Where to publish

1. BMC Genomics (methodology)
2. BMC Bioinformatics (methodology)
3. Scientific Reports
4. GBE (letter)
5. BMC Research Notes

Challenges and pitfalls in the use of partitioned gene counts for homoeologous gene expression and co-expression network analyses

Hu G 1,2, Grover CE 1,2, Arick II MA, Meiling??, Peterson DG, and JF Wendel2

1 These authors contributed equally

2 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011

Introduction

Comparative transcriptomics has brought immeasurable power to many arenas of biology, providing the means to estimate rates of evolution [1–3], conduct phylogenetics [3–5], evaluate overall expression differences between species or populations (cit), uncover trait-related genes [6], and elucidate changes in entire gene networks [7,8]. The significant technical and analytical breakthroughs that have accompanied next-generation sequencing have facilitated both inexpensive and efficient comparisons among species whose lack of existing resources would have previously been an obstacle. Consequently, transcriptome data exist for over 1000 plant species (cit), many of those non-model species, and the number of transcriptomic projects continues to grow.

In the context of comparative transcriptomics, polyploid genomes offer both unique questions and considerations. Recent research suggests that polyploidy in plants is far more prevalent than once thought, acting both historically and more recently to shape plants genomes [9–12]. The vestiges of ancient polyploidy remain apparent in the genes retained [13], and the various effects of more recent polyploidization are becoming well-characterized for several species [14]. Among those are influences on gene expression, including gene loss [11,15–17], transposable element activation/proximity [15,18–23], heterochromatic state (cit), gene dosage (cit), gene regulatory changes [24–28], small RNA populations [29–34], maladaptive interactions among divergent homoeologs (cit?), and network interactions [7,8,35,36]. Changes in gene expression can be immediate or ongoing, biased toward one parental expression level or transgressive, and may exhibit an individual or genome-wide bias toward genome-specific homoeologs [27,28,37–43]. Furthermore, changes among genes and homoeologs may be coordinated in ways that are not completely transparent and which may depend on a host of factors, including homoeologous interactions, network position, gene dosage, and gene connectivity.

Research on expression level differences among polyploids, or between polyploids and their progenitors, can be divided into two broad categories: those that evaluate total expression and those that partition expression among homoeologs. The simplicity of and considerations for each analysis depends largely upon the age, origin, and composition of the polyploids themselves, as well as the resources available for each question being addressed; however, for recent polyploid species whose homoeologs have yet to diverge significantly, the two analyses become distinctly different. Assessing aggregate homoeolog expression for each homoeologous gene pair (or trio, etc) is relatively straight-forward, and may include either (1) summing the expression for individuals in a homoeologous group mapped to a sequenced polyploid genome/transcriptome, or (2) mapping to a sequenced model diploid progenitor, permitting reads from both homoeologs to accumulate on the same targeted gene. These analyses are useful in characterizing total expression level differences between model diploid progenitors and the polyploid species, which can be subsequently classified into their broad categories, (e.g., additive, transgressive, expression level dominant, etc; [27], others). Such analyses have been useful in characterizing patterns of gene expression changes in cotton [28] , coffee [42,44], Arabidopsis [45], Brassica [46], wheat [37], and others.

Expression partitioning among homoeologs becomes more complicated due to the intrinsically near-duplicate nature of homoeologs. As first discussed in Ilut et al (2012), the redundant nature of polyploidy violates the assumption of non-ambiguous mapping for RNA-seq quantification. Despite the ever increasing read-length for next-generation sequencing (NGS) technologies, the potential for ambiguous read mapping still exists and is largely dependent upon: (1) the divergence between subgenomes, (2) the sequencing strategy (read length, and paired- versus single-ended reads), and (3) the resources available to distinguish homoeologous reads. Due to the redundant nature of plant genomes, unique read assignment can be a problem for some genes in all species; however, the problem is magnified in recent polyploids due to the massively redundant nature of those genomes, particularly for those at higher ploidy levels. In addition, because diagnostic single nucleotide polymorphisms (SNPs) between polyploid parents may not be evenly distributed among genes and exonic regions (basic molecular evolution citation here), the proportion of identifiable homoeolog-specific reads will vary among genes in a given polyploid genome. The actual number of nucleotides assignable to specific homoeologs has led to the idea of “effective gene length”, i.e., how much of a gene is uniquely assignable ([47]; Figure 1), which is dependent both on the number of diagnostic SNPs in a given gene, as well as the average read or fragment length of the sequencing strategy.

Several strategies exist for partitioning homoeologous reads, each with its own limitations and practical consequences. PolyCat [48], HyLiTE [49], HANDS [50], and HANDS2 [51] all assign reads based on either an existing or *de novo* generated homoeoSNP index, whereas a similar program, PolyDog [52], relies on competitive mapping to parental or homoeologous genomes/transcriptomes. For those reads that do not cover a diagnostic homoeoSNP, however, these programs are unable to assign the read; the proportion of unassignable reads is determined by the program, the amount of information available for determining homoeoSNPs (e.g., diploid sequences), and the biological system itself (e.g., age of polyploidy and ploidy level). Several pipelines, such as RSEM [53] and Salmon [54], correct for this by using statistical models to infer the appropriate assignment ambiguous reads.

