0.72$ ), followed by the ranked-based binary networks ( $0.54\text{--}0.67$ ) and the Z-statistics-based binary networks ( $0.50\text{--}0.55$ ), respectively. This may suggest that the WGCNA network construction was able to capture more function and/or biologically relevant information.



**Figure 6.** Different inferences of subnetwork topology. The network density of A-subnetwork and D-subnetwork and interconnections between A- and D-subnetworks were shown for both the expected and observed data from the GSNAP-PolyCat (A–C) and EAGLE-RC (E–G) estimation with  $\log_2$ RPKM normalization. (A and E) Rank-based binary network with top 1% connections; (B and F) Z-statistics binary network with connections above Z-score of 2; (C and G) WGCNA network with the power of 12.

Overall, the performance of co-expression network analysis was more affected by network construction methods than by read ambiguity and partitioning methods. In general, either  $\log_2$ RPKM or  $r$ log combined with WGCNA produced the best results for these data, regardless of read assignment method.

### Bioinformatic choices can strongly affect the interpretation of duplicated gene network topology

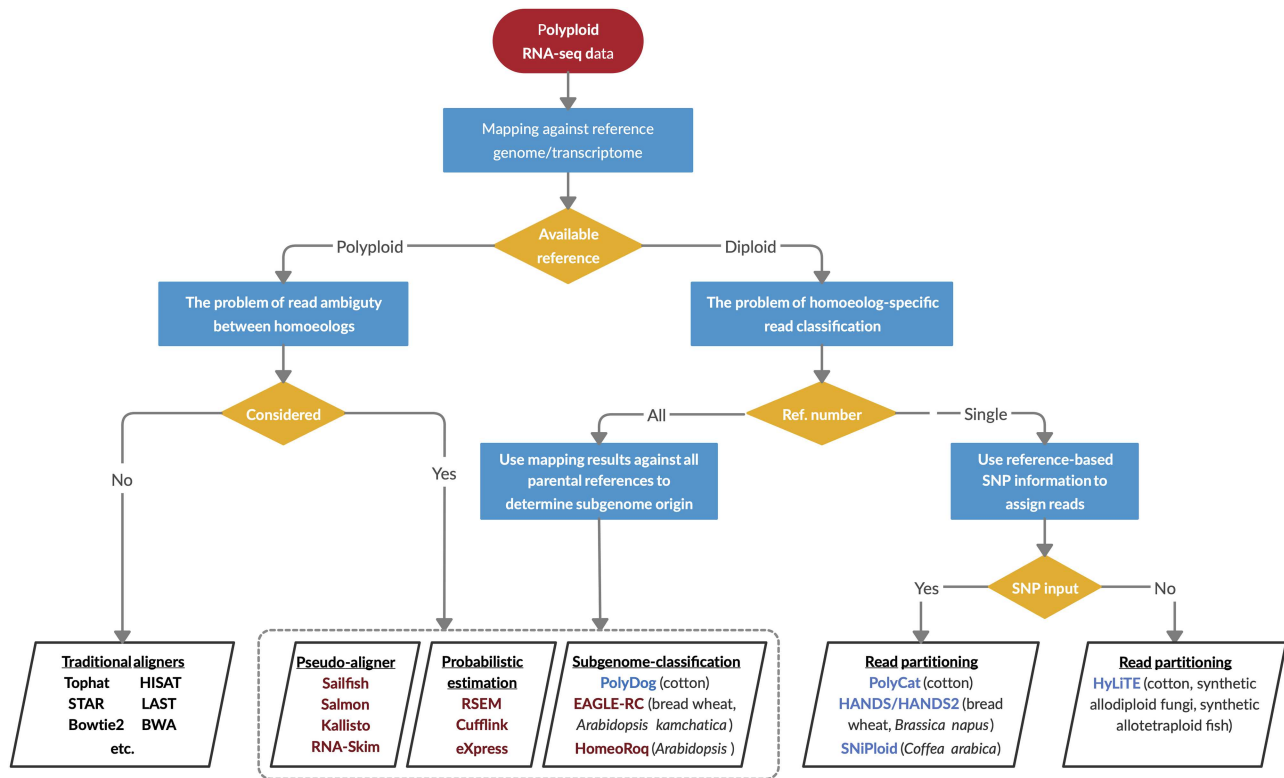
In the context of polyploid gene network, it is of particular interest to compare subnetwork properties within each subgenome and between subgenomes. Taking the GSNAP-PolyCat and EAGLE-RC datasets followed by  $\log_2$ RPKM normalization as examples (Figure 6), both rank-based and WGCNA networks revealed the highest density (mean connections) of the A-subnetwork, followed by that of the D-subnetwork and then of the interconnections between A- and D-subgenomes. In contrast, similar levels of A- and D-subnetwork density were revealed in the Z-statistics-based networks (Figure 6B and F). These results led to opposite conclusions regarding the potential topological asymmetry between two subgenomes. According to the performance assessment above, we believe that the conclusion derived from WGCNA and rank-based binary networks is more reliable; that is, the At genes are more interconnected than are the Dt genes, reflecting the difference in gene regulation between the two subgenomes (i.e. the  $A_2$  and  $D_5$  diploids used generate synthetic AD). In addition, all networks agreed on the much lower density of

