

# Dissertation

submitted to the  
Combined Faculty of Natural Sciences and Mathematics  
of the Ruperto Carola University Heidelberg, Germany  
for the degree of

Doctor of Natural Sciences

presented by  
Zhongyi Wang, M.Sc.  
born in: Hunan, China

Oral examination: 23<sup>rd</sup> November, 2018



Ribosome profiling reveals principles of  
translatome and transcriptome evolution  
in mammalian organs

Referees: Prof. Dr. Henrik Kaessmann  
Prof. Dr. Georg Stoecklin



# Abstract

A primary goal in evolutionary biology is to understand the molecular basis responsible for phenotypic differences between species, most notably between humans and other species. Regulatory mutations affecting gene expression likely underlie most phenotypic changes. Recent evolutionary studies of mammalian transcriptomes have provided initial insights into mammalian gene expression evolution. However, mRNA levels are, in general, limited proxies for protein levels due to a sequence of regulations that succeed transcription. The fact that the evolution of mammalian translomes or proteomes is essentially unexplored has severely limited our understanding of gene expression evolution and its phenotypic implications.

To fill this gap and explore the co-evolution of regulatory processes across the transcriptome and translome layers of gene expression, we generated, in the framework of my thesis project, ribosome profiling (high-throughput sequencing of ribosome-protected fragments) and matched RNA sequencing data for three major mammalian organs (brain, liver, testis) from representatives of all major mammalian lineages (human, macaque, mouse, opossum, platypus) and a bird (chicken), which serves as an evolutionary outgroup.

My analyses identified strong and highly differential patterns of translational buffering among organs, gene classes and chromosomes. Specifically, to assess the extent to which transcriptional changes of individual genes are reflected at the level of protein synthesis, we devised a "translational tuning index" (TTI), and found that translational forces frequently counteracted but rarely boosted transcriptional changes. Expression changes of functionally cooperating genes tend to be balanced by concerted (modular) translational changes to preserve ancestral cellular stoichiometries. Contrary to individual gene compensation, this concerted buffering is more pronounced in brain and liver than in testis. By contrasting the evolutionary dynamics of transcriptomes and translomes, my analyses furthermore revealed that the widespread translational buffering more strongly preserved dosage-sensitive and, especially, housekeeping genes. I also found that translational upregulation acts to globally counterbalance the global dosage reduction that arose in the wake of mammalian sex chromosome differentiation; translational buffering thus represents a novel mechanism for X chromosome dosage compensation.

In summary, my PhD thesis work revealed that fine-tuned translational buffering substantially stabilized gene expression levels during mammalian evolution.



# Zusammenfassung

Ein primäres Ziel in der evolutionsbiologischen Forschung ist es, die molekularen Grundlagen phänotypischer Evolution zu verstehen, vor allem die zwischen Menschen und anderen Arten. Genregulatorische Mutationen, welche Genexpressionsveränderungen hervorrufen, liegen wahrscheinlich den meisten phänotypischen Veränderungen zu Grunde. Bisherige Untersuchungen des Transkriptom haben erste Einsichten in die Genexpressionsevolution von Säugetieren geliefert. Allerdings sind solche mRNA-Studien von begrenztem Wert für die Bestimmung von Proteinmengen, da diese nicht nur auf der Transkriptionsebene, sondern auch in nachfolgenden Expressionsschritten reguliert werden können. Die Tatsache, dass die Evolution von Translatomen und Proteomen bisher nahezu unerforscht geblieben ist, hat das Verständnis der Genexpressionsevolution bisher stark eingeschränkt.

Um diese Lücke zu füllen und die Evolution von genregulatorischen Prozessen sowohl auf der Transkriptom- als auch der Translationsebene zu untersuchen, haben wir, im Rahmen meiner Doktorarbeit, sogenannte "ribosome profiling"-Daten sowie entsprechende RNA-Sequenzierungsdaten für drei wichtige Organe (Gehirn, Leber, Hoden) repräsentativer Säugetiere (Mensch, Rhesusaffe, Maus, Opossum, Schnabeltier) und - für evolutionäre Vergleiche - einem Vogel (Huhn) generiert.

Meine Analysen haben starke und sehr differenzierte Muster translationaler Pufferung von Transkriptomveränderungen identifiziert. Insbesondere habe ich herausgefunden, dass Veränderungen auf der Translationsebene Transkriptomveränderungen häufig entgegengewirkt und selten verstärkt haben. Expressionsveränderungen funktional interagierender Gene, wurden in der Evolution häufig durch konzertierte Translationsänderungen kompensiert, um ursprüngliche Stöchiometrien zu erhalten. Im Gegensatz zur Kompensation einzelner Gene, ist die konzertierte Pufferung stärker im Gehirn als in Leber und im Hoden ausgeprägt. Weitere Analysen haben gezeigt, dass die weitverbreitete translationale Pufferung in der Evolution vor allem die Expression dosisabhängiger Gene und, vor allem, von Haushaltsgenen stark stabilisiert hat. Schließlich konnte ich zeigen, dass translationale Hochregulierungen auf dem X-Chromosom, Gendosisreduzierungen, die während der Evolution auf den Geschlechtschromosomen entstanden sind, entgegengewirkt haben.

Insgesamt habe ich in meiner Doktorarbeit also herausgefunden, dass feinabgestimmte translationale Pufferungsmechanismen Genexpressionsniveaus während der Säugetierevolution substanzial stabilisiert haben.





# Acknowledgements

Over the course of completing my PhD thesis, my time has been incredibly positive, and I am grateful to have been part of the open, collaborative and supportive group. Much of the work presented here would not have been possible, let alone enjoyable, without the guidance, collaboration, and support of many people.

First and foremost, I would like to thank my supervisor Dr. Henrik Kaessmann for giving me the opportunity to work in his lab and for his immense and invaluable contributions to my scientific growth throughout my PhD time. Our interactions have shaped the way I do science, and helped me to focus on crafting impactful analyses and questions.

I would also like to thank Dr. Georg Stoecklin for being my second supervisor, and Dr. David Gattfield for his encouraging words to my progress reports and for serving on my thesis committee, and Dr. Mirko Völkers for agreeing to serve on my thesis committee.

My collaboration with Evgeny Leushkin was a great working experience and I would like to thank him for the constant support and advice he provided me with for the past three years. I also wish to thank Angélica Liechti and Katharina Mößinger, Thoomke Brüning, and Coralie Rummel for their great work in generating ribosome profiling and RNA sequencing data.

I am thankful to Maria Warnefors and Margarida C. Moreira for their continuous help and the things they taught me, and for proofreading my dissertation attentively and providing many illuminating corrections and remarks, which helped me tremendously. Many thanks to Francesco Lamanna and Florent Murat for being great officemates and friends. Thanks to all other past and present members of the Kaessmann group for creating such a great working environment.

Last but not least, I want to thank my grandparents (爷爷奶奶) and parents (爸爸妈妈) for their love, constant aid and for always believing in me and the things I do. I am grateful for all the endless support and encouragement from my wife Nai-Jyuan (乃娟) throughout my PhD and also beyond.



# Contents

<b>Abstract</b> .....	<b>i</b>
<b>Zusammenfassung</b> .....	<b>iii</b>
<b>Acknowledgements</b> .....	<b>v</b>
<b>Contents</b> .....	<b>vii</b>
<b>List of Figures</b> .....	<b>xi</b>
<b>List of Tables</b> .....	<b>xiii</b>
<b>Chapter 1 General introduction</b> .....	<b>1</b>
1.1 The molecular basis of phenotypic evolution .....	1
1.2 RNA sequencing enables comparative transcriptomics.....	3
1.3 mRNA levels are poor predictors of protein levels .....	5
1.4 Limitations in comparative transcriptomics and proteomics .....	6
1.5 Ribosome profiling provides a powerful solution .....	7
1.6 Objectives of this thesis .....	10
<b>Chapter 2 Generation and quality assessment of mammalian translome data</b> .....	<b>11</b>
2.1 Biological samples .....	11
2.2 Ribo-seq protocol optimization and data production.....	12
2.2.1 <i>Implementation of an additional rRNA depletion step</i> .....	12
2.2.2 <i>Ribo-seq and matched RNA-seq data production</i> .....	14
2.3 Gene annotation preparation .....	17
2.3.1 <i>Genome and transcript isoform annotation</i> .....	17
2.3.2 <i>Selection of the dominant splice isoform</i> .....	18
2.3.3 <i>Extraction of orthologous gene sets</i> .....	19
2.3.4 <i>Perfectly aligned coding sequences across species</i> .....	20
2.3.5 <i>Compiling structural RNA sequences for each species</i> .....	21
2.4 Data processing and quality assessment .....	21
2.4.1 <i>Read mapping and processing</i> .....	22

2.4.2 A-site calibration.....	23
2.4.3 Triplet periodicity analysis.....	24
2.4.4 Assessment of reproducibility for both data types.....	26
2.4.5 Expression levels and normalization .....	28
2.4.6 Principal component analysis .....	29
<b>Chapter 3 Widespread translational buffering of individual genes across organs .....</b>	<b>33</b>
3.1 Introduction.....	33
3.2 Methods .....	34
3.2.1 Translational tuning index.....	34
3.2.2 Identification of translational efficiency changes .....	36
3.2.3 Enrichment analysis.....	37
3.3 Results .....	37
<b>Chapter 4 Global patterns of gene expression conservation .....</b>	<b>43</b>
4.1 Introduction.....	43
4.2 Methods .....	44
4.2.1 Gene expression phylogenies for each organ .....	44
4.2.2 Total branch length analysis.....	45
4.2.3 Estimating modularity in gene expression changes .....	46
4.3 Results .....	47
4.3.1 Gene expression phylogenies.....	47
4.3.2 Global patterns of gene expression conservation .....	49
4.3.3 Modular compensation.....	51
<b>Chapter 5 Patterns of expression divergence and compensatory evolution across gene classes .....</b>	<b>55</b>
5.1 Introduction.....	55
5.2 Methods .....	58
5.2.1 Resources for gene class annotations.....	58
5.2.2 Tissue specificity index .....	59
5.3 Results .....	60
5.3.1 Dosage sensitivity and compensatory evolution.....	60
5.3.2 Gene essentiality and compensatory evolution .....	61
5.3.3 Spatial expression characteristics and compensatory evolution .....	63
5.3.4 Gene age and compensatory evolution.....	64
<b>Chapter 6 X chromosome dosage compensation through translational upregulation.....</b>	<b>67</b>

6.1 Introduction .....	67
6.2 Methods .....	71
6.2.1 Extraction of orthologous gene sets .....	71
6.2.2 Normalization of current X and proto-X expression levels .....	71
6.3 Results .....	72
<b>Chapter 7 Discussion and Outlook.....</b>	<b>77</b>
7.1 Discussion.....	77
7.2 Outlook.....	81
7.3 Concluding remarks.....	83
<b>Supplementary Figures.....</b>	<b>85</b>
<b>References.....</b>	<b>101</b>



# List of Figures

Figure 1.1: Two major types of mutations underlying phenotypic differences.....	2
Figure 1.2: Mechanisms involved in the regulation of gene expression.....	5
Figure 1.3: A simplified experimental overview of ribosome profiling.....	8
Figure 2.1: Study of three major organs across five mammals and a bird.....	12
Figure 2.2: Comparison of Ribo-seq libraries prepared with original and optimized protocols.....	13
Figure 2.3: Schematic representation of the algorithm used to select the dominant splice isoform	19
Figure 2.4: Illustration of a perfectly aligned region across species .....	20
Figure 2.5: Overview of the translome and matched transcriptome data .....	21
Figure 2.6: Different read length distributions for Ribo-seq and RNA-seq libraries .....	23
Figure 2.7: Calibration of the A-site of ribosome footprints.....	24
Figure 2.8: Distributions of reads on transcript features and on the three reading frames.....	25
Figure 2.9: Metagene profiles of ribosomal footprints.....	26
Figure 2.10: Correlations between technical replicates for Ribo-seq and RNA-seq data .....	27
Figure 2.11: Correlations between biological replicates for Ribo-seq and RNA-seq data .....	28
Figure 2.12: PCA of two gene expression layers across different organs and species.....	30
Figure 2.13: Correlations of two gene expression layers across different organs and species .....	31
Figure 3.1: Illustration of the translational tuning index (TTI) .....	36
Figure 3.2: Translational versus transcriptional changes for individual genes .....	39
Figure 3.3: Distributions of log <sub>2</sub> -fold changes for the two gene expression layers.....	40
Figure 3.4: TTI distribution for genes of significant TE changes between mouse and chicken .....	41
Figure 4.1: Mammalian gene expression (translatome and transcriptome) phylogenies .....	48
Figure 4.2: Mammalian gene expression phylogeny for brain.....	49
Figure 4.3: Total tree length analyses for translomes and transcriptomes.....	50
Figure 4.4: Expression correlation between macaque and the other four species .....	50
Figure 4.5: Comparison of correlations of expression levels between actual and simulated data ...	52
Figure 4.6 Total tree length analysis for the genes with near-zero TTI.....	53

Figure 5.1: Compensatory evolution of dosage-sensitive and -insensitive genes.....	61
Figure 5.2: Compensatory evolution of essential and nonessential genes .....	62
Figure 5.3: Compensatory evolution of housekeeping and tissue-specific genes.....	64
Figure 5.4: Compensatory evolution of old and recent genes .....	65
Figure 6.1: The evolution of therian sex chromosomes and dosage compensation.....	68
Figure 6.2: Illustration of translational efficiency (TE) .....	71
Figure 6.3: Current and ancestral levels of translation and transcription on the X chromosome .....	75
Figure 6.4: TE ratios of X-linked to autosomal genes across therians.....	75



# List of Tables

Table 1: In-house biotinylated rRNA depletion oligonucleotides for each species ..... 14

Table 2: Tissue-specific genomic annotations for each species..... 18

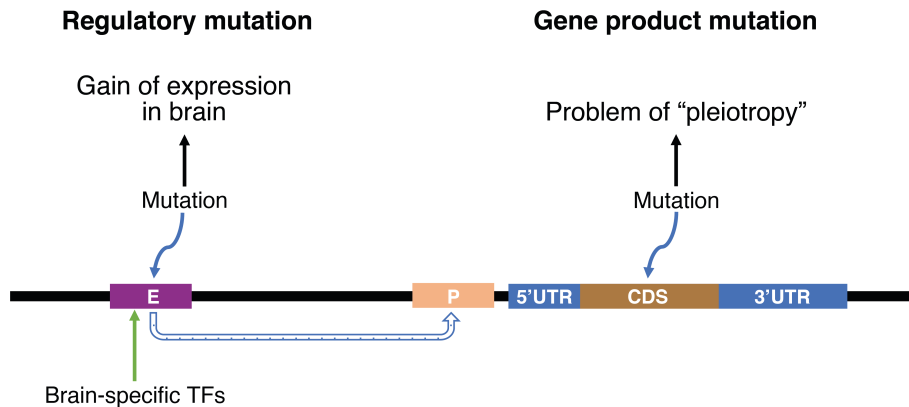


# Chapter 1 General introduction

## 1.1 The molecular basis of phenotypic evolution

A central objective in evolutionary biology is to understand the molecular basis responsible for phenotypic differences between species, and of particular interest are changes underlying distinct mammalian traits, most notably those between humans and other species. For example, mammals have evolved shared traits that include lactation, hair and relatively large brains with unique structures, but also distinct lineage-specific anatomical, physiological and behavioral characteristics relating to differences in reproduction, life span, cognitive abilities and disease susceptibility (Kemp, 2005).

Phenotypic differences between species arise during evolution due to two major types of mutations (Figure 1.1) (Khaitovich et al., 2006). The first class comprises mutations that change the DNA sequence (e.g., substitutions, insertions or deletions) and, as a consequence, the function of the final gene product (i.e., the encoded protein or RNA) (Figure 1.1). For example, the Forkhead box protein P2 that is encoded by *FOXP2* is highly conserved in primates, but contains two non-synonymous substitutions that have been fixed by positive selection (Enard et al., 2002a). It was suggested that both mutations are responsible for the normal development of speech and language that is unique to humans; this conclusion, however, was questioned by a recent study claiming that the finding was based on skewed population sampling (Atkinson et al., 2018). Instead, the authors of the latter study identified an intron region in *FOXP2* that potentially functions as an enhancer associated with human language abilities and which contains several mutations that are only shared in humans but variable between populations (Atkinson et al., 2018).



**Figure 1.1: Two major types of mutations underlying phenotypic differences**

To the left: regulatory changes, for example in enhancer sequence, are more likely responsible for the gain and loss of tissue-specific traits. To the right: mutations in the coding region typically cause pleiotropic effects. E, enhancer; P, promoter; 5' UTR, five prime untranslated region; CDS, coding sequence; 3' UTR, three prime untranslated region; TFs, transcription factors.

The second class comprises regulatory mutations (e.g., in enhancer sequences) that possibly affect transcription, post-transcriptional regulation, translation, or protein degradation (Figure 1.1). Notably, certain gene product sequence alterations that change the function of the protein (e.g., mutations in transcription factors) may also have consequences for gene regulation. By modifying developmental programs, both types of mutations may lead to distinct tissue morphologies, laying the foundation for species- or lineage-specific physiology and behavior. It was postulated almost a half century ago that regulatory mutations affecting gene expression underlie many, or even most, phenotypic differences between closely related species (e.g., human and chimpanzee) (Britten and Davidson, 1969 & 1971; King and Wilson, 1975). This is because these mutations allow for tissue-specific adaptations, whereas changes in the protein or RNA sequence may be more likely to have deleterious pleiotropic consequences, given that they typically affect all tissues in which a gene is expressed (Figure 1.1) (Wray, 2007; Somel et al., 2013; Necsulea and Kaessmann, 2014a). For example, it was recently found that increased expression of the *FZD8* gene causes a faster cell cycle in neural progenitors, which is due to the human-accelerated regulatory enhancer *HARE5* (Boyd et al., 2015). Another study revealed that the loss of limbs in snakes is associated with sequence changes disrupting the function of a limb enhancer of *Sonic*

*hedgehog* (*Shh*) (Kvon et al., 2016). Moreover, adaptive evolution of pelvic reduction in sticklebacks is caused by changes in gene expression that result from the recurrent deletion of a tissue-specific enhancer of the pituitary homeobox transcription factor 1 (*Pitx1*) gene (Shapiro et al., 2004; Chan et al., 2010; Jones et al., 2012).

## 1.2 RNA sequencing enables comparative transcriptomics

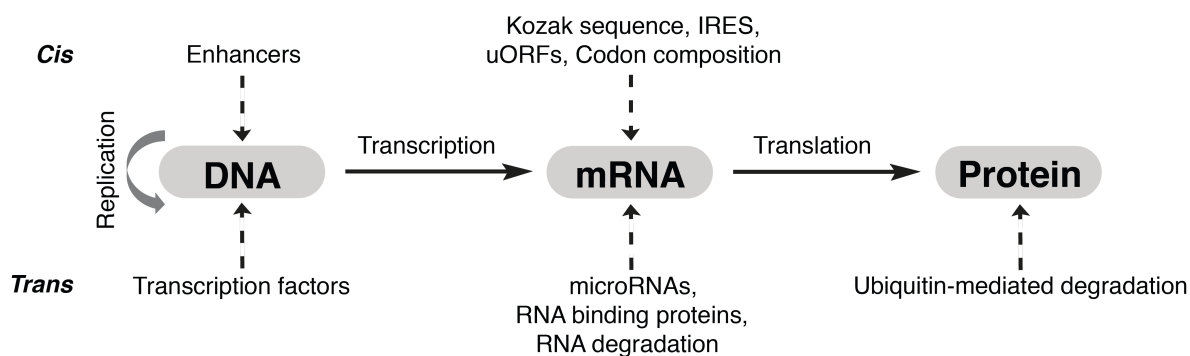
Direct comparisons of gene expression patterns between species can reveal fundamental principles of gene expression evolution and have already received much attention. These efforts have focused on comparisons of transcripts, which represent the first (in the case of protein-coding genes) or final (in the case of noncoding RNA genes) products of genes. Comparisons of mammalian transcriptomes were initially restricted to closely related primates or mice (Enard et al., 2002b; Schadt et al., 2003; Khaitovich et al., 2006), because of the limitations of microarrays, the most suitable technology available at the time. Microarrays require hybridization to species-specific probes, which requires prior gene annotations and especially depends on their sequence (different probe sets for one gene will give different results), thus making interspecies comparisons of individual transcript abundance difficult. Nevertheless, these studies provided initial insights into expression changes relevant for human-specific phenotypes, as well as initial evidence for general principles that govern the evolution of gene expression (Khaitovich et al., 2006; Necsulea and Kaessmann, 2014a).

The advent of high-throughput RNA sequencing (RNA-seq) protocols a decade ago opened the door to unprecedented genome-wide and cross-species transcriptome comparisons by allowing for accurate, sensitive and essentially unbiased assessments of transcript sequences and their expression levels (Mortazavi et al., 2008; Marioni et al., 2008; Wang et al., 2009). The power and utility of RNA-seq for comparative transcriptome investigations was originally demonstrated in studies of humans and a few closely related primates, elucidating patterns of transcript abundance and alternative splicing for these species (Romero et al., 2012). Subsequently, the first

cross-mammalian set of transcriptome data for a range of major organs was established using RNA-seq (Brawand et al., 2011). This study revealed that gene expression trees recapitulate the known phylogeny of the associated species, and that in stark contrast to the testis, RNA levels are highly conserved in nervous tissues during mammalian evolution (Brawand et al., 2011). Mammalian transcriptomes are more similar in homologous tissues from different species than they are from different tissues of the same species (Brawand et al., 2011). Analyses of this dataset further revealed global patterns of protein-coding gene expression change (e.g., rates of expression evolution across mammalian lineages, and chromosomes), general principles that govern the evolution of gene expression (e.g., the dominant role of purifying selection and genetic drift), and selectively driven expression shifts that likely contributed to the specific organ biology of various mammals (e.g., that of human/primate brains) (Brawand et al., 2011). Based on these and complementary data, additional studies explored the evolution of alternative splicing (Barbosa-Morais et al., 2012; Merkin et al., 2012), sex chromosome dosage compensation (Julien et al., 2012; Marin et al., 2017), the relationship of protein sequence divergence and expression divergence (Warnefors et al., 2013), spermatogenic transcriptomes (Soumillon et al., 2013), the role of intron retention in evolutionary adaptation (Braunschweig et al., 2014; Schmitz et al., 2017), the functional evolution of the Y chromosome (Cortez et al., 2014), and the evolution and expression patterns of new genes, such as retrocopies (Carelli et al., 2016) and duplicate genes (Guschanski et al., 2017). Other studies focused on the non-coding portion of the transcriptome, thus illuminating the birth, functionality and evolution of long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) and also the role of miRNAs in dosage compensation (Necsulea et al., 2014b; Washietl et al., 2014; Meunier et al., 2013; Warnefors et al., 2014; Warnefors et al., 2017). Overall, these comparative RNA-seq-based studies have provided many novel insights into the dynamics of evolutionary gene expression changes and associated phenotypic implications in mammals and tetrapods at large (Necsulea and Kaessmann, 2014a).

### 1.3 mRNA levels are poor predictors of protein levels

While mRNA abundances are widely used as a proxy for protein levels, protein-coding gene expression may frequently be regulated on layers that succeed transcription (Vogel et al., 2012; Hershey et al., 2012; McManus et al., 2015; Liu et al., 2016); these include, but are not limited to, post-transcriptional and translational regulations, and protein degradation (Figure 1.2). Consequently, protein abundances may or may not occur in proportion to their relative mRNA levels. Given that it is ultimately protein abundance that matters, transcriptome studies have likely provided an incomplete picture of protein-coding gene expression evolution.



**Figure 1.2: Mechanisms involved in the regulation of gene expression**

According to the central dogma of molecular biology, protein-coding genes are first transcribed into mRNAs, which are then translated into proteins under multilevel and multifactorial governance of regulatory processes. IRES, internal ribosome entry sites; uORFs, upstream open reading frames.

Two methods have been used to assess mRNA-protein correlations. Firstly, one can explore to what extent mRNA level variation propagates to the protein level across different individuals, tissues, conditions or time points. Secondly, one can correlate protein levels with their respective mRNA levels for all or a particular set of genes. Using direct parallel genome-scale measurements of mRNA and protein levels in unperturbed mammalian cells, Schwanhauser *et al.* (2011) showed that while 40% of the variance in protein levels is explained by mRNA levels, translation rate differences account for a large fraction of the remaining variance, with only a small impact of protein degradation variability. Another study using the human Daoy medulloblastoma cell

line revealed that variation in mRNA abundance alone only explains 25–30% of the variation in protein abundance, but combining sequence signatures and mRNA concentration increases it to ~67% (Vogel et al., 2010). In addition, genome-wide correlations between mRNA and protein levels are rather moderate in human cell lines, with Spearman correlation coefficients (Spearman's  $\rho$ ) in the order of 0.4 (Lundberg et al., 2010). Wilhelm *et al.* (2014) claimed that combining both estimated gene-specific translation rates (the ratio of protein to mRNA levels) and mRNA levels can accurately predict the corresponding protein levels in human tissues, reporting high Spearman's  $\rho$  (~0.9) between predictions and measurements across genes. However, this result was challenged by a reanalysis of the data, in which the authors used standard statistical evaluation methods to show that the gene-specific translation rates estimated by Wilhelm et al. (2014) together with RNA levels are insufficient to reliably predict protein levels (median Spearman's  $\rho$  at 0.21) (Fortelny et al., 2017).

#### **1.4 Limitations in comparative transcriptomics and proteomics**

One possible explanation for the poor correlations between mRNA and protein levels, is that evolutionary shifts in mRNA expression due to transcriptional regulatory mutations may be, for example, offset by post-transcriptional mutations that reconstitute (optimal) protein levels. Indeed, initial pioneering work showed that, contrary to mRNA levels, protein abundances have been remarkably preserved over long evolutionary time periods (Spearman's  $\rho = 0.7$  at a divergence of 1 million years), at least for the highly conserved one-to-one (1:1) orthologs investigated (Schrimpf et al., 2009; Weiss et al., 2010; Laurent et al., 2010). Consistent with the notion of compensatory evolution across the different gene expression layers, a recent study, which compared mRNA and protein expression divergence across human, chimpanzee and macaque lymphoblastoid cell lines at a genome-wide scale using state-of-the-art technologies (i.e., quantitative mass spectrometry (MS) for the proteome, RNA-seq for the transcriptome), revealed a sizeable number of genes with significant expression differences between species at the mRNA level



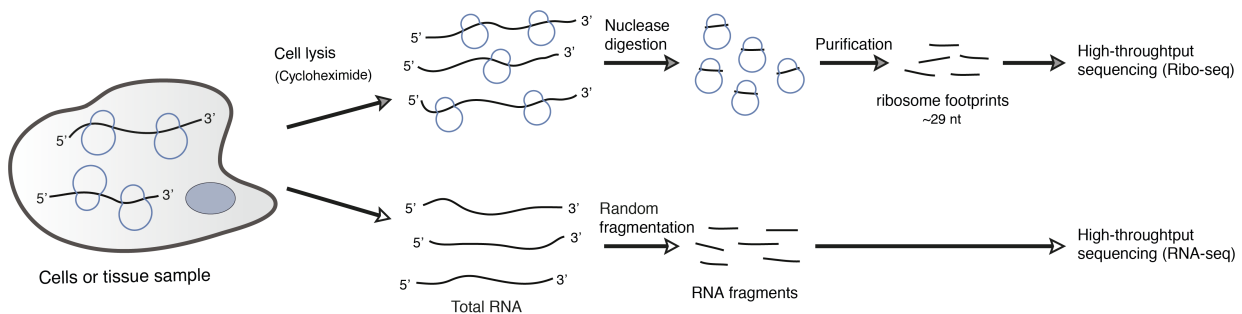
yet little or no difference in protein expression (Khan et al., 2013). They concluded that selective constraints on protein abundances are stronger than those on mRNA levels. It is likely that many interspecies mRNA expression changes in primary tissues are also not propagated to their corresponding protein levels and thus do not alter cellular physiology. Moreover, based on ~1,300 proteins a recent study showed consistent low correlations between mRNA and protein expression levels in two brain regions of human and chimpanzee (Bauernfeind et al., 2015). Altogether, these observations suggest that quantitative genome-wide and cross-species assessments closer to the proteome level are crucial for a better understanding of gene expression change and associated phenotypic evolution. It could thus be considered ideal to directly assess protein abundance across tissues and species.

However, although over the past decade MS technologies have evolved towards higher data quality, they are still limited in their ability to independently determine protein sequence and to match the depth and breadth of coverage that is routinely possible in nucleic acid sequencing experiments, and are overall cumbersome and time consuming (Brar and Weissman, 2015; Liu et al., 2016). Furthermore, cross-species comparisons of protein abundances for primary tissues remain difficult due to data normalization issues. Thus, detailed qualitative and quantitative analyses of mammalian proteomes, which cover most or all genes (highly and lowly expressed), are not readily applicable at a global scale.

## **1.5 Ribosome profiling provides a powerful solution**

The ribosome profiling (or Ribo-seq) technique (Ingolia et al., 2009) provides a powerful solution to this dilemma (Figure 1.3). This highly sensitive and accurate method, which approximates the rate of protein synthesis rather directly, is based on deep sequencing of ribosome-protected mRNA fragments (“ribosome footprints”) and enables genome-wide qualitative and quantitative investigations of translation at single-nucleotide resolution (Brar and Weissman, 2015; Ingolia et al., 2018) (Figure 1.3). It is more robust and reproducible compared with its complementary

version – polysome profiling, in which the transcripts are separated on the basis of the number of bound ribosomes by ultracentrifugation (Arava et al., 2003; Chassé et al., 2017).



**Figure 1.3: A simplified experimental overview of ribosome profiling**

Cells or tissue samples are lysed in the presence of cycloheximide, which globally arrests translating ribosomes on the mRNA. The cytosolic extract, which contains ribosome-bound mRNAs, is subjected to controlled nuclease (typically RNase I) digestion. Ribosome-protected mRNA fragments (ribosome footprints, ~29 nt) are then purified by polyacrylamide gel electrophoresis and subsequently converted to a sequencing library. In parallel, a matched RNA sequencing library is prepared based on the same lysate using the same protocol as for the footprints. Finally, both libraries are sequenced on a high-throughput sequencing platform (e.g., Illumina HiSeq 2500).

Ribosome occupancy (as measured by Ribo-seq) is a much better predictor of protein abundance (as measured by MS) than measurements of mRNA levels (as measured by RNA-seq) (Liu et al., 2017; Cheng et al., 2018), although, given that this technology measures the rates of protein synthesis but not of protein degradation, it does not allow for direct inferences of actual steady-state protein levels. In addition to the quantification of the translated portion of the transcriptome, Ribo-seq allows for a qualitative assessment of the translated portion of the transcriptome, including the rigorous evaluation of genes previously annotated as long noncoding RNAs (lncRNAs) (Ingolia et al., 2011; Guttman et al., 2013; Ingolia et al., 2014b; Chekulaeva and Rajewsky, 2018; Zeng et al., 2018). Moreover, comparisons between inferred rates of protein synthesis and the abundance of mRNAs afford the assessment of the translational efficiency (TE) for each mRNA (i.e., the rate of translation per mRNA molecule), which has the potential to reveal mechanisms underlying the translational control of gene expression (e.g., miRNAs, upstream open reading frames (uORFs)

and RNA modifications such as adenosine N<sup>6</sup> methylation (m<sup>6</sup>A)) (Guo et al., 2010; Bazzini et al., 2012; Bazzini et al., 2014; Johnstone et al., 2016; Janich et al., 2015; Wang et al., 2015a; Slobodin et al., 2017; Peer et al., 2018).

Since the degree to which genetic variants affect the translation or protein levels of their target genes has long been an open question, Ribo-seq has been applied to study the genetic variants affecting RNA, translation or protein levels in primate lymphoblastoid cell lines (Battle et al., 2015; Cenik et al., 2015). Specifically, Battle *et al.* (2015) showed that expression quantitative trait loci (eQTLs) likely have significantly reduced effect sizes on protein levels (protein-level variation), suggesting that the effects of many eQTLs on RNA levels are subsequently attenuated or buffered. Another study revealed that many RNA expression changes were offset at the level of protein synthesis through tuning TEs triggered by genetic variants (Cenik et al., 2015).

Moreover, the power and utility of Ribo-seq for comparative gene expression analyses (translatomes versus transcriptomes) has been demonstrated in studies of yeast (McManus et al., 2014; Muzzey et al., 2014; Artieri and Fraser, 2014; Albert et al., 2014; Wang et al., 2015b), nematodes (Stadler et al., 2013), hybrid mouse cells (Hou et al., 2015) and primate cell lines (Wang et al., 2018), providing initial insights into patterns of transcriptome versus translome evolution. Whereas some studies concluded that translational buffering frequently counteracts but rarely reinforces mRNA expression changes (McManus et al., 2014; Artieri and Fraser, 2014; Wang et al., 2015b; Hou et al., 2015), others reported the conflicting finding that translation mostly reinforces mRNA changes (Muzzey et al., 2014; Albert et al., 2014).

Altogether, while transcriptome studies in mammals have begun to emerge, the evolutionary comparison of mammalian translomes represent, as yet, essentially uncharted territory. Ribo-seq enables large-scale investigations of translome evolution and thus has the potential to provide fundamental novel insights into the contribution of gene expression change to mammalian phenotypic evolution.