Given the complexities inherent in working with polyploid transcriptome data, the salient question becomes to what extent do these challenges affect our ability to derived accurate and meaningful conclusions from homoeolog partitioned transcriptomic data? That is, does the removal or inferred assignment of ambiguous data cryptically modify our illation regarding differential gene expression and coexpression network modification? One can imagine that the former may be inherently more robust to conserved data-loss, i.e., the same gene regions are removed for ambiguity from all samples, whereas analyses relying on intergene comparisons, such as coexpression networks analysis, may be perturbed by differential data loss due to uneven distribution of homoeoSNPs among genes. Here we explore the consequences of read ambiguity and disproportionate representation in polyploid expression datasets in the context of both differential gene expression analysis and network inference, using the cotton genus (*Gossypium*; Figure 0) as an example.

**Materials and methods**

*Data availability and preparation*

Cottonseed transcriptome data containing four developmental time points from the model diploid progenitors of polyploid cotton, *G. arboreum* (A2) and *G. raimondii* (D5), was downloaded from SRA (Table 1; SRA BioProject PRJNA179447) ~~and quality trimmed (QC) via sickle [55]~~. The number of reads was standardized between the two diploids. That is, the three replicates at each timepoint were compared between the A- and D-diploids for the number of reads passing QC, and the minimum number of reads was randomly removed from each diploid replicate to make the number of reads equivalent between the two species for each timepoint (Supplementary Table summary). For each timepoint and replicate, the adjusted diploid A2 and D5 reads were combined to generate *in silico* allopolyploid datasets (ADs) containing an equal number of reads derived from each diploid species.

Five methods for read mapping and homoeolog partitioning were evaluated, i.e., GSNAP-PolyCat [48,56], HyLiTE [49,57], RSEM [53,57], Salmon [54], and Kallisto [58]. Both GSNAP-PolyCat and HyLiTE were developed specifically to partition homoeologous reads given parental genomic and/or transcriptomic sequences as a reference, whereas the others were developed to address more generalized issues of multimapping reads (see below) and mapping speed.

Each pipeline was independently applied to the *in silico* polyploid cottonseed and flowering time datasets, each of which consists of “homoeologs” derived by mixing equal reads from each diploid parent. Parameters for each are summarized below, with detailed parameters given at GITHUB\_REPO. The number of expected reads per gene/homoeolog was assessed by mapping the diploid reads HOW?

*GSNAP-PolyCat*. This pipeline uses the SNP-tolerant capabilities of GSNAP to permit equivocal mapping of reads derived from different species/genomes to a single diploid genome using known SNP differences. PolyCat then uses the same SNP information to diagnose the origin of each read. Here, we mapped reads to the *G. raimondii* reference genome [59] using GSNAP [56] and a previously generated genome-diagnostic SNP-index [48] for SNP tolerant mapping. The resulting alignments were processed by samtools [60] and subsequently partitioned into genome-of-origin via PolyCat [48]. Read counts were tabulated via HTSeq [61] for each species/stage representing the number of reads mapped to the reference genome. ~~using: (1) no SNP index and the reference annotation; (2) a diploid-specific SNP index and the reference annotation; or (3) a diploid-specific SNP index in conjunction with PolyCat and the reference annotation to individually count A- and D-derived reads~~.

*HyLiTE*. The HyLiTE pipeline [49] operates by using data from diploid progenitors in conjunction with the polyploid to automatically identify SNPs diagnostic of parental lineages and partition reads. HyLiTE was applied to the original diploid and *in silico* polyploid data using WHAT PARAMETERS?. Briefly, HyLiTE uses Bowtie2 [57] to map reads to a reference diploid genome and subsequently classifies the polyploid reads based on the SNPs present in the diploid samples. These classified reads are automatically summarized into homoeolog-specific read count tables.

*RSEM*. RSEM [53] is a software package developed to accurately identify and assign reads that map to isoforms. For this software, a transcriptome containing all transcripts of interest must be available. While developed to facilitate isoform analyses, RSEM could also be used to assign reads to homoeologs, as exhibited in maize [35]. As with HyLiTE, RSEM uses Bowtie for read alignment, although it requires a reference transcriptome, which we constructed here using the reference diploid *G. raimondii*genome and the same SNP index [52] used above to derive homoeologous transcripts. RSEM was run using the Bowtie2 [57] option (default: Bowtie) and ANY OTHER PARAMETERS OR JUST DEFAULT?

*Salmon.* Salmon [54] employs a light-weight, quasi-mapping strategy [62] and a two phase estimation of expression with two forms of Bayesian inference [63,64] to first estimate and subsequently refine gene counts. A detailed description of the algorithm can be found here [54]. Salmon was run under default parameters with the same reference transcriptome described above and the option “keepDuplicates” for indexing the transcriptome.

*Kallisto*. Kallisto [58] belongs to the class of read aligners known as “pseudoaligners”, which leverage kmer information to detect the transcripts that could have generated a given read without aligning the read to the transcript specifically. Specifically, Kallisto employs fast hashing of read kmers to a provided transcriptome de Bruijn graph and quickly assign reads to transcripts based on kmer-based metrics. Here, Kallisto was run with default parameters using the same reference transcriptome provided to RSEM and Salmon (above).

