Pervasive concerted evolution in gene expression shapes cell type transcriptomes

Authors: Cong Liang[1,3,7,†], Jacob M. Musser[1,2,9†], Alison Cloutier[8], Richard Prum[2,6] and Günter P. Wagner[1,2,4,5,*]

Affiliations

[1]Yale Systems Biology Institute, 850 West Campus Drive, West Haven, CT-06516.

[2]Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT-06520.

[3]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT-06511.

[4]Department of Obstetrics, Gynecology and Reproductive Sciences, Yale Medical School, 333 Cedar Street, New Haven, CT-06520-8063.

[5]Department of Obstetrics and Gynecology, Wayne State University, 540 E Canfield St, Detroit, MI 48201.

[6]Yale Peabody Museum of Natural History, 170 Whitney Avenue, New Haven, CT-06511.

[7] Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT-06511.

[8]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto ON, Canada M5S 3B2.

[9]Developmental Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

*Correspondence to: gunter.wagner@yale.edu

†These authors contributed equally to this work.

Abstract

Transcriptomic data yields valuable insights into cell type, tissue, and organ evolution. However, interpreting this data requires understanding how transcriptomes evolve. A particularly difficult problem is that cell type transcriptomes may not evolve independently. Here we present a statistical model to estimate the level of concerted transcriptome evolution and apply it to published and new data. The results indicate that tissues undergo pervasive concerted evolution in gene expression. Tissues related by morphology or developmental lineage exhibit higher levels of concerted evolution. Concerted evolution also causes tissues from the same species to be more similar in gene expression to each other than to homologous tissues in another species. This result may explain why some tissue transcriptomes cluster by species rather than homology. Our study illustrates the importance of accounting for concerted evolution when interpreting comparative transcriptome data, and should serve as a foundation for future investigations of cell type evolution.

Introduction

Many organisms, like mammals, share a similar body plan composed of homologous cell and tissue types. Different tissues are built during development by distinct gene expression programs, and homologous tissues bear resemblance across species owing to inheritance of the same gene expression program. This concept of homology has formed a central tenet of two decades of comparative development and molecular research. With the advent of RNAseq, it has fueled comprehensive efforts to investigate and classify homologous cell and tissue types within and across species (Tschopp et al., 2014, Wang et al., 2011, Achim et al., 2015), and identify gene expression differences that underlie phenotypic evolution (Pankey et al., 2014).

However, a key finding of RNAseq studies is that most genes are not uniquely expressed in a single cell or tissue type. For instance, different neuronal cell types express synaptic genes (Sieburth et al., 2005, Ruvinsky et al., 2007, Stefanakis et al., 2015), different muscle types share expression of contractile genes (Steinmetz et al., 2012, Brunet et al., 2016), and all tissues employ 'housekeeping genes' (Eisenberg and Levanon, 2013). A recent study by Stefanakis, Carrera, and Hobert (Stefanakis et al., 2015) showed that synaptic gene expression in different neuronal cell types is controlled in part by the same regulatory elements in the genome. Partial deletion of these regulatory elements simultaneously affects expression across multiple cell types. Thus, it is likely that evolutionary changes occurring in these shared regulatory elements will alter gene expression in multiple cell types, a form of pleiotropy called concerted gene expression evolution (Musser and Wagner, 2015). This elementary biological inference suggests that cell and tissue transcriptomes do not evolve independently, an assumption underlying many statistical methods for the comparative analysis of quantitative data (Felsenstein, 2004).

Concerted evolution poses difficult problems for the analysis of comparative data. Most seriously, it can yield misleading inferences drawn from comparative analyses, such as hierarchical clustering. Concerted evolution can lead, paradoxically, to tissues within a species being more similar in gene expression profile to each other than to a homologous tissue in another species (Fig. 1). This results from concerted evolution in gene expression that occurs after speciation, resulting in species-specific tissue similarities that unite tissues within a species. Thus, concerted gene expression evolution can confound efforts to classify and compare cell types across species using RNAseq. On the other hand, estimates of concerted evolution may

provide novel insight into whether, and how much, different tissues share regulatory information in the genome.

In this study, we develop a statistical model to measure the degree of concerted evolution among different cell and tissue types. Applying this method to published data we find that concerted evolution is pervasive, and varies in accordance with similarity in morphology and development. These findings are consistent with a model in which closely related tissues share more regulatory information than more distantly related cells or tissues. Our results also help explain why some studies have found tissue transcriptomes cluster by species, rather than by homologous tissue type, while others found the opposite pattern of similarity. According to our model, the outcome depends on the rate of concerted evolution relative to the phylogenetic distance of the species compared. High rates of concerted evolution lead to clustering of tissue expression profiles by species, whereas low levels of concerted evolution result in clustering by homologous tissue type. Finally, we show that the degree of concerted evolution decreases with the evolutionary age of a tissue, suggesting that the evolutionary independence, or transcriptomic individuation, of tissues increases over time. Our model thus provides a quantitative tool for studying cell type evolution and individuation.

Results and Discussion

Mathematical model of concerted evolution

To model gene expression evolution, we consider the simplest scenario of two cell types, A and B, evolving in two species, 1 and 2 (Fig. 1B). In this case, the two cell types originated prior to the species' last common ancestor, which we refer to as species 0. We characterize the expression levels in tissue type A by random variables $A_{j,i}$, where the first subscript j represent species index, and the second subscript, i, represent the gene index. To model gene expression evolution we assume a Brownian motion model, where σ^2 represents the variance that is expected to accumulate per unit time. The evolutionary paths of $A_{1,i}$ and $A_{2,i}$ are random walks from their initial state $A_{0,i}$. Taking Δt as the time period from their last common ancestor to present, the probability density of expression change $A_{1,i} - A_{0,i}$ follows a normal distribution with variance $\sigma^2 \Delta t$ (Felsenstein, 2004). As in general Brownian evolution models, Brownian motions of the same trait in two species are uncorrelated.

