New Phytologist

Research

Expression-level support for gene dosage sensitivity in three *Glycine* subgenus *Glycine* polyploids and their diploid progenitors

Jeremy E. Coate[1], Michael J. Song[1], Aureliano Bombarely[2] and Jeff J. Doyle[3]

[1]Department of Biology, Reed College, Portland, OR 97202, USA; [2]Department of Horticulture, Virginia Tech, Blacksburg, VA 24061, USA; [3]School of Integrative Plant Science, Plant Breeding and Genetics Section, Cornell University, Ithaca, NY 14850, USA

Author for correspondence:
Jeremy Coate
Tel: +1 9712958785
Email: jcoate@reed.edu

Received: *2 February 2016*
Accepted: *2 June 2016*

New Phytologist (2016) **212:** 1083–1093
doi: 10.1111/nph.14090

Key words: gene balance hypothesis, gene dosage sensitivity, gene duplication, gene expression, *Glycine*, polyploidy, tandem duplication.

Summary

- Retention or loss of paralogs following duplication correlates strongly with the function of the gene and whether the gene was duplicated by whole-genome duplication (WGD) or by small-scale duplication. Selection on relative gene dosage (to maintain proper stoichiometry among interacting proteins) has been invoked to explain these patterns of duplicate gene retention and loss. In order for gene dosage to be visible to natural selection, there must necessarily be a correlation between gene copy number and gene expression level (transcript abundance), but this has rarely been examined.
- We used RNA-Seq data from seven *Glycine* subgenus *Glycine* species (three recently formed allotetraploids and their four diploid progenitors) to determine if expression patterns and gene dosage responses at the level of transcription are consistent with selection on relative gene dosage.
- As predicted, metabolic pathways and gene ontologies that are putatively dosage-sensitive based on duplication history exhibited reduced expression variance across species, and more coordinated expression responses to recent WGD, relative to putatively dosage-insensitive networks.
- We conclude that selection on relative dosage has played an important role in shaping gene networks in *Glycine*.

Introduction

Gene duplication, whether involving individual genes or through whole-genome duplication (WGD; polyploidy), increases genetic complexity, and is considered a major driver of evolution (Ohno, 1970). Within a few million yr, however, most duplicated genes are silenced and/or removed from the genome (Lynch & Conery, 2000, 2003; Otto & Whitton, 2000; Blanc & Wolfe, 2004; Scannell *et al.*, 2007; Moghe *et al.*, 2014). Prevailing theories to explain patterns of long-term retention of duplicated genes include subfunctionalization (Force *et al.*, 1999; Stoltzfus, 1999; Adams *et al.*, 2003; Innan & Kondrashov, 2010), escape from adaptive conflict (Des Marais & Rausher, 2008), neofunctionalization (Hughes, 1994; Guan *et al.*, 2007; Conant & Wolfe, 2008), selection on absolute gene dosage (Bekaert *et al.*, 2011; Hudson *et al.*, 2011), and selection on relative gene dosage (the gene balance hypothesis (GBH); Papp *et al.*, 2003; Freeling & Thomas, 2006; Freeling, 2009; Birchler & Veitia, 2007, 2010, 2012; Moghe *et al.*, 2014).

Of these, only the GBH provides a clear explanation for the inverse retention patterns commonly observed between duplicates produced by WGD and those produced by small-scale duplication (SSD; e.g., tandem duplicates) (Freeling, 2008, 2009; Birchler & Veitia, 2010; Conant *et al.*, 2014) – namely, that classes of genes with greater retention of duplicates from WGD than the genome-wide average tend to have fewer than average duplicates from SSD and vice versa. Freeling's (2009) concise description of the GBH provides the basis for this relationship:

> This hypothesis postulates that any successful genome has evolved, by many stepwise positive selections, an optimum balance (ratio) of gene products binding with one another to make multisubunit complexes, or, alternatively, balances of gene products involved in multiple steps in regulatory cascades. … The more subunits per complex, or the more steps per cascade, then the more 'connected' (in a network of dependency) are the individual product participants. The more 'connected' the product, the more sensitive the phenotype is expected to be to changes in product concentration. If the gene product's concentration is not optimum, as dictated by the stoichiometry of the complex or cascade, then fitness is lowered and disease ensues. … Such gene product–level changes could occur genetically via over- or underexpression mutants, modifier mutants, and, of course, duplications/deficiencies.

Because SSD typically affects only a single gene, it introduces an imbalance relative to the other genes in a network or complex.

WGD, on the other hand, duplicates every gene simultaneously, maintaining balance at the level of genes. The GBH, therefore, predicts that gene duplicates from WGD will tend to be retained, and paralogs produced by SSDs will tend to be removed, by selection to preserve balance for 'connected' genes (Papp *et al.*, 2003; Thomas *et al.*, 2006; Hakes *et al.*, 2007; Freeling, 2009; Birchler & Veitia, 2012; Moghe *et al.*, 2014). The expected over-retention of paralog pairs, following WGD, in gene ontology (GO) terms and networks populated with genes encoding members of connected proteins has been observed in several paleopolyploids, including yeast (Papp *et al.*, 2003; Davis & Petrov, 2005; Conant, 2014), Arabidopsis (Cannon *et al.*, 2004; Seoighe & Gehring, 2004; Maere *et al.*, 2005; Freeling & Thomas, 2006; Freeling, 2008, 2009), poplar (Freeling, 2008), rice (Tian *et al.*, 2005; Wu *et al.*, 2008) and soybean (Coate *et al.*, 2011; Coate & Doyle, 2013). Conversely, GO terms and networks dominated by genes with both paralogs retained following SSD include fewer protein interactions and fewer core regulatory functions than do genes retaining both copies (homoeologs) following WGD (Maere *et al.*, 2005; Freeling, 2009). Consistent with genomic observations, heterozygous mutations cause greater reductions in fitness if they affect genes that function in complexes than if they affect genes that do not (Papp *et al.*, 2003). Consequently, although other mechanisms have undoubtedly been important in the preservation of some paralogs (Conant, 2014), selection on relative dosage is widely considered to explain much of the evolutionary history of duplicated genes (Freeling, 2009).

Because the GBH accounts for patterns of gene retention by postulating effects of gene copy number on protein stoichiometry, there must be a strong positive correlation between gene dosage and protein amount. This, in turn, requires that transcript abundance must increase with gene dosage in order to increase protein abundance. This leads to a prediction of the GBH at the transcriptional level: every gene in a dosage-sensitive network or complex should exhibit a similar transcriptional response to the change in dosage, regardless of the level of transcription in the initial, preduplication condition. If not, even 'balanced' duplications (e.g. WGD) would alter protein stoichiometry. This prediction has not been tested directly, and so the extent to which expression patterns are consistent with duplication-based inferences of selection on gene dosage remains an open question. A second currently untested expectation of the GBH is that transcript abundances among individuals of the same species should be more similar for an orthologous gene that is part of a dosage-sensitive network or complex than for a gene in an insensitive network or complex. This is because gene flow will bring together, in different individuals, different combinations of alleles for genes belonging to a given network or complex. If allelic expression variance is high across genes in the network or complex, then stoichiometry will be disrupted in combinations involving a high-expressing allele for one gene and a low-expressing allele for a second gene. Assuming that expression levels within networks and protein complexes are evolutionarily conservative, this expectation can be extended to different species as well.

One reason for the lack of studies examining the relationship between duplication history and gene dosage responses is that genome-wide patterns of duplicate retention and loss take millions of yr to emerge (Lynch & Conery, 2000, 2003). Consequently, studies that have examined these patterns have relied on ancient polyploids (e.g. Arabidopsis), for which diploid progenitors are unknown or extinct. As a result, there is no diploid baseline to which expression patterns in the polyploids can be compared.

