View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Inferring regulatory change from gene expression: the confounding effects of tissue scaling

Stephen H. Montgomery[1,2] and Judith E. Mank[1]

[1] Dept. Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

[2] Corresponding author: Stephen.Montgomery@cantab.net

Key words: allometry, isometry, grade-shifts, gene expression, RNA-Seq, regulatory evolution

Running title: Tissue scaling and transcript abundance

Abstract

Comparative studies of gene expression are often designed with the aim of identifying regulatory changes associated with phenotypic variation. In recent years large-scale transcriptome sequencing methods have increasingly been applied to non-model organisms to ask important ecological or evolutionary questions. Although experimental design varies, many of these studies have been based on RNA libraries obtained from heterogeneous tissue samples, for example homogenised whole bodies. Comparisons between groups of samples that vary in tissue composition can introduce sufficient variation in RNA abundance to produce patterns of differential expression that are mistakenly interpreted as evidence of regulatory differences. Here we present a simple model that demonstrates this effect. The model describes the relationship between transcript abundance and tissue composition in a two-tissue system, and how this relationship varies under different scaling relationships. Using a range of biologically realistic variables, including real biological examples, to parameterise the model we highlight the potentially severe influence of tissue scaling on relative transcript abundance. We use these results to identify key aspects of experimental design and analysis that can help to limit the influence of tissue scaling on the inference of regulatory difference from comparative studies of gene expression.

Introduction

A substantial amount of intra- and inter-specific diversity results from regulatory variation. Within species, a single genome can encode multiple distinct phenotypes by varying expression levels for the underlying loci. Examples of regulatory-based phenotypes include social insect castes (Toth *et al.* 2008), some instances of plastic alternative morphs such as dominant and subordinate turkeys (Pointer *et al.* 2013) or territorial, satellite and sneaker males in wrasses (Alonzo *et al.* 2000; Stiver *et al.* 2015), caring and non-caring in beetles (Parker *et al.* 2015), and a substantial proportion of differences between males and females (Moczek & Rose 2009; Khila *et al.* 2012). Similarly, across species or divergent populations, gene regulation provides an important route for the evolution of diversity (Carroll 2008; Stern & Orgogozo 2008) with many adaptive phenotypic changes linked to regulatory evolution (e.g. Shapiro *et al.* 2004; Steiner *et al.* 2007).

Given the importance of regulatory variation in shaping phenotypic diversity, transcriptome analyses based on RNA-Seq methods are increasingly used in evolutionary and ecological studies with the explicit aim of identifying genes that underlie phenotypic variation. These studies assume that differential gene expression is the result of altered transcriptional regulation which lead to phenotypic differences between groups of individuals. In many cases functional validation experiments have demonstrated causative relationships between variation in gene expression and variation in phenotypic development (e.g Abzhanov *et al.* 2006; Khila *et al.* 2012). However, functional validation is often inhibited by the polygenic nature of many traits, or a lack of functional genetics tools for the study species. For the moment at least, interpretation of the results of such studies are largely dependent on the assumption that expression differences have functional importance to the phenotypic variation observed across samples.

However, regulatory differences are not the only source of variation in gene expression in heterogeneous tissue samples. The composition of the tissue sampled for RNA extraction, and subsequent quantification of expression level, is a major source of variation that may undermine the validity of any inferred relationship between differential gene expression and phenotypic variation, but is yet to be scrutinised in any detail.

The design of published expression studies varies substantially. Although recent studies have demonstrated the potential to study gene expression in single cells (Sandberg 2014), these remain limited and most studies are based on larger samples, ranging from comparisons between organs (e.g. Enard *et al.* 2002; Khaitovich *et al.* 2004; Ghalambor *et al.* 2007; Brawand *et al.* 2011; Chen *et al.* 2015; Harrison *et al.* 2015), body parts composed of many constituent tissues such as heads (e.g. Parker *et al.* 2015; Standage *et al.* 2016), or whole body samples (e.g. Kvist *et al.* 2013; Feldmeyer *et al.* 2014; Hollis *et al.* 2014; Immonen *et al.* 2014; Stuglik *et al.* 2014). In all these cases, tissue samples are homogenized before mRNA extraction, purification and sequencing, with the resulting expression levels forming the primary data for comparison.

The homogenization of heterogeneous tissue samples provides one source of non-regulatory variation in estimated expression levels. The composition of these heterogeneous tissues depends on the nature of their constituent parts, the scaling relationships between these constituent parts, and the overall size of the tissue or individual. When comparing expression levels between groups of samples, for example groups of biological replicates of different sexes or different phenotypic morphs, the assumed connection between expression level and gene regulation is only valid if we also assume subcomponents of the tissue sample scale isometrically with total size, and do not vary between the groups under comparison. Numerous biological examples suggest isometry between traits is not the norm (Voje 2016), strongly questioning the validity of how we interpret comparative studies of gene expression.

Under isometric scaling the relationship between two component traits is one-to-one. Any individual, regardless of its total size, will have an equal percentage of its mass given over to its constituent parts. Deviation from isometry means this one-to-one relationship is no longer true (Figure 1, rows 1 to 3). As total size varies, an allometric relationship results in the size of component parts of a tissue sample varying to a greater or lesser degree and, as a result, the proportional size of each tissue component can vary. For example, the effects of both scaling patterns can be illustrated in fiddler crabs with asymmetric claw sizes. The smaller ‘minor’ claw scales isometrically with body size, whereas the larger ‘major’ claw scales with positive allometry, or hyper-allometry (Rosenberg 2002). Hence, as body size

increases the size of the minor claw as a proportion of body mass is constant, whereas the size of the major claw becomes disproportionately larger.

When sampling heterogeneous tissue, different forms of scaling relationships will affect comparative studies of gene expression in different ways. Isometry does not present a problem for studies of gene expression because the proportion of the RNA library attributable to a given tissue is constant (Figure 1, panels A3, A4). Any robust and repeatable change in expression level is therefore likely to be attributed to regulatory variation between the groups under comparison. However, under non-isometric scaling this is no longer the case. If we consider the allometric equation ($y = \alpha x^{\beta}$), isometry assumes the scaling coefficient, β, is one (Figure 1A1, A2 and A3). Under hyper-allometry, or positive allometry, β is greater than one. In this case, as trait x increases in size, trait y increases in size more rapidly (Figure 1B1). As a result, the size of y as a proportion of the total size increases in larger individuals (Figure 1B2 and B3). In contrast, under hypo-allometry, or negative allometry, β is less than one and as trait x increases in size trait y increases more slowly and accounts for a smaller proportion of total size in larger individuals (Figure 1C1, C2 and C3). As the proportions of each sub-tissue in a sample change, expression levels of some genes in RNA-Seq datasets could vary in a way that looks like regulatory variation, but is in fact a sampling artefact.

A further confounding effect arises when groups of samples differ in their scaling coefficient, β, or the scaling constant α (Figure 1D1, E1). For example, variation in α results in 'grade-shifts' between groups of individuals under comparison, for example the two sexes, two phenotypic morphs or two populations or species (Figure 1D1). This is often observed between morphs within species, for example in testis mass between male morphs (e.g. Tomkins & Simmons 2002), or between species, such as in the size of testes under different reproductive ecologies (Harcourt *et al.* 1981) or of different brain components (Barton & Harvey 2000; Barton & Venditti 2014). Grade-shifts are also commonly observed in experimental selection lines and appear to be a major axis of evolvability (e.g. Wilkinson 1993; Emlen 1996; Egset *et al.* 2012; Kotrschal *et al.* 2013). Where these grade-shifts occur, individuals will differ in the proportions of their constituent parts regardless of total size (Figure 1D2 and D3).

Shifts in β are perhaps more rare in nature, possibly due to stronger developmental or functional constraint (e.g. Egset *et al.* 2012), but they do occur between cell or tissue types within tissues and across species (Simmons & Tomkins 1996; Herculano-Houzel *et al.* 2015). The main result of β differences between groups is that the similarity of tissue composition between those groups will vary with total size (Figure 1E1, E2 and E3). This will likely increase variance within a group as well as predictably altering mean transcript abundance between groups. As a result of non-isometric scaling relationships, groups of individuals - be they species, morphs, castes, or sexes - can vary substantially in body or tissue composition. In the case of hyper- and hypo-allometry this can occur in the absence of any functional or developmental reorganization, and is a mere consequence of variation in total size. The proportion, or percentage size, of different tissue components is important for studies of gene expression because RNA-Seq is always a proportional rather than absolute measure of expression level, regardless of sequencing depth. RNA abundance within a sample is therefore directly related to the *proportion* of cells in the sample expressing a gene at a certain level. As a result of this, variation among samples in the proportion of different cell types will alter the proportion of mRNA transcripts in the homogenized tissue pool, and therefore expression level estimates. Expression levels are therefore related to variation in proportions of tissue components (Figure 1, rows 3 and 4) rather than the variation around scaling relationships between those tissues, i.e. 'relative' size (as indicated in Figure 1, row 1). As a result, comparing variation in expression level between samples of homogenized, heterogeneous tissue may partly reflect differences in regulation, but could also reflect differences in composition. Unfortunately, these alternatives are not mutually exclusive, further complicating analysis of expression variation.

Differences in tissue scaling are not problematic to studies of RNA-Seq if the sole aim is to simply identify expressed genes. However, if the aim is to identify loci with altered regulation that underpins phenotypic variation, and then to subsequently study the evolutionary characteristics of those loci, tissue scaling becomes a key concern. This is perhaps more apparent in RNA-Seq analyses based on whole-body or amalgamated body parts because of the obvious potential for variation in the proportion of constituent tissues. However, scaling relationships between cell types within organs can also deviate from isometry and can differ between groups of

individuals or species (e.g. Herculano-Houzel *et al.* 2015). As such, finer-scale preparations may also be affected.

