Evolutionary signatures for unearthing functional elements in the human transcriptome.

by

Jenny Chen

BS, Biomedical Computation, Stanford University (2010)
MS, Biomedical Informatics, Stanford University (2011)

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Health Sciences and Technology:
Bioinformatics and Integrative Genomics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author . . . .
Harvard-MIT Division of Health Sciences and Technology
April 30, 2018

Certified by . . . .
Aviv Regev, PhD
Professor, Department of Biology
Thesis Supervisor

Accepted by . . . .
Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology

Evolutionary signatures for unearthing functional elements in the human transcriptome.

by

Jenny Chen

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on April 30, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Health Sciences and Technology:
Bioinformatics and Integrative Genomics

Abstract

Comparative genomics is a powerful method for identifying functional genetic elements by their evolutionary patterns across species. However, current studies largely focus on analysis of genome sequences. The recent development of RNA-sequencing reveals dimensions of regulatory information previously inaccessible to us by sequence alone. The comparison of RNA-sequencing data across mammals has great potential for addressing two open problems in biology: identifying the regulatory mechanisms crucial to mammalian physiology, and deciphering how gene regulation contributes to the diversity of mammalian phenotypes.

For my thesis, I developed two methodologies for interrogating comparative transcriptomic data for biological inference. First, I developed a framework for quantifying the evolutionary forces acting on gene expression and inferring evolutionarily optimal expression levels. I demonstrate how to use this framework to identify expression pathways underlying conserved, adaptive, and disease states of mammalian biology. Second, I developed novel metrics of transcriptional evolution to evaluate the conservation of long noncoding RNAs. These metrics further reveal that long noncoding RNAs harbor distinct evolutionary signatures, suggesting that they are not a homogenous class of molecules but rather a mixture of multiple functional classes with distinct biological roles.

My thesis work provides fundamental quantitative tools for asking biological questions about transcriptome evolution. These tools provide a pivotal framework for interpreting transcriptional data across species and pave the way for deciphering the regulatory changes that lead to mammalian phenotypic variation.

Thesis Supervisor: Aviv Regev, PhD
Title: Professor, Department of Biology

To my parents, who left behind everything they knew to slog through graduate school so that I could do it by choice.

I didn't say it would be easy, Neo. I just said it would be the truth.

Morpheus

Acknowledgments

Graduate school has been a truly incredible journey filled with excitement, frustration, humility, inspiration, and pride. But most of all, it has been a journey filled with gratitude for everyone who have supported me through it all. In particular, I'd like to thank:

Aviv Regev, for her unwavering faith in my naïve, half-baked first-year ideas, and for giving me the opportunity and resources to carry them out.

Manuel Garber, for teaching me how to turn those half-baked ideas into scientific knowledge; and for all of his non-scientific advice that have been essential to my sanity, including his reminders to swim at Walden, to eat plenty of fruit, and to humidify my orchids.

Alex Shishkin, for being such an amazing biologist that I didn't even notice that my data came from samples that had spilled on the floor and been re-pipetted into the sequencer; and Mitch Guttman, for his masterly writing advice.

Kerstin Lindblad-Toh for making the mammals project possible; Federica di Palma, for supporting the project even from across the Atlantic; Wilfried Haerty, for his insightful comments and encyclopedic knowledge of relevant publications; and Beryl Cummings, for being so open to helping a random grad student on a whim of an idea.

The postdocs of the Regev Lab, for their encouragements to persist, especially Yarden Katz, for helping me find a path to the end of the tunnel; Jimmie Ye, for his many good pieces of advice on navigating academia; Marko Jovanovic, for opening my eyes to the motivating qualities of Arnold; and Carl de Boer for being an unlimited source of gum and guidance.

The graduate students of the Regev Lab – especially Dima Ter-Ovanesyan, Atray Dixit, Rebecca Herbst, and Christoph Muus – who, for a short while, were the only reasons I came into lab each day.

Biyu Li, for being a great after-lab Asian food partner; Noga Rogel, for her willingness to always help me over the finish line; Quinn Sievers, for being my constant sounding board; Akshay Krishnamurthy for his free statistician consults; Lily Xu, for teaching me tenacity; and David Feldman, Anthony Garrity, Jacob de Riba Borrajo, Nav Ranu, and Georgia Lagoudas, for being a wonderful second-floor step-family to grab a coffee or beer with.

My BIG family, for being an excellent group of genomicists to nerd out with, especially Melissa Gymrek, for being my other BIG half in the entering class of 2011; Mike Rooney, for all the cookie breaks; Kendell Clement, for passing along his wit and experience; and Jesse Engreitz, for unwittingly being a great scout for my career path.

My HST family, for being a bedrock of support from the beginning, especially Kelli Xu, whose opinion I can always trust to be right; Andrew Warren, for his willingness to step in with Kelli-like opinions when needed; Vyas Ramanan, for the puns; Nikhil Vadhavkar, for teaching me how to break up undergrad parties; Adam Pan, for being a walking wikipedia; and Ronn Friedlander, for being my 'old' grad student friend and hosting the best parties.

Kiran Musunuru, Anne Giersch, and David Housman for exemplifying how to teach.

John Fernandez and Malvina Lampietti and the Baker house team – including Kristen Covino, Dave Whittleston, Pooya Molavi, Carmen Castaños, Eric Benzschawel, and Rachel Hoffman – who created a place that truly felt like home.

Chris Link, for letting me crash at his place during HST interviews, and for putting up with me ever since.

My dad, for reminding me whenever it felt too overwhelming that "the fights are so bitter because the stakes are so low"; My mom, for her unfettered enthusiasm for everything that I do, big or small; Jackie, for her infectious optimism and joy for life.

All of my academic mentors that predate graduate school and instilled in me a curiosity for the world, especially Ms. Hester; Mr. Kucer; Russ Altman and Daphne Koller for being ahead of the times in creating a computational biology major; Gill Bejerano, for introducing me to comparative genomics; and Aaron Wenger for teaching me the fundamentals.

And finally, I'd like to thank the Wachowskis, Conan O'Brien, Nate Silver, the writers of *Fringe*, Karen O, Ted Chang, and Alan Moore for helping me through difficult times and inspiring me to stay true to myself.

Contents

List of Figures

List of Tables

Introduction

With the completion of the human genome in the early 2000s [1, 2], nearly each of the 3 billion letters that code for human life had been sequenced, but the untangling of how the genomic code operates had just begun. In the past two decades, scientists have embarked on monumental efforts to comprehend the genomic code with initiatives ranging from the cataloging of human genetic variation across thousands of individuals [3], to annotating all biochemically active parts of the human genome across hundreds of cellular contexts [4], to sequencing tens of thousands of patients along with medical data collection in order to understand the genetic impact on disease [5].

In addition to systematically collecting genomic data across the human population, another method that has proven powerful for illuminating the inner workings of the human genome has been comparative genomics, the strategy of comparing genomic data across species to infer biological function. The power of comparative genomics rests upon one of few “truth” known to biology: that the process of DNA replication is imperfect and introduces mutations into the population which then serve as substrates for evolution. Mutations that confer a selective disadvantage are eventually removed from the population while the rare mutations that increase species fitness give rise to novel phenotypes. This constant process of mutation and selection leaves signatures in genomic sequences from which we may infer what roles those sequences play in biology: sequence that has remained constant across millions of years of evolution is inferred to be evolutionary constrained and biologically important, while sequence that appears to have diverged rapidly are hypothesized to play roles in lineage-specific phenotypes.

Since the completion of the human genome, over 270 eukaryotic species have now been sequenced, including over 60 mammalian species [6]. The comparison of these genomes have led to the comprehensive annotation of nearly all mammalian coding genes [7, 8], as well as functional noncoding sequence elements such as enhancers [9–11], miRNA target sites [12,13], and noncoding genes [14–16]. Sequence conservation has also become widely used for interpreting clinical data and highlighting likely pathogenic mutations that may be causal for disease [17, 18]. Finally, comparative genomics has unlocked intriguing evolutionary stories lurking within the human genome such as the identification of 'ultraconserved' elements – stretches of DNA with perfect conservation across 300-400 million years of evolution thought to play essential roles in development [19, 20] – in addition to 'human accelerated regions' – segments of the genome conserved across all mammals except for human, and are thought to code for human-specific traits [21, 22]. (Additionally, comparative genomics has been widely applied across the animal kingdom from plants, to insects, bacteria, and viruses, but here, I focus specifically on mammalian comparative genomics and its relationship to understanding the human genome).

These recent advancements within comparative genomics have been derived primarily from the analysis of DNA across species. Meanwhile, the field of genomics has expanded beyond simply genome sequencing. Combined with a tricks borrowed from molecular bioengineering, next-generation sequencing can now be used to profile a myriad of attributes of the genome including its three-dimensional structure [23], epigenetic modifications [24], and the activity of DNA-binding proteins [25]. Of particular note is the ability to now unbiasedly sequence all RNA molecules produced in a biological sample (**'RNA-sequencing'**, or **RNA-seq**), made possible by the reverse transcription of RNA to cDNA [26–28]. RNA-seq not only reveals the quantity of RNA present in the cell but also transcript structures (e.g., exon/intron boundaries, 5' and 3' gene boundaries), alternative isoform variants, and the transcription of noncoding RNA molecules.

RNA-sequencing opens the door for scientists to investigate the first step of information transfer in the central dogma of molecular biology: how DNA regulates the transcriptional timing, quantity, and structure of RNA. This fundamental process, termed 'gene regulation', governs essentially all cellular activity from the development of an egg to a fetus to the proper control of hormones and metabolites, and yet, very little is understood about how this process is coded for by the genome. The advent of RNA-seq has now enabled scientists to directly profile gene regulation in action, from cataloging expression pathways altered under environmental stresses or disease physiology, to usage of alternative splicing in different cellular contexts, to the detection of novel regulatory RNA transcripts, and much more (see [29] for a comprehensive review on RNA-sequencing studies). Further, the pairing of both genomic and transcriptomic data enables scientists to characterize the noncoding DNA that control these regulatory processes in an effort to elucidate a 'noncoding' code.

Additionally, as a major advancement over its predecessor for RNA profiling – microarray technology – RNA-seq can be applied in an unbiased manner to any species of interest. (Microarray technology is probe-based and requires previous knowledge of the transcriptional sequence composition to measure quantities.) This generalizability of RNA-seq across species paves the way for efficient comparative transcriptomic studies that can address questions that cannot be answered by comparative sequence analysis alone. For example, while comparative sequence analysis can identify proteins and even protein domains that are fundamental to an animal system, comparative transcriptomics now enables the identification of essential co-expression modules, splicing events, and noncoding RNAs as well as the specific cellular contexts (e.g., in which tissue type, at which developmental time) in which these events are most necessary. Such studies would be highly informative for (1) augmenting our understanding of every gene in the human genome, (2) increasing the ability for medical community to identify causal processes behind diseases arising from gene misregulation (e.g., many cancers, autoimmune, and developmental disorders [30]), and (3) informing drug discovery efforts the almost always begin with experimentation on animal models [31].

On the other end, mapping the diversity of transcriptional mechanisms across species would lead to an increased understanding of how different transcriptional processes give rise to new phenotypic traits. This question is especially of interest given the long-standing hypothesis that evolving gene regulation plays a predominant role in creating new traits between closely related species [32]. The hypothesis began to gain traction following one of the earliest comparative genomics studies (that predates even Sanger sequencing and instead relies on electrophoresis techniques), which discovered that human and chimp proteins are more than 99% identical in amino acid similarity [33]. The fact that coding gene conservation is quite high, even across mammals, has now been confirmed with high-throughput sequencing studies [34–36], further motivating scientists to search the noncoding genome for the basis of novel phenotypic innovations. Again, the use of comparative transcriptomics to discover the processes underlying species-specific traits would be highly complementary to comparative sequence efforts that have indeed led to the discovery of a handful of regulatory genes controlling intriguing phenotypes [20, 37–39], but have yet lead to a complete understanding of how and to what extent gene regulation controls major phenotypic differences. Given that noncoding sequences are notoriously difficult to interpret with the current state of knowledge and technologies, comparative RNA-seq studies offer an important intermediate phenotype for mapping the relationship between gene regulation and biological traits.

Despite the many promises of knowledge to be gained by comparative transcriptomics, the field remains young and only a few large-scale studies conducting RNA-seq across mammalian species have been carried out thus far. Still, these initial studies have begun to uncover illuminating, and sometimes unexpected, global patterns of transcriptional evolution that begin to sketch out how regulatory changes, both across species and across tissues, may be contributing to phenotypic evolution. Of the handful of studies that have been completed, three general classes of analyses have emerged: expression level, alternative splicing, and long noncoding RNA analysis. I will describe the current state of the literature for each class below.

Comparative expression analysis

Differential expression analysis is one of the major areas of investigation for understanding the molecular underpinnings of physiological processes. The levels of mRNA present in a cell, which is used as a proxy for protein levels, reveals the major molecular players in different cell types, across developmental timing, or in response to environmental stimuli. Comparative expression analysis studying expression across species began with the introduction of microarray technology in the early 2000s, initially focused on human, mouse, and close human relatives (e.g., chimpanzees, gorillas, orangutans). Because microarrays require species-specific sequence probes for hybridization, these analyses were hampered by difficulties overcoming species-specific experimental errors as well as small numbers of sampled species. In fact, many of the initial comparative microarray studies contain numerous conflicting results regarding questions such as: whether orthologous tissues across species are more similar than nonorthologous tissues within a species (e.g., human liver and mouse liver vs. human liver and human brain) [40–43]; whether expression in human tissues was diverging at an accelerated rate compared to other non-human primates [44–47]; and whether expression was largely evolving by neutral drift or under strong stabilizing pressures [43, 48–51].

The introduction of RNA-seq allowed scientists to profile many species efficiently, without the need for species-specific probes. RNA-seq was quickly carried out in mouse and human [52], as well as across 11 nonhuman primates and 5 additional mammals [53, 54]. With a better resolution of gene expression levels across many more species, the field began to converge on a few observations: transcriptomes of homologous tissues across species are typically more similar than nonorthologous tissues within a species [53, 55]; expression in human tissue does not appear to be diverging at an accelerated rate but rather, is consistent with divergence of other non-human primates [47]; and expression evolution appears to be shaped strongly by stabilizing pressures (discussed in [43, 49, 53] and in depth in Chapter 2).

Furthermore, comparative transcriptomic profiling across different organ systems allows for the analysis of tissue-specific evolutionary pressures: Although transcriptomes of homologous tissues across vertebrates appear to diverge slowly, the rate of divergence varies by tissue. Interestingly, it has been consistently observed that brain expression levels diverges the slowest, despite apparently large cognitive differences across vertebrates, and testis expression diverges most rapidly, perhaps partially due to sexual selection [43, 46, 53].

Finally, a major goal in genomics is to map the relationship between noncoding sequence and its control of expression levels. However, initial analyses have been unable to find strong correlation between expression divergence and noncoding sequence conservation in either proximal promoters [56] or within distal regions of a gene [43]. These results suggest that too many compensatory mutations may be at play to be able to accurately compare regulatory sequences across the profiled species, and point to the need for data from more closely related species to dissect the relationship between noncoding sequence and expression.

Comparative splicing analysis

Given that expression levels across species appear to be highly conserved, alternative splicing offers another avenue of explanation for species-specific differences. Alternative splicing can greatly expand the complexity of protein products available and generate isforms with specific regulatory information in 5' or 3' untranslated regions [57]. Early comparative splicing analysis began like expression profiling, with hybridization arrays applied to small numbers of species. These studies found alternative isoform usage tended to be highly species-specific [58–60], raising interest in characterizing alternative splicing events across species. Additionally, though there inconsistent reports on the exact amount of alternative splicing events within each species, the general trend appears to be that vertebrates utilize alternative splicing much more than invertebrates do, potentially driven by the increase in number of distinct cell types between the two phyla [61–63]. This further suggests important regulatory roles for alternative splicing and marks the importance of investigating its contribution to phenotype.

To date, there have been a handful of RNA-seq studies on alternative splicing variation across primates [54, 64], and vertebrates [62, 65]. These studies have confirmed that there is indeed significant amounts of species-specific alternative splicing events (∼90% similarity in exon usage between humans and chimps [60, 64]; ∼50% similarity between human and mouse [62]). Additionally, both large-scale vertebrate studies showed that alternative splicing events are more similar across nonorthologous tissues within a species rather than across homologous tissues from different species, in marked contrast to comparisons of expression level across species and tissues. However, tissue-specific evolutionary pressures on alternative splicing echo those observed in comparative expression analyses, with brain also appearing to have maintained the most conservation in splicing and testis diverging most rapidly.

Finally, when integrated with genomic sequence data, there have been promising results with identifying correlations of splicing conservation and nearby sequence conservation: strong signatures of purifying selection on intronic splicing regulatory element (ISRE) motifs nearby conserved, alternative exons have been reported, while exons that recently converted from alternative to constitutive have also been shown to have substantially increased turnover of ISREs [65]. These global patterns suggest that there is potential for precisely delineating out the evolution of splice site sequences and the resulting isoform products. Still, very few splicing events underlying novel phenotypic traits are known [66,67], and it is yet to be determined whether the diversity of alternative splicing events across vertebrates play major roles in biological function.

Comparative long noncoding RNA analysis

While the contributions of differential expression and alternative splicing to gene regulation were well-known from the beginnings of molecular biology, another source of regulatory complexity was only be fully uncovered by genome-wide profiling techniques: long noncoding RNAs (lncRNAs). LncRNAs are defined to be non-protein coding genes longer than 200 basepairs. Though a few lncRNAs had been discovered in the pre-genomic era, the pervasive transcription of lncRNAs were first noted in a systematic analysis of transcripts from dense oligonucleotide arrays [68–70], and later confirmed to be transcribed across mammalian genome by sequencing analysis [71, 72]. At least one lncRNA, *XIST*, is quite famous for its role in X-inactivation [73], but the functions of the vast majority of observed lncRNAs are unknown. Comparative transcriptome analyses are playing an important part in the current quest to understand the function of lncRNAs.

A central question about lncRNAs since their discovery is whether these RNA molecules play important biological functions or whether they are simply the result of noisy biological processes [74–76]. Initial analysis of the primary sequence of lncRNAs, in which sequence conservation of lncRNA exons were compared to lncRNA introns and random intergenic regions, suggested that as a class, lncRNAs are modestly conserved [76–79]. However, these initial analyses assumed that lncRNA transcription was common across the compared species. While this assumption is almost always valid in analyzing coding genes, systematic RNA-seq of mammalian species has revealed that in fact, the transcription of lncRNA is often not conserved, despite conservation of underlying sequence (discussed in [80–82] and at length in Chapter 3). Additionally, for the smaller set of lncRNAs that are conserved in transcription, studies of RNA sequence co-evolution have been applied in attempts to help identify conserved secondary structures, but have given way to conflicting results [83,84]. In sum, comparative lncRNA studies have raised questions on whether the majority of lncRNAs truly have function as RNA molecules. However, it should not be overlooked that these studies have also identified a small number of lncRNAs that are both conserved in transcription and show purifying selection on their primary sequence, highlighting useful candidates for downstream experimental analysis.

Comparative transcriptomics remains a young field with many questions and conflicting results that await to be resolved. The limited number of comparative RNA-seq studies is, in part, due to difficulties with tissue collection, especially for protected species such as non-human primates where tissues can only be harvested after natural death. Tenuous results from sparse datasets have been further compounded by inconsistent genome quality across mammalian species that introduce species-specific biases to read mapping, isoform reconstruction, and quantification of transcriptional traits.

While the amount and quality of RNA-seq data and genomic annotations is only increasing with time and will soon become an irrelevant experimental barrier, a final major obstacle for comparative transcriptional studies remains: the lack of consensus on the correct models and methodologies for deriving statistically rigorous conclusions from comparative RNA-seq data. This obstacle largely arises from our lack of understanding of the molecular mechanisms of transcriptional evolution. For example in comparative sequence analysis, where DNA mutational rates are well known, a precise probability of sequence identity between species can be calculated and used as a quantitative measure of purifying selection. In contrast, for metrics of transcriptional conservation (e.g., correlation of expression levels, percent common exon usage, etc.), it is much more difficult to interpret which measurements are simply indicative of neutral evolutionary processes and which represent selective pressures. As a case in point, the observation that the transcription of most lncRNAs are species-specific has led to both interpretations that lncRNAs may be playing major roles in species-specific biology and that lncRNAs must simply be transcriptional noise [74, 75].

For my thesis, I addressed this obstacle and developed statistically rigorous methods to analyze comparative transcriptomic data in order to make inferences about the evolution and function of transcriptional processes. In Chapter 2, I discuss a stochastic framework for analyzing comparative expression data and inferring evolutionary optimal expression distributions. This statistical quantification then enables functional hypotheses about the role of that gene's expression (or mis-expression) in specific tissue contexts. In Chapter 3, I present novel statistical metrics for evaluating the evolution of lncRNA transcriptional structure (e.g., isoform structures, splice sites). I show that these evolutionary metrics reveal that lncRNAs are actually a heterogeneous class of molecules composed of distinct functional classes, each presenting with unique evolutionary signatures. Finally, I present a short story in Chapter 4 that is not a cross-species comparison, but instead, a comprehensive comparison across the human population of alternative splicing in immune cells after viral stimulation. This story highlights how characterizing differential isoform usage with RNA-seq helps solve an evolutionary mystery of why a haplotype linked to Crohn's disease appears to be selected for in the human population.

With these stories, I hope to show that evolutionary signatures left in the human genome give us important clues for understanding the history and function of genes, and that these signatures are not only found in genomic sequence but also in patterns of expression, splicing, and transcriptional evolution. Comparative genomics analyses has great potential as a powerful tool for not only teaching us fascinating stories about our prehistoric past, but also for unearthing the functional elements hidden across the vast human genome.

A quantitative framework for characterizing the evolutionary history of mammalian gene expression

Jenny Chen, Ross Swofford, Jeremy Johnson, Beryl B. Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica di Palma, and Aviv Regev

The evolutionary history of a gene helps to predict its function and relationship to phenotypic traits. While sequence conservation is commonly used to decipher gene function and assess medical relevance, methods for functional inferences from comparative expression data are lacking. Here, I use RNA-sequencing across 7 tissues from 17 mammalian species to show that expression evolution across mammals is accurately modeled by the Ornstein-Uhlenbeck process, a commonly proposed model of continuous trait evolution. I apply this model to identify expression pathways under neutral, stabilizing, and directional selection. I further demonstrate novel applications of this model to quantify the extent of stabilizing selection on a gene's expression, parameterize the distribution of each gene's optimal expression, and detect deleterious expression levels in expression data from individual patients. This work provides a statistical framework for interpreting expression data across species and in disease.

2.1 Background

Comparative genomics has identified and annotated functional genetic elements by their evolutionary patterns across species [9, 11, 21, 37, 85, 86]. Current comparative studies focus primarily on analysis of genomic sequences, using methods based on a well-established theoretical framework developed from observations that neutral sequence diverges linearly across time [87–91]. These methods allow for detection of sequence elements that evolve slower (e.g., due to purifying selection) or faster (e.g., due to positive selection or relaxed selective constraints) than expected under the null model of neutral evolution.

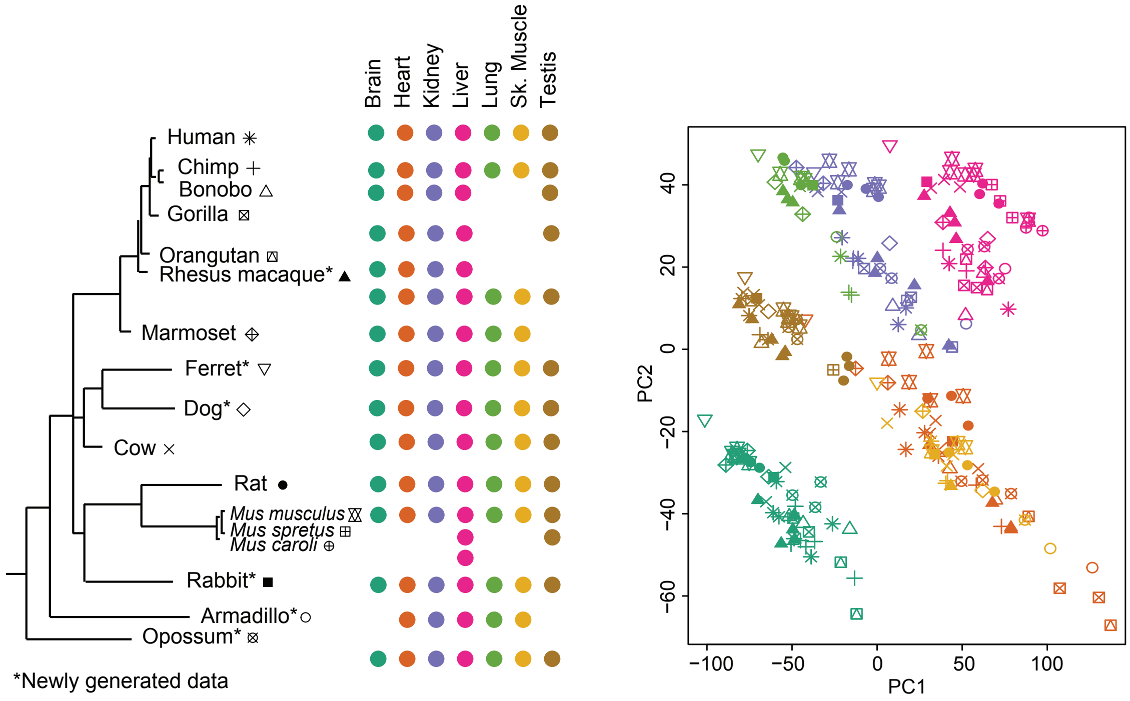
It has long been accepted that divergence of gene regulation, manifested by phenotypic changes in gene expression, also plays a key role in evolution [33,50,92–95]. An evolutionary analysis of gene expression should help interpret gene function and evolutionary processes in ways that cannot be addressed by sequence alone: the extent of stabilizing selection on a gene's expression level in different tissues could reveal the one(s) in which the gene plays the most important role; the strength of evolutionary constraint on a gene's expression level could help interpret expression levels observed in clinical samples; and genes whose expression level is under directional selection can help assess the basis of lineage- and species-specific phenotypes.

While multiple studies have analyzed expression data collected across mammalian species using various heuristic methods for defining conserved and divergent expression levels [43,53, 54,65], there remains no consensus on a quantitative framework for addressing the functional questions related to evolution of expression levels, due in part to a lack of agreement for how to best model expression evolution in mammals. In *Drosophila*, studies have found that unlike sequence evolution, divergence of gene expression levels is not continuously linear across evolutionary time. Instead, it reaches saturation due to stabilizing selective pressures, requiring more sophisticated models than standard neutral drift models [96, 97]. However, initial studies on an evolutionary model for mammalian gene expression have been hampered by small datasets, leading to inconsistent reports on whether the same evolutionary pattern is true within mammals [41, 48, 51, 53], and resulting in conflicting usages of both pure neutral drift models [54] and those that incorporate stabilizing selection [53] in comparative mammalian studies. Moreover, it has not been substantially explored how to use such models, once fit, to draw conclusions on gene function.

2.2 A model for mammalian expression evolution

To systematically explore expression evolution, I compiled a well-sampled dataset across the mammalian phylogeny, spanning 17 species and 7 different tissues (brain, heart, muscle, lung, kidney, liver, testis) (Fig. 2-1a, Methods, Additional file 1). The dataset combines published data for 12 species [53, 65, 98–101] with data for five additional species newly collected here (Fig. 2-1a, asterisks, Methods) to improve phylogenetic coverage. I focused on the 10,899 annotated mammalian one-to-one orthologs [102]. As previously reported [55], expression profiles first cluster by tissue and then by species (Fig. 2-1b), and their hierarchical clustering closely matches the phylogenetic tree (Fig. A-1).

On average, I find that pairwise expression differences between species (Fig. A-2, Methods) saturate with evolutionary time (Fig. 2-2, A-3, A-4), consistent with observed evolutionary trends in *Drosophila*. For example, when comparing each species to the human



(a) Phylogenetic tree of all 17 mammals (symbols, left) marked by tissue types (colored dots, right) for which profiles are included. Asterisk denotes newly generated data.

(b) First two principal components of a principal component analysis of expression profiles of all 230 RNA-seq samples across 17 species (symbols) and 7 tissue types (colors).

Figure 2-1: Overview of mammalian RNA-seq data.

profile, differences initially increasingly diverge with increasing evolutionary distance, but this trend plateaus beyond the primate lineage (43.2 million [M] years) [103] (Fig. 2-2). This relationship is observed in each of the five tissues for which I have expression data for all primates (brain, heart, kidney, liver, testis) (Fig. A-3), and regardless of reference species (Fig. A-4), albeit with varying rates.

2.2.1 Modeling expression evolution with an Ornstein-Uhlenbeck process

The observed pattern of expression divergence corresponds to an Ornstein-Uhlenbeck (OU) process (Fig. 2-3), a stochastic process initially proposed as a model for evolution of general continuous phenotypes by Hansen [104] and has more recently been suggested as an appropriate model specifically for the evolution of gene expression levels in *Drosophila* [96].

In the context of expression levels, the OU process (Fig. 2-3a) is a modification of a

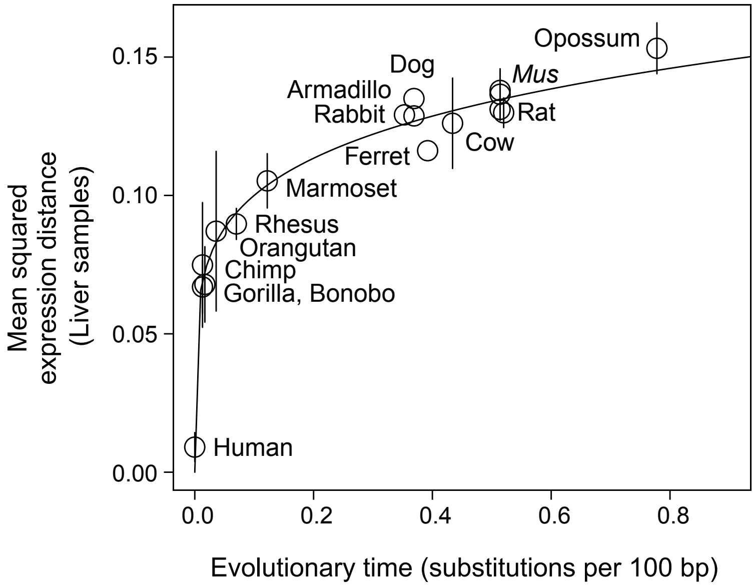
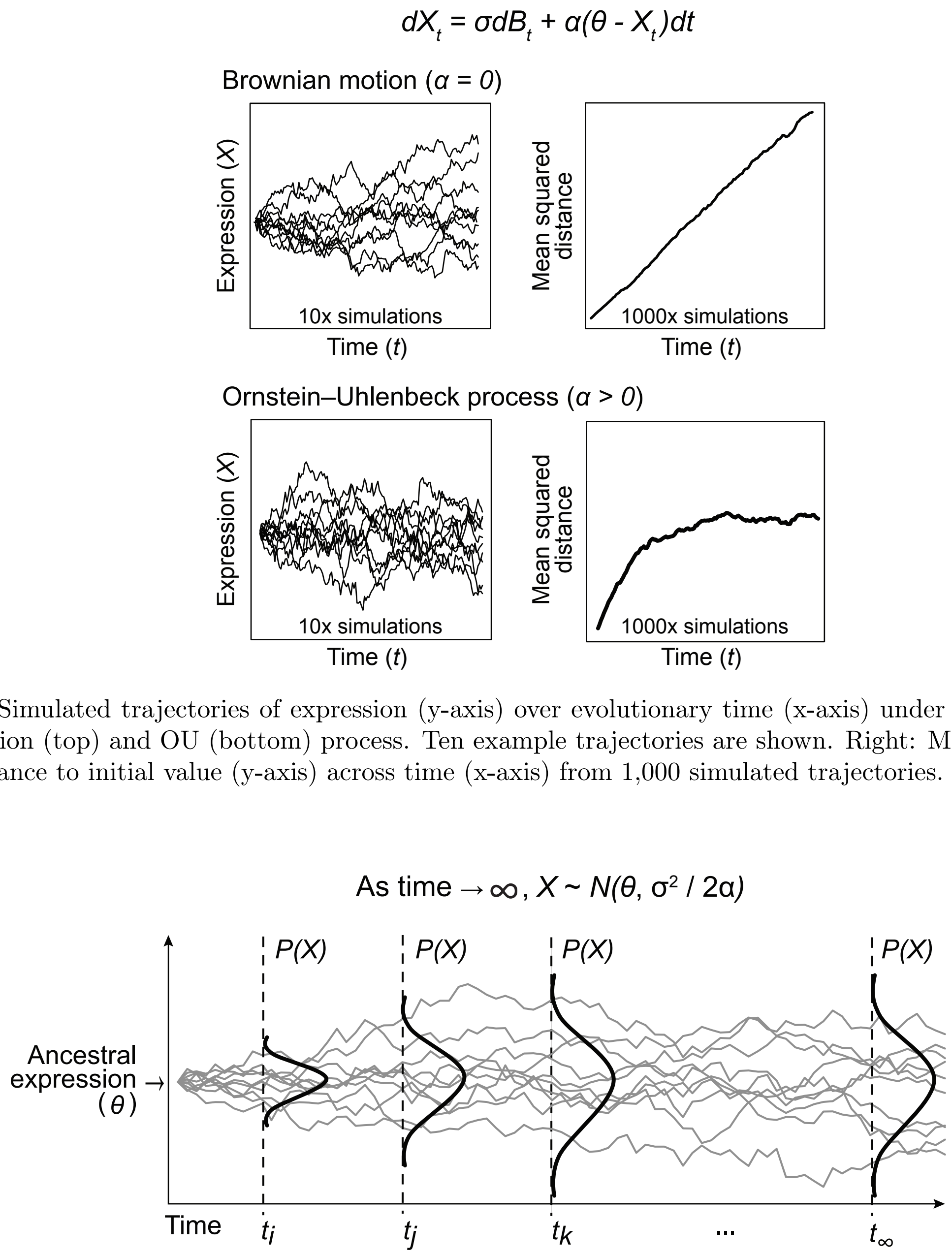


Figure 2-2: Pairwise mean squared expression distances (y-axis) between mammalian and human liver samples across evolutionary time, as estimated by substitutions per 100 basepairs (bp) (x-axis). Error bars: standard deviation of the mean across replicates. Solid line: nonlinear regression fit.

random walk, describing the change in expression (X_t) across time (dt) by:

$$dX_t = \sigma dB_t + \alpha(\theta - Xt)dt$$

where dB_t denotes a Brownian motion process. The model elegantly quantifies the contribution of both drift and selective pressure for any given gene: (1) drift is modeled by Brownian motion with a rate σ (Fig. 2-3a, top), while (2) the strength of selective pressure driving expression back to an optimal expression level θ is parameterized by α (Fig. 2-3a, bottom). The OU process incorporates time information and fully accounts for phylogenetic relationships, thus allowing us to fit individual evolutionary expression trajectories. At longer time scales, the interplay between the rate of drift (σ) and the strength of selection (α) reaches equilibrium and, as time increases to infinity, constrains expression X_t to a stable, normal distribution, with a mean, θ, and variance, $\sigma^2/2\alpha$ (Fig. 2-3b).



(a) Simulated trajectories of expression (y-axis) over evolutionary time (x-axis) under a Brownian motion (top) and OU (bottom) process. Ten example trajectories are shown. Right: Mean squared distance to initial value (y-axis) across time (x-axis) from 1,000 simulated trajectories.

(b) Probability distribution of expression (y-axis) across time (x-axis) under an OU process. The distribution stabilizes as time approaches infinity.

Figure 2-3: Modeling expression evolution using an Ornstein-Uhlenbeck process.

2.3 Functional genomic characterizations using the OU model

Thus far, OU models have primarily been employed for theoretical inferences about fitness gains and selective effects of evolving expression levels [96, 97, 105]. There have also been limited applications of the OU model for detecting selection on expression across smaller mammalian phylogenies and incomplete gene annotations [53, 106]. However, the full power of using the OU model to characterize the evolutionary history of a gene's expression for biological insight has yet to be fully explored.

I thus next developed applications of the OU model to yield biologically interpretable results to evolutionary questions about gene expression levels, gene function, and disease gene discovery. First, for each tissue separately, I estimate from the data the asymptotic distribution of optimal expression for genes under stabilizing selection. I demonstrate that this distribution's OU variance (which I refer to as 'evolutionary variance') accurately characterizes how constrained a gene's expression level is in each tissue. Second, I compare the observed expression level in an individual patient with disease to the optimal level from the model, in order to detect potentially deleterious levels and use those to nominate causal disease genes. Third, I use an extension of the OU model [104, 107] that accounts for the existence of multiple distributions of optimal expression within a phylogeny. I fit this model with a better powered phylogeny and more complete set of gene annotations than previous analyses to identify genetic pathways that may be related to lineage-specific adaptations. I describe each of these applications in turn.

2.3.1 Detecting expression pathways under stabilizing selection

To test whether a gene is under stabilizing selection, I used a likelihood ratio test to compare the fit with no selection ($\alpha = 0$; Brownian motion only; Fig. 2-3a, top) to one with stabilizing selection ($\alpha > 0$, OU process; Fig. 2-3a, bottom, Methods). Because the expression level estimates of lowly expressed genes are associated with high technical variation, their true biological variation across species cannot be accurately inferred [108]. Thus, throughout all subsequent analyses, I focus only on genes expressed over 5 transcripts per million (TPM) (Fig. A-5, Methods). On average, 83% of genes tested (range: 77% - 90%; false discovery rate [FDR] < 0.05) were under stabilizing selection (Fig. 2-4, left, Fig. A-6). Nevertheless, the expression of hundreds of genes within each tissue appeared to be neutrally evolving (Fig. 2-4, right, Fig. A-6).