We represent concerted evolution by introducing a parameter ρ, related to the covariance of gene expression change in two cell types A and B of the same species, $cov(A_{1,i} - A_{0,i}, B_{1,i} - B_{0,i}) = \rho^2 \Delta t$. Analogous to the measure of the rate of Brownian motion σ^2, ρ^2 measures the strength of covariance in evolutionary change. If ρ is larger than zero ($\rho^2 > 0$), cell types A and B will undergo concerted gene expression evolution. On the other hand, if $\rho = 0$ gene expression changes are uncorrelated in tissue A and B and no concerted evolution is happening. Under the Brownian model, the expression levels of gene i in the two descendent species will follow multivariate normal distribution. $(A_{1,i}, A_{2,i}, B_{1,i}, B_{2,i}) \sim \mathcal{N}(E, V)$. The covariance matrix $\mathbf{V}$ is symmetric, and its upper triangular part is shown in the following table (Supplemental method).

$$\mathbf{V} = \begin{pmatrix} \sigma^2 t + var(A_{0,i}) & var(A_{0,i}) & \rho^2 t + cov(A_{0,i}, B_{0,i}) & cov(A_{0,i}, B_{0,i}) \\ & \sigma^2 t + var(A_{0,i}) & cov(A_{0,i}, B_{0,i}) & \rho^2 t + cov(A_{0,i}, B_{0,i}) \\ & & \sigma^2 t + var(B_{0,i}) & var(B_{0,i}) \\ & & & \sigma^2 t + var(B_{0,i}) \end{pmatrix}$$

Without losing much generality, we have made some assumptions for the simplicity of the model in the above matrix: 1) the rate of gene expression evolution, σ, is the same across tissue types, genes and species. 2) the rate of concerted evolution, ρ, is the same in the two descendant species lineages.

To estimate the level of concerted transcriptome evolution, we measure the correlation between expression changes, $\gamma = corr(A_{1,i} - A_{0,i},\ B_{1,i} - B_{0,i})$. This parameter is related to ρ^2 under Brownian model, $\gamma = \frac{\rho^2}{\sigma^2}$.

$$corr(A_{1,i} - A_{2,i},\ B_{1,i} - B_{2,i}) = corr(A_{1,i} - A_{0,i},\ B_{1,i} - B_{0,i}) = \frac{\rho^2}{\sigma^2} = \gamma \qquad (1).$$

As a result, γ is the relative rate of concerted change compared to the overall gene expression change and can be estimated as the empirical correlation between the contrasts of expression vectors for cell type A and B in two species. The approach can be extended to a species tree with more than two species by calculating the independent gene expression contrast vectors on every internal node (Felsenstein, 2004). We show that equation (1) holds for descendent species with different evolutionary divergence times (supplemental method).

Concerted evolution is pervasive across diverse tissues

We first applied our model to three previously published comparative transcriptome datasets: (a) Merkin et al. (2012), containing nine mature tissues from four mammals and chicken, (b) Brawand et al. (2011), containing six mature tissues in ten diverse species of mammal and chicken, and (c) Tschopp et al. (2014), reporting data on limb, genital- and tail buds at early developmental stages in mouse and *Anolis*. These datasets have three advantages. First, all samples for each data set were generated in the same lab and with the same protocol, reducing batch effects. Second, these datasets include a diverse set of both mature and developing tissues, enabling us to determine the extent to which concerted evolution varies depending on tissue relationships. Third, the Brawand and Merkin datasets sampled five tissues in common, allowing us to determine whether estimates of concerted evolution are consistent across datasets and species.

Our results show that concerted evolution is pervasive, and varies depending on the tissues under comparison (Fig. 2; Table S1 and S2). The highest estimates of concerted evolution were from embryonic appendage buds (Fig. 2; points in red), where γ ranged from .81 (early hindlimb and tail buds) to .89 (early hindlimb and genital buds) supporting the conclusion of the authors that genital and limb buds are developmentally highly related. Concerted evolution estimates among the differentiated tissues of the Brawand and Merkin datasets were lower. For each pair of tissues in the Brawand and Merkin datasts we calculated the mean value ($\bar{\gamma}$) of estimates from independent contrasts with split time > 50 million years (see methods, Table S1 and S2). Mean estimates ranged from 0.17 (liver and testes) to 0.71 (forebrain and cerebellum), indicating substantial variation in concerted evolution even among differentiated tissues (Fig. 2). Clearly, forebrain and cerebellum are more similar to each other in terms of cell type composition and gene regulation than liver and testis, supporting the notion that our estimates of concerted evolution reflect genetic relatedness of these tissues. Most tissues exhibited at least some degree of concerted evolution, with ~71% of individual γ estimates significantly greater than expected by chance under permutation ($p < 0.05$). Notable exceptions were tissue comparisons with testes, which dominated the lower end of concerted evolution estimates, and could not be statistically distinguished from a model in which gene expression evolved completely independently between tissues.

In general, concerted evolution was higher between tissues with similar morphology or function. Our highest estimates were among early developing limb, genital, and tail bud tissue.

At this early stage of development, these anlagen largely consist of undifferentiated cells. We also found evidence that concerted evolution in differentiated tissues reflects similarity in morphology and function. For instance, two of the highest measures of concerted evolution between differentiated tissues were for cerebral cortex and cerebellum ($\bar{\gamma}$=.71), and heart and skeletal muscle ($\bar{\gamma}$=.47). Both tissue pairs express common sets of pan-neuronal and contractile genes, respectively. We also found evidence that developmental origin may influence degree of concerted evolution in differentiated tissue. For instance, lung and colon ($\bar{\gamma}$=.50), both endodermal derivatives, were highly correlated relative to other differentiated tissues despite their very distinct functions. More surprising was our finding that spleen (Fig. 2; points in green) also had relatively high estimates of concerted evolution with both lung ($\bar{\gamma}$=.59) and colon ($\bar{\gamma}$=.52). Historically there has been disagreement over whether the mammalian spleen develops from mesoderm or endoderm, but recent evidence has largely favored a mesodermal origin (Brendolan et al., 2007). However, we found mature spleen tissue undergoes more concerted evolution with endodermal derivatives than with tissue derived from the mesoderm, including kidney ($\bar{\gamma}$=.44) and heart ($\bar{\gamma}$=.35). Whether this reflects a common developmental origin, or reflects convergent similarity that occurs after differentiation, is unclear.

We also explored the extent to which our estimates of the relative rate of concerted evolution, γ, depends on the species and dataset used. We found consistent estimates of γ regardless of t divergence time (Hedges et al., 2006) (Fig. 3, ANOVA p-values for dependence of the estimated γ-value on divergence time are all > 0.1). However, estimates of γ derived from recently diverged species pairs (< 50 million years) tended to be lower, contributing to relatively large standard deviation for some tissue pairs (Table S2 and S3). One possible explanation of this finding is that recently diverged species have not accumulated enough gene expression changes to allow for accurate estimation of concerted evolution. Thus, comparisons from recently diverged species may be dominated by observational noise, resulting in underestimates of γ. Finally, we found that estimates of γwere consistent across datasets. The average $\bar{\gamma}$ values, calculated for Merkin and Brawand datasets separately (grey horizontal lines in Fig. 2A; solid line for Brawand dataset, dashed line for Merkin dataset), do not statistically differ (Welch's t-test p-values > .1), indicating that estimates of the relative rate of concerted evolution are robust to variations in data acquisition.