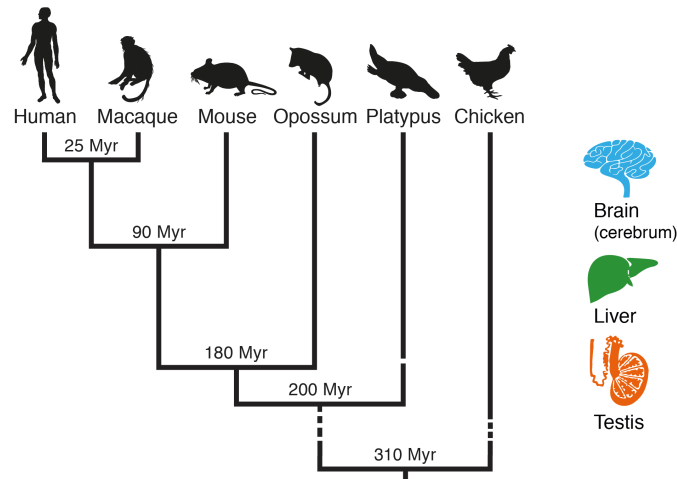
## 1.6 Objectives of this thesis

To fill this gap and explore the co-evolution of regulatory processes across the transcriptome and translome layers of genes expression, Ribo-seq and matched RNA-seq data were generated for three major mammalian organs (brain, liver, testis) from five representatives of the three main mammalian lineages: placental mammals (human, rhesus macaque, mouse); marsupials (grey short-tailed opossum); and egg-laying monotremes (platypus). Corresponding data were generated for a bird (red junglefowl, the progenitor of domestic chicken; henceforth referred to as “chicken”), to be used as an evolutionary outgroup. These unprecedented data allow unique integrated analyses of mammalian gene expression evolution in general and translomes in particular with the following objectives: (i) to assess the extent of transcript abundance changes buffered or reinforced at the level of protein synthesis; (ii) to characterize the global patterns of gene expression evolution and the associated selective forces by contrasting the evolution of translomes with that of transcriptomes; (iii) to investigate the patterns of expression divergence and compensatory evolution across gene classes (e.g., dosage-sensitive and essential genes); (iv) to assess whether X-linked genes are globally upregulated at the translational level following sex chromosome differentiation (Y degeneration) from ancestral autosomes.

# Chapter 2 Generation and quality assessment of mammalian translome and matched transcriptome data

## 2.1 Biological samples

In the framework of my thesis project, we generated Ribo-seq and matched RNA-seq data for the following samples: brain (cerebrum), liver, and testis samples from human (*Homo sapiens*), rhesus macaque (*Macaca mulatta*), mouse (*Mus musculus*, strain: CD-1, RjOrl:SWISS), grey short-tailed opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), and chicken (red junglefowl, *Gallus gallus*) (Figure 2.1). This work complies with all relevant ethical regulations with respect to both human samples and samples for the other mammals. Human samples were obtained from official scientific tissue banks or dedicated companies; informed consent was obtained by these sources from donors prior to death or from next-of-kin. The use of all human samples for the type of work described in this study was approved by an Ethics Screening panel from the European Research Council (ERC) (associated with ERC Consolidator Grant 615253, On-toTransEvol) and local ethics committees; that is, from the Cantonal Ethics Commission Lausanne (authorization 504/12) and Ethics Commission from the Medical Faculty of Heidelberg University (authorization S-220/2017). The use of all other mammalian samples for the type of work in this study was approved by ERC Ethics Screening panels (ERC Starting Grant 242597, SexGenTransEvolution, and ERC Consolidator Grant 615253, On-toTransEvol).



**Figure 2.1: Study of three major organs across five mammals and a bird**

Three major organs from six species were targeted in the framework of my thesis project. A schematic phylogeny of the species and lineages with approximate lineage split times is shown. Myr, million years.

## 2.2 Ribo-seq protocol optimization and data production

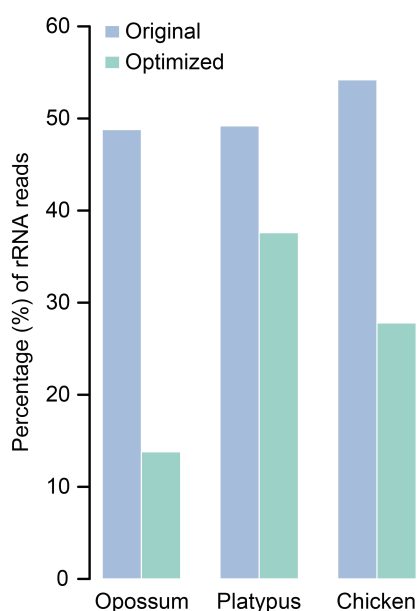
### 2.2.1 Implementation of an additional rRNA depletion step

The Ribo-seq technique has mainly been applied to cells or cell lines (Ingolia et al., 2009; Ingolia et al., 2011; Brar and Weissman, 2015), and only recently was successfully adapted to different mouse tissues in the Gatfield lab (Janich et al., 2015; Castelo-Szekely et al., 2017). This provided initial evidence that their Ribo-seq protocol could be applied to other species, at least for the homologous tissues. In our lab, we possess a large and unique collection of tissue samples, some of which have been stored at  $-80^{\circ}\text{C}$  for several years. Pilot experiments demonstrated that the adapted Ribo-seq technique is applicable to frozen samples across different species<sup>1</sup>.

While the analyses described above testify to the feasibility of applying Ribo-seq to frozen samples across species, the data filtering steps revealed varying degrees of contamination by

<sup>1</sup>Dr. Peggy Janich and Dr. David Gatfield helped to establish the original ribosome profiling method for solid tissues. Dr. Alaaddin Bulak Arpat and Dr. David Gatfield provided guidance during the data production and initial analysis phase.

ribosomal RNAs (rRNAs), a typical and expected challenge in Ribo-seq experiments that results from the nuclease treatment step (mainly the undigested rRNAs from protecting ribosomes) and may substantially reduce the number of usable reads for biological analyses (Ingolia et al., 2009; Brar and Weissman, 2015). Since the rRNA depletion step implemented in the commercial TruSeq Ribo Profile (Mammalian) Library Prep Kit (Illumina) (formerly ARTseq) was based on rRNA sequence information from human, mouse and rat, rRNA contamination is likely to be more pronounced for other distantly related species, presumably due to diverged rRNA sequences. Thus, to ensure a more efficient removal of rRNAs for all species and thus to avoid unnecessary (and financially prohibitive) sequencing efforts to compensate for high levels of rRNA contamination, an additional rRNA depletion step was implemented. Specifically, liver samples from opossum, platypus and chicken were used to optimize the Ribo-seq protocol. Macaque was not included because rRNA genes have only slightly diverged from that of humans. Unsurprisingly, using the original protocol, rRNA reads accounted for ~50% of raw reads for all pilot libraries (Figure 2.2), which is higher than that of mouse libraries (< 40%).



**Figure 2.2: Comparison of Ribo-seq libraries prepared with original and optimized protocols**

The extra step of adding in-house rRNA-depletion oligonucleotides during library preparation substantially reduces the percentage of reads mapping to rRNA compared to using the original commercial TruSeq Ribo Profile (Mammalian) Library Prep kit (Illumina).

I found that the majority of the rRNA reads originated from a small pool of sequences in each species. I then added up the fragments that derived from the same genomic region, meaning they only differ (longer/shorter) at the 5' and/or 3' end. I next pinpointed the rRNA fragments that individually accounted for more than 10% of the total raw reads. This information was then used to design the subtractive hybridization oligonucleotides (Table 1), which were added towards the end of the Ribo-seq protocol, after the step of cDNA circularization (Ingolia et al., 2012). In stark contrast to the data generated with the original protocol, analyses of the test libraries prepared using the optimized protocol based on the same lysates revealed a considerable reduction of rRNA contamination from ~50% to 14-38% (Figure 2.2), which is also much lower than that observed in some other cell or cell line studies (~80%, Ingolia et al., 2009; Battle et al., 2015).

**Table 1: In-house biotinylated rRNA depletion oligonucleotides for each species**

Species	Sequence (5' to 3')
Opossum	CCTGCCGAGGGCGCACCACCGGCCCGTCTCGC
	CCCCGGGGATGCGTGCATTTATCAGA
	AGCCCGTGGACGGTGTGAGGCCGGTAGCG
Platypus	GGTGGTGCGCCGCGACCGGCTCTGGGACGGCTGGGAAG
	GTCGCCTGGATACTCCAGCTAGGAATGATGGAAT
	AGCCCGTGGACGGTGTGAGGCCGGTAGCGGCCCCCG
	CTCCGGGGCTACGCCTGTCTGAGCGTCGCTT
	GCCGTGATCGTATAGTGGTTAGTACTCTGCG
Chicken	GCCGCCGAATACTCCAGCTAGGAATAATGGAATA
	ATCGTCGCCGAATCCCCGGGGCCGAGGGAGAGGAC
	AAGGCCCGGGCGCACCACCGGCCCGTCTCGC
	CTCCGGGGCTACGCCTGCCTGAGCGTCGCTT

### 2.2.2 Ribo-seq and matched RNA-seq data production

The translomes were generated based on the Ribo-seq method established by Ingolia et al. (2012), which has been implemented in the TruSeq Ribo Profile kit (Illumina) (formerly ARTseq) and allows for an additional rRNA depletion step (see Section 2.2.1 for more details)<sup>1</sup>.

<sup>1</sup>Angélica Liechti, Dr. Katharina Mößinger, Thoomke Brüning, and Coralie Rummel generated all of the Ribo-seq and matched RNA-seq data used for my thesis.



Specifically, frozen tissues were treated in 3 volumes of ice-cold lysis buffer (150 mM NaCl, 20 mM Tris-HCl pH7.4, 5 mM MgCl<sub>2</sub>, 5 mM DTT, 100 µg/ml cycloheximide, 1% Triton X-100, 0.5% Sodium deoxycholate, complete EDTA-free protease inhibitors (Roche) and 40 U/ml RNasin plus (Promega)) using a Teflon homogenizer. Lysates were incubated for 10 min on ice and cleared by centrifugation at 3,000 x *g*, 4°C for 3 min. Supernatants were flash-frozen and stored in liquid nitrogen. For absorbance measurements, lysates were gently thawed on ice and the OD<sub>260</sub> was determined using a Nanodrop spectrophotometer (Thermo Fisher Scientific). From the lysate pool, 15 OD<sub>260</sub> were incubated with 650 U RNase I (Ambion) and 5 U Turbo DNase (Ambion) for 45 min at room temperature and gentle agitation. Nuclease digestion was stopped through addition of 8.7 µl SUPERase In RNase Inhibitor (Ambion). Subsequently, lysates were applied to Sephacryl MicroSpin S-400 HR columns (GE Healthcare Life Sciences), pre-washed 3 times with 700 µl polysome buffer (150 mM NaCl, 20 mM Tris-HCl pH7.4, 5 mM MgCl<sub>2</sub>, 5 mM DTT, 100 µg/ml cycloheximide, complete EDTA-free protease inhibitors (Roche)) for 1 min at 600 x *g*, and centrifuged for 2 min at 600 x *g* and 4°C. The flow-through was immediately mixed with 1 ml Qiazol (Qiagen) and ribosome-protected mRNA fragments were purified using the miRNeasy Micro kit (Qiagen) according to the manufacturer's instructions and the concentration of the RNA was determined by Nanodrop.

Prior to library preparation, for each sample a total of 5 µg RNA was subjected to rRNA depletion (Ribo-Zero rRNA Removal kit, Illumina) and subsequently purified using the RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol. The rRNA depleted RNA was separated on a denaturing 15% Urea polyacrylamide gel (Thermo Fisher Scientific) and stained with SYBR-Gold (Thermo Fisher Scientific). Gel slices between 26-34 nt were excised and the RNA was extracted using 450 µl gel extraction buffer (0.5 M Ammonium acetate and 0.05% SDS) for 2 hours at room temperature and gentle agitation. Gel pieces were removed by centrifugation over Spin-X filter tubes (Corning) for 2 min at 15,000 x *g*. RNA was precipitated over night at -20°C in the presence of 1 ml 100% ethanol and 3 µl glycogen. RNA was pelleted for 25 min

and washed with 80% ethanol in a tabletop centrifuge at maximum speed and 4°C. Sequencing libraries were generated using the TruSeq Ribo Profile Library Prep Kit (Illumina). End-repair, 3' adapter, reverse transcription, cDNA purification, and circularization were done according to the manufacturer's instructions.

For opossum, platypus and chicken samples, an additional rRNA depletion step was implemented: first strand cDNAs derived from species specific rRNA contaminants were further depleted after the step of cDNA circularization by hybridization to 5'-biotinylated sense strand oligonucleotides followed by removal of the duplexes through streptavidin affinity as described in Section 2.2.1. PCR amplification of the circularized cDNA product was done using the TruSeq Ribo Profile Library Prep Kit (Illumina) according to the manufacturer's instructions. The final library of 150-200 bp was gel-purified on a 10% polyacrylamide non-denaturing gel (Thermo Fisher Scientific), excised and recovered with 330 µl gel extraction buffer for 1 hour at 37°C and gentle agitation. Gel pieces were removed by centrifugation over Spin-X filter tubes (Corning) for 2 min at 15,000 x *g*. Libraries were precipitated at -20°C for 1 hour in the presence of 525 µl 100% isopropanol and 2 µl glycogen, pelleted for 25 min at 4°C and 15,000 x *g*, washed with 80% ethanol and resuspended in water. Libraries were sequenced on Illumina HiSeq 2500 or (for tests) Illumina MiSeq machines (read lengths: 50 or 100 nucleotides, nt).

In parallel to Ribo-seq library preparation, matched RNA-seq libraries were prepared from the same lysates using TruSeq Ribo Profile Library Prep Kit (Illumina). rRNA was depleted from 5 µg of total RNA with the Ribo-Zero rRNA Removal kit (Illumina) according to the manufacturer's instructions. RNAs were randomly fragmented and converted to a complementary DNA library with TruSeq Ribo Profile Library Prep Kit (Illumina). The concentration and the quality of both the Ribo-seq and RNA-seq libraries were determined using Qubit (Thermo Fisher Scientific) and Fragment Analyzer (Advanced Analyticals) platforms.

## 2.3 Gene annotation preparation

### 2.3.1 Genome and transcript isoform annotation

Given that the quality of genome annotation differs substantially between the studied species and that we aimed for optimal transcript isoform reconstructions for each tissue as a foundation for all analyses in this study, we refined previous annotations from Ensembl (Yates et al., 2016) for each tissue using our previous stranded poly(A)-selected RNA-seq data (Marin et al., 2017; Cardoso-Moreira et al., under review)<sup>1</sup>. Specifically, for each species we downloaded the reference genome from Ensembl release 87 (Yates et al., 2016): hg38 (human), rheMac8 (rhesus macaque), mm10 (mouse), monDom5 (opossum), ornAna1 (platypus), and galGal5 (chicken). For every species-organ combination, the Ensembl annotation was extended using our previous stranded (100 nt, single-end) RNA-seq data (Marin et al., 2017; Cardoso-Moreira et al. under review). Raw reads were first trimmed with cutadapt v1.8.3 (Martin, 2011) to remove adapter sequences and low-quality (Phred score < 20) nucleotides, then reads shorter than 50 nt were filtered out (parameters: `--adapter=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --match-read-wildcards --minimum-length=50 -q 20`). Processed reads were then mapped to the reference transcriptome and genome using Tophat2 v2.1.1 (Kim et al., 2013) (parameters: `--bowtie1 --read-mismatches 6 --read-gap-length 6 --read-edit-dist 6 --read-realign-edit-dist 0 --segment-length 50 --min-intron-length 50 --library-type fr-firststrand --max-insertion-length 6 --max-deletion-length 6`).

We then assembled models of transcripts expressed in each tissue using StringTie v1.3.3 (Pertea et al., 2015) (parameters: `-f 0.1 -m 200 -a 10 -j 3 -c 0.1 -v -g 10 -M 0.5`). Stringent requirements on the number of reads supporting a junction (`-j 3`), minimum gap between alignments to be considered as a new transcript (`-g 10`) and fraction covered by multi-hit reads (`-M 0.5`) were used to avoid merging of independent transcripts and to reduce the noise caused by unspliced or

---

<sup>1</sup>This analysis was designed and performed in collaboration with Dr. Evgeny Leushkin.

incompletely spliced transcripts. We compared the assembled transcript models to the corresponding reference Ensembl annotations using the cuffcompare program v2.2.1 from the cufflinks package (Trapnell et al., 2010). We then combined the newly identified transcripts with the respective Ensembl gene annotation into a single gtf file. We extended the original Ensembl transcriptome annotation by 4.1-18.9 Mbp with novel transcripts and by 26.8-42.0 Mbp with new splice isoforms, providing, as expected, longer total extension for rhesus macaque, opossum, platypus, and chicken than for the well-studied species (i.e., human and mouse) (Table 2).

**Table 2: Tissue-specific genomic annotations for each species**

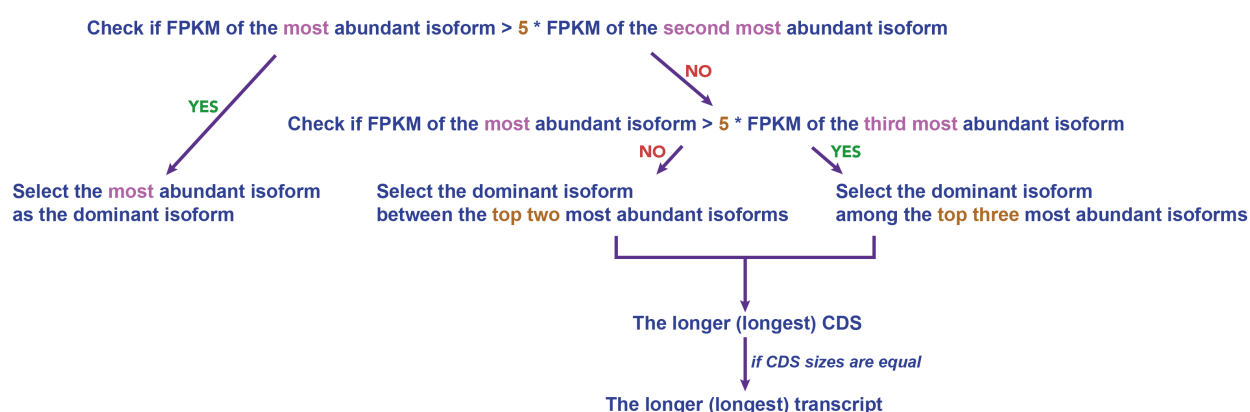
Organism	OTL	NTL (total)	NSF (total)	NTL (brain)	NSF (brain)	NTL (liver)	NSF (liver)	NTL (testis)	NSF (testis)
Human	294,406,476	4,122,449	30,172,016	590,459	9,374,992	633,473	9,872,094	3,149,415	17,946,459
Macaque	103,769,570	14,607,483	39,016,607	4,875,550	21,573,084	2760,546	13,525,542	10,192,411	24,540,108
Mouse	209,740,170	6,911,717	26,781,806	1,182,867	12,856,091	656,829	5,508,795	5,318,906	14,759,290
Opossum	64,969,817	14,970,397	42,010,245	7,094,342	24,278,850	2,735,890	16,759,409	8,767,315	26,859,503
Platypus	42,594,439	16,983,550	31,687,528	7,700,347	20,133,565	5,084,964	15,788,080	11,943,151	20,858,740
Chicken	68,801,956	18,873,114	38,222,455	3,689,553	22,908,096	1,832,002	14,046,304	15,853,413	23,063,594

OTL, original total genome length; NTL, novel transcript length; NSF, new splicing form of existing transcript. The numbers represent the total nucleotide length for each category. Length in base pair (bp).

### 2.3.2 Selection of the dominant splice isoform

Gene expression level estimates may strongly depend on the proper choice of splice isoforms. A previous study based on proteome data suggested that the vast majority of genes have a single dominant splice isoform (Tress et al., 2017), which is not necessarily the longest. In my thesis project, we focused on the dominant isoform, which was identified by taking into account transcript abundances and CDS lengths according to the following criteria (Figure 2.3). For genes with a single annotated isoform, this isoform by definition represents the dominant isoform. For genes with multiple isoforms, I proceed as follows. If the most abundant isoform (i.e., with largest FPKM - fragments per kilobase of transcript per million reads mapped - value based on RNA-seq data) has more than 5 times higher expression level than the second most abundant isoform, then the

most abundant isoform is chosen as the dominant isoform, akin to previous work (González-Porta et al., 2013). Else, I examined if the most abundant isoform has more than 5 times higher expression level than the third most abundant isoform. If so (or if there is no third isoform), I considered the two most abundant isoforms for the final selection step. If not, the final selection was made among the three most abundant isoforms. In the final selection step, the dominant isoform was defined as that with the longest CDS, or, if CDS lengths were the same, the longest transcript.



**Figure 2.3: Schematic representation of the algorithm used to select the dominant splice isoform**

See Section 2.3.2 for more details.

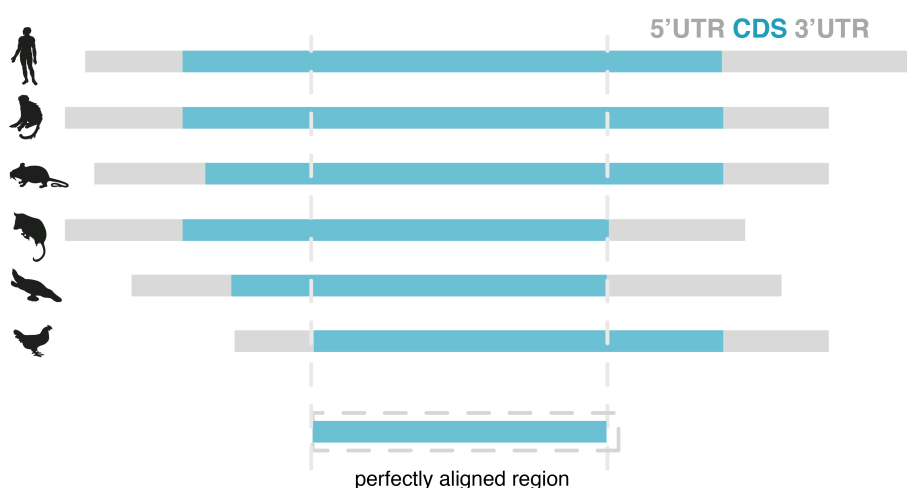
### 2.3.3 Extraction of orthologous gene sets

Gene expression comparisons between species were made based on genes with a 1:1 orthologous relationship across the species investigated in a given analysis. Orthology relationships were extracted from Ensembl, release 87 (Yates et al., 2016). In cases where the dominant splice isoforms of two neighboring genes overlapped in the genome of a species, both genes and their 1:1 orthologs in the other species were removed from all subsequent analyses to avoid read assignment ambiguities. I extracted different sets of 1:1 orthologs for different analyses: 6,327 1:1 orthologs for all six species in this study; 9,325 1:1 orthologs for analyses based on rhesus macaque, mouse, opossum and chicken (i.e., where human was excluded due to the lower number of available replicates and platypus due to the overall low quality of the genome assembly); and

15,668 1:1 orthologs for specific analysis between human and macaque; for the X chromosome dosage compensation analysis I extracted between each of the four therians (human, macaque, mouse, opossum) and chicken (a close outgroup of mammals) 11,876, 10,732, 11,917 and 11,270 1:1 orthologs, respectively.

### 2.3.4 Perfectly aligned coding sequences across species

To ensure that our results and inferences were not affected by potential differences in gene structures between species, key analyses were repeated using only the coding regions of the longest protein-coding isoform of 1:1 orthologs that perfectly align across species (i.e., same length, without any gaps) (Figure 2.4)<sup>1</sup>. Multiple species alignments to human (hg38) obtained from the UCSC site (<http://hgdownload.soe.ucsc.edu/downloads.html>) were used to extract genomic coordinates for sequences that aligned without gaps across all 6 species. This subset of perfectly aligned coding sequences was then used to re-estimate gene expression levels.



**Figure 2.4: Illustration of a perfectly aligned region across species**

See Section 2.3.4 for more details.













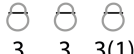


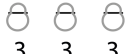


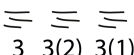
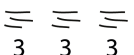
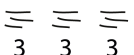
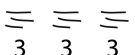
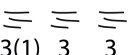
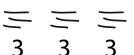
<sup>1</sup>This analysis was designed and performed in collaboration with Dr. Evgeny Leushkin.

### 2.3.5 Compiling structural RNA sequences for each species

To assess how much each library was contaminated by unusable reads generated from structural RNAs, I first collected for each type of major structural RNAs the annotated sequences from multiple public databases. rRNA sequences for each species were retrieved from several sources: Ensembl release 87, SILVA rRNA database v128 (Quast et al., 2013), and NCBI. Transfer RNA (tRNA) sequences were obtained from Ensembl release 87, the genomic tRNA database (gtRNAdb) (Chan and Lowe, 2016), and NCBI. Small nucleolar RNAs (snoRNAs) were downloaded from Ensembl release 87 via BioMart.

## 2.4 Data processing and quality assessment

In total, this resource comprises 54 Ribo-seq and 54 matched RNA-seq libraries (also rRNA-depleted) that were sequenced to a median depth of ~124 million reads and ~112 million reads, respectively (total number of reads: ~14.76 billion) (Figure 2.5). To assess the technical reproducibility of Ribo-seq and RNA-seq protocols, for mouse and chicken liver two technical libraries for each protocol were additionally generated (data not shown in Figure 2.5).

Lineage	Placental mammals			Marsupials	Monotremes	Outgroup
Species	 Human	 Macaque	 Mouse	 Opossum	 Platypus	 Chicken
Tissue						
Ribo-seq #replicate	 3 3 3(1)	 3 3 3	 3 3 3	 3 3 3	 3 3 3	 3 3 3
RNA-seq #replicate	 3 3(2) 3(1)	 3 3 3	 3 3 3	 3 3 3	 3(1) 3 3	 3 3 3

**Figure 2.5: Overview of the translome and matched transcriptome data**

Three biological replicates for each tissue of each species. The numbers in parentheses indicate the number of replicates removed after quality control, otherwise all replicates are used.

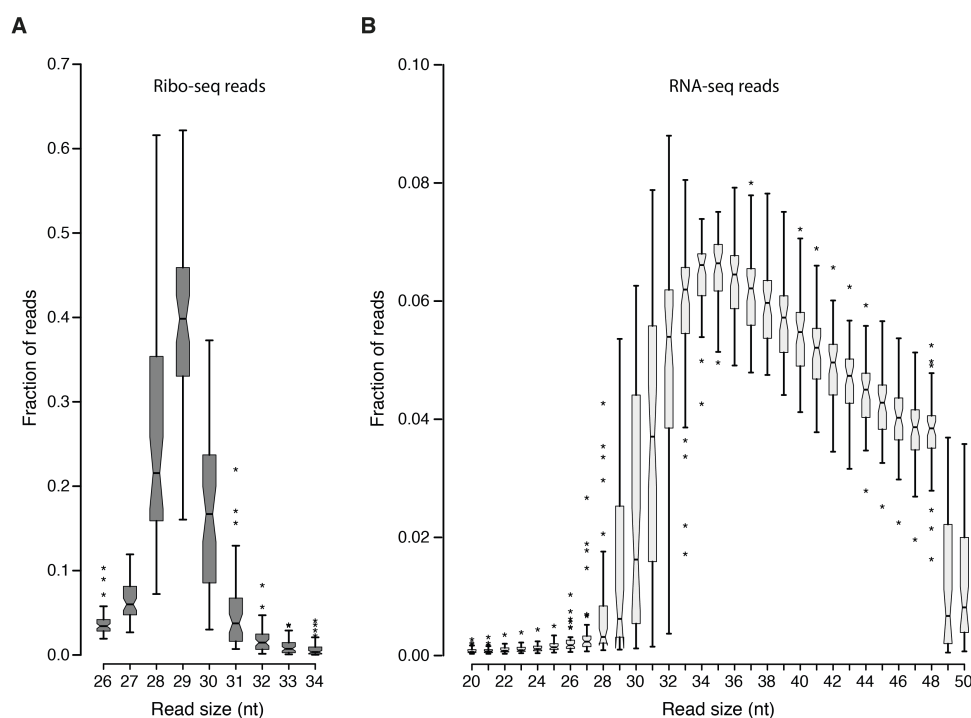
### 2.4.1 Read mapping and processing

Initial quality assessment of the sequencing reads (e.g., average GC, base composition, and variability between clusters) was conducted based on the preliminary quality values produced by the Illumina Casava 1.82 software. Raw reads with known 3' adaptor and low quality bases (Phred score < 20) were trimmed with cutadapt v1.8.3 (Martin, 2011) (parameters: --adapter=AGATCG-GAAGAGCACACGTCTGAACTCCAGTCAC --minimum-length=6 --maximum-length=60 -q 20), and then the clipped reads were sequentially mapped to the index libraries of species-specific rRNAs, human/mouse/rat rRNAs, species-specific tRNAs and species-specific snoRNAs using Bowtie2 v2.3.1 (Langmead and Salzberg, 2012) (parameters: --phred33 -L 20 -N 1 -t --no-unal). I discarded the alignments in each step and kept the unaligned reads. Only reads with specific lengths (26-34 and 20-50 nt for Ribo-seq and RNA-seq reads, respectively) were used in downstream analyses.

After the filtering steps, a total of ~2.72 billion coding sequence reads from 53 Ribo-seq and 50 RNA-seq libraries that passed quality control and that mapped uniquely to the genome and to the dominant transcript isoforms in each organ were used in the downstream analyses (Figure 2.5).

Overall, the Ribo-seq reads (median at ~29 nt, consistent with the biological expectation) are slightly shorter than the RNA-seq reads (median at ~37 nt) (Figure 2.6). To avoid differences in the mappability of reads spanning the exon-exon junction due to read length differences between Ribo-seq and RNA-seq data, RNA-seq reads longer than 29 nt were clipped down to 29 nt to match Ribo-seq reads. Subsequently, the reads were first aligned against organ transcriptomes and then mapped to their respective reference genome with Tophat2 v2.1.1 (Kim et al., 2013) (parameters: --no-novel-juncs --library-type fr-firststrand --read-realign-edit-dist 0 --segment-length 20 --min-anchor-length 5 --min-intron-length 50). Uniquely aligned reads with up to a single mismatch between the query sequence and the reference sequence were accepted. For each gene, only reads that map inside the coding region of its dominant splice isoform were quantified and used.





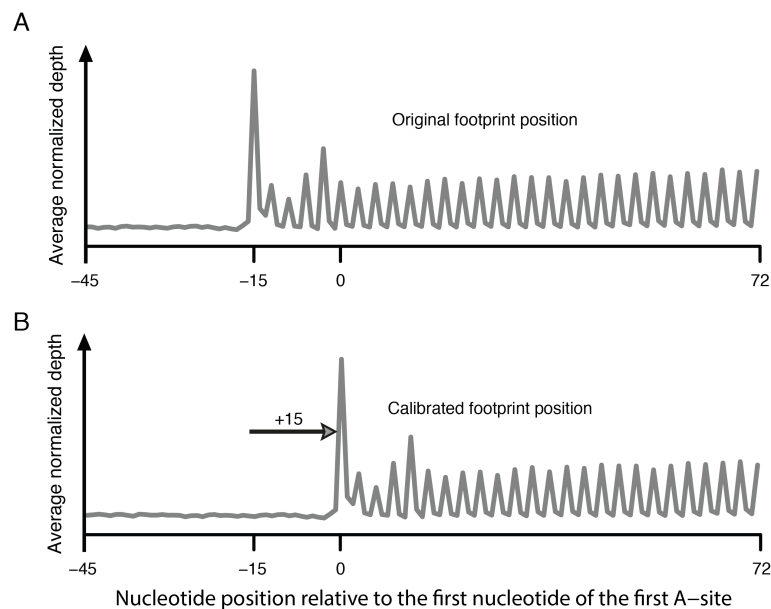
**Figure 2.6: Different read length distributions for Ribo-seq and RNA-seq libraries**

Each boxplot represents the fraction of reads of a particular read size across 53 Ribo-seq (A) or 50 RNA-seq (B) libraries. The median lengths of (A) Ribo-seq reads and (B) RNA-seq reads are ~29 and ~37 nucleotides (nt), respectively.

### 2.4.2 A-site calibration

Instead of assigning the read to the whole sequence to which it corresponds, ribosome footprints were assigned to the first nucleotide position of the ribosomal A-site (aminoacyl-tRNA site) on the basis of the length of each fragment (Ingolia et al., 2011). In order to do this, I took advantage of the empirical observation of the distribution of ribosome footprints; i.e., that there is an increased read density at the beginning of CDSs that represent initiating ribosomes. For reads of a particular size (between 26 nt and 34 nt), I asked how far the 5' end of the reads was from the annotated first A-site (the codon succeeding the start codon, the first P-site). This peak in the distance distribution was then used to adjust the alignment (Figure 2.7A). This method was used to define the distance by which reads of different sizes ought to be adjusted in order to yield A-site mapped reads, and all transcript-mapped reads (not just those overlapping the start codon)

were adjusted accordingly. For all ribosome footprints, the offset between the 5' end of the alignment and the first nucleotide in the A-site is 15 nt (Figure 2.7B). The homogenous A-site offset for different fragment sizes may be partially explained by the fact that the RNA nuclease used in the protocol (i.e., RNase I) has little (if any) sequence specificity, so that its cutting is especially precise at the 5' end of RNA fragments (Jackson and Standart, 2015). Furthermore, given that various analyses in this study compare translation levels and RNA abundances, RNA reads were processed in the same way as ribosome footprints; that is, they were assigned to the 16<sup>th</sup> (offset as +15).