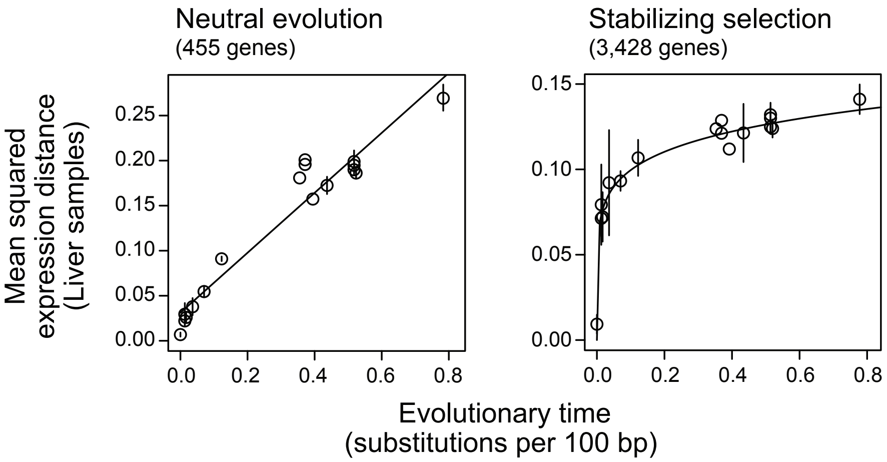


Figure 2-4: Expression divergence patterns of gene expression evolving under neutral evolution or stabilizing selection.

Comparing across tissues, the expression levels of 57% (5,669/8,913) of genes were under stabilizing selection in all tissues in which they were expressed, 39% (2,722) were under stabilizing selection in only some of the tissues where they were expressed, and only 6% (521) were not under stabilizing selection in any of the tissues in the study (Fig. 2-5).

I assessed the sensitivity and specificity to detect genes under expression-stabilizing selection using a jackknifing procedure, where I subsampled to consider phylogenies ranging from 3 to 16 species (Methods). As expected, the number of genes called under stabilizing selection (i.e., rejecting the null hypothesis) increases as more species are included (Fig. A-7a), but does saturate at 14 species. Importantly, the discordance rate (relative to analysis of the full dataset) is very low: less than 1% of genes that are found as under selection with a subsampled phylogeny are found to be neutral (i.e., accepting the null hypothesis) with the full phylogeny (Fig. A-7b).

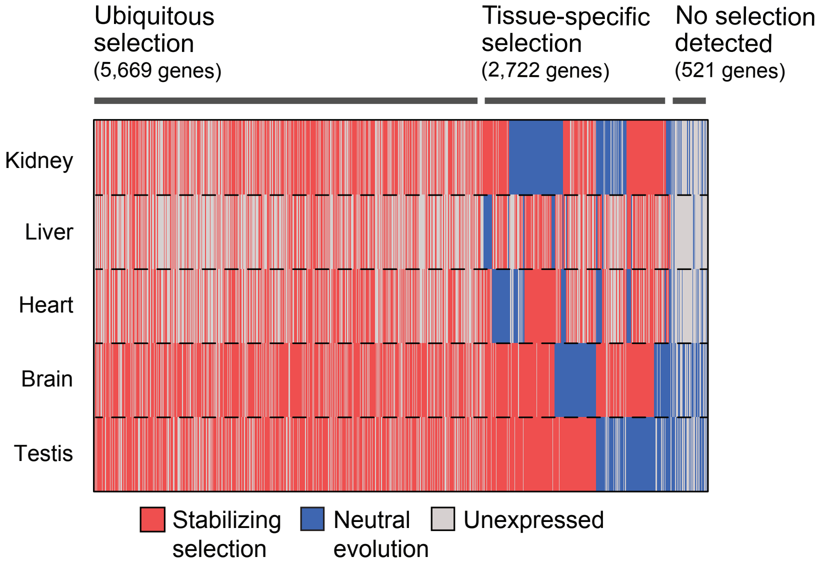


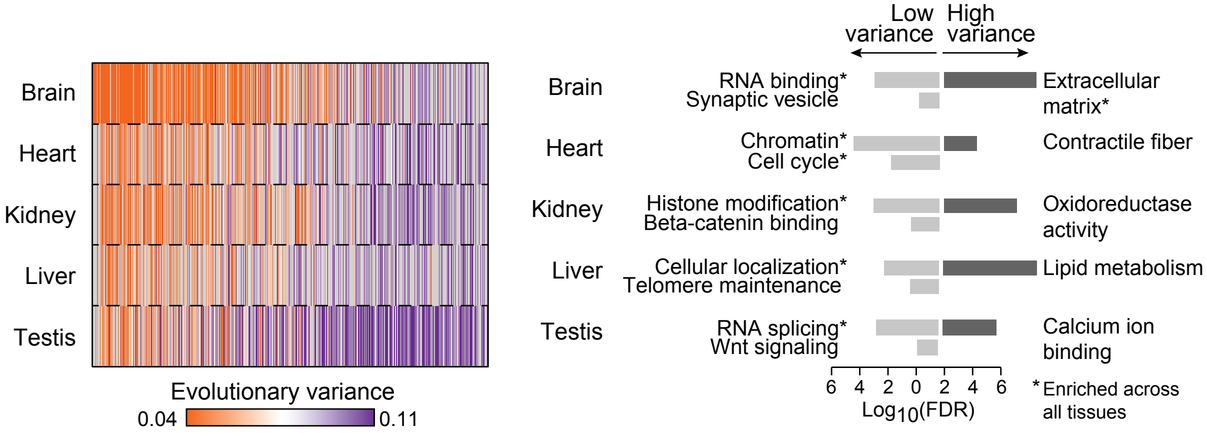
Figure 2-5: Heatmap indicating genes (rows) whose expression is predicted to be evolving under stabilizing selection (red) or neutral evolution (blue) across 5 different tissues (columns). Gray: genes that are expressed < 5 TPM.

2.3.2 Quantifying selective constraint by evolutionary variance

The OU process was considered attractive when initially proposed for modeling expression evolution in *Drosophila* [96] because of its ability to distinguish neutral from stabilizing selection. Given the finding that most mammalian genes are under stabilizing selection, I next explored the ability of the OU model in estimating the stable distribution of gene expression level, which I reasoned is an estimate of the evolutionarily optimal distribution of expression levels. I first investigated the use of the OU model's evolutionary variance as a quantitative measurement of the extent of evolutionary constraint on a gene's expression in each tissue.

The same jackknifing procedure as described above showed that the OU model's evolutionary variance is highly robust to subsampling, as determined by the very low mean squared error (MSE < 0.005) when estimating variance from subsampled phylogenies and with less than 6 species (Fig. A-7c). In fact, the evolutionary variance is far more robust than the simple sample variance used by non-phylogenetic methods (Fig. A-7c). However, because the data compendium is compiled from multiple sources, I do not attempt to interpret absolute values of variance but rather focus on understanding the relative relationship between genes with lower and higher variance.

Brain had the most genes with low variance (most constraint), and testis the least, consistent with previous estimates of rate of expression evolution for those tissues [43,53] (Fig. 2-6a and Fig. A-8). Variance was highly correlated between somatic tissues (mean Pearson's r = 0.84), and less correlated between somatic tissues and testis (mean Pearson's r = 0.55) (Fig. A-9a). For genes expressed across three or more somatic tissues, expression level across tissues was negatively correlated with variance across the tissues (median Pearson's r = -0.27), though the tissue of highest expression only matched tissue of lowest variance in 34.5% (1,673 / 4,840) of genes (Fig. A-9b).



(a) Heatmap of evolutionary variance of expression (orange: low; purple: high) across 8,794 genes (columns) in 5 tissues (rows). Gray: genes < 5 TPM.

(b) Barplot of -log$_{10}$FDR values for significantly enriched gene ontology (GO) categories of low (light gray) and high (dark gray) variance genes within each tissue. Asterisk denotes enrichment across all tissues.

Figure 2-6: Evolutionary variance across tissues and biological processes.

Evolutionary variance and function were strongly associated, consistent with results from previous non-phylogenetic methods that investigated the relationship between cross-species expression variance and gene function [43,109]: across all tissues, genes with low variance were enriched for housekeeping functions (e.g., RNA binding and splicing, chromatin organization, cell cycle), whereas those with high variance were enriched for extracellular proteins (rank-based enrichment test FDR $< 10^{-3}$, Methods). Some processes were enriched in genes with low or high variance only in specific tissues (Fig. 2-6b, Additional file 2): among the processes with tissue-specific low variance were synaptic proteins in brain (FDR $= 1.10 \times 10^{-2}$) and Wnt signaling in testis (FDR $= 1.14 \times 10^{-2}$); processes with high variance included contractile fiber part in heart (FDR $= 5.00 \times 10^{-3}$), oxidoreductase activity in kidney (FDR $= 6.10 \times 10^{-6}$), and lipid metabolism in liver (FDR $= 2.31 \times 10^{-9}$). Thus, estimates of evolutionary variance can be relied on as an indicator of expression constraint and gene function.

2.3.3 Relationship between expression and sequence constraint

I found only a modest correlation between expression and sequence constraint (Pearson's r = -0.25) (Fig. 2-7, Methods). Genes conserved in both expression and sequence were significantly enriched for housekeeping processes (FDR $< 10^{-4}$, Fig. 2-7, Additional file 3), and genes divergent in both were enriched for immune and inflammatory response (FDR $< 10^{-6}$, Fig. 2-7, Additional file 3). More intriguingly, genes conserved in sequence but divergent in expression were enriched in transcriptional regulators (FDR $= 3.10 \times 10^{-5}$), especially those involved in embryonic morphogenesis (FDR $= 9.80 \times 10^{-8}$; e.g., *IRX5*, *HAND2*, *NOTCH1*). Although higher evolutionary variance of expression levels may be impacted by environment, changes in cell type composition, and genetic differences, this analysis supports the hypothesis that divergence in gene regulation without protein sequence divergence can account for species-specific phenotypes.

2.3.4 Detecting deleterious expression levels with evolutionary expression distributions

In analysis of rare diseases, sequence conservation is commonly used to prioritize mutations in genes that are more essential and likely causal for rare diseases when mutated [110–112]. By analogy, I hypothesized that expression conservation should also be predictive of gene essentiality. Indeed, the expression levels of genes that are either essential in culture [113], essential in mice [114], or haploinsufficient in humans [115] had significantly lower evolutionary variance (higher constraint) than their non-essential or haplosufficient counterparts across almost all tissues (Wilcoxon rank-sum test $p < 0.01$, Fig. 2-8a, Methods).

I then examined the variance of disease genes in each of three settings: rare single genes

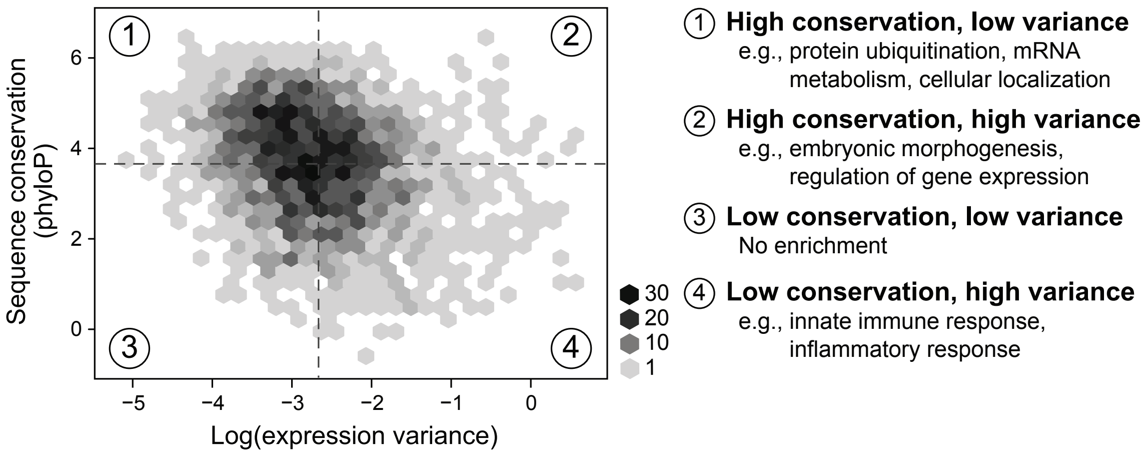
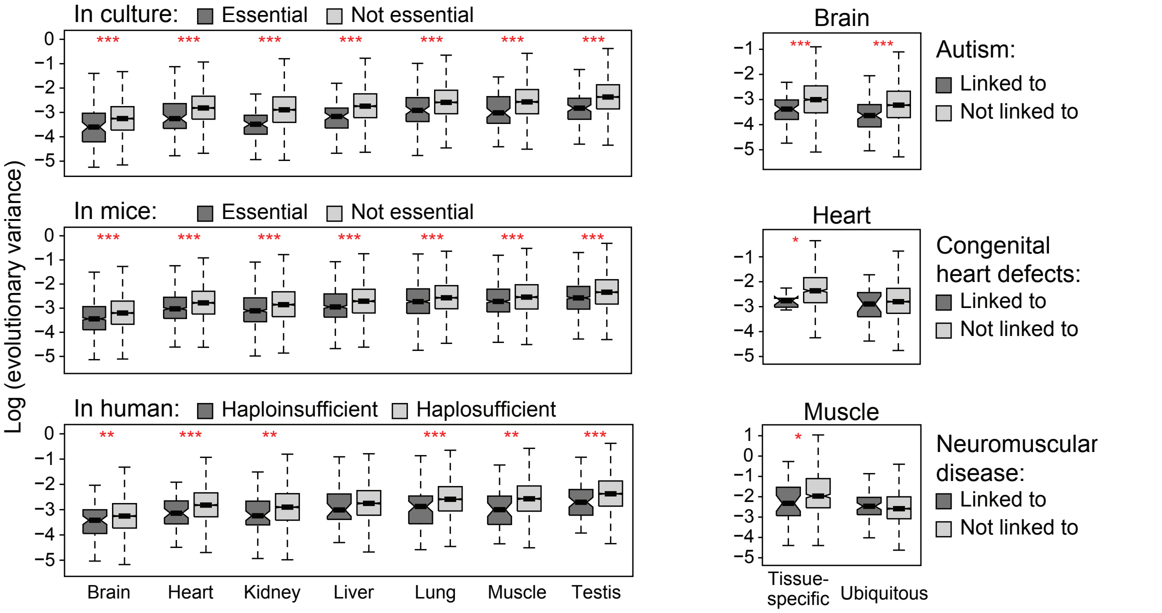


Figure 2-7: Binned scatterplot of evolutionary expression variance (x-axis) vs. sequence conservation (y-axis). Median expression variance and sequence conservation scores are indicated by vertical and horizontal dotted lines, respectively. Enriched GO categories (FDR $< 10^{-3}$) for genes in each quadrant of the scatterplot are listed on the right.

directly linked to non-syndromic autism spectrum disorder (ASD) (brain) [116], congenital heart defects (heart) [117, 118], and neuromuscular disease (skeletal muscle) [119]. In each case, disease genes with tissue-specific expression (Fig. A-10, Methods) consistently exhibited significantly lower variance in the disease-relevant tissue than tissue-specific non-disease genes ($p < 0.05$, Fig. 2-8b). In ASD-linked genes (but not the other two conditions), I also observed significantly lower variance of ubiquitously expressed disease vs. non-disease genes, perhaps related to observed high rates of co-morbidities with ASD [120].

Next, I hypothesized that the parameters of each gene's optimal OU distributions can predict disease genes by highlighting outlier, likely pathogenic, gene expression levels in rare disease patient data. This is analogous to causal disease gene discovery by identifying putatively pathogenic sequence mutations in whole exome sequencing [121–124]. To this end, I obtained RNA-seq of muscle biopsies of 93 patients clinically diagnosed with neuromuscular disease (Methods, Additional file 4). For each patient sample, I calculated a z-score for each gene to assess how they deviate from the evolutionarily optimal fit for that gene's expression in skeletal muscle, with correction for multiple hypothesis testing (Fig. 2-9a, Methods).

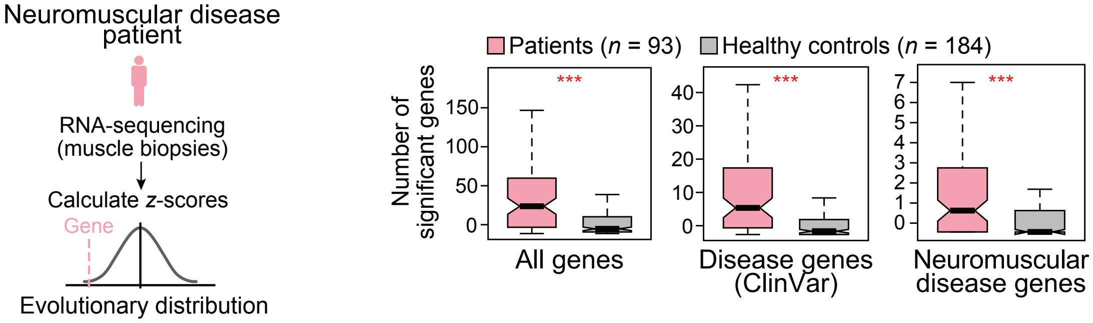
Compared to GTEx muscle samples from 184 healthy people [125], patients had, on average, 3.2-fold more dysregulated genes overall by this measure (Wilcoxon rank sum test $p =$



(a) Evolutionary variance of genes essential in culture (top), essential in mice (middle), and haploinsufficient in human (bottom) (dark gray), and their non-essential or haplosufficient counterparts (light gray) in each of 7 tissues (x-axis). *** denotes $p < 0.001$; ** denotes $p < 0.01$; * denotes $p < 0.05$.

(b) Evolutionary variance of genes linked (dark gray) and not linked (light gray) to high-penetrance disease. Left boxes: genes specifically expressed in disease tissue; right boxes: ubiquitously expressed genes. *** denotes $p < 0.001$; ** denotes $p < 0.01$; * denotes $p < 0.05$.

Figure 2-8: Evolutionary variance of essential and disease-related genes.



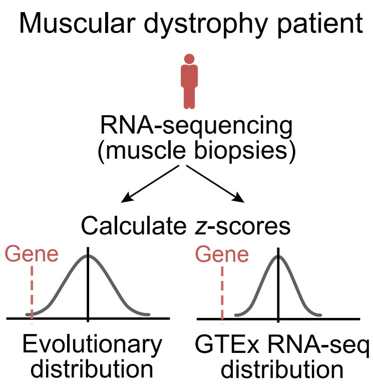
(a) Schematic of method to identify outlier gene expression in RNA-seq data from patient muscle biopsies.

(b) Number of significant genes (FDR < 0.01) from patient data (pink) and healthy controls (gray) when considering all genes (left), all disease genes (middle), and neuromuscular disease genes (right). *** denotes $p < 0.001$.

Figure 2-9: Using evolutionary distributions to identify outlier gene expression in patient RNA-seq data.

2.0×10^{-9}, Fig. 2-9b, left), 3.0-fold more dysregulated muscle-expressed disease genes [126] ($p = 2.1 \times 10^{-10}$, Fig. 2-9b, middle), and 2.0-fold more dysregulated known neuromuscular disease genes ($p = 2.7 \times 10^{-4}$, Fig. 2-9b, right). This suggests that the evolutionary parameters fit by the OU model can be used to detect outlier expression values that are more likely to be deleterious. Importantly, in contrast to methods for differential expression between patient and healthy controls, the test does not require a control population, and can be conducted for a single patient sample.

Finally, I tested whether the OU model could be used to identify the causative gene in rare disease analysis. As a proof of principle, I focused on the subset of 8 patients from the muscle disease cohort who were clinically diagnosed with either Becker or Duchenne muscular dystrophy, including confirmation of absent or decreased dystrophin protein via immunoblotting [119]. To compare this approach to a standard differential expression analysis, I ranked genes by outlier expression with z-scores defined based either on (1) comparison to the mean and variance estimated from the evolutionary data; or (2) comparison to a mean and variance estimated from only healthy GTEx human data (Fig. 2-10a).


Muscular dystrophy patient
RNA-sequencing
(muscle biopsies)
Calculate *z*-scores
Gene
Gene
Evolutionary distribution
GTEx RNA-seq distribution

(a) Two scoring approaches based on evolutionary distributions (left) or GTEx distributions (right).

Evolutionary distribution			GTEX RNA-seq distribution		
# sig. genes	*DMD* -$\log_{10}$FDR	*DMD* rank	# sig. genes	*DMD* -$\log_{10}$FDR	*DMD* rank
2	4.49	1	0	0.65	43
1	2.26	2	0	0.15	75
3	9.13	1	1	3.87	1
32	8.27	1	7	1.73	57
10	6.50	2	22	3.80	10
0	2.96	1	80	0.58	377
5	15.4	1	132	11.7	31
23	7.22	2	250	5.22	50

(b) Results from scoring approaches when using distributions estimated from evolutionary data (left) or GTEx RNA-seq (right). Highlighted row denotes FDR $< 10^{-3}$.

Figure 2-10: Identifying outlier gene expression from RNA-seq data of muscular dystrophy patients.

By the evolutionary data, fewer genes ranked as significant outliers in each patient (median: 4, range: 0 – 32), and *DMD* ranked as either the top or second most significantly aberrantly expressed gene in 6 of 8 patients, each showing significant underexpression (FDR

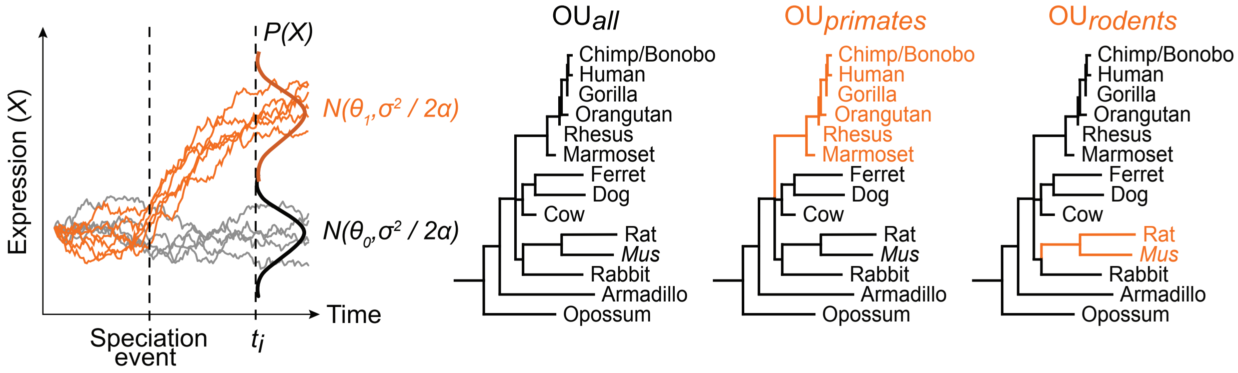
$< 10^{-3}$) (Fig. 2-10b, left). By comparison, scoring in reference to GTEx expression data did not yield such specific results: a median of 14.5 genes were outliers (range: 0 – 250), only 4 of 8 patients were called as significantly underexpressing *DMD* (FDR $< 10^{-3}$) (Fig. 2-10b, right), and its significance in these patients ranked between 1 and 50. Thus, using the OU model's estimate of evolutionary mean and variance of optimal gene expression helps detect gene dysregulation of the actual disease gene and could aid novel disease gene discovery in individual patients, even without any control samples.

2.3.5 Identifying directional selection in gene expression with a multivariate OU model

Finally, I explored the use of the OU framework to detect directional selection in gene expression. I used an extension of the model that accounts for multiple selection regimes across a single phylogeny by modeling the distribution of expression level as a multivariate normal distribution whose mean and variance are estimated for each (predefined) subclade [104, 107] (Fig. 2-11a). A previous application of this extended OU model identified over 9,000 significant expression changes across the mammalian phylogeny [53], but the analysis relied on a smaller phylogeny and thus focused on identifying species-specific shifts in gene expression that are easily confounded by environmental causes or technical effects.

I leveraged the comprehensive phylogenetic coverage of my dataset and focused on detecting shifts in expression consistent in direction and magnitude across entire subclades of two or more mammals, whose samples were collected and sequenced across multiple sources to mitigate non-genetic confounders. I identified 'differential gene expression' across the tree based on the approach suggested by Butler and King 2004 (Methods): I applied the extended model for each gene in each tissue and tested each of three hypotheses: OU_{all}, which models a single optimum for all species, and $\text{OU}_{primates}$ and $\text{OU}_{rodents}$, each modeling two optima, one for the ancestral distribution and one for the distribution within primates (branch length $= 0.12$) or rodents (branch length $= 0.18$), respectively (Fig. 2-11b).

For each gene, I first used a likelihood ratio test between each OU model and the null hypothesis of a Brownian motion model and removed any models against which the neutral model could not be rejected. I then assigned the best OU model using goodness-of-fit



(a) Simulated trajectories of expression (y-axis) over time (x-axis) under a multivariate OU process.

(b) Three tested hypotheses of expression evolution: the univariate OU$_{all}$ model, in which gene expression evolves under a single stabilizing regime across the phylogeny (black), and two multivariate OU models, OU$_{primates}$ and OU$_{rodents}$, in which gene expression evolves under the ancestral regime (black) and a new regime in the specified subclade (orange).

Figure 2-11: Modeling lineage-specific expression changes using a multivariate OU process

tests. In a related method [53, 106], p-values are derived by directly testing the alternative hypothesis $\theta_{subclade} \neq \theta_{ancestral}$ against the null hypothesis $\theta_{subclade} = \theta_{ancestral}$. However, I found that many models are unable to overcome multiple hypothesis correction with this stringent approach, even with my larger phylogeny. Instead, I estimate false discovery rates by shuffling species assignments (Methods) and found that I achieved FDR $< 30\%$ in liver and testis (both subclades), as well as in the primate clade for brain and lung (Fig. A-11). Finally, as a conservative measure, I retained only those genes that also changed at least 2-fold between subclades and had a mean expression level of at least 5 TPM in one of the subclades.

As an example, in liver, I identified 640, 794, and 615 genes with lineage-specific expression changes in primates, rodents, and carnivores, respectively, highlighting specific metabolic processes diverging in regulation in each clade. The expression levels of lineage-specific genes deviated significantly from expectation only if there was no clade-specific selection (Fig. 2-12).

Because of the larger set of differentially expressed genes compared to previous applications, I could identify functional enrichments among lineage-specific genes (Additional file 5). I found primate-specific downregulation of genes related to a number of lipid metabolic

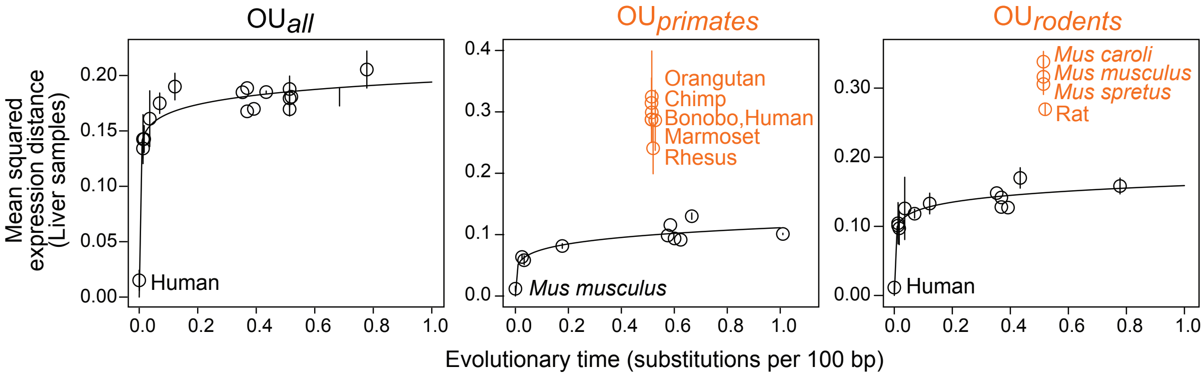


Figure 2-12: Pairwise mean squared expression distances (y-axis) between a reference species (labeled black point) and each of the other mammals in liver samples for genes assigned to each of three tested OU models. Black points: Species evolving under ancestral distribution; Labeled orange points: species evolving under new regime after the lineage split.

processes in the liver (FDR $= 1.88 \times 10^{-11}$). These processes include peroxisomal functions (FDR $= 2.45 \times 10^{-8}$), fatty acid metabolism (FDR $= 1.52 \times 10^{-8}$), and lipid transport (FDR $= 3.36 \times 10^{-3}$) (Fig. 2-13, Additional file 5), and contain known regulators of lipid metabolism such as the LDL receptor (*LDLR*) [127], hepatic lipase (*LIPC*) [128], and the transcription factor *PPAR*-α [129]. Thus, the expression of multiple pathways may have diverged at the ancestral primate branch, consistent with observations that human lipidemia is not well-modeled by mice without further genetic modification [130]. In another example, genes involved in regulation of immune response were downregulated across rodent livers (FDR $= 6.97 \times 10^{-4}$), and in testis, microtubule-based movement genes (FDR $= 2.82 \times 10^{-3}$) and spermatogenesis (FDR $= 2.82 \times 10^{-2}$) were downregulated across primates (Fig. 2-13), reflecting the known rapid evolution of immune- [109, 131, 132] and reproduction-related genes [133, 134].

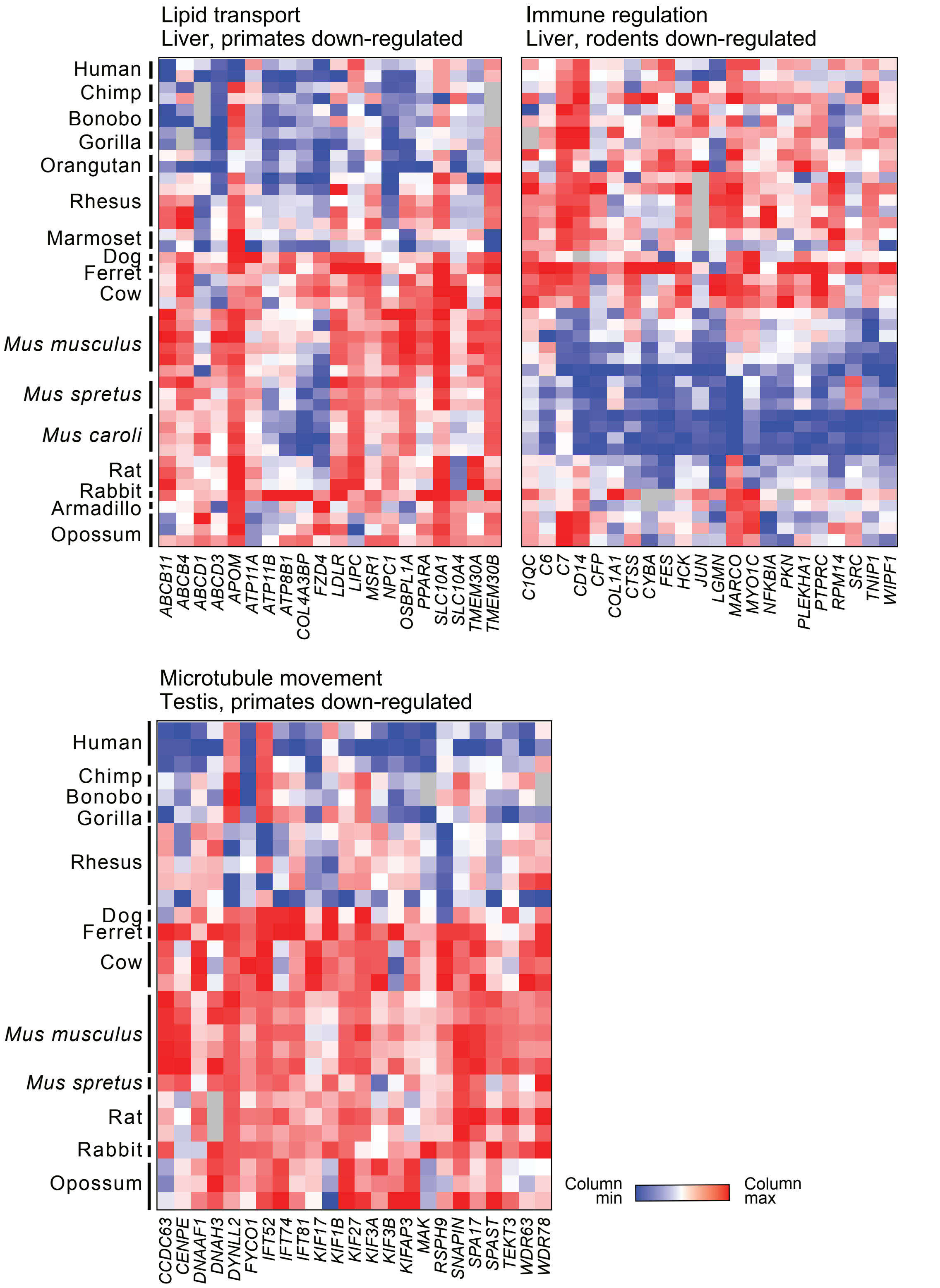


Figure 2-13: Example processes enriched for lineage-specific expression.

2.4 Conclusion

In conclusion, by combining a large dataset of comparative gene expression profiles across mammals with systematic analysis, I showed that gene expression of one-to-one mammalian orthologs is evolving nonlinearly across evolutionary time and is accurately modeled by an OU process. I then show how to use this model to answer three key questions: (1) estimating the distributions of optimal gene expression levels and quantifying the extent of evolutionary constraint on expression, (2) identifying deleterious gene expression in individual patient disease tissue by characterizing outliers relative to a predicted distribution of optimal expression for each gene, and (3) detecting lineage-specific expression using an extension that accounts for multiple distributions of optimal expression. Looking forward, I anticipate that the OU model can be further developed for other biological queries, for example testing for stabilizing selection across pathways of genes or paralog families or estimating ancestral expression states. As shown by my analysis, characterizations of expression across additional tissue types and species under varied developmental and environmental contexts will provide increased power and further insight into the evolution of gene expression, and the relationship between genotype and phenotype.

2.5 Methods

2.5.1 Data collection

The following table summarizes the sources for all data used in this study. For a more detailed table of SRA accession numbers and read alignment statistics, see Additional file 1.

Sources

Illumina Body Map 2.0 at `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/`

Merkin et al. 2012 [65]

Brawand et al. 2011 [53]

Harr and Turner 2010 [98]

Non-human primate reference transcriptome resource (NHPRTR) [99]

Cortez et al. 2014 [100]

Wong et al. 2015 [101]

(Table on next page.)

Species	**Reference genome**	**Brain**	**Heart**	**Kidney**	**Liver**	**Lung**	**Sk. muscle**	**Testis**
Human	hg19	Brawand et al.; Illumina Body Map 2.0	Brawand et al.; Illumina Body Map 2.0	Brawand et al.; Illumina Body Map 2.0	Brawand et al.; Illumina Body Map 2.0	Illumina Body Map 2.0	Illumina Body Map 2.0	Brawand et al.; Illumina Body Map 2.0
Chimp	panTro4	Brawand et al.	Brawand et al.	Brawand et al.	Brawand et al.	NHPRTR	NHPRTR	Brawand et al.
Bonobo	panTro4	Brawand et al.	Brawand et al.	Brawand et al.	Brawand et al.	NHPRTR	NHPRTR	Brawand et al.
Gorilla	gorGor3	Brawand et al.	Brawand et al.	Brawand et al.	Brawand et al.			Brawand et al.
Orangutan	ponAbe2	Brawand et al.	Brawand et al.	Brawand et al.	Brawand et al.			
Macaque	rheMac8	Brawand et al.; Merkin et al.	Brawand et al.; Merkin et al.	Brawand et al.; Merkin et al.	Brawand et al.; Merkin et al.	Merkin et al.	Merkin et al.	Brawand et al.; Merkin et al.
Marmoset	calJac3	Cortez et al.	Cortez et al.	Cortez et al.	Cortez et al.	NHPRTR	NHPRTR	
Mus musculus	mm10	Brawand et al.; Merkin et al.	Brawand et al.,; Merkin et al.	Brawand et al.; Merkin et al.	Brawand et al.; Merkin et al.	Merkin et al.	Merkin et al.	Brawand et al.; Merkin et al.
Mus spretus	mm10				Wong et al.			Harr and Turner
Mus caroli	mm10				Wong et al.			
Rat	rn6	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.
Rabbit	oryCun2	This study	This study	This study	This study	This study	This study	This study
Dog	canFam3	This study	This study	This study	This study	This study	This study	This study
Ferret	musFur1	This study	This study	This study	This study	This study	This study	This study
Cow	bosTau6	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.	Merkin et al.
Armadillo	dasNov3		This study	This study	This study	This study	This study	
Opossum	monDom5	Brawand et al.; This study	Brawand et al.; This study	Brawand et al.; This study	Brawand et al.; This study	This study	This study	Brawand et al.; This study

Table 2.1: Data sources for all samples used in in this study.

Samples for evolutionary dataset

RNA samples from dog and rabbit tissues were commercially obtained from Zyagen. RNA samples from opossum tissues were a kind gift from Paul Samollow (Texas A&M). RNA samples from armadillo tissues were a kind gift from Jason Merkin and Christopher Burge (MIT). All tissue collection was approved by IACUC and carried out in accordance with respective institutional guidelines.

RNA-seq for evolutionary dataset

RNA-seq libraries were prepared as described in [135]. Briefly, 10 μg total RNA was poly-A selected twice using Dynabeads mRNA Purification Kit (Invitrogen, 610.06). Resulting mRNA was DNase treated (Ambion AM1907) and then fragmented using heat. First strand cDNA synthesis was performed using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, 11917-010), supplementing in SuperScript III Reverse Transcriptase (Invitrogen, 18080-093), incorporating SUPERase*In (Ambion, AM2694), and Actinomycin D (USB, 10415). First strand cDNA was cleaned using 1.8X RNAClean XP SPRI beads (Beckman Coulter, A64987). Second-Strand synthesis was performed replacing dTTP with dUTP, and the resulting double-stranded cDNA was cleaned using a MinElute PCR Purification Kit (Qiagen, 28004). Illumina libraries were constructed by repairing the ends of the cDNA, ligating adapters, and cleaning/size-selecting with 0.7x SPRI. Illumina libraries were treated with USER to excise dUTP, and amplified via PCR using Fusion Master mix with GC buffer (NEB, F532S). Samples were sequenced on an Illumina HiSeq 2000 sequencer, to a minimum depth of 35M reads.

2.5.2 Data processing

Genome and transcriptome annotations

All genomes were downloaded from the UCSC Genome Browser [6]. To assemble transcriptomes, Ensembl gene annotations [102] were downloaded from UCSC Table Browser (table ensGene) and converted to sequence using BEDTools [136]. Ortholog annotations were downloaded from Ensembl BioMart (Ensembl Genes 90) [137]. Only genes that met the following criteria were used for this study: (1) no duplications in any of the studied mammals, (2) an ortholog present in either armadillo or opossum (i.e., placental mammal or marsupial outgroup), (3) no more than three gene losses across primates (human, chimp, gorilla, orangutan, macaque, marmoset), (4) no more than one gene loss across glires (mouse, rat, rabbit), and (5) no more than one gene loss across laurasiatherians (cow, dog, ferret).