These findings suggest that the rate of concerted evolution reflect the biological nature of the tissues under comparison, rather than artifacts of genome quality, batch effects, or species divergence time. Notably, different tissue types from the same set of species (e.g. human and mouse) can have very different degrees of concerted evolution (Fig. 2). This implies that concerted evolution does not solely originate from gene expression changes affecting globally shared "housekeeping" genes, for instance metabolic genes. Instead, the concerted evolution signal appears heavily influenced by the nature of the tissues and cells compared, suggesting that the structure of the gene regulatory networks active in these cells affects the strength of concerted evolution. For instance, our finding of relatively high concerted evolution between fore brain and cerebellum is consistent with findings in other animals that common transcription factors regulate pan-neuronal gene expression across different neuronal cell types (Ruvinsky et al., 2007, Stefanakis et al., 2015). With regards to testes, previous studies have shown testes gene expression has a high degree of tissue-specificity (Yue et al., 2014, Lin et al., 2014), is highly divergent among mammals relative to other tissues (Brawand et al., 2011), and is under positive selection (Khaitovich et al., 2005, Khaitovich et al., 2006). Our results suggest a key factor in explaining testes distinct evolutionary history is its ability to evolve gene expression independently. Thus, even though testes share a genome in common with other tissues, it is capable of evolving gene expression with near complete independence.

Concerted evolution alters clustering patterns

Hierarchical clustering of transcriptomes has been utilized as a discovery tool for cell and tissue type homologies across species (Tschopp et al., 2014), and for reconstructing the phylogeny of cell type diversification (Arendt, 2008, Kin et al., 2015). However, cross-species transcriptome comparisons often yield clustering patterns that reflect species identity rather than tissue homology (Lin et al., 2014, Gilad, 2015, Merkin et al., 2012, Tschopp et al., 2014, Brawand et al., 2011). For example, lung and spleen evolved prior to the common ancestor of birds and mammals, yet we found lung and spleen samples from chicken and mouse clustered by species, rather than tissue identity (Fig. 4A). One possibility is that concerted evolution is altering clustering patterns by generating species-specific similarity between tissues that overcomes similarity from homology.

To explore whether concerted evolution is affecting hierarchical clustering, we determined the clustering pattern for every pair of tissue and species in the Brawand, Merkin, and Tschopp datasets. We found that tissues with high levels of concerted evolution cluster by species (Fig. 4B; symbols in red, Table S4). Further, the more distantly related the species, the more likely tissues clustered by species rather than homology. Tissues collected from species isolated less then 50 million years always clustered by tissue homology. However, with more than 50 million years of divergence, nearly all tissue pairs with $\gamma > .5$ clustered by species, rather than homologous tissue. These findings are consistent with the interpretation that concerted evolution is contributing to changes in tissue clustering patterns (Musser and Wagner, 2015). Phylogenetic divergence time also affects clustering because greater divergence time allows for the accumulation of more species-specific similarities via concerted evolution than shorter divergence times, whereas similarity due to homology remains the same or decays. Thus, in cross-species comparisons, transcriptomes reflect both the homology of cell and tissue types, as well as the effect of concerted evolution within each species lineage.

Concerted evolution differs with phylogenetic age of tissue

The observation that tissues exhibit different degrees of concerted evolution that tissues' evolutionary independence may itself evolve. For instance, do cell and tissue types of recent origin exhibit more concerted evolution relative to other tissues? We addressed this question using transcriptomes collected in our lab from developing bird feathers, two different avian scale types, and claws from two avian species, chicken and emu. Feathers are an evolutionary innovation that evolved in bird ancestors, replacing scales across much of the body. Scales are found in all reptiles, and also on the feet of birds. Bird scales include large, asymmetric "scutate" scales, found on the top of the foot, and small, symmetric "reticulate" scales on the sides and plantar surface. Birds also have claws, which develop from distal toe skin, and develop immediately adjacent to both avian scale types. Claws are shared by reptiles and mammals, indicating that they are phylogenetically older than feathers and scales.

We sampled epidermis during early development of these skin appendages in chicken and emu. Concerted evolution estimates (Table 1) ranged from .71 (feathers and claws) to .88 (feathers and scutate scales). Claws, the phylogenetically most ancient skin appendage in this sample, undergo less concerted evolution with respect to the other skin appendages. Feathers,

which evolved most recently, are highly correlated with scutate scales (gamma=.88), but exhibit less concerted evolution with reticulate scales (gamma=.79, t-test on bootstrapping p=0.0009) or claws (gamma=.72, t-test on bootstrapping p=0.0003). Notably, feathers and scutate scales undergo more concerted evolution than is found between the two avian scale types (gamma = .86). These results indicate that the degree of concerted evolution is not determined by the shared anatomical location of scales and claws on the distal hindlimb. Rather, our findings show that feathers evolve in concert with scutate scales. This is consistent with previous hypotheses that feathers and scutate scales are homologous in early development, and are both derived from ancestral archosaur scales (Prum and Williamson, 2001, Harris et al., 2002, Musser et al., 2015, Di-Poï and Milinkovitch, 2016). Thus, our results show that the degree of genetic individuation is increasing over evolutionary time.

	Feather	**Scutate Scale**	**Reticulate Scale**	**Claw**
Feather	1	0.88	0.79	0.71
Scutate Scale		1	0.86	0.72
Reticulate Scale			1	0.71
Claw				1

Conclusions

We analyzed patterns of transcriptome divergence among different cell and tissue types using a stochastic model of gene expression evolution. We found different cell and tissue types undergo extensive concerted gene expression evolution. We interpret the signal for concerted evolution as reflecting the gene regulatory network architecture shared between cell types, with cells sharing more similar gene regulatory networks having a higher rate of concerted evolution than cells with less similar gene regulatory networks. Our findings indicate that cell types exhibit cell type-specific control of gene expression, as well as regulatory information shared across broader classes of cells. This warrants a renewed focus on understanding mechanistically how regulatory information affects the evolution of gene expression in different cells and tissues.

Our study also raises several issues of practical importance for comparative transcriptomics. First, it is clear that many evolutionary changes in gene expression are not

limited to a single cell type. As a result, identifying gene expression changes responsible for phenotypic evolution in a particular cell type will require sampling a range of related tissues, in addition to the tissue of interest. Second, new computational tools are needed to remove concerted evolution signal if hierarchical clustering is to be employed for classifying closely related homologous cell and tissue types. This will be challenging because concerted evolution is occurring to a different extent in different subsets of tissues. For instance, our results show that concerted evolution does not merely reflect species-specific physiological evolution, or "species signal", but instead is influenced by the evolutionary history and individuation of tissues and cell types within the organism. In this regard, it is worth noting the parallels between comparative analysis of transcriptomics and the study of gene evolution. During the 1970s and 80s, it was discovered that hierarchical comparisons of gene sequences within and among individuals and species required understanding the phylogenetic histories of species lineages, gene duplication, and sequence differentiation. Similarly, transcriptomic analyses have now revealed that comparative analysis of gene expression in cell types and tissues requires the recognition of the hierarchical consequences of species phylogeny, cell type origin, and transcriptomic individuation (Fig. 1).

