Supplementary Information for "Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility"

Itai Yanai, Leonid Peshkin, Paul Jorgensen, and Marc W. Kirschner

SUPPLEMENTAL DATA

Figure S1. Microarray data reproducibility and microarray design, related to Figure 1. a) Estimating the reproducibility of the microarray data using technical replicates. Two samples, each of 1ug, from the same total RNA (one *X. laevis* stage 23 embryo) were taken for amplification, cy3 labeling, microarray hybridization, and normalization. The correlation coefficient between the two microarrays is 0.991. **b)** Properties of the microarray designs. Distribution of GC-content of the 43,803 60-mer probes on both microarrays. **c)** Additional properties of the microarray probes. Shown are the distributions of melting temperature, folding score, complexity score, and 3'-position score as given by the OligoWiz2 (Wernersson and Nielsen 2005).

Figure S2. Figure S2. Controls for estimating microarray quality, related to Figure 2. a) Comparison of the microarray data with a previously published set. The Baldessari et al. dataset (Baldessari, Shin et al. 2005) on *X. laevis* was compared for the 2,974 genes common to both in stage 13. We find a correlation of R=0.57 among the datasets. Genes were matched using *X. laevis* accession IDs. **b)** Control for different temperatures. Since *X. laevis* and *X. tropicalis* embryos were isolated in different temperatures (22°C and 28°C, respectively) we tested whether this might introduce a confounding effect. We isolated stage 10 *X. tropicalis* embryos for both temperatures and examined their transcriptomes. A principle components analysis revealed that the Stage 10 embryos cluster together regardless of temperature. The little variation that is observed in the Stage 10 transcriptomes across the second principle component is not greater than that seen in the previous stage. Furthermore the second component captures far less of the variation across these samples than the first, 22% and 57% respectively. We conclude that the difference in temperatures does not confound the cross-species comparisons. The principle components analysis was computed on the set of dynamically expressed genes (see Experimental Procedures).

Figure S3. Expression divergences across all genes and in specific pathways, related to Figure 3. a) Distribution of ED_i for the examined genes. The expression divergence index (ED_i) of all 11,095 orthologs was computed as described in Experimental Procedures. **b)** Expression profiles of members of the Membrane attack complex (MAC). *X. tropicalis* (green) and *X. laevis* (blue) profiles are shown. For each set of profiles the t_1's are indicated. **c)** The mean expression profiles of ribosomal genes in both species. **d)** Expression profiles of the 'heterochronic pathway' gene *lin-41* and *lin-28*.

Figure S4. Heterometry in a developmental pathway, related to Figure 4. a) Relationship between the differences in 3'UTR lengths and differences in gene expression levels. For the 150 genes with the most ED_i differences (Table S1), 26 had known mRNAs for both *X. laevis* and *X. tropicalis* in Refseq (Pruitt and Maglott 2001). The length of the 3'UTR was computed as the differences between the length of the transcript and the location of last exon. The difference in expression level was computed as the difference in the sums of the orthologous profiles. The plot shows for each ortholog pair the difference in 3'UTR as a function of its change in expression level. While for the extreme 3'UTR differences there was an

associated difference in detected expression level, presumably due to more efficient reverse transcription at the start of the amplification protocol (see Experimental Procedures), the dominant majority of the differences in expression levels are not associated with a correlated difference in 3'UTR length. **b)** Expression profiles of the core members in the hedgehog signaling pathway. Profiles are shown in the same format as Fig. 1b.

Figure S5. Supplemental analyses regarding the convergence of the developmental transcriptome, related to Figure 5. In same format as Fig. 5d: **a)** The number of genes with a maternal profiles (66 genes, cluster is shown in dark blue in Fig. 1c) that differ at each stage. **b)** The number of genes different at each stage between *X. laevis* clutch 1 and *X. laevis* clutch 3. **c)** The number of genes different at each stage between *X. tropicalis* clutch 1 and *X. tropicalis* clutch 3.

Table S1. Expression divergence index (ED_i) for all examined orthologs, related to Figure 3. The columns correspond to ensemble ID, ED_i, gene symbol, and description. This table is located in a separate document.

Table S2. Comparative expression profiles for genes with conserved, heterochronic, heterometric, convergent patterns, related to Figures 1, 4, and 6. More information regarding each gene is given in Table S1. This table is located in a separate document.

Table S3. Gene sets with low expression divergence between species, related to Figure 3

