METHOD

A likelihood ratio test for changes in homeolog expression bias

Ronald D. Smith[1], Taliesin J. Kinser[1], Gregory D. Smith[1] and Joshua R. Puzey[1*]

Abstract

Gene duplications are a major source of raw material for evolution and a likely contributor to the diversity of life on earth. Duplicate genes (i.e., homeologs, in the case of a whole genome duplication) may retain their ancestral function, sub- or neofunctionalize, or be lost entirely. A primary way that duplicate genes may evolve new functions is by altering their expression patterns. Comparing the expression patterns of duplicate genes may give clues as to whether any of these evolutionary processes have occurred. Here we develop a likelihood ratio test for the analysis of the expression ratios of duplicate genes across two conditions. We demonstrate an application of this test by comparing homeolog expression patterns of 1,448 homeologous gene pairs using RNA-seq data generated from the leaves and petals of a tetraploid monkeyflower (*Mimulus luteus*). Using simulated data, we show the sensitivity of this test to different levels of homeolog expression bias. While we have developed this method for the analysis of duplicate genes, it can be used for comparing expression patterns of any two genes (or alleles) across any two conditions.

Background

Gene duplications are a major source of raw material for evolution and a likely contributor to the diversity of life on earth [1–9]. Gene duplications are a special type of mutation resulting in the multiplication of intact functional components. These duplicate genes may either retain the ancestral function or individual portions of the gene's ancestral function may be partitioned (i.e., subfunctionalize) or evolve new functions entirely (i.e., neofunctionalize) [10–12]. Duplicate genes may evolve new functions either by changes in the primary coding sequence or altering where and when they are expressed. Previous work has indicated that changes to gene expression and their regulatory networks may be more important, rapid, or flexible than divergence of protein identities in the evolution of sub- and neofunctionlization [13–19].

There are multiple scenarios in which genes can be duplicated, ranging from small regional gene duplications to massive whole genome duplications (WGDs). The term polyploid refers to cells or organisms that have undergone a WGD event and contain more than two paired sets of chromosomes. Each complete set of chromosomes is referred to as a subgenome. Homologous genes located on separate subgenomes are referred to as hom*e*ologs.

WGDs are especially common in plants; indeed, all extant angiosperms (i.e., flowering plants) have at least two rounds of WGD in common [20] and up to 15% of speciation events in angiosperms may have been the product of WGDs [21]. Importantly, all major crops (rice, corn, potato, wheat, etc.) are polyploid [22]. WGD events and the resulting polyploidy are not restricted to plants, but have occurred in both vertebrate and invertebrate lineages as well. For example, the African clawed frog, *Xenopus*, commonly used as an experimental model system and extensively studied in developmental biology, includes species ranging from diploid to dodecaploid [23]. Other examples of polyploids with ancient WGD events include the the zebrafish *Danio rerio* [24], several salmonids [2], and some species of fungi [25]. Interestingly, there exists at least one polyploid mammal [26], a tetraploid rat from Argentina that mediates gene dosage by regulation of ribosomal RNA.

The biological consequences of gene duplications and subfunctionalization are significant and include examples such as the evolution of eyes [27], the evolution of hemoglobins [28], development of heat resistance in plants [29], and insecticide resistance [30]. Given the importance of duplicate genes in evolution, it is natural to ask how we might quantify differences in the activity or function of hom*e*ologous genes. One way

[*] Correspondence: jrpuzey@wm.edu

[1] Departments of Biology and Applied Science, The College of William and Mary, 23187, Williamsburg, VA, USA

Full list of author information is available at the end of the article

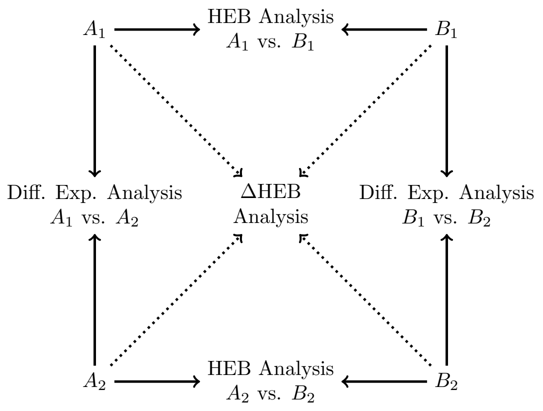

Figure 1: Comparison of the analysis of differential expression, homeolog expression bias (HEB), and homeolog expression bias shift (ΔHEB). A and B represent homeologous genes; the subscripts 1 and 2 denote experimental conditions. Differential expression analysis (vertical solid lines) uses high throughput RNA sequencing (RNA-seq) of one gene in different experimental conditions. HEB analysis (horizontal solid lines) involves two homeologous genes in one condition. ΔHEB analysis (diagonal broken lines) simultaneously compares expression level ratios for two genes and two conditions.

to begin exploring this question is by analyzing gene expression levels.

Genome-wide gene expression levels are commonly quantified using high throughput RNA sequencing (RNA-seq) [31]. In RNA-seq experiments, researchers extract and purify mRNA. This mRNA is reverse transcribed into cDNA, fragmented into smaller pieces, and sequenced using next-generation technology. The resulting millions of sequence reads are then mapped to either a reference genome or reference transcriptome, and the number of sequences mapping to a particular gene is used as an indication of the expression level of that gene.

In *differential expression analysis*, high-throughput RNA-seq data is used to determine if gene expression levels vary under different experimental conditions, or in distinct tissues, etc. Several different approaches to this statistical analysis exist [32], some of which use methods based on maximum likelihood estimation and likelihood ratio tests. A specific implementation of this approach is provided by DE-seq (an R/Bioconductor software package) [33,34].

Homeologous gene pairs frequently have distinguishing sequence differences. Therefore, sequencing reads derived from individual homeologs can be distinguished and expression levels can be determined for each homeolog. The term *homeolog expression bias* (HEB) refers to cases where homeologs are expressed at unequal levels in a single experimental condition [35]. The primary objective of this paper, statistical analysis of *changes* in homeolog expression bias (denoted ΔHEB) is fundamentally different than the statistical analysis of differential expression. For example, statistical evidence for ΔHEB involves simultaneous comparison of count data of *four* genes (two homeologs in two conditions, see Fig. 1) and is not reducible to sequential differential expression tests. Due to the importance of duplicate genes in evolution [1], specifically the ability of duplicate genes to evolve distinct expression profiles (e.g., sub- or neofunctionalize), we have developed a likelihood ratio test for ΔHEB analysis of RNA-seq data. The following sections describe the development of a likelihood ratio test for HEB and ΔHEB. We then use this test to analyze homeolog expression patterns from the leaves and petals of a tetraploid monkeyflower (*Mimulus luteus*). Using simulated data, we show the sensitivity of this test to different levels of homeolog expression bias.