Measuring the relative rate of concerted evolution can provide valuable insights into the evolution and individuation of cell types. Individuation has been proposed as the principal mechanism by which new cell and tissue types evolve (Wagner, 1989, Arendt, 2008, Wagner et al., 2007, Wagner, 2014, Musser and Wagner, 2015, Arendt et al. *In Press*). Individuation is the process by which new cell types gain the ability to evolve independently. This requires evolving new mechanisms for controlling gene expression, such as the evolution of new transcription factor complexes (Wagner, 2014). Arendt (2008) and others (Musser and Wagner, 2015, Arendt et al. *In Press*) have also proposed that individuation typically results in the "splitting" of an ancestral cell type into two descendant, and partially individuated, sister cell types. Liang et al (2015) showed statistical support for this model by demonstrating that cell type transcriptomes have a tree-like pattern of relationships, which is expected if sister cell types originate by bifurcation of an ancestral cell type. Our estimates of concerted evolution in avian skin appendages suggest a similar process may explain the evolution of early feather development. Confirmation of this will require a better understanding of gene expression regulation in feathers and bird scales, as well as comparisons with non-avian reptile scales. In any case, understanding

how cell types arise, and become individuated, is a central challenge of the emerging field of cell type evolution. Our model provides a tool for studying this process.

Methods

Publicly available transcriptome data

Transcript read counts in Merkin et al. (Merkin et al., 2012) were downloaded from GSE41637 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41637). Nine tissues (brain, colon, heart, kidney, liver, lung, muscle, spleen, and testes) in five species (chicken, cow, mouse, rat, and rhesus macaque) are profiled in this work. We followed methods in Merkin et al. (Merkin et al., 2012) to map each transcript to respective genomes (musmus9, rhemac2, ratnor4, bostau4, galgal3) and extract FPKM values for each gene. We then normalized FPKM to TPM (Transcript per million mapped transcripts) by the sum (Wagner et al., 2012).

Normalized gene expression RPKM from Brawand et al. (2011) was downloaded from the supplementary tables of the original publication. (http://www.nature.com/nature/journal/v478/n7369/full/nature10532.html). Six tissues (cerebellum, brain without cerebellum, heart, kidney, liver and testis) in nine mammalian species and chicken are profiled in this work.

Transcript read counts from Tschopp et al. (Tschopp et al., 2014) was downloaded from GSE60373 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60373). Tail bud, forelimb bud, hindlimb bud, and genital bud in early development stage from mouse and anolis were profiled in this work. Read counts were also normalized to TPM values.

Skin appendage at placode stage transcriptome data

To collect skin appendage placode transcriptome data we dissected developing skin from embryos of single comb white leghorn chicken and emu. Fertile chicken eggs, obtained from the University of Connecticut Poultry Farm, and emu eggs, obtained from Songline Emu Farm in Gill, MA, were incubated in a standard egg incubator at 37.8 and 35.5 degrees Celsius, respectively. Skin appendages in birds develop at different times across the embryo. We dissected embryonic placode skin from dorsal tract feathers between Hamburger and Hamilton (1951; H&H) stages 33 and 34, asymmetric (scutate) scales from the dorsal hindlimb tarsometatarsus at H&H stage 37, symmetric (reticulate) scales from the hindlimb ventral

metatarsal footpad at H&H stage 39, and claws from the tips of the hindlimb digits at H&H stage 36. Dissected skin was treated with 10 mg/mL dispase for approximately 12 hours at 4 degrees Celsius to allow for complete separation of epidermis and dermis. RNA was extracted from epidermal tissue using a Qiagen Rneasy kit. Stand-specific polyadenylated RNA libraries were prepared for sequencing by the Yale Center for Genome Analysis using an in-house protocol. For each individual tissue sample we sequenced approximately 50 million reads (one-quarter of a lane) on an Illumina Hiseq 2000. We sequenced two biological replicates for each tissue, with the exception of emu symmetric scales and claws, for which we were only able to obtain one replicate due to limitations in the number of eggs we could acquire.

Reads were mapped to the respective species genome available at the time: WASHUC2 for chicken (Ensembl release 68 downloaded October 4th, 2012) and a preliminary version of the Emu genome generated by the laboratory of Allan Baker at the Royal Ontario Museum and University of Toronto (mapped on December 5th, 2012). Reads were mapped with Tophat2 v2.0.6 using the –GTF option. Mapped reads were assigned to genes using the program HTSeq, requiring that reads map to the correct strand, and using the "intersection-nonempty" option for dealing with reads that mapped to more than one feature.

Estimating the rate of concerted transcriptome evolution

For all datasets we used square root transformed TPM (transcripts per million transcripts) as the measure of gene expression level (Wagner et al., 2012). We used the square root transformation to correct for the heteroscedasticity of transcriptome data (Musser and Wagner, 2015). However, we found some very highly expressed genes, such as mitochondrial genes, still had high variance relative to other genes even after the square root transformation. These were removed to prevent overestimation of concerted evolution. We then averaged sqrt(TPM) values across replicates to generate a representative transcriptome for each tissue in each species.

All transcriptome data were further processed to facilitate cross-species comparison, with the exception of the Brawand dataset which was already normalized between species using a median-scaling procedure on highly conserved genes. For the other datasets we followed the normalization method of Musser and Wagner (Musser and Wagner, 2015), using one-to-one orthologs.

We estimated concerted evolution (γ) for each combination of tissue and species in all datasets by estimating the Pearson correlation of their contrast vectors. We tested the significance of concerted evolution by calculating a p-value from a null distribution of randomly permutated gene expression contrast vectors (Figure S1). One challenge of estimating the p-value is that gene expression values between genes are not independent, for instance because they are co-regulated by the same transcription factors (Pavlicev and Widder, 2015). To account for this, we only subsampled 50 genes for each permutation, which represents a lower bound for the number of independently varying gene clusters identified in single cell transcriptome data (Supplemental methods) (Wagner et al., 2008, Pavlicev et al. *Forthcoming*). We conducted 1000 permutations to generate a null distribution of the correlation coefficient γ. Actual estimates of concerted evolution were considered significantly higher than expected by chance if they fell in the top 5% of the null distribution. We conducted anova tests in R to determine whether concerted evolution values varied depending on time since the last common ancestor of the species used for comparisons. We tested whether mean γ values estimated independently from the Brawand and Merkin datasets for the same tissue pairs were significantly different using Welch's t-test.