inter-subgenome connections than that of within-subgenome connections, indicating that a gene is much more likely to be connected with genes from the same subgenome than with genes from the other subgenome. For other combinations of homoeolog expression estimation, transformation and network construction methods, the measures of subnetwork density are shown in Supplementary Table S5, available online at <https://academic.oup.com/bib>, where WGCNA and rank-based generally support the conclusion of higher density in At versus Dt subnetworks, in contrast to the Z-statistics-based networks.

### Discussion

The duplicated nature of polyploid genomes poses unique challenges for bioinformatics. Presently, we are witnessing an explosion of interest in better understanding these challenges and developing appropriate methodologies and tools for polyploids, for applications as diverse as genome sequence assembly [67], genotyping [68, 69], haplotype phasing [70, 71], population-based trait analysis [72], phylogenetic inference [73, 74] and transcriptomic-based analyses [75, 76] such as *de novo* transcriptome assembly [77] and transcript quantification [30]. Quantification of homoeolog expression is particularly interesting, given the various patterns of duplicate gene expression possible in polyploid species (reviewed in [12]), the interactions among homoeologs in a gene network context [6, 14] and the phenomenon of unbalanced homoeolog expression bias together with its potential long-term consequences for





**Figure 7.** A decision-making diagram to choose the appropriate bioinformatic resources for estimating homoeolog expression levels. When a reference genome or transcriptome is available for the polyploid species, quantification of individual homoeologs is straightforward using the traditional aligners such as TopHat, or the problem of read ambiguity by applying probabilistic estimation methods, pseudo-aligners or subgenome-classification approaches is considered. The latter also applies to the scenario when subgenome references are available from all the diploid progenitors. When the reference is only available for one diploid progenitor, software has been developed for partitioning and quantifying homoeolog-specific reads given the genetic variants such as SNPs between homoeologous sequences. Maroon-colored software, such as RSEM and EAGLE-RC, statistically assigns the subgenome origin for ambiguously mapped reads; blue-colored software such as PolyCat utilizes only unambiguously mapped reads for estimation. The polyploid systems for which they were originally developed are noted in parentheses.

fractionation [78–80]. A number of previous studies have explored the bioinformatics of homoeolog expression profiling [30, 75–77]; however, both the fundamental issue of read ambiguity and the downstream inferences regarding polyploid expression evolution have not been addressed. Here we present a comprehensive analytic workflow to demonstrate the challenges and pitfalls of these analyses (Figure 1), as well as how they are influenced by the extent of read ambiguity in the dataset and how that ambiguity is handled in understanding homoeolog expression and co-expression patterns (Figure 2).

