Supporting Materials for "Comparison of diverse developmental transcriptomes reveals that co-expression of gene neighbors is not evolutionarily conserved"

Itai Yanai and Craig P. Hunter

Figure S1: **Estimation of gene expression profile error rates**. In order to gauge the accuracy of the gene expression profile dataset, we compared the ability of multiple probes for the same gene to yield identical profiles. For this analysis we composed a set of 3,232 *C. elegans* genes for which we had three probes for each gene. The probes are designed as follows: 1) A-probe: the best according to our probe scoring scheme (see Methods); 2) B-probe: located in the 150bp following the stop codon and potentially transcribed as 3'UTR. These were included as they typically have a better signal due to the reverse transcription step of the amplification protocol which enriches for transcripts towards the 3' end (see Methods); and 3) C-probe: the second best probe according to our probe scoring scheme. We set a threshold of an ANOVA P -value$\leq$0.05 for the expression profile for this probe in the following analysis.

For most genes the custom microarray included both an A and B -probe. Thus we examined two questions:

1. How likely is the A-probe profile to be true, given that the B-probe gives a strong signal ($P\leq0.05$) and both the A-probe and B-probe profiles are correlated ($R>0.85$)?
2. How likely is the A-probe profile to be true given that the B-probe profile does not have a strong signal ($P>0.05$)?

We found that for each of these the accuracy is dependent upon the ANOVA P -value (Fig S1). In the figure, genes are grouped according to the two scenarios above . For each group the genes are binned according to P-values of the A-probe and the table below indicates the size of each bin. For each bin we then quantified the fraction of genes in which the A-probe is correlated with the C-probe ($R>0.85$). We note that this

analysis assumes that the C-probe is truth, ignoring the possible error in this measurement and thereby inflating the error rate.

This analysis shows that for *C. elegans* genes with a supporting B-probe, a *P* -value of 0.01 (FDR corrected to 0.0055) is sufficient to ensure 95% accuracy for this scenario. For genes with no signal for the B-probe, the *P*-value must be set to 0.001 (FDR corrected to 0.0004). We note that since the distribution is cumulative the actual accuracy is likely to be much higher (See Figs. S2 and S3) and that we have biased against ourselves by assuming C-probe truth and the error-rate of highest *P*-value category. We set our threshold to 95% accuracy, however the same general results are observed at higher thresholds. The same analysis is repeated in the figure for *C. briggsae*.

Figure S2: **Control for expression data using multiple probes.** We included in our microarrays a set of 795 *C. briggsae* genes for which, in addition to the A-probe (the best one according to our probe scoring scheme, see Methods), four additional probes are present (see Methods). For those additional probes that passed our thresholds (see Fig. S1) we calculated the agreement with the A-probe. Each box corresponds to one of the 360 genes in which both the predefined best probe and at least one other associated probe passed our expression profile thresholds. Profiles in blue correspond to the A-probe. Black and red profiles indicate the profiles of the additional probes where the latter indicate a correlation of <0.85 with the A-probe. Of the 360 genes, 5 (1.4%) have an A-probe that is not consistent with the additional probes, indicated by a red box.

Figure S3: **Comparison of *C. elegans* Agilent-chip dataset with a previously published *C. elegans* Affymetrix-chip dataset(Davis et al. 2005).** The expression ratio of two time-points present in both sets, the 4-cell stage and the 190-cell stage, was computed in the Baugh et al. dataset using the Affymetrix platform(Davis et al. 2005) and in the present *C. elegans* dataset using the Agilent platform. We examined those genes included in our analysis (See Fig S1) and that passed a *P*-value of 0.01 (t-test) on the Affymetrix data and for which the maximum intensity is at least 25, twice the estimated level of noise(Davis et al. 2005). Of the 442 genes, 8 (1.8%, indicated in red) have ratios with a different sign.

Figure S4: **Whole-transcriptome correlations.** The overall correlations among the transcriptomes of the 5 stages match well between the two species. In both species, the 4-cell stage is unique and the adjacent remaining stages are similar. Similarity is computed as the Spearman correlation coefficient of the gene expression profiles subtracted from unity.

Figure S5: **Comparative transcriptomics of nematode embryonic development.** This figure is analogous to Figure 2 starting with the K-means clustering of the *C. briggsae* genes. **A)** For 3,658 *C. briggsae* genes the temporal gene expression profiles are clustered to six general patterns by K-means clustering. **B)** Expression profiles of *C. elegans* genes orthologous to the *C. briggsae* genes shown in **A**. **C)** *C. elegans* orthologs are re-ordered within the *C. briggsae* defined clustered.

Figure S6: **Distributions of strain expression variation by gene class.** Black circles indicate the fraction of genes in each gene class that were detected as having

expression level variation between the Hawaiian (CB4856) and Bristol (N2) *C. elegans* strains at the 4-cell stage. The red line indicates the expected fraction of expression variation based on that of the set of one-to-one orthologs.

Figure S7: **Divergence of gene expression in co-expressed gene neighbors, excluding genes in operons**. Same as Figure 4A but excluding *C. elegans* genes involved in operons, as defined in Wormbase 195.

Figure S8: **Divergence of gene expression in co-expressed gene neighbors, restricting by genomic distance.** Same as Figure 4A but dividing the pairs according to the genomic distance separating them to three categories: <1kb; ≥1kbp,<5kb; and ≥5kb.