Results

Quantifying homeolog expression bias (HEB)

We will write A and B to denote a homeologous gene pair from which RNA-seq data is generated in n biological replicates. Typically, the mean expression levels of the homeologs (denoted $\bar{a}$ and $\bar{b}$) are normalized by gene length and sequencing depth, as when reported in units of RPKM (reads per kilobase of coding sequence per million mapped reads). We define the homeolog expression bias (HEB) of the n replicates as

$$\text{HEB} = \log\left(\bar{b}/\bar{a}\right) = \log\bar{b} - \log\bar{a}\,,$$

a dimensionless quantity with $\text{HEB} = 0$ indicating no bias. If one uses the base 2 logarithm, $\text{HEB} = -3$ indicates 8-fold bias towards homeolog A.

Likelihood ratio test for HEB

After accounting for the possibility of different gene lengths, the statistical test for HEB is essentially a likelihood ratio test for differential expression of a pair of homeologous genes. The goal is to determine whether there is sufficient evidence to reject the null hypothesis (H_0) that there is no bias (i.e., equal expression levels for homeologous genes) in favor of the alternative hypothesis (H_1) that bias is present, i.e., different expression levels for homeologous genes. In mathematical terms, the null hypothesis H_0 corresponds to the parameters (denoted by θ) of a probability model for generating the data being in a specified subset Θ_0 of the parameter space Θ, that is,

$$\begin{aligned} H_0 &: \ \theta \in \Theta_0 \\ H_1 &: \ \theta \in \Theta \backslash \Theta_0\,. \end{aligned}$$

Let $\theta = (\lambda^a, \lambda^b)$ denote the true but unknown expression levels (properly scaled, e.g. in units of RPKM). Assuming positive expression, the parameter space is $\Theta = \{\theta : \lambda^a, \lambda^b \in \mathbb{R}_+\}$. The null ($H_0$) and alternative ($H_1$) hypotheses for the likelihood ratio test for homeolog expression bias are formalized as follows,

$$\begin{aligned} H_0 &: \ (\lambda^a, \lambda^b) \in \{\lambda^a, \lambda^b \in \mathbb{R}_+ : \lambda^a = \lambda^b\} \\ H_1 &: \ (\lambda^a, \lambda^b) \in \{\lambda^a, \lambda^b \in \mathbb{R}_+ : \lambda^a \neq \lambda^b\} . \end{aligned}$$

Equivalently, we let $\omega = \lambda^b/\lambda^a$ denote the ratio of expression levels and drop the superscript indicating the reference homeolog ($\lambda = \lambda^a$). In that case, $\lambda^b = \omega\lambda$ and the hypotheses are written as follows,

$$\begin{aligned} H_0 &: \quad (\lambda, \omega) \in \{\lambda, \omega \in \mathbb{R}_+ : \omega = 1\} \\ H_1 &: \quad (\lambda, \omega) \in \{\lambda, \omega \in \mathbb{R}_+ : \omega \neq 1\} . \end{aligned}$$

Once we specify a probability model for the data $\mathcal{X}$, likelihood functions for each hypothesis, $\mathcal{L}_0(\theta|\mathcal{X})$ and $\mathcal{L}_1(\theta|\mathcal{X})$, can be derived (see next section). For composite hypotheses, the appropriate likelihood ratio test statistic is

$$W(\mathcal{X}) = -2 \ln \frac{\hat{\mathcal{L}}_0}{\hat{\mathcal{L}}_1} = 2 \left(\ln \hat{\mathcal{L}}_1 - \ln \hat{\mathcal{L}}_0 \right) , \qquad (1)$$

where $\hat{\mathcal{L}}_1$ and $\hat{\mathcal{L}}_0$ are the maximized likelihoods,

$$\begin{aligned} \hat{\mathcal{L}}_1 &= \sup\{ \mathcal{L}(\theta|\mathcal{X}) : \theta \in \Theta \} \\ \hat{\mathcal{L}}_0 &= \sup\{\mathcal{L}(\theta|\mathcal{X}) : \theta \in \Theta_0 \} . \end{aligned}$$

A critical value of the test statistic (W_*) is obtained from the Chi-squared distribution with significance level $\alpha = 0.05$. The number of degrees of freedom δ is the difference in the number of free parameters in Θ and Θ_0 (here $\delta = 1$) [36]. The null hypothesis H_0 is rejected in favor of the alternative H_1 when $W(\mathcal{X}) > W_*$.

Probability model for RNA-seq read counts

Denote the lengths of homeologous genes a and b as ℓ^a and ℓ^b (e.g., in kilobases) and let d_i be the sequencing depth (e.g., in millions) of replicate i. The expected number of RNA-seq reads for gene a and replicate i is

$$\mu_i^a = \lambda^a \ell^a d_i = \lambda \ell^a d_i , \qquad (2)$$

where in the second equality we have dropped the superscript for the reference homeolog ($\lambda = \lambda^a$). Similarly, the expected number of RNA-seq reads for gene b and replicate i is

$$\mu_i^b = \lambda^b \ell^b d_i = \omega \lambda \ell^b d_i \qquad (3)$$

where $\omega = \lambda^b/\lambda^a = \lambda^b/\lambda$.

The probability model assumes that the count data for each gene is drawn from a negative binomial distribution,

$$f(x; \mu, r) = \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\mu}{\mu + r} \right)^x \left(\frac{r}{\mu + r} \right)^r ,$$

where μ is the appropriate mean (μ_i^a or μ_i^b in Eqs. 2 and 3). That is, if X_i^a and X_i^b are random variables representing the count data for replicate i of homeologous genes A and B,

$$\begin{aligned} \Pr\{X_i^a = a_i\} &= f(a_i; \lambda \ell^a d_i, r_i) \\ \Pr\{X_i^b = b_i\} &= f(b_i; \omega \lambda \ell^b d_i, r_i) , \end{aligned}$$

where we have used $\mu_i^a = \lambda \ell^a d_i$ and $\mu_i^b = \omega \lambda \ell^b d_i$. In these expressions, the aggregation parameter r_i is obtained from the observed mean-variance relation for all homeolog pairs of the ith experimental replicate (see Appendix 1).

