bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identifying adaptive and plastic gene expression levels using a unified model for expression variance between and within species

Rori V. Rohlfs[1*] and Rasmus Nielsen[1,2]

1 Department of Integrative Biology, University of California Berkeley, CA, USA

2 Center for Bioinformatics, University of Copenhagen, Denmark

Corresponding author:
Rori V. Rohlfs
1005 Valley Life Sciences Building #3140
Berkeley, CA 94720
USA
(510) 643 - 0060
rrohlfs@berkeley.edu

Running title:
Identifying adaptive and plastic expression levels

Keywords:
comparative expression; expression adaptation; plasticity; Ornstein-Uhlenbeck model; population variance

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Thanks to the reduced cost of RNA-Sequencing and other advanced methods for quantifying expression levels, accurate and expansive comparative expression data sets including data from multiple individuals per species are emerging. Comparative genomics has been greatly facilitated by the availability of statistical methods considering both between and within species variation for testing hypotheses regarding the evolution of DNA sequences. Similar methods are now needed to fully leverage comparative expression data. In this paper, we describe the β model which parameterizes the ratio of population to evolutionary expression variance, facilitating a wide variety of analyses, including a test for expression divergence or diversity for a single gene or a class of genes. The β model can also be used to test for lineage-specific shifts in expression level, amongst other applications. We use simulations to explore the functionality of these tests under a variety of circumstances. We then apply them to a mammalian phylogeny of 15 species typed in liver tissue. We identify genes with high expression divergence between species as candidates for expression level adaptation, and genes with high expression diversity within species as candidates for expression level conservation and plasticity. Using the test for lineage-specific expression shifts, we identify several candidate genes for expression level adaptation on the catarrhine and human lineages, including genes possibly related to dietary changes in humans. We compare these results to those reported previously using the species mean model which ignores population expression variance, uncovering important differences in performance.

1 Introduction

Comparative expression studies including variation within species are now emerging (Kalinka et al., 2010; Brawand et al., 2011; Perry et al., 2012), made possible by the developments of both RNA-Seq as a reliable method to quantify expression (Wang et al., 2009), and bioinformatic methods to appropriately normalize expression accounting for species differences, (Bullard et al., 2010; Trapnell et al., 2010; Li and Durbin, 2009; Pickrell et al., 2010).

Initially, comparative expression analyses considered few species and employed adaptations of standard statistical methods (particularly ANOVA) to detect genes with unusually high expression divergence between species given the variance within species (Nuzhdin et al., 2004; Gilad et al., 2006; Khaitovich et al., 2006; Whitehead and Crawford, 2006). These methods typically assume independence between species. However, as more species are considered, the difference in shared evolutionary history between closely and distantly related species increases, and a complex covariance structure emerges. In current comparative expression datasets across larger phylogenies, the assumption of species independence does not hold, necessitating more sophisticated methods taking into account evolutionary relationships (Felsenstein, 1985).

Subsequently, a number of models have been developed to describe the evolution of gene expression under neutral drift, stabilizing selection, or directional selection (Butler and King, 2004; Bedford and Hartl, 2008). These models are used to calculate the expected species average expression levels and expression covariance between species under a particular evolutionary scenario. Likelihood ratio (LR) tests can then be formulated to distinguish neutrality, stabilizing selection, and directional selection, as has been successfully analyzed in a number of datasets (Bedford and Hartl, 2008; Kalinka et al., 2010; Perry et al., 2012). These methods have been augmented to allow for within species expression level variance (Lynch, 1991; Felsenstein, 2008; Hansen and Bartoszek, 2012; Rohlfs et al., 2014).

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We build upon these models to create a the β model, describing expression evolution between species and variance within species. The β model facilitates rigorous analyses investigating hypotheses of expression level conservation, adaptation, and plasticity. Our model allows analyses of expression variance between and within species, analogous to ANOVA style comparative expression methods, but accounting for phylogenetic structure.

The model and parameterization we propose can be used for an expression analogy to classic genetic neutrality tests considering polymorphism and diversity, namely, the HKA (Hudson et al., 1987) and McDonald-Kreitman (McDonald and Kreitman, 1991) tests. In these tests, the amounts of relative amounts polymorphism within species and divergence between species at synonymous and non-synonymous sites within a gene are compared. Under neutrality, the ratio of synonymous to non-synonymous differences will be the same in the polymorphism and divergence data. However, in genes affected by selection, that ratio may be higher or lower for divergence than polymorphism data, depending on the directionality of selection. Analogously, in our model, we parameterize the ratio of within to between species expression variance across genes as β_{shared}. Genes with an unusually high ratio ($\beta_i > \beta_{shared}$) show proportionally high expression divergence and may be subject to directional selection on expression level. Genes with an unusually low ratio ($\beta_i < \beta_{shared}$) show proportionally high expression diversity and may have conserved expression levels which are plastic or environmentally responsive. This test can alternatively be thought of as a phylogenetic ANOVA, as it accounts for varying relationships between species.

By selectively constraining parameters of the β model, a variety of additional tests can be constructed. For example, we can test for unusual species or lineage-specific expression variance, as may be observed under recent relaxation of constraint on expression level, diversifying selection on expression level, increased constraint on gene expression level, or under extreme branch-specific demographic processes. Other tests may be constructed to test for differing expression diversity for groups of individuals within each species, for example, evolutionarily conserved age or sex-specific expression variance. All of these tests could be performed on a particular gene of interest or on a class of genes interest, for example a list of candidate genes could be queried for increased expression diversity in older individuals. In addition to these novel tests, the β model can be used for the same tests as other expression evolution models which discount within-species variance. In particular, the β model can test for lineage-specific shifts in expression level, while taking into account within-species variance.