Alignment and expression quantification

RSEM v1.2.12 [138] was used to align reads to the transcriptome of each species and to quantify TPM of each gene using default parameters.

2.5.3 Estimating divergence rate of gene expression

Quantifying expression difference

To calculate pairwise expression differences between each species ('comparing species') and a reference species (e.g., human in Fig. 2-2, or mouse in Fig. A-4, I applied principal component analysis (PCA) on pairwise gene expression levels ($\log_{10}$[TPM]), considering only genes that were expressed ($>$ 0 TPM) in at least one species. For each tissue and each pair of species, I used the first principal component as the best fit line between the two species' expression profiles. I then defined the pairwise expression difference as the orthogonal distance from the observed expression level in the comparing species to the best fit line. I used PCA rather than a linear regression because PCA accounts for noise in expression values from both species, while the linear regression would only model noise in the comparing species and treat the reference species as an independent variable (Fig. A-2).

Phylogenetic tree

The phylogenetic tree of vertebrate species was downloaded from UCSC Genome Browser at `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz100way/` [6]. Distances between mammals used in this study were extracted using the Environment for Tree Exploration Toolkit [139].

2.5.4 Modeling expression evolution

Fitting linear and nonlinear regression models

Under an OU model, the expected mean squared distance across time follows a power law relationship ($y = ax^k$). To fit this relationship between observed mean squared expression distances (y) and evolutionary time (x), I log-transformed both axes to relate the variables linearly: $log(y) = log(a) + klog(x)$. I then used least squares regression to find coefficients a and k.

For genes whose expression evolution fit better under a Brownian motion model (see below), I used least squares regression to find the best fit line between mean squared expression distances and evolutionary time.

Normalization of gene expression values

Gene expression values ($\log_{10}$[TPM]) were normalized using TMM normalization [140] from the Bioconductor package edgeR [141]. Briefly, TMM normalization assumes that the majority of genes are not differentially expressed (DE) between samples and estimates a scaling factor between a pair of samples, such that the trimmed mean of log expression ratios (trimmed mean of M values [TMM]) is equal to 1. It is reasonable to make the assumption that the majority of genes between pairs of species are not DE, because even between distant mammals such as human and opossum, Pearson's correlation of expression level in a given tissue is > 0.75. Within each tissue, I then use human expression level as a reference to fit a scale factor and normalize all other samples.

Fitting OU process parameters

Brownian motion (BM) and OU models were fit to normalized expression values using the R package ouch [107] with default parameters. P-values for each gene were calculated using a likelihood ratio test comparing the OU (alternative hypothesis) to the BM (null hypothesis) model, and then corrected for multiple hypothesis testing using the Benjamini-Hochberg FDR procedure [142].

Relationship between gene expression level and OU variance

As expected, genes with low expression levels are estimated to have high OU variance, but this is likely largely contributed from technical, rather than true biological, variance [108]. To account for this, I focused only on genes whose estimated OU mean (θ) was over 5 TPM. I chose this cutoff because it removes the majority of the relationship between OU variance and expression level, while preserving the majority of expressed genes for analysis (Fig. A-5). Note that even among genes with TPM $>$ 5, those with higher expression level still have slightly lower OU variance, contrary to expectations of heteroscedasticity.

2.5.5 Estimating robustness of OU process parameters

To test the robustness of the OU model, I used a jackknifing procedure, where I subsampled phylogenies ranging from 3 to 16 species (out of a total of 17 species). For each phylogeny size, I created 10 randomly subsampled phylogenies and then fit the OU model as described above.

2.5.6 Functional annotations, by evolutionary variance

To test for enriched GO categories across genes with low or high evolutionary variance, I used the ranked enrichment test from GOrilla [143]. To avoid biases due to relationship between lowly expressed genes and high evolutionary variance estimates, I only used genes expressed at $>$ 5 TPM.

2.5.7 Functional annotations, by expression variance and sequence conservation

Measuring sequence conservation

Sequence conservation of a gene was defined by mean phyloP score [86] across the coding region of the longest annotated coding transcript of that gene.

GO annotation by sequence and expression conservation

I tested for enriched GO categories across genes in all four categories of high or low evolutionary variance and high or low sequence conservation. For each tissue separately, I defined 'high' or 'low' based on the median evolutionary expression variance and median phyloP score, respectively, and assigned all genes expressed at $>$ 5 TPM to one of four categories. For GO enrichment analysis, where only sets with relatively large numbers of genes are typically enriched at levels that survive multiple hypothesis testing correction, I first unified the genes of each category across all tissues and then used GOrilla [143] to test for enrichments in the combined gene lists. Because gene function is related to evolutionary variance, for the background set I used the appropriate list of all high or low expression variance genes expressed at $>$ 5 TPM.

2.5.8 Evolutionary variance of essential and disease genes

Essential, haploinsufficient, and disease gene sets

The following gene lists were downloaded from the McArthur Lab gene lists repository at `https://github.com/macarthur-lab/gene_lists`: essential in culture, essential in mice, ClinGen haploinsufficient genes, genes with any disease association reported in ClinVar, and neuromuscular disease genes.

Rare, single genes contributing to non-syndromic autism spectrum disorder were downloaded from the SFARI database at `https://gene.sfari.org/` by selecting Category 1 genes (rare single gene variants, disruptions/mutations, and submicroscopic deletions/duplications directly linked to ASD) with a gene score of 1 (high confidence), 2 (strong candidate) or 3 (suggestive evidence).

Genes contributing to congenital heart disease were curated by filtering for genes annotated with 'Congenital heart defects' in OMIM's Morbid Map at `https://omim.org/downloads/` as well as genes associated with congenital heart disease (DOID: 1682) from the MGI Disease Ontology Browser at `http://www.informatics.jax.org/disease`.

Defining tissue-specific genes

Because the dataset consists of closely related tissues (e.g., heart and skeletal muscle, Fig. 2-1b), I did not want to define only genes expressed in a single tissue as tissue-specific. I found that the distribution of number of tissues in which genes are expressed $>$ 5 TPM (Fig. A-10) is somewhat bimodal and, based on visual inspection, defined a cut-off of three or fewer tissues as tissue-specific. The observation that tissue-specific disease genes had lower variance compared to non-disease genes was robust to different cutoffs (data not shown). (However, lower cutoffs result in fewer genes defined tissue-specific, reducing power to achieve statistical significance for downstream analyses.)

2.5.9 Identifying disease genes from neuromuscular disease RNA-seq data

Samples for neuromuscular disease dataset

The cohort of neuromuscular disease patient RNA-seq described in this study is a superset of that described in Cummings et al. 2017 [119] (dbGaP accession phs000655.v3.p1) and 30 additional patients. Tissues were procured under Institutional Review Board (IRB) approved protocols at National Institute of Neurological Disorders and Stroke (Protocol #12-N-0095), Newcastle University (CF01.2011), Boston Children's Hospital (03-12-205R), University College London (08ND17), UCLA (15-001919), and St. Jude Children's Research Hospital (10/CHW/45). Patients were consented to these protocols in clinic visits prior to biopsy. Patient muscle biopsies were collected as described in Cummings et al. 2017.

RNA-seq for neuromuscular disease dataset patient data

RNA-seq from muscle biopsies was performed as described in Cummings et al. 2017 [119]. To minimize technical differences, patient muscle samples were sequenced using the same protocol as in the GTEx project [125], patients sequenced at or above the same coverage as GTEx, and analyzed using identical pipelines. Briefly, muscle biopsies or RNA were shipped frozen from clinical centers via a liquid nitrogen dry shipper and stored in liquid nitrogen cryogenic storage. All samples analyzed with H&E showed muscle quality sufficient to proceed to RNA-seq. RNA was extracted from muscle biopsies via the miRNeasy Mini Kit from Qiagen per kit instructions. All RNA samples were measured for quantity and quality and samples had to meet the minimum cutoff of 250ng of RNA and RNA Quality Score (RQS) of 6 to proceed with library prep. RNA-seq library preparation was performed at the Broad Institute Genomics Platform using the poly-A selection of mRNA with an Illumina TruSeq kit. Paired-end sequencing was performed in the Genomics Platform on Illumina HiSeq 2000 instruments. Read length and sequence coverage information is available in Additional file 4.

GTEx BAM files were downloaded from dbGaP under accession ID phs000424.v6.p1 and realigned after conversion to FASTQ files with Picard SamToFastq. Both patient and GTEx reads were aligned using Star 2-Pass v.2.4.2a [144] using hg19 as the genome reference and Gencode V19 annotations [145]. Duplicate reads were marked with Picard MarkDuplicates (v.1.1099) available at `http://broadinstitute.github.io/picard`.

Detecting outlier expression in patient samples

Genes expression values ($\log_{10}$[TPM]) were first normalized by TMM normalization [140] to the human skeletal muscle expression values used to originally to fit the OU parameters. For each gene in each patient sample, a z-score was calculated using the asymptotic mean and variance estimated from the evolutionary data. Z-scores were only calculated for genes that were assessed to fit better under the OU rather than the BM model (FDR < 0.05, see Fitting OU process parameters) and whose asymptotic mean was estimated to be 5 TPM or higher. Z-scores were converted to p-values and then corrected for multiple hypothesis testing [142]. I used a FDR threshold of 0.01 to initially define significance. I then removed another 330 genes that scored as a significant outlier in more than 25% of the GTEx samples.

As a comparator (Fig. 2-10), z-scores were also calculated using the sample mean and variance estimated from healthy human GTEx samples. To ensure comparability between the two methods, I only calculated z-scores for genes that were not filtered out at any steps during the evolutionary method above.

2.5.10 Detecting lineage-specific expression programs

Within each tissue, OU parameters for each of the three hypotheses (OU_{all}, $\mathrm{OU}_{primates}$, $\mathrm{OU}_{rodents}$) were estimated for each gene as described above. P-values were calculated using a likelihood ratio test comparing each of the OU models to the BM model. Results from each of the three hypotheses were then independently adjusted for multiple hypothesis testing using the Benjamini-Hochberg FDR procedure [142]. For each gene, Akaike and Bayesian Information Criterion scores were calculated on all models that were significant against the null to determine the best fitting model. Both scores were in agreement for the best fitting model in all cases. Unrealistic parameters (optimal expression $\theta > 10^{4.5}$ or $\theta < 0$) were estimated only in very few cases (1.6% of genes tested per tissue, on average) and moreover none of these models was found to be statistically significant against the null hypothesis.

Estimating FDR of lineage-specific expression programs

To estimate the FDR, I performed the same procedure in each tissue using shuffled species assignments (on the same tree topology) and only retained hypotheses that achieved a FDR < 0.30. For additional stringency, I only defined genes as being differentially expressed if the fold change between the estimated means across the lineages (e.g., primate mean vs. ancestral mean) was greater than 2-fold and if the mean reached at least 5 TPM in one of the lineages. I performed GO enrichment analysis on each set of up- and down-regulated genes separately, using a background set of genes with mean expression of at least 1 TPM across all species in the appropriate tissue.

2.5.11 Data availability

Processed expression data and evolutionary expression distributions for all one-to-one mammalian orthologs in each tissue context are available at `https://portals.broadinstitute.org/evee/`.

New RNA-seq data is available under GEO accession GSE106077 at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106077`.

2.5.12 Additional files

Additional file 1: Data accessions. [link]

Additional file 2: Enriched GO annotations, by evolutionary variance. [link]

Additional file 3: Enriched GO annotations, by expression variance and sequence conservation. [link]

Additional file 4: Alignment statistics for neuromuscular disease RNA-seq data. [link]

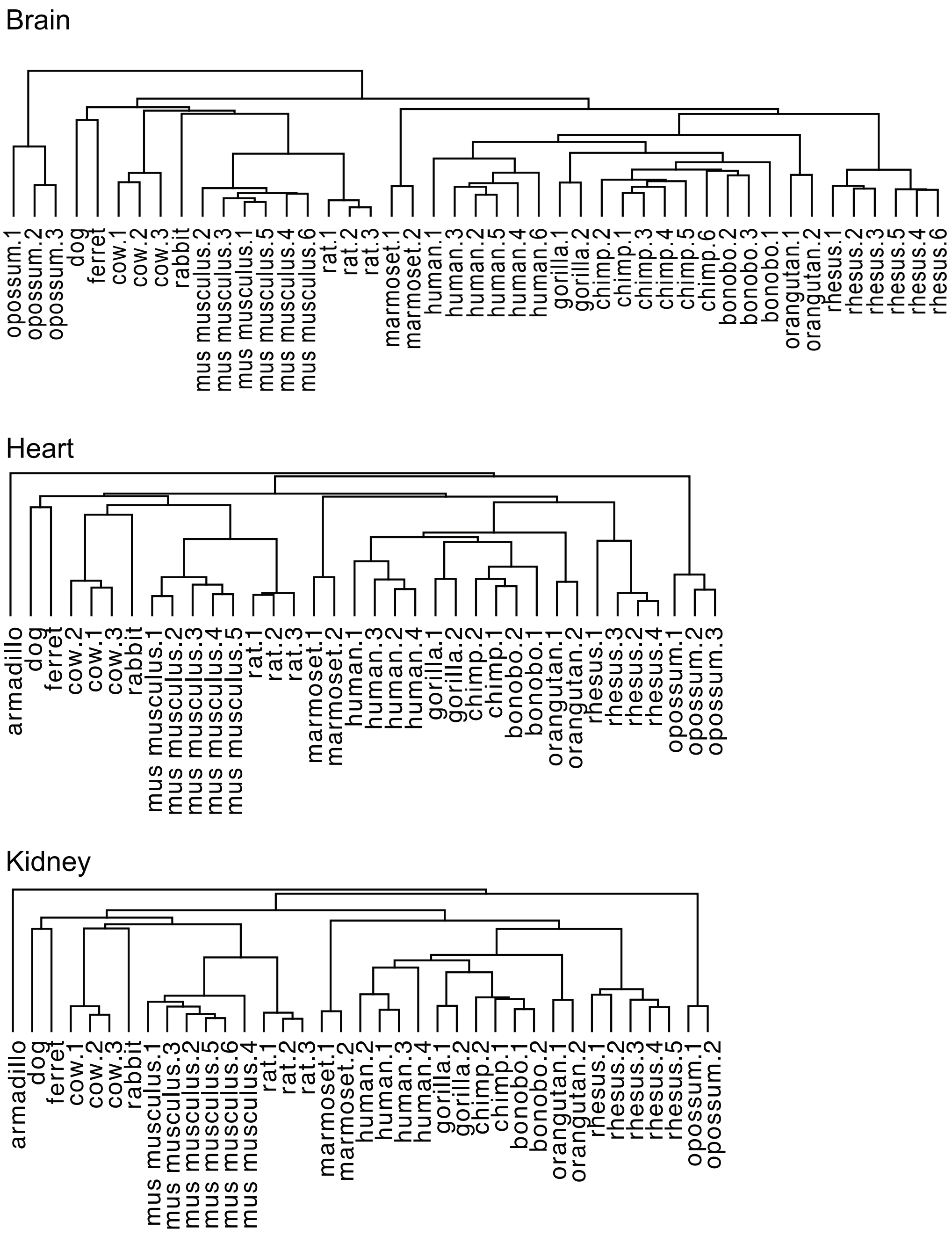
Additional file 5: Enriched GO annotations of genes with lineage-specific expression. [link]

2.5.13 Authors' contributions

JC conceived, designed, and coordinated all aspect of the study, carried out all analyses including software development of EVEE Gene Browser, and wrote the manuscript. JJ carried out mammalian data collection and RS carried out RNA-sequencing. BC participated in designing neuromuscular disease analysis and provided all neuromuscular RNA-seq data and supplementary analyses. NR participated in data analysis. KLT, WH, FDP, and AR participated in designing the study and writing the manuscript. All authors read and approved the manuscript.

Additional thanks to Daniel MacArthur for help with neuromuscular disease dataset; Leslie Gafney for artwork and advise on figures; Hopi Hoekstra, Daniel Hartl, Brian Haas, and Akshay Krishnamurthy for advice on analysis; Quinlan Sievers and Yarden Katz for advice on writing; Eric Weitz and Riccardo Calixte for help with software development; and to the Regev laboratory members for helpful discussions.

Supplementary materials for Chapter 2



(Figure continues on next page.)

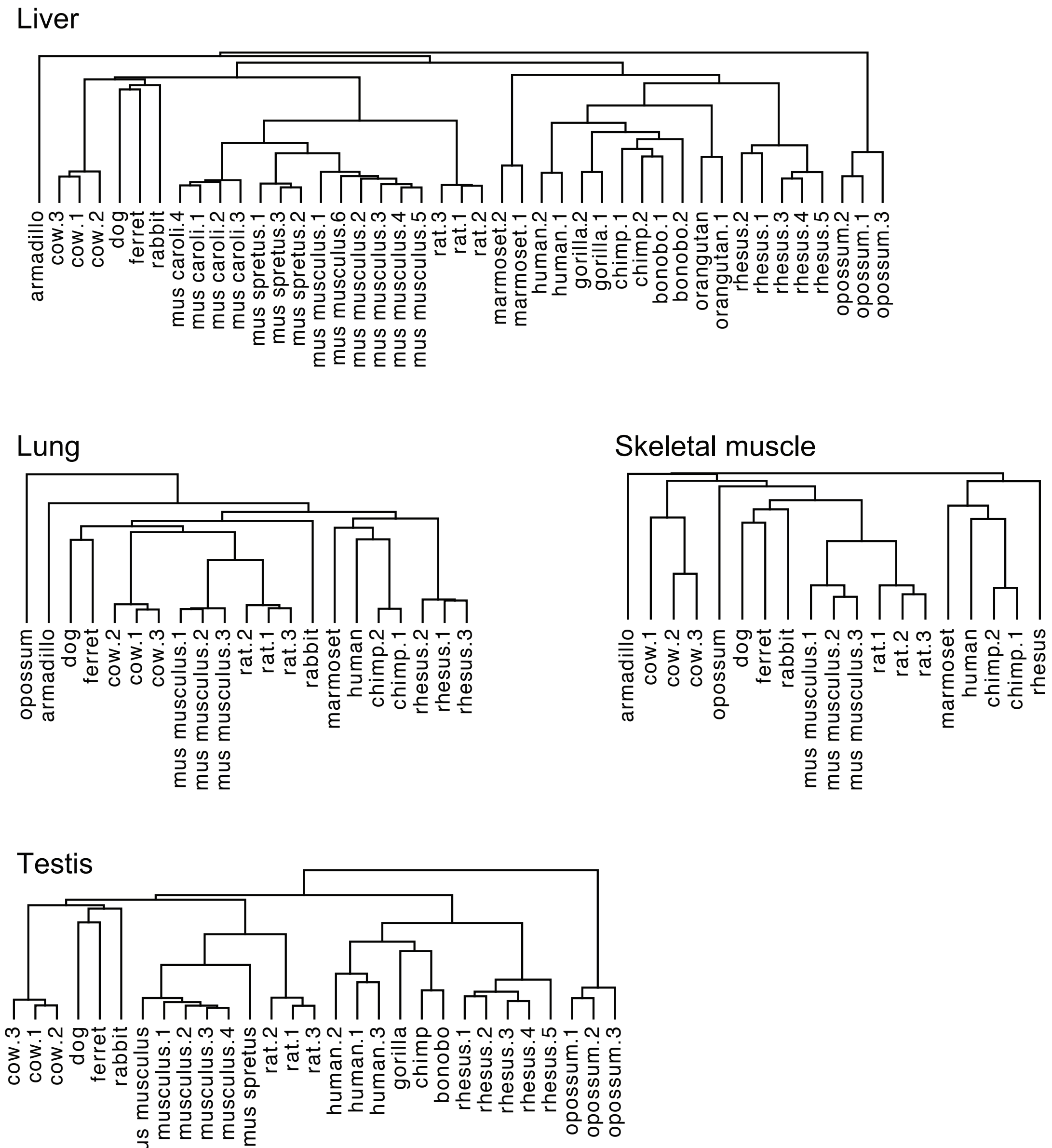


Figure A-1: Dendrograms from hierarchical clustering of gene expression ($\log_{10}$[TPM]) within each of 7 tissue type (label, top) using Pearson's correlation as the distance metric.

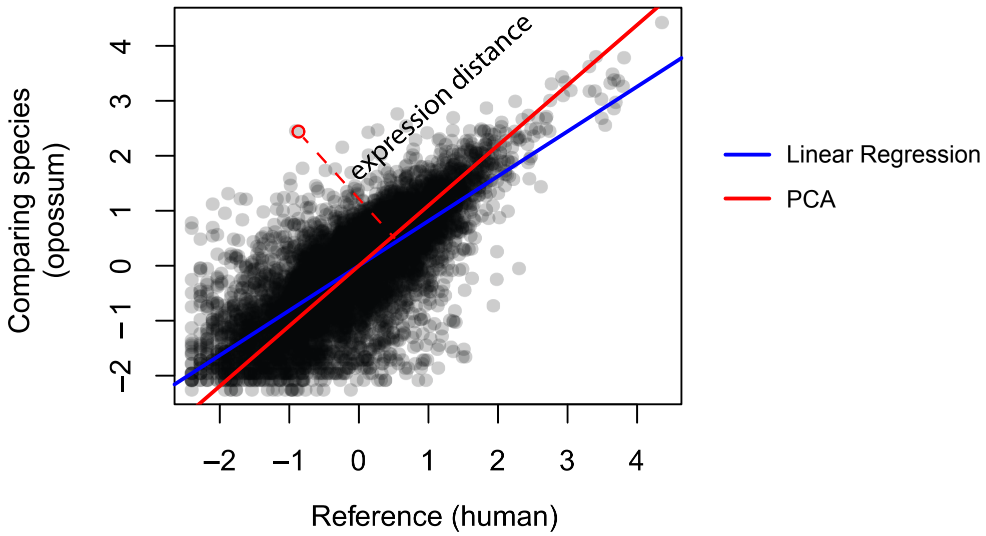
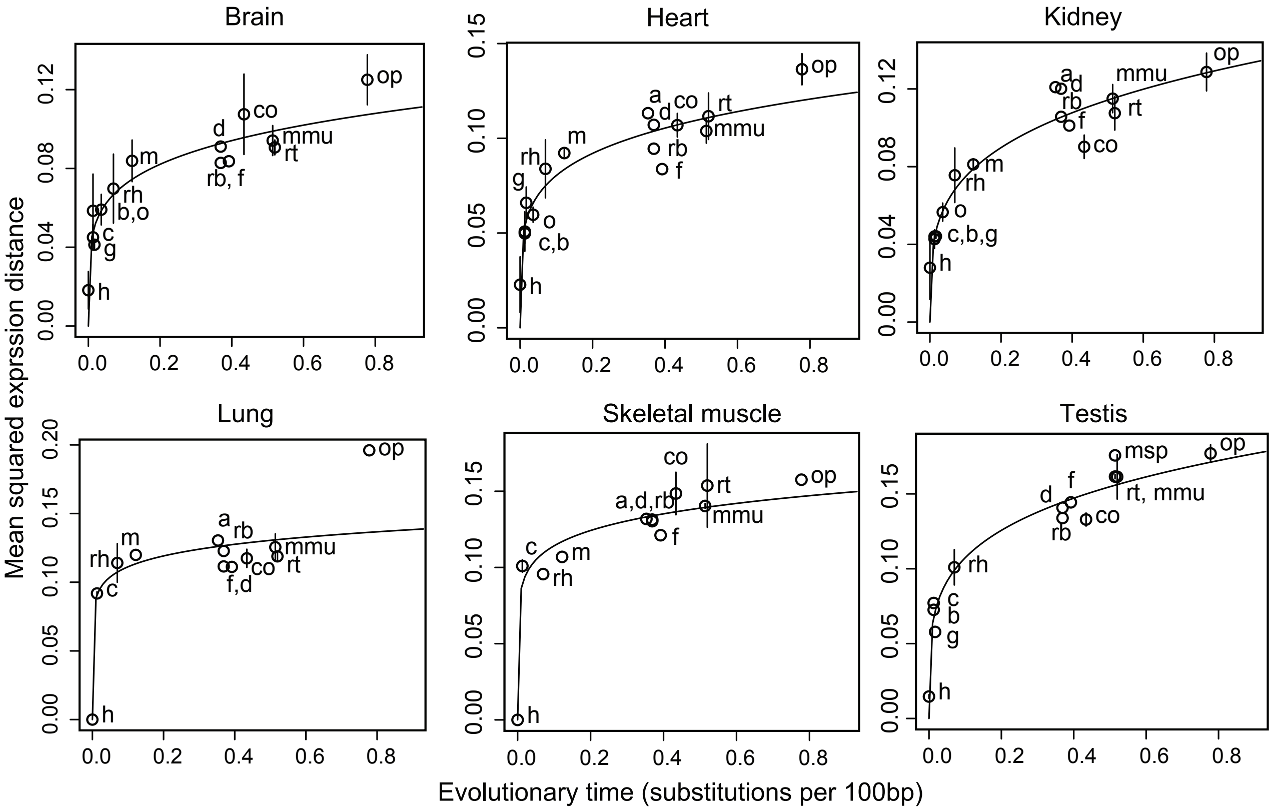
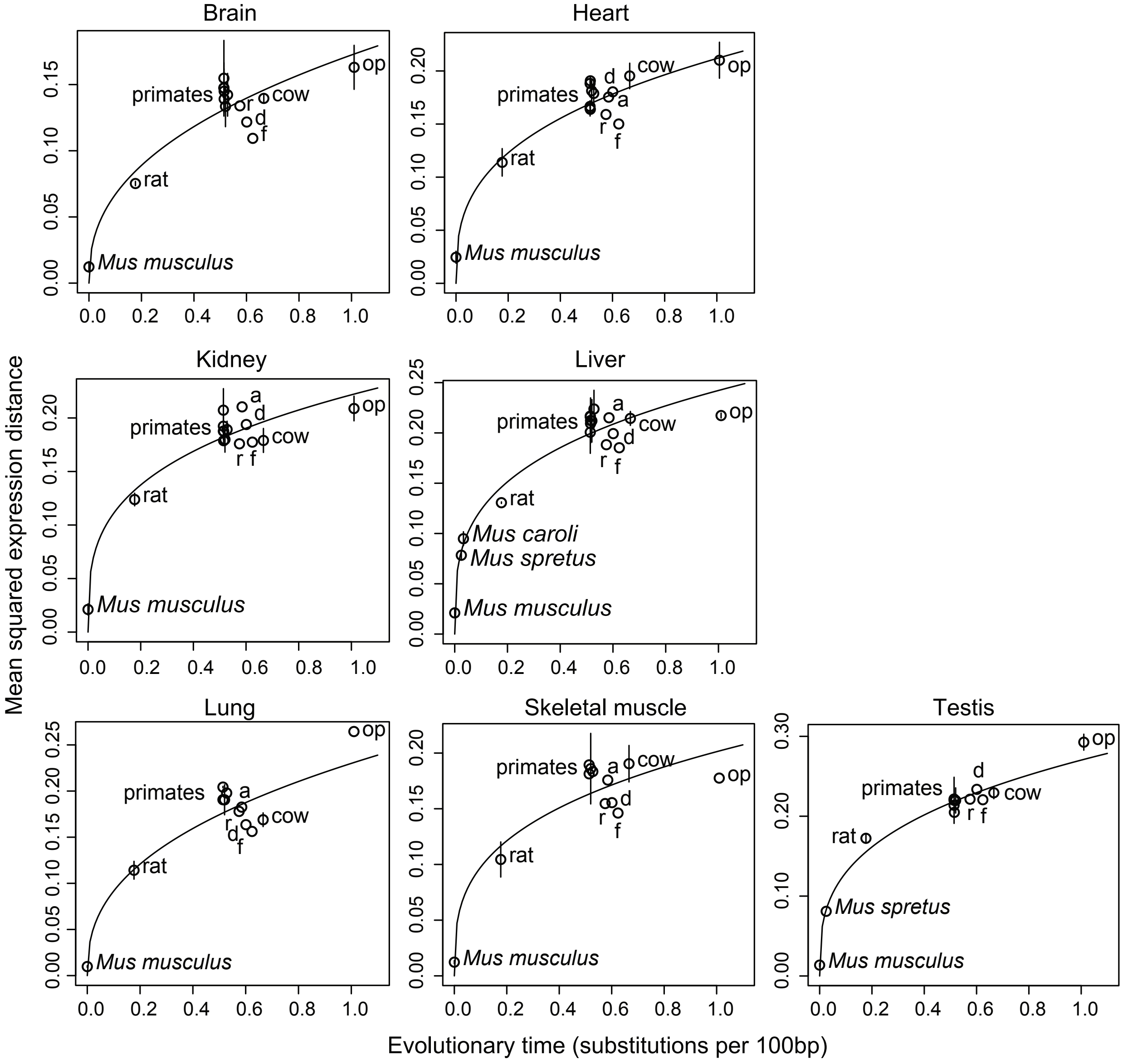


Figure A-2: Illustrative example of unnormalized expression level of all genes from a human (x-axis) and an opossum liver sample (y-axis) and the best fit regression line as calculated by a linear regression (blue line) and PCA (red line). Expression distance is the orthogonal distance to the best fit line from PCA (dashed red line).



h = human; c = chimp; b = bonobo; g = gorilla; o = orangutan; rh = rhesus macaque; m = marmoset; d = dog;
f = ferret; co = cow; a = armadillo; rb = rabbit; rt = rat; mmu = *Mus musculus*; msp = *Mus spretus*; op = opossum

Figure A-3: Pairwise mean squared expression distances (y-axis) between mammals and human for each of six tissue types across evolutionary time, as estimated by substitutions per 100 bp (x-axis). Error bars: standard deviation of the mean across replicates.



primates = human, chimp, bonobo, gorilla, orangutan, rhesus macaque, marmoset;
d = dog; f = ferret; a = armadillo; r = rabbit; op = opossum

Figure A-4: Pairwise mean squared expression distances (y-axis) between mammals and *mouse* for each of seven tissue types across evolutionary time, as estimated by substitutions per 100 basepairs (x-axis). Error bars: standard deviation of the mean across replicates.

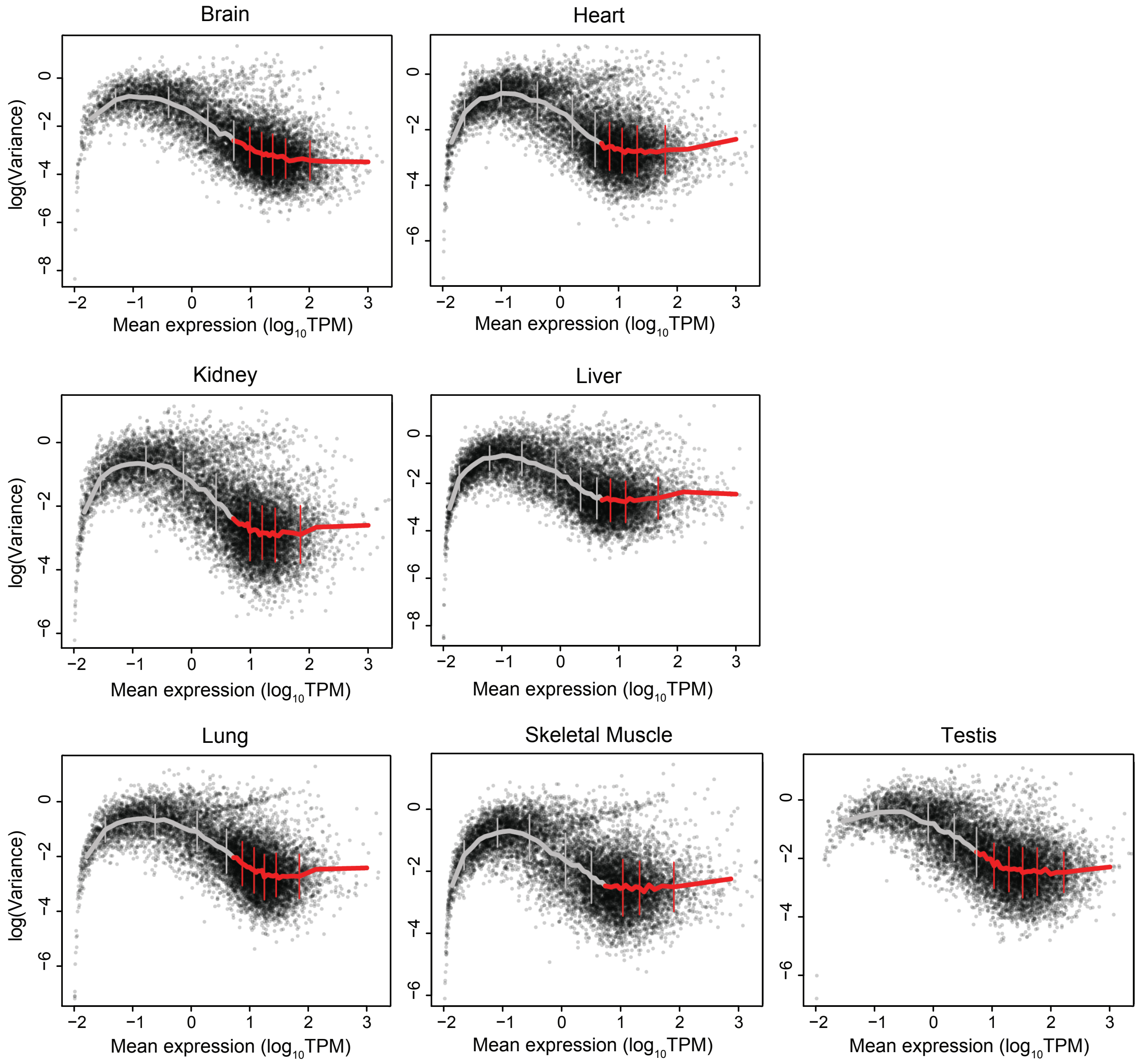


Figure A-5: Evolutionary mean (x-axis, $\log_{10}$[TPM]) and log(evolutionary variance) (y-axis) for each gene as estimated from OU model in each of 7 tissue types. Mean expression was binned into bins of 250 genes, and mean variance for each bin is shown in solid lines (grey/red: genes with expression lower/higher than 5 TPM). Error bars: standard deviation of the mean.

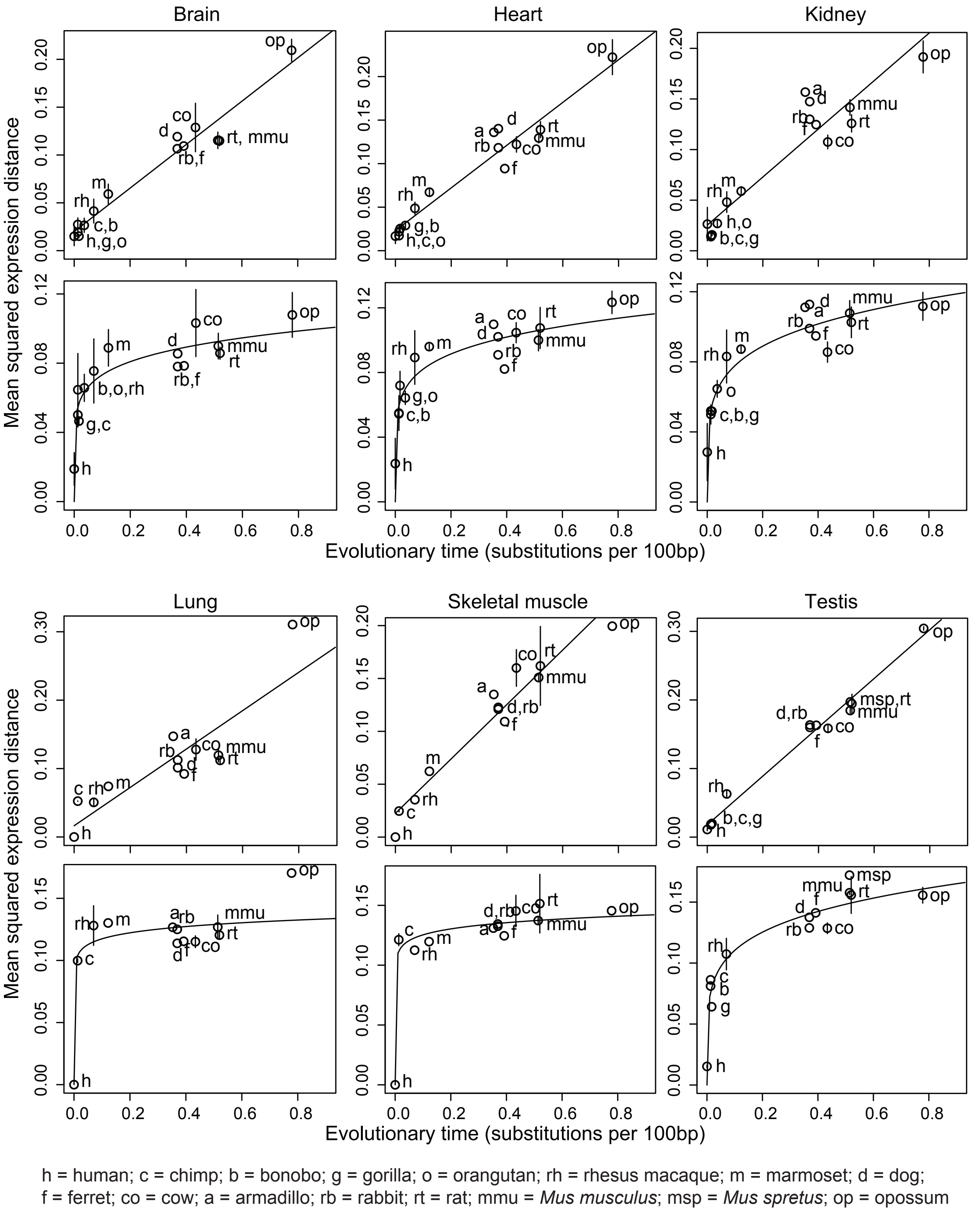
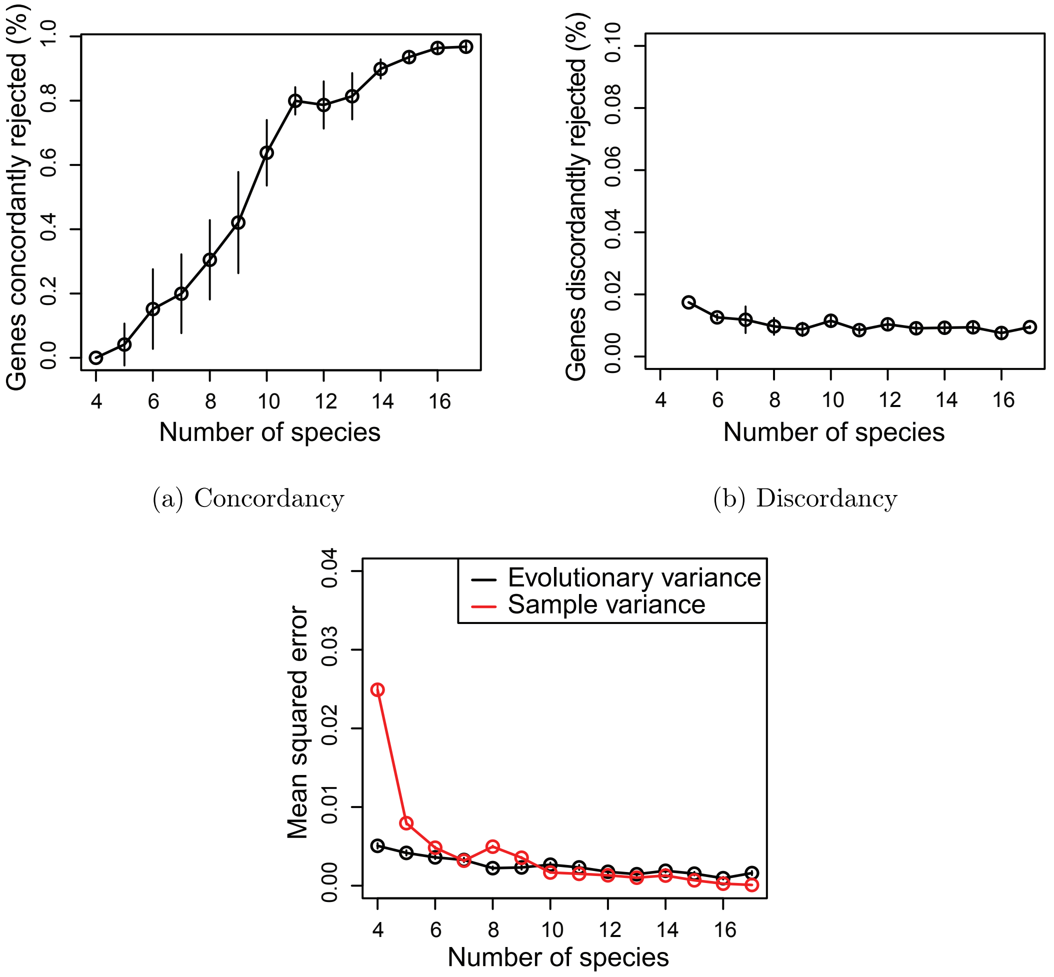


Figure A-6: Pairwise mean squared expression distances for genes whose expression evolution fits better under a BM process (i.e., neutral evolution, top) and for genes whose expression evolution fits better an OU process (i.e., presence of stabilizing selection, bottom).