### Duplication and deficiency: when redundancy renders reads unresolved

In addition to the redundant nature of polyploid genomes, there are a number of biological and technical causes for ambiguous read mapping, including transcripts that are expressed at low levels, sequence homology, small-scale gene duplications and errors in sequencing and annotation. While we can control some of these factors through experimental design (i.e. read length, paired-end sequencing, etc.), the nature of the biological system and the amount/distribution of subgenome divergence, as measured by ‘Ambiguity,’ will influence the ability to accurately assign reads to homoeologs. Although our analysis is limited to the example data from *Gossypium*, the metric of ‘Ambiguity’ can be applied to any other real-world or simulated polyploid systems. For systems that have less divergent subgenomes than

*Gossypium*, the ‘Ambiguity’ values are expected to be higher, and longer read lengths will be required to improve the ability to accurately assign reads. Knowing the range of ‘Ambiguity’ for any specific polyploid system or for a list of genes of interest, we can foresee the use of Figure 2 to query how such a range affects the performance of bioinformatic inferences regarding homoeolog read estimation (B–E) and polyploid expression evolution (F–H).

Among tools that have been devised to estimate homoeolog expression levels under different conditions (e.g. the availability and type of the reference genome; Figure 7), numerous methods exist for handling the subset of reads that are not uniquely assignable, typically either discarding these reads (as in GSNAP-PolyCat and HyLiTE) or statistically assigning the reads (e.g. RSEM, Kallisto and Salmon). Among the seven pipelines evaluated in this study (see Supplementary Table S2 available online at <https://academic.oup.com/bib>), most performed relatively well, achieving >90% success for information retrieval metrics (Table 2). Notably, EAGLE-RC and GSNAP-PolyCat exhibited the best scores for most metrics, representing the subgenome-classification and single-reference-based read partitioning methods for homoeolog expression quantification (Figure 7), respectively. In a previous study, Kuo et al. [30] showed that EAGLE-RC outperforms other methods including STAR, LAST, Kallisto and HomoeoRoq to precisely estimate homoeolog expression in both tetraploid *Arabidopsis kamchatica* and hexaploid wheat. This category of the subgenome-classification approaches, including EAGLE-RC, HomoeoRoq and PolyDog, requires read mapping against the genome reference of each

subgenome separately (similar to RSEM, Kallisto and Salmon, which require individual transcriptome references) in order to determine the better supported homoeolog origin for reads, followed by homoeolog identification. Application of these has an added layer of complexity, as the reference quality and annotation methods often differ between the separately generated parental reference genomes and/or between diploid and polyploid genomes. In addition to these reference-based differences, ortholog and/or homoeolog identification determined by other software (e.g. OrthoFinder or similar) is required before accurate comparisons can be made. Here, for example, due to the assembly and annotation differences between the  $A_2$  and  $D_5$  references, EAGLE-RC assayed about 25% fewer homoeolog gene pairs than did the other pipelines.

Contrary to the high performance of GSNAP-PolyCat, another single-reference-based read partitioning method—HyLiTE—exhibited the least performance in homoeolog quantification, even worse than the Bowtie2 method counting only uniquely mapped reads. This poor performance of HyLiTE can be attributed to the limited amount of SNP information directly generated from the input RNA-seq datasets, while a more comprehensive set of SNP information based on extensive resequencing data was utilized by pipelines including GSNAP-PolyCat, RSEM, Kallisto, Salmon and Bowtie2. Among these latter pipelines, while it is tempting to attribute the improved performance of GSNAP-PolyCat to its partitioning algorithm that discards ambiguous reads, the performance of Bowtie2, which also discards ambiguous reads by considering only uniquely mapped reads, was not better than RSEM, Salmon or Kallisto that statically assign the subgenome origin to ambiguous reads. When ‘Ambiguity’ was low, all pipelines performed similarly well; however, those that statistically assign ambiguous reads (RSEM, Salmon, Kallisto) perform significantly worse for those genes with ‘Ambiguity’ above 20%. This may be due to the noise in the underlying statistics as the relative number of unique reads drops compared to those that will be statistically assigned; that is, any error in statistical inference will be amplified as the number of ambiguous reads begins to outweigh the number of unique reads. This is an important observation for polyploid systems whose subgenomes are more recently diverged. That is, methods which statistically assign ambiguous reads should be used with caution when the divergence between parental genomes is low. For those genomes, GSNAP-PolyCat and HyLiTE will provide a more reliable representation of relative homoeolog read counts, with GSNAP-PolyCat outperforming HyLiTE when a priori homoeoSNP information is available.