Gene set	Mean ED_i relative to mean of all genes	$-\log_{10}(p)$	number of genes
structural constituent of ribosome	0.25	16	56
ribosomal proteins	0.29	17	63
mRNA splice site selection	0.30	3	11
intercalated disc	0.30	2	4
protein amino acid lipidation	0.31	4	13
lipoprotein biosynthetic process	0.31	4	13
RNA dependent atpase activity	0.33	3	13
lipoprotein metabolic process	0.34	3	15
ATP dependent RNA helicase activity	0.35	3	12
protein targeting to mitochondrion	0.35	2	6
RNA helicase activity	0.35	4	19
translation initiation factor activity	0.36	2	15
carbohydrate transmembrane transporter activity	0.36	2	6
mRNA binding activity	0.37	2	8
translational initiation	0.38	2	15
TCA	0.41	2	12
proteasome pathway	0.42	2	19
regulation of cellular pH	0.42	2	7
cellular monovalent inorganic cation homeostasis	0.42	2	7
proteasome	0.44	2	15
intrinsic to endoplasmic reticulum membrane	0.44	3	17
integral to endoplasmic reticulum membrane	0.44	3	17
HDAC pathway	0.48	2	21
ATP dependent helicase activity	0.51	2	20
glycerophospholipid biosynthetic process	0.52	2	17
translation factors	0.52	4	37
mitochondrial membrane part	0.53	3	39
intrinsic to organelle membrane	0.53	3	36
integral to organelle membrane	0.54	3	34
aguirre pancreas	0.54	2	33
phosphoinositide biosynthetic process	0.56	2	15
outer membrane	0.56	2	19
glycerophospholipid metabolic process	0.56	2	25
protein RNA complex assembly	0.58	3	39
phospholipid biosynthetic process	0.58	2	21
organelle outer membrane	0.58	2	18
organelle inner membrane	0.59	3	55
ribonucleoprotein complex biogenesis and assembly	0.59	2	50
ubiquitin mediated proteolysis	0.59	2	20
MAPK cascade	0.59	2	22
mitochondrial inner membrane	0.60	3	48

mitochondrial envelope	0.60	3	69
mitochondrial membrane	0.60	3	63
regulation of gene specific transcription	0.60	2	8
mRNA processing	0.60	3	37
yagi aml prognosis	0.60	2	21
translation	0.61	6	93
helicase activity	0.61	3	38
mRNA splicing	0.61	4	41
circadian exercise	0.61	2	34
membrane lipid biosynthetic process	0.63	2	27
RNA binding	0.63	8	167
mootha voxPhos	0.64	2	55
Eif2 pathway	0.65	2	8
envelope	0.67	3	123
organelle envelope	0.67	3	123
aguirre pancreas chr1	0.69	2	24
ribonucleoprotein complex	0.69	4	91
mitochondrial part	0.69	3	102
RNA processing	0.70	3	121
Toll pathway	0.71	2	25
RNA splicing	0.71	4	72
mRNA metabolic process	0.72	2	55
ATPase activity	0.72	2	76
response to light stimulus	0.72	2	23
protein catabolic process	0.73	2	53
mRNA processing go 0006397	0.73	2	52
organelle membrane	0.74	4	205
ATPase activity coupled	0.74	2	64
macromolecule biosynthetic process	0.74	6	173
mRNA processing reactome	0.74	3	90
nucleolus	0.76	3	71
structural molecule activity	0.77	8	123
RNA polymerase ii transcription factor activity	0.77	2	115
basso regulatory hubs	0.78	2	91
endomembrane system	0.79	2	143
generation of precursor metabolites and energy	0.80	2	78
mitoDB 6 2002	0.81	4	282
mitochondrion	0.81	2	225
endoplasmic reticulum	0.81	3	173
mitochondria	0.82	4	298
flotho casp8ap2 mrd diff	0.82	2	53
RNA metabolic process	0.83	3	537
nuclear part	0.84	3	367
macromolecular complex	0.84	4	588
tarte plasma blastic	0.85	2	228
intracellular organelle part	0.86	3	748
organelle part	0.86	3	752
pgc	0.86	2	242
protein complex	0.87	2	507
nucleobase nucleoside nucleotide and nucleic acid metabolic process	0.87	2	783
nucleus	0.87	3	887
nuclear lumen	0.87	2	238
organelle lumen	0.87	2	287
membrane enclosed lumen	0.87	2	287
cellular biosynthetic process	0.88	2	181
jison sicklecell diff	0.89	3	202
cellular protein metabolic process	0.89	2	653
pyrophosphatase activity	0.90	2	145
biosynthetic process	0.91	2	267

Table S4. Gene sets with high expression divergence between species, related to Figure 3.

Gene set	Mean ED_i relative to mean of all genes	$-\log_{10}(p)$	number of genes
glucocorticoid mineralocorticoid metabolism	4.13	3	4
alternative pathway	3.19	3	5
c21 steroid hormone metabolism	3.17	3	6
oxygen binding	3.00	2	7
icosanoid metabolic process	2.91	3	9
PLCD pathway	2.86	2	3
acetaminophen pathway	2.62	2	2
gpcrs class c metabotropic glutamate pheromone	2.41	2	4
gpcrdb class c metabotropic glutamate pheromone	2.41	2	4
riboflavin metabolism	2.19	2	6
SHH pathway	2.09	2	9
kinase activator activity	2.04	2	6
oxidoreductase activity go 0016705	2.03	2	24
regulation of axonogenesis	1.99	2	6
nicotinate and nicotinamide metabolism	1.97	2	9
serine type endopeptidase inhibitor activity	1.94	2	11
protease inhibitor activity	1.92	2	14
hdaci colon cluster7	1.90	2	10
fatty acid metabolic process	1.88	3	31
regulation of neurogenesis	1.87	2	8
carm1 pathway	1.82	2	9
mouse tissue kidney	1.75	2	3
prostaglandin synthesis regulation	1.70	2	19
histone acetyltransferase activity	1.67	2	11
ross fab m7	1.62	2	32
chemical pathway	1.53	2	17
il10 pathway	1.53	2	6
cell cycle checkpoint ii	1.49	2	6
negative regulation of hydrolase activity	1.48	3	10
insoluble fraction	1.47	2	9
gluconeogenesis	1.47	2	34
glycolysis	1.47	2	34
negative regulation of map kinase activity	1.44	2	12
contractile fiber	1.44	2	16
contractile fiber part	1.44	2	16
bystrykh hsc cis glocus	1.40	2	74
zhan mm cd138 cd2 vs rest	1.37	2	18
passerini adhesion	1.33	2	23
jison sickle cell	1.33	2	21
carboxylesterase activity	1.32	2	14
positive regulation of map kinase activity	1.31	2	22
deaminase activity	1.29	2	7
magnesium ion binding	1.29	2	36
regulation of map kinase activity	1.29	2	36
iritani adprox vasc	1.29	2	92
hsiao liver specific genes	1.28	3	121
bystrykh hsc brain cis glocus	1.28	2	36
negative regulation of catalytic activity	1.27	3	43
regulation of hydrolase activity	1.25	2	41