References and Notes:

ACHIM, K., PETTIT, J. B., SARAIVA, L. R., GAVRIOUCHKINA, D., LARSSON, T., ARENDT, D. & MARIONI, J. C. 2015. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol,* 33**,** 503-9.

ARENDT, D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet,* 9**,** 868-82.

ARENDT, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D. & Wagner, G. P. 2016. Evolution of sister cell types by individuation. *Nat Rev Genet, In Press*

BRAWAND, D., SOUMILLON, M., NECSULEA, A., JULIEN, P., CSARDI, G., HARRIGAN, P., WEIER, M., LIECHTI, A., AXIMU-PETRI, A., KIRCHER, M., ALBERT, F. W., ZELLER, U., KHAITOVICH, P., GRUTZNER, F., BERGMANN, S., NIELSEN, R., PAABO, S. & KAESSMANN, H. 2011. The evolution of gene expression levels in mammalian organs. *Nature,* 478**,** 343-8.

BRENDOLAN, A., ROSADO, M. M., CARSETTI, R., SELLERI, L. & DEAR, T. N. 2007. Development and function of the mammalian spleen. *Bioessays,* 29**,** 166-77.

BRUNET, T., FISCHER, A. H. L., STEINMETZ, P. R. H., LAURI, A., BERTUCCI, P. & ARENDT, D. 2016. The evolutionary origin of bilaterian smooth and striated myocytes. *bioRxiv*.

DI-POÏ, N. & MILINKOVITCH, M. C. 2016. The anatomical placode in reptile scale morphogenesis indicates shared ancestry among skin appendages in amniotes. *Science Advances,* 2.

EISENBERG, E. & LEVANON, E. Y. 2013. Human housekeeping genes, revisited. *Trends Genet,* 29**,** 569-74.

FELSENSTEIN, J. 2004. *Inferring phylogenies,* Sunderland, Mass., Sinauer Associates.

GILAD, Y. M.-M., O. 2015. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research*.

HARRIS, M. P., FALLON, J. F. & PRUM, R. O. 2002. Shh-Bmp2 signaling module and the evolutionary origin and diversification of feathers. *J Exp Zool,* 294**,** 160-76.

HEDGES, S. B., DUDLEY, J. & KUMAR, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics,* 22**,** 2971-2.

KHAITOVICH, P., ENARD, W., LACHMANN, M. & PÄÄBO, S. 2006. Evolution of primate gene expression. *Nat Rev Genet,* 7**,** 693-702.

KHAITOVICH, P., HELLMANN, I., ENARD, W., NOWICK, K., LEINWEBER, M., FRANZ, H., WEISS, G., LACHMANN, M. & PAABO, S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science,* 309**,** 1850-4.

KIN, K., NNAMANI, M. C., LYNCH, V. J., MICHAELIDES, E. & WAGNER, G. P. 2015. Cell-type phylogenetics and the origin of endometrial stromal cells. *Cell Rep,* 10**,** 1398-409.

LIANG, C., CONSORTIUM, F., FORREST, A. R. & WAGNER, G. P. 2015. The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat Commun,* 6**,** 6066.

LIN, S., LIN, Y., NERY, J. R., URICH, M. A., BRESCHI, A., DAVIS, C. A., DOBIN, A., ZALESKI, C., BEER, M. A., CHAPMAN, W. C., GINGERAS, T. R., ECKER, J. R. & SNYDER, M. P. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A,* 111**,** 17224-9.

MERKIN, J., RUSSELL, C., CHEN, P. & BURGE, C. B. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science,* 338**,** 1593-9.

MUSSER, J. M. & WAGNER, G. P. 2015. Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called "species signal". *J Exp Zool B Mol Dev Evol*.

MUSSER, J. M., WAGNER, G. P. & PRUM, R. O. 2015. Nuclear beta-catenin localization supports homology of feathers, avian scutate scales, and alligator scales in early development. *Evol Dev,* 17**,** 185-94.

PANKEY, M. S., MININ, V. N., IMHOLTE, G. C., SUCHARD, M. A. & OAKLEY, T. H. 2014. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proc Natl Acad Sci U S A,* 111**,** E4736-42.

PAVLICEV, M. & WIDDER, S. 2015. Wiring for independence: positive feedback motifs facilitate individuation of traits in development and evolution. *J Exp Zool B Mol Dev Evol,* 324**,** 104-13.

PRUM, R. O. & WILLIAMSON, S. 2001. Theory of the growth and evolution of feather shape. *J Exp Zool,* 291**,** 30-57.

RUVINSKY, I., OHLER, U., BURGE, C. B. & RUVKUN, G. 2007. Detection of broadly expressed neuronal genes in C. elegans. *Dev Biol,* 302**,** 617-26.

SIEBURTH, D., CH'NG, Q., DYBBS, M., TAVAZOIE, M., KENNEDY, S., WANG, D., DUPUY, D., RUAL, J. F., HILL, D. E., VIDAL, M., RUVKUN, G. & KAPLAN, J. M. 2005. Systematic analysis of genes required for synapse structure and function. *Nature,* 436**,** 510-7.

STEFANAKIS, N., CARRERA, I. & HOBERT, O. 2015. Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. *Neuron,* 87**,** 733-50.

STEINMETZ, P. R., KRAUS, J. E., LARROUX, C., HAMMEL, J. U., AMON-HASSENZAHL, A., HOULISTON, E., WORHEIDE, G., NICKEL, M., DEGNAN, B. M. & TECHNAU, U. 2012. Independent evolution of striated muscles in cnidarians and bilaterians. *Nature,* 487**,** 231-4.

TSCHOPP, P., SHERRATT, E., SANGER, T. J., GRONER, A. C., ASPIRAS, A. C., HU, J. K., POURQUIE, O., GROS, J. & TABIN, C. J. 2014. A relative shift in cloacal location repositions external genitalia in amniote evolution. *Nature,* 516**,** 391-4.

WAGNER, G. P. 1989. The Origin of Morphological Characters and the Biological Basis of Homology. *Evolution,* 43**,** 1157-1171.

WAGNER, G. P. 2014. *Homology, genes, and evolutionary innovation*, Princeton University Press.

WAGNER, G. P., KENNEY-HUNT, J. P., PAVLICEV, M., PECK, J. R., WAXMAN, D. & CHEVERUD, J. M. 2008. Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature,* 452**,** 470-2.