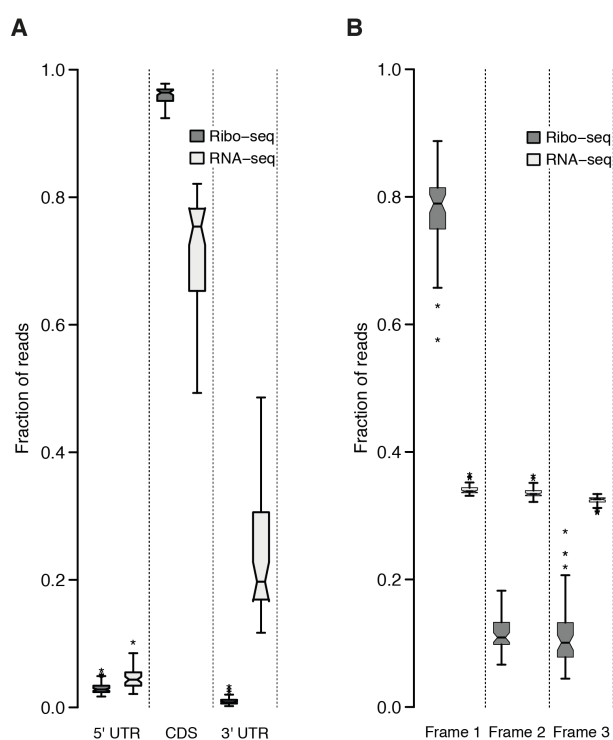
**Figure 2.7: Calibration of the A-site of ribosome footprints**

Metagene profiles of raw (A) and A-site adjusted (B) ribosome footprints proximally aligned to the first nucleotide of the first A-site (the codon succeeding the start codon, i.e., the first P-site). The A-site for each length of ribosome footprints (26-34 nt) was adjusted; offset of the A-site is +15 from the 5' end for all alignments.

### 2.4.3 Triplet periodicity analysis

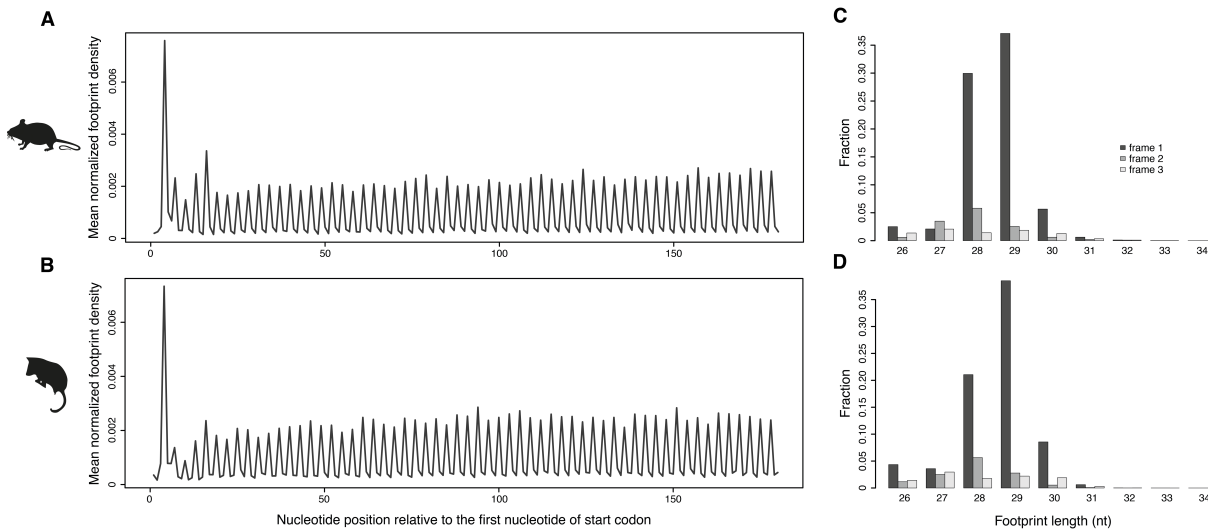
Given that Ribo-seq remains a non-trivial technique, especially when applied to primary organs from different non-model species, I evaluated in detail the quality of the data. Ribosome footprints predominantly mapped to the main coding region, which is consistent with previous work

(Janich et al., 2015) (Figure 2.8A). I then used triplet periodicity to evaluate the quality of the Ribo-seq data, given that it reflects the pattern of genuine translation. Footprint profiles within CDSs were generated by assigning ribosomal A-sites to each nucleotide position of each codon (reading frames 1, 2, and 3). The number of reads mapped to each of the three reading frames was normalized by the total number of reads within the CDS. In sharp contrast to the RNA-seq reads, which mapped evenly to the three codon positions, ~70-85% of the ribosome footprints in each sample mapped periodically to the canonical open reading frame (Figure 2.8B), in agreement with previous work (Janich et al., 2015). The average footprint density of metagene profiles along the CDS faithfully reflects mRNA translocation by codon as translation occurs (Figure 2.9).



**Figure 2.8: Distributions of reads on transcript features and on the three reading frames**

(A) Proportions of Ribo-seq (dark grey boxes) and RNA-seq (light grey boxes) reads mapping to 5' UTRs, CDS, and 3' UTRs, respectively. (B) Distribution of Ribo-seq and RNA-seq reads across the three reading frames in the CDS of dominant splicing isoforms (Frame 1: canonical reading frame).

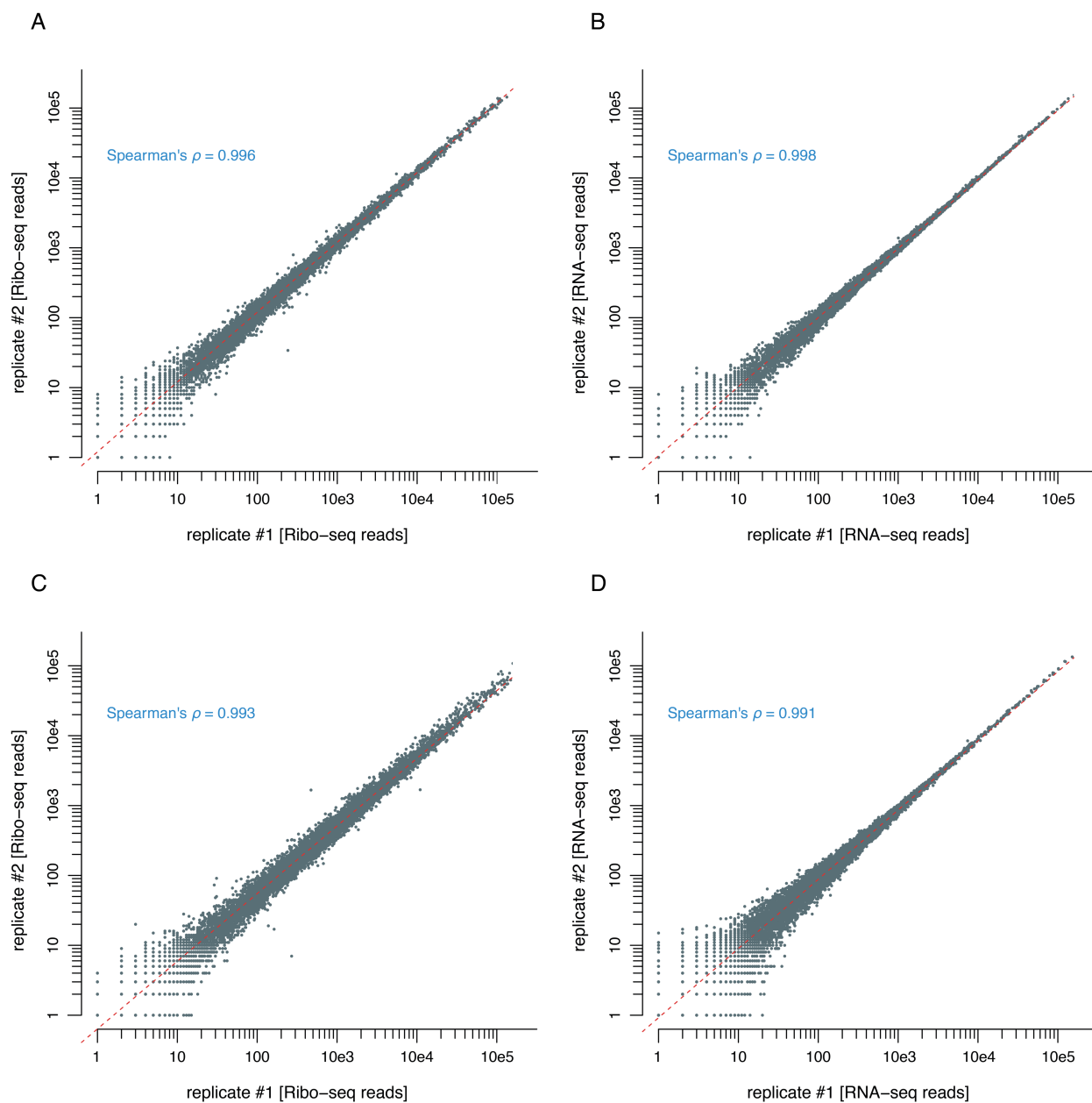


**Figure 2.9: Metagene profiles of ribosomal footprints**

(A and B) Mean normalized density of footprints along the coding region of the dominant isoforms of protein coding genes for Ribo-seq libraries of mouse and opossum brains. The Ribo-seq read (A-site) density for each position is plotted relative to the first nucleotide position of the start codon. (C and D) For the corresponding data in A and B, the fractions of A-sites falling within the three reading frames and for each footprint length (26-34 nt).

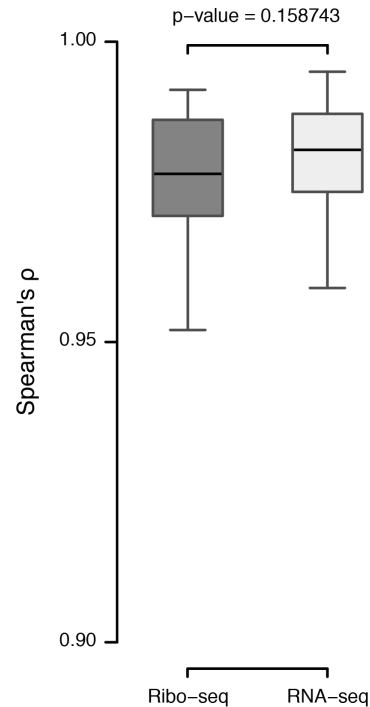
#### 2.4.4 Assessment of reproducibility for both data types

To assess the reproducibility of the Ribo-seq and RNA-seq datasets and its similarity between the two data types, Spearman's rank correlation coefficient ( $\rho$ ) of expression values for genes with a median read count  $> 1$  between each pair of biological replicates and technical replicates (generated for mouse and chicken liver samples) were calculated for the Ribo-seq and RNA-seq data. The high correlation coefficients observed across technical replicates ( $\rho > 0.99$ ) (Figure 2.10) and biological replicates ( $\rho$ : 0.95-0.99, median: 0.98) (Figure 2.11) for both the Ribo-seq and RNA-seq datasets indicate high technical/biological reproducibility (i.e., low technical/biological variation). Notably, the Spearman's  $\rho$  and hence the reproducibilities are statistically indistinguishable between the two data types (Figure 2.10 and Figure 2.11). These observations testify to the high quality of the data and rule out the possibility that observations made in downstream biological analyses are explained by technical differences between the Ribo-seq and RNA-seq datasets (e.g., higher technical variation in the Ribo-seq data than in the RNA-seq data).



**Figure 2.10: Correlations between technical replicates for Ribo-seq and RNA-seq data**

Spearman's  $\rho$  of genes with a median read count  $> 1$  between the two technical replicates was calculated for mouse liver Ribo-seq (A) and RNA-seq (B) data, and for chicken liver Ribo-seq (C) and RNA-seq (D) data.



**Figure 2.11: Correlations between biological replicates for Ribo-seq and RNA-seq data**

For each tissue of each species in this work, Spearman's  $\rho$  were calculated based on genes with median read count  $> 1$  across biological replicates for Ribo-seq (dark grey box) and RNA-seq (light grey box) data. The correlations between the two data types are statistically indistinguishable (Mann-Whitney  $U$  test).

#### 2.4.5 Expression levels and normalization

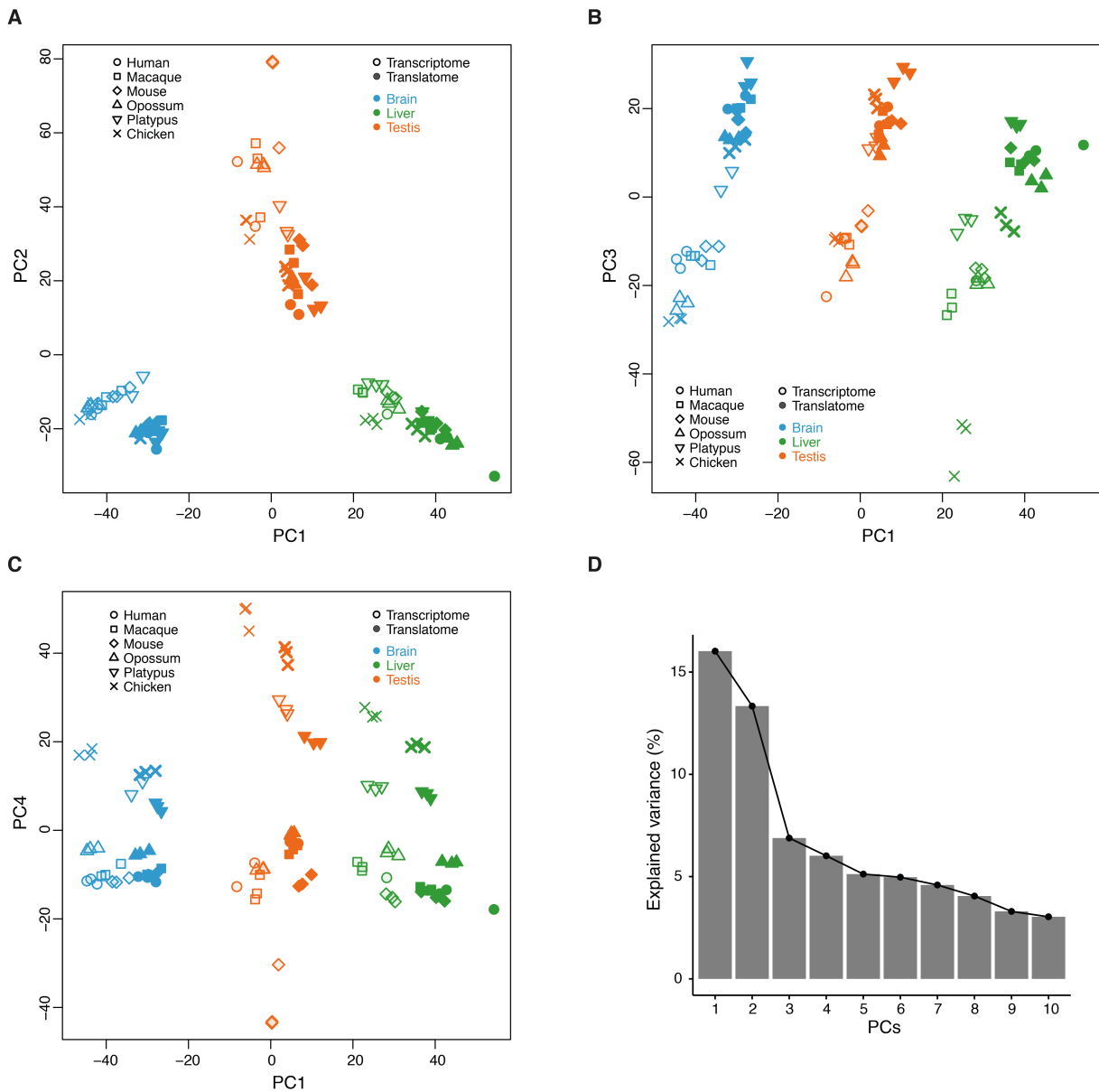
RNA abundance and ribosome occupancy (translation rate) for each gene were measured in fragments per kilobase of CDS per million uniquely CDS-aligning reads (FPKM), a unit which corrects for both feature length and sequencing depth. FPKM based only on the coding region of each locus (i.e., the dominant splice isoform — see above) for both Ribo-seq and RNA-seq libraries was calculated, to exclude biased measurements due to heterogeneous quality of annotations for UTRs across species/tissues and the fact that Ribo-seq reads, contrary to RNA-seq reads, predominantly map to the main coding region. To render the data comparable across species and tissues, translational and transcriptional FPKMs were separately normalized based on our published approach (Brawand et al., 2011). Specifically, among the genes with expression values in the interquartile range, I identified the 1,000 genes that have the most conserved ranks among samples and calculated their median expression levels in each sample. I then derived scaling

factors that adjusted these medians to a common value. Finally, these factors were used to scale expression values of all genes in the samples.

#### 2.4.6 Principal component analysis

To obtain a global overview of the transcriptomes and translomes across species, I performed a principal component analysis (PCA) based on the set of 5,231 robustly expressed (median FPKM > 1 across all RNA-seq libraries, no filtering for Ribo-seq data) 1:1 amniote orthologs, using the *'prcomp'* function in the *'stats'* R package.

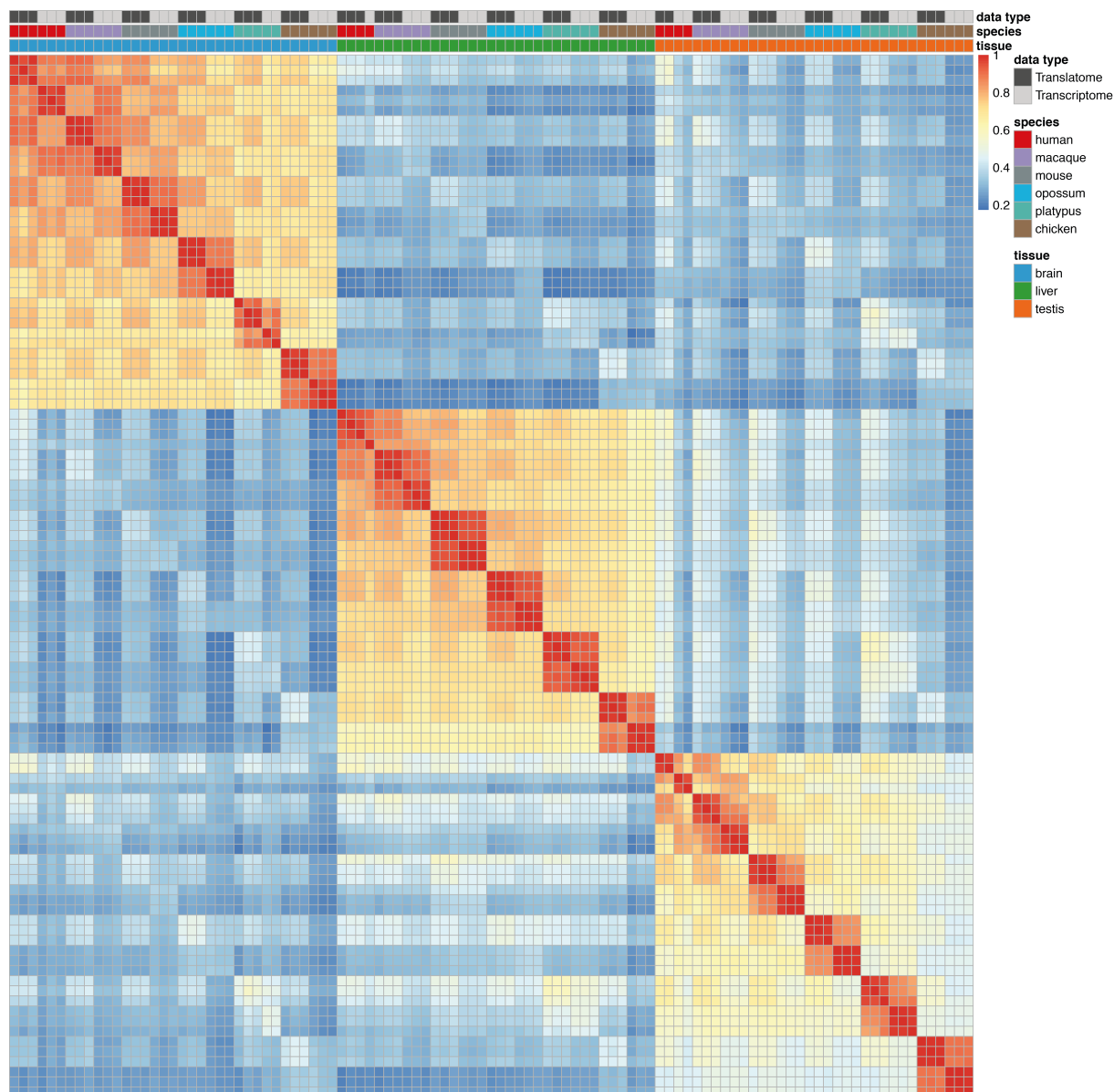
The first principal component (PC1), explaining most gene expression variance, separates the samples by organs (Figure 2.12). This observation is concordant with previous analyses of adult and developmental transcriptomes (Necsulea and Kaessmann, 2014a; Brawand et al., 2011; Cardoso-Moreira et al., under review), and it represents an expected pattern, given that the studied organs originated in common vertebrate ancestors long before the emergence of amniotes (i.e., the mammalian/avian species studied here) and that their principal functions are the same across vertebrates (Necsulea and Kaessmann, 2014a). PC2 separates the germline (testis) and somatic (brain and liver) data (Figure 2.12A), also consistent with previous work (Brawand et al., 2011), while PC3 represents the distinct clustering of the translome and transcriptome data (Figure 2.12B). Finally, PC4 separates the data according to the different species/lineages (Figure 2.12C). The overall highly consistent clustering of the different aspects of the translome and transcriptome data (see also correlation heatmap, Figure 2.13) is a further indicator of the high data quality and provides a firm basis for evolutionary investigations across these two major gene expression layers. The results of the correlation heatmap based on perfectly aligned regions of the set of 6,327 1:1 amniote orthologs (Supplementary Figure 1) are similar to those in Figure 2.13 and thus confirm the robustness of the observations.



**Figure 2.12: PCA of two gene expression layers across different organs and species**

The PCA is based on the set of 5,231 robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 amniote orthologs. The same genes were used for Ribo-seq libraries. **(A to C)** PC1 reflects gene expression variance attributable to differences between organs. **(A)** PC2 separates germline (testis) and somatic (brain and liver) data. **(B)** PC3 represents the distinct clustering of the translatome and transcriptome data. **(C)** PC4 separates the data according to the different species/lineages. **(D)** The scree plot indicates the percentage of variance explained by each of the first 10 PCs.





**Figure 2.13: Correlations of two gene expression layers across different organs and species**

The heatmap of the pairwise correlations (Spearman's  $\rho$ ) is based on the set of 5,231 robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 amniote orthologs. It represents the degree of similarity of gene expression profiles between data types (translatome, transcriptome), species (human, macaque, mouse, opossum, platypus, chicken) and tissues (brain, liver, testis).



# Chapter 3 Widespread translational buffering of individual genes across organs

## 3.1 Introduction

A previous study of orthologous genes in different mammalian species and lineages reported many potentially adaptive individual mRNA expression level shifts during various time points of mammalian evolution (Brawand et al., 2011). However, whether these shifts at the mRNA level can also be seen on the more functionally relevant protein level remained unclear. Given that mRNA expression changes might be buffered at the translational level and that genes might show shifts at both expression layers or at the translational level only, in this chapter we aimed to screen, in parallel, both the transcriptome and translome data for expression shifts using dedicated methods.

By comparing transcript abundance and ribosome occupancy, it is possible to distinguish between different evolutionary scenarios: First, genes show significant expression shifts only at the transcriptional level. These might represent cases where protein levels are completely buffered against changes in mRNA levels and are evolving under compensatory selection pressures (i.e., mRNA expression shifts are offset by compensatory TE shifts).

Second, genes show consistent expression shifts at both levels of gene expression. These would suggest that they experienced changes at the phenotypically relevant protein level and indeed contributed to the phenotypic evolution. In the case of consistent change on both expression layers, we, by also considering translation efficiencies (TE) for these orthologs across species (McManus et al., 2014), established whether divergence is equally pronounced at the mRNA and

translational level (i.e., the change of mRNA abundance drives overall expression divergence), more pronounced at the translational level (i.e., TE changed in the same direction as mRNA abundance divergence; e.g., a higher efficiency in the case of mRNA abundance increase), or less pronounced at the translational level (i.e., TE changed in the opposite direction of mRNA abundance divergence).

Third, by considering TE, we also screened for genes that show selectively driven expression shifts at the translational level but evolved under stabilizing selection at the mRNA level. These interesting cases, which can be explained by lineage-specific translation regulatory changes (i.e., alterations in TE), may thus (in addition to shifts that occurred at both regulatory levels) also represent adaptive shifts that contributed to lineage-specific phenotypic innovation.

## 3.2 Methods

### 3.2.1 Translational tuning index

To assess the extent of translational tuning (e.g., buffering or reinforcement) for individual genes, we devised a translational tuning index (TTI) as follow (Figure 3.1)<sup>1</sup>:

$$TTI = \frac{LFC(Ribosome\ occupancy)}{LFC(RNA\ abundance)} - 1$$

where  $LFC(Ribosome\ occupancy)$  denotes  $\log_2$ -fold changes (LFC) of ribosome occupancy and  $LFC(RNA\ abundance)$  denotes LFC of RNA abundance between two species of interest (e.g., mouse and chicken) for a given tissue.

LFC was calculated separately based on Ribo-seq or RNA-seq read count data using DESeq2 v1.14.1 (Love et al., 2014), an R package that estimates and accounts for biological variability in a statistical test based on a negative binomial distribution under the generalized linear model.

---

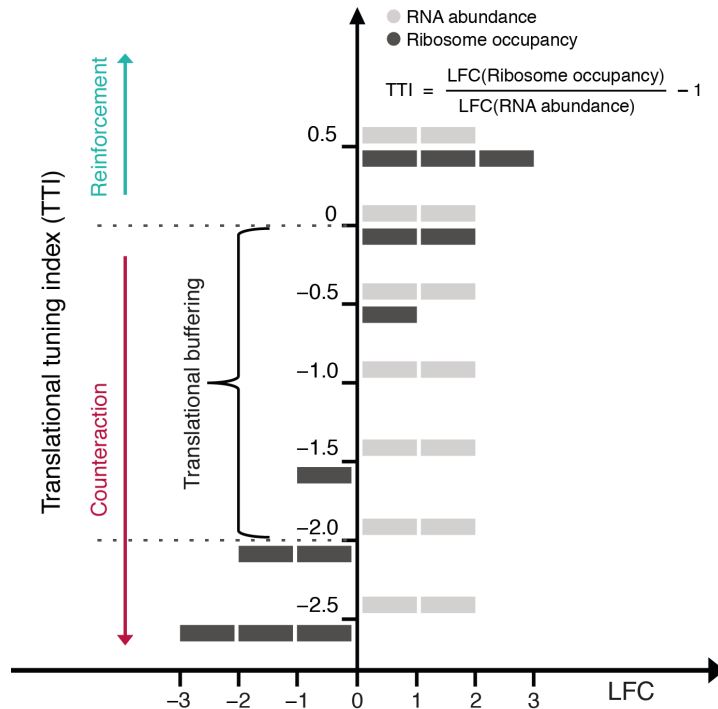
<sup>1</sup>This method was designed in collaboration with Dr. Simon Anders and Dr. Evgeny Leushkin.

Read counts were normalized by their respective gene lengths prior to analyses. The effective library size of each deep-sequencing library was then determined, and raw read counts were normalized by their respective scaling factors so that the median read count was the same for all libraries. LFC and its standard error ( $LFC_{SE}$ ) were then estimated, and p-values were calculated based on the stats ( $LFC/LFC_{SE}$ ) with a Wald test. Finally, p-values were corrected using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). A false discovery rate (FDR) < 5% was used when calculating significance of LFCs.

The  $TTI_{SE}$  was computed to determine whether the translational compensation or reinforcement for a given gene is of statistical significance. For each LFC value, there is an SE provided in the DESeq2 result table and the LFC value falls in the range between  $LFC-SE$  and  $LFC+SE$ . When putting these two extremes of LFC into the TTI formula, two TTI extremes are created. In this formula, there are two LFCs (for RNA abundance and ribosome occupancy data, respectively) involved to compute the TTI, so there are four TTIs obtained by substituting into each of the four combinations of  $LFC\pm SE$  for RNA abundance and  $LFC\pm SE$  for ribosome occupancy. The estimate of SE for TTI ( $TTI_{SE}$ ) for each gene was defined as half the range between the smallest and the largest TTIs. P-values were calculated according to the z-score of TTI ( $TTI/TTI_{SE}$ ) with the '*pnorm*' function in the '*stats*' R package, and were then corrected for multiple hypothesis testing using the BH method with '*p.adjust*' function in the '*stats*' R package. Note that TTI is ill-defined when LFC of RNA abundance is close to zero. Therefore, the analysis of translational tuning was restricted to genes with statistically significant RNA expression differences between species.

Overall, this approach is similar to that developed in a previous study (Bader et al., 2015) but considers  $LFC_{SE}$  when estimating significance of TTIs.

In Figure 3.2, the  $TTI_{SE}$  is reflected by the extent of transparency; the larger the SE, the more transparent the plotted data point. For visualization purposes,  $TTI_{SE}$  was transformed with the formula: square root of 1 over  $TTI_{SE}$ , where  $TTI_{SE}$  is capped at 10, with more extreme values replaced by this value (R code: `sqrtpmin(10, 1/TTI.SE)`).



**Figure 3.1: Illustration of the translational tuning index (TTI)**

A TTI value of 0 for a given gene indicates that the transcriptional change was not altered at the translational level;  $TTI > 0$  implies that the transcriptional change was reinforced at the translational level; and  $TTI < 0$  means that the transcriptional change was counteracted by an opposing translational alteration (e.g.,  $TTI = -1$  indicates that the transcriptional change was completely offset at the translational level). TTI values between -2 and 0 indicate translational buffering of the transcriptional change (i.e., expression divergence is attenuated at the translational level). LFC,  $\log_2$ -fold change.

### 3.2.2 Identification of translational efficiency changes

In conjunction with RNA expression measurements, Ribo-seq enables the estimation of translational efficiency (TE) by capturing a snapshot of the transcriptome-wide ribosome occupancy (Brar and Weissman, 2015). DESeq2 estimates the effective library size of each deep-sequencing library and normalizes raw count data accordingly. Specifically, for the analysis of translation regulation in one tissue, a linear regression was performed to the normalized read counts with log link, as a function of data type (RNA-seq and Ribo-seq) and replicate variables (replicate 1, replicate 2 and replicate 3). Here the coefficient of data type variable (Ribo-seq over RNA-seq) is a measurement of TE. To reveal the differences in translational regulation when comparing between species for the same tissue, I analyzed the interaction term of the data types (Ribo-seq

and RNA-seq) between species (e.g., mouse and chicken). Using a likelihood ratio test, which removes the interaction term between data type and species (*dataType:species*) from the full model as the reduced model, DESeq2 calculates the coefficient of the interaction term as the measurement of changes in TE between species (e.g., mouse and chicken). P-values were adjusted using the BH procedure for multiple testing. The cutoff of FDR < 1% was used when calling significant cases.

*full model = ~ dataType + sample + dataType:species*  
*reduced model = ~ dataType + sample*

### 3.2.3 Enrichment analysis

The weighted-mean method was developed to test whether the mean TTI of genes with significant TE changes (Sig.TE) between species are overall statistically different from that of other genes. When calculating weighted mean of TTI,  $TTI_{SE}$  was also considered, i.e., a gene with a relatively smaller/larger SE is assigned to a relatively larger/smaller weight: 1 over square root of  $TTI_{SE}$  (R code:  $1/\sqrt{TTI.SE}$ ). Weighted means of TTI for both Sig.TE and other genes were computed using the *'weighted.mean'* function in the *'stats'* R package. A permutation test was performed to assess whether the difference between the two weighted means occur by chance: I first randomly picked the same number of genes as Sig.TE and labeled the rest; and then computed the difference between the weighted means of the two groups; this analysis was repeated 10,000 times to get the permutation distribution; finally, a two-sided p-value was calculated.