The plant genus *Glycine* affords a solution to this problem. The ancestor of the entire genus experienced a polyploidy event *c.* 5–13 million yr ago (Ma; Schlueter *et al.*, 2004; Schmutz *et al.*, 2010; Doyle & Egan, 2010; Coate *et al.*, 2011; Fig. 1). Patterns of gene retention and loss from this ancient duplication are apparent in the sequenced genome of soybean (*G. max*; Schmutz *et al.*, 2010; Du *et al.*, 2012). A subsequent burst of genome duplication occurred in the subgenus *Glycine* within the last 500 000 yr (Bombarely *et al.*, 2014), producing an extensive and well-studied polyploid complex (reviewed in Doyle *et al.*, 2004; Sherman-Broyles *et al.*, 2014). This complex includes eight allopolyploid species ($2n = 78$, 80) and one autopolyploid species (*Glycine hirticaulis*; $2n = 80$), derived from various combinations of diploid ($2n = 38$, 40) genomes. Thus, expression profiles can be compared across multiple species sharing the same history of ancient genome duplication, and expression responses to recent WGD can be quantified by comparing the subgenus *Glycine* polyploids to their extant diploid progenitors. This affords the opportunity to assess if expression patterns in the short term and duplicate gene retention patterns in the long term are both consistent with selection on gene dosage, and to do so across what are, effectively, biological replicates of WGD.

Using the same approach as in previous studies in Arabidopsis (Seoighe & Gehring, 2004; Maere *et al.*, 2005; Freeling & Thomas, 2006), we assess metabolic networks and gene

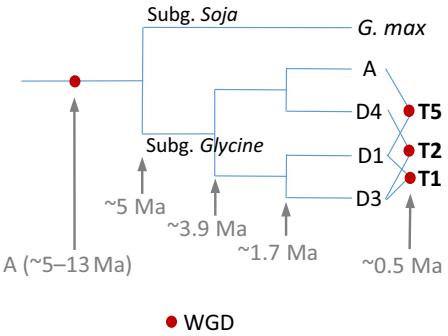

Fig. 1 The *Glycine* study system. The genus *Glycine* includes the cultivated soybean (*G. max*) in subgenus (subg.) *Soja*, as well as over 26 perennial species classified in subg. *Glycine*. All of these species share an ancient whole genome duplication event, designated 'A' (Coate *et al.*, 2011), *c.* 5–13 million yr ago (Ma), but are fully diploidized, with disomic segregation and bivalent formation. Within subg. *Glycine*, several additional independent allopolyploidy events have occurred within the last *c.* 0.5 Myr. The figure shows the diploid species combinations that gave rise to the three allotetraploids (T1, T2 and T5) that were used in this study. Polyploidy events are designated by red circles. Estimated divergence dates are from Bombarely *et al.* (2014).

ontologies in cultivated soybean for the degree to which they are populated by genes retaining duplicates produced by WGD vs SSD, thus identifying those hypothesized to be under selection for relative gene dosage. We then analyze RNA-Seq data for three neoallopolyploid *Glycine* species (termed T1, T2 and T5) and their four diploid progenitors (A, D1, D3 and D4; Fig. 1) to quantify expression patterns, and expression responses to WGD in three independent polyploid lineages.

We test the two corollaries of the GBH stated earlier: that dosage-sensitive metabolic networks and gene ontologies should exhibit coordinated expression responses to WGD, resulting in lower variance in transcript abundances across members than in dosage-insensitive networks and ontologies; and that genes in dosage-sensitive metabolic networks or gene ontologies should exhibit less variation in expression level across individuals within and between species than do genes that are dosage-insensitive.

Our results show that metabolic networks and gene ontologies with duplication histories indicative of relative dosage sensitivity do, in fact, exhibit expression responses consistent with selection on relative gene dosage. We conclude, therefore, that selection on relative dosage plays a pervasive role in shaping observed patterns of duplicate gene fractionation following WGD and SSD.

Materials and Methods

Duplication history in the soybean (*G. max*) genome

Syntenic blocks in soybean and the duplicated genes therein were downloaded from the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication/; Lee *et al.*, 2013). Blocks with mean $K_s \leq 0.40$ were assigned to the A WGD (5–13 Ma), and blocks with $K_s \geq 0.40$ and ≤ 1.2 were assigned to the B WGD (*c.* 54 Ma). Genes located in syntenic blocks that lacked a duplicate were designated as singletons for the corresponding duplication event. Tandem duplicates in soybean were downloaded from CoGe (https://genomevolution.org/CoGe/; Lyons & Freeling, 2008).

Gene expression analyses in perennial soybean (*Glycine* subgenus *Glycine*)

Leaf transcriptomes of three *Glycine* subgenus *Glycine* allotetraploid species (*G. dolichocarpa* (T2), *G. tomentella* T1 (T1) and *G. tomentella* T5 (T5)) and their diploid progenitor species (*G. clandestina* (A), *G. tomentella* D1 (D1), *G. tomentella* D3 (D3), and *G. syndetika* (D4)) were profiled by RNA-Seq (Illumina, San Diego, CA, USA). For each species, two to five accessions (seed descended by selfing from wild-collected individuals) from the CSIRO Division of Plant Industry Perennial *Glycine* Germplasm Collection were sequenced and treated as biological replicates for the species. For each accession, leaflets were pooled from six individuals. Plant growth, tissue collection, RNA extraction, RNA-Seq library construction, read processing, and read mapping were all performed as previously described (Bombarely *et al.*, 2014). Reads were mapped to the soybean (*G. max*) reference genome (assembly Glyma1; http://phytozome.net/soybean.php; Schmutz *et al.*, 2010), and read counts per gene model (v.1.1 gene set) were determined with HTSeq (Anders *et al.*, 2015) using the '–m intersection-noempty' setting. All accessions in Bombarely *et al.* (2014) were included in the present study, with the exceptions of the synthetic tetraploid, A58-1, and its diploid parent (*G. canescens* G1232). For each gene model in soybean, the expression (reads per kilobase per million reads; RPKM) for a given species was calculated as the average of expression estimates across accessions in that species. Only genes with average expression per species > 1 RPKM were used in subsequent analyses. This cutoff was used to avoid cases where polyploid response variance (PRV) was artificially inflated as a result of very small RPKM values in the polyploid or its diploid progenitors. Of 54 175 gene models in the soybean reference genome, we obtained average expression estimates ≥ 1 RPKM for between 24 862 and 27 908 genes per species. An additional 11 702–14 242 genes were expressed at < 1 RPKM and where thus excluded from further analysis.

Gene ontology annotations for soybean gene models (v.1.1) were obtained from the *G. max* gene annotation file (Gmax_189_annotation_info.txt) file at Phytozome (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV9). Soybean metabolic networks (SoyCyc 5.0) were downloaded from the Plant Metabolic Network (Plant Metabolic Network, 2014; http://www.plantcyc.org/). Only GO terms with ≥ 20 genes, and metabolic networks with ≥ five genes were included in subsequent analyses. Where specified, GO terms were filtered using Revigo (Supek *et al.*, 2011) to retain a single representative from GO terms with high semantic similarity (terms with largely overlapping gene sets). Filtering was performed using the whole UniProt database, the SimRel semantic similarity measure, and an allowed similarity of 'small (0.5)'.