Figure S9: **Comparison of correlation with neighbors.** For co-expressed *C. elegans* neighbors (separated in the genome by up to 4 four genes) where the *C. briggsae* orthologs are not neighbors (at least 10 genes away) we examined the *C. briggsae* ortholog that is in the conserved neighborhood. We examined the correlation with the upstream and downstream neighbor and recorded the neighbor with highest maximum intensity. Shown are the distributions of these neighbor correlations compared with a random set of genes with randomized neighbors.

Figure S10: **Gene pair analysis in human mouse tissue data.** Human and mouse data from the Novartis gnf dataset (Su et al. 2004) was used to define a set of 26 analogous tissues. Ensembl defined one-to-one orthologs were invoked and co-expressed gene neighbors were defined as in nematodes; though with an expression

threshold of R>0.4. **A** shows a co-expressed gene pair in human with a conserved expression in mouse. The horizontal lines indicate the human and mouse chromosomes and the vertical bars indicate neighboring genes and one-to-one orthologous relationships. The heatmap images to the right show the tissue expression profiles for the one-to-one orthologs. The x-axis corresponds to 26 common tissues of expression in the gnf dataset. The liver expression (LVR) observed in the human gene pair is conserved in the mouse gene neighbors. **B** shows a pair of human co-expressed genes whose expression is not conserved in mouse. Skeletal muscle and tongue expression is observed in both human genes but of the two mouse genes – not neighbors – only one shows muscle and tongue expression. **C** summarizes the entire observed data in a format analogous to that presented in Figure 4. Starting with co-expressed human genes the mouse orthologs are split into two sets: 1) also neighbors, or 2) not neighbors. Those gene pairs with divergent neighbors are significantly less correlated (P-value<10^{-4}) than pairs with conserved neighborhoods.

Figure S11. **Promoter similarity of orthologs.** Promoter sequence similarity between *C. elegans* and *C. briggsae* orthologs was defined based on the presence or absence of motifs in the upstream sequences. For this, we collected the 500bp upstream sequence for those genes with sufficient intergenic sequence. Based upon this sequence a "motif composition profile" was generated quantifying the occurrences of each of the 600 highest scoring motifs previously identified by a non-alignment based motif detection method (Elemento and Tavazoie 2005). We then computed promoter similarity as the correlation coefficient between the two profiles for each member of the 2,099 one-to-one orthologs. Shown are the distributions of promoter similarities for the gene sets described in Figure 5. Non-essential orthologs with conserved neighborhoods have

more conserved promoters than non-essential genes with non-conserved neighborhoods (*P*-value<10^{-9}). The same is observed for essential orthologs (*P*-value<10^{-3}). Also shown is the distribution of promoter similarities between gene neighbors. While the neighboring genes tend to be similar in expression (Fig. 4) this similarity is not strongly correlated with motifs.

Table S4. **Fraction of essential and non-essential gene pairs among the pairs of co-expressed gene neighbors**

Gene pair	Frequency in *C. elegans*	Frequency in *C. briggsae*
Essential genes	530	422
Non-essential genes	2307	1613
All genes	2837	2035

Table S5. **Genomic arrangement of gene pairs among the pairs of co-expressed gene neighbors**

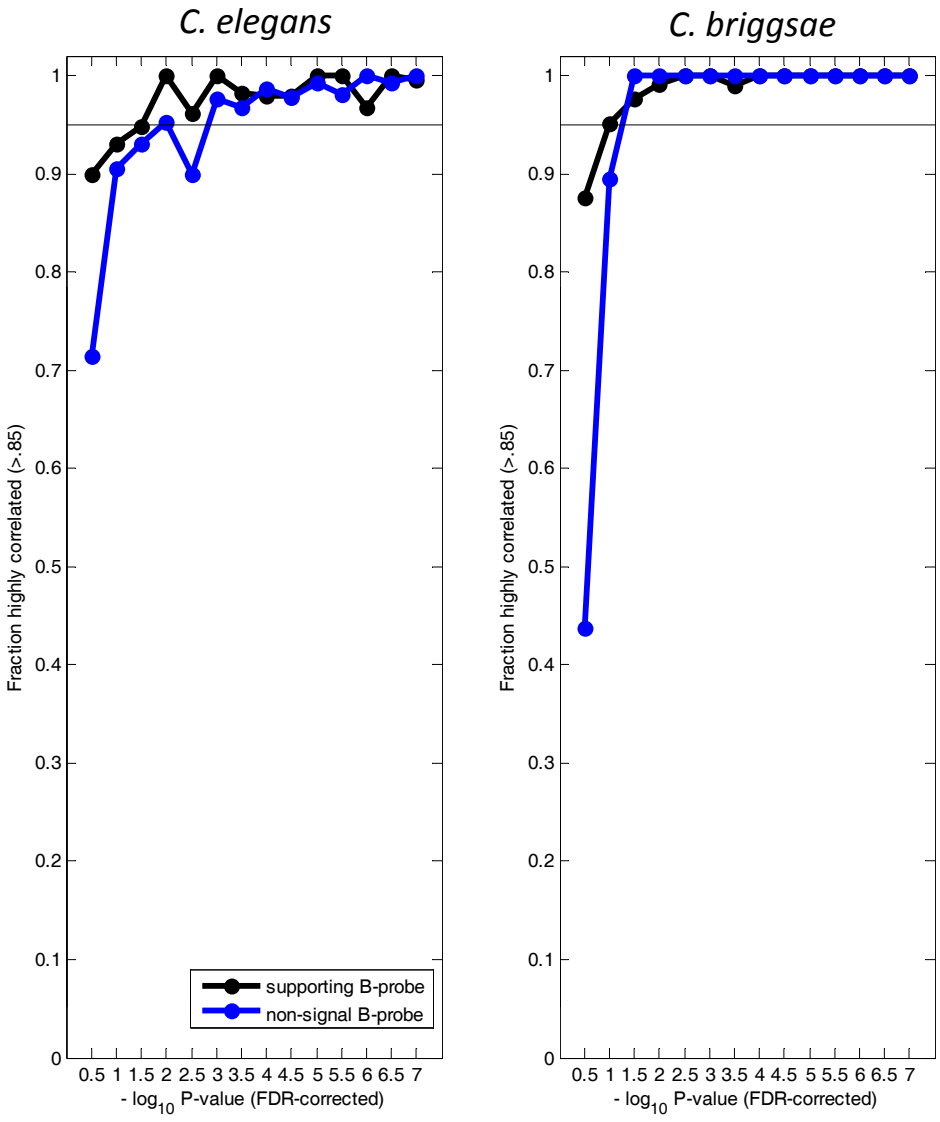
Gene pair	Frequency in *C. elegans*	Frequency in *C. briggsae*
→ →	1488	929
→ ←	733	375
← →	616	427

