Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Song Guo

PhD Program: Life science

Title of Thesis: Using RNA expression as a quantitative molecular phenotype to study human and vertebrate evolution

Supervisor: Prof. Philipp Khaitovich

The thesis document includes the following changes in answer to the external review process.

In general, I got some comments on the descriptions of figures and I revised the following figure legends: Figure 3,9,10,17,18,19,23,25-27,31-33,44,46-53,55-59. More detailed replies to each comment from referee are listed as follows.

Reviewer 1: Prof. Georgii Bazykin

Q1. In Fig. 2 caption, the content of panel B is not explained; is it the same as in A?

A: Panel A and B are the same but was colored by different aspect. Panel A was colored by sequencing batch, B was colored by human population.

Q2. The discussion of Fig. 5 seems to indicate that panel B is somehow different from the remaining panels; is this really implied? I am not sure how to interpret the differences.

A: Panel B is different from other panels, which are based only on population-associated miRNAs. We use the overall pattern (background) to demonstrate that the significant population-associated miRNAs are not random and indicate that most genetic changes are probably neutral, as suggested in most studies.

I revised it on Page 48 as follows: “Similarly, miRNA expression in African American population differed most from the other three based on analysis of 1,008 expressed miRNAs, but with lower expression divergence.”

Q3. Generally, the descriptions of the figures are not always clear. For example, the fig. 17 caption refers to “factors listed on the x-axis” while this axis is in fact not signed.

A: I revised the following figure legends: Figure 3,9,10,17,18,19,23,25-27,31-33,44,46-53,55-59

Reviewer 3: Prof. Michael Lachmann

Q1: Chapter 3 The samples used for the study are very well chosen, coming from the same location and treated equally. There is still a possibility that population grouping of samples correlates with socioeconomic status, cultural upbringing and others. It would have been nice to stress this influence a bit more.

A: Unfortunately, we did not collect socioeconomic status and cultural upbringing-related information in this study. Also, because we used a limited number of samples, we ignored some interesting factors, like diet habits, due to the limitation of statistical power.

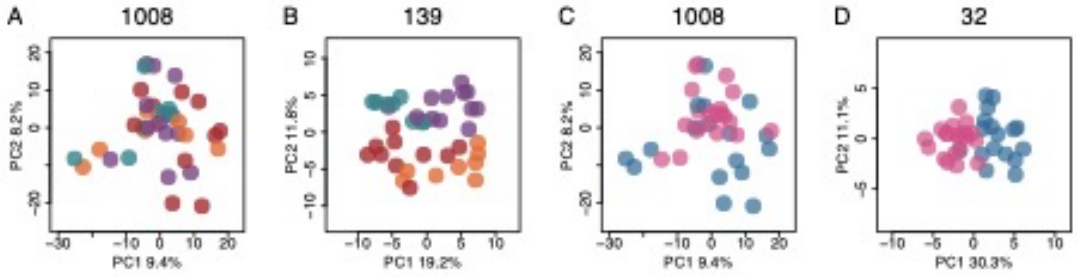
I add more on the discussion part combining with the genetic factors as follows on Page114: "Population has the most substantial influence explaining up to 11% of the total miRNA variance. The overall relative miRNA expression divergence among the four populations studied in the study is similar to their genetic divergence, implying that the presence of single nucleotide polymorphism (SNP) in different miRNA gene regions could be one cause of population difference, as we know that SNP in different miRNA gene regions can affect miRNA function or the efficiency of miRNA biogenesis (Duan, Pak, & Jin, 2007; Sun et al., 2009). We also know that CNV influences gene expression and demonstrates differences between human populations. Using the previous study's list of copy number variable miRNAs, we discovered four miRNAs in CNV regions among our 93 population-associated miRNAs (Marcinkowska, Szymanski, Krzyzosiak, & Kozlowski, 2011). Except for the genetic influence, environments, socioeconomic status, cultural upbringing, and other factors may correlate with population grouping, which we can investigate more precisely in the latter study."

Q2: Page 45: "The results showed population identity and baby's sex are two major factors." Since there is very little difference in % explained variance between baby's sex, first pregnancy, mothers age, BMI, delivery type, I would say "the two largest factors", or maybe say population identity was the major factor, with 7 other factors having a lesser effect, though still above 3% of the variance."

A: I revised it on Page45 as follows: "The results showed that population identity was the major factor, with the other seven factors having a lesser effect, though still above 3% of the variance."

Q3: Page 47, Figure 4. Dots are too big. Is it possible to fix that?

A: I fixed the size of dots and corrected Figure 4 as follows due to a citation mistake.



Q4: Chapter 4 Page 64. "the largest variance (>30%) in gene expression can be explained by different developmental stages (Figure 18)." It is unclear what in figure 18 relates to this statement.

A: I added the sentence on the legend of Figure 18 on Page65 to clarify it as "The proportion of variance explained by each principal component is shown in the axis."

Q5: Page 67. Figure 19, and also in text. I think that calling genes that change their expression in development "develop-related" genes is a bit confusing. That label implies that they control development, not that they are controlled by development. Maybe "development responsive" genes?

A: To keep the consistency of the whole thesis, I changed "development-related gene" to "dynamically changed genes" in the entire text.

Q6: Page 84: "most (85%) precursors are in intronic or intergenic regions" it would be nice to state if this number is consistent with expectations, maybe based on relative size of genomic regions."

A: Main miRNA biogenesis determines that most precursors are in the intronic or intergenic regions. I add a reference to show the consistent expectations on Page83: "precursors are in intronic or intergenic regions, which has been widely observed (Olena & Patton, 2010). The result is also consistent across the eight species."

Q7: Page 87: This is quite a complicated question, with a complicated setup and analysis. I had a hard time keeping track for each analysis which miRNA are used, organ-specific, how many developmental stages, how many species, etc. Maybe a schematic overview that also shows that different analyses done would be beneficial.

A： I use a schema instead of the phylogenetic tree in Figure 33.