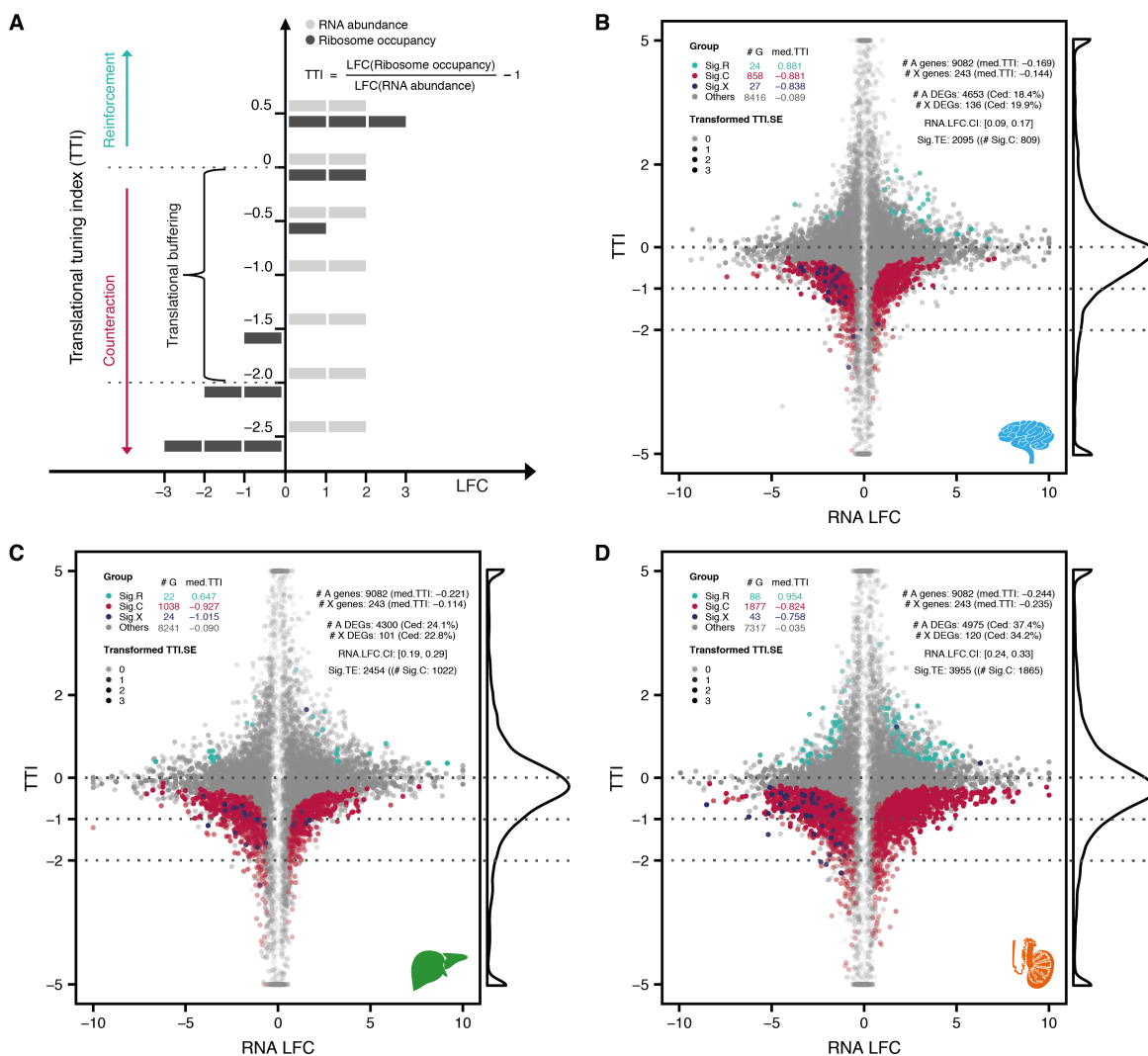
## 3.3 Results

The extent to which evolutionary changes in transcript abundance of individual genes (measured by the RNA-seq data) are reflected at the level of protein synthesis (measured by the Ribo-seq data), was assessed by the translational tuning index (TTI) (Figure 3.1 and Figure 3.2A, Section 3.2.1). A TTI value of 0 for a given gene indicates that the transcriptional change was not altered at the translational level;  $TTI > 0$  implies that the transcriptional change was reinforced at the

translational level; and  $TTI < 0$  means that the transcriptional change was counteracted by an opposing translational alteration (e.g.,  $TTI = -1$  indicates that the transcriptional change was completely offset at the translational level). TTI values between -2 and 0 indicate translational buffering of the transcriptional change (i.e., expression divergence is attenuated at the translational level).

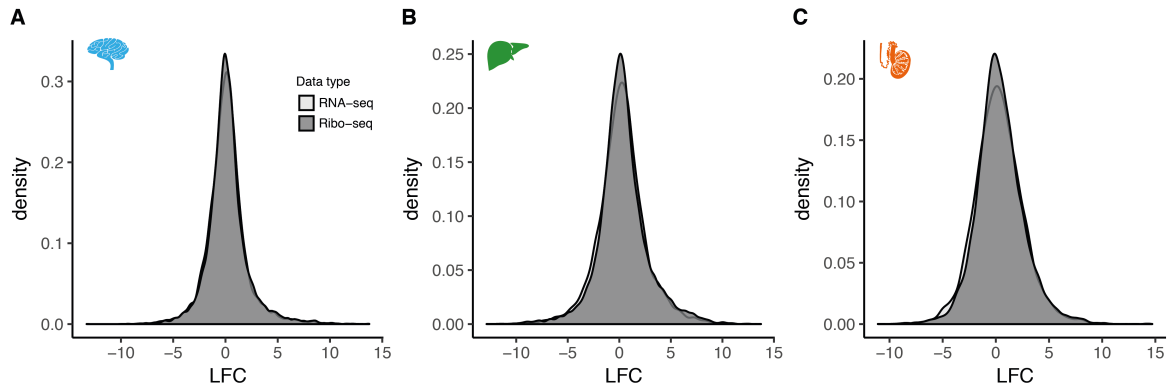
TTI analyses across representative species pairs revealed global shifts of TTI value distributions, with median TTI values for autosomal genes across organs and species pairs being significantly smaller than 0 in all instances (range of median TTI values: -0.36 to -0.05) and, depending on the organ and species pair (i.e., different evolutionary divergence times), small to substantial proportions (range: 1.8% to 40.4%) of significantly compensated transcript abundance changes (Figure 3.2, B to D and Supplementary Figures 2 to 7). I note that the few TTI values below -2 have high variances and therefore likely do not indicate increased expression divergence through overly strong opposing translational changes. I also note that the observed patterns are not explained by technical differences between the Ribo-seq and RNA-seq data (e.g., higher technical variation in the Ribo-seq data than in the RNA-seq data), given that the correlations between replicates for both data types are high and statistically indistinguishable from each other (Figure 2.11), and that the distributions of  $\log_2$ -fold expression level differences are highly similar for the two data types (Figure 3.3).





**Figure 3.2: Translational versus transcriptional changes for individual genes**

(A) Illustration of the translational tuning index (TTI) (LFC,  $\log_2$ -fold change). (B to D) TTI versus transcript (RNA) abundance changes (RNA LFC) for 9,325 1:1 orthologous genes between mouse and chicken (the reference) for brain, liver, and testis, respectively (see Section 3.2.1 for more details). To the left in each plot: the number (# G) and median TTI (med.TTI) of autosomal genes with significant compensation (Sig.C, red) or reinforcement (Sig.R, cyan); the total number of significantly compensated or reinforced X-linked genes (Sig.X, blue) is also indicated. The transformed standard error (SE) of TTI is reflected by the extent of transparency; the larger the SE, the more transparent the plotted data point. To the right in each plot: the numbers of mouse autosomal (A) genes and X-linked (X) genes with 1:1 orthologs in chicken genome, with median TTIs labeled in parentheses; the number of differentially expressed genes (DEGs) detected based on RNA-seq data comparisons, with the proportions of compensated genes among the DEGs shown in parentheses; the 95% confidence interval (CI) of the RNA LFC; the number of genes showing differential TE between mouse and chicken, with the shared number of genes showing significant compensation in parentheses (see also Supplementary Figures 2 to 7). For display purposes, TTIs and RNA LFCs were capped at -5 and 5, and at -10 and 10, respectively, with more extreme values replaced by these values. TTI density distributions for all genes are shown to the right of each scatter plots.



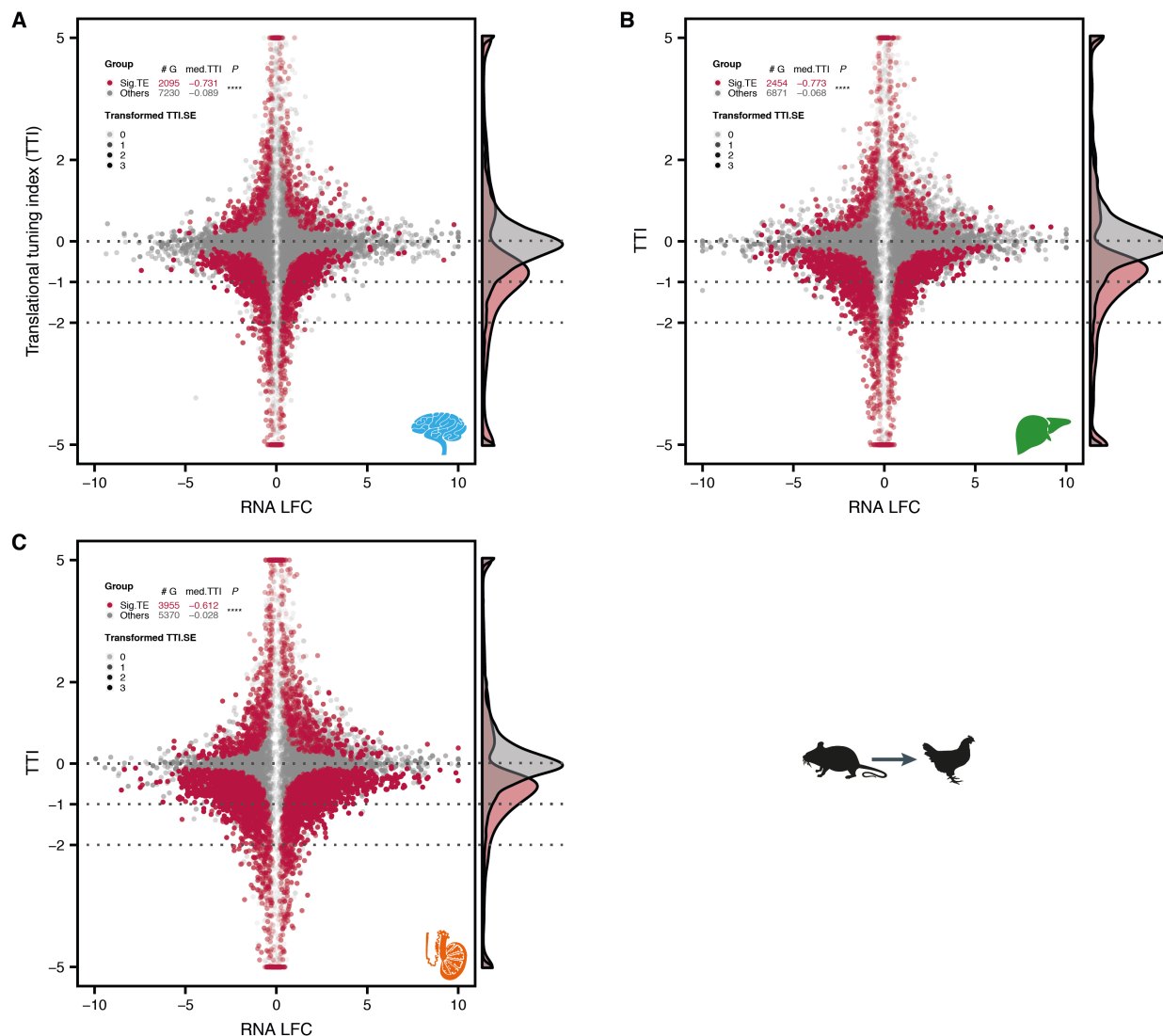
**Figure 3.3: Distributions of log<sub>2</sub>-fold changes for the two gene expression layers**

Distributions of ribosome occupancy changes (log<sub>2</sub>-fold change, LFC) and RNA abundance changes (LFC) for 9,325 1:1 orthologs between mouse and chicken (the reference) for brain (A), liver (B), and testis (C), respectively.

Together, the observations therefore imply an overall extensive buffering of transcriptome divergence through antagonistic changes on the translational layer, as also reflected by numerous corresponding changes of TEs (Figure 3.4 and Supplementary Figures 8 to 13). Notably, the extent of compensation is particularly high in testis; for example, 1,860 of 4,975 (37.4%) genes differentially transcribed between mouse and chicken show significant translational compensation (Figure 3.2, B to D and Supplementary Figures 2 to 7). Thus, translational buffering seems to strongly counteract the particularly rapid evolutionary transcriptome divergences in the adult testis (Brawand et al., 2011), which is at least partly due to an overall reduction of purifying selection at the transcriptional layer because of relaxed transcriptional regulation in dominant spermatogenic cell types (Schmidt, 1996; Kleene, 2001; Soumillon et al., 2013).

In contrast to the many attenuating translational changes, there are generally very few genes where transcriptional changes were significantly reinforced at the level of translation (i.e., TTI > 0) (Figure 3.2, B to D and Supplementary Figures 2 to 7). Thus, mutational changes in transcriptional regulation were apparently rarely further boosted by mutations affecting translational regulation. Overall, the TTI observations suggest that many transcript abundance changes would have been deleterious without compensation at the translational level and therefore reveal an

important role of translational buffering in stabilizing gene expression levels during mammalian organ evolution.



**Figure 3.4: TTI distribution for genes of significant TE changes between mouse and chicken**

(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the larger the SE, the more transparent the plotted data point (Section 3.2.1). For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (see Section 3.2.3 for details) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



---

# Chapter 4 Global patterns of gene expression conservation

## 4.1 Introduction

A major aim of this chapter is to shed light on the evolutionary dynamics of protein synthesis rates in mammals, in particular in light of recent cross-mammalian mRNA expression studies (Khaitovich et al., 2006; Romero et al., 2012; Necsulea and Kaessmann, 2014a). We thus studied for each organ the evolution of global gene expression of 1:1 orthologs at both the transcriptional and translational layers. Evolutionary rates of expression change across mammalian lineages were determined using our previously established approach, which is based on phylogenetic expression distance analyses (Brawand et al., 2011). This analysis revealed commonalities and differences of global evolutionary patterns between the two gene expression layers.

We assessed whether protein synthesis rates are, in general, more conserved during mammalian evolution than mRNA levels, as could be partially predicted from the results of individual genes reported in chapter 3 and the previously reported more pronounced preservation of protein abundances compared to mRNA levels in other species and mammalian cell lines (Schrimpf et al., 2009; Weiss et al., 2010; Laurent et al., 2010; Khan et al., 2013). Evolutionary changes in mRNA levels for many genes may be effectively neutral, if buffered or compensated for at the protein level (Khan et al., 2013).

Previous work revealed different mRNA expression divergence rates across tissues (Brawand et al., 2011). For example, neural tissues evolve slowly and the testis very rapidly in terms of mRNA

expression divergence. It is important to assess to what extent this pattern is recapitulated at the translation level. In particular, it will be interesting to contrast mRNA and translation divergence rates for the testis. The high overall rate of mRNA expression divergence in the testis may at least partly have been facilitated by the nonfunctional transcription and potentially more relaxed purifying selection in this tissue compared to other organs (Soumillon et al., 2013). It is hypothesized that the testis may be less of an outlier when translation rates are compared across species, given that purifying selection may be more pronounced at the level of protein synthesis and that the “noisy” transcription in the testis might be translationally repressed (Kleene, 2001). Conversely, the high conservation of brain expression could be expected to be even more pronounced at the translation level.

## 4.2 Methods

### 4.2.1 Gene expression phylogenies for each organ

For each organ, we only considered genes expressed with median FPKM  $> 1$  in all RNA-seq libraries for that organ. 5,361, 4,630, and 5,237 genes among the 6,327 1:1 amniote orthologs were considered for brain, liver and testis, respectively. We constructed gene expression trees using the neighbor-joining (NJ) approach, based on pairwise distance matrices between all samples for a given organ, following our previous procedure (Brawand et al., 2011). The distance between samples was computed as  $1 - \rho$ , where  $\rho$  is Spearman’s correlation coefficient; unlike Pearson’s correlation coefficient, it is robust to outliers and any potential inaccuracies in the normalization procedure. The NJ trees were constructed using functions in the ‘ape’ package in R (Paradis et al., 2004). The reliability of branching patterns was assessed with bootstrap analyses (1:1 orthologs were randomly sampled with replacement 1,000 times). The bootstrap values are the proportions of replicate trees that share the branching pattern of the majority-rule consensus tree shown in the figures. As noted above, all main biological analyses in my thesis work, including the phylogenetic analysis of gene expression, were performed using the dominant splice

isoform. To verify the robustness of the observations, I also reconstructed the expression phylogenies based on the set of perfectly aligned coding sequences.

#### 4.2.2 Total branch length analysis

I compared the total branch lengths of expression trees across the three organs (brain, liver, and testis) and between data types (RNA-seq and Ribo-seq). Because the X chromosome has evolved in distinct ways due to sex-related selective pressures (Necsulea and Kaessmann, 2014a), only autosomal 1:1 orthologs were used in this analysis. For each organ, I only considered genes with median FPKM  $> 1$  across all RNA-seq libraries. I then calculated for each organ the total tree length by summing up the divergence along the internal (shared) branches leading to the individuals of the different species first for the RNA-seq data. Subsequently, the same genes were used for Ribo-seq total tree length calculations without any further gene filtering. Next, the ratio of translome tree length to transcriptome tree length was computed. The reliability of total tree length estimates was assessed with bootstrap analyses; all genes were randomly sampled with replacement 1,000 times; for each round of bootstrapping, the same procedure of tree calculation was repeated, and the ratio of translome tree length to transcriptome tree length was computed. Finally, the median value of the 1,000 ratios was reported. To examine the conservation and compensation for different gene classes, I repeated the aforementioned procedure for each of the gene sets. Using the full set of orthologous genes as a reference, I calculated the difference between a given gene class and the reference by deducting the median ratio of that gene class from the median ratio of its corresponding reference. Negative/positive values for the difference indicates that the gene class is overall more/less buffered than the genomic background. Mann-Whitney  $U$  tests were employed to estimate the significance for the 1,000 ratios between the gene set and its reference.

### 4.2.3 Estimating modularity in gene expression changes

To estimate to what extent concerted gene expression changes contributed to the conservation of relative gene expression levels, Spearman's  $\rho$  of gene expression values were compared between species observed in the data to those obtained in a simulated scenario<sup>1</sup>. In the simulation, gene expression changes occur independently from each other; therefore, the contribution of concerted changes is removed. To model gene expression changes, we first fit a *LOESS* curve for the difference in median  $\log_e$ -expression values between two species of interest. Based on the expected differences in median  $\log_e$ -expression levels and the estimated error, a simulated set of expression values was generated, taking gene expression values of one of the species as a starting point. The simulated dataset has approximately the same amount of individual changes as in our analyses, but those changes are no longer intrinsically dependent on each other. 1,000 simulations were performed for the species pair mouse-opossum for all three organs and all three replicates.

---

<sup>1</sup>This analysis was designed and performed in collaboration with Dr. Evgeny Leushkin.

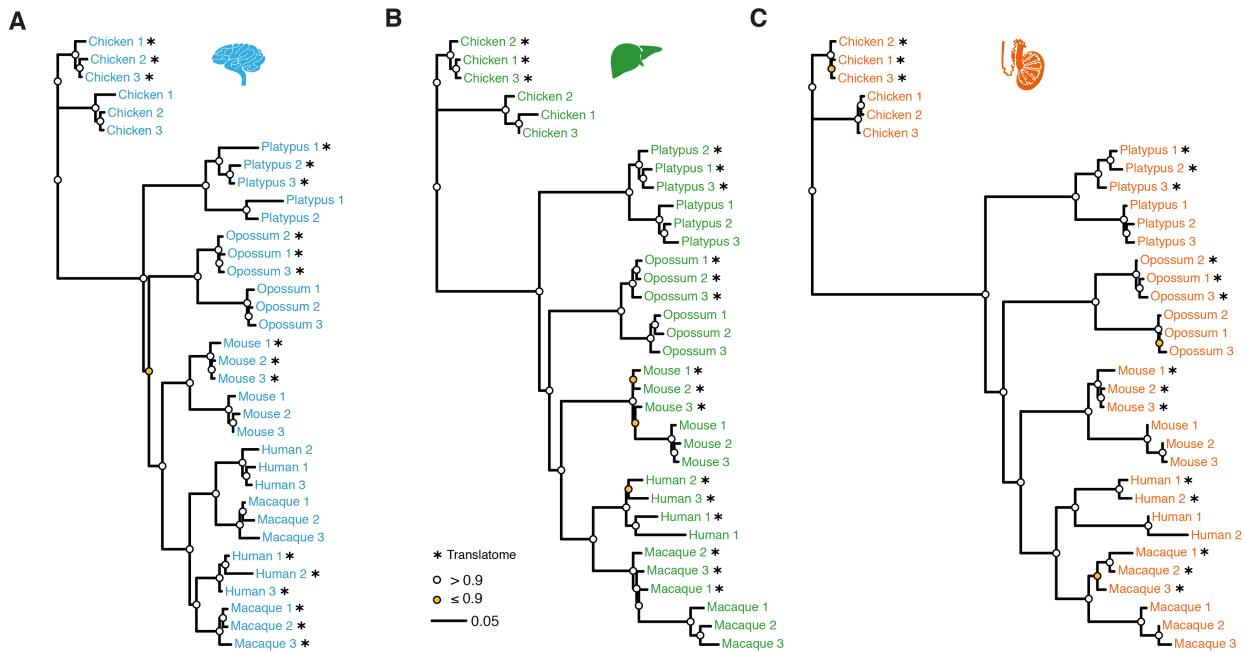


## 4.3 Results

### 4.3.1 Gene expression phylogenies

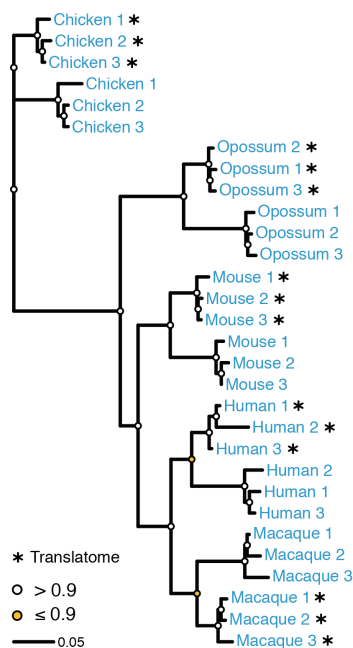
Functionally related groups of genes (modules) may change expression in a concerted manner to preserve ancestral cellular stoichiometries (Wagner et al., 2007; Lalanne et al., 2018). To trace the extent to which the predominantly compensatory evolution of individual genes (see above) is complemented by concerted gene expression changes to shape global patterns of gene expression evolution, expression distance matrices for each organ were built (Brawand et al., 2011), which formed the basis of gene expression trees and an expression level heatmap (Figure 4.1 and Figure 2.13). All trees recapitulate the known mammalian phylogeny; that is, they resolve the three major mammalian lineages (placentals, or eutherians; marsupials; and monotremes) and, within eutherians, group the two primates separately from the rodent (mouse) (Figure 4.1). Consistent with previous transcriptome analyses (Brawand et al., 2011), this suggests that regulatory changes at both expression layers steadily accumulated over evolutionary time, such that present-day RNA abundance levels and protein synthesis rates reflect the evolution of mammalian lineages and species. The second notable pattern is the distinct clustering of the transcriptome and translome data within clades defined by the corresponding species (Figure 4.1). These observations are supported by the trees based on perfectly aligned regions of the set of 6,327 1:1 orthologs (Supplementary Figure 14).

The only exception is the primate brain, where differences in expression between species are even smaller than the differences between the transcriptome and translome. This is likely explained by the slow gene expression evolution of the brain and the relatively short divergence time between the two primates, which makes the difference between the transcriptome and translome data slightly exceed that between the two species for each data type in this tree analysis. However, in trees based on larger numbers of 1:1 orthologs, including more recent genes, the two data types cluster by primate species (Figure 4.2).



**Figure 4.1: Mammalian gene expression (translatome and transcriptome) phylogenies**

(A to C) NJ trees based on pairwise expression distance matrices (1 - Spearman's  $\rho$ ) for 5,361, 4,630 and 5,237 robustly expressed (median FPKM > 1 across RNA-seq libraries) 1:1 amniote orthologs in brain, liver and testis, respectively. Asterisks (\*) at terminal nodes indicate Ribo-seq libraries; the other terminal nodes represent RNA-seq libraries. The divergence in gene expression levels recapitulates the phylogeny of the associated species. Gene expression levels are constrained across species, especially at the translational level, but show tissue-specific variation in the degree of constraint. Bootstrap values (all 1:1 orthologs were randomly sampled with replacement 1,000 times) represent the proportions of replicate trees supporting the branching pattern of the majority-rule consensus tree indicated by circles: white, > 0.9; yellow,  $\leq$  0.9.

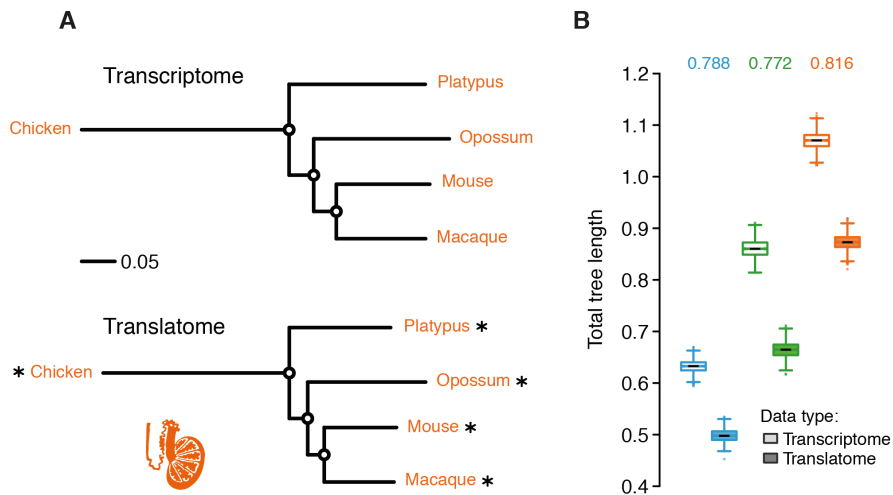


**Figure 4.2: Mammalian gene expression phylogeny for brain**

Human and macaque data are separated by species in the tree constructed based on the set of 9,085 1:1 orthologs shared between five species without platypus. (See Figure 4.1 for more legend details).

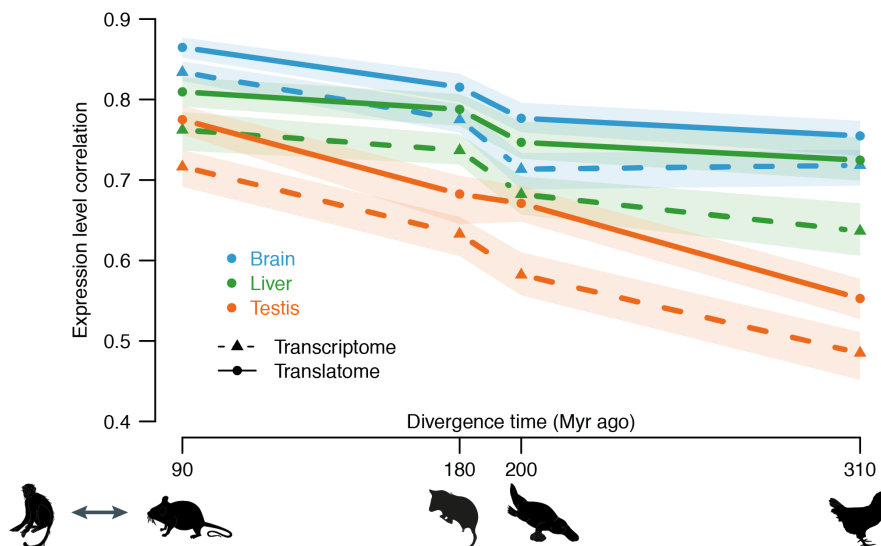
### 4.3.2 Global patterns of gene expression conservation

In line with the TTI analyses of individual genes and the notion that gene expression evolution is slowed down by compensatory changes at the level of translation, the lengths of internal branches in the translato-me data trees are overall shorter than those defined by the transcriptome data (Figure 4.1). To explore this observation in more detail, separate trees for the transcriptome and translato-me data were generated, and their total branch lengths quantified in these trees (Figure 4.3A, Section 4.2.2). These analyses revealed that the branches in the translato-me trees are approximately 18-23% shorter than those in the transcriptome trees (Figure 4.3B). Pairwise comparisons further illustrate that the translato-me has been substantially more preserved than the transcriptome during evolution (Figure 4.4). Notably, the extent of translational compensation is comparable to the marked differences in expression divergence rates between organs (Figure 4.3B and Figure 4.4).



**Figure 4.3: Total tree length analyses for translatomes and transcriptomes**

(A and B) For each organ the total tree length reflects the divergence along the internal (shared) branches leading to each species (See Section 4.2.2 for more details) (human data were excluded from these analyses because the single high-quality library for liver precludes internal branch length estimation). In (A), trees for testis are shown. (B) Comparisons of total tree lengths between organs for the five species. The boxplots (non-filled boxes: transcriptomes, filled: translatomes) reflect bootstrapping values: all genes were randomly sampled with replacement 1,000 times; the ratio of translatome to transcriptome tree lengths was calculated each time; for each organ, the median value of the 1,000 ratios was shown on top of the two boxplots.

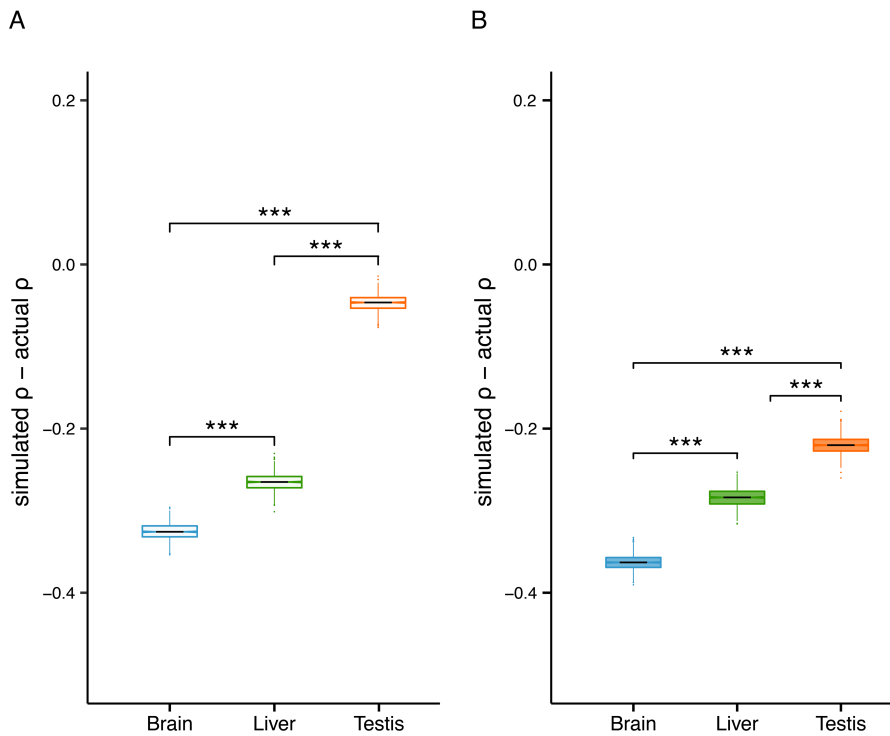


**Figure 4.4: Expression correlation between macaque and the other four species**

Expression correlation (Spearman’s  $\rho$ ) between macaque and the other four species across organs for both gene expression layers. Colored envelopes show ranges of values obtained in 100 bootstrap replicates.

### 4.3.3 Modular compensation

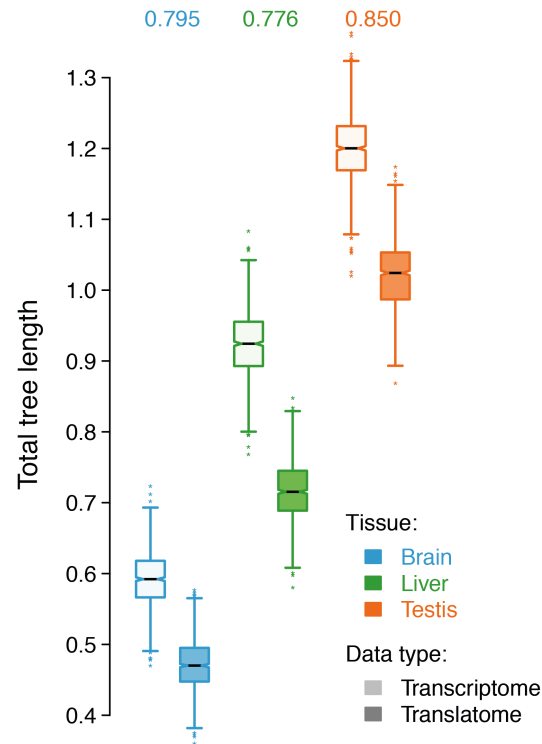
Contrary to the TTI analyses of individual genes in chapter 3, where brain and liver mostly show substantially less compensation than testis (Figure 3.2, B to D and Supplementary Figures 2 to 7), the extent of compensation in the tree analyses, which takes into account how genes change relatively to each other, is similar for the three organs (Figure 4.3B). Given that the various types of genes in the genome interact in various ways (e.g., they function together in the same pathway, regulate each other, and/or are co-regulated) to determine cellular and global tissue functions of a species, we hypothesized that the observed pattern is caused by higher modularity of gene expression changes in the somatic organs when compared to testis. To assess and compare the extent of such modular gene expression changes in the three organs, a simulated evolutionary scenario in which gene expression changes occur independently from each other was considered (Section 4.2.3). This analysis revealed a substantial reduction in correlation coefficients between the simulated datasets compared to the actual data, indicating the presence of modular expression changes in the three organs (Figure 4.5). Consistent with the hypothesis, these reductions are significantly more pronounced in brain and liver than in testis (Figure 4.5), suggesting that gene expression levels changed in an overall more concerted way in the two somatic organs than in testis.



**Figure 4.5: Comparison of correlations of expression levels between actual and simulated data**

Simulated evolutionary scenarios in which gene expression levels change independently from each other show decreased correlations (Spearman's correlation coefficients ( $\rho$ )) between mouse and opossum data. Spearman's  $\rho$  was simulated (see Section 4.2.3 for more method details) 1,000 times for the transcriptomes (A) and translomes (B) of brain, liver and testis, respectively; each time, the Spearman's  $\rho$  was compared to the corresponding Spearman's  $\rho$  calculated based on the actual data. Correlations were calculated for 1:1 orthologous (between macaque, mouse, opossum, and chicken) gene sets robustly expressed (median FPKM between 1 and 500) in liver (5,660 sets), brain (7,241 sets), and testis (6,331 sets). For both data types tissues differ significantly according to a Mann-Whitney  $U$  test (\*\*\*,  $p$ -value < 0.001).

The observations imply that even if individual genes show no signs of translational buffering, they might still contribute to overall gene expression conservation via concerted evolution. Indeed, when genes with near zero TTI values are considered (i.e., without individual compensation), we still observe at least 15% shorter branch lengths in translome than compared to transcriptome trees (Figure 4.6). Notably, modular translational buffering is significantly stronger in liver and brain than in testis (Figure 4.6). Thus, while the overall extent of compensatory evolution at the protein synthesis level is similar across organs, translational buffering of individual genes is more widespread in testis, whereas modular buffering is stronger in the two somatic organs.