Davis, J.C., O. Brandman, and D.A. Petrov. 2005. Protein evolution in the context of Drosophila development. *J Mol Evol* **60:** 774-785.

Elemento, O. and S. Tavazoie. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol* **6:** R18.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Figure S1



Sample sizes

	-log_10 P-value (FDR corrected)														p-value for 95%
	0.5	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**5.5**	**6**	**6.5**	**7**	
Ce - supporting B-probe	10	29	39	72	78	69	112	96	94	91	104	93	92	252	0.0055
Ce - no B-probe	35	63	87	106	110	127	124	152	133	135	155	118	128	326	0.0004
Cb - supporting B-probe	32	82	123	114	126	101	91	69	56	29	34	23	14	32	0.04277
Cb - no B-probe	16	19	22	25	19	27	21	14	9	11	5	1	2	2	0.0105

Figure S2

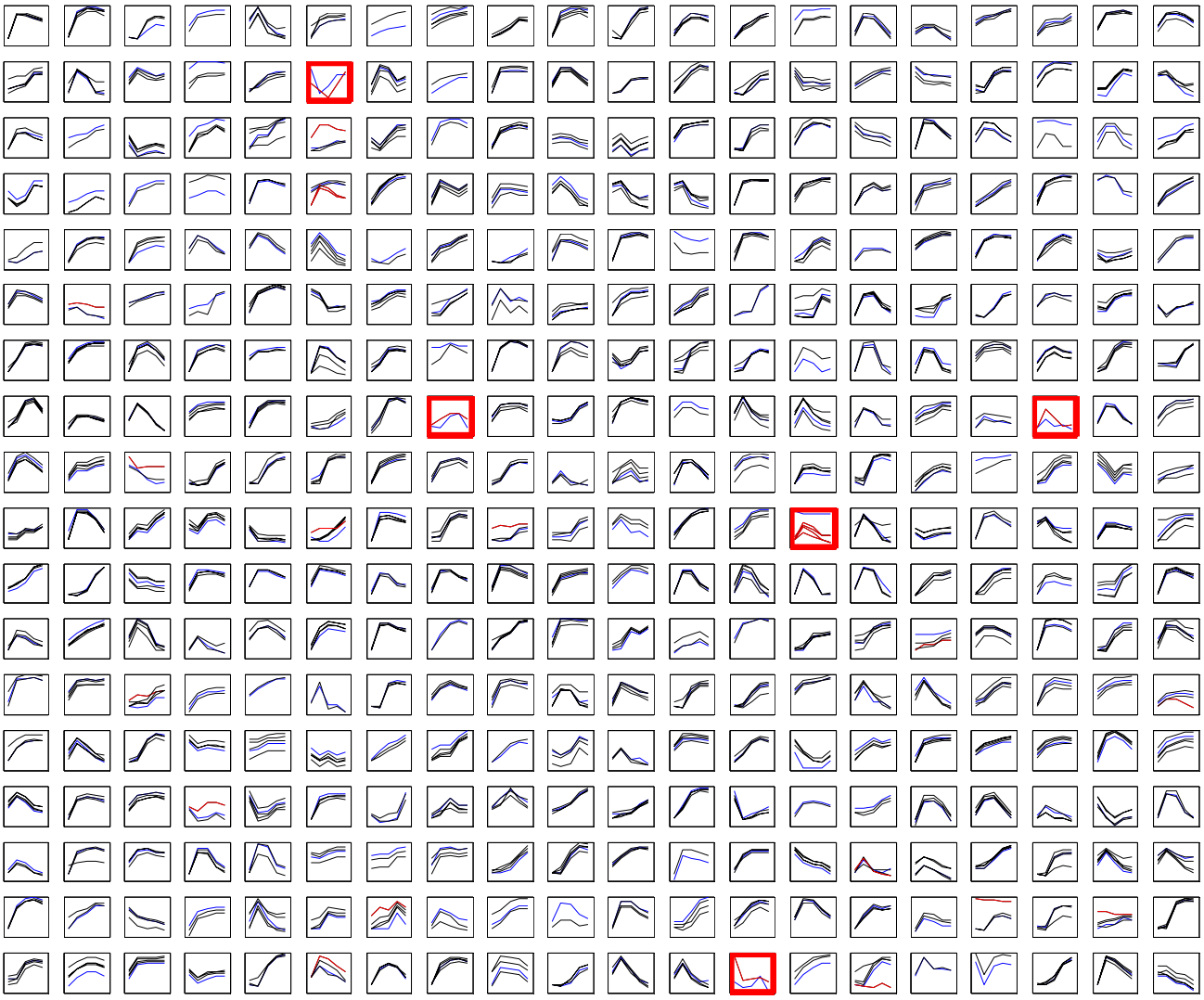


Figure S3

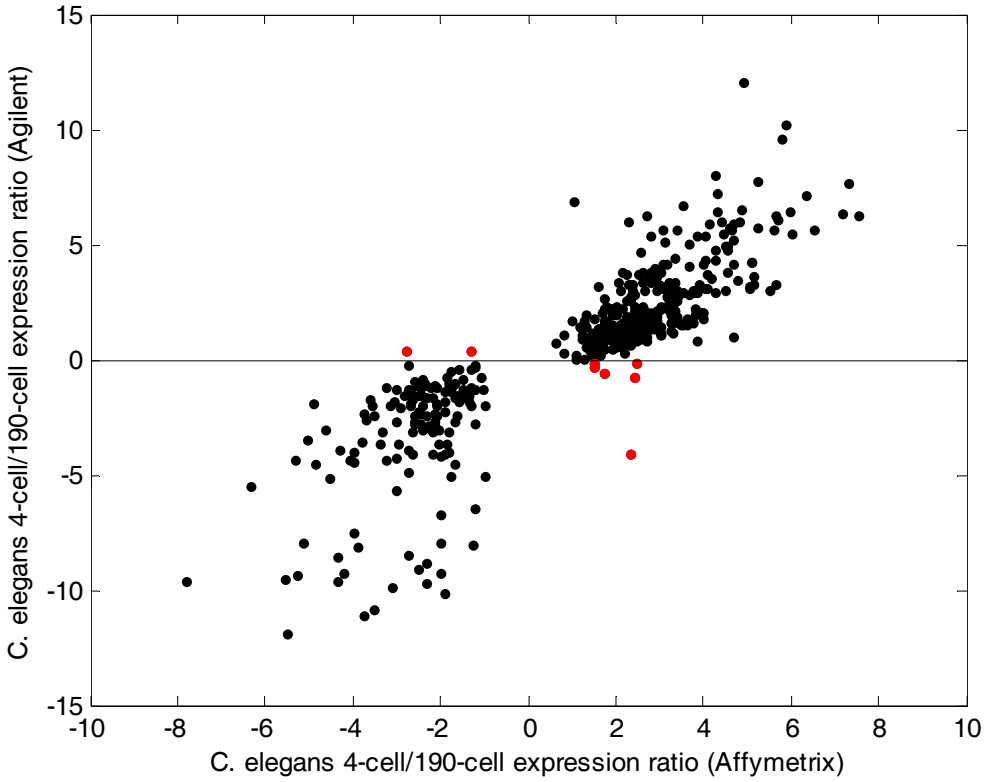


Figure S4

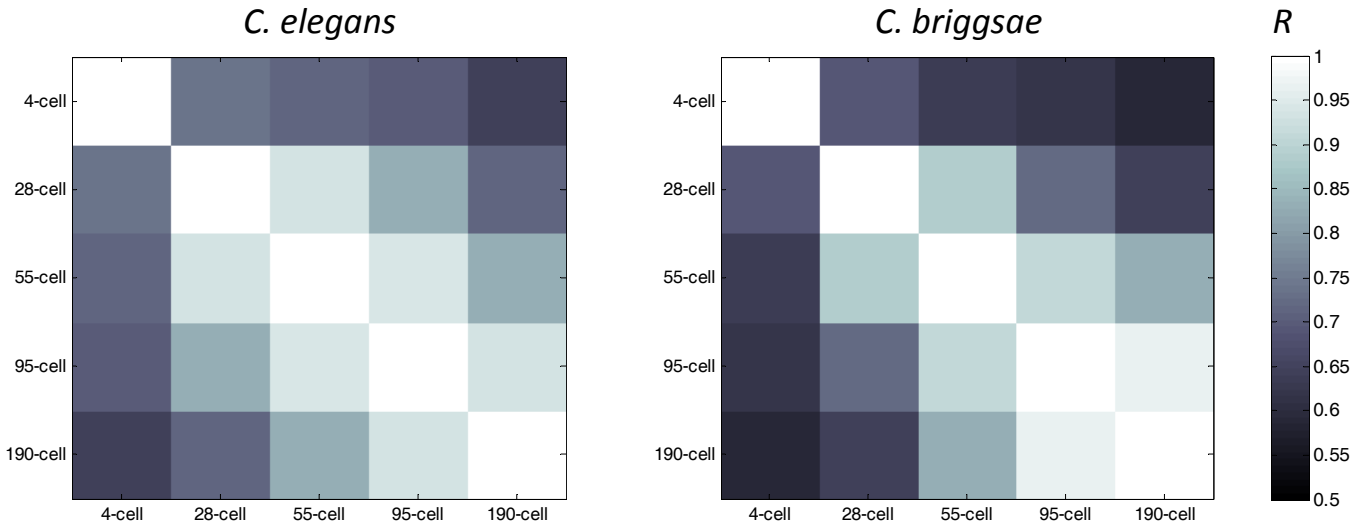


Figure S5