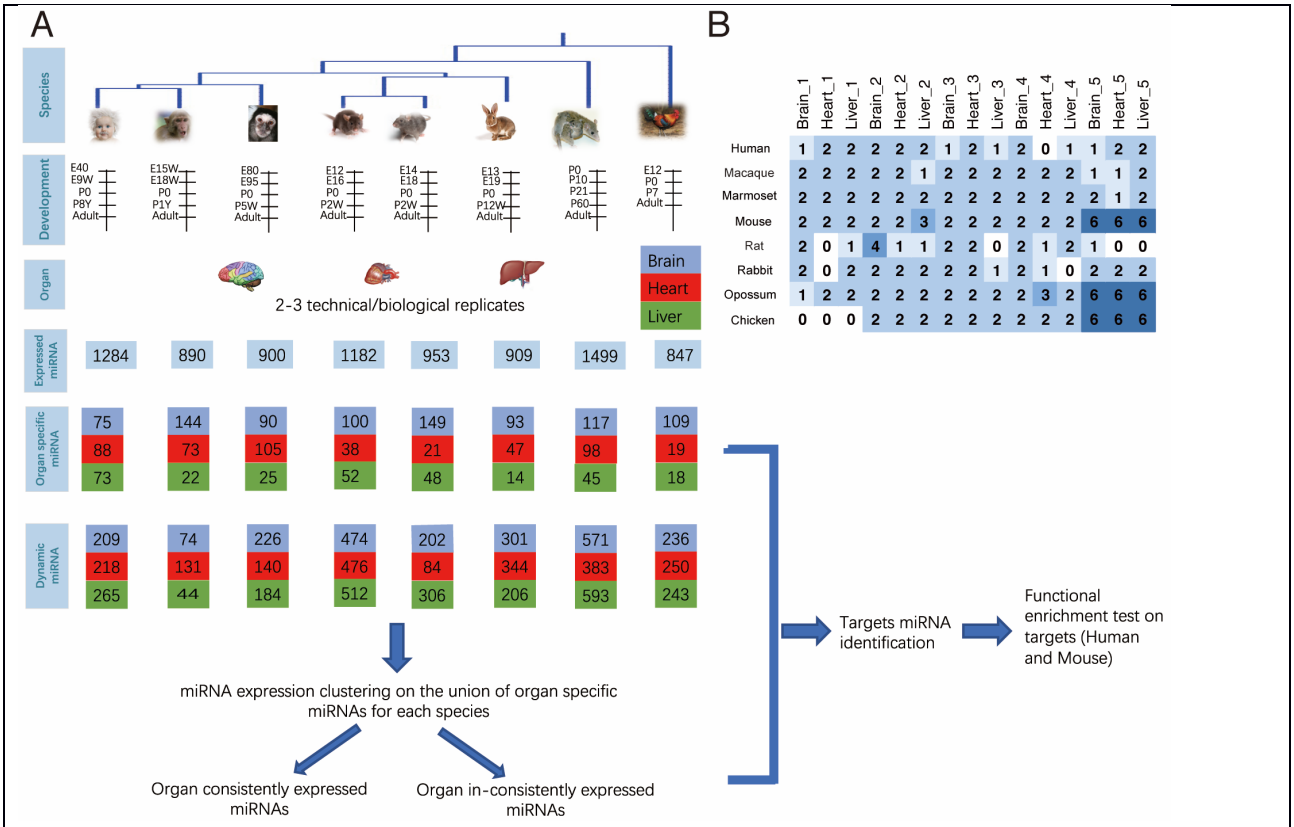


Figure 1 Species phylogeny tree and sample number for each stage

A. Workflow of the study with the illustration of species phylogeny tree, developmental stage. The number on each box represents the number of miRNAs in each identified category. **B.** Sample number on each stage for each tissue. The number of biological samples on each developmental stage of each organ. 1-5 represents the developmental stage from early embryonic development to adulthood.

Q8: Page 88: It would be nice to expand upon the statement: "This observation is consistent with the notion suggested by Chen and Rajewsky, postulating that most organ-specific miRNAs might function to minimize deleterious off-targeting effects and to allow natural selection to eliminate slightly deleterious targets over evolution"

A: I revised this paragraph on Page92-94 as follows, combining with the comments from Prof. Mehmet. Additional analysis based on weak miRNA-target relationships was considered now. "To explore if these organ-specific miRNAs are biologically relevant to organ development, we applied Metascape to perform a functional enrichment test on the set of organ-specific miRNA targets. We used miRNAtap with at least three methods to predict human and mouse miRNA targets, which we then combined with the expression of target genes from Cardoso-Moreira's dataset (Cardoso-Moreira et al., 2019). However, compared to random miRNA-non target interactions, we only found weak repression on miRNA targets for human liver, heart, and mouse liver-specific miRNAs (bootstrap the non-target genes 100 times for each miRNA; One-side Wilcoxon test; $p<0.05$;Figure 46). Further, we defined the genes with the strong repression (Spearman correlation rho< -0.5) as potential targets for organ-specific miRNAs. Overall, we found weak enrichment in GO ontology terms and KEGG pathways on strong down-regulated miRNA targets. However, we discovered that the targets downregulated by mouse liver specific miRNAs are significantly enriched in

biological functions such as dendrite development, brain development, axon guidance, and the Wnt signaling pathway, which makes sense given that those targets are highly expressed in the mouse brain and heart. To see if the lack of relevant function was due to a small number of targets, we included weakly repressed miRNA targets, defined as genes with a reverse correlation of less than 0.1. We would then discover more biological significance, but it would not be conserved between human and mouse, except for the targets of liver-specific miRNAs, which demonstrate biological relevance on brain development in both Humans and Mouse (Figure 47). Our observation shows the majority of targets predicted for each miRNA are only weakly repressed, as reported in previous studies (Baek et al., 2008; Hausser & Zavolan, 2014; Selbach et al., 2008), and we can observe some functional miRNA regulatory effects on a phenotypic consequence from weakly repressed miRNA targets. However, we cannot evaluate the precise regulatory function of miRNAs because we found no significant repression effects on targets and didn't find the highly functional overlaps on targets between Human and Mouse. It could imply that organ-specific miRNAs are more likely to target multiple genes by chance, with possible deleterious consequences (Berezikov, 2011)