Results

Using data from the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication/; Lee *et al.*, 2013) and from CoGe (https://genomevolution.org/CoGe/; Lyons & Freeling, 2008), we categorized genes in the soybean genome as either singletons or duplicates from the A WGD (5–13 Ma; Coate *et al.*, 2011; Fig. 1), and as being tandemly duplicated or not (see the Materials and Methods section). We then examined whether functionally related genes (gene ontologies or SoyCyc metabolic networks; Plant Metabolic Network, 2014) exhibited an inverse relationship between the retention of paralogs produced by tandem duplication and those produced by WGD (high retention of duplicates from one duplication mechanism correlated with low retention of duplicate from the other). As predicted by the GBH and observed in Arabidopsis (Freeling, 2009), we observed an inverse relationship between retention of tandem duplicates and WGD duplicates (homoeologs), both for GO terms and for metabolic networks (linear regression for GO terms, slope = −0.971, $R^2 = 0.5334$, $F = 296.1$, df = 1 and 259, $P < 10^{-16}$; linear regression for metabolic networks, slope = −0.939, $R^2 = 0.5525$, $F = 435.9$, df = 1 and 353, $P < 10^{-16}$; Fig. 2a,b). The pattern observed for GO terms

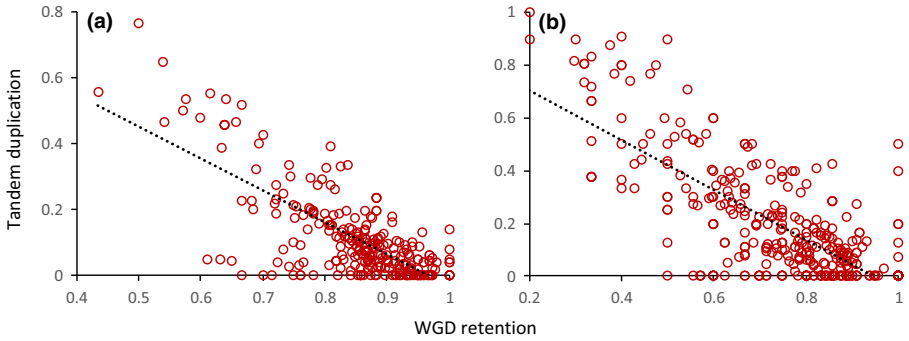

Fig. 2 Retention of paralogs from tandem duplication vs from whole-genome duplication (WGD). (a) Gene ontology terms (GO; terms with 20 or more genes). (b) Metabolic networks (SoyCyc; networks with five or more genes). Each circle represents a distinct GO term or metabolic network. The *y*-axis is the fraction of genes comprising that term that have a tandem duplicate in soybean. The *x*-axis is the fraction of genes comprising that term that have a duplicate in soybean from the A WGD event (5–13 million yr ago (Ma)). Dotted lines are trend lines.

persisted if the GO terms were first filtered based on semantic similarity (Supek *et al.*, 2011) to retain only a single representative from among those that are highly redundant (Supporting Information Fig. S1a).

If this reciprocal pattern is, in fact, driven by differences in dosage sensitivity, metabolic networks or GO terms with high WGD retention and low tandem retention (data points in the lower right corner of Fig. 2a,b) should be the most dosage-sensitive, whereas networks or GO terms with the opposite pattern (data points in the upper left, Fig. 2a,b) should be the least dosage-insensitive. We next examined if expression patterns within the gene groupings at each end of this spectrum were consistent with this inference.

Genes in putatively dosage-sensitive gene networks/ontology terms exhibit more coordinated expression responses to WGD than do genes in dosage-insensitive networks

If a functional class of genes (e.g. a GO) is dosage-sensitive, genes included in the term should also exhibit more highly coordinated expression responses to WGD than do genes in dosage-insensitive networks. To examine if this is generally true, we first divided GO terms and metabolic networks into two classes, those with lower than median WGD retention and higher than median tandem duplication (hereafter referred to as class I; yellow data points in Fig. 3a) and those with higher than median WGD retention and lower than median tandem retention (class II; blue data points in Fig. 3a). To assess degree of coordination in expression response to polyploidy, we utilized RNA-Seq data from three *Glycine* subgenus *Glycine* allotetraploid species and their four diploid progenitors (Fig. 1). To quantify the degree of coordination in polyploid expression responses among genes of a GO term or metabolic network, we divided polyploid expression by midparent diploid expression to determine fold-change in response to polyploidy for each gene (Fig. 3b). We then calculated the standard deviation of fold-changes for all genes in a given GO term or metabolic network (the PRV; Fig. 3b). As illustrated in Fig. 3(b), the PRV of a network is a measure of the degree to which the polyploid expression responses of genes within a network are correlated (low PRV indicates strong coordination of expression responses among genes, and a large PRV indicates weak coordination of expression responses).

As predicted, we observed significantly lower PRV for class II GO terms than for class I GO terms in all three polyploids (Fig. 3c–e; Table 1). This held true after filtering out redundant GO terms using Revigo (Fig. S1b; Table S1). We observed the same relationship for metabolic networks (Fig. S2; Table S2).

It should be noted that because we used median values of WGD retention and tandem duplication (0.87 and 0.07, respectively) as cutoffs for assigning GO terms and metabolic networks to duplication classes, some gene groups were assigned to different classes despite having nearly identical duplication histories (groups near the median values; Fig. 3a). For example, the GO term GO:0015991 (ATP hydrolysis coupled proton transport) has WGD retention of 88.6% and tandem duplication of 6.8% and was therefore assigned to class II, whereas GO term GO:0000287 (magnesium ion binding) has WGD retention of 86.7% and tandem duplication of 7.7%, and was assigned to class I. Thus, if the two duplication classes reflect degree of dosage sensitivity, it is likely that some GO terms or networks assigned to different classes have effectively equivalent dosage sensitivities, obscuring differences between the two classes. This makes the fact that there are significant differences in PRV between the two classes particularly noteworthy.

To mitigate this potential problem, we repeated the analysis after redefining the two classes using more stringent criteria, as follows. To be assigned to class I, a GO term had to have WGD retention in the first quartile of values (≤ 0.81), as well as tandem duplication in the third quartile of values (≥ 0.15), for all GO terms. Conversely, to be assigned to class II, a GO term had to have WGD retention in the third quartile of values (≥ 0.92), as well as tandem duplication in the first quartile of values (≤ 0.03), for all GO terms. Using these criteria, there was a clear disjunction between the two duplication classes (Fig. S3a), leaving 37 class II GO terms and 46 class I GO terms. As expected, with these criteria, PRV was again significantly lower for class II GO terms than for class I GO terms, and the differences in PRV were larger than in the less stringent analysis (Fig. S3b–d; Table S3).

The negative linear relationship between WGD and tandem retention observed for GO terms and metabolic networks (Fig. 2a,b) suggests that dosage sensitivity could be a