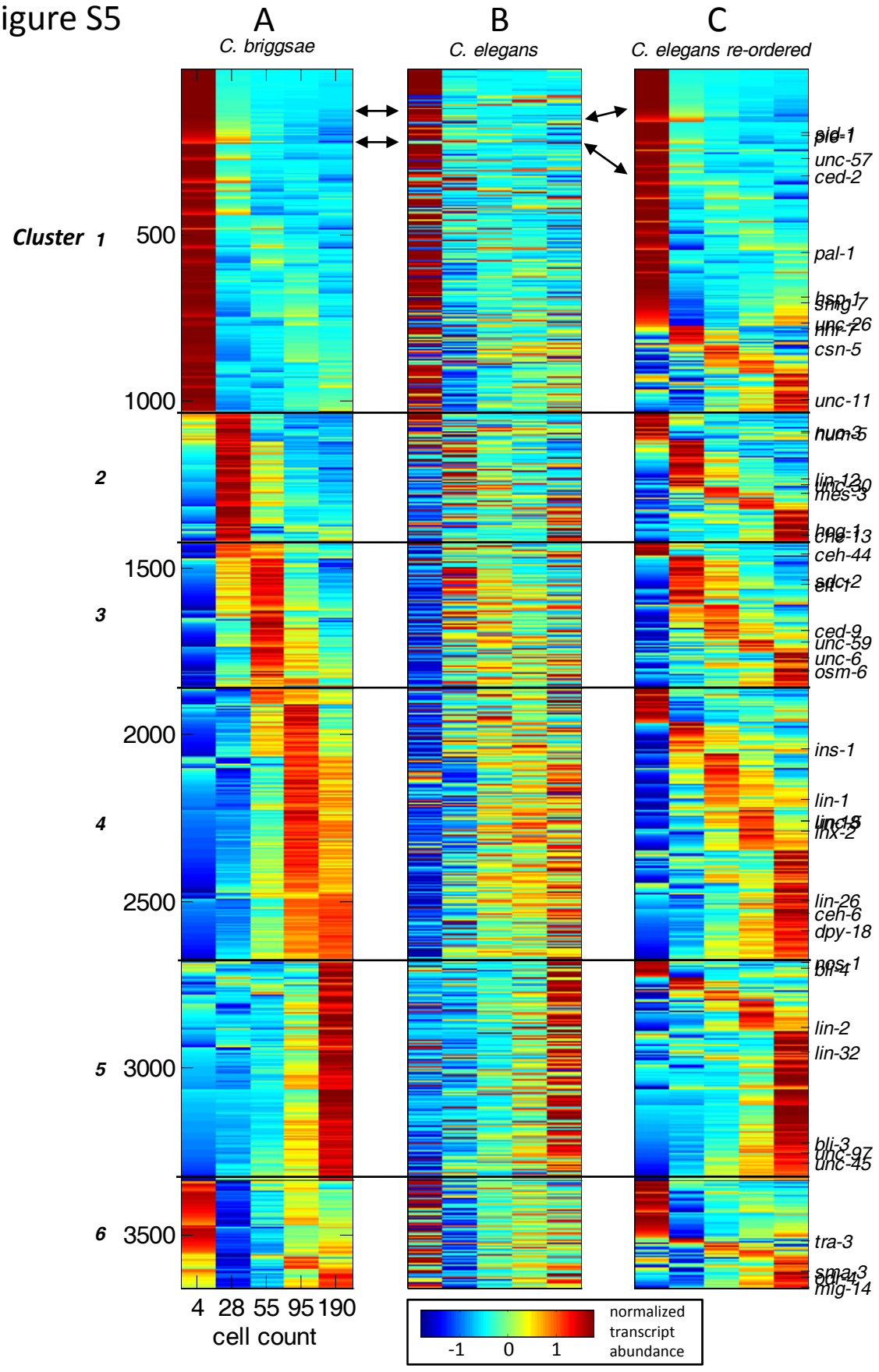

A
C. briggsae
B
C. elegans
C
C. elegans re-ordered
Cluster 1
2
3
4
5
6
500
1000
1500
2000
2500
3000
3500
pid-1
ptc-1
unc-57
ced-2
pal-1
hsp-1
smg-7
unc-26
hnf-7
csn-5
unc-11
hum-5
lin-12
unc-30
mes-3
hpo-1
cnc-13
ceh-44
sdc-2
elt-1
ced-9
unc-59
unc-6
osm-6
ins-1
lin-1
unc-18
inx-2
lin-26
ceh-6
dpy-18
nos-1
bli-4
lin-2
lin-32
bli-3
unc-97
unc-45
tra-3
spna-3
odr-4
mig-14
4 28 55 95 190
cell count
normalized
transcript
abundance
-1 0 1