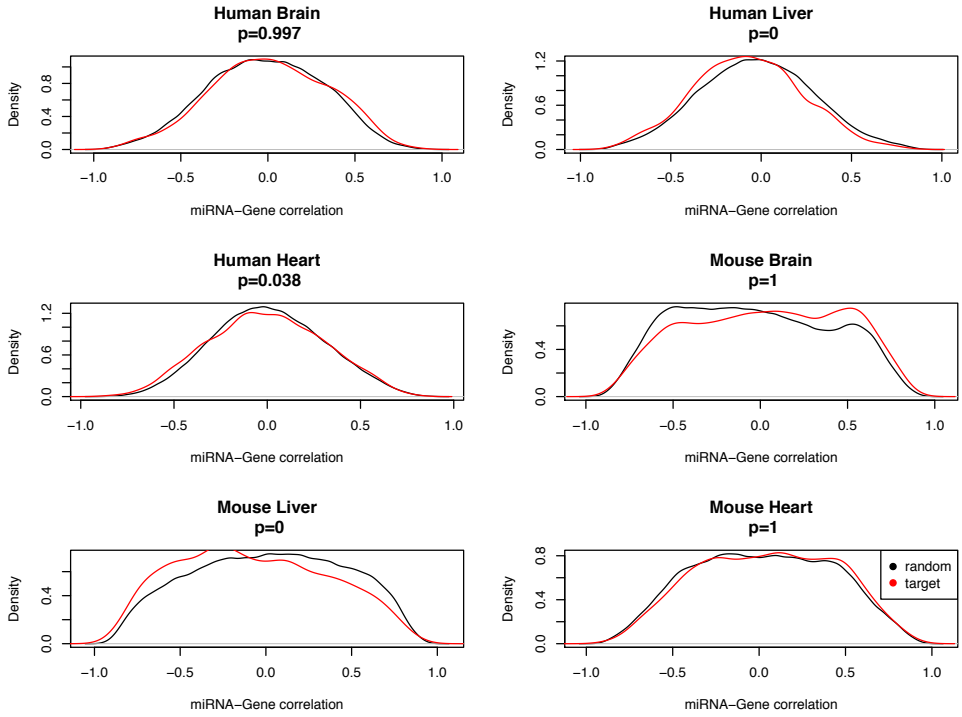


Figure 2. Distributions of miRNA-target correlations

The red line shows the distribution of Spearman correlations between miRNA-target pairs relative to miRNAs and non-target pairs. The black lines depict the random expectation estimated by the union of 100 bootstraps of miRNA-non-target pairs. The p-value based on the one-side Wilcoxon test is listed on the top of each panel. The X-axis indicates the miRNA-gene correlation.

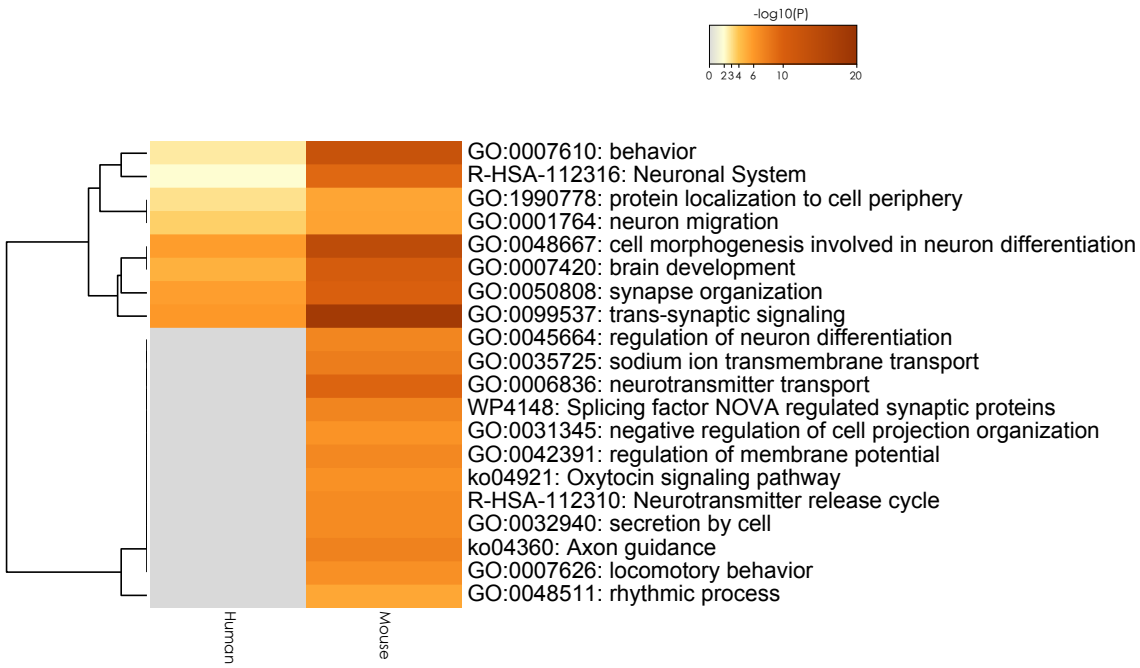


Figure 3. Gene Ontology terms enriched in the targets of liver-specific miRNAs.

Selected terms with the best p-value within each cluster, as the representative terms, are displayed in a dendrogram. The heatmap cells are colored by their enrichment p-values; light cells indicate the lack of enrichment for that term in the corresponding species. The significant terms are hierarchically clustered into a tree based on Kappa-statistical similarities among their gene memberships. 0.3 kappa score was applied as the threshold to cast the tree into term clusters."