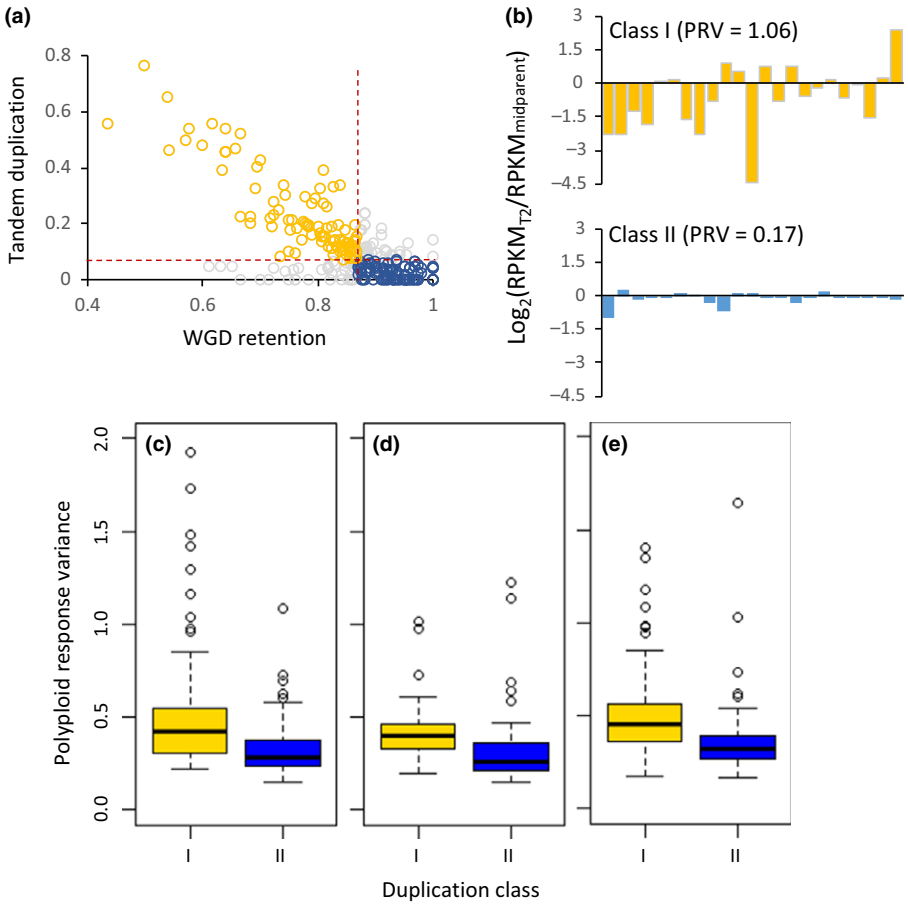

Fig. 3 Polyploid response variances (PRVs) are lower for class II gene ontology (GO) terms (putatively dosage-sensitive based on duplication history) than for class I GO terms (putatively dosage-insensitive based on duplication history). (a) As in Fig. 2, retention of paralogs from tandem duplication vs from whole-genome duplication (WGD) by GO term (terms with 20 or more genes) in soybean. Vertical and horizontal dotted lines indicate the median fraction of genes with polyploid or tandem duplicates, respectively, for all included GO terms (median fraction of genes with tandem duplicates = 0.07; median fraction of genes with A WGD duplicates = 0.87). GO terms with higher than median tandem duplication and lower than median polyploid duplication were designated as class I (yellow circles), and GO terms with lower than median tandem duplication and higher than median polyploid duplication were designated as class II (blue circles). (b) Variation in *Glycine dolichocarpa* (T2) polyploid expression response ($RPKM_{T2}/RPKM_{midparent}$; RPKM, reads per kilobase per million reads) by gene for representative class I and class II GO terms (class I, response to biotic stimulus (GO:0009607); class II, protein phosphatase type 2A complex (GO:0000159)). Each bar represents fold-change in expression in T2 relative to the midparent diploid expression level ($RPKM_{T2}/RPKM_{midparent}$; $\log_2$ scale). PRV was calculated as the standard deviation of fold-change values for all genes in the GO term. (c) PRV by class in *Glycine tomentella* T1 (T1). (d) PRV by class in *G. dolichocarpa* (T2). (e) PRV by class in *G. tomentella* T5 (T5). In (c) and (d) boxplots represent the median (horizontal line inside the box), the first and thirst quartiles (1Q and 3Q; lower and upper box edges, respectively), the larger of the minimum value and 1Q − 1.5× the interquartile range (IQR) (lower whisker), and the smaller of the maximum value and 3Q + 1.5× IQR (upper whisker).

Table 1 Summary statistics from Kruskal–Wallis tests of polyploid response variance (PRV) vs duplication class for all gene ontology (GO) terms with ≥ 20 genes expressed at ≥ 1 read per kilobase per million reads in three *Glycine* subgenus *Glycine* tetraploids (*G. tomentella* T1 (T1), *G. dolichocarpa* (T2), and *G. tomentella* T5 (T5))

	Median PRV				
Species	Class II	Class I	X^2	df	*P*-value
T1	0.28	0.42	32.4	1	1.3×10^{-8}
T2	0.26	0.40	42.1	1	8.6×10^{-11}
T5	0.32	0.45	38.4	1	5.7×10^{-10}

quantitative trait (i.e. there is continuous variation in the extent of dosage sensitivity) rather than a qualitative trait (i.e. GO terms and metabolic networks are either dosage-sensitive or not). Thus, in addition to grouping GO terms into discrete classes based on duplication history (I and II), we also examined whether PRV decreases in a linear fashion as the apparent dosage sensitivity (as indicated by duplication history) increases. To summarize apparent dosage sensitivity based on duplication history, we calculated a putative dosage sensitivity index (DSI) for each GO term or metabolic network as follows:

$$\text{WGD retention } (0-1) - \text{tandem retention } (0-1)$$

Thus, the DSI ranges from −1 to 1, with higher values indicating greater putative dosage sensitivity. We then plotted PRV vs DSI by GO term or metabolic network for each polyploid. As expected, we found that there was a highly statistically significant negative correlation between DSI and

PRV, for both GO terms and metabolic networks (Table S4; Fig. S4).

Putatively dosage-sensitive gene networks/ontology terms exhibit lower average expression level variation within and between species than do dosage-insensitive networks

As already explained, if a functional class of genes (e.g. a GO) is dosage-sensitive, a given gene included in the term should exhibit smaller variation in transcript abundance within species (among alleles) and across species (among orthologs), compared with genes in networks that are dosage-insensitive. To examine if this is generally true, we examined if class II GO terms and metabolic networks exhibit smaller variation in transcript abundance compared with class I GO terms and metabolic networks. To summarize variation in transcript abundance, we calculated the coefficient of variation (CV) for expression levels among accessions within species, and among average expression levels across species for each gene (Fig. 4a). We then averaged expression CVs ('expression variance' or EV) by GO term or metabolic network. As predicted, using the low-stringency cutoffs for assignment to duplication classes, class II GO terms have lower median EV than do class I GO terms (Fig. 4b–i). This pattern was consistent and significant in all seven species as well as across species (Table 2), and persisted after filtering out redundant GO terms using REVIGO (Fig. S1c; Table S5). We observed the same pattern when grouping by SOYCYC metabolic network, with class II networks again having significantly lower EV than do class I networks in all seven species as well as across species (Fig. S5; Table S6). As with PRV, we also repeated the EV analysis for GO terms using stringent cutoffs for assignment to duplication class (Fig. S6; Table S7), and by regressing EV vs DSI (Fig. S7; Table S8). As with PRV, EV was again significantly lower for class II GO terms than for class I GO terms using stringent cutoffs, with the differences in EV being larger than in the less stringent analysis. Similarly, we found a significant negative correlation between EV and DSI for both GO terms and metabolic networks (Table S8, Fig. S7).