The use of one reference sequence is also not without challenges. For example, GSNAP-PolyCat and HyLiTE both appeared to partition more reads than expected to the reference genome used, whereas RSEM, Kallisto and Salmon statistically characterized more reads as A-derived while the reference transcriptome was built on the D-reference gene models. The cause of this discrepancy is unknown, but may include technical differences (e.g. algorithm design) or biological sources of error, such as uncharacterized copy number variation and diploid/subgenome-unique genes. These errors have the potential to influence subsequent conclusions, such as co-expression network characteristics (e.g. the more robust inference of D- versus A-subnetwork topology observed here). These caveats notwithstanding, our results here demonstrate that a single reference genome can reasonably be used where resources are limited. We do, however, envision approaches using the entire polyploid genome or representative diploid genomes will be useful for hybrid and polyploid systems where

quality differences among progenitor reference genomes are negligible and where similar annotation methods are used for each.

While we did not specifically test the consequences of diminishing read depth on the accuracy of read assignment, it is reasonable to assume an inverse relationship between the number of reads per sample and ‘Ambiguity.’ The datasets used here contained relatively high coverage of the transcriptome, averaging approximately 26.6 million reads per library with low ambiguity for most pipelines. Reducing the number of reads per library either will decrease the overall number of reads assigned (for pipelines that discard reads) or will reduce the signal for pipelines that statistically partition reads. Therefore, it is generally advisable to scale the depth of sequencing with ploidy when partitioning homoeolog expression to account for the additional gene space.

### Consequences of inaccurate quantification for inferences of polyploid evolution

Beyond the narrow issue of evaluation of homoeolog quantification, our interest lies in identifying a reasonable set of methods to address biological and evolutionary questions concerning polyploidy. DE is commonly among the first transcriptomic analyses performed, providing a generalized look at the extent of expression divergence. For polyploid species, the relative expression of each subgenome is of particular interest, which may provide insights into homoeolog bias, expression level dominance, *cis-trans* resolution, putative sub-/neo-functionalization of homoeologs and other phenomena [14]. In general, GSNAP-PolyCat and EAGLE-RC best represented the expected DE between homoeologs, followed closely by Kallisto and Salmon (Figure 3). The standout, RSEM, performed significantly more poorly than the rest despite its intended application to isoform quantification; we therefore advise caution when using RSEM for duplicate gene analyses. With respect to DE inference, both DESeq2 and EBSseq resulted in reasonable performance metrics, with the choice likely being the stringency level and parameters preferred.

In contrast to the general robustness of the DE results, homoeolog read ambiguity and the choice of quantification pipelines strongly influence our interpretation of co-expression relationships among genes. In particular, we are interested in detecting coordination among homoeologous genes in polyploids. The most significant error evident is the false detection of positive correlations where none exists. Notably, following the best-performing EAGLE-RC, those methods that discard reads (i.e. GSNAP-PolyCat, HyLiTE and Bowtie2) far outperformed the other methods, particularly for those genes with higher ‘Ambiguity.’ These results were consistent for the two normalization methods tried, i.e. *rlog* and *log2RPKM*, and may indicate a general preference for discarding ambiguous reads when the biological question depends on an accurate assessment of differential co-expression.