Assuming independence of experimental replicates, the likelihood functions $\mathcal{L}_1$ and $\mathcal{L}_0$ are products of the likelihood functions for each observation, that is,

$$\mathcal{L}_1(\mathcal{X}) = \textstyle\prod_{i=1}^n \mathcal{L}_1^i(\mathcal{X}) ,$$

and similarly for $\mathcal{L}_0(\mathcal{X})$, where $\mathcal{X}_i = \{a_i, b_i\}$ indicates the observed read counts for replicate i and $\mathcal{X} = \cup_{i=1}^n \mathcal{X}_i$. The likelihood function for the alternative hypothesis and the ith replicate is

$$\begin{aligned} \mathcal{L}_1^i(\mathcal{X}) = & \frac{\Gamma(r_i + a_i)}{\Gamma(r_i) a_i!} \frac{\Gamma(r_i + b_i)}{\Gamma(r_i) b_i!} \\ & \times \left(\frac{\lambda \ell^a d_i}{\lambda \ell^a d_i + r_i} \right)^{a_i} \left(\frac{\omega \lambda \ell^b d_i}{\omega \lambda \ell^b d_i + r_i} \right)^{b_i} \\ & \times \left(\frac{r_i}{\lambda \ell^a d_i + r_i} \right)^{r_i} \left(\frac{r_i}{\omega \lambda \ell^b d_i + r_i} \right)^{r_i} . \end{aligned} \qquad (4)$$

The likelihood function for the null hypothesis and the ith replicate, $\mathcal{L}_0^i(\mathcal{X})$, is given by Eq. 4 with $\omega = 1$.

Maximum likelihood estimation

Maximum likelihood estimation is performed using the the log-likelihood function corresponding to Eq. 4, namely,

$$\ln \mathcal{L}_1(\mathcal{X}) = \textstyle\sum_i \ln \mathcal{L}_1^i(\mathcal{X}) , \qquad (5)$$

where

$$
\begin{aligned}
\ln \mathcal{L}_1^i(\mathcal{X}) &= \gamma\left(r_i + a_i\right) + \ln\left(a_i!\right) \\
&+ \gamma\left(r_i + b_i\right) + \ln\left(b_i!\right) \\
&+ 2r_i \ln r_i - 2\gamma\left(r_i\right) \\
&+ a_i \ln\left(\lambda \ell^a d_i\right) + b_i \ln\left(\omega\lambda\ell^b d_i\right) \\
&- \left(a_i + r_i\right)\ln\left(\lambda\ell^a d_i + r_i\right) \\
&- \left(b_i + r_i\right)\ln\left(\omega\lambda\ell^b d_i + r_i\right)\big] \qquad (6)
\end{aligned}
$$

and $\gamma(\cdot) = \ln\Gamma(\cdot)$. The log-likelihood function for the null hypothesis ($\ln\mathcal{L}_0$) is given by Eq. 13 with $\omega = 1$.

The log-likelihood function $\ln\mathcal{L}_1(\mathcal{X})$ is maximized by numerically solving for $\hat\lambda$ and $\hat\omega$ leading to zero partial derivatives,

$$0 = \left.\frac{\partial \ln \mathcal{L}_1}{\partial\lambda}\right|_{\hat\lambda,\hat\omega} \qquad (7)$$

$$0 = \left.\frac{\partial \ln \mathcal{L}_1}{\partial\omega}\right|_{\hat\lambda,\hat\omega}, \qquad (8)$$

as described in Appendix 2. The log-likelihood function $\ln\mathcal{L}_0(\mathcal{X})$ is maximized by solving for $\hat\lambda$ leading to

$$0 = \left.\frac{\partial \ln \mathcal{L}_0}{\partial\lambda}\right|_{\hat\lambda}. \qquad (9)$$

The optimal parameter values $\hat\lambda$ and $\hat\omega$ are used to evaluate $\ln\hat{\mathcal{L}}_0(\mathcal{X};\hat\lambda)$, $\ln\hat{\mathcal{L}}_1(\mathcal{X};\hat\lambda,\hat\omega)$, and the test statistic W (see Eq. 1).

Example of the likelihood ratio test for HEB applied to tetraploid *Mimulus luteus*

To demonstrate the application of the likelihood ratio test for HEB, five biological replicates of RNA-seq data were generated from petals of the tetraploid *Mimulus luteus* (monkeyflower), and another five replicates were generated from the leaves (see Appendix 3 for details). We have chosen *M. luteus* because it is a tetraploid with two distinct subgenomes [37]. One of the subgenomes contains genes which more closely resemble the extant diploid *M. guttatus* than the other, so we refer to the homeologs as *M. guttatus-like* and *Other*. Please see [37] for a detailed description of how the homeologs were identified and an overview of the natural history of *M. luteus*.

In this section, we use the likelihood ratio test for HEB to find homeologous gene pairs where one homeolog is expressed at significantly different levels than the other, one tissue at a time. In the section on ΔHEB we develop a likelihood ratio test to determine whether there is a significant difference in the bias between the two tissues.

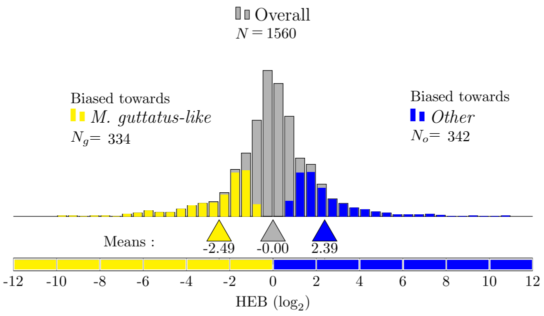

Figure 2: Likelihood ratio test for HEB in petals of *M. luteus*. Of 1,560 testable homeologous gene pairs (gray), a total of 676 show significant bias. Of these, 334 pairs are biased towards the *M. guttatus-like* homeolog (yellow), with a mean HEB of −2.49 (5.6×). 342 pairs are biased towards the *Other* homeolog (blue), with a mean HEB of 2.39 (about 5.2×). The Benajamini-Hochberg correction for multiple testing was applied at significance level $\alpha = 0.05$ (and also in Figure 3–5).

Homeolog expression bias in Mimulus luteus *petals*

Figure 2 shows the result of applying the likelihood ratio test for HEB to the petal data. There are 1,853 homeologous gene pairs in *M. luteus* that can be identified as coming from separate subgenomes. Of these 1,853 homoeologous pairs, 1,560 were testable (measurable expression from each individual homeolog). Of the testable pairs, a total of 676 gene pairs show significant bias (using a significance level of $\alpha = 0.05$, and applying the Benjamini-Hochberg correction [38] to account for multiple testing error). In the 334 pairs biased towards the *M. guttatus-like* homeolog the mean HEB is −2.49 (5.6-fold change). Negative bias throughout this paper will indicate bias towards the *M. guttatus-like* homeolog. In the 342 pairs biased towards the *Other* homeolog, the mean HEB is 2.39 (5.2-fold change).

These results may be indicative of a number of evolutionary processes. For example, one of the homeologs may have become sub- or neofunctionalized in this tissue, or one of the homeologs may simply be losing its function.