If allometric scaling contributes to large differences in gene expression, the central assumption of comparative studies of gene expression, that divergence in expression level reflects divergence in gene regulation, would be difficult to support. However, it is not clear what magnitude of differences we might expect under different scaling scenarios, or how this may vary across different expression levels. Without this knowledge, it is difficult to know when a shift in gene expression is more likely explained by regulatory variation than an effect of scaling, or vice versa. Our goal here is to explore the ways that tissue scaling can influence RNA-Seq studies using a modelling approach, and to offer some suggested guidelines that may facilitate improved interpretation of RNA-Seq studies that aim to study the phenotypic effects of variation in gene regulation.

Materials and methods

A tissue-scaling model of gene expression differences

To explore the effects of allometric scaling on patterns of gene expression we developed a simple model. In this model, a sample is comprised of two tissues, *x* and *y*, which scale with each other according to the allometric equation $y = \alpha x^{\beta}$ where β is the scaling coefficient and α is the scaling constant. The total size of the sample (*S*) is therefore the sum of tissue *y* and tissue *x*:

$$S = y + x = \alpha x^{\beta} + x \qquad \text{[eq. 1]}$$

Within each tissue, we assume the total expression level of an individual gene (*C*) is constant for a given unit of size (e.g. mass or cell number). To reflect the independent regulation of expression level for different genes in tissue types we allow this constant to vary between tissues, and between genes. The number of transcripts for a gene in tissues *x* and *y* are therefore:

$$Transcript\ count\ of\ gene\ a\ in\ tissue\ x = C_{a,x} \times x \qquad \text{[eq. 2]}$$

$$Transcript\ count\ of\ gene\ a\ in\ tissue\ y = C_{a,y} \times \alpha x^{\beta} \qquad \text{[eq. 3]}$$

In a homogenised sample, the total expression will be the sum of eq. 2 and eq 3. However, with current methods, the observed value will be a proportion of the total

transcript count (C_{total}). This is modelled as the average expression of a gene across both tissues (C_m) multiplied total sample size (S) and the number of expressed genes (G):

$$C_{total} = C_m \times [\alpha x^{\beta} + x] \times G \quad \text{[eq. 4]}$$

The relative expression of an individual gene (RE_a) will therefore equal the sum of its abundance in tissues x and y (eq. 2 and eq. 3) divided by the total transcript count (C_{total}; eq. 4):

$$RE_a = \frac{[C_{a,x} \times x] + [C_{a,y} \times \alpha x^{\beta}]}{[\alpha x^{\beta} + x] \times G \times C_m} \quad \text{[eq. 5]}$$

RE_a is easily converted to be equivalent to commonly used measures of relative gene expression such as 'counts per million' (CPM), by simple multiplication:

$$CPM = RE_a \times 10^6 \quad \text{[eq. 6]}$$

CPM is used to compare the expression level of a gene between groups of samples, for example between sexes, morphs, populations or species. Significant shifts in log-transformed *CPM* can be identified using traditional statistics such as *t*-tests or a Mann-Whitney *U* test. The $\log_2$-fold change (*FC*) between two groups is calculated as:

$$FC = log_2(CPM_{group\ 1}) - log_2(CPM_{group\ 2}) \quad \text{[eq. 7]}$$

Using this model we can estimate *FC* between two samples which do not differ in the expression level of gene *a* but that can vary for x (and therefore y and S), α or β, as indicated by the subscript numbers:

$$FC = log_2\left(\frac{[C_{a,x} \times x_1] + [C_{a,y} \times \alpha_1 x_1^{\beta_1}]}{[\alpha_1 x_1^{\beta_1} + x_1] \times G \times C_m} \times 10^6\right) - log_2\left(\frac{[C_{a,x} \times x_2] + [C_{a,y} \times \alpha_2 x_2^{\beta_2}]}{[\alpha_2 x_2^{\beta_2} + x_2] \times G \times C_m} \times 10^6\right)$$

This model was used to investigate the expected effect on *FC* under three scenarios: i) effects of size differences under conserved allometric scaling by varying S between two groups while α and β remain constant, ii) effects of varying the allometric constant (α) between two groups while S and β remain constant, iii) effects of varying the allometric coefficient (β) between two groups while S and α remain constant. In each analysis, β was set according to the range of values (0.1-3.0) observed in over

3,200 datasets recently reviewed by Voje (2016). S was varied by setting different values of x. Across real datasets, values of x and α will vary greatly and depend on the units of measurements used. Generally, however, α is small relative to x. Unless otherwise stated we therefore set x to 10 units and α to 0.1. We also examined how the size of these effects varies with variable levels of tissue-biased expression (measured as $\log_2(C_{a,x})$-$\log_2(C_{a,y})$). In all comparisons we fixed G and C_m, to 10,000 and 5,000 respectively, to reflect raw values of read counts obtained in a recent RNA-Seq dataset (Harrison *et al.* 2015). $C_{a,y}$ was set to 5,000 so that results obtained reflect an 'average gene'. $C_{a,x}$ varied between 0 and 50,000. It is important to note that results obtained for genes limited to, or biased towards, x will be similar, but inverted relative to y-biased genes with a relationship defined by the rearranging the allometric equation for x.

To further explore the practical relevance of these effects we also used our model to simulate expected results using published scaling parameters from real biological data. We chose two examples to reflect the sorts of studies being conducted with real data: i) scaling relationships between soma and testis tissue in different male morphs from four species of insects; ii) scaling relationships between cell types in mammalian brains.

Results

i. Model effects

Effects of size differences under conserved allometric scaling

We modelled the effect of allometric scaling by varying S between two hypothetical groups, keeping α and β constant in order to identify the influence of simple size differences on the relative proportions of sub-tissues on comparative studies of gene expression (Figure 2A). Specifically, we used our model to compare gene expression levels between two groups, where $x = 10$ for group one, and $0.1 < x < 100$ in group two, a ten-fold change in size in both directions. β was fixed at either 0.1, 0.5, 1.0, 1.5 and 2.0 in both groups. As expected, under isometric scaling ($\beta = 1$) *FC* is consistently zero regardless of the magnitude of size differences between the two groups, or the extent of tissue-biased expression. Turning to allometric scaling, we first consider tissue-specific expression ($C_{a,y} = 5{,}000$; $C_{a,x} = 0$) as we anticipated

this would reflect the worst case scenario. The model predicts consistent differences in CPM between groups that increase with greater size differences, or greater deviation from isometry. The effects of negative and positive allometry generally mirror one another, except where extreme positive allometry results in *y* comprising nearly all of *S*, minimizing the influence of tissue-biased expression. The opposite will occur for *x*-specific genes. Large fold-changes (FC ≥1 or <-1) are expected to require relatively large size differences. For example, under strong negative allometry ($\beta = 0.1$) if $x = 10$ for group one, group two requires $x<4.5$ or >22 (a $\log_{10}(S$ ratio) of <0.45 or > 2.19; in Figure 2 the $\log_{10}(S$ ratio) is plotted to compress the variance for visual clarity) to produce a two-fold expression difference (FC ≥1 or <-1). Under strong positive allometry ($\beta = 2$) this occurs only when $x <3.25$ for group two. When the degree of tissue-bias in expression is varied ($C_{a,y}$ = 5,000; $C_{a,x}$ = 0-50,000), increasing tissue-bias in either direction results in larger *FC* (Figure 2B, C). This effect is amplified according to the degree to which β deviates from one. In summary, our model predicts that where the sample differs in mean size between groups under comparison any deviation from isometric scaling could produce difference in transcript abundance.

Effects of varying the allometric constant between groups

We next used our model to assess the impact that differences in the allometric constant between groups have on relative transcript abundance, modelling the expected effects of 'grade-shifts' between groups. This was done by varying α between two groups while *S* and β remained constant (Figure 3, panel A). With *x* set to 10 in both groups and an α of 0.1 in group one, we varied α in group two between 0.1 and 1 (a ten fold range). First considering tissue-specific genes ($C_{a,x}$ = 5,000; $C_{a,y}$ = 0), the model predicts absolute *FC* will increase linearly with the log-ratio of α values. When $\beta <1$, the magnitude of the effect is largely unaffected by variation in β. Where $\beta >1$, the effect is dampened as β increases because the contribution of expression in tissue *y* quickly overwhelms that of tissue *x*. The opposite will occur for *x*-specific genes. Large fold-changes (≥1 or <-1) occur from relatively small shifts in α. Under negative allometry ($\beta <1$), if α is 0.1 in group two, the FC is ≥1 or <-1when $0.05>\alpha>0.2$ in group two (an α ratio <0.5 or >2). Under positive allometry the necessary magnitude of shift in α to produce this size of effect

increases, but the opposite will occur for *x*-specific genes. Finally, when the degree of tissue bias in expression is varied ($C_{a,y}$ = 5,000; $C_{a,x}$ = 0-50,000), tissue-specificity is again always the worst-case scenario. Increasing tissue-bias in either direction produces larger *FC*, an effect amplified by increased variance in α between groups (Figure 4B, C). In summary, our model predicts that differences in allometric constants between groups under comparison can have a large impact on transcript abundance, regardless of the similarity in total size of the tissue sampled.

Effects of varying the allometric coefficient between groups

Finally, we used our model to predict how this will affect patterns of differential expression in scenarios where the total size across two groups is constant but the scaling relationships between their constituent parts differ. We first varied β while *S* and α remain constant setting β = 0.5 (Figure 4A) or 1.5 (Figure 4B) in group one, and $0.1 < \beta < 3$ in group two. We repeated this analysis using different values of *x* to explore how variation in β interacts with variation in size (Figure 4C,D). First, considering tissue-specific genes ($C_{a,y}$ =5,000; $C_{a,x}$ = 0), the model predicts *FC* will increase linearly with β until the contribution of expression in tissue *y* overwhelms that of tissue *x*. The opposite will occur for *x*-specific genes. We find that modest differences in β can produce large *FC* ($\geq$1 or <-1). For example, when *x* = 10 in both groups and β = 0.5 in group one, -1>*FC*>1 when 0.2> β >0.9 in group two (a β ratio of <0.4 or >1.8; Figure 4A). As *x* increases the shift in β necessary to produce this scale of difference decreases; when *x* = 100 it will occur when 0.3> β >0.7 (a β ratio of <0.6 or >0.78), when *x* = 1,000 it will occur when 0.4> β >0.6 (a β ratio of <0.8 or >1.2). Similar results are found regardless of the value set for β in group one. Again, when the degree of tissue-bias in expression is varied ($C_{a,y}$ = 5,000; $C_{a,x}$ = 0-50,000), genes with tissue-specific expression are always most affected. Increasing tissue-bias in either direction produces larger *FC*, an effect amplified by increased variance in β between groups (Figure 4C,D). In summary, any deviation between the scaling exponents governing the scaling relationships between tissue types in two groups will again lead to predictable differences in transcript abundance.