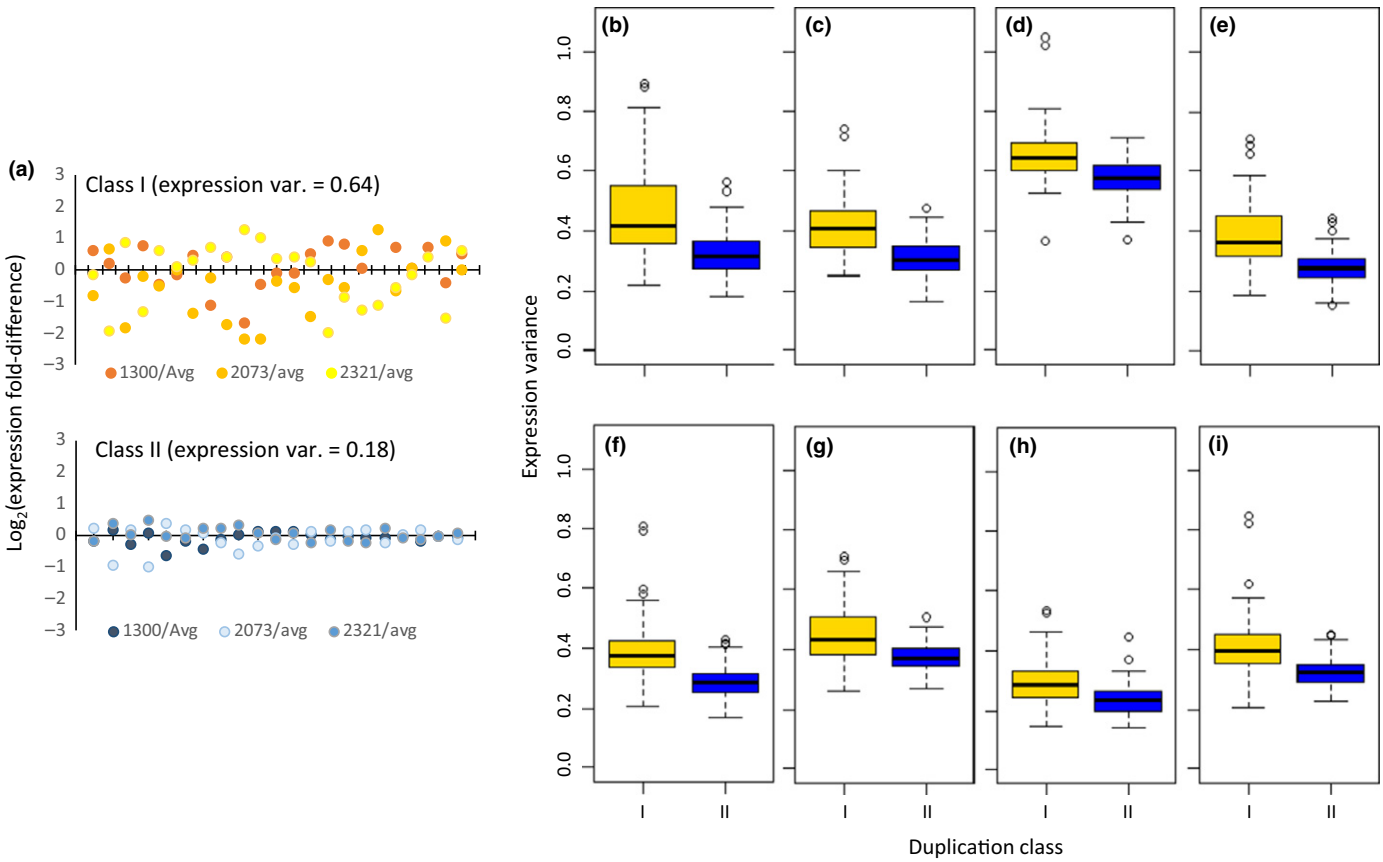

Fig. 4 Expression variances (EVs) are lower for class II gene ontology (GO) terms (putatively dosage-sensitive based on duplication history) than for class I GO terms (putatively dosage-insensitive based on duplication history). (a) Variation in expression levels by accession examples of class I and class II GO terms (class I, response to biotic stimulus (GO:0009607); class II, protein phosphatase type 2A complex (GO:0000159)). Each column of data points represents the expression levels of three D4 diploid accessions for a single gene. Expression levels are displayed as the $\log_2$ fold-difference from the species average. The coefficient of variation (CV) was calculated for each gene, and EV is equal to the average CV for all genes in the GO term. In the subsequent panels, the distribution of expression variances is shown by class in: (b) *G. clandestina* (A) diploids ($n = 3$); (c) *G. tomentella* D1 (D1) diploids ($n = 4$): (d) *G. tomentella* D3 (D3) diploids ($n = 4$); (e) *G. syndetika* (D4) diploids ($n = 3$); (f) *G. tomentella* T1 (T1) tetraploids ($n = 4$); (g) *G. dolichocarpa* (T2) tetraploids ($n = 5$); (h) *G. tomentalla* T5 (T5) tetraploids ($n = 2$); and (i) across *Glycine* subgenus *Glycine* species ($n = 7$). In (b)–(i), boxplots represent the median (horizontal line inside the box), the first and third quartiles (1Q and 3Q; lower and upper box edges, respectively), the larger of the minimum value and 1Q – 1.5× the interquartile range (IQR) (lower whisker), and the smaller of the maximum value and 3Q + 1.5× IQR (upper whisker).

Table 2 Summary statistics from Kruskal–Wallis tests of expression variance vs duplication class for all gene ontology (GO) terms with ≥ 20 genes expressed at ≥ 1 read per kilobase per million reads within and among seven species of *Glycine* subgenus *Glycine* (*G. clandestina* (A), *G. tomentella* D1 (D1), *G. tomentella* D3 (D3), *G. syndetika* (D4), *G. tomentella* T1 (T1), *G. dolichocarpa* (T2) and *G. tomentella* T5 (T5))

	Median EV				
Species*	Class II	Class I	X^2	df	*P*-value
A	0.32	0.42	178.6	1	3.1×10^{-12}
D1	0.30	0.41	137.4	1	7.1×10^{-14}
D3	0.58	0.64	115.9	1	7.3×10^{-11}
D4	0.28	0.36	238.4	1	1.8×10^{-15}
T1	0.28	0.37	193.0	1	2.8×10^{-15}
T2	0.37	0.43	174.9	1	3.6×10^{-11}
T5	0.23	0.28	100.9	1	8.5×10^{-8}
Cross-species	0.32	0.39	95.34	1	8.9×10^{-14}

*Statistics are reported for median expression variance (coefficient of variation) among individuals within species (A, D1, D3, D4, T1, T2 and T5), and among species (cross-species).

Expression patterns provide useful criteria for identifying gene networks whose evolution was shaped by selection on relative gene dosage

The preceding analyses showed that genes in GO terms and metabolic networks predicted to be dosage-sensitive based on duplication history (class II) tend to have lower EVs and PRVs than do genes in GO terms or metabolic networks that do not appear to be dosage-sensitive based on duplication history (class I). This supports the hypothesis that duplication histories are indeed generally indicative of dosage sensitivity. Nonetheless, some class II gene groupings exhibited EVs and/or PRVs similar to class I groupings (Fig. 5a). We reasoned that some gene groupings may have duplication histories indicative of dosage sensitivity as a result of chance or other, unidentified evolutionary processes, and that by combining evidence from both duplication history and gene expression response (EV and PRV), we should be able to refine the set of GO terms and metabolic networks that are in fact dosage-sensitive.

Of the 93 class II GO terms, EV was less than the median EV for class I GO terms in all seven species and across species, and thus consistent with dosage sensitivity, for 56 terms (59%). Similarly, PRV was consistent with dosage sensitivity in all three polyploids for 58 terms (62%). Overall, 70 of 93 terms that were putatively dosage-sensitive based on duplication history also had at least one expression-based metric consistent with dosage sensitivity (75%), and 44 out of 93 (47%) were dosage-sensitive based on both expression-based metrics in all seven species (Fig. 5b; Table S9).