The inference of co-expression network topology, on the other hand, was generally less sensitive to the quantification method, but rather was dependent on method of network construction. This is probably because the multivariate nature of co-expression relationships mitigates the influence of individual and random quantification errors. Intuitively, however, the inferred topology of a subnetwork containing low ‘Ambiguity’ genes should be more reliable than a subnetwork containing higher ‘Ambiguity’ genes. In order to compare network topologies between subgenomes, choosing the appropriate network

construction method becomes critical; otherwise incorrect and even opposite conclusions may be reached (Figure 6). For example, both rank-based binary and WGCNA reconstructions of the present datasets suggest that the A-subnetwork is more tightly interconnected than the D-subnetwork, whereas the less reliable Z-statistic-based binary networks suggest they are equally interconnected.

## Conclusions

In this study, we present an analytical workflow from homoeolog expression quantification to a series of downstream analysis to infer key phenomena of polyploid expression evolution. By examining the extent and consequences of read ambiguity, we demonstrated the potential artifacts that may affect our understanding of duplicate gene expression, such as an overestimation of homoeolog co-regulation and the incorrect inference of subgenome asymmetry in network topology. Such errors may be reduced by mitigating technical factors that influence ambiguity, i.e. sequencing strategy and fundamental resources (i.e. genomes and/or resequencing). While the focus here is on analyses of tetraploids, many of the pipelines suitable for tetraploid analyses can be used for higher-order polyploids, with additional caveats regarding multidimensional read ambiguity and differences in phylogenetic distance among subgenomes. Although the collection of scenarios and methods tested in this study is not comprehensive and may be superseded by those yet to be developed, our work introduces the metric of 'Ambiguity' and designates a set of reasonable practices applicable to other polyploid systems.

### Key Points

- We present an analytical workflow to evaluate a variety of bioinformatic method choices at different stages of polyploid RNA-seq analysis, from homoeolog expression quantification to downstream analysis used to infer key phenomena of polyploid expression evolution.
- We used transcriptomic data from the cotton genus (*Gossypium*) as an example to examine the extent and consequences of homoeolog read ambiguity.
- Our results show that EAGLE-RC and GSNAP-PolyCat outperform other quantification pipelines tested, and their derived expression datasets best represent the expected results in downstream analyses of DE and co-expression network analysis.
- We illuminate the potential artifacts that may affect our understanding of duplicate gene expression, including an overestimation of homoeolog co-regulation and the incorrect inference of subgenome asymmetry in network topology.
- Overall, our work points to a set of reasonable practices that are broadly applicable to the evolutionary exploration of polyploids.

## Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

## Acknowledgements

We thank the National Science Foundation Plant Genome Research Program and Cotton Incorporated for financial support. We also thank ResearchIT for computational support at Iowa State University.

## References

1. Flagel LE, Wendel JF, Udall JA. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* 2012;13:302.
2. Buschiazzi E, Ritland C, Bohlmann J, et al. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol* 2012;12:8.
3. Yang Y, Moore MJ, Brockington SF, et al. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Mol Biol Evol* 2015;32:2001–14.
4. Bombarely A, Coate JE, Doyle JJ. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. *PeerJ* 2014;2:e391.
5. McCormack JE, Hird SM, Zellmer AJ, et al. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* 2013;66:526–38.
6. Gallagher JP, Grover CE, Hu G, et al. Insights into the ecology and evolution of polyploid plants through network analysis. *Mol Ecol* 2016;25:2644–60.
7. Hu G, Koh J, Yoo M-J, et al. Gene-expression novelty in allopolyploid cotton: a proteomic perspective. *Genetics* 2015;200:91–104.
8. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 2019;574:679–85.
9. Jiao Y, Wickett NJ, Ayyampalayam S, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011;473:97–100.
10. Wendel JF. The wondrous cycles of polyploidy in plants. *Am J Bot* 2015;102:1753–6.
11. Jiao Y, Paterson AH. Polyploidy-associated genome modifications during land plant evolution. *Philos Trans R Soc B: Biol Sci* 2014;369:20130355.
12. Grover CE, Gallagher JP, Szadkowski EP, et al. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol* 2012;196:966–71.
13. Yoo MJ, Liu X, Pires JC, et al. Nonadditive gene expression in polyploids. *Annu Rev Genet* 2014;48:485–17.
14. Hu G, Wendel JF. Cis-trans controls and regulatory novelty accompanying allopolyploidization. *New Phytol* 2019;221:1691–700.
15. Hu G, Hovav R, Grover CE, et al. Evolutionary conservation and divergence of gene coexpression networks in *Gossypium* (cotton) seeds. *Genome Biol Evol* 2016;8:3765–83.
16. Pfeifer M, Kugler KG, Sandve SR, et al. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* 2014;345:1250091.
17. Takahagi K, Inoue K, Mochida K. Gene co-expression network analysis suggests the existence of transcriptional modules containing a high proportion of transcriptionally differentiated homoeologs in hexaploid wheat. *Front Plant Sci* 2018;9:1–10.