Tissue scaling can produce false negatives

The previous results focus on false-positives, however it is likely that the same scaling effects will obscure real patterns of group differences in gene expression. To illustrate this effect we used our model to vary $C_{a,x}$ between two groups. In group one $C_{a,x}$ and $C_{a,y}$ were both set to 5,000. In group two $C_{a,y}$ was again set to 5,000 but $C_{a,x}$ was set to either 20,000, 10,000, 5,000, 2,500 or 1,250. This simulates the gain of tissue-biased expression in group two with an inter-group $\log_2$-fold change (FC) for $C_{a,x}$ of 2, 1, 0, -1 and -2 respectively. We first examined the effects of varying the average size of the sample (as described above with $x = 10$ for group one, and $0.1 < x < 100$ in group two) whilst keeping α and β constant. We set the scaling parameters to reflect moderately hyper-allometric scaling. As expected, as the size difference between groups increases, the estimated FC rapidly declines (Figure 5A). Turning next to inter-group differences in α, we set α to 0.1 in group one and varied α in group two between 0.1 and 1, whilst keeping x at 10 and β at 1.5. Again, as the discrepancy between α_1 and α_2 increases, the measured FC decreases exponentially, with even large FC differences in $C_{a,x}$ dropping below and FC of ±0.5 (Figure 5B). Finally, we examined the effects of varying β by keeping β at 1.5 in group 1, and varying β between 0.1 and 3 in group 2. α was set to 0.1 and x was set to 10 in both groups. Again an effect of reduced detected FC is found with increase inter-group differences in scaling parameters. Here, the effect is sigmoidal with an accelerated decline in FC as the β ratio exceeds ~2.5 (Figure 5C). Similar results are obtained with alternative values for the scaling parameters. Together the model demonstrates that with increasing deviation from isometry, or increasing inter-group differences in scaling, the detection of true shifts in gene expression becomes increasingly inaccurate, potentially leading to substantial numbers of false negatives.

ii. Biological examples

Testes size in male morphs

Relative testes size can vary dramatically across species, often in association with reproductive competition imposed by multiple-mating in females (Harcourt *et al.* 1981; Hosken & Ward 2001). In many species multiple male morphs have evolved to exploit alternative reproductive strategies (Gross 1996; Sinervo & Lively 1996). These morphs typically reflect trade-offs in pre and post-copulatory male-male

competition, and by extension, investment in sperm production and testes size. Many studies of gene expression in smaller organisms, such as insects, utilise whole-body samples in order to avoid laborious dissections and/or to obtain sufficient RNA for sequencing. However, as whole-body samples are particularly prone to tissue scaling problems, we explored how differences in testes size might affect results using published scaling parameters from $\log_{10}$-$\log_{10}$ regressions between soma and testis mass for two species of dung beetle, (*Onthophagus taurus* and *O. binodis*), a burrowing bee (*Amegilla dawsoni*), and an earwig (*Forficula auricularia*) (Tomkins & Simmons 2002). Each of these species has two male morphs, one that guards females and one that adopts a 'sneaky' male strategy. We are not aware of any whole-body RNA-Seq analyses based on these particular species, but rather use them as an example of how the composition of the tissue sampled may affect perceived levels of differential expression between groups of individuals without the need to invoke morph-specific regulation of gene expression.

For each pair of morphs, we used the estimated morph-specific values of β and α to parameterise the model (Table S1), and varied the degree of tissue-bias (here, towards the testis) in expression for an average gene by setting $C_{a,y}$ to 5,000 and $C_{a,x}$ to range incrementally between 0 and 50,000, with S set to an realistic body mass. We also extended this range to include increases in $C_{a,y}$ up to 50,000 whilst $C_{a,x}$ was set to 0 (i.e. soma-specific gene expression). We then plotted the estimated $\log_2$-fold change in expression (FC) between the morphs against the degree of tissue-bias ($\log_2(C_{a,x})$-$\log_2(C_{a,y})$). With the exception of *O. taurus*, the difference in gonad-soma scaling between morphs was sufficient to produce $FC \geq 0.5$ for genes modelled as testis-specific, with FC increasing with testis-specificity in expression (Figure 6A).

We further explored how this effect might influence the kind of statistical methods used in real analyses by simulating a modest dataset of 1,000 genes for 5 individuals of each morph using the scaling relationships as described above. Here, $C_{a,x}$ and $C_{a,y}$ for each gene were set as equal, random numbers between 1 and 50,000 with 100 testis-specific genes and 100 soma-specific genes. Across individuals $C_{a,x}$ and $C_{a,y}$ were constrained to be within 10% of expression level of the corresponding gene in the first simulated individual. Under these conditions we would not expect any evidence of significant expression differences between groups because there is no contribution of regulatory variation, as such, all gene expression differences are solely caused by scaling effects. When we plotted expression in both morphs against one

another, the correlations are significant, but show a range of *FC*. Importantly, a proportion of genes is identified as 'significantly differentially expressed' between morphs using standard *t*-tests with no fold-change threshold (Table 1). We next used these data in two multivariate analyses, often utilised in RNA-Seq studies. First we used a Principal Component Analysis (PCA) to compress the variation in the dataset into PCs, we then asked if these PCs are significantly different between morphs using a *t*-test. Second we used hierarchical clustering to test if the simulated data can separate each morph. In three of the four cases the clustering grouped morphs by gene expression and had one PC significantly associated with morph, accounting for 10-16% of variance (Figure 6B-E). Note, these values will depend on the permitted degree of variation in expression of a gene between simulations. However, in each of these analyses the influence of allometry directly reflects differences in the estimated ratio of percentage testis volumes between morphs (Table 1).

Cellular scaling in mammalian brains

Many comparative studies have been conducted across species with the aim of identifying species-specific shifts in gene expression. These may focus on specific organs or tissues, but the scaling relationships among cell types could potentially drive some of the observed patterns. Recently interspecific datasets on the cellular composition of mammalian brain regions have revealed variation in the scaling relationship between neurons and non-neuronal cells between brain regions, and for individual structures across mammalian orders (Herculano-Houzel *et al.* 2015). We used these data to explore how allometric relationships between cell types might affect estimates of relative levels of gene expression across species. Using published data we re-estimated the scaling relationship between neurons and non-neuronal cells for two brain structures, the cerebral cortex and cerebellum, across two mammalian orders, glires and primates, using Phylogenetic Generalised Least Square Regressions (Pagel 1999) (Table S2). We used these scaling parameters to explore how variation in cellular scaling might affect comparative studies of gene expression on brain tissue.

We first examined the effects of varying *S* assuming a conserved allometric relationship between neuron and non-neuronal cell number within each order. By setting x_1 to the minimum and x_2 maximum values of non-neuronal cell number observed in each dataset we asked what size of $\log_2$-fold change (*FC*) in gene expression might be observed when comparing gene expression across species within

each order, at varying levels of cell-bias in gene expression. The results demonstrate moderate *FC* are expected, but their range varies across structures and orders (Figure 7). For the cerebral cortex (Figure 7, panel A), variation in *S* in primates produces more modest *FC* than observed in glires with the largest *FC* (1.49) predicted for genes expressed exclusively in neurons. In contrast, for the cerebellum the pattern is reversed. Primates are predicted to show a greater range of *FC* as *S* varies, with the largest *FC* (-2.8) predicted for genes expressed exclusively in non-neuronal cells (Figure 7, panel B). This difference in pattern between cerebral cortex and cerebellum is most likely related to the pattern of variation in β, which is higher in primates for the cerebral cortex, and higher in glires for the cerebellum.

We next explored how the difference in allometric parameters would affect comparisons of individuals (with constant *S*, set to the approximated midpoint in the overlap in ranges of *x* between groups) in different orders (i.e. under different β and α). For the cerebral cortex the model predicts modest *FC* between the two orders (-0.1<*FC*<0.3) (Figure 7, panel C), whereas for the cerebellum we predict a larger range in *FC*, with *FC* increasing as gene expression becomes increasingly biased towards non-neuronal cells (-0.97<*FC*<0.15) (Figure 7, panel D). The analyses above assume gene expression is related to cell number, independently of cell size, in the Supplementary Information we explore the effects of considering cell type mass, rather than number, which leads to broadly similar conclusions.

Discussion

Our results illustrate that non-isometric scaling relationships between tissue or cell types within groups of samples, and heterogeneity in scaling relationships across groups of samples, may influence inferred patterns of differential expression. This will occur at multiple biological levels, be it organ types within whole body samples, or cell type abundance when specific tissues are targeted for RNA extraction. We illustrated the effects of our model using simulated expression data, which we generated due to the absence of real RNA-Seq data from samples with accompanying morphometric-scaling information. Although a simplification of a complex problem, our model illustrates how the scaling relationships between sub-components of a heterogeneous tissue sample can result in apparent differences in expression without

changes in the regulatory control of a gene. In particular, we highlight the following conclusions:

- Scaling will *always* affect estimates of relative expression except when all components of a sample scale isometrically.
- Even where groups have common allometric scaling relationships, large differences in mean size between groups can lead to the appearance of differential expression. The effect increases with increasing deviation from isometry.
- Small differences in the allometric coefficient (β) or allometric constant (α) between groups can produce large fold-changes in gene expression. The effect is greater with increased deviation in scaling parameters between groups.
- In all cases the effect increases with tissue-bias in expression, and is most pronounced for genes expressed only in one tissue.
- Tissue scaling effects can produce both false positive and false negative detection of differential gene expression between groups.