Homeolog expression bias in Mimulus luteus *leaves*

Next, the likelihood ratio test for HEB was applied to the leaf data (results shown in Fig 3). Of the 1,853 homoeologous pairs, 1,498 were testable. Of this subset, a total of 399 gene pairs show significant bias (using a significance level of $\alpha = 0.05$, and applying the Benjamini-Hochberg correction to account for multiple testing error). In the 199 pairs biased towards the *M. guttatus-like* homeolog the mean HEB is −2.83 (7.1-fold change). In the 200 pairs biased towards the *Other* homeolog, the mean HEB of 2.80 (7.0-fold change).

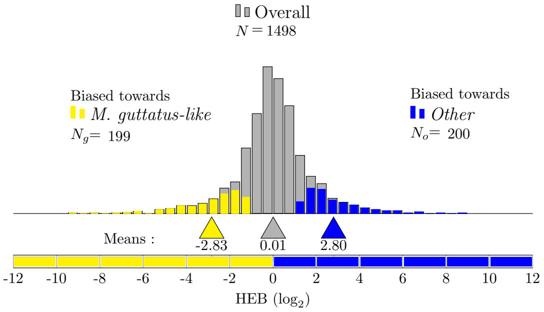

Figure 3: likelihood ratio test for HEB in leaves of *M. luteus*. Of 1,498 testable homeologous gene pairs (gray), a total of 399 show significant bias. 199 pairs are biased towards the *M. guttatus-like* homeolog (yellow) the mean HEB is −2.83 (7.1×). 200 pairs are biased towards the *Other* homeolog (blue), with a mean HEB of 2.80 (7.0×).

Comparison of HEB in M. luteus *petals and leaves*
While the number of significantly biased gene pairs in the two tissues is quite different, i.e., 676 biased pairs in the petals vs. only 399 in the leaves, we cannot simply combine these two results to determine which genes exhibit a significant difference in bias (ΔHEB) between the two tissues. In the following section we derive a statistical test for ΔHEB, and Fig 5 shows the results of this test applied to the leaf and petal data we just saw.

Quantifying changes in homeolog expression bias (ΔHEB)

The statistical analysis of changes in homeolog expression bias (ΔHEB) involves simultaneous comparison of count data of *four* genes (two homeologs in two conditions, see Fig. 1). Suppose A and B represent homeologous genes and RNA-seq data is generated under conditions 1 and 2 in n biological replicates, leading to mean expression levels $\bar{a}_1$, $\bar{a}_2$, $\bar{b}_1$, $\bar{b}_2$. The change in homeolog expression bias (ΔHEB) is defined as

$$\Delta\text{HEB} = \text{HEB}_2 - \text{HEB}_1 = \log\left(\frac{\bar{b}_2/\bar{a}_2}{\bar{b}_1/\bar{a}_1}\right), \qquad (10)$$

where the last equality uses $\text{HEB}_1 = \log \bar{b}_1/\bar{a}_1$ and $\text{HEB}_2 = \log \bar{b}_2/\bar{a}_2$.

Likelihood ratio test for ΔHEB

The likelihood ratio test for ΔHEB is designed to determine whether there is sufficient evidence to reject the null hypothesis (H_0) that homeolog expression bias is the same under two experimental conditions in favor of the alternative hypothesis (H_1) that there is a difference in bias. Following notation similar to the previous section, our hypotheses are

$$\begin{aligned} H_0 &: \theta \in \Theta_0 = \{\lambda_{1|2}^{a|b} \in \mathbb{R}_+ : \lambda_1^b/\lambda_1^a = \lambda_2^b/\lambda_2^a\} \\ H_1 &: \theta \in \Theta\backslash\Theta_0 = \{\lambda_{1|2}^{a|b} \in \mathbb{R}_+ : \lambda_1^b/\lambda_1^a \neq \lambda_2^b/\lambda_2^a\}, \end{aligned}$$

where $\lambda_{1|2}^{a|b}$ is an abbreviation for $\lambda_1^a, \lambda_1^b, \lambda_1^b, \lambda_2^b$. Equivalently,

$$\begin{aligned} H_0 &: \quad \theta \in \Theta_0 = \{\lambda_1, \lambda_2, \omega_1, \omega_2 \in \mathbb{R}_+ : \omega_1 = \omega_2\} \\ H_1 &: \quad \theta \in \Theta\backslash\Theta_0 = \{\lambda_1, \lambda_2, \omega_1, \omega_2 \in \mathbb{R}_+ : \omega_1 \neq \omega_2\}, \end{aligned}$$

where $\omega_1 = \lambda_1^b/\lambda_1^a$, $\omega_2 = \lambda_2^b/\lambda_2^a$, $\lambda_1 = \lambda_1^a$ and $\lambda_2 = \lambda_2^a$. The difference in degrees of freedom of the alternative and null hypotheses is $\delta = 4 - 3 = 1$.

The likelihood functions for the ΔHEB test are similar to those for HEB, though the two different experimental conditions lead to twice as many terms (cf. Eq. 4). The likelihood function for H_1 is

$$\mathcal{L}_1(\mathcal{X}) = \prod_{k=1}^{2}\prod_{i=1}^{n} \mathcal{L}_1^{k,i}(\mathcal{X}) \qquad (11)$$

where $\mathcal{L}_1^{k,i}$, the likelihood function for the ith replicate of the kth condition, has the form of Eq. 4 with appropriate parameters indexed by condition ($a_{k,i}$, $b_{k,i}$, $r_{k,i}^a$, $r_{k,i}^b$, ω_k). The log-likelihood function for H_1 is thus

$$\ln \mathcal{L}_1(\mathcal{X}) = \sum_{k=1}^{2}\sum_{i=1}^{n} \ln \mathcal{L}_1^{k,i}(\mathcal{X}) \qquad (12)$$

where

$$\begin{aligned} \ln \mathcal{L}_1^{k,i}(\mathcal{X}) \; &= \; \gamma\left(r_{k,i} + a_{k,i}\right) + \ln\left(a_{k,i}!\right) \\ &+ \; \gamma\left(r_{k,i} + b_{k,i}\right) + \ln\left(b_{k,i}!\right) \\ &+ \; 2r_{k,i}\ln r_{k,i} - 2\gamma\left(r_{k,i}\right) \\ &+ \; a_{k,i}\ln\left(\lambda\ell^a d_i\right) + b_{k,i}\ln\left(\omega_k\lambda\ell^b d_i\right) \\ &- \; \left(a_{k,i} + r_{k,i}\right)\ln\left(\lambda\ell^a d_i + r_{k,i}\right) \\ &- \; \left(b_{k,i} + r_i\right)\ln\left(\omega_k\lambda\ell^b d_i + r_{k,i}\right)(13) \end{aligned}$$

and $\gamma(\cdot) = \ln\Gamma(\cdot)$. The log-likelihood function for the null hypothesis ($\ln \mathcal{L}_0$) is given by the above expressions with $\omega_1 = \omega_2 = \omega$. The aggregation parameters ($r_{k,i}$) are determined from the data with experimental conditions $k = 1$ and 2 considered separately (cf. Eqs. 17–19).