Q9: Page 91, Figure 44: make sure that each of the histograms has 5 bars, even when proportion is 0. (unless some species are missing stages for some tissues, in which case this should be emphasized more.)

A: I revised the Figure44 on Page90 as follows:

Figure 4. Developmental persistence of organ-specific miRNAs.

Colors indicate the number of stages each organ-specific miRNA was detected in each tissue of each species. 1-5 represents the number of stages each miRNA expressed. Each bar represents the percentage of organ-specific miRNAs.

Q10: Page 94, figure 47: Why isn't this displayed in terms of fractions, since the text on page 93 refers to fractions?

A: I revised the Figure47 (now is Figure49) to display the fractions and mark the number of miRNA on each bar on Page96.

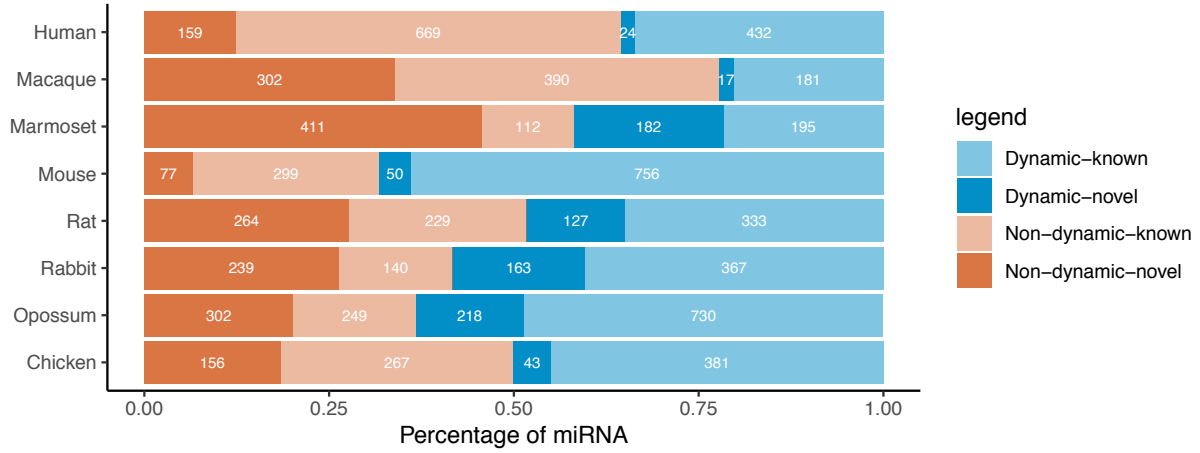


Figure 5. Proportion of total dynamically changed miRNAs.

Blue represents the proportion of dynamically changed miRNAs; Orange represents non-dynamically changed miRNAs; Dark color represents novel miRNAs; light color represents known

miRNAs. The number of miRNAs is marked on each bar. The number is the union of dynamic changes at least on one organ.

Q11: Page 95. "we used non-organ-specific miRNAs with dynamically changed at least one organ", I think it should be "we used non-organ-specific miRNAs which dynamically changed at least one organ"

A: Corrected

Q12: Page 95: The lower part of this paragraph is very unclear. Too much information compressed into too little text. Break it up, explain more.

A: I add more description to explain my idea on Page99 as follows:
"To assess whether organ-consistent miRNAs are more ancient with the importance of some origination mechanisms, we used a BLAST-based phylostratigraphic technique to calculate the evolutionary age of miRNA for each species. Each miRNA was allocated to a phylostratum, representing the miRNA's most ancient phylogenetic node. We used 0 as the most ancient phylogenetic node, which origins from Zebrafish and Fugu."

Q13: Page 98: "whether miRNAs with the same seed, which has the same regulatory functions as usually assumed," I didn't understand this sentence Page 98: Figure 53. Explain this figure more. What is the difference between horizontal and vertical, what is being compared, what can the reader see?

A: I add more description to explain my idea on Page100 as follows:
"Action of miRNA, as an essential posttranscriptional regulator, could be associated with the degradation or translational inhibition of mRNA by binding to the 3' UTR of target genes. This process is dominated by the seed (7-8nt) region of the miRNA. A common thought is miRNAs with similar seed target similar sets of genes and thus similar sets of biological functions or pathways."

Q14: Page 99, 100: text on page 99 says 55A is increasing, but figure legend says A is decreasing.

A: In the text, it says, "the potential targets of organ-consistently increasing miRNAs, while the legend of Figure 55A (now Figure 57A) says miRNA target with decreasing expression pattern, which were identified by the gene expression anti-correlated with miRNA.

Q15: Page 100, Figure 54 is unclear. What do the arc colors indicate? What kind of overlap is plotted?

A: The legend is revised as follows: "

Figure 6. The circus plots showing miRNA target genes overlap between human and mouse by Metascape

On the outside, each arc represents the identity of each gene list. On the inside, each arc represents a gene list, where each gene has a spot on the arc. Dark orange represents the genes that appear in multiple lists, and light orange represents genes that are unique to that gene list. Purple lines link

the same gene that are shared by multiple gene lists. Blue lines link the different genes where they fall into the same ontology term. The greater the number of purple links and the longer the dark orange arcs the greater is the overlap of miRNA target genes among human and mouse. Blue links indicate the amount of functional overlap among the miRNA target genes of human and mouse. **A**. miRNA target genes showing decreasing expression pattern among all three organs. **B**. miRNA target genes showing increasing expression pattern among all three organs."

Q16: Page 101, Figure 55. Remove top trees. Two species always show the same tree. There should be more explanation of what the reader is to infer from the figure. Top is a bit darker, does that say anything? Columns are more similar? What do trees say? Again, increasing, decreasing, organ consistent, etc, are very confusing. It might be good to also have an overall schematic of the analyses.