Here, we explore the performance of two tests: the test for unusual expression divergence or diversity and the test for lineage-specific expression level shifts. We use simulations to describe these tests and formulate expectations under the null hypotheses We then apply the tests to a previously published expression dataset typed which includes 15 mammals. We identify a number of genes with high expression level divergence between species as candidates for expression level adaptation, and genes with high expression level diversity within species as candidates for environmentally-responsive gene expression (plasticity). Using the test for lineage-specific expression shifts, we identify several viable candidate genes for expression adaptation on the catarrhine and human lineages. For comparison, we consider the species mean model described by Bedford and Hartl (2008) and recently used in a number of practical analyses (Bedford and Hartl, 2008; Kalinka et al., 2010; Perry et al., 2012). We compare our results to those obtained using the species mean method, noting important differences, especially for analyses of

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species-specific expression shifts (Perry et al., 2012).

2 Results

2.1 The β model for gene expression evolution and population variance

The evolution of quantitative traits by drift and stabilizing selection has been modeled using an Ornstein-Uhlenbeck (OU) process, which can be thought of as a random walk with a pull towards an optimal value (Hansen, 1997; Butler and King, 2004; Hansen et al., 2008; Bedford and Hartl, 2008; Kalinka et al., 2010). In an OU model of stabilizing selection on gene expression level, the parameter θ_i can be thought of as the optimal expression level for gene i, σ_i^2 the degree of drift acting that expression level, and α_i strength of selective stabilization on that expression level. Over evolutionary time, the stationary variance of species mean expression levels for gene i will be $\frac{\sigma_i^2}{2\alpha_i}$, which we refer to as the evolutionary variance.

More recently, several OU-based models have been augmented to include within-species population level variance (Felsenstein, 2008; Lynch, 1991; Hansen and Bartoszek, 2012; Rohlfs et al., 2014). Accounting for population variance is crucial to distinguish non-phylogenetic variation, total drift, and stabilizing selection (Rohlfs et al., 2014).

The model we describe builds on these OU models for quantitative trait evolution with the additional parameter β which describes the ratio of population to evolutionary variance. Within species j the expression level of any individual k is distributed as $Y_{jk} \sim N(Y_j, \beta\frac{\sigma^2}{2\alpha})$, where Y_j is the species mean expression level determined by the OU process. We call this the β model, which describes a linear relationship between population and evolutionary variance.

In his classic paper, Lande (1976) showed that under an OU model of stabilizing selection, a linear relationship arises between the evolutionary variance and the population variance within species. Additionally, the Poisson nature of RNA-Seq and gene expression itself means that both evolutionary and population expression variance increase with expression mean. With that in mind, our model assumes a linear relationship between evolutionary and population expression variance. That assumption is reflected in the data, which shows a linear relationship between estimated evolutionary variance ($\frac{\hat{\sigma_i^2}}{2\hat{\alpha_i}}$) and estimated population variance ($\hat{\beta}_i\frac{\hat{\sigma_i^2}}{2\hat{\alpha_i}}$) (Figure 1).

The slope of this linear relationship (parameterized by β) should be consistent across genes which have undergone the same evolutionary and demographic processes under stabilizing selection. However, in a gene, i, which have experienced directional selection, driving increased divergence between species, while maintaining low variance within species, β_i would be lower as compared to other genes in the same individuals. Similarly, a gene with plastic expression level may have more variation within species than between as compared to other genes, raising the value of β_i.

2.2 Testing for unusual expression divergence and diversity

Using the β model, we can test each gene for unusual expression divergence between species or diversity within species. This test amounts to a gene expression analogy to the HKA or MK tests comparing the amount of polymorphism within species to the divergence between species to identify deviations from neutrality.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Specifically, the likelihoods of a set of genes can compared under the null model where β for a particular gene i is constrained to the value of β estimated over the entire set of genes (β_{shared}) ($H_0 : \beta_i = \beta_{shared}$) and under the alternative model where β_i is allowed to vary ($H_a : \beta_i \neq \beta_{shared}$). By comparing these likelihoods in a LR test, we can determine if β_i for a particular gene varies significantly from β_{shared} across the genes. A gene where $\beta_i < \beta_{shared}$ has high expression variance between species as compared to within, or high expression divergence. A gene where $\beta_i > \beta_{shared}$ has high expression variance within species as compared to between, or high expression diversity.

2.2.1 Test expectation under the null hypothesis

At the asymptotic limit, the LR test statistic for testing $H_0 : \beta_i = \beta_{shared}$) versus $H_A = \beta_i \neq \beta_{shared}$), $LRT_{\beta_i \neq \beta_{shared}}$ is χ^2_1 distributed under the null hypothesis. However, when applied to small phylogenies, the distribution of $LRT_{\beta_i \neq \beta_{shared}}$ may not be near the asymptotic limit, and may deviate from a χ^2_1 (Boettiger et al., 2012, e.g.,) (see Supplementary Materials). To explore the null distribution of $LRT_{\beta_i \neq \beta_{shared}}$ over different parameter values and phylogeny sizes, we simulated data under the null hypothesis of $\beta_i = \beta_{shared}$ for four sets of parameter values (Supplementary Table 1) based on the median maximum likelihood estimates from the experimental data, under four tree sizes based on the mammalian phylogeny that we subsequently will analyze (Supplementary Figure 1 and Supplementary Materials).

While the null distribution resembles the asymptotically expected χ^2_1 for a phylogeny like the one analyzed here, we observe some minor deviations (Supplementary Figure 2). However, as the size of the phylogeny considered increases, the null distribution approaches a χ^2_1, though it converges more slowly under some parameter values. As in previous studies examining parameter estimates over phylogeny size (Boettiger et al., 2012), we see that the parameter estimates improve with phylogeny size, though some are more easily estimable than others (Supplementary Figures 5-10).