18. Li L, Briskine R, Schaefer R, et al. Co-expression network analysis of duplicate genes in maize (*Zea mays* L.) reveals no subgenome bias. *BMC Genomics* 2016;17:875.
19. Ilut DC, Coate JE, Luciano AK, et al. A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *Am J Bot* 2012;99:383–96.
20. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform* 2011;12:323.
21. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14:417–9.
22. Page JT, Gingle AR, Udall JA. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. *G3* 2013;3:517–25.
23. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. *BMC Genet* 2015;16(Suppl 2):S4.
24. Peralta M, Combes M-C, Cenci A, et al. SNIploid: a utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species. *Int J Plant Genomics* 2013;2013:890123.
25. Duchemin W, Dupont P-Y, Campbell MA, et al. HyLiTE: accurate and flexible analysis of gene expression in hybrid and allopolyploid species. *BMC Bioinform* 2015;16:8.
26. Khan A, Belfield EJ, Harberd NP, et al. HANDS2: accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. *Sci Rep* 2016;6:1–8.
27. Mithani A, Belfield EJ, Brown C, et al. HANDS: a tool for genome-wide discovery of subgenome-specific base-identity in polyploids. *BMC Genomics* 2013;14:653.
28. Akama S, Shimizu-Inatsugi R, Shimizu KK, et al. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid *Arabidopsis*. *Nucleic Acids Res* 2014;42:e46.
29. Kuo T, Frith MC, Sese J, et al. EAGLE: explicit alternative genome likelihood evaluator. *BMC Med Genomics* 2018;11:28.
30. Kuo T, Hatakeyama M, Tameshige T, et al. Homeolog expression quantification methods for allopolyploids. *Brief Bioinform* 2018;1–13. doi: [10.1093/bib/bby121](https://doi.org/10.1093/bib/bby121)
31. Wendel JF, Grover CE. Taxonomy and evolution of the cotton genus, *Gossypium*. *Cotton* 2015;25–44.
32. Joshi NA, Fass JN. Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33) [Software]. 2011. Available at <https://github.com/najoshi/sickle>.
33. Wu TD, Reeder J, Lawrence M, et al. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol* 2016;1418:283–34.
34. Paterson AH, Wendel JF, Gundlach H, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 2012;492:423–7.
35. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
36. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31:166–69.
37. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;9:357–9.
38. Du X, Huang G, He S, et al. Resequencing of 243 diploid cotton accessions based on an updated a genome identifies the genetic basis of key agronomic traits. *Nat Genet* 2018;50:796–802.
39. Bray NL, Pimentel H, Melsted P, et al. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–7.
40. Srivastava A, Sarkar H, Gupta N, et al. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* 2016;32:i192–200.
41. Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nat Biotechnol* 2008;26:897–9.
42. Foulds J, Boyles L, DuBois C, et al. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'13* 2013. arXiv:1305.2452 [cs.LG].
43. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27:861–74.
44. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol* 2011;2:37–63.
45. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;405:442–51.
46. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min* 2017;10:35.
47. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
48. Leng N, Dawson JA, Thomson JA, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 2013;29:1035–43.
49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;57:289–300.
50. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;21:3940–1.
51. Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction. *Math Intelligencer* 2005;27:83–5.
52. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Quart J Roy Meteor Soc* 2002;128:2145–66.
53. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS One* 2018;13:e0206312.
54. Li P, Piao Y, Shon HS, et al. Comparing the normalization methods for the differential analysis of illumina high-throughput RNA-Seq data. *BMC Bioinform* 2015;16:347.
55. McKenzie AT, Katsyov I, Song W-M, et al. DGCA: a comprehensive R package for differential gene correlation analysis. *BMC Syst Biol* 2016;10:106.
56. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 1915;10:507–21.
57. Serin EAR, Nijveen H, Hilhorst HWM, et al. Learning from co-expression networks: possibilities and challenges. *Front Plant Sci* 2016;7:444.
58. Horvath S. *Weighted Network Analysis: Applications in Genomics and Systems Biology*, 2011.
59. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* 2008;9:559.
60. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4 Article17.