The log-likelihood function $\ln \mathcal{L}_1(\mathcal{X})$ used in the analysis of ΔHEB is maximized by numerically solving uncoupled systems of the form of Eqs. 7 and 8 for $(\hat{\lambda}_1, \hat{\omega}_1)$ and $(\hat{\lambda}_2, \hat{\omega}_2)$. The log-likelihood function

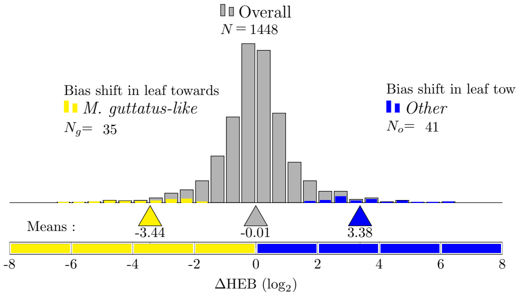

Figure 4: Likelihood ratio test for ΔHEB in the leaves vs. petals of *M. luteus*. Of 1,448 testable homeologous gene pairs (gray), 76 show significant ΔHEB. Of these, 35 are more biased towards the *M. guttatus-like* homeolog in the leaf than in the petal (yellow). 41 gene pairs are more biased towards the *Other* homeolog in the leaf than in the petal (blue).

$\ln \mathcal{L}_0(\mathcal{X})$ is maximized by solving for $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\omega}$ that lead to zero partial derivatives,

$$0 = \left.\frac{\partial \ln \mathcal{L}_0}{\partial \lambda_1}\right|_{\hat{\lambda}_1,\hat{\lambda}_2,\hat{\omega}} \tag{14}$$

$$0 = \left.\frac{\partial \ln \mathcal{L}_0}{\partial \lambda_2}\right|_{\hat{\lambda}_1,\hat{\lambda}_2,\hat{\omega}} \tag{15}$$

$$0 = \left.\frac{\partial \ln \mathcal{L}_0}{\partial \omega}\right|_{\hat{\lambda}_1,\hat{\lambda}_2,\hat{\omega}} . \tag{16}$$

The optimal parameter values are used to evaluate the likelihoods, $\hat{\mathcal{L}}_0(\mathcal{X};\hat{\lambda}_1,\hat{\lambda}_2,\hat{\omega})$ and $\hat{\mathcal{L}}_1(\mathcal{X};\hat{\lambda}_1,\hat{\lambda}_2,\hat{\omega}_1,\hat{\omega}_2)$, and the test statistic W (see Eq. 1).

The numerical solution of these equations was facilitated by transforming these equations in a manner that ensured both parameters are positive and was symmetric with respect to the mean expression levels of homeolog A and B (see Appendix 2).

Example of the likelihood ratio test for ΔHEB applied to real data

Returning to the leaf and petal data from the previous sections on HEB, we now have a total of 1,448 testable pairs (for ΔHEB, each gene in the pair needs to have at least 1 read in both conditions (i.e., tissues)).

Figure 4 shows the results of the likelihood ratio test for ΔHEB. We find a total of 76 gene pairs show significant ΔHEB. Of these, 35 are more biased towards the *M. guttatus-like* homeolog in the leaf than they are in the petal. The remaining 41 gene pairs are more biased towards the *Other* homeolog in the leaf than they are in the petal.

Figure 5 shows a scatter plot of homeolog expression bias (HEB) in leaf and petal. Colored marks indicate gene pairs with statistically significant changes in homeolog expression bias (ΔHEB) (these points correspond to the colored bars in Figure 4). Data points in the top-left and bottom-right quadrants of Figure 5 represent homeologous pairs where one homeolog is more highly expressed in one tissue and its partner is more highly expressed in the other tissue. On the other hand, the top-right and bottom-left quadrants correspond to homeologous pairs where the difference in bias favors the same homeolog but has become more extreme. Finally, all of the marks that are colored blue or yellow show significant change in bias and are candidates for tissue specific sub- or neofunctionalization.

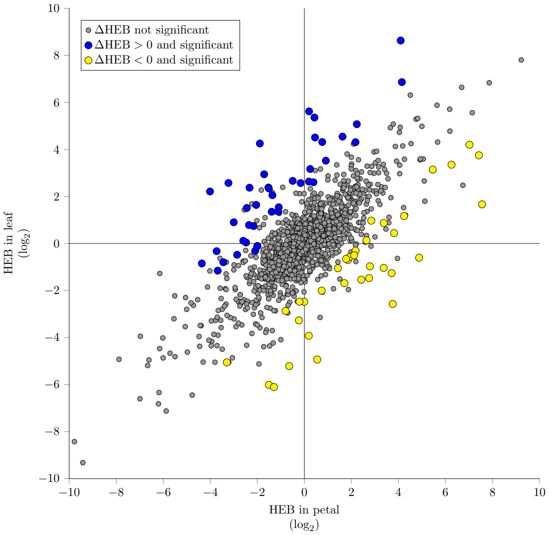

Figure 5: Statistical significance of ΔHEB compared to homeolog expression bias (HEB) in leaf and petal. Yellow and blue indicates homeolog gene pairs with significant ΔHEB. The likelihood ratio test for ΔHEB is distinct from HEB tests in leaf and petal (see text).

Although the change in homeolog expression bias is defined by Eq. 10 as the log-fold change in homeolog expression bias, the intercalation of significant (yellow and blue) and not significant (gray) ΔHEB in Figure 5 makes it clear that statistical evidence for ΔHEB is not reducible to the difference between HEB_{leaf} and $\text{HEB}_{\text{petal}}$ (the vertical or horizontal distance to the line of slope 1 where $\text{HEB}_{\text{leaf}} = \text{HEB}_{\text{petal}}$).

Whether or not ΔHEB can be called significant also depends on differences in sequencing depths, mean expression levels (e.g., lowly expressed genes are more likely to be influenced by shot noise), and ratios of gene lengths. All of these factors are considered simultaneously in the likelihood ratio test presented here. Calling ΔHEB based on sequential HEB results would almost certainly result in a different set of genes being called significant.