Recommendations on how to minimise the influence of tissue scaling when inferring regulatory variation

Differences in relative expression level between groups or across species will reflect a combination of measurement error, drift, selection and variation in tissue composition. We have presented a simple model that suggests variation in tissue composition caused by non-isometric tissue scaling between groups may have strong implications for identifying genes with altered regulation. The size of the effect is dependent on the variability in tissue composition, variability in tissue size, and the properties of scaling relationships between sub-components of the sampled tissue. Although the effect size varies, any consistent effect between groups that is greater than intra-group variation could produce signatures of significant differential gene expression without any underlying regulatory variation. In real datasets the effects are likely to be more complex than presented above, as variation in tissue size will interact with scaling parameters across multiple classes of cell or tissue types.

Recent bioinformatic approaches have been developed to parse expression differences from heterogeneous samples (Gong & Szustakowski 2013; Li & Xie 2013). These approaches can be useful if the goal is to identify heterogeneity in cell type abundance across samples. However, they may have limited scope for ecological

and evolutionary studies. First, they are based on the assumption of conserved regulatory architecture within similar cell types across samples, and may therefore struggle to identify regulatory variation in constituent cells. Second, they require information about transcriptional abundance in 'pure' samples of at least one sub-tissue, and/or data on the proportions of constituent tissue types. This data is unlikely to be available for the majority of ecological studies, and if it were, it would often be a preferable source of the primary sequence data for analysis. In the absence of readily applicable bioinformatics tools we recommend the influence of tissue scaling should be considered in the design and analysis of comparative studies of gene expression. In particular we recommend the following approaches:

1) **Use fold-change thresholds:** Small but consistent effects of tissue scaling may produce significant differences in gene expression when analyzed with standard pairwise statistical tests. Introducing fold-change thresholds when identifying differentially expressed genes will go a long way to reducing the false-positive effects of tissue scaling on downstream analyses. Based on the results described above, a $\log_2$-fold change of 1, as previously used in several studies (e.g. Pointer *et al.* 2013; Harrison *et al.* 2015), would provide an adequate threshold in a range of scenarios. We would recommend higher thresholds when comparing tissues or groups/species with increasingly different phenotypic sizes or compositions. It may also be necessary to consider higher thresholds for tissue-specific genes. Of course, fold-change thresholds do not avoid false negatives, and to combat the false positive inflation it may be necessary to accept an increase in false-negative rate. However, we note that many studies of gene expression have identified genes with considerably higher fold-changes between comparisons than we suggest as a minimal threshold. This is true both for candidate genes (e.g. Palmer *et al.* 2016) and transcriptome-wide analyses (e.g. Brawand *et al.* 2011 see Figure 3). Although sometimes controversial, adopting fold-change thresholds is therefore unlikely to be prohibitive to the inference of altered regulation in sufficiently well powered and well-designed studies.

2) **Know your phenotype:** Many RNA-Seq experiments are conducted with the aim of understanding the molecular basis of divergent phenotypes, be they specific differences in the development of a trait or broad differences in individuals with

different behavioral or ecological strategies. At least a modest understanding of the phenotype in question is necessary to design informative studies of divergence in gene expression. Where possible, more precise tissue sampling will likely produce estimates of relative gene expression that more accurately reflect real variation in gene regulation. In addition to manual dissections, in 'ideal' conditions laser capture micro-dissection may provide a route to more accurate tissue sampling (Espina *et al.* 2006). In the many situations where such an approach is currently infeasible, quantifying variation in the size or composition of tissue to be analyzed may still help improve both experimental design and the interpretation of results. Estimates of scaling parameters between major tissues in the sample, either measured directly from samples for RNA-Seq, or approximated from comparable phenotypic studies, can be used to estimate the fold-change thresholds needed to minimize the effects of tissue scaling and maximize power to detect true signals of regulatory divergence. Technical difficulties in performing dissections while maintaining RNA integrity, small organism size, or simply time and expense required for additional samples, may still prevent collecting data on scaling parameters. In cases such as these, ruling out the contribution of tissue scaling is more difficult, but steps can still be taken to minimize the effect, for example by implementing more conservative fold-change thresholds.

3) **Be wary of tissue-specific genes:** Our model suggests genes with strong tissue- or cell-biased expression will be particularly prone to large changes in expression level caused by tissue scaling, and the most susceptible genes are tissue- or cell-specific. Where possible, genes identified as being differentially expressed in heterogeneous tissue samples should be examined for over-representation of tissue-specific genes in detailed expression databases, such as Flybase (Attrill *et al.* 2015) or the Mouse Atlas (Richardson *et al.* 2014). Of course, this is only possible in model species and their close relatives. It is also worth noting that tissue-biased genes may be more amenable to the action of selection, and/or may have biologically important roles in the phenotype of interest. It may therefore be reasonable to expect tissue-biased genes to be among the most differentially expressed genes in a comparative study using RNA-Seq for multiple reasons.

4) **Be wary of divergence along single principal components:** Multivariate analyses have frequently been applied to gene expression studies to show that different groups of individual samples can be distinguished based on their patterns of gene expression (e.g. Brawand *et al.* 2011; Ghalambor *et al.* 2015). Our analyses suggest this result can be produced solely by differences in tissue composition. The variance accounted for by this effect will depend on the relative balance between within group variation and the effect size of any scaling differences between groups. We expect that in many cases the scaling effects will primarily load on one single Principal Component (see Figure 6). To demonstrate that groups of samples are genuinely distinct in their transcription patterns we recommend requiring isolation across at least two dimensions in any multivariate analysis. We also note that where phenotypic data can be collected, it may be possible to include this in a multivariate analysis of gene expression to control for major differences in tissue composition between groups.

5) **Introduce phenotypic data into neutral models of gene expression:** Although we have focused on pairwise comparisons of groups, the effects of tissue scaling will also affect phylogenetic analysis of gene expression. For example, an Ornstein-Uhlenbeck (OU) model has been proposed as a potential model of expression divergence, facilitating the identification of shifts in expression that were putatively caused be selection (Brawand *et al.* 2011; Rohlfs *et al.* 2014). OU models simulate adaptive optima across a phylogeny with stabilizing selection constraining divergence around these optima (Martins 1994; Beaulieu *et al.* 2012). The presence of multiple optima is interpreted as evidence of variation in selection pressure across species. We suspect that tissue scaling could also produce a pattern of divergence across species which is similar to that predicted under an OU model. Where species in a phylogenetic dataset vary extensively by size, or differ in their scaling relationships, patterns of expression linked to tissue composition may not fit an OU model with a single optimum, giving the appearance of adaptive changes in expression level. Similar effects could be imagined under alternative comparative models which may prove useful for studying gene expression if large enough datasets can be assembled, such as incorporating heterogeneity in evolutionary rate across branches of a phylogeny (Venditti *et al.* 2011). We suggest further exploration of how the effects of tissue

scaling may affect these methods is necessary. If found to be prohibitive, one solution may be to incorporate phenotypic variation in the null model as an explicit error term, as has been done in studies of intraspecific variation (Rohlfs *et al.* 2014), or as a co-factor in the analysis.

6) **Single-cell transcriptome analysis:** Analysis of gene expression within single cells is becoming an increasingly feasible option (Sandberg 2014). Single-cell transcriptomics is free from the complicating effects of scaling between components of a heterogeneous tissue sample making the inference of regulatory change more direct. However, these analyses remain technically difficult partly because they require either cell culture or dissociation of cell aggregates from live-caught samples, and partly because they require many replicates of many cell types to uncover the full regulatory diversity of any single organ. Due to the need for increased amplification steps, single-cell analyses may also require substantial replication to overcome inaccuracy in measuring all but the highest expression ranges. The combination of technical difficulty, cell culture or disaggregation and expense from extra replication may discourage many labs from adopting single-cell analysis for evolutionary or ecological questions, particularly in non-model species. However, as with all next-generation technologies, improvements may soon remove some of these technical barriers leaving sample availability and collection as the primary limiting step.

Conclusion

Comparative analysis of gene expression provides a potentially powerful tool in the evolutionary biologist's toolkit. In an ecological or evolutionary context, most studies utilizing this tool aim to understand the relationship between variation in the regulation of gene expression and phenotypic variation. We have argued that our ability to infer this relationship can be affected by the scaling relationships between sub-tissues of the sample used to obtain RNA. In some scenarios the effect can produce the appearance large fold changes in gene expression. We have presented a simple model to explore whether, and under what scenarios, tissue scaling can produce perceptions of large expression differences without altered gene regulation. Our results suggest that under non-isometric scaling, or when comparing individuals with different scaling relationships, the effects can be moderate to severe. Based on

these analyses, we have suggested a number of experimental and analytical approaches that may go some way to minimising the effects of tissue scaling on down stream analyses of genes with divergent gene expression. The absence of datasets with both gene expression datasets and information on tissue scaling relationships has prevented a full exploration of these effects in real data. The addition these kinds of datasets, potentially derived from experimental mixing of cell cultures, would permit a useful test of our results and may potential provide further improvements on how to analyse expression data derived from heterogeneous tissues. However, we note many of the effects we describe are observable in published work and are most notable where direct comparisons can be made between whole-body and tissue-specific expression datasets. For example, Perry *et al.* (2014) showed that tissue specific sequencing of gonad transcriptomes produce greater numbers of sex-biased genes, consistent with the effects of somatic tissue diluting this signal in whole-body RNA libraries. Although we fully expect comparative studies of gene expression to continue to illuminate the gene-phenotype relationship, we caution against the naïve assumption that all differences in expression level are the result of altered gene regulation.

Acknowledgments

SHM is grateful for funding from an Early Career Research Fellowship from the Leverhulme Trust, and JEM acknowledges support from the European Research Council (grant agreements 260233 and 680951). We thank Fabian Zimmer and Peter Harrison for helpful discussions on how to formulate the model, and Natasha Bloch, Rebecca Dean, Vicencio Oostra and Alison Wright for wider discussions and comments on the manuscript. Finally, we thank the many anonymous reviewers of previous studies who have raised concerns about how tissue scaling influences estimates of gene expression for stimulating this work.

SHM and JEM conceived the project, SHM produced the model and performed the analyses, SHM and JEM wrote the manuscript.

Data Accessibility

This paper has no accompanying data.