Among the 44 GO terms with consistent duplication- and expression-based support for dosage sensitivity (Table S9) there are several terms that have been previously characterized as dosage-sensitive in Arabidopsis based on duplication history alone. These include terms related to protein modification (GO:0000159, GO:0008601, GO:0016301, GO:0006457, GO:0051082), ubiquitination (GO:0006511, GO:0004221), nucleic acid binding (GO:0003676), oxygen evolving complex (GO:0009654), and motors (GO:0003774) (Blanc & Wolfe, 2004; Seoighe & Gehring, 2004; Maere *et al.*, 2005; Li *et al.*, 2016). Conversely, among the class I GO terms with expression-based metrics consistent with a lack of dosage sensitivity are several terms that have been characterized as dosage-insensitive in Arabidopsis (Blanc & Wolfe, 2004; Maere *et al.*, 2005),

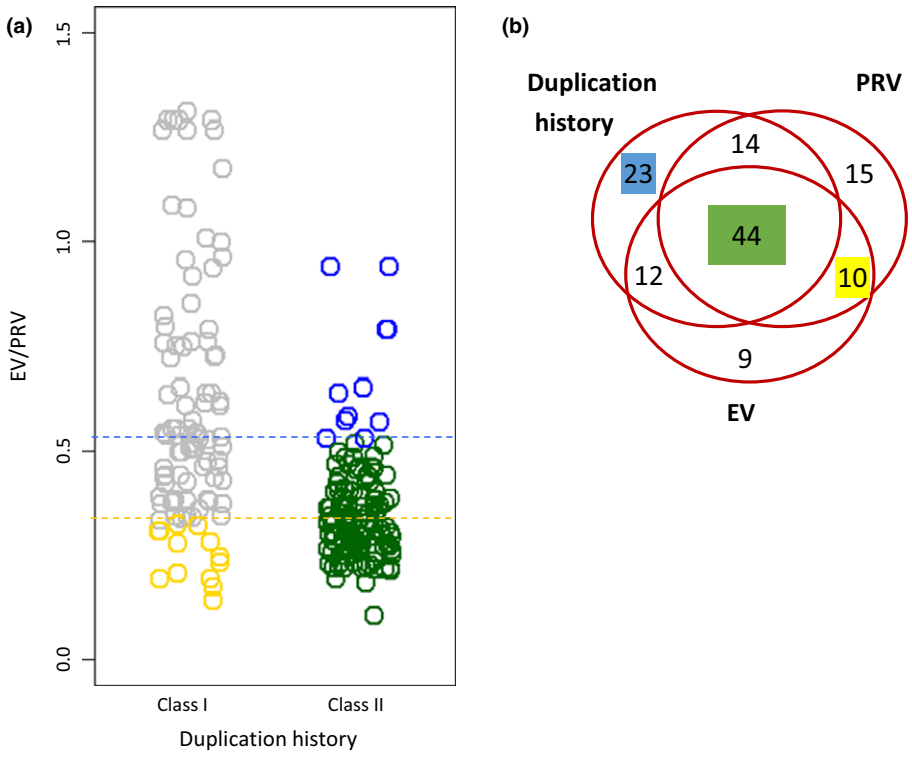

Fig. 5 Evidence supporting dosage sensitivity for gene ontology (GO) terms. (a) A representative distribution of expression variances (EVs) or polyploid response variances (PRVs) by duplication class. Dashed lines indicate the median values for class I (blue) and class II (yellow). Class II GO terms with EV or PRV less than the median value for class I are shown in green. Class II GO terms with EV or PRV values greater than the median value for class I are shown in blue. Class I GO terms with EV or PRV greater than the class II median are shown in gray. Class I GO terms with EV or PRV less than the median value for class II are shown in yellow. (b) Venn diagram of counts of GO terms that were designated as dosage-sensitive according to the above criteria within and across all seven species (*G. clandestina* (A), *G. tomentella* D1 (D1), *G. tomentella* D3 (D3), *G. syndetika* (D4), *G. tomentella* T1 (T1), *G. dolichocarpa* (T2) and *G. tomentella* T5 (T5)).

including terms related to defense (GO:0006952), lipid metabolism (GO:0006629), electron transport (GO:0009055), or, in a multi-species analysis of core eukaryotic gene families (Li *et al.*, 2016), terms related to translation and ribosomes (GO:0006412, GO:0042254, GO:0005840). The fact that our expression-based metrics characterize GO terms as dosage-sensitive or -insensitive in a way that is generally consistent with previous studies lends support to the conclusion that these metrics are, in fact, useful indicators of dosage sensitivity.

Additionally, several GO terms that appeared to be dosage-sensitive based on duplication history but not based on EV or PRV in *Glycine* (Fig. 5a,b; Table S9) have been previously characterized as dosage-insensitive based on duplication history (Blanc & Wolfe, 2004; Maere *et al.*, 2005; Li *et al.*, 2016). These included terms related to stress response (GO:0006950), and response to external stimulus (GO:0009416). These terms may exhibit duplication histories suggestive of dosage sensitivity because of chance or other historical contingency rather than because of selection on relative gene dosage. For example, such genes may have high levels of WGD retention because they are amenable to sub- or neofunctionalization. Note, however, that this would not explain why these genes also exhibit low levels of tandem duplication.

Alternatively, artifacts of automated annotation could result in misleading duplication histories for some GO terms. Genes involved in GO terms, such as those relating to stress responses, probably comprise a mix of rapidly evolving genes (e.g.defense-related genes) and conserved genes (e.g. core transcription factors). GO annotations in soybean were assigned based on homology searches using interpro2go (http://genome.jgi.doe.gov/PhytozomeV9/download/_JAMO/52b9c79e166e730e43a34ef9/Gmax_189_readme.txt?requestTime=1464890800), which could have resulted in an enrichment of highly conserved genes and under-representation of rapidly evolving genes. If the conserved genes are dosage-sensitive and the rapidly evolving genes are not, this would have the effect of elevating the apparent dosage sensitivity, as inferred by duplication history, for the GO term as a whole.

Conversely, we identified several GO terms that do not appear to be dosage-sensitive based on duplication history but whose expression-based metrics are indicative of dosage sensitivity (Fig. 5a,b; Table S9). Notably, several of these terms include genes whose products function in complexes, and that have previously been identified as dosage-sensitive in Arabidopsis (Blanc & Wolfe, 2004; Seoighe & Gehring, 2004; Maere *et al.*, 2005; Freeling, 2009). These include the terms 'signal transduction' (GO:0007165) and 'protein serine/threonine kinase activity' (GO:0004674). This further suggests that duplication history alone may present a misleading picture of dosage sensitivity, and that expression-based metrics are useful to refine and/or validate the extent to which a class of genes is truly dosage-sensitive.

Discussion

Several lines of evidence lend compelling support for the GBH, which, in turn, provides a plausible explanation – selection on relative gene dosage – for the nonrandom patterns of gene duplicate retention and loss observed in ancient polyploids (Papp *et al.*, 2003; Cannon *et al.*, 2004; Seoighe & Gehring, 2004; Davis & Petrov, 2005; Maere *et al.*, 2005; Tian *et al.*, 2005; Freeling & Thomas, 2006; Freeling, 2008, 2009; Wu *et al.*, 2008; Coate *et al.*, 2011; Conant, 2014). However, a critical assumption underlying this hypothesis – namely that there is a tight correlation among gene dosage, transcript abundance and protein abundance (Coate & Doyle, 2010, 2015) – had not been widely tested. If there is a disconnect between gene copy number and transcript abundance, then unbalanced duplications in dosage-sensitive networks would be invisible to selection on relative dosage. Moreover, if the dosage responses of individual genes within a network are uncoordinated, then protein stoichiometry could be disrupted by WGD even though relative gene dosage is preserved.