**Figure 4.6 Total tree length analysis for the genes with near-zero TTI**

Correlation-based total tree length analysis for the genes with near-zero TTI indicates modular compensation. 600 1:1 orthologs with near-zero TTI were extracted from the mouse and chicken comparisons (Figure 3.2, B to D) for brain, liver and testis: namely 300 sets with TTIs greater/smaller than (and the closest to) 0, respectively. (See Figure 4.3 and Section 4.2.2 for more method details).





# Chapter 5 Patterns of expression divergence and compensatory evolution across gene classes

## 5.1 Introduction

The analyses in this chapter aim to further dissect and understand the genomic source of the widespread translational compensation. Specifically, we seek to contrast the evolutionary dynamics of transcriptomes and translomes across different gene types that we hypothesize to show unusual translational buffering patterns, or whose expression evolution was previously shown or predicted to have been subject to specific (strong) selective regimes.

Gene duplication is a major source of phenotypic novelty (Ohno, 1970), and evolution tinkers often with duplicate genes rather than *de novo* genes (Jacob, 1977; Magadum et al., 2013; Guschanski et al., 2017). Detrimental effects resulting from excessive gene dosage, on the other hand, have been the main driving force for the birth of dosage buffering mechanisms (Stenberg and Larsson, 2011). Ohnologs (paralogous gene pairs generated by whole genome duplication (WGD)) (Ohno, 1970) are enriched for dosage-sensitive genes, which is supported by the observations that most ohnologs are unduplicated even in lineages that diverged prior to the WGD event (Maere et al., 2005) and that ohnologs are rarely observed in copy number variants (CNVs, genomic regions that are duplicated or deleted in some individuals of a population) in healthy individuals (Makino and McLysaght, 2010).

CNVs are the most abundant kind of genetic variation per base-pair (Conrad et al., 2010) and have been characterized in many species, especially in humans (Conrad et al., 2010; Rice and McLysaght, 2017b). Although complete depletion of some genes results in no phenotypic changes (Zarrei et al., 2015), many CNVs are associated with human disorders, most notably in brain (Walsh et al., 2008; Cooper et al., 2011; Stefansson et al., 2014). Trisomies, a type of aneuploidy, are chromosomal duplication events that can be conceptually seen as large CNVs. For example, Down syndrome is a genetic disorder caused by an extra chromosome 21 (Antonarakis, 2017). Human chromosome 21 has a depletion of Ohnologs (used as a proxy for dosage sensitivity), which might explain why this chromosome has the most common (least severe) trisomy (Makino and McLysaght, 2010).

Gene expression is often noisy (Munsky et al., 2012), and genes have different levels of tolerance to their dosage changes (Rice and McLysaght, 2017a). Haploinsufficiency occurs when only half of the biologically active form is expressed (Fisher and Scambler, 1994; Bartha et al., 2018). For a subset of genes in the human genome an alteration in gene dosage caused by heterozygous loss-of-function mutations is usually implicated in many diseases, including heart disease, cancers and neuropsychiatric disorders (Glessner et al., 2014; Craddock et al., 2010; de Clare et al., 2011). For example, a recent study demonstrated that human *RELA* haploinsufficiency is linked to autosomal-dominant chronic mucocutaneous ulceration (Badran et al., 2017), and the syndrome of *BACH2*-related immunodeficiency and autoimmunity is associated with haploinsufficiency of *BACH2* (Afzali et al., 2017).

It has also reported that protein complex subunits tend to exhibit gene-dosage sensitivity (Papp et al., 2003). Many haploinsufficient genes in *Saccharomyces cerevisiae* encode subunits of protein complexes (Deutschbauer et al., 2005), and their dosage balance has to be tightly regulated to produce the right amount of complete and active protein complexes (Cardarelli et al., 2011; Veitia and Birchler, 2015). For protein complexes that are stringently regulated in a stoichiometry-dependent manner, even a transient disruption to their relative ratios is linked to noticeable

consequences (Veitia and Birchler, 2010 & 2015). For example, at the translational level (e.g., via adjusting TE), Li et al. (2014) revealed that subunits of FoF1 ATP synthase complex are synthesized in proportion to their stoichiometry. At the post-translational level (e.g., through protein degradation), excess subunits caused by various perturbations were found to be buffered to maintain cellular robustness and phenotypic stability (Ishikawa et al., 2017). Thus, protein complex stoichiometry seems to constrain protein level variation more strongly than RNA abundance changes.

Haploinsufficiency is also regarded as one of the major proxies for gene essentiality, for example, ~3,000 human genes cannot tolerate heterozygous loss-of-function variants, which lead to loss of one of the two alleles and to haploinsufficiency of their respective genes (Lek et al., 2016). Representative genome-wide matrices for the measurement of haploinsufficiency have been devised by Dang *et al.* (2008), Khurana *et al.* (2013), Steinberg *et al.* (2015) and Shihab *et al.* (2017). Several other metrics have also been developed based on the Exome Aggregation Consortium (ExAC) dataset of 60,706 human exomes (Lek et al., 2016) to score gene essentiality, and these scores are highly correlated with one another (Bartha et al., 2018). The fundamental principle of most of the scores is to rank genes according to the strength of purifying selection acting against protein-truncating variants (Bartha et al., 2018). The most used metric of gene essentiality is the probability of being loss-of-function intolerant (pLI score) (Lek et al., 2016; Bartha et al., 2018).

Housekeeping genes are broadly expressed genes that are instrumental in maintaining fundamental cellular functions across tissues of an organism (Eisenberg and Levanon, 2013). The distinct genomic, structural, and evolutionary properties of housekeeping genes compared to tissue-specific genes make them interesting gene types to understanding various aspects of gene expression evolution (Eisenberg and Levanon, 2013), for example, by contrasting the levels of gene expression constraints imposed on housekeeping and tissue-specific genes.

The age of gene origin is also associated with gene expression constraints given that it affects the level of a gene integration into the functional cellular environment (Vishnoi; 2010; Capra et al., 2013;). Compared with younger genes, older genes tend to interact with more transcriptional factors, have more preserved upstream sequences, and house more potential miRNA targets (Warnefors and Eyre-Walker, 2011). Consistent with this notion, the gene connectivity in co-expression networks, gene involvement in complex regulatory networks, gene haploinsufficiency were also found to increase with gene age (Popadin et al., 2014; Rice and McLysaght, 2017a).

Given that the evolution of aforementioned gene classes was previously shown or predicted to have been subject to specific (strong or weak) selective forces, I collected their respective gene sets to study the evolution and patterns of translational buffering.

## 5.2 Methods

### 5.2.1 Resources for gene class annotations

The gene sets underlying the different classes analyzed were retrieved from various sources. I obtained the set of mouse protein complex subunits through the gene ontology term GO:0043234 from org.Mm.eg.db, a dedicated R package for genome-wide gene annotation for mouse (Carlson, 2018). Mouse protein–protein interaction data were downloaded from BioGRID V3.4.156 (Chatr-Aryamontri et al., 2017). Ohnologs (strict set) for mouse were downloaded from the database of Vertebrate Ohnologs (<http://ohnologs.curie.fr/>, (Singh et al., 2015)). Haploinsufficiency scores from Shihab et al. (2017) were used as proxies of the extent of haploinsufficiency for human genes. Human scores were projected to mouse 1:1 orthologs when using mouse as the focal species in specific analyses. Mouse scores of the set of 9,325 1:1 orthologs for the representative species (i.e., macaque, mouse, opossum and chicken) were first ranked from the largest to the smallest values and then used to define two gene subsets (first/last quartile) that were defined as sensitive or insensitive to haploinsufficiency, respectively. Gene essentiality was

defined based on the probability of being loss-of-function intolerant; that is, the pLI score (Lek et al., 2016); the score data were obtained from ExAC release 0.3.1 (<http://exac.broadinstitute.org/>). Finally, the phylogenetic duplication age of each mouse gene was retrieved from a parallel study from our lab (Cardoso-Moreira et al., under review). In that study, on the basis of genomic annotations from Ensembl 69, gene duplication age was assigned based on syntenic alignments across vertebrates and parsimony as previously described (Zhang et al., 2010). Given that the analyses that consider gene duplication age are based on shared 1:1 orthologs among the four representative species (macaque, mouse, opossum and chicken), I focused the age analyses on orthologs that emerged by duplication in the amniote or tetrapod ancestors (genes defined as relatively young) and orthologs that emerged before (i.e., ancestors of jawed vertebrates).

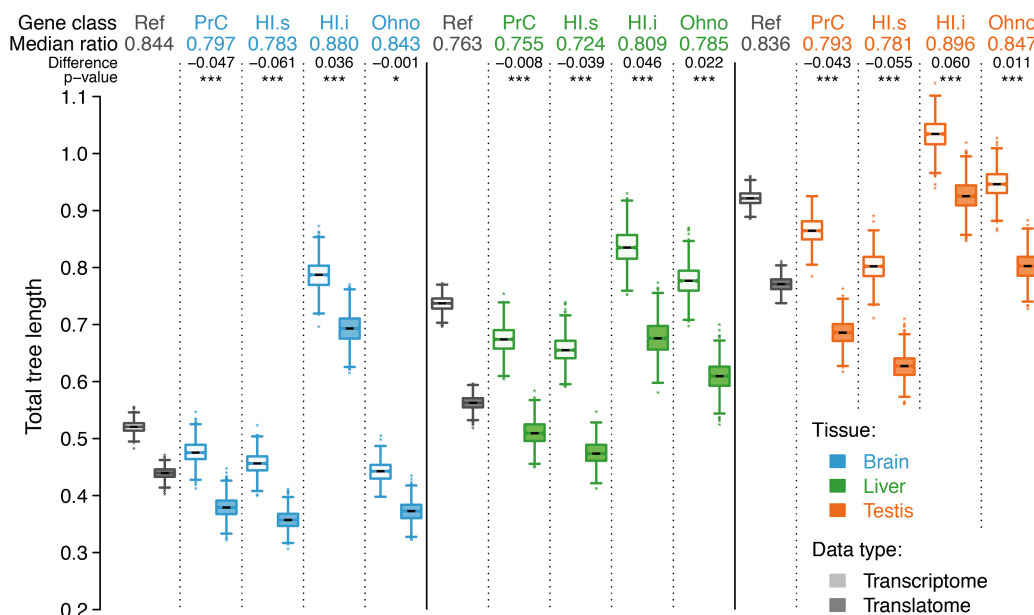
### 5.2.2 Tissue specificity index

To define the sets of both housekeeping and tissue-specific genes for mouse, I relied on the RNA-seq data for five adult mouse tissues (i.e., brain, heart, liver, kidney and testis) obtained from a parallel study (Cardoso-Moreira et al., under review). I did not include cerebellum in this analysis, because including both brain (prefrontal cortex) and cerebellum would reduce the number of brain-specific genes due to the frequently shared gene expression profiles in those two tissues. I assessed gene expression breadth using the tau ( $\tau$ ) tissue specificity index (Yanai et al., 2005). Genes with  $\tau \leq 0.2$  and  $\tau \geq 0.7$  were defined as housekeeping and tissue-specific, respectively.

## 5.3 Results

### 5.3.1 Dosage sensitivity and compensatory evolution

I found that genes that are generally dosage sensitive and/or haploinsufficient show lower evolutionary expression divergence at both the transcriptome and translome level and show significantly stronger buffering than the average gene in the genome (the genomic background) (Figure 5.1), consistent with the functional importance of fine-tuned expression levels for such genes (Rice and McLysaght, 2017a). For example, genes encoding proteins that are assembled into protein complexes and therefore are sensitive to stoichiometrical perturbations (Rice and McLysaght, 2017a) show low expression divergence across both expression layers as well as pronounced translational buffering (Figure 5.1 and Supplementary Figure 15). Generally, genes with high haploinsufficiency scores show strong expression conservation coupled with pronounced translational buffering (Figure 5.1). Conversely, genes with low haploinsufficiency scores show much higher gene expression divergence and less translational compensation compared to high-scoring genes or the genomic background (Figure 5.1). I observed similar patterns across organs, but at different levels of expression divergence that correspond to the overall organ-typical evolutionary rates of expression divergence (Necsulea and Kaessmann, 2014a; Brawand et al., 2011) (Figure 5.1). I also investigated a specific set of dosage sensitive genes (so-called Ohnologs) that duplicated as part of two whole-genome duplication events in the vertebrate ancestor. Consistent with their presumably mainly neural functions (Singh et al., 2015; Guschanski et al., 2017; Roux et al., 2017), I found strong expression conservation and moderate translational buffering for Ohnologs in the brain, whereas in the liver and testis they evolve rapidly at all levels. Altogether, these analyses reveal a strong positive correlation between dosage sensitivity and the extent of evolutionary preservation of gene expression levels.



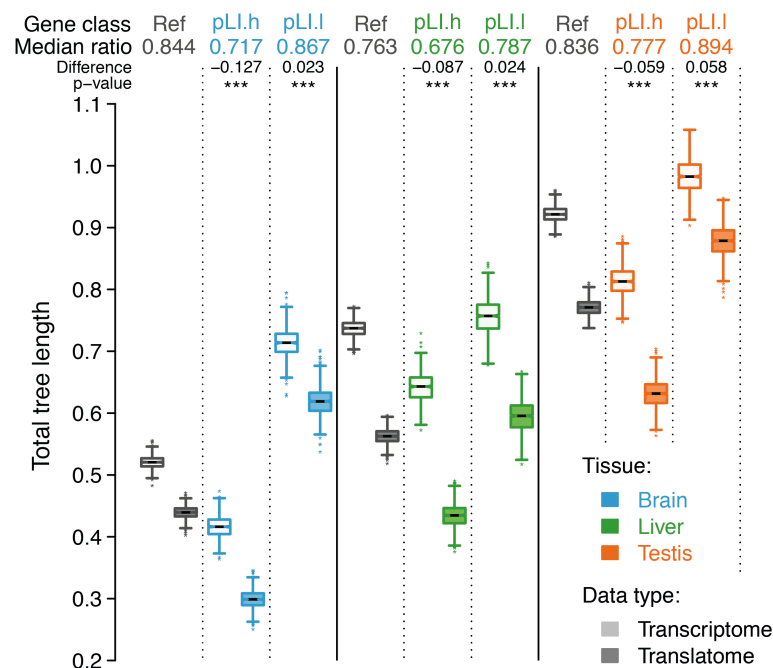
**Figure 5.1: Compensatory evolution of dosage-sensitive and -insensitive genes**

The procedure described in Figure 4.3 and Section 4.2.2 was repeated for each gene class. Distributions of values calculated for all genes are used as references (grey). Abbreviations: Ref, robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs among the four representative species (macaque, mouse, opossum, and chicken) for the respective tissue; PrC, protein complex subunits; HI.s/HI.i, haploinsufficiency sensitive/insensitive genes; Ohno, Ohnologs. Human and platypus were excluded from these analyses because the lower number of replicates for humans and the relatively poor genome (annotation) quality of platypus would have limited the number of 1:1 orthologs in the analyses. The median value of 1,000 ratios between translome and transcriptome tree lengths are indicated for each gene class. The difference between the median ratio of a particular gene set and that of its corresponding reference is also indicated (negative/positive values suggests that a given gene set is overall more/less translationally buffered than genes on average in the genome. Mann-Whitney  $U$  tests are employed to estimate whether a difference is statistically significant; Benjamini-Hochberg-corrected p-values: \*\*\*, < 0.001; \*\*, < 0.01; \* < 0.05; ns (not significant),  $\geq 0.05$ .

### 5.3.2 Gene essentiality and compensatory evolution

I then gauged the relationship between the phenotypic impact of a gene (i.e., how essential its function is for organismal fitness) and the patterns of evolution across both expression layers. To do so, I leveraged several recently established metrics of mutational tolerance (typically within coding sequences) across the genome (Bartha et al., 2018). My analyses revealed a strong relationship between gene essentiality and expression evolution in the three organs; that is, genes

that are highly sensitive to mutations (essential genes) show lower expression divergence together with stronger translational buffering compared to the genomic background. I observed the opposite pattern for genes that are particularly tolerant to mutations (Figure 5.2 and Supplementary Figure 15). It is noteworthy that dosage sensitivity has emerged as a key property of gene essentiality (Bartha et al., 2018). Thus, measures of dosage sensitivity and mutational tolerance of the coding sequence are highly correlated (Bartha et al., 2018), which means that, for example, haploinsufficient genes will not only be selectively constrained in terms of their expression evolution but will typically also be sensitive to coding sequence mutations.



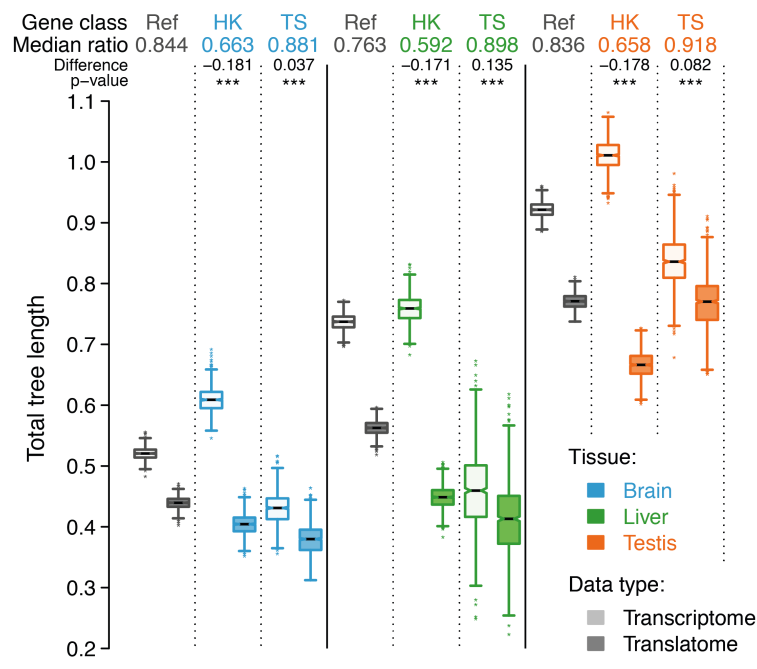
**Figure 5.2: Compensatory evolution of essential and nonessential genes**

The procedure described in Figure 4.3 and Section 4.2.2 was repeated for each gene class. Distributions of values calculated for all genes are used as references (grey). Abbreviations: Ref, robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs among the four representative species (macaque, mouse, opossum, and chicken) for the respective tissue; pLI.h/pLI.l, genes with high/low probability of being loss-of-function intolerant. See Figure 5.1 for more legend details.



### 5.3.3 Spatial expression characteristics and compensatory evolution

I next assessed the relationship between spatial expression characteristics and expression evolution. Notably, I observed that broadly expressed genes and genes previously defined as "housekeeping" (Eisenberg and Levanon, 2013) diverge rapidly at the transcriptome level in all organs, consistent with previous work (Breschi et al., 2016 & 2017), but that this high transcriptome divergence is counterbalanced by the by far strongest translational buffering of all gene categories, which reduces the extent of translome divergence to levels that are similar to (in brain and testis) or even lower (in liver) than those of other constrained gene classes (Figure 5.3 and Supplementary Figure 15). Thus, the broadly expressed (housekeeping) genes also have similar (brain and liver) or lower translome divergence (testis) values than genes expressed specifically or predominantly in one organ (tissue-specific genes), in spite of their rapid transcriptome evolution (Figure 5.3). Notably, in striking contrast to housekeeping genes, tissue-specific genes show the lowest amount of translational buffering of all gene categories (Figure 5.3). My observations are consistent with the low promoter sequence conservation (Farré et al., 2007) and overall simple transcriptional control of housekeeping genes (e.g., they have few enhancer sequences; (Osterwalder et al., 2018)), which may imply less tight transcriptional regulation. They are also in line with the importance of translational regulation and strong translational selection inferred for housekeeping genes in a study of codon usage bias (Ma et al., 2014). Thus, high selective pressures likely drove pronounced translational compensation of housekeeping genes during evolution, potentially facilitated by sophisticated translational regulatory mechanisms. By contrast, the expression evolution of tissue-specific genes seems to have been predominantly shaped by selection affecting transcription.



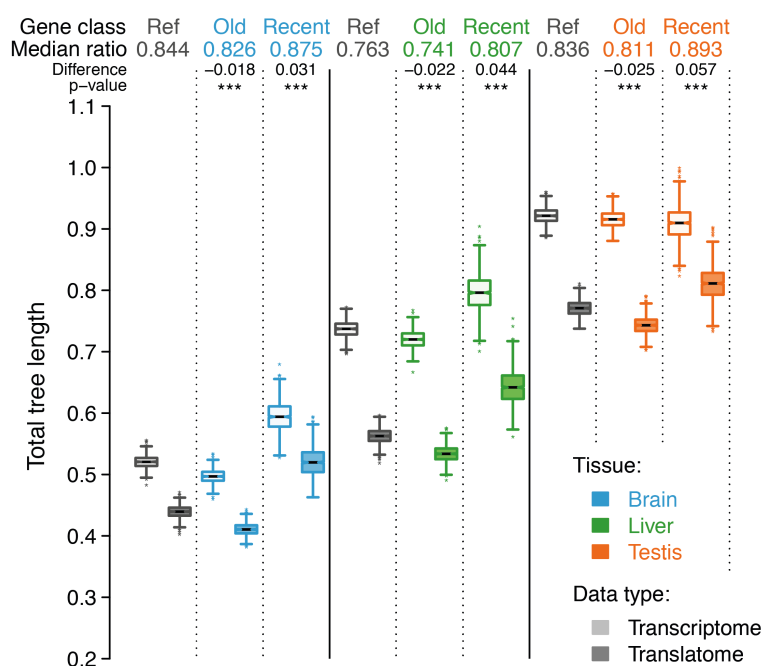
**Figure 5.3: Compensatory evolution of housekeeping and tissue-specific genes**

The procedure described in Figure 4.3 and Section 4.2.2 was repeated for each gene class. Distributions of values calculated for all genes are used as references (grey). Abbreviations: Ref, robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs among the four representative species (macaque, mouse, opossum, and chicken) for the respective tissue; HK, housekeeping genes; TS, tissue-specific genes. See Figure 5.1 for more legend details.

### 5.3.4 Gene age and compensatory evolution

Finally, it has been hypothesized that genes of different ages may show differential divergence dynamics across the two expression layers. For instance, given that new (duplicate) genes are typically functional in later, less constrained developmental stages (Cardoso-Moreira et al., under review) and are less likely to be essential compared to older genes (Chen et al., 2012), they could be expected to show less constrained and less buffered gene expression change. Indeed, when amniote 1:1 orthologs in my analyses are stratified into genes that originated by duplication in amniote or tetrapod ancestors and genes with older vertebrate origins, I found that the older genes show overall lower gene expression divergence than genes in the younger category, and that — contrary to younger genes — older genes show stronger translational buffering than the

genomic background (Figure 5.4). It is also in agreement with the stronger translational selection inferred for old genes than young genes based on another study of codon usage bias (Yin et al., 2016).



**Figure 5.4: Compensatory evolution of old and recent genes**

The procedure described in Figure 4.3 and Section 4.2.2 was repeated for each gene class. Distributions of values calculated for all genes are used as references (grey). Abbreviations: Ref, robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs among the four representative species (macaque, mouse, opossum, and chicken) for the respective tissue; Old/Recent, genes that duplicated in common bony fish ancestor/genes with duplication origins in tetrapod ancestor. See Figure 5.1 for more legend details.



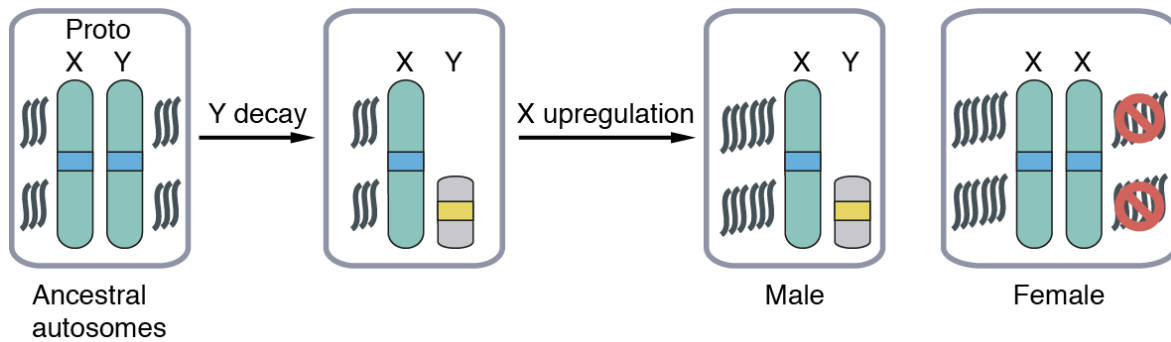
# Chapter 6 X chromosome dosage compensation through translational upregulation

## 6.1 Introduction

Motivated by the pronounced translational buffering observed across organs and gene classes, we explored whether this mechanism has contributed to the compensation of the chromosome-wide dosage reduction due to the massive gene loss on the Y chromosome during sex chromosome evolution (Graves, 2016).

The study of sex determination in vertebrates dates back to 1947, when Alfred Jost published that primary sex determination involves the decision in the gonad to differentiate as a testis or an ovary on the basis of his famous experiments on rabbits (Jost, 1947; Capel, 2017). Many follow-up studies made it clear that male and female therian mammals (placentals and marsupials) share the same complement of autosomes but differ in their sex chromosomes: females carry two X chromosomes (XX) while males bear one X chromosome and one Y chromosome (XY) (Figure 6.1). In contrast, birds have a ZW sex-determination system, wherein females are the heterogametic sex (ZW) and males are the homogametic sex (ZZ). The sex chromosomes of therians are thought to have originated ~180 million years ago from a pair of autosomes, and this notion has been supported by comparative genomics studies that showed that genes on the human X are autosomal in birds (Ezaz et al., 2006; Smith and Voss, 2007), reptiles (Ezaz et al., 2009), and even in egg-laying monotremes (Veyrunes et al., 2008). Further evidence showed that some marsupial autosomal genes are orthologous to genes on the X chromosome (X-linked genes) in human, which suggests that they were only added to the X after the split of placental and marsupial

mammals ~90 million years ago (Graves, 1995). Likewise, genes on the chicken Z chromosome were also found to be orthologous to genes on human chromosome 9 (Nanda et al., 1999), suggesting that sex chromosomes from chickens (birds) and humans (therians) evolved in parallel.



**Figure 6.1: The evolution of therian sex chromosomes and dosage compensation**

Sex chromosomes differentiated from ordinary chromosomes. A potential first step was the acquisition of a sex-determining region Y (*SRY*), which was followed by the accumulation of sexually antagonistic mutations that suppressed the proto-sex chromosomes from recombination. A lack of recombination caused the decay of Y-linked genes (pseudogenization), which led to the accumulation of repetitive DNA on the Y chromosome. It is hypothesized that the decay of the Y chromosome triggered the evolution of dosage compensation on the X chromosome in therians: the only copy of the X chromosome in males doubled its expression while in females one copy of the X was inactivated while the other doubled its expression.

It was proposed in 1967 by Susumu Ohno that imbalanced copy numbers of the X chromosome between males and females following the degeneration of the Y from ancestral autosomes triggered the evolution of X dosage compensation in mammals (Figure 6.1) (Ohno, 1967; Charlesworth, 1978; Disteche et al., 2012; Bachrog, 2013; Brockdorff and Turner, 2015; Graves, 2016; Capel, 2017). That is, to compensate for the initial two-fold reduction of the transcriptional output from the remaining single X in males, X-linked genes were hypothesized to have evolved two-fold higher expression levels, thereby restoring the ancestral transcript levels of the X in males while also maintaining the balance between X-linked and autosomal gene expression in this sex (Ohno's hypothesis) (Ohno, 1967). The resulting overabundance of X-linked transcripts in females, resulting from the combined activity of the two upregulated X chromosomes, was then compensated by the well-known process of X chromosome inactivation (Figure 6.1).

Early microarray data showed X upregulation relative to autosomes (Gupta et al., 2006; Nguyen and Disteche, 2006). However, later evolutionary transcriptome studies made it clear that, globally, X-linked genes only have approximately half of the ancestral expression output in eutherians and thus lack global transcriptional compensation, whereas marsupials, which have the same sex chromosome system (Graves, 2016), do show signatures of widespread upregulation, at least in some organs (Julien et al., 2012; Necsulea and Kaessmann, 2014a). Other studies suggested that Ohno's hypothesis still holds true in eutherians when some genes are discarded in the analysis (Deng et al., 2011; Lin et al., 2011; Kharchenko et al., 2011), but the logic behind the removal of those genes has been challenged (He et al., 2011). While individual genes may not necessarily be upregulated in placental mammals, Julien *et al.* (2012) detected an alternative mechanism of compensation for the two-fold expression reduction of a subset of genes, i.e. a two-fold expression reduction of the autosomal partner genes. Furthermore, some dosage-sensitive genes may have been relocated to autosomes, especially the members of large protein complexes (Bellott et al., 2014; Hughes et al., 2015; Bellott et al., 2017).

Notably, sex chromosome differentiation also triggered the emergence of complete meiotic silencing of sex chromosomes (MSCI) in the male germline (Turner, 2015), and it is clear that mechanisms must have evolved to compensate for this lack of active X-linked gene transcription in the testis during meiosis (Turner, 2015). So far, one mechanism — the generation of autosomal substitute gene copies — has been discovered, which nevertheless only compensates for the silencing of a limited number of key X-linked genes (Turner, 2015).

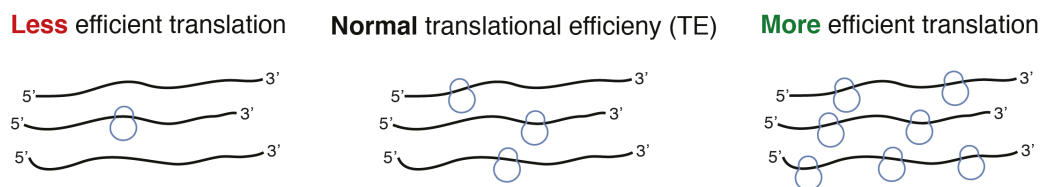
However, all the studies mentioned above studied dosage compensation at the RNA level, which is not a good predictor of protein levels. Chen and Zhang (2015) compiled human proteome and transcriptome data for 22 tissues from several sources to test Ohno's hypothesis and found X dosage compensation to be also absent at the protein level. However, in this study, Ohno's hypothesis was only tested indirectly for each human tissue by assessing the ratio of X-linked and autosomal median expression values. A more direct way would be comparing therian X-linked

genes (current X) with their 1:1 orthologs of an evolutionary outgroup (e.g., chicken) (ancestral proto-X) (Figure 6.1). In addition, their proteome data was of heterogeneous quality and low resolution comparing with RNA-seq data.

Ribo-seq, which generates translome data of much higher data quality and resolution than that of proteome data, also enables the comparisons of ribosome density between X-linked and autosomal genes. Based on mouse cell line Ribo-seq data, Faucillion and Larsson (2015) found that X-linked genes show significantly higher ribosome density than autosomal genes, which to some extent supports Ohno's hypothesis. However, this is again only based on an indirect comparison of expression levels between the present-day X and the present-day autosomes. Furthermore, cell lines might not properly represent the biology of that of primary organs. Thus, the X dosage compensation model in eutherians has so far not been rigorously tested.

In this chapter, I assessed the possibility of a X chromosome-wide form of translational compensation in several mammalian organs. This allows for assessment of whether there are additional mechanisms (increase of TE) (Figure 6.2). I thus used Ribo-seq data to assess whether upregulation has occurred by evaluating if the rate of protein synthesis of X-linked genes overall increased after sex chromosome differentiation to compensate (partially or completely) for the two-fold reduction in gene dose. To do this, I sought to compare translation rates of X-linked genes in therian samples to those from orthologs in an outgroup species with a different sex chromosome system (i.e., chicken), where these genes are located on autosomes, akin to previous studies (Marin et al., 2017; Julien et al., 2012). Given that the autosomal orthologs from outgroup species have been unaffected by sex-related selective forces, they may serve to gauge expression levels of proto-sex chromosomes (i.e., the ancestral autosomes, prior to their differentiation into sex chromosomes) (Necsulea and Kaessmann, 2014a). Thus, for example, a similar translation output from (single) X-linked genes in placental mammals as from their (two) autosomal orthologs in chicken would be indicative of an increase of ribosome occupancy on the X.





**Figure 6.2: Illustration of translational efficiency (TE)**

A comparison between the rate of protein synthesis and the level of mRNA expression makes it possible to determine the TE for each mRNA. The more efficient the translation, the more proteins are produced per RNA molecule.

## 6.2 Methods

### 6.2.1 Extraction of orthologous gene sets

Because Ohno's dosage compensation hypothesis concerns "old" genes that existed before the origin of the mammalian X, I focused on genes for each of the four therians (human, macaque, mouse and opossum) that have 1:1 orthologs in chicken (a close outgroup of mammals), and obtained 11,876, 10,732, 11,917, and 11,270 genes, respectively.

### 6.2.2 Normalization of current X and proto-X expression levels

Prior to any direct comparisons, raw X expression levels in the focal therian species (i.e., human, macaque, mouse or opossum) and chicken were normalized relative to their respective autosomal backgrounds. Briefly, for each library, X expression levels were normalized on the basis of a scaling factor that was derived from adjusting the median expression levels of autosomal 1:1 orthologs across all RNA-seq or all Ribo-seq libraries to a common value (i.e., each median value was divided by the mean of all median values). I next computed the median value of FPKMs averaged over biological replicates for all X-linked genes for each species. Finally, current X to estimated proto-X expression ratios were calculated with median values for each data type (i.e., median value of focal species divided by that of chicken). Statistically significant differences of the FPKMs averaged over biological replicates between species for the same data type were assessed using Mann-Whitney  $U$  tests.