WAGNER, G. P., KIN, K. & LYNCH, V. J. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci,* 131**,** 281-5.

WAGNER, G. P., PAVLICEV, M. & CHEVERUD, J. M. 2007. The road to modularity. *Nat Rev Genet,* 8**,** 921-31.

WANG, Z., YOUNG, R. L., XUE, H. & WAGNER, G. P. 2011. Transcriptomic analysis of avian digits reveals conserved and derived digit identities in birds. *Nature,* 477**,** 583-6.

YUE, F., CHENG, Y., BRESCHI, A., VIERSTRA, J., WU, W., RYBA, T., SANDSTROM, R., MA, Z., DAVIS, C., POPE, B. D., SHEN, Y., PERVOUCHINE, D. D., DJEBALI, S., THURMAN, R. E., KAUL, R., RYNES, E., KIRILUSHA, A., MARINOV, G. K., WILLIAMS, B. A., TROUT, D., AMRHEIN, H., FISHER-AYLOR, K., ANTOSHECHKIN, I., DESALVO, G., SEE, L. H., FASTUCA, M., DRENKOW, J., ZALESKI, C., DOBIN, A., PRIETO, P., LAGARDE, J., BUSSOTTI, G., TANZER, A., DENAS, O., LI, K., BENDER, M. A., ZHANG, M., BYRON, R., GROUDINE, M. T., MCCLEARY, D., PHAM, L., YE, Z., KUAN, S., EDSALL, L., WU, Y. C., RASMUSSEN, M. D., BANSAL, M. S., KELLIS, M., KELLER, C. A., MORRISSEY, C. S., MISHRA, T., JAIN, D., DOGAN, N., HARRIS, R. S., CAYTING, P., KAWLI, T., BOYLE, A. P., EUSKIRCHEN, G., KUNDAJE, A., LIN, S., LIN, Y., JANSEN, C., MALLADI, V. S., CLINE, M. S., ERICKSON, D. T., KIRKUP, V. M., LEARNED, K., SLOAN, C. A., ROSENBLOOM, K. R., LACERDA DE SOUSA, B., BEAL, K., PIGNATELLI, M., FLICEK, P., LIAN, J., KAHVECI, T., LEE, D., KENT, W. J., RAMALHO SANTOS, M., HERRERO, J., NOTREDAME, C., JOHNSON, A., VONG, S., LEE, K., BATES, D., NERI, F., DIEGEL, M., CANFIELD, T., SABO, P. J., WILKEN, M. S., REH, T. A., GISTE, E., SHAFER, A., KUTYAVIN, T., HAUGEN, E., DUNN, D., REYNOLDS, A. P., NEPH, S., HUMBERT, R., HANSEN, R. S., DE

BRUIJN, M., et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature,* 515**,** 355-64.

Acknowledgments: We are thankful for the contributions of Allan Baker, principal investigator of the Emu Genome Project, who died unexpectedly in November 2014. Allan leaves behind a long legacy in the fields of ornithology, evolutionary biology, and avian conservation. He will be greatly missed. We would like to thank Thomas Stewart for his thoughtful and helpful review of the manuscript. CL gratefully acknowledges the receipt of a graduate fellowship from China Scholarship Council and the support of the Raymond and Beverly Sackler Institute. JM gratefully acknowledges funding from the Dillon and Mary Ripley Graduate Fellowship Fund at Yale University and the National Science Foundation Graduate Research Fellowship under grant No. DGE-1122492. The research was supported by the W. R. Coe Fund at Yale University, the John Templeton Foundation grant #54860, NSF grant # 14-001273, and the Yale University Science Development Fund. The opinions expressed in this article are not those of the John Templeton Foundation. RNAseq data for bird skin appendages was produced in the Yale Center of Genome Analysis.

Figure Captions:

Fig. 1 Concerted evolution of cell transcriptomes. (A) Cell type history overlaid with species history. Cell types A and B originate from ancestral cell type O prior to split of species 1 and 2. Cell type lineages show homology relationships for the two cell types in the descendant species (subscripts indicate species). **(B)** Cell type history without species phylogeny illustrates cell type homology. Grey squiggly lines indicate concerted changes to the transcriptome. Concerted transcriptome evolution leads to an increase of transcriptome similarities in cell types of the same species relative to their homologous counterpart in other species. **(C)** Concerted evolution causes the accumulation of species-specific gene expression similarities between tissues in the same organism, potentially resulting in different tissues within a species being more similar to each other than to their homologous counterparts in another species.

Fig. 2 Estimates of the relative rate of concerted evolution. Concerted evolution estimates for Tschopp, Merkin, and Brawand datasets. Tschopp estimates (γ; points in red) are for early

forelimb-, hindlimb-, genital-, and tail buds from mouse and *Anolis*. Merkin and Brawand estimates ($\bar{\gamma}$) are from 10 different mature tissues in 12 species of mammal and chicken. Estimates of concerted evolution range from 0 to 1, with higher values indicating greater concerted evolution. Dotted line indicates cutoff for estimates of γ significantly greater than expected by chance under a model without concerted evolution. Numbers identify concerted evolution estimates for tissues comparisons with spleen (points in green) and cerebellum (points in blue): 1) cerebellum and forebrain, 2) spleen and lung, 3) spleen and colon, 4) spleen and kidney, 5) spleen and forebrain, 6) cerebellum and heart, 7) cerebellum and kidney, 8) spleen and heart, 9) spleen and liver, 10) spleen and skeletal muscle, 11) cerebellum and liver, 12) spleen and testes, 13) cerebellum and testes.

Fig. 3 Effect of tissue type, species divergence, and dataset on estimates of concerted evolution. (A) estimates of the relative rate of concerted evolution (γ) from each species and tissue pair for the five tissues sampled in both Merkin and Brawand datasets (forebrain, heart, kidney, liver and testes). In each panel is a plot for one tissue pairs. The horizontal axis is speciation time in million years at each internal node as shown in 2B and 2C. The vertical axis is the estimated relative concerted transcriptome evolution signal γ. We used independent contrasts of square root TPM of every gene to calculate γ. The estimated concerted evolution signal is consistent in each plot, and for both datasets. The grey lines indicates the mean value $\bar{\gamma}$ of γ (solid line for Merkin data, and dashed line for Brawand data). Colors of points represent empirical p-values estimated from permutation test with 50 random selected genes according to heat scale below. **(B)** The time tree for species analyzed in Brawand dataset. **(C)** The time tree for species analyzed in Merkin dataset. All tissues considered here are much older than the species compared in these two datasets.