In yeast, knocking out one of two alleles caused a 50% reduction in protein abundance for 80% of the 730 genes examined (Springer *et al.*, 2010), and only 3% of genes exhibited complete or near-complete dosage compensation. This suggests that, in yeast, gene expression is generally well coupled to gene dosage, at least in the case of allelic deletions. It is unclear, however, to what extent this holds true in other taxa, or in response to gene or genome duplications. In *Drosophila melanogaster*, for example, 79% of 207 copy number variants showed no change in gene expression (Zhou *et al.*, 2011). In plants, a wide range of dosage responses has been observed in response to aneuploidy (Guo & Birchler, 1994) and WGD (Guo *et al.*, 1996; Coate & Doyle, 2010). For example, 25% of 15 761 genes assayed were dosage-compensated or showed negative dosage effects, and another 7% exhibited $>1:1$ dosage effects, following WGD in one of the subgenus *Glycine* allotetraploids used in the present study (T2; Coate & Doyle, 2010). This variation in dosage responses means that duplication history alone is inadequate to identify genes under selection for relative gene dosage, and raises the possibility that observed patterns of retention and loss are shaped by factors other than selection on relative gene dosage.

Despite the large range of gene dosage responses that have been observed in one of the polyploids studied here (T2; Coate & Doyle, 2010, 2015), we found that expression-based metrics of dosage sensitivity (EV and PRV) were correlated significantly with duplication-based metrics of dosage sensitivity. Thus, in addition to duplication histories, we have demonstrated that expression patterns are generally consistent with dosage sensitivity, adding a new layer of support for the hypothesis that selection on relative dosage drives the observed patterns of gene duplication and loss. In fact, the strength of expression-level support for selection on relative gene dosage is somewhat surprising, given the fact that we were comparing expression-based metrics of dosage sensitivity from recent WGDs (< 1 Ma) with duplication-based metrics from a much older WGD (5–13 Ma). Schnable *et al.* (2012) found that genes retained in duplicate from one WGD have only a 50% chance of retaining duplicates from subsequent WGDs. This emphasizes the point that duplication history may be shaped by chance as well as by selection, and also the fact that any given gene may be under selection for gene dosage

at one point in its evolutionary history but released from these constraints at other times (Schnable *et al.*, 2012; Conant *et al.*, 2014).

Furthermore, the majority of genes (*c.* 70%) from the A WGD event in soybean are still retained in duplicate, which is considerably more than are retained following polyploidy events of similar ages in maize (Schnable *et al.*, 2009) or *Brassica rapa* (Wang *et al.*, 2011). This could suggest that neutral gene losses are ongoing in soybean and that not all of the homoeologs that have persisted in duplicate are being retained by selection. This would tend to obscure the duplication-based signal of dosage sensitivity (Li *et al.*, 2016). In other words, some dosage-insensitive GO terms and metabolic networks were probably assigned to class II by virtue of not having sufficient time to lose WGD duplicates. The inclusion of dosage-insensitive GO terms and/or networks in class II would reduce any actual differences between class I and class II genes in terms of expression variance or polyploid response variance.

Conversely, the recent polyploidy events in subgenus *Glycine* are estimated to have occurred as much as 0.5 Ma (Bombarely *et al.*, 2014), and the diploid species diverged from each other as much as 5 Ma. Dosage constraints are believed to be gradually circumvented over time (Coate *et al.*, 2011; Schnable *et al.*, 2012; Conant *et al.*, 2014; Li *et al.*, 2016), which may have released constraints on cross-species EV and/or PRV for some dosage-sensitive GO terms or metabolic networks in our system.

Finally, gene functional annotations are sometimes incorrect (Yon Rhee *et al.*, 2008; Coate *et al.*, 2011). As a result, some metabolic networks or GO terms are 'contaminated' by incorrectly annotated genes that may have different dosage sensitivities than those of the genes that have been correctly assigned.

Consequently, for all of these reasons, our expression-based tests of selection on gene dosage should be viewed as quite conservative, and the fact that we still observe clear patterns of expression consistent with dosage sensitivity further suggests that selection on relative dosage has a strong and persistent influence on duplicate gene evolution.

Nonetheless, although expression patterns (EV and PRV) generally correlate with duplication patterns as predicted under the GBH, there are exceptions, and this emphasizes the point that some dosage-insensitive gene networks may exhibit duplication histories suggestive of dosage sensitivity, and vice versa, as a result of chance or other historical contingencies, or because there has simply not been sufficient time for differences in selection to resolve themselves at the level of duplication history. In this regard, expression-based metrics of dosage sensitivity (e.g., EV and PRV) could potentially be used to infer dosage sensitivity even in very recent polyploids whose genomes have experienced little or no fractionation subsequent to the duplication event.

Similarly, some gene networks may be dosage-sensitive at the protein level (i.e. under selection to maintain protein stoichiometry), but because protein dosage is decoupled from gene dosage, they are not sensitive to gene dosage (Moghe *et al.*, 2014). For a given protein complex or network, different taxa could maintain protein balance via different mechanisms (e.g. by maintaining balance in gene dosage or by transcriptional, post-transcriptional, or post-translational regulation). This could explain why many GO classes seem to exhibit lineage-specific differences in duplication histories (and, therefore, inferences of gene dosage sensitivity; Barker *et al.*, 2008; Carretero-Paulet & Fares, 2012; Li *et al.*, 2016). For example, Barker *et al.* (2008) found that GO terms that were over-retained following polyploidy in Compositae were quite different from those over-retained following polyploidy in Arabidopsis, and genes encoding ribosomal proteins have been characterized as dosage-sensitive in Arabidopsis but not in poplar or rice (Freeling, 2009) or in soybean (Table S7). In such cases where duplication histories differ across taxa, it is difficult to determine if this was the result of lineage-specific differences in dosage sensitivity, or if duplication history simply failed to reflect dosage sensitivity because of chance or other unknown forms of selection.

By integrating expression-based metrics of dosage sensitivity (EV and PRV) with duplication-based metrics, it should be possible to refine which gene networks have truly experienced selection for relative gene dosage. In this study, of 93 class II GO terms, only 44 also have expression patterns (low EV and PRV) that are uniformly consistent with dosage sensitivity (Table S7). These include terms for proteins that are known to function in complexes, and that have been characterized as dosage-sensitive in previous studies, and we suggest that these represent the most reliable circumscription of dosage-sensitive GO terms in *Glycine*.

Similarly, we identified several GO terms whose duplication histories do not suggest dosage sensitivity (class I), but whose expression patterns are generally consistent with dosage sensitivity. Among these GO terms are proteins that are known to function in complexes, and that have been characterized as dosage-sensitive in previous studies (e.g., protein kinase activity, signal transduction; Table S7). Some of these GO terms may be populated by dosage-sensitive genes, and represent cases in which duplication history alone gives a misleading indication of the dosage sensitivity.

Furthermore, high-level GO terms probably lump together proteins that are dosage-sensitive with others that are not. For example, genes with signal transduction functions have been characterized as dosage-sensitive in previous studies (Blanc & Wolfe, 2004; Seoighe & Gehring, 2004; Maere *et al.*, 2005), and terms related to signal transduction were similarly dosage-sensitive based on all metrics in our study. In contrast, one GO term related to signal transduction ('two-component signal transduction system (phosphorelay)'; GO:0000160) had a duplication history in soybean consistent with dosage sensitivity, but its interspecies EV and PRV in T2 both indicated a lack of dosage sensitivity (Table S2). Thus, classifying broad terms overall as dosage-sensitive or insensitive may be an oversimplification. By combining duplication history with expression-based metrics such as EV, and PRV, we should be able to quantify more reliably, and with greater resolution, dosage sensitivity and the extent to which selection on relative gene dosage has determined the fate of duplicated genes.

Acknowledgements

We thank Loren Reiseburg for the suggestion that genes in dosage-sensitive networks should exhibit reduced EV among individuals.

This work was supported by grants from the US National Science Foundation (0744306, 0939423 and 1257522).