A: Top trees were removed on Figure57-59

Reviewer 4: Mehmet Somel

Suggestions on Discussion related to Chapter 3 results: The reason for population differences in miRNA expression level in the placenta could be interesting to discuss. How likely are these environmental, purely genetic, or both, given what we know about the social groups in question. I think the Discussion could also benefit from some suggestions on child sex-related miRNAs and the distinct functions of their target genes. The text implicates hormones driving miRNA-bias (previous observation), but the observation made with the placenta dataset is that biased-miRNA expression is driving hormonal response. How these are exactly linked could be more extensively discussed.

A: I add more discussion on Page115 as follows:
"Furthermore, 32 female and male newborn-associated miRNAs were shown to have regulatory interactions with several biological processes. However, the targets of these two categories of miRNAs both have hormone-related functions, as evidenced by the findings of a recent study indicating sex hormones can influence miRNA expression. It's worth mentioning, however, that placental tissue can secrete steroid hormones, estrogen, and human placental lactogen, which could skew some of our findings. We could investigate the regulatory functions adjusting hormone related miRNA targets in the future. "

I found the hypothesis that cell lines and placenta differ in the amount of evolutionary constraint on their transcriptomes and this causes population-specific miRNAs to be more prevalent in the former very interesting. Could one think of ways of testing this idea? For instance, perhaps it could be tested by comparing the overall preservation of coding transcriptomes in cell lines versus placenta. An alternative (and not mutually exclusive) hypothesis would be not about cell lines but the tissue origin. Blood tissue, which is the source of lymphoblasts, could possibly be subject to more local adaptation events? Perhaps this could be discussed further.

A: Thanks for this comment, but I don't have very clear idea on it now. I will discuss it with my colleagues further.

Suggestions on Discussion related to Chapter 4 results:Among mammals, brain-specific miRNAs are found to be more common in this work. This may be expected given the brain's tissue complexity and perhaps other factors. But human brain-specific miRNAs are an exception and occur at a modest proportion among organ-specific miRNAs in this dataset. Is this a biological result or a technical fluke, e.g. in the PCA in Fig 46 a human liver sample clusters with brain samples, any may be obscuring brain-specific expression.

A: I would guess the biological problem on sample collection would cause this exception, because we could find the same trend on gene expressions. As we known, we have the difficult to collect human samples, especial for the sample before birth. On the other hand, I have only one biological sample for human brain on stage1,3, and 5, therefore, it's not possible to check whether there was technical problem. But this could be improved when add more biological samples and developmental stages.

The observation regarding genes with decreasing expression levels during embryogenesis being more conserved than genes with increasing expression is nice. Again, I felt there could be further discussion on the reasons for this. For instance, this could perhaps be tied to evolutionary constraints under pleiotropy?

A: I add more discussion on Page116 as follows: " One possible explanation is that genes with a conserved descending pattern have a higher proportion of older genes. It has been postulated that older gene promoters have DNase I hypersensitive sites (DHSs) at earlier developmental stages, allowing gene expression to be active at earlier stages (Gao et al., 2018). Another reason could be that those genes are subject to gene pleiotropy constraints. We found that 88% of mouse genes with descending expression profiles expressed at many stages, but only 48% of mouse genes with ascending expression profiles did. Other vertebrates have made similar observations."

Finally, I much liked the discussion on why negative correlation between miRNA and predicted target mRNA expression levels were frequently not found (p114). One suggestion/question would be to use correlations between miRNA and mRNA clusters (instead of individual mRNA) to increase power in such analysis? Perhaps this may have been already explored -I just wanted to share as a suggestion.

Suggestions on methodology in both Chapters 3 and 4:In general, I felt that more details about methodology could be useful, as I could not follow how some of the results were obtained. At a number of points, why certain methodological choices were made could be more clearly explained. These include the use of polynomial regression to model RNA-seq data but linear regression for miRNA-seq data; why the gray-module was removed from co-expressed gene modules; etc.

A: I add more paragraphs on Page 108 on the methodology part and some minor revises

4.2.6 Gene expression interpretation to mouse developmental stage

To compare the similarity of the expression profiles across developmental stages of nine species, we used the predicted developmental stage alignment presented in Figure 22 to create a

unified alignment of eight species to the mouse developmental curves. To do so, we mapped 33 stages cumulatively interpolated from eight species to the full mouse developmental curve fitted using cubic smoothing spline with ten degrees of freedom. We then compared gene expression curves among nine species based on z-transformed expression of each gene interpolated at these 33 stage points.

4.2.7 Clustering of gene expression profiles in six vertebrate and nine chordate species

To investigate the expression pattern diversity in nine or six species, we used hierarchical clustering (hclust function in R) of z-transformed gene expression trajectories aligned among species with (1 - rho) as the distance measure, where rho is the Spearman correlation coefficient. We chose k equal six, as optimal, based on visual inspection of clusters obtained using different k values.

4.2.8 Clustering all co-expressed gene modules in nine distal species

To obtain complete transcriptomic information across all species, we combined 103,728 dynamically changed mouse-orthologous genes to identify the co-expressed modules by WGCNA with the parameter corType as "biocor" in 21 different modules. The gray module containing 176 genes, which cannot be clustered in any module, was removed for further analysis. Then, the GO and KEGG Pathway Enrichment Test were applied to the expressed genes of each species in each module.

Q1: However, it is estimated that no less than 85% of genetic variation within human populations might be neutral (Lewontin & Krakauer, 1973). are you sure this paper makes such an estimate? in any case, because this paper was about polymorphism in coding loci, it may be better to use a more recent reference for global variation, like some 1000 Genomes papers?