Figure S6

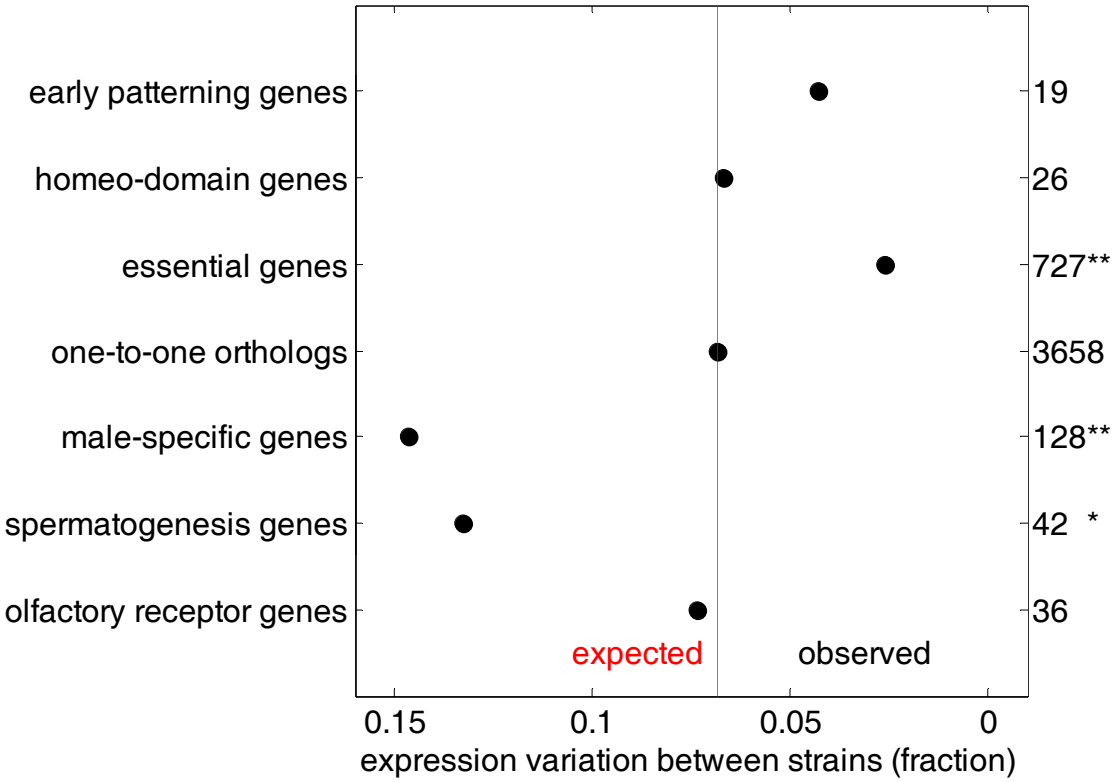


Figure S7

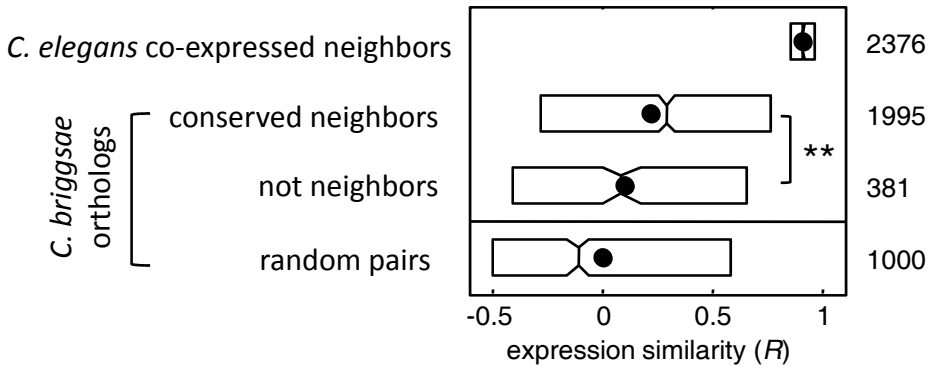


Figure S8