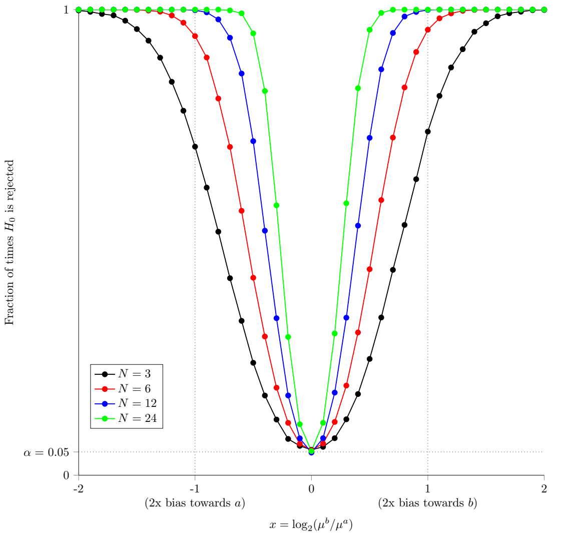

Figure 6: Ability of the likelihood ratio test for HEB to detect different levels of bias. Simulation results show the fraction of times H_0 was rejected for 10,000 trials with the given values of x and n (parameters: $\alpha = 0.05$, $\mu^a = 100$, $r = 10$). A 4-fold change (i.e., $|x| = 2$) is almost always detectable with $n \geq 3$ replicates. Detecting a 2-fold change greater than 95% of the time requires 12 or more replicates.

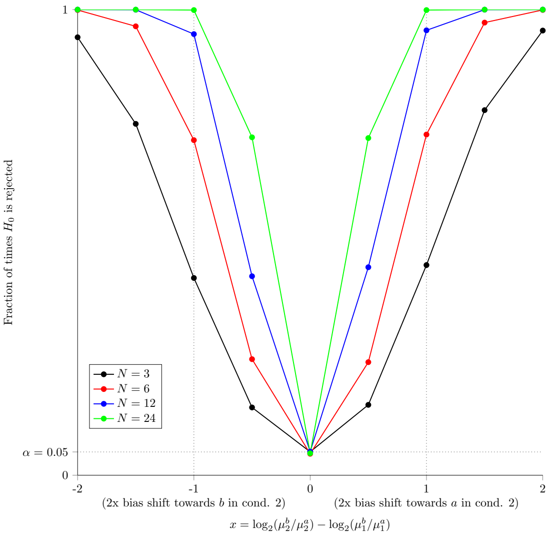

Figure 7: Ability of the likelihood ratio test for ΔHEB to detect different levels of change in bias. Simulation results show the fraction of times H_0 was rejected for 10,000 trials with given values of x and n (parameters: $\alpha = 0.05$, $\mu_1^a = \mu_1^b = \mu_2^a = 100$, $r = 10$). A 4-fold change (i.e., $|x| = 2$) is almost always detectable with $n \geq 6$ replicates. However, detecting a 2-fold change greater than 95% of the time requires 24 or more replicates.

The remaining sections explore how the statistical test for HEB and ΔHEB presented here perform on simulated data.

Validation of the likelihood ratio test for HEB using simulated data

A natural question to ask about HEB and ΔHEB is, "How large does the change in expression levels between homeologs across conditions need to be before we can detect ΔHEB most of the time?". Unsurprisingly, this depends largely on the number of biological replicates.

To explore this question we generated simulated data with one expression level fixed at a constant value, $\mu^a = 100$, and let the other expression level, μ^b, vary; $\mu^b = 2^x\mu^a$, with $x \in [-2, 2]$, in steps of 0.1. For each value of x, we used MATLAB to generate 10,000 sets of data from a negative binomial distribution for $N = 3, 6, 12$ and 24 replicates. We fixed the parameter $r = 10$ for simplicity; this is in the typical range of values we have observed in RNA-seq data.

Fig 6 shows the results of the likelihood ratio test for HEB on this data set. We find that a 4-fold change (i.e $|x| = 2$) is almost always detectable, regardless of the number of replicates. However, detecting a 2-fold change almost all of the time requires at least 12 replicates.

Validation of the likelihood ratio test for ΔHEB using simulated data

To assess the sensitivity of ΔHEB to different levels of bias shift, we created a data set similar to that used for Fig 6. This time, we set 3 of the means equal ($\mu_1^a = \mu_1^b = \mu_2^a = 100$), and let the fourth one vary; $\mu_2^b = 2^x\mu_2^a$, with $x \in [-2, 2]$ in steps of 0.5 (fewer points are generated than in HEB due to longer computational time). Aggregation parameter $r = 10$. For each value of x, 10,000 sets of data were generated from a negative binomial distribution for $n = 3, 6, 12$ and 24 replicates.

Fig 7 shows the results of the likelihood ratio test for ΔHEB on this data set. The results are similar to those for HEB, with the test for ΔHEB being slightly less sensitive than the test for HEB. For ΔHEB, a 4x change in bias is not detected nearly 100% of the time for $N = 3$, but it is for $N \geq 6$. As in HEB, the ability to detect smaller changes increases significantly with the number of replicates.

Conclusions

Gene duplication and polyploidy are extremely important factors in generating the diversity of life on earth. As Ohno stated in his seminal work on gene duplication [1], "Natural selection merely modified while redundancy created" the raw materials necessary for the diversification of life on earth.

In this paper we have developed a robust statistical framework specifically designed for the comparison of duplicate gene expression patterns. Importantly, this technique is consistent and reproducible. Through analysis of simulated data we have shown that these methods perform well, especially given the typically small sample sizes in most RNA-seq experiments. We have shown that the ability to detect small differences in expression levels increases as a function of sample size, a fact which can be used to aid experimental design. Other authors have noted this with traditional differential expression analysis and made similar recommendations [39–41]. Moreover, we have demonstrated the application of the likelihood ratio test for ΔHEB using a real RNA-seq dataset derived from a polyploid plant.

While we have developed this test for the purpose of analyzing changes in expression patterns of homeologous genes, we wish to emphasize that the methods are suitable for the expression analysis of any two genes (they need not be homeologs) across any two conditions. A potential application of this test may be the comparison of expression patterns of duplicate genes following hybridization through comparison of parent and hybrid expression levels (assuming proper internal controls are used) [42]. One exciting application of this method may be the analysis of allele specific expression changes.

Appendix 1: Determining aggregation parameters

Due to the typically small number of replicates in RNA-seq experiments, accurate estimation of the aggregation parameter is not realistic on a gene-by-gene basis [34, 43]. Instead, we use the mean-variance relation of a negative binomial distribution, namely,

$$\sigma^2 = \mu + \frac{1}{r}\mu^2 \,, \tag{17}$$

to compute an aggregation parameter r for each experimental replicate, after rescaling to account for each replicates sequencing depth.