We performed further simulations based on a fish bone (also called fully pectinate or ladder) phylogeny for different numbers of species (Supplementary Figures 3, 11). Again, we see that as the phylogeny size increases, the simulated null distribution more closely matches the asymptotic expectation. It is important to note that the null distribution under a fish bone topology more quickly approaches χ^2_1 than the other topology, because there are more varying branch lengths between species in a fish bone phylogeny. Trait evolution methods are powered by multiple varying branch length differences between species, making a fish bone phylogeny the most informative.

2.2.2 Parametric bootstrap approach for the null distribution

To account for deviations from the asymptotically expected null distributions of $LRT_{\beta_i \neq \beta_{shared}}$, we follow the suggestion of Boettiger *et al.* (2012) and use a parametric bootstrap. That is, for a particular gene, we simulate expression profiles based on the maximum likelihood parameter estimates under the null hypothesis. These simulated expression profiles are then tested for deviation from the null hypothesis to determine the parametric bootstrapped null distribution of $LRT_{\beta_i \neq \beta_{shared}}$,to which the experimental result can be compared.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We performed a parametric bootstrap analysis with 100 simulations for each of the genes simulated under the null hypothesis described above. For each gene, we compared the original test statistic ($LRT_{\beta_i \neq \beta_{shared}}$) to the distribution created by these additional simulations to determine the parametric bootstrapped p-value. The resulting bootstrapped p-values are approximately uniformly distributed between 0 and 1 (Supplemental Figure 12) as expected. Note that generally the parametric bootstrap approach is most effective for accurate parameter estimates; in the presence of biased estimates and a dependence of the distribution of the LR test statistics on parameter values, the parametric bootstrap approach can be biased. It is therefore worthwhile to test the parametric bootstrap before interpreting results based on it.

2.3 Expression divergence and diversity in mammals

2.3.1 Mammalian expression data

We applied the β model to analyze a comparative expression dataset over 15 mammalian species with about four individuals per species (Perry et al., 2012). Of the 15 species typed, five are anthropoids (common marmoset (mr), vervet (ve), rhesus macaque (mc), chimpanzee (ch), human (hu)), five are lemurs (aye-aye (ay), Coquerel's sifaka (sf), black and white ruffed lemur (bw), mongoose lemur (mn), and crowned lemur (cr)), and the remaining five are more distantly related mammals (slow loris (sl), northern treeshrew (ts), house mouse (ms), nine-banded armadillo (ar), and gray short-tailed opossum (op)). Liver tissue from each individual was typed using RNA-Seq (Perry et al., 2012). We consider a subset of 675 genes with no missing data across all species and individuals.

2.3.2 Assessing expression divergence and diversity

We applied the test for unusual expression divergence between species or unusual diversity within species to each gene in the mammalian dataset. The resulting empirical $LRT_{\beta_i \neq \beta_{shared}}$ values increase with departure from $\hat{\beta}_i = \hat{\beta}_{shared}$ (Figure 2). We see much higher values of $LRT_{\beta_i \neq \beta_{shared}}$ for low $\hat{\beta}_i$ than high $\hat{\beta}_i$. This is partially explained by error in β_i estimates, especially for higher values (Supplementary Figures 5, 11). Additionally, under the null hypothesis, some of the observed expression variance may be explained by increasing the estimated evolutionary variance, so power is reduced for genes with high β_i.

We additionally estimated parametric bootstrapped p-values using 1000 simulations for each gene, finding that they roughly follow a uniform distribution with some excess of low p-values (Supplementary Figure 14). We compared those bootstrapped p-values to $LRT_{\beta_i \neq \beta_{shared}}$ and found a clear correlation (Supplementary Figure 15). Using 1000 simulations, the minimum p-value is 0.001, so more simulations would be needed to more accurately assess the degree of departure from the null distribution in the tail of the distribution.

2.3.3 Candidate genes for expression adaptation and plasticity

Genes in the tail of the $LRT_{\beta_i \neq \beta_{shared}}$ distribution with high $\hat{\beta}$ have unusually high population, as compared to evolutionary, variance, which may be indicative of conservation of gene expression level across species, but with plastic gene expression levels responding to individual

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

environmental conditions. Among the most significant high $\hat{\beta}_i$ genes, we see PPIB which has been implicated in immunosuppression (Price et al., 1991; Luban et al., 1993) and HSPA8, a heat shock protein (Daugaard et al., 2007) (Figure 3a). Based on their function, the expression levels of both of these genes are expected to vary depending on environmental inputs such as pathogen load and temperature.

Conversely, genes with low $\hat{\beta}$ have unusually high evolutionary variance as compared to population variance, which is expected in cases of directional selection on expression level. The most extreme outlier with low $\hat{\beta}_i$ is F10, which encodes Factor X, a key blood coagulation protein produced in the liver (Uprichard and Perry, 2002). F10 is highly expressed in armadillo as compared to the other mammals considered (Figure 3b). This observation is likely related to the fact that armadillo blood coagulates twice or five times faster than human blood, perhaps due in part to higher expression of F10 (Lewis and Doyle, 1964). This functional connection to increased F10 expression suggests that the shift may be adaptive.

3 Testing for branch-specific expression level shifts

The β model can be used to formulate hypotheses about branch-specific shifts in the expression of gene i by comparing likelihoods under $H_0 : \theta_i^a = \theta_i^{non-a}$ versus $H_a : \theta_i^a \neq \theta_i^{non-a}$, where θ_i^a is the value of θ_i at all nodes in the shifted lineage(s), a, and θ_i^{non-a} is the value of θ_i at the remaining $(non-a)$ nodes. The corresponding LR test statistic is asymptotically χ_1^2 distributed. The phylogeny used for these analyses seems sufficient to achieve that asymptotic distribution. (Supplementary Figure 16). We performed this test querying expression level shift on both the catarrhine (containing humans, chimpanzees, rhesus macaques, and vervets) and human lineages (Supplementary Tables 2, 3).