*Assessment of homoeolog expression estimates*

Several measures were implemented to evaluate the performance of homoeologous read assignment. Since the *in silico* polyploid reads were derived from combining diploid reads, we expect the number of reads per homoeolog, i.e., AT and DT, to be equivalent to the number of reads for that gene in the diploid source, *G. arboreum* (A2) and *G. raimondii* (D5), respectively. In many cases, the AT and DT reads are distinguishable by homoeoSNPs; however, the programs differ in their treatment of ambiguous reads. For two pipelines, GSNAP-PolyCat and HyLiTE, ambiguous reads are discarded; for the remaining three (i.e., RSEM, Salmon, and Kallisto), the distribution of ambiguous reads is statistically determined [53,54,58]. Therefore, we first evaluated the ***Efficiency***(***Ef***) of read assignment for GSNAP-PolyCat and HyLiTE, which is simply the proportion of reads assigned over the total number of reads. Due to statistical inference of ambiguous read assignment, the remaining three all have ***Ef*** ~ 1**.** We approximated the error in read assignment (***Discrepancy; Di***) by comparing the number of reads recovered versus the number expected (see Table 1 for abbreviations). Here ***Di*** for each subgenome is given by ***Ac(A)* =**  and ***Ac(D)* =** , which is written generally ***Di* =** .

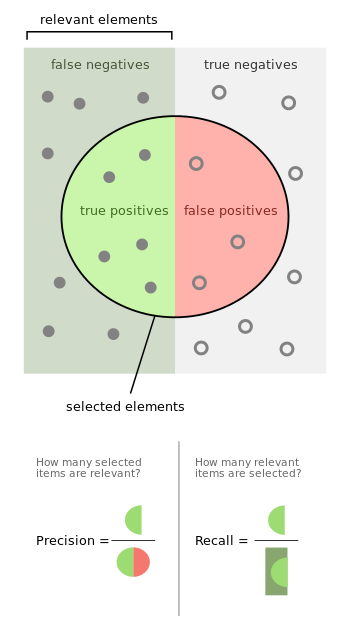
Since our test data are tetraploid and therefore should contain only two classes of reads, we also employed two statistical measures which evaluate the quality of binary classifications, i.e., Precision/Recall [65] and the Matthews correlation coefficient [66]. Precision/recall is used in statistics to evaluate, respectively, (1) the fraction of true positives recovered from all that were retuned positive, and (2) the fraction of true positives recovered from all possible true positives (also known as sensitivity or the True Positive Rate, TPR). Their harmonic mean, or F1 score, provides a generalized measure of accuracy. The general formulas are given as: ***Precision* = , Recall =** , and **F1 =** (TP/FP = true and false positives, TN/FN = true and false negatives).

Here, where the abbreviations (e.g., At.A, D5.A) correspond to “source - mapped to - subgenome”

Table 1: Abbreviations

|  |  |  |  |
| --- | --- | --- | --- |
| Reads Accurately Assigned |  | At.A A2.A |  |
| Reads Incorrectly Assigned |  |  |  |
| Observed |  |  |  |
| Expected |  |  |  |
| True Positive | At.A, Dt.D |  |  |
| False Positive | At.D, Dt.A |  |  |
| True Negative | At.A, Dt.D |  |  |
| False Negative | At.D, Dt.A |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

, i.e., *Efficiency* (***Ef***), *Accuracy* (***Ac***), and *Discrepancy* (***Di***). Therefore, we can consider the total expression per gene (or homoeologous gene set) to be represented by T = A2 + D5 = AT + DT + N. *Efficiency* measures the proportion of total reads assigned to either homoeolog in the polyploid, where ***Ef* =** and is agnostic with respect to the correctness of the partitioning. The assignment *Accuracy* measures the ability to assign reads to the correct genome, *in the absence of confounding (i.e., homoeologous) reads*. We assessed the *Accuracy* by individually running each diploid dataset with the homoeologous reference generated above and evaluating the number of reads correctly assigned, and calculated the combined accuracy as ***Ac* = .** Here, A.obsA and D.obsD are reads correctly assigned (i.e., observed), while A.obsD and D.obsA are misidentified reads. The individual accuracy for each subgenome is given by ***Ac(A)* =**  and ***Ac(D)* =** . The expected values are calculated represented by the sum of the reads correctly *or* incorrectly identified. This is an important distinction from the true number of expected reads because it ignores the rate of nonidentification (i.e., the number of N reads) for HyLiTE and GSNAP-PolyCat, which do not partition ambiguous reads.



• The x-axis showing recall (= sensitivity = TP / (TP + FN))   
• The y-axis showing precision (= positive predictive value = TP / (TP + FP))

The final metric, *Discrepancy*, is a measure of the absolute deviation from the expected counts given perfect efficiency and accuracy (***Di*** = ). In addition to overall measures aggregated for both homoeologs, separate assessments for At and Dt were performed accordingly (Supplemental Table ?). Use Correlations among variables and Multiple Logistic Regression tests to show statistical evidences for the association between performance metrics and explanatory variables. R scripts for performance assessment and below analyses are available at https://github.com/???.

**Supplemental table ?: Definition of performance measures of homoeolog expression estimation.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Aggregated | *At only* | *Dt only* |
| *Efficiency* (*Ef*) |  |  |  |
| *Accuracy* (*Ac*) |  |  |  |
| *Discrepancy* (*Di*) |  |  |  |

*Gene expression analysis*

Two methods were independently used to analyze differential expression (DE) of homoeologs for each pipeline, DESeq2 [67]and EBSeq [68]. While both methods use a negative binomial model to estimate gene-wise dispersion parameters, DESeq2 takes a classical hypothesis testing approach to report nominal p-values, whereas EBSeq accommodates the uncertainty inherent in differential expression (DE)of isoforms (here, homoeologs) using a Bayesian framework and returning posterior probabilities for equal or differential expression. For both methods, the “true” differential expression of homoeologs is represented by the DE of those genes in the parental diploids *G. arboreum* (A2) and *G. raimondii* (D5), which are the source of the synthetic polyploid reads (i.e., homoeologs) here. Here, DE was reported using a false discovery rate α=0.05, with the DESeq2 p-values adjusted using Benjamini–Hochberg [69] and the posterior probability of DE by EBSeq was considered >1- α. Sensitivity (true-positive rate) and specificity (true-negative rate) were calculated for each method. All methods can be found at <https://github.com/Wendellab/homoeologGeneExpression-Coexpression>.