61. Ballouz S, Weber M, Pavlidis P, et al. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* 2016;btw695.
62. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 2004;16:1679–91.
63. Liu S-L, Adams KL. Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the Brassicaceae. *Mol Biol Evol* 2010;27:2817–28.
64. Chaudhary B, Flagel L, Stupar RM, et al. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics* 2009;182:503–17.
65. Liang Z, Schnable JC. Functional divergence between subgenomes and gene pairs after whole genome duplications. *Mol Plant* 2018;11:388–97.
66. Oliver S. Proteomics: guilt-by-association goes global. *Nature* 2000;403:601.
67. Kyriakidou M, Tai HH, Anglin NL, et al. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci* 2018;9:1660.
68. Limborg MT, Seeb LW, Seeb JE. Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Mol Ecol* 2016;25:2117–29.
69. Mason AS. Challenges of genotyping polyploid species. *Methods Mol Biol* 2015;1245:161–8.
70. Motazed E, de Ridder D, Finkers R, et al. TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics* 2018;34:3864–72.
71. Zhang X, Wu R, Wang Y, et al. Unzipping haplotypes in diploid and polyploid genomes. *Comput Struct Biotechnol J* 2019.
72. Bourke PM, Voorrips RE, Visser RGF, et al. Tools for genetic studies in experimental populations of polyploids. *Front Plant Sci* 2018;9:513.
73. Blischak PD, Mabry ME, Conant GC, et al. Integrating networks, phylogenomics, and population genomics for the study of polyploidy. *Annu Rev Ecol Syst* 2018;49:253–78.
74. Jones G, Sagitov S, Oxelman B. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst Biol* 2013;62:467–78.
75. Voshall A, Moriyama EN. Next-generation transcriptome assembly and analysis: impact of ploidy. *Methods* 2019. pii: S1046–2023(18)30380–3. doi: [10.1016/j.ymeth.2019.06.001](https://doi.org/10.1016/j.ymeth.2019.06.001).
76. Payá-Milans M, Olmstead JW, Nunez G, et al. Comprehensive evaluation of RNA-seq analysis pipelines in diploid and polyploid species. *Gigascience* 2018;7:1–18.
77. Chen L-Y, Morales-Briones DF, Passow CN, et al. Performance of gene expression analyses using *de novo* assembled transcripts in polyploid species. *Bioinformatics* 2019;35:4314–4320.
78. Wendel JF, Jackson SA, Meyers BC, et al. Evolution of plant genome architecture. *Genome Biol* 2016;17:37.
79. Freeling M, Woodhouse MR, Subramaniam S, et al. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr Opin Plant Biol* 2012;15:131–9.
80. Edger PP, Smith RD, McKain MR, et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell Online* 2017;29:2150–67.