### 6.3 Results

To evaluate to what extent translational buffering (i.e., upregulation) might have attenuated the X dosage reduction in eutherian somatic organs and/or the complete silencing through meiotic sex chromosome inactivation (MSCI) in meiotic cells, I compared, at both expression layers, current X expression levels in eutherians and marsupials with ancestral (proto-X) expression levels, inferred from expression levels of autosomal orthologs from the outgroup chicken, which has a different sex chromosome system (Necsulea and Kaessmann, 2014a; Julien et al., 2012).

These analyses revealed that current expression levels are more similar to ancestral levels on the translome layer than on the transcriptome layer (Figure 6.3), which is consistent with the notion that dosage compensation may not necessarily occur at the mRNA level, because ultimately it is the protein abundance that matters. The extent of compensatory upregulation varies between organs and species, with several instances of potentially full compensation. For example, in the human and opossum brain, current and ancestral expression levels are highly similar and statistically indistinguishable for the translome, whereas current transcript levels are substantially reduced relative to ancestral ones (Figure 6.3, A and D).

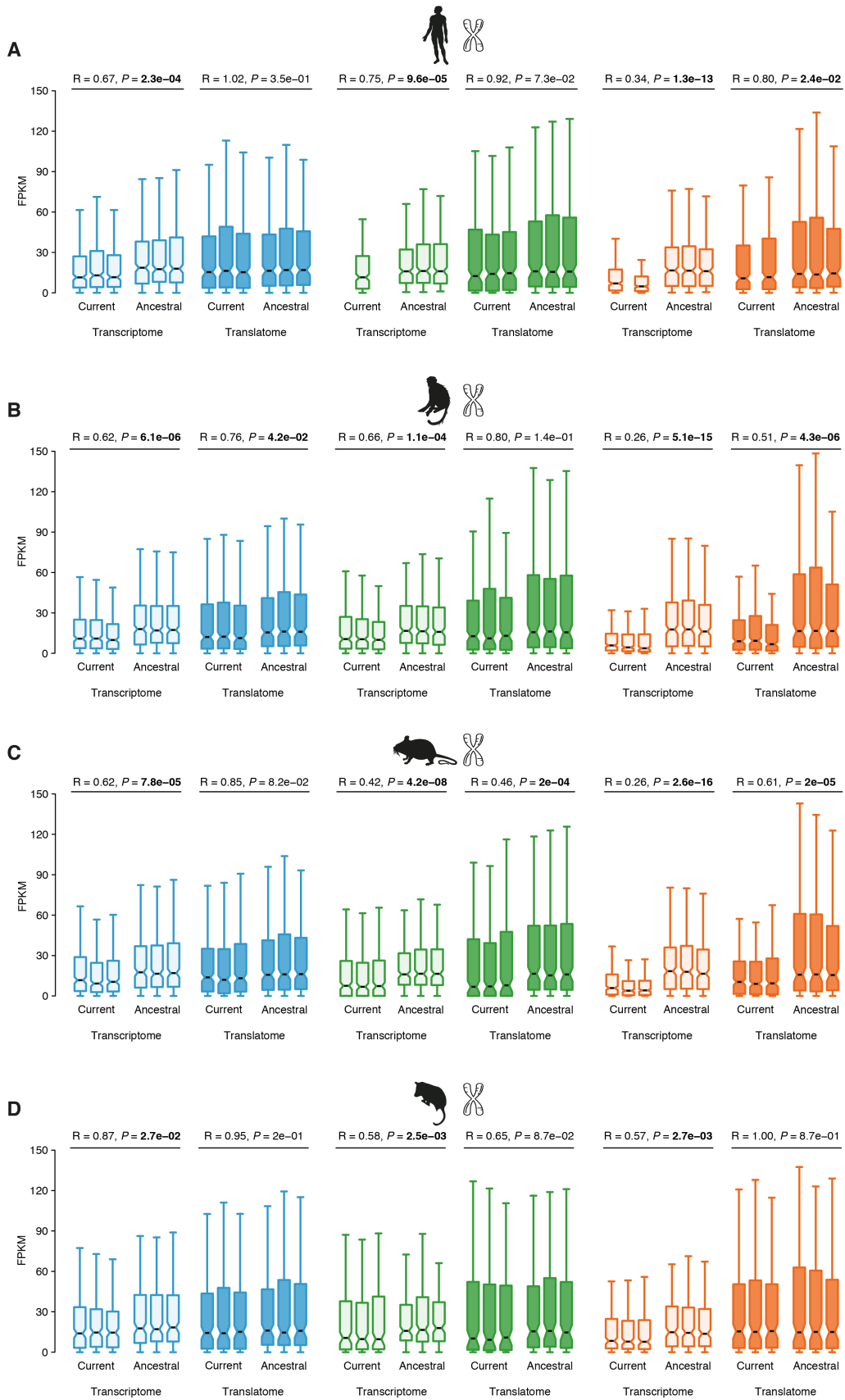
Notably, the by far strongest translational upregulation occurs in the testis, which is dominated by meiotic and post-meiotic cells, where MSCI exerts its effect (Figure 6.3). This suggests that for genes that are instrumental in maintaining normal cellular functions during meiotic and post-meiotic phases of spermatogenesis, their transcripts inherited from pre-meiotic phases underwent a strong translational upregulation to reach sufficiently high protein levels.

Consistent with these observations, I found that TEs are significantly higher on the X chromosome than on autosomes for each eutherian species in the somatic tissues, consistent with the aforementioned cell line study (Faucillion and Larsson, 2015), and are particularly high in the testis (Figure 6.4). It is noteworthy that since marsupials show signatures of chromosome-wide transcriptional upregulation in some somatic organs (including brain and liver) (Julien et al., 2012),

the overall TEs are not significantly different between the X and autosomes, as would be expected if dosage compensation is already achieved at the transcriptional level.

Further analyses, based on the TTI procedures, identified large proportions of individual translationally upregulated (compensated) genes (dark blue points in Figure 3.2, B to D and Supplementary Figures 2 and 7), which may represent particularly dosage sensitive (haploinsufficient) genes. They also highlight the strongly biased directionality of the X dosage (transcript abundance) reduction in mammals relative to the outgroup species (e.g., RNA LFC < 0 between mouse and chicken in Figure 3.2, B to D) and associated compensatory translational upregulation of individual X-linked genes, in particular in the testis (Figure 3.2, B to D and Supplementary Figures 2 and 7).

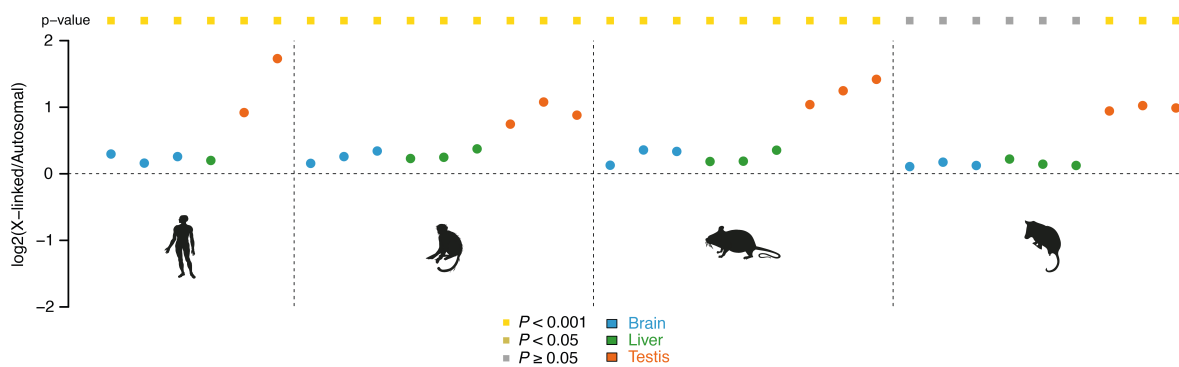
Overall, my analyses unveil an important role of translational upregulation in X chromosome dosage compensation in the soma of eutherians and marsupials as well as a role in counterbalancing MSCI.



(figure legend continued on next page)

### Figure 6.3: Current and ancestral levels of translation and transcription on the X chromosome

(A) Expression levels (transcriptome and translome, respectively) for robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs between human and chicken, the outgroup, which is used as a proxy for the ancestral (i.e., proto-sex chromosomal) expression levels. See Section 6.2.2 for details. (B to D) Same as in (A) but for robustly expressed 1:1 orthologs between macaque and chicken, between mouse and chicken, and between opossum and chicken, respectively. The median value of FPKMs averaged over biological replicates for all robustly expressed (median FPKM > 1) X-linked genes for each species was first calculated. The ratio (R) of current X to proto-X median expression levels is shown for the transcriptome (light boxplots) and translome (dark boxplots). Statistical significance of between-species differences of expression levels (i.e., FPKMs averaged over biological replicates) was assessed using Mann-Whitney  $U$  tests (p-values,  $P$ , for each test are indicated; values in bold indicate  $P < 0.05$ ).



### Figure 6.4: TE ratios of X-linked to autosomal genes across therians

Median X-linked to autosomal TE ratios of robustly expressed genes (median FPKM across RNA-seq libraries > 1) was calculated for each therian sample. For visualization purposes, ratios are plotted on a log<sub>2</sub> scale. Mann-Whitney  $U$  tests were employed to estimate the significance for the TEs between X-linked and autosomal genes; p-values,  $P$ : light yellow squares, < 0.001; dark yellow squares, < 0.05; grey squares, ≥ 0.05.



## Chapter 7 Discussion and Outlook

### 7.1 Discussion

In my thesis project, I explored gene expression evolution across the two major layers of transcription and translation, based on Ribo-seq and matched RNA-seq data across three major organs from representatives of all major mammalian lineages. My analyses uncovered that evolutionary changes in steady-state transcript abundance were frequently counterbalanced and rarely reinforced at the protein synthesis level, suggesting that many transcriptional changes were, at least initially (i.e., without translational compensation), unfavorable. Translational buffering has therefore overall substantially preserved ancestral gene expression outputs during mammalian organ evolution. The few instances of translational reinforcement represent interesting candidates for potentially adaptive (i.e., phenotypically relevant) expression changes.

Initial yeast hybrid work revealed a dominant role of translational buffering as well (McManus et al., 2014; Artieri and Fraser, 2014), although this conclusion was subsequently challenged based on analytical considerations (Albert et al., 2014; Bader et al., 2015), which also apply to a study reporting an excess of compensatory change in hybrid fibroblasts from two mouse strains (Hou et al., 2015). The only real previous between-species comparison in mammals, a comparative study of lymphoblastoid cell lines between human, chimpanzee, macaque across the three main expression layers (transcriptome, translome, proteome), found very little evidence of translational buffering (only 1 to 9 instances, depending on the species comparison) (Wang et al., 2018). The short evolutionary divergence time of the species covered in this study may have limited the emergence of compensatory mutations and/or the power to detect (potentially subtle)

translational changes. Consistently, I also found substantially fewer significant compensation events when I assess buffering in a slowly evolving organ (i.e., brain) from the two most closely related species in my thesis project (i.e., between human and macaque) (Supplementary Figures 7 and 13). Previous yeast work is also consistent with this notion, given that there is less evidence of translational buffering in hybrids of yeast strains from the same species compared to hybrid work based on different species (Schaefer et al., 2018).

My analyses also unveiled differential and apparently fine-tuned patterns of translational buffering across tissues, gene classes/ages and chromosomes. In addition to the translational buffering of individual genes, I discovered that gene expression divergence is overall attenuated by concerted (modular) translational changes of genes. This modular compensation mechanism likely serves to preserve ancestral stoichiometries of functionally interacting proteins, in agreement with the strong buffering of protein complexes (Figure 5.1) and a recent comparative transcriptome study of bacterial pathways (Lalanne et al., 2018). The relative prevalence of these two compensatory mechanisms and their contribution to gene expression preservation substantially differs between organs. Thus, the rapid transcriptome evolution in the testis, due to relaxed purifying selection and positive selection (Necsulea and Kaessmann, 2014a), is strongly offset by individual translational buffering and, probably to a lesser degree, by modular buffering. By contrast, modular buffering represents a more important evolutionary force in brain and liver, which generally show lower divergence rates than the testis with respect to both expression layers.

The extent of overall gene expression divergence and translational buffering also profoundly varies between different gene classes. For example, translational buffering substantially contributes to the generally high expression level conservation of genes that are dosage sensitive and/or essential for organismal fitness, sometimes with organ-specific patterns (e.g., Ohnologs in the brain). Broadly expressed (housekeeping) genes show a remarkable pattern; their rather rapid transcriptome evolution is compensated by the strongest translational buffering of all gene classes. My findings emphasize that higher gene expression layers than that of transcription need to



be considered in future work, for example, when predicting the suitability of the mouse as a model of human gene functions and diseases. Indeed, because of the pronounced transcriptome divergence of broadly expressed genes between human and mouse, it was suggested that the mouse may represent a poorer model for such genes than for more tissue-specific genes (Breschi et al., 2016 & 2017).

Finally, I uncovered translational upregulation as a novel mechanism to globally counterbalance the otherwise potentially detrimental effects of the dosage reduction that arose in the wake of mammalian sex chromosome differentiation. Together with other alternative mechanisms (e.g., downregulation of autosomal partners of X-linked genes, (Necsulea and Kaessmann, 2014a; Julien et al., 2012)), translational compensation may therefore have contributed to the emergence of X inactivation in females. This mechanism, which was originally hypothesized to have emerged due to excessive expression from the combined activity of the two upregulated X chromosomes, has remained enigmatic, given the lack of global transcriptional X upregulation (Necsulea and Kaessmann, 2014a; Julien et al., 2012). I note that a previous human proteome study failed to identify dosage compensation at higher expression layers (Chen and Zhang, 2015), but that work was based on indirect X-to-autosome (rather than on current-to-ancestral) comparisons and the quantitative resolution of proteome data remains limited, as also acknowledged by the authors. Furthermore, my work identifies translational upregulation as a novel and powerful mechanism to counteract the potentially hazardous effects of MSCI, another consequence of sex chromosome differentiation, and is therefore likely key for the functioning of the male germline.

Two potential, nonexclusive mechanisms may underlie the widespread translational buffering of divergent gene expression observed in my study (McManus et al., 2014). First, in the case of genetic compensation, regulatory mutations leading to counterbalancing translational changes might be fixed during evolution. Second, in the case of network robustness, gene regulatory networks might be inherently robust, such that changes in mRNA abundance are immediately and automatically buffered at the translational level by regulatory feedback loops. I hypothesize that

most of the translational buffering observed here is explained by genetic compensation, for three reasons: (i) translational buffering in systems covering short evolutionary time periods seems overall limited (see above); (ii) over longer time periods, with accumulating transcriptional change, network robustness would be expected to reach its limits or would need to adapt to new thresholds; (iii) previous yeast work suggested the presence of transcriptional rather than translational network robustness (Bader et al., 2015); and (iv) a general mechanism that could sense evolutionary mRNA abundance changes and regulate translation accordingly is difficult to envision (Bader et al., 2015). Compensatory mutations might in particular shape rates of translational initiation, given that this step represents the most regulated step of translational control (Sonenberg and Hinnebusch, 2009). Such mutations would increase fitness in individuals with fitness-decreasing transcriptional changes and would therefore be expected to be ultimately fixed in populations and species by positive selection.

Altogether, my work identified strong and highly differential patterns of translational buffering of gene expression divergence in mammalian organs. Given the potential of further buffering at the post-translational level (Vogel et al., 2012; Wang et al., 2018), protein abundance is likely to be conserved across mammals to a remarkable degree, especially for certain gene classes and tissues, a notion that needs to be considered in future investigations of mammalian genome biology. The extensive translome and matched transcriptome data and results provide a resource for such future endeavors and for the exploration of regulatory mechanisms and their evolution across all gene expression layers.

## 7.2 Outlook

Apart from what I have presented above, the extensive data generated in the framework of my thesis project also allow me to explore other aspects:

**The translation and functionality of newly emerged genes.** In the analyses presented above, I have mainly focused on the evolution and patterns of the two gene expression layers of relatively old genes (1:1 orthologs). The “birth” of new genes is fundamental to the evolution of lineage- or species-specific phenotypic traits (Kaessmann, 2010). Many previous studies from our lab have characterized the evolution of new protein-coding genes in mammals (Burki and Kaessmann, 2004; Kaessmann et al., 2009; Kaessmann, 2010; Carelli et al., 2016; Guschanski et al., 2017). However, while these studies have provided many insights into the mechanisms underlying the formation and evolution of new genes, a major challenge has been to solidly support their functionality and thus to distinguish phenotypically relevant new genes from nonfunctional pseudogenes (Kaessmann, 2010). This has been a problem in particular for recently emerged genes for which statistically significant evolutionary signatures indicative of functionality (e.g., the selective preservation of ORFs) are harder to obtain. Proving that a putative new protein-coding gene is not only transcribed but that its mRNA is also translated would lend strong support to its functionality.

**The evolution and functional relevance of uORFs and transcript leaders.** In addition to the main coding region of genes, a sizeable number of genes (in particular those involved in cellular growth and differentiation) may carry upstream open reading frames (uORFs) in their transcript leader sequences, commonly referred to as 5' UTRs, although in the case of the presence of an uORF this designation is misleading (Morris and Geballe, 2000). However, although uORFs have been shown to affect translation of the main coding ORF (Janich et al., 2015; Johnstone et al., 2016), their precise prevalence, genomic distribution, overall functional relevance, and evolutionary dynamics have remained unclear. It would be interesting to determine the frequencies of conserved and lineage-specific uORFs across mammals and tissues and thus assess their turnover rates and

potential functional roles. Together with analyses of implications of uORFs in translation rate changes, we can shed initial light on the overall functional relevance and evolutionary dynamics of uORFs and thus, more generally, on the function of transcript leader sequences. It will likely also reveal individual candidate cases that warrant more detailed experimental characterization and validation.

**Regulatory basis of evolutionary translation change.** Translation can be controlled in many ways, especially at both the translation initiation and elongation steps (Ingolia et al., 2009; Ingolia et al., 2011). We seek to identify regulatory alterations (e.g., uORFs, Kozak sequence, pausing site, RNA modifications) that underlie evolutionary changes in translation efficiencies. We are also interested in assessing the conservation of translation regulatory features, in order to obtain novel insights into their functional relevance in mammals. Thus, we could compare in detail translation patterns and sequences of orthologous genes that we found to have diverged significantly in terms of their overall translation efficiencies in a given mammalian lineage. For example, we could screen for potential translational pause sites across orthologs using our ribosome footprint density plots and typical motifs in associated coding sequences (Ingolia et al., 2011; Zhang et al., 2017; Joazeiro, 2017), and then assess whether these sites changed in the species/lineage displaying the translation efficiency change and may thus underlie efficiency shifts. Similarly, we could assess to what extent gains, losses, or quantitative expression changes of potential uORFs (see above) may have contributed to evolutionary changes of translation efficiencies. Generally, depending on the precise observations that we will make, we could explore various other potential regulatory changes that may underlie changes in translation efficiencies/control. For example, by combining the data generated in this work with our previous data for miRNAs (Meunier et al., 2013; Warnefors et al., 2017), we may be able to assess whether changes in 3' UTRs that lead to changes in target sites of microRNAs contributed to evolutionary changes in translation efficiencies, through miRNA-mediated translational repression (REFS) (Carthew and Sontheimer, 2009; Guo et al., 2010)). Finally, we could assess the extent of conservation of translation features such

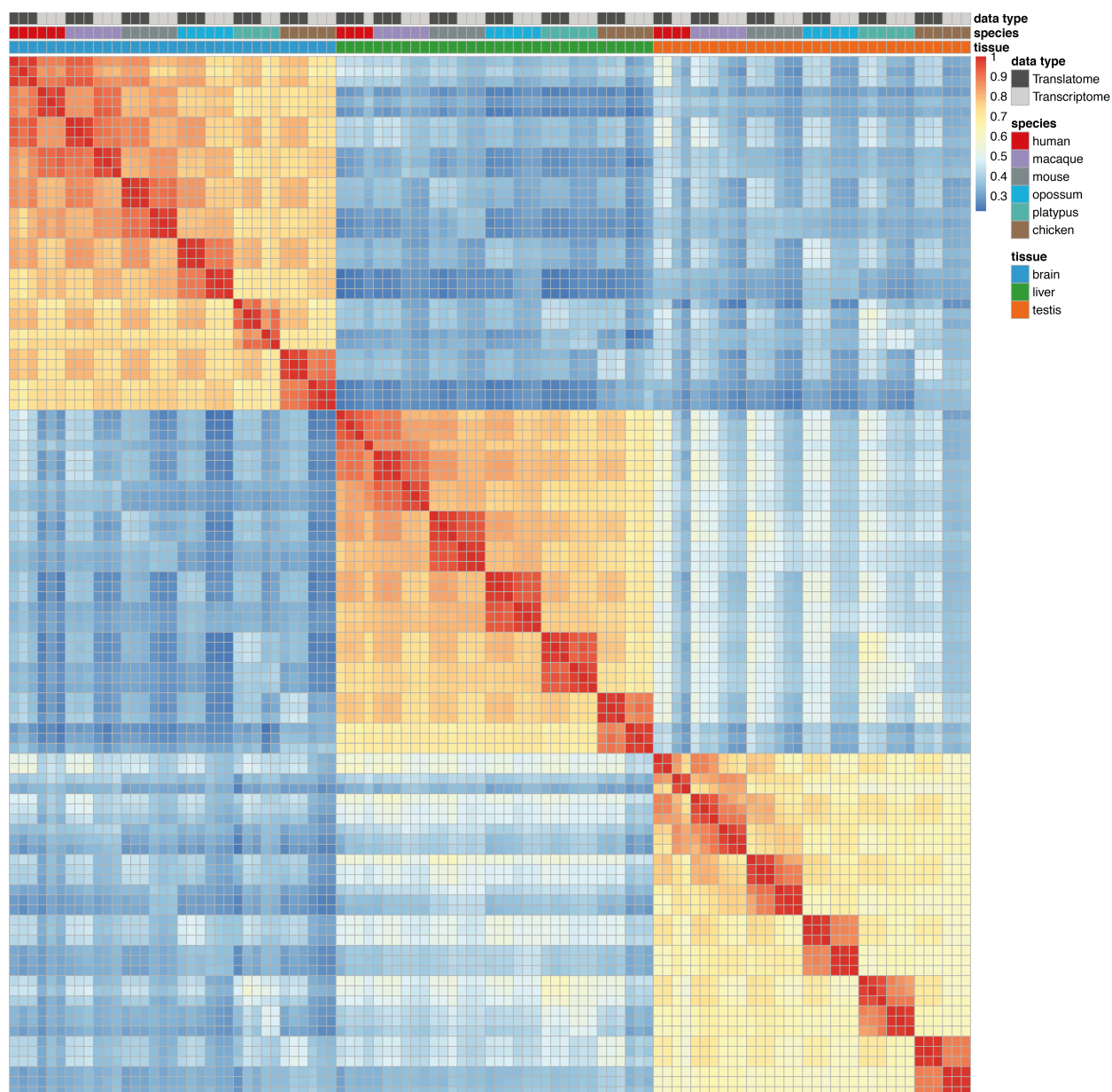
as translational pause sites and uORF translation patterns across orthologous mammalian genes. We may thus identify highly conserved regulatory mechanisms of specific genes that can be further scrutinized experimentally.

### **7.3 Concluding remarks**

In the framework of my thesis project, we generated an extensive and unique set of translome data across all major mammalian lineages using the technique of ribosome profiling. In junction with matched transcriptome data, my analyses addressed long-standing hypotheses pertaining to gene expression change in mammals and its implications for the emergence of species-specific phenotypes. Specifically, translational forces frequently counteracted but rarely boosted transcriptional changes (chapter 3). Expression changes of functionally cooperating genes tend to be balanced by concerted (modular) translational changes to preserve ancestral cellular stoichiometries (chapter 4). The widespread translational buffering more strongly preserved dosage-sensitive and, especially, housekeeping genes (chapter 5). Translational upregulation acts to globally counterbalance the global dosage reduction that arose in the wake of mammalian sex chromosome differentiation (chapter 6). Altogether, fine-tuned translational buffering substantially stabilized gene expression levels during mammalian evolution.



## Supplementary Figures

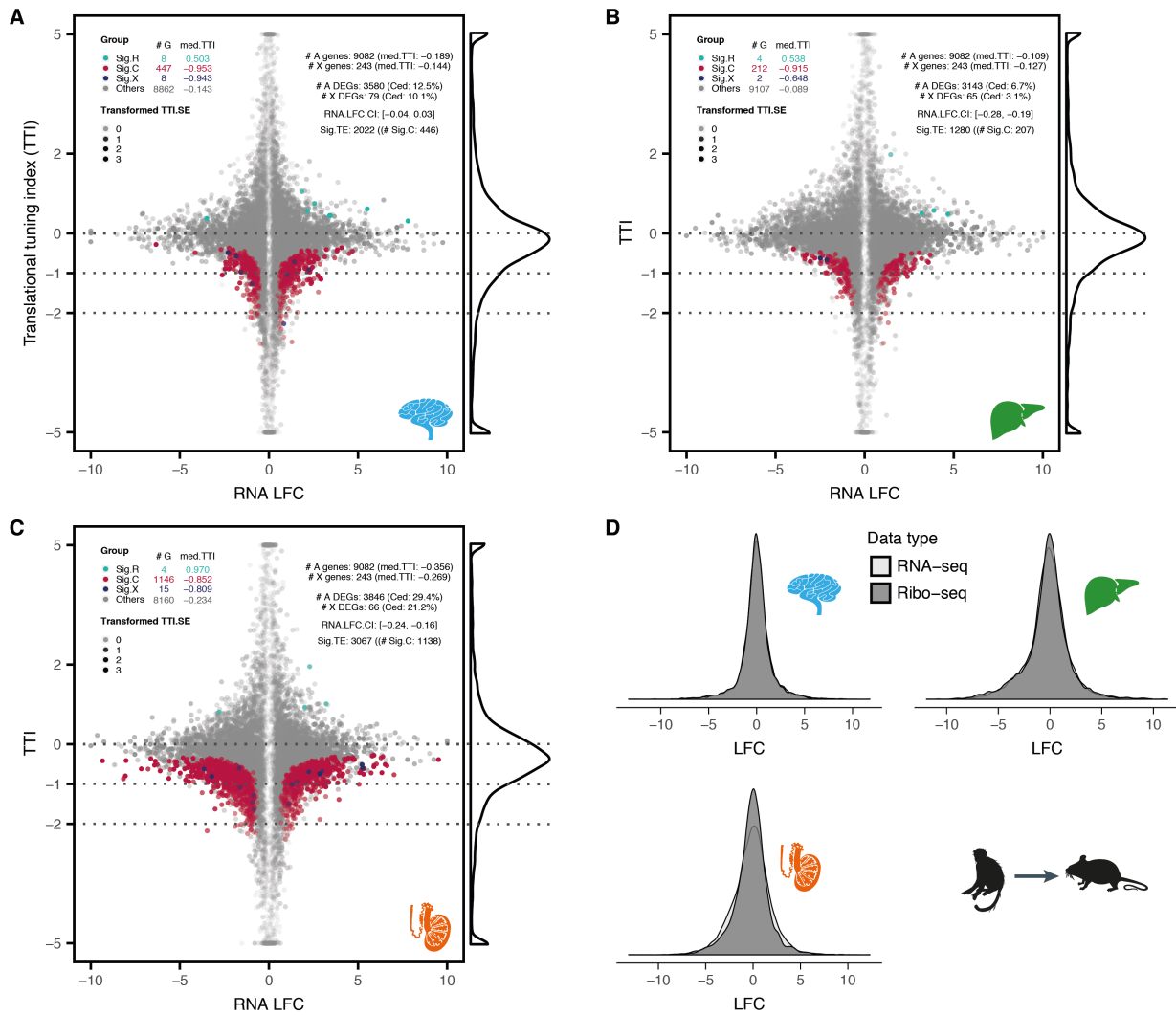


(figure legend continued on next page)

**Supplementary Figure 1: Correlations of two gene expression layers across different organs and species (perfectly aligned region)**

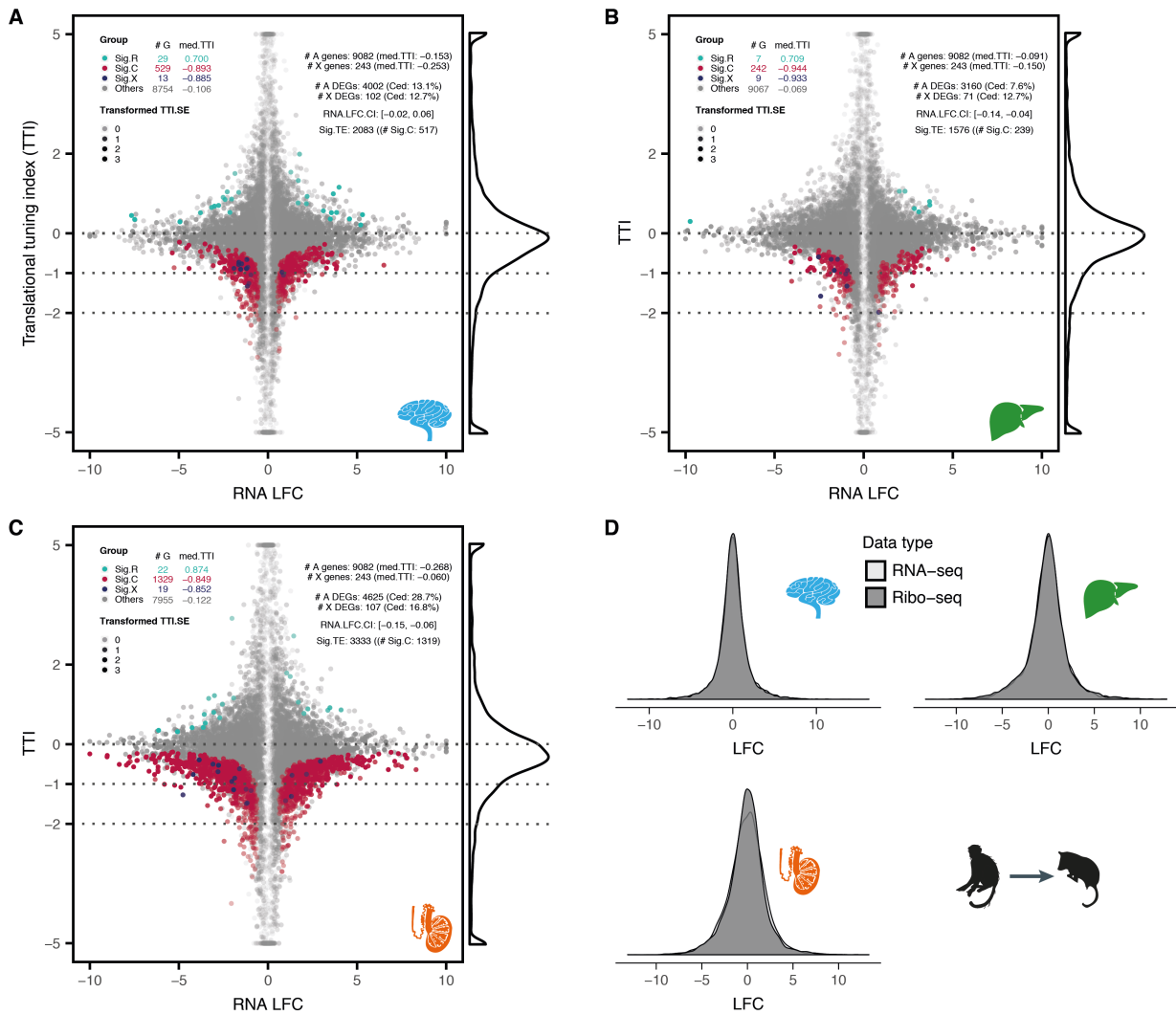
The heatmap of the pairwise correlations (Spearman's  $\rho$ ) is based on the set of 5,149 robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 amniote orthologs (perfectly aligned region). It represents the degree of similarity of gene expression profiles between data types (translatome, transcriptome), species (human, macaque, mouse, opossum, platypus, chicken) and tissues (brain, liver, testis).