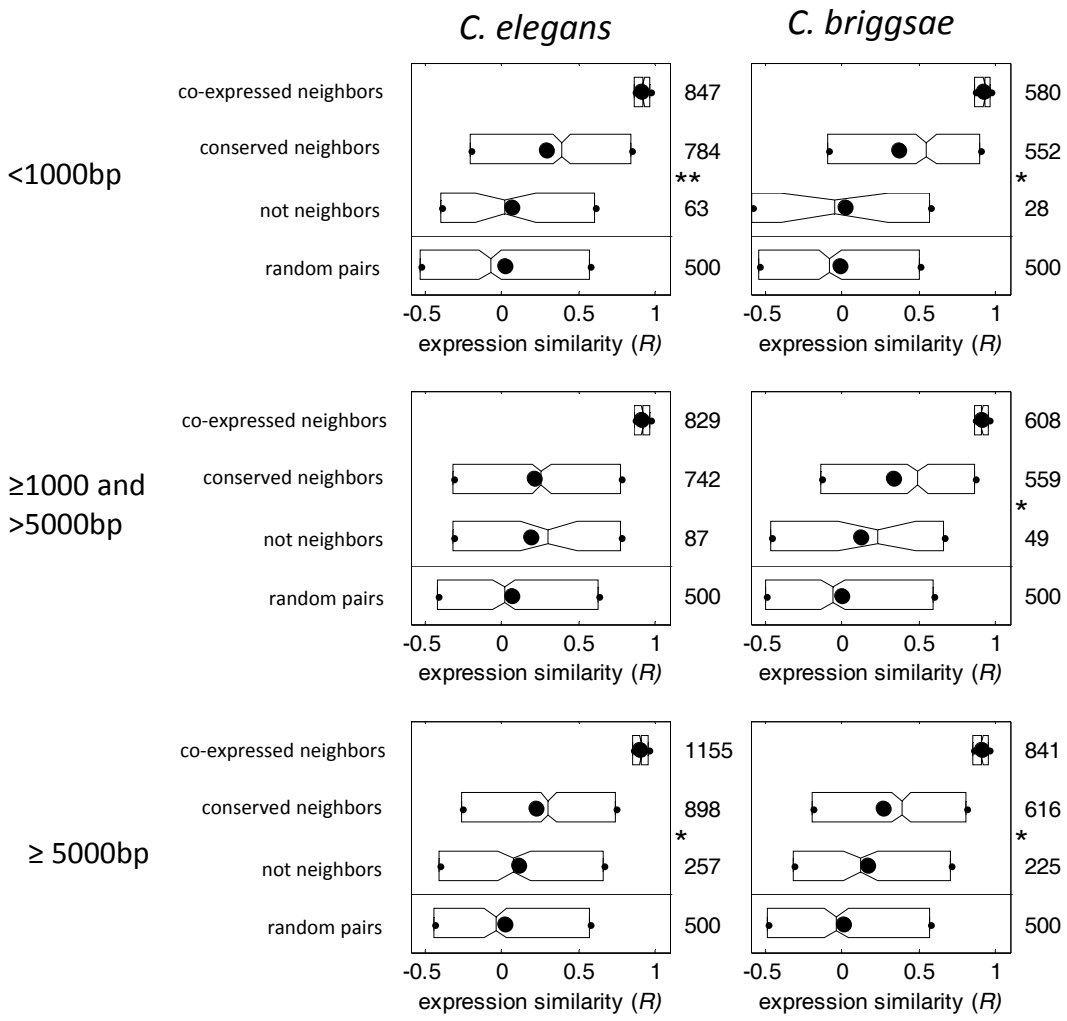


Figure S9

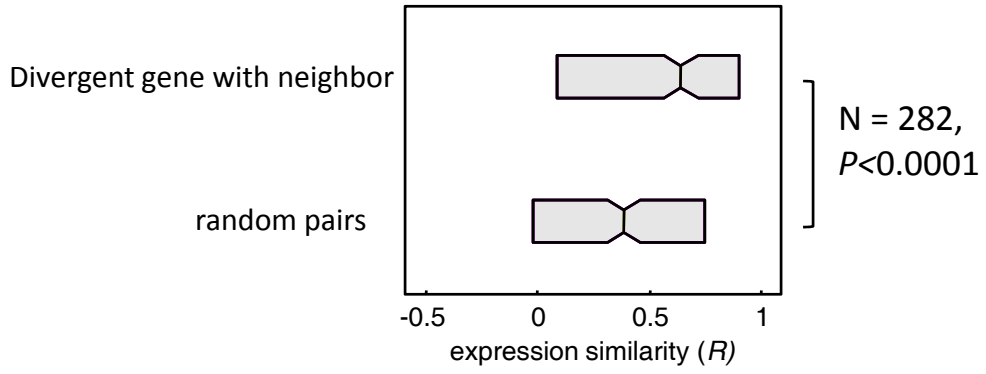
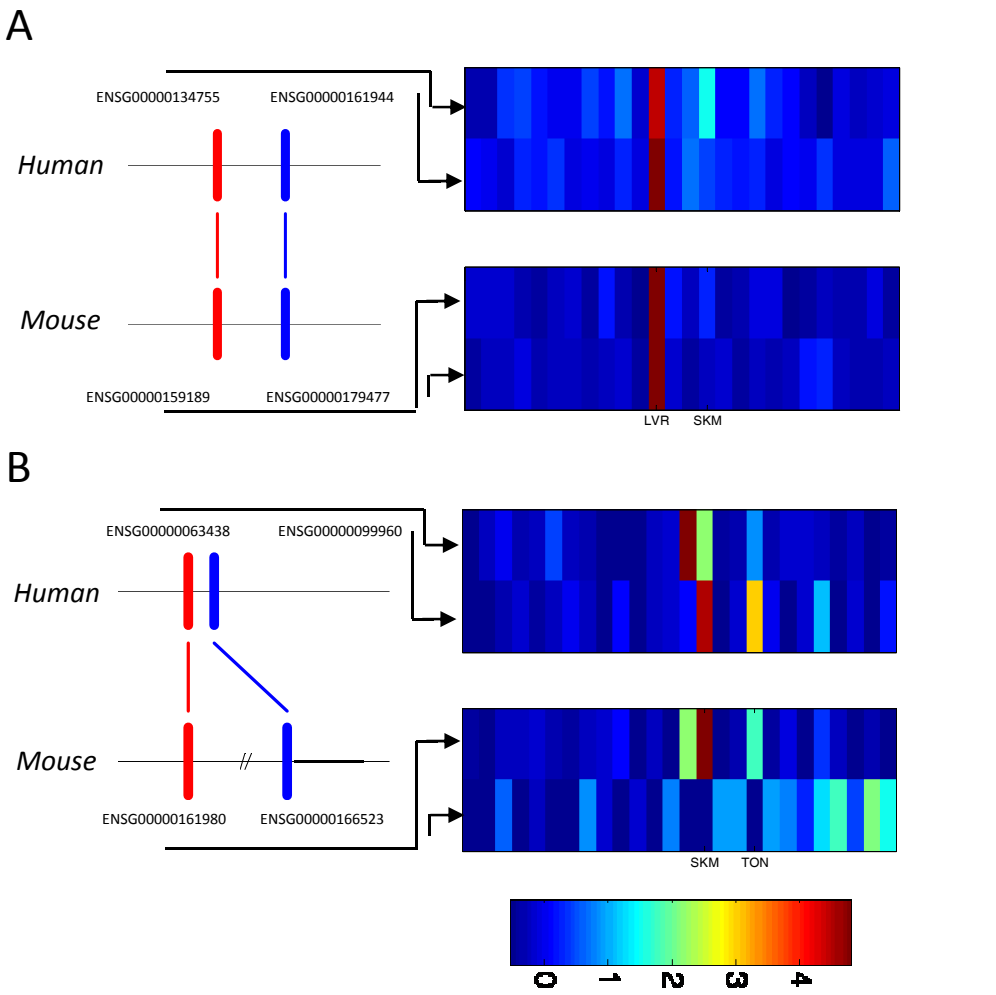