(a) Concordancy

(b) Discordancy

(c) Accuracy

Figure A-7: Percent of genes concordantly rejected (top, left), percent genes discordantly rejected (top, right), and mean squared error of variance (bottom) estimated from OU model (black) or directly from the sample (red) when compared to estimates using the full phylogeny. Each metric is test on increasing number of species in the phylogeny (x-axis). Error bars: standard deviation of the mean across 10 iterations of sampled phylogenies.

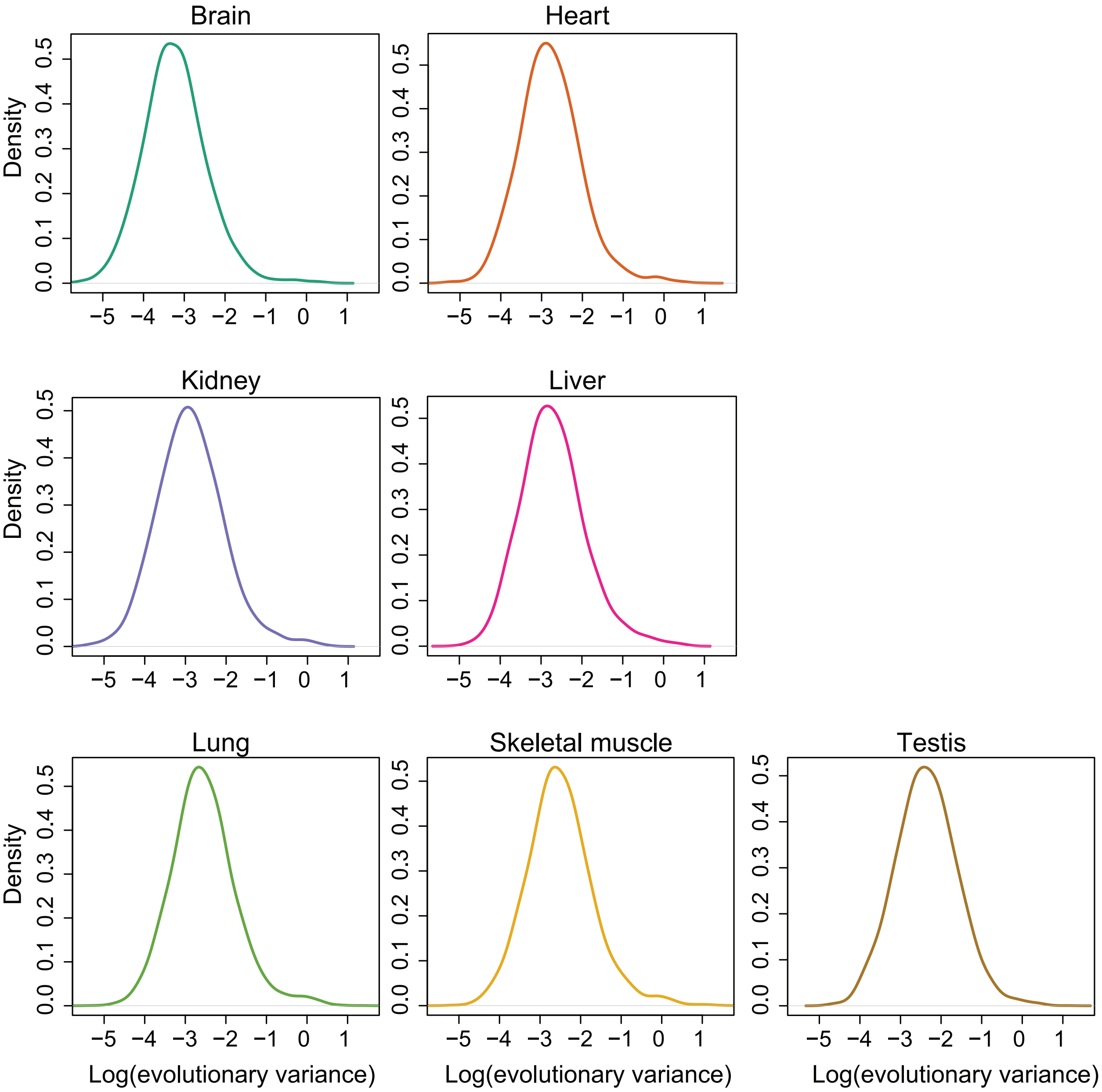
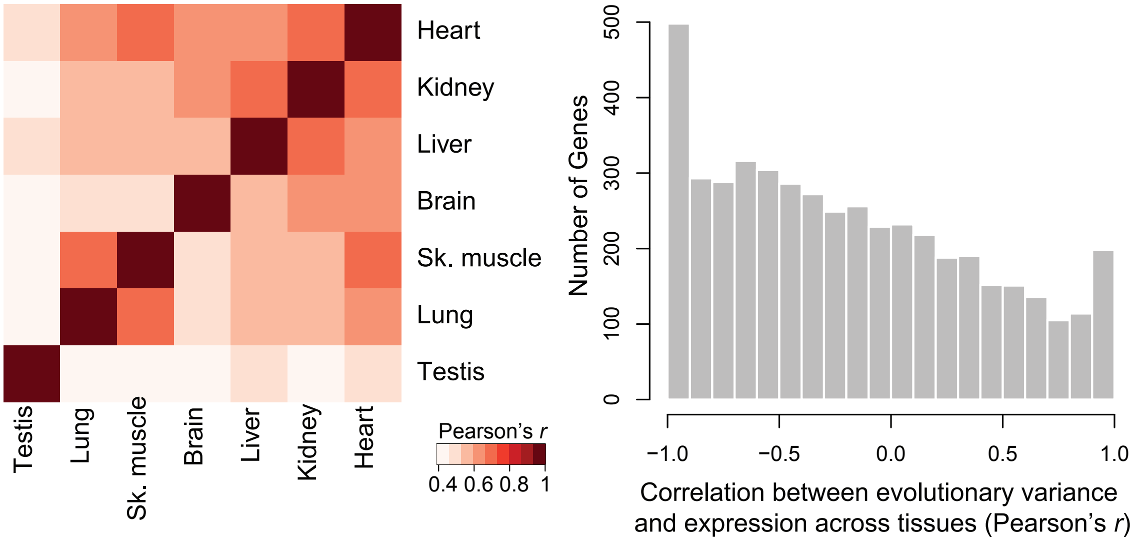


Figure A-8: Distribution of log(evolutionary variance) for all genes in each of 7 tissue types.



(a) Heatmap of Pearson's correlation coefficient (r) of evolutionary variance estimates for all expressed genes (> 5 TPM) between each pair of tissues.

(b) Distribution of Pearson's correlation coefficients (r) between variance and expression across tissues, for genes expressed (> 5 TPM) in 3 or more tissues.

Figure A-9: Correlation between evolutionary variance and expression across tissues.

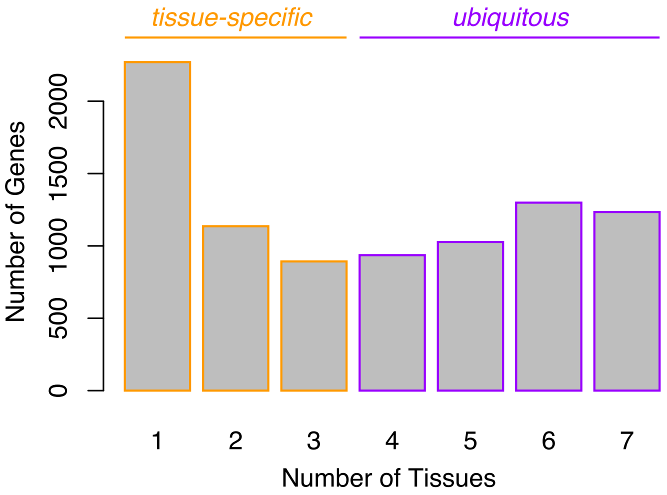


Figure A-10: Histogram of number of tissues in which genes are expressed (> 5 TPM). Genes expressed in 3 or fewer tissues were defined to be tissue-specific.

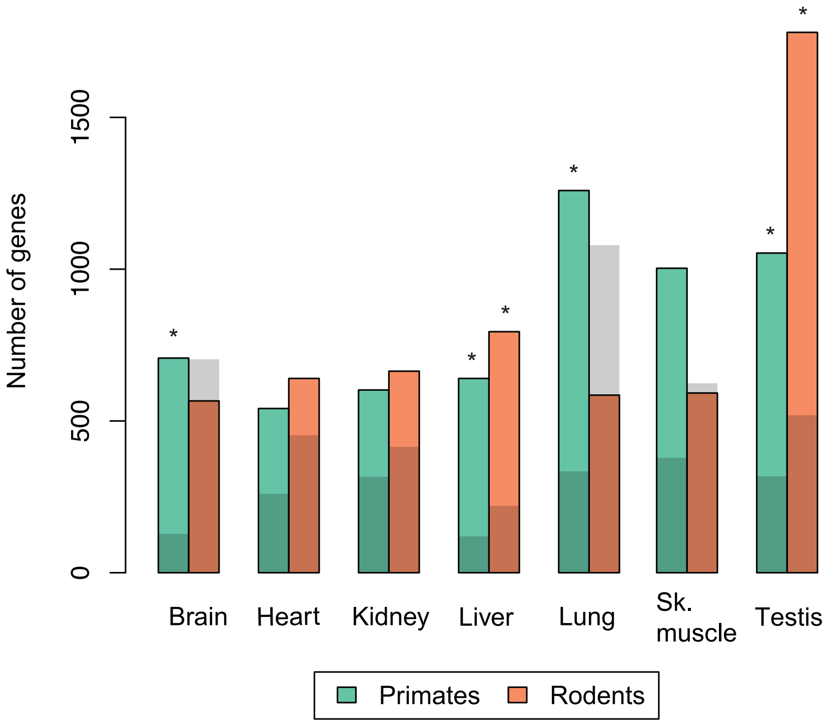


Figure A-11: Number of genes (y-axis) detected to have lineage-specific expression changes across primates (green) and rodents (orange) in each of 7 tissue types (x-axis). Gray shading: number of genes detected by the same analysis when using a shuffled phylogeny. * denotes FDR < 0.30.

Evolutionary analysis across mammals reveals distinct classes of long noncoding RNAs

This chapter is adapted from its initial publication:

Jenny Chen, Alexander A. Shishkin, Xiaopeng Zhu, Sabah Kadri, Itay Maza, Mitchell Guttman, Jacob H. Hanna, Aviv Regev and Manuel Garber. Evolutionary analysis across mammals reveals distinct classes of long noncoding RNAs. *Genome Biology*, February 2016.

Recent advances in transcriptome sequencing have enabled the discovery of thousands of long noncoding RNAs (lncRNAs) across many species. Though several lncRNAs have been shown to play important roles in diverse biological processes, the functions and mechanisms of most lncRNAs remain unknown. Two significant obstacles lie between transcriptome sequencing and functional characterization of lncRNAs: identifying truly noncoding genes from *de novo* reconstructed transcriptomes, and prioritizing the hundreds of resulting putative lncRNAs for downstream experimental interrogation. I present *slncky*, a lncRNA discovery tool that produces a high-quality set of lncRNAs from RNA-sequencing data and further uses evolutionary constraint to prioritize lncRNAs that are likely to be functionally important. My analysis reveals that evolutionary selection acts in several distinct patterns, highlighting that lncRNAs are not a homogenous class of molecules but rather a mixture of multiple functional classes with distinct biological mechanism and/or roles. Further, my novel comparative methods for lncRNAs reveals 233 constrained lncRNAs out of tens of thousands of currently annotated transcripts, whose data is available through the *slncky Evolution Browser* for downstream experimental interrogation.

3.1 Background

Recent advances in transcriptome sequencing have led to the discovery of thousands of lncRNAs, many of which have been shown to play important roles in diverse biological processes from development to immunity, or associated with numerous cancers when misregulated [73, 146–154]. Given the importance of lncRNAs in biology and disease, there is great interest in defining lncRNAs in new experimental systems, disease models, and even primary cancer samples. Yet, despite important progress in RNA-sequencing (RNA-seq), the annotation and computational characterization of lncRNAs from RNA-seq data remains a major challenge.

Current computational approaches for filtering lncRNAs from RNA-seq transcript assemblies are largely based on the absence of evolutionarily signatures of protein-coding potential [78,79,82,155]. Yet, this approach is limited in both sensitivity and specificity: (1) it incorrectly classifies *bona fide* lncRNAs as protein-coding simply because they are conserved; and (2) it incorrectly classifies transcripts as lncRNAs when they are actually fragmented untranslated regions (UTRs) of coding genes, pseudogenes, or members of lineage-specific protein-coding gene family expansions, such as zinc finger proteins or olfactory genes. Other lncRNA cataloging efforts have addressed these issues by incorporating additional filtering criteria along with extensive manual curation [79,155,156] or by performing additional experiments to better capture transcript boundaries (e.g., 5'- or 3'-end sequencing) [82,157]. While these approaches have proven to be extremely valuable, they remain labor-intensive and time-consuming, even for experienced users.

To address this challenge, I developed *slncky*, a method and accessible software package that enables robust and rapid identification of high-confidence lncRNA catalogs directly from RNA-seq transcript assemblies without reliance on evolutionary measures of coding potential. *slncky* goes through several key steps to accurately separate lncRNAs from coding genes, pseudogenes, and assembly artifacts, while also identifying novel proteins including small peptides. When applied to mouse embryonic stem cells (ESCs), *slncky* accurately identifies virtually all well-characterized lncRNAs and performs as well as previous manually curated catalogs.

Downstream of lncRNA annotation, comparative analysis remains an important computational approach to assess potential function of a lncRNA without requiring additional experimental efforts. Despite its importance, few tools for identifying conservation of lncRNAs exist. To address this need, *slncky* additionally incorporates a comparative analysis pipeline with novel metrics especially designed for the study of RNA evolution. Here, I demonstrate the utility of *slncky* by applying it to a comparative study of the ESC transcriptome across human, mouse, rat, chimpanzee, and bonobo, and to previously defined datasets consisting of $>$ 700 RNA-seq experiments across human and mouse. When applying *slncky* to these datasets, I discover hundreds of conserved lncRNAs. Furthermore, my metrics for evaluating transcript evolution show that evolutionary properties divide lncRNAs into separate classes, which each display distinct patterns of selective pressure.

3.2 A method for defining a high-quality set of long noncoding RNAs

Determining a set of lncRNAs from reconstructed annotations involves several steps to ensure that transcripts represent complete transcriptional units and that they are unlikely to encode for a protein. Current methods for defining coding potential, such as PhyloCSF [15] and RNACode [16], rely on codon substitution models which fail in three important cases: (1) they often misclassify noncoding RNAs as protein-coding — including *TUG1*, *MALAT1*, and *XIST* — merely because they are conserved; (2) they fail to identify lineage-specific proteins as coding; and (3) they erroneously identify noncoding elements (e.g., UTR fragments, intronic reads) as lncRNAs. Rather than use codon substitution models, *slncky* implements a set of sensitive filtering steps to exclude fragment assemblies, UTR extensions, gene duplications, and pseudogenes, which are often mischaracterized as lncRNAs, while also avoiding the exclusion of *bona fide* lncRNA transcripts that are often excluded simply because they have high evolutionary conservation.

To achieve this goal, *slncky* carries out the following steps (Fig. 3-1, Methods): (1) *slncky* removes any transcript that overlaps (on the same strand) any portion of an annotated protein-coding gene in the same species; (2) *slncky* leverages the conservation of coding genes and uses annotations in related species to further exclude unannotated protein-coding genes, or incomplete transcripts that align to UTR sequences; and (3) to remove poorly annotated members of species-specific protein-coding gene expansions, *slncky* aligns all identified transcripts to each other and removes any transcript families that shares significant homology with each other. The result is a filtered set of transcripts that retains conserved, noncoding transcripts that may score highly for coding potential, while excluding up to approximately 25% of coding or pseudogenic transcripts normally identified as lncRNAs by traditional approaches.

After removing reconstructions that are likely gene fragments, pseudogenes, or members of gene family expansions, *slncky* searches for novel or previously unannotated coding genes using a method that is less confounded by evolutionary conservation than codon substitu-

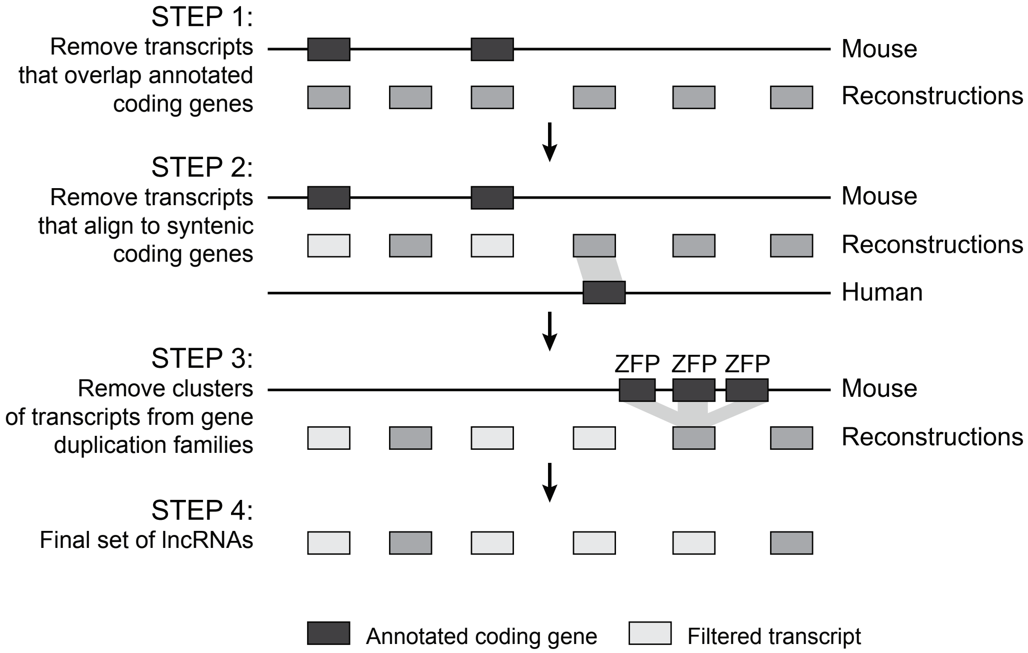

STEP 1:
Remove transcripts that overlap annotated coding genes
Mouse
Reconstructions
STEP 2:
Remove transcripts that align to syntenic coding genes
Mouse
Reconstructions
Human
STEP 3:
Remove clusters of transcripts from gene duplication families
ZFP
ZFP
ZFP
Mouse
Reconstructions
STEP 4:
Final set of lncRNAs
Annotated coding gene
Filtered transcript

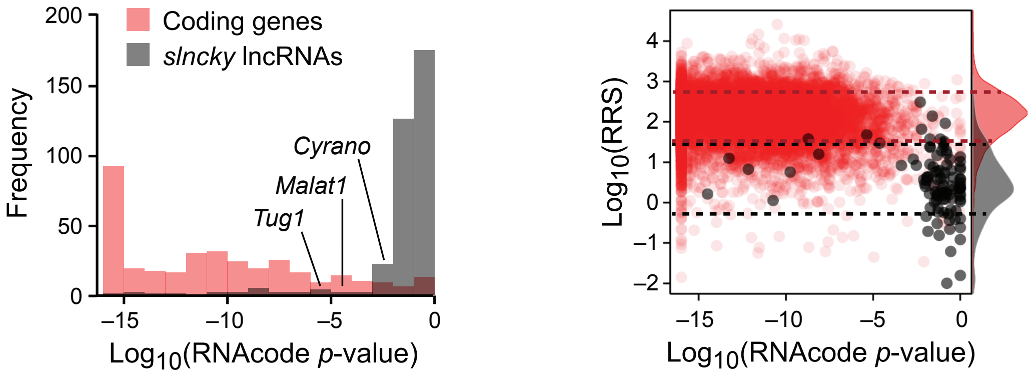
Figure 3-1: Schematic of *slncky*'s filtering pipeline.

tion models (Methods). Specifically, *slncky* aligns orthologous transcripts and analyzes all possible open reading frames (ORFs) that are present in both species. For each ORF, *slncky* computes the ratio of nonsynonymous to synonymous mutations (dN/dS) and excludes all annotations with a significant dN/dS ratio. By requiring the presence of a conserved ORF that is transcribed in multiple species, and by computing the dN/dS ratio across the entire ORF alignment, *slncky* is more specific than conventional coding-potential scoring software which report all high-scoring segments within an alignment.

3.2.1 Validation with mouse embryonic stem cell lncRNAs

Having developed a method to identify lncRNAs directly from RNA-seq data, I sought to characterize its sensitivity and specificity by comparing lncRNAs identified by *slncky* to the well-studied set of lncRNAs expressed in mouse ESCs [78]. To do this, I used RNA-seq libraries from pluripotent cells obtained from three different mouse strains cultured using previously described growing conditions [158,159] (Methods, Table B.1). I then performed *de novo* reconstruction to build transcript models (Methods), and subsequently applied *slncky* to define a set of 408 lncRNAs (Fig. B-1). The analysis also revealed four transcripts — *Apela*, *Tunar*, *1500011K16Rik* (*LINC00116*), and *BC094334* (*LINC00094*) — that contain conserved ORFs with high coding potential (Fig. B-2a, Table B-2b).

Several lines of evidence indicate that the identified set represents *bona fide* lncRNAs: (1) *slncky* recovered all of the 20 well-characterized lncRNAs that are expressed in the pluripotent state (Methods), demonstrating that this stringent approach is still sensitive; (2) Identified lncRNAs contain chromatin modifications of active RNA Polymerase II transcription (Methods), exhibiting similar levels as previous ES catalogs (approximately 70%) [78, 160]; (3) lncRNAs identified by *slncky* have significantly lower evolutionary coding potential scores than protein-coding genes (t-test $p = 1.3 \times 10^{-6}$, Methods) (Fig. 3-2a); (4) *slncky* does not filter out known conserved lncRNAs, such as *Malat1*, *Tug1*, *Miat*, that are often excluded due to significant coding-potential scores (Fig. B-2a, Table B-2c); and (5) Identified lncRNAs have significantly reduced ribosome release scores [161] (Methods), a measure that accurately predicts coding potential from ribosome profiling data, than protein-coding genes (73-fold, t-test $p < 2.2 \times 10^{-16}$) (Fig. 3-2b).



(a) Histogram of $\log_{10}$(p-values) of coding potential as evaluated by RNACode [81] for *slncky* lncRNAs (gray) and coding genes (red).

(b) Scatterplot of $\log_{10}$(p-values) of coding potential (x-axis) and $\log_{10}$(ribosomal-release scores [RRS]) (y-axis) of *slncky* lncRNAs (gray) and coding genes (red). Distributions of RRS are shown along right side of y-axis; Dotted lines denote one standard deviation.

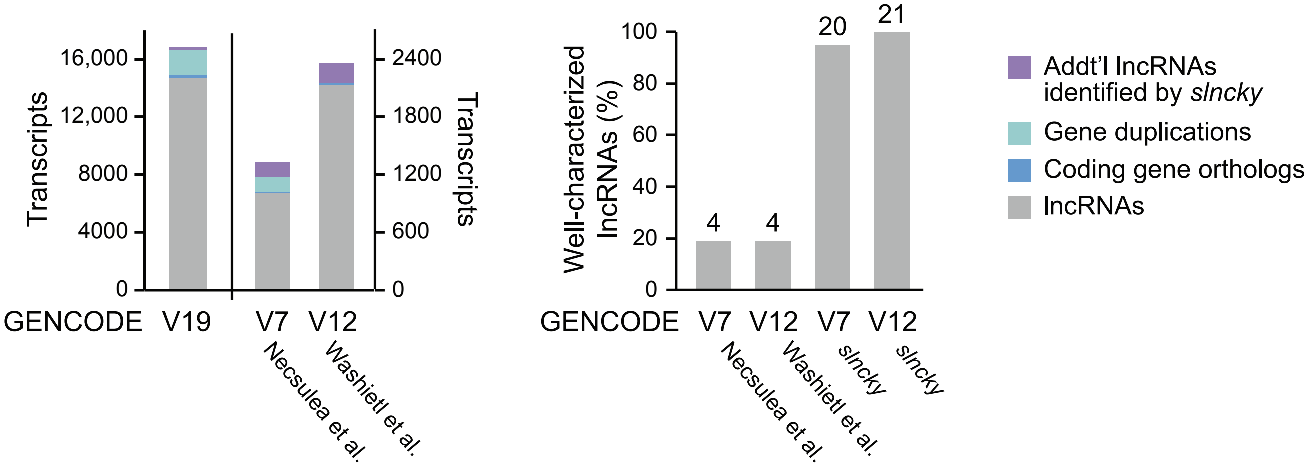
Figure 3-2: Coding potential and ribosomal-release scores of *slncky* lncRNAs and coding genes.

These results demonstrate that *slncky* provides a simple and robust strategy for identifying lncRNAs from a *de novo* transcriptome. Rather than requiring many user-defined parameters, *slncky* learns filtering parameters directly from the data making it useful across many different species, including non-model organisms.

3.2.2 Comparison to previous methods

To verify the scalability and overall utility of *slncky* for defining lncRNAs across multiple datasets in different species, I ran *slncky* on GENCODE's latest comprehensive gene annotation set (V19) totaling 189,020 transcripts, of which 16,482 are annotated as lncRNAs that do not overlap a coding gene [156]. GENCODE is an ideal test case because it represents the current gold standard lncRNA annotation set, as much of its content undergoes extensive manual curation. Applying *slncky*, I identified 14,722 human lncRNA genes. Importantly, these include $> 90\%$ of the lncRNAs identified by GENCODE, with only 136 human (0.9%) annotated protein coding gene, and 83 (0.6%) annotated pseudogenes identified as lncRNAs. Transcripts that are annotated as lncRNAs by GENCODE but not by *slncky* include 1,735 (12%) transcripts that are part of a cluster of duplicated genes, of which 123 (1%) aligned to a known zinc finger protein or olfactory gene. An additional 181 (1%) transcripts were excluded because they aligned significantly to an orthologous protein coding gene in mouse (Fig. 3-3a).

I then compared my filtering strategy with two previously published large-scale comparative studies that were based on GENCODE annotations, Washietl et al. [81] and Necsulea et al. [80] (Methods). For the set of lncRNAs defined by Washietl et al., *slncky* was able to remove 9.6% (156) of the annotations that were likely results of gene duplications and 1.2% (19) that aligned significantly to a mouse coding transcript. In contrast, *slncky* only removed a handful of transcripts ($< 0.1\%$) from the Necsulea et al. dataset. Importantly, *slncky* was much more sensitive as it identified virtually all well-characterized lncRNAs (20/21) compared to only 20% (4/21) by these previous reports (Fig. 3-3b). Finally, I compared *slncky* to a recently published pipeline for filtering reconstructed transcripts from RNA-seq data called PLAR (Hezroni et al. [82]). I found that *slncky* and PLAR performed comparably in removing coding gene orthologs and gene duplications, but *slncky* remained more sensitive in recovering well-characterized transcripts (33/36 recovered by *slncky* compared to 27/36 by PLAR) (Fig. B-3).



(a) Number of transcripts that *slncky* annotates as a lncRNA (gray), removes as gene duplication or coding gene (light and dark blue), and additionally identifies as novel lncRNAs (purple).

(b) Percentage of well-characterized lncRNAs identified in previously published sets compared to *slncky* results. Numbers denote absolute number of lncRNAs.

Figure 3-3: Comparison of previously published sets of lncRNAs to *slncky* results.

Together, these results highlight the power of *slncky* for identifying a high-confidence set of lncRNAs and excluding known artifacts that are often mistaken for lncRNAs. Furthermore, these results demonstrate that *slncky* performs as well as manual curation for defining *bona fide* lncRNAs and can even identify the challenging cases that are often missed by curation efforts.

3.3 A method for studying lncRNA evolution

Having developed a method to define a high-quality set lncRNAs, I sought to study the evolutionary properties of lncRNAs. While comparative genomics has provided important insights for studying proteins, enhancers, and promoters [7,11,162–165], relatively few comparative methods have been developed to study the evolution of lncRNAs. One of the main challenges is that lncRNAs diverge rapidly, accumulating both nucleotide substitutions and insertion/deletion (indel) events, rendering lncRNAs difficult to align with conventional aligners and phylogenetic approaches.

To enable evolutionary analysis of lncRNAs, I implemented a computationally efficient and sensitive strategy to align lncRNAs and characterize their sequence and transcript evolution (Fig. 3-4, Methods). To this end, *slncky* identifies the syntenic genomic region for a lncRNA in the orthologous species. If a transcript exists in a syntenic region, *slncky* aligns the two regions using a sensitive seed-based local pairwise aligner [166]. To avoid the possibility of spurious matches, *slncky* scores each alignment relative to a set of random intergenic regions from the orthologous genome and only keeps alignments that score higher than 95% of the random intergenic sequences.

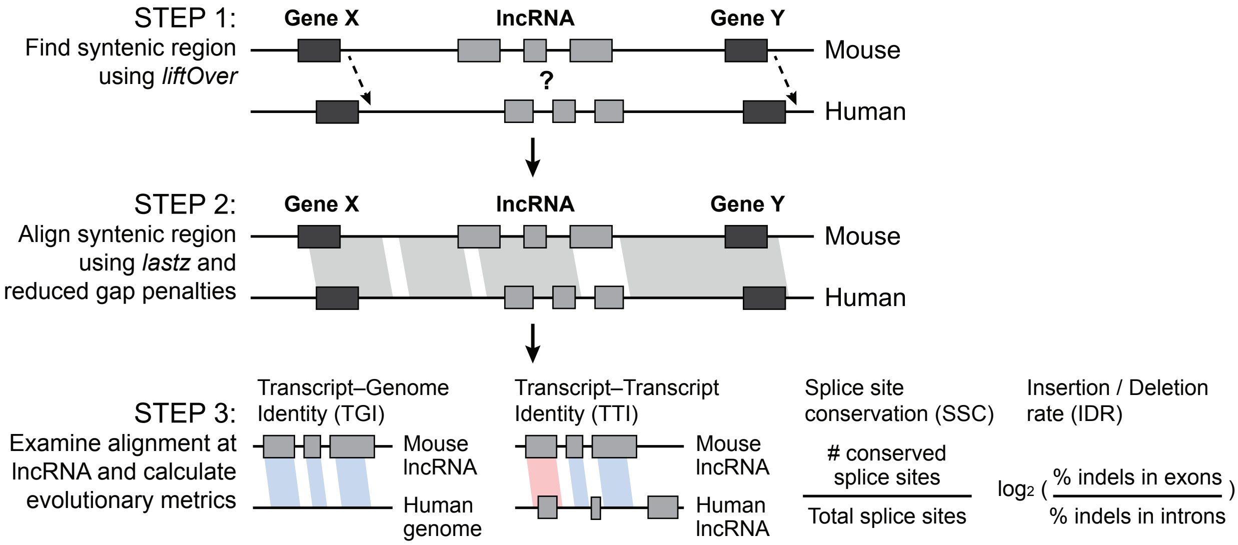

STEP 1:
Find syntenic region using *liftOver*
Gene X
lncRNA
Gene Y
Mouse
?
Human
STEP 2:
Align syntenic region using *lastz* and reduced gap penalties
Gene X
lncRNA
Gene Y
Mouse
Human
STEP 3:
Examine alignment at lncRNA and calculate evolutionary metrics
Transcript–Genome Identity (TGI)
Mouse lncRNA
Human genome
Transcript–Transcript Identity (TTI)
Mouse lncRNA
Human lncRNA
Splice site conservation (SSC)
# conserved splice sites
Total splice sites
Insertion / Deletion rate (IDR)
log2 (% indels in exons / % indels in introns)

Figure 3-4: Schematic of *slncky*'s orthology pipeline and metrics for measuring sequence and transcript evolution.

3.3.1 Novel metrics for quantifying lncRNA evolution

Next, *slncky* characterizes sequence and transcript conservation properties of orthologous lncRNAs. *slncky* calculates four metrics (Fig. 3-4):

1. A **'transcript-genome identity' (TGI)** score, defined as the percent of lncRNA base pairs that align and are identical to a syntenic genomic locus, to characterize how well the transcript sequence is conserved across the two species;

2. A **'transcript-transcript identity' (TTI)** score, defined as the percent of identical, aligning base pairs found in the transcribed, exonic regions of both lncRNAs, to characterize how much of the transcript is transcribed in both species;

3. A **'splice site conservation' (SSC)** score, defined as the percent of splice sites that are conserved across both lncRNAs, to characterize conservation of transcript structure; and

4. An **'insertion/deletion rate' (IDR)**, defined as the $\log_2$ rate of insertion/deletion events in exonic regions relative to intronic regions, to provide an alternative measure of sequence conservation.

3.3.2 Comparison to previous methods

I tested the performance of *slncky*'s orthology finding step by reanalyzing previous studies of lncRNA conservation across mammals [81] and vertebrates [80, 82, 157] (Methods). The approach of aligning the two syntenic loci rather than just the transcripts increases *slncky* sensitivity with very little drop in specificity. In mammals, *slncky* successfully identified the vast majority ($>$ 95%, 1,466/1,521 lncRNAs) of the previously reported orthologous lncRNAs while also finding an additional 121 pairs (8.0%) of homologous human-mouse lncRNAs that were previously reported as species-specific. Similarly, in vertebrates, a four-fold greater evolutionary distance, *slncky* was able to recover 26 of 29 (90%) of the previously defined ancestral lncRNAs; the alignments for the remaining three, although found, are indistinguishable from alignments that can be randomly found across syntenic loci and do not pass the significance threshold. Furthermore, *slncky* identified an additional three pairs of vertebrate conserved lncRNAs.

Together, these results demonstrate that *slncky* provides an efficient, sensitive, and accessible method for detecting and characterizing orthologous lncRNAs across any pair of species, providing an important tool for studying lncRNA evolution or for prioritizing lncRNAs based on evolutionary conservation.