myosin complex	1.23	2	9
caspase activation	1.23	2	14
digestion	1.23	2	14
zpositive regulation of caspase activity	1.19	2	15
receptor binding	1.19	2	146
chesler brain highest variance genes	1.18	2	14
MAPKKK cascade go 0000165	1.17	2	52
cytoskeletal protein binding	1.16	2	90
ray p210 diff	1.14	2	33
spindle pole	1.14	2	11
ichiba gvhd	1.13	2	126
integral to plasma membrane	1.13	2	370
intrinsic to plasma membrane	1.13	2	373
cytoskeletal part	1.12	2	130
ion binding	1.11	2	148
nuclear import	1.07	2	31
protein import into nucleus	1.07	2	31
cytoskeleton	1.07	2	202
identical protein binding	1.07	2	197
passerini signal	1.06	3	194
extracellular region	1.06	2	163
butanoate metabolism	1.03	2	19
spindle	1.01	2	26
extrinsic pathway	1.01	2	9

Table S5. Maternal transcriptome conservations between species, related to Figure 6.

Gene set	Mean ED_i relative to mean of all genes	$-\log_{10}(p)$	number of genes
TCA	0.144238	2	12
Ribosome Biogenesis And Assembly	0.179121	2	11
Aston Oligodendroglia Myelination Subset	0.209767	2	8
Nucleotide Sugar Metabolic Process	0.216904	2	8
Human Tissue Liver	0.228056	3	15
mRNA Binding Activity	0.236843	2	8
Smooth Muscle Contraction Go 0006939	0.26813	2	6
Lian Myeloid Diff Receptors	0.324017	2	10
Hdac Pathway	0.371615	2	21
Coenzyme Metabolic Process	0.402755	2	25
Proteasome pathway	0.415485	2	19
Krebs TCA Cycle	0.418889	2	24
Basal Lamina	0.426465	2	11
Structural Constituent Of Ribosome	0.472452	2	56
Cellular Protein Catabolic Process	0.528254	2	44
Protein Catabolic Process	0.54397	2	53
mRNA Splicing	0.558102	2	41
Ribosomal Proteins	0.562756	2	63
Ribonucleoprotein Complex	0.571653	2	91
Translation Factors	0.580485	2	37

Table S6. Maternal transcriptome divergences between species, related to Figure 6.

Gene set	Mean ED_i relative to mean of all genes	$-\log_{10}(p)$	number of genes
Acetaminophen pathway	6.28658	2	12
Glucocorticoid Mineralocorticoid Metabolism	5.837434	2	11
C21 Steroid Hormone Metabolism	4.857145	2	8
Oxygen Binding	3.885952	2	8
Msppathway	3.62922	2	15
Metabotropic Glutamate Gaba B Like Receptor Activity	3.558569	2	8
Icosanoid Metabolic Process	3.307843	2	6
Sa Mmp Cytokine Connection	3.232155	2	10
Kinase Activator Activity	3.21782	3	21
Monooxygenase Activity	2.940834	2	25
Oxidoreductase Activity Go 0016705	2.767711	2	19
Cell Substrate Adherens Junction	2.757285	2	24
Focal Adhesion	2.757285	2	11
Prostaglandin Synthesis Regulation	2.726806	2	56
Rankl pathway	2.689907	2	44
Rho Guanyl Nucleotide Exchange Factor Activity	2.43933	2	53
Uvb Nhek1 C4	2.345788	2	41
Tall1 pathway	2.246784	2	63
Exocytosis	2.13868	3	91
Positive Regulation Of Immune Response	2.134036	2	37