Fig. 4 Concerted evolution can cause tissues to cluster by species as opposed to homology (A) Hierarchical clustering of mouse and chicken lung and spleen illustrating examples of tissues clustering by species. Clustering is performed with R package pvclust. Numbers at nodes indicate bootstrap support for clusters. **(B)** Levels of concerted evolution versus clustering pattern of tetrads with samples from two species and two homologous tissue types. Higher relative rates of concerted evolution is associated with larger chance of clustering by species.

Table 1: Concerted evolution among avian skin appendages. Gamma estimates were obtained from transcriptomes of early skin appendage development (placodes) in chicken and emu embryos. For each skin appendage type in each species we averaged gene expression across replicates, and calculated gamma using gene expression contrasts smaller than 10 to avoid the influence of highly expressed structural genes.

Figure 1

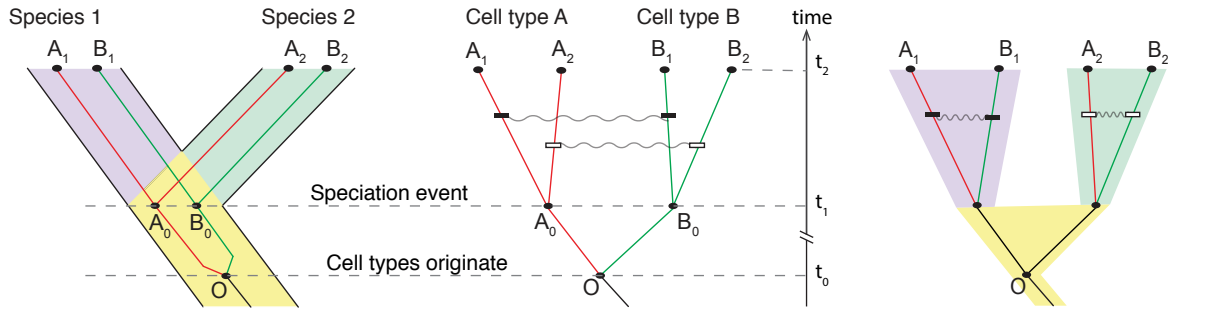

Species 1
Species 2
A1
B1
A2
B2
A0
B0
O
Speciation event
Cell types originate
Cell type A
Cell type B
time
t2
t1
t0
A
Species phylogeny
B
Cell type history
C
Similarity of transcriptomes

Figure 2

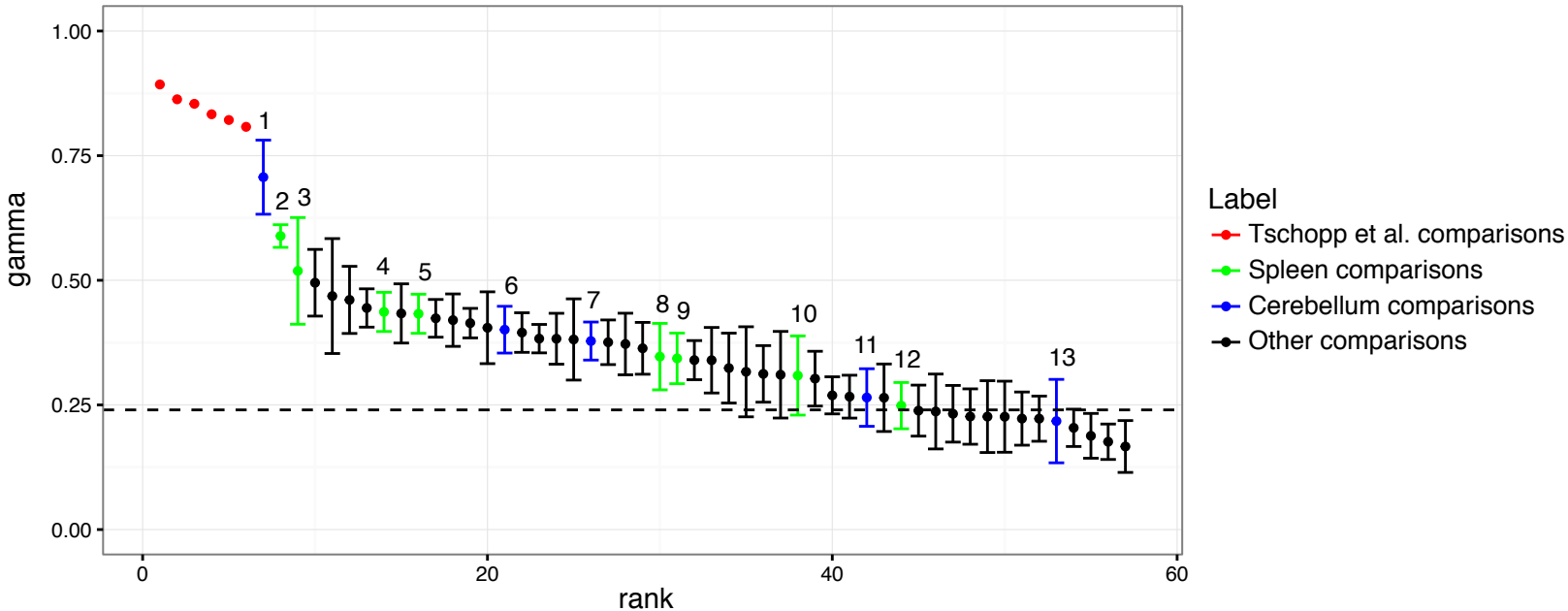

1.00
0.75
0.50
0.25
0.00
gamma
0
20
40
60
rank
1
2
3
4
5
6
7
8
9
10
11
12
13
Label
Tschopp et al. comparisons
Spleen comparisons
Cerebellum comparisons
Other comparisons