3.1 Candidate genes for adaptation on catarrhine and human lineages

In the test for expression shift in catarrhines (cat), we identify a number of interesting outliers (Supplementary Figure 17). The most significant shift is seen in DEXI, showing a clear increase in expression level in catarrhines. High expression of DEXI has recently been shown to be protective against auto-immune diseases type I diabetes and multiple sclerosis (Davison et al., 2012), indicating an important functional role of this increased expression.

Similarly, the test for expression shift on the human (hum) branch revealed interesting outliers (Supplementary Table 3, notably, two genes linked to fat metabolism or obesity. In the extreme tail of the distribution, we detected human-specific increased expression of MGAT1, which aides in metabolism of fatty acids to triglycerides (Yen et al., 2002), and the expression of which has been associated with excess retention of lipids (Lee et al., 2012). Additionally, we see that TBCA, a tubulin cofactor which assists in the folding of β-tubulin (Tian et al., 1996), has increased expression in humans. Given that reduced expression of TBCA through a heterozygous deletion has been associated with childhood obesity in humans (Glessner et al., 2010), it could be that the human-specific increase in TBCA expression assists in metabolism of a higher fat diet. However, in both cases, it is unclear if the increased expression in humans is an evolutionary shift in expression, helping to adapt to a more fat-rich diet, or if the increased

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression in humans is environmentally responding to a fat-rich diet. Expression level studies can only distinguish between these alternatives if the environmental conditions have been controlled between study objects, which may be hard or impossible to achieve when comparing humans to other mammals.

Another gene with a significant expression shift in humans is BCKDK. BCKDK inactivates the branched-chain ketoacid dehydrogenase (BCKD) complex, which catalyzes metabolism of branched-chain amino acids (BCAAs). Nonsense and frameshift mutations in BCKDK have recently been linked to low levels of BCAAs and a phenotype including autism and epilepsy (Novarino et al., 2012). The observed increased human BCKDK expression may be adaptive to slow the metabolism of BCAAs so they can be processed into neurotransmitters (Novarino et al., 2012).

3.2 Comparing results using the β model and species mean model

We compared our results for the expression shift tests to those reported in a side analysis by Perry *et al.* (2012) using the species mean model described by Bedford and Hartl (2008). The distributions of $LRT_{\theta_i^{cat} \neq \theta_i^{non-cat}}$ and $LRT_{\theta_i^{hum} \neq \theta_i^{non-hum}}$ from that analysis deviate substantially from the χ_1^2 distribution expected under the null hypothesis (Supplementary Figure 19). This could be due to a number of possible numerical, optimization, or book-keeping errors, as these methods require a number of important technical considerations. In a comparison of the rank of expression shift test statistics as computed by Perry *et al.* (2012) and as computed using the β model, we see a general lack of correlation with some similarity in the extreme outliers discussed in that paper (Supplementary Figure 20).

As the species mean model is a reduction of the β model where each species has a single "individual" which is set to the species mean and β is fixed at zero, we implemented the species mean model to test for expression shifts on the catarrhine and human lineages. In our implementation, we see that the empirical distribution of test statistics are approximately χ_1^2 distributed with some excess of high values (Supplementary Figure 21) and a much improved correlation to β model test statistics (Figure 4). While both models identify similar genes with branch-specific θ_i shifts, we see much higher correlation between models for a shift on the catarrhine lineage than on the human lineage (Figure 4). Since the species mean model discards individual expression levels in favor of the species mean, it may identify genes where the mean expression appears to have shifted, even if the degree of variance may make that shift seem less extreme. By the same token, the β method may identify genes with a shift that can not be explained by the expected within-species variance. This difference is most pronounced when considering shift of a single species (such as humans) where considering variance within that single species may alter the perception of an expression shift.

Figure 5 shows the three genes with the biggest difference in value of $LRT_{\theta_i^{hum} \neq \theta_i^{non-hum}}$ between the β and species mean models, that is, the genes that are most clearly identified by one model, while missed by the other. The gene TBCA, discussed above as a candidate for diet-associated expression adaptation, is a clear outlier under the β model ($LRT_{\theta_{TBCA}^{hum} \neq \theta_{TBCA}^{non-hum}} = 9.5$), but is less easily identified using the species mean model ($LRT_{\theta_{TBCA}^{hum} \neq \theta_{TBCA}^{non-hum}} = 5.5$).

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Discussion

We have described the β model for gene expression evolution which parameterizes the ratio between population and evolutionary variance as β so that, in addition to more classic tests for selection on gene expression level, diversity to divergence tests may be formulated. We have explored a test for gene-specific β_i, showing that the null distribution of the test statistic $LRT_{\beta_i \neq \beta_{shared}}$ is asymptotically χ^2_1, though depending on the size of the dataset and the value of the parameters, the null distribution may not have converged to the asymptote. We show that in these cases, a parametric bootstrap approach can be used to more accurately assess the significance of $LRT_{\beta_i \neq \beta_{shared}}$ values. Since the parametric bootstrap may be sensitive to variance in parameter estimates, it is prudent to verify its effectiveness on a particular dataset with simulations before using it to interpret data.

The test for gene-specific β_i can be thought of as a phylogenetic ANOVA, or as a gene expression analog to the HKA or MK tests. This enables a previously unavailable line of inquery into gene expression divergence, which may be indicative of expression-level adaptation, and gene expression diversity, which may be indicative of plastic expression levels responding to environmental conditions. The utility in the comparative method in this setting is that it allows us to distinguish between genes which have high variance in expression levels simply because expression of this gene has little effect on fitness, and genes in which expression varies because the gene mediates a plastic response to the environment.