*Gene co-expression analysis*

For data transformation, RPKM followed by log2 transformation were performed with local R script, and regularized logarithm *rlog* transformation were conducted with R package DESeq2 [67]. Based on transformed expression profiles, Pearson correlation coefficient was calculated for each pair of genes, and Fisher’s z-test was used to identify significant differential correlations (DC) with *P* < 0.05. DC gene pairs were classified as having gain of correlation (GOC) or loss of correlation (LOC) against the reference condition. Briefly, based on a threshold of *P* < 0.05 for correlation significance and the sign of correlation in each condition, gene-gene correlations in each condition can be categorized into 3 classes, i.e. significant positive correlation (+), no significant correlation (0), and significant negative correlation (-); thus, a total of 9 classes can be inferred between two conditions. For example, class 0/+ indicates GOC from a not significant correlation (0) to a significant positive correlation (+), and class +/- indicates LOC from a positive correlation to a negative correlation. Identification and classification of DC genes pairs were performed using the R package DGCA [70]. To determine if a specific DC class is significantly enriched, one-sided Fisher's exact tests was applied to evaluate the overlap between significant (*P* < 0.05) and all gene pairs classified into this class.

Differentially co-expressed genes were identified as those significantly over-represented by DC gene pairs. Briefly, the percentage of DC among all possible gene pairs measures the extent of differential co-expression, or the probability *p* to detect “a differential co-expression gene pair”. For a gene observed in *k* differential correlation pairs among all possible pairs *n*, the probability *P* of a “differential co-expression gene” follows the binomial distribution model as follows:

*P* was corrected by the BH method [69] at α = 0.05 for identifying differentially co-expression genes.

*Co-expression network construction*

The construction of a co-expression network begins with a measure of similarity score from pairwise expression profiles for each pair of genes, and the next step is converting the similarity scores to an adjacency matrix that reflects connections between gene nodes in network [71]. By applying a threshold to similarity scores to determine whether a pair of gene should be connected in the network, an adjacency matrix containing only 0 and 1 values is resulted as a binary network. Instead of hard thresholds, weighted gene co-expression network analysis (WGCNA) applies a soft threshold chosen by the scale-free topology criterion to assign a connection weight to each gene pair, which allows the adjacency matrix to present network connections quantitatively [72,73]. Using Pearson correlation coefficient to determine similarity scores, we constructed a series of binary and weighted network for each expression dataset generated from three read estimating pipelines followed by log2 RPKM or *rlog* data transformation. Two types of hard threshold were used to generate binary networks – rank based and Fisher’s Z statistics based thresholds. Based on the ranking of similarity scores, a set of cutoffs (5%, 1%, 0.5% and 0.1%) were applied to select top ranked connections as edges in network. Fisher’s Z transformation normalizes the distribution of Pearson correlation and converts each correlation coefficient to a z-score for significance test. A set of Z cutoffs (1.5, 2.0, 2.5, 3.0) were applied to select significant correlations as edges in network. To construct weighted networks, we also used a test of soft thresholds (0, 12, 24) to test the effect of thresholding.

Binary network construction is a decision problem as the inference algorithm decides whether an edge exists (positive) between a pair of genes or not (negative). Treating the edges inferred from true homoeolog expression data as true label, we can evaluate the performance of binary network construction to recover true edges from estimated homoeolog expression data with a Receiver Operating Characteristic (ROC) curve. Due to the large gene number of polyploid networks (> 60,000 genes), a 10% random sampling of genes was used to compute ROC curves and AUC (area under receiver characteristic curve) measures with 10 repeats. AUC of 0.5 indicates a completely random guess, 0.7 is generally good and 0.9 or above is very high.

*Network topology and functional connectivity assessment*

Node connectivity (*k*) measures the amount of connections to a given node in network, which is calculated as the summation of connected edge numbers or edge weights. In terms of functional connectivity of the network, the ‘Guilt-by-Association’ principle suggests that genes of similar functions are preferentially connected in network. Therefore, a neighbor voting algorithm can be used to classify genes as belonging to a given functional group according to the functionality of their connected genes (i.e. neighborhood). Based upon the known functional labels of genes (here we used GO and KEGG annotations), the voting algorithm acts as a binary classifier to return TRUE or FALSE predictions of functional labels, whose performance can be assessed with a Receiver Operating Characteristic (ROC) curve. The derived metric of AUC (area under receiver characteristic curve) 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 *Gossypium raimondii* reference genome downloaded from Phytozome ([www.phytozome.net/)](http://www.phytozome.net/)). Functional connectivity assessment was performed with R package EGAD [74].