SUPPLEMENTAL EXPERIMENTAL PROCEDUES

Xenopus orthology. *X. tropicalis* sequences for 18,025 genes based upon the genome sequencing project (Hellsten, Harland et al. 2010) were retrieved from Ensembl database (Hubbard, Aken et al. 2009). In the absence of a draft genomic sequence for *X. laevis* we used 35,523 sequences from Unigene and 39,724 sequences from the TIGR indices. We searched for significant blastn alignments among these DNA sequences of ≥100bp, an E-value ≤ 10^{-15}, and spanning ≥30% the length of the shorter aligned sequence. We then searched for a bi-directional best hit (BBH) between each *X. tropicalis* gene (*Xt*) and found such relationships for 10,719 genes. 376 *Xt* genes did not have a BBH but the aligned region was determined to be unique by additional searches and these genes formed additional clusters. These 11,095 clusters with sequences from either species form the primary group of clusters examined in the manuscript. 1,656 *Xt* genes were identified as in-paralogs (more similar to an *Xt* gene than the best aligning *Xl* gene) and were added to the 356 clusters they matched. For the remaining 5,151 *Xt* genes, we did not find an *Xl* hit, or (for 136 genes) a weak hit with 5% identity worse than the alignment with the BBH between the *Xl* gene and *Xt* gene it hits best. This group of 5,151 genes formed 5,015 clusters based upon close similarities (≥95% identity) form the secondary group of clusters.

Microarray probe design. Each cluster is associated with one sequence for either species. In clusters with members from both species (primary group) the sequences are based upon the bi-directional best hit pair. In clusters of the secondary group, the *Xt* gene sequences is used for both species. Sequences were exon separated to avoid assigning probes spanning splice sites that are more prone to misannotation. The exonized *Xl* sequence was inferred by joining *Xl* sequences that aligned to the *Xt* Ensembl exons by blastn. We next identified probes using the OligoWiz2 software (Wernersson and Nielsen 2005). Each probe was selected based upon its score: $S_{probe} = \Delta T_m w_{\Delta T_m} + F w_F + C w_C + P w_P + S w_S$, where S_{probe} is

the score of a 60-mer probe, ΔT_m is the melting temperature score that favors probes with a melting temperature closest to the array average, *F* is a folding score that penalizes probes likely to undergo folding, *C* is a low-complexity score that penalizes probes with common subsequences, *P* is a position score that favors probes closer to the 3' end of the transcript, *S* is a similarity score that penalizes probes that are not also similar to any in-paralogs. The weights were set to: $w_{\Delta T_m} = .25; w_F = .15; w_C = .15; w_P = .25; w_S = .2$. The *Xt* and *Xl* probes were paired and to each an average score was assigned. Probes were then allocated to clusters based upon their scores. Wherever probes pairs scores were similar those that did not overlap were selected. Primary and secondary clusters were assigned three and two probes, respectively, on the 43,803 probe microarray. To fill the microarray, 192 secondary clusters were assigned an additional probe. Figure S9 provides the properties of the probes.

Open-access data browsers. We developed a web-based browser for the gene expression dataset with URL http://kirschner.med.harvard.edu/Xenopus_Transcriptomics.html that provides a summary of expression profiles, averaged over the three clutches, where the two curves reflecting *X. laevis* and *X. tropicalis* profiles are superimposed for direct comparison. Genes are referred to by a 5-digit identifier which is the suffix of the Ensembl gene ID of the *X. tropicalis* ortholog. For example, a gene named "SSH" is identified as 13504 since its ID is ENSXETG00000013504. Since annotations for many Xenopus genes are often incomplete and sometimes inaccurate, it may be necessary to match by sequence similarity the query sequence to the *X. tropicalis* genome.

SUPPLEMENTAL REFERENCES

Baldessari, D., Y. Shin, et al. (2005). "Global gene expression profiling and cluster analysis in Xenopus laevis." Mech Dev **122**(3): 441-75.

Hellsten, U., R. M. Harland, et al. (2010). "The genome of the Western clawed frog Xenopus tropicalis." Science **328**(5978): 633-6.

Hubbard, T. J., B. L. Aken, et al. (2009). "Ensembl 2009." Nucleic Acids Res **37**(Database issue): D690-7.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res **29**(1): 137-40.

Wernersson, R. and H. B. Nielsen (2005). "OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes." Nucleic Acids Res **33**(Web Server issue): W611-5.