In applying the gene specific β_i test to a mammalian dataset, we identified several candidates for expression level adaptation, most notably high expression of F10 in armadillos, which may be linked to their phenotype of rapid blood coagulation. We additionally identified several candidate genes for environmentally-responsive expression levels including PPIB, which helps regulate immunosuppression, and HSPA8, a heat shock protein.

In addition to this novel test for unusual population or evolutionary variance, we used the β model to test for branch-specific shifts in expression level, as had been done previously with the species mean model. We found an increase in DEXI expression in catarrhines, which may have an adaptive role in auto-immune regulation. In humans, we found increased expression of two genes thought to be involved in lipid metabolism (MGAT1 and TBCA) and of BCKDK, the low expression of which has been linked to BCAA (necessary for neurotransmitters) deficiency, epilepsy, and autism.

When comparing our lineage-specific expression shift results to those previously reported using the species mean model, we observed startling differences. We've attributed these differences to a numerical or optimization problem in that original analysis, highlighting the importance of carefully addressing these issues. We performed an additional analysis using the species mean model to create a fair comparison. From that secondary analysis, we observe important differences between the β model and species mean model, most notably when testing for a shift in a single species. By discarding population variance, the species mean model may mistake a mild expression shift attributable to expected within-species variance for an evolutionary shift. We see this illustrated by the identification of an expression shift in humans for TBCA using the β model, but not using the species mean model.

By changing parameter constraints, the β model can be used to test a variety of hypotheses. For example, branch-specific β values, which may be expected under branch-specific tightening

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or relaxation of constraint, or under unusual branch-specific demographic processes. The β model could also be used to test hypotheses of gene class-specific (rather than gene-specific) β values, which may vary based on gene class function. For example, genes involved in stress response may have a higher β value than housekeeping genes.

Like all comparative expression methods, the β method applies to any heritable quantitative trait with environmental components, including metabolomics (Nicholson and Lindon, 2008; Cui et al., 2008; Sreekumar et al., 2009) and genome-wide methylation (Pokholok et al., 2005; Pomraning et al., 2009).

As larger expression and other quantitative trait comparative datasets emerge, the versatile β model and framework described here will facilitate a wide variety of sophisticated analyses.

5 Methods

5.1 Data likelihood under β model

The β model is similar to other OU process based evolutionary models (Butler and King, 2004; Bedford and Hartl, 2008), with the addition of within-species variance in terms of the evolutionary variance. As such, under the β model, expression levels across individuals and species follow a multivariate normal distribution identical to those under species means models at the species level as

$$\begin{aligned} E(Y_i) &= E(Y_p)e^{-\alpha_i t_{ip}} + \theta_i(1 - e^{-\alpha_i t_{ip}}) \\ Var(Y_i) &= \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_{ip}}) + Var(Y_p)e^{-2\alpha_i t_{ip}} \\ Cov(Y_i, Y_j) &= Var(Y_a)exp(-\sum_{k \in l_{ij}} \alpha_k t_k - \sum_{k \in l_{ji}} \alpha_k t_k) \end{aligned}$$

where Y_i is the expression level in species i, Y_p is the species mean expression at internal parental node p, θ_i, σ_i^2, and α_i are the parameter values on the branch leading to node i, t_{ip} is the length of the branch between i and p, Y_a is the expression level at the most recent common ancestor of species i and j, and l_{ij} is the set of nodes in the lineage of Y_i not in the lineage of Y_j (Rohlfs et al., 2014).

This multivariate normal distribution describing the species-level expression is augmented in the β model to include individuals within species, so for an individual k in species i, $Y_{ik} \sim N(Y_i, \beta_i \frac{\sigma_i^2}{2\alpha_i})$. In this way, the within species variance parameter described by Rohlfs *et al.* (Rohlfs et al., 2014) τ^2 is re-parameterized as $\beta_i \frac{\sigma_i^2}{2\alpha_i}$. The entire multivariate normal distribution can be described as

$$\begin{aligned} E(Y_{ik}) &= E(Y_i) \\ Var(Y_{ik}) &= Var(Y_i) + \beta_i \frac{\sigma_i^2}{2\alpha_i} \\ Cov(Y_{ik}, Y_{il}) &= Var(Y_i) \\ Cov(Y_{ik}, Y_{jl}) &= Cov(Y_i, Y_j) \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $i \neq j$ and $k \neq l$. With the distribution of expression levels under a particular set of parameters defined according to this multivariate normal, the likelihood of the data under the model is simply the probability density. Notice that sampling and experimental variance is accounted for (and confounded) in the parameters governing the distribution of $Y_{ik}|Y_i$.

5.2 Maximum likelihood procedures

For the test for individual gene departures from β_{shared}, under the null hypothesis each gene i is governed by parameters θ_i, σ_i^2, and α_i, reflecting the evolutionary process of each gene based on its degree of stabilizing selection and drift. The population variance in all genes is controlled by the single parameter β_{shared}. To more computationally efficiently maximize likelihood over these $3n+1$ parameters, we use a nested structure with Brent's method in the outer loop to maximize over the single parameter β_{shared}, and the BFGS algorithm in the inner loop to optimize over θ_i, σ_i^2, and α_i for each gene. Under the alternative hypothesis, the likelihood of each gene i is maximized using the BFGS algorithm over θ_i, σ_i^2, α_i, and β_i. Then the likelihoods of each individual gene under the null and alternative hypotheses are compared to compute the LR.

Note that in the likelihood maximization under the null hypothesis, likelihoods across genes are assumed to be independent so that for a particular value of β_{shared} the likelihood of a set of genes is simply the product of the likelihoods of each gene. While this assumption is currently typical in this sort of analysis, it leaves something to be desired since the evolution of expression levels of inter-related genes are not independent, and nor are the particular expression levels measured in an individual which may be responding to the environment of that individual. A more rigorous approach would take into account complex correlation structures across genes, as has been outlined for some evolutionary models (Lande and Arnold, 1983; Felsenstein, 1985, 1988; Lynch, 1991).