Results

*Polyploid divergence and read ambiguity: the problem*

The idea that read ambiguity could affect the outcome of homoeolog partitioned RNA-seq data was first considered in soybean, which underwent allopolyploidization circa 5 mya [47]. In that case, the data available were Illumina 36 nucleotide (nt) reads. Given the ulta-short nature of these reads, one can imagine that individual gene coverage by homoeolog partitioned reads (i.e., those reads containing identifiable SNPs) could be low and fragmented relative to the full gene region. Sequencing technologies are trending toward both longer and paired reads, which could alleviate some of the read loss; however, the effects of SNP density and distribution among homoeologs remain pertinent. Consider a scenario where we have the standard 100 nt, single end sequencing reads for an allotetraploid and its model diploid progenitors, whose average exonic divergence is 1.5% and which ranges 0% (no diagnostic SNPs) to 3%. For a given 1 kb homoeologous transcript pair, each covered 10-fold by these 100 nt reads, the number of reads retained (i.e., its presumed relative expression level within that cohort) depends both on SNP number and distribution (Figure 1). For those transcripts containing a single diagnostic SNP (0.1% divergence; Figure 1A), only 20% of the total transcript length (approximately double the read length) will be represented by each homoeolog and only 10% of the reads will be unambiguously mapped. As nucleotide divergence between homoeologs increases, both the proportion of homoeologous transcript covered and the percentage of reads mapped increases at a rate determined by the distribution of the diagnostic SNPs. If, for example, the diagnostic SNPs are evenly distributed (Figure 1B, C) the entire length of each homoeologous transcript is completely (or nearly completely) represented at 0.5% divergence. The read coverage, however, is still reduced by ~50% given that the distance between SNPs (approximately 200nt) is half the read length. That is, because each nucleotide is covered 10x, we can expect that 50 out of the 100 total reads (to give 10x) will contain a diagnostic SNP (5 SNPs x 10 read coverage/nt = 50 reads). At 1% divergence, we would expect near complete read retention if the homoeologous SNPs are evenly distributed due to the read length being approximately equivalent to the distance between SNPs.

This is somewhat reassuring for many polyploid species with a modest amount of divergence, e.g. cotton (citation), Brassica (citation), coffee (citation). If you consider the above, where mean =1.5% divergence over a range of 0% - 3%, then the proportion of genes with near-complete partitioning of homoeologous reads should be approximately 84%. To examine this using biological data, we evaluated percent divergence between the two model diploid progenitors of polyploid cotton (Supplementary Table 1) which diverged approximately 5-10 million years ago (mya). The distribution of divergence for all cotton homoeologs is notable centered on approximately 1.5% median divergence with a slight right tail (Figure 2), which may be attributable to errors in either gene annotation (i.e., including the faster involving intergenic space in gene bodies) or the SNP index (e.g., paralogs, sequencing errors, etc). Indeed, restricting the range in divergence to 0 – 3% reduces the mean divergence from approximately 2% to the median of 1.5% and produces a nearly-bell shaped curve (Figure 2) with a slight left skew, the latter evident both from the graph and the summation that only 78.4% of genes are >=1% different in exon sequence versus the expected 84%.

The proportion of reads retained per gene, however, tells a more complicated story. Given the logic above, we would expect those genes with >1% divergence to have near complete homoeologous read assignment with 100 nt reads; however, biologically-derived data from cotton indicate that uneven distribution of SNPs is both common and of consequence to read assignment. RNA-seq from four time-points in cottonseed development was mapped to the genome, partitioned into homoeologous reads by polyCat, and subsequently filtered for a minimum number of reads (n=10). Using only those genes with homoeoSNP divergences between >0 to 3%, only about 15% of genes have near complete (i.e., >90%) homoeologous read assignment. Most genes exhibited at least 60% homoeologous read assignment, although the proportion of unassigned reads is higher for each category than expected given even coverage. This may be due to multiple reasons, most prominently the uneven distribution of SNPs within genes (Figure 1D, E). Indeed, nearly 25% of the exons in the published cotton D-genome do not contain a diagnostic SNP (45,868 out of 188,048 exons), and 10% of homoeologous SNPS (61,388 out of 634,797) are spaced greater than 100nt apart (Figure 4). Consequently, individual gene representation in the transcriptome pool is dictated by the interaction of the divergence parents (prior to polyploidy) and the distribution of the divergence within and among genes.

*Artificial datasets permit assessment of fidelity in homoeologous read assignment*

We evaluated the (reality) of homoeologous read assignment using several artificial polyploid datasets generated by combining equal numbers of parental reads (here, diploid A2 and D5) and whose “homoeologous” expression ratios should directly mirror the expression of the diploids from which they were derived. SUMMARY OF HOW THESE WERE COMBINED, WHICH LIBRARIES GENERATED THEM, ETC. In each case, we generated three *in silico* polyploid datasets to serve as replicates in evaluating the effects of either (1) non-identification or (2) misidentification of ambiguous reads.

Four pipelines for read partitioning were individually applied to each dataset and assessed for their *Efficiency* in partitioning reads into homoeologs, their *Accuracy* in assigning those reads, and the resulting *Discrepancy*, i.e., the absolute deviation from expected counts (see methods). Two of these pipelines, GSNAP-PolyCat [48,56] and HyLiTE [49,57], were specifically developed to facilitate homoeologous read assignment in polyploid systems; however, the challenge of read assignment and ambiguity is not unique to polyploidy. The RSEM pipeline [53,57] was developed as a method for estimating expression of gene isoforms and was reported as a promising approach to study duplicate gene expression in maize [35]. Salmon [54], on the other hand, addresses the general challenge of read multimapping estimate read counts and abundance uncertainty. Using the artificially created *in silico* datasets described above, we evaluate the performance of these four pipelines and the consequences of their deviation from expected on downstream analyses, such as gene expression and coexpression.