693 # References

694 Abzhanov A, Kuo WP, Hartmann C *et al.* (2006) The calmodulin pathway and
695 evolution of elongated beak morphology in Darwin's finches. *Nature*, **442**, 563–
696 567.

697 Alonzo SH, Taborsky M, Wirtz P (2000) Male alternative reproductive behaviours in
698 a mediterranean wrasse, *Symphodus ocellatus*: Evidence from otoliths for
699 multiple life-history pathways. *Evolutionary Ecology Research*, **2**, 997–1007.

700 Attrill H, Falls K, Goodman JL *et al.* (2015) FlyBase: establishing a Gene Group
701 resource for *Drosophila melanogaster*. *Nucleic Acids Research*, **44**, gkv1046–.

702 Barton RA, Harvey PH (2000) Mosaic evolution of brain structure in mammals.
703 *Nature*, **405**, 1055–8.

704 Barton RA, Venditti C (2014) Rapid evolution of the cerebellum in humans and other
705 great apes. *Current Biology*, **24**, 2440–2444.

706 Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC (2012) Modeling stabilizing
707 selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution.
708 *Evolution*, **66**, 2369–2383.

709 Brawand D, Soumillon M, Necsulea A *et al.* (2011) The evolution of gene expression
710 levels in mammalian organs. *Nature*, **478**, 343–348.

711 Carroll SB (2008) Evo-Devo and an expanding evolutionary synthesis: A genetic
712 theory of morphological evolution. *Cell*, **134**, 25–36.

713 Chen Y-C, Harrison PW, Kotrschal A *et al.* (2015) Expression change in
714 *Angiopoietin-1* underlies change in relative brain size in fish. *Proceedings of the*
715 *Royal Society B: Biological Sciences*, **282**, 20150872.

716 Egset CK, Hansen TF, Le Rouzic a. *et al.* (2012) Artificial selection on allometry:
717 Change in elevation but not slope. *Journal of Evolutionary Biology*, **25**, 938–
718 948.

719 Emlen D (1996) Artificial selection on horn length-body size allometry in the horned
720 beetle *Onthophagus acuminatus* (Coleoptera: Scarabaeidae). *Evolution*, **50**,
721 1219–1230.

722 Enard W, Khaitovich P, Klose J *et al.* (2002) Intra- and interspecific variation in
723 primate gene expression patterns. *Science*, **296**, 340–343.

724 Espina V, Wulfkuhle JD, Calvert VS *et al.* (2006) Laser-capture microdissection.
725 *Nature Protocols*, **1**, 586–603.

726 Feldmeyer B, Elsner D, Foitzik S (2014) Gene expression patterns associated with
727 caste and reproductive status in ants: Worker-specific genes are more derived
728 than queen-specific ones. *Molecular Ecology*, **23**, 151–161.
729 Ghalambor CK, Hoke KL, Ruell EW *et al.* (2015) Non-adaptive plasticity potentiates
730 rapid adaptive evolution of gene expression in nature. *Nature*, **525**, 372–375.
731 Ghalambor C, McKay J, Carroll S, Reznick D (2007) Adaptive versus non-adaptive
732 phenotypic plasticity and the potential for contemporary adaptation in new.
733 *Functional Ecology*, **21**, 394–407.
734 Gong T, Szustakowski JD (2013) DeconRNASeq: A statistical framework for
735 deconvolution of heterogeneous tissue samples based on mRNA-Seq data.
736 Bioinformatics, 29, 1083–1085.
737 Gross MR (1996) Alternative reproductive strategies and tactics: diversity within
738 sexes. *Trends in Ecology & Evolution*, **11**, 92–98.
739 Harcourt AH, Harvey PH, Larson SG, Short R V (1981) Testis weight, body weight
740 and breeding system in primates. *Nature*, **293**, 55–57.
741 Harrison PW, Wright AE, Zimmer F *et al.* (2015) Sexual selection drives evolution
742 and rapid turnover of male gene expression. *Proceedings of the National*
743 *Academy of Sciences*, **112**, 201501339.
744 Herculano-Houzel S, Catania K, Manger PR, Kaas JH (2015) Mammalian brains are
745 made of these: A dataset of the numbers and densities of neuronal and
746 nonneuronal cells in the brain of glires, primates, scandentia, eulipotyphlans,
747 afrotherians and artiodactyls, and their relationship with body mass. *Brain,*
748 *Behavior and Evolution*, **2015**, 145–163.
749 Hollis B, Houle D, Yan Z, Kawecki TJ, Keller L (2014) Evolution under monogamy
750 feminizes gene expression in *Drosophila melanogaster*. *Nature communications*,
751 **5**, 3482.
752 Hosken DJ, Ward PI (2001) Experimental evidence for testes size evolution via sperm
753 competition. *Ecology Letters*, **4**, 10–13.
754 Immonen E, Snook RR, Ritchie MG (2014) Mating system variation drives rapid
755 evolution of the female transcriptome in *Drosophila pseudoobscura*. *Ecology*
756 *and Evolution*, **4**, 2186–2201.
757 Khaitovich P, Muetzel B, She X *et al.* (2004) Regional patterns of gene expression in
758 human and chimpanzee brains. *Genome Research*, **14**, 1462–1473.
759 Khila A, Abouheif E, Rowe L (2012) Function, developmental genetics, and fitness

consequences of a sexually antagonistic trait. *Science*, **336**, 585–589.

Kotrschal A, Rogell B, Bundsen A *et al.* (2013) Artificial selection on relative brain size in the guppy reveals costs and benefits of evolving a larger brain. *Current Biology*, **23**, 168–171.

Kvist J, Wheat CW, Kallioniemi E *et al.* (2013) Temperature treatments during larval development reveal extensive heritable and plastic variation in gene expression and life history traits. *Molecular Ecology*, **22**, 602–619.

Li Y, Xie X (2013) A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. BMC bioinformatics, 14 Suppl 5, S11.

Martins EP (1994) Estimating the Rate of Phenotypic Evolution from Comparative Data. *American Naturalist*, **144**, 193–209.

Moczek AP, Rose DJ (2009) Differential recruitment of limb patterning genes during development and diversification of beetle horns. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8992–8997.

Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Palmer WJ, Duarte A, Schrader M *et al.* (2016) A gene associated with social immunity in the burying beetle N*icrophorus vespilloides*. *Proceedings of the Royal Society B: Biological Sciences*, **238** (1823), 0–7.

Parker DJ, Cunningham CB, Walling CA *et al.* (2015) Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. *Nature Communications*, **6**, 8449.

Perry JC, Harrison PW, Mank JE (2014) The ontogeny and evolution of sex-biased gene expression in drosophila melanogaster. Molecular Biology and Evolution, 31, 1206–1219.

Pointer MA, Harrison PW, Wright AE, Mank JE (2013) Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genetics*, **9**.

Richardson L, Venkataraman S, Stevenson P *et al.* (2014) EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Research*, **42**, 1–10.

Rohlfs R V., Harrigan P, Nielsen R (2014) Modeling gene expression evolution with an extended ornstein-uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution*, **31**, 201–211.

794 Rosenberg MS (2002) Fiddler crab claw shape variation: a geometric morphometric
795 analysis across the genus *Uca* (Crustacea: Brachyura: Ocypodidae). *Biological*
796 *Journal of the Linnean Society*, **75**, 147–162.
797 Sandberg R (2014) Entering the era of single-cell transcriptomics in biology and
798 medicine. *Nature Methods*, **11**, 22–24.
799 Shapiro MD, Marks ME, Peichel CL *et al.* (2004) Genetic and developmental basis of
800 evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**, 717–723.
801 Simmons LW, Tomkins JL (1996) Sexual selection and the allometry of earwig
802 forceps. *Evolutionary Ecology*, **10**, 97–104.
803 Sinervo B, Lively CM (1996) The rock–paper–scissors game and the evolution of
804 alternative male strategies. *Nature*, **380**, 240–243.
805 Standage DS, Berens AJ, Glastad KM *et al.* (2016) Genome, transcriptome, and
806 methylome sequencing of a primitively eusocial wasp reveal a greatly reduced
807 DNA methylation system in a social insect. *Molecular Ecology*, **25**, 1769-1784.
808 Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice
809 produced by two interacting pigmentation genes. *PLoS Biology*, **5**, 1880–1889.
810 Stern DL, Orgogozo V (2008) The loci of evolution: How predictable is genetic
811 evolution? *Evolution*, **62**, 2155–2177.
812 Stiver KA, Harris RM, Townsend JP, Hofmann HA, Alonzo SH (2015) Neural gene
813 expression profiles and androgen levels underlie alternative reproductive tactics
814 in the ocellated wrasse, *Symphodus ocellatus*. *Ethology*, **121**, 152–167.
815 Stuglik MT, Babik W, Prokop Z, Radwan J (2014) Alternative reproductive tactics
816 and sex-biased gene expression: The study of the bulb mite transcriptome.
817 *Ecology and Evolution*, **4**, 623–632.
818 Tomkins JL, Simmons LW (2002) Measuring relative investment: a case study of
819 testes investment in species with alternative male reproductive tactics. *Animal*
820 *Behaviour*, **63**, 1009–1016.
821 Toth AL, Smith CR, Suarez A V., Robinson GE (2008) Genetic and genomic analyses
822 of the division of labour in insect societies. *Nature Reviews Genetics*, **9**, 735–
823 748.
824 Venditti C, Meade A, Pagel M (2011) Multiple routes to mammalian diversity.
825 *Nature*, **479**, 393–6.
826 Voje KL (2016) Scaling of morphological characters across trait type, sex, and
827 environment: A meta-analysis of static allometries. *The American Naturalist*,

187, 89–98.

Wilkinson GS (1993) Artificial sexual selection alters allometry in the stalk-eyed fly *Cyrtodiopsis dalmanni* (Diptera: Diopsidae). *Genetical Research*, **62**, 213–222.