For the test of branch-specific expression shift for a particular gene i, under the null hypothesis the likelihood of each gene i is maximized over θ_i, σ_i^2, α_i, and β_i using the BFGS algorithm. Under the alternative hypothesis, the likelihood of each gene i is maximized over θ_i^a, θ_i^{non-a}, σ_i^2, α_i, and β_i, again, using the BFGS algorithm.

Acknowledgments

We are immensely grateful to the individuals whose RNA samples were used in this study, without which none of this work would be possible. We thank George Perry and colleagues for making their data and results available and for assisting us in their interpretation, Youna Hu and Josh Schraiber and Tyler Linderoth for their valuable discussions on these topics, Alex Safron for his help drawing schematics. This work was supported in part by National Institutes of Health grant 2R14003229-07, and National Science Foundation award 1103767. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Disclosure Declaration

The authors have no conflicts of interest to declare.

Figure Legends

Figure 1. The maximum likelihood estimated per-gene evolutionary variance ($\frac{\hat{\sigma_i^2}}{2\hat{\alpha_i}}$) and population variance ($\hat{\beta}_i \frac{\hat{\sigma_i^2}}{2\hat{\alpha_i}}$) are plotted against each other. The linear regression line is shown.

Figure 2. The test for a gene with β_i varying from $\hat{\beta}_{shared}$ was computed for each gene. Those LR test statistics ($LRT_{\beta_i \neq \beta_{shared}}$) are plotted against the log of the β parameter estimated for each gene ($log(\hat{\beta}_i)$) in a volcano plot. The dashed line indicates the value of $\hat{\beta}_{shared}$.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Each plot shows the expression profile across the 15 species for a genes in the extreme tails of the empirical distribution of the test statistic for a gene-specific β differing from β_{shared} ($LRT_{\beta_i \neq \beta_{shared}}$). (a) shows genes with high $\hat{\beta}_i$ values and (b) shows genes with low $\hat{\beta}_i$ values.

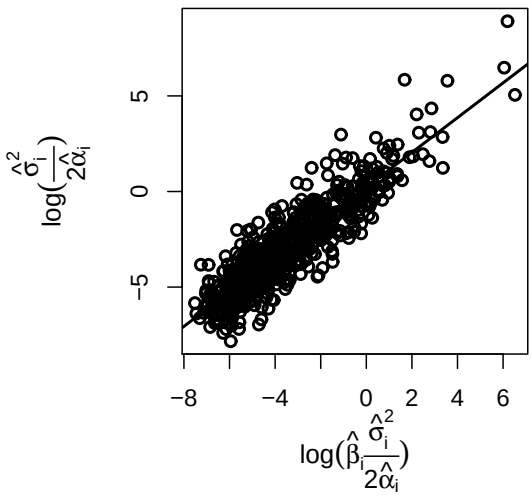
Figure 4. Each plot shows (a) $LRT_{\theta_i^{cat} \neq \theta_i^{non-cat}}$ and (b) $LRT_{\theta_i^{hum} \neq \theta_i^{non-hum}}$ calculated using the β model (y-axes) and species mean model (x-axes) as implemented in this analysis. The line indicates $x = y$.

Figure 5. Each plot shows the expression profile for genes identified with an expression shift in humans by the β model, but not by the species mean (SM) model (top row), and identified by the species mean model, but not by the β model (bottom row). Expression levels in humans are highlighted in pink. Each plot shows $LRT_{\theta_i^{hum} \neq \theta_i^{non-hum}}$ (as LRT) as computed under the β and species mean models.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