We first evaluated capability of each pipeline to recognize and partition polyploid reads, i.e., the *Efficiency*, based on the identification of diagnostic homoeolog SNPs (homoeoSNPs). Both RSEM and Salmon report the highest *efficiency* at 100% read assignment; however, this level of *Efficiency* is achieved by inferring the assignment of ambiguous reads based on the distribution of SNP-assignable reads [53,54]. Between the two polyploid-specific pipelines which rely on the presence of diagnostic homoeoSNPs, GSNAP-PolyCat exhibits a significantly higher *Efficiency* than HyLiTE (Table 1; ***E***=87.7% and 82.2%, respectively; student’s T test p<0.05).

Next, the *Accuracy* of read assignment to the true genome of origin was assessed. Here, GSNAP-PolyCat again out performs the other two pipelines with over 99% correct read assignment (Table 1), followed closely by RSEM and Salmon, both with 98% *Accuracy*. Unlike HyLiTE (***A***=87.3%), these three pipelines use existing homoeoSNP information in assigning reads; however, it is worth noting that HyLiTE does provide the option to include additional genomic information to improve its on-the-fly SNP calling [49]. The prebuilt SNP index used with GSNAP-PolyCat [52] is an ongoing resource derived from rich genomic resequencing and therefore represents the best…

The final metric, *Discrepancy*, measured the absolute difference between observed read assignments and the number expected, which is influenced by both unassignable reads and incorrect assignment. Both RSEM and Salmon exhibited the lowest *Discrepancy* (***D***=5.1%), i.e., overall deviation from the expected number and partitioning of reads. This low measure of *Discrepancy* is largely due to the complete *Efficiency* guaranteed by the RSEM and Salmon algorithms and solely reflects the *Accuracy* metric. GSNAP-PolyCat (***D***=14.5%) exhibited lower *Discrepancy* than from HyLiTE (***D***=22.1%), as expected from the better performance of GSNAP-PolyCat in both *Efficiency* and *Accuracy*.

**Table 1.** Overall and sub-genome assessment of homoeolog expression estimation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **GSNAP-PolyCat** | **HyLiTE** | **RSEM** | **Salmon** |
| *Efficiency* (***E***) | 87.7% | 82.2% | 100.0% | 100.0% |
| - At only | 86.7% | 78.5% | 101.0% | 101.6% |
| - Dt only | 88.7% | 85.8% | 99.1% | 98.5% |
| *Accuracy* (***A***) | 99.4% | 87.3% | 98.0% | 98.0% |
| - At only | 99.6 % | 88.0% | 97.8% | 97.5% |
| - Dt only | 99.2% | 86.8% | 98.2% | 98.4% |
| *Discrepancy* (***D***) | 14.5% | 22.1% | 5.1% | 5.1% |
| - At only | 15.5% | 27.4% | 5.4% | 5.1% |
| - Dt only | 13.6% | 17.3% | 4.8% | 5.1% |

*Technical and biological variables can confound homoeologous read identification*

Besides the technical difference of analytic pipelines, many experimental and biological biases are known to affect the assignment of homoeolog-specific reads, including dependencies on library preparation and sequencing platform, sequencing depth, gene lengths, transcript expression levels, the abundance and distribution of diagnostic SNPs, as well as various technical and biological limitations to infer “true” diagnostic SNPs dependent on the polyploid system being studied (e.g. lack of diploid model, ancestral divergence between progenitor genomes, asymmetric evolutionary rate, gene conversion, etc. These are beyond the scope of this analysis. Do we need to further elaborate and where to?). Using the gene-wise measures of homoeolog expression performance (*Efficiency*, *Accuracy* and *Discrepancy*) as dependable variables, we examined the impact of sequencing type (PE vs SE), gene length, transcript expression, SNP abundance and distribution, and found …fill in Meiling’s anslysis results.

Explanatory variables:

* **geneLenM**: gene length
* **percentageEffectM**: percentage of gene regions that are diagnostic of homoeolog origin; this measure is determined based on homoelog-specific SNPs distribution and RNA-seq read length.
* **expression**: total read count (T), reflecting the combined effects of sequencing depth, expressed transcript abundance. *Given the strongly skewed distribution, I aslo made log2 tranformation of expression*.Comparing performance of two homoeologous expression esti*mation pipelines*

Use Correlations among variables and Multiple Logistic Regression tests to show statistical evidences for the association between performance metrics and explanatory variables. All factors listed seem to be significant according to multiple linear regression. Is it possible to conduct prediction? Based on prediction, we can propose a new gene filter to remove poorly partitioned homoeolog gene pairs.

*Homoeolog read assignment has little effect on the amount of differential expression*