3.4 Studies of mammalian lncRNA evolution

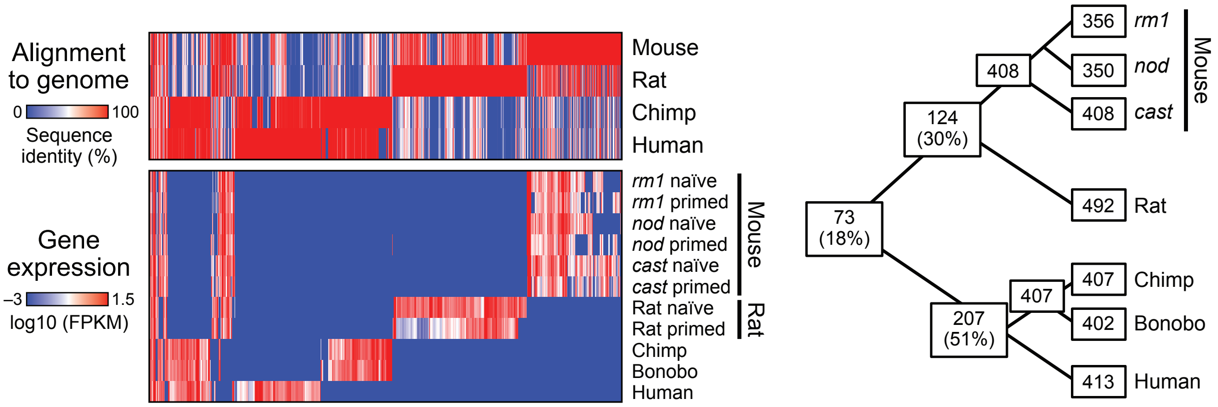
Initial work by others suggests that the expression of lncRNAs is often poorly conserved — with the rate of transcript expression loss occurring faster than loss of its genomic sequence identity across species [80, 81]. While these results provided important insights into the evolution of lncRNAs, these analyses did not fully explore the properties of the conserved lncRNAs. Having developed a method to comprehensively identify and align lncRNAs across species, I sought to further understand the evolutionary properties of lncRNAs. To do this, I used RNA-seq data from ESCs derived from three mouse strains (*129SvEv*, *NOD*, and *castaneous*), rat, and human (Methods). I added additional published RNA-seq data for chimpanzee and bonobo iPS cells [167] (Table B.1). The gene expression across mammalian pluripotent cells shows a similarly high correlation to that previously observed for matched somatic tissues across mammals (Fig. B-4), highlighting the suitability of this set for comparative analysis.

Applying *slncky*, I identified 408 mouse, 492 rat, 407 chimpanzee, and 413 human lncRNAs (Fig. B-1, Additional file 2). Consistent with previous work, I found that lncRNAs are generally expressed only in a single species, despite the fact that most lncRNA loci can be aligned across species (Fig. 3-5a). In all, I found 73 (18%) lncRNAs that are expressed in pluripotent cells across all mammals and are likely to be present prior to the divergence between rodents and primates (Fig. 3-5b, Methods, Additional file 3).

3.4.1 Distinct evolutionary signatures of ESC lncRNAs

Like previous catalogs, the set of pluripotent-expressed lncRNAs fall into different classes: **miRNA host genes**, **snoRNA host genes**, **divergently expressed lncRNAs** that are transcribed in the opposite orientation of a coding gene with which they share a promoter, and a remaining set of **'intergenic' lncRNAs** (lincRNAs). Interestingly, I found that these classes have distinct patterns of sequence and transcript evolution (Fig. 3-6a):

- While the loci of **miRNA host genes** can readily be aligned across species (i.e., have high TGI), their transcript structure have diverged tremendously, with only 8.5% median TTI across humans and mouse.



(a) Sequence identity (top heatmap) and expression level (bottom heatmap) of each lncRNA loci (columns) across syntenic regions of mammalian genomes.

(b) Number of lncRNAs found within each species and at each ancestral node.

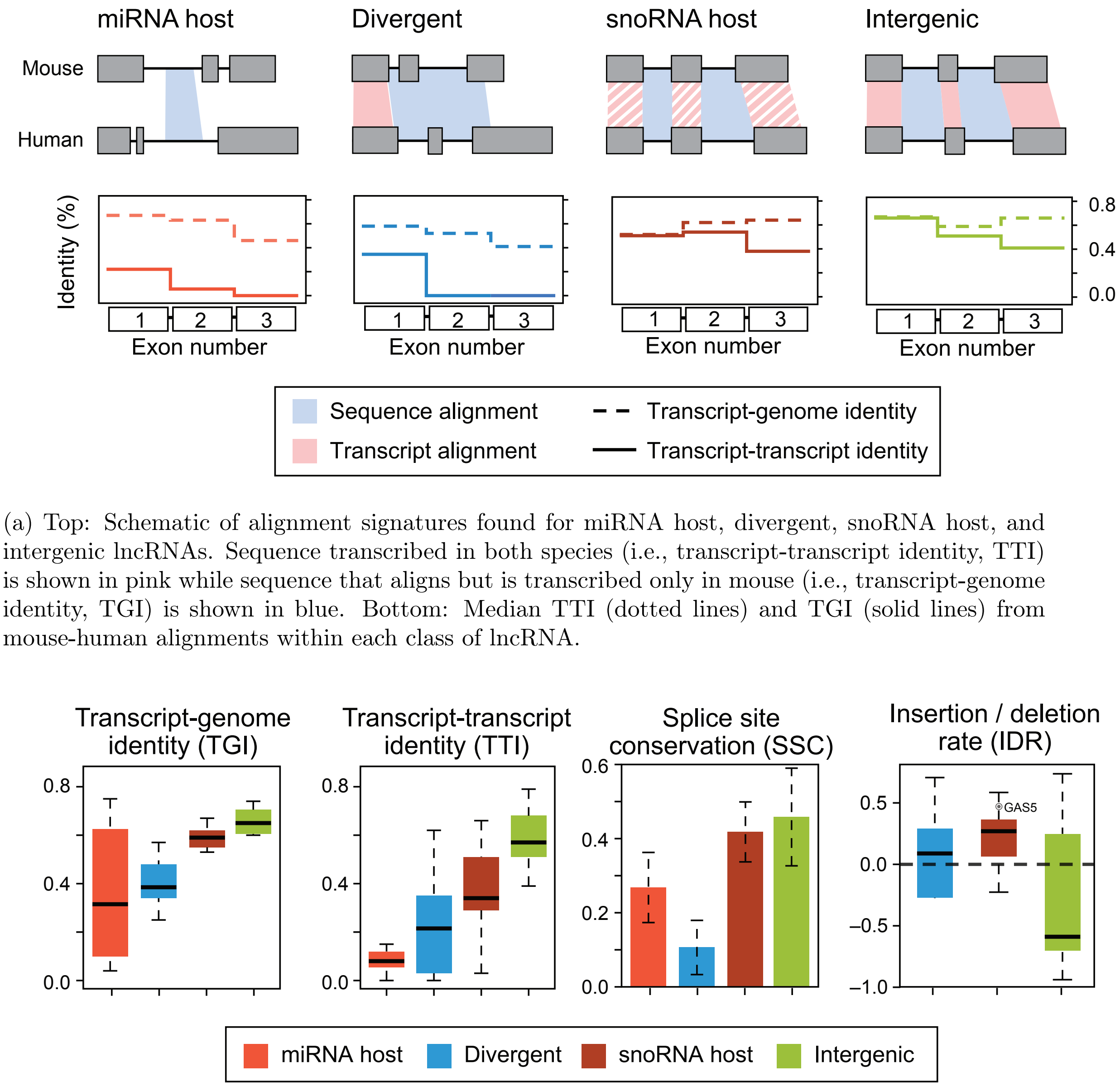
Figure 3-5: Conservation of lncRNA sequence and expression across mammalian ESCs.

- **lncRNAs divergently transcribed** within 500 base pairs of a coding gene have also diverged rapidly in TTI, except for sequence transcribed near the promoter. For these genes, TTI is generally confined to the first exon.

- **snoRNA host transcripts** are very well conserved in both sequence and transcript structure, though they contain an excess of indel events in exons (1.2-fold more) as compared to introns (Fig. 3-6b).

- Finally, **intergenic lncRNAs (lincRNAs)** also have conserved transcript structure but a 1.5-fold reduction in exonic indel events compared to snoRNA hosts (Fig. 3-6b), despite comparable intronic indel rates (Fig. B-5), suggesting that they undergo different selective pressure than host genes. Most of the pluripotent-expressed, well-characterized lncRNAs are found in this class of lincRNAs, which displays high TTI and splice site conservation (SSC). Two notable exceptions to the class of lincRNAs are *FIRRE* and *TSIX*, which have very poor TTI (5% and 0.1%, respectively). Both lincRNAs have been previously reported as ‘conserved in synteny’ only [82, 168], indicating that they may belong to a different class of lincRNAs.

In addition to distinct differences in conservation of transcript structure, I found that the turnover of transcription differ across lncRNA classes: the majority of miRNA host and



(a) Top: Schematic of alignment signatures found for miRNA host, divergent, snoRNA host, and intergenic lncRNAs. Sequence transcribed in both species (i.e., transcript-transcript identity, TTI) is shown in pink while sequence that aligns but is transcribed only in mouse (i.e., transcript-genome identity, TGI) is shown in blue. Bottom: Median TTI (dotted lines) and TGI (solid lines) from mouse-human alignments within each class of lncRNA.

(b) Evolutionary metrics of each class of lncRNA.

Figure 3-6: Evolutionary signatures reveal distinct functional classes of ESC lncRNAs.

snoRNA host genes show conserved transcription across mammals (95% and 87%, respectively), whereas only a small percentage of divergent and intergenic genes show conserved transcription (22% and 7%, respectively, Fig. 3-7).

Some lncRNAs have been proposed to have dual functions and these novel evolutionary metrics allow us to further explore this possibility. For example, *GAS5* is a known snoRNA

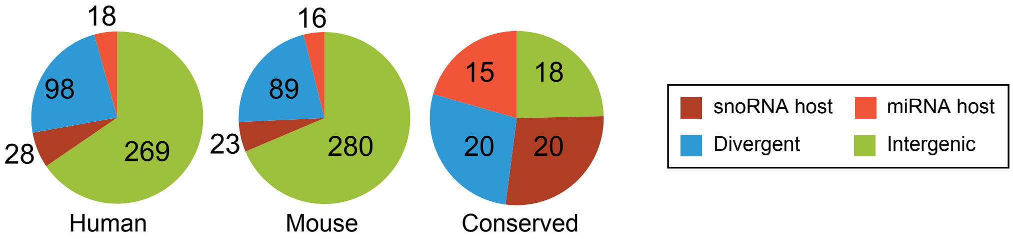


Figure 3-7: Number of mouse, human, and mammalian-conserved transcripts within each lncRNA class.

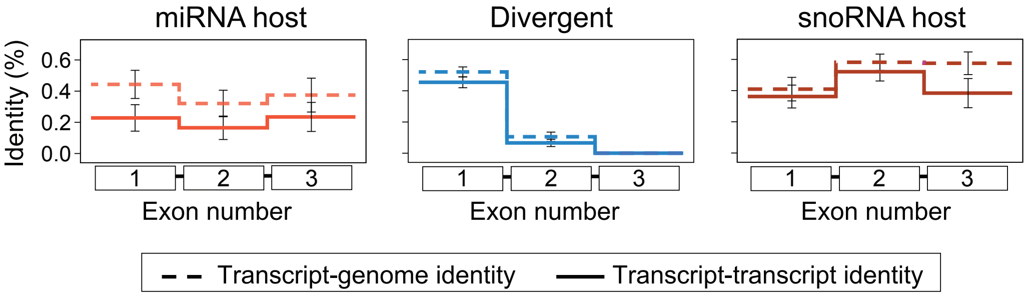
host gene and has also been reported to function as a RNA gene [169]. Interestingly, I found that *GAS5* does not match the evolutionary profile of an intergenic gene but rather has the typical signature of a snoRNA host, with higher indel rates at exons relative to its intronic regions (1.4-fold higher) (Fig. 3-6b, Additional file 3). This suggests that if *GAS5* is truly functional as a noncoding gene, it likely acts through a different mechanism than other intergenic lncRNAs.

I further note that these distinct signatures of evolution are robust enough to identify incorrectly annotated transcripts. For example, based on current annotations, *LINC-PINT* is an 'intergenic' lncRNA as the closest annotated coding gene, *MKLN1*, begins approximately 184 kb downstream [170]. However, its transcriptional conservation pattern is typical of a divergent transcript, with transcriptional identity confined only to its first exon. Closer inspection of expression data from ESCs and other tissues [65] revealed that in fact, an unannotated, alternative transcriptional start site of *MKLN1* begins less than 200 base pairs downstream, suggesting that *LINC-PINT* is in fact a divergently transcribed lncRNA (Fig. B-6).

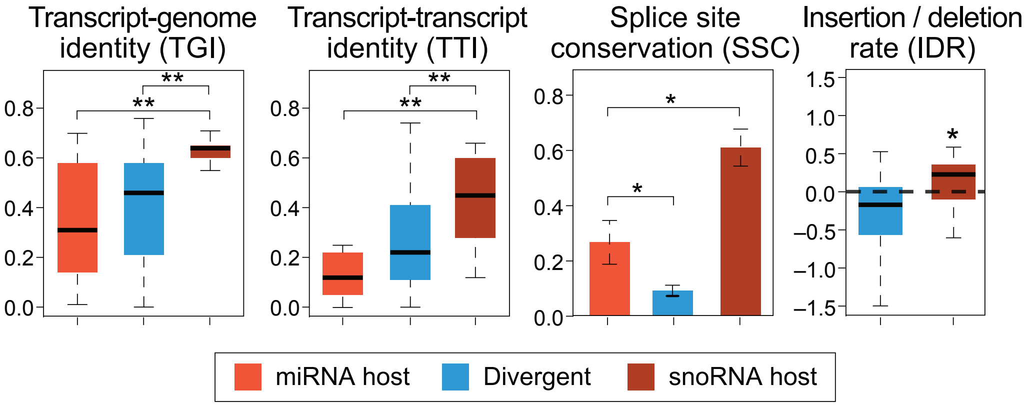
3.4.2 Distinct evolutionary signatures across all annotated lncRNAs

I next sought to extend my evolutionary analysis to larger catalogs of mouse and human lncRNAs (Methods) [80, 81, 156, 171]. Altogether, I searched for candidate orthologs across 251,786 human and 25,335 mouse transcripts corresponding to 56,280 and 15,508 unique lncRNA loci (Fig. B-7) using default parameters of *slncky*. miRNA hosts, divergent lncR-

NAs, and snoRNA host genes show the same distinct evolutionary patterns that I observed in pluripotent cells (Fig. 3-8). Additionally, I found that miRNA hosts that harbor miRNAs inside exonic regions (e.g., *H19* [172]) show a distinct conservation pattern reminiscent of lincRNAs (high TTI and SSC), but without indel-constrained exons (Fig. B-8).



(a) Mean TTI (solid line) and TGI (dotted line) of miRNA host (orange), divergent (blue), and snoRNA host (red) genes from combined lncRNA analysis, recapitulating signatures found in pluripotent lncRNA analysis. Error bars represent standard error of the mean.



(b) Evolutionary metrics of lncRNA classes from combined lncRNA analysis, recapitulating signatures found in pluripotent lncRNA analysis.

Figure 3-8: Evolutionary signatures of all annotated lncRNAs.

Turning my attention specifically to 1,861 candidate orthologous intergenic lncRNAs (lincRNAs), I found that the majority of orthologous pairs did not display signatures of purifying selection and instead had low TTI ($< 30\%$) and no conserved splice sites. Several lines of evidence suggest that the majority of these poorly aligning pairs may not be true orthologs but instead may be transcripts at syntenic loci in different cell types or transcriptional noise. First, applying my orthology-finding pipeline to randomly shuffled transcripts resulted in a similar proportion of syntenic transcripts with low TTI and zero conserved splice sites (Fig. 3-9). Second, though poor alignment metrics could be the result of incomplete reconstructions of lowly expressed lincRNAs, when I performed a similar analysis on a expression-matched set of reconstructed coding transcripts, orthologous pairs have both high TTI and high SSC (Fig. B-9). Third, incorporating human and mouse expression data and limiting the orthology search to only lincRNAs expressed in matched tissues drastically reduced the number of poorly aligning lincRNAs (Fig. B-10).

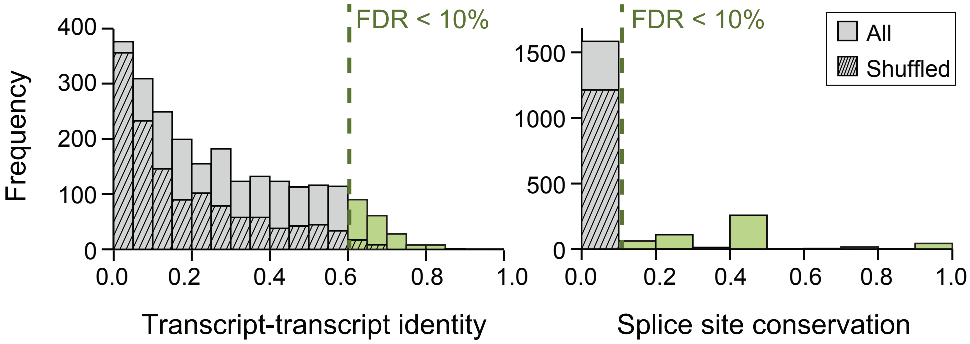


Figure 3-9: Evolutionary metrics of candidate (solid bars) and shuffled (hashed bars) intergenic lncRNAs. Dotted line and green bars denote orthologs with false discovery rate controlled at 10%.

Taken together, I concluded that the majority of syntenic pairs I found were actually unrelated transcripts that have been annotated independently in human and mouse, perhaps in very different cell types, and which have no ancestral relationship. Therefore, I sought to reduce the number of possible spurious lincRNA orthologous pairs by either requiring transcript-transcript identity $>$ 60% or by requiring at least one conserved splice sites, which controls the false discovery rate (FDR) at 10% (3-9). (I also excluded eight intergenic transcripts that contain a conserved ORF between human and mouse with a significant dN/dS ratio and significant coding potential score because they appear to encode for small proteins [Table B.2].) Applying these filtering criteria, 232 pairs of human-mouse lincRNAs orthologs remained with a conservation profile suggestive of high purifying selection at the transcript level (Fig. B-11). However, unlike the pluripotent analysis, the TTI distribution of the filtered lincRNAs was bimodal (Fig. 3-10). Modeling the TTI distribution as two Gaussians, I found 186 (80.1%) lincRNAs with high TTI (mean 65.5% +/- 7.1%) and 46 (19.8%) with low TTI (mean 15.6% +/- 11.7%). This further suggests that selection may operate in two distinct ways: for the majority of lincRNAs, it acts on the full RNA transcript, preserving the transcript sequence, while for a small subset of lincRNAs, the lincRNA sequence may be under positive selection, or perhaps only the act of transcription may be under selective constraint. With the goal of aiding in the study of these human-mouse conserved lincRNAs, I built an easily accessible application available at `https://scripts.mit.edu/~jjenny` as a resource for visually exploring the alignment and conservation properties of these lincRNAs.

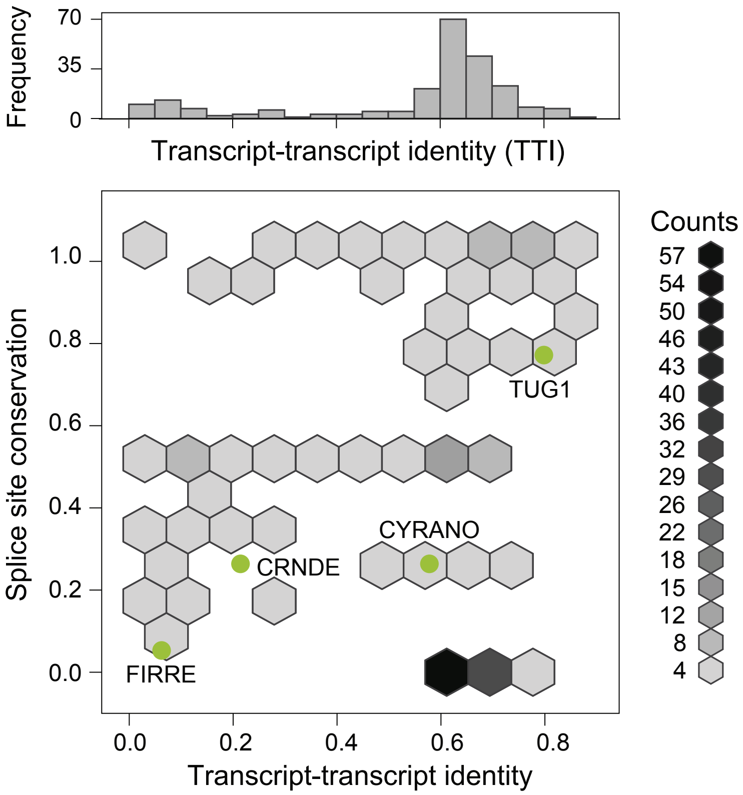
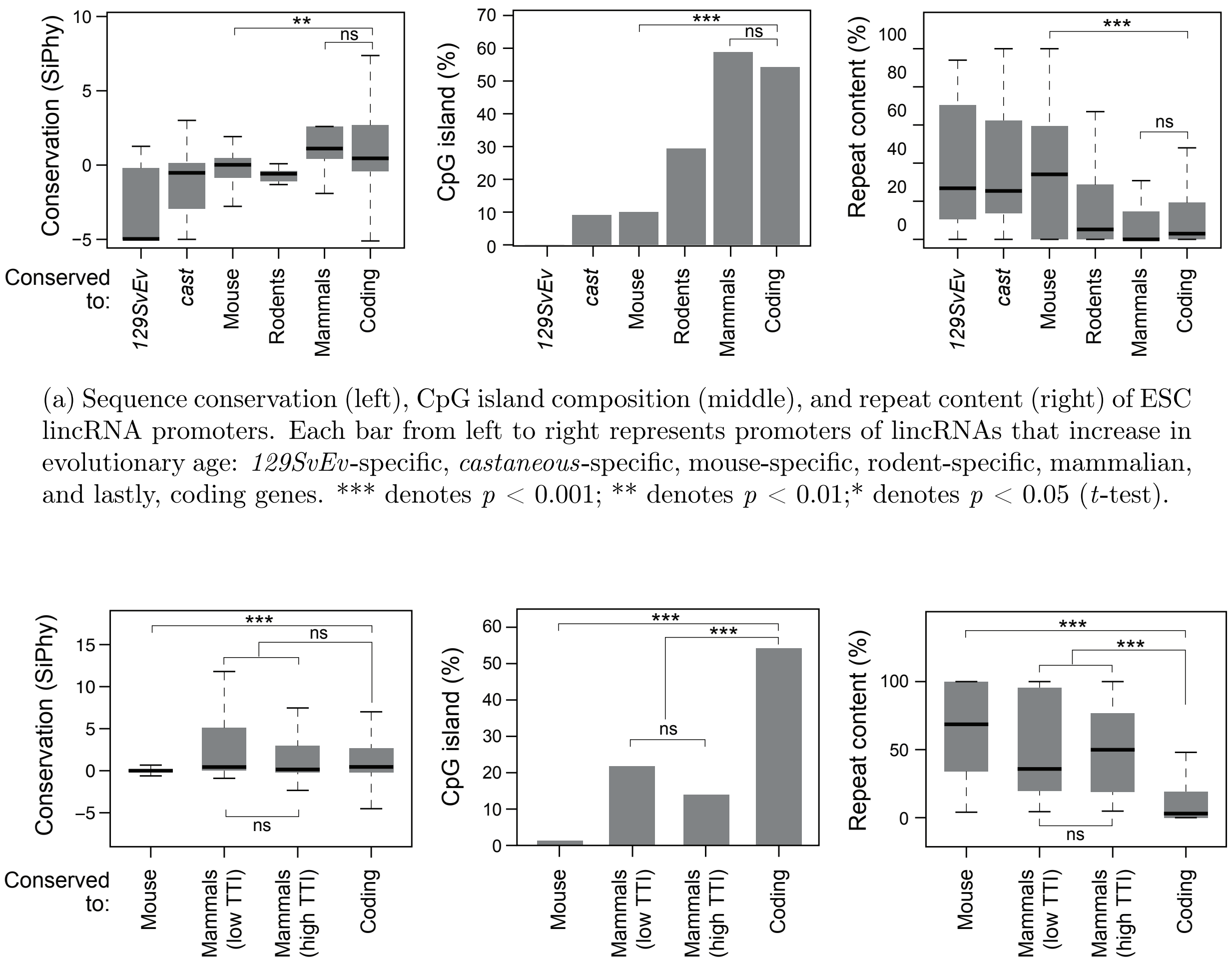


Figure 3-10: Evolutionary metrics of filtered lincRNA orthologs, enriched for true ancestral orthologs. Top: Distribution of filtered TTI. Bottom: Binned scatterplot of TTI and SSC. Overlaid are data points for well-known lncRNAs.

3.4.3 Properties of lineage-specific and conserved intergenic lncRNA promoters

Finally, I sought to understand properties of lincRNAs that explain their conservation or rapid turnover by investigating promoter conservation (Methods). Within the set of



(a) Sequence conservation (left), CpG island composition (middle), and repeat content (right) of ESC lincRNA promoters. Each bar from left to right represents promoters of lincRNAs that increase in evolutionary age: *129SvEv*-specific, *castaneous*-specific, mouse-specific, rodent-specific, mammalian, and lastly, coding genes. *** denotes $p < 0.001$; ** denotes $p < 0.01$;* denotes $p < 0.05$ (t-test).

(b) Same promoter metrics as above for mouse-specific and mammalian-conserved lincRNAs from combined lncRNA catalogs. Mammalian-conserved lincRNAs are split between those with low and high TTI. *** denotes $p < 0.001$; ** denotes $p < 0.01$; * denotes $p < 0.05$ (t-test).

Figure 3-11: Genomic properties of lincRNA promoters of increasing evolutionary age.

pluripotent-expressed lincRNAs (Fig. 3-11a), I found that mammalian-conserved lincRNA promoters have conservation scores comparable to protein coding genes, consistent with previous reports [78, 79], while species-specific lincRNA promoters are indistinguishable from neutral evolution of random intergenic genomic sequence. Conservation also extends to the promoter structure, as I found clear enrichment for CpG islands in conserved lincRNAs, despite comparable CG content (approximately 48%) to that of species-specific lincRNA promoters. In contrast, I found that conservation is negatively correlated with repeat content in lincRNA promoters, and that a significant fraction (30.6%, Fisher's exact test $p =$

1.65×10^{-3}) of species-specific lincRNA promoters contain species-specific endoretroviral K (ERVK) repeat element that appear to be driving transcription. This repeat element is enriched only in promoters of lincRNAs expressed in pluripotent and testis cells (Table B.3), consistent with previous observations that repeat elements are transcribed in ES and germline tissues and silenced in differentiated tissues. I observe that for 60.7% of mouse- or rodent-specific lincRNAs, the time of ERVK integration on the evolutionary tree corresponds exactly with the evolutionary pattern of lincRNA transcription, providing strong evidence that the ERVK element is a primary driver for the origin of the lincRNA. I found corroborating trends of promoter conservation when examining the larger set of lincRNAs from the combined set of annotations (Fig. 3-11b). Importantly, I found no statistical difference in promoter conservation between high and low TTI lincRNA orthologs, suggesting selection for transcriptional control even with poorly aligning orthologs. Together, these results highlight the power of evolutionary analysis to sift through the tens of thousands of annotated lncRNAs to identify a small set of transcripts under selection and likely to be biologically functional. This set serves as an important starting point for downstream experimental interrogation for unraveling the roles and mechanisms of lncRNAs.

3.5 Conclusion

While interest in lncRNAs has intensified, there is still relatively little known about the functions of lncRNAs and much skepticism about what these large number of transcripts mean. The main challenge is that the number of functionally characterized lncRNAs remains a tiny fraction of the total number of lncRNAs that have been annotated. The significant effort required for functional characterization of a single lncRNA compared to its annotation has impeded the functional characterization of the large catalogs of lncRNAs. Accordingly, liberal cataloging efforts have led to a plethora of transcripts defined as lncRNAs that are rarely transcribed or artifacts of transcript assembly, thereby preventing experimental progress. *slncky* provides an important and conservative approach for defining lncRNAs that enriches for *bona fide* lncRNAs. While *slncky* will not necessarily capture every single lncRNA nor will it provide the longest list of possible lncRNAs, it provides a method to define high confidence annotation of lncRNAs from any RNA-seq dataset. This approach will enable meaningful experimental characterization of lncRNAs, making it easier to reconcile the large numbers of defined lncRNAs with the functional roles of these lncRNAs, and providing a consistent standard for evaluating bona fide lncRNAs.

Additionally, evolutionary conservation has long been a confusing feature of lncRNAs. While it is clear that lncRNAs are enriched for conserved genomic sequences, the majority of lncRNAs appear to be transcribed in a species-specific manner, raising questions about whether most of these transcripts are simply byproducts of transcription, with no important biological function. Alternatively, these lncRNA functions may be highly redundant or easily replaceable, in which case evolutionary turnover could be explained by a stochastic evolutionary process where redundant lincRNAs are fixed randomly along the evolutionary tree. Finally, it is possible that many lncRNAs have 'functional orthologs': genes with similar function but no ancestral relationship. For example, evidence of functional orthology was recently reported for *XIST*. Although *XIST* is not found in marsupials, an opossum lncRNA called *RSX* was shown to have similar function. While *RSX* is capable of silencing the X chromosome in mouse, it shares no ancestral relationship with *XIST* [173]. I note that functional orthology cannot be studied with the methods presented here and future work will be needed to explore how many lncRNAs might play such lineage-specific roles or to what extend non-homologous lncRNAs carry similar function.

Despite the rapid turnover in transcription of lncRNAs, I demonstrated that those that are conserved across species can further be categorized into distinct sets based on their evolutionary properties. In particular, I found 232 conserved intergenic lncRNAs that do not host small RNAs in their introns nor are they transcribed from the promoter of a coding gene. Notably, these lincRNAs fall into two sets: one that shows signs of purifying selection in transcript sequence and transcriptional control (i.e., promoter properties), and one that shows selection only for transcriptional control. Furthermore, there are likely many other classes of lncRNAs that cannot be defined by conservation alone. I anticipate that as more cell types and tissues are explored, these annotation and evolutionary approaches will be even more valuable and enable more detailed studies of lncRNA biology.

3.6 Methods

3.6.1 *slncky*

Filtering pipeline for high-quality lncRNAs

slncky filters for lncRNAs in three simple steps. First, *slncky* filters out reconstructed transcripts that overlap coding genes or 'mapped-coding' genes on the same strand, in any amount.

After this step, *slncky* chooses a canonical isoform to represent overlapping transcripts. To do this, *slncky* clusters all transcripts with any amount of exonic overlap into one cluster, and chooses the longest transcript as the canonical isoform.

Next, *slncky* searches for gene duplication events (e.g., zinc finger protein or olfactory gene expansions) by aligning each transcript to every other putative lncRNA transcript using lastz with default parameters [166]. *slncky* then aligns each transcript to shuffled intergenic regions to find a null distribution of alignment scores, repeating this procedure 200 times in order to estimate an empirical p-value. Any alignment with a p-value lower than 0.05 is considered significant. Sets of putative lncRNAs transcript that share significant homology are then merged, creating larger 'duplication clusters'. These transcripts do not necessarily share similarity to a protein-coding gene, though *slncky* will check and report homology to known ZFPs and olfactory genes. *slncky*'s default parameters, which I used in all analyses reported (--min_cluster_size 2), notes and removes any duplication cluster containing two or more transcripts.

Finally, *slncky* removes any transcript that aligns to a syntenic coding gene in another species. (Human and mouse annotations are provided, though users can define their own). First, *slncky* learns a positive distribution by aligning all the transcripts removed in the first filtering step, which overlapped coding genes, to their syntenic coding gene and building an empirical positive score distribution from these alignments. To align genes *slncky* first uses liftOver (--minMatch = 0.1) [174] to determine the syntenic loci in the comparing genome and lastz [166] to perform the alignment across the syntenic region. Using the empirical distribution, *slncky* learns an exonic identity threshold that has an empirical p-value of

0.05. *slncky* repeats the alignment procedure on the putative lncRNAs to syntenic coding genes and filters out any transcripts that align at a higher score than this threshold, even if alignments occur only in UTR or intronic regions. In this way, *slncky* removes unannotated coding genes, pseudogenes, as well use UTR or intronic fragments from incomplete transcript assemblies. To reduce computational cost, whenever more than 250 coding-overlapping genes were filtered out from the first step, only a random subset of 250 transcripts is used to build the positive distribution.

Flagging potentially coding 'lncRNAs'

To find conserved lncRNAs that potentially harbor novel, unannotated protein, *slncky* first aligns putative lncRNAs to syntenic transcripts in a comparing species, using a sensitive noncoding alignment strategy described below. *slncky* then crawls through each significant alignment and reports back any aligned ORF longer than 30 base pairs. Only ORFs that do not contain a frame shift inducing indel in either species are reported. The start codon is defined as 'ATG' and stop codons are defined as 'TAA', 'TAG', or 'TGA'. *slncky* further calculates the ratio of nonsynonymous to synonymous substitutions (dN/dS ratio). For the analyses in this study, I additionally calculated an empirical-value for each dN/dS ratio by aligning 50,000 random intergenic regions and repeating the ORF finding procedure. Because the distribution of dN/dS ratio is dependent on ORF length (Fig. B-2a), I binned ORF lengths by 5 base pair windows and assigned an empirical p-value if I had at least 100 random ORFs within that bin. For long ORFs, for which less than 100 length-matched random ORFs existed, I defined all alignments with dN/dS ratios < 1 as significant.

A sensitive method for aligning orthologous lncRNAs

In searching for conserved lncRNA orthologs, *slncky* first defines the syntenic region of the comparing genome with liftOver (--minMatch = 0.1 --multiple = Y) [174]. If a noncoding transcript exists in the syntenic region, *slncky* then aligns the area 150,000 base pairs upstream to 150,000 base pairs downstream of two syntenic regions. I choose 150,000 base pairs as a general heuristic that is likely to include an easily-alignable coding transcript up- and downstream of the lncRNA, which helps lastz to find a positively scoring alignment.

Importantly, I also found that lncRNAs could only be aligned with a reduced gap-open penalty (--gap = 25,40) because of many small insertions that appear to be well-tolerated by lncRNA transcripts.

To ensure I am not reporting alignments that may occur at random (driven mostly by repetitive elements), I align each lncRNA to shuffled intergenic regions to establish a null distribution and determine the empirical 5% threshold for determining significant alignment scores. Because of the inclusion of flanking regions, it is possible to have a significant alignment in which only the flanking regions align but not the lncRNA transcripts. *slncky* reports these transcripts since it is possible that they are 'syntologs' and carry out orthologous functions but have evolved to a point where they no longer align.

3.6.2 Data collection

Pluripotent cell lines and growth conditions

Naïve 2i/LIF media for mouse and rat (rodent) naïve pluripotent cells was assembled as follows: 500 mL of N2B27 media was generated by including: 240 mL DMEM/ F12 (Biological Industries; custom-made), 240 mL Neurobasal (Invitrogen; 21103), 5 mL N2 supplement (Invitrogen; 17502048), 5 mL B27 supplement (Invitrogen; 17504044), 1 mM glutamine (Invitrogen), 1% non-essential amino acids (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma), penicillin-streptomycin (Invitrogen), and 5 mg/mL BSA (Sigma). Naïve conditions for murine ESCs included 10 μg recombinant human LIF (Peprotech) and small-molecule inhibitors CHIR99021 (CH, 1 μM Axon Medchem) and PD0325901 (PD, 0.75 μM - TOCRIS) referred to as naïve 2i/LIF conditions. Naïve rodent cells were expanded on fibronectin coated plates (Sigma Aldrich). Primed (EpiSC) N2B27 media for murine and rat cells (EpiSCs) contained 8 ng/mL recombinant human bFGF (Peprotech Asia), 20 ng/mL recombinant human Activin (Peprotech), and 1% Knockout serum replacement (Invitrogen). Primed rodent cells were expanded on matrigel (BD Biosciences).

129SvEv (Taconic farms) male primed epiblast stem cell (EpiSC) line was derived from E6.5 embryos previously described in [175]. *129SvEv* naïve ESCs were derived from E3.5 blastocysts. *NOD* naïve ESC and primed EpiSC lines were previously embryo-derived generated and described in [176]. *Castaneous* ESC line was derived from E3.5 in naive 2i/LIF conditions and rendered into a primed cell line by passaging over eight times into primed conditions [177, 178].

Rat naïve iPSC lines were previously described in Hanna et al. [178]. Briefly, rat tail tip derived fibroblasts were infected with a DOX inducible STEMCA-OKSM lentiviral reprogramming vector and M2rtTa lentivirus in 2i/LIF conditions. Established cell lines were maintained on irradiated MEF cells in 2i/LIF independent of DOX. Simultaneously, primed rat pluripotent cells were generated by transferring the rat naïve iPSC cells into primed EpiSC medium for more than eight passages before analysis was conducted.

Naïve human C1 iPSC lines were derived and expanded on irradiated DR4 feeder cells as previously described [158].

RNA-sequencing

RNA-seq libraries were prepared as described in Shishkin et al. [179]. Briefly, 10 μg of total RNA was polyA selected twice using Oligo(dT)25 beads (Life Technologies) and NEB oligo(dT) binding buffer. PolyA-selected RNA was fragmented, repaired, and cleaned using Zymo RNA concentrator-5 kit. A total of 30 ng of polyA-selected RNA per sample were used to make RNA-seq libraries. An adapter was ligated to RNA, RNA was reverse transcribed, and a second adapter was ligated on cDNA. Illumina indexes were introduced during nine cycles of PCR using NEB Q5 Master Mix. Samples were sequenced 100-index-100 on HiSeq2500.

3.6.3 Transcriptome reconstruction and filtering

Transcripts were reconstructed from RNA-sequencing data using Scripture (v3.1, --coverage = 0.2) [78] and multi-exonic transcripts were filtered using *slncky* with default parameters. Annotations of coding genes were downloaded from UCSC (‘coding’ genes from track UCSC Genes, table kgTxInfo) [180] and RefSeq [181]. Mapped coding genes were downloaded from UCSC Transmap database (track UCSC Genes, table transMapAlnUcscGenes) [180]. For the mouse genome, I also included any blat-aligned human coding gene (track UCSC Genes, table blastHg18KG) [180]. As expected, the majority of reconstructed transcripts overlapped an annotated coding or mapped coding gene at $> 95\%$ (Fig. B-1). In the next step, *slncky* aligned each putative lncRNA to every other putative lncRNA to detect duplications of species-specific gene families. Across mouse, rat, and human transcriptomes, I found large clusters (15+ genes) of transcripts sharing significant sequence similarity with each other that also aligned to either zinc finger proteins or olfactory proteins. For unclear reasons, but likely due to the draft status of the assembly which results in collapsed repetitive sequence, I did not find any large clusters of duplicated genes in the chimpanzee genome, and instead found five small clusters of paralogs (Fig. B-1).

Finally, *slncky* aligned the remaining transcripts to syntenic coding genes. For mouse and chimp transcripts, I aligned to syntenic human coding genes and for rat and human transcripts, I aligned to syntenic mouse coding genes. The learned transcript similarity threshold for each pair of comparing species varied as a function of distance between species: the empirical threshold for calling a significant human-chimp alignment was 29.8% sequence similarity while for human-mouse alignments it was approximately 14% (Fig. B-1).