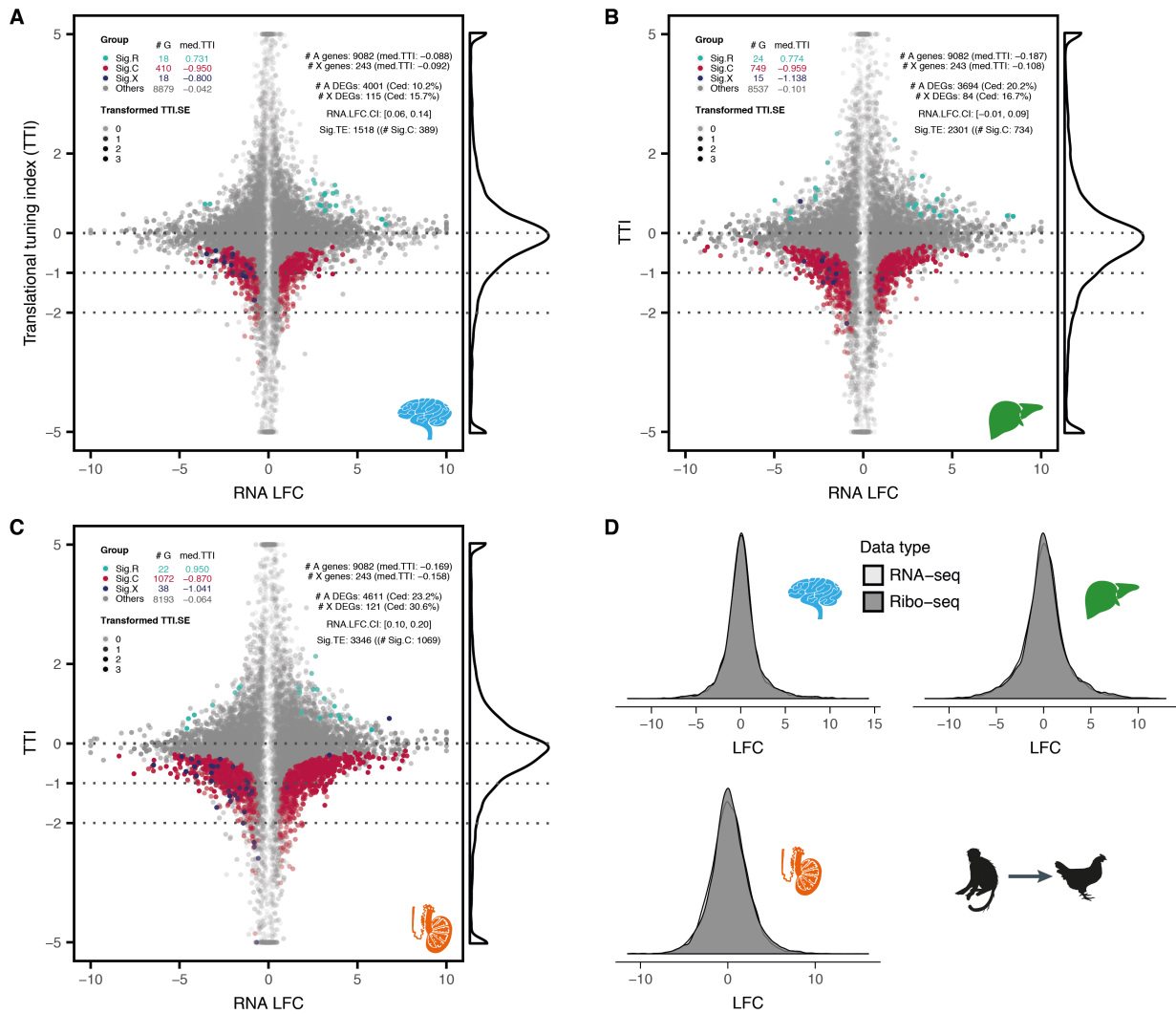
### Supplementary Figure 2: Translational vs. transcriptional changes for individual genes between macaque and mouse

Translational tuning index (TTI) for each of the 9,325 1:1 orthologs between macaque and mouse (the reference) plotted against corresponding transcript (RNA) abundance change ( $\log_2$ -fold change, LFC) for brain (A), liver (B), and testis (C) (See Figure 3.2 for more legend details). (D) Distributions of the LFCs of expression levels between macaque and mouse for RNA-seq and Ribo-seq data for brain, liver, and testis, respectively.



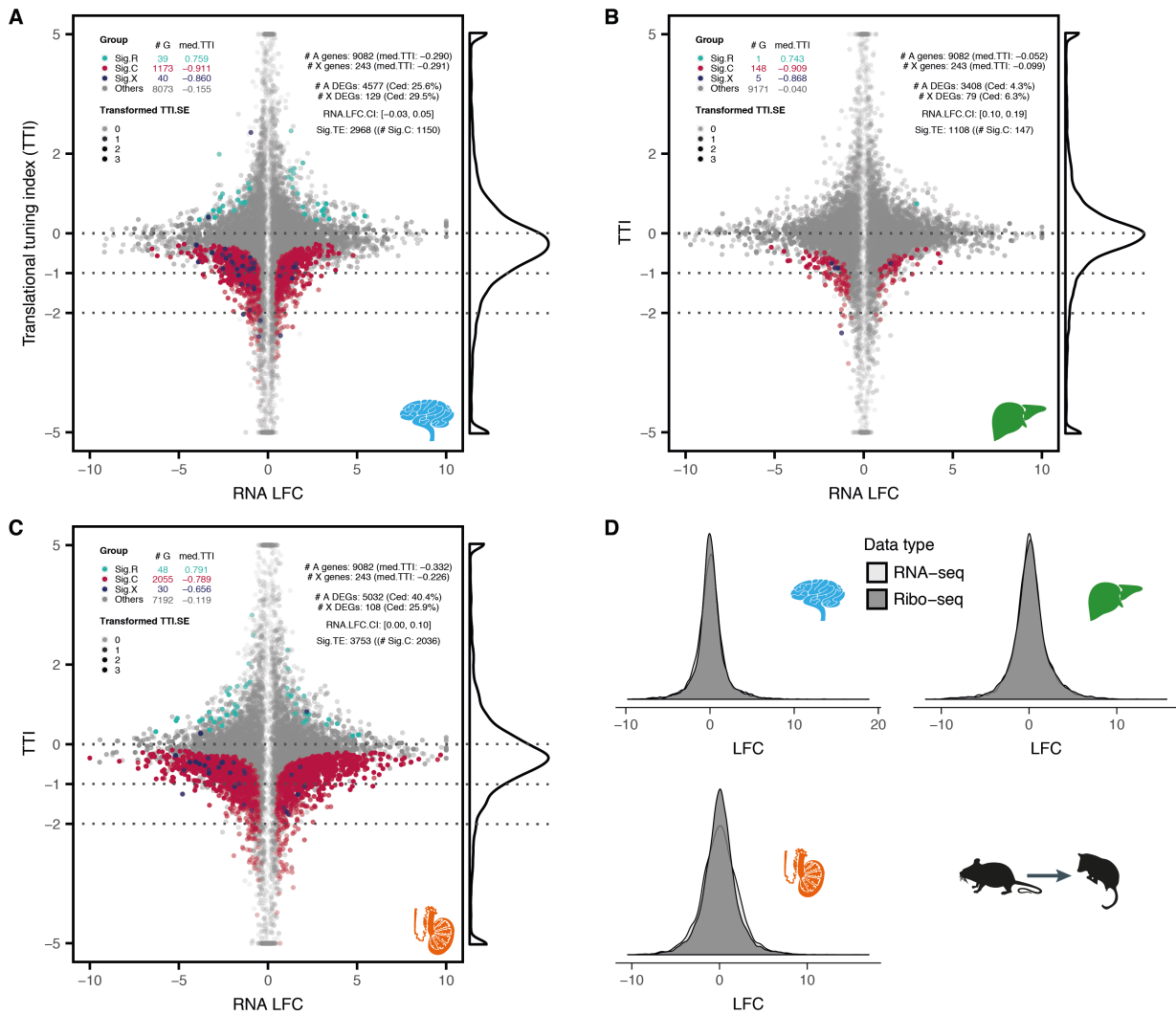
**Supplementary Figure 3: Translational vs. transcriptional changes for individual genes between macaque and opossum**

Translational tuning index (TTI) for each of the 9,325 1:1 orthologs between macaque and opossum (the reference) plotted against corresponding transcript (RNA) abundance change (log<sub>2</sub>-fold change, LFC) for brain (A), liver (B), and testis (C) (See Figure 3.2 for more legend details). (D) Distributions of the LFCs of expression levels between macaque and opossum for RNA-seq and Ribo-seq data for brain, liver, and testis, respectively.



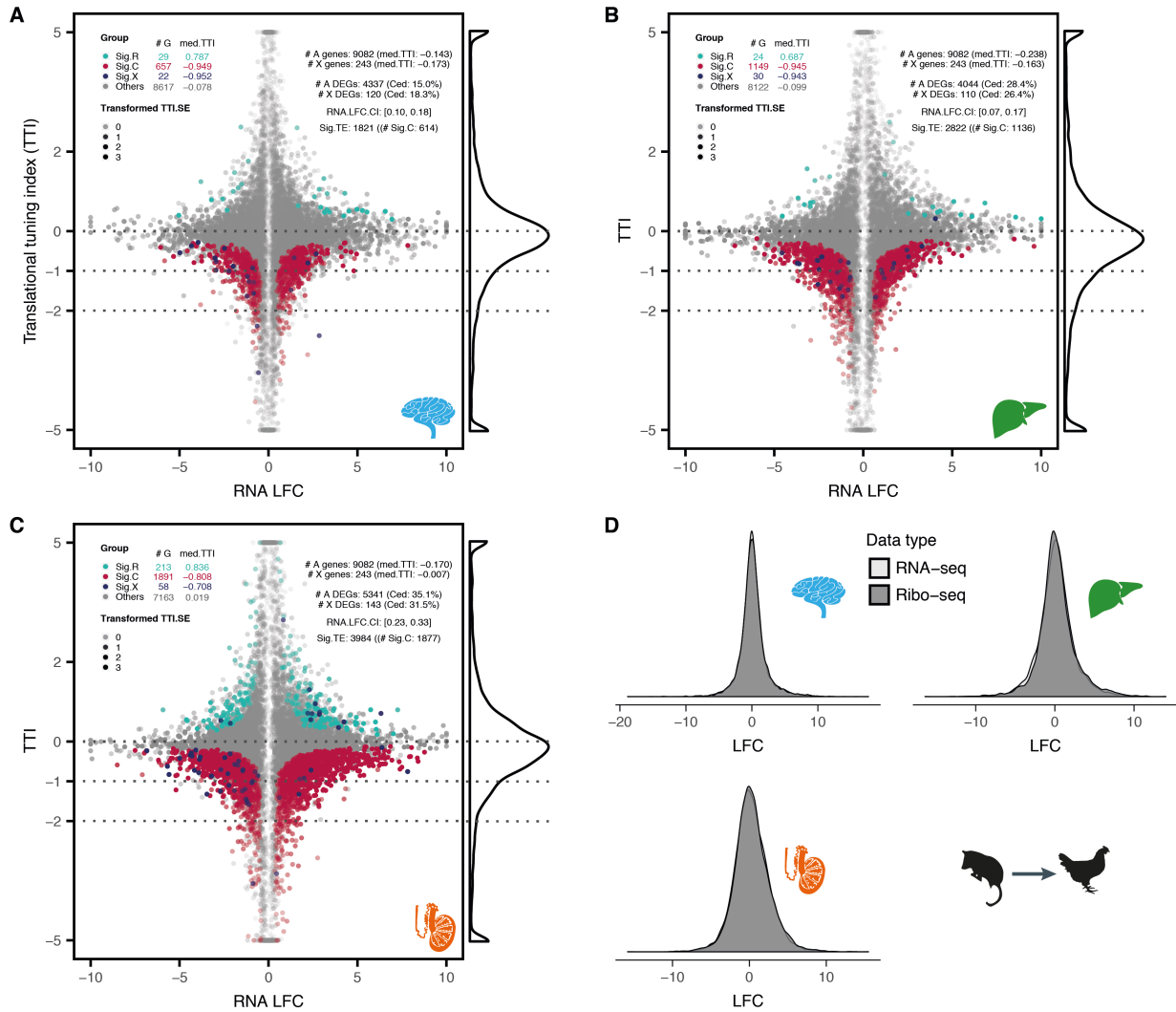
#### Supplementary Figure 4: Translational vs. transcriptional changes for individual genes between macaque and chicken

Translational tuning index (TTI) for each of the 9,325 1:1 orthologs between macaque and chicken (the reference) plotted against corresponding transcript (RNA) abundance change ( $\log_2$ -fold change, LFC) for brain (A), liver (B), and testis (C) (See Figure 3.2 for more legend details). (D) Distributions of the LFCs of expression levels between macaque and chicken for RNA-seq and Ribo-seq data for brain, liver, and testis, respectively.



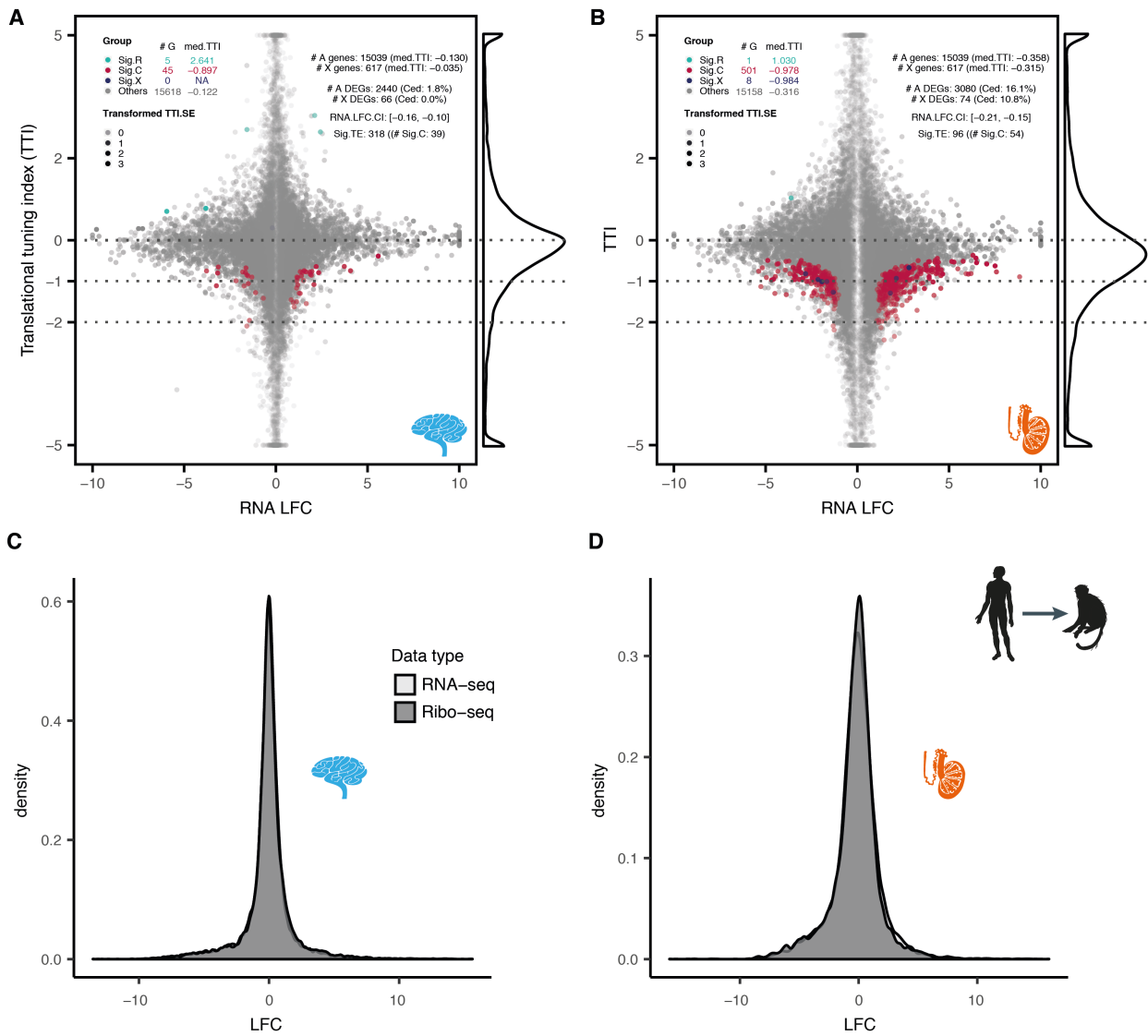
**Supplementary Figure 5: Translational vs. transcriptional changes for individual genes between mouse and opossum**

Translational tuning index (TTI) for each of the 9,325 1:1 orthologs between mouse and opossum (the reference) plotted against corresponding transcript (RNA) abundance change ( $\log_2$ -fold change, LFC) for brain (A), liver (B), and testis (C) (See Figure 3.2 for more legend details). (D) Distributions of the LFCs of expression levels between mouse and opossum for RNA-seq and Ribo-seq data for brain, liver, and testis, respectively.



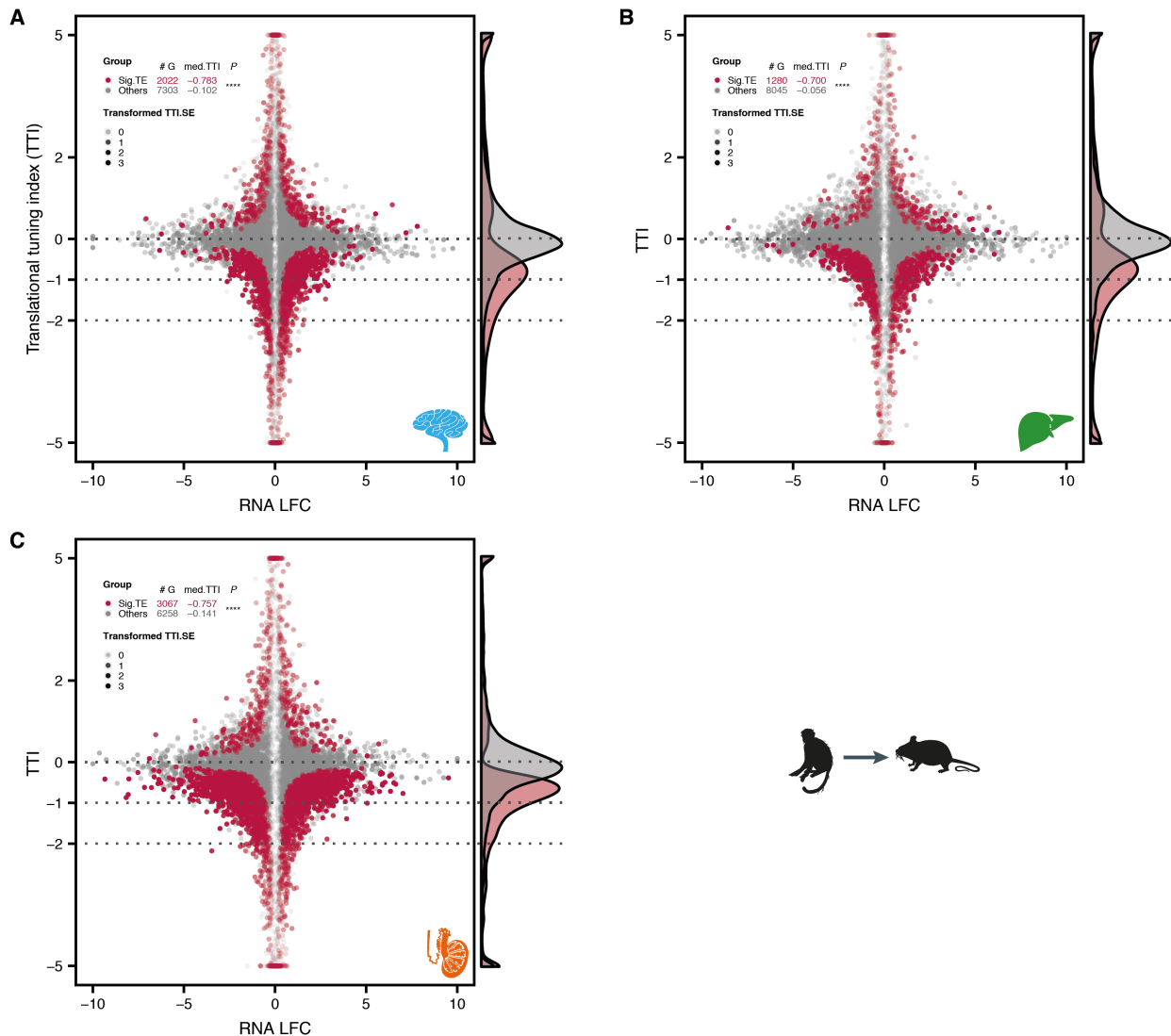
### Supplementary Figure 6: Translational vs. transcriptional changes for individual genes between opossum and chicken

Translational tuning index (TTI) for each of the 9,325 1:1 orthologs between opossum and chicken (the reference) plotted against corresponding transcript (RNA) abundance change ( $\log_2$ -fold change, LFC) for brain (A), liver (B), and testis (C) (See Figure 3.2 for more legend details). (D) Distributions of the LFCs of expression levels between opossum and chicken for RNA-seq and Ribo-seq data for brain, liver, and testis, respectively.



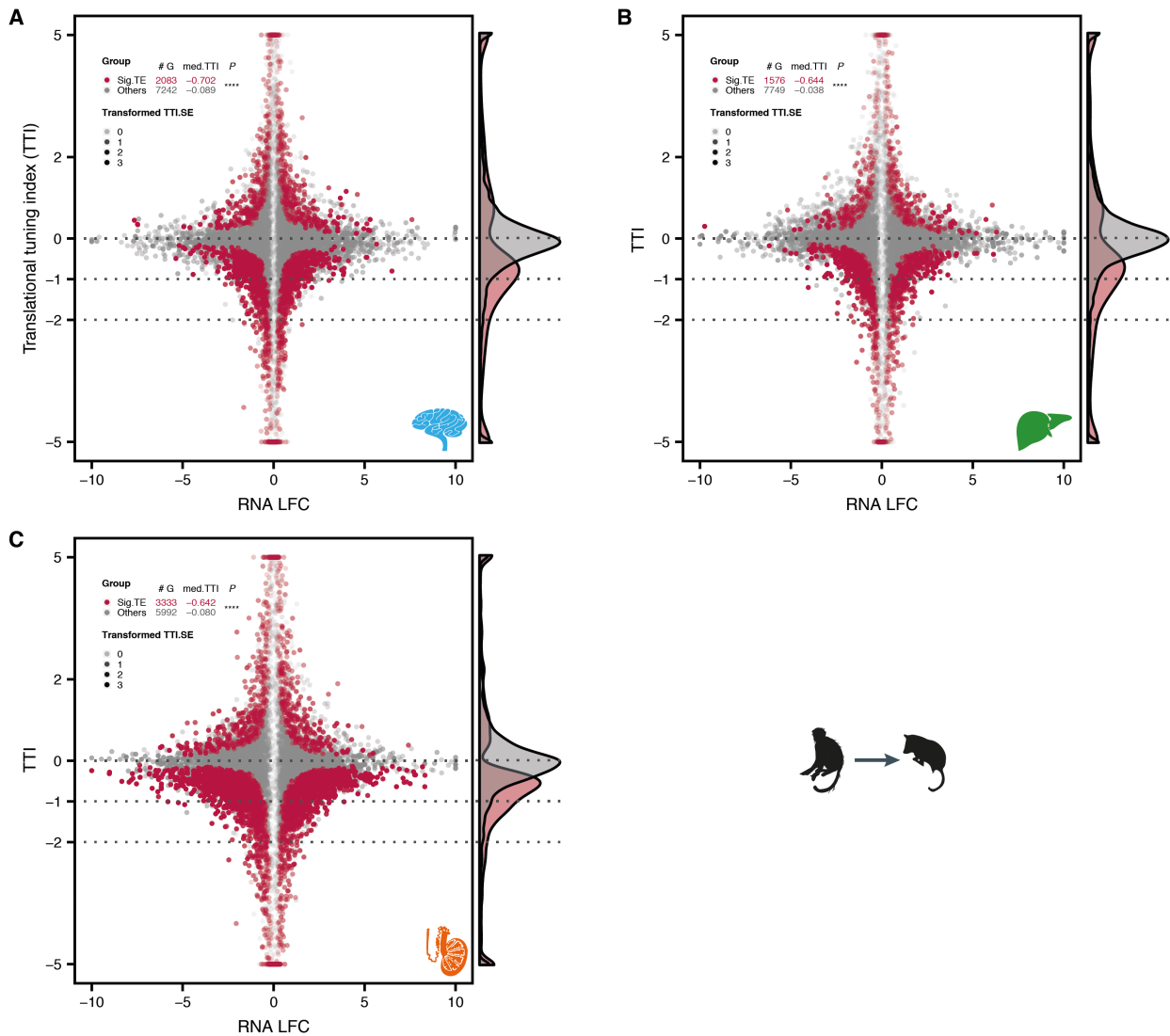
**Supplementary Figure 7: Translational vs. transcriptional changes for individual genes between human and macaque**

Translational tuning index (TTI) for each of the 15,668 1:1 orthologs between human and macaque (the reference) plotted against corresponding transcript (RNA) abundance change (log<sub>2</sub>-fold change, LFC) for brain (A) and testis (B) (See Figure 3.2 for more legend details). Distributions of the LFCs of expression levels between human and macaque for RNA-seq and Ribo-seq data for brain (C) and testis (D). 12 genes on the Y chromosome were not included in this analysis.



### Supplementary Figure 8: TTI distribution for genes of significant TE changes between macaque and mouse

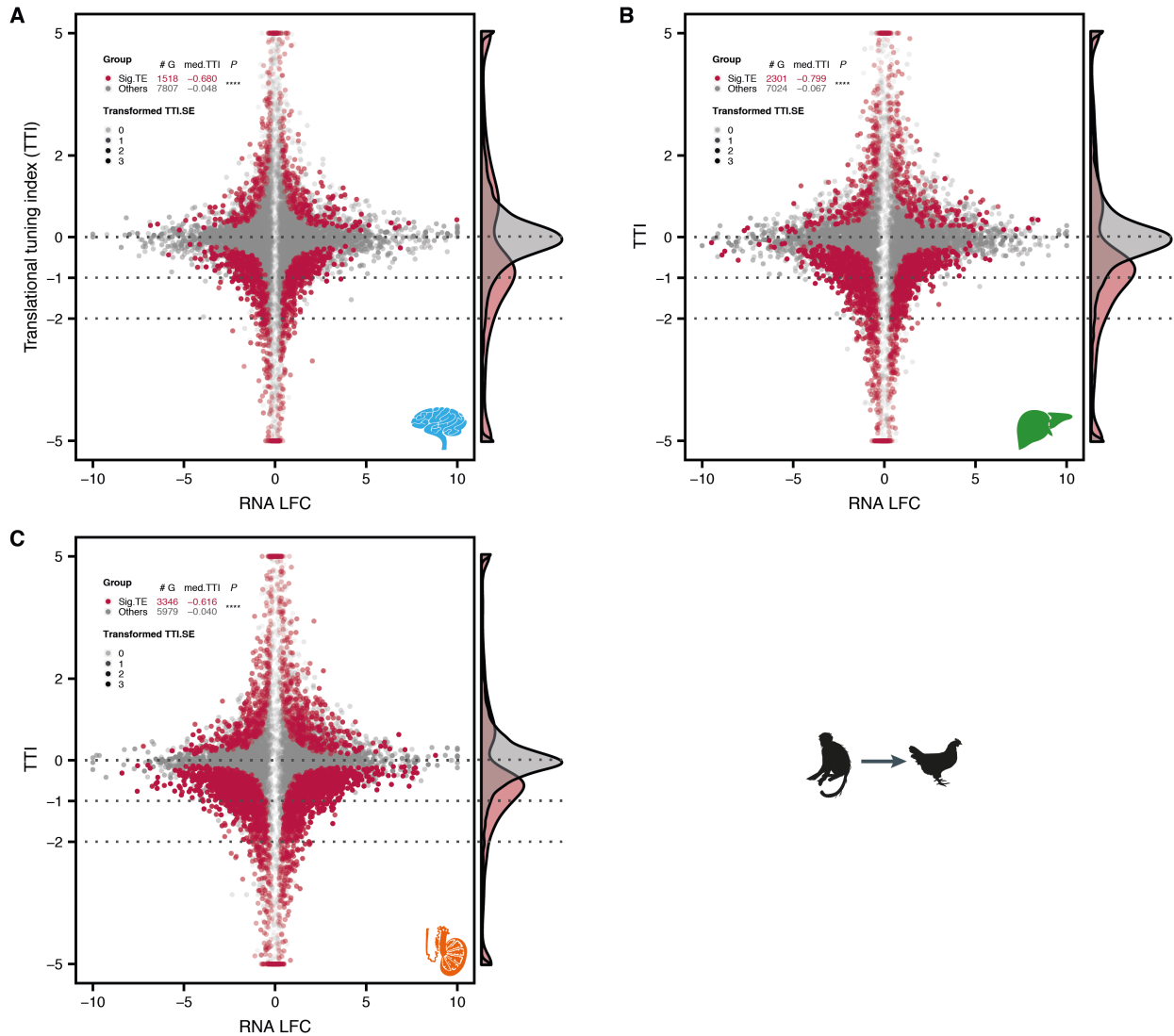
(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



**Supplementary Figure 9: TTI distribution for genes of significant TE changes between macaque and opossum**

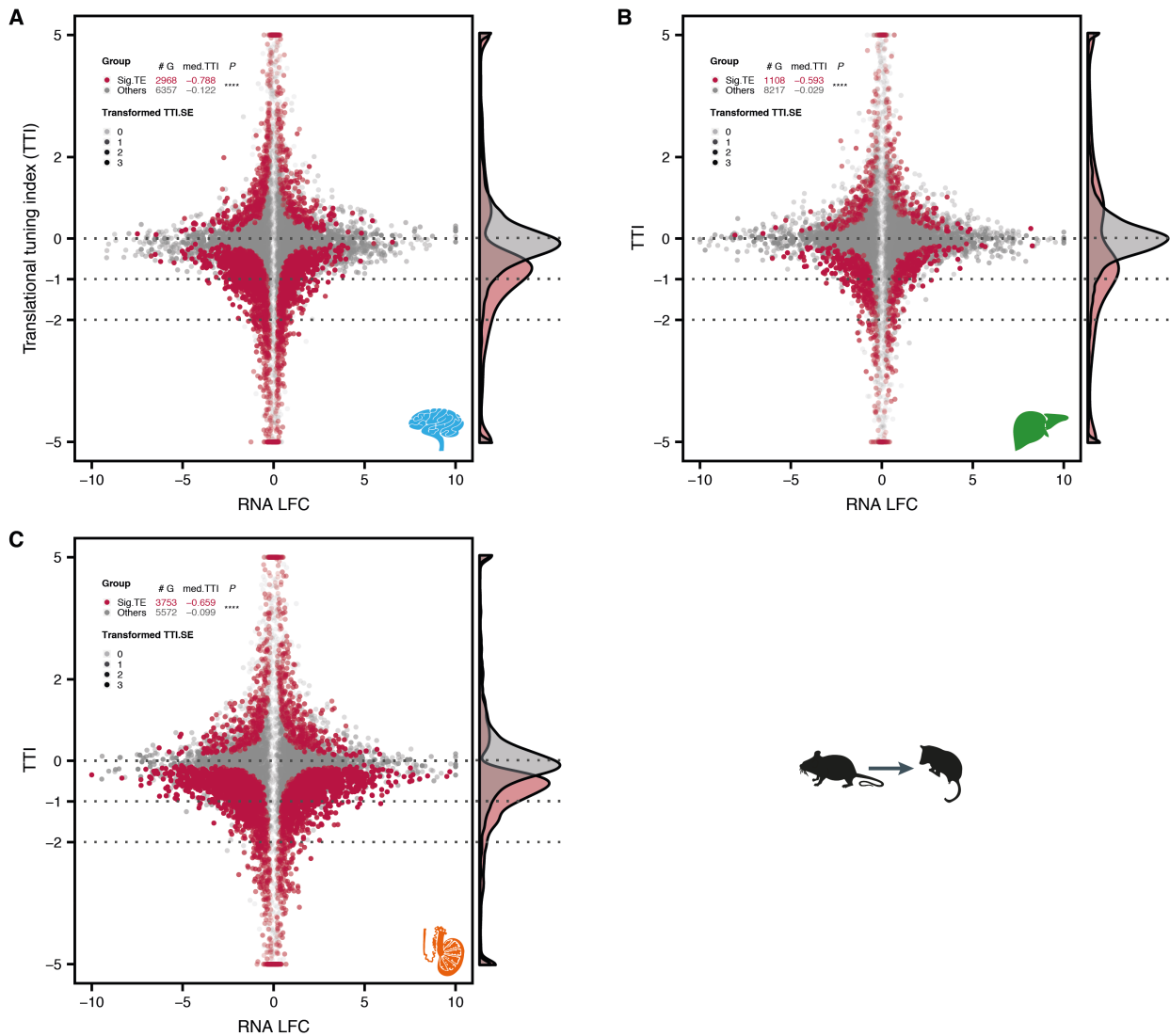
(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .





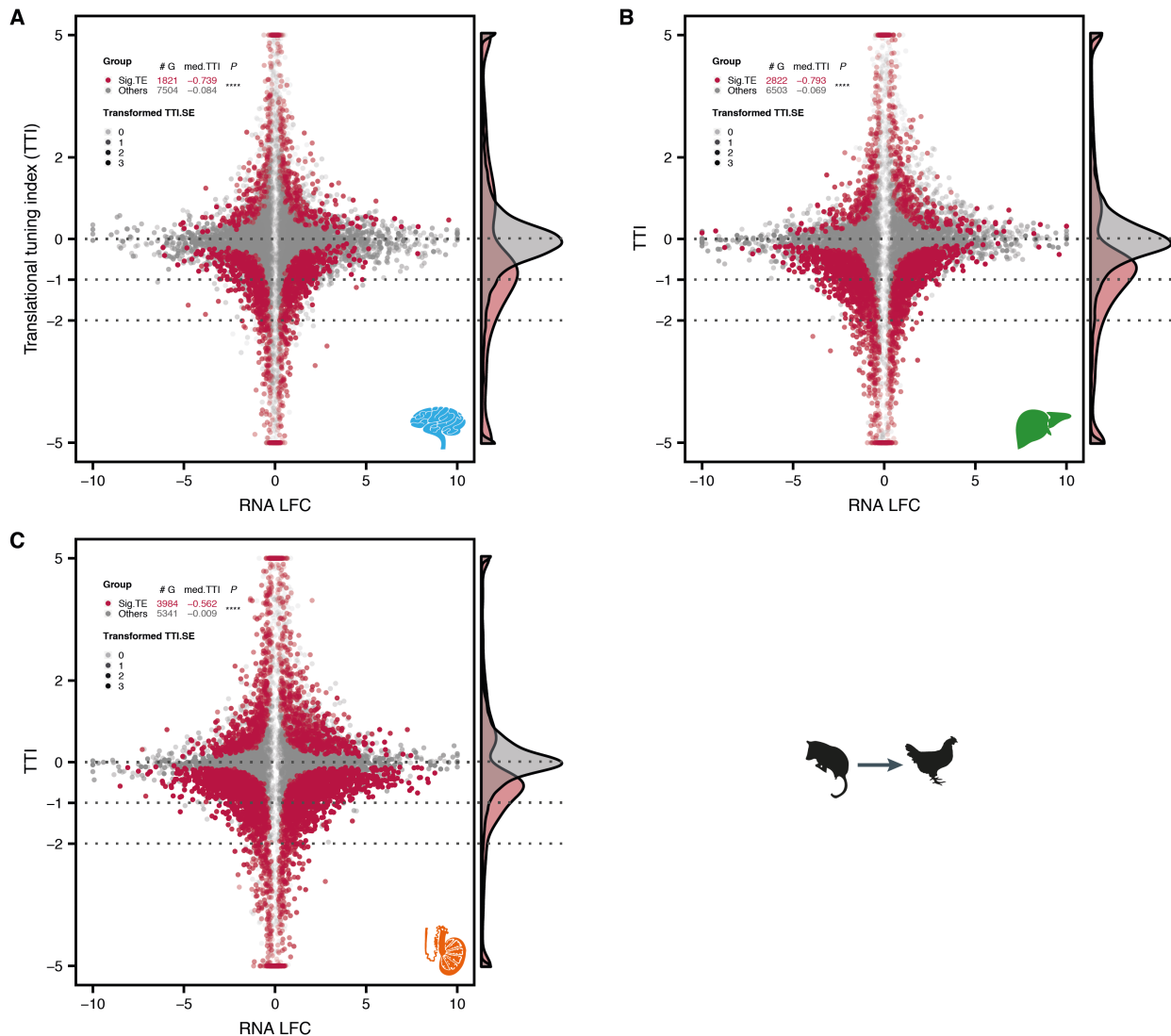
### Supplementary Figure 10: TTI distribution for genes of significant TE changes between macaque and chicken

(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



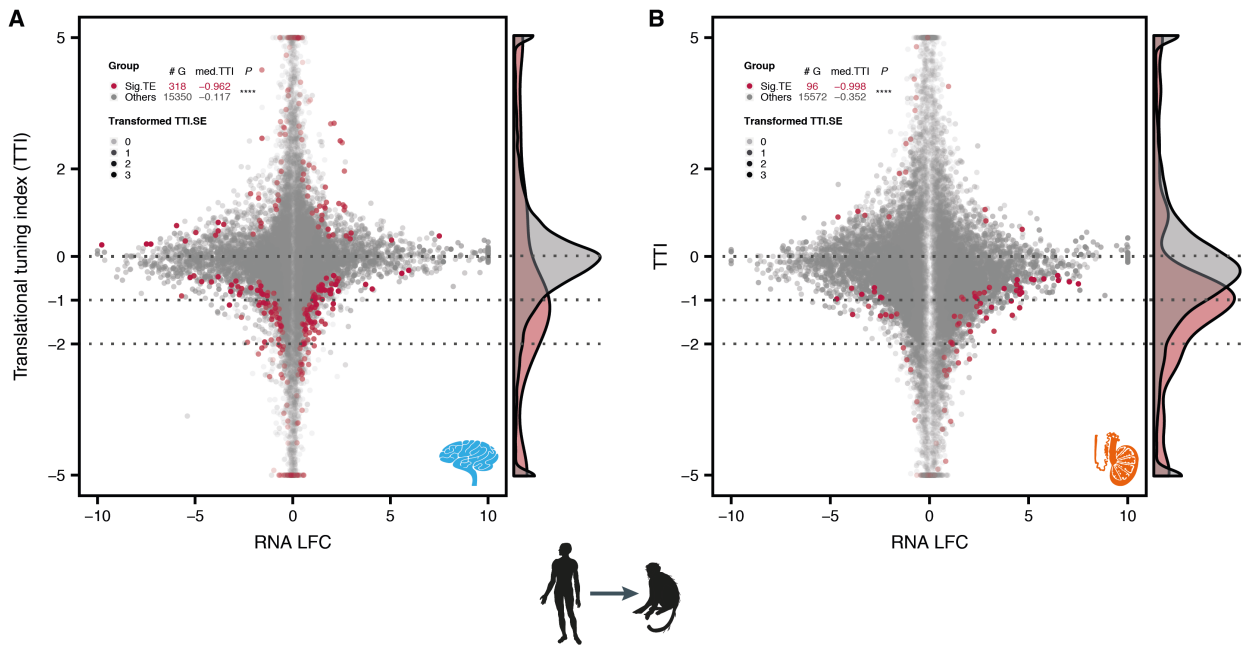
**Supplementary Figure 11: TTI distribution for genes of significant TE changes between mouse and opossum**

(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



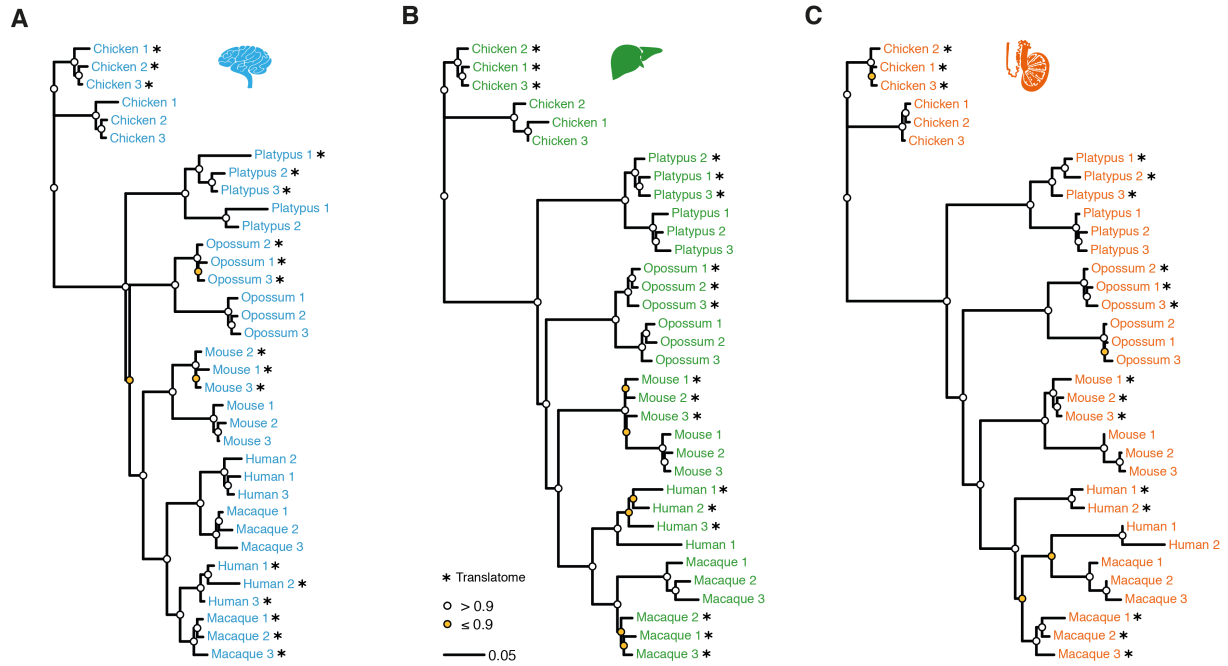
### Supplementary Figure 12: TTI distribution for genes of significant TE changes between opossum and chicken

(A to C) The number (# G) and median TTI (med.TTI) of 9,325 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain, liver and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI.SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



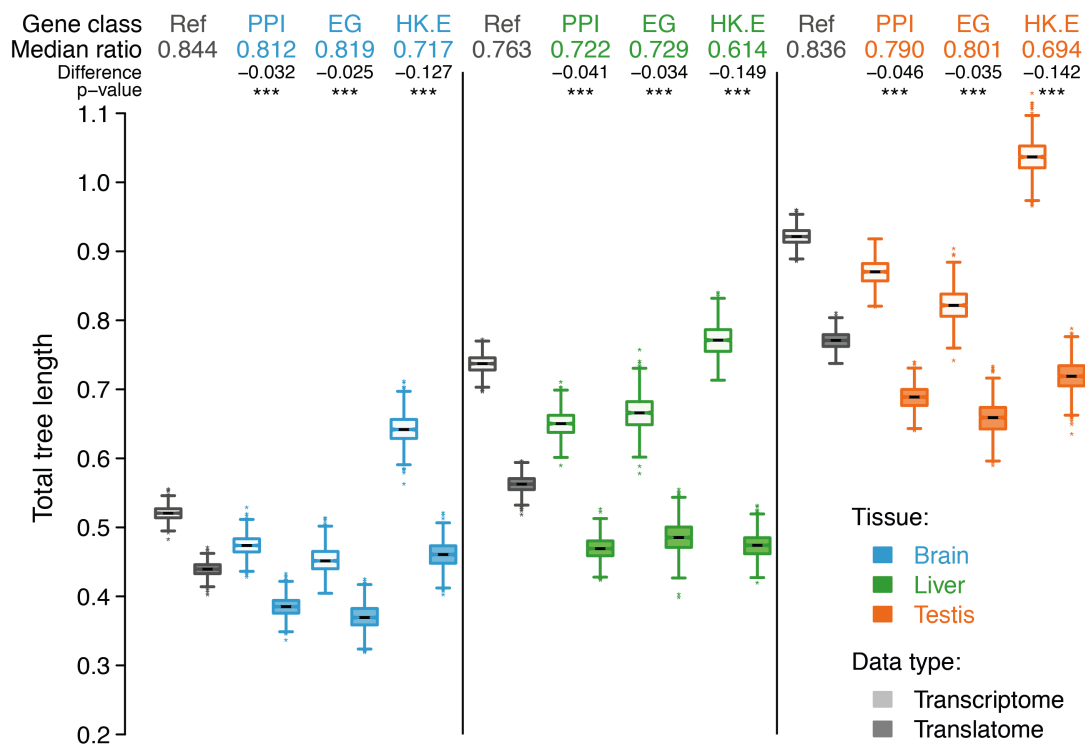
**Supplementary Figure 13: TTI distribution for genes of significant TE changes between human and macaque**

(A and B) The number (# G) and median TTI (med.TTI) of 15,668 1:1 orthologs with significant changes of TE (Sig.TE, red) (see Section 3.2.2 for method details) are shown for brain and testis, respectively. The transformed standard error (SE) of TTI (Transformed TTI,SE) is reflected by the extent of transparency; the bigger the SE, the more transparent the plotted data point. For display purposes, TTIs and RNA LFCs were capped at  $-5$  and  $5$ , and at  $-10$  and  $10$ , respectively, with more extreme values replaced by these values. TTI density distributions for Sig.TE and other genes are shown to the right of each scatter plots. Enrichment analysis (Section 3.2.3) was employed to estimate whether the weighted mean of TTI for Sig.TE is statistically different from that of other genes; p-value,  $P$ : \*\*\*\*,  $< 0.0001$ ; \*\*\*,  $< 0.001$ ; \*\*,  $< 0.01$ ; \*,  $< 0.05$ ; ns (not significant),  $\geq 0.05$ .



**Supplementary Figure 14: Mammalian gene expression (translatome and transcriptome) phylogenies (perfectly aligned region)**

(A to C) NJ trees based on pairwise expression distance matrices ( $1 - \text{Spearman's } \rho$ ) for 5,095, 4,593 and 5,233 robustly expressed 1:1 amniote orthologs (perfectly aligned region) in brain, liver and testis, respectively. See Figure 4.1, A to C for more legend details.



### Supplementary Figure 15: Patterns of expression divergence and compensatory evolution across gene classes

Extended from Figures 5.1 to 5.3. Abbreviations: Ref, robustly expressed (median FPKM > 1 across all RNA-seq libraries) 1:1 orthologs among the four representative species (macaque, mouse, opossum, and chicken) for the respective tissue; PPI, mouse genes encoding proteins that involve in protein-protein interactions downloaded from BioGRID V3.4.156 (Chatr-Aryamontri et al., 2017); EG, mouse essential genes (gene knockout leads to lethality) collected from the database of Mouse Genome Informatics (MGI) (Smith et al., 2018); HK.E, mouse housekeeping genes projected from a set of human housekeeping genes obtained from a previous study (Eisenberg and Levanon, 2013).

---

## References

- Afzali B, Grönholm J, Vandrovцова J, O'Brien C, Sun HW, Vanderleyden I, Davis FP, Khoder A, Zhang Y, Hegazy AN, Villarino AV, Palmer IW, Kaufman J, Watts NR, Kazemian M, Kamenyeva O, Keith J, Sayed A, Kasperaviciute D, Mueller M, et al. 2017. BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency. *Nat. Immunol.* **18**, 813–823.
- Albert FW, Muzzey D, Weissman JS, Kruglyak L. 2014. Genetic influences on translation in yeast. *PLoS Genet.* **10**, e1004692.
- Antonarakis SE. 2017. Down syndrome and the complexity of genome dosage imbalance. *Nat. Rev. Genet.* **18**, 147–163.
- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3889–3894.
- Artieri CG, Fraser HB. 2014. Evolution at two levels of gene expression in yeast. *Genome Res.* **24**, 411–421.
- Atkinson EG, Audesse AJ, Palacios JA, Bobo DM, Webb AE, Ramachandran S, Henn BM. 2018. No Evidence for Recent Selection at *FOXP2* among Diverse Human Populations. *Cell* **174**, 1–12.
- Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124.
- Bader DM, Wilkening S, Lin G, Tekkedil MM, Dietrich K, Steinmetz LM, Gagneur J. 2015. Negative feedback buffers effects of regulatory variants. *Mol. Syst. Biol.* **11**, 785.
- Badran YR, Dedeoglu F, Leyva Castillo JM, Bainter W, Ohsumi TK, Bousvaros A, Goldsmith JD, Geha RS, Chou J. 2017. Human RELA haploinsufficiency results in autosomal-dominant chronic mucocutaneous ulceration. *J. Exp. Med.* **214**, 1937–1947.
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593.
- Bartha I, di Iulio J, Venter JC, Telenti A. 2018. Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62.
- Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667.

- Bauernfeind AL, Soderblom EJ, Turner ME, Moseley MA, Ely JJ, Hof PR, Sherwood CC, Wray GA, Babbitt CC. 2015. Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. *Genome Biol. Evol.* **7**, 2276–2288.
- Bazzini AA, Lee MT, Giraldez AJ. 2012. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**, 233–237.
- Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993.
- Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, Kremitzki C, Fulton RS, Dugan S, Ding Y, Morton D, Khan Z, Lewis L, Buhay C, Wang Q, Watt J, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499.
- Bellott DW, Skaletsky H, Cho TJ, Brown L, Locke D. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat. Genet.* **49**, 387–394.
- Boyd JL, Skove SL, Rouanet JP, Pilaz LJ, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr. Biol.* **25**, 772–779.
- Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16**, 651–664.
- Brar GA. 2016. Beyond the Triplet Code: Context Cues Transform Translation. *Cell* **167**, 1681–1692.
- Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786.
- Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348.
- Breschi A, Djebali S, Gillis J, Pervouchine DD, Dobin A, Davis CA, Gingeras TR, Guigó R. 2016. Gene-specific patterns of expression variation across organs and species. *Genome Biol.* **17**, 151.
- Breschi A, Gingeras TR, Guigó R. 2017. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18**, 425–440.
- Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165**, 349–357.
- Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46**, 111–138.



- Brockdorff N, Turner BM. 2015. Dosage compensation in mammals. *Cold Spring Harb. Perspect. Biol.* **7**, a019406.
- Burki F, Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat. Genet.* **36**, 1061–1063.
- Capel B. 2017. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nat. Rev. Genet.* **18**, 675–689.
- Capra JA, Stolzer M, Durand D, Pollard KS. 2013. How old is my gene? *Trends Genet.* **29**, 659–668.
- Cardarelli L, Maxwell KL, Davidson AR. 2011. Assembly mechanism is the key determinant of the dosage sensitivity of a phage structural protein. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10168–10173.
- Cardoso-Moreira M, et al. 2018. Molecular innovation and conservation across mammalian organ development. under review.
- Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**, 301–314.
- Carlson M. 2018. org.Mm.eg.db: Genome wide annotation for Mouse.
- Carthew RW, Sontheimer EJ. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655.
- Castelo-Szekely V, Arpat AB, Janich P, Gatfield D. 2017. Translational contributions to tissue specificity in rhythmic and constitutive gene expression. *Genome Biol.* **18**, 116.
- Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, Ricci E, Snyder MP. 2015. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25**, 1610–1621.
- Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jónsson B, Schluter D, Bell MA, Kingsley DM. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305.
- Chan PP, Lowe TM. 2016. GtRNADB 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189.
- Charlesworth B. 1978. Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 5618–5622.
- Chassé H, Boulben S, Costache V, Cormier P, Morales J. 2017. Analysis of translation using polysome profiling. *Nucleic Acids Res.* **45**, e15.
- Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M. 2017. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379.

- Chekulaeva M, Rajewsky N. 2018. Roles of Long Noncoding RNAs and Circular RNAs in Translation. *Cold Spring Harb. Perspect. Biol.* pii, a032680.
- Chen WH, Trachana K, Lercher MJ, Bork P. 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol. Biol. Evol.* **29**, 1703–1706.
- Chen X, Zhang J. 2015. No X-chromosome dosage compensation in human proteomes. *Mol. Biol. Evol.* **32**, 1456–1460.
- Cheng Z, Otto GM, Powers EN, Keskin A, Mertins P, Carr SA, Jovanovic M, Brar GA. 2018. Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell* **172**, 910–923.
- Cloutier JM, Turner JM. 2010. Meiotic sex chromosome inactivation. *Curr. Biol.* **20**, R962–R963.
- Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712.
- Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846.
- Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493.
- Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720.
- Dang VT, Kassahn KS, Marcos AE, Ragan MA. 2008. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur. J. Hum. Genet.* **16**, 1350–1357.
- de Clare M, Pir P, Oliver SG. 2011. Haploinsufficiency and the sex chromosomes from yeasts to humans. *BMC Biol.* **9**, 15.
- Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, Schlesinger F, Davis CA, Reinke VJ, Gingeras TR, Shendure J, Waterston RH, Oliver B, Lieb JD, Disteché CM. 2011. Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat. Genet.* **43**, 1179–1185.

- Deng X, Berletch JB, Ma W, Nguyen DK, Hiatt JB, Noble WS, Shendure J, Distèche CM. 2013. Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. *Dev. Cell* **25**, 55–68.
- Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925.
- Distèche CM. 2012. Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560.
- Doan RN, Shin T, Walsh CA. 2018. Evolutionary Changes in Transcriptional Regulation: Insights into Human Behavior and Neurological Conditions. *Annu. Rev. Neurosci.* **41**, 185–206.
- El-Brolosy MA, Stainier DYR. 2017. Genetic compensation: A phenomenon in search of mechanisms. *PLoS Genet.* **13**, e1006780.
- Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574.
- Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002a. Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872.
- Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Pääbo S. 2002b. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343.
- Ezaz T, Stiglec R, Veyrunes F, Marshall Graves JA. 2006. Relationships between vertebrate ZW and XY sex chromosome systems. *Curr. Biol.* **16**, R736–R743.
- Ezaz T, Quinn AE, Sarre SD, O'Meally D, Georges A, Graves JA. 2009. Molecular marker suggests rapid changes of sex-determining mechanisms in Australian dragon lizards. *Chromosome Res.* **17**, 91–98.
- Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* **14**, 1880–1887.
- Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* **8**, R140.
- Faucillion ML, Larsson J. 2015. Increased expression of X-linked genes in mammals is associated with a higher stability of transcripts and an increased ribosome density. *Genome Biol. Evol.* **7**, 1039–1052.
- Fisher E, Scambler P. 1994. Human haploinsufficiency--one for sorrow, two for joy. *Nat. Genet.* **7**, 5–7.
- Fortelny N, Overall CM, Pavlidis P, Freue GVC. 2017. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20.
- Gaidatzis D, Burger L, Florescu M, Stadler MB. 2015. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729.

- Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448.
- Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, Mazaika E, Vardarajan B, Italia M, Leipzig J, DePalma SR, Golhar R, Sanders SJ, Yamrom B, Ronemus M, Iossifov I, Willsey AJ, State MW, Kaltman JR, White PS, et al. 2014. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ. Res.* **115**, 884–896.
- González-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14**, R70.
- Graves JA. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays* **17**, 311–320.
- Graves JA. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat. Rev. Genet.* **17**, 33–46.
- Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840.
- Gupta V, Parisi M, Sturgill D, Nuttall R, Doctolero M, Dudko OK, Malley JD, Eastman PS, Oliver B. 2006. Global analysis of X-chromosome dosage compensation. *J. Biol.* **5**, 3.
- Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Res.* **27**, 1461–1474.
- Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251.
- Hardison RC. 2016. A guide to translation of research results from model organisms to human. *Genome Biol.* **17**, 161.
- He X, Chen X, Xiong Y, Chen Z, Wang X, et al. 2011. He et al. reply. *Nat. Genet.* **43**, 1171–1172.
- Hershey JW, Sonenberg N, Mathews MB. 2012. Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.* **4**, a011528.
- Hou J, Wang X, McShane E, Zauber H, Sun W, Selbach M, Chen W. 2015. Extensive allele-specific translational regulation in hybrid mice. *Mol. Syst. Biol.* **11**, 825.
- Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154.
- Hughes JF, Skaletsky H, Koutseva N, Pyntikova T, Page DC. 2015. Sex chromosome-to-autosome transposition events counter Y-chromosome gene loss in mammals. *Genome Biol.* **16**, 104.

- Ignatiadis N, Klaus B, Zaugg JB, Huber W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580.
- Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223.
- Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802.
- Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550.
- Ingolia NT. 2014a. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213.
- Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014b. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379.
- Ingolia NT. 2016. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**, 22–33.
- Ingolia NT, Hussmann JA, Weissman JS. 2018. Ribosome Profiling: Global Views of Translation. *Cold Spring Harb. Perspect. Biol.* **p11**, a032698.
- Ishikawa K, Makanae K, Iwasaki S, Ingolia NT, Moriya H. 2017. Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. *PLoS Genet.* **13**, e1006554.
- Jackson R, Standart N. 2015. The awesome power of ribosome profiling. *RNA* **21**, 652–654.
- Jacob F. 1977. Evolution and tinkering. *Science* **196**, 1161–1166.
- Jang C, Lahens NF, Hogenesch JB, Sehgal A. 2015. Ribosome profiling reveals an important role for translational control in circadian gene expression. *Genome Res.* **25**, 1836–1847.
- Janich P, Arpat AB, Castelo-Szekely V, Lopes M, Gatfield D. 2015. Ribosome profiling reveals the rhythmic liver transcriptome and circadian clock regulation by upstream open reading frames. *Genome Res.* **25**, 1848–1859.
- Joazeiro CAP. 2017. Ribosomal Stalling During Translation: Providing Substrates for Ribosome-Associated Protein Quality Control. *Annu. Rev. Cell Dev. Biol.* **33**, 343–368.
- Johnstone TG, Bazzini AA, Giraldez AJ. 2016. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht

- AK, Brady SD, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61.
- Jost, A. 1947. Recherches sur la différenciation sexuelle de l'embryon de lapin. *Archs Anat. Microsc. Morph Exp.* **36**, 271–315 (1947).
- Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, Daish T, Grützner F, Kaessmann H. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol.* **10**, e1001328.
- Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10**, 19–31.
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326.
- Kemp TS. 2005. *The Origin and Evolution of Mammals*. Oxford University Press, Oxford.
- Khaitovich P, Enard W, Lachmann M, Pääbo S. 2006. Evolution of primate gene expression. *Nat. Rev. Genet.* **7**, 693–702.
- Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100–1104.
- Kharchenko PV, Xi R, Park PJ. 2011. Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat. Genet.* **43**, 1167–1169.
- Khurana E, Fu Y, Chen J, Gerstein M. 2013. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* **9**, e1002886.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116.
- Kleene KC. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech. Dev.* **106**, 3–23.
- Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS, Plajzer-Frick I, Lee EA, Kato M, Garvin TH, Akiyama JA, Afzal V, Lopez-Rios J, Rubin EM, Dickel DE, Pennacchio LA, Visel A. 2016. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633–642.
- Lalanne JB, Taggart JC, Guo MS, Herzel L, Schieler A, Li GW. 2018. Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell* **173**, 749–761.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359.

- Laurent JM, Vogel C, Kwon T, Craig SA, Boutz DR, Huse HK, Nozue K, Walia H, Whiteley M, Ronald PC, Marcotte EM. 2010. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10**, 4209–4212.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291.
- Li GW, Burkhardt D, Gross C, Weissman JS. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635.
- Lin H, Halsall JA, Antczak P, O'Neill LP, Falciani F, Turner BM. 2011. Relative overexpression of X-linked genes in mouse embryonic stem cells is consistent with Ohno's hypothesis. *Nat. Genet.* **43**, 1169–1170.
- Lin F, Xing K, Zhang J, He X. 2012. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. *Proc. Natl. Acad. Sci. U S. A.* **109**, 11752–11757.
- Liu Y, Beyer A, Aebersold R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550.
- Liu TY, Huang HH, Wheeler D, Xu Y, Wells JA, Song YS, Wiita AP. 2017. Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Syst.* **4**, 636–644.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.
- Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenäs C, Lundberg J, Mann M, Uhlen M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6**, 450.
- Ma L, Cui P, Zhu J, Zhang Z, Zhang Z. 2014. Translational selection in human: more pronounced in housekeeping genes. *Biol. Direct.* **9**, 17.
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* **102**, 5454–5459.
- Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161.
- Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9270–9274.
- Marcel M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12.

- Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Brüning T, Mössinger K, Trefzer T, Conrad C, Kerver HN, Wade J, Tschopp P, Kaessmann H. 2017. Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome Res.* **27**, 1974–1987.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.
- McManus CJ, May GE, Spealman P, Shteyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24**, 422–430.
- McManus J, Cheng Z, Vogel C. 2015. Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. *Mol. Biosyst.* **11**, 2680–2689.
- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599.
- Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Res.* **23**, 34–45.
- Morris DR, Geballe AP. 2000. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628.
- Munsky B, Neuert G, van Oudenaarden A. 2012. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187.
- Muzzey D, Sherlock G, Weissman JS. 2014. Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Res.* **24**, 963–973.
- Nanda I, Shan Z, Scharl M, Burt DW, Koehler M, Nothwang H, Grützner F, Paton IR, Windsor D, Dunn I, Engel W, Staeheli P, Mizuno S, Haaf T, Schmid M. 1999. 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat. Genet.* **21**, 258–259.
- Necsulea A, Kaessmann H. 2014a. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* **15**, 734–748.
- Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014b. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640.
- Nguyen DK, Disteche CM. 2006. Dosage compensation of the active X chromosome in mammals. *Nat. Genet.* **38**, 47–53.



- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J Sninsky J, Adams MD, Cargill M. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170.
- Ohno S. 1967. Sex chromosomes and sex-linked genes. *New York: Springer-Verlag.*
- Ohno S. 1970. Evolution by gene duplication. *Springer-Verlag Berlin Heidelberg.*
- Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, Kato M, Garvin TH, Pham QT, Harrington AN, Akiyama JA, Afzal V, Lopez-Rios J, Dickel DE, Visel A, Pennacchio LA. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243.
- Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197.
- Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290.
- Peer E, Moshitch-Moshkovitz S, Rechavi G, Dominissini D. 2018. The Epitranscriptome in Translation Regulation. *Cold Spring Harb. Perspect. Biol.* pii, a032623.
- Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295.
- Pessia E, Engelstädter J, Marais GA. 2014. The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? *Cell. Mol. Life Sci.* **71**, 1383–1394.
- Popadin KY, Gutierrez-Arcelus M, Lappalainen T, Buil A, Steinberg J, Nikolaev SI, Lukowski SW, Bazykin GA, Seplyarskiy VB, Ioannidis P, Zdobnov EM, Dermitzakis ET, Antonarakis SE. 2014. Gene age predicts the strength of purifying selection acting on gene expression variation in humans. *Am. J. Hum. Genet.* **95**, 660–674.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596.
- Rancati G, Moffat J, Typas A, Pavelka N. 2018. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49.
- Reilly SK, Noonan JP. 2016. Evolution of Gene Regulation in Humans. *Annu. Rev. Genomics Hum. Genet.* **17**, 45–67.
- Rice AM, McLysaght A. 2017a. Dosage-sensitive genes in evolution and disease. *BMC Biol.* **15**, 78.
- Rice AM, McLysaght A. 2017b. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.* **8**, 14366.

- Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166.
- Rohlfsv RV, Harrigan P, Nielsen R. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Mol. Biol. Evol.* **31**, 201–211.
- Rohlfsv RV, Nielsen R. 2015. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Syst. Biol.* **64**, 695–708.
- Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13**, 505–516.
- Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705.
- Roux J, Liu J, Robinson-Rechavi M. 2017. Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. *Mol. Biol. Evol.* **34**, 2773–2791.
- Schadt EE, Monks SA, Drake TA, Lusk AJ, Che N, Colino V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.
- Schaefer B, Sun W, Li YS, Fang L, Chen W. 2018. The evolution of posttranscriptional regulation. *Wiley Interdiscip. Rev. RNA* **31**, e1485.
- Schafer S, Adami E, Heinig M, Rodrigues KE, Kreuchwig F, Silhavy J, van Heesch S, Simate D, Rajewsky N, Cuppen E, Pravenec M, Vingron M, Cook SA, Hubner N. 2015. Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat. Commun.* **6**, 7200.
- Schmidt EE. 1996. Transcriptional promiscuity in testes. *Curr. Biol.* **6**, 768–769.
- Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley MC, Shini S, Lieschke GJ, Wong JJ, Rasko JEJ. 2017. Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol.* **18**, 216.
- Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmström J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, Aebersold R, von Mering C, Hengartner MO. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* **7**, e48.
- Schuerer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, Roma G. 2017. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics* **18**, 442.
- Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342.
- Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723.

- Shihab HA, Rogers MF, Campbell C, Gaunt TR. 2017. HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics* **33**, 1751–1757.
- Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* **34**, 532–544.
- Singh PP, Arora J, Isambert H. 2015. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput. Biol.* **11**, e1004394.
- Slobodin B, Han R, Calderone V, Vrieling JAFO, Loayza-Puch F, Elkon R, Agami R. 2017. Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* **169**, 326–337.
- Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome Database Group. 2018. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.* **46**, D836–D842.
- Somel M, Liu X, Khaitovich P. 2013. Human brain evolution: transcripts, metabolites and their regulators. *Nat. Rev. Neurosci.* **14**, 112–127.
- Sonenberg N, Hinnebusch AG. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745.
- Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS, Kaessmann H. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190.
- Stadler M, Fire A. 2013. Conserved translome remodeling in nematode species executing a shared developmental transition. *PLoS Genet.* **9**, e1003739.
- Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, Bjornsdottir G, Walters GB, Jonsdottir GA, Doyle OM, Tost H, Grimm O, Kristjansdottir S, Snorrason H, Davidsdottir SR, Gudmundsson LJ, Jonsson GF, Stefansdottir B, Helgadottir I, Haraldsson M, et al. 2014. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366.
- Steinberg J, Honti F, Meader S, Webber C. 2015. Haploinsufficiency predictions without study bias. *Nucleic Acids Res.* **43**, e101.
- Stenberg P, Larsson J. 2011. Buffering and the evolution of chromosome-wide gene regulation. *Chromosoma* **120**, 213–225.
- Tahmasebi S, Sonenberg N, Hershey JWB, Mathews MB. 2018. Protein Synthesis and Translational Control: A Historical Perspective. *Cold Spring Harb. Perspect. Biol.* **pii**, a035584.

- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.
- Tress ML, Abascal F, Valencia A. 2017. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.* **42**, 98–110.
- Turner JM. 2015. Meiotic Silencing in Mammals. *Annu. Rev. Genet.* **49**, 395–412.
- Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, Zeeberg BR, Kane D, Weinstein JN, Blume J, Darnell RB. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37**, 844–852.
- Veitia RA, Birchler JA. 2010. Dominance and gene dosage balance in health and disease: why levels matter! *J. Pathol.* **220**, 174–185.
- Veitia RA, Birchler JA. 2015. Models of buffering of dosage imbalances in protein complexes. *Biol. Direct.* **10**, 42.
- Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. 2010. Young proteins experience more variable selection pressures than old proteins. *Genome Res.* **20**, 1574–1581.
- Völker M, Backström N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* **20**, 503–511.
- Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6**, 400.
- Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232.
- Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**, 1365–1374.
- Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. *Nat. Rev. Genet.* **8**, 921–931.
- Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543.
- Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.
- Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. 2015a. N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**, 1388–1399.

- Wang Z, Sun X, Zhao Y, Guo X, Jiang H, Li H, Gu Z. 2015b. Evolution of gene regulation during transcription and translation. *Genome Biol. Evol.* **7**, 1155–1167.
- Wang SH, Hsiao CJ, Khan Z, Pritchard JK. 2018. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol.* **19**, 83.
- Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time. *Genome Biol. Evol.* **3**, 667–673.
- Warnefors M, Kaessmann H. 2013. Evolution of the correlation between expression divergence and protein divergence in mammals. *Genome Biol. Evol.* **5**, 1324–1335.
- Warnefors M, Liechti A, Halbert J, Valloton D, Kaessmann H. 2014. Conserved microRNA editing in mammalian evolution, development and disease. *Genome Biol.* **15**, R83.
- Warnefors M, Mössinger K, Halbert J, Studer T, VandeBerg JL, Lindgren I, Fallahshahroudi A, Jensen P, Kaessmann H. 2017. Sex-biased microRNA expression in mammals and birds reveals underlying regulatory mechanisms and a role in dosage compensation. *Genome Res.* **27**, 1961–1973.
- Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* **24**, 616–628.
- Weiss M, Schimpf S, Hengartner MO, Lercher MJ, von Mering C. 2010. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* **10**, 1297–1306.
- Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587.
- Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216.
- Xiong Y, Chen X, Chen Z, Wang X, Shi S, Wang X, Zhang J, He X. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet.* **42**, 1043–1047.
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659.
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, et al. 2016. Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716.
- Yin H, Ma L, Wang G, Li M, Zhang Z. 2016. Old genes experience stronger translational selection than young genes. *Gene* **590**, 29–34.

- Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183.
- Zeng C, Fukunaga T, Hamada M. 2018. Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. *BMC Genomics.* **19**, 414.
- Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8**, e1000494.
- Zhang S, Hu H, Zhou J, He X, Jiang T, Zeng J. 2017. Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cell Syst.* **5**, 212–220.