A major interest in analyzing polyploid transcriptomes is to identify expression divergence between homoeologous genes, which has become an established approach to study homoeolog expression bias and to infer functional divergence of duplicated genes [27,28]. To evaluate the extent to which estimated (partitioned) homoeolog read counts can faithfully represent the expression divergence exhibited by actual homoeolog expression levels, we conducted differential expression (DE) analysis for the synthetic polyploid datasets using *true* and *estimated* homoeolog expression levels. In combination with the three expression estimating pipelines, two methods of DE analysis were tested as a possible source of variance to infer expression divergence between homoeologs (Figure 1. Workflow). With 11.1% to 44.1% of differential expression detected across sampling conditions (supplemental “s3.DE.summary.txt”), no significant difference was observed between the *true* and *estimated* numbers of DE genes (this is different from seeds-only results, where “fewer DE genes were identified from estimated homoeolog reads counts than those inferred from true expression levels, leading to an approximately 10% under-estimation of homoeolog expression divergence associated with the process of homoeolog read estimation” ) (Analysis of variance followed by Tukey’s HSD post-hoc test; supplemental “s3.DE.performance.pdf” first 2 pages + anova results in chart). By examining the overlap of *true* and *estimated* lists of DE genes, we evaluated the resulted homoeolog expression divergence in terms of sensitivity, specificity, and AUC. As shown in Figure 2 (“s3.DE.performance.pdf” last 3 pages + anova results in chart), Gsnap-Polycat significantly outperforms the other two pipelines. Although similar levels of sensitivity were seen for two DE methods, EBSeq exhibited a higher level of specificity and AUC than DESeq2 (Figure 2). Our results indicated that with default parameter setting and FDR controlled at 5%, DESeq2 analysis are more susceptible to Type I error than EBSeq (i.e., false positive detection of DE).

Some discussion about DE method choice:

[*https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-91*](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-91)

*more about how to dicuss method differencehttp://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176185*

*Homoeolog read assignment may lead to an underestimation of homoeolog expression divergence*

Distinct from DE analysis, differential co-expression or correlation (DC) operates on the level of gene pairs rather than individual genes, which directly measures how expression relationships between genes are rewired across conditions. For each pair of homoeolog genes, we performed DC analysis between the *true* and *estimated* expression profiles, and compared the changes of co-expression relationship associated with different estimating pipelines in conjunction with data transformation methods (log2RPKM and *rlog*). With the lower percentage of co-expression changes, GSNAP\_PolyCat (log2RPKM - 0.15%, *rlog* - 0.10%) and Bowtie2\_HyLiTE (log2RPKM - 0.22%, *rlog* - 0.12%) outperform Bowtie2\_RSEM (RPKM - 3.10%, *rlog* - 2.29%). Considering 1% false discovery rate in identifying DC between homoeologous genes, only the Bowtie2\_RSEM pipeline introduced significant errors (Figure 3, “s5.DC.classes.homoeolog.pdf”, add total percentage to left). Among its misidentified DC relationships, over 80% exhibits a gain of co-expression (GOC) pattern of 0/+; that is, these homoeolog gene pairs were detected with significant positive correlation from the *estimated* expression profiles while no such correlation could be inferred from their *true* expression profiles. Despite that small numbers co-expression changes were found in GSNAP\_PolyCat and Bowtie2\_HyLiTE datasets, this pattern remains significantly enriched (Fig. 3). It is suggested that the process to estimate homoeolog read counts could lead to a significant change in co-expression relationships between homoeologs, which is characterized by an over-estimation of positive correlations, thereby impairing our ability to infer the functional divergence of duplicated genes. For example, a function related example would be convincing.

In addition to homoeolog gene pairs, the identification and classification of DC was also conducted for all possible gene pairs between the *true* and *estimated* expression profiles, resulting in less than 1% of global pairwise co-expression changes. Nothing too interesting here.

**Table ?.** Percentage of differential co-expressed gene pairs.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **GSNAP-PolyCat** | **HyLiTE** | **RSEM** |
| RLD | 0.997% | 0.483% | 0.866% |
| RPKM | 0.778% | 0.411% | 0.766% |

Differential co-expressed genes can be detected as enriched of changed gene pairs. As shown in table below, the highest number of DC genes were resulted from HyLiTE, and lowest from RSEM; this pattern doesn’t make much sense. Taking a closer look of DC genes, more At then Dt genes were found, and DC genes appear to show lower expression and smaller percentage effective region.

**Table ?.** Genes enriched with differential co-expressed gene pairs.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **GSNAP-PolyCat** | **HyLiTE** | **RSEM** |
| RLD | 6980 (10.1%) | 8880 (13.7%) | 5893 (8.6%) |
| RPKM | 7282 (10.5%) | 8696 (13.5%) | 5940 (8.4%) |

*Co-expression network construction*

What network c

*Rank based method is more robust than Z statistics based method to infer binary co-expression network of homoeologs*

The construction of an un-weighted network requires a binary classifier (threshold) to decide whether there is an edge between each pair of genes. Two types of thresholds were chosen for testing - rank based and Fisher’s Z statistics based cutoffs. Based on ROC and AUC:

* Datasets processed from Hylite and PolyCat mapping pipelines are more robust (HyLite better than polycat too) to infer binary edges than those from RSEM: higher AUC with smaller variance
* Correlation rank based methods outperform Z-score based methods:
* Within D sub-genome connections were inferred with higher AUC than those within sub-genome or cross subgenome connections. This bias in performance is likely to be introduced by mapping against D genome reference.
* Top 0.5% and Z=1.5 appear to be better

*Weighted co-expression network inference outperforms binary network inference to*

WGCNA networks were constructed using soft threshold 1, 12 and 24.

**The inference of node connectivity and functional connectivity is quite robust, not significantly affected by homoeolog-specific read assignment.**

**Node connectivity (k) –** the amount of connections to a given node in network.

**Functional connectivity (FC) -** the degree to which genes of similar functions are preferentially connected in network.