Supplemental Figures

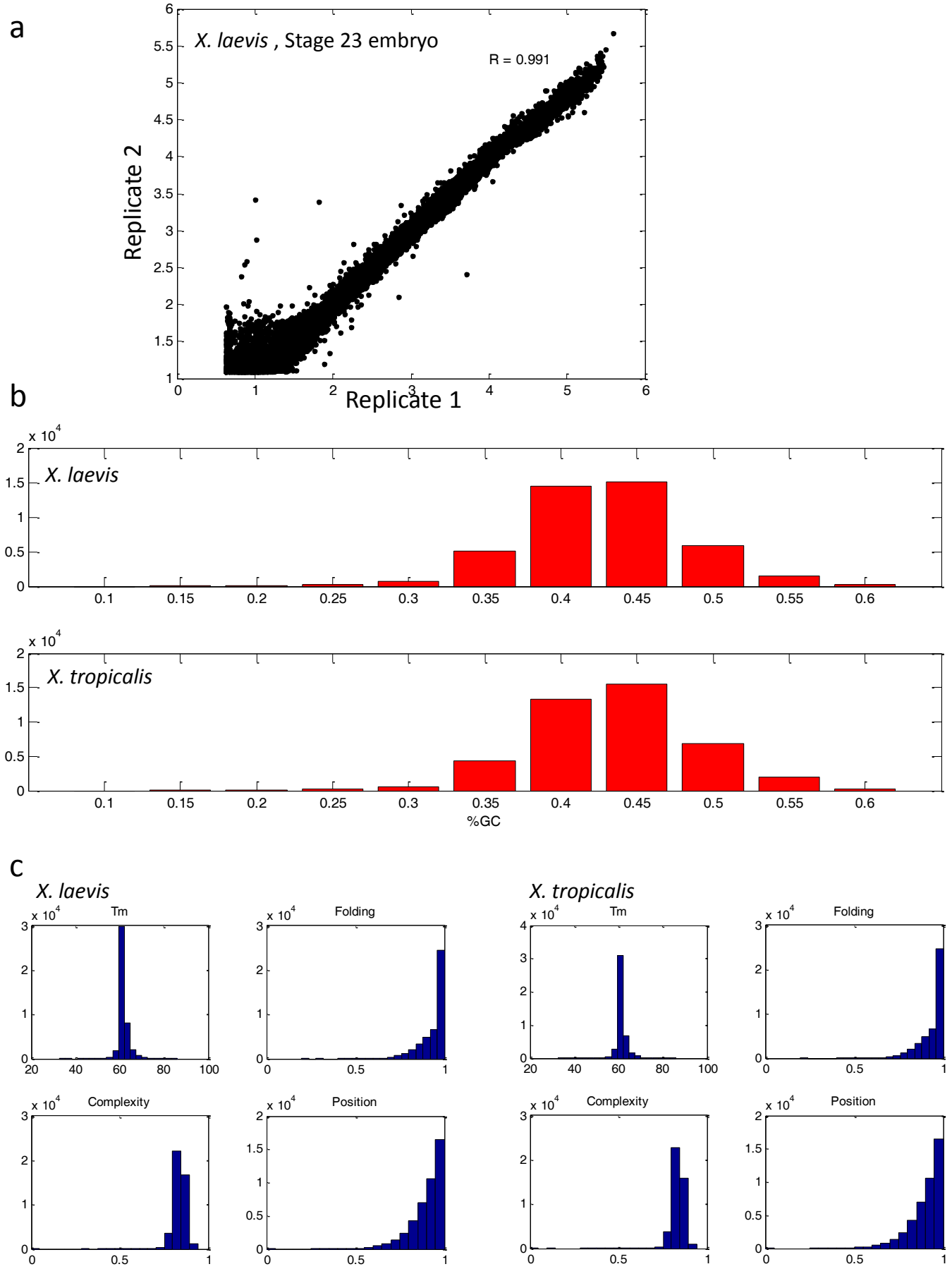
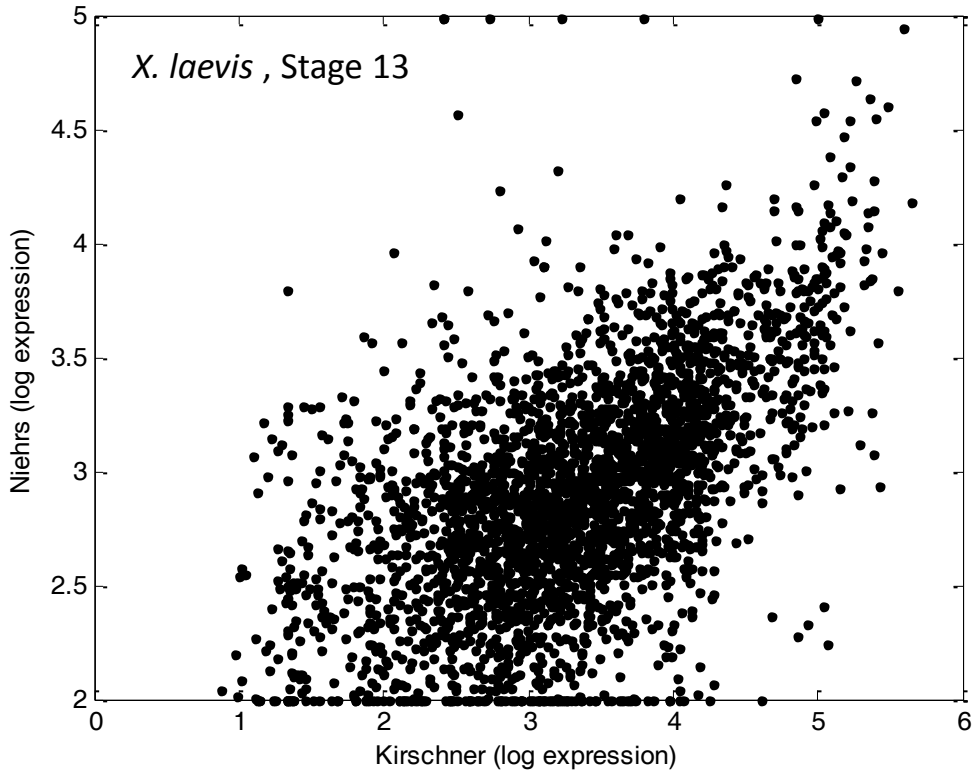