Figure 3

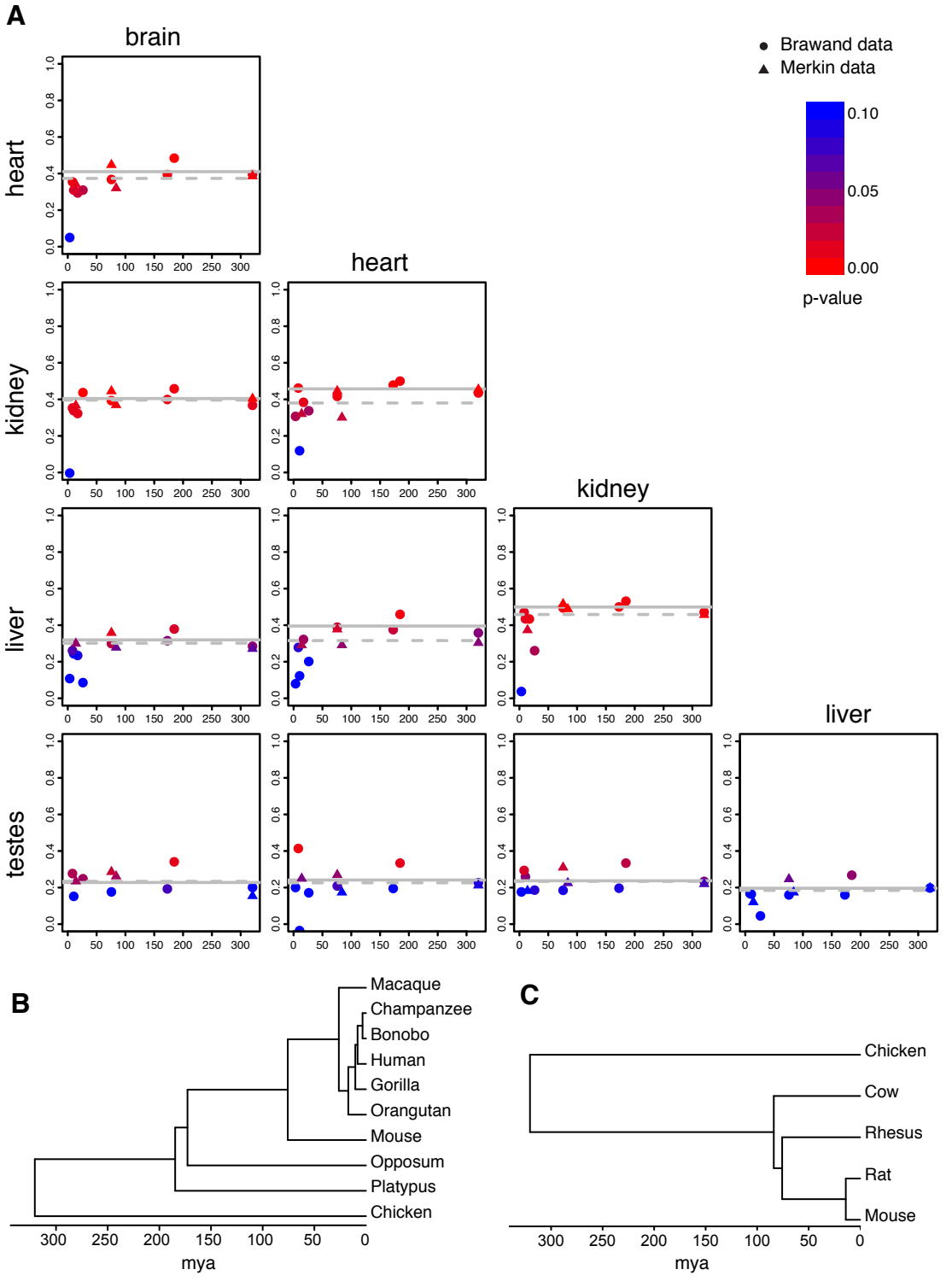
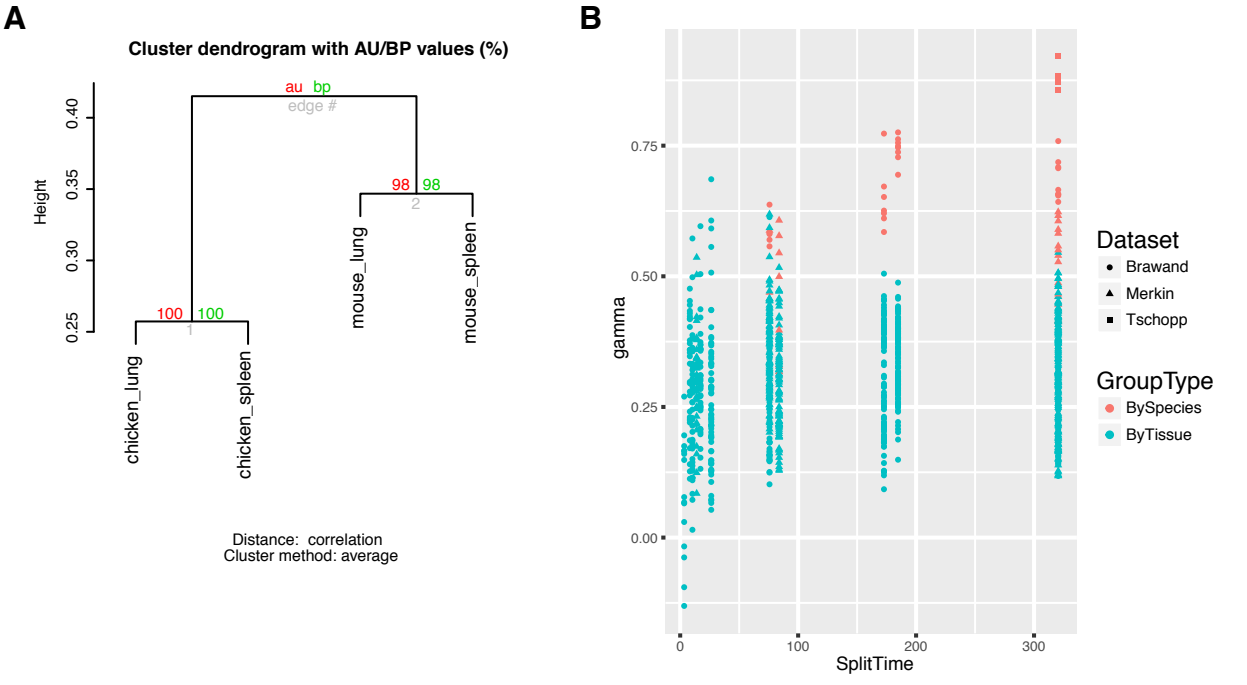

A
brain
heart
kidney
liver
testes
Brawand data
Merkin data
0.10
0.05
0.00
p-value
B
Macaque
Champanzee
Bonobo
Human
Gorilla
Orangutan
Mouse
Opposum
Platypus
Chicken
300 250 200 150 100 50 0
mya
C
Chicken
Cow
Rhesus
Rat
Mouse
300 250 200 150 100 50 0
mya

Figure 4


A
Cluster dendrogram with AU/BP values (%)
au bp
edge #
98 98
2
100 100
1
Height
0.40
0.35
0.30
0.25
chicken_lung
chicken_spleen
mouse_lung
mouse_spleen
Distance: correlation
Cluster method: average
B
gamma
0.75
0.50
0.25
0.00
0
100
200
300
SplitTime
Dataset
Brawand
Merkin
Tschopp
GroupType
BySpecies
ByTissue