A: I revised this with more recent results as follows on page24:
"However, some genetic variations may not affect fitness or phenotypic variation. 1000 genome project has found around 12,000 polymorphic variants in protein-coding region and up to 565,000 variants in the sites of known regulatory regions per genome. Only around 2,000 variants are associated with complex traits or rare diseases in European ancestry samples, but not in other populations (Genomes Project et al., 2015; O'Huallachain, Karczewski, Weissman, Urban, & Snyder, 2012)"

Q2: *reported 2% to 26% genes with differential expression between population pairs (Lappalainen et al., 2013). perhaps would help to mention that the reported percentages actually reflect statistical power and threshold used, so may not be directly comparable.*

A: It was revised on page24 as follows: "A recent study on five populations (89-95 individuals per populations), including CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI), reported 2% to 26% genes with differential expression between population pairs, requiring genes with FDR<0.05 and log2 fold change greater than two (Lappalainen et al., 2013)."

Q3: *Page25 "still leaving 35% of the differences as potentially affected influenced by positive selection" not very clear - consider revising the sentence*

A: It was revised as follows: "however, still 35% of the differences were considered as potentially affected by positive selection."

Q4: *Page28 "most miRNAs are substantially involved in the maintenance of cell differentiation rather than tissue identity" how is this inferred?*

A: I clarify my idea on page29 as follows:" A study showed that ancient miRNA expression is conserved at multiple tissues between protostomes and deuterostomes, including the central nervous system, sensory tissue, muscle, and gut (Christodoulou et al., 2010). It indicated that ancient miRNAs have the functions to establish tissue identity. We guess that most miRNAs might be substantially involved in maintaining cell differentiation rather than tissue identity."

Q5: *Page30. "Assess the relationship between development and evolution"* a bit vague "development and phylogenetic divergence"?

A: It was revised as follows on page30: "Assess the relationship between development and phylogenetic divergence at evolutionary homologous developmental stages among distant species."

Q6: *Page35. "Changes in gene expression level during the developmental process may correspond to developmental time and tissue composition changes" , can it not involve simply loss of expression of a gene in a cell type? that would be not directly related to timing, nor to composition…*

A: It was revised as follows on page35: "Changes in gene expression level during the developmental process may correspond to developmental time, tissue composition changes, cell type compositions and other influences"

Q7: *Page37 often introduce pattern bias when have been analyzed using other computational methods. have been analyzed using other computational methods.*

A: It was revised on page37 as follows: "However, the results often introduce pattern bias when the species studied have distant phylogenetic distances.". I deleted the statement" when have been analyzed using other computational methods" because it was really irrelevant to the context.

Q8: *Page38. "speciation could be considered as a long-term environmental adaptation until some traits fit in a population" this is not clear to me. I don't think speciation can be considered an adaptation directly.do you wish to say something like "Different lineages, in the long-term, accumulate specific adaptations that increase their fitness, and some of these would be reflected as divergence"?*

A: Corrected

Q9: *Page39. "with high conservation" all of them "with high conservation"?*

A: "with high conservation" was not described clearly here, I removed it and revised it on page39 as follows:
"miRNAs are expressed in both plants and animals and are essential for regulating gene expression by facilitating mRNA degradation or translational repression"

Q10: Page46." To reduce other interaction effects" i didn't understand the reasoning here

A: The description was not clear, I did not consider any interaction effects on the model1 based on ANOVA and wanted to estimate the effect of each variable more precisely. So I changed it as follows: "To assess the variables' effects more precisely" on page46.

Q11: Page46. "The four continuous variables modeled first were: maternal body mass index, birth weight of the newborn child, mother's age, and birth length of the child. " why these 4 and not others?

A: Because only these four are continuous variables, I again estimated the effect on other factors using ANOVA.

Q12: *Page46. "the most notable contributors to the miRNA expression variation". but the model does not include other variables, e.g. "first pregnancy". how do you then decide these are the "most notable contributors"?*

A: Figure 3 did not display the effect of all factors, but we did include the other variables. In the model2 only four continuous variables were reduced.

Q13: *Page46 "mean of the variance proportion" again, if this is mean among all 1000 mirnas, please mention. also, would be good to explain how the SD is calculated*

A: The legend of Figure 3 was revised as follows:" Bars represents the mean of the variance proportion explained by the factors across 1008 miRNAs. Error bars represent the standard deviation of the variance."

Q14: Page47: "consistent with their genetic divergence" perhaps say "as estimated using the 1000 Genomes data set (Methods)"

A: It was revised as follows: "The relative miRNA expression divergence among four populations investigated in the study is consistent with their genetic divergence as estimated using the 1000 Genomes data set"

Q15: *Page48 "miRNA expression in African Americans was the most distant from the other populations", would be good to discuss the possible reasons environment, or genetics?*
e.g. did the different population samples show difference in their BMI or other characteristics?

A: Yes, we think genetic diversity may be one potential causal to miRNA expression variance based on 1000 genome data. But we don't have clear evidence due to a lack of genetic data. In addition, the mean BMI of African and South east Asian is higher than the other two populations.
I add more discussion on genetic effect on discussion part of page115 "Population has the most substantial influence explaining up to 11% of the total miRNA variance. The overall relative miRNA expression divergence among the four populations studied in the study is similar to their genetic divergence, implying that the presence of single nucleotide polymorphism (SNP) in different miRNA gene regions could be one cause of population difference, as we know that SNP in different miRNA gene regions can affect miRNA function or the efficiency of miRNA biogenesis (Duan, Pak, & Jin, 2007; Sun et al., 2009). We also know that CNV influences gene expression and demonstrates differences between human populations. Using the previous study's list of copy number variable miRNAs, we discovered four miRNAs in CNV regions among our 93 population-associated miRNAs (Marcinkowska, Szymanski, Krzyzosiak, & Kozlowski, 2011). Except for the genetic influence, environments, socioeconomic status, cultural upbringing, and other factors may correlate with population grouping, which we can investigate more precisely in the latter study."

Q16: Page56 "driving mechanism of miRNA sex-biased expression"driving mechanism or target? your results suggest they're the target, no?

A: Thanks for this comment. Yes, my results suggest the targets are in the hormone involved functions, such as response to steroid hormones and response to estradiol, and I did not check how many targets are hormones, which is worth checking in the future.