Figure S1

a


X. laevis , Stage 13
Niehrs (log expression)
Kirschner (log expression)
5
4.5
4
3.5
3
2.5
2
0
1
2
3
4
5
6

b

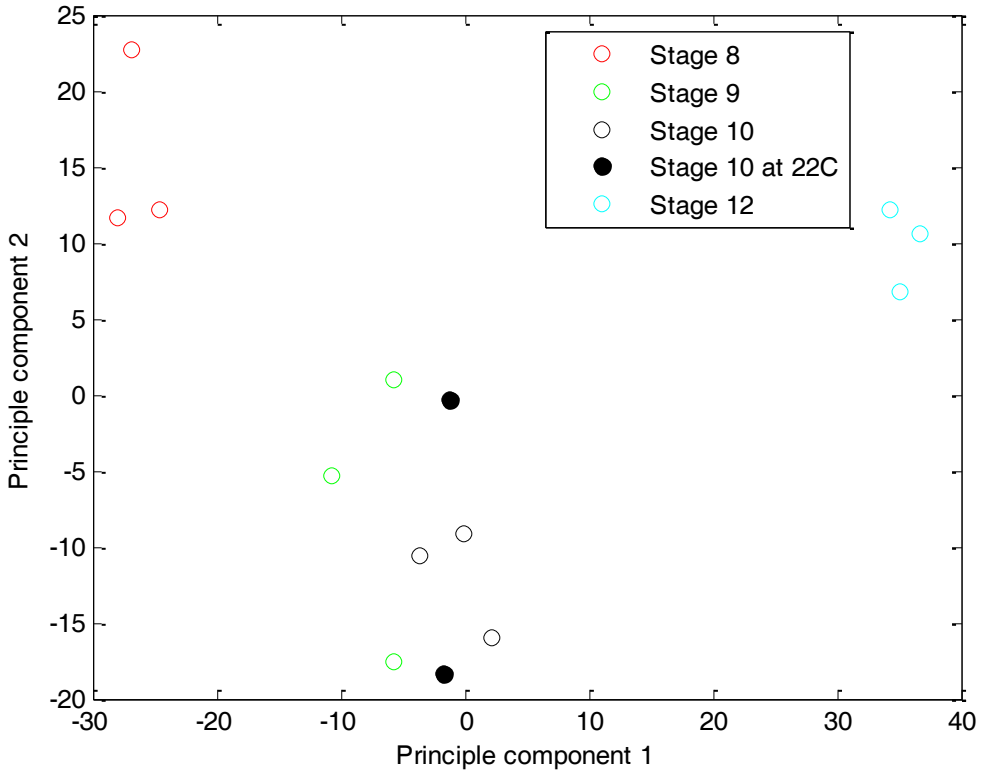
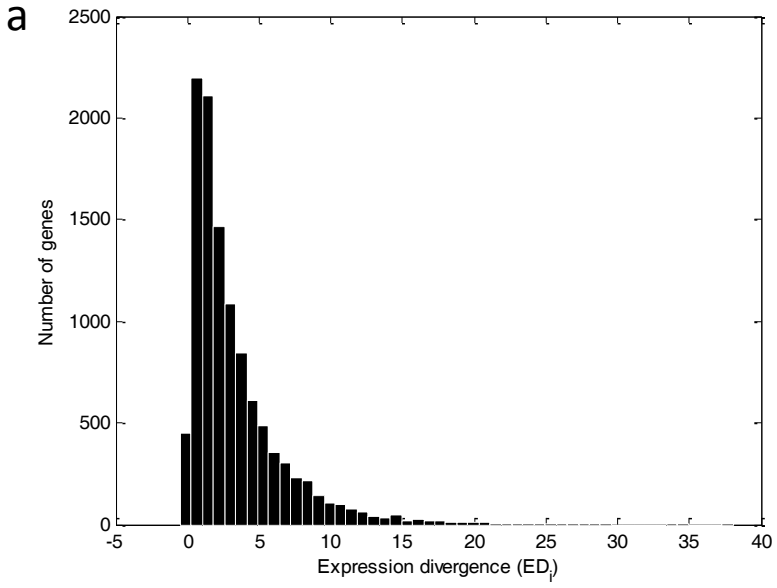
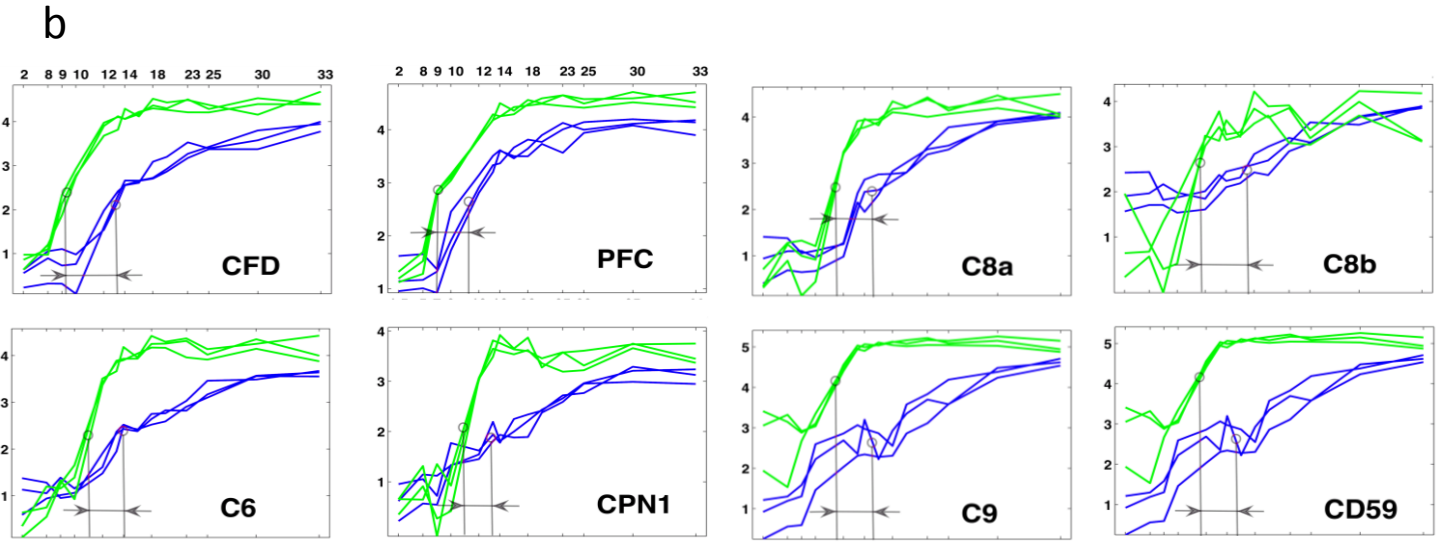