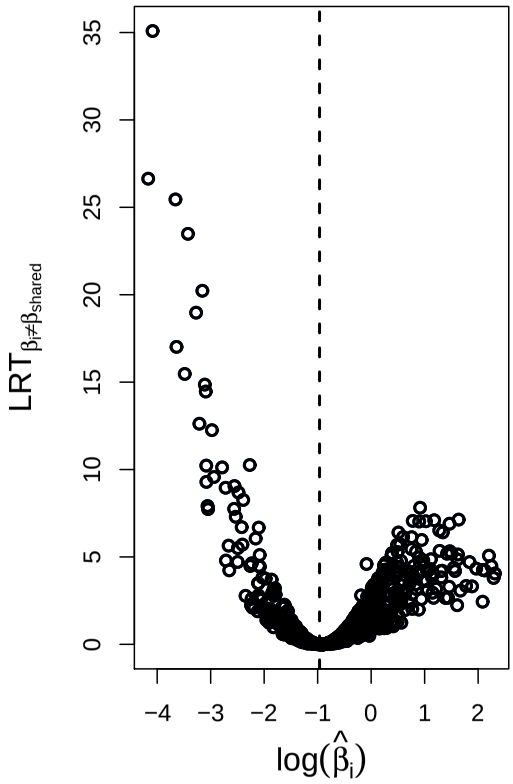
Figures

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

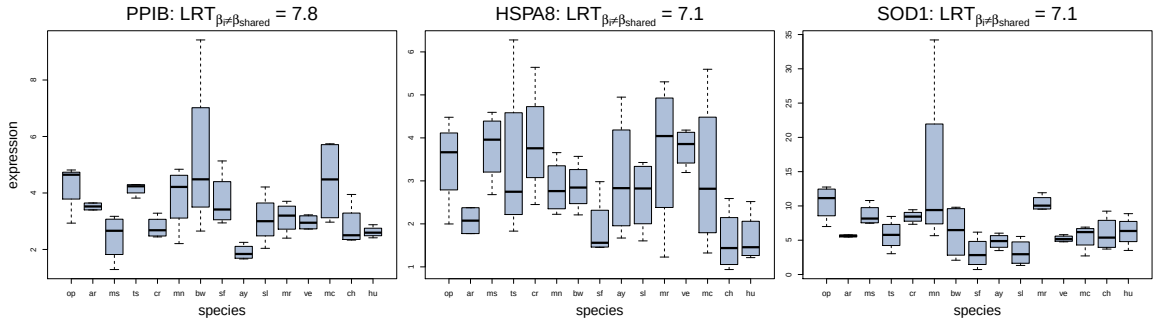
Figure 2



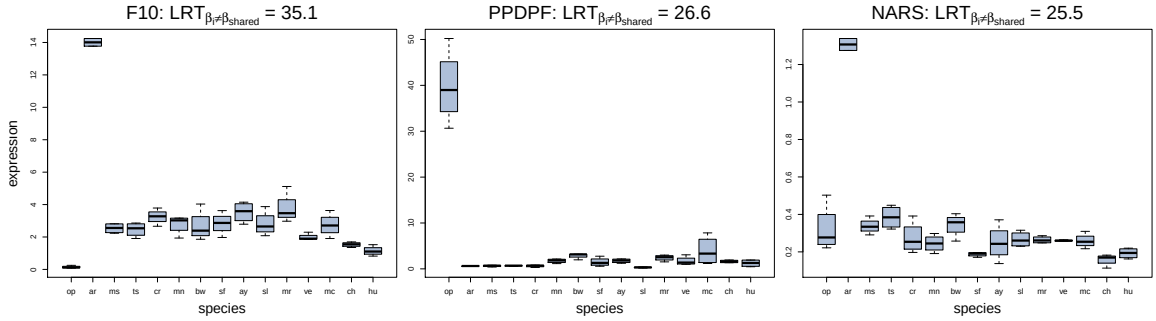
bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3

a



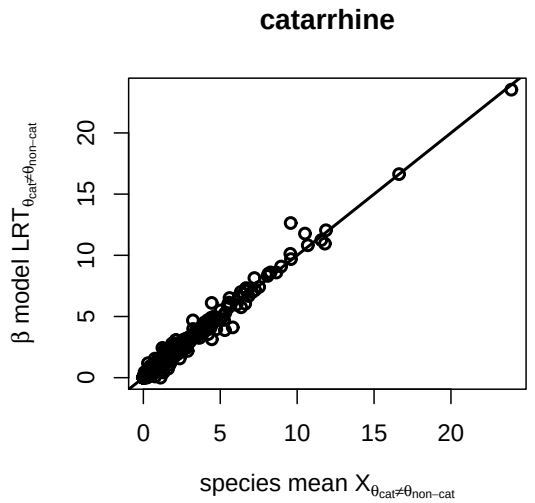
b



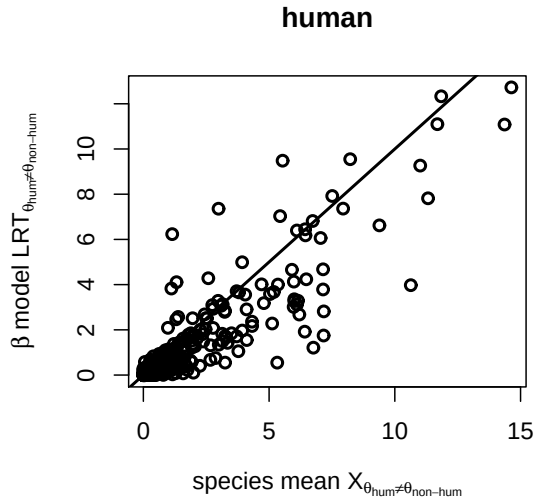
bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4

a

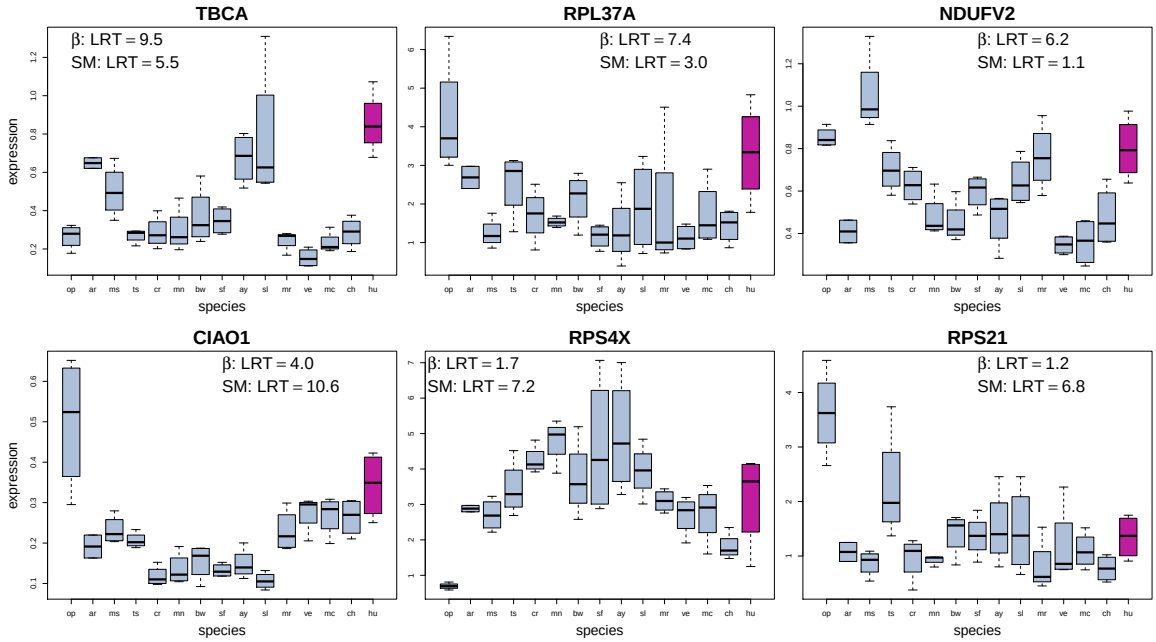


b



bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5



bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Bedford T and Hartl D. 2008. Optimization of gene expression by natural selection. *Proc Natl Acad Scie USA* **106**: 1133–1138.