Single exon lncRNAs

Transcript reconstruction software tends to report thousands of single exon transcripts existing in a RNA-seq library. Previous work suggests that the vast majority of these transcripts are results from incomplete UTR reconstruction, processed pseudogenes, very low expressed regions, and DNA contamination [82]. Although *slncky* filters a great number of these artifacts, I find that especially for single exon transcripts, many spurious reconstructions remain. For this reason, when analyzing single exon genes, I only focused on single-exon lncRNAs that are conserved across species.

3.6.4 Verification of filtered lncRNAs

I verified *slncky*'s lncRNA annotations by comparing *slncky* results with other computational and experimental methods, detailed below.

Well-characterized lncRNAs

To test the sensitivity of lncRNA filtering pipelines, I derived a list of well-characterized lncRNAs. To do this, I first took the intersection of annotated noncoding transcripts from

UCSC [180], RefSeq [181], and GENCODE [145]. I then removed any lncRNA with a generically assigned name (e.g., *LINC00028* or *LOC728716*) as well as generically named snoRNA and miRNA host genes (e.g., *SNHG8* or *MIR4697HG*). Finally, I performed a literature search on the remaining lncRNAs, and kept only those that were specifically experimentally interrogated rather than reported from a large-scale screen. This list of well-characterized lncRNAs is available in Additional file 1.

Chromatin modifications

Raw reads from ChIP-sequencing experiments for H3K4me3 and H3K4me36 histone modifications in mouse embryonic stem cells (E14) were downloaded from Xiao et al. [182] (GSE36114). Reads were mapped to mouse genome (mm9) using Bowtie (v0.12.7) [183] with default parameters. Peaks were called as previously described [184].

Coding potential

I scored coding potential of mouse lncRNAs using RNACode (v0.3) [81] with default parameters and multiple sequence alignments of 29 vertebrate genomes from the mouse perspective [11].

Ribosome release scores

Ribosome profiling data of mouse ESCs (E14) was downloaded from Ingolia et al. [185] (GSE30839). Ribosome release scores (RRS) were calculated as described in [161] using the RRS Program provided by the Guttman Lab.

3.6.5 Reanalysis of previously published lncRNA sets

I compared *slncky*'s annotation of lncRNAs to three different human lncRNA sets: GENCODE V19 'Long noncoding RNA' set [145], a set reported by Necsulea et al. [80] based, in part, on GENCODE V7 annotations, and a set reported by Washietl et al. [81] based on GENCODE V12 annotations. For all three comparisons, I first downloaded the appropriate version of GENCODE's 'Comprehensive' gene annotations and applied *slncky* using default parameters. Because Necsulea et al. and Washietl et al. both focused on lncRNAs expressed in RNA-seq data from Brawand et al. [53], I further scored expression of GENCODE annotations using the same RNA-seq data (using Cufflinks v2.1.1 [186] with default parameters) and constrained my analysis to only robustly expressed lncRNAs (fragments per kilobase transcript per million mapped reads [FPKM] $>$ 10 in any tissue).

3.6.6 Reanalysis of previous studies of lncRNA conservation

I downloaded lncRNA annotations and ortholog tables derived from Necsulea et al. [80] and applied *slncky*'s orthology pipeline to mouse and human lncRNAs using default parameters. I compared the human-mouse orthologs discovered by *slncky* to the list of transcripts that were defined by Necsulea et al. to be ancestral to all Eutherians. I used downloaded FPKM tables from Necsulea et al. to constrain my analysis to pairs in which both transcripts are expressed in corresponding tissues.

To assess the ability of *slncky* to discover lncRNAs of a further evolutionary distance than mouse and human, I downloaded lncRNA and ortholog annotations from [157] and applied *slncky* using more relaxed parameters (--minMatch 0.01, --pad 500000) to search for human-zebrafish and mouse-zebrafish lncRNA orthologs. Note that in both analyses, lncRNA annotations were not filtered by *slncky*'s filtering pipeline prior to the ortholog search so that the results could be directly comparable with the original publication.

3.6.7 Annotating orthologous lncRNAs in mammalian ESCs

I applied *slncky* to the pluripotent RNA-seq data to conduct an evolutionary analysis of lncRNAs across multiple mammalian species. I first searched for orthologous lncRNAs in a pairwise manner between every possible pair of species. Because the reconstruction software I used does not report lowly expressed transcripts that do not pass a significance threshold, and because I removed single-exons in the filtering step, I devised a method to rescue orthologous transcripts that may have been removed in those steps. For each lncRNA, if no orthologous lncRNA was detected by *slncky*, I went back to the original RNA-seq data and forced reconstruction of lowly-expressed and/or single-exon transcripts in the syntenic region. I then re-aligned the lncRNA with these newly reconstructed transcripts and added the transcript to the lncRNA set when a significant alignment was found. I kept only pairs of conserved lncRNAs where a significant alignment was found in both reciprocal searches (e.g., mouse-to-human and human-to-mouse).

Next, given pairs of lncRNA orthologs across all species, I created ortholog groups by greedily linking ortholog pairs. For example, given pairs {A,B} and {B,C}, I assigned {A,B,C} to one orthologous group, even if pairing {A,C} did not exist. Finally, I used Fitch's algorithm [187] to recursively reconstruct the most parsimonious presence/absence phylogenetic tree for each lncRNA and determine the last common ancestor (LCA) in which each lncRNA appeared. In the event a single LCA could not be determined by parsimony, I chose the most recent ancestor as the LCA in order to have conservative conservation estimates. For example, if a lncRNA was found in mouse and rat, but missing in human and chimp, I assigned the LCA to be at the rodent root, rather than at the mammalian root with a loss event at primates.

Annotating matched low expression coding genes

I tested *slncky*'s ability to detect conservation of lowly expressed transcripts by first reconstructing lowly-expressed coding genes known to be conserved across mammalian species from the RNA-seq data. I then binned the set of intergenic lncRNAs by increments of 0.1 $\log_{10}$(FPKM), and sampled a set of 162 coding genes that matched in $\log_{10}$(FPKM) distribution in mouse ESCs. I applied *slncky*'s orthology-finding module to the *de novo* reconstructed coding genes. Repeating the same analysis as was done for lncRNAs, I assigned the LCA of each coding gene. I was able to correctly assign the human-mouse ancestor as the LCA for 134 of 162 (83%) coding genes, providing confidence that I am able to sensitively detect orthologs of lncRNAs, even though they are lowly expressed.

3.6.8 Combined catalog analysis

I downloaded human and mouse lncRNA annotations, where they existed, from RefSeq [80, 181], UCSC [180], GENCODE (v19 and vM1) [79,145], and MiTranscriptome [171]. I filtered lncRNAs and searched for orthologs using *slncky* with default parameters. For overlapping isoforms that belong to the same gene, I chose one canonical ortholog pair that had the highest number of conserved splice sites or, if no splice sites were conserved, the highest transcript-transcript identity. miRNA host and snoRNA host genes were annotated using Ensembl annotations of miRNAs and snoRNAs [188]. Divergent genes were annotated based on distance and orientation of closest UCSC- or RefSeq-annotated coding gene. Orthologous lncRNAs were classified as a miRNA host, divergent, or snoRNA host if the transcript was annotated as such in both species. All other lncRNAs were classified as intergenic.

An orthology search was conducted on shuffled transcripts by collapsing overlapping isoforms to a canonical gene as described above, and shuffling to an intergenic location (that is, not overlapping an annotated coding gene) using shuffleBed [136]. I then carried out the orthology search and alignment exactly as described for lncRNAs. To empirically estimate the expected number of conserved splice sites across shuffled orthologs, I took each pair of true lncRNA orthologs and reshuffled splice sites within the loci such that it was correctly located at donor/acceptor sites (GT, AG), and re-evaluated number of conserved splice sites.

I used distributions resulting from the shuffled orthology search to filter and remove spurious hits from the set of candidate lincRNA orthologs such that the FDR $< 10\%$. I then fitted two Gaussians to the resulting transcript-transcript identity using mixtools [189]. Convergence was reached after 31 iterations of EM and final log-likelihood was 146.64. Each ortholog pair was assigned to a Gaussian based on posterior probability cutoff of 50%.

3.6.9 Promoter properties

I defined promoters to be the 500 base pairs upstream of the lincRNA's transcription start site. I calculated several genomic properties of this region as follows:

SiPhy scores

I calculated average SiPhy score across promoter region as previously described [190] using 29-mammals alignment from mouse perspective [11].

CpG islands

For the analysis of CpG islands, I used annotations provided by the UCSC Genome Browser (assembly mm9, track CpG Islands, table cpgIslandExt).

Repeat elements

I intersected promoter regions with annotations from RepeatMasker [191] and calculated the number of base pairs of a lincRNA promoter belonging to a repeat element as well as percentage of lincRNA promoters harboring each class of repeat element. I then repeated this analysis with random intergenic regions, matched in size and GC content. To find statistically significant deviations in repeat content, I used Fisher's exact test to compare the proportion of species-specific lincRNA promoters containing each repeat element to the proportion of random, GC-matched intergenic regions containing the same element. I reported any repeat element that deviated from random, intergenic regions with a p-value < 0.005 (corrected for number of repeat types tested).

3.6.10 Data availability

Raw and processed RNA-seq data are available under GEO accession GSE64818 at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64818`.

A database of conserved lncRNAs discovered in this analysis is available at `https://scripts.mit.edu/~jjenny`

3.6.11 Software availability

slncky (`http://slncky.github.io`) was developed in Python 2.0 and is freely available as source code distributed under the MIT License. *slncky* was tested on Linux and Mac OS X. The version used in this manuscript is available from DOI: `10.5281/zenodo.44628`.

3.6.12 Additional files

Additional file 1: Curated list of “well-characterized lncRNAs”. [`link`]
Additional file 2: Bed file of lncRNAs discovered from mouse (mm9), human (hg19), chimp/bonobo (panTro4), and rat (rn5). [`link`]
Additional file 3: Excel file of evolutionary metrics of all lncRNAs found to be conserved to the primate and rodent ancestor. [`link`]

3.6.13 Authors' contributions

JC participated in the design and coordination of the study, and carried out all computational analyses and software development of *slncky* and *slncky Evolutionary Browser*. AS carried out all RNA-sequencing work. XZ and SK participated in development of supporting software. IM and JH derived ES cell lines. MG conceived of the study and participated in its design and coordination. JC and MG wrote the manuscript with contributions from MGuttman and AR. All authors read and approved the final manuscript. Additional thanks to Leslie Gaffney for artwork and advise on figures and to the Garber, Lander, and Regev laboratory members for helpful discussions.

Supplementary materials for Chapter 3

Species (Strain)	Assembly	Cell type	Number of sequenced fragments	Number of aligned fragments	SRA accession
Mouse (*120SvEv*)	mm9	naïve ESC	180,535,866	118,386,301	SRR1747435
Mouse (*120SvEv*)	mm9	primed epiSC	180,368,378	110,377,225	SRR1747436
Mouse (*NOD*)	mm9	naïve ESC	141,615,128	94,816,294	SRR1747437
Mouse (*NOD*)	mm9	primed epiSC	177,918,230	102,394,440	SRR1747438
Mouse (*cast*)	mm9	naïve ESC	199,168,080	158,066,464	SRR1747439
Mouse (*cast*)	mm9	primed epiSC	224,000,150	157,372,110	SRR1747440
Rat	rn5	naïve ESC	247,087,648	100,883,4721	SRR1747441
Rat	rn5	primed epiSC	114,987,318	80,516,323	SRR1747442
Chimpanzee	panTro4	iPS	159,906,000	108,736,080	SRR873623* SRR873624* SRR873625* SRR873626*
Bonobo	panTro4	iPS	239,033,834	162,543,008	SRR873626* SRR873629* SRR873628* SRR873627*
Human	hg19	iPS	244,014,732	201,066,988	SRR1747443

Table B.1: RNA-Sequencing libraries used in lncRNA study. Asterix denotes downloaded data.

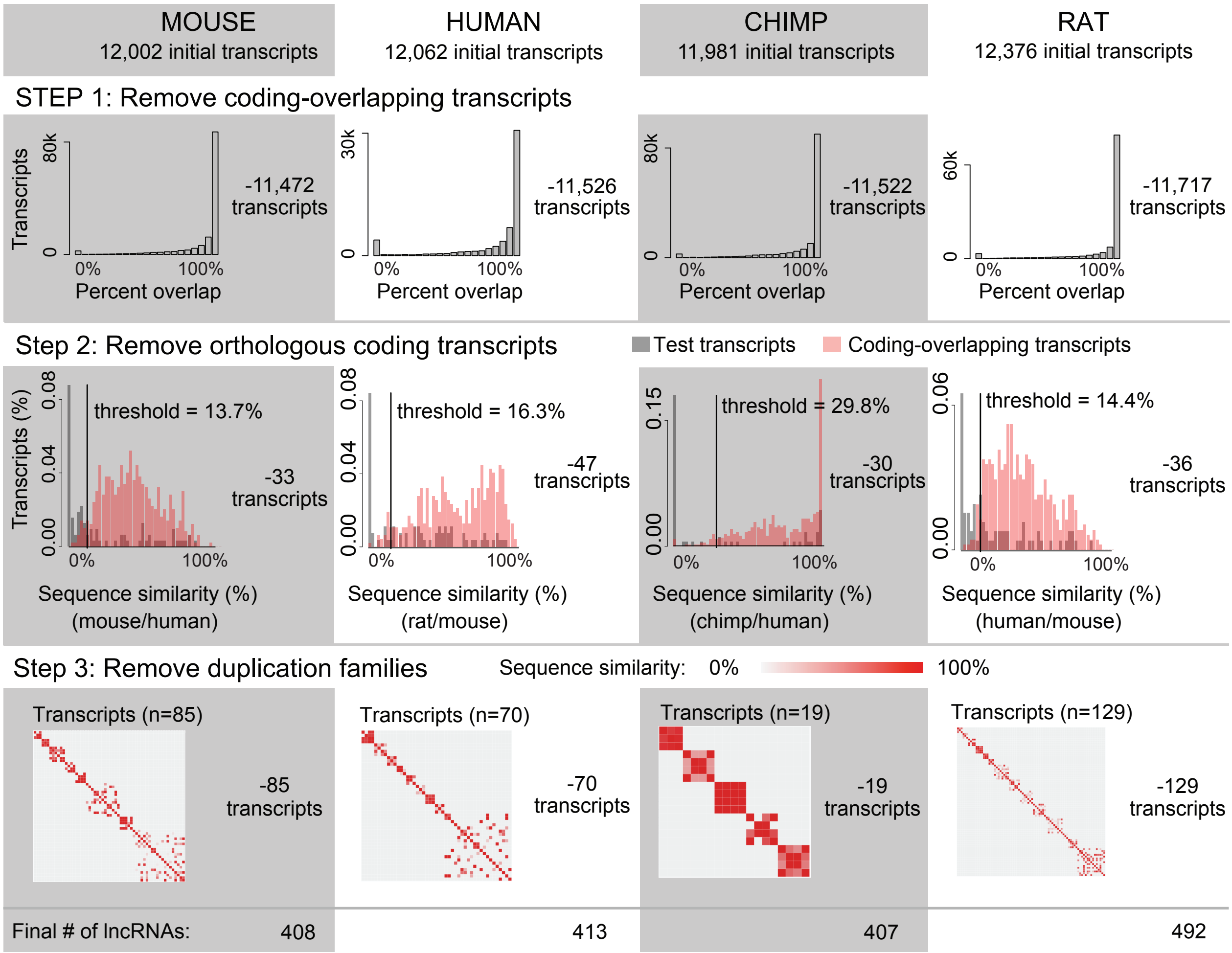
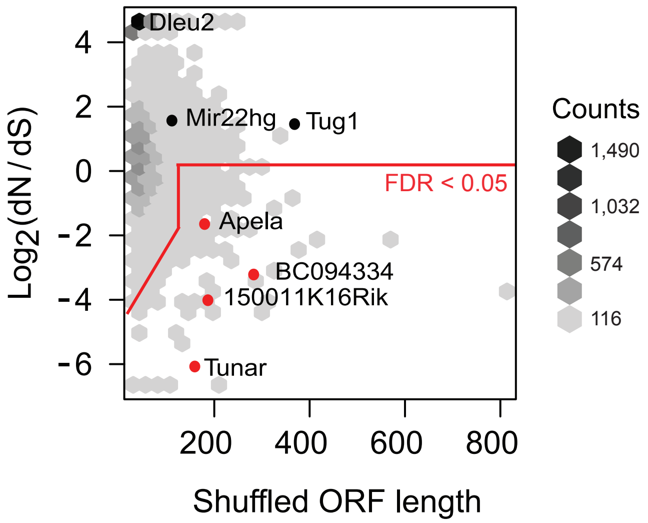


Figure B-1: Statistics from *slncky*'s filtering pipeline applied to mammalian ESC data. Top row: Percent exonic overlap of reconstructed transcripts with annotated coding genes. Number of transcripts removed are shown inside circles. Middle row: Exonic sequence similarity between coding-overlapping transcripts that align to syntenic coding genes (red) and reconstructed transcripts that align to a syntenic coding gene (gray). Distribution of sequence similarity for coding-overlapping transcripts (red) is used to define empirical 5% threshold used for filtering. Bottom row: Heatmap of sequence similarity between reconstructed transcripts that align significantly to each other.



(a) dN/dS ratios (y-axis) of ORFs from shuffled lncRNA alignments, binned by ORF length (x-axis). Red line shows dN/dS cutoff for $p < 0.05$. For long ORFs, where not enough shuffled alignments were available to estimate a p-value cutoff, the dN/dS cutoff is set to 1. Labeled black points are dN/dS ratios of true lncRNAs with significant coding potential scores; labeled red points are annotated lncRNAs with conserved ORFs flagged by *slncky*.

Gene	ORF length (bp)	dN	dS	dN/dS	Coding potential (RNACode p-value)
Tunar	147	0.003	0.22	0.014	1.19e-14
150011K16Rik	171	0.01	0.14	0.059	2.78e-04
BC094334	255	0.02	0.15	0.103	5.90e-10
Apela	165	0.05	0.15	0.308	2.00e-03

(b) dN/dS ratios of annotated lncRNAs, flagged by *slncky* as likely harboring a coding ORF.

Tug1	330	0.04	0.02	2.69	4.34e-06
Malat1	NA	-	-	-	2.28e-04
Cyrano	NA	-	-	-	1.10e-03
Mir22hg	105	0.10	0.04	2.905	6.00e-03
Dleu2	33	0.16	0.00	inf	6.00e-03

(c) dN/dS ratios of known lncRNAs with significant coding potential scores.

Figure B-2: dN/dS ratios of ORFs found in shuffled and real lncRNA alignments.

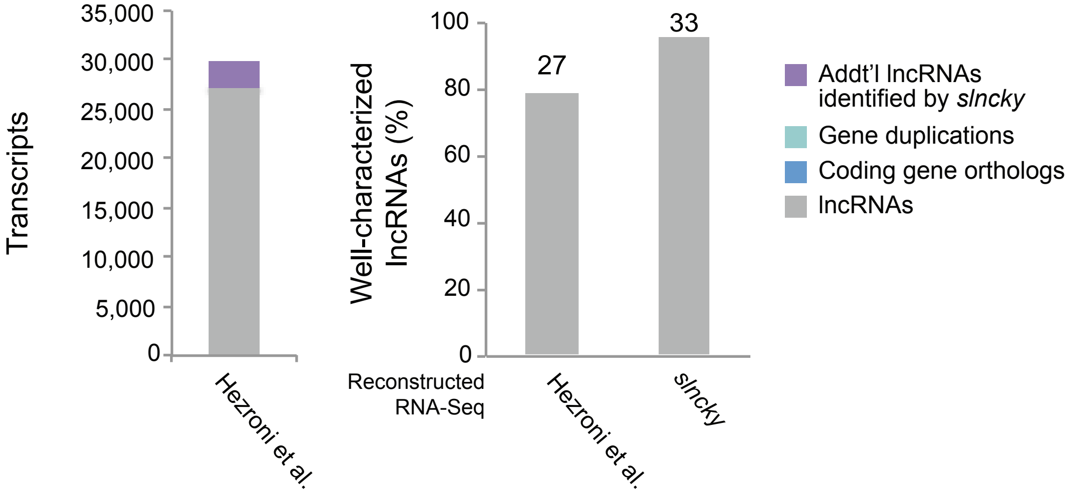


Figure B-3: Comparison of *slncky*'s filtering pipeline with PLAR. Left: When applied to PLAR-filtered lncRNAs, number of transcripts that *slncky* also annotated as a lncRNA (gray), removes as gene duplication or coding (light and dark blue), and additionally identifies as novel lncRNAs (purple). Right: Percentage of well-characterized lncRNAs identified by PLAR compared to *slncky* results. Numbers above bars denote absolute number of lncRNAs.

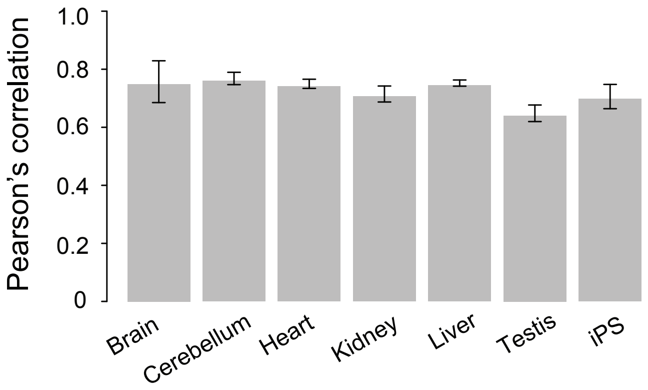


Figure B-4: Pearson's correlation of $\log_{10}$(FPKM) values from expressed genes between mouse and human samples from somatic tissues [65] and iPS data presented in this study.

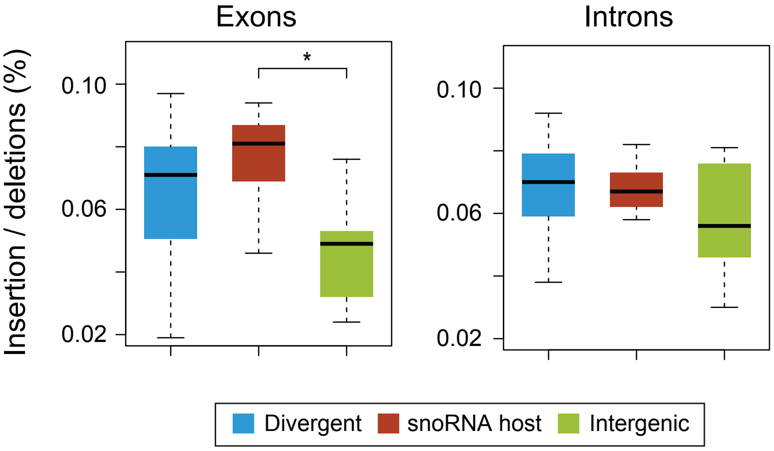


Figure B-5: Percent indels in exons and introns of divergent (blue), snoRNA host (red), and intergenic (green) lncRNAs. * denotes $p < 0.05$ (t-test)

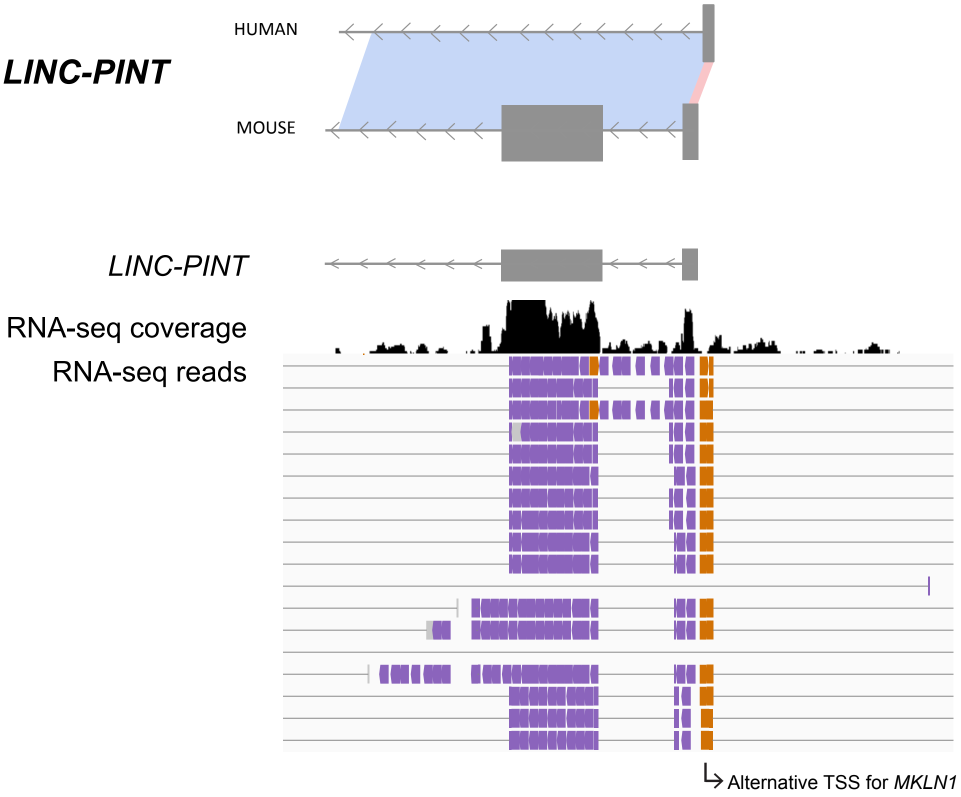


Figure B-6: Alignment profile (top) and RNA-seq alignments (bottom) of the 5' end of *LINC-PINT*. Negative strand reads are shown in purple and positive strand reads in orange. Positive strand reads represent an unannotated, alternative 5' end of *MKLN1*.

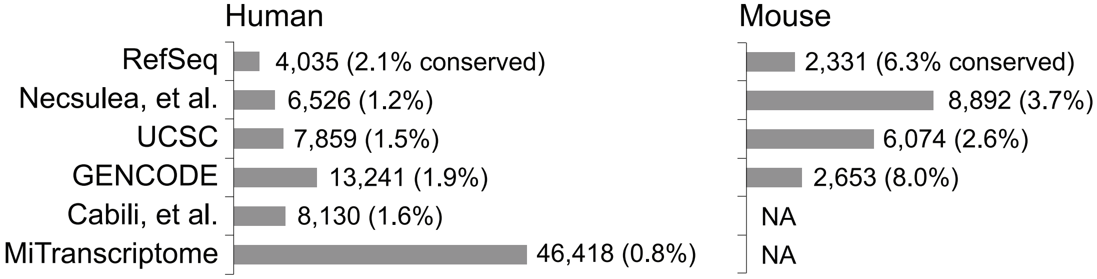
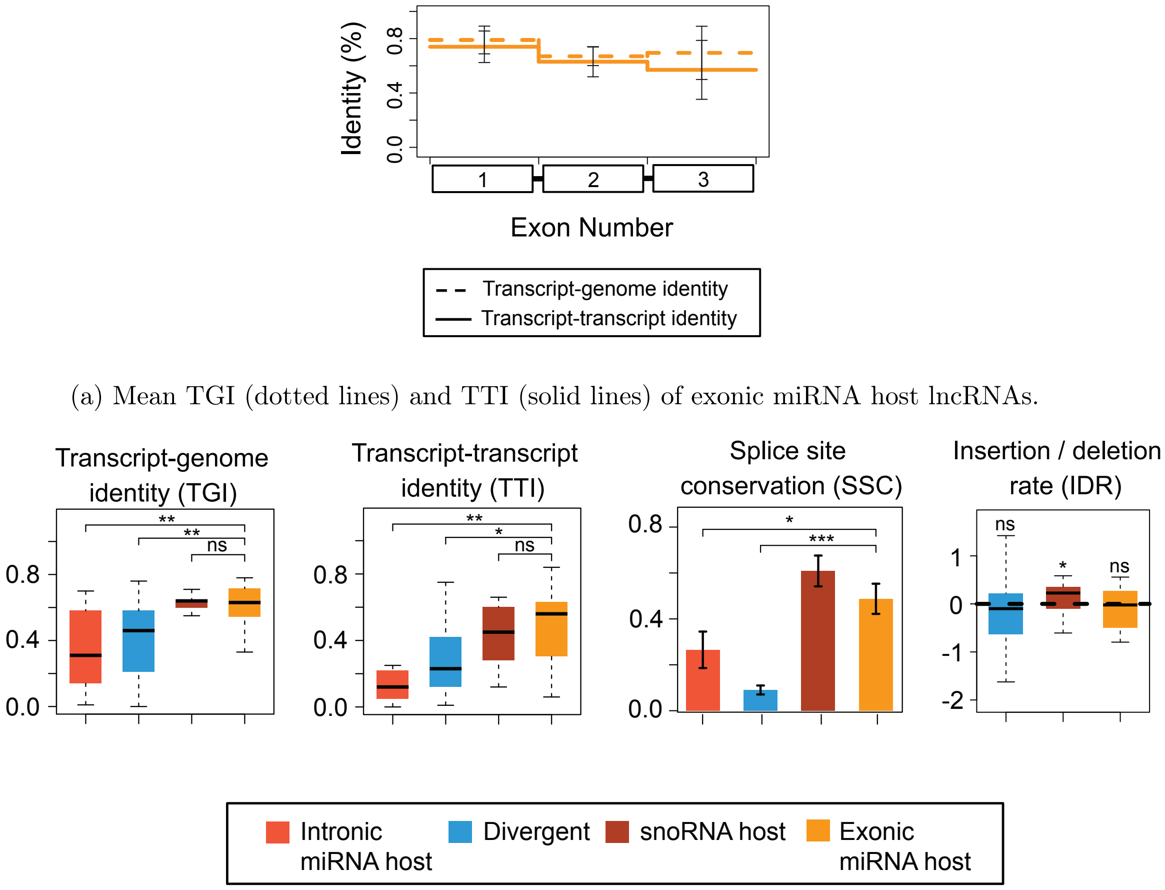


Figure B-7: Number of transcripts found in existing lncRNA catalogs.



(a) Mean TGI (dotted lines) and TTI (solid lines) of exonic miRNA host lncRNAs.

(b) Evolutionary metrics of each lncRNA class from combined analysis compared to exonic miRNA host genes.

Figure B-8: Evolutionary signatures of exonic miRNA host genes

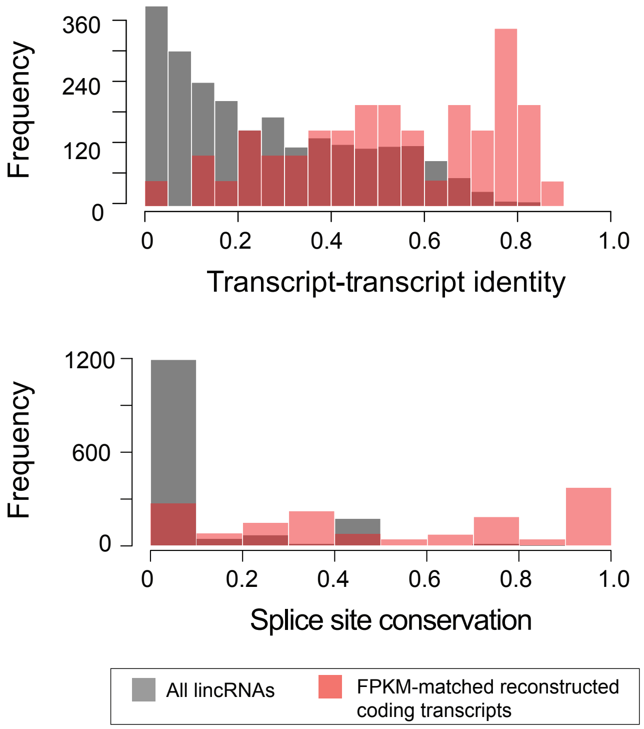


Figure B-9: TTI (top) and splice site conservation (bottom) of all lincRNA orthologs (gray) compared to expression-matched coding genes (red).

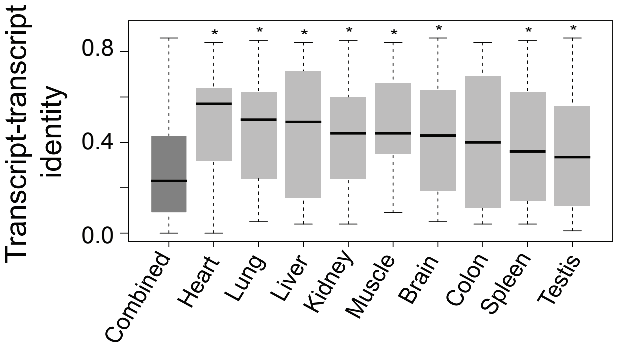
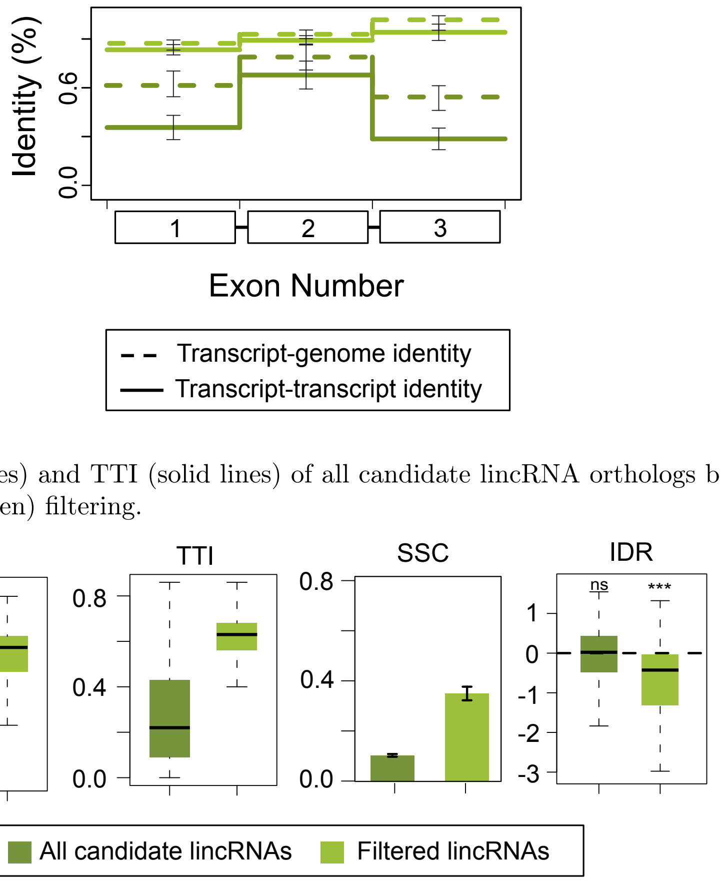


Figure B-10: TTI of all candidate lincRNA orthologs compared to TTI from only lincRNAs expressed in matched tissues of human and mouse. * denotes $p < 0.05$ (t-test when compared to combined TTI)

Gene	ORF length (bp)	dN	dS	dN/dS	Coding potential (RNACode *p*-value)
ENSMUSG00000053724	525	0.07	0.13	0.54	4.18e-11
LINC00948 (*MRLN*)	141	0.04	0.21	0.19	1.33e-08
LINC00890	273	0.01	0.09	0.13	1.03e-08
LOC100507537	108	0.05	0.12	0.46	1.71e-05
CDIPT-AS1	123	0.08	0.10	0.45	4.80e-04
GQ868703	87	0.02	0.06	0.27	5.00e-03
AK136239	60	0.03	0.08	0.38	3.60e-02
AK094929	90	0.01	0.02	0.27	3.60e-02

Table B.2: dN/dS ratios and coding potential scores of transcripts that likely harbor ORFs.



(a) Mean TGI (dotted lines) and TTI (solid lines) of all candidate lincRNA orthologs before (dark green) and after (light green) filtering.

(b) Evolutionary metrics of candidate lincRNA orthologs before (dark green) and after (light green) filtering.

Figure B-11: Evolutionary signatures of candidate and filtered lincRNA orthologs.

	Pluripotent lncRNAs		Necsulea, et al. lncRNAs						
	Mouse specific (n=291)	Conserved (n=48)	ES (n=829)	Brain (n=566)	Heart (n=352)	Kidney (n=828)	Liver (n=254)	Ovary (n=1170)	Testis (n=3379)
L1	9.2e-06	5.1e-02	8.2e-11	8.9e-07	2.9e-04	3.7e-07	7.3e-05	2.0e-16	2.9e-42
Low complexity	3.1e-01	5.8e-01	1.9e-05	6.0e-03	1.8e-03	1.7e-04	1.0e-01	2.8e-07	1.1e-06
Simple repeat	1.0e+00	4.3e-01	1.1e-01	1.8e-02	6.1e-02	7.6e-01	9.2e-01	2.9e-05	4.0e-07
Alu	3.1e-01	1.0e+00	4.2e-02	4.8e-01	1.3e-01	8.0e-03	5.7e-01	1.6e-02	2.6e-03
MaLR	1.0e+00	5.9e-02	1.8e-02	6.7e-01	1.5e-01	6.9e-01	8.0e-01	6.9e-01	1.6e-10
ERVK	1.7e-03	1.0e+00	4.1e-03	5.3e-02	7.9e-01	1.8e-02	8.7e-01	7.2e-02	1.7e-04
B4	1.3e-01	1.0e+00	7.1e-01	1.0e+00	7.2e-01	1.4e-02	5.6e-01	1.4e-01	3.0e-01
B2	3.8e-02	1.0e+00	3.1e-01	6.8e-02	7.9e-01	1.0e+00	4.9e-01	8.4e-01	4.1e-01
ERV1	3.0e-01	1.0e+00	3.2e-01	7.2e-01	4.9e-02	1.4e-01	3.8e-01	4.6e-01	7.0e-02

Table B.3: Enrichment and depletion of repeat elements in lncRNA promoters. Fisher's exact test *p*-values from comparing proportion of repeat element in lncRNA promoters to the proportion observed in GC-matched, random intergenic regions. Red text denotes enrichment and blue denotes depletion.

Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of *ERAP2* transcripts under balancing selection

Jenny Chen*, Chun Jimmie Ye*, Alexandra-Chloé Villani, Meena Subramaniam, Rachel E. Gate, Tushar Bhangale, Mark N. Lee, Towfique Raj, Raktima Raychowdhury, Weibo Li, Noga Rogel, Selina H. Imboywa, Portia I. Chipendo, Cristin McCabe, Michelle H. Lee, Irene Y. Frohlich, Barbara E. Stranger, Philip L. De Jager, Aviv Regev, Tim Behrens, Nir Hacohen

*Authors contributed equally to this work.

This chapter is not a comparative analysis across species, but instead, an analysis of the regulatory control of isoform usage across the human population. Still, this study highlights how investigations into transcriptional regulation can help answer questions about unexplained evolutionary signatures observed in the human genome.

While the impact of common genetic variants on transcript abundance in response to cellular stimuli has been analyzed in depth, less is known about how stimulation modulates the genetic control of isoform usage. Using RNA-sequencing profiles of monocyte-derived dendritic cells from 243 individuals, we uncovered thousands of unannotated isoforms synthesized in response to viral infection or stimulation with Type 1 interferon. We identified more than a thousand single nucleotide polymorphisms associated with isoform usage (isoQTLs), many of which are independent of expression QTLs (eQTLs) for the same gene. Compared to eQTLs, isoQTLs are enriched for splice sites and untranslated regions, and depleted of upstream sequences. We specifically examine the *ERAP2* locus, where the major haplotype is under balancing selection, though associated with Crohn's disease risk. At baseline and following Type 1 interferon stimulation, the major haplotype is associated with absence of *ERAP2* expression; but in response to influenza infection, the major haplotype results in the expression of two previously uncharacterized, alternatively transcribed, spliced, and translated short isoforms. Thus, genetic variants at a single locus could modulate independent gene regulatory processes in the innate immune response, and in the case of *ERAP2*, may confer a historical fitness advantage in response to virus, but increase risk for autoimmunity in the modern environment.

4.1 Background

An important aspect of eukaryotic gene regulation is the control of alternative gene isoforms. This is achieved through several mechanisms at the transcript level, including: alternative promoters for transcription initiation, alternative splicing of pre-messenger RNA, alternative polyadenylation, and selective degradation of isoforms. These processes regulate the relative abundances of multiple coding and non-coding RNAs from the same underlying DNA sequence, often resulting in altered function of the protein products in response to developmental or environmental changes [192–195].

A case in point is the role of alternative isoform usage in the human immune response.

For example, studies have shown that alternative splicing is critical across many immune processes, such as B cell functions reflected in the balance between IgM and IgD immunoglobulin isoforms [196], naïve and memory T cell functions controlled by *CD45* isoforms [197], and innate immune responses to pathogens regulated by different isoforms of *MYD88* [198]. Genetic variants that affect isoform usage have been associated with immune disorders [199] including the association of systemic lupus erythematosus with common variants in a splice site at the *IRF5* locus [200].

Previous studies have identified shared and divergent transcriptional programs in the antibacterial and antiviral response of innate immune cells [201, 202], with genetic variation imparting stimulation specific effects on gene expression [202–205]. While genetic maps of alternative splicing are beginning to emerge, most notably in lymphoblastoid cell lines [206, 207], across healthy human tissues [208, 209], and in macrophages stimulated with bacteria [210], variability in isoform usage across individuals and its genetic basis in the human antiviral response have not been studied.

Here, we integrate RNA-sequencing (RNA-seq) with dense genotyping to systematically investigate the genetic control of isoform usage in monocyte derived dendritic cells (MoDCs) at rest, and in response to influenza-infection or Type 1 interferon. Because the Type 1 interferon pathway is known to be engaged by a broad array of microbial products, our study design is unique in allowing the separation of the universal and influenza-specific effects on the interferon-induced response. Since the human transcriptome has never been annotated under these conditions, we first used *de novo* assembly to catalog and quantify all synthesized isoforms in resting and stimulated cells. Then, by harnessing the natural transcriptomic and genetic variation in the ImmVar cohort [202, 211, 212], we mapped genetic variants associated with isoform usage (isoQTLs). Systematic characterization of isoQTLs, especially in comparison to eQTLs, provides mechanistic insights into the genetic control of different aspects of gene regulation and enables the functional interpretation of loci associated with immune disease and under natural selection.

4.2 Alternate isoform usage in anti-viral response

We used paired-end RNA-seq to profile the transcriptomes of primary MoDCs from healthy donors at rest (n = 99), and following stimulation with either influenza ΔNS1 (a strain engineered to maximize the IFNβ-induced host response to infection by the deletion of a key virulence factor [213]) (n = 250) or interferon beta (IFNβ), a cytokine that stimulates anti-viral effectors (n = 227). A total of 552 pass-filter samples (out of 576) – 84 from all three conditions, 127 from both stimulation conditions, and 46 from only one condition – were analyzed (Additional file 1). To define the corpus of transcripts in human dendritic cells at rest and in response to stimulation, including previously unannotated transcripts, we assembled the transcriptome *de novo* in each sample (individual-condition pair), retained only expressed isoforms ($>$ 5 transcripts per million in any sample), then combined isoforms across all samples to enable direct comparisons between conditions. Overall, we identified 35,411 transcripts: 18,644 present in resting cells, and 29,841 (flu) and 25,127 (IFNβ) present in stimulated cells. The 35,411 transcripts correspond to 15,123 genes, including 8,338 previously unannotated transcription start sites (TSSs) (corresponding to 5,414 genes), 16,062 previously unannotated splice sites (corresponding to 11,704 isoforms and 6,703 genes), and 1,653 previously unannotated transcripts (corresponding to 1,281 genes) (Additional file 2).

Compared to IFNβ induction, flu infection elicited a prominent change in isoform usage independent of gene expression, estimated as the ratio of isoform abundance over total gene abundance. Relative to baseline, the usage of 3x as many isoforms (4,937 vs. 1,651) were altered in flu-infected compared to IFNβ-stimulated cells (beta regression, FDR < 0.01, isoform abs(log$_2$[fold change]) > 1, Additional file 3). In response to both conditions, more than 50% of isoforms with differential usage were previously unannotated, highlighting the inadequacy of current annotations in describing the full diversity of gene isoforms in the human antiviral response. Of the differentially expressed genes with more than one isoform, 36% (flu, 1326/3680) and 22% (IFNβ, 433/1995) had at least one isoform that differed in usage (Fig. 4-1), suggesting independent regulatory mechanisms that control overall gene abundance and specific isoform usage in response to stimuli.

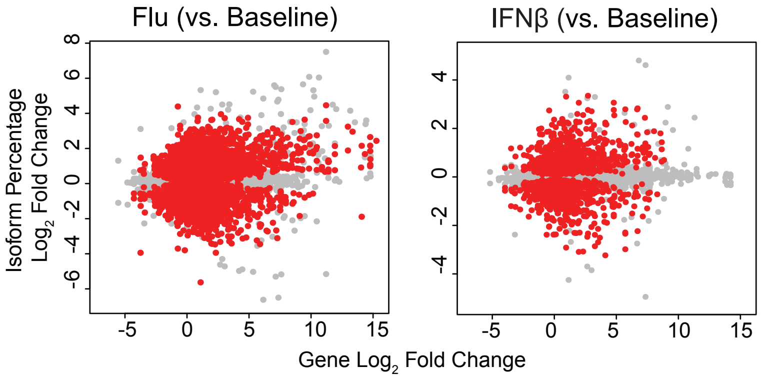


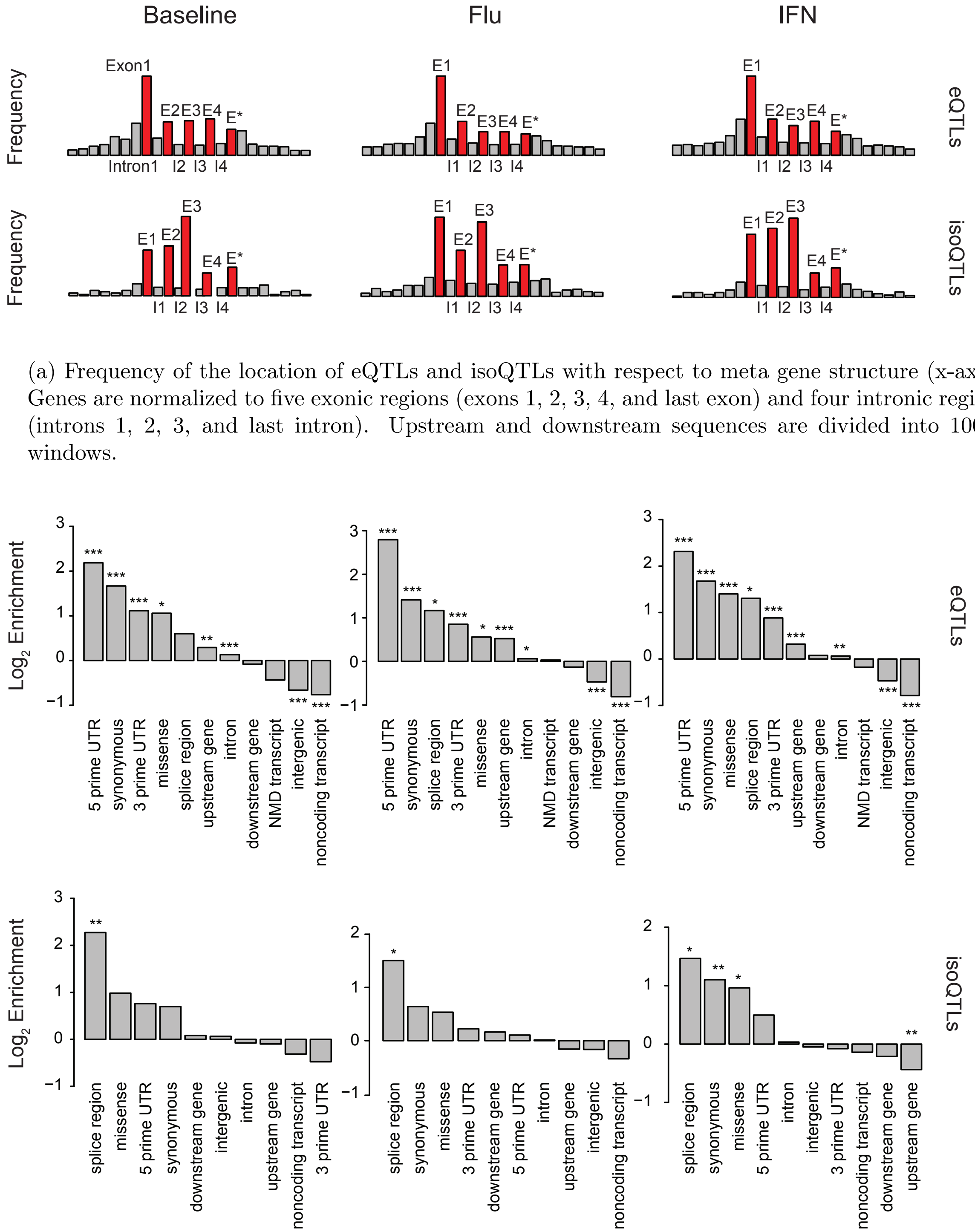
Figure 4-1: Scatterplot of fold change of overall gene abundance (x-axis) versus fold change in isoform percentage (y-axis) in flu-infected (left) and IFNβ-simulated (right) cells compared to baseline. Each dot represents one isoform. Isoforms that significantly differed in their usage percentages (beta regression, FDR < 0.01) are highlighted in red.

4.3 Genetic control of alternative isoform usage

We assessed whether common genetic variants, known to affect gene expression [202], could affect isoform usage in both resting and stimulated MoDCs. We associated over 10 million (M) imputed variants with two transcriptional traits, isoform percentage and total gene abundance, to identify isoform usage quantitative trait loci (isoQTLs) and expression quantitative trait loci (eQTLs), respectively. After adjusting for unwanted variation from latent effects (Fig. C-1), we identified 2,393 isoforms corresponding to 1,345 genes (linear regression, permutation FDR < 0.05, Additional file 4) with local isoQTLs (+/- 500 kilobases [kb] of TSS) and 8,350 genes (linear regression, permutation FDR < 0.05, Additional file 5) with local eQTLs in at least one condition. A substantial proportion of leading isoQTL SNPs (58% baseline, 42% flu, 39% IFNβ) were not significant eQTLs, suggesting that the genetic control of isoform usage and overall gene abundance are largely independent.

4.3.1 Regulatory features of genetic determinants of isoform usage

Genetic variants could modulate isoform usage through several mechanisms including perturbing the usage of alternate promoters, splice sites, or regulatory elements in the untranslated regions (UTRs). We compared the *cis* properties of isoQTLs and eQTLs to identify the mechanisms by which each class of variants acts. When normalized by exon and intron lengths, leading SNPs for isoQTLs were enriched across the entire gene body (Fig. 4-2a), in distinct contrast with leading SNPs for eQTLs, which were enriched near TSS and transcription end site. Further, when compared to a set of SNPs matched for allele frequency and distance to TSS, leading SNPs for isoQTLs were most enriched for splice sites (4.8x baseline, 2.8x flu, 2.7x IFNβ), synonymous (1.6x baseline, 1.6x flu, 2.1x IFNβ) and missense variants (2.0x baseline, 1.4x flu, 1.9x IFNβ), and 5' (1.7x baseline, 1.1x flu, 1.4x IFNβ) and 3' (1.5x flu, 1.4x IFNβ) UTRs (Fig. 4-2b). These results suggest that genetic variants associated with isoform usage likely do so via *cis* regulatory sequences that modulate alternative splicing and transcript stability.



(a) Frequency of the location of eQTLs and isoQTLs with respect to meta gene structure (x-axis). Genes are normalized to five exonic regions (exons 1, 2, 3, 4, and last exon) and four intronic regions (introns 1, 2, 3, and last intron). Upstream and downstream sequences are divided into 100kb windows.

(b) Fold enrichment of genomic annotations of eQTLs and isoQTLs compared to a background set of SNPs matched for distance to TSS and allele frequency. * denotes adjusted $p < 0.05$; ** denotes adjusted $p < 0.01$; *** denotes adjusted $p < 0.001$.

Figure 4-2: Genomic properties of eQTLs and isoQTLs.

4.3.2 Genetic control of alternative isoform usage in responses to virus and interferon

To assess how the genetic control of isoform usage differs in response to stimuli, we analyzed 84 donors whose cells were assayed in all three conditions to enable equally powered comparisons across conditions. Genetic variants imparted stronger effects on isoform usage in baseline and IFNβ-infected cells than in flu-infected cells as indicated by more isoQTLs detected (815 in baseline, 784 in IFNβ, and 427 in flu, permutation FDR < 0.05) and an increase in the proportion of variance of isoform usage explained (R^2_{iso}) by the associated variants (Fig. C-2). The correlation of R^2_{iso} was lowest between flu-infected and baseline cells (Pearson's $\rho_{flu,baseline} = 0.46$ compared to $\rho_{IFN,baseline} = 0.69$ and $\rho_{IFN,flu} = 0.66$) suggesting flu-specific genetic control of isoform usage independent of Type 1 interferon signaling. Isoforms with higher R^2_{iso} in stimulated cells are upregulated in response to stimuli, suggesting that the activation of specific gene regulatory programs that control isoform usage are sensitive to genetic effects unobserved in inactive states (Fig. 4-3).

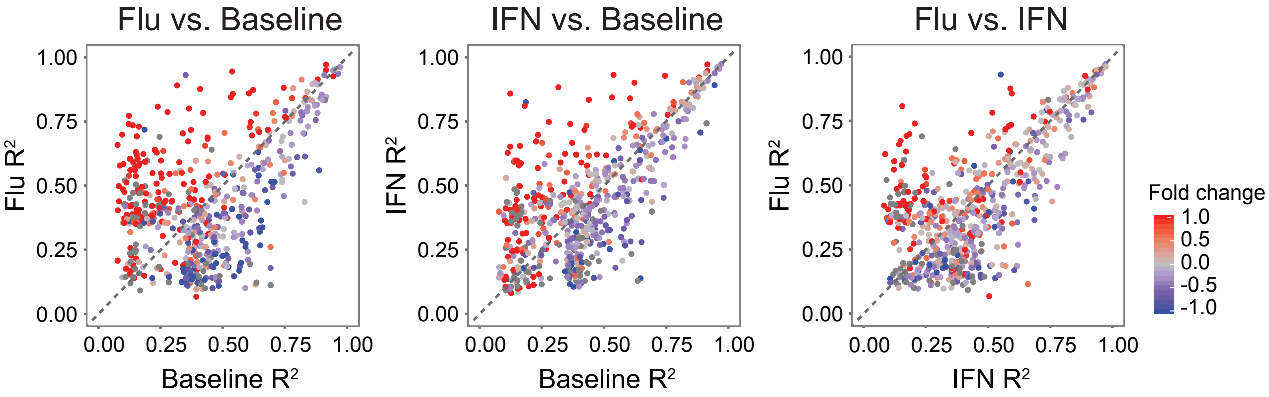


Figure 4-3: Correlation of effect sizes (R^2) for isoQTLs between pairs of conditions. Transcripts are colored by differential expression (red: up-regulated, blue: down-regulated).

To directly assess how stimulation modifies the effects of individual genetic variants on isoform usage, we mapped SNPs associated with the difference in isoform usage between conditions, herein referred to as response-isoQTLs (r-isoQTLs). Compared to resting cells, we identified 74 (flu) and 22 (IFNβ) significant r-isoQTLs corresponding to 50 and 14 genes (permutation FDR < 0.1, Additional file 6). Amongst the 13 genes that share r-isoQTLs in both stimulated conditions was *IFI44L* (Fig. 4-4, left), a type 1 interferon-stimulated gene that has been shown to have moderate effects in inhibiting human hepatitis virus replication *in vitro* [214] and whose splicing has been shown to be influenced by the most significant r-isoQTL (rs1333973) [215]. Among the 37 genes that have r-isoQTLs in flu-infected but not interferon-stimulated cells was *ZBP1* (Fig. 4-4, right), a sensor of influenza infection that triggers cell death and inflammation and contributes to virus-induced lethality [216]. While influenza-infected and interferon-stimulated cells are expected to share some r-isoQTLs reflecting a common gene regulatory program (as interferons are induced by viral infection), influenza-specific r-isoQTLs confer genetic control of previously unknown viral sensing pathways independent of downstream effector (type-1 interferon) signaling.

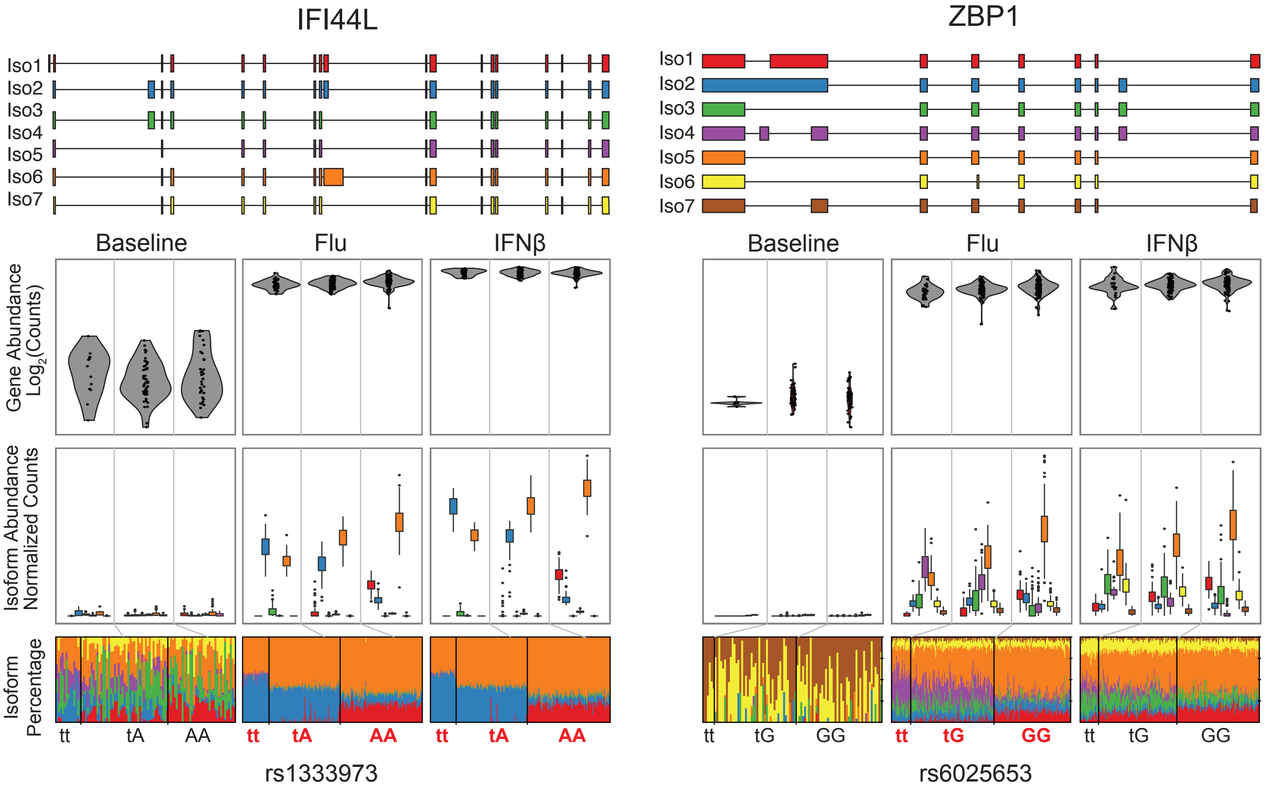


Figure 4-4: *De novo* reconstructed transcript structure (top panel) and box-whisker plots (bottom three panels) between transcript quantitative traits (y-axis: $\log_2$(normalized gene counts), normalized isoform counts, or isoform percentage) and genotype (x-axis) for 2 genes (*IFI44L* and *ZBP1*) with r-isoQTLs. Significant r-isoQTLs are highlighted with red text.

4.4 Influenza-specific regulation of *ERAP2* isoforms under balancing selection

Finally, we specifically examined the genetic control of *ERAP2* transcripts in the human antiviral response because of the known role of *ERAP2* in antigen presentation [217] and because of its peculiar evolutionary history. The *ERAP2* locus is characterized by two frequent and highly differentiated (40 SNPs in perfect linkage disequilibrium) haplotypes observed in every major human population (B: 53% and A: 47%) (Fig. C-3). The major allele (G) of rs2248374, a splice-site variant tagging Haplotype B, creates an alternate 3' donor splice site inducing the splicing of an extended exon 10 with two premature termination codons35. As a result, transcripts from Haplotype B are degraded by nonsense-mediated decay resulting in one of the most significant eQTLs and isoQTLs in most tissues and cell types [202, 206, 208, 212]. Intriguingly, while Haplotype B is associated with increased risk for Crohn's disease [218], and it is also maintained by long term balancing selection (between 1.4M [219] and 5.1M years [220]). This raises the important question: in what environmental condition does balancing selection act to maintain the seemingly loss-of-function (LOF), disease-causing haplotype in humans?

In resting and IFNβ-stimulated cells, we confirmed the known genetic association of rs2248374^G allele with lower *ERAP2* expression (Fig. 4-5). Remarkably, while the overall abundance of *ERAP2* was elevated in stimulated conditions, two previously uncharacterized short isoforms (*ERAP2*/Iso2, *ERAP2*/Iso3, Fig. 4-5, C-4) were transcribed from Haplotype B only in flu-infected and not IFNβ-stimulated cells, resulting in the partial rescue of *ERAP2* expression. The short isoforms differed from the constitutive full-length isoform (*ERAP2*/Iso1 transcribed from Haplotype A) by the initiation of transcription at exon 9 and the alternate splicing of an extended exon 10, and differed from each other by alternative splicing at a secondary splice site at exon 15. The initiation of transcription at exon 9 results in an alternate in-frame translation start site at exon 11 thus rendering the premature termination codon in exon 10 inactive.

The influenza-dependent genetic control of *ERAP2* isoform usage is further supported by (1) correlation between overall flu transcript abundance, a proxy for degree of infection,

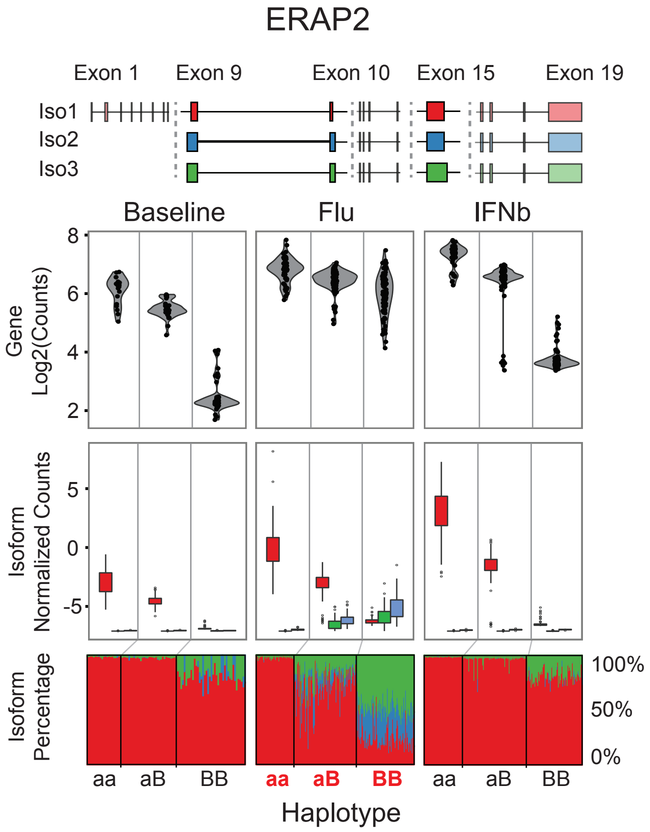
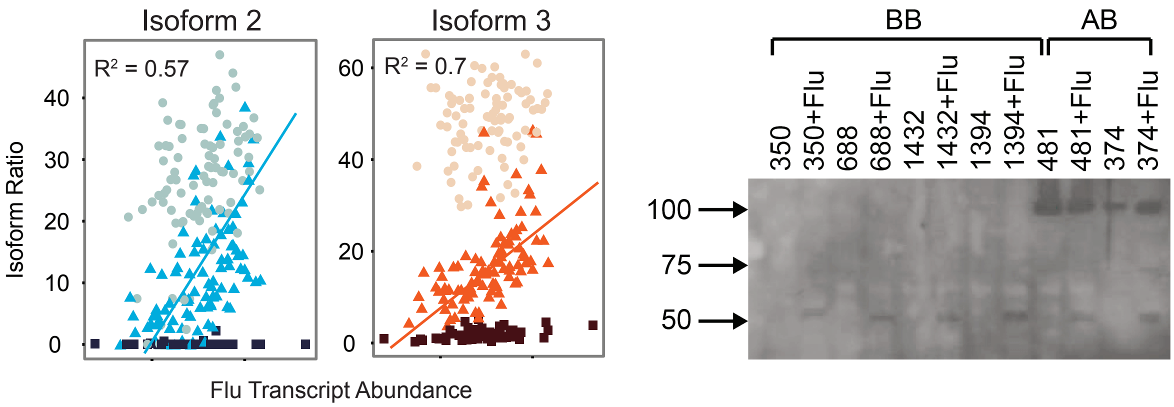


Figure 4-5: Transcript structure (top panel) and box-whisker plots (3 bottom panels) between *ERAP2* transcript quantitative traits (y-axis: $\log_2$(normalized gene counts), normalized isoform counts, or isoform percentage) and genotype (x-axis).

and *ERAP2*/Iso2 ($R^2 = 0.57$) and *ERAP2*/Iso3 ($R^2 = 0.70$) (Fig. 4-6a), (2) evidence of alternative translation starting at exon 11 as shown by the detection of flu-specific protein isoforms (50 kilodaltons) in flu-infected cells from Haplotype B homozygotes and heterozygotes (Fig. 4-6b), and (3) evidence of transcription of *ERAP2*/Iso2 or *ERAP2*/Iso3 (marked by an extended exon 10) in monocyte derived macrophages infected by H3N2 over a time course (fluomics, GEO: GSE97672) (Fig. C-5).

The complex genetic signals at the *ERAP2* locus is consistent with three perfectly linked variants on Haplotype B affecting *ERAP2* transcription and splicing in response to viral stimulation independent of Type 1 interferon signaling. rs2548538, an intronic variant that overlaps chromatin marks from LCLs [221], likely causes alternate transcript initiation at exon 9. rs2248374^G, the known splice site mutation, creates an alternate preferred splice site



(a) Correlation between ERAP2/Iso2 (blue) and ERAP2/Iso3 (orange) usage percentage (y-axis) and overall flu transcript abundance (x-axis) segregated by genotype (aa: squares, aB: triangles, BB: circles).

(b) Western blot of MoDCs before and after flu-infection from 5 Haplotype B homozygotes and 2 heterozygotes. Full length ERAP2 protein isoform is expected at 120 kDas. Two flu-specific ERAP2 protein isoforms are expected at 49 and 29 kDas.

Figure 4-6: Evidence for influenza-specific regulation of *ERAP2* isoforms.

resulting in alternative splicing of an extended exon 10. rs2549797^G, a splice-site mutation that creates a competing alternate splice site, results in ∼40% of the transcripts with an extended exon 15. The signature of natural selection, the previous disease associations, and the viral specific transcription suggest a critical antiviral role for the short *ERAP2* isoforms that could also result in an overactive auto-inflammatory response in Crohn's disease.

4.5 Conclusion

Although maps of genetic variants associated with overall transcript abundance have been generated in many tissue types, the genetic control of alternate isoform usage has not been extensively studied. Using *de novo* transcript reconstruction, we found a large number of previously uncharacterized transcripts in human dendritic cells, especially in response to influenza and interferon stimulation, indicating that the current reference human transcriptome is far from complete. We further found genetic variants (isoQTLs) associated with alternate isoform usage are widespread, approximately half of which are not associated with the overall abundance of the corresponding gene, indicative of independent genetic control of gene regulation at most loci of the genome.

IsoQTLs, like eQTLs, can affect gene expression at other loci in the genome suggesting important downstream effects on gene regulation. Different genetic variants in a locus could also affect multiple facets of gene regulation in response to stimulation in establishing transcriptome diversity and susceptibility to disease. This was clearly demonstrated at the ERAP2 locus where multiple variants on the Crohn's disease-associated haplotype lead to differential expression and splicing of the transcript in response to influenza. Previous experimental evidence has shown that full length *ERAP2* is a prototypical aminopeptidase that heterodimerizes with *ERAP1* [217] to perform peptide trimming during MHC class I presentation. The lack of the aminopeptidase domain in the flu-specific *ERAP2* isoforms suggests that it could interact with *ERAP1* to negatively influence antigen presentation or adopt previously unknown immunological function. Altogether, this dataset can help elucidate the mechanisms underlying disease alleles by providing deeper molecular data for each gene in baseline and inflammation.

4.6 Methods

4.6.1 Data collection

Study subjects

Donors were recruited from the Boston community and gave written informed consent for the studies. Individuals were excluded if they had a history of inflammatory disease, autoimmune disease, chronic metabolic disorders or chronic infectious disorders. Donors were between 18 and 56 years of age (mean 29.9 years).

Preparation and stimulation of primary human monocyte-derived dendritic cells

Influenza A (PR8 ΔNS1) was prepared as described in [213]. Recombinant human IFNβ was obtained from PBL Assay Science (Piscataway, NJ). Antibodies used were anti-IRF1 (sc-497x; Santa Cruz Biotechnology; Dallas, TX), anti-STAT2 (sc-476x; Santa Cruz Biotechnology) and anti-IRF9 (sc-10793x; Santa Cruz Biotechnology).

As previously described in [202], 35 - 50 mL of peripheral blood from fasting subjects was collected between 7:30 - 8:30 am. The blood was drawn into sodium heparin tubes and peripheral blood mononuclear cells (PBMCs) were isolated by Ficoll-Paque (GE Healthcare Life Sciences; Uppsala, Sweden) centrifugation. PBMCs were frozen in liquid N_2 in 90% FBS (Sigma-Aldrich; St. Louis, MO) and 10% DMSO (Sigma-Aldrich). Monocytes were isolated from PBMCs by negative selection using the Dynabeads Untouched Human Monocytes kit (Life Technologies; Carlsbad, CA) modified to increase throughput and optimize recovery and purity of $CD14^{+}CD16^{lo}$ monocytes: the FcR Blocking Reagent was replaced with Miltenyi FcR Blocking Reagent (Miltenyi; Bergisch Gladbach, Germany); per mL of Antibody Mix, an additional 333 μg biotinylated anti-CD16 (3G8), 167 μg biotinylated anti-CD3 (SK7) and 167 μg biotinylated anti-CD19 (HIB19) antibodies (Biolegend; San Diego, CA) were added; the antibody labeling was modified to be performed in 96-well plates; and Miltenyi MS Columns or Multi-96 Columns (Miltenyi) were used to separate magnetically-labeled cells from unlabeled cells in an OctoMACS Separator or MultiMACS M96 Separator (Miltenyi) respectively. The number of PBMCs and monocytes was estimated using CellTiter-Glo Luminescent Cell Viability Assay (Promega; Madison, WI). A subset of the isolated monocytes was stained with PE-labeled anti-CD14 (M5E2; BD Biosciences; Franklin Lakes, NJ) and FITC-labeled anti-CD16 (3G8; Biolegend), and subjected to flow cytometry analysis using an Accuri C6 Flow Cytometer (BD Biosciences). A median of 94% CD14+ cells and 99% CD16lo cells was obtained. The remaining monocytes were cultured for seven days in RPMI (Life Technologies) supplemented with 10% FBS, 100 ng/mL GM-CSF (R&D Systems; Minneapolis, MN) and 40 ng/mL IL-4 (R&D Systems) to differentiate the monocytes into monocyte-derived dendritic cells (MoDCs). 4×10^4 MoDCs were seeded in each well of a 96-well plate, and stimulated with influenza virus for 10 hours, 100 U/mL IFNβ for 6.5 hours or left unstimulated. Cells were then lysed in RLT buffer (Qiagen; Hilden, Germany) supplemented with 1% β-mercaptoethanol (Sigma-Aldrich).

DNA extraction and genotyping

As previously described [202], genomic DNA was extracted from 5 mL whole blood (DNeasy Blood & Tissue Kit; Qiagen), and quantified by Nanodrop. Each subject was genotyped using the Illumina Infinium Human OmniExpress Exome BeadChips, which includes genome-wide genotype data as well as genotypes for rare variants from 12,000 exomes as well as common coding variants from the whole genome. In total, 951,117 SNPs were genotyped, of which 704,808 SNPs are common variants (Minor Allele Frequency [MAF] > 0.01) and 246,229 are part of the exomes. The genotype success rate was greater than or equal to 97%.

RNA isolation and sequencing

RNA from all samples was extracted using the RNeasy 96 kit (Qiagen, cat. #74182), according to the manufacturer's protocols. 576 total samples were sequenced (99 baseline, 250 influenza infected, and 227 interferon stimulated). 552 pass filter samples (94 baseline, 243 influenza, and 215 interferon) were sequenced to an average depth of 38M 76 basepair paired end reads using the Illumina TruSeq kit with 86% mapping to transcriptome and 97% mapping to the genome (Additional file 1).

4.6.2 Data processing

Adjusting for expression heterogeneity

We empirically determined the number of principal components to adjust for each stimulation condition and either overall gene abundance or isoform percentage. Because of the smaller number of individuals in the baseline study, the number of principal components adjusted is fewer (Fig. C-1). Because the isoform percentage implicitly adjusts for confounders that affect overall gene abundance and isoform abundance levels (i.e., other eQTLs), the number of adjusted PCs is also fewer (Fig. C-1).

Transcriptome reconstruction

After aligning reads to the genome, we reconstructed transcriptomes for each sample individually using StringTie [222] using default parameters and quantified the abundances of annotated transcripts using kallisto [223]. For genes expressed at > 5 transcripts per million (TPM) in any sample, we removed isoforms expressed at < 5 TPM across all samples. In order to preserve isoforms that may be uniquely expressed in a single condition (e.g., baseline, flu, IFN), transcriptomes within the same conditions were first merged before transcriptomes across all three conditions were merged, using cuffcompare [186]. As a final step, cuffmerge [186] (--overhang-tolerance 0) was used to remove redundant isoforms.

Transcriptome quantification

Differential expression testing was carried out with sleuth [224] using 100 bootstraps per sample. Gene-level quantification was estimates by summing isoform counts and differential expression testing was carried out with DESeq2 [225].

Beta regression

Differential isoform ratio testing was carried out in R using beta regression package betareg [226] and p-values were calculated using likelihood ratio test and adjusted with a false discovery rate adjustment [142].

DNA genotyping

We applied rigorous subject and SNP quality control (QC) that includes (1) gender misidentification, (2) subject relatedness, (3) Hardy-Weinberg Equilibrium testing, (4) use concordance to infer SNP quality, (5) genotype call rate, (6) heterozygosity outlier, (7) subject mismatches. In the European population, we excluded 1,987 SNPs with a call rate $< 95\%$, 459 SNPs with Hardy-Weinberg equilibrium p-value $< 10^{-6}$, 234 SNPs with a MisHap p-value $< 10^{-9}$, and 63,781 SNPs with MAF $< 1\%$ from (a total of 66,461 SNPs excluded). In the African-American population, we excluded 2,161 SNPs with a call rate $< 95\%$, 298 SNPs with Hardy-Weinberg equilibrium p-value $< 10^{-6}$, 50 SNPs with a MisHap p-value $< 10^{-9}$, and 17,927 SNPs with MAF $< 1\%$ from (a total of 20,436 SNPs excluded). In the East Asian population, we excluded 1,831 SNPs with a call rate $< 95\%$, 213 SNPs with Hardy-Weinberg equilibrium p-value $< 10^{-6}$, 47 SNPs with a MisHap p-value $< 10^{-9}$, and 84,973 SNPs with MAF $< 1\%$ from (a total of 87,064 SNPs excluded). After QC, 52 subjects across all three populations and approximately 18,000 – 88,000 SNPs in each population were filtered out from our analysis.

Underlying genetic stratification in the population was assessed by multi-dimensional scaling using data from the International HapMap Project [227] (CEU, YRI and CHB samples) combined with IBS cluster analysis using the Eigenstrat 3.0 software [228].

The quality control of the genotyping data were performed using PLINK [229].

