**Possible target journals**

* Scientific Report - 5.228 <http://www.nature.com/srep/publish>
* BMC Genomics: METHODOLOGY ARTICLE - 3.867 <https://bmcgenomics.biomedcentral.com/submission-guidelines>
* BMC Bioinformatics – 2.435 <https://bmcbioinformatics.biomedcentral.com/submission-guidelines>
* GBE – 4.098: letter <https://academic.oup.com/gbe/pages/General_Instructions>

**Title**

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

**Background**

Although numerous tools have been developed for co-expression network analysis originally based on microarray datasets, reasonable practices remain poorly defined for RNA-seq co-expression analysis.

Our focus is to bring attention to challenges and technical biases

During RNA-seq read mapping for polyploid species, it is challenging to differentiate and assign reads between homoeolgous genes. Although many tools have been developed for this purpose, the factors affecting their performance and downstream expression and coexpression analysis remain poorly understood.

**Outline of main results**

Technically

* Method choice of estimation pipeline matters
* Genetic features and expression variation matter too, can we make prediction??
* Small biases are introduced into network inferences
* WGCNA out performs binary network approaches

Evolution

**Evolutionary dynamics of duplicated gene networks in allopolyploids**

* **Allopolyploidization increases co-expression connections between sub-genomes.**
* **No global bias between the two sub-genomes in terms of gain or loss of edges.**
* **Holds promise for understanding regulatory architecture of complex genomes and phenotypes**

Seed biology

* Increased connection by domestication mainly contributed by D –genome gain of connectivity

**Results and Discussion**

To assess homoeolog-based analyses in a controlled setting, we constructed an *in silico* polyploid dataset (ADs) by combining equal amounts of parental diploid A2 and D5 reads, whose orthologous gene expressions thus represent the true values of homoeolog expression in ADs. From raw RNA-seq datasets, we compared three RNA-seq read mapping and homoeolog expression estimating pipelines in combination with different method choices at multiple stage of our analysis as illustrated in Figure 1. as a possible source of variance for downstream analyses. Once the best combination determined we analyze real allopolyploid networks (TM1 and Yuc) in contrast to the synthetic ADs, with variance from technical in mind, to undercover the network rewiring accompany polyploidy. trueADs vs obsADs

*Method choice, genetic and expression variation of homoeolog gene pairs confound the assignment of homoeolog read counts*

Begin with read mapping and characterizing homoeolog-specific reads, we assessed the performance of three pipelines, among which GSNAP-PolyCat and Bowtie2-HyLiTE were developed for polyploid systems ([Page et al. 2013](#_ENREF_5); [Duchemin et al. 2015](#_ENREF_1)), while Bowtie2-RSEM, as a tool specialized at estimating gene isoform expression levels, was also reported as a promising approach to study duplicated gene expression in maize ([Li et al. 2016](#_ENREF_4)).

We first evaluated the pipeline capability to recognize polyploid reads containing diagnostic homoeolog SNPs, and showed that GSNAP-PolyCat exhibits a significantly higher *Efficiency* than Bowtie2-HyLiTE (Table 1; ***E***=87.7% and 82.2%, respectively; student’s T test p<0.05). Instead of reporting homoeolog-specific read counts based on alignment, Bowtie2-RSEM outputs the maximum likelihood estimates and inherently ensures 100% assignment *Efficiency* of totalreads mapped to homoeologous genes.

Next, the *Accuracy* of read assignment to the true sub-genome origin was evaluated to compare the pipeline performance; GSNAP-PolyCat out performs the other two pipelines with over 99% correct read assignment (Table 1). Noting that GSNAP-PolyCat and Bowtie2-RSEM both used a prebuilt homoeolog-specific SNP index from rich genomic resources ([Page and Udall 2015](#_ENREF_6)) to infer sub-genome origin, the lowest accuracy of Bowtie2-HyLite (***A***=87.3%) is most likely attributed to its built-in procedure of SNP calling from supplied transcriptomes. Although HyLiTE provides the option to include additional genomic information to improve SNP calling ([Duchemin et al. 2015](#_ENREF_1)), the pipeline doesn’t support the usage of a prebuilt SNP index as PolyCat and RSEM.

A third metric of *Discrepancy* was used to measure the absolute differences between observed and expected expression values, which can be attributed to both unassignable reads and incorrect assignment. Thus, the lower level of *Discrepancy* was resulted from from GSNAP-PolyCat (***D***=14.5%) than from Bowtie2-HyLiTE (***D***=22.1%), as expected from the better performance of GSNAP-PolyCat in assignment *Efficiency* and *Accuracy*. The lowest measure of *Discrepancy* was seen for Bowtie2-RSEM (***D***=5.1%), solely reflecting the expression difference caused incorrect read estimation, given that complete read assignment was guaranteed by its algorithm.

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **GSNAP-PolyCat** | **Bowtie2-HyLiTE** | **Bowtie2-RSEM** |
| *Efficiency* (***E***) | 87.7% | 82.2% | 100.0% |
| - At only | 86.7% | 78.5% | 101.0% |
| - Dt only | 88.7% | 85.8% | 99.1% |
| *Accuracy* (***A***) | 99.4% | 87.3% | 98.0% |
| - At only | 99.6 % | 88.0% | 97.8% |
| - Dt only | 99.2% | 86.8% | 98.2% |
| *Discrepancy* (***D***) | 14.5% | 22.1% | 5.1% |
| - At only | 15.5% | 27.4% | 5.4% |
| - Dt only | 13.6% | 17.3% | 4.8% |

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 ([Grover et al. 2012](#_ENREF_2); [Yoo et al. 2014](#_ENREF_7)). To evaluate to what extent the estimated homoeolog read counts can faithfully represent the expression divergence as exhibited by actual homoeolog expression levels, we conducted differential expression (DE) analysis for the synthetic ADs dataset using its *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** | **Bowtie2-HyLiTE** | **Bowtie2-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** | **Bowtie2-HyLiTE** | **Bowtie2-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

Whether the topological properties of network modules are reproducible?

We calculate several module preservation statistics as described in ([Langfelder et al. 2011](#_ENREF_3)) 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).

Duchemin W, Dupont PY, Campbell MA, Ganley AR, Cox MP (2015) HyLiTE: accurate and flexible analysis of gene expression in hybrid and allopolyploid species. BMC bioinformatics 16:8. doi:10.1186/s12859-014-0433-8

Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF (2012) Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol 196 (4):966-971. doi:10.1111/j.1469-8137.2012.04365.x

Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is my network module preserved and reproducible? PLoS computational biology 7 (1):e1001057. doi:10.1371/journal.pcbi.1001057

Li L, Briskine R, Schaefer R, Schnable PS, Myers CL, Flagel LE, Springer NM, Muehlbauer GJ (2016) Co-expression network analysis of duplicate genes in maize (*Zea mays* L.) reveals no subgenome bias. BMC Genomics 17 (1):875. doi:10.1186/s12864-016-3194-0

Page JT, Gingle AR, Udall JA (2013) PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 3 (3):517-525. doi:10.1534/g3.112.005298

Page JT, Udall JA (2015) Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet 16 Suppl 2:S4. doi:10.1186/1471-2156-16-S2-S4

Yoo MJ, Liu X, Pires JC, Soltis PS, Soltis DE (2014) Nonadditive gene expression in polyploids. Annu Rev Genet 48:485-517. doi:10.1146/annurev-genet-120213-092159