In brief, let x_j^i denote the count data for the jth pair of homeologous genes obtained for experimental replicate $i \in \{1, 2, \ldots, n\}$. For each of the n replicates, we produce an auxiliary data set $(y_{k,j}^i)$ by rescaling the count data for all replicates as though each were obtained in an experiment with the sequencing depth of replicate k,

$$y_{k,j}^i = \frac{d_k}{d_i} x_j^i \,. \tag{18}$$

For each gene (j), we compute a scaled mean $(\mu_{k,j})$ and variance $(\sigma_{k,j}^2)$ of $y_{k,j}^i$ over replicates (i). To obtain the aggregation parameter r_k, we perform a nonlinear least squares fit of the observed mean-variance relation across all genes. That is, r_k minimizes the sum of squares error,

$$E = \sum_j \left(\sigma_{k,j}^2 - \mu_{k,j} - \frac{1}{r_k}\mu_{k,j}^2 \right)^2 \,. \tag{19}$$

Appendix 2: Numerical scheme for maximum likelihood estimation

For the analysis of both HEB and ΔHEB, parameter values maximizing the likelihood functions $\hat{\mathcal{L}}_0$ and $\hat{\mathcal{L}}_1$ were obtained using the built-in MATLAB command `fsolve` applied to Eqs. 7–9 and 14–16. In both cases, the numerical procedure was facilitated by changing variables from (λ, ω) to (v, y) through

$$\begin{aligned} \lambda &= e^{v-y} \\ \omega &= e^{2y} \,, \end{aligned}$$

that is, $v = \ln \lambda + y$ and $y = (\ln \omega)/2$. This ensures positivity of λ and ω and leads to a system of equations that is symmetric in $\lambda^a \leftrightarrow \lambda^b$. The new variable v is the logarithm of the geometric mean of the expression levels $\lambda^a = \lambda$ and $\lambda^b = \omega\lambda$,

$$v = \ln \sqrt{\lambda^a \lambda^b} = \ln \sqrt{\lambda \cdot \omega\lambda} \,,$$

that is, $\lambda^a = \lambda = e^{v-y}$ and $\lambda^b = \omega\lambda = e^{v+y}$. The transformed partial derivatives used to maximize the log-likelihood $\ln \mathcal{L}_1$ (Eqs. 7–8) are

$$0 = \frac{\partial \ln \mathcal{L}_1}{\partial v} = \sum_i B_i(v, y) + A_i(v, y) \tag{20}$$

$$0 = \frac{\partial \ln \mathcal{L}_1}{\partial y} = \sum_i B_i(v, y) - A_i(v, y) \tag{21}$$

where

$$A_i(v, y) = a_i - \frac{(a_i + r_i)e^{v-y}\ell^a d_i}{e^{v-y}\ell^a d_i + r_i} \tag{22}$$

$$B_i(v, y) = b_i - \frac{(b_i + r_i)e^{v+y}\ell^b d_i}{e^{v+y}\ell^b d_i + r_i} \,. \tag{23}$$

The transformed partial derivative used to maximize $\ln \mathcal{L}_0$ are found by substituting $y = 0$ in Eq. 20,

$$0 = \frac{\partial \ln \mathcal{L}_0}{\partial v} = \sum_i B_i(v, 0) + A_i(v, 0) \,.$$

For the analysis of ΔHEB, the partial derivatives used to maximize $\ln \mathcal{L}_1$ are two uncoupled systems of the form of Eq. 20–23, one for each experimental condition ($k = 1$ and 2),

$$\begin{aligned} 0 &= \frac{\partial \ln \mathcal{L}_1}{\partial v_k} = \sum_i B_{k,i}(v_k, y_k) + A_{k,i}(v_k, y_k) \\ 0 &= \frac{\partial \ln \mathcal{L}_1}{\partial y_k} = \sum_i B_{k,i}(v_k, y_k) - A_{k,i}(v_k, y_k) \end{aligned}$$

where

$$\begin{aligned} A_{k,i}(v,y) &= a_{k,i} - \frac{(a_{k,i} + r_{k,i})e^{v-y}\ell^a d_i}{e^{v-y}\ell^a d_i + r_{k,i}} \\ B_{k,i}(v,y) &= b_{k,i} - \frac{(b_{k,i} + r_{k,i})e^{v+y}\ell^b d_i}{e^{v+y}\ell^b d_i + r_{k,i}} \, . \end{aligned}$$

In the case of the null hypothesis $y_2 = y_1 = y$ we numerically solve a system of three equations, including

$$0 = \frac{\partial \ln \mathcal{L}_0}{\partial v_k} = \sum_i B_{k,i}(v_k, y) + A_{k,i}(v_k, y)$$

for $k = 1$ and 2. These are coupled via

$$0 = \frac{\partial \ln \mathcal{L}_0}{\partial y} = \sum_k \sum_i B_{k,i}(v_k, y) - A_{k,i}(v_k, y) \, .$$

Appendix 3: Experimental methods

Plant tissues were collected from second generation inbred *M. luteus*. All plants were grown in a greenhouse under a 16 hour light regiment at 21 °C and 30% humidity. Petal tissue was collected from the corolla of a flower bud near blooming, and leaf tissue came from young leaves adjacent to the stem apical meristem. Five replicates of each tissue type were collected, at the same time of day, from different individuals. Approximately 100 - 200 mg of plant tissue was immediately placed into liquid nitrogen. RNA was extracted by grinding frozen tissue with pestles in PureLink® Plant RNA Reagent from Ambion™. Column isolation of RNA was subsequently performed using Direct-zol™ RNA MiniPrep Plus Kit from Zymo Research. Libraries were constructed using KAPA Stranded mRNA-Seq Kit. During library construction, sequence specific Illumina TruSeq® adapters were added to distinguish each library. Using an Agilent 2100 Bioanalyzer, average fragment lengths were determined to be between 230 and 300 bp. Libraries were then pooled and sequenced by the Duke Center for Genomic and Computational Biology on an Illumina HiSeq 2500 instrument. The resulting reads (50 base pair, single end) were mapped to the *M. luteus* genome using bowtie2 [44] with the `--local-very-sensitive` option. Reads to exonic regions were counted using `htseq-count` [45] with the default settings (minimum alignment quality of 10 on the phred scale).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
All four co-authors were involved in conception of the work; data analysis and interpretation; and drafting, critical review and final approval of the article. Data collection (TK and JRP). Mathematical framework and numerical scheme (RDS and GDS). TK supervised by JRP. RDS was jointly supervised by JRP and GDS.

Acknowledgements
The work was supported in part by National Science Foundation Grant DMS 1121606 to GDS, the Biomathematics Initiative at The College of William & Mary, and a W&M Summer Research grant to JRP.

Author details
[1]Departments of Biology and Applied Science, The College of William and Mary, 23187, Williamsburg, VA, USA. [2].