Genotype imputation

To accurately evaluate the evidence of association signal at variants that are not directly genotyped, we used the BEAGLE software (version: 3.3.2) [230] to imputed the post-QC genotyped markers using reference haplotype panels from the 1000 Genomes Project (version 3) [3], which contain a total of 37.9M SNPs in 1,092 individuals with ancestry from West Africa, East Asia, and Europe. For subjects of European and East Asian ancestry, we used haplotypes from Utah residents (CEPH) with Northern and Western European ancestry (CEU), and combined panels from Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT), respectively. For imputing African American subjects, we used a combined haplotype reference panel consisting of CEU and Yoruba in Ibadan, Nigeria (YRI). For the admixed African American population, using two reference panels substantially improves imputation performance. After genotype imputation, we filtered out low MAF SNPs (MAF < 0.01), which resulted in 7.7M, 6.6M, 12.7M common variants in European, East Asian and African American, respectively. This set of genotyped and imputed markers was used for all the subsequent association analysis.

QTL mapping

QTL mapping was performed using the Matrix eQTL [231] package using empirically determined number of principal components as covariates for each analysis (Fig.C-1). For *cis* QTLs, isoform usage or overall gene abundance were regressed against all genetics variants with a MAF $> 5\%$ in a 1 megabase (+/- 500 kb) window. Empirical p-values were calculated by comparing the nominal p-values with null p-values determined by permuting each isoform/gene 1,000 times [232]. False discovery rates were calculated using the qvalue [233] package as previously described [234].

QTL annotation

QTLs were annotated using Variant Effect Predictor and Ensembl release 79 [235]. Exonic and intronic locations of QTLs were determined using UCSC's canonical transcripts (table knownCanonical) as a reference [235]. Enrichments were calculated against background set of SNPs that were matched in allele frequency (binned by 4%) and distance to nearest transcription start site (binned by 10 kb).

GWAS associations

The GREGOR suite [236] was used for calculating the enrichment of eQTLs and isoQTLs containing a GWAS loci across baseline, flu, and IFNβ stimulations. GWAS associations for disease with FDR < 0.1 are reported.

Estimating flu transcript abundance

Flu transcript abundance was estimated by running RSEM [138] on a custom reference of the influenza PR8 genome.

4.6.3 Experimental validation

ERAP2 Western Blot

Protein extracts were fractionated by SDS-PAGE (4-12% Bis-Tris gel, Thermo scientific, NP0335BOX) and transferred to PVDF membrane (BioRad, cat. #162-0177). After blocking with 2% BSA in TBST (Tris buffered saline containing 0.1% tween-20) for 1 hour, membranes were incubated with primary antibody (either ERAP2, R&D Systems, cat# AF3830, 1:3000) or b-actin (Abcam, cat. #ab6276, 1:15,000) overnight at 4C. Membranes were then washed and incubated with a 1:5000 dilution of HRP conjugated secondary antibody (either donkey anti-goat from Santa Cruz Biotech cat. #sc2020, or with goat anti-mouse from Jackson immune Research cat. #115-035-146) for 1 hour. Membranes were washed and developed with ECL system (VWR, cat. #89168-782) according to the manufacturer's protocol.

4.6.4 Data availability

Processed RNA-sequencing data is available under GEO accession GSE92904. Raw fastq data is available from dbGAP under accession phs000815.v1.p1.

4.6.5 Additional files

Additional file 1: Sequencing statistics. [link]
Additional file 2: Reconstructed transcriptome annotations. [link]
Additional file 3: Differential isoform usage analysis results. [link]
Additional file 4: eQTL results. [link]
Additional file 5: isoQTL results. [link]
Additional file 6: r-isoQTL results. [link]

4.6.6 Authors' contributions

CJY conceived of the study, participated in its design and coordination, and carried out all genotyping and QTL analysis. JC participated in the design of the study, and carried out transcriptome reconstruction, differential isoform usage analysis, and QTL annotations. ACV performed experimental validations. MS, REG, and NR participated in data analysis. TBhangale, MNL, TR, RR, WL, SHI, PIC, CM, MHL, and IYF, participated in data collection and sequencing. BES, PLDJ, AR, TBehrens, and NH participated in conception, design, and coordination of the study. The manuscript was written by CJY and JC with input from AR and NH.

Additional thanks to Geo Pertea, Brian Haas, Sean Simmons, and members of the Regev, Hacohen, and Ye laboratory for helpful discussions.

Supplementary materials for Chapter 4

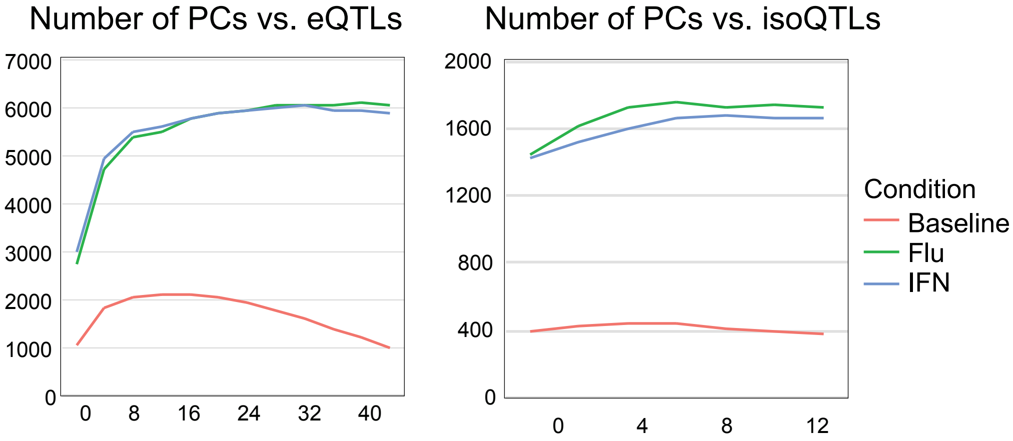


Figure C-1: Empirically determined number of principal components (PCs) to adjust for eQTLs (left) and isoQTLs (right) in each of three conditions.

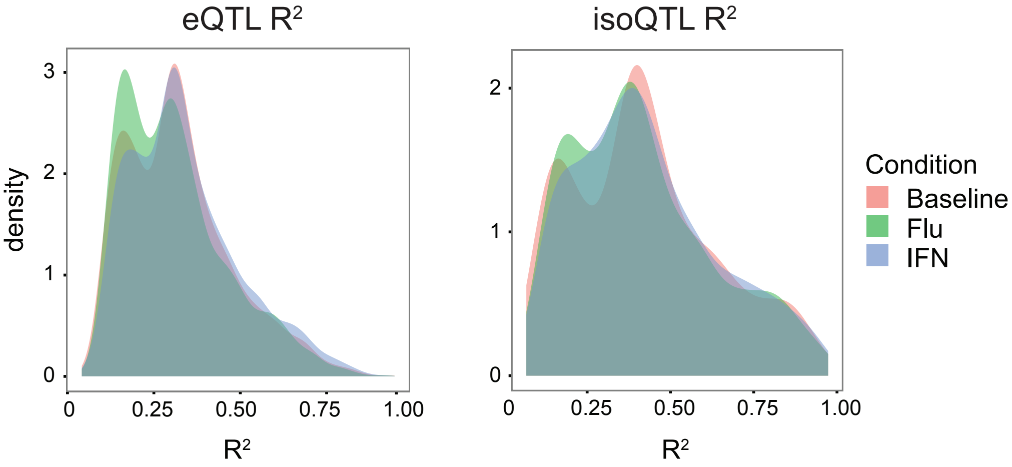


Figure C-2: Effect size (R^2) distribution of eQTLs and isoQTLs at each of three conditions.

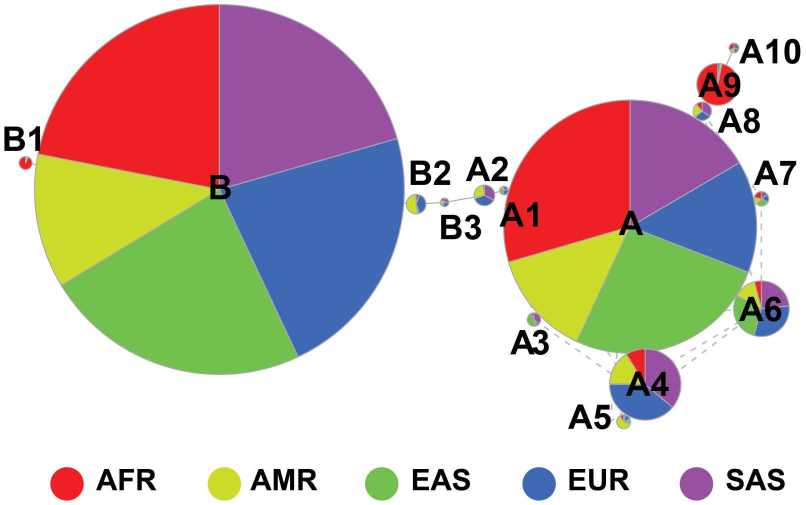


Figure C-3: Distribution of *ERAP2* haplotypes based on human populations from 1000 Genomes Project. AFR: African; AMR: American; EAS: East Asian; EUR: European; SAS: South Asian.

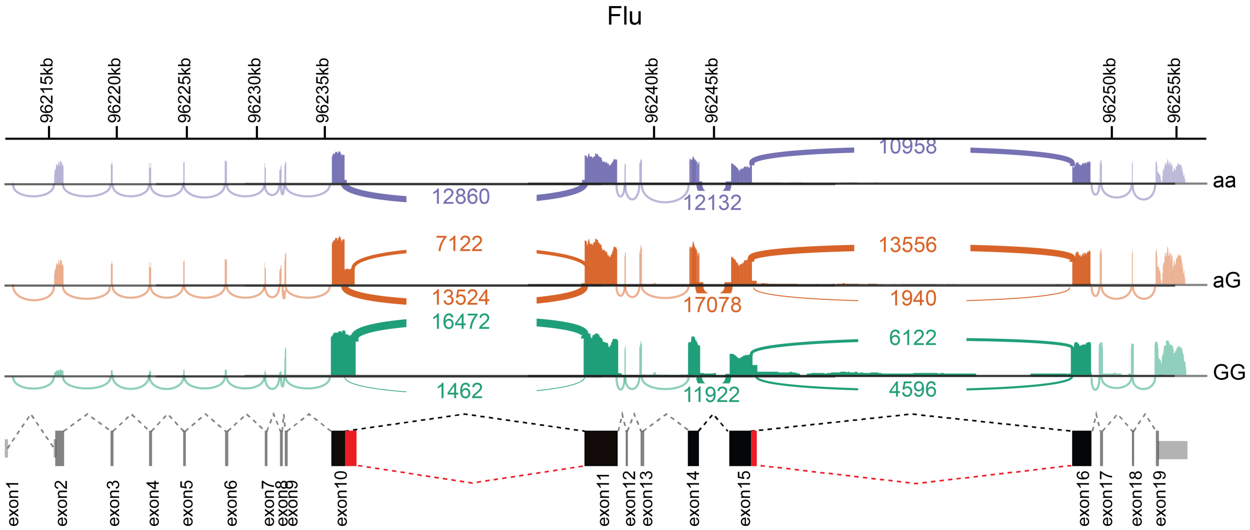


Figure C-4: Diagram of *ERAP2* isoform abundances in response to influenza as a function of patient genotype.

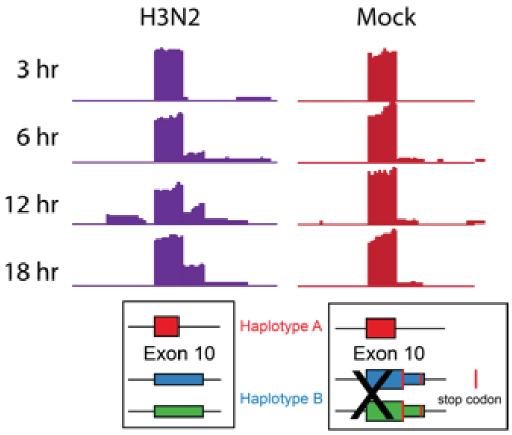


Figure C-5: *ERAP2* isoform abundances in monocyte derived macrophages infected by H3N2 over a time course (left) compared to mock infections (right).

Discussion

Comparative transcriptomics holds great promise for studying and understanding the contributions of gene regulation and expression to biology and physiology. The evolutionary signatures contained within comparative transcriptomic data harbor evolutionary stories that provide abundance of clues to each gene's role across a diversity of species phenotypes. However, extracting knowledge from comparative transcriptomic data requires accurate models transcriptional evolution and statistically principled analysis methods to isolate the evolutionary signal from the noise.

While the field of comparative genomics is now decades old, many of the comparative genomics methods that have been developed and widely accepted are targeted at comparative sequence analysis are not appropriate for analyzing transcriptional properties that (1) are often continuous traits (as opposed to the discrete nature of DNA sequence), and (2) evolve under mechanisms that are still yet to be understood. As a result, comparative transcriptomics studies frequently suffer from shallow analyses that give way to global patterns but cannot test specific hypotheses, or make incorrect assumptions that lead to erroneous interpretations.

Here, I present quantitative methods for analyzing comparative transcriptomic data in order to test precise hypotheses about whether expression levels and transcriptional structures are evolving under unique selective pressures, indicative of biological function. First, I present a framework that utilizes the stochastic Ornstein-Uhlenbeck process to model optimal gene expression levels as probabilistic distributions that, in turn, enable statistical testing for deleterious expression in disease tissues and directional selection along specific phylogenetic branches. Second, I present novel metrics for evaluating the conservation of lncRNAs that reveal unique evolutionary histories behind the heterogeneous class of genes.

Both studies highlight the unsuitability of simply applying models and metrics used in sequence analysis to transcriptional traits. Unlike the widespread neutral evolution of base-pair substitutions, expression is evolving nonlinearly under strong purifying selection and requires a more complex model than a simple random walk for accurate hypothesis testing; and unlike coding genes, which are almost always conserved in transcription, lncRNAs turn over rapidly in transcription and their conservation cannot be accurately quantified by sequence conservation alone. However, with the proper analytical tools, both studies also highlight the power of using evolutionary signatures to refine hypotheses about the operations and mechanisms of the genome. Though it was initially hypothesized that gene expression differences would explain many species-specific traits, analysis of the evolution of gene expression reveals that the majority of genes' expression levels are under strong purifying selection. While this observation can be utilized to improve the annotation of essential expression pathways and distributions of optimal expression levels, it also leaves open the question of where species-specific traits are arising from. Additionally, the initial discovery of lncRNAs was followed by speculations as to their fundamental regulatory roles in cellular biology, but the observation that they turn over rapidly across species point to the idea that many of them may not play significant functional roles. Further, the evolutionary signatures of lncRNA transcription suggest that those that are conserved are comprised of different classes of genes that may have distinct biological functions.

The studies presented here represent just the beginning of what comparative transcriptomics can teach us. So far, we have only profiled transcriptional profiles from a small number of differentiated, steady-state adult organs from mammalian species. However, the analysis of splicing events in human immune cells before and after viral stimulation demonstrates that many regulatory events are only observable in dynamic states. It may be that the majority of species-specific phenotypic diversity arises from transcriptional differences that occur in development or in response to environmental stimuli. While these tissue contexts are extremely difficult to obtain directly from higher mammals, current bioengineering advancements in *in vitro* cellular reprogramming and organoid development now permits us to begin to investigate transcriptional evolution in these dynamic contexts [237–239]. Looking forward, the resolution of transcriptional data can be increased not only through expanded repertoires of tissues profiled, but also through current progress in single cell RNA-sequencing technologies which will further facilitate the decomposition of cell type proportional differences from regulatory differences - an aspect of transcriptional variation that cannot be addressed with current bulk RNA-seq methods.

Additionally, my thesis work focused only on characterizing the transcriptional results of regulatory evolution, with limited analyses integrating genomic regulatory data or noncoding sequence. In part, this is due to the immense difficulties analyzing comparative regulatory data: the few comparative regulatory studies that have been conducted thus far have found that regulatory sequence and enhancers evolve rapidly and cannot be easily traced across currently profiled species [240, 241]. To address this limitation, data collection is needed across more species, ideally from a dense phylogeny in which noncoding sequences and enhancer elements have not yet degraded beyond the point where ancestral events cannot be inferred. The successful integration of both comparative regulatory and transcriptomic data would be a huge methodological advancement for mapping noncoding sequences to phenotypic traits.

Finally, comparative transcriptomic analysis suffers greatly from our lack of understanding of the biological mechanisms by which transcription is controlled. Without this knowledge, it is near impossible to develop proper models of transcriptional evolution to fully explain the complexities of comparative genomics data. The studies in this thesis present two strategies for overcoming this issue: In the first study, I utilized a model that globally fits the data well, but doesn't necessarily represent the mechanism by which expression actually evolves (i.e., by a bounded random walk process). In the second study, I used permutation-based methods to estimate the significance of observations, which avoids the need for a model, but can lead to results that are highly sensitive to the permutation strategy. While both strategies have been informative to our understanding of transcriptional evolution, they are far from being optimal tools for interrogating comparative transcriptional data. In this regard, comparative analysis could benefit greatly from directed evolution studies that help elucidate how mutations affect transcription from generation to generation, in both the presence and absence of selective pressure. For example, a few studies have already begun to suggest that random sequences code for transcription and transcription factor binding much more frequently than initially expected [242, 243] which has implications for what a null model for functional transcription should be. Experimental efforts to develop models of transcriptional evolution will only serve to better refine our data collection and data analysis methods for even more accurate characterization of transcriptional evolution.

The 21^{st} century is an exhilarating time to be a genomic scientists, with a dizzying array of high-throughput profiling technology, and even mammalian genome editing technology, available for conducting experiments and collecting data. Amidst the sea of genomic data, however, still lays an uncompleted mission of comprehending genome biology to the extent such that we can easily read, interpret, and predictably edit genomic sequence. The incredible existing diversity of species and animal phenotypes, generated by a universal DNA code, presents us with a rich sampling of unique genotype-phenotyping pairings, and offers us an opportunity to use comparative methodologies to deduce the roles and mechanisms of genomic elements. Understanding of how to identify and interpret evolutionary signatures harbored by both coding and noncoding DNA sequence remains a powerful strategy for deciphering genomic properties. As genomic science moves beyond simply whole-genome sequencing and into probing uncharted depths of transcriptomic and epigenomic traits, so too, must comparative methodologies advance in order to complement and inform experimental strategies, refine scientific models of how the genome operates, and contribute towards fully decoding the complexities of the genome.

Afterword

I wrote this essay at the end of my fifth year of graduate school, shortly after I made it out of the "valley of despair". I wrote it as a reminder to myself of a very important leg of my PhD journey and I offer it now for any future graduate students who may find themselves in a similar struggle.

In my second year of grad school, I fell off my bike. After making the incredibly stupid decision to bike home in the dark without bike lights, I hit a curb at full speed, flipped over the handlebars, and crashed into the sidewalk. A stranger came running out of her house yelling "Are you okay??" I sobbed a bit from the shock; she rubbed my back and gave me bandaids and Neosporin for my wounds. The next day, a friend who saw my bruises gave me very good advice. She said, "If you want to be able to ride your bike again, get back on that bike as soon as possible. It will be one of the hardest thing you'll ever do, but the longer you wait, the less likely you'll ever get on a bike again. Just start out slow." She was right. Getting back on the bike was terrifying and panic-inducing, but in time, I went from biking slower than I could walk to biking like nothing had ever happened.

In the fourth year of grad school, I fell off my proverbial PhD bike. After spending years analyzing a dataset, the only conclusion I could reach was a less-than-exciting negative result that was difficult to publish. This time, it was less of a crash and more of a very, very, very slow derailment into a ditch. This time, people were less helpful. Many people pretended like I hadn't fallen off. "Just keep riding!" they yelled at me with two thumbs up, as they quickly sped by. Others told me I really should have seen that ditch coming when I picked such a busy advisor. Some people told me that I'd come far enough already and shouldn't be whining about being in a ditch when there were bikeless children in Africa.

Not everyone was terribly unhelpful. A few good friends came and hung out with me in the ditch, but of course I couldn't expect them to stay forever. One astute friend peered in and remarked, "That looks a lot deeper than the other ditches I've seen. Maybe you should try to find professional help for getting out of a ditch that size." The professionals couldn't tell me exactly how to get out of the ditch, but reassured me that it was okay to be angry about being in it, even if there were bikeless children in Africa. The most helpful friends encouraged me to get back on my bike, regain my confidence, and get outta the goddamn ditch.

Like the process of getting back on my actual bike, I started out slow. I put my paper on a preprint server and received emails from appreciative scientists who were grateful to hear about the negative result sooner than later. I contributed my expertise to other projects going on in the lab and brainstormed with younger grad students about what projects to pursue. I volunteered at a local middle school to teach genetics and evolution and rediscovered why I entered science to begin with. Many months later, I finally found myself able to think about embarking on a new science project without becoming immobilized with panic and dread.

Every year, I TA a genetics course and tell the incoming graduate students, "Life is a journey, not a race." Now, I understand that I forgot to add "Life is a journey, not a race, but either way you might fall off your proverbial PhD bike and it will be really confusing because it won't exactly hurt when you fall and no one will run over asking if you're okay, but months later you'll find yourself overwhelmed and terrified at the mere thought of having to complete a thesis." Perhaps a better message would be, "Life is a journey, not a race, and if you find that you've fallen off your bike at any point in this journey, just get back on that bike as soon as possible. It will be one of the hardest thing you'll ever do, but the longer you wait, the less likely you'll ever get on a bike again. Just start out slow."

Bibliography

[1] Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

[2] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[3] Siva, N. 1000 genomes project. *Nat Biotechnol* **26**, 256 (2008).

[4] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).

[5] Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

[6] Tyner, C. *et al.* The ucsc genome browser database: 2017 update. *Nucleic Acids Res.* **45**, D626–34 (2017).

[7] Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).

[8] Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108–117 (2003).

[9] Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).

[10] Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).

[11] Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

[12] Lewis, B. P., Shih, I.-H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).

[13] Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).

[14] Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19428–19433 (2007).

[15] Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–82 (2011).

[16] Washietl, S. *et al.* RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**, 578–594 (2011).

[17] Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

[18] Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).

[19] Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).

[20] Dickel, D. E. *et al.* Ultraconserved enhancers are required for normal development. *Cell* **172**, 491–499.e15 (2018).

[21] Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, e168 (2006).

[22] Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130025 (2013).

[23] van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* (2010).

[24] Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

[25] Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).

[26] Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).

[27] Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).

[28] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

[29] Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

[30] Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

[31] Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18**, 425–440 (2017).

[32] López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* **9**, 583–593 (2008).

[33] King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

[34] Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

[35] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).

[36] Mikkelsen, T. S. *et al.* Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).

[37] McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011).

[38] Kvon, E. Z. *et al.* Progressive loss of function in a limb enhancer during snake evolution. *Cell* **167**, 633–642.e11 (2016).

[39] Indjeian, V. B. *et al.* Evolving new skeletal traits by cis-regulatory changes in bone morphogenetic proteins. *Cell* **164**, 45–56 (2016).

[40] Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 4465–4470 (2002).

[41] Yanai, I., Graur, D. & Ophir, R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* **8**, 15–24 (2004).

[42] Liao, B.-Y. & Zhang, J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* **23**, 530–540 (2006).

[43] Chan, E. T. *et al.* Conservation of core gene expression in vertebrate tissues. *J. Biol.* **8**, 33 (2009).

[44] Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).

[45] Cáceres, M. *et al.* Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 13030–13035 (2003).

[46] Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**, 1850–1854 (2005).

[47] Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. & White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**, 242–245 (2006).

[48] Khaitovich, P. *et al.* A neutral model of transcriptome evolution. *PLoS Biol.* **2**, E132 (2004).

[49] Gilad, Y., Oshlack, A. & Rifkin, S. A. Natural selection on gene expression. *Trends Genet.* **22**, 456–461 (2006).

[50] Whitehead, A. & Crawford, D. L. Variation within and among species in gene expression: raw material for evolution. *Mol. Ecol.* **15**, 1197–1211 (2006).

[51] Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K. & Gilad, Y. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* **4**, e1000271 (2008).

[52] Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17224–17229 (2014).

[53] Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

[54] Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* **22**, 602–610 (2012).

[55] Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16**, 287 (2015).

[56] Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M. & Barkai, N. On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* **4**, 159 (2008).

[57] Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).

[58] Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34**, 177–180 (2003).

[59] Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2850–2855 (2005).

[60] Calarco, J. A. *et al.* Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.* **21**, 2963–2975 (2007).

[61] Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* **35**, 125–131 (2007).

[62] Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).

[63] Chen, L., Bush, S. J., Tovar-Corona, J. M., Castillo-Morales, A. & Urrutia, A. O. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Mol. Biol. Evol.* **31**, 1402–1413 (2014).

[64] Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).

[65] Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).

[66] Gracheva, E. O. *et al.* Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature* **476**, 88–91 (2011).

[67] Terai, Y., Morikawa, N., Kawakami, K. & Okada, N. The complexity of alternative splicing of hagoromo mRNAs is increased in an explosively speciated lineage in east african cichlids. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12798–12803 (2003).

[68] Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).

[69] Rinn, J. L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17**, 529–540 (2003).

[70] Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).

[71] Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

[72] Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).

[73] Ballabio, A. *et al.* Deletions of the steroid sulphatase gene in "classical" x-linked ichthyosis and in x-linked ichthyosis associated with kallmann syndrome. *Hum. Genet.* **77**, 338–341 (1987).

[74] Wang, J. *et al.* Neutral evolution of 'non-coding' complementary DNAs. *Nature* **431** (2004).

[75] Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5 (2006).

[76] Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007).

[77] Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* **10**, R124 (2009).

[78] Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).

[79] Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).

[80] Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).

[81] Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* (2014).

[82] Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).

[83] Somarowthu, S. *et al.* HOTAIR forms an intricate and modular secondary structure. *Mol. Cell* **58**, 353–361 (2015).

[84] Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2017).

[85] Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).

[86] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

[87] Harris, H. Enzyme polymorphisms in man. *Proc. R. Soc. Lond. B Biol. Sci.* **164**, 298–310 (1966).

[88] Lewontin, R. C. & Hubby, J. L. A molecular approach to the study of genic heterozygosity in natural populations. II. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics* **54**, 595–609 (1966).

[89] Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

[90] Kimura, M. & Ohta, T. Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469 (1971).

[91] Jukes, T. H. & King, J. L. Deleterious mutations and neutral substitutions. *Nature* **231**, 114–115 (1971).

[92] Pierce, V. A. & Crawford, D. L. Phylogenetic analysis of glycolytic enzyme expression. *Science* **276**, 256–259 (1997).

[93] Wang, D., Marsh, J. L. & Ayala, F. J. Evolutionary changes in the expression pattern of a developmentally essential gene in three drosophila species. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 7103–7107 (1996).

[94] Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9721–9726 (1999).

[95] Fraser, H. B., Moses, A. M. & Schadt, E. E. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2977–2982 (2010).

[96] Bedford, T. & Hartl, D. L. Optimization of gene expression by natural selection. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1133–1138 (2009).

[97] Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).

[98] Harr, B. & Turner, L. M. Genome-wide analysis of alternative splicing evolution among mus subspecies. *Mol. Ecol.* **19 Suppl 1**, 228–239 (2010).

[99] Pipes, L. *et al.* The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res.* **41**, D906–14 (2013).

[100] Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).

[101] Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* **25**, 167–178 (2015).

[102] Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).

[103] Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32**, 835–845 (2015).

[104] Hansen, T. F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**, 1341–1351 (1997).

[105] Nourmohammad, A. *et al.* Adaptive evolution of gene expression in drosophila. *Cell Rep.* **20**, 1385–1395 (2017).

[106] Rohlfs, R. V. & Nielsen, R. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Syst. Biol.* **64**, 695–708 (2015).

[107] Butler, M. A. & King, A. A. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist* **164**, 683–695 (2004).

[108] Silvestro, D., Kostikova, A., Litsios, G., Pearman, P. B. & Salamin, N. Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods Ecol. Evol.* **6**, 340–346 (2015).

[109] Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).

[110] Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* **23**, 1063–1068 (2013).

[111] Jordan, D. M. *et al.* Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524**, 225–229 (2015).

[112] Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.* **17**, 405–424 (2015).

[113] Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).

[114] Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9**, e1003484 (2013).

[115] Rehm, H. L. *et al.* ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

[116] Banerjee-Basu, S. & Packer, A. SFARI gene: an evolving database for the autism research community. *Dis. Model. Mech.* **3**, 133–135 (2010).

[117] Amberger, J. S. & Hamosh, A. Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Curr. Protoc. Bioinformatics* **58**, 1.2.1–1.2.12 (2017).

[118] Blake, J. A. *et al.* Mouse genome database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45**, D723–D729 (2017).

[119] Cummings, B. B. *et al.* Improving genetic diagnosis in mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).

[120] Schieve, L. A. *et al.* Concurrent medical conditions and health care use and needs among children with learning and behavioral developmental disabilities, national health interview survey, 2006-2010. *Res. Dev. Disabil.* **33**, 467–476 (2012).

[121] Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19096–19101 (2009).

[122] Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).

[123] O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).

[124] Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare mendelian disorders. *JAMA* **312**, 1880–1887 (2014).

[125] Lonsdale, J. *et al.* The Genotype-Tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

[126] Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).

[127] Brown, M. S. & Goldstein, J. L. Receptor-mediated endocytosis: insights from the lipoprotein receptor system. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 3330–3337 (1979).

[128] Guerra, R., Wang, J., Grundy, S. M. & Cohen, J. C. A hepatic lipase (LIPC) allele associated with high plasma concentrations of high density lipoprotein cholesterol. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4532–4537 (1997).

[129] Kersten, S. Integrated physiology and systems biology of PPARα. *Mol Metab* **3**, 354–371 (2014).

[130] von Scheidt, M. *et al.* Applications and limitations of mouse models for understanding human atherosclerosis. *Cell Metab.* **25**, 248–261 (2017).

[131] Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).

[132] Areal, H., Abrantes, J. & Esteves, P. J. Signatures of positive selection in toll-like receptor (TLR) genes in mammals. *BMC Evol. Biol.* **11**, 368 (2011).

[133] Torgerson, D. G., Kulathinal, R. J. & Singh, R. S. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol. Biol. Evol.* **19**, 1973–1980 (2002).

[134] Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. Positive darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2509–2514 (2001).

[135] Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nat. Methods* **7**, 709–715 (2010).

[136] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

[137] Kinsella, R. J. *et al.* Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database* bar030 (2011).

[138] Li, B. & Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

[139] Huerta-Cepas, J., Serra, F. & Bork, P. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–8 (2016).

[140] Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol.* **11**, R25 (2010).

[141] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).

[142] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

[143] Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

[144] Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

[145] Harrow, J. *et al.* GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).

[146] Greider, C. W. & Blackburn, E. H. A telomeric sequence in the RNA of tetrahymena telomerase required for telomere repeat synthesis. *Nature* **337**, 331–337 (1989).

[147] Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570–1573 (2005).

[148] Guan, Y. *et al.* Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* **13**, 5745–5755 (2007).

[149] Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113–1117 (2010).

[150] Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).

[151] Prensner, J. R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).

[152] Ellis, B. C., Molloy, P. L. & Graham, L. D. CRNDE: A long Non-Coding RNA involved in CanceR, neurobiology, and DEvelopment. *Front. Genet.* **3**, 270 (2012).

[153] Flockhart, R. J. *et al.* BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res.* **22**, 1006–1014 (2012).

[154] Carpenter, S. *et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–792 (2013).

[155] Pauli, A. *et al.* Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577–591 (2012).

[156] Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**, 1775–1789 (2012).

[157] Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).

[158] Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proceedings of the National Academy of Sciences* **107**, 9222–9227 (2010).

[159] Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282–286 (2013).

[160] Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).

[161] Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).

[162] Bafna, V. & Huson, D. H. The conserved exon method for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 3–12 (2000).

[163] Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**, S140–8 (2001).

[164] Pachter, L., Alexandersson, M. & Cawley, S. Applications of generalized pair hidden markov models to alignment and gene finding problems. *J. Comput. Biol.* **9**, 389–399 (2002).

[165] Wenger, A. M. *et al.* PRISM offers a comprehensive genomic approach to transcription factor function prediction. *Genome Res.* **23**, 889–904 (2013).

[166] Harris, R. S. *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, Pennsylvania State University (2007).

[167] Marchetto, M. C. N. *et al.* Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**, 525–529 (2013).

[168] Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA firre. *Nat Struct Mol Biol.* **21**, 198–206 (2014).

[169] Smith, C. M. & Steitz, J. A. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell. Biol.* **18**, 6897–6909 (1998).

[170] Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**, e01749–e01749 (2013).

[171] Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

[172] Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10**, 28–36 (1990).

[173] Grant, J. *et al.* Rsx is a metatherian RNA with xist-like properties in x-chromosome inactivation. *Nature* **487**, 254–258 (2012).

[174] Hinrichs, A. S. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

[175] Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).

[176] Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).

[177] Guo, G. *et al.* Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).

[178] Hanna, J. *et al.* Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell* **4**, 513–524 (2009).

[179] Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods* **12**, 323–325 (2015).

[180] Karolchik, D. *et al.* The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–70 (2014).

[181] Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–63 (2014).

[182] Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392 (2012).

[183] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

[184] Garber, M. *et al.* A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47**, 810–822 (2012).

[185] Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).

[186] Trapnell, C. *et al.* Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat Protoc* **7**, 562–78 (2012).

[187] Fitch, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416 (1971).

[188] Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–55 (2014).

[189] Benaglia, T., Chauveau, D., Hunter, D. & Young, D. mixtools: An r package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).

[190] Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–62 (2009).

[191] Smit, A. F. A., Hubley, R. & Green, P. Repeatmasker. http://www. repeatmasker. org (1996).

[192] Black, D. L. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem* **72**, 291–336 (2003).

[193] Maquat, L. E. Nonsense-mediated mrna decay: splicing, translation and mrnp dynamics. *Nat Rev Mol Cell Biol* **5**, 89–99 (2004).

[194] Matlin, A. J., Clark, F. & Smith, C. W. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**, 386–98 (2005).

[195] Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6 (2008).

[196] Enders, A. *et al.* Zinc-finger protein zfp318 is essential for expression of igd, the alternatively spliced igh product made by mature b lymphocytes. *Proc Natl Acad Sci U S A* **111**, 4513–8 (2014).

[197] Berard, M. & Tough, D. F. Qualitative differences between naive and memory t cells. *Immunology* **106**, 127–138 (2002).

[198] Martinez, N. M. & Lynch, K. W. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol Rev* **253**, 216–36 (2013).

[199] Xiong, H. Y. *et al.* Rna splicing. the human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

[200] Graham, R. R. *et al.* Three functional variants of ifn regulatory factor 5 (irf5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci U S A* **104**, 6758–63 (2007).

[201] Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–63 (2009).

[202] Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).

[203] Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1204–1209 (2012).

[204] Fairfax, B. P. & Knight, J. C. Genetics of gene expression in immunity to infection. *Curr Opin Immunol* **30**, 63–71 (2014).

[205] Quach, H. *et al.* Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656 e17 (2016).

[206] Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).

[207] Li, Y. I. *et al.* Rna splicing is a primary link between genetic variation and disease. *Science* **352**, 600–4 (2016).

[208] Consortium, G. T. Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–60 (2015).

[209] Rivas, M. A. *et al.* Human genomics. effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–9 (2015).

[210] Nedelec, Y. *et al.* Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 e21 (2016).

[211] Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–23 (2014).

[212] Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human t cell activation. *Science* **345**, 1254665 (2014).

[213] Shapira, S. D. *et al.* A physical and regulatory map of host-influenza interactions reveals pathways in h1n1 infection. *Cell* **139**, 1255–67 (2009).

[214] Schoggins, J. W. *et al.* A diverse range of gene products are effectors of the type i interferon antiviral response. *Nature* **472**, 481–5 (2011).

[215] Lalonde, E. *et al.* Rna sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res* **21**, 545–54 (2011).

[216] Kuriakose, T. *et al.* Zbp1/dai is an innate sensor of influenza virus triggering the nlrp3 inflammasome and programmed cell death pathways. *Sci Immunol* **1** (2016).

[217] Saveanu, L. *et al.* Concerted peptide trimming by human erap1 and erap2 aminopeptidase complexes in the endoplasmic reticulum. *Nat Immunol* **6**, 689–97 (2005).

[218] Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–24 (2012).

[219] Andres, A. M. *et al.* Balancing selection maintains a form of erap2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet* **6**, e1001157 (2010).

[220] Cagliani, R. *et al.* Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to hiv-1 infection. *Hum Mol Genet* **19**, 4705–14 (2010).

[221] Consortium, E. P. An integrated encyclopedia of dna elements in the human genome. *Nature* **489**, 57–74 (2012).

[222] Pertea, M. *et al.* Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat Biotechnol* **33**, 290–5 (2015).

[223] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nat Biotechnol* **34**, 525–7 (2016).

[224] Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of rna-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687–690 (2017).

[225] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

[226] Cribari-Neto, F. & Zeileis, A. Beta regression in r. *Journal of Statistical Software* **34**, 1–24 (2010).

[227] International HapMap, C. The international hapmap project. *Nature* **426**, 789–96 (2003).

[228] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–9 (2006).

[229] Purcell, S. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007).

[230] Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am J Hum Genet* **98**, 116–26 (2016).

[231] Shabalin, A. A. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics* **28**, 1353–8 (2012).

[232] Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–71 (1994).

[233] Bass, J., Dabney, A. & Robinson, D. qvalue: Q-value estimation for false discovery rate control. http://github.com/jdstorey/qvalue (2015).

[234] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440–5 (2003).

[235] Karolchik, D. *et al.* The ucsc table browser data retrieval tool. *Nucleic Acids Res* **32**, D493–6 (2004).

[236] Schmidt, E. M. *et al.* Gregor: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–6 (2015).

[237] Wunderlich, S. *et al.* Primate iPS cells as tools for evolutionary analyses. *Stem Cell Res.* **12**, 622–629 (2014).

[238] Mora-Bermúdez, F. *et al.* Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife Sciences* **5**, e18683 (2016).

[239] Otani, T., Marchetto, M. C., Gage, F. H., Simons, B. D. & Livesey, F. J. 2D and 3D stem cell models of primate cortical development identify Species-Specific differences in progenitor behavior contributing to brain size. *Cell Stem Cell* **18**, 467–480 (2016).

[240] Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).

[241] Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

[242] Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. https://www.biorxiv.org/content/early/2017/08/07/111880 (2017).

[243] de Boer, C., Sadeh, R., Friedman, N. & Regev, A. Deciphering cis-regulatory logic with 100 million random promoters. https://www.biorxiv.org/content/early/2017/11/25/224907 (2018).