Stage 8
Stage 9
Stage 10
Stage 10 at 22C
Stage 12
Principle component 2
Principle component 1
25
20
15
10
5
0
-5
-10
-15
-20
-30
-20
-10
0
10
20
30
40

Figure S2


a
Number of genes
Expression divergence (ED$_i$)


b
CFD
PFC
C8a
C8b
C6
CPN1
C9
CD59

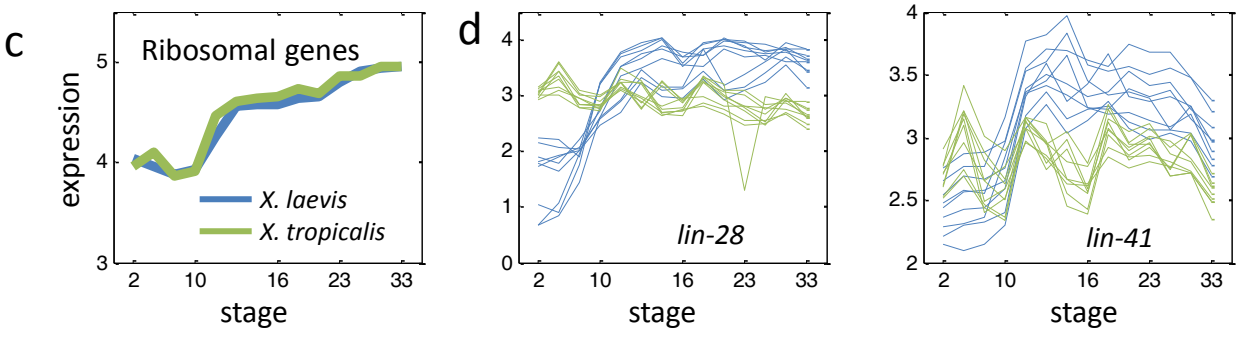
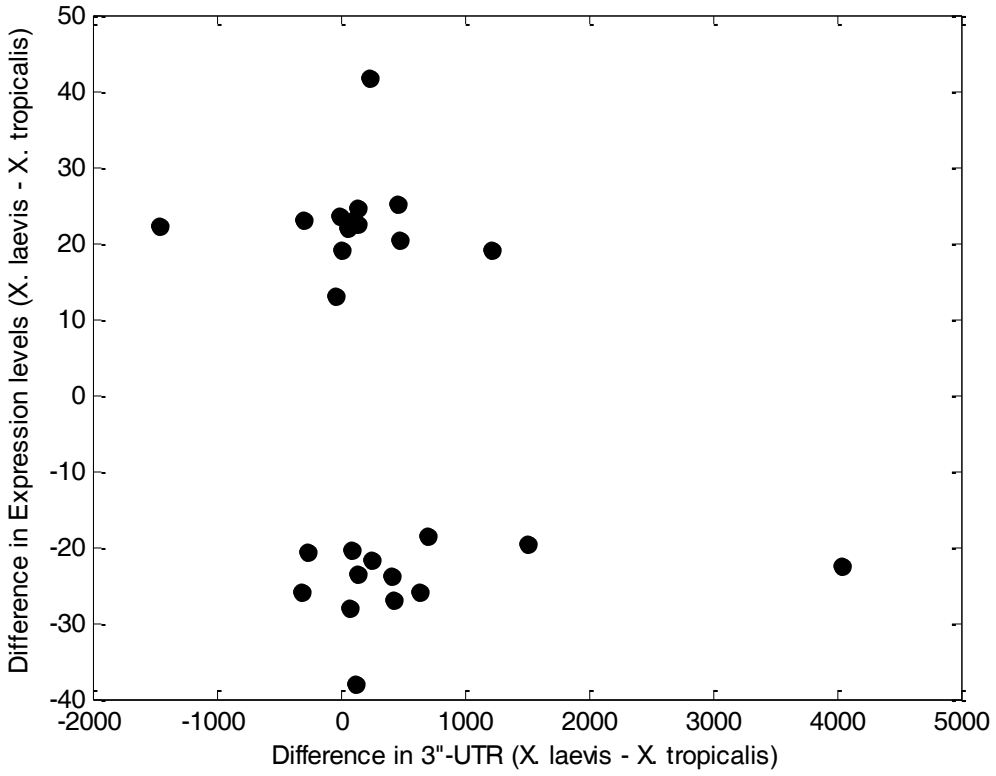