Tables

Table 1. Results of the simulated data sets based on scaling parameters between male morphs of multiple insects

	Pearson correlation		log_2-fold change (N)			differentially expressed[1]	
Species	**r**	**p**	**mean**	**minimum**	**maximum**	**p < 0.05**	**p < 0.001**
Onthophagus binodis	0.995	<0.001	-0.233	-2.239	0.007	121 (111)	106 (103)
Onthophagus taurus	0.999	<0.001	0.003	-0.001	0.026	1 (0)	0 (0)
Forficula auriculaira	0.998	<0.001	-0.104	-0.866	0.014	166 (126)	119 (96)
Amegilla dawsoni	0.999	<0.001	-0.05	-0.473	0.002	107 (101)	79 (31)

[1] numbers in parentheses are after Bonferoni correction for multiple tests.

Figure legends

Figure 1. **Types of scaling relationships and how they shape proportional size.** Here we show a hypothetical comparison between two groups of individuals which may differ in size and which are comprised of two tissues. In each scenario, row 1 shows the relationship between tissue A and total size for individuals from two groups (red and blue). The scaling relationships are determined by the allometric equation $y = \alpha x^{\beta}$, where β is the scaling coefficient and α is the scaling constant. Row 2 shows illustrative examples of individuals from each group imagining tissue A as gonad size. Note, this is only an example and components tissues can be any aspect of morphology. Row 3 shows an illustration of how the proportion of tissue A (coloured) varies between groups as a result of the scaling relationship and differences in mean size. Row 4 shows the effects these proportional differences might have on relative gene expression, illustrated with box whisker plots.

Figure 2. **Effects of size differences under conserved allometric scaling.** A) Effects of comparing two groups with different total sizes under alternative scaling coefficients, β. The $\log_2$-fold change is plotted against the ratio of the total size of two groups. In this comparison $x = 10$ in group one and varied x in group two between 0.1 and 100. Effects of comparing two groups with different levels of tissue-biased expression B) under hyper-allometry ($\beta = 2$) and C) under hypo-allometry ($\beta = 0.1$). In B and C coloured lines indicate comparisons where expression of gene *a* is set to 5,000 in component *y* and it's expression in component *x* is varied as indicated in the colour key. The black dashed line indicates a comparison where expression of gene *a* is set to 0 in component *y* and 5,000 in component *x*. Dashed grey lines indicated a FC of ±1, often used as a threshold of significant difference in expression.

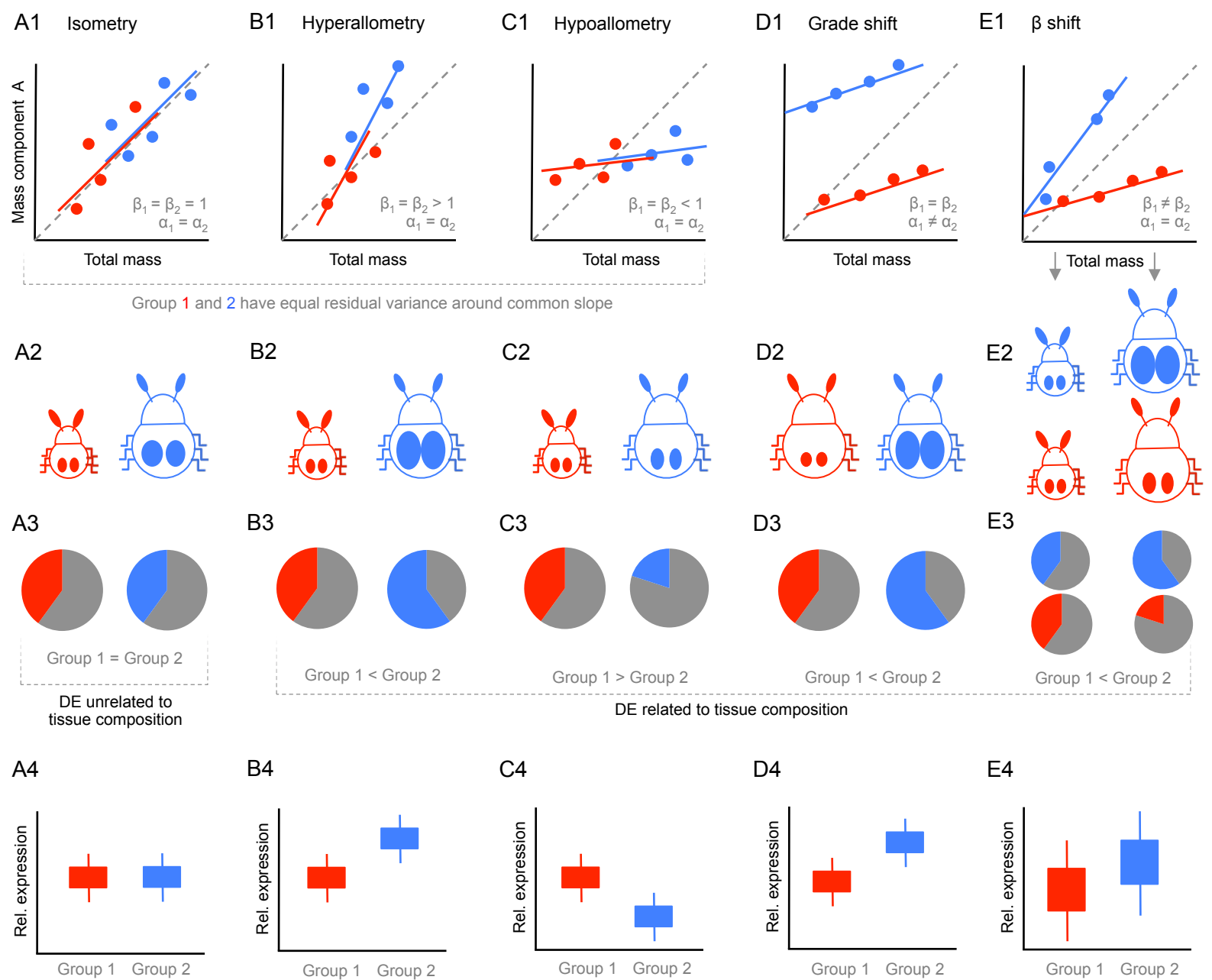
Figure 3. **Effects of varying the allometric constant between groups.** A) Effects of comparing two groups with different scaling constants, α, across different shared scaling coefficients (β), with α in group one set to 0.1 and varying α in group two between 0.1 and 10. The effects of comparing two groups with different α across different levels of tissue-biased expression B) under hyper-allometry ($\beta = 1.5$) and C) under hypo-allometry ($\beta = 0.5$). In B and C coloured lines indicate comparisons where

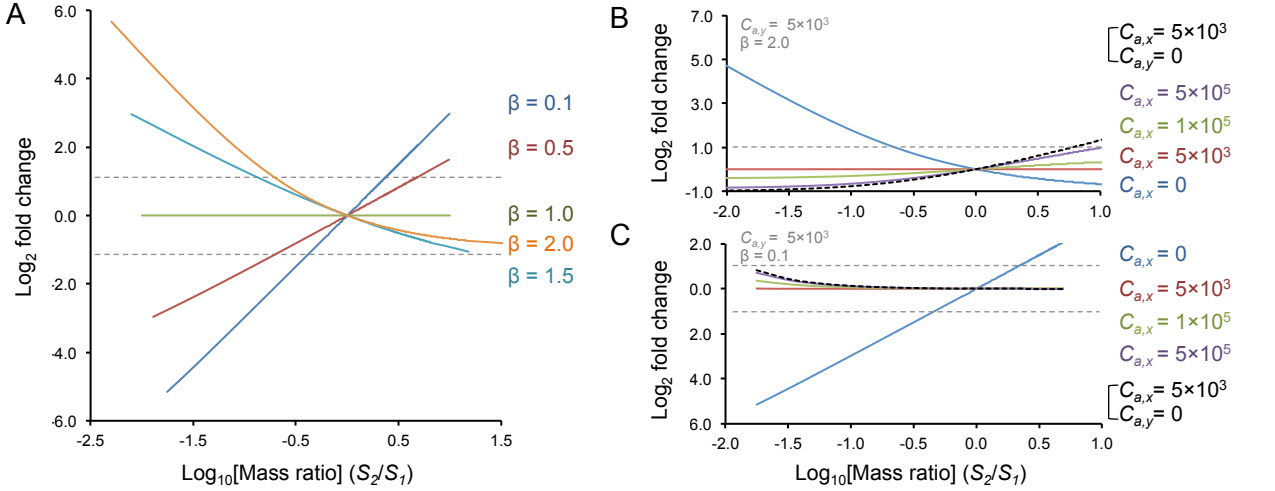
893 expression of gene *a* is set to 5,000 in component *y* and it's expression in component
894 *x* is varied as indicated in the colour key. The black dashed line indicates a
895 comparison where expression of gene *a* is set to 0 in component *y* and 5,000 in
896 component *x*. The $\log_2$-fold change is plotted against the ratio of the α of each group.
897 Dashed grey lines indicated a FC of ±1.
898
899 Figure 4. **Effects of varying the allometric coefficient between groups.** Effects of
900 comparing two groups with different scaling coefficients, *β*, across different units of
901 size (*x*) with A) *β* in group one set to 0.5 and varying *β* in group two between 0.1 and
902 3, and B) *β* in group one set to 1.5 and varying *β* in group two between 0.1 and 3.
903 Effects of comparing two groups with different levels of tissue-biased expression with
904 C) *β* in group one set to 0.5 and varying *β* in group two between 0.1 and 3 and D) *β* in
905 group one set to 1.5 and varying *β* in group two between 0.1 and 3. In C and D
906 coloured lines indicate comparisons where expression of gene *a* is set to 5,000 in
907 component *y* and it's expression in component *x* is varied as indicated in the colour
908 key. The black dashed line indicates a comparison where expression of gene *a* is set
909 to 0 in component *y* and 5,000 in component *x*. The $\log_2$-fold change is plotted against
910 the ratio of the *β* of each group. Dashed grey lines indicated a FC of ±1, often used as
911 a threshold of significant difference in expression.
912
913 Figure 5. **Tissue scaling effects can mask true positives.** A) Effects of non-isometric
914 but conserved scaling on the detection of a differentially expressed gene. Two groups
915 were modelled with conserved scaling constant, *α* (0.1), and scaling coefficient, *β*
916 (1.5), values but different total sizes. The estimated $\log_2$-fold change is plotted against
917 the mass ratio, setting *x* in group one to be 10, and varying *x* in group two between 0.1
918 and 100. B) Effects of 'grade-shifts', or group differences in α, on the detection of a
919 differentially expressed gene. Two groups were modelled with conserved sizes (*x* =
920 10) and *β* (1.5) values but different α values. The estimated $\log_2$-fold change is plotted
921 against the mass ratio, setting α in group one to be 0.1, and varying *x* in group two
922 between 0.1 and 10. C) Effects of group differences in *β* on the detection of a
923 differentially expressed gene. Two groups were modelled with conserved sizes (*x* =
924 10) and α (0.1) values but different α values. The estimated $\log_2$-fold change is plotted
925 against the mass ratio, setting *β* in group one to be 1.5, and varying *β* in group two
926 between 0.1 and 3. In each case expression of gene *a* in subcomponent *y* is 5,000. In