According to “Guilt-by-Association” principle, higher FC suggests more reasonable network topology

1. AUCs
2. Correlation of functional connectivity between expected and observed dataset node dregress
3. Correlation of node degree between A vs D

*5. Assessment of network topology and functional connectivity*

* Node degree or node connectivity (k) – summation of all connections (binary-1, weighted- numeric edge weight) to a given node.
* Node maximum adjacency ratio (MAR) - measures whether a node has connectivity because of many weak connections (small) or because of strong but few connections (high).
* Functional connectivity (FC) - For thousands of functional gene sets (GO, KEGG), we evaluate the degree to genes sharing similar functions are connected in network. Further compared between subgenomes, and with respect to expected dataset.
* Module membership (Mm) – Given module structure derived from the expected dataset, we applied EGAD algorithm to evaluate the probability that the observed dataset can rebuild the expected modules. This was tested for At and Dt separately, followed by comparison to inspect any subgenome bias. (I am hoping to see all high values, boxplot)

6. comparing diploid to polyploid

Co-expression analysis

To compare the co-expression gene networks (CGNs), we first characterize CGNs with several graph variable including …, and then applied a multivariate approach to the matrix of graph variables based on Principle Component Analysis (PCA). Leal et al (2014) showed that the closeness of CGNs projected on the principle component space is an indicative of similarity among CGNs. of

Leal et al. (2014), Construction and comparison of gene co-expression networks shows complex plant immune responses. PeerJ 2:e610; DOI 10.7717/peerj.610 [75]

Whether the topological properties of network modules are reproducible?

We calculate several module preservation statistics as described in [76] to determine whether the properties of the *true* network are preserved in the observed network.

Traditional cluster validation (or quality) statistics can be split into four broad categories: cross-tabulation, density, separability, and stability statistics

* Cross-tabulation based module preservation statistics (both ref and test module assignments needed) – mp$accuracy
  + Co-clustering: the proportion of pairs of members in ref module q that also cluster in test module q’
  + Accuracy and the related Fisher exact test p-value: the proportion of ref module members that are present in the corresponding test module, which has the highest number of objects common to both the reference and the test module.

3 types of network based module preservation statistics: 1) density based, 2) separability based, and 3) connectivity based preservation statistics.

* Density based preservation statistics - whether module nodes remain highly connected in the test network
  + Mean adjacency
  + Mean correlation matrix
  + Mean KME
  + propVarExpl
* Connectivity based preservation statistics －whether the connectivity pattern between nodes in the reference network is similar to that in the test network.
  + ( cor: kIM , cor: kME, cor: kMEall , cor: cor), and between the
* Separability statistics - whether network modules remain distinct (separated) from one another in the test network
  + separability: ave, separability: ME).

*Differential Gene Expression with Effective Length*

Thoughts about analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **set** | **Reference (exons only)** | **SNP-index** | **PolyCat** | **What to expect for mapped reads?** |
| 1 | Full-length gene model | No | No | Without considering At SNPs, Dt and N reads were mapped to D5 reference. |
| 2 | Full-length gene model | Yes | No | Considering SNPs, At. Dt and N reads were mapped, giving their aggregate total. |
| 3 | Full-length gene model | Yes | Yes | Mapped and partitioned. |
| Comparisons among 1-3 check the performance of mapping and reads partitioning, providing some foundation for further discussion of 4 and 5. | | | | |
| 4 | Effective-gene-length model: theoretical | Yes | Yes | Mapped, partitioned, and corrected. |
| 5 | Effective-gene-length model: empirical | yes | Yes | Mapped, partitioned, and corrected. |
| Comparisons among 3-5 check the performance of effective length correction. | | | | |

Analysis of A2 diploid

* Lists of differential expression during fiber development can be compared.
  + My expectation is set 1 is most different from 2 and 3, because the A-genome specific reads were not well mapped.
  + If set 2 (un-partitioned reads representing At + N) and set 3 (use partitioned At reads) provide similar results, that means DE analysis is quite robust, not affected much by read counts (or say expression levels). This is expected according to the algorithm of DESeq, which correct significant threshold by expression levels.
  + If above is true, comparisons among set 3-5 should not result into much differences either. Again, gene-by-gene DE analysis should be robust.
* WGCNA Network analysis of seed development for each set
  + Set 2 network should be the “real” network, based on all reads mapped to each gene. The accuracy and performance can be evaluated by comparing to set 2 network.
  + I expect set 1 to give the worst network, due to the mapping problem
  + Set2 vs set3 will let us know whether and how the variation in SNP content and distribution affect co-expression relationships.
  + Set4 and set5 each compared to set2, see if either one is closer to the real set2 network than set3. See if we have a clear winner, or both work fine. But if neither works better than set 3, we will have new questions.

Analysis of *in silico* allopolyploid (ADs)

* Equal amount of A2 and D5 reads were combined for each sample (per dpa per rep) to generate this *in silico* allopolyploid seed transcriptomes. Given that the “real” subgenome transcriptomes can be derived from A2 and D5 diploid transcriptomes, we can examine which method re-construct the subgenome transcriptomes At and Dt closest to “real” results.
* Differential expression
  + Should we look at developmental changes as for A2 diploid, or check for homoeolog bias (A2 vs D5 and At vs Dt)? The formal question doesn’t require homoeolog partition, not much difference from the A2 example above. The latter one shouldn’t matter that much for comparing homoeolog expression, but we can check and just say one sentence about it. And then focus on network analysis below?
* WGCNA co-expression network analysis, 74000 genes X 12 samples
  + Set2 for A2 and D5 diploid reads for constructing “real network”.
  + Set 3-5 for At and Dt reads for constructing ADs netwroks.
  + Which ADs network is best representing the “real” network”.