c
Ribosomal genes
expression
X. laevis
X. tropicalis
stage
d
lin-28
lin-41
stage
stage

Figure S3

a


Difference in Expression levels (X. laevis - X. tropicalis)
Difference in 3"-UTR (X. laevis - X. tropicalis)
50
40
30
20
10
0
-10
-20
-30
-40
-2000
-1000
0
1000
2000
3000
4000
5000

b

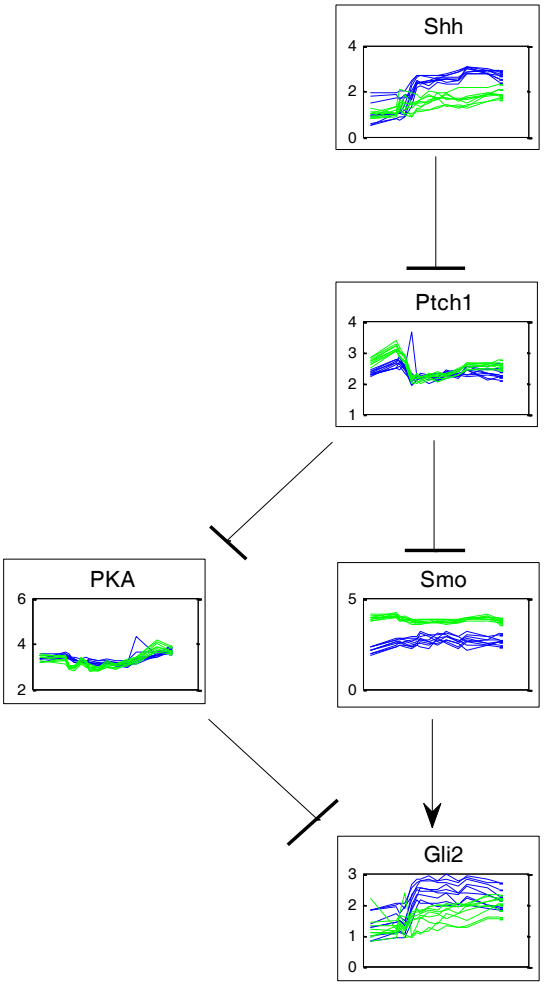

Shh
Ptch1
PKA
Smo
Gli2

Figure S4

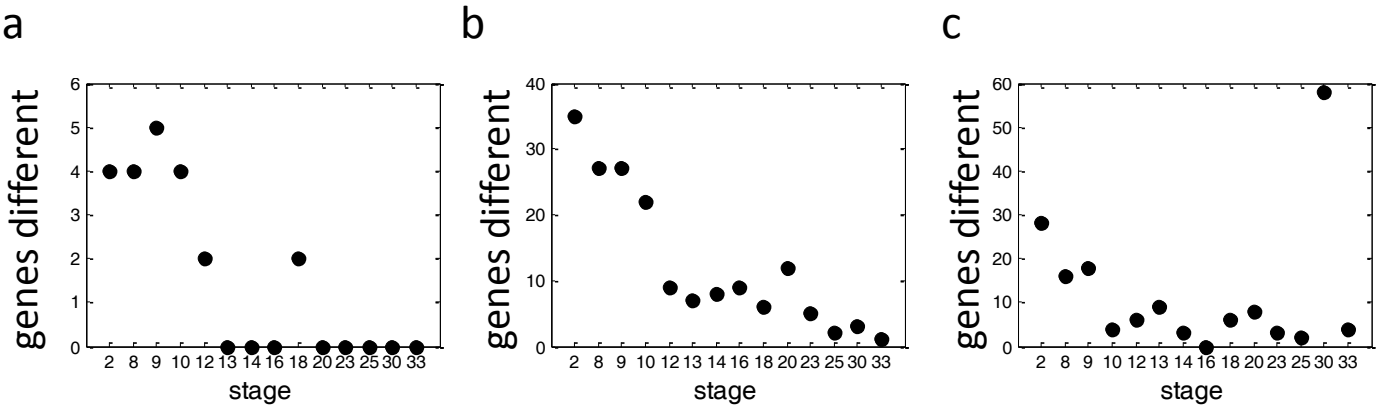

a
b
c
genes different
stage
0
1
2
3
4
5
6
10
20
30
40
50
60
2 8 9 10 12 13 14 16 18 20 23 25 30 33

Figure S5