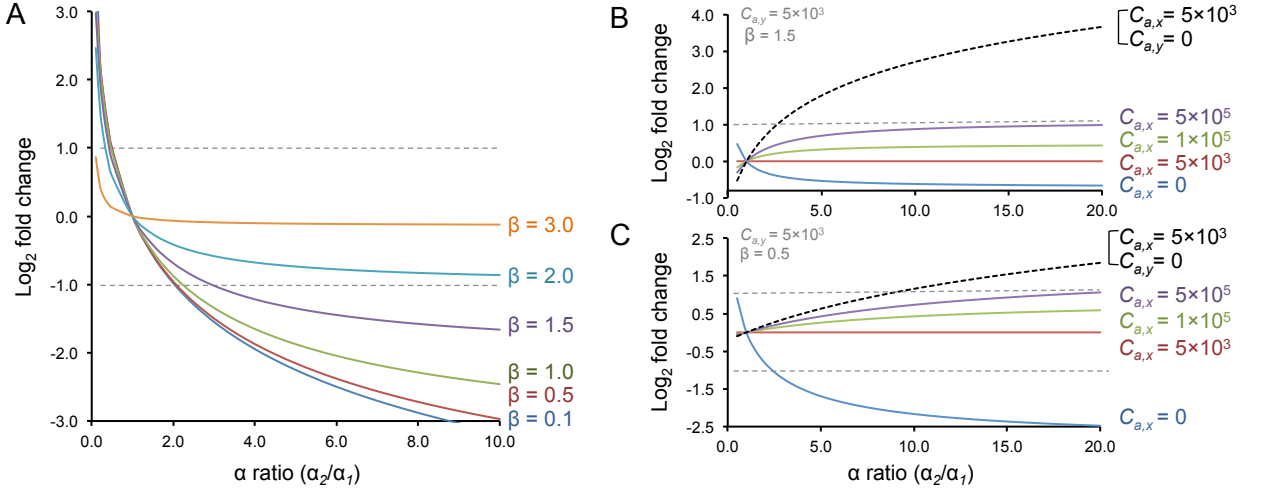
group one expression of *a* in *x* is 5,000 but expression of *a* in *x* varies in group two taking values of either 20,000, 10,000, 5,000, 2,500 or 1,250 (representing $\log_2$-fold change values of 2, 1, 0, -1 and -2 respectively).

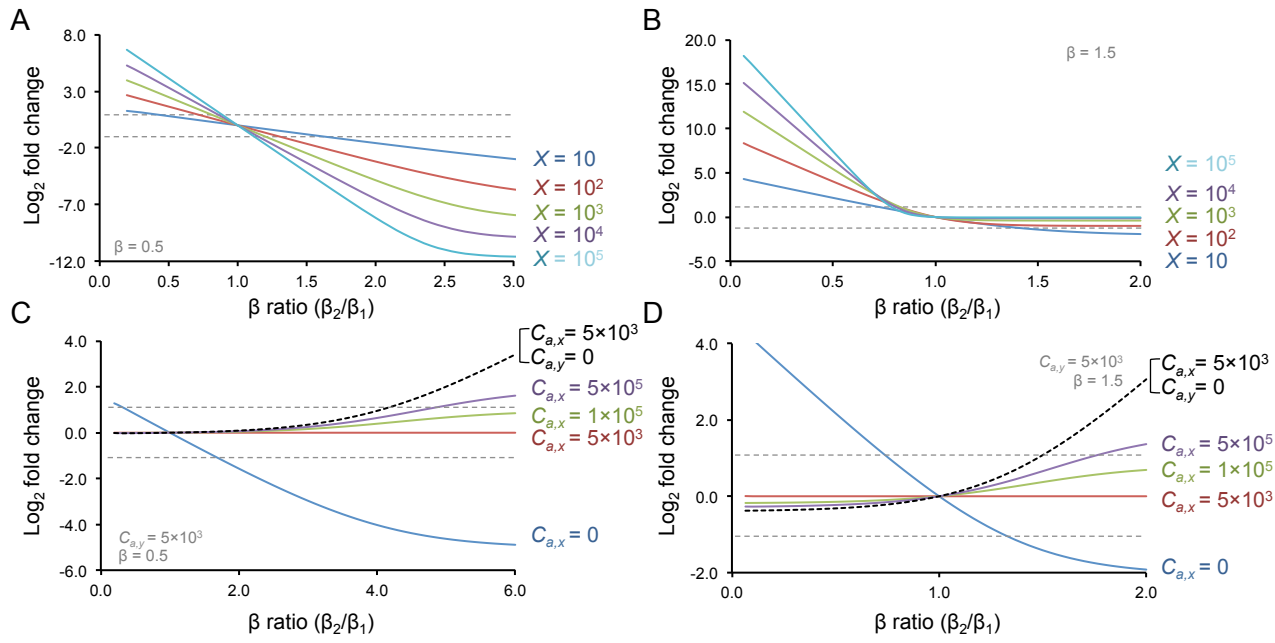
Figure 6. **Predicted differences in relative expression level between male morphs of multiple species of insect based on testis~soma scaling**. A) Predicted fold-change in expression across different levels of tissue-biased expression ($C_{a,x}$ = gonad expression, $C_{a,y}$ = soma expression). B-E) Results of Principal Component Analyses (B1-E1) and hierarchical clustering (B2-E2) using simulated datasets from the model paramterised using testis~soma scaling relationships for *O. taurus* (B), *A. dawsoni* (C), *F. auricularia* (D) and *O. binodis* (E). In the PCAs, we plot the PC significantly associated with morph type (indicated by *) against PC1. Colours indicate different categories of male morph.

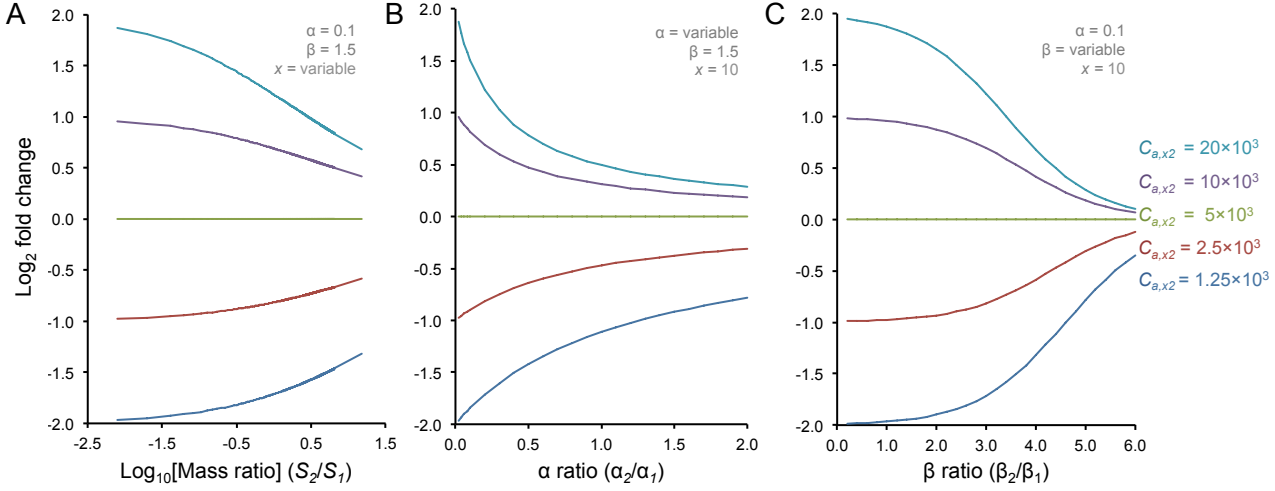
Figure 7. **Predicted differences in relative expression level between or within primates and glires based on scaling relationships between neuron number and non-neuronal cell number in the cerebral cortex and cerebellum.** A-B) Predicted fold-change between two groups representing the smallest and largest individuals within primates (blue) and glires (red) assuming conserved, order-specific scaling relationships and varying levels of tissue-biased expression. A) Results for cerebral cortex and B) results for cerebellum. C-D) Predicted differences in gene expression between two group of individuals, one with glire-scaling relationships and one with primate-scaling relationships, but which have an equal, constant size. Results show the predicted fold-change across different levels of tissue bias for C) the cerebral cortex, and D) the cerebellum.