Author contributions

J.E.C., M.J.S. and J.J.D. planned and designed the research; J.E.C. and M.J.S. collected the data; J.E.C., M.J.S. and A.B. analyzed the data; and J.E.C., M.J.S. and J.J.D. wrote the manuscript.

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences* **100**: 4649–4654.

Anders S, Pyl PT, Huber W. 2015. HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**: 2445–2455.

Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell Online* **23**: 1719–1728.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell Online* **19**: 395–402.

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist* **186**: 54–62.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences* **109**: 14746–14753.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**: 1679–1691.

Bombarely A, Coate JE, Doyle JJ. 2014. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. *Peer J* **2**: e391.

Cannon S, Mitra A, Baumgarten A, Young N, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology* **4**: 10.

Carretero-Paulet L, Fares MA. 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular Biology and Evolution* **29**: 3541–3551.

Coate JE, Doyle JJ. 2010. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biology and Evolution* **2**: 534–546.

Coate JE, Doyle JJ. 2013. Genomics and transcriptomics of photosynthesis in polyploids. In: Chen ZJ, Birchler JA, eds. *Polyploid and hybrid genomics*. Ames, IA, USA: John Wiley & Sons, 153–169.

Coate J, Doyle J. 2015. Variation in transcriptome size: are we getting the message? *Chromosoma* **124**: 27–43.

Coate JE, Schlueter J, Whaley A, Doyle J. 2011. Comparative evolution of photosynthetic genes in response to polyploid and non-polyploid duplication. *Plant Physiology* **155**: 2081–2095.

Conant GC. 2014. Comparative genomics as a time machine: how relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Molecular Biology and Evolution* **31**: 3184–3193.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19**: 91–98.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* **9**: 938–950.

Davis J, Petrov D. 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends in Genetics* **21**: 548–551.

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**: 762–765.

Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. 2004. Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): a study of contrasts. *Biological Journal of the Linnean Society* **82**: 583–597.

Doyle JJ, Egan AN. 2010. Dating the origins of polyploidy events. *New Phytologist* **186**: 73–85.

Du J, Tian Z, Sui Y, Zhao M, Song Q, Cannon SB, Cregan P, Ma J. 2012. Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. *Plant Cell Online* **24**: 21–32.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dynamics* **4**: 25–40.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome segmental, or by transposition. *Annual Review of Plant Biology* **60**: 433–453.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16**: 805–814.

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* **175**: 933–943.

Guo M, Birchler JA. 1994. Trans-acting dosage effects on the expression of model gene systems in maize aneuploids. *Science (Washington DC)* **266**: 1999–2002.

Guo M, Davis D, Birchler JA. 1996. Dosage effects on gene expression in a maize ploidy series. *Genetics* **142**: 1349–1355.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology* **8**: R209.

Hudson CM, Puckett EE, Bekaert M, Pires JC, Conant GC. 2011. Selection for higher gene copy number after different types of plant gene duplications. *Genome Biology and Evolution* **3**: 1369–1380.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **256**: 119–124.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**: 97–108.

Lee T-H, Tang H, Wang X, Paterson AH. 2013. PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research* **41**: D1152–D1158.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**: 326–344.

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

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* **3**: 35–44.

Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences: how to usefully compare plant genomes. *Plant Journal* **53**: 661–673.

Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, de Peer YV. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences, USA* **102**: 5454–5459.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S-H. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish *Raphanus raphanistrum* and three other Brassicaceae species. *Plant Cell* **26**: 1925–1937.

Ohno S. 1970. *Evolution by gene duplication*. New York, NY, USA: Springer.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* **34**: 401–437.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.

Plant Metabolic Network (PMN). 2014. *Summary of Glycine max, version 7.0* [WWW document] URL http://pmn.plantcyc.org/organism-summary?object=SOY, on www.plantcyc.org [accessed 22 June 2014].

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast

species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences* **104**: 8397–8402.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**: 868–876.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.

Schnable JC, Wang X, Pires JC, Freeling M. 2012. Escape from preferential retention following repeated whole genome duplications in plants. *Frontiers in Plant Science* **3**: 94.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends in Genetics* **20**: 461–464.

Sherman-Broyles S, Bombarely A, Powell AF, Doyle JL, Egan AN, Coate JE, Doyle JJ. 2014. The wild side of a major crop: soybean's perennial cousins from Down Under. *American Journal of Botany* **101**: 1651–1665.

Springer M, Weissman JS, Kirschner MW. 2010. A general lack of compensation for gene dosage in yeast. *Molecular Systems Biology* **6**: 368.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* **49**: 169–181.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**: e21800.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**: 934–946.

Tian C-G, Xiong Y-Q, Liu T-Y, Sun S-H, Chen L-B, Chen M-S. 2005. Evidence for an ancient whole-genome duplication event in rice and other cereals. *Yi Chuan Xue Bao* **32**: 519–527.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F *et al.* 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* **43**: 1035–1039.

Wu Y, Zhu Z, Ma L, Chen M. 2008. The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. *Molecular Biology and Evolution* **25**: 1003–1006.

Yon Rhee S, Wood V, Dolinski K, Draghici S. 2008. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics* **9**: 509–515.

Zhou J, Lemos B, Dopman EB, Hartl DL. 2011. Copy-number variation: the balance between gene dosage and expression in *Drosophila melanogaster*. *Genome Biology and Evolution* **3**: 1014–1024.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 REVIGO-filtered gene ontology analyses.

Fig. S2 Polyploid response variances (PRVs) are lower for class II metabolic networks (putatively dosage-sensitive based on duplication history) than for class I metabolic networks (putatively dosage-insensitive based on duplication history).

Fig. S3 Polyploid response variances (PRVs) are lower for class II GO terms than for class I GO terms (stringent definitions).

Fig. S4 Polyploid response variance (PRV) decreases with increasing values of putative dosage sensitivity index (DSI).

Fig. S5 Expression variances (EVs) are lower for class II metabolic networks (putatively dosage-sensitive based on duplication history) than for class I metabolic networks (putatively dosage-insensitive based on duplication history).

Fig. S6 Expression variances (EVs) are lower for class II GO terms than for class I GO terms (stringent cutoffs).

Fig. S7 Expression variance (EV) decreases with increasing values of putative dosage sensitivity index (DSI).

Table S1 Summary statistics from Kruskal–Wallis tests of polyploid response variance (PRV) vs duplication class for all gene ontology (GO) terms with ≥ 20 genes expressed at ≥ 1 RPKM, after removing redundant GO terms by REVIGO

Table S2 Summary statistics from Kruskal–Wallis tests of polyploid response variance (PRV) vs duplication class for SOYCYC metabolic networks

Table S3 Summary statistics from Kruskal–Wallis tests of polyploid response variance (PRV) vs duplication class (using stringent cutoffs*) for gene ontology (GO) terms

Table S4 Summary statistics from linear regressions of polyploid response variance (PRV) vs putative dosage sensitivity index (DSI) for SOYCYC metabolic networks

Table S5 Summary statistics from Kruskal–Wallis tests of expression variance (EV) vs duplication class for all gene ontology (GO) terms with ≥ 20 genes expressed at ≥ 1 RPKM, after removing redundant GO terms by REVIGO

Table S6 Summary statistics from Kruskal–Wallis tests of expression variance (EV) vs duplication class for SOYCYC metabolic networks

Table S7 Summary statistics from Kruskal–Wallis tests of expression variance (EV) vs duplication class (using stringent cutoffs) for gene ontology (GO) terms

Table S8 Summary statistics from linear regressions of expression variance (EV) vs putative dosage sensitivity index (DSI) for gene ontology (GO) terms

Table S9 Expression variance (EV) and polyploid response variance (PRV) for gene ontology terms assigned to duplication class I or II

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.