Q17: *Page57." largest characterized imprinted miRNA cluster, located on chromosome 19 (C19MC) and expressed almost exclusively in the placenta " imprinted on both maternal and paternal sides, or just one?*

A: C19MC is only active on the paternally inherited allele

Q18: Page57. "Significant negative correlation with the mother's BMI"did other miRNAs not show such correlation? did the different population samples show difference in their BMI?

A: Half of miRNAs show a negative correlation with the mother's BMI (cor<0), but miRNAs in the C19MC cluster show a more significant negative correlation with BMI, shown in the comparison in Figure 15A.

Q19: Page64: the largest variance (>30%) you mean PC1s? PC2's are also related to devo state - you could sum their percentages as well, perhaps?

A: It was revised on page63 as follows:" Similarly, based on principal component analysis of all samples for each species, a large variance (>50%) in gene expression can be explained by different developmental stages (Figure 18; the sum of the variation from PC1 and PC2)."

Q20:Page67 "and functionally conserved between species"
how was this understood? if it was a result from Hu et al 2017, perhaps say more explicitly?

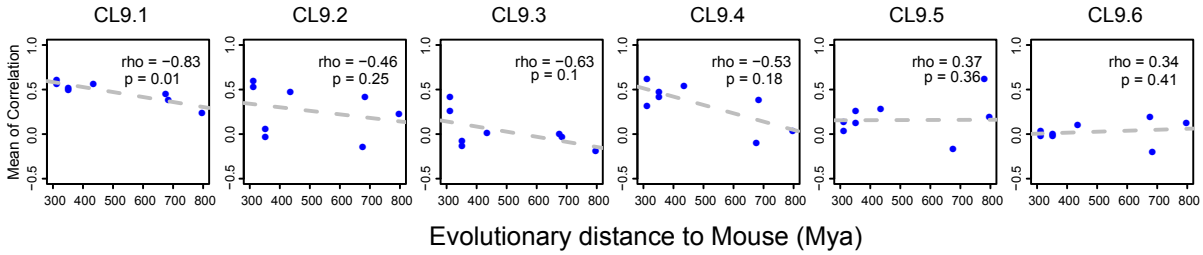
A: I revised the paragraph on page 66 as follows: "We defined genes preferentially expressed at a particular developmental stage as stage-associated genes in each species following the method described in (H. Hu et al., 2017; J. J. Li et al., 2014). The stage-associated genes are potentially related to organ development, such as growth and reproduction in early embryonic stage. For each species, we required the gene on the stage with FPKM (fragments per kilobase of exon model per million reads mapped) > 2 and Z-score > 1.5. These criteria guaranteed that the genes are more highly expressed in one stage than in others and are expressed distinguishable from the background. On average, we identified 65%-90% of expressed genes as stage-associated genes for given species."

Q21:Figure 21 i didn't understand how this is related to Fig 20, perhaps more information in the text or figure legend could help?

A: Figure21 repeat the same approach as in Figure 20 and uses random genes instead of stage-associated genes to test gene similarity in their transcriptome characteristics.

Q22:Figure 25 i'd suggest also including info about the significance of the correlations. it makes little sense that rho>0, and those cases are probably not significant (random) anyway

A: I revised the figure as follows:



Q23: Page75: "all 103,728 genes from nine species"i didn't exactly get this: this 103,000 include the sum of genes from all 9 species?e.g. FOXP2 in mouse and FOXP2 in chicken are counted as 2?perhaps explain a bit more?

A: Yes, for example, FOXP2 in mouse and FOXP2 in chicken are counted as two genes because even ortholog genes have different expressions in each species. I add more explanation on the legend of Figure 27 as follows: "Figure 7. Clustering of developmental gene expression profiles based on all 103,728 genes with annotated mouse orthologs. These 103,728 genes are the union of all dynamically changed mouse genes and mouse orthologs."

Q24:Page77 consistent with previously reported results that evolutionarily older genes tend to be expressed at earlier developmental stages i didn't understand where you show this result

A: The genes with descending expression pattern (CL20.2+CL20.3) were enriched on housekeeping genes shown in Figure 29; here, I used housekeeping genes to indicate older genes. But I did not check sequence conservation or gene age here.

Q25: Figure 30 sorry but i didn't understand the analysis steps here.
do you analyze data from each species separately (if so, perhaps indicate how many genes were included in each genes), cluster them (all or development-related ones?), and perform GO enrichment on the clusters (all clusters or some?)?

A: The test in Figure 30 is the same as in Figure 28 but performs GO enrichment test instead. The test was performed on each cluster for each species independently. Here we specifically displayed the results only on GO terms related to "development" and show that "development-related" biological functions are not enriched on the genes with similar expression profiles in multiple species.

Q26: P80 genes expressed in three mouse Organs, just "genes expressed" or genes in CL8.2 in Fig 31?

A: expressed genes in CL8.2 in Figure 31.

Q27: Figure37. it could have been easier for explaining if you had turned the age-scale upside down, so that "age" increases with time. perhaps the age-scale could also be normalized?

A: Age 0 indicates the miRNA first appeared at the branch of zebrafish. 0-12 each indicates a different phylogeny branch, so the age scale can not be normalized. The corresponding phylogeny and age are listed in supplementary Table S5.

Q28: Page93 35%~68% of miRNAs show dynamic changes in at least one organ for each species. Only 22% of macaque miRNAs show dynamic changes. why only macaque? perhaps discuss briefly?

A: It was revised on page95 as follows: "Only 22% of macaque miRNAs show dynamic changes (Figure 49). We also observed that there are fewer dynamic changed mRNAs than other species. But we did not find significant difference in RNA quality and sequencing reads coverage. We guess that may be due to improper sample collection."

Q29: Page98 but no significant expression profile conservation for miRNAs with organ-inconsistent developmental expression patterns (Figure 53B). in fact there is some overlap between liver-up mirnas in different species, right? could one test this using a statistic for clustering and a permutation test?