References

1. Ohno, S.: Evolution by Gene Duplication. Springer-Verlag
2. Otto, S.P., Whitton, J.: Polyploid incidence and evolution. Annu. Rev. Genet. **34**, 401–437 (2000)
3. Wendel, J.F.: Genome evolution in polyploids. Plant Molecular Biology **42**, 225–249 (2000)
4. Crow, K.D., Wagner, G.P.: What is the role of genome duplication in the evolution of complexity and diversity? Mol. Biol. Evol. **23**(5), 887–892 (2006)
5. Proulx, S.R.: Multiple routes to subfunctionalization and gene duplicate specialization. Genetics **190**, 737–751 (2012)
6. McLysaght, A., Hokamp, K., Wolfe, K.H.: Extensive genomic duplication during early chordate evolution. Nature genetics **31**(2), 200–204 (2002)
7. Dehal, P., Boore, J.L.: Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3**(10), 314 (2005)
8. Spring, J.: Genome duplication strikes back. Nature genetics **31**(2), 128–130 (2002)
9. Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B., Salt, D.E.: Polyploids exhibit higher potassium uptake and salinity tolerance in arabidopsis. Science **341**(6146), 658–659 (2013)
10. Sémon, M., Wolfe, K.H.: Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proceedings of the National Academy of Sciences **105**(24), 8333–8338 (2008)
11. Rastogi, S., Liberles, D.A.: Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology **5**(28) (2005). doi:10.1186/1471-2148-5-28
12. Taylor, J.S., Raes, J.: Duplication and divergence: The evolution of new genes and old ideas. Annu. Rev. Genet. (38), 615–643 (2004)
13. Smet, R.D., de Peer, Y.V.: Redundancy and rewiring of genetic networks following genome-wide duplication events. Current Opinion in Plant Biology **15**, 168–176 (2012)
14. Blanc, G., Wolfe, K.H.: Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. The Plant Cell **16**(7), 1679–1691 (2004)
15. Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C., Ragan, M.A.: Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. Genome Research **19**(8), 1404–1418 (2009). doi:10.1101/gr.086827.108
16. Huminiecki, L., Wolfe, K.H.: Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome research **14**(10a), 1870–1879 (2004)
17. Makova, K.D., Li, W.-H.: Divergence in the spatial pattern of gene expression between human duplicate genes. Genome research **13**(7), 1638–1645 (2003)

18. Gu, Z., Nicolae, D., Lu, H.H., Li, W.-H.: Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in genetics **18**(12), 609–613 (2002)
19. Qiu, Y., Liu, S.-L., Adams, K.L.: Frequent changes in expression profile and accelerated sequence evolution of duplicated imprinted genes in *arabidopsis* . Genome biology and evolution **6**(7), 1830–1842 (2014)
20. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., *et al.*: Ancestral polyploidy in seed plants and angiosperms. Nature **473**(7345), 97–100 (2011)
21. Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., Rieseberg, L.H.: The frequency of polyploid speciation in vascular plants. Proceedings of the national Academy of sciences **106**(33), 13875–13879 (2009)
22. Renny-Byfield, S., Wendel, J.F.: Doubling down on genomes: Polyploidy and crop plants. American Journal of Botany **101**(10), 1711–1725 (2014)
23. Kobel, H.R., Pasquier, L.D.: Genetics of polyploid *Xenopus*. Trends in Genetics **2**, 310–315 (1986)
24. Wolfe, K.H.: Yesterday's polyploids and the mystery of diploidization. Nature Reviews Genetics **2**, 333–341 (2001)
25. Albertin, W., Marullo, P.: Polyploidy in fungi: evolution after whole-genome duplication. Proc. R. Soc. B **279**, 2497–2509 (2012). doi:10.1098/rspb.2012.0434
26. Gallardo, M.H., González, C.A., Cebrián, I.: Molecular cytogenetics and allotetraploidy in the red vizcacha rat Tympanoctomys barrerae (Rodentia, Octodontidae). Genomics **88**, 214–221 (2006)
27. Rivera, A.S., Pankey, M.S., Plachetzki, D.C., Villacorta, C., Syme, A.E., Serb, J.M., Omilian, A.R., Oakley, T.H.: Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. BMC evolutionary biology **10**(1), 123 (2010)
28. Hardison, R.C.: Evolution of hemoglobin and its genes. Cold Spring Harbor perspectives in medicine **2**(12), 011627 (2012)
29. Hu, C., Lin, S.-y., Chi, W.-t., Charng, Y.-y.: Recent gene duplication and subfunctionalization produced a mitochondrial GrpE, the nucleotide exchange factor of the Hsp70 complex, specialized in thermotolerance to chronic heat stress in *arabidopsis*. Plant physiology **158**(2), 747–758 (2012)
30. Remnant, E.J., Good, R.T., Schmidt, J.M., Lumb, C., Robin, C., Daborn, P.J., Batterham, P.: Gene duplication in the major insecticide target site, rdl, in *drosophila melanogaster*. Proceedings of the National Academy of Sciences **110**(36), 14705–14710 (2013)
31. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**, 57–63 (2009)
32. Soneson, C., Delorenzi, M.: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics **14**(91) (2013)
33. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(550) (2014). doi:10.1186/s13059-014-0550-8
34. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology **11**(R106) (2010)
35. Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., Wendel, J.F.: Homoeolog expression bias and expression level dominance in allopolyploids. New Phytologist **196**, 966–971 (2012)
36. Wilks, S.S.: The large sample distribution of the likelihood ratio test for testing composite hypotheses. Ann. Math Statist. **9**(1), 60–62 (1938). doi:10.1214/aoms/1177732360
37. Edger, P.P., Smith, R.D., McKain, M.R., Cooley, A.M., Vallejo-Marin, M., Yuan, Y., Bewick, A.J., Ji, L., Platts, A.E., Bowman, M.J., et al.: Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140 year old naturally established neo-allopolyploid monkeyflower. bioRxiv, 094797 (2016)
38. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289–300 (1995)
39. Liu, Y., Zhou, J., White, K.P.: RNA-seq differential expression studies: more sequence or more replication? Bioinformatics **30**(3), 301–304 (2013)
40. Schurch, N.J., Schofield, P., Gierlinski, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., Barton, G.J.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA **22**(6), 839–851 (2016)
41. Roulin, A., Auer, P.L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R.W., Jackson, S.A.: The fate of duplicated genes in a polyploid plant genome. The Plant Journal **73**(1), 143–153 (2013)
42. Coate, J.E., Doyle, J.J.: Variation in transcriptome size: are we getting the message? Chromosoma **124**(1), 27–43 (2015)
43. Robinson, M.D., Smyth, G.K.: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**(2), 321–332 (2008)
44. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4), 357–359 (2012)
45. Anders, S., Pyl, P.T., Huber, W.: HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics, 638 (2014)