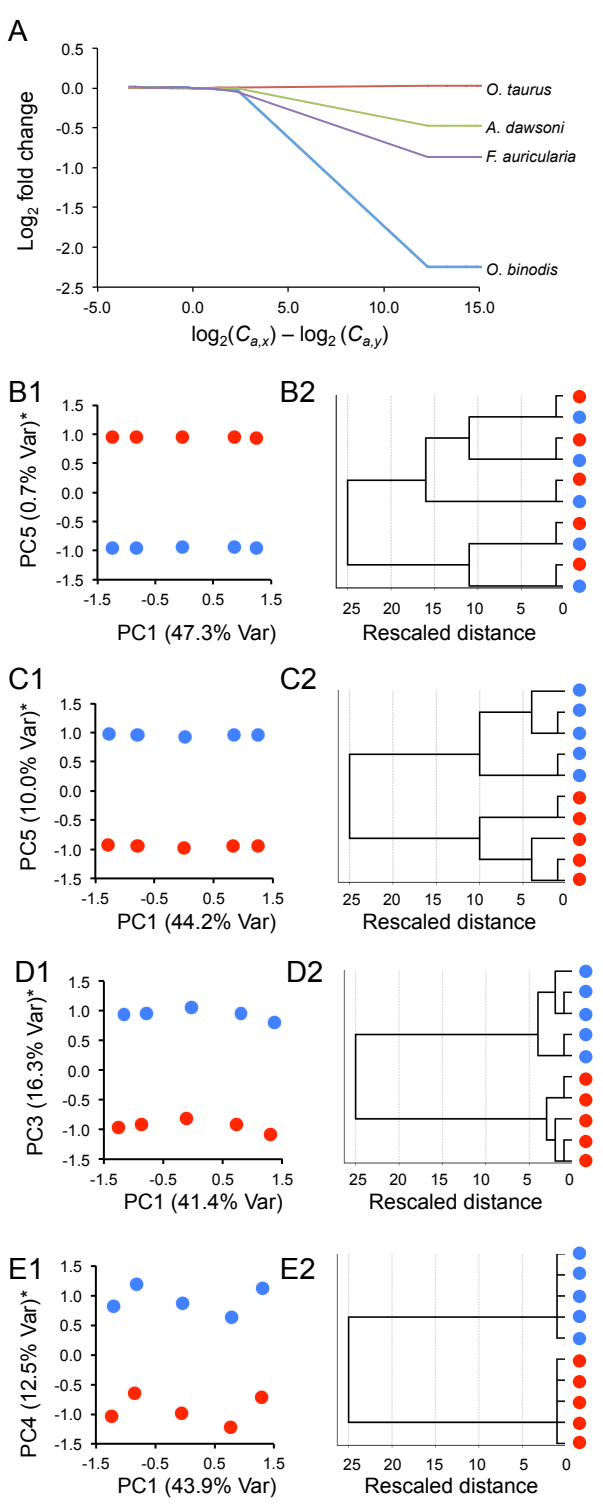

A1
Isometry
B1
Hyperallometry
C1
Hypoallometry
D1
Grade shift
E1
β shift
Mass component A
β1 = β2 = 1
α1 = α2
β1 = β2 > 1
α1 = α2
β1 = β2 < 1
α1 = α2
β1 = β2
α1 ≠ α2
β1 ≠ β2
α1 = α2
Total mass
Total mass
Total mass
Total mass
Total mass
Group 1 and 2 have equal residual variance around common slope
A2
B2
C2
D2
E2
A3
B3
C3
D3
E3
Group 1 = Group 2
Group 1 < Group 2
Group 1 > Group 2
Group 1 < Group 2
Group 1 < Group 2
DE unrelated to tissue composition
DE related to tissue composition
A4
B4
C4
D4
E4
Rel. expression
Rel. expression
Rel. expression
Rel. expression
Rel. expression
Group 1
Group 2
Group 1
Group 2
Group 1
Group 2
Group 1
Group 2
Group 1
Group 2

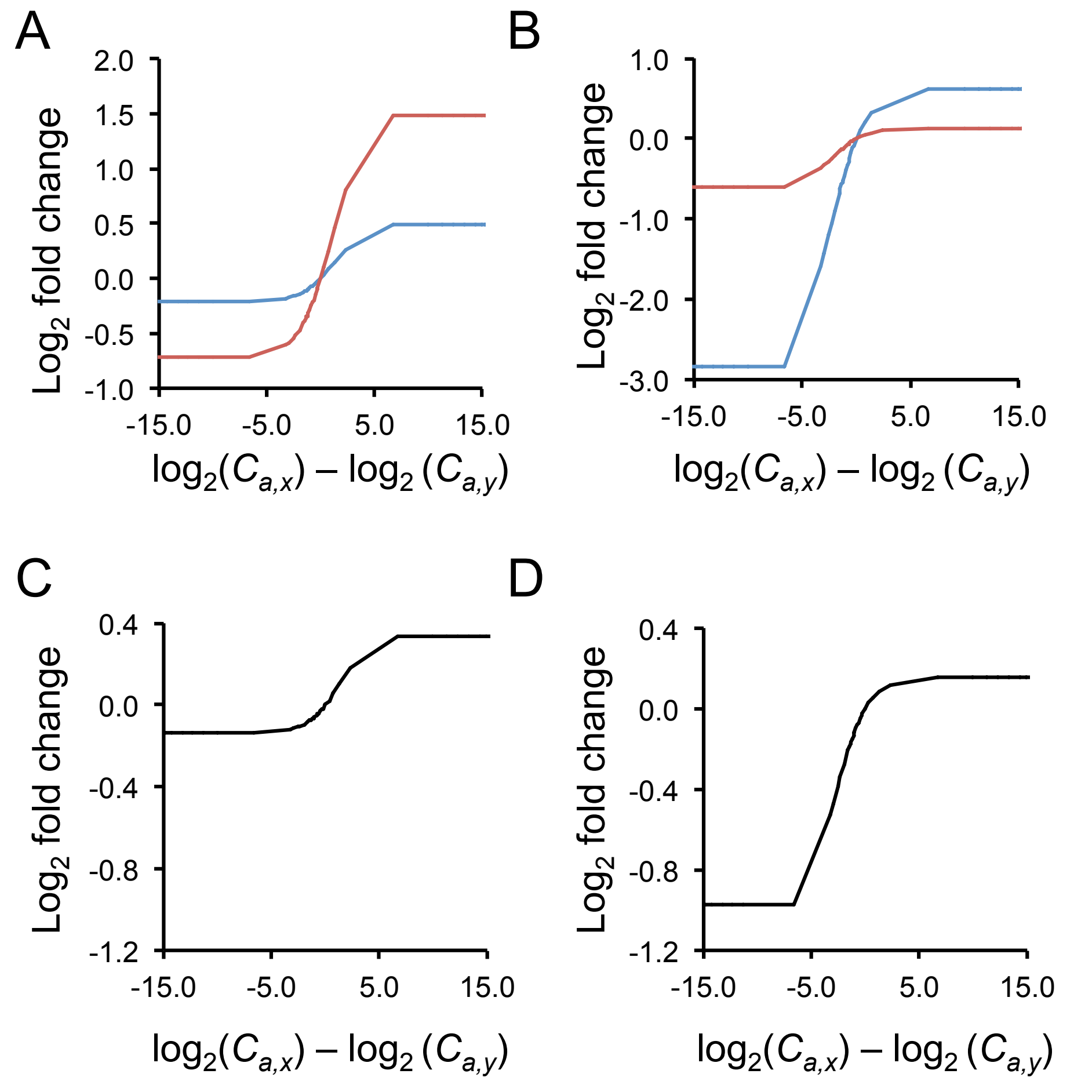

A
Log2 fold change
Log10[Mass ratio] (S2/S1)
β = 0.1
β = 0.5
β = 1.0
β = 2.0
β = 1.5
B
Ca,y = 5×10^3
β = 2.0
Ca,x = 5×10^3
Ca,y = 0
Ca,x = 5×10^5
Ca,x = 1×10^5
Ca,x = 5×10^3
Ca,x = 0
C
Ca,y = 5×10^3
β = 0.1
Ca,x = 0
Ca,x = 5×10^3
Ca,x = 1×10^5
Ca,x = 5×10^5
Ca,x = 5×10^3
Ca,y = 0


A
Log2 fold change
α ratio (α2/α1)
β = 3.0
β = 2.0
β = 1.5
β = 1.0
β = 0.5
β = 0.1
B
Ca,y = 5×10^3
β = 1.5
Ca,x = 5×10^3
Ca,y = 0
Ca,x = 5×10^5
Ca,x = 1×10^5
Ca,x = 5×10^3
Ca,x = 0
C
Ca,y = 5×10^3
β = 0.5
Ca,x = 5×10^3
Ca,y = 0
Ca,x = 5×10^5
Ca,x = 1×10^5
Ca,x = 5×10^3
Ca,x = 0


A
Log2 fold change
β ratio (β2/β1)
β = 0.5
X = 10
X = 10^2
X = 10^3
X = 10^4
X = 10^5
B
β = 1.5
C
Ca,x = 5×10^3
Ca,y = 0
Ca,x = 5×10^5
Ca,x = 1×10^5
Ca,x = 5×10^3
Ca,x = 0
Ca,y = 5×10^3
β = 0.5
D
Ca,y = 5×10^3
β = 1.5


A
α = 0.1
β = 1.5
x = variable
B
α = variable
β = 1.5
x = 10
C
α = 0.1
β = variable
x = 10
Log2 fold change
Log10[Mass ratio] (S2/S1)
α ratio (α2/α1)
β ratio (β2/β1)
Ca,x2 = 20×10^3
Ca,x2 = 10×10^3
Ca,x2 = 5×10^3
Ca,x2 = 2.5×10^3
Ca,x2 = 1.25×10^3


A
Log2 fold change
O. taurus
A. dawsoni
F. auricularia
O. binodis
log2(Ca,x) – log2 (Ca,y)
B1
PC5 (0.7% Var)*
PC1 (47.3% Var)
B2
Rescaled distance
C1
PC5 (10.0% Var)*
PC1 (44.2% Var)
C2
Rescaled distance
D1
PC3 (16.3% Var)*
PC1 (41.4% Var)
D2
Rescaled distance
E1
PC4 (12.5% Var)*
PC1 (43.9% Var)
E2
Rescaled distance


A
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
Log2 fold change
-15.0
-5.0
5.0
15.0
log2(Ca,x) – log2 (Ca,y)
B
1.0
0.0
-1.0
-2.0
-3.0
Log2 fold change
-15.0
-5.0
5.0
15.0
log2(Ca,x) – log2 (Ca,y)
C
0.4
0.0
-0.4
-0.8
-1.2
Log2 fold change
-15.0
-5.0
5.0
15.0
log2(Ca,x) – log2 (Ca,y)
D
0.4
0.0
-0.4
-0.8
-1.2
Log2 fold change
-15.0
-5.0
5.0
15.0
log2(Ca,x) – log2 (Ca,y)