Figure S10


A
ENSG00000134755
ENSG00000161944
Human
Mouse
ENSG00000159189
ENSG00000179477
LVR
SKM
B
ENSG00000063438
ENSG00000099960
Human
Mouse
ENSG00000161980
ENSG00000166523
SKM
TON
0
1
2
3
4

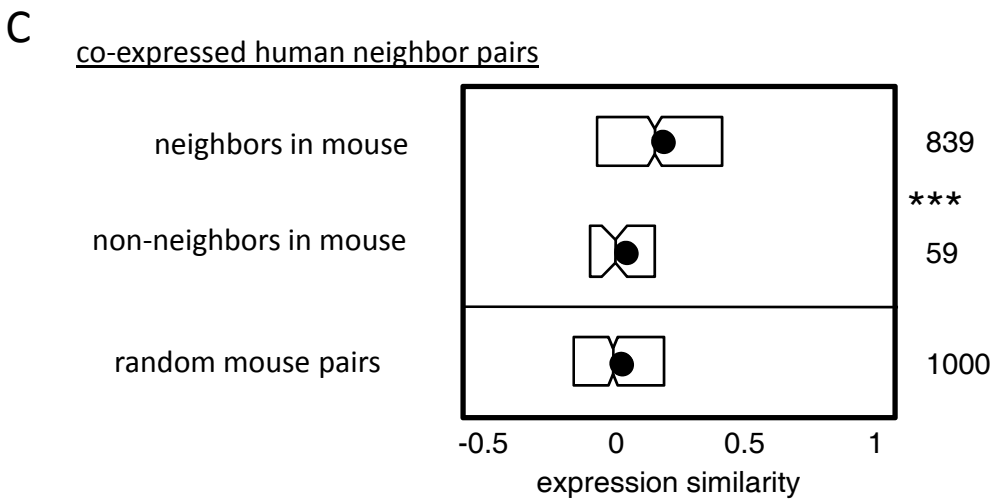
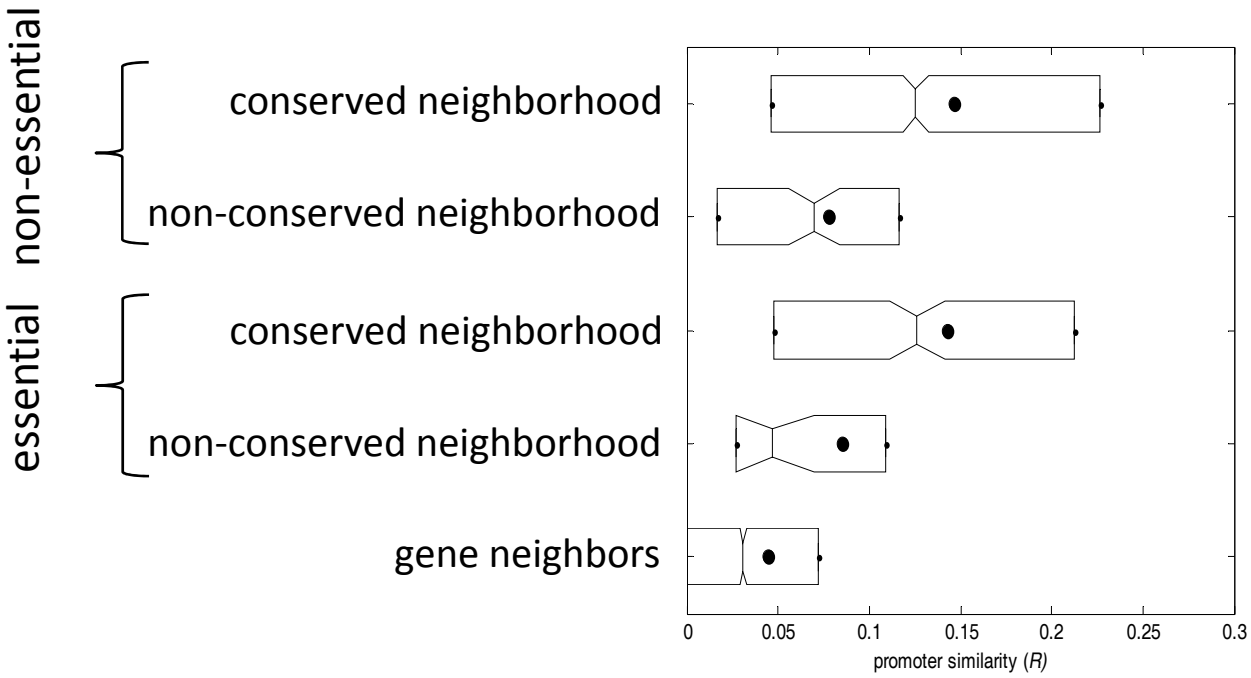

C
co-expressed human neighbor pairs
neighbors in mouse
non-neighbors in mouse
random mouse pairs
839
***
59
1000
-0.5
0
0.5
1
expression similarity

Figure S11


non-essential
essential
conserved neighborhood
non-conserved neighborhood
conserved neighborhood
non-conserved neighborhood
gene neighbors
0
0.05
0.1
0.15
0.2
0.25
0.3
promoter similarity (R)