Boettiger C, Coop G, and Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. *Evolution* **66**: 2240–2251.

Brawand D, Soumillon M, Necsulea A, Julien P, Csrdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al.. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Bullard J, Purdom E, Hansen K, and Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.

Butler M and King A. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *American Naturalist* **164**: 683–695.

Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, and Markley JL. 2008. Metabolite identification via the madison metabolomics consortium database. *Nature Biotechnology* **26**: 162–164.

Daugaard M, Rohde M, and Jttel M. 2007. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. {*FEBS*} *Letters* **581**: 3702 – 3710.

Davison LJ, Wallace C, Cooper JD, Cope NF, Wilson NK, Smyth DJ, Howson JM, Saleh N, Al-Jeffery A, Angus KL, et al.. 2012. Long-range dna looping and gene expression analyses identify dexi as an autoimmune disease candidate gene. *Human Molecular Genetics* **21**: 322–333.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* **125**: 1–15.

Felsenstein J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19**: 445–471.

Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist* **171**: 713–725.

Gilad Y, Oshlack A, Smyth G, Speed T, and White K. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**: 242–245.

Glessner JT, Bradfield JP, Wang K, Takahashi N, Zhang H, Sleiman PM, Mentch FD, Kim CE, Hou C, Thomas KA, et al.. 2010. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *American Journal of Human Genetics* **87**: 661–666.

Hansen T. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**: 1341–1351.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hansen T and Bartoszek K. 2012. Interpreting the evolutionary regressions: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* **61**: 413–425.

Hansen T, Pienaar J, and Orzack S. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* **62**: 1965–1977.

Hudson R, Kreitman M, and Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.

Kalinka A, Varga K, Gerrard D, Preibisch S, Corcoran D, Jarrells J, Ohler U, Bergman C, and Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**: 811–816.

Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, and Pääbo S. 2006. Functionality of intergenic transcription: An evolutionary comparison. *PLoS Genetics* **2**: e171.

Lande R and Arnold SJ. 1983. Measurement of selection on correlated characters. *Evolution* **37**: 1210–1226.

Lee Y, Ko E, Kim J, Kim E, Lee H, Choi H, Yu J, Kim H, Seong J, Kim K, et al.. 2012. Nuclear receptor ppar-regulated monoacylglycerol o-acyltransferase 1 (mgat1) expression is responsible for the lipid accumulation in diet-induced hepatic steatosis. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 13656–13661.

Lewis JH and Doyle AP. 1964. Coagulation, protein and cellular studies on armadillo blood. *Comparative Biochemistry and Physiology* **12**: 61 – 66.

Li H and Durbin R. 2009. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics* **25**: 1754–1760.

Luban J, Bossolt KL, Franke EK, Kalpana GV, and Goff SP. 1993. Human immunodeficiency virus type 1 gag protein binds to cyclophilins a and b. *Cell* **73**: 1067 – 1078.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45**: 1065–1080.

McDonald J and Kreitman M. 1991. Adaptive protein evolution at the *adh* in *drosophila*. *Nature* **351**: 652–654.

Nicholson J and Lindon J. 2008. Systems biology: Metabolomics. *Nature* **455**: 1054–1065.

Novarino G, El-Fishawy P, Kayserili H, Meguid NA, Scott EM, Schroth J, Silhavy JL, Kara M, Khalil RO, Ben-Omran T, et al.. 2012. Mutations in bckd-kinase lead to a potentially treatable form of autism with epilepsy. *Science* **338**: 394–397.

Nuzhdin S, Wayne M, Harmon K, and McIntyre L. 2004. Common pattern of evolution of gene expression level and protein sequence in *drosophila*. *Molecular Biology and Evolution* **21**: 1308–1317.

bioRxiv preprint doi: https://doi.org/10.1101/004374; this version posted April 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Perry G, Melsted P, Marioni J, Wang Y, Bainer R, Pickrell J, Michelini K, Zehr S, Yoder A, Stephens M, et al.. 2012. Comparative RNA sequencing reveals substantial genetic variation in endagered primates. *Genome Research* **22**: 602–610.

Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, and Pritchard J. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–772.

Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, et al.. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**: 517–527.

Pomraning KR, Smith KM, and Freitag M. 2009. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**: 142–150.

Price ER, Zydowsky LD, Jin MJ, Baker CH, McKeon FD, and Walsh CT. 1991. Human cyclophilin b: a second cyclophilin gene encodes a peptidyl-prolyl isomerase with a signal sequence. *Proceedings of the National Academy of Sciences* **88**: 1903–1907.

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

Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, et al.. 2009. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**: 910–914.

Tian G, Huang Y, Rommelaere H, Vandekerckhove J, Ampe C, and Cowan NJ. 1996. Pathway leading to correctly folded β-tubulin. *Cell* **86**: 287 – 296.

Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M, Salzberg S, and Wold BJ Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511–515.

Uprichard J and Perry DJ. 2002. Factor x deficiency. *Blood Reviews* **16**: 97 – 110.

Wang Z, M G, and M S. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63.

Whitehead A and Crawford D. 2006. Variation within and among species in gene expression: raw material for evolution. *Molecular Ecology* **15**: 1197–1211.

Yen CLE, Stone SJ, Cases S, Zhou P, and Farese RV. 2002. Identification of a gene encoding mgat1, a monoacylglycerol acyltransferase. *Proceedings of the National Academy of Sciences* **99**: 8512–8517.