A: Yes, we can see some overlap between species, but there was no significant overlapping for any pair of organ-inconsistent miRNAs when I did fisher's exact test pair wisely

Q30: Page105 nominal p-value <0.01 (permutation p-value< 0.05) were both criteria applied, or at least one?

A: nominal p-value <0.01 and permutation p-value< 0.05 are exact same criteria

Q31: Page104. it would also be good to explain why you use different models (polynomial vs linear) for RNA-seq vs miRNA-seq data

A: It was revised on page106 as follows: "For each gene, we selected the best regression model with 11-17 developmental stages (by rank)" and "We defined developmental changes in miRNA expression levels using linear regression models because there are only 3-5 developmental stages available for miRNA."

Q32: Page111 Newly emerged organ-specific miRNAs do not play evident regulatory roles in developmental regulation in the corresponding organ,
Organ-specific miRNAs not showing clear regulatory roles - this is deduced from lack of GO enrichment, is that right? You could instead say "miRNAs not showing clear regulatory roles focused on specific functions". Because the current statement suggests there is no regulatory effect at all. This could also be tested by lack of negative correlation between miRNA and predicted targets, but I think you don't do that.

A: I revised my observation on page113 as follows:
"Organ-specific miRNAs, which trend to be evolutionary young, do not show clear regulatory roles focused on specific functions in organ development, consistent with proposed gradual evolution of regulatory engagement of newly evolved miRNAs."

And I revised the paragraphs on Page92-94, adding the analysis on the weak negative correlation between miRNA and predicted targets.

Q33: Page113 a group of genes with gradually increasing developmental expression was substantially less conserved among species and more enriched in tissue specific genes.
Could you perhaps tie this observation (decreasing genes higher conserved than increasing genes) to evolutionary constraints under pleiotropy?

A: I add some discussions under your suggestion on page155:
" One possible explanation is that genes with a conserved descending pattern have a higher proportion of older genes. It has been postulated that older gene promoters have DNase I hypersensitive sites (DHSs) at earlier developmental stages, allowing gene expression to be active at earlier stages (Gao et al., 2018). Another reason could be that those genes are subject to gene pleiotropy constraints. We found that 88% of mouse genes with descending expression profiles expressed at many stages, but only 48% of mouse genes with ascending expression profiles did. Other vertebrates have made similar observations."

Reviewer 5: Prof. Shuhua Xu
Q1: The thesis title sounds to me is a too big topic, is there some way to make it more specific?
I understand the RNA data are the focus of the thesis in terms of computational analysis and view angle of the questions, however, I would be happy to read some more discussions on the impact or influence of genetic variations. I believe it is applicable given the DNA-level data are also available for some tissues studied in the thesis.

A: I agree with your comments, the title of this thesis is really big. But it may be hard to change now. Discussions related on genetic variations are added as follows:
On page114: “Population has the most substantial influence explaining up to 11% of the total miRNA variance. The overall relative miRNA expression divergence among the four populations studied in the study is similar to their genetic divergence, implying that the presence of single nucleotide polymorphism (SNP) in different miRNA gene regions could be one cause of population difference, as we know that SNP in different miRNA gene regions can affect miRNA function or the efficiency of miRNA biogenesis (Duan, Pak, & Jin, 2007; Sun et al., 2009). We also know that CNV influences gene expression and demonstrates differences between human populations. Using the previous study's list of copy number variable miRNAs, we discovered four miRNAs in CNV regions among our 93 population-associated miRNAs (Marcinkowska, Szymanski, Krzyzosiak, & Kozlowski, 2011). Except for the genetic influence, environments, socioeconomic status, cultural upbringing, and other factors may correlate with population grouping, which we can investigate more precisely in the latter study.”

On Page119: “Furthermore, in my study, I oberseved that there are less overlap of tissue-speicfic miRNAs and its regulatory function among vetebrates. Similar observation were found in a study of teleost fish investigating gene expression variation among tissues, individuals, and populations. The results showed that half of the genes were differentially expressed among individuals within a population-tissue group and that only a small subset (31%) of tissue-specific differences were consistent across all three populations. It suggested that many tissue-specific differences in gene expression are unique to a single population and are therefore unlikely to contribute to fundamental distinctions between tissue types (Whitehead & Crawford, 2005). However, both of studies did not look at the effect of genetic diversity on gene and miRNA expression levels across tissues. An analysis of copy number variation (CNV) in multiple somatic tissues from six unrelated people revealed a significant amount of intraindividual genomic variation between tissues, demonstrating that somatic tissues can be genetically diverse. 79% of these events have an impact on genes (O'Huallachain et al., 2012). The Genotype-Tissue Expression (GTEx) project that includes genotype, gene expression, histological and clinical data for 449 human individuals across 42 distinct tissues discovered that cis-acting genetic variants tend to affect either most tissues or a small number of tissues (G. T. Consortium et al., 2017). Accoordingly, to understand the causal of phenotype difference among tissues, individuals, or populations, we would integrate genetic information in the future studies.”

Q2: Since the sample size is often not large for the data used in the thesis, which I understand the situation, some evaluation of the statistical power for different sample size might be helpful. Is there any batch effects in the data given they might be generated from different batches, technical platforms, with different reagent, or different time?

A: Batch and other technical factors need to be considered in the data process. In part 3.2.2, I described how to design samples for sequencing and how to remove batch effects. However, I did not describe mRNA data procession in detail for the study in Chapter 4 because I utilized the processed expression data from previous publications. But before I started my analysis, I evaluated the potential factors on expression according to the information I had. For miRNA evo-devo data, all samples of eight species are randomly arranged in different sequencing runs, and some biological samples have two different reagents. However, I did not observe the batch effects here. Overall, my

analysis is more interested in the expression profiles along with development instead of absolute expression value. Some bias could be avoided by the library preparing and sequencing randomly.

Q3: The thesis is overall well-written, but there is still some room to improve the language. I would suggest these issues can be addressed after the candidate do oral defense.